SIAM. J. MATH. ANAL.
Vol. 23, No. 1, pp. 1-19, January 1992

001

WEAK CONVERGENCE OF INTEGRANDS AND THE YOUNG MEASURE REPRESENTATION*

DAVID KINDERLEHRER† AND PABLO PEDREGAL‡

Abstract. Validity of the Young measure representation is useful in the study of microstructure of ordered solids. Such a Young measure, generated by a minimizing sequence of gradients converging weakly in L^p, often needs to be evaluated on functions of the pth power polynomial growth. A sufficient condition for this evaluation is given in terms of the variational principle. The principal result, Theorem 2.1, concerns lower semicontinuity of functionals integrated over arbitrary sets. The question arose in the numerical analysis of equilibrium configurations of crystals with rapidly varying microstructure, whose specific application is treated elsewhere. Several applications are given. Of particular note, Young measure solutions of an evolution problem are found.

Key words. weak convergence, lower semicontinuity, Young measure, calculus of variations

AMS(MOS) subject classifications. 35Q53, 35K15, 73C50

1. Introduction.

For a lower semicontinuous functional of the form[1]

$$\Phi(v)=\int_\Omega \varphi(\nabla v)\,dx, \qquad v\in H^{1,p}(\Omega;\mathbb{R}^m),$$

the convergence property

$$\Phi(u^k)\to\Phi(u) \quad \text{and} \quad u^k\to u \quad \text{in } H^{1,p}(\Omega;\mathbb{R}^m) \text{ weakly}$$

for a particular sequence (u^k) does not by itself inform us of the behavior of the sequence $(\varphi(\nabla u^k))$. Here we show that if φ is nonnegative and has polynomial growth, then $(\varphi(\nabla u^k))$ is weakly convergent in $L^1(\Omega)$ to $\varphi(\nabla u)$. A consequence is that the Young measure generated by (∇u^k) represents the weak limit of a sequence $(\psi(\nabla u^k))$ when ψ is dominated by φ, which we explain below. Our interest in this question arose in the attempt to estimate convergence properties of numerical methods for functionals which are not lower semicontinuous, where φ plays the role of the relaxed density. Validity of the Young measure representation is useful knowledge in the study of the microstructure of ordered solids; cf. Ball and James [5], [6], Chipot and Kinderlehrer [10], Ericksen [18]-[29], Fonseca [31]-[34], James [35], James and Kinderlehrer [36], Kinderlehrer [37], Kinderlehrer and Pedregal [38], Matos [41], and Pedregal [45], [46]. We do not give any explicit applications to the numerical analysis in this paper except to say that our results confirm the validity of the Young measure representation for the limits of the approximations generated by finite element methods when the energy density has appropriate polynomial growth at infinity. We refer to [9], [11]-[14] for discussions of these developments.

The proof of this and related facts is based on a method of Acerbi and Fusco [1] and subsequent application of the Dunford and Pettis criterion for weak convergence

* Received by the editors July 9, 1990; accepted for publication (in revised form) May 8, 1991. This research was supported by the National Science Foundation and the Air Force Office of Scientific Research through DMS 87-18881 and the Army Research Office through DAAL03 88 K 0110 at the University of Minnesota.

† Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

‡ Department de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain.

[1] Research Group on Transitions and Defects in Ordered Materials, Department of Mathematics and Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

in L^1. Weak convergence of a sequence (f^k) in L^1 is sufficient but not necessary to give sense to the Young measure representation. Ball and Zhang [8] use the Chacon biting lemma to study this question under hypotheses weaker than ours.

The proofs of our results are in §§ 1-3. Three applications are given in §§ 4-6. The example of constraint management in § 4 is a generalization of a result of Müller [44]; cf. also Zhang [51]. In § 5 we give a discussion of the Young measure representation when surface energies are present in the system; cf. [39]. Both of these applications use the convergence property above, or (1.3) below, without assuming that the functional is being driven to a minimum. An application to an evolution problem is given in § 6, where it is shown how Young measure solutions may be found. This builds on some recent work of Slemrod [47]. Useful discussions of Young measures are given by Young [50] and Tartar [48], [49], and more recently by Ball [3] and Evans [30]. One consequence of our considerations is that they lead to a notion of Young measures generated by functions whose gradients are in L^p for finite p [45]. We begin with a description of our principal results. Below, $\mathbb{M}$ denotes the $m\times n$ matrices.

THEOREM 1.1. *Let φ be continuous and quasi-convex and satisfy*

$$0\leqq \varphi(A)\leqq C(1+|A|^p), \qquad A\in\mathbb{M}, \tag{1.1}$$

where $1\leqq p\leqq\infty$. Suppose that

$$u^k\to u \quad \textit{in } H^{1,p}(\Omega) \textit{ weakly and} \tag{1.2}$$

$$\int_\Omega \varphi(\nabla u)\,dx=\lim_{k\to\infty}\int_\Omega \varphi(\nabla u)\,dx. \tag{1.3}$$

Then there is a subsequence (not relabeled) of the (u^k) such that

$$\varphi(\nabla u^k)\to\varphi(\nabla u) \quad \textit{in } L^1(\Omega) \textit{ weakly.}$$

THEOREM 1.2. *Let φ be continuous and quasi-convex and satisfy*

$$0\leqq \varphi(A)\leqq C(1+|A|^p), \qquad A\in\mathbb{M},$$

where $1\leqq p\leqq\infty$. If $u^k\to u$ in $H^{1,p}(\Omega)$ weakly, then

$$\int_E \varphi(\nabla u)\,dx\leqq\liminf_{k\to\infty}\int_E \varphi(\nabla u^k)\,dx \tag{1.4}$$

for every (measurable) $E\subset\Omega$.

We wish to discuss Theorem 1.2 briefly, prior to giving the proof. We shall give a direct proof based on the method of Acerbi and Fusco [1]. This will provide both an efficient self-contained proof and will expose the limitations of the method.

The case of Theorem 1.2 with $p=\infty$ is automatic since $\{\varphi(\nabla u^k)\}$ are uniformly bounded in this case. Indeed, choose M with the property

$$\|\varphi(\nabla u^k)\|_{L^\infty(\Omega)}\leqq M \quad \text{for all } k.$$

Given E, let U be an open neighborhood of E with $|U-E|<\varepsilon$. Now U is the union of countably many cubes $\{D_j\}$ with disjoint interiors and for each D_j, (1.4) holds [15], [16], [45]. Hence

$$\begin{aligned}\int_U \varphi(\nabla u)\,dx&\leqq\sum\int_{D_j}\varphi(\nabla u)\,dx\\&\leqq\sum\liminf\int_{D_j}\varphi(\nabla u^k)\,dx\\&\leqq\liminf\int_U\varphi(\nabla u^k)\,dx.\end{aligned}$$

Finally, we have that

$$\int_E \varphi(\nabla u)\,dx \leqq \liminf \int_E \varphi(\nabla u^k)\,dx + 2M\varepsilon.$$

Thus, if $u_k \to u$ in $H^{1,\infty}(\Omega)$ weak*, then

$$\int_E \varphi(\nabla u)\,dx \leqq \liminf \int_E \varphi(\nabla u^k)\,dx \tag{1.5}$$

for any measurable $E \subset \Omega$.

The case $p=1$ for Theorems 1.1 and 1.2 is easy and will not be discussed.

To illustrate how the preceding results apply to the Young measure representation, let us recall this notion. Let $C_0(\mathbb{R}^m)$ denote the continuous functions on $\mathbb{R}^m$ vanishing at ∞. Given a sequence $f^k \in L^1(\Omega; \mathbb{R}^m)$, $k = 1, 2, 3, \cdots$, for any $\psi \in C_0(\mathbb{R}^m)$, the composed sequence $(\psi(f^k))$ admits a subsequence weak* convergent in $L^\infty(\Omega)$; namely, for some $\bar{\psi} \in L^\infty(\Omega)$,

$$\psi(f^{k'}) \to \bar{\psi} \quad \text{in } L^\infty(\Omega) \text{ weak*}.$$

The association of ψ to $\bar{\psi}$ is linear, which, after some argument (cf. [3]) employing the boundedness in L^1 of the sequence (f^k), permits us to assert the existence of a family $\nu = (\nu_x)_{x \in \Omega}$ of probability measures with the property that

$$\bar{\psi}(x) = \int_{\mathbb{R}^m} \psi(\lambda)\,d\nu_x(\lambda) \quad \text{in } \Omega \text{ a.e.}$$

For the validity of the representation it is sufficient that

$$\psi(f^{k'}) \to \bar{\psi} \quad \text{in } L^1(\Omega) \text{ weakly.} \tag{1.6}$$

The family ν is the Young measure or parametrized measure generated by $(f^{k'})$; cf. Young [50].

If $f^k \in L^p(\Omega; \mathbb{R}^m)$, $k = 1, 2, 3, \cdots$, satisfy

$$\|f^k\|_{L^p(\mathbb{R}^n)} \leqq C$$

and generate a Young measure $\nu = (\nu_x)_{x \in \Omega}$, it may be verified that whenever $\psi \in C(\mathbb{R}^m)$ satisfies

$$\lim_{|\lambda| \to \infty} \frac{|\psi(\lambda)|}{1 + |\lambda|^p} = 0,$$

the sequence $\psi(f^k)$ converges weakly in L^1 and the representation (1.6) is valid. When $f^k = \nabla u^k$ arises as a minimizing sequence in a variational principle for φ satisfying (1.7) or W satisfying (1.12) below, it is less than obvious that we are entitled to use the Young measure to evaluate the limit energy or other integrals of (∇u^k) with exactly pth power growth. Our results, Theorem 1.3 and Corollary 1.4, affirm the validity of the Young measure representation:

Introduce the Banach space, for $p \geqq 1$ fixed,

$$E = \left\{ \psi \in C(\mathbb{M}) \colon \sup_{\mathbb{M}} \frac{|\psi(A)|}{|A|^p + 1} < \infty \right\}.$$

THEOREM 1.3. *Let φ be quasi-convex and satisfy, for some constants $C \geqq c > 0$,*

$$\max\{c|A|^p - 1, 0\} \leqq \varphi(A) \leqq C(1 + |A|^p), \qquad A \in \mathbb{M}, \tag{1.7}$$

where $1 \leqq p < \infty$. *Suppose that*

$$u^k \to u \quad in\ H^{1,p}(\Omega)\ weakly\ and \tag{1.8}$$

$$\int_\Omega \varphi(\nabla u)\,dx = \lim_{k\to\infty} \int_\Omega \varphi(\nabla u^k)\,dx. \tag{1.9}$$

Let $\nu = (\nu_x)_{x\in\Omega}$ *be a Young measure generated by* (∇u^k). *Then for any* $\psi \in E$, *the sequence*

$$\psi(\nabla u^k) \to \bar{\psi} \quad in\ \sigma(L^1(\Omega), L^\infty(\Omega)),$$

where

$$\bar{\psi}(x) = \int_{\mathbb{M}} \psi(A)\,d\nu_x(A) \quad in\ \Omega\ a.e. \tag{1.10}$$

Further, consider $W \in C(\mathbb{M})$ satisfying

$$\max\{c|A|^p - 1, 0\} \leqq W(A) \leqq C(1+|A|^p), \qquad A \in \mathbb{M}, \tag{1.11}$$

for some $p \geqq 1$ and $0 < c \leqq C$. Let

$$A_\Omega(y_0) = \{v \in H^{1,p}(\Omega)\colon v = y_0 \text{ on } \partial\Omega\}, \quad \text{where } y_0 \in H^{1,p}(\Omega).$$

COROLLARY 1.4. *Let* W *satisfy* (1.11). *Suppose that* $(u^k) \subset A_\Omega(y_0)$ *satisfies*

$$\lim_{k\to\infty} \int_\Omega W(\nabla u^k)\,dx = \inf_{A_\Omega(y_0)} \int_\Omega W(\nabla v)\,dx, \tag{1.12a}$$

$$u^k \to u \quad in\ H^{1,p}(\Omega)\ weakly. \tag{1.12b}$$

Let $\nu = (\nu_x)_{x\in\Omega}$ *be a Young measure generated by* (u^k). *Then for any* $\psi \in E$, *the sequence*

$$\psi(\nabla u^k) \to \bar{\psi}\ in\ \sigma(L^1(\Omega), L^\infty(\Omega)),$$

where

$$\bar{\psi}(x) = \int_{\mathbb{M}} \psi(A)\,d\nu_x(A) \quad in\ \Omega\ a.e. \tag{1.13}$$

In particular, the $(W(\nabla u^k))$ *converges to a limit energy density* $\bar{W}$ *in* $\sigma(L^1(\Omega), L^\infty(\Omega))$ *where*

$$\bar{W}(x) = \int_{\mathbb{M}} W(A)\,d\nu_x(A) \quad in\ \Omega\ a.e. \tag{1.14}$$

Note above that if $p = 1$, we are not assured of a sequence (u^k) satisfying the hypothesis (1.12b). A version of Corollary 1.4 has also been proved independently by Matos [42] who obtains an improved class E by combining Ekeland's lemma with the reverse Hölder inequality, although the convergence is then restricted to $\sigma(L^1(\Omega'), L^\infty(\Omega'))$ for $\Omega' \subset\subset \Omega$.

Note that a particular consequence of Theorem 1.3 is that the sequence $\{|M \cdot \nabla u_k|^p\}$, for a constant matrix M, converges weakly in $L^1(\Omega)$, although not to $|M \cdot \nabla u|^p$. Another consequence concerns the relaxation of W, or its quasi convexification; cf. [7], [15], [16], for example. Assume that $p > 1$. The integrand

$$W^{\#}(F) = \inf_{H_0^{1,\infty}(\Omega)} \frac{1}{|\Omega|} \int_\Omega W(F + \nabla\zeta)\,dx \tag{1.15}$$

is quasi-convex and relaxes the variational principle (1.8) in the sense that

$$\inf_{A_\Omega(y_0)} \int_\Omega W(\nabla v)\,dx = \inf_{A_\Omega(y_0)} \int_\Omega W^{\#}(\nabla v)\,dx.$$

Obviously a minimizing sequence for (1.12) is also a minimizing sequence for the functional with the integrand $W^{\#}$. For a given F, the infimum in (1.15) may or may not be realized, but given a minimizing sequence $u^k(x) = Fx + \zeta^k(x) \in H^{1,p}(\Omega; \mathbb{R}^m)$,

$$W^{\#}(F)|\Omega| = \lim_{k\to\infty} \int_\Omega W(\nabla u^k)\, dx.$$

Let $\mu = (\mu_x)_{x\in\Omega}$ be a Young measure generated by (u^k). We may assume that μ_x is independent of $x \in \Omega$, although we pass over the details of that here. Applying Corollary 1.4, we obtain in particular that

$$W^{\#}(F) = \int_{\mathbb{M}} W(A)\, d\mu(A), \tag{1.16}$$

so the infimum is attained in a Young measure fashion. Moreover, the inequality $W^{\#} \leqq W$ ensures that

$$\operatorname{supp} \mu \subset \{A\colon W(A) = W^{\#}(A)\}.$$

Of course, if σ is any other Young measure generated by some sequence of the form $(v^k) \subset H^{1,p}(\Omega, \mathbb{R}^m)$ with $v^k = Fx$ on $\partial\Omega$, then

$$\int_{\mathbb{M}} W(A)\, d\mu(A) \leqq \int_{\mathbb{M}} W(A)\, d\sigma(A),$$

so μ satisfies an ersatz minimizing principle as well.

2. Proof of Theorem 1.2. Our aim is to give a proof of the second result. Theorem 1.1 will be a corollary of it. For this we adopt a technique of Acerbi and Fusco, which has an important ingredient from a paper of F.-C. Liu [40]. The technique uses these facts from Acerbi and Fusco.

LEMMA 2.1. *Let $G \subset \mathbb{R}^n$ have $|G| < \infty$. Assume that $\{M_k\}$ is a sequence of subsets of G such that for some $\varepsilon > 0$*

$$|M_k| > \varepsilon \quad \textit{for all } k.$$

Then there is a subsequence k_j for which

$$\bigcap M_{k_j} \neq \varnothing.$$

LEMMA 2.2. *Let $\{f_k\}$ be a sequence bounded in $L^1(\Omega)$. Then for each $\varepsilon > 0$, there is a triple $(A_\varepsilon, \delta, S)$, where $A_\varepsilon \subset \Omega$ with $|A_\varepsilon| < \varepsilon$, $\delta > 0$, and S is an infinite subset of the natural numbers, such that*

$$\int_D |f_k|\, dx < \varepsilon$$

whenever $D \cap A_\varepsilon = \varnothing$ and $|D| < \delta$ for all $k \in S$.

For any $v \in C_0^\infty(\mathbb{R}^n)$, we set

$$M^*v(x) = M(|v(x)|) + M(|\nabla v(x)|),$$

where

$$Mf(x) = \sup_{r>0} \frac{1}{|B_r|} \int_{B_r(x)} |f(z)|\, dz$$

is the maximal function of f. It is well known that if $v \in C_0^\infty(\mathbb{R}^n)$, then $M^*v \in C(\mathbb{R}^n)$ and

$$\|M^*v\|_{L^p(\mathbb{R}^n)} \leqq C(n, p)\|v\|_{H^{1,p}(\mathbb{R}^n)}, \qquad 1 < p \leqq \infty, \tag{2.1}$$

and, in particular, for any $\lambda > 0$,

$$|\{M^*v \geqq \lambda\}| \leqq C(n,p)\lambda^{-p}\|v\|^p_{H^{1,p}(\mathbb{R}^n)}, \qquad 1<p<\infty. \tag{2.2}$$

LEMMA 2.3. *Let* $v \in C_0^\infty(\mathbb{R}^n)$ *and* $\lambda > 0$. *Set* $H^\lambda = \{M^*v < \lambda\}$. *Then*

$$\frac{|v(x)-v(y)|}{|x-y|} \leqq C(n)\lambda, \qquad x, y \in H^\lambda, \tag{2.3}$$

where $C(n)$ *depends only on* n.

We shall also make use of the well-known fact that a Lipschitz function defined on a subset of $\mathbb{R}^n$ may be extended to all of $\mathbb{R}^n$ without increasing its Lipschitz constant.

Proof of Theorem 1.2. We regard u^k and u as extended to functions in $H^{1,p}(\mathbb{R}^n)$ with norms controlled by their $H^{1,p}(\Omega)$ norms. Let $\varepsilon > 0$.

Step 1. Since the functional of (1.4) is continuous in $H^{1,p}(\mathbb{R}^n)$ in the norm topology, because of the upper bound on φ, we may find $z, z^k \in C_0^\infty(\mathbb{R}^n)$ with

$$\int_{\mathbb{R}^n} |\varphi(\nabla u) - \varphi(\nabla z)|\, dx < \varepsilon, \tag{2.4}$$

$$\int_{\mathbb{R}^n} |\varphi(\nabla u^k) - \varphi(\nabla z + \nabla z^k)|\, dx < \varepsilon, \tag{2.5}$$

and

$$\|u - u^k - z^k\|_{H^{1,p}(\mathbb{R}^n)} < \frac{1}{k}.$$

Thus $z^k \to 0$ in $H^{1,p}(\mathbb{R}^n)$ weakly and

$$\|z^k\|_{H^{1,p}(\mathbb{R}^n)} \leqq M < \infty. \tag{2.6}$$

Set

$$H^\lambda = \{M^*z < \lambda\} \quad \text{and} \quad H_k^\lambda = \{M^*z^k < \lambda\}.$$

According to Lemma 2.3, we may find ζ^k, $\eta \in H^{1,\infty}(\mathbb{R}^n)$ such that $\zeta^k = z^k$ on H_k^λ and $\eta = z$ on H^λ with

$$\|\nabla \zeta^k\|_{L^\infty(\mathbb{R}^n)} \leqq \|\nabla z^k\|_{L^\infty(H_k^\lambda)} \leqq \lambda$$

and

$$\|\zeta^k\|_{H^{1,\infty}(\mathbb{R}^n)} \leqq C(n)\lambda,$$

and the same for η. After extraction of a subsequence we may suppose that

$$\zeta^k \to \zeta \quad \text{in } H^{1,\infty}(\mathbb{R}^n) \text{ weak*}.$$

We apply Lemma 2.2 to the sequence $\{M^*(z^k)^p\}$. By (1.2) and (2.1) these functions are bounded in $L^1(\Omega)$. So, given $\varepsilon' > 0$, there is a triple $(A_{\varepsilon'}, \delta, S)$ with $|A_{\varepsilon'}| < \varepsilon'$ and

$$\int_D M^*(z^k)^p\, dx < \varepsilon'$$

whenever $D \cap A_{\varepsilon'} = \varnothing$ with $|D| < \delta$ and $k \in S$.

Now let $G = \{\zeta \neq 0\}$. Since the $z^k \to 0$ in $L^p(\mathbb{R}^n)$ in norm, we may assume that $z^k \to 0$ pointwise almost everywhere in Ω. Thus if we set $G_0 = G \backslash \{x \in \Omega\colon z^k(x) \to 0\}$, then $|G_0| = |G|$. We write G_0 as a union,

$$G_0 = (G_0 \cap H_k^\lambda) \cup (G_0 \cap (\mathbb{R}^n - H_k^\lambda)).$$

By (2.2),

$$|G_0 \cap (\mathbb{R}^n - H_k^\lambda)| \leqq C\lambda^{-p}M^p \quad \text{for all } k. \tag{2.7}$$

This implies that

$$|G_0| = |G| \leqq 2C\lambda^{-p}M^p. \tag{2.8}$$

Otherwise, if

$$|G_0| > 2C\lambda^{-p}M^p,$$

then

$$|G_0 \cap H_k^\lambda| > C\lambda^{-p}M^p,$$

by (2.7). Applying Lemma 2.1, there would be a subsequence k_j such that

$$G_0 \cap \left(\bigcap H_{k_j}^\lambda\right) \neq \varnothing,$$

and for x in this intersection,

$$\zeta(x) = \lim \zeta^k(x) = \lim z^k(x) = 0,$$

which contradicts the definition of the set G. Hence (2.8) holds.

Step 2. Since $\varphi(\nabla u) \in L^1(\Omega)$, we may find $\sigma, 0 < \sigma < \varepsilon$, and λ large enough that

$$\int_{A_\sigma \cup (\Omega - H^\lambda) \cup G} \varphi(\nabla u)\, dx < \varepsilon; \tag{2.9}$$

cf. (2.8) above. Let $E \subset \Omega$ be measurable and assume a subsequence of the u^k chosen (but not relabeled) so that

$$\lim \int_E \varphi(\nabla u^k)\, dx = \liminf \int_E \varphi(\nabla u^k)\, dx.$$

Put

$$\alpha_k = \int_E \varphi(\nabla u^k)\, dx.$$

Since $\varphi \geqq 0$, by (2.5)

$$\begin{aligned} \alpha_k &\geqq \int_{E \cap H^\lambda \cap H^{\lambda,k} \cap (\Omega - A_\sigma)} \varphi(\nabla u^k)\, dx \\ &\geqq -\varepsilon + \int_{E \cap H^\lambda \cap H^{\lambda,k} \cap (\Omega - A_\sigma)} \varphi(\nabla z + \nabla z^k)\, dx. \end{aligned}$$

But $\nabla z = \nabla \eta$ and $\nabla z^k = \nabla \zeta^k$ in $H^\lambda \cap H^{\lambda,k}$ so that

$$\begin{aligned} \alpha_k &\geqq -\varepsilon + \int_{E \cap H^\lambda \cap H^{\lambda,k} \cap (\Omega - A_\sigma)} \varphi(\nabla \eta + \nabla \zeta^k)\, dx \\ &= -\varepsilon + \int_{E \cap H^\lambda \cap (\Omega - A_\sigma)} \varphi(\nabla \eta + \nabla \zeta^k)\, dx \\ &\quad - \int_{E \cap H^\lambda \cap (\Omega - H^{\lambda,k}) \cap (\Omega - A_\sigma)} \varphi(\nabla \eta + \nabla \zeta^k)\, dx \\ &= -\varepsilon + \beta_k - \gamma_k. \end{aligned}$$

Since $\nabla(\eta+\zeta^k)$ is uniformly bounded and φ is quasi-convex, by the remark (1.5) we have that for K sufficiently large

$$\beta_k+\varepsilon \geqq \int_{E\cap H^\lambda\cap(\Omega-A_\sigma)} \varphi(\nabla\eta+\nabla\zeta)\,dx.$$

We now inspect γ_k. Using the bounds on $\nabla\eta$ and $\nabla\zeta^k$, and choosing λ large enough,

$$\begin{aligned}\gamma_k &\leqq C(1+\lambda^p)|(\Omega-H^\lambda_k)\cap(\Omega-A_\sigma)|\\ &\leqq C|\Omega-H^\lambda_k|+\int_{(\Omega-H^{\lambda,k})\cap(\Omega-A_\sigma)} CM^*(z^k)^p\,dx\\ &\leqq C\varepsilon+C\sigma\leqq 2C\varepsilon.\end{aligned}$$

Consequently, for k sufficiently large,

$$\alpha_k\geqq -C\varepsilon+\int_{E\cap H^\lambda\cap(\Omega-A_\sigma)} \varphi(\nabla\eta+\nabla\zeta)\,dx. \tag{2.10}$$

Step 3. Again using the positivity of φ, from (2.10),

$$\alpha_k\geqq -C\varepsilon+\int_{E\cap H^\lambda\cap(\Omega-A_\sigma)\cap(\Omega-G)} \varphi(\nabla\eta+\nabla\zeta)\,dx.$$

Since $\zeta=0$ in $\Omega-G$, we have that $\nabla\zeta=0$ in $\Omega-G$, so, since $\eta=z$ in H^λ, we deduce that

$$\begin{aligned}\alpha_k &\geqq -C\varepsilon+\int_{E\cap H^\lambda\cap(\Omega-A_\sigma)\cap(\Omega-G)} \varphi(\nabla\eta)\,dx\\ &\geqq -C\varepsilon+\int_{E\cap H^\lambda\cap(\Omega-A_\sigma)\cap(\Omega-G)} \varphi(\nabla z)\,dx.\end{aligned}$$

By (2.4) and (2.9),

$$\begin{aligned}\alpha_k &\geqq -(1+C)\varepsilon+\int_{E\cap H^\lambda\cap(\Omega-A_\sigma)\cap(\Omega-G)} \varphi(\nabla u)\,dx\\ &\geqq -(1+C)\varepsilon+\int_E \varphi(\nabla u)\,dx-\int_{E\cap[A_\sigma\cup(\Omega-H^\lambda)\cup G]} \varphi(\nabla u)\,dx\\ &\geqq -(2+C)\varepsilon+\int_E \varphi(\nabla u)\,dx.\end{aligned}$$

Since $\varepsilon>0$ is arbitrary, the theorem is proved.

3. Proofs of the other results.

Proof of Theorem 1.1. This follows from the Dunford-Pettis criterion. Assume that the sequence $(\varphi(\nabla u^k))$ is not $\sigma(L^1, L^\infty)$ relatively compact. Then for some $\varepsilon>0$ and every $\delta>0$, there is an $A_\delta\subset\Omega$ and an integer k_δ such that $|A_\delta|<\delta$ and

$$\int_{A_\delta} \varphi(\nabla u^{k_\delta})\,dx>\varepsilon.$$

Since $\varphi(\nabla u)\in L^1(\Omega)$, there is a $\delta_0>0$ such that if $|E|<\delta_0$, then

$$\int_E \varphi(\nabla u)\,dx<\varepsilon. \tag{3.1}$$

Let us choose in particular $\delta_j = 2^{-j}\delta_0$. Then there is a sequence A_j, $|A_j| < \delta_j$, and k_j such that

$$\int_{A_j} \varphi(\nabla u^{k_j})\, dx > \varepsilon \quad \text{for all } j.$$

Let $E = \bigcup A_j$, so $|E| \leqq \delta_0$ and (3.1) holds. Thus

$$\varepsilon \leqq \int_E \varphi(\nabla u^{k_j})\, dx \leqq \int_\Omega \varphi(\nabla u^{k_j})\, dx - \int_{\Omega - E} \varphi(\nabla u^{k_j})\, dx.$$

Letting $k_j \to \infty$, we have by Theorem 1.2 and the hypothesis (1.3) that

$$\begin{aligned} \varepsilon &\leqq \int_\Omega \varphi(\nabla u)\, dx - \int_{\Omega - E} \varphi(\nabla u)\, dx \\ &= \int_E \varphi(\nabla u)\, dx < \varepsilon, \end{aligned}$$

a contradiction.

Proof of Theorem 1.3. The proof of Theorem 1.3 also follows by the Dunford–Pettis criterion, using Theorem 1.1.

Proof of Corollary 1.4. A minimizing sequence for the functional

$$\int_\Omega W(\nabla v)\, dx$$

is also a minimizing sequence for its relaxation

$$\int_\Omega W^{\#}(\nabla v)\, dx$$

whose integrand $W^{\#}$ is quasi-convex and satisfies (1.11). We apply Theorem 1.3 to $W^{\#}$.

4. Constraint management in a limit case. Certain variational principles in elasticity constrain the admissible variations $v \in H^{1,p}(\Omega; \mathbb{R}^n)$, where $\Omega \subset \mathbb{R}^n$, to satisfy

$$\det \nabla v > 0 \quad \text{in } \Omega \text{ a.e.}$$

In the limit case $p = n$, $\det \nabla v \in L^1(\Omega)$ for $v \in H^{1,n}(\Omega; \mathbb{R}^n)$ but it is not necessarily integrable to any higher power. Thus it is not automatic that if $u^k \to u$ in $H^{1,n}(\Omega; \mathbb{R}^n)$ weakly, that $\det \nabla u^k \to \det \nabla u$ in $L^1(\Omega)$ weakly. In fact, without additional requirements, this condition does not hold. The reader may refer to the counterexamples in Ball and Murat [7]. However, much is known about this situation, as we summarize below.

First of all, the determinant is a null Lagrangian, that is, if $u, v \in H^{1,n}(\Omega; \mathbb{R}^n)$ and $u|_{\partial\Omega} = v|_{\partial\Omega}$, then

$$\int_\Omega \det \nabla u\, dx = \int_\Omega \det \nabla v\, dx. \tag{4.1}$$

Assume that $u^k, u \in H^{1,n}(\Omega; \mathbb{R}^n)$ and

$$u^k \to u \quad \text{in } H^{1,n}(\Omega; \mathbb{R}^n) \text{ weakly.} \tag{4.2}$$

Then for a subsequence of the (u^k), not relabeled (cf., e.g., [2]),

$$\det \nabla u^k \to \det \nabla u \quad \text{in } D'(\Omega). \tag{4.3}$$

Very recently, Müller [44] showed that if (4.2) holds and $\det \nabla u^k \geqq 0$, then

$$\det \nabla u^k \to \det \nabla u \quad \text{in } L^1_{\text{loc}}(\Omega) \text{ weakly.} \tag{4.4}$$

We give a slight generalization of Müller's result, and also an independent proof of it. With it, alternate proofs of some results in elasticity may be given, for example, some of those in Zhang [51].

THEOREM 4.1. *Let* $u^k, u \in H^{1,n}(\Omega; \mathbb{R}^n)$ *satisfy*

$$u^k \to u \quad \text{in } H^{1,n}(\Omega; \mathbb{R}^n) \text{ weakly,} \tag{4.5}$$

$$\det \nabla u^k \geqq 0 \quad \text{in } \Omega \text{ a.e.,} \tag{4.6}$$

$$u^k|_{\partial\Omega} = u_0|_{\partial\Omega}, \tag{4.7}$$

where $u_0 \in H^{1,n}(\Omega; \mathbb{R}^n)$ *is fixed. Then*

$$\det \nabla u^k \to \det \nabla u \text{ in } L^1(\Omega) \text{ weakly} \tag{4.8}$$

Proof. First of all, $u = u_0$ on $\partial\Omega$. From (4.6), we deduce that

$$\int_\Omega \zeta \det \nabla u \, dx \geqq 0 \quad \text{whenever } 0 \leqq \zeta \in C_0^\infty(\mathbb{R}^n),$$

thus $\det \nabla u \geqq 0$ in Ω almost everywhere. By (4.1),

$$\int_\Omega \det \nabla u \, dx = \int_\Omega \det \nabla u^k \, dx = \int_\Omega \det \nabla u_0 \, dx, \quad \text{for all } k. \tag{4.9}$$

Now let

$$\varphi(A) = \max\{\det A, 0\}, \quad A \in \mathbb{M},$$

which is continuous, quasi-convex, and satisfies

$$0 \leqq \varphi(A) \leqq C(1 + |A|)^n, \qquad A \in \mathbb{M}.$$

Then $\varphi(\nabla u^k) = \det \nabla u^k$ and $\varphi(\nabla u) = \det \nabla u$, so, trivially, by (4.9),

$$\int_\Omega \varphi(\nabla u) \, dx = \lim_{k\to\infty} \int_\Omega \varphi(\nabla u^k) \, dx.$$

Consequently, by Theorem 1.1, possibly for a subsequence which we do not relabel,

$$\det \nabla u^k \to \det \nabla u \quad \text{in } L^1(\Omega) \text{ weakly.} \qquad \square$$

The idea of Theorem 4.1 is that the sequence (u^k) may arise as a minimizing sequence for some variational principle subject to (4.6). Additional information then follows from the theorem.

5. Application to functionals with surface energies. We consider a simple situation where cooperative bulk and surface energies are minimized. Let $\Omega \subset \mathbb{R}^n$ have smooth boundary Γ and set

$$E(v) = \int_\Omega W(\nabla v) \, dx + \int_\Gamma \tau(\nabla v, \nu) \, dS, \qquad v \in C^1(\bar{\Omega}; \mathbb{R}^m), \tag{5.1}$$

where ν denotes the exterior normal to Γ. The infimum of E over $C^1(\bar{\Omega}; \mathbb{R}^m)$ is not necessarily the sum of the infima of its two summands, so we envision an application of our results when (1.3) will hold for each of the two terms but where these quantities will not be the unrestricted infima of their portions of the functional.

Assume that W is continuous and satisfies, for some $p>1$ and $C \geqq c>0$,

$$\max\{c|A|^p-1,0\} \leqq W(A) \leqq C(1+|A|^p), \qquad A \in \mathbb{M}. \tag{5.2}$$

About τ we assume that it is continuous and, for some $s>1$,

$$\begin{aligned} &0 \leqq \tau(A,\nu), \\ &c(|A_{\tan}|^s-1) \leqq \tau(A,\nu) \leqq C(|A|^s+1), \end{aligned} \qquad A \in \mathbb{M}, \tag{5.3}$$

where $A_{\tan}=A(\mathbb{1}-\nu\otimes\nu)$ is the tangential part of A.

For a fixed $\nu \in \mathbb{S}^{n-1}$, let $D' \subset \{x\cdot\nu=0\}$ be a domain and let dx' denote the $(n-1)$-Lebesgue measure on D'. By $D'\times(-r,r)$, $r>0$, we abbreviate the name of the set

$$\{x \in \mathbb{R}^n \colon x'=(\mathbb{1}-\nu\otimes\nu)x \in D' \text{ and } |x\cdot\nu|<r\}.$$

Let $[E]$ denote the $(n-1)$-dimensional Lebesgue measure of E. We define

$$\begin{aligned} &\tau^{\#}(F,\nu)=\inf_{C'}\frac{1}{[D']}\int_{D'}\tau(F+\nabla\zeta,\nu)\,dx', \qquad (F,\nu)\in\mathbb{M}\times\mathbb{S}^{n-1}, \\ &C'=C_0^1(D'\times(-r,r)). \end{aligned} \tag{5.4}$$

We always suppose that $[\partial D']=0$. Clearly $\tau^{\#}\geqq 0$ and is independent of r. The relaxation of the functional E is given by

$$E^{\#}(v)=\int_\Omega W^{\#}(\nabla v)\,dx+\int_\Gamma \tau^{\#}(\nabla v,\nu)\,dS, \qquad v\in C^1(\bar{\Omega};\mathbb{R}^m), \tag{5.5}$$

where $W^{\#}(A)$ is the ordinary quasi convexification of W and $\tau^{\#}$ is defined by (5.4). A special property of $\tau^{\#}$ is that

$$\tau^{\#}(A,\nu)=\tau^{\#}(A_{\tan},\nu), \qquad A\in\mathbb{M},$$

which implies that

$$c(|A_{\tan}|^s-1)\leqq\tau^{\#}(A,\nu)\leqq C(|A|^s+1), \qquad A\in\mathbb{M}, \tag{5.6}$$

and that $\tau^{\#}$ is well defined on $H^{1,s}(\Gamma;\mathbb{R}^m)$. An easy generalization of [39] tells us that

$$\inf_{C^1(\bar{\Omega})}E(v)=\inf_V E^{\#}(v), \qquad V=H^{1,p}(\Omega;\mathbb{R}^m)\times H^{1,s}(\Gamma;\mathbb{R}^m). \tag{5.7}$$

Let $(u^k)\subset V$ be a minimizing sequence for E. Then (u^k) is a minimizing sequence for $E^{\#}$, which is bounded in V. Suppose that $u\in V$ and $u^k\to u$ in V weakly. By lower semicontinuity,

$$\begin{aligned} &E^{\#}(u)=\lim_{k\to\infty}E^{\#}(v)=\inf_{C^1(\bar{\Omega})}E(v)=\inf_V E^{\#}(v) \text{ and} \\ &\int_\Omega W^{\#}(\nabla u)\,dx=\lim_{k\to\infty}\int_\Omega W^{\#}(\nabla u^k)\,dx, \\ &\int_\Gamma \tau^{\#}(\nabla_{\tan}u,\nu)\,dS=\lim_{k\to\infty}\int_\Gamma\tau^{\#}(\nabla u^k,\nu)\,dS. \end{aligned} \tag{5.8}$$

We may apply Theorem 1.3, or a slight generalization of it in the case of $(\tau^{\#}(\nabla u^k,\nu))$, to deduce that

$$\begin{aligned} &W^{\#}(\nabla u^k)\to W^{\#}(\nabla u) \quad \text{in } L^1(\Omega) \text{ weakly}, \\ &\tau^{\#}(\nabla u^k,\nu)\to\tau^{\#}(\nabla_{\tan}u,\nu) \quad \text{in } L^1(\Gamma) \text{ weakly}. \end{aligned}$$

If $\mu=(\mu_x)_{x\in\Omega}$ denotes a Young measure generated by (∇u^k), we have the limit energy representations

$$\bar{W}(x)=W^{\#}(\nabla u(x))=\int_{\mathbb{M}} W(A)\,d\mu_x(A),\qquad x\in\Omega,$$

$$\bar{\tau}(x)=\tau^{\#}(\nabla_{\tan}u(x),\nu(x))=\int_{\Gamma}\tau^{\#}(A,\nu(x))\,d\mu_x(A),\qquad x\in\Gamma,$$

and

$$\int_\Omega \bar{W}(x)\,dx+\int_\Gamma \bar{\tau}(x)\,dS=\inf_{C^1(\bar{B})}E(v).$$

6. Measure valued solutions of an evolution problem. Some of our methods may be employed to study measure valued solutions of evolution problems. A more extensive treatment is given by Slemrod [47]; here we wish to explain merely how such solutions may come about. For further developments we refer to Demoulini [17]. To fix the ideas, we consider a scalar case. Suppose that $\varphi\in C^1(\mathbb{R}^n)$ satisfies

$$\begin{aligned}&\max(c|a|^2-1,0)\leqq\varphi(a)\leqq C(|a|^2+1),\\ &|\nabla\varphi(a)|\leqq C|a|,\end{aligned}\qquad a\in\mathbb{R}^n,\tag{6.1}$$

where $0<c\leqq C$. Let $q(a)=\nabla\varphi(a)$. Our interest is in solutions, possibly Young measures, which in some sense satisfy

$$-\operatorname{div}\bar{q}+\frac{\partial u}{\partial t}=0\quad\text{in }\Omega\times\mathbb{R}^+,\tag{6.2}$$

$\mathbb{R}^+=(0,\infty)$, subject to appropriate boundary conditions.

To render this more precise, let us agree that $\nu=(\nu_{x,t})_{(x,t)\in\Omega\times\mathbb{R}^+}$ is a Young measure solution of (6.2) provided that the following condition holds.

CONDITION 6.1. *ν is a family of probability measures and*

$$u\in L^\infty(\mathbb{R}^+;H_0^1(\Omega))\quad\textit{with }\frac{\partial u}{\partial t}\in L^2(\Omega\times\mathbb{R}^+)$$

which satisfy

$$-\operatorname{div}\bar{q}+\frac{\partial u}{\partial t}=0\quad\textit{in }H^{-1}(\Omega\times\mathbb{R}^+),\tag{6.3}$$

$$u|_{t=0}=u_0,\tag{6.4}$$

where

$$\begin{aligned}\bar{q}(x,t)&=\int_{\mathbb{R}^n}q(a)\,d\nu_{x,t}(a)\\ \nabla u(x,t)&=\int_{\mathbb{R}^n}a\,d\nu_{x,t}(a)\end{aligned}\qquad\textit{in }\Omega\times\mathbb{R}^+\textit{ a.e.}\tag{6.5}$$

Above, $u_0\in H_0^1(\Omega)$ is given. Moreover, we shall impose the condition that ν is a Young measure in the sense that it meets the following condition.

CONDITION 6.2. *ν is generated by a sequence (∇u^h), $h>0$, where*

$$u^h\in L^\infty(\mathbb{R}^+;H_0^1(\Omega)).\tag{6.6}$$

The equation (6.5) means that

$$\int_0^\infty \int_\Omega \left(\bar{q} \cdot \nabla \zeta + \frac{\partial u}{\partial t} \zeta \right) dx\, dt = 0 \quad \text{for } \zeta \in H_0^1(\Omega \times \mathbb{R}^+). \tag{6.7}$$

We shall give an outline of the proof of the following theorem.

THEOREM 6.1. *Assume* (6.1) *about* φ. *Then there exists a Young measure solution* $\nu = (\nu_{x,t})_{(x,t) \in \Omega \times \mathbb{R}^+}$ *of*

$$-\operatorname{div} \bar{q} + \frac{\partial u}{\partial t} = 0 \quad \textit{in } \Omega \times \mathbb{R}^+,$$

satisfying (6.3)-(6.6). *In addition,*

$$\operatorname{supp} \nu_{x,t} \subset \{a \in \mathbb{R}^n \colon \varphi(a) = \varphi^{**}(a)\} \quad \textit{in } \Omega \times \mathbb{R}^+ \ \textit{a.e.}, \tag{6.8}$$

where φ^{**} *is the convexification of* φ.

Recall that if $\varphi \in C^1(\mathbb{R}^n)$, then $\varphi^{**} \in C^1(\mathbb{R}^n)$, whence

$$q(a) = q^{**}(a) \quad \text{in } \{a \in \mathbb{R}^n \colon \varphi(a) = \varphi^{**}(a)\},$$

where $q^{**}(a) = \nabla \varphi^{**}(a)$. Note also that φ^{**} satisfies (6.1). Hence the following corollary.

COROLLARY 6.2. *Assume* (6.1) *about* φ *and let* $\nu = (\nu_{x,t})_{(x,t) \in \Omega \times \mathbb{R}^+}$ *be a Young measure solution satisfying* (6.8). *Then* ν *is a solution of the relaxed problem*

$$-\operatorname{div} \bar{q}^{**} + \frac{\partial u}{\partial t} = 0 \quad \textit{in } \Omega \times \mathbb{R}^+. \tag{6.9}$$

The constructed solution has some additional properties which we shall describe in the sequel.

Step 1. An equilibrium problem. Let $w \in H_0^1(\Omega)$ and $h > 0$ and consider

$$\Phi(v) = \Phi_h(v) = \int_\Omega \left(\varphi(\nabla v) + \frac{1}{2h} |v - w|^2 \right) dx, \qquad v \in H_0^1(\Omega), \tag{6.10}$$

$$\Phi^{**}(v) = \int_\Omega \left(\varphi^{**}(\nabla v) + \frac{1}{2h} |v - w|^2 \right) dx, \qquad v \in H_0^1(\Omega), \tag{6.11}$$

where φ^{**} is the convexification of φ. By a known relaxation theorem (cf. [16]),

$$I = \inf\nolimits_{H_0^1(\Omega)} \Phi(v) = \inf\nolimits_{H_0^1(\Omega)} \Phi^{**}(v). \tag{6.12}$$

Now let (v^k) be a minimizing sequence for $\Phi(v)$. We may assume there is a $u \in H_0^1(\Omega)$ such that

$$v^k \to u \quad \text{in } H_0^1(\Omega) \text{ weakly as } k \to \infty.$$

By lower semicontinuity,

$$\Phi(v^k) \to \Phi^{**}(u) \quad \text{as } k \to \infty,$$

and by the Rellich theorem,

$$\int_\Omega |v^k - w|^2\, dx \to \int_\Omega |u - w|^2\, dx \quad \text{as } k \to \infty.$$

Hence

$$\int_\Omega \varphi^{**}(\nabla u)\, dx = \lim\nolimits_{k \to \infty} \int_\Omega \varphi^{**}(\nabla v^k)\, dx = \lim\nolimits_{k \to \infty} \int_\Omega \varphi(\nabla v^k)\, dx.$$

Hence, by Theorem 1.1,

$$\varphi^{**}(\nabla v^k) \to \varphi^{**}(\nabla u) \quad \text{in } L^1(\Omega) \text{ weakly},$$
$$\varphi(\nabla v^k) \to \varphi^{**}(\nabla u) \quad \text{in } L^1(\Omega) \text{ weakly}.$$

Denoting by $\nu = (\nu_x)_{x \in \Omega}$ the Young measure generated by (∇v^k),

$$\begin{gathered} \operatorname{supp} \nu \subset \{a \in \mathbb{R}^n : \varphi(a) = \varphi^{**}(a)\}, \\ \varphi^{**}(\nabla u) = \bar{\varphi} = \bar{\varphi}^{**} \quad \text{and} \quad \bar{q} = \bar{q}^{**} \quad \text{in } \Omega \text{ a.e.}, \end{gathered} \tag{6.13}$$

where

$$\bar{\psi}(x) = \int_{\mathbb{R}^n} \psi(a)\, d\nu_x(a) \quad \text{in } \Omega \text{ a.e.}$$

In fact, the Young measure representation holds for any $\psi \in E$, where

$$E = \left\{ \psi \in C(\mathbb{R}^n \times \Omega) : \sup_{\mathbb{R}^n \times \Omega} \frac{|\psi(a, x)|}{|a|^2 + 1} < \infty \right\}.$$

We may now apply the technique developed in [10] to discuss stable Young measure minimizers of variational principles; cf. § 5. The idea here is to observe that

$$\psi_\varepsilon(a, x) = \varphi(a + \varepsilon \nabla \zeta(x)) \leqq C(1 + |a|^2), \qquad a \in \mathbb{R}^n,$$

when $\zeta \in C_0^\infty(\Omega)$, $-1 \leqq \varepsilon \leqq 1$. Hence $\psi_\varepsilon \in E$, so

$$\begin{aligned} \int_\Omega \bar{\varphi}\, dx &\leqq \lim_{k\to\infty} \int_\Omega \varphi(\nabla v^k + \varepsilon \nabla \zeta)\, dx \\ &= \int_\Omega \int_{\mathbb{R}^n} \varphi(a + \varepsilon \nabla \zeta(x))\, d\nu_x(a)\, dx. \end{aligned}$$

This equation may be differentiated with respect to ε. As a consequence, we may write an equilibrium equation

$$\int_\Omega \left(\bar{q} \cdot \nabla \zeta + \frac{1}{h}(u - w)\zeta \right) dx = 0 \quad \text{for } \zeta \in H_0^1(\Omega). \tag{6.14}$$

Finally, the Young measure representation provides us with an elementary estimate for $\bar{q}$. Indeed, using the estimates of (6.1) and (6.13),

$$\begin{aligned} \int_\Omega |\bar{q}|^2\, dx &\leqq \int_\Omega \int_{\mathbb{R}^n} |q(a)|^2\, d\nu_x(a)\, dx \\ &\leqq C \int_\Omega \int_{\mathbb{R}^n} |a|^2\, d\nu_x(a)\, dx \\ &\leqq C \int_\Omega \int_{\mathbb{R}^n} (\varphi(a) + 1)\, d\nu_x(a)\, dx \\ &= C \int_\Omega (\varphi^{**}(\nabla u) + 1)\, dx. \end{aligned} \tag{6.15}$$

Step 2. Approximate solution. Let $u_0 \in H_0^1(\Omega)$ be given and $h > 0$. We define a sequence of Young measure solutions $\nu^{h,j}$ and underlying functions $u^{h,j}$ by setting

$$\nu^{h,0} = \delta_{\nabla u_0} \quad \text{and} \quad u^{h,0} = u_0$$

and $\nu^{h,j+1}$ is the solution of (6.12) with $w=u^{h,j}$ and $u^{h,j+1}$ its underlying function. We are then in possession of the energy densities

$$\varphi^{**}(\nabla u^{h,j})=\langle \nu^{h,j},\varphi\rangle=\langle \nu^{h,j},\varphi^{**}\rangle \tag{6.16}$$

and the flux densities

$$\bar{q}^{h,j}=\langle \nu^{h,j},q\rangle=\langle \nu^{h,j},q^{**}\rangle. \tag{6.17}$$

Let $I^{h,j}=[hj,h(j+1))$, $\chi^{h,j}=\chi_{I^{h,j}}$, the characteristic function of $I^{h,j}$, and

$$\lambda^{h,j}(t)=\begin{cases}\dfrac{t}{h}-j, & hj\leqq t\leqq h(j+1),\\ 0, & \text{otherwise.}\end{cases}$$

Set

$$u^h(x,t)=\textstyle\sum_j \chi^{h,j}\{(1-\lambda^{h,j}(t))u^{h,j}(x)+\lambda^{h,j}(t)u^{h,j+1}(x)\}\in L^\infty(\mathbb{R}^+;H_0^1(\Omega)) \tag{6.18}$$

and

$$\nu_{x,t}^h=\textstyle\sum_j \chi^{h,j}(t)\nu_x^{h,j}\in E'. \tag{6.19}$$

Now from (6.18),

$$\frac{\partial u^h}{\partial t}=\frac{1}{h}(u^{h,j+1}-u^{h,j})\quad\text{and}\quad \bar{q}^h=\langle \nu^h,q\rangle=\textstyle\sum_j \bar{q}^{h,j}\chi^{h,j} \tag{6.20}$$

comprise a solution of

$$-\operatorname{div}\bar{q}^h+\frac{\partial u^h}{\partial t}=0\quad\text{in } H^{-1}(\Omega),\quad\text{for each } t,$$

from which it is elementary to check that

$$\int_0^\infty\int_\Omega\left(\bar{q}^h\cdot\nabla\zeta+\frac{\partial u^h}{\partial t}\zeta\right)dx\,dt=0\quad\text{for } \zeta\in H_0^1(\Omega\times\mathbb{R}^+). \tag{6.21}$$

Step 3 (Estimates). Uniform estimates are available for $u^h\in L^\infty(\mathbb{R}^+;H_0^1(\Omega))$ and $\partial u^h/\partial t\in L^2(\Omega\times\mathbb{R}^+)$. To begin, $u^{h,j}$ is admissible in the variational principle for $u^{h,j+1}$, so

$$\int_\Omega\left(\varphi^{**}(\nabla u^{h,j+1})+\frac{1}{2h}|u^{h,j+1}-u^{h,j}|^2\right)dx\leqq\int_\Omega\varphi^{**}(\nabla u^{h,j})\,dx.$$

Hence

$$\int_\Omega\varphi^{**}(\nabla u^{h,j})\,dx\leqq\int_\Omega\varphi^{**}(\nabla u_0)\,dx=M^2 \tag{6.22}$$

and

$$\frac{1}{2h}\textstyle\sum_j|u^{h,j+1}-u^{h,j}|^2\leqq\displaystyle\int_\Omega\varphi^{**}(\nabla u_0)\,dx=M^2. \tag{6.23}$$

Since φ^{**} satisfies (6.1), the inequality (6.22) tells us that

$$\|\nabla u^{h,j}\|_{L^2(\Omega)}\leqq M. \tag{6.24}$$

By convexity of the L^2 norm and (6.24) we have that

$$\|u^h\|_{L^\infty(\mathbb{R}^+;H_0^1(\Omega))} \leqq M. \tag{6.25}$$

Rearranging a little in (6.23) and noting (6.20),

$$\int_0^\infty \int_\Omega \left|\frac{\partial u^h}{\partial t}\right|^2 dx\,dt \leqq M^2. \tag{6.26}$$

Introduce the function

$$w^h(x,t) = \textstyle\sum_j u^{h,j}(x)\chi^{h,j}(t) \in L^\infty(\mathbb{R}^+; H_0^1(\Omega)). \tag{6.27}$$

Then (6.24) implies that

$$\|w^h\|_{L^\infty(\mathbb{R}^+;H_0^1(\Omega))} \leqq M. \tag{6.28}$$

Finally, we wish to estimate $\bar{q}^h$ using (6.15), which provides the estimate

$$\|\bar{q}^h\|_{L^\infty(\mathbb{R}^+;L^2(\Omega))} \leqq C \int_\Omega (\varphi^{**}(\nabla u^{h,j})+1)\,dx \leqq C(M^2+1). \tag{6.29}$$

Step 4 (Passage to the limit). We let $h \to 0$. From the estimates (6.25), (6.26), (6.28), and (6.29), we may extract a subsequence of h as $h \to 0$ and

$$\begin{aligned}
&\nu = (\nu_{x,t})_{(x,t)\in\Omega\times\mathbb{R}^+} \in E' && \text{with supp } \nu \subset \{\varphi(a) = \varphi^{**}(a)\} \text{ and } \nu \text{ is a Young measure,}\\
&w \in L^\infty(\mathbb{R}^+; H_0^1(\Omega)) && \text{with } \nabla w = \langle \nu, a\rangle,\\
&\bar{q} \in L^\infty(\mathbb{R}^+; L^2(\Omega)) && \text{with } \bar{q} = \langle \nu, q\rangle = \langle \nu, q^{**}\rangle,\\
&u \in L^\infty(\mathbb{R}^+; H_0^1(\Omega)) && \text{with } \frac{\partial u}{\partial t} \in L^2(\Omega\times\mathbb{R}^+),
\end{aligned}$$

which satisfy

$$\int_0^\infty \int_\Omega \left(\bar{q}\cdot\nabla\zeta + \frac{\partial u}{\partial t}\zeta\right) dx\,dt = 0 \quad \text{for } \zeta \in H_0^1(\Omega\times\mathbb{R}^+). \tag{6.30}$$

In fact, (6.30) above holds for $\zeta \in L^\infty(\mathbb{R}^+; H_0^1(\Omega))$. We remark that ν is a Young measure but it is not generated by the sequence (∇u^h) of (6.18), but rather by a diagonal subsequence of the functions which generate the (ν^h) of (6.19). Although $\nu \in E'$, we have not verified the Young measure representation for $\psi \in E$, although, as we mentioned in the introduction, under these circumstances, whenever $\psi \in C(\mathbb{R}^m)$ satisfies

$$\lim_{|a|\to\infty} \frac{|\psi(a)|}{1+|a|^2} = 0,$$

the sequence $\psi(\nabla v^k)$ converges weakly in L^1 and the representation is valid.

It remains to show that the Young measure ν and the limit function u are connected. We claim that $u = w$. In fact, we shall show that $\nabla u = \nabla w$ by means of an easy lemma.

LEMMA 6.3. *Let* $(f^{h,j}) \subset$ *bounded set of* $L^2(\Omega)$ *for* $h > 0$ *and* $j = 1, 2, 3, \cdots$, *and set*

$$f^h(x,t) = \textstyle\sum_j f^{h,j}(x)\chi^{h,j}(t),$$

$$g^h(x,t) = \textstyle\sum_j \chi^{h,j}\{(1-\lambda^{h,j}(t))f^{h,j}(x) + \lambda^{h,j}(t)f^{h,j+1}(x)\},$$

where $\chi^{h,j}$ is the characteristic function of $[hj, h(j+1))$ and

$$\lambda^{h,j}(t)=\begin{cases} t/h-j, & hj\leqq t\leqq h(j+1),\\ 0, & \text{otherwise.}\end{cases}$$

Suppose that

$$f^h\to f \quad \text{and} \quad g^h\to g \quad \text{in } L^2_{\text{loc}}(\Omega\times\mathbb{R}^+) \text{ weakly.}$$

Then $f=g$.

Proof. It suffices to show that

$$\int_0^\infty\int_\Omega f\zeta\,dx\,dt=\int_0^\infty\int_\Omega g\zeta\,dx\,dt$$

for $\zeta\in C_0^\infty(\Omega)$ of the form $\zeta(x,t)=w(x)z(t)$. Let $z^{h,j}=z(hj)$ and

$$\zeta^h(x,t)=w(x)\sum_j z^{h,j}\chi^{h,j}(t),$$
$$\xi^h(x,t)=w(x)\sum_j\chi^{h,j}(t)\{(1-\lambda^{h,j}(t))z^{h,j}+\lambda^{h,j}(t)z^{h,j-1}\}.$$

It is elementary to check that $\zeta^h\to\zeta$ and $\xi^h\to\zeta$ uniformly since z is smooth. Since

$$\int_0^\infty\int_\Omega f^h\xi^h\,dx\,dt=\int_0^\infty\int_\Omega g^h\zeta^h\,dx\,dt,$$

the lemma follows. □

From the lemma we may write a generalized Fourier law for the solution, which is really much weaker than the property that supp $\nu\subset\{\varphi(a)=\varphi^{**}(a)\}$. If $\varphi^{**}(0)=0$, then

$$\bar{q}^{**}\cdot\nabla u\geqq 0.$$

Acknowledgment. We are delighted to thank J. Matos and S. Demoulini for many helpful conversations. We also thank M. Gurtin and L. Tartar for their interest in this work.

REFERENCES

[1] E. ACERBI AND N. FUSCO (1984), *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86, pp. 125–145.

[2] J. M. BALL, (1977), *Constitutive inequalities and existence theorems in nonlinear elastostatics*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. I, R. Knops, ed., Pitman Res. Notes in Math. 17, Longman Scientific & Technical, Harlow, UK, pp. 187–241.

[3] ——— (1989), *A version of the fundamental theorem for Young measures*, PDE's and Continuum Models of Phase Transitions, M. Rascle, D. Serre and M. Slemrod, eds., Lecture Notes in Physics 344, Springer-Verlag, New York, pp. 207–215.

[4] ——— (1990), *Sets of gradients with no rank-one connections*, J. Math. Pures Appl, 69, pp. 241–259.

[5] J. M. BALL AND R. JAMES (1987), *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100, pp. 15–52.

[6] ——— (1989), *Proposed experimental tests of a theory of fine microstructure and the two well problem.*

[7] J. M. BALL AND F. MURAT (1984), *$W^{1,p}$-quasiconvexity and variational problems for multiple integrals*, J. Funct. Anal., 58, pp. 225–253.

[8] J. M. BALL AND K. ZHANG (1990), *Lower semicontinuity of multiple integrals and the biting lemma*, Proc. Roy. Soc. Edinburgh Sect. A, 114, pp. 367–379.

[9] M. CHIPOT AND C. COLLINS, (1990), *Numerical approximation in variational problems with potential wells*, Institute for Mathematics and Its Applications, Minneapolis, MN, preprint.

[10] M. CHIPOT AND D. KINDERLEHRER (1988), *Equilibrium configurations of crystals*, Arch. Rational Mech. Anal., 103, pp. 237-277.

[11] M. CHIPOT, (1991), *Numerical analysis of oscillations in nonconvex problems*, to appear.

[12] C. COLLINS AND M. LUSKIN (1989), *The computation of the austenitic-martensitic phase transition*, in PDE's and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Physics 344, Springer-Verlag, New York, pp. 34-50.

[13] ——— (1991), *Numerical modeling of the microstructure of crystals with symmetry-related variants*, Proc. ARO US-Japan Workshop on Smart/Intelligent Materials and Systems, Technomic.

[14] C. COLLINS, D. KINDERLEHRER, AND M. LUSKIN (1991), *Numerical approximation of the solution of a variational problem with a double well potential*, SIAM J. Numer. Anal., 28, pp. 321-332.

[15] B. DACOROGNA (1982), *Weak continuity and weak lower semicontinuity of nonlinear functionals*, Lecture Notes in Math. 922, Springer-Verlag, New York.

[16] ——— (1989), *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York.

[17] S. DEMOULINI, Ph.D. thesis, University of Minnesota, Minneapolis, MN.

[18] J. L. ERICKSEN (1979), *On the symmetry of deformable crystals*, Arch. Rational Mech. Anal., 72, pp. 1-13.

[19] ——— (1980), *Some phase transitions in crystals*, Arch. Rational Mech. Anal., 73, pp. 99-124.

[20] ——— (1981), *Changes in symmetry in elastic crystals*, IUTAM Sympos. Finite Elasticity, D. E. Carlson and R. T. Shield, eds., M. Nijhoff, the Netherlands, pp. 167-177.

[21] ——— (1981), *Some simpler cases of the Gibbs phenomenon for thermoelastic solids*, J. Thermal Stresses, 4, pp. 13-30.

[22] ——— (1982), *Crystal lattices and sublattices*, Rend. Sem. Mat. Univ. Padova, 68, pp. 1-9.

[23] ——— (1983), *Ill posed problems in thermoelasticity theory*, in Systems of Nonlinear Partial Differential Equations, J. Ball, ed., D. Reidel, Boston, MA, pp. 71-95.

[24] ——— (1984), *The Cauchy and Born hypotheses for crystals*, in Phase Transformations and Material Instabilities in Solids, M. Gurtin, ed., Academic Press, pp. 61-78.

[25] ——— (1986), *Constitutive theory for some constrained elastic crystals*, Internat. J. Solids Structures, 22, pp. 951-964.

[26] ——— (1986), Stable equilibrium configurations of elastic crystals, Arch. Rational Mech. Anal., 94, pp. 1-14.

[27] ——— (1987), *Twinning of crystals I*, in Metastability and Incompletely Posed Problems, S. Antman, J. L. Ericksen, D. Kinderlehrer, and I. Müller, eds., IMA Vol. Math. Appl. 3, Springer-Verlag, New York, pp. 77-96.

[28] ——— (1988), *Some constrained elastic crystals*, in Material Instabilities in Continuum Mechanics, J. Ball, ed., Oxford University Press, London, pp. 119-136.

[29] ——— (1989), *Weak martensitic transformations in Bravais lattices*, Arch. Rational Mech. Anal., 107, pp. 23-36.

[30] L. C. EVANS (1990), *Weak convergence methods for nonlinear partial differential equations*, CBMS Regional Conf. Ser. Math. 74, Conf. Board Math. Sci., Washington, DC.

[31] I. FONSECA (1985), *Variational methods for elastic crystals*, Arch. Rational Mech. Anal., 97, pp. 189-220.

[32] ——— (1988), *The lower quasiconvex envelope of the stored energy function for an elastic crystal*, J. Math. Pures Appl., 67, pp. 175-195.

[33] ——— (1991), *Lower semicontinuity of surface energies*, Carnegie Mellon University, preprint.

[34] ——— (1991), *The Wulff Theorem revisited*, Proc. Roy. Soc. London Ser. A., 432, pp. 125-145.

[35] R. D. JAMES (1988), *Microstructure and weak convergence*, Proc. Sympos. Material Instabilities in Continuum Mechanics, Heriot-Watt, J. M. Ball, ed., Oxford University Press, London, pp. 175-196.

[36] R. D. JAMES AND D. KINDERLEHRER (1989), *Theory of diffusionless phase transitions*, in PDE's and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Physics 344, Springer-Verlag, New York, pp. 51-84.

[37] D. KINDERLEHRER (1988), *Remarks about the equilibrium configurations of crystals*, Proc. Sympos. Material Instabilities in Continuum Mechanics, Heriot-Watt, J. M. Ball, ed., Oxford University Press, London, pp. 217-242.

[38] D. KINDERLEHRER AND P. PEDREGAL, *Remarks about Young measures supported on two wells*, to appear.

[39] D. KINDERLEHRER AND G. VERGARA-CAFFARELLI (1989), *The relaxation of functionals with surface energies*, Asymptotic Anal., 2, pp. 279-298.

[40] F.-C. LIU (1977), *A Luzin type property of Sobelov functions*, Indiana Univ. Math. J., 26, pp. 645-651.

[41] J. MATOS, *The absence of fine microstructure in α-β quartz*, to appear.

[42] ——— (1990), Ph.D. thesis, University of Minnesota, Minneapolis, MN.

[43] C. B. MORREY, JR. (1966), *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, New York.

[44] S. MÜLLER (1991), *Higher integrability of determinants and weak convergence in L^1*, preprint.

[45] P. PEDREGAL (1989), Ph.D. thesis, University of Minnesota, Minneapolis, MN.

[46] ——— (1989), *Weak continuity and weak lower semicontinuity for some compensation operators*, Proc. Roy. Soc. Edinburgh Sect. A, 113, pp. 267–279.

[47] M. SLEMROD (1991), *Dynamics of measure valued solutions to a backward-forward heat equation*, to appear.

[48] L. TARTAR (1983), *The compensated compactness method applied to systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., D. Reidel, Boston, MA.

[49] ——— (1984), *Étude des oscillations dans les équations aux dérivées partielles nonlinéaries*, Lecture Notes in Physics 195, Springer-Verlag, Berlin, pp. 384–412.

[50] L. C. YOUNG (1969), Lectures on Calculus of Variations and Optimal Control Theory, W. B. Saunders, Philadelphia, PA.

[51] K. ZHANG (1990), *Biting theorems for Jacobians and their applications*, Anal. Nonlinéaire, 7, pp. 345–366.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 20-28, January 1992

REMARQUES SUR L'EXISTENCE GLOBALE POUR LE SYSTEME DE NAVIER-STOKES INCOMPRESSIBLE*

J. Y. CHEMIN†

Résumé. Le but de cet article est d'appliquer une estimation d'énergie à perte démontrée avec la technique du découpage dyadique de l'espace des fréquences à la démonstration de résultats d'existence globale de solutions suffisamment régulières des équations de Navier-Stokes avec un terme de force.

Abstract. The goal of this paper is the application of an energy estimate with loss of derivative, proved with the Littlewood-Paley theory, to prove some results of global existence of smooth enough solutions of the Navier-Stokes equations with an exterior force.

Mots-clefs. mécanique des fluides (visqueux), Littlewood-Paley (théorie de)

Codes matières AMS. 35L60, 76A02

Introduction. Nons nous intéressons dans ce travail au mouvement des particules d'un fluide incompressible visqueux. Ce mouvement est décrit par le système de Navier-Stokes relatif à un champ de vecteurs défini sur tout $\mathbf{R}^d$, d valant ici 2 ou 3, que nous rappelons:

$$(S_\varepsilon) \qquad \begin{aligned} &\partial_t v + v.\nabla v - \varepsilon\Delta v = -\nabla p + \nabla^\perp V,\\ &\operatorname{div} v(t,\cdot)=0 \quad \text{à tout instant } t \text{ positif ou nul},\\ &v|_{t=0}=v_0. \end{aligned}$$

Ici, $v(t,x)$ désigne la vitesse d'une particule située au point x à l'instant t, $p(t,x)$ la pression dans le fluide au point x à l'instant t, $V(t,x)$ le potentiel (ici une matrice antisymétrique), dont dérive la force extérieure au point x à l'instant t, $\nabla^\perp V$ le vecteur de jème coordonnée $\sum_i \partial_i V_j^i$, et ε la viscosité du fluide, qui est supposée être une constante strictement positive. Les données du problème sont bien sûr le potentiel V et le champ des vitesses v_0.

Le but de ce travail est de démontrer des résultats d'existence globale en regardant le système de Navier-Stokes, non pas comme un système parabolique, mais comme un système hyperbolique amélioré par le terme de viscosité. Ce point de vue permet de retrouver des résultats bien connus, comme par exemple le Théorème 0.1 relatif à la dimension deux et le Théorème général d'unicité 2.2, et d'en démontrer un nouveau, le Théorème 0.2. Dans cette optique, l'existence locale en temps pour des données régulières, par exemple, $v_0 \in H^s$ et $V \in L^1_{\text{loc}}[(0,+\infty[;H^{s+1})$, $s > d/2+1$, résulte de la théorie classique des systèmes hyperboliques (voir, par exemple, [1]), le terme de viscosité étant ignoré grâce à son signe. Cette théorie repose sur la démonstration d'estimations d'énergie. La première estimation d'énergie vérifiée par une solution assez régulière du système (S_ε) provient de l'identité suivante:

$$(0.1)\ (|v(t,\cdot)|_0)^2-(|v_0|_0)^2+2\varepsilon\int_{[0,t]}(|\nabla v(\tau,\cdot)|_0)^2\,dt = 2\int_{[0,t]}(\nabla^\perp V(\tau,\cdot)|v(\tau,\cdot))\,d\tau.$$

En utilisant le fait que $2(\nabla^\perp V(\tau,\cdot)|v(\tau,\cdot)) \leqq \varepsilon^{-1}(|V(\tau,\cdot)|_0)^2+\varepsilon(|\nabla v(\tau,\cdot)|_0)^2$, nous

* Received by the editors June 8, 1990; accepted for publication (in revised form) March 20, 1991.

† Centre de Mathematiques, Unité de Recherche Associée D 0169, Ecole Polytechnique, F-91128 Palaiseau Cedex, France.

obtenons trivialement l'estimation bien connue suivante:

$$(E) \qquad (|v(t,\cdot)|_0)^2+\varepsilon\int_{[0,t]}(|\nabla v(\tau,\cdot)|_0)^2\,d\tau \leqq (|v_0|_0)^2+\varepsilon^{-1}\int_{[0,t]}(|V(\tau,\cdot)|_0)^2\,d\tau.$$

L'idée de l'article est d'appliquer des méthodes d'énergies pour estimer l'évolution, non plus nécessairement de la norme L^2 du champ des vitesses v, mais de la norme H^s, pour un s assurant l'existence globale d'une solution suffisamment régulière pour être unique dans sa classe. L'outil permettant de réaliser ce programme est une estimation d'énergie avec perte dans les espaces de Sobolev homogènes sur le terme $(v.\nabla a|a)$. Cette estimation, ne mettant en jeu qu'un faible niveau de régularité du champ de vecteurs v, introduit des pertes de régularité que le terme de viscosité permet d'absorber.

La structure de l'article sera la suivante. Dans le premier paragraphe, on exposera la démonstration de l'inégalité d'énergie avec perte de régularité, démonstration qui repose sur la caractérisation en couronnes dyadiques des espaces de Sobolev homogènes et inhomogènes. Nous donnerons de plus une première application de ce résultat sous la forme d'une condition nécessaire de non-existence globale de solutions régulières. Dans le deuxième paragraphe, nous utiliserons cette estimation pour démontrer un lemme d'explosion, un théorème d'unicité et un lemme de convergence qui, appliqués au cas de la *dimension deux d'espace*, permettent de redémontrer fort aisément le théorème bien connu suivant.

THÉOREME 0.1. *Soit s un réel postif ou nul; si v_0 est un champ de vitesses de $\mathbf{R}^2$ à coefficients L^2, et V un potentiel-vecteur de $\mathbf{R}^2$ à coefficients $L^2_{\mathrm{loc}}([0,+\infty[;H^s)$; alors le système (S_ε) admet une unique solution dans l'espace $C([0,+\infty[;L^2)\cap C(]0,+\infty[;H^s)\cap L^2_{\mathrm{loc}}(]0,+\infty[;H^{s+1})$.*

Remarquons que ce théorème est tout-à-fait classique, voir, par exemple, [4].

Dans le troisième et dernier paragraphe, nous appliquerons ce qui précède au cas de la *dimension trois d'espace.* Désignons par H^s (respectivement, $\underline{H}^s$) l'espace de Sobolev (respectivement, homogène), et par $|u|_s$ (respectivement, $\|u|_s$) la quantité $\int(1+|\xi|^2)^s|\hat{u}(\xi)|^2\,d\xi$ (respectivement, $\int|\xi|^{2s}|\hat{u}(\xi)|^2\,d\xi$); nous démontrerons alors le théorème suivant.

THÉOREME 0.2. *Il existe un réel C_0 strictement postif tel que, si s est un réel supérieur ou égal à $\frac{1}{2}$, si les données v_0 et V sont telles que*

(i) $v_0\in H^{1/2}$ *et* $V\in L^2([0,+\infty[;H^{1/2})\cap L^2_{\mathrm{loc}}(]0,+\infty[;H^s)$,

(ii) $\|v_0|_{1/2}+\varepsilon^{-1/2}(\int_{[0,+\infty[}(\|V(t,\cdot)|_{1/2})^2\,dt)^{1/2}\leqq C_0\varepsilon$;

alors (S_ε) admet une unique solution dans $C([0,+\infty[;\,H^{1/2})\cap L^2([(0,+\infty[;\,H^{3/2})\cap C(]0,+\infty[;\,H^s)\cap L^2_{\mathrm{loc}}(]0,+\infty[;\,H^{s+1})$.

Remarques. Ce Théorème 0.1 a été démontré dans [3] par des méthodes de semi-groupes en supposant que V est une fonction Höldérienne de $[0,+\infty[$ à valeurs dans H^1 et en remplaçant, dans la condition (ii) du Théorème 0.2, la quantité $(\int_{[0,+\infty[}(\|V(t,\cdot)|_{1/2})^2\,dt)^{1/2}$ par $\mathrm{Sup}_{t\in[0,+\infty[}\,t^{3/4}\|V(t,\cdot)|_1$.

Enfin, observons que les quantités intervenant dans le membre de gauche de (ii) sont invariantes par changement d'échelle, ainsi que $\mathrm{Sup}_{t\in[0,+\infty[}\,t^{3/4}\|V(t,\cdot)|_1$.

Nous utiliserons dans tout ce travail les notations suivantes:

- Nous désignerons par ϕ une partition dyadique de l'unité, i.e., une fonction de $C_0^\infty(\mathbf{R}^d\backslash\{0\};[0,1])$ telle que, pour tout $\xi\in\mathbf{R}^d\backslash\{0\}$, on ait $1=\sum_{q\in\mathbf{Z}}\phi(2^{-q}\xi)$; de plus, on posera $\psi=1-\sum_{q\geqq 0}\phi(2^{-q}.)$; remarquons que ψ appartient à $C_0^\infty(\mathbf{R}^d)$. Nous désignerons par Δ_q l'opérateur $\phi(2^{-q}D)$, par S_q l'opérateur $\psi(2^{-q}D)$ et par $\underline{N}$ un entier tel que $0\notin\mathrm{Supp}\,\psi(2^{-\underline{N}}.)+\mathrm{Supp}\,\phi$;
- Si s est un réel, nous désignerons parΛ^s l'opérateur $(\mathrm{Id}-\Delta)^{s/2}$.

Enfin, nous allons rappeler la caractérisation des espaces de Sobolev homogènes et inhomogènes à l'aide du découpage dyadique de l'espace des fréquences. Il est clair que $C^{-1} \leqq \sum_{q \in \mathbf{Z}} (\phi(2^{-q}\xi))^2 \leqq C$ et que $C^{-1} \leqq (\psi(\xi))^2 + \sum_{q \geqq 0} (\phi(2^{-q}\xi))^2 \leqq C$; le fait que $C^{-1} \| \phi(2^{-q}D)u|_s \leqq 2^{qs} |\phi(2^{-q}D)u|_0 \leqq C \| \phi(2^{-q}D)u|_s$ assure

$$\begin{aligned} &C^{-1}(\|u|_s)^2 \leqq \sum_{q \in \mathbf{Z}} 2^{2qs}(|\phi(2^{-q}D)u|_0)^2 \leqq C(\|u|_s)^2, \\ (0.2) \qquad &C^{-1}(|u|_s)^2 \leqq \sum_{q \leqq 0} 2^{2qs}(|\phi(2^{-q}D)u|_0)^2 + (|\psi(D)u|_0)^2 \leqq C(|u|_s)^2. \end{aligned}$$

Pour des démonstrations complètes, nous renvoyons le lecteur à [1].

1. Inégalité d'énergie à perte; Application à la dimension deux d'espace. Nous allons énoncer et démontrer une inégalité d'énergie avec perte, en utilisant la décomposition de Bony d'un produit.

LEMME 1.1. *Soient v un champ de vecteurs à divergence nulle et s, r, r', r'' quatre réels tels que $r + r' + r'' = d/2 + 1 + 2s$, $r + r' > 0$ et $0 \leqq r < d/2 + 1$; alors il existe une constante C telle que*

(i) $(\Delta^{s/2}(v.\nabla a) | \Delta^{s/2}a) \leqq C|v|_r \|a|_{r'} \|a|_{r''} + |a|_r \|v|_{r'} \|a|_{r''}$,

(ii) $(\Lambda^s(v.\nabla a) | \Lambda^s a) \leqq C|v|_r |a|_{r'} a|_{r''} + |a|_r |v|_{r'} |a|_{r''}$.

Si, de plus $r' < d/2 + 1$, alors, nous avons,

(iii) $(\Lambda^s(v.\nabla a) | \Lambda^s a) \leqq C|v|_r |a|_{r'} |a|_{r''}$.

Démonstration. L'idée de base est d'écrire $(\Delta^{s/2}(v.\nabla a) | \Delta^{2/2}a)$ comme suit:

$$(1.1) \qquad (\Delta^{s/2}(v.\nabla a) | \Delta^{s/2}a) = \sum_{q,q',p,p;j} (\Delta^{s/2}\Delta_q(\Delta_p(v^j)\partial_j \Delta_{p'}(a)) | \Delta^{s/2}\Delta_{q'}(a)).$$

Il est bien clair, d'après la formule de Plancherel, qu'il existe un entier N tel que $|q - q'| \leqq N$. En utilisant la décomposition de Bony, il vient

$$\begin{aligned} &(\Delta^{s/2}(v.\nabla a) | \Delta^{s/2}a) = \sum_1^3 T^i, \quad \text{avec} \\ &T^1 = \sum_{q,q',p;j} (\Delta^{s/2}\Delta_q(S_{p-\underline{N}}(v^j)\partial_j \Delta_p(a)) | \Delta^{s/2}\Delta_{q'}(a)), \\ (1.2) \qquad &T^2 = \sum_{q,q',p;j} (\Delta^{s/2}\Delta_q(\Delta_p(v^j)\partial_j S_{p-\underline{N}}(a)) | \Delta^{s/2}\Delta_{q'}(a)), \\ &T^3 = \sum_{q,q',p,p',|p-p'| \leqq \underline{N};j} (\Delta^{s/2}\Delta_q(\Delta_p(v^j)\partial_j \Delta_{p'}(a)) | \Delta^{s/2}\Delta_{q'}(a)). \end{aligned}$$

Nous allons maintenant majorer chaque T^i. D'après la définition de T^1, des manipulations algébriques très simples et une intégration par parties assurent, grâce à la nullité de la divergence du champ de vecteurs v

$$\begin{aligned} &T^1 = \sum_{q,q',p;j} ([\Delta^{s/2}\Delta_q, S_{p-\underline{N}}(v^j)]\partial_j \Delta_p(a)) | \Delta^{s/2}\Delta_{q'}(a)) \\ (1.3) \qquad &+ \frac{1}{2} \sum_{q',p;j} ((S_{q'-\underline{N}} - S_{p-\underline{N}})(v^j)\partial_j \Delta^{s/2}\Delta_p(a)) | \Delta^{s/2}\Delta_{q'}(a)). \end{aligned}$$

Vu que $\|(S_{q'-\underline{N}} - S_{p-\underline{N}})(v^j)\|_{L^\infty} \leqq 2^{-p(r-d/2-1)}$, nous sommes donc ramenés à étudier la commutation entre l'opérateur de convolution $\Delta^{s/2}\Delta_q$ et la multiplication par $S_{p-\underline{N}}(v^j)$, ce qui se fait grâce à la formule suivante:

$$(1.4) \quad \Delta^{s/2}\Delta_q(S_{p-\underline{N}}(v^j)\partial_j \Delta_p(a)) = 2^{qd} \int h_s(2^q(x-y))(S_{p-\underline{N}}(v^j)\partial_j \Delta_p(a))(y)\, dy,$$

où h_s désigne la transformée inverse de Fourier de $|\xi|^s \phi(\xi)$.

Une formule de Taylor à l'ordre 1 avec reste intégral assure

$$
\begin{aligned}
\Delta^{s/2}\Delta_q(S_{p-\underline{N}}(v^j)\partial_j\Delta_p(a)) &= S_{p-\underline{N}}(v^j)\partial_j\Delta^{s/2}\Delta_q(\Delta_p(a)) \\
&\quad +2^{qd}\sum_{1\leqq i\leqq d}\int_{[0,1]}\int (x^i-y^i)h_s(2^q(x-y)) \\
&\quad \cdot S_{p-\underline{N}}(\partial_i v^j)(y+t(y-x))\partial_j\Delta_p(a)(y)\,dy\,dt.
\end{aligned}
\tag{1.5}
$$

Il ressort de la caractérisation des espaces H^s que, comme $r<d/2+1$, nous avons $\|S_{p-\underline{N}}(\partial_i v^j)\|_{L^\infty}\leqq C2^{-p(r-d/2-1)}|v|_r$. La définition de h et le fait que $|p-q|\leqq N'$ assure alors

$$
|[\Delta^{s/2}\Delta_q, S_{p-\underline{N}}(v^j)]\partial_j\Delta_p(a)|_0\leqq c_p 2^{-p(r+r'+d/2-1)}|v|_r|a|_{r'}\operatorname*{Sup}_{1\leqq i\leqq d}\|x^i h\|_{L^1},
\tag{1.6}
$$

avec $(c_p)_{p\in\mathbb{N}}\in 1^2(\mathbb{N})$. D'où $T^1\leqq C|v|_r\|a|_{r'}\|a|_{r''}$.

Comme $r<d/2+1$, $\|\partial_j S_{p-\underline{N}}(a)\|_{L^\infty}\leqq C2^{-p(r-d/2-1)}|a|_r$. Le fait que $T^2\leqq C|a|_r\|v|_{r'}\|a|_{r''}$ résulte alors simplement de l'existence d'un entier N' tel que $|p-q|$ et $|q-q'|$ soient majorés par N'.

La nullité de la divergence du champ de vecteurs v permet de se ramener à majorer $(\Delta_\mathrm{q}(\Delta_\mathrm{p}(\mathrm{v}^\mathrm{j})\Delta_{\mathrm{p}'}(\mathrm{a}))|\Delta_{\mathrm{q}'}(\mathrm{a}))$. Il est clair qu'il existe un entier N' tel que $|q-q'|$ soit majoré par N'; d'autre part, vu que $|p-p'|\leqq \underline{N}$, le support de la transformée de Fourier de $\Delta_p(v^j)\Delta_{p'}(a)$ est inclus dans une boule de rayon $\mathrm{C}2^\mathrm{p}$, il existe un entier $\underline{N}'$ tel que $p\geqq q-\underline{N}'$. Le fait que $|\Delta_p(v^j)\Delta_{p'}(a)|_0\leqq c_p c'_p 2^{-p(r+r'-d/2)}\|v|_r\|a|_{r'}$ et la relation entre r, r', r'' et s assurent alors

$$
T^3\leqq\sum_{p\geqq q-\underline{N}'} c_q 2^{(q-p)(r+r')}\|v|_r\|a|_{r'}\|a|_{r''},
\tag{1.7}
$$

où c_q est le terme général d'une série sommable; d'où le (i) du lemme, $r+r'$ étant supposé strictement postif.

La démonstration dans le cas des espaces de Sobolev inhomogènes, strictement analogue si l'on pose $\Delta_q=0$ si $q\leqq -2$ et $\Delta_q=\psi(D)$ si $q=-1$, est laissée au lecteur.

Remarque. Nous avons utilisé l'hypothèse de nullité de la divergence du champ de vecteurs v uniquement pour assurer le résultat lorsque $r+r'>0$. Il est clair, d'après la démonstration que le lemme est encore vrai sans hypothèse de nullité de la divergence du champ de vecteurs v en supposant alors $r+r'>1$.

LEMME 1.2. *Soit s un réel strictement supérieur à $d/2+1$, nous considèrons un potentiel-vecteur V dans $L^2_{\mathrm{loc}}([0,+\infty[;H^s)$ et v une solution du système (S_ε) qui soit dans $C([0,T^*[;H^s)$ pour un $s>d/2+1$, l'intervalle $[0,T^*[$ étant l'intervalle maximal de définition de la solution. Nous avons alors l'alternative suivante*:

$$
\textit{ou bien } T^*=+\infty,\quad \textit{ou bien } v\notin L^2([0,T^*[;H^{d/2}).
$$

Démonstration: Nous allons démontrer que, si la solution v appartient à $L^2([0,T^*[;H^{d/2})$; alors elle appartient à $L^\infty([0,T^*[;H^s)$, ce qui interdit à T^* d'être fini. En posant $g_s(t)=(\|v(t,\cdot)|_s)^2$, nous avons l'inégalité d'énergie

$$
\begin{aligned}
g'_s(t)+2\varepsilon(\|\nabla v(t,\cdot)|_s)^2 &= 2(\Delta^{s/2}v.\nabla v)(t,\cdot)|\Delta^{s/2}v(t,\cdot)) \\
&\quad +2(\Delta^{s/2}\nabla^\perp V(t,\cdot)|\Delta^{s/2}v(t,\cdot)).
\end{aligned}
\tag{1.8}
$$

En faisant une intégration par parties, il vient

$$
2(\Delta^{s/2}\nabla^\perp V(t,\cdot)|\Delta^{s/2}v(t,\cdot))\leqq 2\varepsilon^{-1}(\|V(t,\cdot)|_s)^2+\varepsilon/2(\|v(t,\cdot)|_{s+1})^2.
\tag{1.9}
$$

Puis, en appliquant le Lemme 1.1(i), avec $r=d/2$, $r'=s$ et $r''=s+1$, nous avons $2(\Delta^{s/2}v(t,\cdot))\leqq C(|v(t,\cdot)|_{d/2}\|v(t,\cdot)|_{s}\|v(t,\cdot)|_{s+1}$. Il vient alors

$$2(\Delta^{s/2}(v.\nabla v)(t,\cdot)|\Delta^{s/2}v(t,\cdot))\leqq(C^2/2\varepsilon)(|v(t,\cdot)|_{d/2})^2g_s(t)+\varepsilon(\|v(t,\cdot)|_{s+1})^2. \tag{1.10}$$

Il résulte alors de (1.9) que

$$g_s'(t)+\varepsilon(|v(t,\cdot)|_{s+1})^2\leqq((C^2/2\varepsilon)(|v(t,\cdot)|_{d/2})^2+\varepsilon/2)g_s+2\varepsilon^{-1}(\|V(t,\cdot)|_s)^2. \tag{1.11}$$

Une intégration standard entraine alors l'importante estimation d'énergie suivante:

$$\begin{aligned} g_s(t)\leqq g_s(0)\exp\left(\int_{[0,t]}((C^2/2\varepsilon)(|v(\tau,\cdot)|_{d/2})^2\,d\tau\right)\\ +2\varepsilon^{-1}\int_{[0,t]}(\|V(\tau,\cdot)|_s)^2\exp\left(\int_{[\tau,t]}((C^2/2\varepsilon)(|v(\dagger,\cdot)|_{d/2})^2\,d\dagger\right)d\tau. \end{aligned} \tag{1.12}$$

Cette estimation, jointe au fait que v appartienne à $L^2([0,T^*[;H^{d/2})$ et V à $L^2([0,T^*[;H^s)$ assure, avec l'aide de l'estimation (0.1), que v est dans $L^\infty([0,T^*[;H^s)$; d'où le lemme.

Remarques. Nous pouvons démontrer le même résultat en supposant que le potentiel-vecteur V appartient à $L^1_{\text{loc}}([0,+\infty[;H^{s+1})$.

L'estimation (1.12) ci-dessus est valable pour tout réel positif s.

2. Un lemme de produit; Application à un théorème d'unicité. Nous allons énoncer et démontrer un lemme de produit utilisant la décomposition de Bony d'un produit de deux distributions tempérées.

LEMME 2.1. (i) *Soit v un champ de vecteurs à divergence nulle, si r et r' sont deux réels tels que $r<d/2$, $r'<d/s+1$ et $r+r'\geqq d/2$; alors, nous avons*

$$|v.\nabla a|_{r+r'-d/2-1}\leqq C|v|_r|a|_{r'};$$

(ii) *Soit $r>1-d/2$, nous avons* $|ab|_{r-1}\leqq C(|a|_{d/2-1}|b|_r+|a|_r|b|_{d/2-1})$.

Démonstration. Nous utilisons la décomposition d'un produit en somme de deux paraproduits et d'un reste introduite par Bony dans [2]; c'est à dire que nous écrivons

$$\begin{aligned} v.\nabla a=\sum_{q\geqq 0;j}S_{q-N}(v^j)\partial_j\Delta_q(a)+\sum_{q\geqq 0;j}S_{q-N}(\partial_ja)\Delta_q(v^j)\\ +\sum_{q,q'\geqq 1,|q-q'|\leqq N;j}\Delta_{q'}(v^j)\partial_j\Delta_q(a). \end{aligned} \tag{2.1}$$

Le fait que la divergence du champ de vecteurs v soit nulle permet d'écrire

$$\begin{aligned} v.\nabla a=\sum_{q\geqq 0;j}S_{q-N}(v^j)\partial_j\Delta_q(a)+\sum_{q\geqq 0;j}S_{q-N}(\partial_ja)\Delta_q(v^j)\\ +\sum_{q,q'\geqq -1,|q-q'|\leqq N;j}\partial_j(\Delta_{q'}(v^j)\Delta_q(a)). \end{aligned} \tag{2.2}$$

La caractérisation des espaces de Sobolev inhomogènes donnée en (0.2) permet alors de conclure, une fois observé que, vue la localisation du support de la transformée de Fourier, nous avons $\|\Delta_q(a)\|_{L^\infty}\leqq C2^{qd/2}|\Delta_q(a)|_0$. La démonstration du point (ii) est analogue.

Nous pouvons maintenant énoncer et démontrer le théorème d'unicité suivant.

THÉORÈME 2.2. *Le système (S_ε) admet au plus une solution dans l'espace $L^\infty[(0,T[;H^{d/2-1})\cap L^2([(0,T[;H^{d/2})$.*

Démonstration. Nous procèdons de manière très classique en supposant l'existence de deux solutions v et w pour une même donnée initiale v_0 dans $H^{d/2-1}$, et nous estimons l'évolution en temps de la différence. Nous avons

$$\partial_t(v-w)+v.\nabla(v-w)=-\nabla p(v,w)+(v-w).\nabla w. \tag{2.3}$$

Posons $\delta_\lambda(t)=(|\psi(\lambda D)(v-w)(t,\cdot)|_{d/2-1})^2$, $\delta(t)=\operatorname{Sup}_{\tau\in[0,t]}(|(v-w)(t,\cdot)|_{d/2-1})^2$ et $\Delta(t)=\operatorname{Sup}_{\tau\in[0,t]}|v(\tau,\cdot)|_{d/2-1}$. Il résulte de (1.3), en appliquant l'opérateur $\psi(\lambda D)\Lambda^{d/2-1}$ et en faisant le produit scalaire avec $\psi(\lambda D)\Lambda^{d/2-1}(v-w)$ que

$$\begin{aligned}\delta'_\lambda(t)=&-2(\psi(\lambda D)\Lambda^{d/2-1}(v.\nabla(v-w))(t,\cdot)|\psi(\lambda D)\Lambda^{d/2-1}(v-w)(t,\cdot))\\&+2(\psi(\lambda D)\Lambda^{d/2-1}((v-w).\nabla w)(t,\cdot)|\psi(\lambda D)\Lambda^{d/2-1}(v-w)(t,\cdot))\\&-2\varepsilon(|\psi(\lambda D)\nabla(v-w)(t,\cdot)|_{d/2-1})^2.\end{aligned} \tag{2.4}$$

Le Lemme 1.1 appliqué avec $r=s=d/2$ et $r'=r''=d/2$ assure, pour presque-tout t,

$$\begin{aligned}&(\psi(\lambda D)\Lambda^{d/2-1}(v.\nabla(v-w))(t,\cdot)|\psi(\lambda D)\Lambda^{d/2-1}(v-w)(t,\cdot))\\&\leqq C|v(t,\cdot)|_{d/2}|(v-w)(t,\cdot)|_{d/2-1}|(v-w)(t,\cdot)|_{d/2}.\end{aligned} \tag{2.5}$$

D'après le Lemme de produit 2.1 (i), appliqué avec $r=d/2-1$ et $r'=d/2$, il vient

$$\begin{aligned}&(\psi(\lambda D)\Lambda^{d/2-1}((v-w).\nabla w)(t,\cdot)|\psi(\lambda D)\Lambda^{d/2-1}(v-w)(t,\cdot))\\&\leqq C|w(t,\cdot)|_{d/2}|(v-w)(t,\cdot)|_{d/2-1}|(v-w)(t,\cdot)|_{d/2}.\end{aligned} \tag{2.6}$$

Il résulte de (2.4-2.6) que nous avons, pour presque-tout t,

$$\begin{aligned}&\delta'_\lambda(t)+\varepsilon(|\psi(\lambda D)\nabla(v-w)(t,\cdot)|_{d/2-1})^2\\&\leqq C(|v(t,\cdot)|_{d/2}+|w(t,\cdot)|_{d/2})|(v-w)(t,\cdot)|_{d/2-1}|(v-w)(t,\cdot)|_{d/2}.\end{aligned} \tag{2.7}$$

Or, le fait que $|(v-w)(t,\cdot)|_{s+1}\leqq C(|\nabla(v-w)(t,\cdot)|_s+|(v-w)(t,\cdot)|_s)$ et que $2ab\leqq C^{-1}\varepsilon^{-1}a^2+C\varepsilon b^2$, assure que nous avons, pour presque-tout t,

$$\delta(t)\leqq((C^2/\varepsilon)\left(\int_{[0,t]}(|v(\tau,\cdot)_{,d/2})^2\,d\tau+\int_{[0,t]}(|w(\tau,\cdot|_{d/2})^2\,d\tau\right)+\varepsilon t)\,\delta(t). \tag{2.8}$$

Vu que v et w appartiennent à $L^2([0,T];H^{d/2})$, nous obtenons le résultat en prenant T assez petit.

On va maintenant aborder la démonstration du Théorème 0.1. Dans un premier temps, nous régularisons la donnée initiale v_0 et le potentiel-vecteur V en considérant une suite $(v_{0,n})_{n\in\mathrm{N}}$ (respectivement, $(V_n)_{n\in\mathrm{N}}$) de H^∞ (respectivement, de $L^2_{\mathrm{loc}}([0,+\infty[,H^\infty))$ telle que $v_{0,n}$ (respectivement, V_n) tende vers v_0 dans L^2 (respectivement, V dans $L^2_{\mathrm{loc}}([0,+\infty[,L^2))$.

Le théorème d'existence d'une solution locale en temps pour les systèmes hyperboliques symétriques s'appliquent ici. En effet, comme le terme de viscosité n'apparait pas grâce à son signe, l'estimation hyberbolique linéaire standard (voir, par exemple, [1]) assure, pour le linéarisé $L_vw=\partial_tw+v.\nabla w-\varepsilon\Delta w-\nabla\Delta^{-1}(\operatorname{tr}(dv\,dw))$, l'estimation suivante:

$$\begin{aligned}&\text{si }\lambda\geqq C_s\|v\|_{L^\infty([0,T],H^s)},\text{ pour tout }r>-s+d/2,\text{ si }w\text{ est assez régulière,}\\&\|e^{-\lambda t}w\|_{L^\infty([0,T],H^s)}\leqq|w(0)|_s+\|e^{-\lambda t}L_vw\|_{L^\infty([0,T],H^s)}.\end{aligned} \tag{2.9}$$

En suivant la démarche empruntée par Alinhac et Gérard dans [1], nous obtenons l'existence d'une solution dans $C([0,T];H^s)$ vérifiant l'estimation (2.9) pour le linéarisé. Nous utilisons alors un shéma itératif standard défini par $L_{v^n}v^{n+1}=\nabla^\perp V$ et

$v^{n+1}_{|t=0}=v_0$. L'estimation (2.9) entraine que, pour T assez petit, la suite $(v_n)_{n\in\mathbf{N}}$ est une suite bornée de $C([0, T]; H^s)$. En suivant toujours la démarche usuelle, nous observons que $L_{v^n}(v^{n+1}-v^n)=L_{v^n-v^{n+1}}v^{n+1}$, l'estimation (2.9) assurant alors que, pour t assez petit, la suite $(v_n)_{n\in\mathbf{N}}$ est une suite de Cauchy dans $C([0, T]; H^{s-1})$.

Nous disposons alors, pour tout n, d'une solution v_n définie a priori sur un intervalle de temps $[0, T_n[$. Le Lemme 1.2 assure que ces solutions existent globalement en temps; reste maintenant à démontrer, en s'inspirant de la démonstration du théorème d'unicité, que la suite $(v_n)_{n\in\mathbf{N}}$ est de Cauchy dans $L^\infty_{\text{loc}}([0, +\infty[;L^2)$. Cela va résulter du lemme suivant, valable en toute dimension.

LEMME 2.3. *Soient q un réel de l'intervalle $[d/2-1, +\infty[$, et T un réel strictement positif; si $(v_n)_{n\in\mathbf{N}}$ est une suite de solution du système (S_ε) sur l'intervalle $[0, T]$ pour les données $v_{0,n}$ et V_n telle que:*

(i) *la suite $(v_{0,n})_{n\in\mathbf{N}}$ soit de Cauchy dans l'espace H^q,*

(ii) *la suite $(V_n)_{n\in\mathbf{N}}$ soit de Cauchy dans l'espace $L^2([0, T];H^q)$,*

(iii) *la suite $(v_n)_{n\in\mathbf{N}}$ soit bornée dans l'espace $L^2([0, T];H^{d/2})$;*

alors, la suite $(v_n)_{n\in\mathbf{N}}$ est de Cauchy dans l'espace $C([0, T];H^q)$.

Démonstration. Nous allons étudier l'évolution en temps de $\delta_{n,m}(t)=(|v_n-v_m)(t,\cdot)|_q)^2$. Le système (S_ε) assure que

$$\begin{aligned}\delta'_{n,m}(t)=&-2(\Lambda^q(v_n.\nabla(v_n-v_m))(t,\cdot)|\Lambda^q(v_n-v_m)(t,\cdot))\\ &+2(\Lambda^q((v_n-v_m).\nabla v_m).\nabla v_m)(t,\cdot)|\Lambda^q(v_n-v_m)(t,\cdot))\\ &-2\varepsilon(|\nabla(v_n-v_m)(t,\cdot)|_q)^2+2(\Lambda^q\nabla^\perp(V_n-V_m)(t,\cdot)|\Lambda^q(v_n-v_m)(t,\cdot)).\end{aligned}\tag{2.10}$$

En appliquant le Lemme 1.1 (ii) avec $r=d/2$, $r'=q+1$ et $s=r''=q$, il vient, comme $q\geqq d/2-1$,

$$\begin{aligned}&-2(\Lambda^q(v_n.\nabla(v_n-v_m))(t,\cdot)|\Lambda^q(v_n-v_m)(t,\cdot))\\ &\qquad\leqq C|v_n(t,\cdot)|_{q+1}|(v_n-v_m)(t,\cdot)|_{q+1}|(v_n-v_m)(t,\cdot)|_q.\end{aligned}\tag{2.11}$$

Comme

$$\begin{aligned}&2(\Lambda^q((v_n-v_m).\nabla v_m)(t,\cdot)|\Lambda^q(v_n-v_m)(t,\cdot))\\ &\qquad=2(\Lambda^{q-1}((v_n-v_m).\nabla v_m)(t,\cdot)|\Lambda^{q+1}(v_n-v_m)(t,\cdot)),\end{aligned}$$

l'application du Lemme de produit 2.1 (ii) avec $r=q$, il vient, grâce au fait que q est plus grand que $d/2-1$,

$$\begin{aligned}&2(\Lambda^q((v_n-v_m).\nabla v_m)(t,\cdot)|\Lambda^q(v_n-v_m)(t,\cdot))\\ &\qquad\leqq C|v_m(t,\cdot)|_{q+1}|(v_n-v_m)(t,\cdot)|_q|(v_n-v_m)(t,\cdot)|_{q+1}.\end{aligned}\tag{2.12}$$

Il s'agit donc maintenant, d'après (2.11) et (2.12), de majorer la quantité

$$\Delta_{n,m}(t)=C(|v_n(t,\cdot)|_{q+1}+|v_m(t,\cdot)|_{q+1})|(v_n-v_m)(t,\cdot)|_{q+1}|(v_n-v_m)(t,\cdot)|_q.\tag{2.13}$$

En utilisant le fait que $|(v_n-v_m)(t,\cdot)|_{q+1}\leqq|\nabla(v_n-v_m)(t,\cdot)|_q+|(v_n-v_m)(t,\cdot)|_q$, nous obtenons

$$\begin{aligned}\Delta_{n,m}(t)\leqq\;& C(|v_n(t,\cdot)|_{q+1}+|v_m(t,\cdot)|_{q+1})(\varepsilon^{-1}(|v_n(t,\cdot)|_{q+1}+|v_m(t,\cdot)|_{q+1})+1)\delta_{n,m}(t)\\ &+(\varepsilon/2)(|\nabla(v_n-v_m)(t,\cdot)|_q)^2.\end{aligned}\tag{2.14}$$

Enfin, une intégration par parties standard assure

$$\begin{aligned}&2(\Lambda^q\nabla^\perp(V_n-V_m)(t,\cdot)|\Lambda^q(v_n-v_m)(t,\cdot))\\ &\qquad\leqq\varepsilon^{-1}(|(V_n-V_m)(t,\cdot)|_q)^2+\varepsilon(|\nabla(v_n-v_m)(t,\cdot)|_q)^2.\end{aligned}\tag{2.15}$$

Il résulte alors de (2.11-2.15) que l'on a

$$
\begin{aligned}
&\delta_{n,m}(t)+(\varepsilon/2)\int_{[0,t]}(|\nabla(v_n-v_m)(\tau,\cdot)|_q)^2\,d\tau\\
&\leqq\delta_{n,m}(0)\exp C(\varepsilon,T)\int_{[0,T]}((|v_m(t,\cdot)|_{q+1})^2+(|v_n(t,\cdot)|_{q+1})^2+1)\,d\tau\\
(2.16)\qquad&+\int_{[0,t]}\varepsilon^{-1}(|V_n-V_m)(t,\cdot)|_q)^2\exp(C(\varepsilon,T)\\
&\cdot\int_{[\tau,t]}(|v_m(\dagger,\cdot)|_{q+1})^2+(|v_n(\dagger,\cdot)|_{q+1})^2+1)\,d\dagger)\,d\tau,
\end{aligned}
$$

d'où le lemme.

Revenons à la démonstration du Théorème 0.1. Comme, d'après l'identité d'énergie (E), la suite $(v_n)_{n\in\mathbf{N}}$ est bornée dans $L^2([0,T[;H^1)$, le Lemme 2.3 assure le Théorème 0.1 pour $s=0$.

Pour démontrer le Théorème 0.1 dans tout les cas, on va procéder par une récurrence très simple; supposons que la solution v soit telle que $v\in C([0,+\infty[;H^r)\cap L^2_{\text{loc}}([0,+\infty[;H^{r+1})$, pour $r<s$. Pour tout réel α strictement positif, il existe un réel t_0 de l'intervalle $]0,\alpha[$ tel que $v(t_0,\cdot)$ appartienne à H^{r+1}. Nous considérons alors une suite $(v_n)_{n\in\mathbf{N}}$ de solution régularisée construite comme au début de la démonstration du cas $s=0$. L'estimation d'énergie (1.12) assure que la suite $(v_n)_{n\in\mathbf{N}}$ est bornée dans $L^2_{\text{loc}}([t_0,+\infty[;H^{\text{Inf}(s,r+1)+1})$. Le Lemme 2.3 alors le résultat.

3. Existence globale en dimension 3. Le but de cette partie est la démonstration du Théorème 0.2. Observons le comportement du système (S_ε) par changement d'échelle. Si v est une solution du système (S_ε) avec données v_0 et f; alors, pour tout réel λ strictement postif, $v_\lambda(t,x)=\lambda v(\lambda^2t,\lambda x)$ est solution du systéme (S_ε) avec données $v_{0,\lambda}=\lambda v_0(\lambda x)$ et $V_\lambda(t,x)=\lambda^2V(\lambda^2t,\lambda x)$. Nous avons très facilement les égalités suivantes:

(3.1)

$$|v_{0,\lambda}|_0=\lambda^{-1/2}|v_0|_0\quad\text{et}\quad\left(\int_{[0,+\infty[}(|V_\lambda(t,\cdot)|_0)^2\,dt\right)^{1/2}=\lambda^{-1/2}\left(\int_{[0,+\infty[}(|V(t,\cdot)|_0)^2\,dt\right)^{1/2},$$

$$\|v_{0,\lambda}|_{1/2}=\|v_0|_{1/2}\quad\text{et}\quad\left(\int_{[0,+\infty[}(\|V_\lambda(t,\cdot)|_{1/2})^2\,dt\right)^{1/2}=\left(\int_{[0,+\infty[}(\|V(t,\cdot)|_{1/2})^2\,dt\right)^{1/2}.$$

Il en résulte qu'il suffit de démontrer le Théorème 0.2 en supposant

$$|v_0|_{1/2}+\varepsilon^{-1/2}\left(\int_{[0,+\infty[}(|V(t,\cdot)|_{1/2})^2\,dt\right)^{1/2}\leqq C_0\varepsilon. \tag{3.2}$$

Nous procèdons comme pour la démonstration du Théorème 0.1; nous commencons par régulariser les données v_0 et V en considérant une suite $(v_{0,n})_{n\in\mathbf{N}}$ de H^∞ (respectivement, $(V_n)_{n\in\mathbf{N}}$) telle que $v_{0,n}$ tende vers v_0 dans $H^{1/2}$ (respectivement, V_n tende vers V dans $L^2([0,+\infty[;H^{1/2}))$ et que

$$|v_{0,n}|_{1/2}+\varepsilon^{-1/2}\left(\int_{[0,+\infty[}(|V_n(t,\cdot)|_{1/2})^2\,dt\right)^{1/2}\leqq C_0\varepsilon.$$

L'étape essentielle consiste en la démonstration d'estimations a priori sur les solutions v_n du système (S_ε) pour les données initiales $v_{0,n}$ et V_n assurant que ces solutions existent globalement en temps; il suffit ensuite d'appliquer le Lemme 2.3.

Nous allons désigner par v une solution quelconque du système (S_ε) avec données v_0 et V vérifiant (3.2) telles que v_0 soit H^∞ et V soit $L^2_{\text{loc}}([0,+\infty[;H^\infty)$. La solution v est définie sur l'intervalle maximal $[0, T^*[$. Etudions la fonction $g(t)=(\|v(t,\cdot)|_{1/2})^2+(|v(t,\cdot)|_0)^2$; nous avons

$$\begin{aligned}g'(t)&\leqq 2(\Delta^{1/4}(v.\nabla v)|\Delta^{1/4}v)-2\varepsilon(\Delta^{1/4}\nabla v(t,\cdot)|\Delta^{1/4}\nabla v(t,\cdot))\\&\quad+2(\Delta^{1/4}\nabla^\perp V(t,\cdot)|\Delta^{1/4}v(t,\cdot))\\&\quad-2\varepsilon(\nabla v(t,\cdot)|\nabla v(t,\cdot))+2(\nabla^\perp V(t,\cdot)|v(t,\cdot)).\end{aligned}\tag{3.3}$$

D'après le Lemme 1.1(i), il existe une constante $C>0$ telle que l'on ait:

$$2(\Delta^{1/4}(v.\nabla v)(t,\cdot)|\Lambda^{1/4}v(t,\cdot))\leqq C|v(t,\cdot)|_{1/2}(\|v(t,\cdot)|_{3/2})^2.\tag{3.4}$$

De plus, $2(\nabla^\perp V(\tau,\cdot)|v(\tau,\cdot))\leqq\varepsilon^{-1}(|V(\tau,\cdot)|_0)^2+\varepsilon(\|v(t,\cdot)|_1)^2$; de même, nous avons $2(\Delta^{1/4}\nabla^\perp V(t,\cdot)|\Delta^{1/4}v(t,\cdot))\leqq\varepsilon^{-1}(\|V(t,\cdot)|_{1/2})^2+\varepsilon(\|v(t,\cdot)|_{3/2})^2$; nous obtenons alors

$$g'(t)+\varepsilon(\|v(t,\cdot)|_{3/2})^2\leqq(C|v(t,\cdot)|_{1/2}-2\varepsilon)(\|v(t,\cdot)|_{3/2})^2+\varepsilon^{-1}(|V(t,\cdot)|_{1/2})^2.\tag{3.5}$$

Il en résulte que, tant que $|v(t,\cdot)|_{1/2}$ reste inférieur à ε/C, nous avons

$$g'(t)+\varepsilon(\|v(t,\cdot)|_{3/2})^2\leqq\varepsilon^{-1}(|V(t,\cdot)|_{1/2})^2.\tag{3.6}$$

Il en résulte, par une intégration immédiate, que, tant que $|v(t,\cdot)|_{1/2}$ reste inférieur à ε/C,

$$(|v(t,\cdot)|_{1/2})^2+\varepsilon\int_{[0,t]}(\|v(\tau,\cdot)|_{3/2})^2\,d\tau\leqq(|v_0|_{1/2})^2+\varepsilon^{-1}\int_{[0,+\infty[}(|V(\tau,\cdot)|^2_{1/2}\,d\tau.\tag{3.7}$$

Donc, si $|v_0|_{1/2}+\varepsilon^{-1/2}\int_{[0,+\infty[}(|V(\tau,\cdot)|^2_{1/2}\,d\tau)^{1/2}\leqq\varepsilon/2C$, alors l'estimation (3.7) est vraie sur tout l'intervalle $[0,T^*[$. Le lemme d'explosion 1.2 assure alors que $T^*=+\infty$. L'estimation (3.7) est donc valable sur tout $\mathbf{R}^+$. En utilisant une récurrence strictement analogue à celle concluant la démonstration du Théorème 0.1, nous obtenons le Théorème 0.2 via le Lemme 2.3.

REFERENCES

[1] S. Alinhac et P. Gérard, *Opérateurs pseudo-différentiels et théorème de Nash-Moser*, Orsay Publications Universitaires Scientifiques, Paris, 1989.

[2] J.-M. Bony, *Calcul symbolique et propagation des singularités dans les équations aux dérivées partielles non linéaires*, Ann. Sci. Ecole Norm. Sup. (4), 14 (1981), pp. 209-246.

[3] H. Fujita et T. Kato, *On the Navier-Stokes initial value problem* I, Arch. Rational Mech. Anal., 16 (1964), pp. 269-315.

[4] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1984.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 29-54, January 1992

003

ON UNIFORM DIFFERENCE SCHEMES FOR SECOND-ORDER SINGULAR PERTURBATION PROBLEMS IN BANACH SPACES[*]

A. ASHYRALYEV[†] AND H. O. FATTORINI[‡]

Abstract. The paper considers finite difference approximations of arbitrary order $O(\tau^n)$ to certain nonhomogeneous singular perturbation problems involving a small coefficient ε^2 in the higher derivative, where τ is the length of the discretization step in t and n is an arbitrary integer fixed in advance. Approximation is uniform with respect to ε.

Key words. singular perturbation, finite difference schemes

AMS(MOS) subject classifications. 65N05, 35R99

1. Introduction. We consider in this paper four singular perturbation problems involving second-order differential equations in a Banach space E. The equations are

$$\varepsilon^2 u''(t; \varepsilon) = Au(t, \varepsilon) + f(t), \tag{1.1}$$

$$\varepsilon^2 u''(t, \varepsilon) + u'(t, \varepsilon) = Au(t, \varepsilon) + f(t), \tag{1.2}$$

$$\varepsilon^2 u''(t, \varepsilon) - iu'(t, \varepsilon) = Au(t, \varepsilon) + f(t), \tag{1.3}$$

$$\varepsilon^2 u''(t, \varepsilon) + u'(t, \varepsilon) = (\varepsilon^2 A + B)u(t, \varepsilon) + f(t), \tag{1.4}$$

where A, B are linear, generally unbounded operators in E. In each case, the usual initial conditions are given:

$$u(0) = u_0, \qquad u'(0) = u_1. \tag{1.5}$$

The *singular perturbation problem* related to these equations is that of showing that

$$u(t, \varepsilon) \to u(t), \tag{1.6}$$

where $u(t)$ is the solution of the *limit equation* obtained by setting $\varepsilon = 0$ in (1.1), (1.2), (1.3), (1.4). The corresponding limit equations are

$$Au(t) + f(t) = 0, \tag{1.7}$$

$$u'(t) = Au(t) + f(t), \tag{1.8}$$

$$u'(t) = iAu(t) + if(t), \tag{1.9}$$

$$u'(t) = Bu(t) + f(t). \tag{1.10}$$

Equations (1.8), (1.9), and (1.10) come with the initial condition

$$u(0) = u_0, \tag{1.11}$$

while (1.7) is not associated with any initial condition. The singular perturbation problem associated with (1.1) is called *elliptic* since in typical applications A is a

[*] Received by the editors September 5, 1989; accepted for publication (in revised form) March 20, 1991.

[†] Turkmen State University, Department of Mathematics, Lenin Prospekt 31, Ashkhabad, 744000 Union of Soviet Socialist Republics. During the preparation of this paper, this author was a visitor at the University of California at Santa Barbara under the auspices of the International Research and Exchange Board.

[‡] University of California, Department of Mathematics, Los Angeles, California 90024. The work of this author was supported in part by National Science Foundation grant DMS-87011877.

(linear) elliptic operator. The singular perturbation problem associated with (1.2) is called *parabolic* [11], [14]. The *Schrödinger* perturbation problem is associated with (1.3), ([12], [14]) and the *hyperbolic* problem corresponds to (1.4) ([13], [14]); a reason for the latter terminology is that one of the main applications of (1.4) is to the hyperbolic partial differential equation $\varepsilon^2(u_{tt}-a(x)u_{xx})+(u_t+b(x)u_x)=f(x,t)$, which appears in problems of traffic flow (see [13]). The parabolic singular perturbation problem appears in the study of oscillations in viscous media and the Schrödinger singular perturbation problem in the proof that the nonrelativistic limit of the Klein-Gordon equation is the Schrödinger equation. For further details and bibliography on these applications, see [14], especially Chapters VI and VIII.

Equations like (1.1) could be handled as first-order systems in a suitable "phase space" as in [14, Chap. III], or [18], thus reducing the problem to one involving first-order rather than second-order equations. However, the fact that the limit equations of the second-order equations are of first order makes it simpler and more efficient to deal directly with the equations rather than with the corresponding systems. The same comment applies to (1.2), (1.3), and (1.4). Moreover, the finite difference schemes obtained by reduction to first-order systems are different than those implemented in the second-order equation.

A (strong) solution of any of the equations (1.1)-(1.4) is a function twice continuously differentiable in the norm of E such that $u(t)\in D(A)$ and the equation is satisfied everywhere. A minimal assumption on the initial value problem for any of the equations (1.1)-(1.4) for fixed ε is *well posedness*: there exists a dense subspace D of E such that (a) for every $u_0, u_1\in D$ the initial value problem has a strong solution $u(t;\varepsilon)$ in $t\geqq 0$, and (b) arbitrary strong solutions $u(t;\varepsilon)$ depend continuously on their initial data u_0, u_1 uniformly on compact subsets. In each of the four cases the equation is reduced by elementary transformations to the equation

$$v''(t)=Av(t); \tag{1.12}$$

thus the well-posedness requirement will be transferred to (1.12). It is known [23], [8] that the well-posedness condition is satisfied if and only if A is the infinitesimal generator of a strongly continuous cosine function $\{C(t);\ -\infty<t<\infty\}$. The solution of the inhomogeneous equation

$$v''(t)=Av(t)+g(t) \tag{1.13}$$

is given by

$$v(t)=C(t)v(0)+S(t)v'(0)+\int_0^t S(t-s)g(s)\,ds, \tag{1.14}$$

where

$$S(t)u=\int_0^t C(s)u\,ds. \tag{1.15}$$

The operator valued functions $C(\cdot)$, $S(\cdot)$ are called the *solution operators* of (1.12). In general, $v(t)$ is a strong solution only under suitable assumptions on $v(0)$, $v'(0)$, and $g(s)$ (for example, $v(0)$, $v'(0)\in D(A)$, $g(\cdot)$ is twice continuously differentiable). For arbitrary $v(0)$, $v'(0)$, and locally integrable $g(\cdot)$ the (strongly continuous) function (1.14) satisfies the equation (1.13) and the second initial condition (1.5) only in the sense of distributions and is called a *weak solution* (or simply a *solution*) of (1.12). It follows from a uniqueness argument that every solution of (1.13) admits the representation (1.14).

The solution operators $C(\cdot)$, $S(\cdot)$ can be extended to all t setting $C(t)=C(-t)$, $S(t)=-S(-t)$; we have $C(0)=I$, $S(0)=0$,

$$(1.16)\qquad \begin{aligned} &(S(t)u)'=C(t) \qquad (u\in E),\\ &(C(t)u)'=AS(t)u=S(t)Au \qquad (u\in D(A)),\end{aligned}$$

$$(1.17)\qquad C(s+t)+C(s-t)=2C(s)C(t),$$

$$(1.18)\qquad S(s+t)+S(s-t)=2S(s)C(t)$$

for $-\infty<s,\ t<\infty$ (the last two equations are called the *cosine* and *sine* functional equation, respectively). Finally, $C(\cdot)$ and $S(\cdot)$ grow exponentially at infinity:

$$(1.19)\qquad \|C(t)\|,\ \|S(t)\|\leqq Me^{\omega|t|} \qquad (-\infty<t<\infty)$$

for suitable M, ω. The resolvent $R(\lambda^2;\ A)=(\lambda^2I-A)^{-1}$ exists for $\operatorname{Re}\lambda>\omega$. (For proof of these and other properties, see [23] or [14].)

We consider in this paper discretizations (that is, approximation by finite difference schemes) of the initial value problems (1.1)-(1.4), which yield approximations to the solution $u(t,\ \varepsilon)$ of arbitrarily high order in the discretization step τ, uniformly in ε for $\varepsilon>0$. These discretizations are based on reduction of each equation to (1.13) and on the simplest difference scheme for this equation (introduced in [21]), which is examined in detail in § 3 and can be considered a direct generalization of the Courant-Friedrichs-Lewy difference scheme for the one-dimensional wave equation in [4]. This difference scheme is directly applied to (1.13) via Taylor expansions of the right-hand side about mesh points $k\tau$. Other treatments of finite difference schemes for second-order equations are in [3], [16], [20], and [22]. A general study of discretization schemes for linear convolution equations (possibly more general than differential) is in [10], although without rates of convergence.

Uniform difference schemes for ordinary differential equations were considered by numerous authors: see [19] and especially [5] and bibliography there. See also [7], [1], and [2] for a related problem for a partial differential equation. In this connection, we note that there is another equally natural way to approximate the solution of singular perturbation problems as $\varepsilon\to 0$, which is first to use asymptotic expansions and then to discretize them, as done in [17] and [6]. For a treatment of asymptotic expansions for operator equations see [11], [14] for the parabolic problem and [13] for the hyperbolic problem.

Since the elliptic singular perturbation problem is not included in [14] (and has probably not been studied in this level of generality) we include in § 2 a few simple convergence results, although convergence of $u(t,\varepsilon)$ to the solution of (1.7) is not essential (uniform approximation holds even in cases where $u(t,\ \varepsilon)$ is not convergent). A treatment of convergence of $u(t,\ \varepsilon)$ for (1.2) and (1.3) (respectively, (1.4)) can be found in [14] (respectively, [13]).

2. The elliptic singular perturbation problem. The initial value problem is

$$(2.1)\qquad \varepsilon^2u''(t;\ \varepsilon)=Au(t;\ \varepsilon)+f(t),$$

$$(2.2)\qquad u(0;\ \varepsilon)=u_0, \qquad u'(0;\ \varepsilon)=u_1.$$

The assumptions on A have been stated in § 1. We assume in addition that 0 belongs to the resolvent set $\rho(A)$ of A, that is, that A^{-1} exists and is bounded (this can always be insured by a translation). In this case, the solution $u(t)$ of the limit equation (1.7) is given by

$$(2.3)\qquad u(t)=-A^{-1}f(t).$$

Set $v(t)=u(\varepsilon t;\ \varepsilon)$. Then $v(\cdot)$ satisfies the initial value problem

$$v''(t)=Av(t)+f(\varepsilon t), \tag{2.4}$$

$$v(0)=u_0, \qquad v'(0)=\varepsilon u_1. \tag{2.5}$$

Expressing $v(t)$ by means of (1.14) we obtain the formula

$$u(t;\ \varepsilon)=C(t/\varepsilon)u_0+\varepsilon S(t/\varepsilon)u_1+\varepsilon^{-1}\int_0^t S((t-s)/\varepsilon)f(s)\,ds. \tag{2.6}$$

In general, $u(t,\ \varepsilon)$ does not converge as $\varepsilon\to 0$.

Using formulas (1.16) and assuming that $f(t)$ is continuously differentiable as many times as necessary we obtain, integrating by parts repeatedly,

$$\begin{aligned}\int_a^b S((b-s)/\varepsilon)f(s)\,ds=&-\sum_{j=1}^{m}\varepsilon^{2j-1}\{C((b-s)/\varepsilon)A^{-j}f^{(2j-2)}(s)\}\Big|_a^b\\ &-\sum_{j=1}^{m-1}\varepsilon^{2j}\{S((b-s)/\varepsilon)A^{-j}f^{(2j-1)}(s)\}\Big|_a^b\\ &+\varepsilon^{2m-1}\int_a^b C((b-s)/\varepsilon)A^{-m}f^{(2m-1)}(s)\,ds.\end{aligned} \tag{2.7}$$

Integrating by parts once again,

$$\begin{aligned}\int_a^b S((b-s)/\varepsilon)f(s)\,ds=&-\sum_{j=1}^{m}\varepsilon^{2j-1}\{C((b-s)/\varepsilon))A^{-j}f^{(2j-2)}(s)\}\Big|_a^b\\ &-\sum_{j=1}^{m}\varepsilon^{2j}\{S((b-s)/\varepsilon)A^{-j}f^{(2j-1)}(s)\Big|_a^b\\ &+\varepsilon^{2m}\int_a^b S((b-s)/\varepsilon)A^{-m}f^{(2m)}(s)\,ds.\end{aligned} \tag{2.8}$$

We justify these formulas under minimal hypotheses on f. Let $f(\cdot)$ be an E-valued function defined in $a\leqq t\leqq b$. Denote by $H^{1,p}(a,b;\ E)$ the space of all functions $f(\cdot)$ such that there exists $g(\cdot)\in L^p(a,b;\ E)$ and $u\in E$ such that

$$f(t)=u+\int_a^t g(s)\,ds \qquad (a\leqq t\leqq b).$$

Obviously, $f(\cdot)$ is absolutely continuous and has a derivative $f'(t)=g(t)$ almost everywhere.

For $r=1,2,\cdots$ we denote by $H^{r,p}(a,b;\ E)$ the space of all E-valued $r-1$ times continuously differentiable functions $u(\cdot)$ such that $u^{(r-1)}(\cdot)\in H^{1,p}(a,b;\ E)$.

Let ω be the constant in (1.19). It is shown in [8], [14] that if $\lambda\geqq\omega$ then fractional powers $(\lambda^2I-A)^\alpha$ can be defined for all α, $-\infty<\alpha<\infty$. These fractional powers satisfy the additivity property $(\lambda^2I-A)^{\alpha+\beta}=(\lambda^2I-A)^\alpha(\lambda^2I-A)^\beta$. Moreover, $(\lambda^2I-A)^{-\alpha}$ is bounded for $\lambda>\omega$ and $\alpha>0$ and for arbitrary λ, $\mu\geqq\omega$, $-\infty<\alpha<\infty$, $D((\lambda^2I-A)^\alpha)=D(\mu^2I-A)^\alpha)$ and $(\lambda^2I-A)^\alpha-(\mu^2I-A)^\alpha$ is bounded.

LEMMA 2.1. *Let $f(\cdot)$ be an E-valued function defined in $a\leqq t\leqq b$ and let $\lambda\geqq\omega$, $0\leqq\alpha\leqq 1$, m, r be positive integers. Then the following statements are equivalent*:

$$(A^{-m}(\lambda^2I-A)^{-\alpha}f)(\cdot)\in H^{r,1}(a,b;\ E), \tag{2.9}$$

$$((\lambda^2I-A)^{-m-\alpha}f)(\cdot)\in H^{r,1}(a,b;\ E) \tag{2.10}$$

and independent of λ for $\lambda>\omega$.

That (2.9) and (2.10) are equivalent follows from the equalities

$$\begin{aligned} A^{-m}(\lambda^2I-A)^{-\alpha}-(\lambda^2I-A)^{-m-\alpha} &= (A^{-m}-(\lambda^2I-A)^{-m})(\lambda^2I-A)^{-\alpha} \\ &= (I-A^{m}(\lambda^2I-A)^{-m})A^{-m}(\lambda^2I-A)^{-\alpha} \\ &= -(I-(\lambda^2I-A)^{m}A^{-m})(\lambda^2I-A)^{-m-\alpha} \end{aligned}$$

with both $(I-A^m(\lambda^2I-A)^{-m})$ and $(I-(\lambda^2I-A)^mA^{-m})$ bounded. As for independence of λ, we note that, for $\beta>0$ and $\lambda>\omega$,

$$(\mu^2I-A)^{-\alpha}=((\mu^2I-A)^{-\alpha}(\lambda^2I-A)^{\alpha})(\lambda^2I-A)^{-\alpha}$$

with $(\mu^2I-A)^{-\alpha}(\lambda^2I-A)^{\alpha}$ bounded (see [14]).

It is known ([7], [14, Chap. III]) that if $\lambda\geqq\omega$ and $\alpha<\frac{1}{2}$ then $S(t)E\subset D(((\lambda^2I-A)^{\alpha})$ and

$$t\to(\lambda^2I-A)^{\alpha}S(t) \tag{2.11}$$

is a strongly continuous function for all t. This may not be true for $\alpha=\frac{1}{2}$; for instance, if E is the space $C_0(-\infty,\infty)$ of all continuous functions on the line that tend to zero at $\pm\infty$ and $A=d^2/dx^2$ with maximal domain, A is the infinitesimal generator of the cosine function

$$C(t)u(x)=\{u(x+t)+u(x-t)\}/2$$

and (2.11) is not bounded for $\alpha=\frac{1}{2}$ (see [14]). However, (2.11) is strongly continuous for $\alpha=\frac{1}{2}$ if $E=L^p(\Xi,\Sigma,\mu)$ for $1<p<\infty$, (Ξ,Σ,μ) a measure space, in particular, if E is a Hilbert space.

LEMMA 2.2. (a) *Let $r\geqq1$ be odd and let $f(\cdot)$ be an E-valued function such that, for some $\lambda\geqq\omega$ and $\alpha<\frac{1}{2}$,*

$$((\lambda^2I-A)^{-j-\alpha}f)(\cdot)\in H^{2j,1}(a,b;E)\qquad(1\leqq j\leqq(r-1)/2), \tag{2.12}$$

$$(A^{-j}f)(\cdot)\in H^{2j-1,1}(a,b;E)\qquad(1\leqq j\leqq(r+1)/2). \tag{2.13}$$

Then

$$\begin{aligned} &\int_a^b S((b-s)/\varepsilon)f(s)\,ds \\ &\quad=-\sum_{j=1}^{(r+1)/2}\varepsilon^{2j-1}\{(A^{-j}f)^{(2j-2)}(b)-C((b-a)/\varepsilon))(A^{-j}f)^{(2j-2)}(a)\} \\ &\qquad+\sum_{j=1}^{(r-1)/2}\varepsilon^{2j}S((b-a)/\varepsilon)(A^{-j}f)^{(2j-1)}(a) \\ &\qquad+\varepsilon^{r}\int_a^b C((b-s)/\varepsilon)(A^{-(r+1)/2}f)^{(r)}(s)\,ds \\ &\quad=I_r(a,b,\varepsilon;f)+\varepsilon^{r}\int_a^b C((b-s)/\varepsilon)(A^{-(r+1)/2}f)^{(r)}(s)\,ds \end{aligned} \tag{2.14}$$

(r odd).

(b) *Let $r\geqq2$ be even and $f(\cdot)$ be an E-valued function such that, for some $\lambda\geq\omega$ and $\alpha<\frac{1}{2}$,*

$$((\lambda^2I-A)^{-j-\alpha})f(\cdot)\in H^{2j,1}(a,b;E)\qquad(1\leqq j\leqq r/2), \tag{2.15}$$

$$(A^{-j}f)(\cdot)\in H^{2j-1,1}(a,b;E)\qquad(1\leqq j\leqq r/2). \tag{2.16}$$

Then

$$\begin{aligned}\int_a^b & S((b-s)/\varepsilon)f(s)\,ds \\ &= -\sum_{j=1}^{r/2} \varepsilon^{2j-1}\{(A^{-j}f)^{(2j-2)}(b) - C((b-a)/\varepsilon))(A^{-j}f)^{(2j-2)}(a)\} \\ &\quad + \sum_{j=1}^{r/2} \varepsilon^{2j} S((b-a)/\varepsilon)(A^{-j}f)^{(2j-1)}(a) \\ &\quad + \varepsilon^r \int_a^b S((b-s)/\varepsilon)(A^{-r/2})f^{(r)}(s)\,ds \\ &= I_r(a, b, \varepsilon; f) + \varepsilon^r \int_a^b S((b-s)/\varepsilon)(A^{-r/2})f^{(r)}(s)\,ds \end{aligned} \tag{2.17}$$

$(r\ even)$.

If $E = L^p(\Xi, \Sigma, \mu)$ $(1 < p < \infty)$, (Ξ, Σ, μ) a measure space, we may take $\alpha = \frac{1}{2}$ in (2.12) and (2.15).

Proof. We extend $f(\cdot)$ setting $f(t) = f(a)$ for $t \leqq a$ and $f(t) = f(b)$ for $t \geqq b$ and define $f_n = \phi_n * f$, where $\{\phi_n\}$ is a δ-sequence of scalar test functions. Since f_n is infinitely differentiable, (2.7) can be applied. To take limits in the first summation, we observe that $(A^{-j}f_n)^{(2j-1)}(\cdot) \to (A^{-j}f)^{(2j-1)}(\cdot)$ in $L^1(a, b; E)$, so that $(A^{-j}f_n)^{(2j-2)}(\cdot) \to (A^{-j}f)^{(2j-2)}(\cdot)$ uniformly and, in particular, $(A^{-j}f_n)^{(2j-2)}(a) \to (A^{-j}f)^{(2j-2)}(a)$. For the second summation, we note that

$$\begin{aligned} & S((b-a)/\varepsilon)(A^{-j}f_n)^{(2j)}(s) \\ &\quad = ((\lambda^2 I - A)^\alpha S((b-a)/\varepsilon))(A^{-j}(\lambda^2 I - A)^{-\alpha} f_n)^{(2j)}(s), \end{aligned} \tag{2.18}$$

deducing that $S((b-a)/\varepsilon)(A^{-j}f_n)^{(2j)}(\cdot) \to S((b-a)/\varepsilon)(A^{-j}f)^{(2j)}(\cdot)$ in $L^1(a, b; E)$. Finally, we can take limits in the integral since, due to the last condition (2.12), $(A^{-(r+1)/2}f_n)^{(r)}(\cdot) \to (A^{-(r+1)/2}f)^{(r)}(\cdot)$ in the space $L^1(a, b; E)$.

The proof of (b) is similar. The treatment of the boundary terms is exactly the same. As for the integral, we write

$$\begin{aligned} & S((b-s)/\varepsilon)(A^{-r/2}f_n)^{(r)}(s) \\ &\quad = ((\lambda^2 I - A)^\alpha S((b-s)/\varepsilon))(A^{-r/2}(\lambda^2 I - A)^{-\alpha} f_n)^{(r)}(s) \end{aligned} \tag{2.19}$$

and use the last condition (2.16).

Example 2.3. Let $E = L^2(0, \pi)$, $A = d^2/dx^2$ with maximal domain determined by the boundary conditions

$$u(0) = u(\pi) = 0.$$

The spectral decomposition of A is

$$Au(x) = A \sum_{n=1}^{\infty} c_n \sin nx = -\sum_{n=1}^{\infty} n^2 c_n \sin nx \tag{2.20}$$

and that of the cosine function $C(t)$ generated by A is

$$C(t)u(x) = \{u(x+t) + u(x-t)\}/2 = \sum_{n=1}^{\infty} c_n \cos nt \sin nx. \tag{2.21}$$

Since we are in a Hilbert space and $\omega = 0$, we may use conditions (2.12) or (2.15) for $\lambda = 0$ and $\alpha = \frac{1}{2}$. An E-valued function

$$f(t) = \sum_{n=1}^{\infty} c_n(t) \sin nx \tag{2.22}$$

satisfies (2.12) and (2.13) if and only if all the coefficients $c_n(t)$ are $r-1$ times continuously differentiable with $c_n^{(r-1)}(t)$ absolutely continuous and such that the functions

$$\{\Sigma \, |n^{-2j-1} c_n^{(2j)}(t)|^2\}^{1/2} \qquad (1 \leqq j \leqq (r-1)/2), \tag{2.23}$$

$$\{\sum |n^{-2j} c_n^{(2j-1)}(t)|^2\}^{1/2} \qquad (1 \leqq j \leqq (r+1)/2) \tag{2.24}$$

belong to $L^1(a, b; E)$ or, equivalently, if all the functions

$$\{\sum |n^{-j-1} c_n^{(j)}(t)|^2\}^{1/2} \qquad (1 \leqq j \leqq r) \tag{2.25}$$

belong to $L^1(a, b; E)$. Conditions (2.15) and (2.16) reduce also to (2.25).

Remark 2.4. The assumption that $f(\cdot)$ is an E-valued function in Lemma 2.2 can be weakened: it is enough that, for some $\alpha < \frac{1}{2}$, $f(\cdot)$ satisfy (2.12) or (2.15) for $j = 0$. If $E = L^p(\Xi, \Sigma, \mu)(1 < p < \infty)$, we may take $\alpha = \frac{1}{2}$. In Example 2.3, this amounts to abandoning the requirement that $\{\sum |c_n(t)|^2\}^{1/2}$ be summable and require only that (2.25) hold for $j = 0$ as well.

Formulas (2.14) and (2.17) (used for $a = 0$, $b = t$) provide asymptotic series up to any power of ε for the solution $u(t, \varepsilon)$ of (2.1). For instance, formula (2.17) for $r = 2$ yields

$$\begin{aligned} u(t, \varepsilon) = {} & -A^{-1}f(t) + C(t/\varepsilon)(u_0 + A^{-1}f(0)) \\ & + \varepsilon S(t/\varepsilon)(u_1 + (A^{-1}f)'(0)) \\ & + \varepsilon \int_0^t S((t-s)/\varepsilon)(A^{-1}f)''(s)\, ds \end{aligned} \tag{2.26}$$

or, formula (2.14) with $r = 3$,

$$\begin{aligned} u(t, \varepsilon) = {} & -A^{-1}f(t) + C(t/\varepsilon)(u_0 + A^{-1}f(0)) \\ & + \varepsilon S(t/\varepsilon)(u_1 + (A^{-1}f)'(0)) \\ & - \varepsilon^2 (A^{-2}f)''(t) + \varepsilon^2 C(t/\varepsilon)(A^{-2}f)''(0) \\ & + \varepsilon^2 \int_0^t C((t-s)/\varepsilon)(A^{-2}f)'''(s)\, ds. \end{aligned} \tag{2.27}$$

These two formulas are used in the following result.

THEOREM 2.5. *Assume that*

$$\|C(t)\|, \|S(t)\| \leqq M \qquad (-\infty < t < \infty). \tag{2.28}$$

(a) *Let* $f(\cdot)$ *satisfy*

$$((\lambda^2 I - A)^{-1-\alpha})f(\cdot) \in H^{2,1}(0, T; E), \qquad (A^{-1}f)(\cdot) \in H^{1,1}(0, T; E), \tag{2.29}$$

for some $\alpha < \frac{1}{2}$. *Then*

$$u(t, \varepsilon) = -A^{-1}f(t) + C(t/\varepsilon)(u_0 + A^{-1}f(0)) + 0(\varepsilon) \qquad (\varepsilon \to 0) \tag{2.30}$$

uniformly in $0 \leqq t \leqq T$.

(b) *Let* (2.28) *hold, and assume in addition that*

$$(2.31)\qquad ((\lambda^2 I - A)^{-1-\alpha})f(\cdot) \in H^{2,1}(0, T; E), \qquad (A^{-2}f)(\cdot) \in H^{3,1}(0, T; E)$$

for some $\alpha < \frac{1}{2}$. *Then*

$$(2.32)\qquad \begin{aligned} u(t, \varepsilon) &= -A^{-1}f(t) + C(t/\varepsilon)(u_0 + A^{-1}f(0)) \\ &\quad + \varepsilon S(t/\varepsilon)(u_1 + (A^{-1}f)'(0)) + 0(\varepsilon^2) \qquad (\varepsilon \to 0) \end{aligned}$$

uniformly in $0 \leqq t \leqq T$.

(c) *If* $E = L^p(\Xi, \Sigma, \mu)$ $(1 < p < \infty)$, *then* (2.32) *is true under the only assumption that* $\alpha = \frac{1}{2}$ *in* (2.31).

(d) *If E is a Hilbert space then* (2.31) *is true under the only assumption that* $\alpha = \frac{1}{2}$ *in* (2.31).

The proof of Theorem 2.5 (a) and (b) follows from (2.26) and (2.27), respectively; we use the fact that the (strongly continuous operator valued) function

$$(2.33)\qquad (\lambda^2 I - A)^{\alpha} S(\cdot)$$

is uniformly bounded in $-\infty < s < \infty$, which follows from an analogue of formula (6.18) in [14] for $\alpha < \frac{1}{2}$. To show (c) we note that if $E = L^p(\Xi, \Sigma, \mu)$ $(1 < p < \infty)$, then (2.33) is a strongly continuous operator valued function in $-\infty < s < \infty$ also for $\alpha = \frac{1}{2}$. However, the hypotheses that $E = L^p(\Xi, \Sigma, \mu)$ does not insure uniform boundedness of (2.33) [14, Chap. III]; thus (2.30) does not follow. However, if E is a Hilbert space, the first condition (2.28) implies that

$$C(t) = K \cosh(tB) K^{-1},$$

where B is self-adjoint with $B \leqq 0$ and K is self-adjoint and invertible [9], [14]. Moreover, the second condition (2.28) implies that $B \leqq -\rho I$ with $\rho > 0$ [14, Chap. IV]. Thus, uniform boundedness of (2.33) for $\alpha = \frac{1}{2}$ can be directly proved using the functional calculus.

COROLLARY 2.6. *Assume that* (2.28) *holds and that* $(A^{-1}f)(\cdot) \in H^{1,1}(0, T; E)$. *Then*

$$(2.34)\qquad u(t, \varepsilon) = -A^{-1}f(t) + C(t/\varepsilon)(u_0 + A^{-1}f(0)) + 0(1) \qquad (\varepsilon \to 0)$$

uniformly in $0 \leqq t \leqq T$.

Proof. We use this time formula (2.14) for $r = 1$, which yields

$$(2.35)\qquad \begin{aligned} u(t, \varepsilon) &= -A^{-1}f(t) + C(t/\varepsilon)(u_0 + A^{-1}f(0)) \\ &\quad + \varepsilon S(t/\varepsilon)u_1 + \int_0^t C((t-s)/\varepsilon)(A^{-1}f)'(s)\, ds \end{aligned}$$

if $(A^{-1}f)(\cdot) \in H^{1,1}(0, T; E)$. We construct a sequence $\{f_n(\cdot)\}$, each f_n satisfying the conditions of Theorem 2.5 (a) (or, for that matter, smooth as in the proof of Lemma 2.2) and such that

$$\int_0^T \|(A^{-1}f)(s) - (A^{-1}f_n)(s)\|\, ds \to 0 \qquad (n \to \infty).$$

Now, let $u_n(t, \varepsilon)$ denote the solution of (2.1)-(2.2) with $f = f_n$ and choose $\delta > 0$. We select n so large that $\|u(t, \varepsilon) - u_n(t, \varepsilon)\| \leqq \delta/2$ uniform with respect to ε, and then let $\varepsilon \to 0$ making use of Theorem 2.5 (a).

Remark 2.7. Formulas such as (2.14) and (2.17) can be extended to the case where A does not have an inverse (this will be useful in § 7). The formula corresponding to (2.14) is.

$$\begin{aligned}\int_a^b & S((b-s)/\varepsilon)A^{(r+1)/2}f(s)\,ds \\ &= -\sum_{j=1}^{(r+1)/2} \varepsilon^{2j-1}\{(A^{(r+1-2j)/2}f)^{(2j-2)}(b) \\ &\qquad\qquad -C((b-a)/\varepsilon))(A^{(r+1-2j)/2}f)^{(2j-2)}(a)\} \\ &\quad + \sum_{j=1}^{(r-1)/2} \varepsilon^{2j}S((b-a)/\varepsilon)(A^{(r+1-2j)/2}f)^{(2j-1)}(a) \\ &\quad +\varepsilon^r\int_a^b C((b-s)/\varepsilon)f^{(r)}(s)\,ds,\end{aligned}\tag{2.36}$$

which is valid for any $f(\cdot)$ such that

$$f(t)\in D(A^{(r+1)/2}) \quad \text{a.e. in } a\leqq t\leqq b,\tag{2.37}$$

$$(A^{(r+1-2j)/2}f)(\cdot)\in H^{2j,1}(a, b; E) \qquad (0\leqq j\leqq (r+1)/2).\tag{2.38}$$

A similar extension can be made of (2.17):

$$\begin{aligned}\int_a^b & S((b-s)/\varepsilon)A^{r/2}f(s)\,ds \\ &= -\sum_{j=1}^{r/2} \varepsilon^{2j-1}\{(A^{(r-2j)/2}f)^{(2j-2)}(b) \\ &\qquad\qquad -C((b-a)/\varepsilon))(A^{(r-2j)/2}f)^{(2j-2)}(a)\} \\ &\quad + \sum_{j=1}^{r/2} \varepsilon^{2j}S((b-a)/\varepsilon)(A^{(r-2j)/2}f)^{(2j-1)}(a) \\ &\quad +\varepsilon^r\int_a^b S((b-s)/\varepsilon)f^{(r)}(s)\,ds\end{aligned}\tag{2.39}$$

under the hypotheses that

$$f(t)\in D(A^{r/2}) \quad \text{a.e. in } a\leqq t\leqq b,\tag{2.40}$$

$$(A^{(r-2j)/2}f)(\cdot)\in H^{2j,1}(a, b; E) \qquad (0\leqq j\leqq r/2).\tag{2.41}$$

3. The difference equation. We use for (2.1) the difference scheme

$$\varepsilon^2\tau^{-2}(u_{k+1}-2u_k+u_{k-1}) = 2(\tau/\varepsilon)^{-2}(C(\tau/\varepsilon)-I)u_k+f_k \qquad (k=1, 2, \cdots, N-1)\tag{3.1}$$

on (the interior points of) the grid $0, \tau, 2\tau, \cdots, N\tau = T$. We first treat the homogeneous case ($f_k=0$) and take $\varepsilon=1$, so that the difference equation is

$$u_{k+1}+u_{k-1} = 2C(\tau)u_k \qquad (k=1, 2, \cdots, N-1).\tag{3.2}$$

Following Piskarev [21] we define two operator valued solutions $\{C_k(\tau)\}$, $\{S_k(\tau)\}$ ($k\geqq 0$) of (3.2) specifying the initial conditions

$$C_0(\tau)=C_1(\tau)=I, \quad S_0(\tau)=0, \quad S_1(\tau)=I.\tag{3.3}$$

These solutions (called the *solution operators* of (3.2)) are uniquely defined and can be explicitly computed (recursively).

If $\{u_k\}$ is the solution of (3.2) satisfying the initial conditions

$$u_0 = \tilde{u}_0, \qquad \tau^{-1}(u_1 - u_0) = \tilde{u}_1, \tag{3.4}$$

then

$$u_k = C_k(\tau)\tilde{u}_0 + \tau S_k(\tau)\tilde{u}_1 \qquad (k = 0, 1, \cdots, N). \tag{3.5}$$

We check inductively using the cosine functional equation (1.17) that the operators $\{C_k(\tau)\}$, $\{S_k(\tau)\}$ are given by the following formulas:

$$\begin{aligned} C_k(\tau) &= 2C((k-1)\tau) - 2C((k-2)\tau) + \cdots + 2(-1)^{k+2}C(\tau) + (-1)^{k+1}I \\ &= (-1)^{k+1}I + 2(-1)^{k+1}\sum_{j=1}^{k-1}(-1)^j C(j\tau) \qquad (k \geqq 2), \end{aligned} \tag{3.6}$$

$$\begin{aligned} S_k(\tau) &= 2C((k-1)\tau) + 2C((k-3)\tau) + \cdots + 2C(3\tau) + 2C(\tau) \\ &= 2\sum_{j=0}^{(k-2)/2} C((2j+1)\tau) \qquad (k \geqq 2 \text{ even}), \end{aligned} \tag{3.7}$$

$$\begin{aligned} S_k(\tau) &= 2C((k-1)\tau) + 2C((k-3)\tau) + \cdots + 2C(2\tau) + I \\ &= I + 2\sum_{j=1}^{(k-1)/2} C(2j\tau) \qquad (k \geqq 3 \text{ odd}). \end{aligned} \tag{3.8}$$

(Formula (3.7) for $k = 2$ is $S_2(\tau) = 2C(\tau)$.) Thus, the following two estimates result from (1.19):

$$\|C_k(\tau)\| \leqq Mk\, e^{k\omega\tau}, \quad \|S_k(\tau)\| \leqq Mk\, e^{k\omega\tau} \qquad (k = 0, 1, \cdots, N). \tag{3.9}$$

The estimate for $\|C_k(\tau)\|$ falls short of establishing stability in the strict sense for solutions of (3.2) since the only bound on k is $k \leqq N = T/\tau$ and the right-hand side will tend to infinity as $\tau \to 0$; the corresponding estimate for $\|S_k(\tau)\|$ is sufficient since only $\tau S_k(\tau)$ is used and $k\tau \leqq T$. However, the first estimate (3.9) establishes "stability" in a sense acceptable in numerical analysis; see [15, p. 32].

We show below that (3.9) cannot in general be improved.

Example 3.1. We consider the operator A in Example 2.5. Formula (3.6) translates into

$$C_k(\tau) \sum_{n=1}^{\infty} c_n \sin nx = \sum_{n=1}^{\infty} F_n(k, \tau) c_n \sin nx, \tag{3.10}$$

where the functions $F_n(k, \tau)$ are given by

$$\begin{aligned} F_n(k, \tau) &= (-1)^{k+1}\left(1 + 2\sum_{j=1}^{k-1}(-1)^j \cos(jn\tau)\right) \\ &= \cos(k - \tfrac{1}{2})n\tau / \cos(n\tau/2), \end{aligned} \tag{3.11}$$

(see [24]) which is only $O(k)$ if $n\tau \approx \pi$.

We note that, in view of the characterization (2.21) of the cosine function $C(t)$, the difference equation (3.2) is the classical Courant-Friedrichs-Lewy difference scheme

$$u(x, (k+1)\tau) + u(x, (k-1)\tau) = u(x+\tau, k\tau) + u(x-\tau, k\tau) \tag{3.12}$$

(see [4]). For a stability analysis of (3.12) confirming the estimates (3.9) in this particular case, see [15, p. 31].

Remark 3.2. Formulas (3.6)-(3.8) for $\{C_k(\tau)\}$, $\{S_k(\tau)\}$ are not especially transparent. They become much more obvious when there exists a strongly continuous group $\{U(t); -\infty < t < \infty\}$ such that

$$C(t) = (U(t) + U(-t))/2. \tag{3.13}$$

Proceeding formally, we try

$$C_k(\tau) = AU(k\tau) + BU(-k\tau), \tag{3.14}$$

where A, B are unknown operator coefficients. Using the initial conditions $C_k(0) = C_k(\tau) = I$ we obtain the formal expression

$$\begin{aligned} C_k(\tau) &= U(k\tau)(I + U(\tau))^{-1} + U(-k\tau)(I + U(\tau))^{-1} \\ &= \{U(k\tau) + U((1-k)\tau)\}(I + U(\tau))^{-1} \\ &= \sum_{j=1-k}^{k-1} (-1)^{k-1-j} U(j\tau). \end{aligned} \tag{3.15}$$

In the same way we obtain

$$S_k(\tau) = -\{U((k+1)\tau) - U((1-k)\tau)\}(I - U(2\tau))^{-1} = \sum_{j=1-k}^{k-1} U(j\tau). \tag{3.16}$$

The final expressions for $C_k(\tau)$ and $S_k(\tau)$ are easily obtained from (3.15) and (3.16) and can be checked directly in (3.2).

We deal next with the inhomogeneous equation (3.1) for $\varepsilon = 1$:

$$u_{k+1} + u_{k-1} = 2C(\tau)u_k + \tau^2 f_k \qquad (k = 1, 2, \cdots, N-1). \tag{3.17}$$

Obviously, we may define f_0, f_N arbitrarily, since these values of f_k do not play a role in (3.17). We define $f_0 = 0$ and define a sequence $\{u_k\}$ starting with $u_0 = 0$ and continuing with

$$u_k = \tau^2 \sum_{j=0}^{k-1} S_{k-j}(\tau) f_j \qquad (k = 1, \cdots, N). \tag{3.18}$$

Obviously, we have $u_0 = 0$, $u_1 = 0$. Moreover,

$$u_2 + u_0 = u_2 = \tau^2 S_2(\tau) f_0 + \tau^2 f_1 = \tau^2 f_1 = 2C(\tau)u_1 + \tau^2 f_1,$$

and, due to the sine functional equation (1.18),

$$\begin{aligned} u_{k+1} + u_{k-1} &= \tau^2 \sum_{j=0}^{k} S_{k+1-j}(\tau) f_j + \tau^2 \sum_{j=0}^{k-2} S_{k-1-j}(\tau) f_j \\ &= \tau^2 f_k + \tau^2 \sum_{j=0}^{k-1} (S_{k+1-j}(\tau) + S_{k-1-j}(\tau)) f_j \\ &= \tau^2 f_k + 2C(\tau)\left(\tau^2 \sum_{j=0}^{k-1} S_{k-j}(\tau) f_j\right) \\ &= 2C(\tau)u_k + \tau^2 f_k \qquad (k \geqq 2). \end{aligned}$$

THEOREM 3.3. *The solution of the inhomogeneous equation* (3.17) *with initial conditions* (3.4) *is given by* $u_0 = \tilde{u}_0$,

$$u_k = C_k(\tau)\tilde{u}_0 + \tau S_k(\tau)\tilde{u}_1 + \tau^2 \sum_{j=0}^{k-1} S_{k-j}(\tau) f_j \qquad (k = 1, 2, \cdots, N) \tag{3.19}$$

and

$$\begin{aligned} \|u_k\| &\leqq Mk\, e^{k\omega\tau} \|\tilde{u}_0\| + Mk\tau\, e^{k\omega\tau} \|\tilde{u}_1\| \\ &\quad + M(T(T+\tau)/2)\, e^{k\omega\tau} \max_{1 \leqq j \leqq k-1} \|f_j\| \qquad (k = 0, 1, \cdots, N). \end{aligned} \tag{3.20}$$

Proof. Formula (3.19) has already been established. We use the second formula (3.9) to estimate:

$$\begin{aligned} \left\| \tau^2 \sum_{j=0}^{k-1} S_{k-j}(\tau) f_j \right\| &\leqq M\tau^2 \sum_{j=0}^{k-1} (k-j)\, e^{(k-j)\omega\tau} \|f_j\| \\ &\leqq \left(M\, e^{k\omega\tau} \tau^2 \sum_{j=1}^{k} j \right) \max_{1 \leqq j \leqq k-1} \|f_j\| \\ &\leqq M\, e^{k\omega\tau} \tau k \tau (k+1)/2 \max_{1 \leqq j \leqq k-1} \|f_j\| \\ &\leqq M(T(T+\tau)/2)\, e^{k\omega\tau} \max_{1 \leqq j \leqq k-1} \|f_j\|. \end{aligned} \tag{3.21}$$

This ends the proof of Theorem 3.3.

Equation (3.1) is reduced to (3.17), writing it in the form

$$\tau^{-2}(u_{k+1} - 2u_k + u_{k-1}) = 2\tau^{-2}(C(\tau/\varepsilon) - I)u_k + \varepsilon^{-2} f_k$$

or, equivalently,

$$u_{k+1} + u_{k-1} = 2C(\tau/\varepsilon)u_k + \tau^2(\varepsilon^{-2} f_k) \qquad (k = 1, 2, \cdots, N-1). \tag{3.22}$$

Using Theorem 3.3

$$u_k = C_k(\tau/\varepsilon)\tilde{u}_0 + \tau S_k(\tau/\varepsilon)\tilde{u}_1 + \tau^2 \sum_{j=0}^{k-1} S_{k-j}(\tau/\varepsilon)(\varepsilon^{-2} f_j); \tag{3.23}$$

thus

$$\begin{aligned} \|u_k\| &\leqq Mk\, e^{k\omega\tau/\varepsilon} \|\tilde{u}_0\| + Mk\, e^{k\omega\tau/\varepsilon} \|\tilde{u}_1\| \\ &\quad + M(T(T+\tau)/2)\, e^{k\omega\tau/\varepsilon} \max_{0 \leqq j \leqq k-1} \|\varepsilon^{-2} f_j\|. \end{aligned} \tag{3.24}$$

We shall use (3.24) for $\omega = 0$ (that is, we assume that (2.28) holds). In this case we obtain

$$\|u_k\| \leqq Mk \|\tilde{u}_0\| + Mk\tau \|\tilde{u}_1\| + M((T^2+1)/2) \max_{0 \leqq j \leqq k-1} \|\varepsilon^{-2} f_j\| \tag{3.25}$$

for $\tau \leqq 1/T$.

Remark 3.4. Discretization of the second-order equation $u''(t) = Au(t)$ by means of the difference scheme (3.2) runs into problems absent in the case of the first-order equation $u'(t) = Au(t)$. For this equation the corresponding difference scheme is

$$\tau^{-1}(u_{k+1} - u_k) = \tau^{-1}(S(\tau) - I)u_k$$

or

$$u_{k+1} = S(\tau)u_k \qquad (k = 0, 1, \cdots, N-1), \tag{3.26}$$

where $S(t)$ is the semigroup generated by A. The solution operator of (3.26) is $S_k(\tau)=S(k\tau)$, thus the difference equation is just as stable as the differential equation: if

$$(3.27) \qquad \|S(t)\| \leqq Me^{\omega t} \qquad (t \geqq 0)$$

then

$$(3.28) \qquad \|S_k(\tau)\| \leqq Me^{\omega k\tau} \qquad (k=1,2,\cdots).$$

There is no exact analogue for the second-order equation; the first bound (1.19) for the solution operators of the differential equation results in the weaker estimate (3.9) for the discrete propagator $C_k(\tau)$. As shown in Example 3.1, this is unavoidable.

Remark 3.5. We note another discrepancy between the first- and second-order case. The first-order singularly perturbed initial value problem is

$$(3.29) \qquad \varepsilon u'(t;\varepsilon)=Au(t;\varepsilon)+f(t),$$

$$(3.30) \qquad u(t;\varepsilon)=u_0 \qquad (t \geqq 0),$$

whose explicit solution is

$$(3.31) \qquad u(t;\varepsilon)=S(t/\varepsilon)u_0+\varepsilon^{-1}\int_0^t S((t-s)/\varepsilon)f(s)\,ds \qquad (t\geqq 0),$$

where $S(t)$ is the semigroup generated by A. The corresponding discretization of (3.29) is

$$(3.32) \qquad u_{k+1}=S(\tau/\varepsilon)u_k+\tau(\varepsilon^{-1}f_k).$$

The solution of (3.32) corresponding to the initial condition u_0 is

$$(3.33) \quad u_k=S(k\tau/\varepsilon)u_0+\tau\sum_{j=0}^{k-1}S((k-j-1)\tau/\varepsilon)(\varepsilon^{-1}f_j) \qquad (k=0,1,\cdots,N).$$

Assume that the semigroup $S(t)$ has negative exponential growth

$$(3.34) \qquad \|S(t)\| \leqq Me^{-\omega t} \qquad (t\geqq 0)$$

with $\omega>0$. Then we obtain from (3.31) and the fact that

$$\varepsilon^{-1}\int_0^\infty e^{-\omega s/\varepsilon}\,ds=\omega^{-1}$$

the estimate

$$(3.35) \qquad \|u(t)\| \leqq M\|u_0\|+\omega^{-1}M\max_{0\leqq t\leqq T}\|f(t)\|.$$

On the other hand, (3.33) yields the estimate

$$(3.36) \qquad \|u_k\| \leqq M\|u_0\|+CM\max_{0\leqq j\leqq n}\|f_j\|,$$

where C is the maximum of the function

$$\sigma\sum_{k=0}^{\infty}e^{-k\omega\sigma}=\sigma/(1-e^{-\omega\sigma})$$

in $0\leqq\sigma\leqq\infty$. Hence the a priori bounds for the differential equation and the difference equation are of the same stability type. The situation is different for the second-order equation (2.1) and its discretization (3.1). Since there are no cosine functions of negative exponential growth, on the one hand we cannot get rid of the factor ε^{-1} estimating

(2.6); on the other hand, the estimate (3.15) corresponding to the difference equation contains a factor ε^{-2} on the right-hand side instead of the factor ε^{-1} resulting from (2.5). This determines that we should prefer the differential equation to the difference equation in estimates for approximate solutions.

4. Higher-order estimates. We study in this section the approximation of the solution of the differential equation (2.1) by that of the difference equation (3.1). We assume that $\omega = 0$, that is, that (2.28) holds.

The method used below (and in the next sections) can be outlined as follows. All estimations will be based on formula (2.35) for solutions of the initial value problem (2.1)-(2.2), that is,

$$\begin{aligned} u(t, \varepsilon) &= -A^{-1}f(t) + C(t/\varepsilon)(u_0 + A^{-1}f(0)) \\ &\quad + \varepsilon S(t/\varepsilon)u_1 + \int_0^t C((t-s)/\varepsilon)(A^{-1}f)'(s)\, ds \end{aligned} \tag{4.1}$$

with $(A^{-1}f)(\cdot) \in H^{1,1}(0, T; E)$. Assume that $f_{\tau,n}(\cdot)$ is a function such that $(A^{-1}f_{\tau,n})(\cdot) \in H^{1,1}(0, T; E)$ as well, and that

$$\int_0^T \|(A^{-1}f)'(s) - (A^{-1}f_{\tau,n})'(s)\|\, ds \leqq C\tau^n. \tag{4.2}$$

Then, if $u_{\tau,n}(t; \varepsilon)$ is the solution of the initial value problem (2.1)-(2.2), with the same initial conditions and $f = f_{\tau,n}$ we have

$$\|u(t; \varepsilon) - u_{\tau,n}(t; \varepsilon)\| \leqq C\tau^n \qquad (0 \leqq t \leqq T). \tag{4.3}$$

We shall take the approximants $f_{\tau,n}$ to be piecewise polynomial functions in the intervals of a grid $0, \tau, 2\tau, \cdots, (N-1)\tau$, $N\tau = T$. We use Taylor polynomials in each interval $k\tau \leqq t < (k+1)\tau$, although (see Remark 4.6) other approximations can be used, for instance, interpolation polynomials of Newton-Cotes type. The particular form of the $f_{\tau,n}$ will make possible to obtain $u_{\tau,n}(t, \varepsilon)$ at the gridpoints $k\tau$ by means of the difference equation (3.1) with a particular right-hand side $\{f_k(\varepsilon)\}$ and initial conditions $u_0(\varepsilon)$, $u_1(\varepsilon)$ involving the values of $f(\cdot)$ and of its derivatives at gridpoints.

We begin by deducing this difference equation for a solution of the initial value problem (2.1)-(2.2) with an arbitrary right-hand side $f(\cdot)$. Let $t_k = k\tau$ for $k = 1, 2, \cdots, N$, with $N\tau = T$. We compute $u(t, \varepsilon)$ using (4.1) and replace the sequence $\{u(t_k, \varepsilon)\}$ in the difference equation (3.1) (or, equivalently, (3.22)). It follows from the cosine functional equation (1.17) and the sine functional equation (1.18) that both $\{C(t_k/\varepsilon)u_0\}$ and $\{S(t_k/\varepsilon)u_1\}$ satisfy the homogeneous difference equation. Accordingly, $\{u(t_k, \varepsilon)\}$ satisfies

$$\begin{aligned} &u(t_{k+1}, \varepsilon) + u(t_{k-1}, \varepsilon) - 2C(\tau/\varepsilon)u(t_k, \varepsilon) \\ &\quad = -A^{-1}f((k+1)\tau) + 2C(\tau/\varepsilon)f(k\tau) - A^{-1}f((k-1)\tau) \\ &\quad\quad + \int_0^{k\tau} \{C(((k+1)\tau - s)/\varepsilon) + C(((k-1)\tau - s)/\varepsilon) \\ &\quad\quad\quad -2C(\tau/\varepsilon)C((k\tau - s)/\varepsilon)\}(A^{-1}f)'(s)\, ds \\ &\quad\quad + \int_{k\tau}^{(k+1)\tau} C(((k+1)\tau - s)/\varepsilon)(A^{-1}f)'(s)\, ds \\ &\quad\quad - \int_{(k-1)\tau}^{k\tau} C(((k-1)\tau - s)/\varepsilon)(A^{-1}f)'(s)\, ds \\ &\quad = \tau^2(\varepsilon^{-2}f_k(\varepsilon)) \qquad (k = 1, 2, \cdots, N-1). \end{aligned} \tag{4.4}$$

Since the first integral vanishes due to the sine functional equation (1.18), we have

$$\begin{aligned} f_k(\varepsilon) = &-\varepsilon^2\tau^{-2}\{A^{-1}f((k+1)\tau)+2C(\tau/\varepsilon)A^{-1}f(k\tau)-A^{-1}f((k-1)\tau)\} \\ &+\varepsilon^2\tau^{-2}\int_{k\tau}^{(k+1)\tau} C(((k+1)\tau-s)/\varepsilon)(A^{-1}f)'(s)\,ds \\ &-\varepsilon^2\tau^{-2}\int_{(k-1)\tau}^{k\tau} C(((k-1)\tau-s)/\varepsilon)(A^{-1}f)'(s)\,ds. \end{aligned} \tag{4.5}$$

This solution satisfies the initial conditions

$$u(0,\varepsilon)=\tilde{u}_0=u_0, \tag{4.6}$$

$$\begin{aligned} \tau^{-1}(u(\tau,\varepsilon)-u(0,\varepsilon)) = &-A^{-1}f(\tau)+A^{-1}f(0) \\ &+\tau^{-1}(C(\tau/\varepsilon)-I)(u_0+A^{-1}f(0))+\varepsilon\tau^{-1}S(\tau/\varepsilon)u_1 \\ &+\tau^{-1}\int_0^\tau C((\tau-s)/\varepsilon)(A^{-1}f)'(s)\,ds=\tilde{u}_1(\varepsilon), \end{aligned} \tag{4.7}$$

where u_0, u_1 are the initial values (2.2) for the differential initial value problem.

Denote by $B^n(a, b; E)$ the space of all E-valued functions $g(\cdot)$ having derivatives of order $\leqq n$ with $g^{(n)}(s)$ bounded. We approximate $g(\cdot)$ by its Taylor polynomial of order $n-1$ in each interval $k\tau \leqq t \leqq (k+1)\tau$,

$$\begin{gathered} g_{\tau,n}(t)=\sum_{r=1}^{n-1}\frac{1}{r!}(t-k\tau)^r g^{(r)}(k\tau) \\ (k\tau\leqq t<(k+1)\tau,\ k=0,1,\cdots,N-1). \end{gathered} \tag{4.8}$$

It follows from the remainder formula applied in each interval that

$$\|g(t)-g_{\tau,n}(t)\|\leqq C\tau^n \qquad (0\leqq t\leqq T). \tag{4.9}$$

These considerations will be applied to $g(\cdot)=A^{-1}f(\cdot)$, where $f(\cdot)$ is the function on the right-hand side of (2.1). Assuming that $A^{-1}f(\cdot)\in B^n(a, b; E)$, we will have

$$\|(A^{-1}f)(t)-(A^{-1}f)_{\tau,n}(t)\|\leqq C\tau^n \tag{4.10}$$

so that (4.2), a fortiori (4.3), will hold. To figure out explicitly the difference equation satisfied by $\{u_{\tau,n}(t_k,\varepsilon)\}$ it is enough to compute $f_k(\varepsilon)$ given by (4.4) and $\tilde{u}_1(\varepsilon)$ given by (4.7). This amounts to finding explicit expressions for the operators

$$P_{r,b}(a,b,\varepsilon)u=\int_a^b C((b-s)/\varepsilon)u(s-a)^r\,ds \tag{4.11}$$

and

$$P_{r,a}(a,b,\varepsilon)u=\int_a^b C((s-a)/\varepsilon)u(s-a)^r\,ds \tag{4.12}$$

for arbitrary a, b, ε. For related purposes, we shall also make use of the operators

$$Q_{r,b}(a,b,\varepsilon)u=\int_a^b S((b-s)/\varepsilon)u(s-a)^r\,ds, \tag{4.13}$$

$$Q_{r,a}(a,b,\varepsilon)u=\int_a^b S((s-a)/\varepsilon)u(s-a)^r\,ds. \tag{4.14}$$

To compute $Q_{r,b}(a, b, \varepsilon)$ we use Lemma 2.2 for $f(s) = (s-a)^r$; for r even we use (2.14) for $r+1$ (so that the integral term drops out) and for r odd we use (2.17) for $r+1$. To compute the $Q_{r,a}(a, b, \varepsilon)$ we switch a and b in (2.14) and (2.17) and use $f(s) = (s-b)^r$. The $P_{r,b}$ and $P_{r,a}$ are reduced to the $Q_{r,a}$ and the $Q_{r,b}$ by means of an integration by parts. The first few $P_{r,a}$, $P_{r,b}$, $Q_{r,a}$, $Q_{r,b}$ are

$$P_{0,a}(a, b, \varepsilon) = P_{0,b}(a, b, \varepsilon) = \varepsilon S((b-a)/\varepsilon), \tag{4.15}$$

$$Q_{0,a}(a, b, \varepsilon) = Q_{0,b}(a, b; \varepsilon) = \varepsilon A^{-1}C((b-a)/\varepsilon) - \varepsilon A^{-1}, \tag{4.16}$$

$$\begin{aligned} P_{1,a}(a, b, \varepsilon) &= \varepsilon(b-a)S((b-a)/\varepsilon) - \varepsilon^2 A^{-1}C((b-a)/\varepsilon) + \varepsilon^2 A^{-1}, \\ P_{1,b}(a, b, \varepsilon) &= \varepsilon^2 A^{-1}C((b-a)/\varepsilon) - \varepsilon^2 A^{-1}, \end{aligned} \tag{4.17}$$

$$\begin{aligned} Q_{1,a}(a, b, \varepsilon) &= \varepsilon A^{-1}C((b-a)/\varepsilon) - \varepsilon^2 A^{-1}S((b-a)/\varepsilon), \\ Q_{1,b}(a, b, \varepsilon) &= -\varepsilon A^{-1}(b-a) + \varepsilon^2 A^{-1}S((b-a)/\varepsilon). \end{aligned} \tag{4.18}$$

We note that the computation of the $P_{r,a}$ can be reduced to the computation of the $P_{r,b}$ by writing $(s-a)^r = ((s-b)+(b-a))^r$ and using the binomial formula. The same observation applies to the $Q_{r,a}$ and $Q_{r,b}$.

THEOREM 4.1. *Assume the boundedness condition* (2.28) *is satisfied, and let* $u(t, \varepsilon)$ *be the solution of* (2.1)-(2.2) *with* $A^{-1}f(\cdot)$ *continuously differentiable and*

$$(A^{-1}f(\cdot))' \in B^n(0, T; E) \tag{4.19}$$

$(n \geqq 1)$. *Then the solution* $\{u_k(\varepsilon)\}$ *of the difference equation* (3.22) *with right-hand side*

$$\begin{aligned} f_k(\varepsilon) = {} & -\varepsilon^2\tau^{-2}\{A^{-1}f((k+1)\tau) + 2C(\tau/\varepsilon)A^{-1}f(k\tau) - A^{-1}f((k-1)\tau)\} \\ & + \varepsilon^2\tau^{-2}\sum_{r=0}^{n-1}\frac{1}{r!}P_{r,(k+1)\tau}(k\tau, (k+1)\tau, \varepsilon)(A^{-1}f)^{(r+1)}(\kappa\tau) \\ & - \varepsilon^2\tau^{-2}\sum_{r=0}^{n-1}\frac{1}{r!}P_{r,(k-1)\tau}((k-1)\tau, k\tau, \varepsilon)(A^{-1}f)^{(r+1)}((k-1)\tau) \end{aligned} \tag{4.20}$$

and initial conditions

$$u_0(\varepsilon) = \tilde{u}_0 = u_0, \tag{4.21}$$

$$\begin{aligned} \tau^{-1}(u_1(\varepsilon) - u_0(\varepsilon)) = \tilde{u}_1 = {} & -A^{-1}f(\tau) + A^{-1}f(0) \\ & + \tau^{-1}(C(\tau/\varepsilon) - I)(u_0 + A^{-1}f(0)) \\ & + \varepsilon\tau^{-1}S(\tau/\varepsilon)u_1 \\ & + \tau^{-1}\sum_{r=0}^{n-1}\frac{1}{r!}P_{r,\tau}(0, \tau, \varepsilon)(A^{-1}f)^{(r+1)}(0) \end{aligned} \tag{4.22}$$

satisfies

$$\|u_k(\varepsilon) - u(t_k, \varepsilon)\| = 0(\tau^n) \qquad (k = 1, 2, \cdots, N) \tag{4.23}$$

as $\tau \to 0$, *uniformly with respect to* ε.

Remark 4.2. A number of variants of Theorem 4.1 can be established using the same methods. For instance, approximations that satisfy

$$\|u_k(\varepsilon) - u(t_k, \varepsilon)\| = 0(\varepsilon\tau^n) \tag{4.24}$$

can be obtained by starting with formula (2.26) and requiring that $(A^{-1}f)''(\cdot) \in B^n(0, T; E)$. In this case, the $P_{r,a}$, $P_{r,b}$ are replaced by the $Q_{r,a}$, $Q_{r,b}$.

Remark 4.3. Since the right-hand sides of (4.15) and (4.16) contain the operators A^{-1}, $C(\tau/\varepsilon)$, and $S(\tau/\varepsilon)$, the finite difference schemes proposed in this section are practical only if these can be explicitly computed. In general, A^{-1}, $C(\tau/\varepsilon)$, and $S(\tau/\varepsilon)$ will have to be replaced by suitable approximations (in the case of a hyperbolic equation, approximations to $C(\tau/\varepsilon)$ and $S(\tau/\varepsilon)$ may be constructed discretizing the space variables), and the resulting scheme will be independent of ε inasmuch as these approximations are independent of ε. Thus, Theorem 4.1 provides only a framework or "template" for construction of practical discretizations. (The same applies to the difference schemes proposed in §§ 6–8.)

Example 4.4. To illustrate the application of the results to a particular example, we consider a Banach space analogue of the operator in Example 2.3. Let $E = C(0, \pi)$ be the space of all continuous functions in the interval $0 \leqq x \leqq \pi$ satisfying

$$u(0) = u(\pi) = 0 \tag{4.25}$$

endowed with the supremum norm, and let $A = d^2/dx^2$ with maximal domain. Then A generates a cosine function $C(t)$ given by

$$C(t)u(x) = (u(x+t) + u(x-t))/2, \tag{4.26}$$

where $u(x)$ has been continued to $-\infty < x < +\infty$ as a 2π-periodic function odd about $x = 0$ and $x = \pi$. We have

$$S(t)u(x) = \frac{1}{2}\int_{x-t}^{x+t} u(\xi)\, d\xi, \tag{4.27}$$

$$A^{-1}u(x) = \int_0^x (x-\xi)u(\xi)\, d\xi - \frac{x}{\pi}\int_0^\pi (\pi-\xi)u(\xi)\, d\xi. \tag{4.28}$$

Assume we wish to construct an approximation of order $0(\tau^2)$ to the solution of the initial value problem (2.1)–(2.2) by means of the difference scheme (3.22). Theorem 4.1 will hold if $f(t)(x) = f(x, t)$ is three times differentiable with respect to t uniformly with respect to x in $0 \leqq x \leqq \pi$. Taking advantage of periodicity, the integral (4.27) can be approximated independently of ε, while (4.28) does not depend on ε. Approximation of sufficiently high degree will of course need smoothness assumptions with respect to x on $u_0(x)$, $u_1(x)$.

Remark 4.5. Theorem 4.1 actually holds under milder hypotheses; for instance, an approximation of order $0(\tau)$ can be obtained under the only assumption that $(Af)'(t)$ is Lipschitz continuous.

Remark 4.6. Approximations different from the piecewise Taylor polynomial (4.8) can also be employed, and may in fact be of easier application in practice. For instance, we may use in each interval $k\tau \leqq t \leqq (k+1)\tau$ polynomials interpolating $f(t)$ at $k\tau$, $(k+1)\tau$ and $n-1$ interior points as in Newton-Cotes integration formulas. The advantage of this method is that evaluation of derivatives of $f(\cdot)$ is unnecessary. The expressions on the right-hand side of the difference equation (3.22) are computed using the repeated integration-by-parts formulas (2.14) and (2.17).

5. The parabolic singular perturbation problem. The assumptions on the initial value problem

$$\varepsilon^2 u''(t, \varepsilon) + u'(t, \varepsilon) = Au(t, \varepsilon) + f(t), \tag{5.1}$$

$$u(0, \varepsilon) = u_0, \qquad u'(0, \varepsilon) = u_1 \tag{5.2}$$

are those in § 1. We study (5.1)-(5.2) by reducing it to the elliptic singular perturbation problem. If $u(t, \varepsilon)$ is a strong solution of (5.1)-(5.2) and

$$v(t;\, \varepsilon)=e^{t/2\varepsilon^2}u(t, \varepsilon), \tag{5.3}$$

then $v(\cdot;\, \varepsilon)$ satisfies the initial value problem

$$\varepsilon^2 v''(t;\, \varepsilon)=(A+(1/4\varepsilon^2)I)v(t;\, \varepsilon)+e^{t/2\varepsilon^2}f(t), \tag{5.4}$$

$$v(0;\, \varepsilon)=u_0, \qquad v'(0;\, \varepsilon)=(1/2\varepsilon^2)u_0+u_1. \tag{5.5}$$

We treat (5.4)-(5.5) using the techniques of the elliptic singular perturbation problem. Denoting by $C(t;\, \varepsilon)$ the cosine function generated by $A+(1/4\varepsilon^2)I$, the difference equation associated with the differential equation (5.4) is

$$v_{k+1}+v_{k-1}=2C(\tau/\varepsilon;\, \varepsilon)v_k+\tau^2(\varepsilon^{-2}f_k). \tag{5.6}$$

Hence, the difference equation for

$$\{u_k\}=\{e^{-k\tau/2\varepsilon^2}v_k\} \tag{5.7}$$

is

$$e^{(k+1)\tau/2\varepsilon^2}u_{k+1}+e^{(k-1)\tau/2\varepsilon^2}u_{k-1}=2C(\tau/\varepsilon;\, \varepsilon)\, e^{k\tau/\varepsilon^2}u_k+\tau^2\, e^{k\tau/2\varepsilon^2}(\varepsilon^{-2}f_k). \tag{5.8}$$

This equation can be written in the form

$$\begin{aligned} &\tau^{-2}(u_{k+1}-2u_k+u_{k-1})+\tau^{-1}(1-e^{-\tau/\varepsilon^2})\tau^{-1}(u_k-u_{k-1}) \\ &\qquad =\tau^{-2}(2C(\tau/\varepsilon;\, \varepsilon)\, e^{-\tau/2\varepsilon^2}-(1+e^{-\tau/\varepsilon^2})I)u_k+\varepsilon^{-2}\, e^{-\tau/2\varepsilon^2}f_k, \end{aligned} \tag{5.9}$$

which shows better its relation with (5.1). Initial conditions are

$$u_0=\tilde{u}_0, \qquad \tau^{-1}(e^{\tau/2\varepsilon^2}u_1-u_0)=\tilde{u}_1. \tag{5.10}$$

To obtain estimates for $\{u_k\}$ we use the explicit solution formula (3.19) for $\{v_k\}$, denoting by $C_k(t/\varepsilon;\, \varepsilon)$, $S_k(t/\varepsilon;\, \varepsilon)$ the discrete propagators of the difference equation (5.6). The result is

$$\begin{aligned} u_k&=e^{-k\tau/2\varepsilon^2}C_k(\tau/\varepsilon;\, \varepsilon)\tilde{u}_0+e^{-k\tau/2\varepsilon^2}\tau S_k(\tau/\varepsilon;\, \varepsilon)\tilde{u}_1 \\ &\quad +\tau^2\sum_{j=0}^{k-1} e^{-(k-j)\tau/2\varepsilon^2}S_{k-j}(\tau/\varepsilon;\, \varepsilon)(\varepsilon^{-2}f_j), \end{aligned} \tag{5.11}$$

where $\tilde{u}_0$, $\tilde{u}_1$ are the initial conditions (5.10). To estimate this expression we use the bound

$$\|C(t;\, \varepsilon)\|\leqq M\exp{(\omega^2+1/4\varepsilon^2)^{1/2}t} \tag{5.12}$$

(see [14, Chap. VI]). Since

$$(\omega^2+1/4\varepsilon^2)^{1/2}=(1/2\varepsilon)(1+4\omega^2\varepsilon^2)^{1/2}\leqq(1/2\varepsilon)(1+2\omega^2\varepsilon^2) \tag{5.13}$$

we obtain

$$\begin{aligned} e^{-k\tau/2\varepsilon^2}\|C_k(\tau/\varepsilon;\, \varepsilon)\|&\leqq Mk\, e^{k\tau\omega^2}, \\ e^{-k\tau/2\varepsilon^2}\|S_k(\tau/\varepsilon;\, \varepsilon)\|&\leqq Mk\, e^{k\tau\omega^2}. \end{aligned} \tag{5.14}$$

Accordingly,

$$\begin{aligned} \tau^2\sum_{j=0}^{k-1} e^{-j\tau/2\varepsilon^2}\varepsilon^{-2}\|S_{k-j}(\tau/\varepsilon;\, \varepsilon)\|&\leqq \varepsilon^{-2}M\, e^{k\tau\omega^2}\tau^2\sum_{j=0}^{k-1}(k-j) \\ &\leqq \varepsilon^{-2}M\, e^{k\tau\omega^2}\tau k\tau(k+1)/2, \end{aligned}$$

which yields

$$\begin{aligned}\|u_k\| &\leqq Mk\, e^{k\tau\omega^2}\|\tilde{u}_0\| + Mk\tau\, e^{k\tau\omega^2}\|\tilde{u}_1\| \\ &+M((T(T+\tau)/2)\, e^{k\tau\omega^2} \max_{1\leqq j\leqq k-1} \|\varepsilon^{-2}f_j\|.\end{aligned} \tag{5.15}$$

6. Higher-order estimates for the parabolic problem. We study the approximation of the solution of the differential equation (5.1) by that of the difference equation (5.8).

Using (2.6) we obtain

$$\begin{aligned}u(t;\,\varepsilon) &= e^{-t/2\varepsilon^2}C(t/\varepsilon;\,\varepsilon)u_0 + e^{-t/2\varepsilon^2}S(t/\varepsilon;\,\varepsilon)((1/2\varepsilon)u_0+\varepsilon u_1) \\ &+\varepsilon^{-1}\int_0^t S((t-s)/\varepsilon;\,\varepsilon)\, e^{-(t-s)/2\varepsilon^2}f(s)\, ds.\end{aligned} \tag{6.1}$$

Using (5.12) and (5.13) we obtain

$$\begin{aligned}\|S(t;\,\varepsilon)\| &\leqq M(\omega^2+1/4\varepsilon^2)^{-1/2}(\exp(\omega^2+1/4\varepsilon^2)^{1/2}t-1) \\ &\leqq 2\varepsilon M \exp(\omega^2+1/4\varepsilon^2)^{1/2}t \leqq 2\varepsilon M \exp(1/2\varepsilon+\omega^2\varepsilon)t.\end{aligned} \tag{6.2}$$

Hence we deduce from (6.1) that

$$\|u(t;\,\varepsilon)\| \leqq 2M e^{\omega^2 t}(\|u_0\|+\varepsilon^2\|u_1\|)+2M e^{\omega^2 T}\int_0^T \|f(t)\|\, dt. \tag{6.3}$$

A slightly better estimation (with a factor M instead of $2M$) can be obtained by different means (see [14, Chap. VI]).

The arguments in this section are very much the same as those in § 4. Assuming that $f(\cdot)\in L^1(0,\, T;\, E)$ and $f_{\tau,n}(\cdot)\in L^1(0,\, T;\, E)$ is a function such that

$$\int_0^T \|f(s)-f_{\tau,n}(s)\|\, ds \leqq C\tau^n, \tag{6.4}$$

if $u_{\tau,n}(t,\,\varepsilon)$ is the solution of the initial value problem (5.1)-(5.2) with the same initial conditions and $f=f_{\tau,n}$ it follows from (6.3) that

$$\|u(t,\,\varepsilon)-u_{\tau,n}(t,\,\varepsilon)\| \leqq C\tau^n \qquad (0\leqq t\leqq T). \tag{6.5}$$

We deduce the difference equation for a solution $u(t,\,\varepsilon)$ of the initial value problem (5.1)-(5.2) with an arbitrary right-hand side $f(\cdot)$. Taking into account that $v(t;\,\varepsilon)$, given by (5.3), satisfies the differential equation (5.4), a computation similar to that in (4.4) (and the fact that $S(-t)=-S(t)$) reveals that $\{u(t_k,\,\varepsilon)\}$ is a solution of the difference equation

$$\begin{aligned}&e^{(k+1)\tau/2\varepsilon^2}u(t_{k+1},\,\varepsilon)+e^{(k-1)\tau/2\varepsilon^2}u(t_{k-1},\,\varepsilon)-2C(\tau/\varepsilon;\,\varepsilon)\, e^{k\tau/2\varepsilon^2}u(t_k,\,\varepsilon) \\ &=\varepsilon^{-1}\, e^{(k+1)\tau/2\varepsilon^2}\int_{k\tau}^{(k+1)\tau} S(((k+1)\tau-s)/\varepsilon;\,\varepsilon)\, e^{-((k+1)\tau-s)/2\varepsilon^2}f(s)\, ds \\ &-\varepsilon^{-1}\, e^{(k-1)\tau/2\varepsilon^2}\int_{(k-1)\tau}^{k\tau} S(((k-1)\tau-s)/\varepsilon;\,\varepsilon)\, e^{-((k-1)\tau-s)/2\varepsilon^2}f(s)\, ds \\ &=\varepsilon^{-1}\, e^{(k+1)\tau/2\varepsilon^2}\int_{k\tau}^{(k+1)\tau} S(((k+1)\tau-s)/\varepsilon;\,\varepsilon)\, e^{-((k+1)\tau-s)/2\varepsilon^2}f(s)\, ds \\ &+\varepsilon^{-1}\, e^{(k-1)\tau/2\varepsilon^2}\int_{(k-1)\tau}^{k\tau} S((s-(k-1)\tau)/\varepsilon;\,\varepsilon)\, e^{(s-(k-1)\tau)/2\varepsilon^2}f(s)\, ds.\end{aligned} \tag{6.6}$$

Initial conditions must be given in the form (5.10):

$$u(0, \varepsilon) = \tilde{u}_0 = u_0, \tag{6.7}$$

$$\begin{aligned} &\tau^{-1}(e^{\tau/2\varepsilon^2}u(\tau, \varepsilon) - u(0, \varepsilon)) \\ &\quad = \tau^{-1}(C(\tau/\varepsilon; \varepsilon) - I)u_0 + \tau^{-1}S(\tau/\varepsilon; \varepsilon)((1/2\varepsilon)u_0 + \varepsilon u_1) \\ &\quad + (\varepsilon\tau)^{-1}\, e^{\tau/2\varepsilon^2} \int_0^\tau S((\tau - s)/\varepsilon; \varepsilon)\, e^{-(\tau-s)/2\varepsilon^2} f(s)\, ds = \tilde{u}_1. \end{aligned} \tag{6.8}$$

We assume that $f(\cdot)$ belongs to $B_n(a, b; E)$ and use a piecewise Taylor polynomial approximation $f_{\tau,n}(\cdot)$ as in § 4:

$$\begin{gathered} f_{\tau,n}(t) = \sum_{r=1}^{n-1} \frac{1}{r!}(t - k\tau)^r f^{(r)}(k\tau) \\ (k\tau \leqq t < (k+1)\tau,\ k = 0, 1, \cdots, N-1), \end{gathered} \tag{6.9}$$

which yields the approximation

$$\|f_{\tau,}(t) - f_{\tau,n}(t)\| \leqq C\tau^n \qquad (0 \leqq t \leqq T). \tag{6.10}$$

Let $u_{\tau,n}(t, \varepsilon)$ be the solution of the initial value problem (5.1)-(5.2) with $f = f_{\tau,n}$. The function $u_{\tau,n}(t, \varepsilon)$ is obtained at gridpoints $k\tau$ by means of the difference equation (6.6) with initial conditions (6.7)-(6.8), where $f(\cdot)$ is replaced by $f_{\tau,n}(\cdot)$. To figure out explicitly all coefficients in terms of the values of $f(\cdot)$ and its derivatives at gridpoints it is enough to find explicit expressions for the operators

$$Q^p_{r,a}(a, b; \varepsilon) = \int_a^b S((s-a)/\varepsilon; \varepsilon)\, e^{(s-a)/2\varepsilon^2}(s-a)^r\, ds, \tag{6.11}$$

$$Q^p_{r,b}(a, b; \varepsilon) = \int_a^b S((b-s)/\varepsilon; \varepsilon)\, e^{-(b-s)/2\varepsilon^2}(s-a)^r\, ds. \tag{6.12}$$

Just as in the case of the operators $P_{r,a}$, $Q_{r,a}$, $P_{r,b}$, $Q_{r,b}$ in § 4, the computation of $Q^p_{r,a}$ reduces to that of $Q^p_{r,b}$, writing $(s-a)^r = ((s-b)+b-a))^r$ and applying the binomial formula. Thus we limit ourselves to the latter; we explicitly calculate $Q^p_{0,b}(a, b, \varepsilon)$ and $Q^p_{1,b}(a, b, \varepsilon)$ and give a recursion formula from which each $Q^p_{r,b}(a, b; \varepsilon)$ can be computed. For $r = 0$ we use (2.17) for $r = 2$ and $f(s) = e^{-(b-s)/2\varepsilon^2}u$, keeping in mind that $A + (1/4\varepsilon^2)I$ is the infinitesimal generator of $C(t; \varepsilon)$:

$$\begin{aligned} &\int_a^b S((b-s)/\varepsilon; \varepsilon)\, e^{-(b-s)/2\varepsilon^2}\, ds \\ &\quad = -\varepsilon(A + (1/4\varepsilon^2)I)^{-1} + \varepsilon C((b-a)/\varepsilon; \varepsilon)(A + (1/4\varepsilon^2)I)^{-1}\, e^{-(b-a)/2\varepsilon^2} \\ &\quad + (1/2)S((b-a)/\varepsilon; \varepsilon)(A + (1/4\varepsilon^2)I)^{-1}\, e^{-(b-a)/2\varepsilon^2} \\ &\quad + (1/4\varepsilon^2) \int_a^b S((b-s)/\varepsilon; \varepsilon)(A + (1/4\varepsilon^2)I)^{-1}\, e^{(b-s)/2\varepsilon^2}\, ds. \end{aligned}$$

Applying $A + (1/4\varepsilon^2)I$ to both sides, putting integrals on the left-hand side and then applying A^{-1}, we obtain

$$\begin{aligned} Q^p_{0,b}(a, b; \varepsilon) &= \int_a^b S((b-s)/\varepsilon; \varepsilon)\, e^{-(b-s)/2\varepsilon^2}\, ds \\ &= -\varepsilon A^{-1} + \varepsilon C((b-a)/\varepsilon; \varepsilon)A^{-1}\, e^{-(b-a)/2\varepsilon^2} \\ &\quad + (1/2)S((b-a)/\varepsilon; \varepsilon)A^{-1}\, e^{-(b-a)/2\varepsilon^2}. \end{aligned} \tag{6.13}$$

The computation of $Q^p_{1,b}$ is similar, with $f(s)=e^{-(b-s)/2\varepsilon^2}(s-a)u$:

$$\begin{aligned}\int_a^b & S((b-s)/\varepsilon;\,\varepsilon)\, e^{-(b-s)/2\varepsilon^2}(s-a)\,ds \\ &= -\varepsilon(A+(1/4\varepsilon^2)I)^{-1}(b-a)+\varepsilon^2 S((b-a)/\varepsilon;\,\varepsilon)(A+(1/4\varepsilon^2)I)^{-1}\,e^{-(b-a)/2\varepsilon^2} \\ &\quad +(1/4\varepsilon^2)\int_a^b S((b-s)/\varepsilon;\,\varepsilon)(A+(1/4\varepsilon^2)I)^{-1}\,e^{-(b-s)/2\varepsilon^2}(s-a)\,ds \\ &\quad +\int_a^b S((b-s)/\varepsilon;\,\varepsilon)(A+(1/4\varepsilon^2)I)^{-1}\,e^{-(b-s)/2\varepsilon^2}\,ds.\end{aligned}$$

This leads to the formula

$$Q^p_{1,b}(a,b,\varepsilon)=-\varepsilon A^{-1}(b-a)+\varepsilon^2 S((b-a)/\varepsilon)A^{-1}\,e^{-(b-a)/2\varepsilon^2}+A^{-1}Q^p_{0,b}(a,b,\varepsilon). \tag{6.14}$$

In the same way we obtain

$$\begin{aligned}\int_a^b & S((b-s)/\varepsilon;\,\varepsilon)\, e^{-(b-s)/2\varepsilon^2}(s-a)^r\,ds \\ &= -\varepsilon(A+(1/4\varepsilon^2)I)^{-1}(b-a)^r \\ &\quad +(1/4\varepsilon^2)\int_a^b S((b-s)/\varepsilon;\,\varepsilon)(A+(1/4\varepsilon^2)I)^{-1}\,e^{-(b-s)/2\varepsilon^2}(s-a)^r\,ds \\ &\quad +r\int_a^b S((b-s)/\varepsilon;\,\varepsilon)(A+(1/4\varepsilon^2)I)^{-1}\,e^{-(b-s)/2\varepsilon^2}(s-a)^{r-1}\,ds \\ &\quad +r(r-1)\varepsilon^2\int_a^b S((b-s)/\varepsilon;\,\varepsilon)(A+(1/4\varepsilon^2)I)^{-1}\,e^{-(b-s)/2\varepsilon^2}(s-a)^{r-2}\,ds.\end{aligned}$$

Hence

$$\begin{aligned}Q^p_{r,b}(a,b;\,\varepsilon)&=-\varepsilon A^{-1}(b-a)^r+rA^{-1}Q^p_{r-1,b}(a,b,\varepsilon) \\ &\quad +r(r-1)\varepsilon^2A^{-1}Q^p_{r-2,b}(a,b;\,\varepsilon),\end{aligned} \tag{6.15}$$

from which all the $Q^p_{r,b}$ can be calculated inductively.

THEOREM 6.1. *Let $u(t,\varepsilon)$ be the solution of* (5.1)-(5.2) *with $f(\cdot)\in B^n(a,b;E)$. Then, if $\{u_k(\varepsilon)\}$ is the solution of the difference equation* (5.8) *with right-hand side*

$$\begin{aligned}f_k(\varepsilon)&=\varepsilon\tau^{-2}\,e^{\tau/2\varepsilon^2}\sum_{r=0}^{n-1}\frac{1}{r!}\,Q^p_{r,(k+1)\tau}(k\tau,(k+1)\tau,\varepsilon)f^{(r)}(k\tau) \\ &\quad -\varepsilon\tau^{-2}\,e^{-\tau/2\varepsilon^2}\sum_{r=0}^{n-1}\frac{1}{r!}\,Q^p_{r,(k-1)\tau}((k-1)\tau,k\tau,\varepsilon)f^{(r)}((k-1)\tau)\end{aligned} \tag{6.16}$$

and initial conditions

$$u_0(\varepsilon)=u_0, \tag{6.17}$$

$$\begin{aligned}\tau^{-1}(e^{\tau/2\varepsilon^2}u_1(\varepsilon)-u_0(\varepsilon))&=\tau^{-1}(C(\tau/\varepsilon;\,\varepsilon)-I)u_0+\tau^{-1}S(\tau/\varepsilon;\,\varepsilon)((1/2\varepsilon)u_0+\varepsilon u_1) \\ &\quad +(\varepsilon\tau)^{-1}\,e^{\tau/2\varepsilon^2}\sum_{r=0}^{n-1}\frac{1}{r!}\,Q^p_{r,\tau}(0,\tau,\varepsilon)f^{(r)}(0).\end{aligned} \tag{6.18}$$

Then $\{u_k(\varepsilon)\}$ satisfies

$$\|u_k(\varepsilon)-u(t_k;\,\varepsilon)\|=0(\tau^n)\qquad(k=1,2,\cdots,N) \tag{6.19}$$

as $\tau\to 0$, uniformly with respect to ε.

7. The Schrödinger singular perturbation problem. The treatment of the initial value problem

$$\varepsilon^2 u''(t, \varepsilon) - iu'(t, \varepsilon) = Au(t, \varepsilon) + f(t), \tag{7.1}$$

$$u(0, \varepsilon) = u_0, \qquad u'(0, \varepsilon) = u_1 \tag{7.2}$$

is similar to that of (5.1)-(5.2). Formally, the initial value problem can be reduced to (5.1)-(5.2) as follows: if $u(t, \varepsilon)$ is a solution of (7.1)-(7.2), then the function $\tilde{u}(t, \varepsilon)$ "defined" by $\tilde{u}(t, \varepsilon) = u(-it, i\varepsilon)$ is a solution of the equation $\varepsilon^2 \tilde{u}''(t, \varepsilon) + \tilde{u}'(t, \varepsilon) = A\tilde{u}(t, \varepsilon) + f(-it)$ with initial conditions $u(0, \varepsilon) = u_0$, $\tilde{u}'(0, \varepsilon) = -iu_1$. Using this formal correspondence, all formulas in §§ 5 and 6 have analogues here: of course, independent proofs must be provided.

Assuming that $u(t, \varepsilon)$ is a solution of the initial value problem (7.1)-(7.2), we set

$$v(t; \varepsilon) = e^{-it/2\varepsilon^2} u(t, \varepsilon). \tag{7.3}$$

Then $v(\cdot; \varepsilon)$ satisfies the initial value problem

$$v''(t; \varepsilon) = (A - (1/4\varepsilon^2)I)v(t; \varepsilon) + e^{-it/2\varepsilon^2} f(t; \varepsilon), \tag{7.4}$$

$$v(0; \varepsilon) = u_0, \qquad v'(0; \varepsilon) = (-i/2\varepsilon^2)u_0 + u_1. \tag{7.5}$$

Denoting by $C(t; i\varepsilon)$ the cosine function generated by $A - (1/4\varepsilon^2)I$, the difference equation associated to (7.4) is

$$v_{k+1} + v_{k-1} = 2C(\tau/\varepsilon; i\varepsilon)v_k + \tau^2(\varepsilon^{-2} f_k).$$

Thus, the difference equation for $\{u_k\} = \{e^{ik\tau/2\varepsilon^2} v_k\}$ is

$$\begin{aligned} &e^{-i(k+1)\tau/2\varepsilon^2} u_{k+1} + e^{-i(k-1)\tau/2\varepsilon^2} u_{k-1} \\ &\quad = 2\tau^{-2} C(\tau/\varepsilon; i\varepsilon)\, e^{-ik\tau/\varepsilon^2} u_k + \tau^2\, e^{-ik\tau/2\varepsilon^2}(\varepsilon^{-2} f_k), \end{aligned} \tag{7.6}$$

which can be rewritten in the form

$$\begin{aligned} &\tau^{-2}(u_{k+1} - 2u_k + u_{k-1}) + \tau^{-1}(1 - e^{i\tau/\varepsilon^2})\tau^{-1}(u_k - u_{k-1}) \\ &\quad = \tau^{-2}(2C(\tau/\varepsilon; i\varepsilon)\, e^{i\tau/2\varepsilon^2} - (1 + e^{i\tau/\varepsilon^2})I) + \varepsilon^{-2}\, e^{i\tau/2\varepsilon^2} f_k, \end{aligned} \tag{7.7}$$

showing its direct relationship with (7.1). Initial conditions are

$$u_0 = \tilde{u}_0, \qquad \tau^{-1}(e^{-i\tau/2\varepsilon^2} u_1 - u_0) = \tilde{u}_1. \tag{7.8}$$

Estimates for the solution of (7.6)-(7.8) can be obtained directly from the explicit solution,

$$\begin{aligned} u_k = {} & e^{ik\tau/2\varepsilon^2} C_k(\tau/\varepsilon; i\varepsilon)\tilde{u}_0 + e^{ik\tau/2\varepsilon^2} \tau S_k(\tau/\varepsilon; i\varepsilon)\tilde{u}_1 \\ & + \tau^2 \sum_{j=0}^{k-1} e^{i(k-j)\tau/\varepsilon^2} S_{k-j}(\tau/\varepsilon; i\varepsilon)(\varepsilon^{-2} f_j). \end{aligned} \tag{7.9}$$

We shall estimate this expression in the setting used in [12] (see also [14, Chap. VII]) for treatment of the Schrödinger singular perturbation problem; we assume that E is a Hilbert space and that

$$A = A_0 + B, \tag{7.10}$$

where A_0 is a self-adjoint operator bounded above and B is a bounded operator. In this situation there exist constants M, ω such that

$$\|C(t; i\varepsilon)\| \leqq M e^{\omega\varepsilon|t|} \qquad (-\infty < t < \infty), \tag{7.11}$$

$$\|S(t; i\varepsilon)\| \leqq M\varepsilon\, e^{\omega\varepsilon|t|} \qquad (-\infty < t < \infty) \tag{7.12}$$

(see [12] or [14, p. 242]). Using (7.11) and formulas (3.6), (3.7), and (3.8) for the discrete propagators we obtain

$$\|C_k(\tau/\varepsilon;\, i\varepsilon)\| \leqq Mk\, e^{\omega k\tau}, \qquad \|S_k(\tau/\varepsilon;\, i\varepsilon)\| \leqq Mk\, e^{\omega k\tau} \tag{7.13}$$

and we estimate (7.9) in the same way as (5.11):

$$\|u_k\| \leqq Mk\, e^{k\tau\omega}\|\tilde{u}_0\| + Mk\tau\, e^{k\tau\omega}\|\tilde{u}_1\| + M(T(T+\tau)/2)\, e^{k\tau\omega} \max_{1\leq j\leq n} \|\varepsilon^{-2} f_j\|. \tag{7.14}$$

The estimates below are similar to those for the parabolic problem. We begin by writing the solution of (7.4) using (7.3) and (2.6):

$$\begin{aligned} u(t;\, \varepsilon) &= e^{it/2\varepsilon^2} C(t/\varepsilon;\, i\varepsilon)u_0 + e^{it/2\varepsilon^2} S(t/\varepsilon;\, i\varepsilon)((-i/2\varepsilon)u_0 + \varepsilon u_1) \\ &\quad + \varepsilon^{-1} \int_0^t S((t-s)/\varepsilon;\, i\varepsilon)\, e^{i(t-s)/2\varepsilon^2} f(s)\, ds. \end{aligned} \tag{7.15}$$

Using (7.11) and (7.12) we obtain the analogue of (6.3),

$$\|u(t;\, \varepsilon)\| \leqq 2M e^{\omega t}(\|u_0\| + \varepsilon^2\|u_1\|) + 2M e^{\omega T} \int_0^T \|f(t)\|\, dt. \tag{7.16}$$

We proceed exactly as in §§ 5 and 6. Assuming that $f(\cdot) \in L^1(0, T;\, E)$ and $f_{\tau,n}(\cdot) \in L^1(0, T;\, E)$ is a function such that (6.4) holds, then, if $u_{\tau,n}(t, \varepsilon)$ is the solution of the initial value problem (7.1)-(7.2) with the same initial conditions and $f = f_{\tau,n}$, it follows from (7.16) that (6.5) holds.

We deduce the difference equation for a solution $u(t, \varepsilon)$ of the initial value problem (7.1)-(7.2) with an arbitrary right-hand side $f(\cdot)$. Taking into account that $v(t, \varepsilon)$ given by (7.3) satisfies the differential equation (7.4), a computation similar to (6.6) shows that $\{u(t_k;\, \varepsilon)\}$ solves

$$\begin{aligned} & e^{-i(k+1)\tau/2\varepsilon^2} u(t_{k+1}, \varepsilon) + e^{-i(k-1)\tau/2\varepsilon^2} u(t_{k-1}, \varepsilon) - 2C(\tau/\varepsilon;\, i\varepsilon)\, e^{-ik\tau/\varepsilon^2} u(t_k, \varepsilon) \\ & + \varepsilon^{-1}\, e^{-i(k+1)\tau/2\varepsilon^2} \int_{k\tau}^{(k+1)\tau} S(((k+1)\tau - s)/\varepsilon;\, i\varepsilon)\, e^{i((k+1)\tau - s)/2\varepsilon^2} f(s)\, ds \\ & + \varepsilon^{-1}\, e^{-ik\tau/2\varepsilon^2} \int_{(k-1)\tau}^{k\tau} S(((k-1)\tau - s)/\varepsilon;\, i\varepsilon)\, e^{i(k\tau - s)/2\varepsilon^2} f(s)\, ds \end{aligned} \tag{7.17}$$

with initial conditions

$$u(0, \varepsilon) = \tilde{u}_0 = u_0, \tag{7.18}$$

$$\begin{aligned} & \tau^{-1}(e^{-i\tau/2\varepsilon^2} u(\tau, \varepsilon) - u(0, \varepsilon)) \\ & = \tau^{-1}(C(\tau/\varepsilon;\, i\varepsilon) - I)u_0 + \tau^{-1} S(\tau/\varepsilon;\, i\varepsilon)((-i/2\varepsilon)u_0 + \varepsilon u_1) \\ & \quad + (\varepsilon\tau)^{-1}\, e^{-i\tau/2\varepsilon^2} \int_0^\tau S((\tau - s)/\varepsilon;\, i\varepsilon)\, e^{i(\tau - s)/2\varepsilon^2} f(s)\, ds. \end{aligned} \tag{7.19}$$

We use the same piecewise Taylor polynomial approximation (6.9). To figure out explicitly all coefficients in terms of the values of $f(\cdot)$ and of its derivatives at gridpoints, we find explicit formulas for the operators

$$Q^s_{r,a}(a, b, \varepsilon) = \int_a^b S((s-a)/\varepsilon;\, i\varepsilon)\, e^{-i(s-a)/2\varepsilon^2} (s-a)^r\, ds, \tag{7.20}$$

$$Q^s_{r,b}(a, b, \varepsilon) = \int_a^b S((b-s)/\varepsilon;\, i\varepsilon)\, e^{i(b-s)/2\varepsilon^2} (s-a)^r\, ds. \tag{7.21}$$

The computation of $Q^s_{r,a}$ reduces to that of $Q^s_{r,b}$ in the same way as in the parabolic problem; also, as in § 6 we limit ourselves to computing $Q^s_{0,b}$ and $Q^s_{1,b}$ and to give a recursion formula from which each $Q^s_{r,b}$ can be computed. There is a difference, however, between the computations here and those in § 6: the infinitesimal generator of the cosine function $C(t;\, i\varepsilon)$ is $A-(1/4\varepsilon^2)I$, which may not have an inverse. Thus, rather than using formula (2.14) we use (2.36) with the function

$$f(t)=e^{i(b-s)/2\varepsilon^2}(s-a)^r u,$$

which satisfies the necessary assumptions for use of (2.36) if $u\in D(A^m)$ for sufficient large m. We omit the details and only state the final results:

$$\begin{aligned} Q^s_{0,b}(a,b,\varepsilon) &= \int_a^b S((b-s)/\varepsilon;\, i\varepsilon)\, e^{i(b-s)/2\varepsilon^2}\, ds \\ &= -\varepsilon A^{-1}+\varepsilon C((b-a)/\varepsilon;\, i\varepsilon)A^{-1}\, e^{i(b-a)/2\varepsilon^2} \\ &\quad -(1/2)S((b-a)/\varepsilon;\, \varepsilon)A^{-1}\, e^{i(b-a)/2\varepsilon^2}, \end{aligned} \tag{7.22}$$

$$\begin{aligned} Q^s_{1,b}(a,b,\varepsilon) &= \int_a^b S((b-s)/\varepsilon;\, \varepsilon)\, e^{i(b-s)/2\varepsilon^2}(s-a)\, ds \\ &= -\varepsilon A^{-1}(b-a)+\varepsilon^2 S((b-a)/\varepsilon;\, \varepsilon)A^{-1}\, e^{i(b-a)/2\varepsilon^2} \\ &\quad +A^{-1}Q^s_{0,b}(a,b,\varepsilon), \end{aligned} \tag{7.23}$$

$$\begin{aligned} Q^s_{r,b}(a,b,\varepsilon) &= \int_a^b S((b-s)/\varepsilon;\, \varepsilon)\, e^{i(b-s)/2\varepsilon^2}(s-a)^r\, ds \\ &= -\varepsilon A^{-1}(b-a)^r - irA^{-1}Q^s_{r-1,b}(a,b,\varepsilon) \\ &\quad +r(r-1)\varepsilon^2 A^{-1}Q^s_{r-2,b}(a,b;\,\varepsilon). \end{aligned} \tag{7.24}$$

THEOREM 7.1. *Let $u(t,\varepsilon)$ be the solution of* (7.1)-(7.2) *with $f(\cdot)\in B^n(a,b;\,E)$. Then the solution $\{u_k(\varepsilon)\}$ of the difference equation* (7.6) *with right-hand side*

$$\begin{aligned} f_k(\varepsilon) &= \varepsilon\tau^{-2}\, e^{-i\tau/2\varepsilon^2}\sum_{r=0}^{n-1}\frac{1}{r!}\, Q^s_{r,(k+1)\tau}(k\tau,(k+1)\tau,\varepsilon)f^{(r)}(k\tau) \\ &\quad +\varepsilon\tau^{-2}\, e^{i\tau/2\varepsilon^2}\sum_{r=0}^{n-1}\frac{1}{r!}\, Q^s_{r,(k-1)\tau}((k-1)\tau,k\tau,\varepsilon)f^{(r)}((k-1)\tau) \end{aligned} \tag{7.25}$$

and initial conditions

$$u_0(\varepsilon)=u_0, \tag{7.26}$$

$$\begin{aligned} \tau^{-1}(e^{-i\tau/2\varepsilon^2}u_1(\varepsilon)-u_0(\varepsilon)) &= \tau^{-1}(C(\tau/\varepsilon;\, i\varepsilon)-I)u_0+\tau^{-1}S(\tau/\varepsilon;\, i\varepsilon)((-i/2\varepsilon)u_0+\varepsilon u_1) \\ &\quad +(\varepsilon\tau)^{-1}\, e^{-i\tau/2\varepsilon^2}\sum_{r=1}^{n-1}\frac{1}{r!}\, Q^s_{r,\tau}(0,\tau,\varepsilon)f^{(r)}(0). \end{aligned} \tag{7.27}$$

Then $\{u_k(\varepsilon)\}$ satisfies

$$\|u_k(\varepsilon)-u(t_k;\,\varepsilon)\|=0(\tau^n)\qquad (k=1,2,\cdots,N) \tag{7.28}$$

as $\tau\to 0$, uniformly with respect to ε.

8. The hyperbolic singular perturbation problem. We can write the hyperbolic singular perturbation problem (1.4)-(1.5) as an "ε-dependent parabolic singular perturbation problem":

$$\varepsilon^2 u''(t;\,\varepsilon)+u'(t;\,\varepsilon)=A(\varepsilon)u(t,\varepsilon)+f(t), \tag{8.1}$$

$$u(0,\varepsilon)=u_0,\qquad u'(0,\varepsilon)=u_1, \tag{8.2}$$

where $A(\varepsilon)=\varepsilon^2A+B$. Accordingly, the treatment in §§ 5 and 6 can be extended with few modifications.

We state below the necessary information on (8.1), (8.2) from [14]. Assume that, for each $\varepsilon>0$ the function $A(\varepsilon)$ generates a cosine function and denote by $C(t;\varepsilon)$ the cosine function generated by $A(\varepsilon)+(1/4\varepsilon^2)I$ and by $S(t;\varepsilon)$ the corresponding sine function. All the basic estimates in § 6 can be extended if we assume the bounds

$$\|C(t;\varepsilon)\|\leqq M\exp(1/2\varepsilon+\omega\varepsilon)t \qquad (t\geqq 0), \tag{8.3}$$

$$\|S(t;\varepsilon)\|\leqq M\varepsilon\exp(1/2\varepsilon+\omega\varepsilon)t \qquad (t\geqq 0). \tag{8.4}$$

In fact, (8.3) implies

$$e^{-k\tau/2\varepsilon^2}\|C_k(\tau/\varepsilon;\varepsilon)\|\leqq Mk\,e^{k\tau\omega} \tag{8.5}$$

(see (5.14)) and the corresponding bound for the solution of the difference equation (5.8) with initial condition (5.10):

$$\|u_k\|\leqq 2Mk\,e^{k\tau\omega}\|\tilde{u}_0\|+2Mk\tau\,e^{k\tau\omega}\|\tilde{u}_1\|+M(T(T+\tau)/2)\,e^{k\tau\omega}\max_{1\leqq j\leqq n}\|\varepsilon^{-2}f_j\|. \tag{8.6}$$

The solution of the initial value problem (8.1)-(8.2) is given by (6.1), and we obtain on the basis of (8.3)-(8.4) the bound

$$\|u(t;\varepsilon)\|\leqq 2Me^{\omega t}(\|u_0\|+\varepsilon^2\|u_1\|)+2Me^{\omega T}\int_0^T\|f(t)\|\,dt \tag{8.7}$$

in the same way (6.3) is obtained. See [13] for more details on the hyperbolic singular perturbation problem. In particular, the key estimates (8.3) and (8.4) hold if E is a Hilbert space, A is a self-adjoint operator such that

$$(Au,u)\leqq-\kappa(u,u) \qquad (u\in D(A)), \tag{8.8}$$

and B is a closed, densely defined operator with adjoint B^* densely defined, satisfying $D(B)\supseteq D(Q)$, $D(B^*)\supseteq D(Q)$ (Q the unique positive square root of $(-A)^{1/2}$) and such that

$$\|Bu\|\leqq\|Qu\|,\quad \|B^*u\|\leqq\|Qu\| \qquad (u\in D(Q)), \tag{8.9}$$

$$\operatorname{Re}(Bu,u)\leqq\omega(u,u),\quad \operatorname{Re}(B^*u,u)\leqq\omega(u,u) \qquad (u\in D(Q)). \tag{8.10}$$

For other sets of hypotheses implying (8.3)-(8.4) see [13]. We assume that $A(\varepsilon)^{-1}$ exists and is bounded, which is the case for instance if $\omega<0$ in (8.10). The computations in § 6 for the operators $Q^p_{r,a}$, $Q^p_{r,b}$ apply without changes to the operators

$$Q^H_{r,a}(a,b;\varepsilon)=\int_a^b S((s-a)/\varepsilon;\varepsilon)\,e^{(s-a)/2\varepsilon^2}(s-a)^r\,ds,$$

$$Q^H_{r,b}(a,b;\varepsilon)=\int_a^b S((b-s)/\varepsilon;\varepsilon)\,e^{-(b-s)/2\varepsilon^2}(s-a)^r\,ds.$$

The result corresponding to Theorem 6.1 can be stated and proved in exactly the same way, thus we omit the details.

REFERENCES

[1] A. Ashyralyev, *On uniform difference schemes of higher order of approximation for evolutional equations with a small parameter*, to appear.

[2] A. Ashyralyev and P. E. Sobolevskii, *Interpolation theory of linear operators and stability of difference schemes*, Dokl. Akad. Nauk SSSR, 275 (1984), pp. 879-881. (In Russian.) English translation: Soviet Math. Dokl., 29 (1984), pp. 365-367.

[3] G. A. Baker, V. A. Dougalis, and S. M. Serbin, *An approximation theorem for second-order evolution equations*, Numer. Math., 35 (1980), pp. 127-142.

[4] R. Courant, K. O. Friedrichs, and H. Lewy, *Über die partiellen Differenzengleichungen der mathematischen Physik*, Math. Ann., 100 (1928), pp. 32-74.

[5] E. P. Doolan, J. J. H. Miller, and W. H. A. Schilders, *Uniform Numerical Methods for Problems with Initial and Boundary Layers*, Boole Press, Dublin, 1980.

[6] B. F. Esham, *Asymptotics and an asymptotic Galerkin method for hyperbolic-parabolic singular perturbation problems*, SIAM J. Math. Anal., 18 (1987), pp. 762-776.

[7] P. A. Farrell, *Sufficient conditions for uniform convergence of a class of difference schemes for a singularly perturbed problem*, IMA J. Numer. Anal., 7 (1987), pp. 459-472.

[8] H. O. Fattorini, *Ordinary differential equations in linear topological spaces*, I, II, J. Differential Equations, 5 (1969), pp. 72-105; 6 (1969), pp. 79-104.

[9] ———, *Uniformly bounded cosine functions in Hilbert space*, Indiana Univ. Math. J., 20 (1970), pp. 411-425.

[10] ———, *Convergence and approximation theorems for vector valued distributions*, Pacific J. Math., 105 (1983), pp. 77-114.

[11] ———, *Singular perturbation and boundary layer for an abstract Cauchy problem*, J. Math. Anal. Appl., 97 (1983), pp. 529-571.

[12] ———, *On the Schrödinger singular perturbation problem*, SIAM J. Math. Anal., 16 (1985), pp. 1000-1019.

[13] ———, *The hyperbolic singular perturbation problem: An operator-theoretic approach*, J. Differential Equations, 70 (1987), pp. 1-41.

[14] ———, *Second Order Linear Differential Equations in Banach Spaces*, Notas de Mat. 108, North-Holland, Amsterdam, 1985.

[15] G. E. Forsythe and W. A. Wasow, *Finite-difference Methods for Partial Differential Equations*, John Wiley, New York, 1960.

[16] R. H. W. Hoppe, *Discrete approximations to cosine operator functions*, SIAM J. Numer. Anal., 19 (1982), pp. 1110-1128.

[17] G. H. Hsiao and K. E. Jordan, *A numerical treatment for parabolic equations with a small parameter*, SIAM J. Math. Anal., 14 (1983), pp. 507-521.

[18] S. G. Krein, *Linear Differential Equations in Banach Spaces*, Izdatelstvo Nauka, Moscow, 1967. (In Russian.)

[19] J. J. H. Miller, *Optimal uniform difference schemes for linear initial-value problems*, Comm. Math. Appl., 12B (1986), pp. 1209-1215.

[20] L. A. Minin, *Finite Difference Schemes of Second Order Accuracy for Hyperbolic Equations: Numerical Solution of Boundary Value Problems and Integral Equations*, Tartu Gos. University, Tartu, 1981, pp. 20-22. (In Russian.)

[21] C. Piskarev, *Discretization of abstract hyperbolic equations*, Tartu Riikliku Ülikooli Toimetised (Acta et Commentationes Universitatis Tartuensis) Matemaatika-ja mekhaanikaalaseid töid, 500 (1979), pp. 3-23. (In Russian.)

[22] P. E. Sobolevskii and L. M. Cebotareva, *Approximate solutions of the Cauchy problem for an abstract hyperbolic equation by the method of lines*, Izv. Vyss. Ucebn. Zaved. Mat. 5, 180 (1977), pp. 103-116. (In Russian.)

[23] M. Sova, *Cosine operator functions*, Rozprawy Mat., 49 (1966), pp. 1-47.

[24] A. Zygmund, *Trigonometric Functions*, Cambridge University Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 55-71, January 1992

UNIQUE CONTINUATION FOR THE KORTEWEG-DE VRIES EQUATION*

BINGYU ZHANG†

Abstract. Unique continuation problems are considered for the Korteweg-de Vries (KdV) equation

$$u_t + uu_x + u_{xxx} = 0, \qquad -\infty < x, \quad t < +\infty.$$

By using the inverse scattering transform and some results from the Hardy function theory, it is proven that if $u \in L^\infty_{\mathrm{loc}}(R, H^s(R))(s > \frac{3}{2})$ is a solution of the KdV equation, then it cannot have compact support at two different moments unless it vanishes identically. In addition, it is shown under certain conditions that if u is a solution of the KdV equation, then u must vanish everywhere if it vanishes on two horizontal half lines in the x-t space. This implies that the solution u must vanish everywhere if it vanishes on an open subset in the x-t space. As a consequence of the Miura transformation, the above results for the KdV equation are also true for the modified Korteweg-de Vries equation

$$v_t - 6v^2 v_x + v_{xxx} = 0, \qquad -\infty < x, \quad t < +\infty.$$

Key words. unique continuation, Korteweg-de Vries equation, inverse scattering transform, Hardy function

AMS(MOS) subject classification. 35Q20

1. Introduction. Let L be an evolution operator acting on functions defined on some connected open set Ω of $R^{n+1} = R^n_x \times R_t$. L is said to have the unique continuation property if every solution u of $Lu = 0$ that vanishes on some nonempty open set $\Omega_1 \subset \Omega$ vanishes in the horizontal component of Ω_1 in Ω, i.e., in $\{(x, t) \in \Omega;\ \exists x_1, (x_1, t) \in \Omega_1\}$ (cf. [4], [8], and [13]).

Saut and Scheurer [12], [13] considered some dispersive operators in one space dimension of the type

$$L = iD_t + \alpha i^{2k+1} D^{2k+1} + R(x, t, D)$$

where $\alpha \neq 0$, $D = 1/i\, \partial/(\partial x)$, $D_t = 1/i\, \partial/(\partial t)$, and

$$R(x, t, D) = \sum_{j=0}^{2k} r_j(x, t) D^j, \qquad r_j \in L^\infty_{\mathrm{loc}}(R, L^2_{\mathrm{loc}}(R)).$$

They proved that if $u \in L^2_{\mathrm{loc}}(R, H^{2k+1}_{\mathrm{loc}}(R))$ is a solution of $Lu = 0$, which vanishes in some open set Ω_1 of $R_x \times R_t$, then u vanishes in the horizontal component of Ω_1.

As a consequence of uniqueness of the solution of the KdV equation in $L^\infty_{\mathrm{loc}}(R, H^3(R))$, their result immediately yields the following theorem.

THEOREM 1.1. *If $u \in L^\infty_{\mathrm{loc}}(R, H^3(R))$ is a solution of the* KdV *equation*

$$u_t + uu_x + u_{xxx} = 0$$

and vanishes on an open set of $R_x \times R_t$, then

$$u(x, t) = 0 \quad \textit{for } x \in R, \quad t \in R.$$

In this paper, we consider various unique continuation properties of the KdV equation. By using the inverse scattering transform we prove that a solution u of the KdV equation, which decays sufficiently fast at $\pm\infty$, vanishes everywhere if it vanishes on two different horizontal half lines in the x-t space. In addition, we prove that a class of generalized solution of the KdV equation or solutions of the KdV equation in

* Received by the editors January 9, 1990; accepted for publication (in revised form) March 8, 1991.

† Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221-0025.

$L^{\infty}_{\text{loc}}(R;\, H^s(R))(s>\frac{3}{2})$ cannot have compact support at two different moments in the x-t space unless it vanishes identically. The methods we use are different from those of Saut and Scheurer. However, with a little stronger condition made on the solution u, our results recover Theorem 1.1 as a corollary.

The results obtained for the KdV equation are also true for the modified KdV equation

$$v_t - 6v^2 v_x + v_{xxx} = 0$$

as we see by applying the Miura transformation

$$u = -\tfrac{1}{6}(v^2 + v_x),$$

where v is a solution of the modified KdV equation and u is a solution of the KdV equation

$$u_t + uu_x + u_{xxx} = 0.$$

The paper is organized as follows: In § 2, we provide a sketch of the inverse scattering transform and cite some results from Deift and Trubowitz [3] which play important roles in the proof of our main results. In § 3, we prove some unique continuation properties for the linearized KdV equation (Airy equation) by using the Fourier transform. In § 4, we prove our main unique continuation results for the KdV equation. We will see that the inverse scattering transform in the nonlinear problem plays the same role as that of the Fourier transform in the linear problem.

2. The inverse scattering transform. We use the following notation to denote the Fourier and the inverse Fourier transform:

$$\hat{f}(y) = \frac{1}{\pi}\int_{-\infty}^{+\infty} e^{2ixy} f(x)\, dx$$

and

$$\check{f}(x) = \int_{-\infty}^{+\infty} e^{-2ixy} f(y)\, dy.$$

Let H^2_+ be the Hardy space of functions $h(k)$ analytic in $\operatorname{Im} k > 0$ with

$$\sup_{b>0} \int_{-\infty}^{+\infty} |h(a+ib)|^2\, da < \infty.$$

Such a function $h(k)$ assumes boundary values

$$h(a) = \lim_{\varepsilon \to 0} H(a + i\varepsilon)$$

in $L^2(R)$ and it is well known that

$$H^2_+ = \{h(k) \in L^2(R) \colon \operatorname{supp} \hat{h} \subset (-\infty, 0)\}.$$

Similarly, H^2_- denotes the Hardy space of functions analytic in $\operatorname{Im} k < 0$ with

$$\sup_{b<0} \int_{-\infty}^{+\infty} |h(a+ib)|^2\, da < \infty$$

and

$$H^2_- = \{h(k) \in L^2(R) \colon \operatorname{supp} \hat{h} \subset (0, \infty)\}.$$

Consider the Schrödinger operator

$$\mathcal{S}(f) = -f'' + qf - k^2 f, \tag{1}$$

where k is a complex number and q, called a potential of (1), is a real function of $x \in R$ satisfying

$$\int_{-\infty}^{+\infty} (1+|x|)|q(x)|\,dx < \infty. \tag{2}$$

Let f_1 and f_2 be two solutions of $\mathscr{S}(f)=0$ such that

$$f_1 \sim e^{ikx} \quad \text{as } x \to \infty,$$
$$f_2 \sim e^{-ikx} \quad \text{as } x \to -\infty.$$

Set

$$m_1 = e^{-ikx} f_1, \qquad m_2 = e^{ikx} f_2.$$

Then

$$m_1'' + 2ikm_1' = qm_1, \tag{3}$$
$$m_2'' - 2ikm_2' = qm_2 \tag{4}$$

with $m_1 \to 1$ as $x \to \infty$ and $m_2 \to 1$ as $x \to -\infty$.

Define (for all real $k \neq 0$)

$$\frac{1}{T(k)} = 1 - \frac{1}{2ik}\int_{-\infty}^{+\infty} q(t)m_1(t,k)\,dt, \tag{5}$$

$$\frac{R_1(k)}{T(k)} = \frac{1}{2ik}\int_{-\infty}^{+\infty} e^{-2ikt}q(t)m_2(t,k)\,dt, \tag{6}$$

$$\frac{R_2(k)}{T(k)} = \frac{1}{2ik}\int_{-\infty}^{+\infty} e^{2ikt}q(t)m_1(t,k)\,dt, \tag{7}$$

and

$$S(k) = \begin{pmatrix} T(k) & R_2(k) \\ R_1(k) & T(k) \end{pmatrix}.$$

Here $T(k)$ is called the *transmission coefficient* and $R_1(k)$ is called the *right reflection coefficient*. The operator $\mathscr{S}(f)$ has only finite many simple negative eigenvalues, listed as

$$-\beta_n^2 < -\beta_{n-1}^2 < \cdots < -\beta_1^2, \tag{8}$$

which are called *bound states* of the potential q for $\mathscr{S}(f)$. The corresponding eigenfunctions $\psi_n, \cdots, \psi_1$ are assumed to satisfy

$$\int_{-\infty}^{+\infty} |\psi_k|^2\,dx = 1, \qquad k = 1, 2, \cdots, n.$$

Let

$$c_j = \lim_{x\to\infty} e^{-\beta_j x}\psi_n(x), \qquad j = 1, 2, \cdots, n. \tag{9}$$

They are called *normalizing coefficients.*

The following propositions are cited from [3].

PROPOSITION 2.1. *There is a constant c independent of x and k such that*

$$|m_j(x,k) - 1| < e^{c/|k|}\frac{c}{|k|}, \qquad j = 1, 2$$

for $k \neq 0$ and $\operatorname{Im} k \geqq 0$.

PROPOSITION 2.2. *The transmission coefficient* $T(k)$ *is meromorphic in* $\operatorname{Im} k > 0$ *with a finite number of simple poles* $i\beta_1, \cdots, i\beta_n$, $\beta_j > 0$, $T(k)$ *is continuous in* $\operatorname{Im} k \geqq 0$, $k \neq 0$, $i\beta_1, \cdots, i\beta_n$, *and*

$$T(k) = 1 + O\left(\frac{1}{k}\right) \quad \textit{as } |k| \to \infty, \quad \operatorname{Im} k \geqq 0.$$

Moreover, if q *is supported on a half line, then* $T(k)$ *is meromorphic in the entire* k*-plane.*

PROPOSITION 2.3. *Let* q *be the potential of the Schrödinger operator* $\mathcal{S}$ *defined in* (1).

(i) *If* q *has* N*th-order derivatives which are in* $L^1(R)$, *then*

$$R_1(k) = O\left(\frac{1}{k^{N+1}}\right) \quad \textit{as } |k| \to \infty, \quad k \textit{ real.}$$

(ii) *If* q *is supported on a left half line, then* $R_1(k)$ *is meromorphic in the upper half plane.*

(iii) *If* q *has compact support, then* $R_1(k)$ *is meromorphic in the entire plane.*

PROPOSITION 2.4. (Removing all bound states). *Assume that* q *has* n *bound states, listed as*

$$-\beta_n^2 < -\beta_{n-1}^2 < \cdots < -\beta_1^2.$$

Define inductively

$$q(x, 0) = q(x),$$

$$q(x, -n) = q(x, -(m-1)) - 2\frac{d^2}{dx^2}\log f_1(x, i\beta_{n-m+1}; -(m-1))$$

for $1 \leqq m \leqq n$.

Then $q(x, -n)$ *has no bound states and its right reflection coefficient is*

$$R_1(k, -n) = (-1)^n\left(\prod_{j=1}^{n}\frac{k - i\beta_j}{k + i\beta_j}\right)R_1(k). \tag{10}$$

In addition, if $\operatorname{supp} q \subset (a, b)$, *then* $\operatorname{supp} q(x, -n) \subset (a, b)$ *where* a *is a finite number or* $-\infty$.

PROPOSITION 2.5. (Trace formula). *If* q *has no bound states, then*

$$q(x) = \frac{2i}{\pi}\int_{-\infty}^{+\infty} kR_1(k)\, e^{2ikx} m_1^2(x, k)\, dk. \tag{11}$$

PROPOSITION 2.6. *Assume that* q *has no bound states. If* q *has a support lying to the left of some point, then*

$$\sup\{\textit{support of } q\} = \inf\left\{x\colon \int_{-\infty}^{+\infty} kR_1(k)\, e^{2ikt}\, dk = 0 \textit{ for all } t > x\right\}.$$

Consider the initial value problem for the KdV equation

$$u_t + uu_x + u_{xxx} = 0, \qquad u(x, 0) = q(x), \tag{12}$$

for $x \in R$ and $t \geqq 0$ where $q \in H^4(R)$ and

$$\int_{-\infty}^{+\infty} (1 + |x|)|q(x)|\, dx < \infty.$$

It is well known that the initial value problem (12) can be solved by using the inverse scattering transform (cf. [2], [9], and [16] for more details). The basic steps of the solution method are as follows.

Step 1. With the initial data $u(x,0)=q(x)$ as a potential, solve

$$-f''+qf=\lambda f$$

to get the scattering data

$$\lambda_j(0)=-\beta_j^2, \qquad c_j(0)=c_j$$

and

$$R_1(k,0)=R_1(k), \qquad T(k,0)=T(k)$$

as in (8), (9), (6), and (5).

Step 2. Treat t as a parameter and consider $u(x,t)$, the solution of (12), as a potential to solve

$$-f''+u(x,t)f=\lambda f.$$

The remarkable fact is that we can obtain the scattering data corresponding to $u(x,t)$ without knowing what $u(x,t)$ is. In fact,

$$\lambda_j(t)=\lambda_j(0), \quad c_j(t)=c_j(0)\, e^{4\beta_j^3 t}, \qquad j=1,2,\cdots,n \tag{13}$$

and

$$R_1(k,t)=R_1(k,0)\, e^{8ik^3 t}, \qquad T(k,t)=T(k,0). \tag{14}$$

Step 3. Let

$$B(\xi,t)=\sum_{j=1}^{n} c_j^2(0)\, e^{8\beta_j^3 t-\beta_j\xi}+\frac{1}{2\pi}\int_{-\infty}^{+\infty} R_1(k,0)\, e^{i(8k^3t+k\xi)}\, dk \tag{15}$$

and use it to define the Gel'fand-Levitan equation

$$K(x,y;t)+B(x+y;t)+\int_x^{\infty} B(x+z;t)K(z,y;t)\, dz=0 \tag{16}$$

for all time t.

Step 4. Solve (16) to get $K(x,y;t)$. Then the desired solution of (12) is

$$u(x,t)=-2\frac{d}{dx}K(x,x;t).$$

3. The linear KdV equation. Consider the initial value problem for the linear KdV equation:

$$u_t+u_{xxx}=0, \qquad u(x,0)=q(x) \tag{17}$$

for $x\in R$, $t\in R$.

Formally, by using the Fourier transform, we obtain

$$u(x,t)=\frac{1}{\pi}\int_{-\infty}^{+\infty} e^{2ikx}\, e^{ik^3t}\tilde{q}(k)\, dk. \tag{18}$$

If $q\in L^2(R)$, the $u(x,t)$ defined in (18) belongs to $L^2(R)$ for any t and is called a mild solution of (17). Obviously, if q is smooth enough, then $u(x,t)$ is a classical solution of (17).

THEOREM 3.1. *Let $u(x, t)$ be a mild solution of* (17). *If there exist $t_1 < t_2$ such that*

$$\operatorname{supp} u(\cdot, t_j) \subset (-\infty, \alpha), \qquad j = 1, 2 \tag{19}$$

or

$$\operatorname{supp} u(\cdot, t_j) \subset (\alpha, \infty), \qquad j = 1, 2 \tag{20}$$

for some $\alpha \in R$, then

$$u(x, t) \equiv 0, \qquad -\infty < x, \quad t < +\infty.$$

To prove Theorem 2.1 we need the following lemma.

LEMMA 3.1. *Assume $0 \neq q \in L^2(R)$ and* $\operatorname{supp} q \subset (-\infty, 0)$. *Let*

$$H(b) = \int_{-\infty}^{0} e^{2bx} q(x)\, dx.$$

Then there exist $b_n > 0$ such that $b_n \to \infty$ as $n \to \infty$ and

$$|H(b_n)| \geqq \alpha_1 e^{-\beta_1 b_n}, \qquad n = 1, 2, \cdots \tag{21}$$

where α_1 and β_1 are positive constants independent of n.

Proof of Lemma 3.1. Let

$$F(z) = \int_{-\infty}^{\infty} e^{-2izx} q(x)\, dx \qquad (\operatorname{Im} z > 0)$$

and

$$f(z) = F\left(i \frac{1+z}{1-z}\right) \qquad (|z| < 1).$$

Then f belongs to the Hardy space H^2 in the unit disc and

$$H(b) = F(ib).$$

Consider the canonical factorization of f (cf. [10, Chap. 17]):

$$f(z) = B(z) S(z) Q(z)$$

where B is a Blaschke product,

$$B(z) = z^k \prod_n \frac{z_n - z}{1 - \bar{z}_n z} \frac{|z_n|}{z_n},$$

z_n are zeros of $H(z)$ in the unit disk,

$$S(z) = \exp\left\{ -\int_{-\pi}^{\pi} \frac{e^{i\psi} + z}{e^{i\psi} - z}\, d\mu(\psi) \right\}$$

is a singular inner function (μ is a positive singular measure, possibly zero), and

$$Q(z) = \exp\left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i\psi} + z}{e^{-\psi} - z} \log |f(e^{i\psi})|\, d\psi \right\}$$

is an outer function. If $z = r e^{i\theta}$, then

$$\operatorname{Re}\left[\frac{e^{i\psi} + z}{e^{i\psi} - z}\right] = P_r(\theta - \psi)$$

$$= \frac{1 - r^2}{1 - 2r \cos(\theta - \psi) + r^2},$$

the Poisson kernel, and

$$P_r(\theta-\psi)<\frac{1+r}{1-r} \qquad (0\leqq r<1).$$

Thus,

$$\log|S(r)|\geqq -\frac{1+r}{1-r}\|\mu\|$$

($\|\mu\|=\mu(-\pi,\pi)$) and

$$\log|Q(r)|\geqq -\frac{1+r}{1-r}\int_{-\pi}^{\pi}\log^-|f(e^{i\psi})|\,d\psi$$

where $\log=\log^+-\log^-$. So there exists $C<\infty$ such that

$$(1-r)\log|S(r)Q(r)\|\geqq -C \qquad (0<r<1).$$

As for the Blaschke product B, it is known that

$$\overline{\lim_{r\to 1_-}}\sup(1-r)\log|B(r)|=0$$

[14, Lemma 2.3]. Hence there exists $r_n\to 1_-$ as $n\to\infty$ such that

$$\lim_{n\to\infty}(1-r_n)\log|B(r_n)|=0$$

and

$$(1-r_n)\log|f(r_n)|\geqq -3C, \qquad n=1,2,\cdots.$$

Let

$$r_n=\frac{b_n-1}{b_n+1}.$$

Then

$$b_n=\frac{1+r_n}{1-r_n}, \qquad 1-r_n=\frac{2}{b_n+1},$$

and

$$\frac{2}{b_n+1}\log|F(ib_n)|=(1-r_n)\log|f(r_n)|\geqq -3C.$$

Thus

$$|F(ib_n)|\geqq e^{-3/2C(b_n+1)}$$

for all n.

Choosing $\alpha_1=e^{-3C/2}$ and $\beta_1=3C/2$, we have

$$|H(b_n)=|F(ib_n)|\geqq \alpha_1\,e^{-\beta_1 b_n}, \qquad n=1,2,\cdots.$$

The proof is completed. □

Proof of Theorem 3.1. Without loss of generality, assume that $\alpha=0$ and $t_1=0$. Then

$$u(x,t_1)=u(x,0)=q(x)$$

and

$$u(x, t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} e^{2ikx}\, e^{8ik^3x} \check{q}(k)\, dk.$$

For the case in which u satisfies (19), we have

$$\int_{-\infty}^{+\infty} e^{2ikx}\, e^{8ik^3t_2} \check{q}(k)\, dk = 0 \quad \text{for } x > 0. \tag{22}$$

In addition, since $e^{8ik^3t_2}\check{q}(k) \in L^2(R)$, we have $e^{8ik^3t_2}\check{q}(k) \in H_+^2$. Thus, by the definition of the Hardy function,

$$\sup_{b>0} \int_{-\infty}^{+\infty} |e^{8i(a+ib)^3t_2} \check{q}(a+ib)|^2\, da < \infty. \tag{23}$$

On the other hand,

$$\begin{aligned} I(b) &\equiv \int_{-\infty}^{+\infty} |e^{8i(a+ib)^3t_2} \check{q}(a+ib)|^2\, da \\ &= e^{16b^3t_2} \int_{-\infty}^{+\infty} e^{-48a^2bt_2} |\check{q}(a+ib)|^2\, da \\ &\geqq e^{16b^3t_2} \int_0^1 e^{-48a^2bt_2} |\check{q}(a+ib)|^2\, da \\ &\geqq e^{16b^3t_2}\, e^{-48bt_2} \int_0^1 \left| \int_{-\infty}^0 e^{2bx}\, e^{-2iax} q(x)\, dx \right|^2 da \\ &\geqq e^{16b^3t_2}\, e^{-48bt_2} \left| \int_0^1 \int_{-\infty}^0 e^{2bx}\, e^{-2iax} q(x)\, dx\, da \right|^2 \\ &\geqq e^{16b^3t_2}\, e^{-48bt_2} \left| \int_{-\infty}^0 e^{2bx} \left(\int_0^1 e^{-2iax}\, da \right) q(x)\, dx \right|^2 \\ &= e^{16b^3t_2}\, e^{-48bt_2} \left| \int_{-\infty}^0 e^{2bx} h(x)\, dx \right|^2, \end{aligned}$$

where

$$h(x) = \frac{1 - e^{-2ix}}{2ix} q(x).$$

If q is not identical zero, by Lemma 2.1, there exist b_n such that

$$\left| \int_{-\infty}^0 e^{2b_nx} h(x)\, dx \right| \geqq \alpha_1\, e^{-\beta_1 b_n}, \qquad n = 1, 2, \cdots,$$

where $b_n \to \infty$ as $n \to \infty$. We have

$$I(b_n) \geqq \alpha_1^2\, e^{16b_n^3t_2}\, e^{-48b_nt_2}\, e^{-2\beta_1 b_n} \quad \text{for } n = 1, 2, \cdots$$

and

$$\lim_{n\to\infty} I(b_n) = \infty,$$

which is in contradiction with (23). Hence, we must have

$$q(x) = 0 \quad \text{for } x \in R$$

and

$$u(x, t) \equiv 0, \qquad -\infty < x, \quad t < +\infty.$$

The case where u satisfies assumption (20) reduces to the case that we have just treated if we let

$$t' = -t, \qquad x' = -x.$$

The proof is completed. □

COROLLARY 3.1. *Let u be a mild solution to* (17). *If u vanishes on an open subset in x-t space, then u vanishes everywhere.*

Proof. Without loss of generality, we assume that

$$u(x, t) = 0 \quad \text{for } \alpha < x < \beta, \quad t_1 < t < t_2,$$

for some $\alpha, \beta \in R$ and $t_1, t_2 \in R$. Define

$$u_1(x, t) = \begin{cases} u(x, t) & \text{for } x \leqq \gamma, \\ 0 & \text{for } x > \gamma \end{cases}$$

and

$$u_2(x, t) = \begin{cases} 0 & \text{for } x \leqq \gamma, \\ u(x, t) & \text{for } x > \gamma \end{cases}$$

where $\gamma = (\alpha + \beta)/2$.

Both u_1 and u_2 are mild solutions of the linear KdV equation for $t_1 < t < t_2$ and

$$\operatorname{supp} u_1(x, t) \subset (-\infty, \gamma), \qquad \operatorname{supp} u_2(x, t) \subset (\gamma, \infty)$$

for $t_1 < t < t_2$. Hence, by Theorem 2.1,

$$u_2(x, t) = u_1(x, t) = 0 \quad \text{for } x \in R, \quad t_1 < t < t_2,$$

which implies that u vanishes everywhere. The proof is completed. □

4. The Korteweg–de Vries equation. Consider the initial value problem for the KdV equation

$$(24) \qquad u_t + uu_x + u_{xxx} = 0, \qquad u(x, 0) = q(x).$$

It is known that if $q \in H^s(R)$ for $s > 2$, then (24) has a unique solution $u \in C(R; H^s(R))$ and for $\frac{3}{2} < s < 2$, equation (24) has a unique solution $u \in C(-T, T; H^s(R))$ where T depends on the initial value q (cf. [1], [5], and [11]).

For the solutions of (24), we have the following proposition.

PROPOSITION 4.1. *Assume that $u \in C(R; H^s(R))$, $s > \frac{3}{2}$, is a solution of the* KdV *equation and $u(x, 0) = q(x)$ satisfies that*

(i) $\int_{-\infty}^{+\infty} |q'(x)|\, dx < \infty$, $\int_{-\infty}^{+\infty} (1 + |x|) q(x)|\, dx < \infty$;

(ii) $\operatorname{supp} q \subset (-\infty, \alpha)$ *for some* $\alpha \in R$;

(iii) *q has no bound states if q is considered as a potential for* (1).

Then

$$u(x, t) = 0 \quad \textit{for } x \in R, \quad t \in R,$$

if there is a $t^* > 0$ such that

$$\operatorname{supp} u(\cdot, t^*) \subset (-\infty, \alpha).$$

Proof. Without loss of generality, we assume that $\alpha = 0$. According to the steps solving the KdV equation by the inverse scattering transform, $u(x, t^*)$ has no bound states if $q(x)$ has no bound states. Let $R_1(k)$ be the right reflection coefficient of q and $R_1(k, t^*)$ be the right reflection coefficient of $u(\cdot, t^*)$. Then by (14),

$$R_1(k, t^*) = R_1(k)\, e^{8ik^3t^*}.$$

From Proposition 2.3 and 2.6, we have

$$\int_{-\infty}^{+\infty} k^2 |R_1(k)|^2\, dk < \infty \tag{25}$$

and

$$\int_{-\infty}^{+\infty} kR_1(k)\, e^{2ikx}\, dx = 0 \quad \text{for } x \geqq 0. \tag{26}$$

Similarly, we have

$$\int_{-\infty}^{+\infty} k^2 |R_1(k, t^*)|^2\, dk = \int_{-\infty}^{+\infty} k^2 R_1(k)|^2\, dk < \infty \tag{27}$$

and

$$\int_{-\infty}^{+\infty} kR_1(k, t^*)\, e^{2ikx}\, dk = 0 \quad \text{for } x \geqq 0. \tag{28}$$

Thus,

$$kR_1(k) \in H_+^2, \qquad kR_1(k, t^*) \in H_+^2$$

and

$$\sup_{b>0} \int_{-\infty}^{+\infty} |a + ib|^2 |R_1(a + ib, t^*)|^2\, da < \infty. \tag{29}$$

On the other hand,

$$\begin{aligned} I(b) &\equiv \int_{-\infty}^{\infty} |a + ib|^2 |R_1(a + ib, t^*)|^2\, da \\ &= \left(\int_{-\infty}^{+\infty} (a^2 + b^2) |R_1(a + ib)|^2\, e^{-48a^2bt^*}\, da \right) e^{16b^3t^*} \\ &\geqq \left(\int_0^1 (a^2 + b^2) |R_1(a + ib)|^2\, e^{-48a^2bt^*}\, da \right) e^{16b^3t^*} \\ &\geqq \left(\int_0^1 (a^2 + b^2) |R_1(a + ib)|^2\, da \right) e^{16b^3t^* - 48bt^*}. \end{aligned}$$

Note that

$$R_1(k) = \frac{T(k)}{2ik} \int_{-\infty}^{+\infty} e^{-2ikt} q(t) m_2(t, k)\, dt.$$

By Propositions 2.1 and 2.2,

$$R_1(k) \sim \frac{1}{2ik} \int_{-\infty}^{+\infty} e^{-2ikt} q(t)\, dt$$

as $|k| \to \infty$, $\operatorname{Im} k \geqq 0$. Hence,

$$I(b) \geqq C e^{16b^3 t^* - 48bt^*} \int_0^1 |\check{q}(a+ib)|^2\, da,$$

where C is a constant independent of b.

If q is not identically zero, by Lemma 3.1 and the same argument as used in the proof of Theorem 3.1, there exist b_n with $b_n \to \infty$ such that

$$\lim_{n\to\infty} I(b_n) = \infty,$$

which is in contradiction with (29). Hence q must be identically zero which implies, by the uniqueness of the solution to (24), that

$$u(x,t) \equiv 0, \qquad -\infty < x, \quad t < +\infty.$$

The proof is completed. □

THEOREM 4.1. *Suppose $u \in C(R, H^s(R))$, $s > \frac{3}{2}$, is a solution of the* KdV *equation*

$$u_t + uu_x + u_{xxx} = 0.$$

If there exist $t_1, t_2 \in R$ with $t_1 < t_2$ such that

$$\text{(30)} \qquad \operatorname{supp} u(\cdot, t_j) \subset (-\infty, \alpha), \qquad j = 1, 2,$$

and

$$\text{(31)} \qquad \int_{-\infty}^{+\infty} (1+|x|)|u(x,t_1)|\, dx < \infty, \qquad \int_{-\infty}^{+\infty} |u_x(x,t_1)|\, dx < \infty,$$

or

$$\text{(32)} \qquad \operatorname{supp} u(\cdot, t_j) \subset (\alpha, \infty), \qquad j = 1, 2,$$

and

$$\text{(33)} \qquad \int_{-\infty}^{+\infty} (1+|x|)|u(x,t_2)|\, dx < \infty, \qquad \int_{-\infty}^{+\infty} |u_x(x,t_2)|\, dx < \infty,$$

then

$$u(x,t) = 0 \quad \textit{for } x \in R, \quad t \in R.$$

Proof. Without loss of generality, we assume that $t_1 = 0$. We only consider the case wherein u satisfies assumption (30) and (31). The other case will reduce to this case if we make transform

$$x' = -x, \qquad t' = -t.$$

Suppose $u(x,0) = q(x)$. If $q(x)$ has no bound states, then our assertion is Proposition 4.1, which we have proved. Hence we assume that q has n bound states

$$-\beta_n^2 < \cdots < -\beta_1^2.$$

Let

$$q(x,0) = q(x)$$

and

$$q(x, m) = q(x, -(m-1)) - 2\frac{d^2}{dx^2}\log f_1(x, i\beta_{n-m+1}; -(m-1)).$$

Inductively, we obtain $q(x, -n)$, which has no bound states. In addition,

$$\operatorname{supp} q(x, -n) \subset (-\infty, \alpha)$$

and the right reflection coefficient of $q(x, -n)$ is

$$R_1(k, -n) = (-1)^n \left(\prod_1^n \frac{k - i\beta_j}{k + i\beta_j} \right) R_1(k). \tag{34}$$

Let $v(x, t)$ be the solution of

$$v_t + vv_x + v_{xxx} = 0, \qquad v(x, 0) = q(x, -n). \tag{35}$$

Claim.

$$\operatorname{supp} v(\cdot, t_2) \subset (-\infty, \alpha). \tag{36}$$

If (36) is true, then from Proposition 3.1 we have

$$q(x, -n) = 0 \quad \text{for } x \in R.$$

It follows that

$$R_1(k, -n) = (-1)^n \left(\prod_{j=1}^n \frac{k - i\beta_j}{k + i\beta_j} \right) R_1(k) = 0.$$

Thus

$$R_1(k) = 0, \qquad k \text{ real}.$$

It implies that q has only bound states and has no continuous spectrum if we consider q as a potential of (1). Hence u is a pure n-soliton solution of the KdV equation. This is in contradiction with assumption (30) since a pure n-soliton solution of the KdV equation cannot have support on a half line for any $t \in R$. The contradiction implies, via the uniqueness of the solution of the KdV equation in $C(R, H^s(R))$, that

$$q(x) = 0, \qquad x \in R$$

and

$$u(x, t) = 0, \qquad x \in R, \quad t \in R.$$

The theorem will be established when we prove the claim.

Proof of the claim. As a potential for (1), $v(x, t_2)$ has no bound states since $v(x, 0) = q(x, -n)$ has no bound states. By (14) and (34), the right reflection coefficient corresponding to $v(x, t_2)$ is

$$R_1(k, v) = (-1)^n \prod_{j=1}^n \frac{k - i\beta_j}{k + i\beta_j} R_1(k)\, e^{8ik^3 t_2}. \tag{37}$$

Let $p(x) = u(x, t_2)$. Then $p(x)$ has n bound states

$$-\beta_n^2 < \cdots < -\beta_1^2.$$

Let $p(x, -n)$ be defined as in Proposition 2.4 such that $p(x, -n)$ has no bound states and its right reflection coefficient is

$$R_1(k, p(x, -n)) = (-1)^n \prod_{j=1}^{n} \frac{k - i\beta_j}{k + i\beta_j} R_1(k)\, e^{8ik^3 t_2}.$$

It is clear that

$$R_1(k, p(x, -n)) = R_1(k, v).$$

By assumption (30) and Proposition 2.4,

$$\operatorname{supp} p(\cdot\,, -n) \subset (-\infty, \alpha).$$

Proposition 2.6 implies that

$$\int_{-\infty}^{+\infty} kR_1(k, p(x, -n))\, e^{2ikt}\, dk = 0 \quad \text{for } t \geqq \alpha.$$

Hence

$$\int_{-\infty}^{+\infty} kR_1(k, v)\, e^{2ikt}\, dk = 0 \quad \text{for } t \geqq \alpha.$$

Using Proposition 2.6 again, we arrive at

$$\operatorname{supp} v(\cdot\,, t_2) \subset (-\infty, \alpha).$$

The proof is completed. □

Let $\sigma(x) = (1+|x|^2)^{1/2}$. For $r, s \in R$ we denote by $H_r^s(R)$ the completion of the Schwartz space $S(R)$ of rapidly decreasing infinitely differentiable functions on R, in the norm $\|u\|_{r,s} = |\sigma^r F^{-1} \sigma^s F u|_2$, where $|\cdot|_2$ denotes the standard norm of $L^2(R)$ and F^{-1}, F are the inverse Fourier transform and the Fourier transform, respectively. For $r \in R$ and $s \in N \cup \{0\}$, we denote by $W_r^s(R)$ the completion of the space $S(R)$ in the norm $\|u\|^* = (\sum_{|\alpha| \leqq s} |\sigma^r D^\alpha u|_2^2)^{1/2}$. For $s \in N \cup \{0\}$, $H_r^s(R) = W_r^s(R)$ (cf. [15]).

Let

$$S_r^s(R) = H^s(R) \cap H_r^0(R).$$

Then

$$S_r^s(R) \subset [H_r^0(R), H_0^r(R)]_\theta = H_{(1-\theta)r}^{\theta s}(R)$$

for any $0 < \theta < 1$ (cf. [15]). It has been proved by Tsutsumi [15] that if $q \in S_r^s$ with $r \geqq 0$ and $s \geqq \max\{2r, \frac{1}{2} + \varepsilon\}$, then the initial value problem (24) has a unique solution $u(x, t)$ in $L_{\text{loc}}^\infty(R; S_r^s(R))$. It is clear that if $q \in S_r^{2r}(R)$ with $r > \frac{3}{2}$, then $(1+|x|)q(x) \in L^1(R)$ and $q_x \in L^1(R)$. Hence, if $q \in S_r^{2r}$ with $r > \frac{3}{2}$, the solution u of (24) satisfies

$$\int_{-\infty}^{+\infty} (1+|x|)|u(x, t)|\, dx < \infty, \qquad \int_{-\infty}^{+\infty} |u_x(x, t)|\, dx < \infty$$

for any $t \in R$.

From Theorem 4.1 we have the following corollaries.

Corollary 4.1. *Assume that* $q \in S_r^{2r}$ *with* $r > \frac{3}{2}$. *If there exist* $t_1, t_2 \in R$ *such that for some* $\alpha \in R$,

$$\operatorname{supp} u(\cdot\,, t_j) \subset (-\infty, \alpha), \qquad j = 1, 2,$$

or

$$\operatorname{supp} u(\cdot\,, t_j) \subset (\alpha, \infty), \qquad j = 1, 2,$$

where u is the solution of (24), *then*

$$u(x,t)=0 \quad \textit{for } x\in R, \quad t\in R.$$

COROLLARY 4.2. *Assume that* $q\in S_r^{2r}$ *with* $r>\frac{3}{2}$. *If the solution of* (24) *vanishes on an open set of* $R_x\times R_t$, *then it vanishes everywhere.*

Corollary 4.1 is directly from Theorem 4.1. The proof of Corollary 4.2 is similar to the proof of Corollary 3.1.

COROLLARY 4.3. *Assume that* $q\in S_r^{2r}$ *with* $r>\frac{3}{2}$. *If the solution u of* (24) *satisfies*

$$u(\alpha,t)=0, \quad u_x(\alpha,t)=0, \quad u_{xx}(\alpha,t)=0$$

for some $\alpha\in R$ *and* $t\in(t_1,t_2)$, *then*

$$u(x,t)=0 \quad \textit{for } x\in R, \quad t\in R.$$

Proof. Let

$$u_1(x,t)=\begin{cases} u(x,t) & \text{for } x<\alpha, \\ 0 & \text{for } x\geqq\alpha \end{cases}$$

for $t\in(t_1,t_2)$ and

$$u_2(x,t)=\begin{cases} 0 & \text{for } x<\alpha, \\ u(x,t) & \text{for } x\geqq\alpha \end{cases}$$

for $t\in(t_1,t_2)$. Then both u_1 and u_2 are solutions of the KdV equation for $t\in(t_1,t_2)$. By Theorem 4.1,

$$u_1(x,t)=u_2(x,t)=0 \quad \text{for } x\in R, \quad t\in(t_1,t_2)$$

which implies that

$$u(x,t)=0 \quad \text{for } x\in R, \quad t\in R.$$

Remark. Corollaries 4.2 and 4.3 may also be derived from Saut and Scheurer's results in the same way as Theorem 1.1.

Now we consider a class of generalized solutions of the KdV equation.

DEFINITION. $u\in L^2_{\text{loc}}(R\times R)$ $(L^2_{\text{loc}}((0,T)\times R))$ is called a generalized solution of the KdV equation if for any $\phi\in C_0^\infty(R\times R)$ $(\phi\in C_0^\infty((0,T)\times R))$,

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(u\phi_t+\frac{1}{2}u^2\phi_x+u\phi_{xxx}\right)dx\,dt=0$$

$$\left(\int_0^T\int_{-\infty}^{+\infty}\left(u\phi_t+\frac{1}{2}u^2\phi_x+u\phi_{xxx}\right)dx\,dt=0\right).$$

S. N. Kruzhkov and A. V. Faminskii [7] have proved the following theorem.

THEOREM (Kruzhkov and Faminskii). *If* $q\in L^2(R)$ *and satisfies*

$$\int_0^\infty z^{3/2}q^2(x)\,dx<\infty,$$

then the KdV *equation has a unique generalized solution u satisfying*

$$\operatorname*{ess\,sup}_{t\in(0,T)}\left\{\int_{-\infty}^{+\infty}u^2(x,t)\,dx+\int_0^\infty x^{3/2}u^2(x,t)\,dx\right\}<\infty$$

and

$$\lim_{t\to0,t\in E_T}\int_{-\infty}^{+\infty}u(x,t)w(x)\,dx=\int_{-\infty}^{+\infty}q(x)w(x)\,dx$$

for any $w \in C_0^\infty(R)$, where $E_T = (0, T) \backslash \Omega$ and Ω is a subset of measure zero in $(0, T)$ such that for any $t \in E_T$ $u(x, t)$ is defined almost everywhere on R.

In addition, if for some $\varepsilon > 0$ and integer $p \geqq 0$ and $q \geqq 0$ such that

$$\int_0^\infty x^{3p+q+1/2+\varepsilon} q^2(x)\, dx < \infty,$$

then the solution u possesses continuous derivatives $\partial^{k+l} u(x, t)/\partial x^p \partial t^k$ for $0 \leqq k \leqq p$, $0 \leqq l \leqq q$.

THEOREM 4.2. *Let u be a generalized solution of the* KdV *equation on $R \times R$ and for any $T > 0$,*

$$\operatorname*{ess\ sup}_{t \in (-T, T)} \left\{ \int_{-\infty}^{+\infty} u^2(x, t)\, dx + \int_0^\infty x^{3/2} u^2(x, t)\, dx \right\} < \infty.$$

If there exist $t_1 < t_2$ such that for any $w(x) \in C_0^\infty(R)$,

$$\lim_{t \to t_1^+} \int_{-\infty}^{+\infty} u(x, t) w(x)\, dx = \int_{-\infty}^{+\infty} u(x, t) w(x)\, dx$$

and for some $-\infty < a < b < \infty$,

$$\operatorname{supp} u(x, t_j) \subset (a, b), \qquad j = 1, 2$$

then

$$u(x, t) = 0 \quad \textit{for } x \in R, \quad t \in R.$$

Proof. Let

$$q(x) = u(x, t_1), \qquad p(x) = u(x, t_2).$$

Choose $T > t_2$. Then u is a solution of

$$u_t + u u_x + u_{xxx} = 0, \qquad u(x, t_1) = q(x)$$

for $x \in R$ and $t \in (t_1, T)$. Using the compact support of q with the Kruzhkov and Faminskii Theorem, we have $u \in C^\infty((t_1, T) \times R)$. Especially, $u(\cdot, t_2) = p(x) \in C^\infty(R)$, which implies $p \in S(R)$ since p has compact support. According to [15], there exists a $v(x, t) \in S(R)$ for any $t \in R$, which solves the KdV equation and $v(x, t_2) = p(x)$. By uniqueness, we have

$$v(x, t) = u(x, t) \quad \text{for } x \in R, \quad t \in R.$$

Applying Theorem 4.1, we obtain

$$u(x, t) = 0 \quad \text{for } x \in R, \quad t \in R.$$

The proof is completed. □

Similarly, we have the following theorem.

THEOREM 4.3. *Let $u \in L^\infty_{\text{loc}}(R, H^s(R))$, $s > \frac{3}{2}$, be a solution of the* KdV *equation. If there exist $t_1 < t_2$ such that for some $a, b \in R$,*

$$\operatorname{supp} u(\cdot, t_j) \subset (a, b), \qquad j = 1, 2,$$

then

$$u(x, t) = 0 \quad \textit{for } x \in R, \quad t \in R.$$

Proof. Let $p(x) = u(x, t_2)$ and $q(x) = u(x, t_1)$. Then u solves

$$u_t + uu_x + u_{xxx} = 0, \qquad u(x, t_1) = q(x)$$

for $x \in R$, $t \in (t_1, \infty)$.

Since q has compact support,

$$\int_{-\infty}^{+\infty} q^2 e^{2bx}\, dx < \infty \quad \text{for some } b > 0.$$

Thus, by a result in [6], $u(x, t) \in C^\infty(R)$ for any $t > t_1$. In particular, $p(x) = u(x, t) \in C^\infty(R)$, which implies $p \in S(R)$ since $p(x)$ has compact support. Therefore, $u(x, t) \in S(R)$ for any $t \in R$ [15]. Then Theorem 4.1 yields

$$u(x, t) = 0, \qquad x \in R, \quad t \in R.$$

The proof is completed. □

Remark. Consider the modified KdV equation

$$v_t - 6v^2 v_x + v_{xxx} = 0.$$

Let

$$u = -\tfrac{1}{6}(v^2 + v_x). \tag{38}$$

Then u solves

$$u_t + uu_x + u_{xxx} = 0.$$

Equation (38) corresponds to the Miura transformation. Using this transformation, all the results we have obtained for the KdV equation are also true for the modified KdV equation. For instance, if v solves the modified KdV equation and has compact support at two different times, then

$$u = -\tfrac{1}{6}(v^2 + v_x)$$

solves the KdV equation and has compact support at two different times. By Theorem 4.2, u vanishes identically. Hence

$$v^2(x, t) + v_x(x, t) = 0 \quad \text{for any } x \text{ and } t,$$

which implies that v vanishes identically.

Acknowledgment. The author thanks Professors J. L. Bona, D. L. Russell, J. C. Saut, and M. C. Shen for helpful conversations related to the work presented here.

REFERENCES

[1] J. L. BONA AND R. SMITH, *The initial value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London, 278 (1975), pp. 555-604.

[2] A. COHEN, *Existence and regularity for solutions of the Korteweg-de Vries equation*, Arch. Rational Mech. Anal., 71 (1979), pp. 143-175.

[3] P. DEIFT AND E. TRUBOWITZ, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121-251.

[4] D. JERISON, *Unique continuation and Carleman-type inequalities*, in Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J. L. Lions, eds., Pitman Res. Notes Math., Ser. 166, Boston, 1988, pp. 121-131.

[5] T. KATO, *On the Korteweg-de Vries equation*, Manuscripta Math., 28 (1979), pp. 80-99.

[6] ———, *On the Cauchy problems for the (generalized) Korteweg-de Vries equations*, Stud. Appl. Math., Adv. in Math. Suppl. Stud., 8 (1983), pp. 93-128.

[7] N. Kruzhkov and A. V. Faminskii, *Generalized solutions to the Cauchy problems for the Korteweg-de Vries equation*, Math. USSR-Sh, 48 (1984), pp. 93–138.

[8] S. Mizohata, *Unicité du prolongement des solutions pour quelques opératurs différentiels paraboliques*, Mem. Coll. Sci. Univ. Sér A, 31 (1958), pp. 219–239.

[9] R. Miura, *The Korteweg-de Vries equation: A survey of results*, SIAM Rev., 18 (1976), pp. 416–459.

[10] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[11] J. L. Saut and R. Temam, *Remarks on the* KdV *equation*, Israel J. Math., 24 (1976), pp. 78–87.

[12] J. L. Saut and B. Scheurer, *Un théorème dè prolongement unique pour de opérateurs elliptiques dont les coefficients ne sont par localement bornés*, C. R. Acad. Sci. Paris Sér. A, 290 (1980), pp. 589–599.

[13] J. L. Saut and B. Scheurer, *Unique continuation for some evolution equations*, J. Differential Equations, 66 (1987), pp. 118–139.

[14] H. Shapiro and A. L. Shields, *Usual topological properties of the Nevanlinna class*, Amer. J. Math., 97 (1976), pp. 915–936.

[15] M. Tsutsumi, *Weighted Sobolev spaces and rapidly decreasing solutions of some nonlinear dispersive wave equations*, J. Differential Equations, 42 (1981), pp. 260–281.

[16] S. Tanaka, *The Korteweg-de Vries equation: construction of solutions in terms of scattering data*, Osaka J. Math., 11 (1974), pp. 49–59.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 72-80, January 1992

005

A FAMILY OF STABLE EQUILIBRIA IN BIFURCATION WITH SPHERICAL SYMMETRY*

E. BARANY†‡ AND I. MELBOURNE†§

Abstract. It is shown that asymptotically stable branches of equilibria may generically bifurcate from a spherically symmetric solution that loses stability to spherical harmonics of order l for any odd l. The problem is reduced to one of evaluation of certain Clebsch-Gordan coefficients. The evaluation uses methods from the quantum mechanical theory of angular momentum.

Key words. steady-state bifurcation, spherical symmetry, asymptotic stability, Clebsch-Gordan coefficients

AMS(MOS) subject classifications. 58F14, 34C35

Introduction. Steady-state bifurcation with spherical symmetry appears in several physical contexts; see Chossat, Lauterbach, and Melbourne [1990] and the references therein. Generically, the problem reduces to a bifurcation equation on a $(2l+1)$-dimensional space spanned by the spherical harmonics of order l (Ihrig and Golubitsky [1984]; Golubitsky, Stewart, and Schaeffer [1988]). The induced action of $O(3)$ is the natural action on the space of spherical harmonics.

In this paper we establish the existence of a family of stable equilibria occurring in generic steady-state bifurcation with $O(3)$ symmetry in the cases where l is odd, $l \geqq 3$. Previously, branches of stable equilibria had only been computed in the specific cases $l = 1, 3$, and 5 (see Chossat et al. [1990]). In the notation and terminology of the above references, the stable solutions have (maximal) isotropy D_{2l}^d.

Up to a point, the methods and results parallel those of Chossat and Lauterbach [1989], who demonstrated the *instability* of the so-called *axisymmetric* solutions for all odd $l \geqq 3$. (The axisymmetric solutions are also unstable for even l, as shown by Ihrig and Golubitsky [1984].) Using the group-theoretical techniques of the above references, it is possible to reduce questions of existence and stability of certain branches of equilibria to conditions on the Taylor coefficients of a vector field with $O(3)$-symmetry.

Where this paper differs from prior work in this area is in its emphasis on methods from the quantum mechanical theory of angular momentum; see, for example, Biedenharn and Louck [1981] or Wigner [1959]. In particular, we are able to obtain general expressions for the relevant Taylor coefficients by evaluating certain Clebsch-Gordan coefficients. We note that Chossat [1983] used similar techniques to establish his example of submaximal branching. See also Barany [1988].

It transpires that these techniques are more flexible and efficient than the algorithm of Sattinger [1979] which is used, for example, in Chossat and Lauterbach [1989] and Chossat et al. [1990]. The result proved in this paper was easy to conjecture on the basis of previous calculations (especially Chossat et al. [1990] for $l = 3$ and 5). We expect that the techniques used here will lead to a variety of new results for bifurcation with spherical symmetry, and this will be the subject of future work.

* Received by the editors July 30, 1990; accepted for publication March 25, 1991.

† Department of Mathematics, University of Houston, Houston, Texas 77204-3476.

‡ The research of this author was supported in part by Texas Advanced Research Program grant ARP-1100.

§ The research of this author was supported in part by National Science Foundation Defense Advanced Research Projects Agency grant DMS-8700897.

The remainder of this paper falls into two sections. In the first section we set up the equivariant bifurcation theory of Golubitsky, Stewart, and Schaeffer [1988] in the context of spherical symmetry, Ihrig and Golbitsky [1984] (see also Chossat et al. [1990]). In particular, we explain how the natural representations of $O(3)$ arise and define the family of branches of equilibria that will be proven stable. We also perform the calculations which reduce the computation of existence and stability to evaluation of certain cubic order Taylor coefficients. It is shown that the bifurcating equilibria are stable if the branching is supercritical (the equilibria exist to the right of the bifurcation point) and $l-1$ eigenvalues $\mu_0, \cdots, \mu_{l-2}$ are negative simultaneously.

In § 2 we compute the required Taylor coefficients by evaluating certain Clebsch-Gordan coefficients. It is known that there are $[l/3]+1$ independent cubic mappings with $O(3)$-symmetry. These mappings can be ordered with the first cubic being $|x|^2x$. The second mapping is the one that drives the stability result in the cases $l=3$ and 5 considered in Chossat et al. [1990]. It is natural to concentrate on this second mapping for general l. On setting the coefficients of the remaining $[l/3]-1$ terms equal to zero, we obtain the following expressions for the μ_m:

$$\mu_m = \beta(l^2-m^2), \qquad 0 \leqq m \leqq l-3,$$

$$\mu_{l-2} = \beta(4l-6),$$

where β is the coefficient of the second equivariant mapping. Thus the eigenvalues all have the same sign, sgn(β). Since the eigenvalues depend continuously on the coefficients of the equivariant mappings, we obtain stability of the equilibria for an open set of symmetric vector fields.

1. Steady-state bifurcation with $O(3)$-symmetry. In this section we begin by providing the necessary group-theoretic background for the analysis of steady-state bifurcation with $O(3)$-symmetry. See Golubitsky, Stewart and Schaeffer [1988] for the general context of bifurcation with symmetry, and Ihrig and Golubitsky [1984] and Chossat, Lauterbach, and Melbourne [1990] for more details in the specific case of this section. We then proceed to compute the existence and stability of equilibria with isotropy D_{2l}^d in terms of Taylor coefficients of a vector field with $O(3)$ symmetry. There are three topics which we discuss to the extent that they concern us in this paper: (i) reduction to an absolutely irreducible representation, (ii) existence of equilibria, and (iii) stability of equilibria.

(i) *Reduction.* Consider the ODE

$$\frac{dz}{dt} = F(z, \lambda)$$

where $F: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is $O(3)$-*equivariant*, that is,

$$F(\gamma z, \lambda) = \gamma F(z, \lambda) \quad \text{for all } \gamma \in O(3), \quad z \in \mathbb{R}^n, \quad \lambda \in \mathbb{R}.$$

Suppose that $F(0, \lambda) \equiv 0$ so that $z=0$ is a trivial solution. This solution is asymptotically stable if all eigenvalues of the linearization $L = (d_zF)_{0,0}$ have negative real part, and unstable if any eigenvalue has positive real part. We say that a *steady-state bifurcation* takes place at $(z, \lambda) = (0, 0)$ if the trivial solution loses stability by an eigenvalue crossing the imaginary axis at zero. A Lyapunov-Schmidt or center manifold reduction leads to reduced equations on $V = \ker(L)$. These reduction methods can also be applied to more complicated evolution equations on infinite-dimensional Banach spaces. From now on we assume that the reduction has been performed and that $\mathbb{R}^n = V$.

The theory for equivariant steady-state bifurcation established by Golubitsky et al. [1988] implies that generically $O(3)$ acts absolutely irreducibly on V (absolute irreducibility means that the only commuting matrices L: $V \to V$ are scalar multiples of the identity). For each positive integer l there are precisely two $(2l+1)$-dimensional absolutely irreducible representations of $O(3)$, which we now describe.

Write $O(3)=SO(3)\oplus\{\pm I\}$. For each l, there is a $(2l+1)$-dimensional irreducible representation of $SO(3)$ on the space V_l of spherical harmonics of order l. These are the only irreducible representations of $SO(3)$ and moreover they are absolutely irreducible. To each of these representations, there correspond two absolutely irreducible representations of $O(3)$, $-I$ acting either trivially or as minus the identity. The "natural" representation is the one on the space of spherical harmonics V_l, and it has $-I$ acting trivially when l is even and nontrivially when l is odd. It is the natural representation that typically occurs in physical applications.

A more concrete description of the representations of $O(3)$ can be found, for example, in Sattinger [1979] and Chossat et al. [1990], or in Beidenharn and Louck [1981]. Complexified coordinates are given by $\{z_{-l}, \cdots, z_0, \cdots, z_l\}$ with the *reality condition*

$$z_{-m}=(-1)^m \bar{z}_m. \tag{1.1}$$

We can only describe explicitly the action of one maximal torus in $SO(3)$, which we denote by $SO(2)$, and the corresponding copy of $O(2)$ (also in $SO(3)$). This copy of $O(2)$ contains rotations θ in a plane and a rotation κ which restricts to a reflection in the plane of rotation. The action of $O(2)$ is given by

$$\theta \cdot z_m = e^{im\theta} z_m, \tag{1.2a}$$

$$\kappa \cdot z_m = (-1)^{m+l} z_{-m} = (-1)^l \bar{z}_m. \tag{1.2b}$$

We also record the action of $-I$:

$$-I \cdot z_m = (-1)^l z_m. \tag{1.2c}$$

Finally, we comment on the structure of the $O(3)$-equivariant mapping up to cubic order in the case of l odd. Since the action of $O(3)$ is absolutely irreducible, the only commuting linear maps are scalar multiples of the identity. There is a steady-state bifurcation at $\lambda=0$, so after reparametrization we may assume that the linear term has the form λz. Since l is odd, $-I$ acts as minus the identity, and there are no equivariant mappings of even order.

We now turn to the cubic mappings. It is not difficult to determine the restrictions given by the maximal torus $SO(2)$ described above. If we write a general cubic equivariant C in components $(C_{-l}, \cdots, C_l)$, then

$$C_m(z)=\sum_{\substack{i\leqq j\leqq k\\ i+j+k=m}} (ijk) z_i z_j z_k. \tag{1.3}$$

The reality condition (1.1) gives

$$C_{-m}(z)=(-1)^m \overline{C_m(z)}.$$

The main problem lies in determining the restrictions placed on the coefficients (ijk) by the remaining two dimensions of $SO(3)$. In § 2 we find sufficiently many of the restrictions to prove our main result.

In what follows it suffices to truncate F at cubic order. Existence and stability of the equilibria with isotropy D^d_{2l} is determined by the truncated vector field $\lambda z + C(z)$.

(ii) *Existence.* Let $z \in V$ and define the *isotropy subgroup* of z to be

$$\Sigma_z = \{\gamma \in O(3) \mid \gamma z = z\}.$$

The *fixed-point subspace* Fix (Σ) of an isotropy subgroup is defined to be

$$\text{Fix}\,(\Sigma) = \{z \in V \mid \sigma z = z \text{ for all } \sigma \in \Sigma\}.$$

An $O(3)$-equivariant map $F\colon V \times \mathbb{R} \to V$ restricts to a map

$$F|_{\text{Fix}\,(\Sigma)\times\mathbb{R}}\colon \text{Fix}\,(\Sigma) \times \mathbb{R} \to \text{Fix}\,(\Sigma). \tag{1.4}$$

Thus the problem of finding solutions with isotropy Σ reduces to a problem in Fix (Σ). In particular, when dim Fix $(\Sigma) = 1$, the equivariant branching lemma guarantees the existence of a unique branch of equilibria with isotropy Σ.

In this paper, we restrict to the case of l odd and consider the isotropy subgroup D_{2l}^d which is generated by $-I\cdot(\pi/l)$ and κ where the first element is the composition of $-I$ and a rotation in $SO(2)$ through angle π/l. It follows from (1.2) that the action of D_{2l}^d on the complexified space of spherical harmonics is given by

$$-I\cdot(\pi/l)\cdot z_m = -e^{im\pi/l}\,z_m, \tag{1.5a}$$

$$\kappa \cdot z_m = -\bar{z}_m. \tag{1.5b}$$

If we write $z_m = x_m + iy_m$ then it is easy to check that

$$\text{Fix}\,(D_{2l}^d) = \{iy_l\}. \tag{1.6}$$

In particular, dim Fix $(D_{2l}^d) = 1$ and the equivariant branching lemma guarantees a unique branch of solutions in Fix (D_{2l}^d). For $l \geqq 3$, D_{2l}^d is the largest subgroup of $O(3)$ that fixes iy_l, and so it is an isotropy subgroup (Ihrig and Golubitsky [1984]). Hence there exist solutions with isotropy D_{2l}^d.

Restricting F as in (1.4) and using (1.6), we can compute the branching equation for equilibria with isotropy D_{2l}^d by solving $F_l(iy_l) = 0$. In the notation of (1.3) we obtain

$$\lambda = (-lll)y_l^2 + O(y_l^4). \tag{1.7}$$

(iii) *Stability.* The principle of (orbital) linearized stability implies that in order to establish asymptotic stability of a branch of solutions (z, λ) it is sufficient to show that the eigenvalues of the Jacobian $(dF)_{z,\lambda}$ all have negative real part (with the exception of eigenvalues that are forced to vanish by the action of the group).

The main group-theoretic tool for computing the eigenvalues of such a matrix is the *isotypic decomposition.* Recall that the equivariant vector field F satisfies

$$F(\gamma z, \lambda) = \gamma F(z, \lambda) \quad \text{for all } \gamma \in O(3), \quad z \in V, \quad \lambda \in \mathbb{R}.$$

Differentiation of this identity yields

$$(dF)_{\gamma z,\lambda}\gamma = \gamma(dF)_{z,\lambda}. \tag{1.8}$$

Taking $\gamma \in \Sigma$, we see that $(dF)_{z,\lambda}$ commutes with elements of Σ when $z \in \text{Fix}\,(\Sigma)$.

Now we can decompose the space of spherical harmonics into a direct sum:

$$\{iy_l\} \oplus \{x_l\} \oplus \{z_{l-1}, z_{-(l-1)}\} \oplus \cdots \oplus \{z_1, z_{-1}\} \oplus \{z_0\}. \tag{1.9}$$

Each of the summands is invariant under the action of D_{2l}^d. Moreover, using (1.5) it is easy to check that D_{2l}^d acts absolutely irreducibly and nonisomorphically on each factor. Thus (1.9) is the isotypic decomposition for D_{2l}^d and, by (1.8), $(dF)_{iy_l,\lambda}$ maps each summand (or isotypic component) into itself. It follows that the eigenvalues of $(dF)_{iy_l,\lambda}$ can be computed by restriction to each isotypic component.

General theory guarantees that the eigenvalue in $\{iy_l\}$ is -2λ where λ is as computed in (1.7). We call the corresponding eigenvalue the *branching eigenvalue.* This eigenvalue is stable if and only if the branch of equilibria bifurcates supercritically (that is, exists for positive λ). There are three eigenvalues that are forced to be zero by the group. These lie in $\{x_l\}$ and $\{z_{l-1}, z_{-(l-1)}\}$. It is easy to check the eigenvalues in $\{iy_l\}$ and $\{x_l\}$ by explicit computations.

It remains to compute the eigenvalues in $\{z_0\}$ and $\{z_m, z_{-m}\}$, $m=1,\cdots,l-2$. We will include the case $m=l-1$ since the structure of the calculations is identical. Since D^d_{2l} acts absolutely irreducibly on each $\{z_m, z_{-m}\}$ it follows that the corresponding eigenvalues are double. Hence we need only compute

$$\begin{aligned}\mu_m &= \frac{\partial F_m}{\partial z_m}(iy_l, \lambda), \qquad m=0,\cdots,l-1,\\ &= \lambda + (-lml)z_{-l}z_l + \cdots \big|_{z_l=iy_l}\\ &= [(-lll)-(-lml)]y_l^2 + O(y_l^4).\end{aligned}$$

Of course, we already have that $\mu_{l-1}=0$ and so $(-lll)=(-l, l-1, l)$. In § 2 we show that the expressions $[(-lll)-(-lml)]$, $0 \leqq m \leqq l-2$, are simultaneously negative for a nonempty open set of cubic equivariant mappings C.

2. Evaluation of the eigenvalues. Recall from § 1 that we must show that there is a cubic mapping for which the $l-1$ expressions

$$(2.1) \qquad \mu_m = (-lll)-(-lml), \qquad 0 \leqq m \leqq l-2$$

are simultaneously negative. Here, $(-lml)$ appears in the mth component of the bifurcation equations and is the coefficient of $z_{-l}z_m z_l$.

It is known that there are $r=[l/3]+1$ linearly independent equivariant cubic mappings $c_1, \cdots, c_r$ (see Lévy-Leblond and Lévy-Nahas [1965] and Chossat and Lauterbach [1989] for independent and different proofs of this fact). Hence we may write the general equivariant cubic as

$$C = \sum \alpha_i C_i, \qquad \alpha_1, \cdots, \alpha_r \in \mathbb{R}.$$

The first cubic C_1 can be taken to be $|x|^2x$ and appears for all l. The expressions μ_m in (2.1) are easily seen to be independent of C_1. When $l=3$ and 5, there is a second mapping C_2, and it transpires that the expressions in (2.1) are all positive multiples of α_2 (Chossat et al. [1990]). Hence the D^d_{2l} solutions are stable when $\alpha_2 < 0$.

Now, the C_i can be ordered, and we will show that, for all l, on setting $\alpha_i = 0$, $i \geqq 3$, the expressions in (2.1) are positive multiples of α_2. Hence we have stability when $\alpha_2 < 0$, $\alpha_i = 0$, $i \geqq 3$. By continuity, we have stability for an open set of equivariant cubics.

We now describe our techniques for computing nonlinear equivariant mappings. The space of pth order equivariant mappings can be identified with the space of p-linear maps that (a) transform like the lth representation of $O(3)$, and (b) are symmetric, that is, fixed by S_p, the symmetric group on p symbols. Hence the first step is to take the p-fold tensor product of the lth representation of $O(3)$ and reduce simultaneously to that representation and by S_p. The expansion coefficients appearing in the tensor products are called Clebsch-Gordan (CG) coefficients; see, for example, Biedenharn and Louck [1981], Miller [1972], or Sattinger [1979].

Where our methods differ from those of Sattinger is in the use of results from the theory of angular momentum in quantum mechanics to compute CG coefficients. The advantage of these methods is that it is possible to compute each coefficient individually, with each independent equivariant mapping taken one at a time. In physics there is a wealth of literature on these computations, and we shall quote many formulas. The reader is referred to Biedenharn and Louck [1981] or Wigner [1959] for a more detailed discussion and derivation of these formulas.

There is one possible cause of confusion that we wish to clear up from the outset. The results we quote apply to a conventionally normalized form of the CG coefficients. These may differ from other "CG coefficients" by a factor that depends only on the representation of $O(3)$ and the independent equivariant mapping being evaluated. This factor may be absorbed into α_2, and so does not affect the bifurcation theory.

The remainder of this section will fall into two subsections: (i) reduction of the three-fold tensor product, and (ii) evaluation of the required Clebsch–Gordan coefficients.

(i) Consider $z^{(L)} \in V_L$ in the basis of spherical harmonics of order L,

$$z^{(L)} = \sum_{m=-L}^{L} z_m^{(L)} Y_L^m. \tag{2.2}$$

If z is the direct sum $z = \oplus z^{(L)}$ over all nonnegative integers L, then

$$z_m^{(L)} = \int d\Omega\, \overline{Y_L^m} z$$

are the expansion coefficients. In order to compute the quadratic equivariant mappings, we wish to know how the expansion coefficients $(z^2)_m$ for the square of z will depend on the original z_m. We have the following integral over three spherical harmonics:

$$\int d\Omega\, \overline{Y_L^m} Y_{l_1}^{m_1} Y_{l_2}^{m_2} = \left(\frac{(2l_1+1)(2l_2+1)}{4\pi(2L+1)}\right)^{1/2} C_{0,\,0,\,0}^{l_1, l_2, L} C_{m_1, m_2, m}^{l_1,\ l_2,\ L}. \tag{2.3}$$

The $C_{m_1, m_2, m}^{l_1,\ l_2,\ L}$ are CG coefficients that satisfy the conventional choice of normalization of the highest-weight coefficients $C_{m_1, m_2, L}^{l_1,\ l_2,\ L}$ where $L = m_1 + m_2$; see Biedenharn and Louck [1981]. To simplify the notation, we define

$$N_{l_1 l_2}^L = \left(\frac{(2l_1+1)(2l_2+1)}{4\pi(2L+1)}\right)^{1/2} C_{0,\,0,\,0}^{l_1, l_2, L}. \tag{2.4}$$

Then from (2.2) and (2.3), we can write

$$Y_L^{m_1} Y_L^{m_2} = \sum_{L=0}^{2l} N_{ll}^L C_{m_1, m_2, m_1+m_2}^{l,\ \ l,\ \ L} Y_L^{m_1+m_2}. \tag{2.5}$$

In deriving (2.5) we have used two properties of the CG coefficients, namely, that the coefficient $C_{m_1, m_2, m}^{l_1,\ l_2,\ L}$ vanishes unless $m = m_1 + m_2$, and that L satisfies the "triangle inequality" $|l_1 - l_2| \leqq L \leqq l_1 + l_2$.

Now squaring (2.2) and using (2.5) we can identify the quadratic expansion coefficients from the lth representation that transform as in the Lth representation. Let $z_m^{(l);2;(L)}$ be shorthand for $(z^{(l)})_m^{2(L)}$. Then

$$z_m^{(l);2;(L)} = \sum_{m_1+m_2=m} N_{ll}^L C_{m_1, m_2, m}^{l,\ \ l,\ \ L} z_{m_1}^{(l)} z_{m_2}^{(l)}. \tag{2.6}$$

Taking $L = l$ in (2.6) gives the (unique) quadratic equivariant $z^{(l);2;(l)}$. This equivariant vanishes for odd l, as can be seen from the property

$$C_{m_1, m_2, m}^{l_1,\ l_2,\ l} = (-1)^{l_1+l_2-l} C_{m_2, m_1, m}^{l_2,\ l_1,\ l}. \tag{2.7}$$

In order to generate the cubic order equivariants we iterate the process and consider the three-fold tensor product $z^{(l);3;(l)}$ obtained from the product of $z^{(l)}$ with its square

$$z^{(l);2} = \sum_{L=0}^{2l} \sum_{m=-L}^{L} z_m^{(l);2;(L)} Y_L^m. \tag{2.8}$$

We must first reduce to the $(2l+1)$-dimensional representation of $O(3)$. Observe that several terms in this product transform like this representation: one for each of the $(2l+1)$ values of L in (2.8). It is possible to show that the first $[l/3]+1$ even values of L generate the independent equivariant cubics. Since this fact is not required to prove our main result, we defer the proof to a later paper.

Writing $z^{(l);3;(l)}(L)$ as the part of the cubic that transforms as l and comes from L, we get

$$z_m^{(l);3;(l)}(L) = N_{ll}^L N_{lL}^l \sum_{m_3+M=m} \sum_{m_1+m_2=M} C_{m_3,\, M,\, m}^{l,\ L,\ l} C_{m_1,\, m_2,\, M}^{l,\ l,\ L} z_{m_1}^{(l)} z_{m_2}^{(l)} z_{m_3}^{(l)}. \tag{2.9}$$

It follows from (2.7) that the sum in (2.9) vanishes when L is odd. From now on we assume that L is even (shortly we shall specialize to the case $L=2$). To obtain the (ijk) coefficients defined in § 1, we reduce by S_3; that is, we symmetrize (2.9) over the subscripts m_1, m_2, and m_3. There are three cases:

i, j, k all unequal:

$$\begin{aligned}(ijk)_L = 2N_{ll}^L N_{lL}^l (&C_{i,j+k,\,i+j+k}^{l,L,\ l} C_{j,\,k,j+k}^{l,l,\ L} + C_{j,\,i+k,\,i+j+k}^{l,L,\ l} C_{i,\,k,\,i+k}^{l,l,\ L}\\ &+ C_{k,\,i+j,\,i+j+k}^{l,\ L,\ l} C_{i,j,\,i+j}^{l,l,L}),\end{aligned} \tag{2.10a}$$

$i \neq j = k$:

$$(ijj)_L = N_{ll}^L N_{lL}^l (C_{i,2j,\,i+2j}^{l,L,\,l} C_{j,j,2j}^{l,l,L} + 2C_{j,\,i+j,\,i+2j}^{l,L,\ l} C_{i,j,\,i+j}^{l,l,L}), \tag{2.10b}$$

$i = j = k$:

$$(iii)_L = N_{ll}^L N_{lL}^l C_{i,2i,3i}^{l,L,\,l} C_{i,\,i,2i}^{l,l,L}. \tag{2.10c}$$

(ii) We now explicitly compute the coefficients $(-lml)_2$, $0 \leqq m \leqq l$, as required (in fact, there is some redundancy, as we already know that $(-l,l-1,l)=(-lll)$). The choice $L=2$ is motivated by the results for $l=3$ and 5, and simplifies the calculations due to the (obvious) property

$$C_{m_1,\, m_2,\, m}^{l_1,\ l_2,\ L} = 0 \quad \text{for } |m| > L. \tag{2.11}$$

In particular, if the common factor $2N_{ll}^2 N_{l2}^l$ is absorbed into the coefficient α_2 of the cubic equivariant, the equations in (2.10) become

$$\begin{aligned}(-lml)_2 &= C_{m,0,m}^{l,\ 2,l} C_{-l,l,0}^{l,\ \ l,2}, & 0 \leqq m \leqq l-3,\\ (-lml)_2 &= C_{m,0,m}^{l,\ 2,l} C_{-l,l,0}^{l,\ \ l,2} + C_{l,\,m-l,\,m}^{l,2,\ \ \ l} C_{-l,\,m,\,m-l}^{l,\ \ l,\ 2}, & m = l-2,\, l-1,\\ (-lll)_2 &= C_{l,0,l}^{l,2,l} C_{-l,l,0}^{l,\ \ l,2}.\end{aligned}$$

A further simplification can be obtained by transforming the CG coefficients using the following two properties:

$$C_{m_1,\, m_2,\, m}^{l_1,\ l_2,\ l} = (-1)^{l_1-m_1} \left(\frac{2l+1}{2l_2+1}\right)^{1/2} C_{m_1,\,-m,\,-m_2}^{l_1,\ l,\ \ l_2},$$

$$C_{m_1,\, m_2,\, m}^{l_1,\ l_1,\ L} = (-1)^L C_{-m_2,\,-m_1,\,-m}^{l_1,\ \ l_1,\ \ L}.$$

Absorbing the common square root (with $l_2=2$) into α_2, we have

$$(2.12)\qquad \begin{aligned} &(-lml)=(-1)^{m+1}C^{l,\,l,\,2}_{l,-l,0}C^{l,\,l,\,2}_{m,-m,0}, && 0\leqq m\leqq l-3,\\ &(-lml)=(C^{l,\,l,\,2}_{l,-m,l-m})^2+(-1)^{m+1}C^{l,\,l,\,2}_{l,-l,0}C^{l,\,l,\,2}_{m,-m,0}, && m=l-2,\,l-1,\\ &(-lll)=(C^{l,\,l,\,2}_{l,-l,0})^2. \end{aligned}$$

At this point, we have reduced the problem to the evaluation of two types of CG coefficients. In particular, it is sufficient to consider "Racah's first form"

$$(2.13)\qquad \begin{aligned} C^{l_1,\,l_2,\,l}_{m_1,m_2,m}&=\delta_{m,m_1+m_2}\left(\frac{(2l+1)(l_1+l_2-l)!(l_1-m_1)!(l_2-m_2)!(l-m)!(l+m)!}{(l_1+l_2+l+1)!(l+l_1-l_2)!(l-l_1+l_2)!(l_1+m_1)!(l_2+m_2)!}\right)^{1/2}\\ &\quad\cdot\sum_t(-1)^{l_1-m_1+t}\left(\frac{(l_1+m_1+t)!(l+l_2-m_1-t)!}{t!(l-m-t)!(l_1-m_1-t)!(l_2-l+m_1+t)!}\right) \end{aligned}$$

where summation is over those t for which all factorials are of nonnegative numbers. Specializing, we find that

$$(2.14a)\qquad C^{l,\,l,\,2}_{m,-m,0}=2(-1)^{m+1}\left(\frac{5(2l-2)!}{(2l+3)!}\right)^{1/2}(3m^2-l^2-l),$$

$$(2.14b)\qquad C^{l,\,l,\,2}_{l,-l+1,1}=\left(\frac{5(2l-2)!}{(2l+3)!}\right)^{1/2}\sqrt{12l}(2l-1),$$

$$(2.14c)\qquad C^{l,\,l,\,2}_{l,-l+2,2}=\left(\frac{5(2l-2)!}{(2l+3)!}\right)^{1/2}\sqrt{24l(2l-1)}.$$

The derivation of these formulas is straightforward. The only contributions from (2.13) in (2.14a) come from $t=0$, 1, and 2. For (2.14b, c) the only contributions come from $t=0$. On substituting the expressions in (2.14) into (2.12) and then into (2.1), we see that

$$\begin{aligned} &\mu_m=12l(2l-1)(l^2-m^2)\alpha_2, \qquad 0\leqq m\leqq l-3,\\ &\mu_{l-1}=12l(2l-1)(4l-6)\alpha_2. \end{aligned}$$

Hence the μ_m are positive multiples of α_2 as required. It is also easy to verify that $\mu_{l-1}=0$ as expected.

REFERENCES

E. Barany [1988], *Algebraic aspects of broken symmetry: irreducible representations of SO(3)*, Ph.D. thesis, Ohio State University, Columbus, OH.

L. C. Biedenharn and J. D. Louck, [1981], *Angular Momentum in Quantum Physics*, Encyclopedia Math. Appl. 8, Addison-Wesley, Reading, MA.

P. Chossat [1983], *Solutions avec symétrie diédrale dans les problèmes de bifurcation invariants par symétrie sphérique*, CR Acad. Sci. Paris, Sér. I, 297, pp. 639-642.

P. Chossat and R. Lauterbach [1989], *The instability of axisymmetric solutions in problems with spherical symmetry*, SIAM J. Math. Anal., 20 pp. 31-38.

P. Chossat, R. Lauterbach, and I. Melbourne [1990], *Steady-state bifurcation with O(3)-symmetry*, Arch. Rational Mech. Anal., 113, pp. 313-376.

M. Golubitsky, I. N. Stewart, and D. G. Schaeffer [1988], *Singularities and Groups in Bifurcation Theory*, Vol. 2, Appl. Math. Sci. 69, Springer-Verlag, New York.

E. Ihrig and M. Golubitsky [1984], *Pattern selection with $O(3)$ symmetry*, Phys. D, 13, pp. 1–33.

J. M. Lévy-Leblond and M. Lévy-Nahas [1965], *Symmetrical coupling of three angular momenta*, J. Math. Phys., 6, pp. 1372–80.

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

D. H. Sattinger [1979], *Group Theoretic Methods in Bifurcation Theory*, Lecture Notes in Math. 762, Springer-Verlag, Berlin.

E. P. Wigner [1959], *Group Theory*, Academic Press, New York.

QUALITATIVE ANALYSIS OF MYELINATED NERVE FIBERS WITH POINT-NODE FITZHUGH–NAGUMO DYNAMIC SYSTEM*

PEI-LI CHEN[†]

Abstract. The equations for the membrane potentials in a point-node myelinated axon fibers model take the form

$$\begin{aligned}
&U_t = U_{xx} - GU, && x \in (0,1) \mod (1),\\
&W = 0, && x \in (0,1) \mod (1),\\
&U_t = M[U_x]_x + F(U) - W, && x = 0 \mod (1),\\
&W_t = \sigma U - \gamma W, && x = 0 \mod (1),\\
&[U]_x = 0, && x = 0 \mod (1),
\end{aligned}$$

where $U = (u_1, u_2, \cdots, u_n)^t$, $W = (w_1, w_2, \cdots, w_n)^t$, $M = \hat{\lambda} I - \alpha B$ and the model dynamics are of FitzHugh–Nagumo type. In this paper, two new results for this model are presented.

In the first result it is shown that this model has two nontrivial solutions and the contracting rectangle technique is used to show that one of these solutions is stable. The second result gives an existence proof for the Cauchy problem associated with this model.

Key words. myelinated axon fibers, FitzHugh–Nagumo, contracting rectangle, reaction diffusion equations

AMS(MOS) subject classifications. 35A05, B40, 92A05

1. Introduction.

Grindrod and Sleeman [9, p. 119] define a myelinated nerve axon as follows: "A myelinated nerve axon consists of axoplasm surrounded by a long cylindrical membrane which is in turn surrounded by a sheath of lipoprotein, called myelin, formed by condensation of Schwann cell membranes. The sheath, a fatty layered tissue, insulates the axon from the external ionic fluid." In each myelin segment, the potential $u(x,t)$ is governed by a diffusion equation of the form

$$u_t = u_{xx} - gu, \tag{1.1}$$

in which the spatial variable x is measured along the axon and the constant $g = 1/R$, where R represents the axoplasmic resistivity.

Again, quoting Grindrod and Sleeman [9, p. 119]: "At approximately millimeter intervals, there are small gaps called nodes of Ranvier. These nodes expose the extracellular fluid to the excitable axon membrane." At the nodes the axon membrane is selectively permeable to the charged ions within the axoplasm and outer ionic fluid. Here the potential $u(x,t)$ is governed by a nonlinear diffusion equation of the form

$$u_t = u_{xx} - J, \tag{1.2}$$

where J represents the ionic current through the membrane.

In the early 1950s, Hodgkin and Huxley published a series of papers on the unmyelinated axon model based on the giant squid axon. In [13] they proposed a system

*Received by the editors August 1, 1989; accepted for publication (in revised form) April 4, 1991. This work was done while the author was a Morrey Assistant Professor in the Department of Mathematics at the University of California at Berkeley, Berkeley, California 94720.

[†]Department of Mathematics, Southern Illinois University at Carbondale, Carbondale, Illinois 62901-4408.

of differential equations, each of which is in the general form of (1.2). Because of the analytic complexity of the Hodgkin–Huxley model, most work on the model, particularly on qualitative behavior of the solutions, has been numerical. To gain some insight into the mathematical phenomena involved in the nerve conduction process FitzHugh [6] and Nagumo, Arimoto, and Yoshizawa [15] introduced a simpler prototype system of the form

$$\begin{aligned} u_t &= u_{xx} + f(u) - w, \\ w_t &= \sigma u - \gamma w, \end{aligned} \tag{1.3}$$

where u is still the membrane potential and w represents a recovery process. The current-voltage relation $-f(u)$ is an "N-shaped" function, sketched in Fig. 1, and σ, γ are recovery constants.

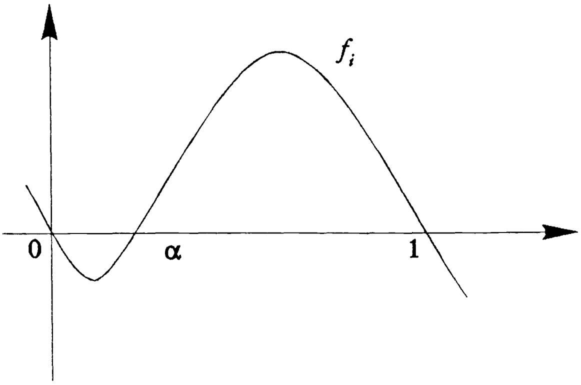


FIG. 1. *Each of $-f_i$ is an N-shaped function.*

Questions of existence and qualitative behavior of solutions to (1.3) have been pursued very actively by many mathematicians (see, for example, [8]–[12], [16], [17]). Grindrod and Sleeman [10] used a contracting rectangle technique (see [4] for a general reference) to give a qualitative analysis of unmyelinated nerve fibers.

Myelinated axons, especially myelinated axon fibers, which are more prevalent in human anatomy than unmyelinated axons, have also been modeled. However, fewer analytical results have been obtained for these models. Some asymptotic behavior was considered in simpler models (e.g., Bell [1], Bell and Cosner [2], Chen and Bell [3], and Grindrod and Sleeman [9]).

The main purpose of this paper is to consider questions related to the equations for the point-node myelinated axon fibers (see [2] and [3] for references). These equations are

$$U_t = U_{xx} - GU, \qquad x \in (0,1) \mod (1), \tag{1.4}$$

$$W = 0, \qquad x \in (0,1) \mod (1), \tag{1.5}$$

$$U_t = M[U_x]_x + F(U) - W, \qquad x = 0 \mod (1), \tag{1.6}$$

$$W_t = \sigma U - \gamma W, \qquad x = 0 \mod (1), \tag{1.7}$$

$$[U]_x = 0, \qquad x = 0 \mod (1), \tag{1.8}$$

where $U = (u_1, u_2, \cdots, u_n)^t \in C^1(\mathbb{R}^+ \times \mathbb{R}, \mathbb{R}^n)$, $W \in C^1(\mathbb{R}^+ \times \mathbb{Z}, \mathbb{R}^n)$, $F(U) = (f_1(u_1), \cdots, f_n(u_n))^t$, $[U_x]_x = U_x(x+, t) - U_x(x-, t)$, $G = \text{Diag}\{g_1, \cdots, g_n\}$ and $M = (\hat{\Lambda} I - \alpha B)$. Here each u_i is the potential of a membrane, each $-f_i$ is an N-shaped function as sketched in Fig. 1, each $g_i > 0$, $\hat{\Lambda}$ and α are positive constants satisfying $\hat{\Lambda} \gg \alpha > 0$, $\sigma = \text{Diag}\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$, $\gamma = \text{Diag}\{\gamma_1, \gamma_2, \cdots, \gamma_n\}$, where σ_i, γ_i are positive constants $(i = 1, 2, \cdots, n)$, I is the $n \times n$ identity matrix, and B is the adjacency matrix for the graph, which is defined as follows.

Following [10] we define Γ to be a graph on n vertices (see Fig. 2) where the adjacency matrix $B = (b_{ij})$, $i, j = 1, \cdots, n$ is defined by

$$(1.9) \qquad b_{ij} = \begin{cases} 0 & \text{if } i = j, \\ 1 & \text{if } i \neq j \text{ and there is an edge} \\ & \qquad \text{in } \Gamma \text{ which connects } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases}$$

Thus B is a symmetric matrix and describes the geometry of the graph Γ. "This concept is often used by chemists where it is sometimes called the topology matrix and is used to describe and classify hydrocarbon molecules. By the previous construction, if Γ is the graph on n vertices, representing n-fibers, the adjacency matrix, B of Γ, describes the ability of separate fibers to interact." See [10, p. 6].

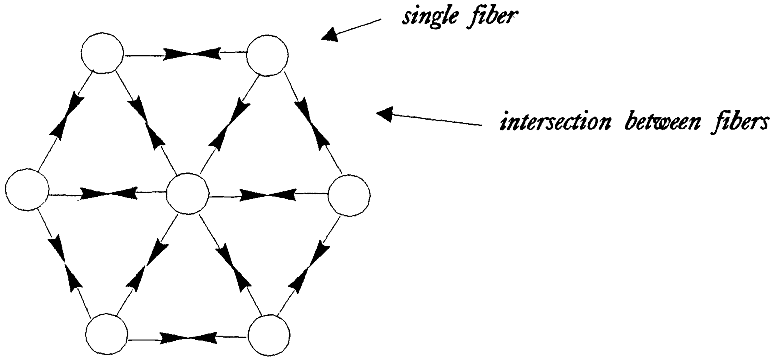


FIG. 2. *Configuration of parallel fibers in cross section.*

Since $\hat{\Lambda} \gg \alpha > 0$ the matrix M is positive definite. Let $\lambda_1, \cdots, \lambda_n$ denote its eigenvalues. It is obvious that $\lambda_i > 0$, $i = 1, \cdots, n$. Let C be the unitary matrix whose columns are the orthonormalized eigenvectors of M, that is, $M = C\Lambda C^t$, where $\Lambda = \text{Diag}\{\lambda_1, \cdots, \lambda_n\}$. Setting $V(x,t) = C^t U(x,t)$ and premultiplying the system (1.4)–(1.8) by C^t, V satisfies the following equations:

$$\begin{aligned}
&(1.10) && V_t = V_{xx} - GV, && x \in (0,1) \mod (1), \\
&(1.11) && V_t = \Lambda[V_x]_x + C^t F(CV) - C^t W, && x = 0 \mod (1), \\
&(1.12) && W_t = \sigma CV - \gamma W, && x = 0 \mod (1), \\
&(1.13) && [V]_x = 0, && x = 0 \mod (1).
\end{aligned}$$

We define the operator G_1 by

$$G_1\begin{pmatrix} V \\ W \end{pmatrix} = \begin{pmatrix} C^tF(CV) - C^tW \\ \sigma CV - \gamma W \end{pmatrix}, \qquad x = 0 \mod (1), \tag{1.14}$$

$$G_1\begin{pmatrix} V \\ W \end{pmatrix} = \begin{pmatrix} -GV \\ 0 \end{pmatrix}, \qquad x \in (0,1) \mod (1), \tag{1.15}$$

and the matrix P by

$$P = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \quad \text{if } x = 0 \mod (1),$$

$$P = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad \text{if } x \in (0,1) \mod (1).$$

System (1.10)–(1.13) can then be rewritten as

$$\begin{pmatrix} V \\ W \end{pmatrix}_t = P \cdot \begin{pmatrix} V \\ W \end{pmatrix}_{xx} + G_1\begin{pmatrix} V \\ W \end{pmatrix}, \qquad x \in (0,1) \mod (1) \tag{1.16}$$

$$\begin{pmatrix} V \\ W \end{pmatrix}_t = P \cdot \begin{pmatrix} [V_x]_x \\ 0 \end{pmatrix} + G_1\begin{pmatrix} V \\ W \end{pmatrix}, \qquad x = 0 \mod (1). \tag{1.17}$$

In §2, by using the Implicit Function Theorem, it is shown that for certain values of the parameters there exist nonzero steady state solutions to (1.16) and (1.17). In §3, we show that there exist contracting rectangles around some special steady state solutions. In §4, we use this result to prove that one of the steady state solutions is approximately stable. In §5, the existence proof of solutions of the system (1.4)–(1.8) is given for the Cauchy problem (see [5] and [3] for references).

2. Steady state solutions. Let (U, W) be the steady state solution of (1.4)–(1.8). In this case, U satisfies the equations

$$U_{xx} - GU = 0, \qquad x \in (0,1) \mod (1), \tag{2.1}$$

$$M[U_x]_x + F(U) - \sigma\gamma^{-1}U = 0, \qquad x = 0 \mod (1). \tag{2.2}$$

The general solution of (2.1) can be expressed as

$$u_i = \frac{[{}_{k+1}u_i \sinh\sqrt{g_i}(x-k) + {}_ku_i \sinh\sqrt{g_i}(k+1-x)]}{\sinh\sqrt{g_i}} \qquad \text{for } x \in (k, k+1),$$

where ${}_ku_i = u_i(k)$ for all $k \in \mathbb{Z}$. Obviously, $(U, W) = (0, 0)$ is a trivial steady state solution. The jump of the first derivative of u_i on each node $x = k$ is given by

$$[u_{ix}]_k = G_i({}_{k+1}u_i - 2\cosh\sqrt{g_i}\ {}_ku_i + {}_{k-1}u_i), \tag{2.3}$$

where $G_i = \sqrt{g_i}/\sinh\sqrt{g_i}$. Let $\hat{G} = \mathrm{Diag}\{G_1, G_2, \cdots, G_n\}$, $\mathrm{Chg} = \mathrm{Diag}\{\cosh\sqrt{g_1}, \cosh\sqrt{g_2}, \cdots, \cosh\sqrt{g_n}\}$, ${}_kU = ({}_ku_1, {}_ku_2, \cdots, {}_ku_n)^t$. Using this notation, (2.3) can be rewritten as

$$[U_x]_k = \hat{G}({}_{k+1}U - 2\mathrm{Chg}_kU + {}_{k-1}U). \tag{2.4}$$

Substituting (2.4) into (2.2), we see that U satisfies

$$\hat{G}({}_{k+1}U - 2\mathrm{Chg}_kU + {}_{k-1}U) + A_\lambda F({}_kU) - A_\lambda\sigma\gamma^{-1}{}_kU = 0, \tag{2.5}$$

$$x = k \in \mathbb{Z},$$

where the matrix $A_\lambda = C\Lambda^{-1}C^t = M^{-1}$. We will now look for a special solution of the form

$$ {}_kU = z \quad \text{for all } k \in \mathbb{Z}, $$

where $z = (z_1, z_2, \cdots, z_n)^t$. In this case, (2.5) is equivalent to

$$ 2\hat{G}(1 - \mathrm{Chg})z + A_\lambda F(z) - A_\lambda \sigma\gamma^{-1} z = 0. $$

In order to find a nonzero solution of the above equation, let us define the functions

$$ \begin{aligned} H_1(z; G, \sigma\gamma^{-1}) &= F(z) - I(z; G, \sigma\gamma^{-1}), \\ I(z; G, \sigma\gamma^{-1}) &= (A_\lambda^{-1} 2\hat{G}(\mathrm{Chg} - I) + \sigma\gamma^{-1})z, \end{aligned} \tag{2.6} $$

where $A_\lambda^{-1} = C\Lambda C^t = M$. Then

$$ D_z H_1(z; G, \sigma\gamma^{-1}) = \begin{pmatrix} f_1' & 0 & 0 & \cdots & 0 \\ 0 & f_2' & 0 & \cdots & 0 \\ 0 & 0 & f_3' & \cdots & 0 \\ & \cdots & & \cdots & \\ 0 & 0 & 0 & \cdots & f_n' \end{pmatrix} - \sigma\gamma^{-1} I - 2A_\lambda^{-1}\hat{G}(\mathrm{Chg} - I). \tag{2.7} $$

Let $\hat{f}_i(z_i) = f_i(z_i) - \sigma_i/\gamma_i z_i$ for $i = 1, 2, \cdots, n$. Then for σ_i/γ_i small there exists $\alpha_1^{(i)}, \alpha_2^{(i)}$ such that

$$ \hat{f}_i(\alpha_1^{(i)}) = \hat{f}_i(\alpha_2^{(i)}) = 0 $$

for $i = 1, 2, \cdots, n$ (cf. Fig. 3). This is equivalent to

$$ H_1(\alpha_1; 0, \sigma\gamma^{-1}) = H_1(\alpha_2; 0, \sigma\gamma^{-1}) = 0, \tag{2.8} $$

where $\sigma\gamma^{-1}$ is fixed and $\alpha_j = (\alpha_j^{(1)}, \alpha_j^{(2)}, \cdots, \alpha_j^{(n)})^t$, $j = 1, 2$. By the definition of $\hat{f}_i$ and $\hat{f}_i'$ at α_1 and α_2, there exists a $g^* > 0$ such that $D_z H_1(\alpha_1; G, \sigma\gamma^{-1})$ is positive definite for $0 < g_i < g^*$ and $D_z H_1(\alpha_2; G, \sigma\gamma^{-1})$ is negative definite for $0 < g_i < g^*$, $i = 1, 2, \cdots, n$. Using the Implicit Function Theorem there exist smooth manifolds $\alpha_1 = \alpha_1(G)$ and $\alpha_2 = \alpha_2(G)$ in a neighborhood $U_0 = \prod_1^n(-q, q)$ of $G = (0, 0, \cdots, 0)^t$ such that

$$ \begin{aligned} &H_1(\alpha_1(G), G, \sigma\gamma^{-1}) = 0, \\ &H_1(\alpha_2(G), G, \sigma\gamma^{-1}) = 0, \qquad G \in U_0. \end{aligned} $$

Hence the following theorem has been proved.

THEOREM 2.1. *The differential equations* (2.1)–(2.2) *have at least two superthreshold* (*nonzero*) *solutions, provided the positive constants* g_i *and* σ_i/γ_i *are small* ($i = 1, 2, \cdots, n$).

If we define $U_{\alpha_i}(x) = (u_{1\alpha_i^{(1)}}(x), u_{2\alpha_i^{(2)}}(x), \cdots, u_{n\alpha_i^{(n)}}(x))^t$ then U_{α_1} and U_{α_2} are two superthreshold solutions with $u_{j\alpha_i^{(j)}}(k) = \alpha_i^{(j)}$ for all $k \in \mathbb{Z}$, where $i = 1, 2$; $j = 1, 2, \cdots, n$.

3. Contracting rectangle. In this section and the next, the idea of "invariant region" or "contracting rectangle" (see [17] and [10]) will be used to reveal the behavior of solutions of (1.4)–(1.8) around the steady state solutions $(U, W) = (0, 0)$ and $(U, W) = (U_{\alpha_2}, \sigma\gamma^{-1}U_{\alpha_2})$. In fact, these two solutions are "attractors," that is, if any solution of (1.4)–(1.8) is close enough to one of the two steady states, then it must be attracted to the steady state solution.

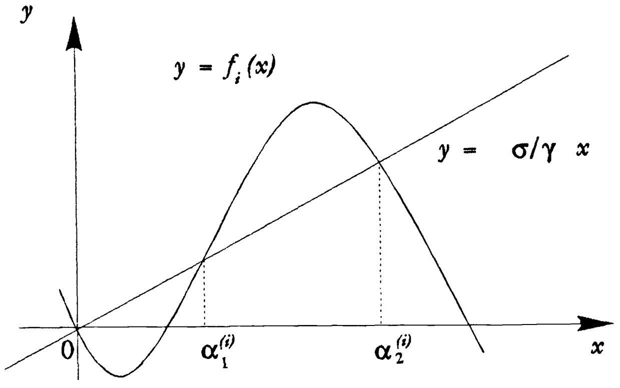


FIG. 3. $\hat{f}_i(\alpha_1^{(i)}) = \hat{f}_i(\alpha_2^{(i)}) = 0$.

DEFINITION 3.1. Let H denote a vector field over $\mathbb{R}^{2n}$ and let S be a bounded convex set in $\mathbb{R}^{2n}$, with boundary ∂S. S is contracting for H if, for every $w \in \partial S$ and the outward normal vector $n(w)$ of ∂S at w, we have

$$H(w) \cdot n(w) < 0.$$

If S is of the form

$$S = S_1 \times S_2, \tag{3.1}$$

where $S_1 = \pi_{j=1}^n[-\overline{u}_j, \overline{u}_j]$ and $S_2 = \pi_{j=1}^n[-\overline{z}_j, \overline{z}_j]$, then S is called a contracting rectangle.

The boundary of S consists of $4n$ "faces" which are denoted as follows.

$S \cap \{(U, W) : u_i = \overline{u}_i\}$ is the upper ith face,
$S \cap \{(U, W) : u_i = -\overline{u}_i\}$ is the lower ith face,
$S \cap \{(U, W) : z_i = \overline{z}_i\}$ is the upper $(n+i)$th face,
$S \cap \{(U, W) : z_i = -\overline{z}_i\}$ is the lower $(n+i)$th face.

3.1. Contracting rectangle around the trivial solution $(U, W) = (0, 0)$. Define $H(U, W) = G_1(U, W)$ as in (1.14) and (1.15). Then either $x \in \mathbb{Z}$ or $x \in \mathbb{R}\backslash\mathbb{Z}$.

3.1.1. Case for $x = 0$ mod(1). Fixing $x = n \in \mathbb{Z}$, the vector field $G_1 : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ defined by (1.14) is the same as (1.1) of [10]; hence we can use the same results as in §1 of [10] where $C = (c_{ij})$ and we define $\hat{C} = (\hat{c}_{ij})$ where $\hat{c}_{ij} = |c_{ij}|$, for $i, j = 1, \cdots, n$.

THEOREM 3.2 (Grindrod and Sleeman [10]). *We assume that the functions f_i $(i = 1, 2, \cdots, n)$ satisfy the condition that*

$$\min_i\{a_i\} > r \max_i\{\sigma_i/\gamma_i\}, \tag{3.2}$$

where $a_i = -f_i'(0)$, $i = 1, \cdots, n$, and $r = \rho(\hat{C}^t\hat{C})$. Then there exists a rectangle S^* of the form (3.1) *and a constant $\kappa_1 > 0$ such that $G_1(U, W) \cdot n(U, W) < -\kappa_1\tau$ on $\partial(\tau S^*)$ for any outward normal $n(U, W)$ to $\partial(\tau S^*)$ at (U, W) and any $\tau \in (0, 1)$.*

3.1.2. Case for $x \in \mathbb{R}\backslash\mathbb{Z}$. Here $(U, 0)$ is on the ∂S if and only if there exists some $u_i = \overline{u}_i$ (or $u_i = -\overline{u}_i$). By (1.15),

$$G_1(U, W) = \begin{bmatrix} -GU \\ 0 \end{bmatrix}.$$

We only consider the case that $(U, 0)^t$ is on the upper ith face. In this case $n(U, 0) = (\ell_i, 0)^t$, where $\ell_i = (0, \cdots, 0, 1, 0, \cdots, 0)^t$ and $G_1(U, 0) \cdot n(U, 0) = -g_i\overline{u}_i$. We have now proved the following lemma.

LEMMA 3.3. *For any rectangle S given by* (3.1), *$\tau \in (0, 1)$, and $\kappa_2 = \min_i\{g_i\}$, we have that*

$$G_1(U, 0) \cdot n(U, 0) \leq -\kappa_2\tau \tag{3.3}$$

on $\partial(\tau S)$ for any outward normal vector $n(U, 0)$ to $\partial(\tau S)$ at $(U, 0)$.

3.2. Contracting rectangle near steady state solution. $(U_{\alpha_2}(x), \sigma\gamma^{-1}U_{\alpha_2}(x))$ Let (U, W) be a solution of (1.4)–(1.8) near the steady state solution $(U_{\alpha_2}, \sigma\gamma^{-1}U_{\alpha_2})$. Let us define the functions

$$V_1(x, t) = U(x, t) - U_{\alpha_2}(x), \qquad x \in (0, 1) \mod (1), \tag{3.4}$$

$$Z_1(n, t) = W(n, t) - \sigma\gamma^{-1}U_{\alpha_2}(n), \qquad n = 0 \mod (1). \tag{3.5}$$

These functions then satisfy the following differential equations

$$V_{1t} = V_{1xx} - GV_1, \qquad x \in (0, 1) \mod (1), \tag{3.6}$$

$$Z_1 = 0, \qquad x \in (0, 1) \mod (1), \tag{3.7}$$

$$V_{1t} = M[V_{1x}]_x + F(U_{\alpha_2} + V_1) - F(U_{\alpha_2}) - Z_1, \qquad x = 0 \mod (1), \tag{3.8}$$

$$Z_{1t} = \sigma V_1 - \gamma Z_1, \qquad x = 0 \mod (1), \tag{3.9}$$

$$[V_1]_x = 0, \qquad x = 0 \mod (1). \tag{3.10}$$

3.2.1. Case for $x = 0$ mod(1). Now $F(U) - F(U_{\alpha_2}) = DF(U_{\alpha_2} + \theta V_1)V_1$, where $0 < \theta < 1$. From before we have that

$$DF(U_{\alpha_2}) = \begin{bmatrix} f_1'(u_{1\alpha_2^{(1)}}) & 0 & \cdots & 0 \\ 0 & f_2'(u_{2\alpha_2^{(2)}}) & \cdots & 0 \\ 0 & 0 & \cdots & f_n'(u_{n\alpha_2^{(n)}}) \end{bmatrix} \tag{3.11}$$

is negative definite. In fact, we can find $\epsilon_1 > 0$ such that $f_i'(u_{i\alpha_2^i}(x)) < -\epsilon_1$ for all $x \in \mathbb{R}$. By the smoothness of the function $F(Y)$, there is a constant $\delta_0 > 0$ such that the matrix

$$DF(U_{\alpha_2}(x) + Y) \tag{3.12}$$

is negative definite for any $Y = (y_1, y_2, \cdots, y_n)^t \in S'$, where $S' = \prod_{i=1}^n[-\delta_0, \delta_0]$. In this case we can find $\eta_0 > 0$, such that

$$\eta_0 = \min_{\substack{1 \leq i \leq n \\ x \in \mathbb{R}, y \in S'}} \{\lambda_i(u_{i\alpha_2^i}(x) + y_i)\},$$

where $-\lambda_i$ $(i = 1, 2, \cdots, n)$ are eigenvalues of the matrix (3.12). Let

$$\hat{D}_0 = \begin{bmatrix} -\eta_0 & & \bigcirc \\ & -\eta_0 & \\ & & \ddots & \\ \bigcirc & & & -\eta_0 \end{bmatrix}, \tag{3.13}$$

$V = C^tV_1$, and $Z = Z_1$. The vector field near $(U_{\alpha_2}, \sigma\gamma^{-1}U_{\alpha_2})$ is then given by

$$G_1(V,Z) = \begin{bmatrix} C^t(F(U_{\alpha_2} + CV) - F(U_{\alpha_2})) - C^tZ \\ \sigma CV - \gamma Z \end{bmatrix}. \tag{3.14}$$

If we linearize this near $(U_{\alpha_2}, \sigma\gamma^{-1}U_{\alpha_2})$ we have that

$$\hat{G}_{1L}(V,Z) = \begin{bmatrix} -\hat{D}_0 & -C^t \\ \sigma C & -\gamma \end{bmatrix} \begin{bmatrix} V \\ Z \end{bmatrix}.$$

We use the same argument as in the contracting rectangle around $(U,W) = (0,0)$ and arrive at the following lemma.

LEMMA 3.4. *Let $-\eta_0 < -\hat{\gamma}\delta_1$, where $\hat{\gamma} = \rho(\hat{C}^t\hat{C})$, $\delta_1 = \max_i\{\sigma_i/\gamma_i\}$, and C is the orthogonal matrix with $MC = \Lambda C$ and $\hat{C} = (|C_{ij}|)_{n\times n}$. Then there exists a rectangle S^{**} and a constant κ_3 such that for $\tau \in (0,1)$*

$$G_1(V,Z)\cdot n(V,Z) < -\kappa_3\tau$$

for $(V,Z) \in \partial(\tau S^*)$.

3.2.2. Case for $x \in \mathbb{R}/\mathbb{Z}$. This is the same as in case 3.1.2.

4. Asymptotic stability.

Let us consider a more general system of the form

$$\begin{aligned}
&(4.1) && U_t = U_{xx} - GU, && x \in (0,1) \ \text{mod}\ (1),\\
&(4.2) && W \equiv 0, && x \in (0,1) \ \text{mod}\ (1),\\
&(4.3) && U_t = \Lambda[U_x]_x + F_1(U,W), && x = 0 \ \text{mod}\ (1),\\
&(4.4) && W_t = F_2(U,W), && x = 0 \ \text{mod}\ (1),\\
&(4.5) && [U]_n = 0, && x = 0 \ \text{mod}\ (1),
\end{aligned}$$

where $U \in C(\mathbb{R}^+ \times \mathbb{R}, \mathbb{R}^n)$, $W \in C(\mathbb{R}^+ \times \mathbb{Z}, \mathbb{R}^n)$. $F_i \in C^1(\mathbb{R}^{2n}, \mathbb{R}^n)$ $(i = 1,2)$; matrices Λ and G are the same as before.

The contracting rectangles allow us to define nonlinear functionals, which are decreasing functions of time [4], [10], [17] for some solutions of the differential equations (4.1)–(4.5). Here the functionals to be considered are those associated with the rectangles $S = S_1 \times S_2$ such that the origin is in the interior of S.

Let $|\cdot|_{S_1}$ be the norm on $\mathbb{R}^n$ defined by S_1 in the usual way:

$$|V|_{s_1} = \inf\{t \geq 0 : V \in tS_1\}. \tag{4.6}$$

Thus $|V|_{s_1}$ is the smallest multiple of S_1 containing V. Similarly, the norm $|Z|_{s_2}$ is defined in the obvious way:

$$|Z|_{s_2} = \inf\{t \geq 0 : Z \in tS_2\}. \tag{4.7}$$

If we now define the continuous functions $V_{s_i} : BC \to \mathbb{R}$ $(i = 1,2)$ by

$$V_{s_1}(V) = \sup_{x\in\mathbb{R}} |V(x)|_{s_1} \quad \text{and} \quad V_{s_2}(Z) = \sup_{n\in\mathbb{Z}} |Z(n)|_{s_2}, \tag{4.8}$$

then we can now state the following lemma.

LEMMA 4.1. *Let S be defined as in Definition* 3.1 (*or $S = S^* \cap S^{**}$ if necessary*). *Suppose that (U,W) is a solution of* (4.1)–(4.5) *for $|t - T| < \delta$, and that (U_0, W_0) is a steady state solution of* (4.1)–(4.5). *Let $V = U - U_0$, $Z = W - W_0$ with $V_{s_1}(V(T)) = \tau_1 < 1$, $V_{s_2}(Z(T)) = \tau_2 < 1$, $V \in C_0(\mathbb{R})$, and $Z \in c_0$. Suppose that there exist $\eta_1, \eta_2 > 0$ such that for any $V \in \partial(\tau_1 S_1)$ or $Z \in \partial(\tau_2 S_2)$ with $Z \neq 0$ and $\overline{n}(V,Z)$ normal to*

$\partial(\tau_1 S_1 \times \tau_2 S_2)$ at (V, Z), we have that $(F_1(U, W) - F_1(U_0, W_0)) \cdot \overline{n}(V, Z) < -\tau_1\eta_1$ or $(F_2(U, W) - F_2(U_0, W_0)) \cdot \overline{n}(V, Z) < -\tau_2\eta_2$; then

$$\overline{D}V_{s_1}(V(T)) \leq -(\eta/L)V_{s_1}(V(T)) \tag{4.9}$$

or

$$\overline{D}V_{s_2}(Z(T)) \leq -(\eta/L)V_{s_2}(Z(T)), \tag{4.10}$$

where L is the length of the shortest side of S, $\eta = \min\{\eta_1, \eta_2, \min_{1\leq j\leq n}\{g_j\overline{u}_j, g_j\overline{z}_j\}\}$.

Proof. Let S be defined as in Definition 3.1. If $V_{s_1}(V(T)) = \tau_1$, then by (4.6)

$$-\tau_1\overline{u}_j \leq v_j(T, x) \leq \tau_1\overline{u}_j \quad \text{for all } x \in \mathbb{R} \quad (j = 1, \cdots, n).$$

We say that $V(T, x)$ is in the jth upper face if $v_j(T, x) = \tau_1\overline{u}_j$, with an analogous definition for the lower face. Now, if $V(T, x) \in \partial\tau_1 S_1$, then there is a subset $J \subset \{1, \cdots, n\}$ such that $V(T, x)$ is on one of the jth faces if and only if $j \in J$. If $V(T, \overline{x})$ is in the jth upper face, then $v_j(T, x) \leq \tau_1\overline{u}_j$ for all x near $x = \overline{x}$. In this case, if $\overline{x} \in \mathbb{R}/\mathbb{Z}$, then $\partial_{xx}v_j(T, \overline{x}) \leq 0$ and

$$\partial_t v_j(T, \overline{x}) = \partial_{xx}v_j(T, \overline{x}) - g_j v_j(T, \overline{x}) \leq -g_j\tau_1\overline{u}_j;$$

if $\overline{x} = k \in \mathbb{Z}$, then $v_j(T, \overline{x})$ is the local maximum and $[v_x]_k \leq 0$. Therefore

$$\partial_t v_j(T, k) < -\eta_1\tau_1,$$

and in this case

$$v_j(T + h, \overline{x}) < \tau_1(\overline{u}_j - \eta_1 h) \quad \text{for small } h.$$

By the continuity of V, this holds for all x in a neighborhood of $\overline{x}$. A similar result holds for the lower faces.

Let $X = \{x : V(T, x) \in \partial\tau_1 S_1\}$. Since $V \in C_0(\mathbb{R})$ then X is a compact set in $\mathbb{R}$, and by the above computation there is an open set $\Omega \subset \mathbb{R}$ which contains X such that, for $\theta \in \Omega$, we have that for small h

$$V(T + h, \theta) \in (1 - h\eta_1 L^{-1})\tau_1 S_1,$$

where $L = \max_{1\leq j\leq n}\{\overline{u}_j, \overline{z}_j\}$. Now $V \in C\{(T - \delta, T + \delta)|C_0(\mathbb{R})\}$ and for $x \in \mathbb{R}\backslash\Omega$, $V(T, x) \subset \operatorname{int}(\tau_1 S_1)$. Thus there is an $h_0 > 0$ and a compact set $K_1 \subset \operatorname{int}(\tau_1 S_1)$ for which $V(T + h, x) \subset K_1$ for all $|h| < h_0$. Hence for sufficiently small h,

$$V_{s_1}(V(T + h)) \leq \tau_1(1 - h\eta_1/L),$$

so that

$$(V_{s_1}(V(T + h)) - V_{s_1}(V(T)))/h \leq -\eta/L \; V_{s_1}(V(T)),$$

and (4.9) has been proved.

Let us turn our attention to $Z(t, k)$. Now $V_{s_2}(Z(T, k)) = \tau_2$ so that if $Z(T, k) \in \partial(\tau_2 S_2)$, say in the $(n + j)$th upper face, then $z_j(T, k) = \tau_2\overline{z}_j$ for some k and

$$\partial_t z_j(T, k) = F_2(U(T, k), W(T, k)) - F_2(U_0(k), W_0(k)) < -\eta_2\tau_2.$$

This makes $z_j(T + h, k) < \tau_2(\overline{z}_j - \eta_2 h)$ for some h. Let $\overline{N} = \{k : Z(T, k) \in \partial\tau_2 S_2\}$. Since $Z \in c_0$, $\overline{N}$ is a bounded set in $\mathbb{Z}$ (provided $\overline{u}_j$ or $-\overline{z}_j$ are not the nonzero roots of F_1 and F_2; in fact, it is valid when $\overline{u}_j$ and $\overline{z}_j$ are small). We can then find a neighborhood $\pi \supset \overline{N}$ such that for $\theta \in \pi$,

$$Z(T + h, \theta) \subset (1 - h\eta_2/L)\tau_2 S_2 \quad \text{for small } h.$$

Let $k \in \mathbb{Z}/\pi$. Since $Z(T, k) \subset \operatorname{int}(\tau_2 S_2)$ and $Z \in C((T - \delta, T + \delta), c_0)$ there is an $h_0 > 0$ and a compact set $K_2 \subset \operatorname{int}(\tau_2 S_2)$ for which $Z(T + h, k) \subset K_2$ for all $|h| < h_0$. Thus for sufficiently small h, $Z(T + h, k) \subset (1 - h\eta_2/L)\tau_2 S_2$ for all $k \in \mathbb{Z}$. Hence

$$(V_{s_2}(Z(T + h)) - V_{s_2}(Z(T)))/h \leq -\eta/L \; V_{s_2}(Z(T)). \qquad \square$$

By Lemma 4.1, we can easily obtain the following result.

THEOREM 4.2. *The steady state solutions* $(U, W) = (0, 0)$ *and* $(U, W) = (U_{\alpha_2}, \sigma\gamma^{-1}U_{\alpha_2})$ *are attractors of* (1.4)–(1.8), *provided* (3.2) *is valid.*

Proof. By Lemma 4.1,

$$\begin{aligned} &V_{s_1}(U - Q)(t) \le e^{-\frac{\eta_3}{L}t}V_{s_1}(U - Q)(0), \\ &V_{s_2}(W - \sigma\gamma^{-1}Q)(t) \le e^{-\frac{\eta_3}{L}t}V_{s_2}(W - \sigma\gamma^{-1}Q)(0), \end{aligned}$$

where $\eta_3 = \min\{\eta, \kappa_1, \kappa_2, \kappa_3\}$, $(Q, \sigma\gamma^{-1}Q)$ is one of the steady states $(0, 0)$ or $(U_{\alpha_2}, \sigma\gamma^{-1}U_{\alpha_2})$ (note that $S = S^* \cap S^{**}$ is fixed and C is an orthogonal matrix). □

Remark 4.3. The above theorem tells us that if the initial values $(U(x, 0), W(k, 0))$ are close enough to one of the steady states with their difference belonging to $C_0(\mathbb{R})$, then the solution $(U(x, t), W(k, t))$ must approach the steady state uniformly in the norm of $L^\infty(\mathbb{R})$ as $t \to \infty$.

5. Existence. We will now give an existence proof for the problem defined by (1.4)–(1.8). We note that the standard literature deals with the existence of solutions (at least) in L^2. However, our steady state solutions $U_{\alpha_i(x)}$ are not in L^2. To overcome this problem we introduce a new function $(V, Z) = (U - Q, W - \sigma\gamma^{-1}Q)$ near $(Q, \sigma\gamma^{-1}Q)$ which is the nontrivial steady state solution of (1.4)–(1.8). We then prove the existence of a solution (V, Z) of (5.1)–(5.4) in some subspace of L^2. This result establishes the existence of the solution $(U, W) = (Q + V, \sigma\gamma^{-1}Q + Z)$ to (1.4)–(1.8), although neither U nor W belongs to L^2.

Let us start with the system (1.10)–(1.13). Suppose that $(Q, \sigma\gamma^{-1}Q)$ is a steady state solution of (1.10)–(1.13) and that $(\hat{U}, \hat{Z})$ satisfies (1.10)–(1.13). Using the above idea, we define new functions

$$\begin{array}{ll} u = \hat{U} - Q, & x \in (0,1) \mod (1), \\ v = \hat{U} - Q, & x = 0 \mod (1), \\ w = \hat{Z} - \sigma\gamma^{-1}Q, & x = 0 \mod (1). \end{array}$$

Then (u, v, w) satisfy the following equations:

$$\begin{array}{lll} (5.1) & u_t = u_{xx} - Gu, & x \in (0,1) \mod (1), \\ (5.2) & v_t = \Lambda[u_x]_x + C^t[F(C(Q + v)) & \\ & \qquad - F(CQ)] - C^t w, & x = 0 \mod (1), \\ (5.3) & w = \sigma Cv - \gamma w, & x = 0 \mod (1), \\ (5.4) & [u]_x = 0, & x = 0 \mod (1). \end{array}$$

Let

$$H = \{(p, q, r) : p \in L^2(\mathbb{R}, \mathbb{C}^n),\ q \in \ell^2(\mathbb{Z}, \mathbb{C}^n),\ r \in \ell^2(\mathbb{Z}, \mathbb{C}^n)\}$$

with the inner product given by

$$\langle (p_1, q_1, r_1), (p_2, q_2, r_2)\rangle_H = \int_{\mathbb{R}} \langle \Lambda p_1, p_2\rangle + \sum_{i \in \mathbb{Z}} (\langle q_1, q_2\rangle^i + \langle r_1, r_2\rangle^i),$$

where $\langle \Lambda p_1, p_2\rangle(x) = (\lambda_1 p_1^1 \bar{p}_2^1 + \lambda_2 p_1^2 \bar{p}_2^2 + \cdots + \lambda_n p_1^n \bar{p}_2^n)(x)$,

$\langle q_1, q_2\rangle^i = (q_1^1 \bar{q}_2^1 + q_1^2 \bar{q}_2^2 + \cdots + q_1^n \bar{q}_2^n)(i)$, for $i \in \mathbb{Z}$, $\Lambda = \mathrm{Diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$.

H is a complete Hilbert space. Let X be a subspace of H given by

$$X = \Big\{(p, q, r) \in H : p, p', p'' \in L^2(\mathbb{R}, \mathbb{C}^n), \{[p']_i\} \in \ell^2, \\ \lim_{x \to i} p(x) = q(i) \text{ for each } i \in \mathbb{Z}\Big\}.$$

Then X is dense in H. Let

$$Y = \{(p,q,r) \in H : p, p' \in L^2(\mathbb{R}, \mathbb{C}^n), \ \lim_{x \to i} p(x) = q(i)\}$$

with the inner product given by

$$\begin{aligned}\langle (p_1,q_1,r_1),(p_2,q_2,r_2)\rangle_Y = \int_{\mathbb{R}} &(\langle \Lambda p_1, p_2\rangle + \langle \Lambda p_1', p_2'\rangle) \\ &+ \sum_{i\in\mathbb{Z}} (\langle q_1, q_2\rangle^i + \langle r_1, r_2\rangle^i).\end{aligned}$$

Y is a complete space and, in fact, $Y = H^1 \times S \times \ell^2$, where

$$S = \{q \in \ell^2 : \text{ there exists a } p \in H^1 \text{ with } \lim_{x\to i} p(x) = q(i)\}.$$

Then the three spaces satisfy

$$X \subset Y \subset H \tag{5.5}$$

and

$$\|(u,v,w)\|_H \le \|(u,v,w)\|_Y \quad \text{for any } (u,v,w) \in Y. \tag{5.6}$$

Define the linear operator $A : X \to H$ by

$$A(u,v,w) = (-u_{xx} + u, \Lambda[-u_x]_x + v, w), \tag{5.7}$$

where $u(i,t) = v^i(t)$. It is easy to verify that $\langle A(u_1,v_1,w_1),(u_2,v_2,w_2)\rangle_H = \langle (u_1,v_1,w_1), A(u_2,v_2,w_2)\rangle_H$ for any $(u_i,v_i,w_i) \in X$ $(i = 1,2)$. That is, A is a self-adjoint operator. In order to prove existence, we will need the following lemmas.

LEMMA 5.1. *The operator A defined by* (5.7) *with* dom $(A) = X$ *is a closed operator from X to H with $A^{-1} : H \to H$ bounded.*

Proof. For any $(u,v,w) \in X$ we have, by the definition of A,

$$A(u,v,w) = (p,q,r) \in H.$$

To show that A has a bounded inverse, we need to find A^{-1}. Fix $(p,q,r) \in H$, then for any $(x,y,z) \in Y$ we have that

$$\begin{aligned}\langle (x,y,z),(p,q,r)\rangle_H &\le \|(x,y,z)\|_H \|(p,q,r)\|_H \\ &\le \|(x,y,z)\|_Y \|(p,q,r)\|_H.\end{aligned}$$

$\langle (x,y,z),(p,q,r)\rangle_H$ defines a bounded linear functional on Y. By the Riesz Representation Theorem, there exists a unique $(u,v,w) \in Y$ such that

$$\langle (x,y,z),(p,q,r)\rangle_H = \langle (x,y,z),(u,v,w)\rangle_Y \quad \text{for any } (x,y,z) \in Y. \tag{5.8}$$

Thus for any given $(p,q,r) \in H$, we can define $B : H \to Y$, where

$$B(p,q,r) = (u,v,w). \tag{5.9}$$

We need to show that $B = A^{-1}$ and B is bounded. This is equivalent to showing that

1°. $\|B\| < \infty$,

2°. $B(H) \subset X$,

3°. Equation (5.9) is valid if and only if $A(u,v,w) = (p,q,r) \in H$, that is, $B = A^{-1}$.

By (5.8) and (5.9),

$$\begin{aligned}\|B(p,q,r)\|_Y^2 &= \langle (u,v,w),(u,v,w)\rangle_Y = \langle (u,v,w),(p,q,r)\rangle_H \\ &\le \|(u,v,w)\|_H \|(p,q,r)\|_H \le \|B(p,q,r)\|_Y \|(p,q,r)\|_H.\end{aligned}$$

Therefore $\|B(p,q,r)\|_Y \leq \|(p,q,r)\|_H$. Hence B is a bounded operator with $\|B\| \leq 1$.

To show 2°, the conclusion of Lemma 15.5 of [7, §I.15] will be used. The bilinear form for the operator $Lw = -w''$ is given by $\beta[x,y] = \int_{[j-1,j]}\langle x', y'\rangle$. Thus Lemma 15.5 in [7] can be rewritten equivalently in the following way.

Assume that $u \in H^1(a,b)$ and that for all $\phi \in C_0^\infty(a,b)$,

$$|\beta[u,\phi]| \leq k_1\|\phi\|_{L^2}. \tag{5.10}$$

Then for any subdomain (a_1,b_1) of (a,b) with $[a_1,b_1] \subset (a,b)$, u' belongs to $H^1((a_1,b_1))$ and

$$(\|u'\|_{L^2} + \|u''\|_{L^2})|_{(a_1,b_1)} \leq K(c + \|u\|_{L^2} + \|u'\|_{L^2})|_{(a,b)},$$

where K is a constant depending only on c, (a_1,b_1), and (a,b).

To prove condition (5.10), we choose $\phi \in C_0^\infty(\mathbb{R},\mathbb{C}^n)$ with $\operatorname{supp}\phi \subset [a,b] \subset (j-1,j)$. Let $\psi = (\psi^k)$, $\eta = (\eta^k)$ be such that $\psi^k = \eta^k = 0$ for all $k \in \mathbb{Z}$; then

$$(\phi,\psi,\eta) \in Y$$

and (5.8) yields

$$\int_{(j-1,j)}\langle \phi', \Lambda u'\rangle = \int_{(j-1,j)}\langle \phi, \Lambda(p-u)\rangle. \tag{5.11}$$

By Holder's inequality,

$$\begin{aligned}\left|\int_{(j-1,j)}\langle\phi',u'\rangle\right| &\leq \frac{1}{\min\limits_{1\leq i\leq n}\lambda_i}(\|(u,v,w)\|_H + \|(p,q,r)\|_H)\|\phi\|_{L^2(j-1,j)}\\ &= \frac{1}{\min\limits_{1\leq i\leq n}\lambda_i}(\|B(p,q,r)\|_H + \|(p,q,r)\|_H)\|\phi\|_{L^2(j-1,j)}\\ &\leq \frac{1}{\min\limits_{1\leq i\leq n}\lambda_i}2\|(p,q,r)\|_H\|\phi\|^2_{L^2(j-1,j)}.\end{aligned}$$

This establishes (5.10). Therefore,

$$u''|_{(a,b)} \in L^2(a,b) \quad \text{for any } [a,b] \subset (j-1,j).$$

Taking the complex conjugate of (5.11),

$$-\int_{(a,b)}\langle u'', \Lambda\phi\rangle = \int_{(a,b)}\overline{\langle \Lambda\phi', u'\rangle} = \int_{(a,b)}\langle p-u, \Lambda\phi\rangle.$$

The above identity is true for any $\phi \in C_0^\infty((a,b),\mathbb{C}^n)$. By choosing suitable ϕ and by (5.8) we also can conclude that $u'', u' \in L^2(\mathbb{R})$. Thus 2° holds.

To show that 3° holds we will show that it holds for each component.

Since C_0^∞ is dense in L^2, for any $h \in L^2[(a,b),\mathbb{C}^n]$,

$$\int_{(a,b)}\langle -u'', h\rangle = \int_{(a,b)}\langle p-u, h\rangle.$$

Therefore $-u'' + u = p$ on (a,b); hence $-u'' + u = p$ on $(j-1,j)$. This is the first component in (5.7).

Now $L^2[(j-1,j)] \subset L^1[(j-1,j)]$, so that u' is absolutely continuous on $(j-1,j)$ and

$$u'(b) - u'(b) = \int_{(a,b)} u'' \quad \text{for any } (a,b) \subset (j-1,j).$$

Let $a \to j-1+$ or $b \to j-$; then $u'(j-1+)$ and $u'(j-)$ are well defined. Next, let us verify the last two components in (5.7).

Let $y^k_{j,i} = \delta_{jk} e_i$, where $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)^t$ is the ith base vector in $\mathbb{R}^n$. We define the function

$$x_{j,i} = \begin{cases} \frac{x-j+h}{h} e_i, & x \in (j-h, j), \\ -\frac{x-j-h}{h} e_i, & x \in (j, j+h), \\ 0 & \text{otherwise}, \end{cases}$$

where $i = 1, 2, \cdots, n$. Then $(x_{j,k}, y_{j,i}, 0) \in Y$. Substituting this into (5.8) we see that

(5.12)
$$\begin{aligned} \frac{1}{h}\int_{(j-h,j)} \lambda_i u_i' + \frac{1}{h}\int_{(j-h,j)} \lambda_i (x-j+h) u_i - \frac{1}{h}\int_{(j,j+h)} \lambda_i u_i' \\ - \frac{1}{h}\int_{(j,j+h)} \lambda_i (x-j-h) u_i + v^j_i \\ = \frac{1}{h}\int_{(j-h,j)} \lambda_i (x-j+h) p_i - \frac{1}{h}\int_{(j,j+h)} \lambda_i (x-j-h) p_i + q^j_i, \end{aligned}$$

where $u = (u_1, \cdots, u_n)^t$, $p = (p_1, \cdots, p_n)^t$, etc.

Now

$$\frac{1}{h}\int_{(j-h,j)} \lambda_i u_i' = \frac{\lambda_i}{h}(u_i(j) - u_i(j-h)) \to \lambda_i u_i'(j-) \quad \text{as } h \to 0,$$
$$\frac{1}{h}\int_{(j,j+h)} \lambda_i u_i' \to \lambda_i u_i'(j+) \quad \text{as } h \to 0,$$

and $\left|\frac{1}{h}\int_{(j-h,j)} \lambda_i (x-j\pm h) k\right| \le \int_{(j-h,j)} \lambda_i |k| \to 0$ as $h \to 0$ for any $k \in L^2(\mathbb{R})$. Letting $h \to 0$ in (5.12) we see that

$$\lambda_i u_i'(j-) - \lambda_i u_i'(j+) + v^j_i = q^j_i, \qquad i = 1, 2, \cdots, n,$$

which is the second component of (5.7). Now choose $(0, 0, z_{j,i}) \in Y$ with $z^k_{j,i} = \delta_{jk} e_i$, where $z_{i,i}(k) = z^k_{j,i}$. Substituting this into (5.8), we have that

$$\langle (0, 0, z_{j,i}), (u, v, w) \rangle_H = \langle (0, 0, z_{j,i}), (p, q, r) \rangle_Y,$$

which implies that $w^i_j = r^i_j$, $i = 1, 2, \cdots, n$. Thus we have obtained the third component of (5.7). Therefore we have that if $B(p, q, r) = (u, v, w)$ then $(u, v, w) \in X$ and $A(u, v, w) = (p, q, r)$. We now need to show the converse. That is, if $A(u, v, w) = (p, q, r)$, then $B(p, q, r) = (u, v, w)$. Thus (5.8) is valid.

By the definition of A,

$$(p, q, r) = (-u_{xx} + u, \Lambda[-u_x] + v, w).$$

Choose any $(\hat{x},\hat{y},\hat{z}) \in Y$. Taking the inner product of the above expression with $(\hat{x},\hat{y},\hat{z})$ we arrive at

$$\begin{aligned}
\langle(\hat{x},\hat{y},\hat{z}),(p,q,r)\rangle_H &= \int \langle\Lambda\hat{x},p\rangle + \sum_{i\in\mathbb{Z}}(\langle\hat{y}^i,q^i\rangle + \langle\hat{z}^i,r^i\rangle) \\
&= \int (-\langle\hat{x},\Lambda u_{xx}\rangle + \langle\hat{x},\Lambda u\rangle) + \sum_{i\in\mathbb{Z}}\{-\langle\hat{y}^i,\Lambda[u_x^i]\rangle + \langle\hat{y}^i,v^i\rangle + \langle\hat{z}^i,w^i\rangle\} \\
&= \int (\langle\hat{x}_x,\Lambda u_x\rangle + \langle\hat{x},\Lambda u\rangle) \\
&\quad + \sum_{i\in\mathbb{Z}}\{\langle\hat{y}^i,\Lambda[u_x^i]\rangle - \langle\hat{y}^i,\Lambda[u_x^i]\rangle + \langle\hat{y}^i,v^i\rangle + \langle\hat{z}^i,w^i\rangle\} \\
&= \langle(\hat{x},\hat{y},\hat{z}),(u,v,w)\rangle_Y.
\end{aligned}$$

Here we used the fact that $\lim_{s\to i}\hat{x}(s) = \hat{y}^i$ and $u_x\big|_{i-}^{i+} = -u_x(i-) + u_x(i+)$. Thus condition 3° holds and the proof is complete. □

To apply semigroup theory to the nonlinear problem we need another lemma, which is a special case of Theorem V3.2 of [14]. Define

$$\begin{aligned}
\theta(A) &= \{\langle A(u,v,w),(u,v,w)\rangle_H | \\
&\qquad (u,v,w) \in \operatorname{dom} A \text{ with } \|(u,v,w)\|_H = 1\}, \\
\Gamma &= \operatorname{cl}(\theta(A)), \\
\Delta &= \mathbb{C}\backslash\Gamma.
\end{aligned}$$

LEMMA 5.2. *Suppose that A is closed and Δ is connected. For $\lambda \in \Delta$, $A-\lambda$ has nullity zero and constant deficiency. If the deficiency is zero, then Δ is contained in the resolvent set of A and*

$$\|(\lambda - A)^{-1}\| \le 1/\operatorname{dis}(\lambda,\Gamma) \qquad \textit{for } \operatorname{Re}\lambda \le 0. \tag{5.13}$$

Let us check the linear operator A to see if it satisfies the conditions in the lemma so that (5.13) is valid.

By Lemma 5.1 we know that A is closed. Choose any $(u,v,w) \in X$ with $\|(u,v,w)\|_H = 1$. Then

$$\begin{aligned}
\langle A(u,v,w),(u,v,w)\rangle_H &= \int_{\mathbb{R}} \langle -\Lambda u'' + \Lambda u, u\rangle \\
&\quad + \sum_{i\in\mathbb{Z}}\{\langle -\Lambda[u']_i + v^i, v^i\rangle + \langle w^i, w^i\rangle\} \\
&= \int_{\mathbb{R}} (\langle u',\Lambda u\rangle + \langle u,\Lambda u\rangle) + \{\|v\|_{\ell^2}^2 + \|w\|_{\ell^2}^2\} \\
&\quad + \sum_{j\in\mathbb{Z}}\{\langle\Lambda u'(j+1-),u(j)\rangle \\
&\quad - \langle\Lambda u'(j+),u(j)\rangle\} + \langle\Lambda[u']_j,u^j\rangle \\
&= \|(u,v,w)\|_Y \\
&\ge \|(u,v,w)\|_H = 1.
\end{aligned}$$

Hence $\theta(A) \subset [1,\infty)$, $\Gamma \subset [1,\infty)$, and $\Delta = \mathbb{C}/\Gamma$ is connected in $\mathbb{C}$.

By the definition of $\theta(A)$, we know that $A-\lambda$ has nullity zero for any $\lambda \in \Delta$. If $\operatorname{Re}\lambda \le 0$, it is very easy to check that

$$L_\pm = \ker(A^* - \overline{\lambda} \mp i) = \emptyset.$$

Hence (5.13) is valid.

To prove that the solution of (5.1)–(5.4) exists, rewrite (5.1)–(5.4) as follows:

$$\begin{aligned}&U_t - U_{xx} + U = (I-G)U, && x \in (0,1) \mod (1),\\ &V_t - \Lambda[U_x]_x + V = V + C^t(F(C(Q+V)) && \\ &\qquad\qquad\qquad\qquad - F(CQ)) - C^t W, && x = 0 \mod (1),\\ &W_t + W = \sigma CV + (I-\gamma)W, && x=0 \mod (1),\\ &[U]_x = 0, && x = 0 \mod (1),\end{aligned} \tag{5.14}$$

with initial conditions given by

$$(U,V,W)\big|_{t=0} = (\phi,\psi,\eta).$$

Let

$$\hat{F}(t,(U,V,W)) = \big((I-G)U, V + C^t\big(F(C(Q+V)) - F(CQ)\big) - C^t W,\\ \sigma CV + (I-\gamma)W\big)^t.$$

Then (5.14) can be simply written as

$$\frac{d}{dt}(U,V,W) + A(U,V,W) = \hat{F}(t,(U,V,W)), \tag{5.15}$$

with

$$(U,V,W)\big|_{t=0} = (\phi,\psi,\eta). \tag{5.16}$$

Next, we use the following lemma (due to Sobolevski's theorem; see [7, §11.16]) to prove that (5.15) and (5.16) have a solution.

LEMMA 5.3. *Let A be a closed linear operator on a Banach space E such that* (5.13) *holds. Suppose that $\hat{F}(t,p)$ is a function on $[0,T_0]\times E$ so that for some constants $\alpha, \eta \in (0,1)$ and for any $R > 0$ there exists a constant $C(R)$ for which*

$$\begin{aligned}&\|\hat{F}(t_1, A^{-\alpha}p_1) - \hat{F}(t_2, A^{-\alpha}p_2)\|_E\\ &\qquad \le C(R)(|t_1 - t_2|^\eta + \|p_1 - p_2\|_E),\end{aligned} \tag{5.17}$$

where $t_1, t_2 \in [0, t_0]$, $p_1, p_2 \in E$ with $\|p_1\|_E, \|p_2\|_E < R$. Then for any $p_0 \in dom\ A$ and each $R > \|A_0^\alpha p_0\|_E$, there exists a $t^* = t^*(R, \|A_0^\alpha p_0\|_E) > 0$ such that the initial value problem

$$\frac{dp}{dt} + Ap = \hat{F}(t,p), \qquad p(0) = p_0 \tag{5.18}$$

has a unique solution on $[0,t^*]$. Furthermore, if there exists a constant $R' > 0$ such that for any solution p of (5.18) *in $[0,T_1]$, $T_1 \le T_0$, we have that $\|Ap\|_E < R'$. Then we may choose $R > R'$ and thus apply the local existence assertion to $[0,t^*]$, $[t^*, 2t^*]$, and so on, until $[0,T_0]$ is exhausted.*

Since the operator A in (5.18) was shown to be closed in X and satisfies (5.13), we need only to establish (5.17) for the function $\hat{F}$ in (5.15) to conclude the local existence of a solution to (5.1)–(5.4).

Let $(X_k, Y_k, Z_k) = A^{-\alpha}(U_k, V_k, W_k)$ for $k = 1, 2$. Then

$$\begin{aligned}
&\|\hat{F}(t_1, A^{-\alpha}(U_1, V_1, W_1)) - \hat{F}(t_2, A^{-\alpha}(U_2, V_2, W_2))\|_H \\
&\quad = \|\hat{F}(t_1, (X_1, Y_1, Z_1)) - \hat{F}(t_2, (X_2, Y_2, Z_2))\|_H \\
&\quad = \|((I - G)(X_1 - X_2), \\
&\qquad\qquad Y_1^i - Y_2^i + C^t[F(C(Q^i + Y_1^i)) - F(C(Q^i + Y_2^i))] \\
&\qquad\qquad - C^t(Z_1^i - Z_2^i), \sigma C(Y_1^i - Y_2^i) + (I - \gamma)(Z_1^i - Z_2^i))\|_H \\
&\quad \leq \|(I - G)(X_1 - X_2), 0, 0)\|_H + \|(0, Y_1^i - Y_2^i + C^t[F(C(Q^i + Y_1^i)) \\
&\qquad - F(C(Q^i + Y_2^i))] - C^t(Z_1^i - Z_2^i), 0)\|_H \\
&\qquad + \|(0, 0, \sigma C(Y_1^i - Y_2^i) + (I - \gamma)(Z_1^i - Z_2^i))\|_H.
\end{aligned}$$

Now C is an orthogonal matrix so that

$$\begin{aligned}
&\|((I - G)(X_1 - X_2), 0, 0)\|_H \\
&\quad \leq \max_{1\leq i\leq n} |1 - g_i| \, \|(X_1 - X_2, Y_1 - Y_2, Z_1 - Z_2)\|_H, \\
&\|(0, 0, \sigma C(Y_1^i - Y_2^i) + (I - \gamma)(Z_1^i - Z_2^i))\|_H \\
&\quad \leq \big(\max_i\{\sigma_i\} + \max_i |1 - \gamma_i|\big)\|(X_1 - X_2, Y_1 - Y_2, Z_1 - Z_2)\|_H, \\
&\big\|\big(0, Y_1^i - Y_2^i + C^t[F(C(Q^i + Y_1^i)) - F(C(Q^i + Y_2^i))] - C^t(Z_1^i - Z_2^i), 0\big)\|_H \\
&\qquad\qquad \leq \|(0, Y_1^i - Y_2^i, 0)\|_H + \|(0, C^t(Z_1^i - Z_2^i), 0)\|_H \\
&\qquad\qquad\qquad + \|(0, C^t[F(CQ^i + CY_1^i) - F(CQ^i + CY_2^i)], 0)\|_H \\
&\qquad\qquad \leq (2 + \max_{\substack{1\leq i\leq n \\ y\in\mathbb{R}}} |f_i'(y)|)\|(X_1 - X_2, Y_1 - Y_2, Z_1 - Z_2)\|_H.
\end{aligned}$$

Therefore

$$\begin{aligned}
\|\hat{F}(t_1, A^{-\alpha}(U_1, V_1, W_1)) - \hat{F}(t_2, A^{-\alpha}(U_2, V_2, W_2))\|_H& \\
\leq C(R)\|(X_1 - X_2, Y_1 - Y_2, Z_1 - Z_2)\|_H&,
\end{aligned} \tag{5.19}$$

where

$$C(R) = \max_{1\leq i\leq n} |1 - g_i| + \max_{1\leq i\leq n} \sigma_i + \max_{1\leq i\leq n} |1 - \gamma_i| + 2 + \max_{\substack{1\leq i\leq n \\ y\in\mathbb{R}}} \{|f_i'(y)|\}.$$

Hence $\hat{F}$ satisfies condition (5.17), and the system (5.15)–(5.16) has unique solution $P = (U, V, W)$ with the initial value $P_0 = (U_0, V_0, W_0)$ at $t = 0$ satisfying $\|A^\alpha P_0\|_H \leq R$.

To prove global existence, we must show that $\|A(U, V, W)\|_H$ is bounded for any solution $P = (U, V, W)$. By (5.19),

$$\|\hat{F}(U, V, W)\|_H \leq C(R)\|(U, V, W)\|_H. \tag{5.20}$$

If $\|(U, V, W)\|_H$ and $\|(U_t, V_t, W_t)\|_H$ are bounded, then by (5.15)–(5.16), we can see that $\|A(U, V, W)\|_H$ is bounded.

Let

$$E(t) = (\|(U, V, W)\|_H^2 + \|(U, V, W)_t\|_H^2)/2.$$

Then $E(t)$ is a Lyapunov function and its derivative is given by

$$\begin{aligned}
E'(t) &= \int_{\mathbb{R}} \{(\langle U, \Lambda U_t\rangle + \langle U_t, \Lambda U_{tt}\rangle) \\
&\qquad + \sum_{\mathbb{Z}} (\langle V_t, V_{tt}\rangle + \langle V_t, V\rangle + \langle W, W_t\rangle + \langle W_t, W_{tt}\rangle)\} \\
&= \int_{\mathbb{R}} \{\langle U, \Lambda U_{xx} - \Lambda GU\rangle + \langle U_t, (\Lambda U_{xx} - \Lambda GU)_t\rangle\} \\
&\qquad + \sum_{i\in\mathbb{Z}} \{\langle V_t^i, \Lambda[V_{tx}]_i + C^t(DF)CV_t^i - C^tW_t^i\rangle \\
&\qquad\qquad + \langle V^i, \Lambda[U_x]_i + C^tDF(CQ^i + \theta CV^i)CV^i - C^tW^i\rangle \\
&\qquad\qquad + \langle W^i, \sigma CV^i - \gamma W^i\rangle + \langle W_t^i, \sigma CV_t^i - \gamma W_t^i\rangle\} \\
&= J_1(t) + J_2(t),
\end{aligned}$$

where

$$\begin{aligned}
J_1(t) &= -\int_{\mathbb{R}} \{\langle U, \Lambda GU\rangle + \langle U_t, \Lambda GU_t\rangle + \langle U_x, \Lambda U_x\rangle + \langle U_{xt}, \Lambda U_{xt}\rangle\}, \\
J_2(t) &= \sum_{i\in\mathbb{Z}} \{\langle V_t^i, C^t(DF)CV_t^i\rangle - \langle V_t^i, C^tW_t^i\rangle \\
&\qquad + \langle V^i, C^tDF(CQ^i + \theta CV^i)CV^i\rangle - \langle V^i, C^tW^i\rangle \\
&\qquad + \langle W^i, \sigma CV^i\rangle - \langle W^i, \gamma W^i\rangle + \langle W_t^i, \sigma CV_t^i\rangle - \langle W^i, \gamma W_t^i\rangle\}.
\end{aligned}$$

Notice that if $B(x) = (b_{ij}(x))_{n\times n}$ is any bounded matrix then for any vector $u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$,

$$\langle u, Bv\rangle \le \frac{n^2}{2} \max_x |b_{ij}(x)|(\|u\|^2 + \|v\|^2).$$

DF is a bounded matrix. Thus we can find a constant $K_1 > 0$ such that

$$J_2(t) \le K_1 \sum_{i\in\mathbb{Z}} (\|V^i\|^2 + \|V_t^i\|^2 + \|W_t^i\|^2 + \|W^i\|)^2.$$

Letting $K_2 = 2\max\{K_1, \{\lambda_i g_i, \lambda_i : i = 1, 2, \cdots, n\}\}$, we have that

$$E'(t) \le K_2 E(t).$$

Therefore

$$E(t) \le E(0)e^{K_2 t} \le E(0)e^{K_2 T_0} \quad \text{for } 0 \le t \le T_0. \tag{5.21}$$

By the definition of $E(t)$, there is a constant $R_1(T_0)$, such that

$$\|(U, V, W)\|_H \le R_1(T_0) \text{ and } \|(U, V, W)_t\|_H \le R_1(T_0).$$

By (5.15) and (5.20), we can find an $R' > 0$ such that

$$\|A(U, V, W)\|_H < R'.$$

Combining this with Lemma 5.3 we have just proved the following theorem.

THEOREM 5.4. *If* $p_0 = (U, V, W)|_{t=0} \in X$ *and* $DF(u) = diag\{f_1'(u_1), f_2'(u_2), \cdots, f_n'(u_n)\}$ *is bounded, then the problem* (1.4)–(1.8) *has a unique solution* $(U, V, W) \in X$ *for all* $t > 0$.

REFERENCES

[1] J. Bell, *Some threshold results for models of myelinated nerves*, Math. Biosci., 54 (1980), pp. 181–190.

[2] J. Bell and C. Cosner, *Threshold conditions for two diffision models suggested by nerve impulse conduction*, SIAM J. Appl. Math., 46 (1986), pp. 844–855.

[3] P. Chen and J. Bell, *Global solutions and long time behavior to a FitzHugh–Nagumo myelinated axon model*, SIAM J. Math. Anal., 20 (1989), pp. 567–581.

[4] K. Chueh, C. Conley, and J. Smoller, *Positive invariant regions for systems of nonlinear parabolic equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[5] C. Cosner, *Existence of global solutions to a model of myelinated nerve axon*, SIAM J. Math. Anal., 18 (1987), pp. 703–710.

[6] R. FitzHugh, *Computation of impulse initiation and saltatory conduction in a myelinated nerve fiber*, Biophys. J., 2 (1962), pp. 11–21.

[7] A. Friedman, *Partial Differential Equations*, Holt, Rinehart & Winston, New York, 1969.

[8] J. M. Greenberg, *A note on the Nagumo equation*, Quart. J. Math., 24 (1973), pp. 307–314.

[9] P. Grindrod and B. D. Sleeman, *A model of a myelinated nerve axon: Threshold behavior and propagation*, J. Math. Biol., 23 (1985), pp. 119–135.

[10] ———, *Qualitative analysis of reaction-diffusion systems modeling unmyelinated nerve fibers*, preprint, April 1984.

[11] S. P. Hastings, *The existence of periodic solutions to Nagumo's equation*, Quart. J. Math., 25 (1974), pp. 369–378.

[12] ———, *The existence of homoclinic and periodic orbits for the FitzHugh–Nagumo equations*, Quart. J. Math., 27 (1976), pp. 123–134.

[13] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 500–544.

[14] T. Kato, *Perturbation Theory for Linear Operators*, Grundehren Math. Wiss. Einzeldarstellungen 132, Springer-Verlag, New York, Berlin, 1966.

[15] J. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proc. Inst. Radio Engineers, 50 (1962), pp. 2061–2070.

[16] J. Rauch, *Global existence for FitzHugh–Nagumo equations*, Comm. Partial Differential Equations, 1 (1976), pp. 609–621.

[17] R. Rauch and J. A. Smoller, *Qualitative theory of the FitzHugh–Nagumo equations*, Adv. in Math., 27 (1978).

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 99-149, January 1992

007

STABILITY AND HOPF BIFURCATION OF STEADY STATE SOLUTIONS OF A SINGULARLY PERTURBED REACTION-DIFFUSION SYSTEM*

ROBERT GARDNER†

Abstract. This paper presents a stability analysis of steady state solutions of a singularly perturbed reaction-diffusion system which arises as a model for predator-prey interactions. This is the first illustration of the application of a topological invariant for systems of boundary value problems called the *stability index*, recently introduced by C. Jones and the author, which counts the multiplicity of unstable eigenvalues of the linearized equations. The main results are as follows:

(i) It is shown that the spectrum σ of the linearized operator is approximated by the spectrum σ_R of a certain reduced operator defined in the small parameter limit.

(ii) Under additional assumptions on the parameters σ_R is characterized in two cases. In the first case stability is obtained for all interval sizes L, while in the second case, countably many Hopf bifurcations occur as L is increased.

Key words. stability, reaction-diffusion, steady states

AMS(MOS) subject classifications. 35B30, 35K55

1. Introduction.

We will study the diffusive predator-prey system with Dirichlet conditions

$$(1.1)\qquad \begin{aligned} &u_t = u_{xx} + f(u, v), && u(\pm L, t) = 0,\\ &v_t = \varepsilon^2 v_{xx} + g(u, v), && v(\pm L, t) = 0, \end{aligned}$$

$$(1.2)\qquad \begin{aligned} &f(u, v) = h(u) - uv = u(b-u)(u-a) - uv,\\ &g(u, v) = m(u-\gamma)v - v^2. \end{aligned}$$

Here u and v are the population densities of prey and predator species, respectively. The parameters m, a, b, γ are all positive; we assume further that $0 < a < \gamma < b$. In the absence of diffusion and boundary conditions, the kinetic equations associated with (1.1) admit a stable rest point at the origin and interior to the positive quadrant, either a stable rest point or a stable limit cycle, depending on the values of m and γ.

The object of this paper is to discuss the stability of steady state solutions of (1.1), i.e., functions $(U(x), V(x))$ satisfying

$$(1.3)\qquad \begin{aligned} &\ddot{U} + f(U, V) = 0, && U(\pm L) = 0,\\ &\varepsilon^2 \ddot{V} + g(U, V) = 0, && V(\pm L) = 0, \end{aligned}$$

wherein both components are positive; such solutions will be called *positive solutions.* The existence of a global continuum of such solutions which bifurcates as ε decreases from a solution in which $v \equiv 0$ was proved by Conway, Gardner, and Smoller [2]; see also Li [11] and Li and Lloyd [12]. The asymptotic behavior and in certain cases the uniqueness of this branch as $\varepsilon \to 0$ was discussed by Dancer [3]. In brief, we show that if ε is sufficiently small and if the other parameters are suitably chosen, then in one case, such solutions are linearly stable for all interval sizes L for which positive solutions

* Received by the editors August 13, 1990; accepted for publication (in revised form) April 29, 1991. This research was partially supported by National Science Foundation grants DMS 8802468 and DMS 8922384.

† Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003.

exist, while in a second case, the positive solution undergoes countably many Hopf bifurcations as L increases. Thus we show that if $\mathscr{L}$ is the linearized operator about such a solution, then the spectrum of $\mathscr{L}$ either lies entirely in the stable half plane $\operatorname{Re}\lambda < 0$, or, as L is increased, pairs of complex conjugate eigenvalues cross the imaginary axis $\operatorname{Re}\lambda = 0$. A precise statement of the main results is given in § 5.

It is well known for this class of equations that nonlinear stability and bifurcation are correctly predicted by the linearized analysis (see Henry [10]). In regard to large amplitude solutions of systems such as (1.3), the difficult aspect of this procedure is to determine the number and location of any unstable eigenvalues of $\mathscr{L}$. Some new methods for analyzing this question for boundary value problems were recently introduced by Gardner and Jones [7]. Similar methods with applications were previously introduced by Alexander, Gardner, and Jones [1], and by Gardner and Jones [8], [9] for traveling waves; this paper presents the first illustration of how these methods work in the stability analysis of a specific boundary value problem. In either setting, the main constructions are a complex vector bundle $\mathscr{E}(K)$ over a real 2-sphere, and an analytic function $D(\lambda)$ of the eigenvalue parameter, called the Evans function. $D(\lambda)$ is essentially the Wronskian of certain distinguished solutions of the linearized equations; its roots, counting order, coincide with the eigenvalues of $\mathscr{L}$, counting multiplicity. The main result in [7] asserts that if $K \subset \mathbb{C}$ is a simple closed curve which is disjoint from the spectrum of $\mathscr{L}$, then a certain topological invariant $c_1(\mathscr{E}(K))$, called the first Chern number of $\mathscr{E}(K)$, is an integer which is equal to the number of eigenvalues of $\mathscr{L}$ inside K (counting multiplicity). In many situations, including the problem at hand, it is possible to show that all potentially unstable eigenvalues of $\mathscr{L}$ lie in some large but fixed contour K containing $\lambda = 0$ in its interior. For such K, the underlying solution will be stable (unstable) if this invariant is equal to (greater than) zero; we therefore call $c_1(\mathscr{E}(K))$ the *stability index*.

The main concern here is to illustrate the application of the stability index to the problem described above. However, we shall briefly describe some of the main points of the general theory. The principal observation is that the candidates for eigenfunctions are the solutions of the linearized equations (written as a first-order system) which satisfy the boundary conditions at $x = -L$. These solutions span certain n-dimensional subspaces $\Phi(x, \lambda)$ of $\mathbb{C}^{2n}$ which are parametrized both by $x \in [-L, L]$ and by the eigenvalue parameter λ. Another way to say this is that they form a vector bundle over the base space $[-L, L] \times \mathbb{C}$. If λ is restricted to lie in K, it turns out that the twisting of these vector spaces over $[-L, L] \times K$ determines the number of eigenvalues of $\mathscr{L}$ inside K. The actual construction of $\mathscr{E}(K)$ consists of taking certain algebraic quotients of vector spaces associated with $\Phi(x, \lambda)$, and then gluing on suitable fibers over the "caps," $\{\pm L\} \times K^\circ$, where K° is the interior of K. The result is a bundle $\mathscr{E}(K)$ over a base space B,

$$B = \{-L\} \times K^\circ \cup [-L, L] \times K \cup \{L\} \times K^\circ,$$

which is topologically a 2-sphere. The quotient and gluing procedures are performed in such a manner that c_1 of the resulting bundle $\mathscr{E}(K)$ coincides with the winding number of the curve $D(K)$, and therefore measures the number of eigenvalues of $\mathscr{L}$ inside K. These ideas were motivated and developed at length in [1], [7] and [8], and the interested reader is referred to these papers for additional details.

These tools are of particular significance when the underlying solution is constructed by a singular perturbation procedure. In this setting the vector spaces $\Phi(x, y)$, and hence, $\mathscr{E}(K)$, frequently admit a direct sum decomposition where the summands are related to certain fast or slow behavior of solutions of the linearized equations. In other words, $\mathscr{E}(K)$ admits a Whitney sum decomposition

$$\mathscr{E}(K)=\bigoplus_{\mathrm{i}=1}^{\mathrm{k}}\mathscr{E}_{\mathrm{i}}(\mathrm{K})$$

of subbundles $\mathscr{E}_i(K)$ of $\mathscr{E}(K)$. The additive property of c_1,

$$c_1(\mathscr{E}(K))=\sum_{i=1}^{k}c_1(\mathscr{E}_i(K)),$$

provides a mechanism for computing the stability index. In particular, the topological character of c_1 permits the passage to a limit as the small parameter, say ε, tends to zero. This permits us to identify certain reduced eigenvalue problems obtained directly from the equations with $\varepsilon=0$, which, working backwards, can then be used to predict the number and location of the eigenvalues of the perturbed problem with $\varepsilon>0$. In the problem at hand, the bundle $\mathscr{E}(K)$ is two-dimensional. A Whitney sum decomposition,

$$\mathscr{E}(K)=\mathscr{E}_1(K)\oplus\mathscr{E}_2(K),$$

will be obtained where $\mathscr{E}_1(K)$ and $\mathscr{E}_2(K)$ are complex line bundles over B, which correspond, respectively, to fast and slow behavior in the linearized equations. It turns out that $\mathscr{E}_1(K)$ is trivial, so that $c_1(\mathscr{E}_1(K))=0$. The bundle $\mathscr{E}_2(K)$ is more complicated to analyze. In the limit of the small parameter ε at $\varepsilon=0$, the reduced eigenvalue problem associated with $\mathscr{E}_2(K)$ is

$$\ddot{P}=\left[(\lambda-\bar{f})-\frac{\bar{f}_v\bar{g}_u}{\lambda-\bar{g}_v}\right]P,\qquad P(-L)=P(L)=0, \tag{1.4}$$

where the partials $\bar{f}_u, \bar{f}_v, \bar{g}_u, \bar{g}_v$ are evaluated at the singular limit

$$(\bar{U}(x),\ \bar{V}(x))=\lim_{\varepsilon\to0}(U(x,\varepsilon),\ V(x,\varepsilon)). \tag{1.5}$$

The second half of the paper is devoted to the spectral analysis of problem (1.4) and is independent of the previous material. Even though (1.4) is a scalar equation, it is complicated by the fact that the eigenvalue parameter λ enters into the potential in a nonlinear, nonstandard manner; this is because it is really an eigenvalue problem for a system. Furthermore, it has variable coefficients. This is therefore not a routine computation, and we found it necessary to make some additional assumptions about the parameters in order to characterize its spectrum. Our main hypothesis is that the parameter m, which is the slope of $g=0$ in Fig. 2.2, is large. The principal tool that is used is Sturm–Liouville theory; in particular, we analyze the associated projectivized problem on $\mathbb{C}P^1$ which, in the local coordinate $z=\dot{P}/P$, is just

$$\dot{z}=G(x,\lambda)-z^2,\qquad z(-L)=\infty, \tag{1.6}$$

where $G(x,\lambda)$ is the potential in (1.4). In the stable case we show that λ with $\operatorname{Re}\lambda\geqq0$ is not an eigenvalue by locating certain invariant regions for (1.6) which keep the solution of (1.6) away from the boundary conditions $z=\infty$ at $x=+L$. In the unstable case we employ a topological argument to prove the existence of pure imaginary eigenvalues for large L. The exact eigenvalue count is then obtained by proving that $\operatorname{Re}\lambda'(L)>0$ whenever $\lambda(L)$ is a pure imaginary eigenvalue. We remark that there is an advantage to using the stability index in bifurcation problems. In particular, since the Evans function $D^{\varepsilon}(\lambda)$ for the perturbed problem uniformly approximates the Evans function $D^*(\lambda)$ for the reduced problem, and since they are both analytic, the transverse crossing of eigenvalues for the reduced problem implies transversality for the perturbed problem. Indeed, this is how we prove transversality here when $\varepsilon > 0$. This idea was introduced by Terman [14] in his stability analysis of the combustion equations.

We finally remark that Dancer has concurrently presented a stability analysis of solutions of the same equations using quite different methods (see [4]). It is also likely that the singular limit eigenvalue problem (SLEP) method could be used as well; see [6].

2. Preliminaries.

2.1. Positive solutions. The equations

$$\ddot{U}+f(U,V)=0, \qquad \varepsilon^2\ddot{V}+g(U,V)=0 \tag{2.1}$$

admit branches of solutions with $V(x)\equiv 0$. The u-component then satisfies the scalar equation

$$\ddot{U}+h(U)=0, \qquad U(-L)=U(L)=0. \tag{2.2}$$

There is a critical interval size L_h such that (2.2) has only the zero solution $U\equiv 0$ for $L<L_h$, and three solutions $0<U_1(x)<U_2(x)$ for $L>L_h$ (see [13]). This is obtained through an analysis of the "time map" associated with (2.2),

$$T(p)=\int_0^{U(p)} \frac{du}{\sqrt{2H(u)-H(U(p))}}$$

where $T(p)$ becomes infinite as $p\to 0, \bar{p}$ and has a unique local minimum at some $p^*\in(0,\bar{p})$; thus $L_h=2T(p_*)$ (see Fig. 2.1). The zero solution and $U_2(x)$ are stable, while $U_1(x)$ is unstable relative to the evolution equation associated with (2.2).

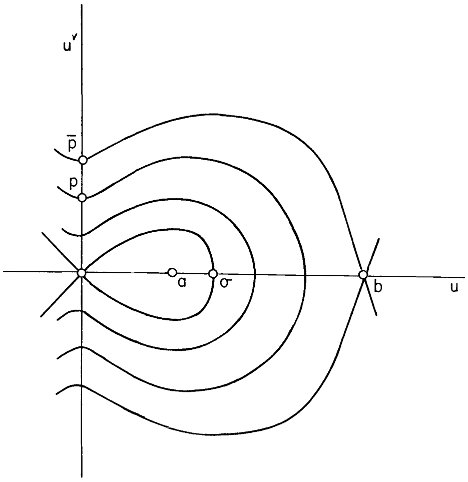


FIG. 2.1

In [2], Conway, Gardner, and Smoller show that if $L>L_h$ is fixed, then as a solution of (1.1) the solution $(U_2(x),0)$ is stable if ε is sufficiently large, but that as ε is decreased, a positive steady solution bifurcates from $(U_2(x),0)$, and stability is transferred to the bifurcating branch. However, the stability of the bifurcating branch far from the bifurcation point was left unresolved.

In a pair of papers, Dancer [3], [4] characterized several interesting properties of the branch of positive solutions of (2.1). In particular, he showed that the bifurcating

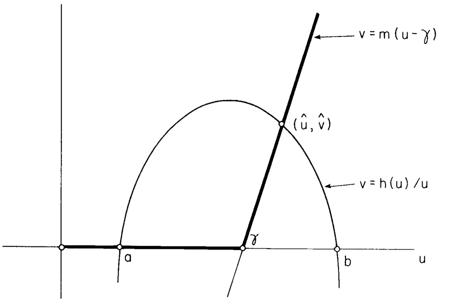


FIG. 2.2

branch $(U(x, \varepsilon), V(x, \varepsilon))$ continues globally as $\varepsilon \to 0$. Furthermore, he calculated the limit in (1.5), by showing that $(\bar{U}(x), \bar{V}(x))$ satisfies a certain reduced equation, which is obtained from (2.1) by setting $\varepsilon = 0$. The second equation $g = 0$ is then algebraic, and is satisfied whenever $v = 0$ or $v = m(u - \gamma)$. Let $p(u)$ denote the Lipschitz continuous function

$$p(u) = \begin{cases} 0 & \text{if } u \leqq \gamma, \\ m(u-\gamma) & \text{if } u \geqq \gamma. \end{cases}$$

The correct reduced equation for (2.1) when $\varepsilon = 0$ is then

$$\ddot{U} + r(U, m, \gamma) = 0, \qquad U(-L) = U(L) = 0, \tag{2.3}$$

where

$$r(U;\, m, \gamma) = (f(U), p(U)).$$

Since $a < \gamma < b$ it is easily seen that $r(U;\, m, \gamma)$ is qualitatively a cubic with roots at zero, a, and $\hat{u} \in (a, b)$, (see Fig. 2.2). In [2], the reduced problem (2.3) was studied by the time map procedure mentioned above in connection with (2.2). It was shown there that there is a unique $L_r > L_h$ such that (2.3) has exactly three solutions $0 < U_1(x;\, m, \gamma) < U_2(x;\, m, \gamma)$ for $L > L_r$ and only the zero solution for $L < L_r$. As in the case of (2.2), the zero solution and U_2 are (linearly) stable relative to the evolution equation associated with (2.3). Furthermore, branch $(U(x, \varepsilon), V(x, \varepsilon))$, which bifurcates from $U_2(x), 0)$ as ε decreases, tends to the limit

$$U_2(x;\, m, \gamma), p(U_2(x;\, m, \gamma))$$

as $\varepsilon \to 0$ (see Dancer [3]). This seems to suggest that $(U(x, \varepsilon), V(x, \varepsilon))$ is also an attractor. This, as it turns out, is not always the case. The reason for this is that the correct linearized eigenvalue problem for (2.1) when $\varepsilon = 0$ is (1.4), which is quite different from the equation obtained by linearizing (2.3) at $U_2(x;\, m, \gamma)$.

We summarize these results in the following theorem.

THEOREM 2.1. *There exists a positive branch of solutions* $(U(x, \varepsilon), V(x, \varepsilon))$ *of* (1.3) *for all* $\varepsilon \in (\varepsilon_0, 0)$ *for some* $\varepsilon_0 > 0$. *The limit* $(\bar{U}(x), \bar{V}(x))$ *as* $\varepsilon \to 0$ *of this branch exists. Furthermore,* $\bar{V}(x) = p(U_2(x;\, m, \gamma))$ *is Lipschitz continuous, and* $U_2(x;\, m, \gamma) = \bar{U}(x)$ *is the stable positive solution of* (2.3). *The limit is uniform in x.*

3. The linearized equations.

3.1. Notation. The linearization of (1.3) about $(U(x, \varepsilon), V(x, \varepsilon))$ yields a second-order linear operator $\mathscr{L}$ with domain $X = [H^2(-L, L) \cap H_0^1(-L, L)]^2$ given by

$$\mathscr{L}\begin{pmatrix} P \\ R \end{pmatrix} = \begin{pmatrix} \ddot{P} + f_u P + f_v R \\ \varepsilon^2 \ddot{R} + g_u P + g_v R \end{pmatrix}. \tag{3.1}$$

The partials f_u, f_v, g_u, and g_v are all evaluated at $(U(x, \varepsilon), V(x, \varepsilon))$. The stability of the solution is determined by the spectrum of $\mathscr{L}$; hence, we consider the eigenvalue problem

$$\mathscr{L}\begin{pmatrix} P \\ R \end{pmatrix} = \lambda \begin{pmatrix} P \\ R \end{pmatrix}, \quad \begin{pmatrix} P \\ R \end{pmatrix} \in X. \tag{3.2}$$

In order to implement the theory in [5] it is convenient to rewrite (3.2) as a first-order system:

$$\begin{aligned} \dot{P} &= Q, & \varepsilon \dot{R} &= S, \\ \dot{Q} &= (\lambda - f_u)P - f_v R, & \varepsilon \dot{S} &= -g_u P + (\lambda - g_v)R. \end{aligned} \tag*{$(3.3)_s$}$$

It will be desirable to express $(3.3)_s$ in vector form. To this end let $Y = (P, Q, R, S)^t \in \mathbb{C}^4$ and let $A(x, \lambda, \varepsilon)$ be the coefficient matrix obtained by multiplying the last two equations in $(3.3)_s$ by ε^{-1}; thus (3.3) can be expressed as

$$\dot{Y} = A(x, \lambda, \varepsilon) Y. \tag*{$(3.3)_s$}$$

In this scaling, the transition layer in the derivative of u when $u = \gamma$ is of width $\mathcal{O}(\varepsilon)$. Since this system focuses on the slow behavior for $u \neq \gamma$ we shall call $(3.3)_s$ the *slow system.*

It will also be convenient to rescale x by setting $\xi = \varepsilon^{-1} x$ and introducing fast variables,

$$p, q, r, s \text{ (at } \xi) = P, Q, R, S \text{ (at } \varepsilon\xi).$$

The equations for p, q, r, s are then easily seen to be

$$\begin{aligned} p' &= \varepsilon q, & r' &= s, \\ q' &= \varepsilon[(\lambda - f_u)p - f_v r], & s' &= -g_u p + (\lambda - g_v) r. \end{aligned} \tag*{$(3.3)_f$}$$

The partials are now evaluated at $(u(\xi, \varepsilon), v(\xi, \varepsilon)) = (U(\xi\varepsilon, \varepsilon), V(\xi\varepsilon, \varepsilon))$. Setting $y = (p, q, r, s)^t \in \mathbb{C}^4$ and $a(\xi, \lambda, \varepsilon)$ to be the coefficient matrix in $(3.3)_f$ we obtain the equivalent vector form

$$y' = a(\xi, \lambda, \varepsilon) y. \tag*{$(3.3)_f$}$$

The range of ξ is now the interval $|\xi| \leqq L_\varepsilon$ where $L_\varepsilon = L\varepsilon^{-1}$.

The boundary conditions for Y are that $P, R = 0$ at $x = \pm L$. Let U be the two-dimensional subspace of $\mathbb{C}^4$:

$$U = \{(0, Q, 0, S)^t \colon Q, S \in \mathbb{C}\}. \tag{3.4}$$

A geometric condition for λ to be an eigenvalue is that $Y(x) \in U$ at $x = \pm L$. In the fast scaling, the condition is that $y(\xi) \in U$ at $\xi = \pm L_\varepsilon$.

3.2. A crude estimate for $\sigma(\mathscr{L})$. Let $K \subset \mathbb{C}$ be the simple curve depicted in Fig. 3.1. More precisely K is the union of four segments $l_A \cup l_+ \cup l_* \cup l_-$, where l_* is the vertical segment connecting $\pm i\delta$ along the imaginary axis, l_A is the indicated portion of the circle $|\lambda| = A$, and l_+, l_- are segments connecting $i\delta, -i\delta$ to $|\lambda| = A$ which make an angle ϕ with the imaginary axis. Let $L_\pm$ be the ray with initial point at $\pm i\delta$ which coincides with $l_\pm$, and define

$$S_\phi = \{\lambda \in \mathbb{C} \colon \lambda \text{ lies to the right of } L_+ \cup l_* \cup L_-\}. \tag{3.5}$$

LEMMA 3.1. *There exists $\phi > 0$ sufficiently small and $A > 0$ sufficiently large such that if $\lambda \in S_\phi$ with $|\lambda| \geqq A$, then λ is* not *an eigenvalue of $\mathscr{L}$. The constants ϕ and A are independent of ε for $0 < \varepsilon \leqq \varepsilon_0$ for some $\varepsilon_0 > 0$.*

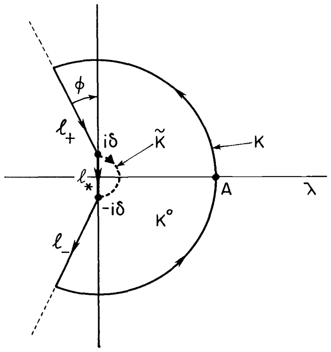


FIG. 3.1

We shall give only a brief sketch of the proof of this lemma since it follows with minor modifications from arguments in [1] and [8]. Let $\delta = |\lambda|^{-1/2}$ and $\psi = \arg \lambda$, and change scales in $(3.3)_s$ by setting $x = \delta z$. In suitably rescaled variables the equations are now

$$(3.6) \qquad \begin{aligned} \dot{P} &= Q, & \varepsilon \dot{R} &= S, \\ \dot{Q} &= (e^{i\psi} - \delta^2 f_u)P - \delta^2 f_v R, & \varepsilon \dot{S} &= -\delta^2 g_u P + (e^{i\psi} - \delta^2 g_v)R. \end{aligned}$$

If ψ is such that $\lambda \in S_\phi$, the coefficient matrix of (3.6) is uniformly hyperbolic as *both* ε and δ tend to zero. There is a pair of uniformly slow eigenvalues as $\delta \to 0$ that tend to $\pm e^{i\psi/2}$, and a pair of uniformly fast eigenvalues that tend to $\pm\varepsilon^{-1} e^{i\psi/2}$ as $\delta \to 0$. The crucial observation is that the fast eigenvalues remain uniformly separated from the slow eigenvalues as both ε and δ tend to zero. This is used to control the behavior of the two-dimensional subspace $\Phi(x, \lambda, \varepsilon)$ satisfying the boundary conditions at $x = -L$. This is accomplished by considering the flow induced by (3.6) on the space of complex lines in $\mathbb{C}^4$ or $\mathbb{C}P^3$ (see also § 5, below). The eigenvector e_f associated with the fast unstable eigenvalue $\varepsilon^{-1} e^{i\psi/2}$ of (3.6) represents a point in $\mathbb{C}P^3$ which is essentially an attracting rest point that is uniform in its attraction property as both ε and δ tend to zero. It turns out that Φ contains a one-dimensional subspace of solutions which is rapidly attracted to the span of e_f under the flow induced by (3.6) on $\mathbb{C}P^3$. This line of solutions therefore remains bounded away from the subspace U of boundary conditions for all $|x| \leqq L$.

The slow subspace of (3.6) is spanned by the two eigenvectors e_s and f_s associated with the eigenvalues tending, respectively, to $e^{i\psi/2}$ and $-e^{i\psi/2}$ as $\delta \to 0$. It follows from an invariant manifold theorem due to Fenichel (see [5]) applied to the projectivized flow on $\mathbb{C}P^3$ induced by (3.6) that there is a slow, two-dimensional subspace $\Psi(x, \lambda, \varepsilon)$ of solutions of the perturbed problem (3.6) that remains close to the slow subspace, Span$\{e_s, f_s\}$ for all x. More precisely, there is an invariant manifold $\hat{\mathcal{M}}$ for the projectivization of (3.6) that remains near the image of Span$\{e_s, f_s\}$ in $\mathbb{C}P^3$. The subspace Ψ is $\pi^{-1}\hat{\mathcal{M}}$ where $\pi : \mathbb{C}^4 \to \mathbb{C}P^3$ is the projection map. Since $\Phi(-L, \lambda) = U$ and U contains a slow direction, it follows that the intersection $\Phi \cap \Psi$ is one-dimensional. It follows that Φ contains a slow line of solutions that is rapidly attracted to the slow unstable direction spanned by e_s, since the latter is an attractor for the projectivized flow on $\mathbb{C}P^3$ relative to the flow in the invariant manifold $\hat{\mathcal{M}}$.

It follows that Φ contains two distinct solutions; the fast solution is attracted to the fast unstable direction e_f and the slow solution is attracted to the slow unstable direction e_s. In particular, Φ will lie near the space spanned by e_f and e_s at $x = L$. It is then immediate that $\Phi(L, \lambda, \epsilon)$ is transverse to the space U of boundary conditions for all small ε and δ. Additional details can be found in [1], [6], [7].

Remark. Lemma 3.1 shows that any potentially unstable eigenvalue λ of $\mathscr{L}$, i.e., any λ for which $\operatorname{Re}\lambda \geqq 0$, necessarily lies interior to the compact loop K, where K is independent of ε.

3.3. Construction of $\mathscr{E}$(K, ε). We can now describe the construction of the bundle $\mathscr{E}(K, \varepsilon)$. It will be convenient to use the slow system in $(3.3)_s$. Let

$$u_1 = (0, 0, 0, 1)^t, \qquad u_2(\gamma) = (0, 1, 0, \gamma)^t,$$

so that $U = \operatorname{Span}\{u_1, u_2(\gamma)\}$ is the subspace of $\mathbb{C}^4$ representing the boundary conditions. The parameter γ will need to be chosen carefully later. We will denote $u_2(0)$ simply by u_2. Let $V = U^{\perp}$, so that $V = \operatorname{Span}\{v_1, v_2\}$, where

$$v_1 = (0,0,1,0)^t, \qquad v_2 = (1, 0, 0, 0)^t.$$

The vector bundle $\mathscr{E}(K, \varepsilon)$ will be specified by a triple (E, B, π), where E is the total space, B is the base space, and $\pi : E \to B$ is the projection map. The base space B is homeomorphic to a real 2-sphere. It is defined as the union $B = B_l \cup B_* \cup B_r$, where

$$\begin{aligned} B_l &= \{-L\} \times \overline{K^\circ} && \text{(left cap)},\\ B_* &= [-L, L] \times K && \text{(sides)},\\ B_r &= \{+L\} \times \overline{K^\circ} && \text{(right cap)}. \end{aligned}$$

The total space E will be a certain subset of the trivial bundle $B \times [\mathbb{C}^4/V \times \mathbb{C}^4/U] \approx B \times \mathbb{C}^4$. The fibers over the caps are defined to be

$$E_l = B_l \times [\mathbb{C}^4/V \times \{\bar{0}\}], \qquad E_r = B_r \times [\{\bar{0}\} \times \mathbb{C}^4/U].$$

Here, $\bar{0}$ will be used to denote the zero element of either $\mathbb{C}^4/U$ or $\mathbb{C}^4/V$; which space is intended will be clear from the context.

It remains to define E over B_*. To this end, let $\Phi_i(x, \lambda, \varepsilon)$, $i = 1, 2$, be the solutions of $(3.3)_s$ satisfying

$$(3.7) \qquad \Phi_1(-L, \lambda, \varepsilon) = u_1, \qquad \Phi_2(-L, \lambda, \varepsilon) = u_2(\gamma).$$

Given $b = (x, \lambda) \in B$ let $x(b) = x$. We define sections $\chi_i : B_* \to B_* \times \mathbb{C}^4/V \times \mathbb{C}^4/U$ by

$$\chi_i(b, \varepsilon) = \left(\frac{L - x(b)}{2L}\Phi_i(b, \varepsilon) + V, \Phi_i(b, \varepsilon) + U\right).$$

The fibers of E over B_* are defined to be

$$E_* = \{(b, \operatorname{Span}\{\chi_1(b, \varepsilon), \chi_2(b, \varepsilon)\}) : b \in B_*\}.$$

Definition. The bundle $\mathscr{E} = \mathscr{E}(K, \varepsilon)$ is defined to be (E, B, π), where

$$E = E_l \cup E_* \cup E_r,$$

$$\pi^{-1}(b) = \begin{cases} \mathbb{C}^4/V \times \{\bar{0}\} & \text{if } b \in B_l,\\ \operatorname{Span}\{\chi_1(b, \varepsilon), \chi_2(b, \varepsilon)\} & \text{if } b \in B_*,\\ \{\bar{0}\} \times \mathbb{C}^4/U & \text{if } b \in B_r. \end{cases}$$

In order to check that this is indeed a bundle we need to check the local triviality condition where the caps fit onto the sides, i.e., over $\partial B_l \cap B_*$ and $\partial B_r \cap B_*$. This is clear for the former set by virtue of the definition of Φ_i, and hence, χ_i at $x = -L$. In particular, Φ_1 and Φ_2 span U at $x = -L$ and U complements V. At $x = +L$, Φ_1 and

Φ_2 may indeed fail to complement U; this occurs precisely when λ is an eigenvalue. Thus we see that if λ is *not* an eigenvalue of $\mathscr{L}$ for all $\lambda \in K$ then $\mathscr{E}(K, \varepsilon)$ forms a bundle over B (see [7] for further details). Although this is generally the case for most curves K, $\mathscr{E}$ depends on the parameter ε. Since we shall need to let $\varepsilon \to 0$ to compute $c_1(\mathscr{E}(K, \varepsilon))$, such a generic result would not be particularly useful. Thus we shall need to locate curves K as in Fig. 3.1 which are disjoint from the spectrum of $\mathscr{L}$ for all sufficiently small ε. It follows from Lemma 3.1 that if ϕ, A are suitably chosen then the portion of K along which $|\lambda| = A$ always satisfies this condition. It is the portion of K near the imaginary axis where all the action occurs. We shall check this condition in § 4 by using the reduced problem (1.4) defined at $\varepsilon = 0$ to obtain sufficient control over the solutions $\Phi_1(x, \lambda, \varepsilon)$ and $\Phi_2(x, \lambda, \varepsilon)$ used in the definition of $\mathscr{E}(K, \varepsilon)$ to approximate them at $x = L$. These solutions will also play an important role in obtaining the Whitney sum decomposition

$$\mathscr{E}(K, \varepsilon) = \mathscr{E}_1(K, \varepsilon) \oplus \mathscr{E}_2(K, \varepsilon),$$

which will be used later to compute $c_1(\mathscr{E}(K, \varepsilon))$.

3.4. Critical eigenvalues. A *critical eigenvalue* λ_ε of $\mathscr{L}$ is a branch of eigenvalues that tend to zero as $\varepsilon \to 0$. Such eigenvalues typically occur in the linearization about solutions with transition layers (see Fujii and Nishiura [6]). Here the singular limit $(\bar{U}(x), \bar{V}(x))$ is *continuous*, and it happens that there are no critical eigenvalues. More precisely, we have the following lemma, which generalizes a result of Dancer [3, Lemma 2].

LEMMA 3.2. *There exists* $\varepsilon_0 > 0$ *and* $\delta > 0$ *such that* $\{\lambda: \operatorname{Re} \lambda \geqq 0$ *and* $|\lambda| \leqq \delta\}$ *is in the resolvent set of* $\mathscr{L}$ *for all* $\varepsilon \in (0, \varepsilon_0]$.

Proof. Suppose to the contrary that there exists a sequence $\varepsilon_n \to 0$ for which there exist $\lambda_n \to 0$ with $\operatorname{Re} \lambda_n \geqq 0$ and an associated sequence of eigenfunctions (P_n, R_n) of $\mathscr{L}$ associated with the eigenvalues λ_n. We shall drop the subscript n and normalize the eigenfunction so that $\|P\|^2_{L^2} + \|R\|^2_{L^2} = 1$. The proof consists of showing that (i) (P, R) converges to a limit $(\bar{P}, \bar{R})$ in an appropriate sense, where $\bar{P}$ satisfies the variational equation obtained by linearizing (2.3) about $\bar{U}(x)$, and (ii) $\bar{P}$ does not vanish identically. Together, these imply that $\bar{U}(x)$ is a linearly degenerate solution of (2.3). However, from [2], it is known that $\bar{U}(x)$ is a linearly nondegenerate solution of [2.3], yielding the desired contradiction.

To prove (i) we note that from the first equation in (3.2), $\{P\}$ is uniformly bounded in $H^2 \cap H^1_0$ and therefore some subsequence converges weakly in H^2 and therefore strongly in H^1_0 to an H^2 limit $\bar{P}$. In particular, P converges to $\bar{P}$ uniformly on $[-L, L]$. Since $\{R\}$ is bounded in L^2 some further subsequence converges weakly in L^2 to a limit $\bar{R}$ with $\|\bar{R}\|_{L^2} \leqq 1$. From the second equation in (3.2) we have for all $\psi \in C^\infty_0(-L, L)$ that

$$\int_{-L}^{L} [\varepsilon^2 \ddot{\psi} R + \psi(g_u P + (g_v - \lambda))]\, dx = 0.$$

It follows that $(g_v - \lambda)R$ is weakly convergent in L^2 to both $\bar{g}_v \bar{R}$ and to $-\bar{g}_u \bar{P}$; it follows that $\bar{R} = -(\bar{g}_u / \bar{g}_v)\bar{P}$ whenever $\bar{g}_v \neq 0$. Since

$$\bar{g}_v = -m|\bar{U}(x) - \gamma|$$

is only zero at two values of x in $(-L, L)$ it follows that $\bar{R} = -(\bar{g}_u / \bar{g}_v)\bar{P}$ almost everywhere. Now take the weak formulation of the first equation in (3.2) and pass to the limit as $\varepsilon \to 0$; from the above, we have that

$$\int_{-L}^{L} \bar{P}\ddot{\psi} + \psi(\bar{f}_u - \bar{f}_v \bar{g}_u / \bar{g}_v)\bar{P}\, dx = 0$$

for all $\psi \in (_0^{-\infty}(-L, L))$. Note that the potential term is just dr/dU at $\bar{U}(x)$, where r is the nonlinear term in (2.3). It follows that $\bar{P} \in H^2 \cap H_0^1$ is a smooth solution of

$$\bar{P}^{\cdot\cdot} + r_U(\bar{U}(x)m, \gamma)\bar{P} = 0, \qquad \bar{P}(-L) = \bar{P}(L) = 0.$$

Next, we show that $\bar{P}$ is not identically zero. Let z^* denote the complex conjugate of z and let $\gamma = RR^*$. A simple computation shows that

$$\varepsilon^2 \ddot{\gamma} - 2\varepsilon^2 |\dot{R}|^2 = -2g_u \operatorname{Re} PR^* + 2 \operatorname{Re}(\lambda - g_v)\gamma.$$

Suppose that $\gamma(x)$ achieves a positive maximum at $x_* \in (-L, L)$. It follows that the left-hand side of the above is nonpositive at x_* so that

$$\operatorname{Re}(\lambda - g_r)\gamma \leqq g_u \operatorname{Re} PR^*.$$

We now use an estimate for g_v for small ε, namely, that

$$g_v = -2V(x, \varepsilon) + m(U(x, \varepsilon) - \gamma) \leqq -\tfrac{1}{2}V(x, \varepsilon)$$

for all $x \in [-L, L]$ (see Dancer [3, Lemma 1 and display (20)]). Since $\gamma \geqq 0$ and $\operatorname{Re} \lambda \geqq 0$, we have from the above that at x_*,

$$\tfrac{1}{2}V\gamma \leqq g_u \operatorname{Re} PR^*.$$

Since $g_u = mV$ it follows that

$$\gamma(x_*) \leqq 2m|P(x_*)|\gamma(x_*)^{1/2}$$

so that $\|R\|_\infty \leqq 2m\|P\|_{L^\infty}$. If $\bar{P}$ vanished identically, it would follow that $\|P\|_{L^\infty}$ converges to zero, since P converges to $\bar{P}$ uniformly. By the above this would imply that R converges to zero uniformly, contradicting our normalization $\|P\|_{L^2}^2 + \|R\|_{L^2}^2 = 1$.

As a consequence of Lemmas 3.1 and 3.2, it follows that K can only intersect $\sigma(\mathscr{L})$ along the segments l_- and l_+ in Fig. 3.1. We shall rule out this possibility later. Assuming this for the moment so that $\mathscr{E}(K, \varepsilon)$ is well defined, we see from Lemma 3.2 that the segment l_* of K can be deformed through a family of curves $l_*(\gamma)$, $0 \leqq \gamma \leqq 1$ lying inside the semicircle $\{\operatorname{Re} \lambda \geqq 0, |\lambda| = \delta\}$ with endpoints at $\lambda = \pm i\delta$ for all $\gamma \in [0, 1]$ in such a manner that (i) $l_* = l_*(0)$ and $\tilde{l} = l_*(1)$ is the semicircle $|\lambda| = \delta$, $\operatorname{Re} \lambda \geqq 0$, and (ii) $\sigma(\mathscr{L})$ does not intersect the curve K_γ obtained by replacing l_* with $l_*(\gamma)$. It follows that the bundles $\mathscr{E}(K_\gamma, \varepsilon)$ are all well defined and isomorphic so that $c_1(\mathscr{E}(K_\gamma, \varepsilon))$ is independent of $\gamma \in [0, 1]$. Thus if $\tilde{K} = K_1$ is the curve depicted in Fig. 3.1, then

$$c_1(\mathscr{E}(K, \varepsilon)) = c_1(\mathscr{E}(\tilde{K}, \varepsilon)).$$

In the following sections we shall work with $\mathscr{E}(\tilde{K}, \varepsilon)$ rather than the original bundle. This, as we shall see below, is essential because the fast-slow structure of $(3.3)_s$ breaks down near $\lambda = 0$. This splitting is used crucially in obtaining the Whitney sum decomposition of the bundle and in computing the stability index.

3.5. Projective space. One of our principal goals will be to characterize the behavior of the solutions $\Phi_i(x, \lambda, \varepsilon)$ of $(3.3)_s$ (see (3.7)) over the interval $[-L, L]$. More precisely, we need to characterize the behavior of the *span* of these solutions. An important tool for analyzing the behavior of one-dimensional subbundles of $\mathscr{E}$ is the flow induced by $(3.3)_s$ on the space of complex lines. In general, given a linear flow on $\mathbb{C}^n$ there is an induced flow on the space of complex lines in $\mathbb{C}^n$. The space of complex lines in $\mathbb{C}^n$ is called *complex projective space* and is denoted by $\mathbb{C}P^{n-1}$. $\mathbb{C}P^{n-1}$ is a compact manifold which is obtained from $\mathbb{C}^n$ by the equivalence relation $y_1 \sim y_2$ for $y_1, y_2 \in \mathbb{C}^n \backslash \{0\}$ if $y_1 = \alpha y_2$ for some $\alpha \in \mathbb{C} \backslash \{0\}$. If $\pi : \mathbb{C}^n \to \mathbb{C}P^{n-1}$ is the projection map we also use the notation $\hat{y}$ for the image $\pi(y)$ in $\mathbb{C}P^{n-1}$ of a point $y \in \mathbb{C}^n \backslash \{0\}$.

Given a linear system on $\mathbb{C}^n$,

$$y' = Ay,$$

there is an induced (nonlinear) flow on $\mathbb{C}P^{n-1}$, which we denote by

$$\hat{y}' = \hat{A}(\hat{y}).$$

It is easily checked that if e is an eigenvector of A then $\hat{e}$ is a critical point of the vector field $\hat{A}(\hat{y})$. Furthermore, if μ is the eigenvalue of A associated with e, then the eigenvalues of $d\hat{A}(\hat{e})$ are $\mu_i - \mu$, where $\{\mu_i\}$ are the remaining eigenvalues of A. In particular, if μ_1 is the eigenvalue of the largest real part and $Ae_1 = \mu_1, e_1$, then $\hat{e}_1$ is an *attracting* rest point of $\hat{A}(\hat{y})$. We shall frequently use this observation.

3.6. The spectrum of the coefficient matrix. In order to analyze the projectivization of $(3.3)_f$ we shall need to characterize the spectrum of its coefficient matrix, $a(\xi, \lambda, \varepsilon)$. We shall use the notation $\bar{f}_u(\xi)$, etc., to denote the partial derivatives of f and g at the singular limit $\bar{u}(\xi)$, $\bar{v}(\xi) = \bar{U}(x)$, $\bar{V}(x)$.

LEMMA 3.3. *For all $\lambda \in \tilde{K} \cup \tilde{K}^\circ$ and all sufficiently small ε the matrices $a(\xi, \lambda, \mu)$ have eigenvalues $\mu_i(\xi, \lambda, \varepsilon)$, $1 \leqq i \leqq 4$ with associated eigenvectors $e_i(\xi, \lambda, \varepsilon)$. For all $|\xi| \leqq L_\varepsilon$, μ_1 and μ_4 are $\mathcal{O}(1)$ as $\varepsilon \to 0$ and satisfy*

$$\operatorname{Re} \mu_4(\xi, \lambda, \varepsilon) < 0 < \operatorname{Re} \mu_1(\xi, \lambda, \varepsilon).$$

In particular, μ_1 and μ_4 tend to nonzero limits μ_{1*}, μ_{4*} as $\varepsilon \to 0$, where

$$\mu_{1*}(\xi, \lambda) = \sqrt{\lambda - \bar{g}_v(\xi)}, \qquad \mu_{4*}(\xi, \lambda) = -\sqrt{\lambda - \bar{g}_v(\xi)}. \tag{3.8}$$

The associated eigenvectors tend to limits $e_{i*}(\xi, \lambda)$ as $\varepsilon \to 0$, where

$$e_{1*}(\xi, \lambda) = (0, 0, 1, \mu_{1*}(\xi, \lambda))^t, \qquad e_{4*}(\xi, \lambda) = (0, 0, 1, \mu_{4*}(\xi, \lambda))^t. \tag{3.9}$$

The slow eigenvalues μ_2 and μ_3 are of order ε as $\varepsilon \to 0$. They may coalesce for some values of ξ; however, for ξ near $\pm L_\varepsilon$ (so that $\bar{u}(\xi)$, $\bar{v}(\xi)$ is near $(0, 0)$), they satisfy

$$\mu_2(\xi, \lambda, \varepsilon) = \sqrt{\bar{g}(\xi, \lambda)}\varepsilon + \mathcal{O}(\varepsilon^2), \qquad \mu_3(\xi, \lambda, \varepsilon) = -\sqrt{\bar{g}(\xi, \lambda)}\varepsilon + \mathcal{O}(\varepsilon^2) \tag{3.10}$$

where

$$\bar{g}(\xi, \lambda) = \lambda - \bar{f}_u(\xi) - \frac{\bar{f}_v(\xi)\bar{g}_u(\xi)}{\lambda - \bar{g}_v(\xi)}. \tag{3.11}$$

The associated eigenvectors tend to limits e_{2*}, e_{3*} as $\varepsilon \to 0$, where

$$\begin{aligned} e_{2*}(\xi, \lambda) &= \left(1, \sqrt{\bar{g}(\xi, \lambda)}, \frac{\bar{g}_u(\xi)}{\lambda - \bar{g}_v(\xi)}, 0\right)^t, \\ e_{3*}(\xi, \lambda) &= \left(1, -\sqrt{\bar{g}(\xi, \lambda)}, \frac{\bar{g}_u(\xi)}{\lambda - \bar{g}_v(\xi)}, 0\right)^t. \end{aligned} \tag{3.12}$$

The proof of a similar lemma for the same system (where ε occurs in the first rather than the second equation) can be found in [8], and will therefore be omitted. We remark that

$$\bar{g}_v(\xi) = -m|\bar{u}(\xi) - \gamma| \leqq 0;$$

hence for $\lambda \in K \cup K^\circ$, where $\tilde{K}$ is the curve depicted in Fig. 3.1 $|\arg(\lambda - \bar{g}_v(\xi))| \leqq (\pi/2) + \phi$. It follows that $\operatorname{Re}(\lambda - \bar{g}_v(\xi))^{1/2}$ is uniformly positive for all ξ and all such λ. This would not be true had we worked with the original K, since when $\bar{u} = \gamma$ and λ is real and nonpositive, $\operatorname{Re}(\lambda - \bar{g}_v(\xi)) \leqq 0$. Thus the fast-slow splitting breaks down here. The reason that this problem is controllable is that the singular limit is *continuous* at this point, so that the linearization of the reduced equation provides a good approximation for small λ. Problems where the singular limit has transition layers are much more delicate (e.g., see [8]).

3.7. The fast subbundle. We can now characterize the one-dimensional subbundle of $\mathscr{E}(K, \varepsilon)$ associated with the solution $\Phi_1(x, \lambda, \varepsilon)$ of $(3.3)_s$ which satisfies

$$\Phi_1(-L, \lambda, \varepsilon) = (0, 0, 0, 1)^t.$$

It will be convenient to change to the fast scaling by setting

$$\phi_i(\xi, \lambda, \varepsilon) = \Phi_i(\xi\varepsilon, \lambda, \varepsilon).$$

In this scaling the coefficient matrix $a(\xi, \lambda, \varepsilon)$ in $(3.3)_f$ has coefficients which depend on ξ through the presence of the partials f_u, f_v, g_u, g_v evaluated at $u(\xi, \varepsilon)$, $v(\xi, \varepsilon)$. We claim that $a_\xi = \mathcal{O}(\varepsilon)$; to this end we show that $u', v' = \mathcal{O}(\varepsilon)$. The equations for (u, v) in this scaling are

$$u' = \varepsilon w, \quad v' = z, \quad w' = -\varepsilon f(u, v), \quad z' = -g(u, v). \tag{3.13}$$

Since $w(\xi) = W(x) = U'(x)$ remains uniformly bounded, $|u'(\xi)|$ is uniformly of order ε for all ξ. Also, let $\gamma = v - p(u)$; the equation for γ is then

$$\gamma' = z - p'(u(\xi, \varepsilon))\varepsilon w(\xi, \varepsilon).$$

It follows that z remains of order ε. If this were not the case, say $|z(\xi_0)| > \delta > 0$ for some $\delta > 0$ independent of ε, the condition $|z(\xi)| > \delta/2$ would then hold on some uniform interval $|\xi - \xi_0| \leqq d$ about ξ_0, since $|z'| \leqq K$ for some $K > 0$ and all ξ. It would then follow that $|\gamma(\xi_0 + d)| \geqq \delta d/2 - \mathcal{O}(\varepsilon)$. This contradicts the fact that $\gamma(\xi, \varepsilon) \to 0$ as $\varepsilon \to 0$ uniformly in ε.

We summarize this in a lemma.

LEMMA 3.4. *There exists $M > 0$ independent of ε such that $|a_\xi(\xi, \lambda, \varepsilon)| \leqq M\varepsilon$ for all $|\xi| \leqq L_\varepsilon$, $\lambda \in \tilde{K} \cup \tilde{K}^\circ$, and $\varepsilon \in (0, \varepsilon_0]$.*

We now use this to control the solution $\phi_1(\xi, \lambda, \varepsilon)$ of $(3.3)_f$ and its projectivization $\hat{\phi}_1(\xi, \lambda, \varepsilon)$, satisfying

$$\hat{y}' = \hat{a}(\hat{y}; \xi, \lambda, \varepsilon), \tag{3.14}$$

together with the condition that $\phi_1 = u_1$ at $\xi = -L_\varepsilon$. In particular, we shall use the parametrized collection of frozen coefficient systems:

$$\hat{y}' = \hat{a}(\hat{y}; \eta, \lambda, \varepsilon). \tag*{$(3.15)_\eta$}$$

The basic idea is that the family $(3.15)_\eta$ admits a smooth curve of *attracting* rest points $\hat{e}_1(\eta, \lambda, \varepsilon)$. This follows from the remarks concerning projectivized flows in § 3.5 together with the limiting forms (3.8) and (3.9) of $\mu_1(\eta, \lambda, \varepsilon)$ and $e_1(\eta, \lambda, \varepsilon)$. In particular, these estimates imply that if $B(\eta, \lambda, \varepsilon)$ is the matrix $d\hat{a}(\hat{e}_1(\eta, \lambda, \varepsilon), \eta, \varepsilon)$ and μ is any eigenvalue of $B(\eta, \lambda, \varepsilon)$, then there exists $\sigma > 0$ independent of η and ε such that $\operatorname{Re} \mu < -\sigma$.

We can now use the "elephant trunk" lemma of [8]. The idea is to take $N(\eta, \lambda, \varepsilon)$ to be an attracting neighborhood of $\hat{e}_1(\eta, \lambda, \varepsilon)$ for $(3.15)_\eta$ and to form a tube $\Omega^+(\lambda, \varepsilon)$ in the augmented space $\mathbb{R}\times\mathbb{C}P^3$,

$$\Omega^+(\lambda, \varepsilon)=\{(\eta, \hat{y}): |\eta|\leqq L_\varepsilon \quad \text{and} \quad \hat{y}\in N(\eta, \lambda, \varepsilon)\}.$$

Using the uniform bound $\operatorname{Re} \sigma B(\eta, \lambda, \varepsilon)\leqq -\sigma$ on the spectrum of B, we see that the neighborhoods $N(\eta, \lambda, \varepsilon)$ can be chosen to contain a ball about $\hat{e}_1(\eta, \lambda, \varepsilon)$ of uniform radius $\delta>0$. (For definiteness, fix some metric ρ on $\mathbb{C}P^3$; δ can then be set relative to ρ and σ.)

We now have all the ingredients for the elephant trunk lemma in [8]. In particular, we have the following.

LEMMA 3.5. *There exists $\varepsilon_0>0$ such that for all $\varepsilon\in(0, \varepsilon_0]$, the set $\Omega^+(\lambda, \varepsilon)$ is positively invariant for the augmented flow*

$$\xi'=1, \qquad \hat{y}'=\hat{a}(y; \xi, \lambda, \varepsilon), \tag{3.16}$$

relative to the interval $|\xi|\leqq L_\varepsilon$, i.e., if $(\bar{\xi}, \hat{y}_0)\in\Omega^+(\lambda, \varepsilon)$, then the solution $(\xi, \hat{y}(\xi))$ with this data remains in $\Omega^+(\lambda, \varepsilon)$ for all $\xi\in[\bar{\xi}, L_\varepsilon]$.

This lemma is an immediate consequence of our Lemma 3.4 and Lemma 4.1 in [8]. We can now state the main result of this section.

THEOREM 3.6. *There exists $\varepsilon_0>0$ and $d>0$ independent of ε such that for all $\varepsilon\in(0, \varepsilon_0]$ and $\lambda\in\tilde{K}\cup\tilde{K}^\circ$, the solution $(\xi, \hat{\phi}_1(\xi, \lambda, \varepsilon))$ of* (3.16) *lies interior to $\Omega^+(\lambda, \varepsilon)$ for all $\xi\geqq -L_\varepsilon+d$.*

Proof. Note that from (3.8) and (3.9) we have that

$$\phi_1(-L_\varepsilon, \lambda, \varepsilon)=u_1=\tfrac{1}{2}(\lambda-\bar{g}(-L_\varepsilon))^{-1/2}[e_1(-L_\varepsilon, \lambda, \varepsilon)-e_4(-L_\varepsilon, \lambda, \varepsilon)]+\mathcal{O}(\varepsilon).$$

Thus at $\xi=-L_\varepsilon$, ϕ_1 is close to the fast subspace, i.e., the subspace spanned by e_1 and e_4. Furthermore, it has a nontrivial projection in the e_1 direction for all $\lambda\in\tilde{K}\cup\tilde{K}^\circ$ and for all $\varepsilon\in(0; \varepsilon_0]$. Let $y(\xi, \lambda, \varepsilon)$ denote the solution of the frozen system,

$$y'=a(-L_\varepsilon, \lambda, \varepsilon)y, \qquad y(-L_\varepsilon)=u_1,$$

so that $\hat{y}(\xi, \lambda, \varepsilon)$ is the associated solution of $(3.15)_{-L_\varepsilon}$. Since $\operatorname{Re}\mu_1(-L_\varepsilon, \lambda, \varepsilon)>\operatorname{Re}\mu_i(-L, \lambda, \varepsilon)$ for $i=2, 3, 4$, it follows that there exists $d>0$ depending only on $\operatorname{Re}\mu_1-\operatorname{Re}\mu_2=\mathcal{O}(1)$ such that $\hat{y}(-L_\varepsilon+d, \lambda, \varepsilon)\in N(-L_\varepsilon, \lambda, \varepsilon)$. By Lemma 3.4,

$$|e_1(\xi, \lambda, \varepsilon)-e_1(-L_\varepsilon, \lambda, \varepsilon)|\leqq K\varepsilon$$

for some $K>0$ and all $\xi\in[-L_\varepsilon, -L_\varepsilon+d]$. Hence $\hat{y}(-L_\varepsilon+d, \lambda, \varepsilon)$ lies interior to $N(-L_\varepsilon+d, \lambda, \varepsilon)$, too. Finally, it follows from Lemma 3.4 and Gronwall's inequality that

$$\rho(\hat{\phi}_1(\xi, \lambda, \varepsilon), \hat{y}(\xi, \lambda, \varepsilon))\leqq K\varepsilon$$

for some $K>0$, $\varepsilon\in(0, \varepsilon_0]$, and all $\xi\in[-L_\varepsilon, -L_\varepsilon+d]$. (Recall that ρ is a fixed metric on $\mathbb{C}P^3$.) Hence $\hat{\phi}_1(-L_\varepsilon+d, \lambda, \varepsilon)\in N(-L_\varepsilon+d, \lambda, \varepsilon)$ for sufficiently small ε_0, which, together with Lemma 3.5, proves the theorem.

We remark that Theorem 3.6 implies that $\hat{\phi}_1(L_\varepsilon, \lambda, \varepsilon)$ is near $\hat{e}_1(L_\varepsilon, \lambda, \varepsilon)$. Since $e_1(L_\varepsilon, \lambda, \varepsilon)$ is transverse to the boundary conditions U, this implies that $\phi_1(\xi, \lambda, \varepsilon)$ is *not* an eigenfunction of $\mathcal{L}$ for all $\lambda\in\tilde{K}\cup\tilde{K}^\circ$. Hence an eigenvalue λ of $\mathcal{L}$ only can arise through an eigenfunction $\phi_2(\xi, \lambda, \varepsilon)$ associated with slow behavior.

We also remark that a similar procedure can be used to determine the behavior of any solution $\hat{y}(\xi, \lambda, \varepsilon)$ of (3.14) that is near $\hat{e}_4(L_\varepsilon, \lambda, \varepsilon)$ when $\xi=L_\varepsilon$. The same considerations together with Lemmas 3.4 and 3.5 enable us to define a *negatively* invariant set $\Omega^-(\lambda, \varepsilon)$ in $[-L_\varepsilon, L_\varepsilon]\times\mathbb{C}P^3$ containing the curve $(\eta, \hat{e}_4(\eta, \lambda, \varepsilon))$ in its interior. The proof simply requires a time reversal.

Corollary 3.7. *For all* $\varepsilon \in (0, \varepsilon_0]$ *and* $\lambda \in K \cup K^\circ$, *there exists a negatively invariant neighborhood* $\Omega^-(\lambda, \varepsilon)$ *of* $(\eta, \hat{e}_4(\eta, \lambda, \varepsilon))$ *for the system* (3.16) *relative to the interval* $|\eta| \leqq L_\varepsilon$.

3.8. The slow subbundle. Next, we shall define and characterize the behavior of the slow solution $\Phi_2(x, \lambda, \varepsilon)$ of $(3.3)_s$. This will be accomplished by proving the existence of a certain two-dimensional space of solutions of $(3.3)_s$ which remains uniformly near the slow eigenspace of the coefficient matrix in $(3.3)_s$; for a precise definition, see § 3.9. Recall that Φ_2 is the solution of $(3.3)_s$ satisfying

$$\Phi_2(-L, \lambda, \varepsilon) = (0, 1, 0, \gamma)^t = u_2(\gamma),$$

where γ is a free parameter. Although an arbitrary choice of γ would suffice in defining the bundle $\mathscr{E}(\tilde{K}, \varepsilon)$ we shall require a judicious choice of γ in order that Φ_2 be uniformly approximated by the slow reduced equation (1.4) for all $x \in [-L, L]$. The reason for this is that the slow subspace is hyperbolic, so that most solutions near the slow subspace will be rapidly pushed into a fast direction in either forward or backward time.

To define Φ_2, let $\Phi_2^\circ(x, \lambda, \varepsilon)$ be the solution with $\gamma = 0$, and let $\Phi(x, \lambda, \varepsilon)$ be the span of $\Phi_1(x, \lambda, \varepsilon)$ and $\Phi_2^\circ(x, \lambda, \varepsilon)$, so that $\Phi(-L, \lambda, \varepsilon) = U$. Next, let $\Psi_i(x, \lambda, \varepsilon)$ be the solution of $(3.3)_s$ satisfying

$$\Psi_i(L, \lambda, \varepsilon) = E_i(L, \lambda, \varepsilon),$$

for $i = 2, 3, 4$, where $E_i(x, \lambda, \varepsilon) = e_i(\xi, \lambda, \varepsilon)$ are as in Lemma 3.3. Let $\Psi(x, \lambda, \varepsilon)$ be the span of the Ψ_i's, $i = 2, 3, 4$. Then Ψ is three-dimensional and Φ is two-dimensional. Furthermore, by Theorem 3.6, $\hat{\Phi}_1(L, \lambda, \varepsilon) = \hat{\phi}_1(L_\varepsilon, \lambda, \varepsilon) \in N(L_\varepsilon, \lambda, \varepsilon)$ and therefore lies near $\hat{E}_1(L, \lambda, \varepsilon)$ in $\mathbb{C}P^3$. It follows that $\Phi_1(L, \lambda, \varepsilon)$ is transverse to $\Psi(L, \lambda, \varepsilon) = \operatorname{span}\{E_i(L, \lambda, \varepsilon); i = 2, 3, 4\}$, and thus, that the subspaces $\Psi(L, \lambda, \varepsilon)$ and $\Phi(L, \lambda, \varepsilon)$ intersect transversely. Therefore, there is a unique choice of γ such that $\Phi_2 \in \Phi \cap \Psi$ for all $(x, \lambda, \varepsilon)$ for some $\Phi_2 \neq 0$. From the definition of Φ_2 we see that

$$\begin{aligned} &\Phi_2(-L, \lambda, \varepsilon) = u_2(\gamma) \in U, \\ (3.17)\qquad &\Phi_2(+L, \lambda, \varepsilon) \in \operatorname{Span}\{E_i(L, \lambda, \varepsilon)\colon i = 2, 3, 4\}. \end{aligned}$$

Note that $\gamma = \gamma(\varepsilon)$ must tend to zero as $\varepsilon \to 0$, since if this were not the case, $u_2(\gamma)$ would have a component in the E_1 direction, which is uniformly bounded away from zero as $\varepsilon \to 0$. By the next result, Lemma 3.8, this would contract the behavior of Φ_2 at $x = L$.

3.9. Hyperbolicity of the slow subbundle. The behavior of $\hat{\Phi}_2$ for all $x \in [-L, L]$ is determined by the conditions in (3.17). This can be seen by introducing the *slow subspace* $\Sigma_s(x, \lambda, \varepsilon)$ associated with $A(x, \lambda, \varepsilon)$. By Lemma 3.3, the fast eigenvectors E_1 and E_4 are simple, and hence well defined for all $x \in [-L, L]$. However, it is possible that the slow eigenvalues μ_2, μ_3 in (3.10) coalesce or become pure imaginary for certain $x \in [-L, L]$ so that the associated eigenvectors E_2 and E_3 may depend discontinuously on $(x, \lambda, \varepsilon)$. However, there is a well-defined two-dimensional (generalized) eigenspace $\Sigma_s(x, \lambda, \varepsilon)$ of $A(x, \lambda, \varepsilon)$ that does depend smoothly on the parameters $(x, \lambda, \varepsilon)$. Furthermore, there exists a basis, $E_2(x, \lambda, \varepsilon)$, $E_3(x, \lambda, \varepsilon)$ for $\Sigma_s(x, \lambda, \varepsilon)$ that depends smoothly on $(x, \lambda, \varepsilon)$. For parameter values such as $x = -L, L$ where μ_2 and μ_3 are simple, E_2 and E_3 can be taken to be eigenvectors. When μ_2 and μ_3 coincide it may be necessary to choose E_2 and E_3 to be vectors other than eigenvectors in order to make them continuous functions of $(x, \lambda, \varepsilon)$.

LEMMA 3.8. *Let $\eta > 0$ be given. There exists $\varepsilon_o > 0$ such that for all $x \in [-L, L]$, $\varepsilon \in (0, \varepsilon_o]$, $\lambda \in \tilde{K} \cup \tilde{K}^\circ$,*

$$\rho(\hat{\Phi}_2(x, \lambda, \varepsilon), \hat{\Sigma}_s(x, \lambda, \varepsilon)) < \eta. \tag{3.18}$$

Proof. Select $y \in \pi^{-1}\hat{\Phi}_2(x, \lambda, \varepsilon)$ with $\|y\| = 1$ in the Euclidean norm. Since the fast eigenvectors $E_1(x, \lambda, \varepsilon)$ and $E_4(x, \lambda, \varepsilon)$ are well defined for all $(x, \lambda, \varepsilon)$, we can express y in the form

$$y = c_1E_1 + c_4E_4 + E_s$$

for appropriate $E_s \in \Sigma_s(x, \lambda, \varepsilon)$. Since $\|y\| = 1$, there exists $d > 0$ depending only on η and the metric ρ such that if

$$|c_1|, |c_4| < d,$$

then (3.18) holds for all $(x, \lambda, \varepsilon)$. Suppose then that this condition fails to hold at some $(x_o, \lambda, \varepsilon)$, say, for example, $|c_1(x_o, \lambda, \varepsilon)| \geqq d$. Let $\xi_o = \varepsilon^{-1}x_o$ and consider the frozen system

$$y' = a(\xi_o, \lambda, \varepsilon)y, \qquad y(\xi_o) = y. \tag{3.19}$$

The eigenvalue of largest real part of the frozen system is $\mu_1(\xi_o, \lambda, \varepsilon)$; since $\operatorname{Re}(\mu_i - \mu_1)$, $i \neq 1$, is uniformly negative for all $(\xi, \lambda, \varepsilon)$, there exists $l > 0$ depending only on $\operatorname{Re}(\mu_i - \mu_1)$, $i = 2, 3, 4$ such that

$$(\xi_o + l, \hat{y}(\xi_o + l)) \in \Omega^+(\lambda, \varepsilon),$$

where $y(\xi)$ is the solution of (3.19) and $\Omega^+(\lambda, \varepsilon)$ is the positively invariant neighborhood of $(\xi, \hat{e}_1)$ constructed in Theorem 3.6. Let $\phi_2(\xi, \lambda, \varepsilon) = \Phi_2(\varepsilon\xi, \lambda, \varepsilon)$ be suitably renormalized so that $\phi_2 = y$ at $\xi = \xi_o$. It follows from standard continuous dependence theorems, together with Lemma 3.4, that $\hat{\phi}_2(\xi, \lambda, \varepsilon)$ is uniformly approximated by $\hat{y}(\xi)$ on the compact interval $[\xi_o, \xi_o + l]$, so that

$$(\xi_o + l, \hat{\phi}_2(\xi_o + l, \lambda, \varepsilon)) \in \Omega^+(\lambda, \varepsilon)$$

for sufficiently small ε. This contradicts the second statement in (3.17), provided that $\xi_o + l \leqq L_\varepsilon = \varepsilon^{-1}L$.

In the event that $|c_4| \geqq d$, we obtain $l < 0$ independent of ε such that

$$(\xi_o + l, \hat{\phi}_2(\xi_o + l, \lambda, \varepsilon)) \in \Omega^-(\lambda, \varepsilon),$$

where $\Omega^-(\lambda, \varepsilon)$ is the negatively invariant neighborhood of $(\xi, \hat{e}_4)$ of Corollary 3.7, provided that $\xi_o + l \geqq -L_\varepsilon$, the argument is the same as for $|c_1| \geqq d$ after a time reversal. This contradicts the first statement in (3.17) at $x = -L$.

It only remains to treat the case where $\xi_o + l$ is exterior to the interval $|\xi| \leqq L_\varepsilon$. Near each boundary we need only consider one case. For ξ_o near $-L_\varepsilon$ (respectively, $+L_\varepsilon$) the case that requires further discussion is the one where $|c_4| \geqq d$ (respectively, $|c_1| \geqq d$). For ξ_o near $-L_\varepsilon$ we still use the solution $y(\xi)$ of the frozen system (3.19) to approximate $\phi_2(\xi, \lambda, \varepsilon)$ in backward time. We can no longer conclude that $\hat{\phi}_2$ enters $\Omega^-(\lambda, \varepsilon)$ in backward time, since the interval $[\xi_o + l, \xi_o]$ may no longer be entirely contained in $|\xi| \leqq L_\varepsilon$. However, the E_4-component is the most rapidly growing component of the soution $y(\xi)$ in backward time; it therefore follows from the initial condition that the E_4-component of $y(\xi)/\|y\|$ remains uniformly bounded away from zero on $\{-L_\varepsilon \leqq \xi \leqq \xi_o\}$.

This implies that $\hat{y}(-L_\varepsilon)$ is uniformly bounded away from $\hat{\Sigma}_s(-L, \varepsilon, \lambda)$, and, in particular, $\hat{u}_2(\gamma)$, the initial value for $\hat{\phi}_2$ at $\xi = -L_\varepsilon$ in (3.17). By Lemma 3.4, $\hat{y}(\xi)$ uniformly approximates $\hat{\phi}_2$ on this interval; hence the first statement in (3.17) would be violated. The second case, where $|c_1| \geqq d$ for ξ_o near L_ε, is treated similarly.

3.10. Approximation by the slow reduced equation. We can now show that the slow solution $\hat{\Phi}_2$ is approximated by a certain solution of the slow reduced equation (1.4). To this end, rewrite (1.4) as a system

$$\dot{\Gamma}_* = B_*(x, \lambda)\Gamma_*, \tag{3.20}$$

where $\Gamma_* = (P_*, Q_*)^t$ and

$$B_*(x, \lambda) = \begin{pmatrix} 0 & 1 \\ \bar{g}(x, \lambda) & 0 \end{pmatrix};$$

here, $\bar{g}(x, \lambda)$ is as in (3.11). Since (3.20) is linear it induces a flow on $\mathbb{C}P^1$ which we denote by

$$\hat{\Gamma}^{\cdot}_* = \hat{B}_*(\hat{\Gamma}_*; x, \lambda). \tag{3.21}$$

It will be convenient to cover $\mathbb{C}P^1$ $(=S^2)$ with two coordinate patches: $w_* = P_*/Q_*$ and $z_* = Q_*/P_*$. The equations for w_* and z_* are easily seen to be

$$\dot{w}_* = 1 - \bar{g}(x, \lambda)w_*^2, \tag{3.22}$$

$$\dot{z}_* = \bar{g}(x, \lambda) - z_*^2. \tag{3.23}$$

The solution of (3.21) of interest is the solution $\hat{\Gamma}_*(x, \lambda)$ representing the boundary condition at $x = -L$, i.e., the solution satisfying the initial condition

$$\pi^{-1}\hat{\Gamma}_*(-L, \lambda) = \text{Span}\,(0, 1).$$

In terms of the coordinate w_* this condition is

$$w_*(-L, \lambda) = 0.$$

LEMMA 3.9. *Let*

$$\Phi_2(x, \lambda, \varepsilon) = (P, Q, R, S) \quad \textit{at } (x, \lambda, \varepsilon)$$

and let

$$\Gamma(x, \lambda, \varepsilon) = (P, Q)^t \quad \textit{at } (x, \lambda, \varepsilon).$$

Given $\sigma > 0$, there exists $\varepsilon_o > 0$ such that

$$\rho(\hat{\Gamma}(x, \lambda, \varepsilon), \hat{\Gamma}_*(x, \lambda)) < \sigma$$

for all $\lambda \in \tilde{K} \cup \tilde{K}^\circ$, $x \in (-L, L]$, and $\varepsilon \in (0, \varepsilon_o]$; here ρ is a fixed metric on $\mathbb{C}P^1$.

Proof. We will show that $\hat{\Gamma}(x, \lambda, \varepsilon)$ satisfies a system of the form

$$\hat{\Gamma}^{\cdot} = \hat{B}(\hat{\Gamma}; x, \lambda, \varepsilon),$$

where $\hat{B}(\hat{\Gamma}; x, \lambda, \varepsilon)$ is uniformly approximated by $\hat{B}_*(\hat{\Gamma}; x, \lambda)$ for all $\hat{\Gamma}$ in $\mathbb{C}P^1$, $x \in [-L, L]$, and $\lambda \in \tilde{K} \cup \tilde{K}_o$. The lemma then follows from standard continuous dependence theorems for flows.

The first task is to choose E_2 and E_3 so that they depend smoothly on $(x, \lambda, \varepsilon)$. The only difficulty occurs when $\mu_2 = \mu_3$. Since the eigenvectors of matrices with nonsimple eigenvalues can depend discontinuously on parameters, some care must be taken here. This possibility indeed arises for certain values of x. However, it is easily checked that when this occurs, the generalized null space of $A(x, \lambda, \varepsilon)$ is spanned by vectors of the form

$$E_2=(1,0,J(x,\lambda),0)^t+\mathcal{O}(\varepsilon), \qquad E_3=(0,1,0,0)^t+\mathcal{O}(\varepsilon), \tag{3.24}$$

where

$$J(x,\lambda)=\frac{\bar{g}_u(x)}{\lambda-\bar{g}_v(x)}.$$

The computation is straightforward though tedious and will be omitted. Note that if E_2 and E_3 are chosen as in (3.24) they span the same space as e_{2*}, e_{3*} in (3.12), modulo $\mathcal{O}(\varepsilon)$ terms. This is the required basis for $\Sigma_s(x, \lambda, \varepsilon)$.

Next, select

$$Y=(P,Q,R,S)^t \in \pi^{-1}\hat{\Phi}_2(x,\lambda,\varepsilon)$$

to be a vector of unit length. For some coefficients α_i we have that

$$Y=\sum_{i=1}^{4}\alpha_i E_i(x,\lambda,\varepsilon).$$

Let $d\in(0,1)$ be given. Since $\Sigma_s(x,\lambda,\varepsilon)$ is spanned by E_2, E_3 at (x,λ,ε), it follows that there exists $\varepsilon_o>0$, $\eta=\eta(d)$ such that (3.18) implies that $|\alpha_1|\leqq d$ and $|\alpha_4|\leqq d$ for $\lambda\in\tilde{K}\cup\tilde{K}^\circ$ and $\varepsilon\in(0,\varepsilon_o]$. It then follows from (3.24) that

$$\begin{aligned} P&=\alpha_2+\mathcal{O}(d), & Q&=\alpha_3+\mathcal{O}(d),\\ R&=\alpha_2 J(x,\lambda)+\mathcal{O}(d), & S&=\mathcal{O}(d). \end{aligned} \tag{3.25}$$

Since Y is a unit vector there exists $k>0$ depending only on $|J|$ and on d such that $\max\{|P|,|Q|\}\geqq k$. Suppose first that $|Q|\geqq k$. Then $(P/Q,\ R/Q,\ S/Q)$ provides a coordinate patch on $\mathbb{C}P^3$ at $\hat{Y}$. If $w=P/Q$, the equation for w is then

$$\dot{w}=1-(\lambda-f_u)w^2+f_vRP/Q^2.$$

From (3.25) and $|Q|\geqq k$ we have that

$$RP/Q^2=J(x,\lambda)w^2+\mathcal{O}(d).$$

Noting that

$$\begin{aligned}\bar{g}(x,\ )&=\lambda-\bar{f}_u-\frac{\bar{f}_v\bar{g}_u}{\lambda-\bar{g}_v}\\ &=\lambda-\bar{f}_u-\bar{f}_vJ(x,\lambda)\end{aligned}$$

and that f_u, f_v are uniformly approximated by $\bar{f}_u, \bar{f}_v$ for small ε, it follows from the above that

$$w'=1-\bar{g}(x,\lambda)w^2+\mathcal{O}(d).$$

Hence if $|Q|\geqq k$, $\hat{\Gamma}$ satisfies an equation which in the local coordinate w uniformly approximates the reduced equation for w_*, (3.22). In the event that $|P|\geqq k$, a similar computation leads to the equation

$$\dot{z}=\bar{g}(x,\lambda)-z^2+\mathcal{O}(d),$$

which uniformly approximates (3.24). Hence $\hat{\Gamma}(x,\lambda,\varepsilon)$ is uniformly approximated by $\hat{\Gamma}_*(x,\lambda)$ for $\varepsilon\in(0,\varepsilon_o]$ and $\lambda\in\tilde{K}\cup\tilde{K}^\circ$.

3.11. The reduced slow subbundle. Suppose for the moment that the eigenvalue problem in (1.4) is well understood. In particular, suppose that we can locate curves $\tilde{K}$ as in Fig. 3.1 that are disjoint from the spectrum of (1.4). Later, in § 4, we shall indicate how this can be done in certain situations. The results of the preceding section suggest defining a reduced line bundle $\mathscr{E}_{2*}(\tilde{K})$ whose fibers are defined via the reduced system (3.20). To this end let

$$\tilde{U}=\{(0,Q)\colon Q\in\mathbb{C}\},\qquad \tilde{V}=\{(P,0)\colon P\in\mathbb{C}\};$$

so that $\tilde{U}=\operatorname{Span}(0,1)^t$ represents the boundary conditions for (3.20) and $\mathbb{C}^2=\tilde{U}\oplus\tilde{V}$. Define a section $\xi_*\colon B_*\to B_*\times[\mathbb{C}^2/\tilde{V}\times\mathbb{C}^2/\tilde{U}]$ by

$$\tilde{\xi}_*(b)=\left(b,\frac{L-x(b)}{2L}\Gamma_*(b)+\hat{V},\Gamma_*(b)+\tilde{U}\right)$$

for all $b\in B_*$. Over the caps, put

$$\tilde{\xi}_l(b)=(b,(0,1)^t+\tilde{V}_o,\bar{0})\qquad(b\in B_L),$$
$$\tilde{\xi}_r(b)=(b,\bar{0},\tilde{E}_{2*}(L,\lambda)+\tilde{U}_o)\ (b\in B_R),$$

where $\tilde{E}_{2*}(L,\lambda)=(1,\sqrt{\bar{g}(L,\lambda)})$. Finally, define the fibers of $\tilde{\mathscr{E}}_{2*}(K)$ by

$$\pi^{-1}(b)=\begin{cases}\operatorname{Span}\tilde{\xi}_l(b), & b\in B_l,\\ \operatorname{Span}\tilde{\xi}_*(b), & b\in B_*,\\ \operatorname{Span}\tilde{\xi}_r(b), & b\in B_r.\end{cases}\tag{3.26}$$

Since we are assuming at this point that $\tilde{K}$ is disjoint from the eigenvalues of (1.4) it follows that $\tilde{\mathscr{E}}_{2*}(\tilde{K})$ is a line bundle over B.

Next we define another line bundle $\mathscr{E}_{2*}(\tilde{K})$ which turns out to be the limit of $\mathscr{E}_2(\tilde{K},\varepsilon)$ as $\varepsilon\to0$. To this end, embed the solution $\Gamma_*(x,\lambda)$ of (3.20) satisfying $\Gamma_*(-L,\lambda)=(0,1)^t$ into $\mathbb{C}^4$ by setting for $b\in B_*$

$$\Phi_{2*}(b)=(P_*(b),Q_*(b),J(b)P_*(b),0)^t.\tag{3.27}$$

Next define a map $\xi_*\colon B_*\to\mathbb{C}^4/V\times\mathbb{C}^4/U$ by setting

$$\chi_{2*}(b)=\left(\frac{L-x(b)}{2L}\Phi_{2*}(b)+V,\Phi_{2*}(b)+U\right).$$

Next, define $\xi_l(b)$, $\xi_r(b)$ over the caps by setting

$$\xi_l(b)=(u_2+V,\bar{0}),\qquad \xi_r(b)=(\bar{0},E_{2*}(b)+U).$$

The bundle $\mathscr{E}_{2*}(\tilde{K})$ is then the line bundle over B, $\pi\colon E\to B$, where

$$\pi^{-1}(b)=\begin{cases}(b,\operatorname{Span}\xi_l(b)), & b\in B,\\ (b,\operatorname{Span}\chi_{2*}(b)), & b\in B_*,\\ (b,\operatorname{Span}\xi_r(b), & b\in B_r.\end{cases}\tag{3.28}$$

It is easily seen that this indeed forms a bundle. The local triviality condition is immediate over $\bar{B}_l\cap B_*$. This condition is also satisfied over $\bar{B}_r\cap B_*$. This is seen by noting that $\mathbb{C}^4/U$ identifies vectors with the same first and third components. From (3.12) and (3.27) we see that

$$\Phi_{2*}(b)=P_*(b)E_{2*}(b)+(0,Q_*(b)-P_*(b)\sqrt{\bar{g}(b)},0,0)^t;$$

thus $\Phi_{2*}(b)$ is a scalar multiple of $E_{2*}(b)$ mod U.

Finally we note that the bundle $\tilde{\mathscr{E}}_{2*}(\tilde{K})$ is equivalent to $\mathscr{E}_{2*}(\tilde{K})$. This can be seen by noting that the sections $\tilde{\xi}_l, \tilde{\xi}_*, \tilde{\xi}_r$ and ξ_l, χ_{2*}, ξ_r can be used to construct an explicit bundle isomorphism by mapping $\tilde{\xi}_*(b)$ onto $\xi_*(b)$ and extending linearly. Since isomorphic bundles have the same Chern numbers we have proved the following result.

LEMMA 3.10. *The line bundles* $\tilde{\mathscr{E}}_{2*}(\tilde{K})$ *and* $\mathscr{E}_{2*}(\tilde{K})$ *defined, respectively, by* (3.26) *and* (3.28) *are isomorphic, and* $c_1(\tilde{\mathscr{E}}_{2*}(\tilde{K}))=c_1(\mathscr{E}_{2*}(\tilde{K}))$.

3.12. The Whitney sum decomposition. The Whitney sum decomposition

$$\mathscr{E}(\tilde{K},\varepsilon)=\mathscr{E}_1(\tilde{K},\varepsilon)\oplus\mathscr{E}_2(\tilde{K},\varepsilon) \tag{3.29}$$

can now be specified more precisely. Let

$$\bar{\Phi}_i(\lambda,\varepsilon)=\Phi_i(L,\lambda,\varepsilon)+U, \qquad i=1,2,$$

$$\bar{e}_i(\lambda,\varepsilon)=E_i(L,\lambda,\varepsilon)+U, \qquad i=1,2.$$

Then $\{\bar{\phi}_1,\bar{\phi}_2\}$ and $\{\bar{e}_1,\bar{e}_2\}$ both form bases for $\mathbb{C}^4/U$. It follows that there exists a nonsingular matrix $f^\varepsilon\in Gl(2;\mathbb{C})$ such that

$$\bar{\phi}_j(\lambda,\varepsilon)=f^\varepsilon_{j,1}(\lambda)\bar{e}_1(\lambda,\varepsilon)+f^\varepsilon_{j,2}(\lambda)\bar{e}_2(\lambda,\varepsilon) \tag{3.30}$$

for $j=1,2$. Theorem 3.6 and Lemma 3.8 provide certain information about $f^\varepsilon_{i,j}(\lambda)$ as $\varepsilon\to 0$, namely, there exists $k>0$ such that for all $\lambda\in\tilde{K}$ and $\varepsilon\in(0,\varepsilon_o]$,

$$\begin{aligned}&|f^\varepsilon_{1,1}(\lambda)|,\ |f^\varepsilon_{2,2}(\lambda)|\geqq k,\\ &\lim_{\varepsilon\to 0}f^\varepsilon_{1,2}(\lambda)=\lim_{\varepsilon\to 0}f^\varepsilon_{2,1}(\lambda)=0,\ \text{uniformly.}\end{aligned} \tag{3.31}$$

Now f^ε is only defined for λ near $\tilde{K}$. In order to obtain the Whitney sum decomposition in (3.29) we need to find a trivialization $B_R\times\{\bar{0}\}\times\mathbb{C}^4/U$ over the entire right cap that coincides with (3.30) over $\tilde{K}\times\{L\}$. Let

$$\bar{\gamma}_1(\lambda,\varepsilon)=\bar{e}_1(\lambda,\varepsilon)+\frac{f^\varepsilon_{1,2}(\lambda)}{f^\varepsilon_{1,1}(\lambda)}\bar{e}_2(\lambda,\varepsilon),$$

$$\bar{\gamma}_2(\lambda,\varepsilon)=\bar{e}_2(\lambda,\varepsilon)+\frac{f^\varepsilon_{2,1}(\lambda)}{f^\varepsilon_{2,2}(\lambda)}\bar{e}_1(\lambda,\varepsilon).$$

Then $\{\bar{\gamma}_1,\bar{\gamma}_2\}$ is a basis for $\mathbb{C}^4/U$ for $\lambda\in\tilde{K}$. By (3.31) the second term in each expression tends to zero uniformly as $\varepsilon\to 0$ for $\lambda\in\tilde{K}$. Let $K_1\subset\tilde{K}^\circ$ be a simple closed curve close enough to $\tilde{K}$ so that $f^\varepsilon(\lambda)$ is well defined in the region between K_1 and $\tilde{K}$ for $\varepsilon\in(0,\varepsilon_o]$. Let $\psi(\lambda)$ be a smooth function satisfying

$$\psi(\lambda)=\begin{cases}1 & \text{for } \lambda\in\tilde{K},\\ 0 & \text{for } \lambda \text{ inside } K_1.\end{cases}$$

Finally, let

$$\bar{g}_1(\lambda,\varepsilon)=\bar{e}_1(\lambda,\varepsilon)+\frac{f^\varepsilon_{1,2}(\lambda)}{f^\varepsilon_{11}(\lambda)}\bar{e}_2(\lambda,\varepsilon)\psi(\lambda),$$

$$\bar{g}_2(\lambda,\varepsilon)=\bar{e}_2(\lambda,\varepsilon)+\frac{f^\varepsilon_{21}(\lambda)}{f^\varepsilon_{22}(\lambda)}\bar{e}_1(\lambda,\varepsilon)\psi(\lambda);$$

then $\{\bar{g}_1(\lambda,\varepsilon),\bar{g}_2(\lambda,\varepsilon)\}$ is a basis for $\mathbb{C}^4/U$ for all $\lambda\in\tilde{K}\cup\tilde{K}^\circ$ and $\varepsilon\in(0,\varepsilon_o]$, which equals $\{\bar{\gamma}_1(\lambda,\varepsilon),\bar{\gamma}_2(\lambda,\varepsilon)\}$ when $\lambda\in\tilde{K}$.

Now let $\chi_1(b, \varepsilon)$ and $\chi_2(b, \varepsilon)$ be the sections of $\mathscr{E}(\tilde{K}, \varepsilon)$ restricted to the sides B_*, which were used to define $\mathscr{E}(\tilde{K}, \varepsilon)$. The summands $\mathscr{E}_i(\tilde{K}, \varepsilon)$ are triples (E_i, B, π_i), where

$$\pi_i^{-1}(b) = \begin{cases} \text{Span}\,\{(u_i + V, \bar{0})\} & \text{if } b \in B_l, \\ \text{Span}\,\chi_i(b, \varepsilon) & \text{if } b \in B_*, \\ \text{Span}\,\bar{g}_i(b, \varepsilon) & \text{if } b \in B_r, \end{cases} \tag{3.32}$$

where $u_2 = u_2(\gamma)$ is as in § 3.8. It follows from the above that $\mathscr{E}_i(\tilde{K}, \varepsilon)$ forms a line bundle $i = 1, 2$; by construction, we see that (3.29) holds.

We remark that the above construction of the summands is somewhat different than in the case of traveling waves in which the spatial domain is infinite. The difference between the two is that, in the latter case, the behavior of each $\hat{\phi}_i$ is precisely determined at $\xi = +\infty$ by the equations. In the case of boundary value problems, each $\hat{\phi}_i$ is only approximately determined at $x = +L$.

We also remark that the function $f^\varepsilon(\lambda)$, which is called the *gluing map* of $\mathscr{E}(\tilde{K}, \lambda)$, plays a central role in the general theory in [1] and [7] and, in particular, in characterizing the first Chern number of $\mathscr{E}(\tilde{K}, \lambda)$. Briefly, this is seen by noting that f^ε is a map from $\tilde{K} \approx S^1$ into $Gl(2; \mathbb{C})$, and so represents a class in $\pi_1(Gl(2; \mathbb{C})) \cong \mathbb{Z}$. It turns out that $c_1(\mathscr{E}(\tilde{K}))$ can be characterized explicitly as the winding number of the curve $\det f^\varepsilon(\tilde{K})$. Furthermore, the relation between $c_1(\mathscr{E}(\tilde{K}, \varepsilon))$ and the number of eigenvalues of $\mathscr{L}$ inside $\tilde{K}$ is obtained by proving that the Evans function $D(\lambda)$ is a nonzero, constant multiple of $\det f^\varepsilon$. See [7] for additional details.

3.13. Continuation to the reduced problem. We can now prove the main theorem of this section. We now return to our original notation wherein K denotes the curve in Fig. 3.1 passing through the origin and $\tilde{K}$ is the curve in the figure with the origin in its exterior. All results of the previous section were proved for the bundle $\mathscr{E}(\tilde{K}, \varepsilon)$.

THEOREM 3.11. *Suppose that $\lambda \in K$ in Fig.* 3.1 *is* not *an eigenvalue of* (1.4) *for all $\lambda \in K$. There exists ε_o such that for all $\varepsilon \in (0, \varepsilon_o]$ the number of eigenvalues inside K of the perturbed problems* (3.2) *counting multiplicity is equal to the number of eigenvalues inside K of the reduced problem* (1.4) *counting multiplicity.*

Proof. The main result in [7] asserts that the multiplicity of eigenvalues of (3.2) (respectively, (1.4)) inside K is equal to $c_1(\mathscr{E}(K, \varepsilon))$ (respectively, $c_1(\tilde{\mathscr{E}}_{2*}(K))$). The topological character of c_1 together with our previous approximation theorems will enable us to equate these two quantities.

To begin with, Lemma 3.2 in § 3.4 implies that (3.2) has no spectrum inside the semicircle, $\text{Re}\,\lambda \geqq 0$, $|\lambda| = \delta$, with δ independent of ε. The same statement is also valid for the spectum of (1.4); this is because $U(x)$ is a linearly nondegenerate solution of (2.3) and (1.4) is precisely the variational equation for (2.3) at $U(x)$ when $\lambda = 0$. Even though the potential in (1.4) is singular for $\lambda < 0$, the result follows from a similar but simpler limiting argument for sequences of eigenvalues $\lambda \to 0$ with $\text{Re}\,\lambda \geqq 0$ as that used in the proof of Lemma 3.2; it will therefore be omitted. We can therefore deform the curve K in Fig. 3.1 to $\tilde{K}$ via the homotopy K_γ to conclude that

$$c_1(\mathscr{E}(K, \varepsilon) = c_1(\mathscr{E}(\tilde{K}, \varepsilon)), \qquad c_1(\tilde{\mathscr{E}}_{2*}(K)) = c_1(\tilde{\mathscr{E}}_{2*}(\tilde{K})).$$

The additive property of c_1 together with (3.29) implies that

$$c_1(\mathscr{E}(\tilde{K}, \varepsilon)) = c_1(\mathscr{E}_1(\tilde{K}, \varepsilon)) + c_1(\mathscr{E}_2(\tilde{K}, \varepsilon)).$$

The proof will be complete if it can be shown that

(i) $c_1(\mathscr{E}_1(\tilde{K}, \varepsilon)) = 0$ for $\varepsilon \in (0, \varepsilon_o]$,

(ii) $c_1(\mathscr{E}_2(\tilde{K}, \varepsilon)) = c_1(\tilde{\mathscr{E}}_{2*}(\tilde{K}))$ for $\varepsilon \in (0, \varepsilon_o]$.

To prove (i) we show that $\mathscr{E}_1(\tilde{K}, \varepsilon)$ is a trivial bundle for all such ε. To this end we show that there is a global, nonvanishing section of $\mathscr{E}_1(\tilde{K}, \varepsilon)$. The existence of such a section is an immediate consequence of Theorem 3.6, which asserts that $\hat{\Phi}_1(L, \lambda, \varepsilon)$ lies near $\hat{E}_1(L, \lambda, \varepsilon)$ for all $\lambda \in \tilde{K} \cup \tilde{K}^\circ$. It follows that the entries $f^\varepsilon_{1,1}(\lambda)$ and $f^\varepsilon_{1,2}(\lambda)$ of the gluing map are well defined and continuous for *all* $\lambda \in \tilde{K} \cup \tilde{K}^\circ$. Thus the characterization (3.30) of $\bar{\phi}_1(\lambda, \varepsilon)$ is valid for all such λ, and we are justified in replacing $\bar{g}_1(b, \varepsilon)$ with $\bar{\phi}_1(b, \varepsilon)$ in (3.32) when $i = 1$. Since $\bar{\phi}_1$ continuously extends $\chi_1(b)$ to all of $\tilde{K} \cup \tilde{K}^\circ$, this provides the desired nonvanishing section.

To prove (ii) we need to show that $\mathscr{E}_2(\tilde{K}, \varepsilon)$ approaches $\mathscr{E}_{2*}(\tilde{K})$ as $\varepsilon \to 0$. Since c_1 is a homotopy invariant, these two bundles would then have to have the same first Chern number; (ii) then follows from Lemma 3.10.

In order to specify what it means for two line bundles to be close together, it is useful to introduce a certain map $\hat{e} : B \to \mathbb{C}P^{n-1}$ associated with a given line bundle (E, B, π) whose fibers are complex lines in $\mathbb{C}^n$. Given such a bundle, $\hat{e}(b)$ is the point in the compact metric space $\mathbb{C}P^{n-1}$ associated with the complex line $\pi^{-1}(b)$. Thus the notion of a complex line bundle over B is equivalent to the specification of a continuous map $\hat{e} : B \to \mathbb{C}P^{n-1}$. More precisely, let $\Gamma_1(\mathbb{C}^n)$ be the canonical bundle over $\mathbb{C}P^{n-1}$, i.e., the fiber in $\Gamma_1(\mathbb{C}^n)$ over a point $\hat{e} \in \mathbb{C}P^{n-1}$ is the complex line in $\mathbb{C}^n$ associated with $\hat{e}$. We then have the commutative diagram

$$\begin{tikzcd}
E \arrow[r, "\hat{e}^*"] \arrow[d, "\pi"] & \Gamma_1(\mathbb{C}^n) \arrow[d] \\
B \arrow[r, "\hat{e}"] & \mathbb{C}P^{n-1};
\end{tikzcd}$$

the pullback $\hat{e}^*\Gamma_1(\mathbb{C}^n)$ is then isomorphic to E. Thus two complex line bundles (E_i, B, π_i) over the same base space are close togeher if their associated maps $\hat{e}_i : B \to \mathbb{C}P^{n-1}$ are close in the topology of $\mathbb{C}P^{n-1}$.

Let $\hat{e}_2(b, \varepsilon)$ and $\hat{e}_{2*}(b)$ be the maps into $\mathbb{C}P^3$ associated with $\mathscr{E}_2(\tilde{K}, \varepsilon)$ and $\mathscr{E}_{2*}(\tilde{K})$, respectively. We need to show that $\hat{e}_2(b, \varepsilon) \to \hat{e}_{2*}(b)$ uniformly in $\mathbb{C}P^3$ for $b \in B$ as $\varepsilon \to 0$. Over the caps B_l and B_r this is immediate, since $u_2(\gamma) \to u_2(0) = u_2$ and $E_2(L, \lambda, \varepsilon) \to E_{2*}(L, \lambda)$ as $\varepsilon \to 0$. The convergence of $\hat{e}_2(b, \varepsilon)$ to $\hat{e}_{2*}(b)$ over the sides B_* will be established if it can be shown that $\hat{\Phi}_2(x, \lambda, \varepsilon)$ converges uniformly to $\hat{\Phi}_{2*}(x, \lambda)$, where Φ_{2*} is as in (3.27). This follows immediately from the form (3.27) for Φ_{2*}, Lemma 3.9, and the expression (3.25) for a unit vector $Y \in \pi^{-1}\hat{\Phi}_2(x, \lambda, \varepsilon)$. In particular, Lemma 3.9 states that if

$$Y = (P, Q, R, S) \in \pi^{-1}\hat{\Phi}_2(x, \lambda, \varepsilon),$$

$$Y_* = (P_*, Q_*, R_*, S_*) \in \pi^{-1}\hat{\Phi}_{2*}(x, \lambda),$$

then

$$(P, Q)^\wedge \to (P_*, Q_*)^\wedge \quad (\text{in } \mathbb{C}P^1).$$

Thus we may select the representatives Y and Y_* so that (P, Q) approaches (P_*, Q_*) in $\mathbb{C}^2$. Finally, (3.25) implies that

$$P = J(x, \lambda)R + \mathcal{O}(d), \qquad P_* = J(x, \lambda)R_*,$$

where the first equation holds for arbitrary d and sufficiently small ε depending on d. Thus $\hat{Y} \to \hat{Y}_*$ in $\mathbb{C}P^3$ uniformly for $x \in (-L, L]$.

4. Analysis of the reduced equation. We have now shown that the spectrum σ_ε of the perturbed operator $\mathscr{L}$ is approximated by the spectrum of the reduced problem (1.4), which we shall denote by σ_R. The spectral analysis of (1.4) is complicated by the fact that it has variable coefficients and that the eigenvalue parameter λ occurs in a nonstandard, nonlinear manner. In this section we shall present a detailed analysis of the spectrum of (1.4) in the special case that the parameter $\mu = \sqrt{m}$ is sufficiently large. In this case the line $v = m(u - \gamma)$ is almost vertical. In terms of the singular limit $(\bar{U}(x), \bar{V}(x))$ there is a transition layer when $|x| = x_\gamma$ is such that $\bar{U}(x_\gamma) = \gamma$; on the interval $|x| \leqq x_\gamma$ the solution remains close to the critical point $(\hat{u}, \hat{v})$ in Fig. 2.2. Hence the solution breaks up into two distinct parts, the outer layers, on the intervals $x_\gamma \leqq |x| \leqq L$, and the inner layer, on the $|x| \leqq x_\gamma$. It turns out that it is the inner layer that is critical in determining the spectrum of (1.4).

4.1. Asymptotics of $U(x; u)$ for large μ. Let $U(x, \mu)$ denote the maximal stationary solution $U_2(x, m, \gamma)$ of the singular limit equation (2.3), where $\mu^2 = m$ and γ is regarded as fixed. Thus $U(x, \mu)$ solves

$$\ddot{U} + r(U, \mu) = 0, \qquad U(\pm L) = 0, \tag{4.1}$$

where

$$r(U, \mu) = h(U) - \mu^2 U(U - \gamma)_+. \tag{4.2}$$

The function $r(U; \mu)$ is qualitatively a cubic which approaches $h(U)$ as $\mu \to 0$. As $\mu \to \infty$ the largest root of $h(U; \mu)$, which we denote by $\bar{u}(\mu)$, tends to γ, The phase plane in (4.1) with $U \geqq \gamma$ and large μ is indicated in Fig. 4.1. The stable and unstable manifolds of the rest point at $(\bar{u}(\mu), 0)$ have slope μ. Upon comparison with Fig. 2.1, which depicts the phase plane for (4.1) with $U \leqq \gamma$, we obtain a condition for the existence of $U(x, \mu)$ for all large μ, namely, that

$$\sigma < \gamma < b, \tag{4.3}$$

where σ is the equal area point of r,

$$\int_0^\sigma r(U; \mu)\, dU = 0;$$

hence if (4.3) holds, σ is independent of γ. A simple calculation shows that

$$\sigma = \frac{2}{3}(a+b) - \left[\frac{4}{9}(a+b)^2 - 2ab\right]^{1/2}. \tag{4.4}$$

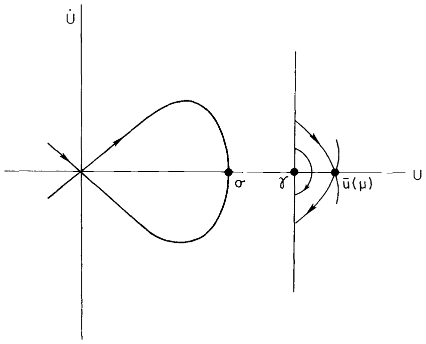


FIG. 4.1

Let $U(x, p, \mu)$ be the solution of (4.1) that satisfies

$$U(0, p, \mu) = 0, \qquad \dot{U}(0, p, \mu) = p, \tag{4.5}$$

so that $U(x, \mu) = U(x - L, p, \mu)$ for some $p = p(L)$. An important tool in our analysis of the asymptotics of $U(x, \mu)$ as $\mu \to \infty$ is the time map for solutions of (4.1):

$$T(p, \mu) = \int_0^{\hat{U}} \frac{dU}{\sqrt{2p^2 - 2R(U, \mu)}} = \int_0^{\hat{U}} \frac{dU}{\sqrt{2R(\hat{U}, \mu) - 2R(U, \mu)}},$$

where $\hat{U} = U(p, \mu)$ is the value of $U(x, p, \mu)$ when $\dot{U} = 0$, and

$$R(U, \mu) = \int_0^U r(s, \mu)\, ds.$$

Thus $T(p, \mu)$ is the time it takes $U(x, p, \mu)$ to hit $\dot{U} = 0$. The properties of the time map for this equation were discussed in [2, Appendix]. In particular, the graph of $T(p, \mu)$ is as indicated in Fig. 4.2: it has a unique local minimum L_o at some positive $p = p_o$, and it has vertical asymptotes at $p = 0$, $p = \bar{p}$. For $L > L_o$ the solution $U(x, \mu)$ is $U(x + L, p, \mu)$, where $U(x, p, \mu)$ is the solution for which $T(p, \mu) = L$ and $p \in (p_o, \bar{p})$. This uniquely determines $U(x, \mu)$ for each $L > L_o$.

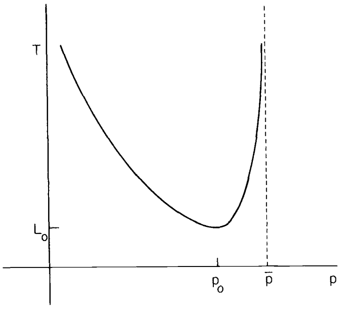


FIG. 4.2

We are interested in the asymptotics of T as $\mu \to \infty$. To this end it is useful to express T as

$$T(p, \mu) = \alpha(p) + \beta(p, \mu),$$

where $\alpha(p)$ is defined by the equation

$$U(\alpha(p), p, \mu) = \gamma,$$

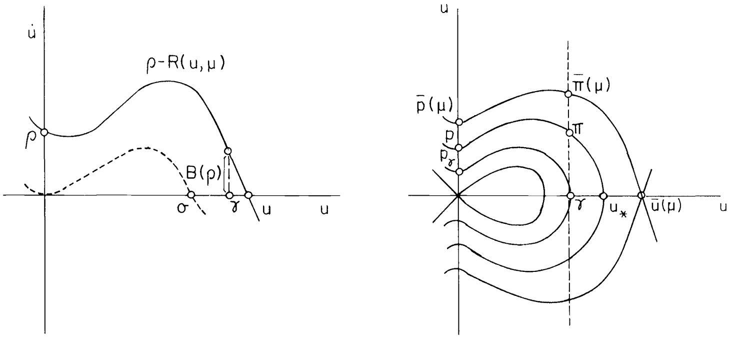
and $\beta = T - \alpha$. Note that since $r(U, \mu)$ is actually independent of μ when $U \leqq \gamma$, $\alpha(p)$ is also independent of μ. Thus $\alpha(p)$ is the time map for the portion of the solution where $0 \leqq U \leqq \gamma$, and β is the time map for the remainder $\gamma \leqq U \leqq \hat{U}(p, \mu)$.

Since β depends implicitly on p it is convenient to introduce another parameter,

$$\pi = \pi(p) = \dot{U}(\alpha(p), p, \mu).$$

Let $\rho = p^2$ and $B = \pi^2$; it can then be seen from Fig. 4.3(a) that $B = B(\rho)$ is monotone increasing in ρ, so that $\pi(p) = B(p^2)^{1/2}$ is monotone increasing in p. Let

$$\bar{p}(\mu) = \sqrt{2R(\bar{u}(\mu), \mu)}, \quad \bar{\pi}(\mu) = \pi(\bar{p}(\mu)), \quad p_\gamma = \sqrt{2R(\gamma, \mu)};$$



(a) (b)

FIG. 4.3

then $\bar{p}(\mu) > p_\gamma$ for all μ and $\bar{p}(\mu) \to p_\gamma$ as $\mu \to \infty$ since $\bar{u}(\mu)$ tends to γ for large μ. More precisely,

$$\bar{u}(\mu) - \gamma = \frac{k(\mu)}{\mu^2},$$

where $k(\mu)$ tends to some constant $k_o > 0$ as $\mu \to \infty$ (at the rate μ^{-2}). We also see from the figure that $0 \leqq \pi \leqq \bar{\pi}(\mu)$.

Now let $p = P(\pi)$ be the inverse function for $\pi = \pi(p)$. We can then express $T(p, \mu)$ as $\hat{T}(\pi, \mu)$, where

$$\hat{T}(\pi, \mu) = T(p, \mu) = T(P(\pi), \mu) = a(\pi) + b(\pi, \mu)$$

for $0 \leqq \pi \leqq \bar{\pi}(\mu)$, where the partial time maps are

$$a(\mu) = \alpha(P(\pi)), \qquad b(\pi, \mu) = \beta(P(\pi), \mu).$$

We now regard $L > L_o$ and γ as fixed and set

$$\Pi(\mu) = \dot{U}(-x_\gamma(\mu), \mu),$$

where $-x_\gamma(\mu) < 0$ is the negative value of x, where $U(x, \mu) = \gamma$, so that $x_\gamma(\mu) = \alpha(p)$ for some p. In order to characterize the behavior of $U(x, \mu)$ for large μ we shall need to estimate how close $\Pi(\mu)$ is to $\bar{\pi}(\mu)$. This will provide estimates for $|U(x, \mu) - \bar{u}(\mu)|$ for $|x| \leqq x_\gamma(\mu)$.

THEOREM 4.1. *Let p_γ be the value of p where the* maximum *value of $U(x, p, \mu)$ on $0 \leqq x \leqq T$ is γ (see Fig. 4.3b), and let $\alpha_o = \alpha(p_\gamma)$. Fix $L \geqq \alpha_o + \delta$ for $\delta > 0$, and let $q > 1$ be given. There exists $K = K(q, \delta) > 0$ such that*

$$(4.6) \qquad \begin{aligned} |U(x, \mu) - \bar{u}(\mu)| &\leqq K\mu^{-1(q+3)/2}, \\ |V(x, \mu) - \bar{v}(\mu)| &\leqq K\mu^{-(q-1)/2}, \end{aligned}$$

where $V(x, \mu) = \mu^2(U(x, \mu) - \gamma)_+$ uniformly on the interval

$$(4.7) \qquad \begin{cases} |x| \leqq x_*(\mu), \\ x_*(\mu) = x_\gamma(\mu) - K \log \mu/\mu. \end{cases}$$

Remark. Note that (4.6) implies that $V(x, \mu)$ has a transition layer at $|x| = x_\gamma(\mu)$ of width $\mathcal{O}(\log \mu/\mu)$. This also implies that $a(\Pi(\mu))$ tends to $\alpha_o < L$ as $\mu \to \infty$, so that for large μ, the amount of time $U(x, \mu)$ spends in the inner segment is approximately $2(L - \alpha_o)$.

Proof. Let $w = u - \bar{u}(\mu)$; for $u \geqq \gamma$ the equation for w is

$$\begin{aligned}\ddot{w} &= 2k\mu^2 w + 3l\mu^2 w^2 + 4\mu^{-2} w^3\\ &= \mu^2 g(w, \mu),\end{aligned}$$

where

$$\left.\begin{aligned}k &= k(\mu) \to k_o > 0\\ l &= l(\mu) \to l_0 > 0\end{aligned}\right. \quad \text{as } \mu \to \infty.$$

Let $w(x)$ be the solution of the above equation with initial conditions

$$w(0) = \gamma - \bar{u}(\mu) = -\frac{a}{\mu^2}, \qquad \dot{w}(0) = \pi_1 = \bar{\pi}(\mu) - \mu^{-q},$$

where $a = a(\mu)$ tends to $a_o > 0$ as $\mu \to \infty$. We estimate the first time $T = T(\pi_1)$ it takes for $w(x; \mu)$ to turn around, i.e., $\dot{w}(T(\pi_1), \mu) = 0$. Let $w_1 = w(T(\pi_1), \mu)$; then it follows that $w_1 < 0$ and that

$$\frac{\dot{w}^2}{2} = \mu^2 G(w, \mu) + A, \tag{4.8}$$

where

$$G(w, \mu) = \int_o^w g(s, \mu)\, ds.$$

Evaluating (4.8) at $x = 0$ gives

$$A = \frac{1}{2}(\bar{\pi}(\mu) - \mu^{-q})^2 - \mu^2 G\left(-\frac{a}{\mu^2}, \mu\right).$$

But $\bar{\pi}^2/2 = \mu^2 G(w(0), \mu)$ so that $\bar{\pi} = b\mu^{-1} + \mathcal{O}(\mu^{-2})$ where $b = a\sqrt{2k}$, and

$$A = -\mu^{-q}\bar{\pi} + \mathcal{O}(\mu^{-2q}).$$

We also have that at $x = T(\pi_1)$,

$$0 = \mu^2 G(w_1, \mu) + A, \tag{4.9}$$

so that

$$T(\pi_1) = \frac{1}{\mu\sqrt{2}} \int_{-a^2/\mu^2}^{w_1} \frac{dw}{[G(w, \mu) - G(w_1, \mu)]^{1/2}}.$$

In the following, K will denote a generic positive constant depending on q. Now from (4.9) we have that

$$\begin{aligned}G(w, \mu) &= kw_1^2 + lw_1^3 + \mu_1^{-4} w^4\\ &= A\mu^{-2}\\ &= \mu^{-q}\bar{\pi}(\mu)\mu^{-2} + \mathcal{O}(\mu^{-2q-2})\\ &= K\mu^{-(q+3)},\end{aligned}$$

so that

$$|w_1| \leqq K\mu^{-(q+3)/2}. \tag{4.10}$$

The time map $T(\pi_1)$ can be expressed as

$$T(\Pi_1)=\frac{1}{\mu\sqrt{2}}\int_{-a/\mu^2}^{w_1}\frac{dw}{[k(w^2-w_1^2)+l(w^3-w_1^3)+\mu^{-2}(w^4-w_1^4)]^{1/2}}$$

$$=\frac{1}{\mu\sqrt{2}}\int_{1}^{-a\mu^{-2}/w_1}\frac{ds}{[k(s^2-1)+lw_1(s^3-1)+\mu^{-2}w_1^2(s^4-1)]^{1/2}},$$

where $s=w/w_1$. Since $q>1$ we see from (4.10) that the upper limit s_* is positive and that

$$s_*=-a\mu^{-2}/w_1=\mathcal{O}(\mu^{(q-1)/3})\to+\infty$$

as $\mu\to\infty$. Thus $T(\pi_1)$ can be expressed as

$$T(\pi_1)=\frac{1}{\mu\sqrt{2}}\left(\int_1^2+\int_2^{s_*}\right).$$

The first integral is clearly bounded independently of μ and therefore contributes a term of order μ^{-1} to $T(q)$. Also, for $2\leqq s\leqq s_*$ we have that

$$|w_1 s|,\ w_1^2 s^2\leqq K\mu^{-2},$$

so that for such s,

$$[k(s^2-1)+lw_1(s^3-1)+w_1^2\mu^{-2}(s^4-1)^{-1/2}\leqq Ks^{-1}.$$

We therefore have that

$$T(\Pi_1)\leqq\frac{1}{\mu}\left[K+K\log\left|\frac{a}{\mu^2 w_1}\right|\right],$$

which together with (4.10) yields

$$T(\pi_1)\leqq K\frac{\log\mu}{\mu} \tag{4.11}$$

for some $K=K(q)$.

We have now proved that

$$\bar{\pi}(\mu)\geqq\Pi(\mu)\geqq\pi_1=\bar{\pi}(\mu)-\mu^{-q},$$

where $\Pi(\mu)=\dot{U}(-x_\gamma(\mu),\mu)$. For $\pi\in[\pi_1,\bar{\pi}(\mu)]$ let $w(x,\pi)$ be the solution of the differential equation satisfying

$$w(0,\pi)=-\frac{a}{\mu^2},\qquad \dot{w}(0,\pi)=\pi,$$

and let

$$T(\pi)=\frac{1}{\mu\sqrt{2}}\int_{-a/\mu^2}^{w_1}[G(w,\mu)-G(w_*(\pi),\mu]^{-1/2}\,dw,$$

where $w_*(\pi)$ is the solution of

$$0=G(w,\mu)+\left[\frac{\pi^2}{2}-G\left(-\frac{a}{\mu^2},\mu\right)\right].$$

$T(\pi)$ is therefore the time map for the portion of the solution $w(x,\pi)$ with values starting at $-a/\mu^2$ and ending at w_1. Since $\pi>\pi_1$ it follows that $w_*(\pi)>w_1$; thus the integrand in $T(\pi)$ is smooth on $[-a/\mu^2, w_1]$. We can therefore compute $T'(\pi)$ by differentiating under the integral sign to obtain

$$T'(\pi)=\frac{1}{\mu 2^{3/2}}\int_{-a/\mu^2}^{w_1} g(w_*(\pi))w'_*(\pi)[G(w,\mu)-G(w_*(\pi)),\mu]^{-3/2}\,dw.$$

Now $g(w,\mu)=kw+\mathcal{O}(w^2)$ so that $g(w_*(\pi))>0$. Also, it is easily seen from Fig. 4.3 that $w'_*(\pi)\geqq 0$, since $w_*(\pi)=u_*-\gamma$, where u_* is as in the figure. It follows that $T'(\pi)\leqq 0$, so that $T(\pi)\leqq T(\pi_1)$ for all $\pi\in[\pi_1,\bar{\pi}]$; by (4.11) we now have that

$$T(\Pi(\mu))\leqq K\frac{\log\mu}{\mu}. \tag{4.12}$$

Now let

$$\begin{aligned} W(x,\mu)&=w(x,\Pi(\mu))\\ &=U(x-x_\gamma(\mu),\mu)-\bar{u}(\gamma)\end{aligned}$$

be the solution of the boundary value problem with $L>\alpha_0+\delta$ translated so that $W=-a/\mu^2$ at $x=0$, and let $P(\mu)=P(\Pi(\mu))$ so that

$$L=\alpha(P(\mu))+\beta(P(\mu),\mu).$$

Since $x_\gamma(\mu)=\alpha(P(\mu))$ it follows that $W(x,\mu)$ is monotone increasing on the interval $0\leqq x\leqq\beta(P(\mu),\mu)$. Since $\Pi(\mu)\in[\pi_1,\bar{\pi}(\mu)]$ it follows that $\alpha(P(\mu))$ tends to α_o as $\mu\to\infty$. For $L>\alpha_o+\delta$ it therefore follows that $\beta(P(\mu),\mu)$ is bounded away from zero for large μ. Finally, let $\hat{x}=T(\Pi(\mu))$; it follows from (4.10) and (4.12) that $\hat{x}\leqq x_*$ in (4.7), and that

$$0\geqq W(\hat{x},\mu)=w_1\geqq -K^{-(q+3)/2}. \tag{4.13}$$

Furthermore, $W(x,\mu)$ is monotone increasing on the interval $[\hat{x},\beta(P(\mu),\mu)]$; thus (4.13) holds for all x in this interval. Since $U(x,\mu)=U(-x,\mu)$, this is equivalent to (4.6), (4.7).

4.2. The linearized singular limit equation. We can now analyze the reduced linearized equation

$$\ddot{P}=G(x,\lambda,\mu)P, \tag{4.14}$$

where

$$G(x,\lambda,\mu)=\lambda-\bar{f}_u-\frac{\bar{f}_v\bar{g}_u}{\lambda-\bar{g}_v},$$

and the partials are evaluated at the singular limit $\bar{U}(x,\mu)$, $\bar{V}(x,\mu)$ of Theorem 2.1. It will be convenient to express the variables in (4.14) in real and imaginary parts; thus let

$$\begin{cases} \lambda=\alpha+i\beta,\\ G(x,\lambda,\mu)=G_R(x,\lambda,\mu)+iG_I(x,\lambda,\mu),\\ G_R=\alpha-h'(U)+V+\dfrac{\mu^2UV(\alpha+V)}{(\alpha+V)^2+\beta^2},\\ G_I=\beta\left[1-\dfrac{\mu^2UV}{(\alpha+V)^2+\beta^2}\right],\end{cases} \tag{4.15}$$

where $(U,V)=(U(x,\mu),V(x,\mu))$. The values of G when the singular limit is near the rest point $(\bar{u}(\mu),\bar{v}(\mu))$ are of particular importance; thus let

$$\bar{G}(\lambda,\mu)=\bar{G}_R(\lambda,\mu)+i\bar{G}_I(\lambda,\mu) \tag{4.16}$$

denote the expressions in (4.15) wherein (U, V) are replaced by $(\bar{u}(\mu), \bar{v}(\mu))$. For simplicity we shall denote the latter by $(\bar{u}, \bar{v})$.

The existence of pure imaginary eigenvalues $\lambda = i\beta$ is of particular importance. Let

$$\bar{g}_R(\beta, \mu) = \bar{G}_R(i\beta, \mu), \qquad \bar{g}_I(\beta, \mu) = \bar{G}_I(i\beta, \mu); \tag{4.17}$$

the graphs of $\bar{g}_R$ and $\bar{g}_I$ as functions of β have a crucial role in determining when such eigenvalues exist. Since they must occur in conjugate pairs it suffices to consider $\beta \geqq 0$. Let

$$\beta_I(\mu) = [\mu^2\bar{u}\bar{v} - \bar{v}^2]^{1/2}; \tag{4.18}$$

then $\bar{g}_I(\beta, \mu) = 0$ for $\beta \geqq 0$ if and only if $\beta = 0$ or $\beta_I(\mu)$. Similarly, let $\Gamma_R = \bar{u}\bar{v}^2/(h'(\bar{u}) - \bar{v})$, and define

$$\beta_R(\mu) = [\Gamma_R\mu^2 - \bar{v}^2]^{1/2}, \tag{4.19}$$

so that $\bar{g}_R(\beta, \mu) = 0$ for $\beta \geqq 0$ only at $\beta = \beta_R(\mu)$. If β_R is infinite or imaginary, then g_R has no real roots. Typical graphs of $\bar{g}_R$ and $\bar{g}_I$ are depicted in Fig. 4.4.

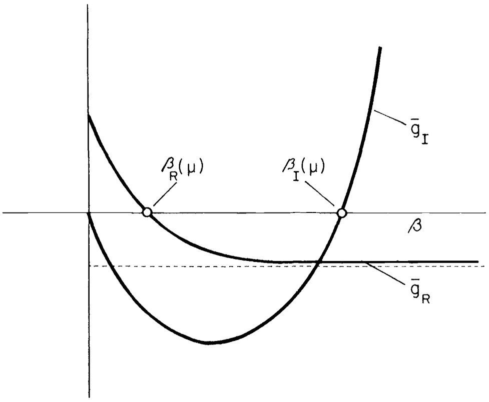


FIG. 4.4

Observe that $\beta_R(\mu)$ and $\beta_I(\mu)$ are both $\mathcal{O}(\mu)$ for large μ and that $\bar{g}_R(\beta, \mu)$ has a horizontal asymptote at the value $-h'(\bar{u}) + \bar{v}$.

It turns out that the potential for an instability depends on the relative position of $\beta_I(\mu)$ and $\beta_R(\mu)$. We therefore distinguish two cases:

Case S (stable case). $\beta_I(\mu) < \beta_R(\mu)$, or $\beta_R(\mu)$ is imaginary.

Case U (unstable case). $\beta_R(\mu) < \beta_I(\mu)$.

It can be seen from Fig. 4.4 that $\bar{g}_R(\beta, \mu) \lesseqgtr 0$ precisely when $\beta \gtreqless \beta_R(\mu)$. Thus if $g_R(\beta_I(\mu), \mu)$ is positive (negative) then we are in Case S (Case U). The transition from Case S to Case U therefore occurs when $\bar{g}_R(\beta_I(\mu), \mu) = 0$; this condition leads to the equation $-h'(\bar{u}) + 2\bar{v} = 0$. Recall that as $\mu \to \infty$, $(\bar{u}, \bar{v})$ tends to (γ, v_γ), where $v_\gamma = (b - \gamma)(\gamma - a)$. Since we are interested in the asymptotics as $\mu \to \infty$, the asymptotic critical condition is that $-h'(\gamma) + 2v_\gamma = 0$. This equation is easily solved for γ (see (1.2)); the solution is $\gamma = \gamma_0 = \sqrt{ab}$. We have now proved the following lemma.

LEMMA 4.2. *Suppose that* $\gamma_o = \sqrt{ab}$. *Then if* $\gamma > \gamma_o$ (*respectively,* $\gamma < \gamma_0$), *Case S* (*respectively, Case U*) *holds for all sufficiently large* μ.

Remark. The condition $\gamma < \sigma$ (see (4.3)) imposed previously on γ was needed to ensure the existence of the singular solution $U(x, \mu)$ of (4.1). If $\gamma_o < \sigma$ then Case S

necessarily occurs for all relevant γ. On the other hand, if $\gamma_o > \sigma$ then Case U occurs whenever $\sigma < \gamma < \gamma_o$. If we regard a as fixed and $b \gg a$ as a large parameter, then an inspection of the asymptotic behavior of the expression (4.4) for σ as $b \to \infty$ shows that σ remains uniformly bounded for large b. Since $\gamma_o = \mathcal{O}(b^{1/2})$ we see that $\sigma < \gamma_o$ whenever $a \ll b$.

Let $P(x, \lambda)$ denote the solution of (4.14) satisfying

$$P(-L, \lambda) = 0, \qquad \dot{P}(-L, \lambda) = 1. \tag{4.20}$$

Then λ is the spectrum σ_R of the slow reduced equation if and only if $P(L, \lambda) = 0$. The following lemma gives another characterization of σ_R.

LEMMA 4.3. *Let* $\bar{x} \in (0, L]$ *and let* $P(x, \lambda)$ *be the solution of* (4.14), (4.20). *Then* $\lambda \in \sigma_R$ *if and only if*

$$P(\bar{x}, \lambda) = -P(-\bar{x}, \lambda), \qquad \dot{P}(\bar{x}, \lambda) = P(-x, \lambda) \tag*{(4.21)$_-$}$$

or

$$P(\bar{x}, \lambda) = P(-\bar{x}, \lambda), \qquad \dot{P}(\bar{x}, \lambda) = -P(-\bar{x}, \lambda). \tag*{(4.21)$_+$}$$

Proof. Suppose first that $\lambda \in \sigma_R$ so that $P(x, \lambda)$ is an eigenfunction. We will first show that $P(x, \lambda) = \pm P(-x, \lambda)$. Since $U(x, \lambda) = U(-x, \lambda)$, $G(x, \lambda, \mu)$ inherits a similar symmetry, so that $P(\pm x, \lambda)$ are both eigenfunctions. It follows from the fact that $\dim P = 1$ that all eigenvalues are simple, whence $P(-x, \lambda) = cP(x, \lambda)$ for some constant c. Since $P(x, \lambda) \not\equiv 0$ it follows that

$$|P(0, \lambda)| + |\dot{P}(0, \lambda)| \neq 0.$$

If $P(0, \lambda) \neq 0$ then $c = 1$, while if $P(0, \lambda) = 0$, $\dot{P}(0, \lambda) \neq 0$, so that in this case $c = -1$. Thus precisely one of the relations $(4.21)_\pm$ hold for arbitrary $\bar{x} \in (0, L]$.

Now suppose that $(4.21)_-$ holds. Let $\tilde{P}(x, \lambda) = -P(-x, \lambda)$, and define

$$P_*(x, \lambda) = \begin{cases} P(x, \lambda) & \text{for } -L \leqq x \leqq \bar{x}, \\ \tilde{P}(x, \lambda) & \text{for } \bar{x} \leqq x \leqq L \end{cases}$$

by $(4.21)_-$ it follows that $P_*(x, \lambda)$ is a C^2 solution of (4.14) which satisfies the boundary conditions both at $x = -L$ and $x = L$, so that $\lambda \in \sigma_R$. The argument is similar in the event that $(4.21)_+$ holds.

In order to study the eigenvalue problem for (4.14) it will be convenient to projectivize (4.14) to get an equation on $\mathbb{C}P^1 = S^2$. To this end set $z = \dot{P}/P$ and $w = P/\dot{P}$ to obtain local coordinates covering $\mathbb{C}P^1$. The equations for z and w are then

$$\dot{z} = G(x, \lambda, \mu) - z^2 \tag{4.22}$$

and

$$\dot{w} = 1 - G(x, \lambda, \mu) w^2, \tag{4.23}$$

respectively. Let

$$z = \sigma + i\tau, \qquad w = s + it;$$

then (4.22) and (4.23) are equivalent to

$$\begin{aligned} \dot{\sigma} &= G_R(x, \lambda, \mu) - \sigma^2 + \tau^2, \\ \dot{\tau} &= G_I(x, \lambda, \mu) - 2\sigma\tau, \end{aligned} \tag{4.24}$$

and

$$\begin{aligned} \dot{s} &= 1 - G_R(x, \lambda, \mu)(s^2 - t^2) + 2G_I(x, \lambda, \mu)st, \\ \dot{t} &= -G_I(x, \lambda, \mu)(s^2 - t^2) - 2G_R(x, \lambda, \mu)st. \end{aligned} \tag{4.25}$$

Condition (4.20) is equivalent to $w=0$ or $z=\infty$ at $x=-L$. *In the following, $z(x, \lambda)$ will denote the solution of* (4.22) *satisfying $z=\infty$ at $x=-L$.* The dependence of z on the parameters μ and L will be suppressed for now. In the bifurcation analysis presented later, the dependence of z on L will be denoted explicitly by writing $z=z(x, \lambda, L)$. The two conditions $(4.21)_{\pm}$ can be incorporated into a single condition,

$$z(\bar{x}, \lambda)=-z(-\bar{x}, \lambda) \tag{4.26a}$$

or equivalently,

$$w(\bar{x}, \lambda)=-w(-\bar{x}, \lambda). \tag{4.26b}$$

By allowing $\bar{x}$ to approach zero in (4.26a) and (4.26b) we obtain the equivalent condition

$$\lambda \in \sigma_R \quad \text{if and only if } z(0, \lambda)=0 \quad \text{or} \quad z(0, \lambda)=\infty. \tag{4.27}$$

There are two distinct ideas that will be used in determining when λ is an eigenvalue. In order to show that λ *cannot* be an eigenvalue, we shall locate certain positively invariant sets for (4.24) that prevent the solution $z(x, \lambda, \mu)$ from satisfying (4.26). In the unstable case, we shall require a topological argument in the λ, z plane to demonstrate the existence of unstable eigenvalues. We first consider the stable case.

THEOREM 4.4. *Suppose that $\sigma<\gamma_o<\gamma$ and that $\mu=\sqrt{m}$ is sufficiently large. Then the solution $z(x, \lambda)$ of* (4.22) *satisfying $z(-L, \lambda)=\infty$ does not satisfy* (4.27) *for any $\lambda \in \mathbb{C}$ with* $\operatorname{Re} \lambda \geqq 0$.

Proof. Since eigenvalues occur in conjugate pairs it suffices to consider only $\beta=\operatorname{Im} \lambda \geqq 0$. We first need to determine the behavior of $z(x, \lambda)$ for x near $-L$. Since $w(-L, \lambda)=0$ it follows from (4.25) that

$$s(x)=(x+L)+\mathcal{O}(x+L)^2.$$

Now for $x \leqq -x_\gamma(\mu)$, $G_I(x, \lambda, \mu)=\beta \geqq 0$. Suppose first that $\beta>0$; the t equation in (4.25) together with the above yields

$$t(x, \lambda)=-\beta(x+L)^3+\mathcal{O}(x+L)^5,$$

so that $t(x)<0$ for x near $-L$ and $x+L>0$. Since $(\sigma+i\tau)=(s-it)(s^2+t^2)^{-1}$ it follows that $\sigma(x, \lambda)>0$ and $\tau(x, \lambda)>0$ for such x. Also, $\sigma(x, \lambda)=\mathcal{O}(x+L)^{-1}$ while $\tau(x, \lambda)=\mathcal{O}(x+L)$, so that $z=\sigma+i\tau$ is asymptotic to the positive σ-axis from above as x approaches $-L$ from above. If $\beta=0$ then $t\equiv\tau\equiv 0$ so that $z(x, \lambda)=\sigma(x, \lambda, \mu)=(x+L)^{-1}+\mathcal{O}(1)$ as $x \to -L^+$.

Let $\alpha_o \geqq 0$ satisfy

$$\alpha_0 \geqq \max\{h'(u)-v\colon 0 \leqq u \leqq b, 0 \leqq v \leqq (a+b)^2/4\};$$

we will consider the following four regions in the nonnegative λ plane separately:

$$\mathrm{I}=\{\lambda\colon \alpha>\alpha_0, \beta \geqq 0\},$$

$$\mathrm{II}=\{\lambda\colon 0 \leqq \alpha \leqq \alpha_0, \beta \geqq K\mu\},$$

$$\mathrm{III}=\{\lambda\colon 0 \leqq \alpha \leqq \alpha_0, \beta_o \leqq \beta \leqq K\mu\},$$

$$\mathrm{IV}=\{\lambda\colon 0 \leqq \alpha \leqq \alpha_o, 0 \leqq \beta \leqq \beta_o\},$$

where β_o and K are positive and independent of μ.

Suppose first that $\lambda \in I$ and let $\Sigma_+=\{z\colon \sigma>0\}$. We will show that $z(x, \lambda)$ remains finite and inside Σ_+ for all $x \in (-L, L]$. This has been demonstrated above for $-L<x$ and x near $-L$. Let $\bar{x} \in (-L, L]$ be the smallest value of x for which the assertion fails to be true. If $z(\bar{x}, \lambda)=\infty$ then $w(\bar{x}, \lambda)=0$. It then follows from (4.25) that

$$s(x, \lambda)=(x-\bar{x})+\mathcal{O}(x-\bar{x})^2,$$

so that $s(x, \lambda) < 0$ for $x < \bar{x}$; since s and σ have the same sign, this contradicts the minimality of $\bar{x}$. Suppose then that $\sigma(\bar{x}, \lambda) = 0$. It follows from (4.24) that $\dot{\sigma} = G_R(\bar{x}, \lambda, \mu) + \tau^2$; our condition that $\alpha > \alpha_0$ implies that G_R, and hence $\dot{\sigma}$ is positive at $x = \bar{x}$, again contradicting the minimality of $\bar{x}$. Hence $z(L, \lambda)$ is finite and nonzero, so that λ is not an eigenvalue.

Now suppose that $\lambda \in \mathrm{II}$. From (4.15) it is evident that if $\beta \geqq K\mu$ and if K is large then $G_{\mathrm{I}}(x, \lambda, \mu) > 0$ for all x and all sufficiently large μ. Let $T_+ = \{z\colon \tau > 0\}$; we claim that $z(x, \lambda)$ remains finite and inside T_+ for all $x \in (-L, L]$. By the expansion obtained earlier this clearly holds for $x > -L$ and x near $-L$. Let $\bar{x}$ be the smallest $x > -L$ such that either condition fails. If $z(\bar{x}, \lambda) = \infty$ then $w(\bar{x}, \lambda) = 0$. Since $G_{\mathrm{I}} > 0$ for all x and $\lambda \in \mathrm{II}$ it follows from (4.25) that $s(x, \lambda) = (x - \bar{x}) + \mathcal{O}(x - \bar{x})^2$ and that $t(x, \lambda) = -G_{\mathrm{I}}(\bar{x}, \lambda, \mu)(x - \bar{x})^3 + \mathcal{O}(x - \bar{x})^4$ so that $\tau = -t(s^2 + t^2)^{-1} < 0$ for $x < \bar{x}$, contradicting the minimality of $\bar{x}$. Finally, if $\tau(\bar{x}, \lambda) = 0$ then $\dot{\tau} = G_{\mathrm{I}}$ is positive at $\bar{x}$, again contradicting the minimality of $\bar{x}$. Thus $\lambda \in \mathrm{II}$ is not an eigenvalue.

Before proceeding to III, we shall need to give a finer estimate for K, namely, that

$$K\mu < \beta_R(\mu) \tag{4.28}$$

for large μ. Recall that K is chosen so that $G_I(x, \lambda, \mu) > 0$ for all x and $\beta \geqslant K\mu$. Since $(\bar{u}, \bar{v})$ are the maxima of (U, V) for $|x| \leqq L$, we see that

$$G_{\mathrm{I}}(x, \lambda, \mu) = \beta\left[1 - \frac{\mu^2 UV}{(\alpha + V)^2 + \beta^2}\right]$$
$$\geqq \beta\left[1 - \frac{\mu^2 \bar{u}\bar{v}}{V^2 + \beta^2}\right].$$

Hence the condition on G_{I} will hold if

$$\beta \geqslant \mu\sqrt{\bar{u}\bar{v} - V^2/\mu^2},$$

so that K can be any constant larger than $(\bar{u}\bar{v})^{1/2}$. Since we are in Case S we see from (4.18), (4.19) that

$$\beta_{\mathrm{I}}(\mu) = \sqrt{\mu^2\bar{u}\bar{v} - \bar{v}^2} < \sqrt{\Gamma_R\mu^2 - \bar{v}^2} = \beta_R(\mu)$$

for large μ, so that $\sqrt{\bar{u}\bar{v}} < \Gamma_R^{1/2}$. Thus if K is chosen so that

$$\sqrt{\bar{u}\bar{v}} < K < \Gamma_R^{1/2},$$

G_{I} will be positive for all x and $\beta \geqq K\mu$, and also (4.28) will be satisfied.

Now suppose that $\lambda \in \mathrm{III}$, so that $0 \leqq \alpha \leqq \alpha_0$ and $\beta_0 \leqq \beta \leqq K\mu$. We first consider the behavior of $z(x, \lambda)$ on the interval $-L \leqq x \leqq -x_\gamma(\mu)$. Since $V(x, \mu) \equiv 0$ on this interval we see from (4.15) that $G_R = \alpha - h'(U) + V$ is uniformly bounded and in fact, independent of β, while $G_{\mathrm{I}} \equiv \beta$. Thus (4.24) can be expressed as

$$\dot{\sigma} = G_R - \sigma^2 + \tau^2, \qquad \dot{\tau} = \beta - 2\sigma\tau.$$

If we regard β as a large parameter and G_R as $\mathcal{O}(1)$ then the phase plane for this system is as depicted in Fig. 4.5 for each $x \in [-L, -x_\gamma(\mu)]$. In particular, the vector field $(\dot{\sigma}, \dot{\tau})$ is nearly vertical in a uniform neighborhood of the origin for large β. Consider the region Σ depicted in the figure. The diagonal edge of Σ near the origin has slope -1; it is then easily seen that $(\dot{\sigma}, \dot{\tau})$ points strictly into Σ along $\partial\Sigma$ so that Σ is positively invariant for (4.24) relative to the interval $[-L, -x_\gamma(\mu)]$ provided that $\beta \geqq \beta_o$ and that β_o is sufficiently large. Note that this condition is independent of μ. Since the right edge of Σ can be extended arbitrarily close to $\sigma = +\infty$, it follows from our characterization of $z(x, \lambda)$ for x near $-L$ that $z(x, \lambda)$ lies in Σ for some $x \in (-L, -x_\gamma(\mu)]$.

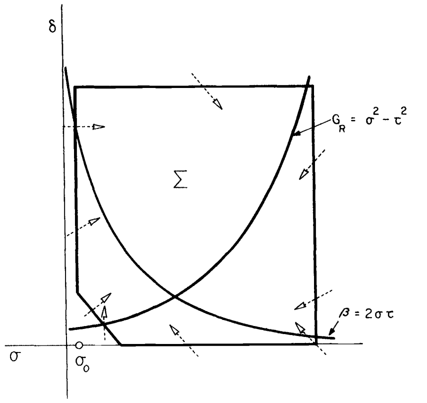


FIG. 4.5

At $x = -x_\gamma(\mu)$ we see that $\sigma(x, \lambda) \geqq \sigma_o > 0$ where $\sigma_o > 0$ is the lower bound for σ in Σ, which depends only on β_o. From Fig. 4.4, we see that $G_R(x, \lambda, \mu)$ is uniformly bounded from below for large μ, say

$$\min_{|x| \leqq L} G_R(x, \lambda, \mu) \geqq -A^2$$

for some $A > 0$ independent of λ and μ. Thus

$$\dot{\sigma} = G_R - \sigma^2 + \tau^2 \geqq -(A^2 + \sigma^2).$$

Thus for $x \geqq x_\gamma$ we obtain the estimate

$$\sigma(x, \lambda) \geqq A \tan\left[-A(x + x_\gamma) + \tan^{-1}(\sigma_o / A)\right]. \tag{4.29}$$

From the above, we see that there exists $x_1 \in (-x_\gamma, 0)$ and $\sigma_1 > 0$, both independent of μ, such that $\sigma(x_1, \lambda) \geqq \sigma_1$ for all $\lambda \in \mathrm{III}$ and all sufficiently large μ. We now use (4.6) and (4.7) of Theorem 4.1 to approximate $G(x_1, \lambda, \mu)$ by $\bar{G}(\lambda, \mu)$. For $\lambda \in \mathrm{III}$, the coefficients of U and V in the expressions (4.15) for G_R and G_I have factors of order at most μ^2. It therefore follows from Theorem 4.1 that

$$|G(x, \lambda, \mu) - \bar{G}(\lambda, \mu)| \leqq K\mu^{-(q-7)/2}$$

for all x satisfying (4.7). Fix $q > 7$ so that x_1 satisfies (4.7) for all large μ. We therefore have that $G(x, \lambda, \mu)$ is uniformly approximated by $\bar{G}(\lambda, \mu)$ on the interval $|x| \leqq |x_1|$. Since $\bar{G}(\lambda, \mu)$ is uniformly positive for $\lambda \in \mathrm{III}$, it follows that $G(x, \lambda, \mu) > 0$ for $|x| \leqq |x_1|$, for all $\lambda \in \mathrm{III}$ and all large μ. We can now use the same argument that was used for $\lambda \in \mathrm{I}$ to show that $z(x, \lambda)$ remains finite and inside the right half plane Σ_+ on the interval $|x| \leqq |x_1|$. This violates the condition (4.26) with $\bar{x} = |x_1|$, so that by Lemma 4.3, $\lambda \in \mathrm{III}$ is not an eigenvalue.

Finally, suppose that $\lambda \in \mathrm{IV}$. We can no longer construct the positively invariant region Σ as in case III. The first step will be to show that $\sigma(x, \lambda)$ remains bounded from below uniformly for $x \in [-L, -x_1]$ for some $x_1 \in (0, x_\gamma)$, $\lambda \in \mathrm{IV}$, and for all large μ. The second step will be to show that $z(x, \lambda)$ is rapidly attracted to a certain rest point of the interior layer equation

$$\dot{z} = \bar{G}(\lambda, \mu) - z^2$$

on the interval $|x| \leqq x_1$. This rest point has positive real part and is of order $\sqrt{\mu}$; thus $z(x, \lambda)$ violates the eigenvalue condition (4.26) with $\bar{x} = x_1$ for large μ.

We first consider the case $0<\beta \leqq \beta_o$. We will show that (σ, τ) remains finite and $\tau(x, \lambda)$ remains positive for all $x \in (-L, -x_\gamma(\mu)]$ and $\lambda \in \mathrm{IV}$. Since $G_I \equiv \beta$ on this interval it follows that τ remains positive for as long as (σ, τ) remains finite. However, by expanding (s, t) about a point $\bar{x}$ where $(s, t)=(0,0)$ it can be seen that $\tau(x, \lambda)$ must change sign from negative to positive as x crosses $\bar{x}$. Hence (σ, τ) remains finite with $\tau>0$ on this interval. Furthermore, a similar argument implies that (s, t) remains finite on this interval, since otherwise (σ, τ) would vanish at some $\bar{x}$. Since $\dot{\tau}=\beta$ at $\bar{x}$ it would again follow that τ changes from negative to positive as x crosses $\bar{x}$, providing a contradiction. It follows that (s, t) is uniformly bounded and continuous and uniformly bounded away from $(0,0)$ on compact subintervals of $(-L, -x_\gamma(\mu)]$. Since $\sigma(x, \lambda) \to +\infty$ as $x \to -L^+$ it follows that

$$\underline{\sigma} = \operatorname*{Min}_{\substack{\lambda \in \mathrm{IV} \\ -L \leqq x \leqq -x_\gamma(\mu)}} \sigma(x, \lambda)$$

is finite. Furthermore, since U, V are actually independent of μ for $x \leqq -x_\gamma(\mu)$ and $V(x, \mu) \equiv 0$ there, it can be seen from (4.15) that $G(x, \lambda, \mu)$—and hence $\sigma(x, \lambda)$—is independent of μ on this interval, so that $\underline{\sigma}$ is independent of μ. Finally, since $\dot{\tau}=\beta-2\sigma\tau$ it follows that

$$0 \leqq \tau \leqq \bar{\tau}$$

on $[-L, -x_\gamma(\mu)]$ where $\bar{\tau}$ depends only on $\underline{\sigma}$.

We next note that the lower bound $\underline{\sigma}$ for $\sigma(x, \lambda)$ can be extended to the interval $(-L, -x_1]$ for some $x_1<x_\gamma(\mu)$. In particular, the term in G_R which becomes significant for $x>-x_\gamma(\mu)$ is always positive (see (4.15)). Hence precisely the same estimate (4.29) as was used in case III to bound σ from below for $x \geqq -x_\gamma(\mu)$ applies here as well.

Now fix $x_1<x_\gamma(\mu)$ as above; for sufficiently large μ it follows that

$$x_1<x_\gamma(\mu)-\frac{K \log \mu}{\mu},$$

where K is as in Theorem 4.1. Thus for $|x| \leqq x_1$, $G(x, \lambda, \mu)$ is within $\mu^{-(q-7)/2}$ of $\bar{G}(\lambda, \mu)$. Since $0 \leqq \alpha \leqq \alpha_o$ and $0 \leqq \beta \leqq \beta_o$ where α_o, β_o are independent of μ, it follows that

$$\bar{G}_R(\lambda, \mu)>L^2\mu^2$$

for some $L>0$ and all large μ (see (4.15)). By the previous estimate it then follows that

$$G_R(x, \lambda, \mu)>L^2\mu^2$$

for $|x| \leqq x_1$. Thus $\sigma(x, \lambda)$ satisfies the differential inequality

$$\dot{\sigma} \geqq L^2\mu^2-\sigma^2$$

for $|x| \leqq x_1$. Furthermore, by our estimate of σ from below on $(-L, -x_1]$ it follows that

$$\sigma(x_1, \lambda)>\underline{\sigma}.$$

If $|\sigma| \leqq \mu L$ we can integrate the differential inequality to obtain

$$\sigma(x, \lambda) \geqq \mu L \frac{\Gamma-e^{-2\mu L(x+x_1)}}{\Gamma+e^{-2\mu L(x+x_1)}},$$

where

$$\Gamma=\frac{\mu L+\underline{\sigma}}{\mu L-\underline{\sigma}}.$$

Thus for $x>-x_1$, $\sigma(x,\lambda)$ rapidly becomes positive and of order μL. Since $\sigma(-x_1,\lambda)\geqq\underline{\sigma}$ it follows that $z(x,\lambda)$ cannot satisfy the eigenvalue condition (4.26) with $\bar{x}=x_1$. Thus $\lambda\in\mathrm{IV}$ is not an eigenvalue, completing the proof.

We next consider Case U, wherein the graphs of $\bar{g}_I$ and $\bar{g}_R$ have the aspect depicted in Fig. 4.6a.

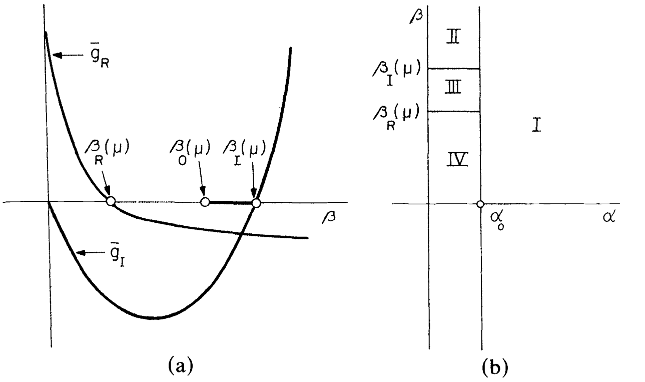


FIG. 4.6

The next theorem shows that for large L, σ_R contains pairs of complex conjugate eigenvalues with positive real part. The proof is somewhat lengthy; we therefore give a brief sketch of the main ideas first. The positive λ-quadrant is again divided into four distinct regions, as depicted in Fig. 4.6b. Here, region IV is defined by taking

$$\beta_0(\mu)=\beta_I(\mu)-p>\beta_R(\mu), \tag{4.30}$$

where $p>0$ is independent of μ, to be determined below. The first part of the proof, Lemma 4.6, shows that σ_R is disjoint from regions I, II, and III. The techniques here for the most part are the same as in Theorem 4.4.

The next step, Lemma 4.7, treats an approximate eigenvalue problem for pure imaginary eigenvalues on an interior interval $|x|\leqq l$ obtained by replacing $G(x,\lambda,\mu)$ by the constant potential $\bar{G}(\lambda,\mu)$. Thus we consider the problem

$$\dot{z}=\bar{G}(i\beta,\mu)-z^2,\qquad z(-l)=(1+i)\beta^{1/2}/2^{1/2}. \tag{4.31a}$$

The approximate eigenvalue problem consists of finding a solution of (4.31) that satisfies

$$z(0)=0\quad\text{or}\quad\infty. \tag{4.31b}$$

Problem (4.31) will have solutions at certain distinguished $\bar{l}_n^j$ and $\bar{\beta}_n^j$ with $j=0$ or ∞ and $n\geqq0$, satisfying

$$\begin{aligned}&\bar{\beta}_n^o<\bar{\beta}_n^\infty<\bar{\beta}_{n+1}^o, &&\bar{\beta}_n^j\in[\beta_o,\beta_I],\\ &\bar{l}_n^o<\bar{l}_n^\infty<\bar{l}_{n+1}^o, &&\bar{l}_n^j\to\infty\quad\text{as } n\to\infty.\end{aligned} \tag{4.32}$$

This determines branches $\bar{\lambda}(l)$ of approximate eigenvalues satisfying $\bar{\lambda}(\bar{l}_n^j)=i\beta_n^j$. It can be shown by direct computation that $\operatorname{Re}\bar{\lambda}'(\bar{l}_n^j)>0$. The transverse crossing of each eigenvalue shows that approximate problem (4.31) has precisely $2(2n+1)$ unstable eigenvalues for $\bar{l}_n^o<l<\bar{l}_n^\infty$ and $2(2n+2)$ unstable eigenvalues for $\bar{l}_n^\infty<l<\bar{l}_{n+1}^o$.

The last step consists of using (4.31) to approximate σ_R. To this end, we note that the interior interval $|x|\leqq x_\gamma(\mu)$ of the nonlinear problem is implicitly a function of L, which we denote by writing $x_\gamma = x_\gamma(\mu, L)$. This suggests the introduction of a parameter $l = x_\gamma(\mu, L)$ for the size of the interior interval. In order to obtain the uniform approximation of G by $\bar{G}$, we need to introduce

$$x_*(\mu, L) = x_\gamma(\mu, L) - K \log \mu/\mu,$$

where K is as in Theorem 4.1. It follows from the remarks preceding Theorem 4.1 that x_γ, and hence x_*, is a monotone function of L. We can therefore invert this relation by writing

$$L = X_*(\mu, l).$$

The proof is completed by showing that (4.31) provides a good approximation to (4.22) for λ along the portion of the imaginary axis satisfying (4.32). This provides a sequence of exact, purely imaginary eigenvalues $i\beta_n^j$ at intervals $L_n^j = X_*(\mu, l_n^j)$, where β_n^j and l_n^j are approximated by $\bar{\beta}_n^j$ and $\bar{l}_n^j$. Finally, we show that the eigenvalues of the exact problem cross the imaginary axis transversally too, with Re $\lambda'(L_n^j) > 0$. This provides an exact eigenvalue count in Re $\lambda \geqq 0$. We summarize this in the following theorem.

THEOREM 4.5. *Suppose that $\sigma < \gamma < \gamma_o$, so that Case U holds. There exist critical interval sizes L_n^j, $j = 0$ or ∞, $n \geqq 0$ with $L_n^{\circ} < L_n^{\infty} < L_{n+1}^{\circ}$, and there exists μ_o and $N(\mu_o)$ such that if $\mu > \mu_o$ then* (i) *$\sigma_R \cap \{$Re $\lambda \geqq 0\}$ is empty for $L < L_o^{\circ}$, and* (ii) *for $L_n^{\circ} < L < L_n^{\infty}$* (respectively, $L_n^{\infty} < L < L_{n+1}^{\circ}$). *$\sigma_R$ contains exactly* $(2n+1)$ (*respectively,* $(2n+2)$) *pairs of complex conjugate eigenvalues, for $0 \leqq n \leqq N-1$. If $\lambda(L)$ is the branch of eigenvalues with* Re $\lambda(L_n^j) = 0$, *then* Re $\lambda'(L_n^j) > 0$.

Proof. The proof will proceed in a sequence of lemmas, as outlined above.

LEMMA 4.6. *There exists $p > 0$ independent of μ such that if $\beta_o(\mu)$ is as in* (4.30), *then $\sigma_R \cap \{$I $\cup$ II $\cup$ III$\} = \varnothing$, where* I, II, III *are the regions depicted in Fig.* 4.6b.

Proof. The proof that I $\cap\, \sigma_R = \varnothing$ is the same as in Theorem 4.4. For $\lambda \in$ II, we claim that $G_{\rm I}(x, \lambda, \mu) > 0$ for $|x| \leqq L$ and all large μ; the proof then proceeds as in case II of the previous theorem. From (4.15), $G_{\rm I}$ will be positive if

$$\frac{\mu^2 UV}{(\alpha + V)^2 + \beta^2} < 1,$$

which will in turn be true if

$$\frac{\mu^2 UV}{(\alpha + V)^2 + \beta^2} < 1.$$

From (4.18), the latter inequality is equivalent to

$$\bar{v}(\mu)^2 - V^2 < \mu^2(\bar{u}(\mu)\bar{v}(\mu) - UV)$$

(recall that $\bar{u}$, $\bar{v}$ are the maxima of U, V). In order to establish this for large μ, note that for such μ we have that

$$\bar{v}(\mu) + V < \mu^2 \bar{u}(\mu).$$

Since $V \leqq \bar{v}(\mu)$ and $U \leqq \bar{u}(\mu)$, the previous inequality is easily seen to imply that the first inequality holds for $|x| \leqq L$.

Finally, suppose that $\lambda \in$ III. We shall split III into three regions:

$$A = \{\lambda\colon 0 \leqq \alpha \leqq \alpha_o, 0 \leqq \beta \leqq \beta_*\},$$

$$B = \{\lambda\colon 0 \leqq \alpha \leqq \alpha_o, \beta_* \leqq \beta \leqq \beta_R(\mu)\},$$

$$C = \{\lambda\colon 0 \leqq \alpha \leqq \alpha_o, \beta_R(\mu) \leqq \beta \leqq \beta_0(\mu)\},$$

where β_* is large, but independent of μ. For $\lambda \in A$ the argument is the same as in case IV of Theorem 4.4.

We next show that $\lambda \in B$ is not an eigenvalue. For the most part, this is proved as in case III of Theorem 4.4. In particular, the region Σ in Fig. 4.5 is positively invariant for $x \in [-L, -x_\gamma(\mu)]$, and a lower bound $\sigma(x, \mu) \geqq \sigma_o > 0$ for $-x_\gamma(\mu) \leqq x \leqq -x_1$ is obtained for $x_1 < x_\gamma(\mu)$ independent of μ, exactly as before. If $\bar{G}_R(\lambda, \mu) > 0$, then $G(x, \lambda, \mu)$ is uniformly positive for $|x| \leqq x_1$ and large μ, and the proof is completed as in the previous theorem. However, since $\bar{G}_R(i\beta_R(\mu), \mu) = 0$ it is no longer true that $\bar{G}_R(\lambda, \mu)$ will be uniformly positive for $\lambda \in B$, so that $G_R(x, \lambda, \mu)$ may in fact become negative for some $|x| \leqq x_1$. For $\lambda \in B$ near $i\beta_R(\mu)$ the important observation is that $\bar{G}_I(i\beta_R(\mu), \mu)$ is negative and of order μ, as can be seen from (4.19) and (4.15) so that $|G_I| \gg |G_R|$ for λ near $i\beta$. Thus the vector field $(\dot{\sigma}, \dot{\tau})$ is nearly vertical for (σ, τ) near the origin. Consider the region Γ depicted in Fig. 4.7 to the right of the union of line segments consisting of the two vertical rays connected by a line segment with fixed positive slope. By the above remarks, Γ will be positively invariant for $(\dot{\sigma}, \dot{\tau})$ for all $|x| \leqq x_1$ since $(\dot{\sigma}, \dot{\tau})$ is nearly vertical near the origin. Since $\sigma(-x_1, \lambda) > \sigma_o > 0$ it follows that $z(-x_1, \lambda) \in \Gamma$. The usual argument now shows that $z(x, \lambda)$ remains finite and inside Γ for $|x| \leqq x_1$. Finally, as in part IV of Theorem 4.4, $\sigma(-x_1, \lambda)$ is positive and of order $\mu^{1/2}$. It then follows from the above that for large μ, $-z(-x_1, \lambda) \notin \Gamma$, so that $z(x_1, \lambda) \neq -z(-x_1, \lambda)$. Thus $\lambda \in B$ is not an eigenvalue.

Next, suppose that $\lambda \in C$, so that

$$\beta_R(\mu) \leqq \beta \leqq \beta_I(\mu) - p. \tag{4.32}$$

For all λ with $0 \leqq \alpha \leqq \alpha_o$ and $\beta \leqq \beta_R(\mu)$, there exists M independent of μ such that $|G_R(x, \lambda, \mu)| \leqq M$ or all $|x| \leqq L$ (see (4.15)). Let

$$x_*(\mu) = x_\gamma(\mu) - K \log \mu/\mu, \tag{4.33}$$

where $K = K(q)$ is chosen as in Theorem 4.1 with $q > 7$. (For the moment, we shall suppress the dependence of x_* on L.) For $|x| \leqq x_*(\mu)$, $G(x, \lambda, \mu)$ is uniformly approximated by $\bar{G}(\lambda, \mu)$. Now $\bar{G}_I(\lambda, \mu)$ is large and negative for $\lambda \in C$. It is easily seen from (4.15) that, given $N > 0$, there exists $p > 0$ independent of μ such that $\bar{G}_I(\lambda, \mu) < -N$ for $\lambda \in C$ with p as in (4.32). Now, given the uniform bound M for $|G_R(\lambda, \mu)|$ for $\lambda \in C$, set $N = N(M)$ so that the region Γ in Fig. 4.7 is positively invariant for $\dot{z} = z^2 - \bar{G}(\lambda, \mu)$. In particular, choosing N large relative to M ensures that $(\dot{\sigma}, \dot{\tau})$ is nearly vertical near the origin. For $|x| \leqq x_*(\mu)$ it follows that Γ will also be positively invariant for (4.24) for large μ. The proof that $\lambda \notin \sigma_R$ is now obtained in the same

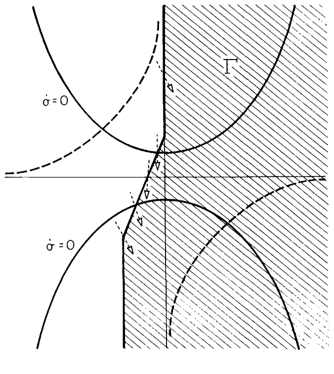


FIG. 4.7

manner as for $\lambda \in B$, by showing that $z(x_*(\lambda), \mu) \neq -z(-x_*(\mu), \lambda)$. The proof will then be complete if it can be shown that $z(-x_*(\mu), \lambda)$ lies in Γ. This fact will be proved in the next lemma, where a sharper estimate for z at $-x_*$ is obtained.

LEMMA 4.7. *For $\lambda \in \mathrm{IV} \cup C$ (see Fig. 4.6b), the solution $z(x, \lambda)$ of (4.22) satisfies*

$$|z(x_*(\mu), \lambda) - \Gamma\beta^{1/2}| \leqq C \log \mu,$$

where $\Gamma = (1+i)/\sqrt{2}$, $C > 0$ is independent of β and μ, and $x_*(\mu)$ is chosen as in (4.33).

Proof. The proof consists of two estimates; first, we approximate z on the outer layer $-L \leqq x \leqq -x_\gamma(\mu)$, and next, on the interval from $-x_\gamma$ to $-x_*$, which is of length $K \log \mu/\mu$. For λ in this range, α is $\mathcal{O}(1)$ and β is $\mathcal{O}(\mu)$, so that β and μ are of the same order of magnitude; in the following, we shall treat them as equivalent parameters.

Since $z = \infty$ at $-L$ it is convenient to work with $w = z^{-1}$, which satisfies (4.23) together with the condition $w = 0$ at $-L$. Let

$$\zeta(y) = \beta^{1/2} w(-L + \beta^{-1/2} y),$$

so that ζ satisfies

$$\dot{\zeta} = 1 - (G/\beta)\zeta^2, \qquad \zeta(0) = 0. \tag{4.34}$$

For $x \in [-L, x_\gamma(\mu)]$ and $\lambda \in \mathrm{III} \cup C$ we have that

$$G/\beta = i + \mathcal{O}(\beta^{-1}).$$

Dropping the $\mathcal{O}(\beta^{-1})$ terms in (4.34) we obtain an exact solution

$$\zeta_*(y) = \frac{1}{\Gamma} \frac{1 - e^{-2\Gamma y}}{1 + e^{-2\Gamma y}},$$

which for large $y > 0$ is exponentially close to Γ^{-1}. Let $\Delta(y) = \zeta(y) - \zeta_*(y)$; then Δ satisfies the equation

$$\dot{\Delta} = -(G/\beta)(\zeta + \zeta_*)\Delta + \mathcal{O}(\beta^{-1}), \qquad \Delta(0) = 0. \tag{4.35}$$

Let $c(y)$ be the coefficient of Δ on the right-hand side of (4.35), so that

$$c(y) = (-i + \mathcal{O}(\beta^{-1}))(\Delta + 2\zeta_*).$$

Since ζ_* is close to Γ^{-1} for large y and $\operatorname{Im} \Gamma^{-1} = -2^{-1/2}$ it follows that given $\gamma \in (0, \sqrt{2})$ there exists $\varepsilon > 0$ and $y_1 > 0$ such that for $|\Delta| \leqq \varepsilon$, $y \geqq y_1$, and large β, we have that $\operatorname{Re} c(y) < -\gamma$. Note that y_1 depends only on $\zeta_*(y)$ and hence is independent of β.

We first estimate $\Delta(y)$ on $0 \leqq y \leqq y_1$. To this end let $D(y)$ be the maximum of $|\Delta(t)|$ on $0 \leqq t \leqq y$. Then from (4.35) we obtain

$$D(y) \leqq K\beta^{-1} \int_o^y e^{tD(t)}\, dt$$

for some constant $K > 0$. We claim that $D(y) \leqq 1$ on $[0, y_1]$. Let $y_* \in [0, y_1]$ be the smallest $y > 0$ for which $D(y_*) = 1$. It then follows from the above that

$$D(y_*) \leqq K\beta^{-1}(e^{y_*} - 1);$$

since y_1 and y_* are bounded independently of β, the right-hand side can be made strictly less than 1 by choosing β large. Hence $y_* \geqq y_1$ and $D(y) \leqq \tilde{K}\beta^{-1}$ for $\tilde{K} = K(e^{y_1} - 1)$ on $[0, y_1]$. Thus for large β we will have that $D(y_1) < \varepsilon$, so that $\operatorname{Re} c(y_1) < -\gamma$.

Set β so large that $\tilde{K}(L+\gamma^{-1})\beta^{-1}<\varepsilon$ where $L=\int_0^{y_1}|\exp\int_t^{y_1}c(r)\,dr|\,dt$. For all $y\geqq y_1$ for which $\operatorname{Re} c(y)<-\gamma$ we have for $y_1\leqq t\leqq y$ that

$$\left|\exp\int_t^y c(r)\,dr\right|\leqq e^{-\gamma(y-t)}$$

for such y. It then follows that

$$\begin{aligned}|\Delta(y)|&\leqq K\beta^{-1}\int_0^y\left|\exp\int_t^y c(r)\,dr\right|dt\\ &\leqq K\beta^{-1}\int_0^{y_1}\left|\exp\int_t^{y_1}\operatorname{Re} c(r)\,dr\right|dt+K\beta^{-1}\int_{y_1}^y e^{-\gamma(y-t)}\,dt\\ &\leqq\tilde{K}\beta^{-1}(L+\gamma^{-1}).\end{aligned}$$

Thus $|\Delta(y)|<\varepsilon$ for such y. It follows that $|\Delta(y)|\leqq K\beta^{-1}\leqq\varepsilon$ for all $y\in[0,\beta^{1/2}(L-x_\gamma(\mu))]$ and some (new) $K>0$. We therefore have that

$$|\zeta(y)-\Gamma^{-1}|\leqq K\beta^{-1}+(e^{-2^{1/2}y}).$$

Returning to z it then follows that

$$z(x,\lambda,\mu)=\beta^{1/2}\Gamma+\mathcal{O}(\beta^{-1/2})$$

for $x\in[-L,-x_\gamma(\mu)]$.

We finally estimate z on the interval $[-x_\gamma,-x_*]$, where x_* is as in (4.33). On this interval, $G_I(x,\lambda,\mu)$ undergoes a rapid transition from β to $\bar{G}_{\mathrm{I}}(\lambda,\mu)$, which is negative and $\mathcal{O}(1)$, and $|G_R|$ remains uniformly $\mathcal{O}(1)$. Let $z_\gamma=z(x_\gamma,\lambda,\mu)$ and set $\Delta=z-z_\gamma$, so that $\Delta(x_\gamma)=0$ and

$$\dot{\Delta}=E(x)-2z_\gamma\Delta-\Delta^2,$$

where $E(x)=G-z_\gamma^2$ is of order β on $[-x_\gamma,-x_*]$. From the previous estimate, $\operatorname{Re} z_\gamma=\sigma_\gamma$ is close to $\beta^{1/2}/\sqrt{2}$, and so σ_γ is large and positive. We then obtain

$$|\Delta(x)|\leqq\int_{-x_\gamma}^x e^{-\sigma_\gamma(x-t)}|E(t)|\,dt+\int_{-x_\gamma}^x e^{-\sigma_\gamma(x-t)}\,|\Delta(t)|^2\,dt.$$

Let $D(x)$ be the maximum of $|\Delta(x)|$ on $[-x_\gamma,-x]$; we then have for some $C>0$ that $|E(x)|\leqq C\beta$, so that

$$D(x)\leqq C\beta\sigma_\gamma^{-1}[1-e^{-\sigma_\gamma(x+x_\gamma)}]+\sigma_\gamma^{-1}[1-e^{-\sigma_\gamma(x+x_\gamma)}]D(x)^2.$$

Since $\sigma_\gamma=\beta^{1/2}/\sqrt{2}+\mathcal{O}(\beta^{-1/2})$ it follows for some constant $C>0$ and $x\in[-x_\gamma,-x_*]$,

$$\begin{aligned}|1-e^{-\sigma_\gamma(x+x_\gamma)}|&\leqq C\sigma_\gamma\log\mu/\mu\\ &=C\log\mu/\mu^{1/2}\end{aligned}$$

since β and μ are of the same order. For large μ, we can therefore arrange that

$$0\leqq\sigma_\gamma^{-1}[1-e^{-\sigma_\gamma(x+x_\gamma)}]<C\log\mu/\mu$$

for $x\in[-x_\gamma,-x_*]$ and some $C>0$. From the above inequality for $D(x)$ we then obtain

$$D(x)<C\log\mu+[C\log\mu/\mu]D^2(x)\tag{4.36}$$

for some $C>0$ independent of μ and for $x\in[-x_\gamma,-x_*]$. Choose μ so large that $C^2\log^2\mu/\mu<1/4$. Since $D(x_\gamma)=0$ and (4.36) holds for $x\in[-x_\gamma,-x_*]$, it easily follows that $D(x) \leqq 2C \log \mu$ for all such x. Since z_γ and $\Gamma\beta^{1/2}$ differ by an error of order $\beta^{-1/2}$, the proof of Lemma 4.6 is complete.

We shall also require the following estimate for $z_\beta(x, \lambda)$ on the interval $[-L, -x_*]$.

LEMMA 4.8. *For $\lambda \in \mathrm{III}$ and $x \in [-L, -x_*]$, the solution $z(x, \lambda)$ of (4.22) satisfies $|z_\beta(x, \lambda)| \leqq C\beta^{-1/2}$.*

The proof, which uses the fact that $|G_\beta|$ is uniformly $\mathcal{O}(1)$ on this interval for all such λ, is similar to and simpler than that of Lemma 4.7, and will therefore be omitted.

Finally we describe $\sigma_R \cap \mathrm{IV}$. We have now determined $z(x, \lambda)$ on the interval $[-L, -x_*(\mu)]$; the crucial estimate is

$$z(-x_*(\mu), \lambda) = \Gamma\beta^{1/2} + \mathcal{O}(\log \mu).$$

On the interior interval $|x| \leqq x_*(\mu)$, the potential $G(x, \lambda, \mu)$ in (4.15) is uniformly approximated by the autonomous potential, $\bar{G}(\lambda, \mu)$ in (4.16). For $\lambda \in \mathrm{III}$, $\bar{G}(\lambda, \mu)$ is uniformly bounded for all μ. Thus the projectivized, nonautonomous flow on $\mathbb{C}P^1$ (as given by (4.22), (4.23) in local coordinates) is uniformly approximated by the autonomous projectivized flow, which in local coordinates is given by

$$\dot{z} = \bar{G}(\lambda, \mu) - z^2, \tag{4.37}$$

$$\dot{w} = 1 - \bar{G}(\lambda, \mu) w^2 \tag{4.38}$$

on the interval $|x| \leqq x_*(\mu)$. By the remark following Lemma 4.3 we have that $\lambda \in \sigma_R$ if and only if $z(0, \lambda) = 0$ or ∞ (see (4.27)). We therefore see that λ will be an approximate eigenvalue of (4.22) whenever there exists a solution $\bar{z}(x, \lambda)$ of (4.7) satisfying

$$\bar{z}(-x_*(\mu), \lambda) = z(-x_*(\mu), \lambda),$$
$$\bar{z}(0, \lambda) = 0 \quad \text{or} \quad \infty.$$

Since (4.37) is autonomous, this suggests introducing a new parameter l for the size of the interior interval. Ultimately, l will have to be related to the entire interval size L. Recall that x_* depends implicitly on L, which we have suppressed until now. The precise relation follows from the estimates in § 4.1, namely,

$$x_*(\mu; L) = L - \alpha(p_\gamma) + \mathcal{O}(\log \mu / \mu),$$

so that $x_*(\mu, L)$ and L essentially differ by a constant. Our strategy will be to first treat the autonomous eigenvalue problem consisting of (4.37) together with the condition

$$\begin{aligned} &\bar{z}(0, \lambda) = z(-x_*(\mu; L), \lambda), \\ &\bar{z}(l, \lambda) = 0 \quad \text{or} \quad \infty. \end{aligned} \tag{4.39}$$

The values $l = \bar{l}$ for which (4.37), (4.39) has a solution will then determine approximate eigenvalues of the exact equations provided that $\bar{L}$ is chosen so that $\bar{l} = x_*(\bar{L}, \mu)$. Exact eigenvalues will then be obtained by a transversality argument. This strategy is made feasible by Lemma 4.7, which estimates $z(-x_*(\mu; L), \lambda)$. In fact, even though this point depends on L, it is of the form

$$z(-x_*(\mu; L), \lambda) = \Gamma\beta^{1/2} + \mathcal{O}(\log \mu);$$

it follows that the error induced in the solution of (4.37) due to the $\log \mu$ term is of order $\log \mu / \mu^{1/2}$, which can be seen by explicitly solving the equation. It follows that the values $l = \bar{l}$ for which (4.37), (4.39) have solutions depending on L only up to an error of order $\log \mu / \mu^{1/2}$. Hence in locating approximate eigenvalues it is sufficient to drop the $\mathcal{O}(\log \mu)$ error terms in the above and treat z at $-x_*$ as essentially $\Gamma\beta^{1/2}$.

We first consider the existence of purely imaginary eigenvalues $\lambda = i\beta$. Let $\bar{z}(x, i\beta)$ be the solution of (4.37) with initial value $\bar{z}(0, \beta, \mu) = \Gamma\beta^{1/2}$. Let $\gamma = \sqrt{G(i\beta, \mu)}$, so that for $\beta < \beta_I(\mu)$, γ is a repeller and $-\gamma$ is an attractor for (4.37). It follows that the solution $\bar{z}(x, i\beta)$ must remain exterior to a neighborhood of γ for all $x \geqq 0$ and must eventually tend to $-\gamma$ as $x \to +\infty$. Since this is actually a flow on $\mathbb{C}P^1$, $\bar{z}$ may blow up in finite time; by using (4.38) in place of (4.37) at such a point $x = \bar{x}$, the usual argument applied to (4.25) shows that

$$\lim_{x \to \bar{x}^-} \operatorname{Re} \bar{z} = -\infty, \qquad \lim_{x \to \bar{x}^+} \operatorname{Re} \bar{z} = +\infty,$$

and that $\operatorname{Im} z$ tends to zero as $x \to \bar{x}$ through positive (respectively, negative) values as $x \to \bar{x}^-$ (respectively, $x \to \bar{x}^+$). Hence if $\bar{z}$ blows up at $\bar{x}$, it is asymptotic to the negative σ-axis from above for $x < \bar{x}$ and to the positive σ-axis from below for $x > \bar{x}$. We therefore see that since $\operatorname{Im} z \leqq 0$ is positively invariant for the flow induced by (4.37), (4.38) on the Riemann sphere, including the point at infinity, there exists a uniquely determined finite time $\bar{l}(\beta) > 0$ for which

$$\operatorname{Im} \bar{z}(\bar{l}(\beta), i\beta) = 0$$

(or where $\bar{z} = \infty$), provided that $\beta < \beta_I(\mu)$. Note that when $\beta = \beta_I(\mu)$, $\bar{G}_I = 0$ so that all solutions of (4.37) are periodic; the particular solution $\bar{z}$ never crosses $\operatorname{Im} z = 0$ and winds about the center γ infinitely often in the clockwise direction. Given $N > 0$ select $\beta_1(\mu) \in (\beta_o(\mu), \beta_I(\mu))$ such that $\bar{z}(x, \beta_1)$ winds about γ exactly N times on the interval $[0, \bar{l}(\beta_1)]$. (See Fig. 4.8.) (More precisely, if we connect the points $\bar{z}$ at $x = 0$ and $x = l(\beta_1)$ with a line segment, the resulting closed curve has winding number $-N$ with respect to γ.) Suppose that the interval $[\beta_o(\mu), \beta_1(\mu)]$ contains no eigenvalues. It follows from our previous estimates that $\operatorname{Re} \bar{z}(\bar{l}(\beta_o), i\beta) > 0$ since such λ also lie in region C. In view of our characterization of $\bar{z}$ near a point $\bar{x}$ where blowup occurs and our assumption that $\sigma_R \cap [i\beta_o i\beta_1] = \varnothing$, it follows that

$$0 < \operatorname{Re} \bar{z}(\bar{l}(\beta, \mu), \beta) < \infty$$

for all $\beta \in [\beta_o, \beta_1]$.

Let

$$D = \{(x, i\beta) \colon \beta_o(\mu) \leqq \beta \leqq \beta_1(\mu), 0 \leqq x \leqq \bar{l}(\beta)\}$$

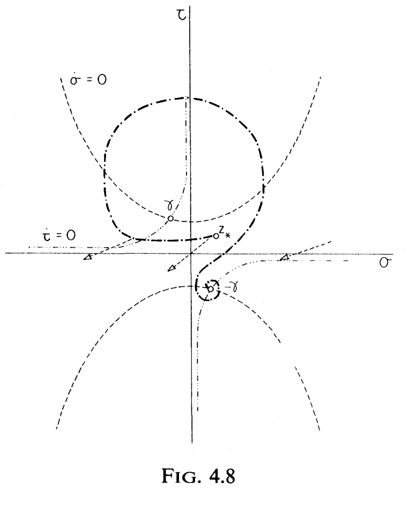


FIG. 4.8

so that D is a 2-cell. The solution $\bar z$ maps D into the hemisphere of $\mathbb{C}P^1$ corresponding to the $\operatorname{Im} z \geqq 0$; since $\bar z$ never hits the point γ for $(x, i\beta) \in D$ we may regard $\bar z$ as a map from D into the punctured disk in $\mathbb{C}P^1$ consisting of $\operatorname{Im} z \geqq 0$ minus the point γ. Furthermore, since we are assuming that $[\beta_o, \beta_1]$ contains no eigenvalues, the (real) curve $\bar z(\bar l(\beta, \mu), i\beta)$ remains in the ray $0 < \operatorname{Re} z < \infty$ for all $\beta \in [\beta_o, \beta_1]$. The behavior of the map $\bar z$ as a map from ∂D into the punctured disk is therefore as depicted in Fig. 4.9; in particular, our assumption that there are no eigenvalues imply that the image of the curve BC lies in the ray $\operatorname{Im} z = 0$, $\operatorname{Re} z > 0$, and therefore has zero winding with respect to the deleted point γ. Thus $\bar z(\partial D)$ represents a class $-N\bar 1$ in $\pi_1(S^1)$, where $\bar 1$ is an appropriate generator for $\pi_1(S^1)$ and $-N$ is the winding of segment (3) about γ, as defined earlier. This provides a contradiction, since $\bar z$ extends to a continuous map from $D^2 \to S^1$, so that $\bar z: \partial D \to S^1$ is homotopic to a constant map. Hence our assumption that $[\beta_o, \beta_1]$ contained no eigenvalues must have been false.

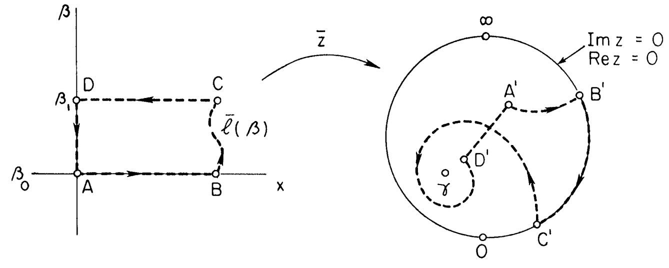


FIG. 4.9

The above argument actually shows that with N as above, $\bar z(\bar l(\beta), i\beta)$ must pass through *both* zero and ∞ at least N times. We shall see next that each crossing must be transverse in the counterclockwise direction in Fig. 4.9, so that the point C^1 must cross zero and ∞ in the counterclockwise direction as β is decreased from β_1 to β_o. Hence, there are *exactly* $2N$ values of β where approximate eigenvalues occur, which we denote by $\bar\beta_n^o, \bar\beta_n^\infty$, $n = 0, \cdots, N-1$. We shall also see below that the time $\bar l(\beta)$ is monotonically increasing with β, near each $\bar\beta_n^j$, and that the approximate eigenvalues occur at interior interval sizes

$$\bar l_n^o = \bar l(\bar\beta_n^o), \qquad \bar l_n^\infty = \bar l(\bar\beta_n^\infty),$$

with

$$\bar l_n^o < \bar l_n^\infty < \bar l_{n+1}^o. \tag{4.40}$$

Let $Z(\beta) = \bar z(l(\beta), i\beta)$, so that $Z(\beta)$ is real and $Z(\bar\beta_n^j) = j$ for $j = 0, \infty$. We first consider the case where β is near $\bar\beta_n^o$. The solution $\bar z(x, i\beta)$ of (4.37) with initial data $z_* = \Gamma\beta^{1/2}$ has the implicit form

$$\log\frac{\gamma+\bar z}{\gamma-\bar z} = 2\gamma\bar l + \log\frac{\gamma+z_*}{\gamma-z_*}, \tag{4.41}$$

$$z_* = z(-x_*(\mu, L), \beta, \mu);$$

however, some care must be used in specifying the branch of the log for β near β_n^j. To this end let $z = re^{i\phi}$ and define

$$\operatorname{Log} z = \log r + \phi i$$

with $-\pi < \phi \leqq \pi$. From Fig. 4.8 it is evident that for $x \leqq \bar{l}(\beta)$, $\bar{z}(x, \beta)$ remains in $\operatorname{Re} z \geqq 0$ and winds $n \geqq 0$ times about the rest point at γ in the *clockwise* direction. It follows that $(\gamma+\bar{z})/(\gamma-\bar{z})$ winds n times in the *counterclockwise* direction about zero, so that the angle of $(\gamma+\bar{z})/(\gamma-\bar{z})$ must increase by $2\pi n$ on $[0, \bar{l}(\beta)]$. Also, z_* is of order $\beta^{1/2}$, so that $(\gamma+z_*)/(\gamma-z_*) = -1+\mathcal{O}(\beta^{-1/2})$. We therefore obtain

$$\operatorname{Log} \frac{\gamma+\bar{z}}{\gamma-\bar{z}}\bigg|_{x=l(\beta)} = 2\gamma\bar{l} - (2n+1)\pi i + \mathcal{O}(\beta^{-1/2})$$

with $n \geqq 0$. Now for $\bar{z}$ near zero, we have that $(\gamma+\bar{z})/(\gamma-\bar{z}) = 1+2\bar{z}\gamma^{-1}+\mathcal{O}(\bar{z}^2)$; recalling that $\gamma^2 = \bar{G}$ we then have for such $\bar{z}$ that

$$\bar{z} = \bar{G}\bar{l} - (n+\tfrac{1}{2})\pi i\gamma + \mathcal{O}(\beta^{-1/2}) + \mathcal{O}(\beta-\beta_n^\circ)^2.$$

The defining condition for $\bar{l}(\beta)$ is that $\operatorname{Im} \bar{z}(\bar{l}, \beta) = 0$. Let $\gamma = \gamma_R + i\gamma_I$ and $\bar{Z}(\beta) = \bar{z}(\bar{l}(\beta), i\beta)$; we then obtain

$$\bar{l}(\beta) = (n+\tfrac{1}{2})\pi\gamma_R/\bar{G}_I + \mathcal{O}(\beta^{-1/2}) + \mathcal{O}(\beta-\beta_n^\circ)^2,$$

$$\bar{Z}(\beta) = \bar{G}_R\bar{l} + (n+\tfrac{1}{2})\pi\gamma_I + \mathcal{O}(\beta^{-1/2}) + \mathcal{O}(\beta-\beta_n^\circ)^2.$$

From Lemma 4.8 we have that $|z_{*\beta}| = \mathcal{O}(\beta^{-1/2})$; it therefore follows that

$$\bar{l}'(\bar{\beta}_n^\circ) = \left(n+\frac{1}{2}\right)\frac{\gamma_{R\beta}\bar{G}_I - \gamma_R\bar{G}_{I\beta}}{\bar{G}_I^2}.$$

It can be seen from (4.15) that $\bar{G}_{I\beta} > 0$ and is $\mathcal{O}(1)$ for $\beta \in [\beta_o, \beta_I]$; it is also easily seen that $|G_{R\beta}| = \mathcal{O}(\beta^{-1})$. Since $\gamma = \sqrt{G}$, $\gamma_R < 0$ and $\gamma_I > 0$, so that the second term in $\bar{l}'$ is positive. A simple computation shows that

$$\gamma_{R\beta} = \frac{\bar{G}_{I\beta}\gamma_I}{2|\gamma|^2} > 0, \qquad \gamma_{I\beta} = \frac{\bar{G}_{I\beta}\gamma_R}{2|\gamma|^2} < 0; \tag{4.42}$$

thus the first term in $\bar{l}'$ is negative. In order to determine the sign of $\bar{l}'$, let $\bar{G} = |\bar{G}|e^{i\phi}$, so that $\phi \in (\pi, 3\pi/2)$, and

$$\gamma = |\bar{G}|^{1/2}\left(-\sin\frac{\phi-\pi}{2} + i\cos\frac{\phi-\pi}{2}\right).$$

Let $a = |\bar{G}_R|/|\bar{G}|$ and $b = |\bar{G}_I|/|\bar{G}|$; we then have that

$$\gamma_R = -|\bar{G}|^{1/2}\left[\frac{1-a}{2}\right]^{1/2}, \qquad \gamma_I = |G|^{1/2}\left[\frac{1+a}{2}\right]^{1/2}. \tag{4.43}$$

Combining the above we obtain

$$\begin{aligned}\bar{l}'(\beta) &= \frac{(n+\frac{1}{2})\pi\bar{G}_{I\beta}|\bar{G}|^{1/2}}{\bar{G}_I^2}\left[\frac{\bar{G}_I\gamma_I}{2|\bar{G}|} - \gamma_R\right] \\ &= \frac{(n+\frac{1}{2})\pi\bar{G}_{I\beta}|\bar{G}|^{1/2}}{\bar{G}_I^2}\left[-\frac{b}{2}\sqrt{\frac{1+a}{2}} + \sqrt{\frac{1+a}{2}}\right].\end{aligned}$$

The sign of $\bar{l}'(\beta)$ is therefore determined by the quantity in brackets. Since $a^2+b^2=1$ and $0 \leqq a, b \leqq 1$, we have that the latter quantity is positive if and only if

$$1 > \frac{(1+a)^2}{4},$$

which clearly is valid for all $a \in (0, 1)$. We have therefore established that $\bar{l}'(\bar{\beta}_n^{\circ}) > 0$. We also have that

$$\bar{Z}'(\bar{\beta}_n^{\circ}) = \bar{G}_{R\beta}\bar{l} + \bar{G}_R\bar{l}' + (n+\tfrac{1}{2})\pi\gamma_{I\beta} + \mathcal{O}(\beta^{-1/2}),$$

since $G_{R\beta} = \mathcal{O}(\beta^{-1})$, so that from the above and (4.42), $\bar{Z}'(\bar{\beta}_n^{\circ}) < 0$.

Next, we characterize the behavior of $\bar{l}(\beta)$ and $\bar{z}(l, i\beta)$ for β near $\bar{\beta}_n^{\infty}$. To this end introduce the local coordinate $\zeta = \bar{G}/z$ on $\mathbb{C}P^1$; the equation for $\bar{\zeta}$ is then

$$\dot{\bar{\zeta}} = \bar{G} - \zeta^2, \qquad \bar{\zeta}(0, \beta) = \bar{G}/z_* = \zeta_*.$$

This is the same equation as for z; only the initial condition ζ_* is different. Here $\zeta_* = \mathcal{O}(\beta^{-1/2})$ is near zero. Since $\zeta(\bar{l}(\bar{\beta}_n^{\infty}), i\beta_n^{\infty}) = 0$, this suggests introducing the time $\hat{l}(\beta)$ (defined for β near $\bar{\beta}_n^{\infty}$) for which $\operatorname{Im} \zeta(\hat{l}, i\beta) = 0$. The formula for $\bar{\zeta}$ is therefore given implicitly by (4.40) with $\bar{z}$, z_*, and $\bar{l}$ replaced by $\bar{\zeta}$, ζ_*, and $\hat{l}$, respectively. The main difference is that $\bar{\zeta}$ and ζ_* are both near zero. Since $\bar{\zeta}$ must wind in the clockwise direction about γ it follows that $(\gamma+\zeta)/(\gamma-\zeta)$ must wind $(n+1)$ times about the origin in the counterclockwise direction as x increases from zero to $\hat{l}(\beta)$. It follows that the equation for $\bar{\zeta}$ and $\hat{l}$ is

$$\operatorname{Log}\frac{\gamma+\bar{\zeta}}{\gamma+\bar{\zeta}} = 2\gamma\hat{l} - 2\pi(n+1)i + \mathcal{O}(\beta^{-1/2}).$$

Let $\bar{\zeta}_R(\beta) = \bar{\zeta}(\hat{l}(\beta), \beta)$; we then obtain

$$\bar{\zeta}_R(\beta) = \bar{G}\hat{l} - \pi(n+1)\gamma i + \mathcal{O}(\beta^{-1/2}) + \mathcal{O}(\beta - \bar{\beta}_n^{\infty})^2.$$

The argument now proceeds exactly as before to show that $\hat{l}'(\bar{\beta}_n^{\infty}) > 0$ and that $\zeta'_R(\bar{\beta}_n^{\infty}) < 0$.

The behavior of $\bar{z}(x, \beta)$ can now be determined. Let $w = z^{-1}$ so that $w = \zeta/\bar{G}$. Since $\bar{\zeta}'_R(\bar{\beta}_n^{\infty}) < 0$ it follows that $w(\hat{l}(\beta), i\beta)$ lies along the span of $1/\bar{G}$; since $\bar{\zeta}_R(\bar{\beta}_n^{\infty}) < 0$, its behavior for $\beta_1 < \bar{\beta}_n^{\infty} < \beta_2$ is as depicted in Fig. 4.10. It is easily seen that this is equivalent to the point $\bar{z}(\hat{l}(\beta), i\beta)$ moving in the clockwise direction through the point at infinity in Fig. 4.9 as β increases through $\bar{\beta}_n^{\infty}$. From Fig. 4.10 it is evident that $\bar{z}(\bar{l}(\beta), i\beta)$ also moves in the clockwise direction about the equation $\operatorname{Im} z = 0$. Also from Fig. 4.10, we see that

$$\bar{l}(\beta_1) < \hat{l}(\beta_1), \qquad \bar{l}(\beta_2) > \hat{l}(\beta_2);$$

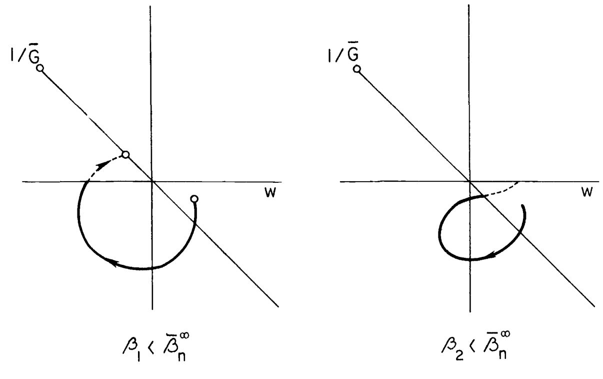

1/Ḡ
w
$\beta_1 < \bar{\beta}_n^{\infty}$
1/Ḡ
w
$\beta_2 < \bar{\beta}_n^{\infty}$

FIG. 4.10

hence

$$\hat{l}(\beta_2)-\hat{l}(\beta_1)<\bar{l}(\beta_2)-\bar{l}(\beta_1),$$

and since $\hat{l}'(\bar{\beta}_n^\infty)>0$ it follows that $\bar{l}(\beta)$ is strictly monotone increasing for β near β_n^∞.

We also need to show that the critical intervals $\bar{l}_n^j$ satisfy (4.40). This is easily established from our formulae for the $\bar{l}_n^j$, namely,

$$\bar{l}_n^o=\pi\left(n+\frac{1}{2}\right)\frac{\gamma_R}{\bar{G}_I}\bigg|_{\beta=\bar{\beta}_n^o}+\mathcal{O}(\beta^{-1/2}),$$

$$\bar{l}_n^\infty=n\pi\frac{\gamma_R}{\bar{G}_I}\bigg|_{\beta=\bar{\beta}_n^\infty}+\mathcal{O}(\beta^{-1/2}).$$

In order to establish (4.40) it suffices to show that the function $(\gamma_R/\bar{G}_I)$ is monotone increasing in β. This will be the case if $(\gamma_R\bar{G}_{I\beta}-\gamma_R\bar{G}_I)>0$; from (4.42), (4.43), and a bit of algebra, this is equivalent to

$$\left[-\frac{1}{2}\left(1+\frac{|\bar{G}_R|}{|\bar{G}|}\right)+1\right]\frac{|\bar{G}_I|}{|\bar{G}|}>0,$$

which, since $|\bar{G}_R|<|\bar{G}|$, is clearly valid.

Our analysis of (approximate) imaginary eigenvalues is now complete. It remains to describe the multiplicity of all unstable (or neutrally stable) approximate eigenvalues for a given l. To this end we take β near $\bar{\beta}_n^j$ and l near $\bar{l}_n^j$, and reverse their roles, i.e., we treat l as the independent variable. Our goal is to show that the purely imaginary eigenvalue $\bar{\lambda}_n^j=i\bar{\beta}_n^j$ lies along a branch of eigenvalues $\bar{\lambda}_n^j(l)$ for l near $\bar{l}_n^j$; in the following, we shall drop the sub- and superscripts. This is accomplished by the implicit function theorem. Moreover, we shall show that $\alpha'(\bar{l}_n^j)>0$, where $\alpha(l)=\operatorname{Re}\lambda(l)$. This implies that eigenvalues can only cross from the left (stable) half plane into the right (unstable) half plane, and that the crossing is transverse. Hence, the unstable eigenvalue count is

$$\begin{array}{ll}0 & \text{for } l<\bar{l}_o^o,\\ 2(2n+1) & \text{for } \bar{l}_n^o<l<\bar{l}_n^\infty,\\ 2(2n+2) & \text{for } \bar{l}_n^\infty<l<\bar{l}_{n+1}^\infty\end{array}$$

(counting complex conjugates).

The computation proceeds as follows for the case $j=0$. The case $j=\infty$ is treated similarly using the variable $\zeta=\bar{G}/z$, and will be omitted. Let $\bar{z}(x,\lambda)$ be the solution of (4.37) satisfying $\bar{z}(0,\lambda)=\Gamma\beta^{1/2}$, so that $\bar{z}(\bar{l}_n^o,i\bar{\beta}_n^o)=0$. Assume the existence of a curve $\bar{\lambda}(l)$ of eigenvalues with $\bar{\lambda}(\bar{l}_n^o)=i\bar{\beta}_n^o$, we see that $\lambda(l)$ must satisfy $\bar{z}(l,\lambda(l))=0$, so that at $l=\bar{l}_n^o$,

$$\bar{\lambda}'(l)=-\frac{\bar{z}_l}{\bar{z}_\lambda}.$$

These partials can be computed directly from (4.41), yielding $\bar{z}_\lambda=\frac{1}{2}\bar{G}_\lambda l$ and $\bar{z}_l=\bar{G}$ (modulo terms of order $\beta^{-1/2}$). From the display preceding (4.15) we see that $\bar{G}_\lambda=1+\mathcal{O}(\beta^{-2})$, so that $\bar{\lambda}'(l)=-2\bar{G}/l+\mathcal{O}(\beta^{-1/2})$. Since $\bar{G}=\lambda-\bar{f}_u+\mathcal{O}(\beta^{-2})$, by taking real parts we finally obtain $\bar{\alpha}'(\bar{l}_n^o)=2\bar{f}_u/l+\mathcal{O}(\beta^{-1/2})$, where the partial is evaluated at the rest point $(\bar{u},\bar{v})$. Since we are in Case U it turns out that $(\bar{u},\bar{v})$ must lie on the left branch of $v=(b-u)(u-a)$ in Fig. 2.2, so that $\bar{f}_u>0$. We have therefore proved that $\bar{\alpha}'(\bar{l}_n^o)>0$.

We now return to the exact equation (4.22). There are two issues that need to be addressed: (i) the outer interval size L must be determined in some manner by the inner interval size l, and (ii) the autonomous potential $\bar{G}$ must be replaced by $\bar{G}(x, \lambda, \mu)$ on the interval $|x| \leqq x_*(\mu; L)$. We first determine interval sizes $\bar{L}^j_n$ that determine approximate, purely imaginary eigenvalues on the full interval $|x| \leqq L$. To this end, note that the inner interval size $x_*(\mu, L) = x_\gamma(\mu, L) - K \log \mu/\mu$ of the underlying solution depends on L through $x_\gamma(\mu, L)$, and that by the time map estimates § 4.1, we have that

$$x_\gamma(\mu, L) = L - \alpha(p_\alpha) + \mathcal{O}(\mu^{-1}).$$

Furthermore, it also follows from the time map estimates of § 4.1 that $x_{*L} = x_{\gamma L} = 1 + \mathcal{O}(\mu^{-1})$. In particular, this is proved by differentiating $T = -x_\gamma + L$ in (4.5) with $\hat{U} = \gamma$ with respect to L and noting that the constant $R(\gamma, \mu) = \mu^{-2} g(\mu, L)$ where g_L is uniformly bounded. Thus $x_*(\mu, L)$ can be made to monotonically sweep through any desired range of values by suitably varying L.

Next, we explicitly denote the L dependence of the exact solution z by writing $z = z(x, \lambda, L)$ for the solution of (4.22) satisfying $z = -\infty$ at $x = -L$ and $w = w(x, \beta, L)$ for z^{-1}. We shall now approximate z and w by $\bar{z}$ and $\bar{w} = \bar{z}^{-1}$, respectively. Recall that the overlapping coordinate patches $|z| \leqq 2$ and $|w| \leqq 2$ cover the Riemann sphere $\mathbb{C}P^1$.

LEMMA 4.9. *There exists $K > 0$ depending only on L such that*

(i) $$|z(x, \lambda, L) - \bar{z}(x + x_*(\mu, L), \lambda)| \leqq K \log \mu/\mu$$

on $|x| \leqq x_*(\mu, L)$ whenever $|\bar{z}| \leqq 2$;

(ii) $$|w(x, \lambda, L) - \bar{w}(x + x_*(\mu, L), \lambda)| < K \log \mu/\mu$$

on $|x| \leqq x_*(\mu, L)$ whenever $|\bar{w}| \leqq 2$.

Proof. Since $\bar{w}$ and w are $\mathcal{O}(\beta^{-1/2})$ at $x = -x_*$ we begin with the w-estimate. Let $x_1 \in [-x_*, x_*]$ be the largest $x > x_*$ such that $|\bar{w}| \leqq 2$. Let $\Delta = w(x, \lambda, L) - \bar{w}(x + x_*, \lambda)$; by Lemma 4.7 we have that $|\Delta(-x_*)| = \mathcal{O}(\log \mu/\mu)$, so that

$$\dot{\Delta} = -(G - \bar{G})w^2 - \bar{G}(w + \bar{w})\Delta,$$

$$\Delta(-x_*) = \mathcal{O}(\log \mu/\mu).$$

Recall that G and $\bar{G}$ are uniformly bounded for $\lambda \in \mathrm{IV}$ of Fig. 4.6, so that for $|w|$, $|\bar{w}| \leqq 2$ the coefficient of Δ is uniformly bounded, by some constant $C > 0$. By Theorem 4.1 $|G - \bar{G}|$ is uniformly of order $\mu^{-(q-7)}$ for $|x| \leqq x_*$, so that for $q = 8$ this term is of order μ^{-1}. An application of Gronwall's inequality yields the estimate in (ii) where K depends only on $\exp[C(x_1 - x_*)]$.

At $x = x_1$, $\bar{w} = 2$ so that $|\bar{z}| \leqq \frac{1}{2}$. Let $x_2 \in [x_1, x_*]$ be the largest x such that $|\bar{z}(x_2, \lambda)| \leqq 2$. We apply a similar estimate to the equation satisfied by $z - \bar{z}$ to obtain estimate (i) on $[x_1, x_2]$. Continuing in this manner we obtain at least one of the estimates in (i) or (ii) for all $|x| \leqq x_*$.

Lemma 4.9 can be stated more compactly in terms of the projection $\pi: \mathbb{C}^2 \to \mathbb{C}P^1$. For $\Gamma \in \mathbb{C}^2 \backslash \{0\}$ let $\hat{\Gamma} = \pi(\Gamma) = \operatorname{span} \Gamma$ and with a slight abuse of notation let $\hat{z} = \pi(1, z)$ (respectively, $\hat{w} = \pi(w, 1)$), so that for $w = z^{-1}$, $\hat{z} = \hat{w}$. Also, let $\hat{0} = \pi((1, 0))$ and $\hat{\infty} = \pi(0, 1)$. If ρ is a fixed metric on $\mathbb{C}P^1$, then for $|x| \leqq x_*(\mu, L)$ we have that

$$\rho(\hat{z}(x, \lambda, L), \hat{\bar{z}}(x + x_*(\mu, L), \lambda)) \leqq K \log \mu/\mu.$$

Let L_o be determined by the equation $\bar{l}^{\infty}_{N-1} = x_*(\mu, L_o)$. We first determine $\delta > 0$ so that $\hat{\bar{\bar{Z}}}(\beta) = \hat{\bar{z}}(\bar{l}(\beta), i\beta)$ is uniformly bounded away from $\hat{0}$ and $\hat{\infty}$ for $|\beta - \bar{\beta}^j_n| \geqq \delta$,

$j=0, \infty$, and $n=0, \cdots, N-1$. For $j=0$ this follows from $\bar{Z}'(\bar{\beta}_n^\circ)<0$ and for $j=\infty$, from $\bar{W}'(\bar{\beta}_n^\infty)>0$ for $\bar{W}=\bar{Z}^{-1}$. For $|\beta-\bar{\beta}_n^\circ|\geqq\delta$ and $x\geqq 0$ there exists $\varepsilon>0$ such that

$$\rho(\bar{z}^{\wedge}(x, i\beta), \hat{0})>\varepsilon.$$

This is because $z(x, i\beta)$ crosses $\operatorname{Im} z=0$ exactly once, at $x=\bar{l}(\beta)$, at which point $\operatorname{Re}\bar{z}\neq 0$ by our condition on β. For x near zero, $\bar{z}^{\wedge}(x, i\beta)$ starts near $\hat{\infty}$, however, since $w'=1+\mathcal{O}(w^2)$ for w near zero, there exists a small $x_1>0$ such that $\rho(\bar{z}^{\wedge}(x_1, i\beta), \hat{\infty})>0$. Using our condition on β, we have that

$$\rho(\bar{z}^{\wedge}(x, i\beta\ ), \hat{\infty})\geqq\varepsilon$$

for $x\geqq x_1$. Now by Lemma 4.9, we have that

$$\rho(\hat{z}(0, i\beta, L), \bar{z}^{\wedge}(x_*(\mu, L), i\beta))\leqq K\log\mu/\mu.$$

For $\beta\leqq\bar{\beta}_{N-1}^\infty+\delta$ we then have that $\hat{z}(0, i\beta, L)\not\in\{\hat{0}, \hat{\infty}\}$ for all L. Finally, for $L\leqq L_o$ it follows that $\bar{l}_n^j>\bar{l}_{N-1}^\infty$ for all $n\geqq N$ and all j. We then have that the orbit segment $\bar{z}^{\wedge}(x, i\beta)$ on $x_1\leqq x\leqq x_*(\mu, L)$ lies uniformly in the region $\{\operatorname{Im} z\geqq\varepsilon\}^{\wedge}$ for some $\varepsilon>0$. It once again follows that $\hat{z}(0, i\beta, L)\not\in\{0, \hat{\infty}\}$ for large μ.

We next show that there is an exact imaginary eigenvalue $i\beta_n^j$ at an interval size L_n^j near $\bar{L}_n^j$, where $\bar{L}_n^j$ is determined by the equation $\bar{l}_n^j=x_*(\mu, \bar{L}_n^j)$. To this end, note that for fixed $\beta\in[\beta_n^\circ-\delta, \beta_n^\circ+\delta]$, $\dot{\bar{z}}(\bar{l}(\beta), i\beta)=\bar{G}+\mathcal{O}(z^2)$, so that $\bar{z}$ crosses $\operatorname{Im} z=0$ transversally at $x=\bar{l}(\beta)$. It follows from Lemma 4.9 that for such β there exists a unique $L=L(\beta)$ such that $\operatorname{Im} z(0, i\beta, L(\beta))=0$. Furthermore, we have by Lemma 4.9 that $|x_*(\mu, L(\beta))-\bar{l}(\beta)|\leqq K\log\mu/\mu$. Now let $\beta_\pm=\beta_n^\circ\pm\delta$; since $\bar{Z}'(\beta_n^\circ)<0$ we have that there exists $c>0$ such that

$$\bar{Z}(\beta_+)<-c\delta<c\delta<\bar{Z}(\beta_-). \tag{4.44}$$

For $\beta=\beta_-$ and β_+ we have that

$$|z(0, i\beta, L(\beta))-\bar{z}(x_*(\mu, L(\beta), i\beta))|<K\log\mu/\mu.$$

Since $x_*(\mu, L(\beta))=\bar{l}(\beta)+\mathcal{O}(\log\mu/\mu)$ we have from (4.44) that

$$\operatorname{Re}\bar{z}(x_*(\mu, L(\beta_+), i\beta_+))<0<\operatorname{Re}\bar{z}(x_*(\mu, L(\beta_-), i\beta_-)).$$

Combining the last two inequalities we see that $z(0, i\beta, L(\beta))$ is positive at $\beta=\beta_+$ and negative at $\beta=\beta_-$ for large μ. Thus there exists $\beta_n^\circ\in[\beta_-, \beta_+]$ such that $z(0, i\beta_n^\circ, L(\beta_n^\circ))=0$. The critical interval size is therefore $L_n^\circ=L(\beta_n^\circ)$. The argument for β near $\beta_n^{-\infty}$ is similar, only now we have to use the variable $\zeta=G/z$ rather than w in order to get the transverse crossing of $\operatorname{Im}\zeta=0$; with this modification, the proof is the same as for $\beta=\bar{\beta}_n^\circ$.

In order to complete the proof we need to show that $Z'(\beta_n^\circ)>0$ and that $W'(\beta_n^\infty)<0$; this will imply that $i\beta_n^j$ is the unique exact imaginary eigenvalue for $|\beta-\beta_n^{-j}|\leqq\delta$; combining this with our previous result that $i\beta$ is not an eigenvalue for $|\beta-\beta_n^j|\geqq\delta$, $n\leqq N-1$, and $L\leqq L_o$, we see that $i\beta_n^j$ are the only imaginary eigenvalues that occur for $L\leqq L_o$. Finally, we need to show that if $\lambda(L)$ is a branch of eigenvalues with $\lambda(L_n^j)=i\beta_n^j$, then $\operatorname{Re}\lambda'(L_n^j)>0$.

The crucial ingredient needed in proving the above assertions will be a description of the dependence of $z(x, \lambda, L)$ on the parameters λ and L. To this end, let

$$\Delta(x, \lambda, L)=z(x, \lambda, L)-\bar{z}(x+x_*(\mu, L), \lambda),$$

$$\delta(x, \lambda, L)=\zeta(x, \lambda, L)-\zeta(x+x_*(\mu, L), \lambda),$$

where $\zeta=\bar{G}/z$ and $\zeta=\bar{G}/\bar{z}$. We then have the following lemma.

LEMMA 4.10. *Given $q > 0$ sufficiently large in Theorem 4.1 there exists $p > 0$ and $K > 0$ such that for $\lambda \in \mathrm{IV}$,*

(i) $|\Delta_\lambda| \leqq K(\mu^{-p} + \mu^{-1/2})$ *for* $|x| \leqq x_*(\mu, L)$, *whenever* $|\bar{z}| \leqq 2$, *and* $|\delta_\lambda| \leqq K(\mu^{-p} + \mu^{-1/2})$ *for* $|x| \leqq x_*(\mu, L)$ *whenever* $|\zeta| \leqq 2$.

(ii) $|\Delta_L| \leqq K\mu^{-1/2}$ *for* $|x| \leqq x_*(\mu, L)$ *whenever* $|\bar{z}| \leqq 2$ *and* $|\delta_L| \leqq K\mu^{-1/2}$ *for* $|x| \leqq x_*(\mu, L)$ *whenever* $|\zeta| \leqq 2$.

Proof. (i) We extend $\bar{z}(0, \lambda)$ analytically by defining $\bar{z}(0, \lambda) = \sqrt{\lambda}$, since $\Gamma\beta^{1/2} = \sqrt{i\beta}$. Thus $\bar{z}\lambda = \lambda^{-1/2} = \mathcal{O}(\mu^{-1/2})$. Next, we estimate z_λ at $x = x_*$. Note that $G_\lambda = 1 + \mathcal{O}(\mu^{-2})$; the equation for w_λ is

$$\dot{w}_\lambda = -G_\lambda w^2 - Gww_\lambda, \qquad w_\lambda(-L, \lambda, L) \equiv 0.$$

Since G is approximately $i\beta$ and w is approximately $\bar{\Gamma}\beta^{-1/2}$ on $[-L, -x_\gamma]$, Re Gw is positive and of order $\beta^{1/2}$. It follows that $|w_\lambda| \leqq K\mu^{-3/2}$ on $[-L, -x_\gamma]$, so that $|z_\lambda| \leqq K\mu^{-1/2}$ for such x. On $[-x_\gamma, -x_*]$ the change in z_λ is of order $\log \mu/\mu$, so that $|z_\lambda| \leqq K\mu^{-1/2}$ at $-x_*$. We therefore have that $|\Delta_\lambda| \leqq K\mu^{-1/2}$ at $x = -x_*$. This also yields the estimate $|\delta_\lambda| \leqq K\mu^{-3/2}$ at $x = -x_*$.

In order to obtain the desired estimate on the interior layer $|x| \leqq x_*(\mu, L)$, we use the equation

$$\dot{\delta}_\lambda = G_\lambda - \bar{G}_\lambda - 2[\zeta\delta_\lambda + \bar{\zeta}_\lambda\delta]$$

for δ whenever $|\bar{\zeta}| \leqq 2$, and the equation

$$\dot{\Delta}_\lambda = G_\lambda - \bar{G}_\lambda - 2[z\Delta_\lambda + \bar{z}_\lambda\Delta]$$

whenever $|\bar{z}| \leqq 2$ is in the proof of Lemma 4.9. By Theorem 4.1 we have that $|G_\lambda - \bar{G}_\lambda| \leqq K\mu^{-(q-7)/2}\mu^{-2} = K\mu^{-(q-3)/2}$. Combining this with our initial estimate $|\delta_\lambda| \leqq K\mu^{-3/2}$ at $-x_*$, together with Lemma 4.9 to estimate δ itself, we obtain the desired estimate for Δ_λ and δ_λ as in Lemma 4.9.

(ii) We first need to estimate z_L on the outer layer $[-L, -x_\gamma]$. The equation for w_L is

$$\dot{w}_L = G_L - 2ww_L, \qquad w_L(-L, \lambda, L) = 0.$$

Clearly, G_L is determined by U_L and V_L (see (4.15)) and on $[-L, -x_\gamma]$, $V \equiv 0$, so that G_L is determined by U_L here. It is not difficult to check using the time map (4.5) that $U_L = \mathcal{O}(1)$ on $[-L, -x_\gamma]$. Now, by the outer estimate of Lemma 4.7, we have that $|w - \bar{\Gamma}\beta^{-1/2}| \leqq K\mu^{-3/2}$; using this to approximate w in the equation for w_L we obtain the estimate $|w| \leqq K\mu^{-3/2}$ by the usual Gronwall estimate. Since $z_L = -w_L/w^2$ we obtain $|z_L| \leqq K\mu^{-1/2}$.

In the interval $[-x, x_*]$ the partial $V_L = \mu^2 U_L$ also needs to be estimated. In order to estimate U_L here, we use the time map

$$x - x_\gamma = \int_\gamma^U [C - R(s, \mu)]^{-1/2}\, ds,$$

where $C = C(L) = R(\gamma, \mu) + \mu^{-2}g(L, \mu)$ and $|g_L|$ is bounded. Differentiation with respect to L yields

$$U_L = \frac{1}{2}[C - R(U, \mu)]^{1/2}\mu^{-2}g_L \int_\gamma^U [C - R(s, \mu)]^{-3/2}\, ds.$$

On the interval $[-x, -x_*]$, $|U - \gamma| \leqq K\mu^{-2}$; it then follows that

$$|C - R(U, \mu)| = \mathcal{O}(1)|\mu^{-2} + \mu^2(U - \gamma)|.$$

Using this in the above expression for U_L yields the estimate $|U_L| \leqq K\mu^{-3}$, so that $|V_L| \leqq K\mu^{-1}$ and $|G_L| \leqq K\mu^{-1}$.

Using the above estimate for $|G_L|$ in $[-x_\gamma, x_*]$ in the equation

$$\dot{z}_L = G_L - 2zz_L,$$

together with our previous estimate for z_L at x_γ and the estimate of Lemma 4.7 for z on $[-x_\gamma, -x_*]$, we obtain

$$\begin{aligned} |z_L(-x_*, \lambda, L)| &\leqq K\mu^{-1/2} + K \log \mu/\mu^2 \\ &= K\mu^{-1/2}. \end{aligned}$$

Now $\bar{z}(0, \lambda)$ is independent of L, so that the above estimate for z_L at $-x_*$ yields the estimate $|\Delta_L| \leqq K\mu^{-1/2}$ and $|\delta_L| \leqq K\mu^{-3/2}$ at $x = -x_*$.

The equations for Δ_L are

$$\begin{aligned} \dot{\Delta}_L &= G_L - 2z\Delta_L - 2\bar{z}_L\Delta, \\ \dot{\delta}_L &= G_L - 2\zeta\delta_L - 2\bar{\zeta}_L\delta, \end{aligned}$$

since $\bar{G}_L \equiv 0$. The estimates for $|\Delta_L|$ and $|\delta_L|$ on $|x| \leqq x_*(\mu, L)$ now follow from the estimate for δ_L at $x = -x_*(\mu, L)$ together with the estimates of Lemma 4.9 for Δ and $\delta = (w - \bar{w})/\bar{G}$. In particular, as in Lemma 4.9, we estimate $|\delta_L|$ when $|\bar{\zeta}| \leqq 2$ and $|\Delta_L|$ when $|\bar{z}| \leqq 2$. The details are similar to previous arguments and will be omitted.

We can now complete the proof of Theorem 4.5. We first note that for β near $\bar{\beta}_n^\circ$, evaluation of (i) of Lemma 4.9 at $x = 0$ and $L = L(\beta)$, together with the defining conditions

$$\operatorname{Im} \bar{z}(\bar{l}(\beta), i\beta) = 0, \qquad \operatorname{Im} z(0, i\beta, L(\beta)) = 0$$

for $\bar{l}(\beta)$ and $L(\beta)$ leads to the estimate

$$|\bar{l}(\beta) - x_*(\mu, L(\beta))| \leqq K \log \mu/\mu.$$

The estimate in (ii) of Lemma 4.9 leads to a similar estimate for β near $\bar{\beta}_n^\infty$. The equation for $L(\beta)$ leads to the formulae

$$L'(\beta) = -\operatorname{Im} iz_\beta/\operatorname{Im} z_L, \qquad \bar{l}'(\beta) = -\operatorname{Im} i\bar{z}_\beta/\operatorname{Im} \dot{\bar{z}}.$$

Lemma 4.10 provides the estimate

$$\begin{aligned} z_\beta(0, i\beta, L) &= \bar{z}_\beta(x_*(\mu, L), i\beta) + \mathcal{O}(\mu^{-1/2}), \\ z_L(0, i\beta, L) &= \dot{\bar{z}}(x_*(\mu, L), i\beta)x_{*L} + \mathcal{O}(\mu^{-1/2}). \end{aligned}$$

Since $x_{*L} = 1 + \mathcal{O}(\mu^{-1})$ we have from the above and our previous estimate for $|\bar{l} - x_*|$ that

$$\begin{aligned} z_\beta(0, i\beta, L) &= \bar{z}_\beta(\bar{l}(\beta), i\beta) + \mathcal{O}(\mu^{-1/2}), \\ z_L(0, i\beta, L) &= \dot{\bar{z}}(\bar{l}(\beta), i\beta) + \mathcal{O}(\mu^{-1/2}). \end{aligned}$$

Using these in the above we obtain the estimate

$$|L'(\beta) - \bar{l}'(\beta)| \leqq K\mu^{-1/2}.$$

We had previously shown that $\bar{l}'(\bar{\beta}_n^\circ) > 0$, from which we obtain $L'(\beta_n^\circ) > 0$. This finally leads to the estimate

$$\begin{aligned} \operatorname{Re}(iz_\beta + z_L L'(\beta)) &= \operatorname{Re}(i\bar{z}_\beta + \dot{\bar{z}}\bar{l}'(\beta)) + \mathcal{O}(\mu^{-1/2}) \\ &= \bar{Z}'(\beta) + \mathcal{O}(\mu^{-1/2}). \end{aligned}$$

We had shown earlier that $\bar{Z}'(\bar{\beta}_n^\circ)<0$; it then follows that $Z(\beta)=z(0, i\beta, L(\beta))$ has a negative derivative at $\beta=\beta_n^\circ$. The proof that $\zeta(0, i\beta, L(\beta))$ has a positive derivative near β_n^∞ is the same as the above for $\beta=\beta_n^\circ$. We have therefore established that the only purely imaginary eigenvalues that occur for $L\leqq L_o$ are $i\beta^j_n$, which occurs at a uniquely determined interval size $L=L_n^j$.

It only remains to show that there is a branch $\lambda(L)$ of eigenvalues for L near L_n^j with $\lambda(L_n^j)=i\beta_n^\circ$ for which $\operatorname{Re}\lambda'(L_n^j)>0$. We again provide the argument only for $j=0$; we therefore need to solve the equation $z(0, \lambda, L)=0$. We have that a solution exists for $(\lambda, L)=(i\beta_n^\circ, L_n^\circ)$, and that if $\lambda(L)$ is such a branch, then

$$z_\lambda\lambda'(L)+z_L=0;$$

by Lemma 4.10 we then obtain

$$\lambda'(L)=-z_L/z_\lambda=-\dot{\bar{z}}/\bar{z}_\lambda+\mathcal{O}(\mu^{-1/2})$$

where we have used that $z_L=\dot{\bar{z}}+\mathcal{O}(\mu^{-1})$. Since $\bar{z}_\lambda=\bar{G}_\lambda l/2\neq 0$, $\lambda(L)$ exists by the implicit function theorem. Moreover, we have already shown that

$$\operatorname{Re}(-\dot{\bar{z}}/\bar{z}_\lambda)>0,$$

which completes the proof of Theorem 4.5.

5. Concluding remarks. We conclude by combining the results of §§ 3 and 4 to obtain some rigorous results on the stability and Hopf bifurcation of the perturbed problem with $\varepsilon>0$. Recall that by Lemma 4.2, Case S (respectively, Case U) holds if $\gamma>\sqrt{ab}$ (respectively, $\gamma<\sqrt{ab}$) and μ is sufficiently large.

THEOREM 5.1. *Suppose that Case S holds and that μ is large. Then there exists $\varepsilon_o>0$ such that the perturbed solution $(U(x, \varepsilon), V(x, \varepsilon))$ of* (1.3) *is a stable solution of* (1.1) *for $0<\varepsilon<\varepsilon_0$ for all $L>L_h$.*

Proof. By Theorem 3.11 it suffices to show that (1.4) has no eigenvalues inside K in Fig. 3.1. Theorem 4.4 implies that (1.4) has no eigenvalues in $\operatorname{Re}\lambda\geqq 0$ in Case S. Since the spectrum of (1.4) is discrete, (1.4) will have no spectrum inside K if the angle ϕ in Fig. 3.1 is small enough.

THEOREM 5.2. *Suppose that Case U holds and that μ is large. Let the angle ϕ in Fig.* 3.1 *be $\phi=0$, so that $c_1(\mathscr{E}(K, \varepsilon))$ measures the number of eigenvalues of $\mathscr{L}$ in* $\operatorname{Re}\lambda\geqq 0$. *Then there exists $\varepsilon_o>0$ and interval sizes $L_n^j(\varepsilon)$ which approach L_n^j as $\varepsilon\to 0$ such that for $0<\varepsilon<\varepsilon_o$,*

(i) *$(U(x, \varepsilon), V(x, \varepsilon))$ is stable for $L<L_o^\circ(\varepsilon)$;*

(ii) *For L near $L_n^j(\varepsilon)$, the perturbed operator $\mathscr{L}$ admits a branch of eigenvalues $\lambda_\varepsilon(L)$ such that $\lambda_\varepsilon(L_n^j(\varepsilon))=i\beta_n^j(\varepsilon)$ for some $\beta_n^j(\varepsilon)$ near β_n^j, and* $\operatorname{Re}\lambda_\varepsilon'(L_n^j(\varepsilon))>0$. *Hence a Hopf bifurcation occurs as L crosses each $L_n^j(\varepsilon)$.*

(iii) *Let $N>0$ be given. The stability index $c_1(\mathscr{E}(K, \varepsilon))$ is* $2(2n+1)$ *for $L_n^\circ(\varepsilon)<L<L_n^\infty(\varepsilon)$ and* $2(2n+2)$ *for $L_n^\infty(\varepsilon)<L<L_{n+1}^\circ(\varepsilon)$, for all $n\leqq N$ and all sufficiently small $\varepsilon>0$.*

Proof. For $L<L_o^\circ(\varepsilon)$ the stability proof is the same as in Case S. In order to prove (ii) consider L near L_n^j and λ near $i\beta^j_n$. Let C be a small circle about $i\beta^j_n$, so that for L near L_n^j, C is disjoint from σ_R, and for each such L, there is a unique eigenvalue $\lambda_*(L)\in\sigma_R$ inside C, so that $\lambda_*(L_n^j)=i\beta_n^j$ and $\operatorname{Re}\lambda_*'(L_n^j)>0$. For such C we may therefore form the bundles $\mathscr{E}_{2*}(C)$ and $\tilde{\mathscr{E}}_{2*}(C)$ for the reduced problem as in § 3.11; for each L near L_n^j, $c_1(\mathscr{E}_{2*}(C))=c_1(\tilde{\mathscr{E}}_{2*}(C))=1$. Using the same procedure as was used to decompose $\mathscr{E}(K, \varepsilon)$, we form the perturbed bundle $\mathscr{E}(C, \varepsilon)$ and decompose it into a fast and slow summand

$$\mathscr{E}(C, \varepsilon)=\mathscr{E}_1(C, \varepsilon)\oplus\mathscr{E}_2(C, \varepsilon),$$

where, as in the case of K, $c_1(\mathscr{E}_1(C, \varepsilon))=0$, and the fibers of the slow summand $\mathscr{E}_2(C, \varepsilon)$ are uniformly near those of $\mathscr{E}_{2*}(C)$ over the whole base space. Now each bundle $\mathscr{E}_2(C, \varepsilon)$ and $\mathscr{E}_{2*}(C)$ depends on the parameter L; by rescaling x and introducing L as a parameter into the equations, it is easily seen that the solutions, and hence the bundles, depend analytically on L. This is seen by employing standard theorems on the analytic dependence of solutions of o.d.e.'s on parameters. Each bundle has its own Evans function, which we shall denote by $D^\varepsilon(\lambda, L)$ and $D^*(\lambda, L)$, respectively. By the previous remark, both D and D^* depend analytically on both λ and L. Since $\mathscr{E}_2$ is approximated by $\mathscr{E}_{2*}$, it follows that $D^\varepsilon(\lambda, L)$ approaches $D^*(\lambda, L)$ for fixed L and $\lambda \in C$. By the Cauchy integral formula, they are therefore also close over the interior of C. By Rouche's theorem it follows that D and D^* have the same number of roots inside C counting multiplicity, so that $D^\varepsilon(\lambda, L)$ has a simple root $\lambda_\varepsilon(L)$ for each L. Now $\operatorname{Re} \lambda'_*(L^j_n)>0$ so that $\operatorname{Re} \lambda_*(L)$ is positive (respectively, negative) for $L_+ > L^j_n + \delta$ (respectively, $L_- < L^j_n - \delta$) for fixed, small $\delta > 0$. Since $D^\varepsilon(\lambda, L)$ uniformly approximates $D^*(\lambda, L)$ for λ inside C, it follows that

$$\operatorname{Re} \lambda^\varepsilon(L_-) < 0 < \operatorname{Re} \lambda^\varepsilon(L_+)$$

so that $\lambda^\varepsilon(L) = i\beta^j_n(\varepsilon)$ at some $L = L^j_n(\varepsilon)$ near L^j_n. Again using the analyticity of D^ε and D^* in both λ and L, it follows from the Cauchy integral formula and the uniform approximation of D^ε by D^* that D^ε_λ and D^ε_L approach D^*_λ and D^*_L as $\varepsilon \to 0$. Hence

$$\begin{aligned}\lim_{\varepsilon\to 0} \operatorname{Re} \lambda'_\varepsilon(L^j_n(\varepsilon)) &= \lim_{\varepsilon\to 0} \operatorname{Re} -D^\varepsilon_\lambda / D^\varepsilon_L \\ &= \operatorname{Re} \lambda'_*(L^j_n) > 0,\end{aligned}$$

so that $\operatorname{Re} \lambda'_\varepsilon(L^j_n(\varepsilon)) > 0$ for small $\varepsilon > 0$.

We can now complete the proof. Let $\delta > 0$ be so small that $\operatorname{Re} \lambda_*(L^j_n \pm \delta) \neq 0$ for all n, j with $n \leqq N-1$. If L is such that $|L - L^j_n| \geqq \delta$ for all n, j then σ_R has no imaginary eigenvalues for this L. We therefore have by Theorem 3.11 that the eigenvalue count for the perturbed operator $\mathscr{L}$ inside K is given by that of the reduced problem (1.4), which, by Theorem 4.5, is as specified in (iii) of Theorem 5.2. Finally, for $|L - L^j_n| \leqq \delta$ we see from the previous paragraph concerning the behavior of the branch $\lambda_\varepsilon(L)$ inside C that the eigenvalue count inside K increases precisely by 2 as L crosses $L^j_n(\varepsilon)$. More precisely, we form the curve $K_1 = \partial(K^\circ \cup C^\circ \cup \bar{C}^\circ)$, where $\bar{C}^\circ$ is the complex conjugate of C°. By previous remarks, we may form the bundle $\mathscr{E}_2(K_1, \varepsilon)$, which will be well defined for L near L^j_n. If $K_2 = \partial(K^\circ \backslash (C^\circ \cup \bar{C}^\circ))$ then K_2 winds about all the eigenvalues with strictly positive real part, while C (respectively, $\bar{C}$) winds about $i\beta^j_n(\varepsilon)$ (respectively, $-i\beta^j_n(\varepsilon)$). We then have that

$$c_1(\mathscr{E}_2(K_1, \varepsilon)) = c_1(\mathscr{E}_2(K_2, \varepsilon)) + c_1(\mathscr{E}_2(C, \varepsilon)) + c_1(\mathscr{E}_2(\bar{C}, \varepsilon));$$

hence the eigenvalue count is precisely as specified in (iii) for all L.

We finally remark that the direction of bifurcation has not been determined. It seems likely, however, that the asymptotic methods of § 4 could be used to compute the higher-order asymptotics of the bifurcating solutions needed in such a computation. We conjecture that each Hopf bifurcation is supercritical, so that the first one at $L = L^\circ_o(\varepsilon)$ gives rise to a stable periodic solution.

REFERENCES

[1] J. ALEXANDER, R. GARDNER, AND C. JONES, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167-212.

[2] E. CONWAY, R. GARDNER, AND J. SMOLLER, *Stability and bifurcation of steady state solutions of predator-prey equations*, Adv. in Appl. Math., 3 (1982), pp. 288-334.

[3] E. N. DANCER, *On positive solutions of some pairs of differential equations* II, J. Differential Equations, 60 (1985), pp. 236-258.

[4] ———, *On uniqueness and stability for solutions of singularly perturbed predator-prey type equations with diffusion*, preprint.

[5] N. FENICHEL, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193-226.

[6] H. FUJII AND Y. NISHIURA, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 481-514.

[7] R. GARDNER AND C. JONES, *A stability index for steady state solutions of boundary value problems*, J. Differential Equations, 91 (1990), pp. 181-203.

[8] ———, *Stability of travelling wave solutions of diffusive predator-prey systems*, Trans. Amer. Math. Soc., to appear.

[9] ———, *Travelling waves of a perturbed diffusion equation arising in a phase-field model*, Indiana Univ. Math. J., 38 (1991), pp. 1197-1222.

[10] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[11] L. LI, *Coexistence theorems for steady states for predator-prey interacting systems*, Trans. Amer. Math. Soc., 305 (1988), pp. 143-165.

[12] L. LI AND M. R. LLOYD, *A numerical behavior of positive solutions to elliptic predator-prey systems over large regions*, Differential Equations and Appl., to appear.

[13] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[14] D. TERMAN, *Stability of planar wave solutions of a combustion model*, SIAM J. Math. Anal., 21 (1990), pp. 1139-1172.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 150–180, January 1992

008

MULTIPARAMETER BIFURCATION OF A PREDATOR-PREY SYSTEM*

J. HAINZL†

Abstract. For a two-dimensional predator-prey system, proposed by Bazykin and depending on several parameters, a complete local bifurcation analysis with respect to all parameters is achieved. The major part of the paper is devoted to the unfolding of a degenerate codimension-2 bifurcation occurring for a one-dimensional subset of parameters. The main problem here consists in studying parameter dependent integrals which are not algebraic.

Key words. predator-prey system, bifurcation, unfolding, stability, parameter dependent integral, limit cycle, homoclinic orbit

AMS(MOS) subject classifications. 34C05, 34C25

1. Introduction.

The starting point of this paper is the predator-prey system

$$
\begin{aligned}
\dot{x}_1 &= \alpha_1 x_1 - \frac{\alpha_{12} x_1 x_2}{1+\beta_1 x_1} - \alpha_{11} x_1^2, \\
\dot{x}_2 &= -\alpha_2 x_2 + \frac{\alpha_{21} x_1 x_2}{1+\beta_1 x_1} - \alpha_{22} x_2^2
\end{aligned}
\tag{1}
$$

with $x_1(t), x_2(t)$ the prey and predator populations, and positive parameters $\alpha_i, \alpha_{ij}, \beta_1$. This system was first proposed by Bazykin in [2], [3], where a detailed discussion concerning ecological motivation and possible behavior of solutions is presented. A rigorous treatment of (1) has been carried out in [9]. There, besides some results on Hopf bifurcation, we mainly aimed at basic questions such as the location and stability of equilibria and the existence of periodic orbits. In the present paper, this former investigation is supplemented by a fairly complete bifurcation analysis of (1) depending on all the parameters. Apart from a regular codimension-2 bifurcation which occurs in a two-dimensional parameter region, we are primarily concerned with the unfolding of a degenerate codimension-2 bifurcation occurring in a one-dimensional parameter region. This interesting bifurcation phenomenon can be observed at parameter values where three equilibria coalesce to a single one (cusp point).

The paper is organized as follows. In §2 we first summarize those facts from [9] that are needed in the present context. We introduce the manifold $\mathcal{M}$ in the parameter space, describing the set of equilibria of (1); it can be reduced to a cusp surface with two folds C_+, C_- and a cusp point C_0. Then we characterize those parameters for which codimension-2 bifurcation occurs on $C_+ \cup C_-$ (regular case) or at C_0 (degenerate case). A transformation defined in §3 shows that the corresponding linearization is of nilpotent type $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. In §4, the regular case is treated and, using known methods, a complete unfolding of this bifurcation problem is achieved. The subsequent §§5–12 are entirely devoted to the degenerate case. The main task here is to investigate the quotient of parameter dependent integrals which, contrary to the cases found in the literature, are not algebraic. Various analytical results, partly supported by numerical

*Received by the editors September 7, 1990; accepted for publication March 20, 1991.

†Fachbereich Mathematik, Universität Kassel, Heinrich-Plett-Straße 40, D-3500 Kassel, West Germany.

calculations, finally yield the complete unfolding of this degenerate bifurcation, which is presented in §12. Some remarks conclude the paper.

2. Preliminaries and auxiliary results. Let us first summarize some simplifications and results from [9]. By rescaling x_1, x_2, t suitably, the parameters α_1, α_{12}, and β_1 in (1) adopt the value 1, and without restriction, we may assume $\alpha_{21} > \alpha_2$. In [9] it turned out to be useful to introduce instead of $x_1, x_2, \alpha_2, \alpha_{11}, \alpha_{21}, \alpha_{22}$ the variables x, y and parameters $\alpha, \beta, \gamma, \delta$ defined by

$$(2)\qquad \begin{aligned} x &:= \alpha_{21}^{-1}(1+x_1), & y &:= \alpha_{21}^{-2}x_2\,, \\ \alpha &:= \alpha_{21}^{-1}, & \beta &:= (\alpha_{21}-\alpha_2)^{-1} \\ \gamma &:= \alpha_{11}^{-1}\alpha_{21}^{-1}(1+\alpha_{11}), & \delta &:= \alpha_{22}^{-1}\alpha_{21}^{-2}(\alpha_{21}-\alpha_2)\,. \end{aligned}$$

System (1) then turns into

$$(3)\qquad \begin{aligned} \dot{x} &= -(x-\alpha)\cdot\left[\frac{x-\gamma}{\gamma-\alpha}+\frac{y}{\alpha x}\right], \\ \dot{y} &= -\frac{y}{\beta\delta}\left(y-\delta+\frac{\beta\delta}{x}\right), \end{aligned} \qquad x \geq \alpha, y \geq 0$$

and, as justified in [9], the variation of the remaining four parameters may be restricted to

$$(4)\qquad 0<\alpha<\beta<\gamma\,, \qquad \delta>0\,.$$

The only equilibria of (3) are two trivial saddles (α,0) and (γ,0), which are ignored throughout the paper, and the intersection points of the two curves

$$(5)\qquad y=\frac{\alpha x(\gamma-x)}{\gamma-\alpha} \text{ (parabola) and } y=\delta-\frac{\beta\delta}{x} \text{ (hyperbola).}$$

Depending on the parameters there exist at least one and at most three such equilibria ($\overline{x}$,$\overline{y}$). Introducing the variable ζ instead of $\overline{x}$, the dependence of these equilibria

$$(6)\qquad (\overline{x},\overline{y}) = (\zeta\,, \alpha(\gamma-\alpha)^{-1}\zeta(\gamma-\zeta))$$

on the parameters is best characterized by identifying these equilibria with the points of a four-dimensional manifold $\mathcal{M}$ in $(\alpha,\beta,\gamma,\delta,\zeta)$-space, defined by the equation

$$(7)\qquad \delta=\alpha\zeta^2(\gamma-\zeta)(\gamma-\alpha)^{-1}(\zeta-\beta)^{-1}\,, \qquad 0<\alpha<\beta<\zeta<\gamma\,.$$

(Note that (7) cannot be solved for ζ uniquely. Thus, for convenience, we shall consider $(\alpha,\beta,\gamma,\zeta)$ instead of $(\alpha,\beta,\gamma,\delta)$ as parameters.) For any fixed (α,β), $\mathcal{M}$ reduces to a two-dimensional surface of cusp type in (γ,δ,ζ)- space, with upper fold C_+, lower fold C_-, and cusp point C_0. A qualitative picture of this surface is given in Fig. 1.

In [9] it was shown that

$$(8)\qquad C_0 = (\overline{\gamma},\overline{\delta},\overline{\zeta}) := (9\beta, 27\alpha\beta^2(9\beta-\alpha)^{-1}, 3\beta)\,;$$

moreover, for increasing $\gamma \in (9\beta,\infty)$, we have

(9) On wC_-, ζ is strictly decreasing from 3β to 2β,
(10) On wC_+, $\frac{\zeta}{\gamma}$ is strictly increasing from $\frac{1}{3}$ to $\frac{1}{2}$.

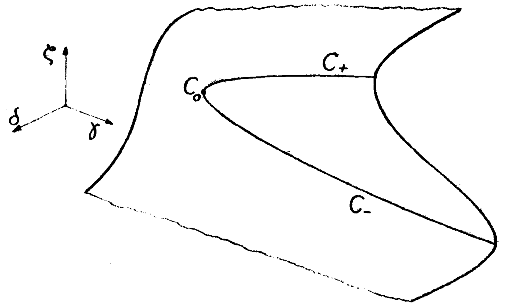


FIG. 1. *The surface $\mathcal{M}$.*

In [9], the stability of the equilibria represented by (7) has been investigated via linearization. Denoting the corresponding characteristic equation by

$$\lambda^2 + p\lambda + q = 0 , \tag{11}$$

we obtain for p, q the following functions on $\mathcal{M}$:

$$p = \frac{1}{\beta} - \frac{1}{\zeta} + \frac{\alpha}{\gamma - \alpha}\left(\frac{\gamma}{\zeta} + \frac{2\zeta}{\alpha} - \frac{\gamma}{\alpha} - 2\right) , \tag{12}$$

$$q = \frac{\alpha}{\gamma - \alpha}\left(\frac{1}{\beta} - \frac{1}{\zeta}\right)\left(\frac{\gamma}{\zeta} + \frac{2\zeta}{\alpha} - \frac{\gamma}{\alpha} - 2\right) + \left(\frac{1}{\alpha} - \frac{1}{\zeta}\right) \cdot \frac{\alpha(\gamma - \zeta)}{\zeta(\gamma - \alpha)} . \tag{13}$$

Furthermore, we know that $q = 0$ exactly on $C_- \cup C_0 \cup C_+, q < 0$ on the middle sheet of $\mathcal{M}$ (corresponding to saddles) and $q > 0$ on the remainder of $\mathcal{M}$. In the points of $C_+ \cup C_-$, saddle-node bifurcation occurs (with δ as the varying parameter). Moreover, the points on $\mathcal{M}$ yielding $p = 0, q > 0$ were investigated for Hopf bifurcation in [9].

In the present paper, we are interested in bifurcations occurring at points where both p and q vanish, corresponding to equilibria with double eigenvalue zero. Such points must be located on the curve

$$\mathcal{C} := C_- \cup C_0 \cup C_+ . \tag{14}$$

By virtue of (8)–(10), $\mathcal{C}$ can be parametrized by $\zeta \in (2\beta, \infty)$. Since $q = 0$ on $\mathcal{C}$, we obtain from (13) (see [9]):

$$2\zeta^2 = \zeta(\gamma + 3\beta) - 2\beta\gamma \text{ on } \mathcal{C}. \tag{15}$$

The following lemma shows that, in a neighborhood of any point on $\mathcal{C}$, we may regard (p, q) instead of (γ, ζ) as parameters of the surface $\mathcal{M}$.

LEMMA 1. *For any fixed $(\alpha, \beta), 0 < \alpha < \beta$, we have*

$$\Delta := p_\gamma q_\zeta - p_\zeta q_\gamma > 0 \quad on\ \mathcal{C} . \tag{16}$$

Proof. From (12), (13) we calculate the partial derivatives

$$\begin{aligned} p_\gamma &= -\zeta^{-1}(\gamma - \alpha)^{-2}(2\zeta - \alpha)(\zeta - \alpha) , \\ p_\zeta &= \zeta^{-2}(\gamma - \alpha)^{-1}(\gamma - \alpha - \alpha\gamma + 2\zeta^2) , \\ q_\gamma &= -\zeta^{-2}\beta^{-1}(\gamma - \alpha)^{-2}(\zeta - \alpha)\left[\zeta(2\zeta - 3\beta - \alpha) + 2\alpha\beta\right] , \\ q_\zeta &= \beta^{-1}\zeta^{-3}(\gamma - \alpha)^{-1}\left[2\zeta^3 - \zeta(\alpha\gamma + 3\alpha\beta + 2\beta\gamma) + 4\alpha\beta\gamma\right] . \end{aligned} \tag{17}$$

Then, using (15) repeatedly, a lengthy calculation yields

$$\Delta = \zeta^{-2}(\gamma-\alpha)^{-3}(\zeta-2\beta)^{-1}\{2(\zeta-\beta)^3 + (\beta-\alpha)\underbrace{[\zeta(6\zeta-6\beta-5\alpha)+2\beta^2+\alpha^2]}_{=:A}\}$$
$$+\beta^{-1}\zeta^{-4}(\gamma-\alpha)^{-2}\{2(\zeta-\beta)^3+(\beta-\alpha)\underbrace{[3\zeta^2-\zeta(6\beta+\alpha)+2\beta^2+2\alpha\beta]}_{=:B}\}.$$

Since $\zeta > 2\beta$ on C, we obviously obtain $A > 0$; B is a quadratic polynomial in ζ whose zeros are less than 2β, whence $B > 0$, too. This yields $\Delta > 0$. □

Using this lemma, we can prove the following.

LEMMA 2. *On the curve C, p is strictly increasing from $(2\beta)^{-1}(1+\alpha-2\beta)$ to β^{-1}, as ζ runs through the interval $(2\beta,\infty)$.*

Proof. From (15) we obtain

$$\gamma =: \gamma(\zeta) = \zeta(\zeta-2\beta)^{-1}(2\zeta-3\beta) \quad \text{on } C. \tag{18}$$

Since (α,β) is fixed, p and q in (12), (13) are functions of (γ,ζ) only. Eliminating γ by (18), we thus have

$$p = p(\gamma(\zeta),\zeta) =: \tilde{p}(\zeta), \qquad q = q(\gamma(\zeta),\zeta) \equiv 0 \quad \text{on } C.$$

Differentiation with respect to ζ yields $p_\gamma \cdot \gamma' + p_\zeta = \tilde{p}', q_\gamma \cdot \gamma' + q_\zeta = 0$. From this and Lemma 1 we obtain by Cramer's rule $\tilde{p}'q_\gamma = -\Delta < 0$, and since $q_\gamma < 0$ (see (17)), we finally get $\tilde{p}'(\zeta) > 0$ for each $\zeta \in (2\beta,\infty)$. The remaining assertion follows easily from (12) and (18). □

LEMMA 3. *Let $0 < \alpha < \beta$ be arbitrarily fixed.*

(a) *If $2\beta \le 1+\alpha$, we have $p > 0$ everywhere on C.*

(b) *If $2\beta > 1+\alpha$, there is just one point $P \in C$ satisfying $p = 0$. Moreover, we have*

$$P\begin{cases} \in C_- \\ = C_0 \\ \in C_+ \end{cases} \quad \text{if } T(\alpha,\beta) := 9\beta^2 - 3\beta(\alpha+6) + 2\alpha \begin{cases} <0 \\ =0 \\ >0 \end{cases}.$$

(*See Fig.* 2.)

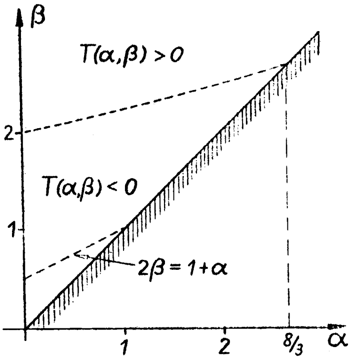


FIG. 2. *Illustration to Lemma* 3.

Proof. From (8), (12) we calculate

$$p = -\frac{1}{3\beta(9\beta-\alpha)}[9\beta^2 - 3\beta(\alpha+6)+2\alpha] \quad \text{in } C_0. \tag{19}$$

Now, all the assertions follow from Lemma 2. □

3. Another transformation. As was already mentioned above, we will consider only neighborhoods of points of $\mathcal{M}$ satisfying $p = q = 0$. By Lemma 3, such a point exists if and only if

$$2\beta > 1 + \alpha\,. \tag{20}$$

Therefore, in addition to the requirement $0 < \alpha < \beta$, we assume inequality (20) throughout the rest of the paper. Let $P = P(\alpha,\beta)$ denote the unique point of $\mathcal{M}$ yielding $p = q = 0$. Then the parameters (γ,ζ) corresponding to P satisfy equation (18) (from $q = 0$), and by exploiting $p = 0$, we obtain from (12) and (18) the additional condition

$$\alpha = \zeta\frac{N_2}{N_1} \quad \text{in } P = P(\alpha,\beta) \tag{21}$$

where

$$N_1 := \beta^2\zeta + 3\beta\zeta - \zeta^2 - 2\beta^2 = (\gamma-\alpha)^{-1}(\zeta-2\beta)^{-1}\zeta^2\beta^2(\zeta-\beta) > 0, \tag{22}$$

$$N_2 := \beta^2\zeta + 5\beta\zeta - 2\zeta^2 - 3\beta^2 > 0\,. \tag{23}$$

Now we consider parameters $(\alpha,\beta,\gamma,\zeta)$ such that (α,β) satisfies (20) and (γ,ζ) defines a point on the surface $\mathcal{M}$ which is sufficiently close to $P(\alpha,\beta)$. Then the parameter dependent affine transformation

$$x = \zeta + \left(\frac{1}{\alpha} - \frac{1}{\zeta}\right)X, \qquad y = \frac{\alpha\zeta(\gamma-\zeta)}{\gamma-\alpha} + \left(\frac{1}{\beta} - \frac{1}{\zeta} - p\right)X - Y \tag{24}$$

turns system (3) into the following system for $X(t), Y(t)$:

$$\dot{X} = Y + \Phi(X,Y)\,, \qquad \dot{Y} = -qX - pY + \Psi(X,Y)\,. \tag{25}$$

Here we calculate

$$\begin{aligned}\Phi(X,Y) &= \alpha\zeta(\zeta-\alpha)^{-1}[\Theta(x,y) - (\gamma-\alpha)^{-1}(x-\zeta)^2]\,,\\ \Psi(X,Y) &= \Theta(x,y)\cdot[1 + \alpha\beta^{-1}(\zeta-\alpha)^{-1}(\zeta-\beta) - \alpha\zeta(\zeta-\alpha)^{-1}p]\\ &\quad - \alpha(\zeta-\alpha)^{-1}(\gamma-\alpha)^{-1}[\beta^{-1}(\zeta-\beta) - \zeta p](x-\zeta)^2\\ &\quad + \alpha^{-1}\beta^{-1}\zeta^{-2}(\gamma-\zeta)^{-1}(\gamma-\alpha)(\zeta-\beta)\\ &\quad \times [y - \alpha\zeta(\gamma-\alpha)^{-1}(\gamma-\zeta)]^2\,,\end{aligned} \tag{26}$$

with

$$\Theta(x,y) := \zeta^{-1}(\gamma-\alpha)^{-1}x^{-1}(x-\zeta)[\alpha(\gamma-\zeta)(x-\zeta) - (\gamma-\alpha)y + \alpha\zeta(\gamma-\zeta)]\,.$$

On the right-hand side of (26), (x,y) has to be expressed in terms of (X,Y) according to (24). Φ and Ψ are at least quadratic with respect to (X,Y) and real-analytic (rational) functions of $X, Y, \alpha, \beta, \gamma, \zeta$; by virtue of Lemma 1 they are also real-analytic functions of $X, Y, \alpha, \beta, p, q$ for (p,q) sufficiently small.

The linearization of system (25) at the point $(X,Y,p,q) = (0{,}0{,}0{,}0)$ is $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. In what follows, we have to investigate (25) in a neighborhood of this point, which corresponds to $P(\alpha,\beta) \in \mathcal{M}$. For this purpose, we need the second and some higher-order derivatives of Φ,Ψ with respect to (X,Y), evaluated at $(X,Y,p,q) = (0{,}0{,}0{,}0)$. For simplicity, we do not explicitly express Φ,Ψ in terms of (p,q), but retain the parameters (γ,ζ) which then have to satisfy (18) and (21). In the following formulae,

γ and α have been eliminated by (18) and (21). The calculations yield (we omit the arguments)

$$\Psi_{XX} = -2N_2^{-1}\beta^{-3}\zeta^{-4}(\zeta-\beta)^4(\zeta-3\beta)\,, \tag{27}$$

$$\Psi_{XY} = -\beta^{-1}\zeta^{-3}N_2^{-1}(\zeta-\beta)[2(\zeta-\beta)^2-\beta^2\zeta]\,, \tag{28}$$

$$\Phi_{XX} = -2\beta^{-2}\zeta^{-3}N_2^{-1}(\zeta-\beta)^3(\zeta-2\beta)\,. \tag{29}$$

LEMMA 4. *Assume* (20), *and let* $P = P(\alpha,\beta)$ *be defined as above. Then we have* $\Phi_{XX} < 0$ *and* $\Psi_{XY} < 0$. *Moreover,* Ψ_{XX} *is* $> 0, = 0, < 0$, *whenever* $P \in C_-$, $P = C_0$, $P \in C_+$, *respectively.*

Proof. The assertions on Ψ_{XX} and Φ_{XX} follow from (8)–(10). It remains to show that $R := 2(\zeta-\beta)^2 - \beta^2\zeta$ is always positive. For $P = C_0$ we have $\zeta = 3\beta$; hence $R = \beta^2(8-3\beta)$. This is positive, since by (19)

$$\alpha = \frac{9\beta(\beta-2)}{3\beta-2} \quad \text{in the case where } P = C_0\,, \tag{30}$$

and thus, to obtain $0 < \alpha < \beta$, we necessarily have $2 < \beta < \frac{8}{3}$. Now, the admissible domain $0 < \alpha < \beta$, $1+\alpha < 2\beta$ in the (α,β)-plane is connected (see Fig. 2), hence it is enough to show that R is nowhere zero. Assuming $\beta^2\zeta = 2(\zeta-\beta)^2$ and inserting this into (21), we obtain $\alpha = \beta$, which is forbidden. □

Since $\Psi_{XX} \neq 0$ for $P \in C_- \cup C_+$, we shall not need any higher derivatives in this regular case. But the degeneracy $\Psi_{XX} = 0$ in case $P = C_0$ necessitates the calculation of some more and higher-order derivatives. Using (30) and $\zeta = 3\beta$ in (27)–(29) and also in calculating the additional derivatives written down below, we finally obtain the following derivatives.

Derivatives required in the degenerate case $P = C_0$:

$$\begin{aligned}
\Phi_{XX} &= -\frac{16}{81\beta^3(\beta-2)}, & \Phi_{XY} &= \frac{1}{9\beta^2}\,, \\
\Psi_{XY} &= -\frac{2(8-3\beta)}{81\beta^3(\beta-2)}, & \Psi_{YY} &= \frac{4}{27\beta^2(\beta-2)}\,, \\
\Phi_{XXX} &= -\frac{16}{3^5\beta^5(\beta-2)}, & \Psi_{XXX} &= -\frac{32(3\beta-4)}{3^7\beta^6(\beta-2)^2}\,, \\
\Psi_{XXY} &= -\frac{16(3\beta-4)}{3^7\beta^5(\beta-2)^2}, & \Psi_{XXXX} &= \frac{2^9(3\beta-4)}{3^{10}\beta^8(\beta-2)^3}\,,
\end{aligned} \qquad 2 < \beta < \tfrac{8}{3}\,. \tag{31}$$

4. The regular case $P \in C_- \cup C_+$**.** Throughout this section the parameters (α,β) are assumed to satisfy

$$0 < \alpha < \beta\,, \quad 1+\alpha < 2\beta\,, \quad \alpha \neq \frac{9\beta(\beta-2)}{3\beta-2} \tag{32}$$

(see Fig. 2) such that, by Lemma 3, the unique point $P = P(\alpha,\beta)$ is lying on $C_+ \cup C_-$. In a neighborhood of P, system (25) adopts the form

$$\begin{aligned}
\dot{X} &= Y + a_{11}X^2 + a_{12}XY + a_{22}Y^2 + \mathcal{O}_3\,, \\
\dot{Y} &= -qX - pY + b_{11}X^2 + b_{12}XY + b_{22}Y^2 + \mathcal{O}_3\,.
\end{aligned} \tag{33}$$

Here, the terms $\mathcal{O}_3$ are power series in (X, Y, p, q) with powers $X^iY^jp^kq^l$ satisfying $i+j+k+l \geq 3$ and $i+j \geq 2$, and with coefficients depending analytically on (α,β). a_{ij}, b_{ij} are given by Taylor's formula, e.g., $a_{11} = \frac{1}{2}\Phi_{XX}$, the derivatives being evaluated at $(X,Y,p,q) = (0,0,0,0)$. By Lemma 4, $b_{11} = \frac{1}{2}\Psi_{XX} \neq 0$; hence for $q \neq 0$ we may rescale in the following way:

$$\begin{aligned} &\varepsilon := |q b_{11}^{-1}|^{\frac{1}{2}}\,, \qquad p =: -|b_{11}|^{\frac{1}{2}}\varepsilon\tau\,, \\ &X =: \varepsilon^2 y_1\,, \qquad Y =: |b_{11}|^{\frac{1}{2}}\varepsilon^3 y_2\,, \\ &t =: |b_{11}|^{-\frac{1}{2}}\varepsilon^{-1}\widetilde{t}. \end{aligned} \tag{34}$$

For definiteness we now assume $q < 0$. (As for the case $q > 0$, see Remark 1 at the end of this section.) From (33), we then derive the following system for $y_1(\widetilde{t}), y_2(\widetilde{t})$ with small parameters ε, τ:

$$\begin{aligned} \dot{y}_1 &= y_2 + \varepsilon a_{11}|b_{11}|^{-\frac{1}{2}}y_1^2 + \mathcal{O}_4\,, \\ \dot{y}_2 &= y_1 + e y_1^2 + \tau y_2 + \varepsilon b_{12}|b_{11}|^{-\frac{1}{2}}y_1 y_2 + \mathcal{O}_4\,, \qquad e := \operatorname{sgn} b_{11}\,. \end{aligned} \tag{35}$$

Here, and in what follows, the terms $\mathcal{O}_4$ are analytic in all variables, at least of second order both with respect to (y_1, y_2) and (ε, τ). Such systems have been treated in the literature; see, e.g., [10]. To establish the standard situation we transform (y_1, y_2) into (U, V) in such a way that the two equilibria of (35), located at $(0,0)$ and near $(-e, 0)$, turn into $(U, V) = (0,0)$ and $(-e, 0)$ exactly. We can choose this transformation as follows:

$$\begin{aligned} U &= y_1 + e b_{12}|b_{11}|^{-\frac{1}{2}}\varepsilon y_1^2 y_2 + \mathcal{O}_4\,, \\ V &= y_2 + \varepsilon a_{11}|b_{11}|^{-\frac{1}{2}}y_1^2 + \mathcal{O}_4 \end{aligned} \tag{36}$$

and obtain the transformed system

$$\begin{aligned} \dot{U} &= V + \varepsilon e b_{12}|b_{11}|^{-\frac{1}{2}}U[2V^2 + U^2(1 + eU)] + \mathcal{O}_4\,, \\ \dot{V} &= U + eU^2 + \tau V + \varepsilon|b_{11}|^{-\frac{1}{2}}UV[2a_{11} + b_{12}(1 - eU - 2U^2)] + \mathcal{O}_4\,. \end{aligned} \tag{37}$$

The unperturbed system (37) with $\varepsilon = \tau = 0$

$$\dot{U} = V\,, \qquad \dot{V} = U + eU^2 \tag{38}$$

is Hamiltonian with the first integral

$$H(U,V) = \tfrac{1}{2}V^2 - \tfrac{1}{2}U^2 - \tfrac{1}{3}eU^3\,. \tag{39}$$

Thus for system (38), there exists a homoclinic orbit filled up with periodic orbits encircling $(-e, 0)$. To find out for which parameters (ε, τ) the perturbed system (37) still has periodic orbits or a homoclinic orbit, we use the method and results from [10]. Calculating the derivative $\dot{H}(U,V)$ of H along solutions of (37) we obtain

$$\begin{aligned} \dot{H}(U,V) = \tau V^2 + \varepsilon|b_{11}|^{-\frac{1}{2}}\{&(2a_{11} + b_{12})UV^2 - eb_{12} \\ &\times [(3U^2 + 4eU^3)V^2 + U^4(1 + eU)^2]\} + \mathcal{O}_4\,. \end{aligned} \tag{40}$$

Now, for $0 < b < 1$, let γ_b denote the half orbit of (38) in the upper half plane with initial point $(-eb, 0)$ and endpoint $(-ec, 0)$, where, owing to (39), $c = c(b) > 1$ is given by

$$c = \tfrac{1}{4}\left(3 - 2b + \sqrt{12b(1-b) + 9}\,\right). \tag{41}$$

Then it is known from [10] that the orbit of (37) through $(-eb, 0)$ is periodic if and only if

$$\tau = -\frac{B(b)}{A(b)}\varepsilon + \mathcal{O}(\varepsilon^2)\,, \tag{42}$$

where $A(b)$,$B(b)$ are the following integrals determined by $\dot{H}(U,V)$:

$$A(b) := \int_{\gamma_b} V\,dU\,,$$

$$(43)\qquad B(b) := |b_{11}|^{-1/2} \int_{\gamma_b} \{(2a_{11} + b_{12})UV^2 - eb_{12}[(3U^2 + 4eU^3)V^2 + U^4(1+eU)^2]\}\frac{dU}{V}\,.$$

To simplify $B(b)$, we apply partial integration, using $dV/dU = U(1+eU)/V$:

$$\int_{\gamma_b} U^4(1+eU)^2\frac{dU}{V} = \int_{\gamma_b} U^3(1+eU)\frac{dV}{dU}\,dU = -\int_{\gamma_b}(3U^2+4eU^3)V\,dU\,,$$

thus obtaining

$$(44)\qquad B(b) = |b_{11}|^{-\frac{1}{2}}(2a_{11}+b_{12})\int_{\gamma_b} UV\,dU\,.$$

Lemma 4 implies $2a_{11} = \Phi_{XX} < 0$, $b_{12} = \Psi_{XY} < 0$. Hence we get

$$(45)\qquad k := -|b_{11}|^{-\frac{1}{2}}(2a_{11}+b_{12}) > 0\,,$$

and introducing

$$(46)\qquad S(b,e) := \frac{\int_{\gamma_b} UV\,dU}{\int_{\gamma_b} V\,dU}\,,\qquad (e = \operatorname{sgn} b_{11})$$

we obtain

$$(47)\qquad -\frac{B(b)}{A(b)} = k\;S(b,e)\,.$$

Now, from [11] or [5] we gather that $S(b,-1)$ has positive derivative in $0 < b < 1$ and

$$(48)\qquad \lim_{b\to 0+} S(b,-1) = \frac{6}{7}\,,\qquad \lim_{b\to 1-} S(b,-1) = 1;$$

moreover, $S(b,1) = -S(b,-1)$. Via (42), this yields existence and uniqueness of periodic orbits for (37), if and only if the small parameters (ε,τ), $\varepsilon > 0$, approximately satisfy the sector condition $\frac{6}{7}k < \frac{\tau}{\varepsilon} < k$ in case $e=-1$, and $-k < \frac{\tau}{\varepsilon} < -\frac{6}{7}k$ in case $e = 1$. Beyond that, we now can state the following theorem.

THEOREM 1. (a) *Unfolding of the bifurcation at $P \in C_+$ (upper fold): Assume $0 < \alpha < \beta$ and $\alpha < 9\beta(\beta-2)(3\beta-2)^{-1}$. Then, in the (ε,τ) half plane $\varepsilon > 0$, a neighborhood $\mathcal{U}$ of the origin is subdivided by two curves*

$$(49)\qquad L_1 : \tau = \tfrac{6}{7}k\varepsilon + \mathcal{O}(\varepsilon^2),\qquad L_2 : \tau = k\varepsilon + \mathcal{O}(\varepsilon^2)$$

into sectors I, II, III *such that the local flow of the original system* (3) *with parameters corresponding to $(\varepsilon,\tau) \in \mathcal{U}$ is qualitatively given in Fig.* 3. *(Note that the third equilibrium, far away towards the "southwest," is not involved in this statement.) In particular, a periodic orbit only exists in sector* II; *it is unique and unstable. On L_1, there exists an unstable homoclinic orbit, and the points on L_2 are origins of subcritical Hopf bifurcation with respect to the parameter τ (ε fixed).*

(b) *Unfolding of the bifurcation at $P \in C_-$ (lower fold): Assume $0 < \alpha < \beta$, $1 + \alpha < 2\beta$ and $\alpha > 9\beta(\beta-2)(3\beta-2)^{-1}$ (see Fig.2). Then, in the (ε,τ) half plane $\varepsilon > 0$, a neighborhood $\mathcal{V}$ of the origin is subdivided by two curves*

$$(50)\qquad L_3 : \tau = -\tfrac{6}{7}k\varepsilon + \mathcal{O}(\varepsilon^2)\,,\qquad L_4 : \tau = -k\varepsilon + \mathcal{O}(\varepsilon^2)$$

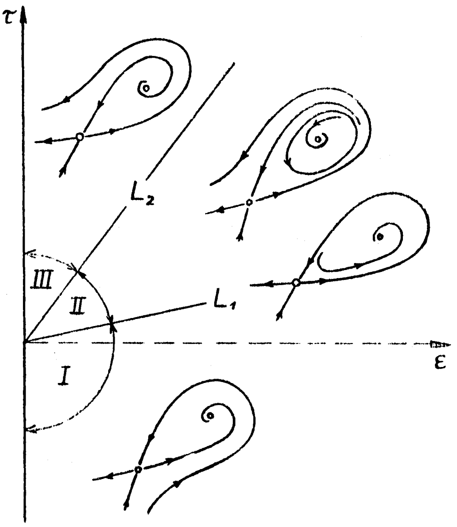


FIG. 3. *Unfolding at $P \in C_+$.*

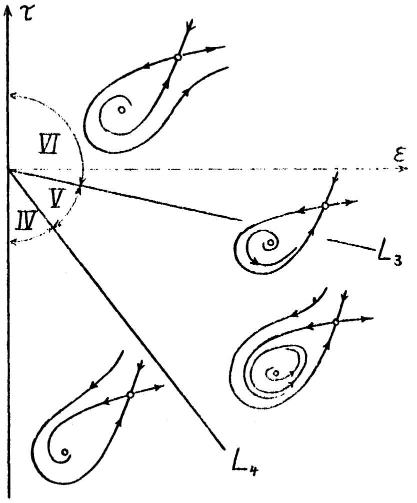


FIG. 4. *Unfolding at $P \in C_-$.*

into sectors IV, V, VI *such that the local flow of the original system* (3) *with parameters corresponding to* $(\varepsilon, \tau) \in \mathcal{V}$ *is qualitatively given in Fig.* 4. (*Again, the third equilibrium, now far away towards "northeast," is not involved.*) *In particular, a periodic orbit only exists in sector* V; *it is unique and asymptotically stable. On L_3, there exists a homoclinic orbit which is asymptotically stable* (*from inside*). *The points on L_4 are origins of supercritical Hopf bifurcation with respect to the parameter τ.*

Proof. Existence and uniqueness of periodic orbits has already been settled. The existence of a homoclinic orbit again is known from the literature [10], [4], [5]. The statements on Hopf bifurcation are easily verified. Thus, we still have to prove the assertions concerning stability. As for the homoclinic orbits, we apply [5, p. 357]. Since the linearized right-hand side of (37) has trace τ, the assertion follows from $\tau > 0$ on L_1 and $\tau < 0$ on L_3. To prove the stability results for the periodic orbits, we refer to [1, Thm. (23.9)]. (We note that [7, Thm. 4.1] could be applied, too.) Having set $\tau := k\varepsilon\, S(b,e) + \mathcal{O}(\varepsilon^2)$ in the right-hand side of (37) which will then be denoted by $f_\varepsilon(U,V)$, this system has a T-periodic orbit Γ through $(U,V) = (-eb, 0)$, described by $(U(t),V(t))$, $0 \le t \le T$. To calculate

$$D := \int_0^T \operatorname{div} f_\varepsilon(U(t),V(t))\,dt \tag{51}$$

we proceed similarly to the derivation of (44), obtaining

$$D = -2ke\,\varepsilon\left[S(b,e)\int_{\gamma_b} V^{-1}dU - \int_{\gamma_b} V^{-1}UdU\right] + \mathcal{O}(\varepsilon^2)\,. \tag{52}$$

Setting

$$C(b) := \int_{\gamma_b} UVdU$$

we calculate the derivatives with respect to b (see also [5])

$$C'(b) = (b^2 - b)\int_{\gamma_b} V^{-1}UdU\,, \qquad A'(b) = (b^2 - b)\int_{\gamma_b} V^{-1}dU\,.$$

Thus (52), (46) yield

$$\begin{aligned} D &= 2ke\varepsilon(b-b^2)^{-1}[S(b,e)A'(b) - C'(b)] + \mathcal{O}(\varepsilon^2) \\ &= -2ke\varepsilon(b-b^2)^{-1}\underbrace{A(b)S'(b,e)}_{>0} + \mathcal{O}(\varepsilon^2)\,. \end{aligned}$$

Hence, $D > 0$ if $e = -1$, and $D < 0$ if $e = 1$. In virtue of the theorem in [1] mentioned above, the statements on the stability of Γ are now proved. □

Remarks. (1) Before establishing equations (35) we assumed $q < 0$. This means that the origin $(U,V) = (0,0)$ of system (37) corresponds to a point on the middle sheet of the cusp surface $\mathcal{M}$. It is easy to see that the alternative assumption $q > 0$ would essentially amount to a translation by the vector $\binom{e}{0}$ in the (U,V)-plane. Then the origin $(U,V) = (0,0)$ would correspond to a point on the lower sheet of $\mathcal{M}$ (if $e = 1$), respectively, upper sheet (if $e = -1$); but no new insights would be attained.

(2) Returning to the parameters (p,q), $q < 0$, via (34), we get the relation

$$\frac{\tau}{\varepsilon} = |b_{11}|^{\frac{1}{2}}\frac{p}{q}\,, \tag{53}$$

but an admissible range of (ε, τ)

$$0 < \delta \le \delta_0, \qquad |\tau| \le \tau_0$$

corresponds to the admissible range $|p| \le \sqrt{|q|}\tau_0$, which is not the intersection of a full neighborhood of (0,0) with the lower half plane $q < 0$. This drawback could be overcome by a modified scaling similar to the one described in the degenerate case (see §11). We do not go into details here.

(3) The phase portraits illustrated in Figs. 3 and 4 may easily be carried over to the cusp surface $\mathcal{M}$ in the (δ,γ,ζ)-space. Using (53) and Lemma 2, we obtain a corresponding subdivision of the intersection of a neighborhood of $P = P(\alpha,\beta)$ with the middle sheet of $\mathcal{M}$. Figure 5, respectively, Fig. 6, shows this subdivision projected parallel to the negative δ-axis. Here, I' corresponds to I, etc., and the curves L_2', L_4' are just the projections parallel to the ζ-axis (onto the middle sheet of $\mathcal{M}$) of the curve $p = 0$, $q > 0$.

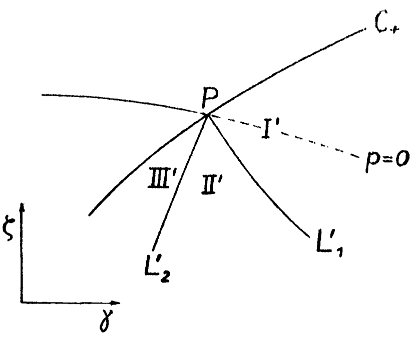


FIG. 5. *See Remark* (3).

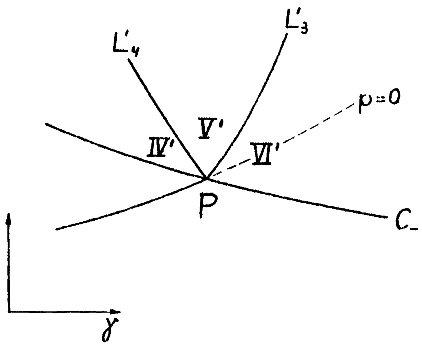


FIG. 6. *See Remark* (3).

5. The degenerate case $P = C_0$; Transformation to normal form. Beginning with the main objective of the paper, we henceforth assume

$$\alpha = 9\beta(\beta-2)(3\beta-2)^{-1}, \qquad 2 < \beta < \tfrac{8}{3}, \tag{54}$$

which means that the point $P = P(\alpha,\beta)$ yielding $p = q = 0$ is now the cusp point C_0. Lemma 4 yields $\Psi_{XX} = 0$; hence system (25) has to be expanded beyond quadratic terms. It turns out that the following expansion will suffice:

$$\begin{aligned}\dot X &= Y + a_{11}X^2 + a_{12}XY + a_{22}Y^2 + a_{111}X^3\\ &\quad + a_{112}X^2Y + a_{122}XY^2 + a_{222}Y^3 + \mathcal{O}_4,\\ \dot Y &= -qX - pY + b_1pX^2 + b_2qX^2 + b_{12}XY + b_{22}Y^2 + b_{111}X^3\\ &\quad + b_{112}X^2Y + b_{122}XY^2 + b_{222}Y^3 + b_{1111}X^4 + \mathcal{O}_4.\end{aligned} \tag{55}$$

Here, $\mathcal{O}_4$ means terms of at least fourth order with respect to (X,Y); the coefficients in (55), well defined by Taylor expansion of Φ,Ψ at $(X,Y) = (0,0)$, are real-analytic functions of the parameters (β,p,q), α being eliminated by (54). To transform (55) into normal form, we choose the near identity transformation

$$\begin{aligned} X &= w_1 + c_{11}w_1^2 + e_{111}w_1^3 + c_{112}w_1^2w_2 \,, \\ Y &= w_2 + d_{11}w_1^2 + d_{12}w_1w_2 + d_{111}w_1^3 \\ &\quad + d_{112}w_1^2w_2 + d_{122}w_1w_2^2 + d_{222}w_2^3 \,, \end{aligned} \tag{56}$$

where the coefficients are obtained from those in (55), evaluated at $(p,q) = (0,0)$. Writing $\widetilde{a}_{11} := a_{11}|_{p=q=0}$ etc., we have

$$\begin{aligned} c_{11} &= \tfrac{1}{2}(\widetilde{a}_{12} + \widetilde{b}_{22}) \,, \quad d_{11} = -\widetilde{a}_{11} \,, \quad d_{12} = \widetilde{b}_{22} \,, \\ c_{111} &= \tfrac{1}{6}\left(2\widetilde{b}_{22}^2 + 3\widetilde{a}_{12}\widetilde{b}_{22} + \widetilde{a}_{12}^2 + 2\widetilde{a}_{112} + \widetilde{b}_{122}\right) , \\ c_{112} &= \tfrac{1}{2}(\widetilde{a}_{122} + \widetilde{b}_{222}) \,, \quad d_{111} = -\widetilde{a}_{11}\widetilde{b}_{22} - \widetilde{a}_{111} \,, \\ d_{112} &= \widetilde{b}_{22}^2 + \tfrac{1}{2}\widetilde{b}_{122} \,, \quad d_{122} = \widetilde{b}_{222} \,, \quad d_{222} = -\widetilde{a}_{222} \,. \end{aligned} \tag{57}$$

We then obtain the normal form system for (w_1, w_2):

$$\begin{aligned} \dot{w}_1 &= w_2 + pg_{11} + qg_{12} + h_1 \,, \\ \dot{w}_2 &= -qw_1 - pw_2 + Aw_1w_2 + Bw_1^3 + Cw_1^2w_2 \\ &\quad + Dpw_1^2 + Eqw_1^2 + Fw_1^4 + pg_{21} + qg_{22} + h_2 \,. \end{aligned} \tag{58}$$

Here, g_{ij} and h_i are real-analytic functions of (w_1, w_2, p, q, β), with (w_1, w_2, p, q) in a neighborhood of $(0,0,0,0)$; moreover, with respect to (w_1, w_2), all terms of g_{ij} are at least of order 2 and all terms of h_i are at least of order 4, other than those already written down in (58) explicitly. As for the coefficients $A, \cdots, F$, a straightforward calculation using (57) yields

$$\begin{aligned} A &= 2\widetilde{a}_{11} + \widetilde{b}_{12}, \qquad B = -\widetilde{a}_{11}\widetilde{b}_{12} + \widetilde{b}_{111} \,, \\ C &= -\widetilde{a}_{11}\widetilde{b}_{22} + \tfrac{1}{2}\widetilde{a}_{12}\widetilde{b}_{12} + \tfrac{1}{2}\widetilde{b}_{12}\widetilde{b}_{22} + 3\widetilde{a}_{111} + \widetilde{b}_{112} \,, \\ D &= \widetilde{a}_{11} + \widetilde{b}_1, \qquad E = \tfrac{1}{2}(\widetilde{b}_{22} - \widetilde{a}_{12}) + \widetilde{b}_2 \,, \\ F &= -\tfrac{1}{2}\widetilde{a}_{11}\widetilde{b}_{12}(\widetilde{a}_{12} + \widetilde{b}_{22}) + \widetilde{a}_{11}^2\widetilde{b}_{22} - \widetilde{b}_{12}\widetilde{a}_{111} \\ &\quad + \tfrac{3}{2}\widetilde{a}_{12}\widetilde{b}_{111} + \tfrac{1}{2}\widetilde{b}_{22}\widetilde{b}_{111} - \widetilde{a}_{11}\widetilde{b}_{112} + \widetilde{b}_{1111} \,. \end{aligned} \tag{59}$$

While $\widetilde{a}_{ij}$, $\widetilde{b}_{ij}$,$\widetilde{a}_{111}$,$\widetilde{b}_{111}$, $\widetilde{b}_{112}$, and $\widetilde{b}_{1111}$ are determined by (31), $\widetilde{b}_i$ are calculated as follows:

$$\begin{aligned} \widetilde{b}_1 &= \tfrac{1}{2}\Psi_{XXp}|_{(X,Y,p,q)=(0,0,0,0)} = \tfrac{1}{2}(\Psi_{XX\gamma}\gamma_p + \Psi_{XX\zeta}\zeta_p) \,, \\ \widetilde{b}_2 &= \tfrac{1}{2}\Psi_{XXq}|_{(X,Y,p,q)=(0,0,0,0)} = \tfrac{1}{2}(\Psi_{XX\gamma}\gamma_q + \Psi_{XX\zeta}\zeta_q) \,. \end{aligned} \tag{60}$$

On the right-hand side, Ψ is considered as a function of (X, Y, γ, ζ), with derivatives to be evaluated at $(0, 0, 9\beta, 3\beta)$, and the derivatives $\gamma_p, \gamma_q, \zeta_p, \zeta_q$ are calculated by using (54), putting $\gamma = 9\beta$, $\zeta = 3\beta$ in (17), and then inverting the matrix

$$\begin{pmatrix} p_\gamma & p_\zeta \\ q_\gamma & q_\zeta \end{pmatrix} .$$

In this way we obtain

$$\gamma_p = 0 \,, \quad \gamma_q = -\tfrac{27}{2}\beta^3 \,, \quad \zeta_p = \tfrac{18\beta^2}{16-3\beta} \,, \quad \zeta_q = -\tfrac{9\beta^2(3\beta+2)}{16-3\beta} \,, \tag{61}$$

and from (26)

$$\Psi_{XX\gamma} = \frac{4(9\beta^2 - 16)}{3^7\beta^6(\beta - 2)}, \qquad \Psi_{XX\zeta} = \frac{2(9\beta^2 - 72\beta + 80)}{3^6\beta^5(\beta - 2)} \,. \tag{62}$$

From (59) we now compute

$$
(63)\quad \begin{aligned}
&A=-\frac{2(16-3\beta)}{81\beta^3(\beta-2)}<0, \qquad B=-\frac{2^6}{3^8\beta^6(\beta-2)^2}<0,\\
&C=\frac{1}{3^7\beta^5(\beta-2)^2}(9\beta^2-132\beta+224)<0,\\
&D=\frac{2(9\beta^2-60\beta+16)}{81\beta^3(\beta-2)(16-3\beta)}<0,\\
&E=\frac{9\beta^3-36\beta^2+32\beta+64}{54\beta^3(\beta-2)(16-3\beta)}>0, \qquad F=\frac{32(20-9\beta)}{3^{11}\beta^8(\beta-2)^3}.
\end{aligned}
$$

6. Scaling; an auxiliary "unperturbed" system. From now on we assume $q\neq 0$. Then, in system (58), the variables (w_1,w_2,t) and the parameters (p,q) may be rescaled as follows:

$$
(64)\quad \varepsilon=\sqrt{|q|}, \quad p=-\varepsilon\tau, \quad w_1=\varepsilon A^{-1}u, \quad w_2=\varepsilon^2A^{-1}v, \quad t=\varepsilon^{-1}\tilde{t}.
$$

This leads to the following system for $u(\tilde{t})$,$v(\tilde{t})$ with small parameters (ε,τ), $\varepsilon>0$ (henceforth we omit the tilde in $\tilde{t}$):

$$
(65)\quad \begin{aligned}
\dot{u}&=v+R_1(u,v,\varepsilon,\tau),\\
\dot{v}&=-fu+\tau v+uv-au^3+\varepsilon bu^2v+\tau cu^2-\varepsilon fdu^2\\
&\quad+\varepsilon eu^4+R_2(u,v,\varepsilon,\tau).
\end{aligned}
$$

Here we have the following: For any fixed $\beta\in(2,\frac{8}{3}), R_i$ are bounded functions for (u,v) in an arbitrary bounded domain and (ε,τ) in a sufficiently small neighborhood of (0,0); moreover, R_i are real-analytic functions of (u,v,ε,τ), whose lowest terms are at least quadratic both with respect to (u,v) and (ε,τ). Finally, $f:=\operatorname{sgn} q$, and the coefficients $a,\cdots,e$ are given by

$$
(66)\quad \begin{aligned}
&a=-\frac{B}{A^2}=\frac{16}{(16-3\beta)^2}>0, \qquad b=\frac{C}{A^2}=\frac{3\beta(9\beta^2-132\beta+224)}{4(16-3\beta)^2}<0,\\
&c=-\frac{D}{A}=\frac{9\beta^2-60\beta+16}{(16-3\beta)^2}<0,\\
&d=-\frac{E}{A}=\frac{3(9\beta^3-36\beta^2+32\beta+64)}{4(16-3\beta)^2}>0, \qquad e=\frac{F}{A^3}=\frac{12\beta(9\beta-20)}{(16-3\beta)^3}.
\end{aligned}
$$

We first investigate the unperturbed system for $U(t)$, $V(t)$, resulting from (65) by setting $(\varepsilon,\tau)=(0{,}0)$:

$$
(67)\quad \begin{aligned}
\dot{U}&=V,\\
\dot{V}&=-fU+UV-aU^3,
\end{aligned} \qquad f=\operatorname{sgn} q.
$$

This system is invariant under $(t,U,V)\mapsto(-t,-U,V)$, and it has the first integral $I(U,V)$ given by

$$
(68)\quad I(U,V):=\frac{1}{2}\ln N+\frac{1}{\sigma}\arctan\frac{4aV-(f+aU^2)}{\sigma(f+aU^2)},
$$

where

$$
(69)\quad \sigma:=\sqrt{8a-1}=\frac{1}{16-3\beta}(-9\beta^2+96\beta-128)^{\frac{1}{2}}>0,
$$

$$
(70)\quad \begin{aligned}
N=N(U,V)&:=2aV^2-V(f+aU^2)+(f+aU^2)^2\\
&=(8a)^{-1}\{(4aV-f-aU^2)^2+\sigma^2(f+aU^2)^2\}.
\end{aligned}
$$

We calculate

$$\frac{\partial I(U,V)}{\partial U} = \frac{2aU}{N}(-V+f+aU^2), \qquad \frac{\partial I(U,V)}{\partial V} = \frac{2aV}{N}. \tag{71}$$

Owing to $I(-U,V) = I(U,V)$, the phase portraits are symmetric with respect to the V-axis. In case $f = 1$ ($q > 0$), system (67) has the only equilibrium $(U,V) = (0,0)$, and we easily check that each orbit is periodic (see Fig. 7). For $f = -1$ ($q < 0$), there are three equilibria $(U,V) = (0,0)$ (hyperbolic saddle point) and $(\pm a^{-1/2}, 0)$ (source and sink), and the phase portrait is qualitatively shown in Fig. 8. In particular, there is a homoclinic orbit, and all orbits intersecting the positive V-axis are periodic.

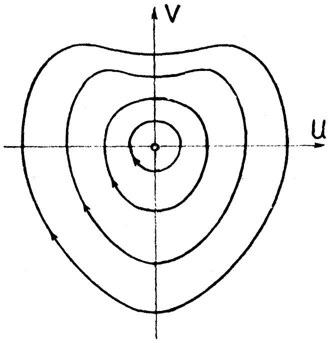


FIG. 7. *Phase portrait of system* (67), $f = 1$.

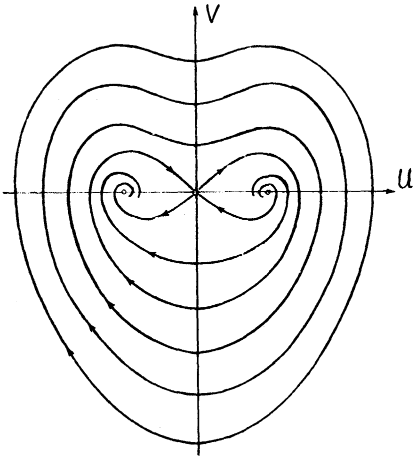


FIG. 8. *Phase portrait of system* (67), $f = -1$.

7. Looking for periodic orbits of system (65). Now, returning to the perturbed system (65), we have to investigate whether some of the periodic orbits of (67) or the homoclinic orbit in case $f = -1$ do survive for certain parameters $(\varepsilon, \tau) \neq (0,0)$.

Let $\dot{I}(u,v)$ denote the derivative of $I(u,v)$ along solutions of (65). Using (71), we obtain

$$\begin{aligned}\dot{I}(u,v) &= \frac{2av}{N(u,v)}\{\tau v + \varepsilon bu^2v + (\tau c - \varepsilon fd)u^2 + \varepsilon eu^4\} \\ &\quad + \frac{2a}{N(u,v)}\{u(-v + f + au^2)R_1(u,v,\varepsilon,\tau) + vR_2(u,v,\varepsilon,\tau)\}\,.\end{aligned} \tag{72}$$

For arbitrary $r > 0$, let $(U(t), V(t))$ be the (periodic) solution of (67) with $(U(0), V(0)) = (0, r)$ and $(U(\pm t_0),\ V(\pm t_0)) = (0, -r_0)$ for some $t_0 > 0$ and $r_0 > 0$ depending on r (t_0 minimal). Analogously, let $(u(t), v(t))$ denote the solution of (65) with $(u(0), v(0)) = (0, r)$, intersecting the negative v-axis for the first positive time t_1 at $(0, -r_1)$ and for the first negative time $-t_2$ at $(0, -r_2)$. Then we obtain

$$\begin{pmatrix}u(t)\\ v(t)\end{pmatrix} = \begin{pmatrix}U(t)\\ V(t)\end{pmatrix} + \mathcal{O}(\sqrt{\varepsilon^2+\tau^2}) \quad \text{uniformly in } |t| \le \max\{t_0, t_1, t_2\}\,,$$

and hence from (72), again uniformly in this t-interval:

$$\begin{aligned}\dot{I}(u(t), v(t)) = \frac{2aV(t)}{N(U(t), V(t)}&\{\tau V(t) + \varepsilon bU^2(t)V(t) \\ &\quad + (\tau c - \varepsilon fd)U^2(t) + \varepsilon eU^4(t)\} \\ + \mathcal{O}(\varepsilon^2 + \tau^2) &=: H(t) + \mathcal{O}(\varepsilon^2 + \tau^2)\,.\end{aligned} \tag{73}$$

Using (71), we see that $(u(t), v(t))$ is periodic if and only if

$$I(0, -r_1) = I(0, -r_2)\,. \tag{74}$$

Since $t_1 - t_0 = \mathcal{O}(\sqrt{\varepsilon^2 + \tau^2})$, we conclude

$$\begin{aligned}I(0, -r_1) - I(0, r) &= \int_0^{t_1} \dot{I}(u(t), v(t))\, dt \\ &= \int_0^{t_0} H(t)dt + \mathcal{O}(\varepsilon^2 + \tau^2)\,,\end{aligned} \tag{75}$$

and similarly,

$$I(0, -r_2) - I(0, r) = \int_0^{-t_0} H(t)\, dt + \mathcal{O}(\varepsilon^2 + \tau^2)\,. \tag{76}$$

The symmetry of (67) yields $U(-t) = -U(t)$, $V(-t) = V(t)$ and hence, $H(-t) = H(t)$, which again implies

$$\int_0^{-t_0} H(t)\, dt = -\int_0^{t_0} H(t)\, dt\,. \tag{77}$$

Thus, setting

$$K(r, \varepsilon, \tau) := I(0, -r_1) - I(0, -r_2) = 2\int_0^{t_0} H(t)\, dt + \mathcal{O}(\varepsilon^2 + \tau^2)\,, \tag{78}$$

(74) is equivalent to $K(r, \varepsilon, \tau) = 0$, or, by the implicit function theorem, to

$$\tau = -\frac{K_\varepsilon(r, 0, 0)}{K_\tau(r, 0, 0)}\,\varepsilon + \mathcal{O}(\varepsilon^2)\,, \tag{79}$$

whenever the denominator does not vanish. The quotient

$$Q(r) := -\frac{K_\varepsilon(r,0,0)}{K_\tau(r,0,0)} = -\frac{\int_0^{t_0} \frac{V(t)}{N(U(t),V(t))}[bU^2(t)V(t) - fdU^2(t) + eU^4(t)]dt}{\int_0^{t_0} \frac{V(t)}{N(U(t),V(t))}[V(t) + cU^2(t)]dt} \tag{80}$$

can be expressed by line integrals along the curve

$$\Gamma = \Gamma(r) := \{(U(t), V(t)) | 0 \leq t \leq t_0\}, \tag{81}$$

resulting in

$$Q(r) = -\frac{\int_\Gamma N^{-1}(U,V)[bU^2V - fdU^2 + eU^4]dU}{\int_\Gamma N^{-1}(U,V)[V + cU^2]dU}. \tag{82}$$

Moreover, in the case $f = -1$, let Γ_0 denote the homoclinic semi-orbit of (67) in the half plane $U \geq 0$ (see Fig. 8). Then, by arguments similar to the preceding ones (see [10], [4]), we find the condition for a homoclinic orbit of (65) to exist. Here, the role of $(u(t), v(t))$ above is played by those pieces of the local unstable, respectively, stable, manifold of the saddle (0,0) of (65), which start out into the upper half plane. We may summarize with the following lemma.

LEMMA 5. *Let r, $0 < r \leq \bar{r}$, be such that $Q(r)$ exists. Then the solution of* (65) *passing through* $(0, r)$ *is periodic if and only if $\tau = \tau(\varepsilon)$ is a well-defined real-analytic function*

$$\tau = Q(r)\varepsilon + \mathcal{O}(\varepsilon^2). \tag{83}$$

Moreover, in case $f = -1$, system (65) *has a homoclinic orbit joining the saddle* (0,0) *to itself, if and only if $\tau = \tau(\varepsilon)$ is a well-defined* (*real-analytic*) *function*

$$\tau = Q_0\varepsilon + \mathcal{O}(\varepsilon^2) \tag{84}$$

where

$$Q_0 := -\frac{\int_{\Gamma_0} N^{-1}(U,V)[bU^2V + dU^2 + eU^4]dU}{\int_{\Gamma_0} N^{-1}(U,V)[V + cU^2]dU} \tag{85}$$

and $N(U, V)$ is given by (70) *with $f = -1$.* □

Remark. As will be seen in §12, the denominator of Q_0 does not vanish.

8. Manipulating the integrals involved in $Q(r)$. We now have to investigate $Q(r)$ as a function of $r > 0$. For this purpose, we first apply partial integration to the individual integrals in (82) in such a way that the boundary terms are zero. Using

$$\frac{dV}{dU} = U - \frac{U}{V}(f + aU^2), \quad \frac{dN}{dU} = \frac{UN}{V}, \quad \frac{dU}{V} = dt, \tag{86}$$

we calculate (all the integrals occurring subsequently do exist; we write $\int$ instead of $\int_\Gamma$)

$$\int \frac{U^2}{N}dU = -\frac{1}{3}\int U^3 \frac{d}{dU}\left(\frac{1}{N}\right) dU = \frac{1}{3}\int \frac{U^4 dU}{NV} > 0, \tag{87}$$

$$\begin{aligned} \int \frac{U^2V}{N}dU &= \int \frac{U^2V^2}{N}\frac{dU}{V} \\ &= -\frac{1}{3}\int U^3 \frac{d}{dU}\left(\frac{V}{N}\right)dU = \frac{1}{3}\int \frac{U^4}{NV}(f + aU^2)dU > 0, \end{aligned} \tag{88}$$

$$\int \frac{V}{N} dU = \int \frac{V^2}{N} \frac{dU}{V} = \int \frac{U^2}{VN}(f + aU^2) dU > 0, \tag{89}$$

$$\int \frac{U^4}{N} dU = \frac{1}{5} \int \frac{U^6}{NV} dU = \frac{3}{5a} \int \frac{U^2 V}{N} dU - \frac{f}{5a} \int \frac{U^4}{NV} dU > 0. \tag{90}$$

At this stage, it seems useful to introduce the following "polar" coordinates (ϱ, φ) in the (U, V)-plane:

$$f + aU^2 - 4aV = f\varrho \sin\varphi, \qquad \sigma(f + aU^2) = f\varrho\cos\varphi. \tag{91}$$

We calculate

$$N(U,V) = \frac{\varrho^2}{8a}, \qquad I(U,V) = \ln\frac{\varrho}{\sqrt{8a}} - \frac{\varphi}{\sigma}. \tag{92}$$

Defining, for convenience, the variable s by

$$r =: -fs, \qquad r > 0 \tag{93}$$

the integration curve $\Gamma = \Gamma(r)$, defined in (81) and also given by $I(U,V) = I(0,r)$, is described in terms of (ϱ, φ) as follows:

$$\varrho = \sqrt{8a} G(s) e^{\sigma^{-1}\varphi}, \qquad \varphi_0 \le \varphi \le \varphi_1 \quad \text{(logarithmic spiral)} \tag{94}$$

where

$$G(s) := \sqrt{2as^2 + s + 1}\exp(-\sigma^{-1}\arctan[\sigma^{-1}(4as + 1)])\,; \tag{95}$$

$$\varphi_0 = \varphi_0(s) := \arctan[\sigma^{-1}(4as + 1)] \tag{96}$$

and $\varphi_1 = \varphi_1(s)$ are consecutive zeros of the function

$$h^2(\varphi, s) := f(\sigma^{-1}\sqrt{8a}G(s)\cos\varphi e^{\sigma^{-1}\varphi} - 1) \tag{97}$$

which, for fixed s, is positive between φ_0 and $\varphi_1 > \varphi_0$ (for $f = -1$, see Fig. 9). Let $h(\varphi, s)$ denote the positive square root of $h^2(\varphi, s)$ in $\varphi_0 \le \varphi \le \varphi_1$. On $\Gamma = \Gamma(r)$, we then get

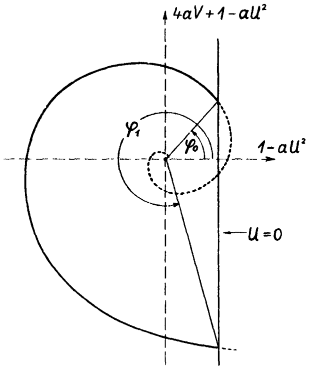


FIG. 9. *Illustrating* φ_0, φ_1 *in case* $f = -1$.

$$(98)\qquad \begin{aligned} &U = \frac{1}{\sqrt{a}}h(\varphi, s)\,, \qquad V = \frac{2f}{\sigma\sqrt{8a}}G(s)(\cos\varphi - \sigma\sin\varphi)e^{\sigma^{-1}\varphi}\,,\\ &dU = \frac{\sqrt{2}fG(s)}{\sigma^2 h(\varphi, s)}(\cos\varphi - \sigma\sin\varphi)e^{\sigma^{-1}\varphi}d\varphi\,,\\ &\frac{dU}{V} = \frac{2\sqrt{a}}{\sigma h(\varphi, s)}d\varphi\,. \end{aligned}$$

Together with (87)–(94), this yields

$$(99)\qquad \begin{aligned} \int_\Gamma \frac{U^2}{N}dU &= \frac{2}{3a^{3/2}\sigma G^2(s)}\int_{\varphi_0}^{\varphi_1} h^3(\varphi, s)e^{-2\sigma^{-1}\varphi}d\varphi > 0\,,\\ \int_\Gamma \frac{U^2V}{N}dU &= \frac{4\sqrt{2}}{3fa\sigma^2 G(s)}\int_{\varphi_0}^{\varphi_1} h^3(\varphi, s)\cos\varphi e^{-\sigma^{-1}\varphi}d\varphi > 0\,,\\ \int_\Gamma \frac{V}{N}dU &= \frac{4\sqrt{2}}{f\sigma^2 G(s)}\int_{\varphi_0}^{\varphi_1} h(\varphi, s)\cos\varphi e^{-\sigma^{-1}\varphi}d\varphi > 0\,,\\ \int_\Gamma \frac{U^4}{N}dU &= \frac{4\sqrt{2}}{5fa^2\sigma^2 G(s)}\int_{\varphi_0}^{\varphi_1} h^3(\varphi, s)\cos\varphi e^{-\sigma^{-1}\varphi}d\varphi\\ &\quad - \frac{2}{5fa^{5/2}\sigma G^2(s)}\int_{\varphi_0}^{\varphi_1} h^3(\varphi, s)e^{-2\sigma^{-1}\varphi}d\varphi > 0\,. \end{aligned}$$

Now, let us introduce the variable

$$(100)\qquad \xi = \xi(s) := \ln G(s) + \sigma^{-1}\arctan\sigma^{-1}$$

yielding

$$(101)\qquad \begin{aligned} &\lim_{s\to 0}\xi(s) = 0\,,\ \lim_{s\to\pm\infty}|s|^{-1}e^{\xi} = \sqrt{2a}\exp[\sigma^{-1}(\arctan\sigma^{-1} \mp \tfrac{\pi}{2}]\,,\\ &\frac{d\xi}{ds} = \frac{2as}{2as^2 + s + 1}\,,\quad \xi(s) > 0 \quad \text{for } s \neq 0\,. \end{aligned}$$

Then the following functions $H_i(\xi)$, $i = 1, 2, 3$, are well defined:

$$(102)\qquad \begin{aligned} H_1(\xi) &:= \frac{f}{G(s)}\int_{\varphi_0}^{\varphi_1} h(\varphi, s)\cos\varphi e^{-\sigma^{-1}\varphi}d\varphi\,,\\ H_2(\xi) &:= \frac{1}{G(s)}\int_{\varphi_0}^{\varphi_1} h^3(\varphi, s)\cos\varphi e^{-\sigma^{-1}\varphi}d\varphi\,,\\ H_3(\xi) &:= \frac{\sigma}{\sqrt{8a}G^2(s)}\int_{\varphi_0}^{\varphi_1} h^3(\varphi, s)e^{-2\sigma^{-1}\varphi}d\varphi\,. \end{aligned}$$

Obviously, from (99) we get

$$(103)\qquad H_1(\xi) > 0\,,\quad fH_2(\xi) > 0\,,\quad H_3(\xi) > 0 \quad \text{for all } \xi > 0\,.$$

Using (99) and (102), we obtain from (82)

$$(104)\qquad Q(r) = -f\frac{(5ab + 3e)H_2(\xi) - (5ad + 3e)H_3(\xi)}{15a^2H_1(\xi) + 5acH_3(\xi)}\,;$$

hence, by (66)

$$(105)\qquad Q(r) = \frac{3f}{5}\frac{\lambda H_2(\xi) + \mu H_3(\xi)}{48H_1(\xi) + \nu H_3(\xi)}$$

with coefficients

$$\begin{aligned}\lambda &:= \beta(9\beta^2+12\beta-40) > 0,\\ \mu &:= -9\beta^3+108\beta^2-200\beta+80 > 0, \qquad (106)\\ \nu &:= 9\beta^2-60\beta+16 < 0.\end{aligned}$$

Furthermore, we easily calculate

$$(107)\qquad f^{-1}(H_1(\xi)-H_3(\xi)) = \frac{\sigma}{\sqrt{8a}G^2(s)}\int_{\varphi_0}^{\varphi_1} h(\varphi,s)e^{-2\sigma^{-1}\varphi}d\varphi > 0,$$

$$(108)\qquad f^{-1}(H_1(\xi)+H_2(\xi)) = \frac{\sqrt{8a}}{\sigma}\int_{\varphi_0}^{\varphi_1} h(\varphi,s)\cos^2\varphi\, d\varphi > 0.$$

Owing to (107), the denominator in (105) may be written as

$$\begin{aligned}(109)\qquad \mathcal{D} &:= 48H_1(\xi)+\nu H_3(\xi)\\ &= (48+\nu)H_1(\xi) - \frac{\nu f\sigma}{\sqrt{8a}G^2(s)}\int_{\varphi_0}^{\varphi_1} h(\varphi,s)e^{-2\sigma^{-1}\varphi}d\varphi.\end{aligned}$$

Now, $\nu<0$ and

$$(110)\qquad 48+\nu = (3\beta-16)(3\beta-4) < 0 \quad \text{for } 2<\beta<\tfrac{8}{3};$$

hence

$$(111)\qquad \mathcal{D}<0 \quad \text{for } f=-1,\ \ \xi>0.$$

9. Differential equations and limit relations for $H_i(\xi)$. Next, we investigate the functions $H_i(\xi)$. As a first step, we derive a system of differential equations for these functions, which correspond to the Picard–Fuchs differential equations considered in [11].

From (100), (101) we obtain

$$(112)\qquad \frac{dG}{d\xi} = G, \qquad \frac{\partial h}{\partial s}\frac{ds}{d\xi} = \frac{f\sqrt{8a}G(s)\cos\varphi e^{\sigma^{-1}\varphi}}{2\sigma h(\varphi,s)}.$$

Moreover, it follows from (96), (97) that φ_0 and, by the implicit function theorem, also φ_1 are real-analytic functions of s satisfying $h(\varphi_0(s),s)\equiv h(\varphi_1(s),s)\equiv 0$. Using these properties, we can easily calculate the derivatives of the functions $H_i(\xi)$.

LEMMA 6. *Let us define the additional function*

$$(113)\qquad H_0(\xi) := \frac{\sqrt{8a}}{\sigma}\int_{\varphi_0}^{\varphi_1} h^{-1}(\varphi,s)\cos^2\varphi\, d\varphi.$$

Then H_1, H_2, H_3 satisfy the following linear system of differential equations:

$$\begin{aligned}H_1' &= -H_1+\tfrac{1}{2}H_0(\xi),\\ H_2' &= \tfrac{3}{2}H_1+\tfrac{1}{2}H_2, \qquad (114)\\ H_3' &= \tfrac{3}{2}H_1-2H_3.\end{aligned}$$

Proof. We shall only prove the second equation, the rest being similar. Using (102), (108), and (112) we obtain

$$\begin{aligned}H_2' &= -\frac{dG}{d\xi}\frac{1}{G^2}\int_{\varphi_0}^{\varphi_1} h^3(\varphi,s)\cos\varphi e^{-\sigma^{-1}\varphi}d\varphi + \tfrac{3}{2}\frac{f\sqrt{8a}}{\sigma}\int_{\varphi_0}^{\varphi_1} h(\varphi,s)\cos^2\varphi\, d\varphi\\ &= -H_2+\tfrac{3}{2}(H_1+H_2) = \tfrac{3}{2}H_1+\tfrac{1}{2}H_2.\end{aligned}$$ □

Next, we investigate the limits of $H_i(\xi)$ as ξ tends to ∞, respectively, to zero. By the way, recall the relation between r, s, and ξ given in (93) and (100).

LEMMA 7 (The limit $|s| \to \infty$).

(a) $\lim\limits_{r\to\infty} \sqrt{|s|}H_i(\xi(s)) = \frac{\sqrt{2}}{\sqrt{\sigma}}e^{-\frac{\pi}{4\sigma}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{\frac{3}{2}} \varphi e^{-\frac{\varphi}{2\sigma}} d\varphi =: L_1, \qquad i = 0,1,3.$

(b) $\lim\limits_{r\to\infty} \frac{H_2(\xi(s))}{\sqrt{|s|}} = \frac{4\sqrt{2}af}{\sigma^{\frac{3}{2}}}e^{\frac{\pi}{4\sigma}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{\frac{5}{2}} \varphi e^{\frac{\varphi}{2\sigma}} d\varphi =: f\, L_2.$

Proof. From (95)–(97) we obtain for $r \to \infty$

$$\frac{\sigma h^2(\varphi, s)}{G(s)\sqrt{8a}e^{\sigma^{-1}\varphi}} \longrightarrow f\cos\varphi\,, \qquad \frac{G(s)}{|s|} \longrightarrow \sqrt{2a}\, e^{\frac{\pi}{2\sigma}}\,,$$

$$(\varphi_0, \varphi_1) \longrightarrow \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \quad \text{in case } f = \ \ 1\,,$$

$$(\varphi_0, \varphi_1) \longrightarrow \left(\frac{\pi}{2}, \frac{3\pi}{2}\right) \quad \text{in case } f = -1\,.$$

If $f = 1$, the assertion for H_1,H_2,H_3 now follows easily from (102). For H_0, the integrand is unbounded, but a careful investigation also yields the assertion in case $f = 1$. The case $f = -1$ is treated similarly, and a final translation $\varphi = \psi + \pi$ then gives the desired result. □

LEMMA 8 (The limit $s \to 0$ ($\xi \to 0$)). *If $f = 1$, the interval $[\varphi_0, \varphi_1]$ shrinks to the point $\chi := \arctan\sigma^{-1}$. If $f = -1$, $[\varphi_0, \varphi_1]$ tends to the interval $[\chi, \vartheta]$, where $\vartheta > \chi$ is the zero next to χ of the function*

$$g(\varphi) := (1 - \sigma^{-1}\sqrt{8a}\cos\varphi e^{\sigma^{-1}(\varphi-\chi)})^{\frac{1}{2}} \geq 0\,. \tag{115}$$

The functions H_i satisfy the following limit relations:

(a) *Case* $f = 1$:

$$\lim_{\xi\to 0} H_0(\xi) = k := \frac{\pi\sigma^2}{4\sqrt{2a}}\,, \tag{116}$$

$$\lim_{\xi\to 0} \frac{H_1(\xi)}{\xi} = \frac{k}{2}\,, \tag{117}$$

$$\lim_{\xi\to 0} \frac{H_i(\xi)}{\xi^2} = \frac{3k}{8}\,, \qquad i = 2,3\,. \tag{118}$$

(b) *Case* $f = -1$:

$$\lim_{\xi\to 0} H_1(\xi) = -e^{\sigma^{-1}\chi} \int_{\chi}^{\vartheta} g(\varphi)\cos\varphi e^{-\sigma^{-1}\varphi} d\varphi =: l_1 > 0\,, \tag{119}$$

$$\lim_{\xi\to 0} H_2(\xi) = e^{\sigma^{-1}\chi} \int_{\chi}^{\vartheta} g^3(\varphi)\cos\varphi e^{-\sigma^{-1}\varphi} d\varphi =: l_2 < 0\,, \tag{120}$$

$$\lim_{\xi\to 0} H_3(\xi) = \frac{\sigma}{\sqrt{8a}} e^{2\sigma^{-1}\chi} \int_{\chi}^{\vartheta} g^3(\varphi) e^{-2\sigma^{-1}\varphi} d\varphi =: l_3 > 0\,, \tag{121}$$

$$\lim_{s\to 0} \frac{H_0(\xi(s))}{-\ln s} = \frac{\sigma^2}{4\sqrt{2a}}\,. \tag{122}$$

Proof. The assertions concerning the interval $[\varphi_0, \varphi_1]$ follow from (95)–(97), which show that $\tan\chi = \sigma^{-1}$, $G(0) = e^{-\sigma^{-1}\chi}$, and $h^2(\varphi, s)$ tends to $-fg^2(\varphi)$ if $\varphi \in [\varphi_0, \varphi_1]$. Then, using $h^2(\varphi_i(s), s) \equiv 0$, $i = 0, 1$, the expansion of

$$f_i(\psi, s) := h^2(\varphi_i(s) + \psi, s) \tag{123}$$

at $(\psi, s) = (0,0)$ yields

$$f_0(\psi, s) = -4af\sigma^{-2}\psi[\sigma s + \psi + \mathcal{O}(\psi^2 + s^2)] \tag{124}$$
$$=: 4af\sigma^{-2}\psi F(\psi, s),$$

$$f_1(\psi, s) = f\sigma^{-1}\psi\Big[1 - \sigma\tan\vartheta + \mathcal{O}(\sqrt{\psi^2 + s^2})\Big] \tag{125}$$

with $1 - \sigma\tan\vartheta < 0$. Let us now consider the case $f = 1$ $(s < 0)$. We obtain from (124) $\varphi_1(s) - \varphi_0(s) = -\sigma s + \mathcal{O}(s^2) =: \delta(s)$, and further, by the Weierstrass preparation theorem [5], $F(\psi, s) = (\delta(s) - \psi)/q(\psi, s)$, where the functions F, δ, and q are real-analytic, $q(\psi, s) = 1 + \mathcal{O}(\sqrt{\psi^2 + s^2})$. From (113) we now calculate

$$\begin{aligned} H_0(\xi) &= \sqrt{2}\int_0^{\delta(s)} \frac{\cos^2(\varphi_0(s) + \psi)}{\sqrt{\psi F(\psi, s)}} d\psi \\ &= \sqrt{2}\cos^2\chi \int_0^{\delta(s)} \frac{1 + \mathcal{O}(\sqrt{\psi^2 + s^2}\,)}{\sqrt{\psi(\delta(s) - \psi)}} d\psi \\ &= \frac{\sigma^2}{4\sqrt{2}a}\left(\underbrace{\int_0^{\delta(s)} \frac{d\psi}{\sqrt{\psi(\delta(s) - \psi)}}}_{=\pi} + \int_0^{\delta(s)} \frac{\mathcal{O}(\sqrt{\psi^2 + s^2}\,)}{\sqrt{\psi(\delta(s) - \psi)}} d\psi\right) \\ &= \ \frac{\pi\sigma^2}{4\sqrt{2}a} + \mathcal{O}(s)\,, \end{aligned}$$

which proves (116). Since, trivially, $H_i(\xi) \to 0$, $i = 1, 2, 3$, we can successively calculate the one-sided derivatives of $H_i(\xi)$ at $\xi = 0$ using (114):

$$\lim_{\xi \to 0} \frac{H_i(\xi)}{\xi} = H_i'(0) = \begin{cases} \frac{k}{2} & \text{if } i = 1, \\ 0 & \text{if } i = 2, 3\,, \end{cases}$$

which proves (117); then for $i = 2, 3$,

$$\lim_{\xi \to 0} \frac{H_i'(\xi)}{\xi} = H_i''(0) = \tfrac{3}{2}H_1'(0) = \frac{3k}{4}\,.$$

This obviously proves (118). Now consider the case $f = -1$ $(s > 0)$. Formulae (119)–(121) are simple consequences of the definition (102). Thus, we only have to prove (122). From (124), (125) we see that, as $s \to 0$, the integrand in H_0 becomes critical at the lower limit $\varphi_0(s)$ while it remains well behaved near the upper limit $\varphi_1(s)$. Therefore, we subdivide the interval $[\varphi_0(s), \varphi_1(s)]$ into $[\varphi_0(s), \varphi_0(s) + \eta]$ and $[\varphi_0(s) + \eta\,, \varphi_1(s)]$ with some $\eta > 0$, independent of s and sufficiently small. It is easy to show that the integral over $[\varphi_0(s) + \eta\,, \varphi_1(s)]$ is uniformly bounded with respect to s. Thus it suffices to investigate the integral over $[\varphi_0(s), \varphi_0(s) + \eta]$:

$$I := \frac{\sqrt{8a}}{\sigma}\int_{\varphi_0}^{\varphi_0 + \eta} h^{-1}(\varphi, s)\cos^2\varphi\, d\varphi = \sqrt{2}\int_0^{\eta} \frac{\cos^2(\varphi_0(s) + \psi)\, d\psi}{\sqrt{\psi[\sigma s + \psi + \mathcal{O}(\psi^2 + s^2)]}}\,.$$

Since, if $\psi \geq 0$, $s > 0$,

$$\begin{aligned} \sigma s + \psi + \mathcal{O}(\psi^2 + s^2) &= (\sigma s + \psi)(1 + \frac{\mathcal{O}(\psi^2 + s^2)}{\sigma s + \psi}) \\ &= (\sigma s + \psi)(1 + \mathcal{O}(\sqrt{\psi^2 + s^2}))\,, \end{aligned}$$

we obtain

$$I = \frac{\sigma^2}{4\sqrt{2}a} \int_0^{\eta} \frac{1 + \mathcal{O}(\sqrt{\psi^2 + s^2})}{\sqrt{\psi(\sigma s + \psi)}}\, d\psi\,.$$

Writing

$$I_1 := \int_0^{-\frac{1}{\ln s}} \frac{1 + \mathcal{O}(\sqrt{\psi^2 + s^2})}{\sqrt{\psi(\sigma s + \psi)}}\, d\psi\,,\quad I_2 := \int_{-\frac{1}{\ln s}}^{\eta} \frac{1 + \mathcal{O}(\sqrt{\psi^2 + s^2})}{\sqrt{\psi(\sigma s + \psi)}}\, d\psi\,,$$

we calculate

$$\begin{aligned} I_1 &= \left[1 + \mathcal{O}\left(\frac{1}{|\ln s|}\right)\right] \int_0^{-\frac{1}{\ln s}} \frac{d\psi}{\sqrt{\psi(\sigma s + \psi)}} \\ &= \left[1 + \mathcal{O}\left(\frac{1}{|\ln s|}\right)\right] \ln\left(\frac{2 + \sigma s|\ln s| + 2\sqrt{1 + \sigma s|\ln s|}}{\sigma s|\ln s|}\right); \end{aligned}$$

hence,

$$\lim_{s\to 0} \frac{I_1}{|\ln s|} = \lim_{s\to 0} \frac{\ln(s|\ln s|)}{\ln s} = 1\,.$$

Similarly, we obtain

$$\begin{aligned} I_2 &= \left[1 + \mathcal{O}(\sqrt{\eta^2 + s^2})\right] \ln \frac{(2\eta + \sigma s)\ln s + 2\ln s\sqrt{\eta^2 + \sigma s \eta}}{\sigma s \ln s - 2 - 2\sqrt{1 - \sigma s \ln s}} \\ &= \mathcal{O}(\ln|\ln s|) = o(|\ln s|)\,. \end{aligned}$$

This completes the proof of (122) and thus, of the whole lemma. □

10. Properties of $Q(r)$; Monotonicity in case $f = -1$. Let us now return to the quotient $Q = Q(r)$ in (105). Though inaccurate, we shall write $Q(r) = Q(s) = Q(\xi)$ for simplicity. Using the preceding lemmata and (103), (106), (110), we immediately obtain Lemma 9.

LEMMA 9. (a) *In the case $f = 1$, the numerator $\mathcal{N}$ of Q*

$$\mathcal{N} := \lambda H_2(\xi) + \mu H_3(\xi) \tag{126}$$

is positive for all $\xi > 0$, while the denominator $\mathcal{D}$, defined in (109), *satisfies the limit relations*

$$\lim_{\xi\to 0} \frac{\mathcal{D}}{\xi} = \frac{3\sqrt{2}\pi\sigma^2}{a}\,,\qquad \lim_{s\to -\infty} \sqrt{|s|}\mathcal{D} = (48 + \nu)L_1 < 0\,. \tag{127}$$

Thus, contrary to the case $f = -1$ (see (111)*), $\mathcal{D} = \mathcal{D}(\xi)$ has at least one zero in $0 < \xi < \infty$ with change of sign.*

(b) *For any $f = \pm 1$, the quotient Q satisfies*

$$\lim_{r\to\infty} \frac{Q}{|s|} = \frac{3\lambda}{5(48 + \nu)}\,\frac{L_2}{L_1} < 0\,. \tag{128}$$

(c) *In the case $f = 1$, we have*

$$\lim_{\xi\to 0} \frac{Q}{\xi} = \frac{3}{5} \cdot \frac{\lambda + \mu}{64} = \frac{3}{8}(3\beta^2 - 6\beta + 2) > 0\,. \tag{129}$$

(d) *In the case $f = -1$, we obtain from* (109) *$48l_1 + \nu l_3 < 0$, and hence*

$$\lim_{\xi\to 0} Q = -\frac{3}{5} \cdot \frac{\lambda l_2 + \mu l_3}{48 l_1 + \nu l_3}\,. \tag{130}$$

□

Remark. The quotient L_2/L_1 occurring in (128) can be expressed by the gamma function. From formula 19 in [6, p. 138] we calculate

$$\frac{L_2}{L_1} = \frac{20\pi^2\sigma(9\sigma^2+1)e^{\frac{\pi}{2\sigma}}}{(1+25\sigma^2)(1+2\sinh^2\frac{\pi}{4\sigma})}|\Gamma\left(\frac{1}{4}+\frac{i}{4\sigma}\right)|^{-4}. \qquad \square$$

From Lemma 9 we see that, for $f=1$, $Q(r)$ is unbounded at some point $r=r_0>0$. Further results on the behavior of Q can be received from the derivative $\frac{dQ}{d\xi}$. Using (114), a straightforward calculation yields

$$\begin{aligned} Z :=& \frac{5f}{3}(48H_1+\nu H_3)^2\frac{dQ}{d\xi} = 72(\lambda+\mu)H_1^2 + \lambda\left(72-\frac{3}{2}\nu\right)H_1H_2 \\ &- \left(48\mu-\frac{3}{2}\lambda\nu\right)H_1H_3 + \frac{5}{2}\lambda\nu H_2H_3 - 24H_0(\lambda H_2+\mu H_3). \end{aligned} \tag{131}$$

From this expression, combined with Lemmata 7 and 8, we obtain

$$\left.\begin{aligned} &(132) \qquad \lim_{\xi\to\infty} Z = f\lambda(48+\nu)L_1L_2 \\ &(133) \qquad \lim_{r\to\infty}\frac{1}{r}\frac{dQ}{d\xi} = \lim_{r\to\infty}\frac{dQ}{dr} = \frac{3\lambda}{5(48+\nu)}\frac{L_2}{L_1} < 0 \end{aligned}\right\} \text{ for } f=\pm1,$$

$$\lim_{\xi\to0}\frac{dQ}{d\xi} = \lim_{r\to0}\frac{1}{2ar}\frac{dQ}{dr} = \frac{3(\lambda+\mu)}{320} > 0 \quad \text{if } f=1, \tag{134}$$

$$\lim_{r\to0}\frac{1}{-r\ln r}\frac{dQ}{dr} = \frac{18\sqrt{2}\sigma^2}{5}\frac{\lambda l_2+\mu l_3}{(48l_1+\nu l_3)^2} \quad \text{if } f=-1. \tag{135}$$

Let us now investigate the case $f=-1$ more closely. The integrals involved in l_2, l_3 (see (120), (121)) can be calculated numerically, and it turns out that

$$\lambda l_2 + \mu l_3 < 0 \quad \text{for each } \beta \in (2, \tfrac{8}{3}). \tag{136}$$

Thus, by (135),

$$\lim_{\xi\to0}\frac{dQ}{d\xi} = -\infty \qquad (f=-1), \tag{137}$$

and the function Z, defined in (131), satisfies

$$\lim_{\xi\to0} Z = \infty, \quad \lim_{\xi\to\infty} Z = -\lambda(48+\nu)L_1L_2 > 0 \qquad (f=-1). \tag{138}$$

This suggests trying to show that $Z>0$, or equivalently, $dQ/d\xi<0$ for all $\xi>0$ and $f=-1$. For this purpose we look for positive constants c_1, c_2 such that

$$H_2(\xi)+c_1H_3(\xi) < 0, \quad c_2H_3(\xi)-H_1(\xi) < 0 \quad \text{for all } \xi>0 \; (f=-1). \tag{139}$$

Note that, from (107), we necessarily have $c_2<1$. Moreover, it seems useful to introduce

$$P_1 := \frac{\sigma}{\sqrt{8aG^2(s)}}\int_{\varphi_0}^{\varphi_1} h^{-1}(\varphi,s)e^{-2\sigma^{-1}\varphi}d\varphi > 0, \tag{140}$$

$$P_2 := \frac{\sigma}{\sqrt{8aG^2(s)}}\int_{\varphi_0}^{\varphi_1} h(\varphi,s)e^{-2\sigma^{-1}\varphi}d\varphi > 0. \tag{141}$$

Then (107) and a short calculation yield

$$H_3 - H_1 = P_2 \qquad (f=-1), \tag{142}$$

$$H_0 - 2H_1 + H_3 = P_1 \qquad (f=\pm1). \tag{143}$$

Now, eliminating H_0 from Z by (143), and using (142) and the inequalities

$$H_1^2 > c_2 H_1 H_3 \,, \quad -H_1 H_3 > c_1^{-1} H_1 H_2 \qquad (f = -1), \tag{144}$$

we obtain

$$\begin{aligned} Z =&72(\lambda + \mu)H_1^2 + 24\mu H_3^2 + (24\lambda - \tfrac{3}{2}\lambda\nu)H_1H_2 - (96\mu - \tfrac{3}{2}\lambda\nu)H_1H_3 \\ &+ \lambda(24 + \tfrac{5}{2}\nu)H_2H_3 - 24P_1(\lambda H_2 + \mu H_3) \\ >&[72c_2(\lambda + \mu) - 72\mu + \tfrac{3}{2}\lambda\nu]H_1H_3 + (48\lambda + \lambda\nu)H_1H_2 \\ &+ 24\mu H_3 P_2 + \lambda(24 + \tfrac{5}{2}\nu)H_2P_2 - 24P_1(\lambda h_2 + \mu H_3) \,. \end{aligned} \tag{145}$$

Since $c_2 < 1$, we have $72c_2(\lambda + \mu) - 72\mu + \frac{3}{2}\lambda\nu < \frac{3}{2}\lambda(48 + \nu) < 0$. Thus, by (144) we further obtain

$$Z > c_1^{-1}\lambda \underbrace{(48 + \nu)H_1H_2}_{>0} \cdot W(c_1, c_2) + R \,, \tag{146}$$

where

$$W(c_1, c_2) := c_1 - \frac{72(\lambda + \mu)}{\lambda(48 + \nu)}(c_2 - 1) - \tfrac{3}{2} \,, \tag{147}$$

$$R := \underbrace{24\mu H_3 P_2}_{>0} + \underbrace{\frac{\lambda}{2}(48 + 5\nu)H_2P_2}_{>0} - 24P_1(\lambda H_2 + \mu H_3) \,. \tag{148}$$

Before continuing here, we have to determine the constants c_1, c_2. For this purpose, we use the following lemma.

LEMMA 10. *Let l_1, l_2, l_3 be the integrals defined in* (119)–(121), *where $\beta \in (2, \frac{8}{3})$ is arbitrarily fixed.*

(a) *If $c_1 \geq \frac{3}{2}$ is such that $l_2 + c_1 l_3 < 0$, then*

$$H_2(\xi) + c_1 H_3(\xi) < 0 \quad \textit{for all } \xi > 0 \qquad (f = -1) \,.$$

(b) *If c_2 satisfies $\frac{1}{3} \leq c_2 < 1$ and $c_2 l_3 - l_1 < 0$, then*

$$c_2 H_3(\xi) - H_1(\xi) < 0 \quad \textit{for all } \xi > 0 \qquad (f = -1) \,.$$

Proof. (a) Setting $F(\xi) := e^{-\frac{\xi}{2}}(H_2(\xi) + c_1 H_3(\xi))$ and using (114), (142), we calculate

$$\begin{aligned} \frac{dF}{d\xi} &= \tfrac{1}{2}e^{-\frac{\xi}{2}}[3(1 + c_1)H_1(\xi) - 5c_1H_3(\xi)] \\ &< \tfrac{1}{2}e^{-\frac{\xi}{2}}(3 - 2c_1)H_3(\xi) \leq 0 \,. \end{aligned}$$

Since $\lim_{\xi \to 0+} F(\xi) = l_2 + c_1 l_3 < 0$, this implies $F(\xi) < 0$ for all $\xi > 0$ and thus proves the assertion.

(b) Using a similar trick, we set $\eta := 2 - 1/(2c_2)$ and $\widetilde{F}(\xi) := e^{\eta\xi}(c_2 H_3(\xi) - H_1(\xi))$. From (114) and (143) we obtain

$$\frac{d\widetilde{F}}{d\xi} = -\tfrac{1}{2}e^{\eta\xi}\left[P_1 + \frac{1}{c_2}\underbrace{(-3c_2^2 + 4c_2 - 1)}_{\geq 0}H_1\right] < 0 \,.$$

Now the assertion follows as in the proof of (a). □

Numerical calculations have shown $l_2 + 6l_3 < 0$ and $\frac{1}{2}l_3 - l_1 < 0$. Thus, by the preceding lemma, (139) holds for

$$c_1 = 6 \,, \qquad c_2 = \tfrac{1}{2} \,.$$

Now a simple estimation yields $W(6,\frac{1}{2}) > 0$. Moreover, from (106) we see that $\mu < 6\lambda$, which, together with $H_2 + 6H_3 < 0$, gives $\lambda H_2 + \mu H_3 < 0$. Thus $R > 0$, too, and hence $Z > 0$. We have proved the following lemma.

LEMMA 11. *In the case $f = -1$, $dQ/d\xi$ is negative for all $\xi > 0$ and arbitrarily fixed $\beta \in (2, \frac{8}{3})$. Moreover, $\lim_{s\to 0+} Q =: M < 0$.*

Let us now consider the case $f = 1$. It seems difficult to attain analytical results on the behavior of Q beyond those already stated in Lemma 9 and in (133), (134). But extensive numerical computations (carried out by Frank Rehrmann) produced the following results.

(149) *Numerical results for $f = 1$, $\beta \in (2, \frac{8}{3})$.* The denominator $\mathcal{D}$ of $Q(s)$ vanishes at exactly one point $s_0 < 0$. For $s_0 < s < 0$, $Q(s)$ is strictly decreasing from ∞ to 0. Moreover, $dQ(s)/ds$ vanishes at exactly one point $s_1 < s_0$; $Q(s)$ is strictly increasing in $(-\infty, s_1)$ from $-\infty$ to $m := Q(s_1) < 0$, and strictly decreasing in (s_1, s_0) from m to $-\infty$.

Figure 10 shows the graph of $Q(s)$ for $\beta = 2{,}3$ ($f = \pm 1$). For other values of β, $Q(s)$ behaves quite similarly: As β increases from 2 to $\frac{8}{3}$, s_0 ranges from $\approx -2{,}2$ to $\approx -1{,}9$, s_1 from $\approx -2{,}93$ to $\approx -2{,}95$, m from ≈ -41 to ≈ -34, and $M := \lim_{s\to 0+} Q(s)$ from $\approx -64{,}3$ to $\approx -18{,}8$.

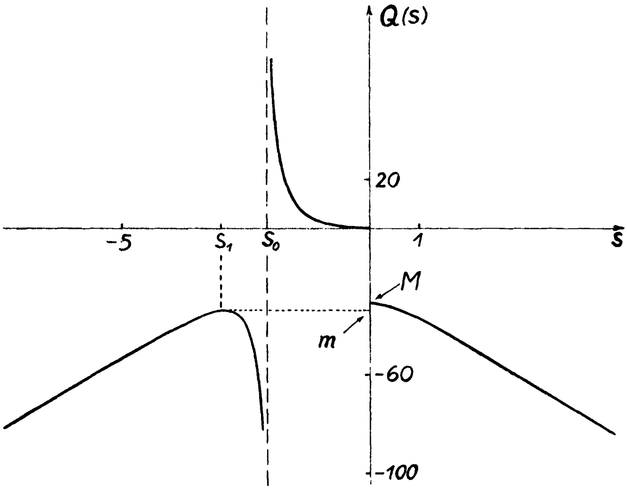


FIG. 10. *The quotient $Q(s)$ for $\beta = 2,3$.*

11. A modified scaling. Before pronouncing the main result on the unfolding of the degenerate bifurcation at the cusp point, we will briefly revisit the scaling chosen in (64). This scaling is unsatisfactory in a double sense. First, an admissible range of (ε, τ)

(150) $$0 < \varepsilon \le \varepsilon_0\,, \quad |\tau| \le \tau_0 \quad (\varepsilon_0, \tau_0 \text{ sufficiently small})$$

leads to an admissible range

(151) $$|p| \le \sqrt{|q|}\,\tau_0\,, \qquad 0 < |q| \le \varepsilon_0^2$$

in the (p,q)-plane which, obviously, is not just a full neighborhood of (0,0) with p-axis removed. The second drawback refers to the admissible range of (u,v) and, thus, of r. While the results of §§6 and 7 are valid only for (u,v) in a bounded domain and (ε,τ)

sufficiently small, Lemma 7 contains results for $r \to \infty$ and thus, for arbitrarily large v. To remedy this we follow the same procedure as in [4, p. 78]: In (64), we replace the scaling $\varepsilon = \sqrt{|q|}$ by

$$\varepsilon = \kappa^{-1}\sqrt{|q|}\,, \qquad 0 < \kappa \leq 1 \tag{152}$$

with a new parameter κ. Repeating the former calculations, it turns out that all the formulae and definitions in §§6–8 remain valid, if we simply replace f, where it occurs explicitly, by $f\kappa^2$. Moreover, in (65), the functions $R_i(u, v, \varepsilon, \tau)$ become real-analytic functions $R_i(u, v, \varepsilon, \tau, \kappa)$, at least quadratic both with respect to (u, v) and (ε, τ), which are again bounded for (u, v) in an arbitrary bounded domain, $0 < \kappa \leq 1$, and (ε, τ) as in (150). Now, the admissible range of (p, q) is

$$|p| \leq \kappa^{-1}\sqrt{|q|}\,\tau_0\,, \qquad 0 < |q| \leq \kappa^2\varepsilon_0^2\,. \tag{153}$$

Therefore, given $|q| \in (0, \varepsilon_0^2]$, if κ varies in $[\varepsilon_0^{-1}\sqrt{|q|}, 1]$, then p covers the interval $[0, \varepsilon_0\tau_0]$; hence, (p, q) covers a full neighborhood of (0,0) except $q = 0$. As for the second drawback mentioned above, we note that after the new scaling, all the former statements are again valid for any bounded region $\mathcal{B} : |u| < u_0, |v| < v_0$. In the original scaling, $\mathcal{B}$ corresponds to $|u| < \kappa^{-1}u_0, |v| < \kappa^{-2}v_0$, and since $\kappa \to 0$ is allowed, the original statements are valid for any $(u, v) \in \mathbb{R}^2$.

12. The unfolding in the cusp point.

From the results in the preceding sections we now derive the following theorem concerning the unfolding of the degenerate codimension-2 bifurcation in the cusp point. Recall that for this phenomenon to occur, the parameters (α, β) have to satisfy (54) and, given β, the parameters (γ, ζ) must be sufficiently close to $(9\beta, 3\beta)$, or equivalently, (p, q) must be sufficiently close to (0,0).

THEOREM 2. *In a neighborhood $\mathcal{U}$ of the origin in the (p, q)-parameter plane there exist curves*

$$\begin{aligned} L_1 &: p = Mq + \mathcal{O}(|q|^{\frac{3}{2}})\,, \quad q < 0 \\ L_2 &: p = -mq + \mathcal{O}(q^{\frac{3}{2}})\,, \quad q > 0 \end{aligned} \qquad (m, M \textit{ as in } (149) \textit{ and Lemma } 11)$$

and a subdivision of $\mathcal{U}$ into regions I–V *generated by L_1, L_2, the p-axis, and the positive q-axis* (see Fig. 11) *such that for system* (58) *the following statements hold locally, i.e., in a neighborhood of $(w_1, w_2) = (0,0)$ (for an illustration of these statements, see Fig.* 12):

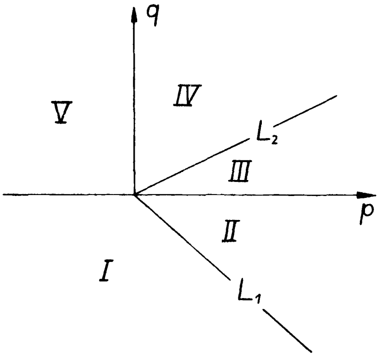


FIG. 11. *Subdivision of $\mathcal{U}$.*

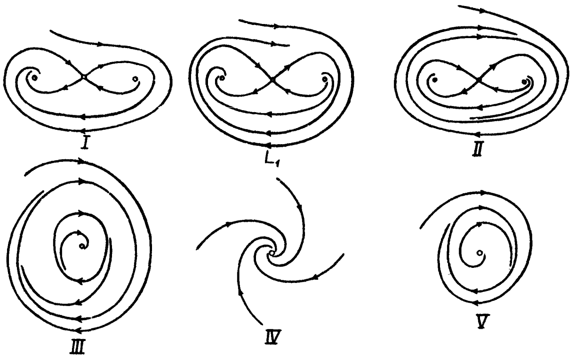


FIG. 12. *Phase portraits for the regions of Fig.* 11.

(a) $(p,q) \in$ I: *There is no periodic orbit, but there exist two heteroclinic orbits joining the saddle* (0,0) *to the sink in the left half plane.*

(b) $(p,q) \in L_1$: *There is a homoclinic orbit joining* (0,0) *to itself and surrounding the two other equilibria; it is asymptotically stable* (*from outside*).

(c) $(p,q) \in$ II: *There is exactly one periodic orbit; it encircles all three equilibria and is asymptotically stable. It approaches the homoclinic orbit just mentioned, if* (p,q) *tends to a point on* L_1. *Its size* (*e.g., its diameter*) *is of order* $\tilde{p}$ *as* (p,q) *tends to the point* $(\tilde{p},0)$. *Moreover, there exist two heteroclinic orbits joining the source in the right half plane to the saddle* (0,0).

(d) $(p,q) \in$ III: *There are exactly two periodic orbits; the outer one is asymptotically stable, while the inner one is unstable. As* (p,q) *tends to* $(\tilde{p},0)$, *the size of the outer periodic orbit is of order* $\tilde{p}$, *while the size of the inner one decreases to zero linearly with* q. *On the other hand, if* (p,q) *crosses the curve* L_2 *at* $(\tilde{p},\tilde{q})$, *the two periodic orbits first coalesce to a single periodic orbit—with size of order* $\tilde{q}$*—and then disappear.*

(e) $(p,q) \in$ IV: *No periodic orbit exists. As* (p,q) *approaches and crosses the positive* q*-axis transversely, supercritical Hopf bifurcation occurs.*

(f) $(p,q) \in$ V: *Exactly one periodic orbit exists; it is asymptotically stable. It shrinks down to a point as* (p,q) *tends either to* $(0,\tilde{q})$ *or to* $(\tilde{p},0)$, $(\tilde{q} > 0,\ \tilde{p} < 0)$. *More precisely, its size decreases like* $\sqrt{-p}$ *in the first limit, and like* q *in the second one.*

Proof. First we note that for $q \neq 0$, system (58) is equivalent to (65); hence we may use all the results on system (65) derived in the preceding sections. Moreover, by (64), the curves L_1, L_2 correspond to the curves $\tau = Q(s)\varepsilon + \mathcal{O}(\varepsilon^2)$ (see (83)) with $s \to 0+$ and $s = s_1$, respectively. We also note that any (nonconstant) periodic orbit of (65) must surround all three equilibria and, therefore, crosses the positive v-axis. The assertions on the precise number of period orbits now follow from Lemma 5 combined with Lemma 11 in case $q < 0$, while in case $q > 0$, the analytical results of Lemmata 5 and 9 do not suffice for their proof; we also need the numerical results (149). As for the size of the periodic orbits starting at $(w_1, w_2) = (0, A^{-1}|q|r)$, most statements are easy consequences of (64) and the limit relations (133), (134). But the two assertions concerning their shrinkage of order q again use the numerical results (149). The existence of a homoclinic orbit follows from Lemma 5, since Q_0 in (85) is easily seen to coincide with M. The fact that this homoclinic orbit is the limit of the periodic orbits in II if (p,q) tends to L_1 is proved along the same lines as analogous results in [10, p. 339] and [4, p. 70]. To prove the asymptotic stability of the homoclinic orbit we apply [5, p. 357]: Inserting $f = -1$ and $\tau = Q_0\varepsilon + \mathcal{O}(\varepsilon^2)$ into the right-hand side of (65), the trace of its linearization at $(u,v) = (0,0)$ is $Q_0\varepsilon + \mathcal{O}(\varepsilon^2)$, which is negative since $Q_0 = M < 0$ and $\varepsilon > 0$. Hence, the homoclinic orbit is asymptotically stable. Next we prove the stability assertions for the periodic orbit $\bar{\gamma}$ of system (65) starting at $(u,v) = (0,\bar{r})$, where $\bar{r} > 0$, $\varepsilon > 0$ are fixed and $\tau = Q(\bar{r})\varepsilon + \mathcal{O}(\varepsilon^2)$. Let $r > 0$ be sufficiently close to $\bar{r}$, and recall the definitions of $I(U,V)$, $N(U,V)$, $H(t)$, and $K = K(r,\varepsilon,\tau)$ in (68), (70), (73), (78). Obviously, $K(\bar{r},\varepsilon,\tau) = 0$, and from $\partial I(U,V)/\partial V = 2a\frac{V}{N} < 0$ for $V < 0$, we easily see that $\bar{\gamma}$ is asymptotically stable (respectively, unstable) if $(r-\bar{r})K < 0$ (respectively, > 0), or a fortiori if $dK/dr|_{r=\bar{r}} < 0$ (respectively, > 0). Now, using the previous notation we obtain

$$K(r,\varepsilon,\tau) = 4a\varepsilon\left\{ Q(\bar{r}) \underbrace{\int_{\Gamma} N^{-1}(V+cU^2)dU}_{=:\,J_1(r)} \right.$$
$$\left. + \underbrace{\int_{\Gamma} N^{-1}(bU^2V - fdU^2 + eU^4)dU}_{=:\,J_2(r)} \right\} + \mathcal{O}(\varepsilon^2)\,.$$

Owing to $Q(r) = -J_2(r)/J_1(r)$ (see (82)), we calculate

$$\frac{dK}{dr}|_{r=\bar{r}} = -4a\varepsilon J_1(\bar{r}) \cdot \frac{dQ}{dr}|_{r=\bar{r}} + \mathcal{O}(\varepsilon^2)\,.$$

The desired stability properties now follow from (111), Lemma 9, Lemma 11, and—in case $f = 1$—the supplementary numerical results (149). As for the heteroclinic orbits, we assume $(p,q) \in \mathrm{I} \cup \mathrm{II}$ and $r > 0$ arbitrarily fixed. By essentially repeating the preceding calculation, we obtain

$$K(r,\varepsilon,\tau) = 4a\underbrace{J_1(r)}_{<0}(\tau - \varepsilon Q(r)) + \mathcal{O}(\varepsilon^2 + \tau^2)\,.$$

Now, in case $(p,q) \in \mathrm{I}$, we have $\tau - \varepsilon Q(r) > 0$ and thus $K(r,\varepsilon,\tau) < 0$. If $(p,q) \in \mathrm{II}$, there exists $\tilde{r} > 0$ sufficiently small such that $\tau - \varepsilon Q(\tilde{r}) < 0$ (see Lemma 11); hence $K(\tilde{r},\varepsilon,\tau) > 0$. By (71), this implies the existence of a positively invariant compact region $\mathcal{P}$ in case $(p,q) \in \mathrm{I}$, and a negatively invariant compact region $\mathcal{N}$ in case $(p,q) \in \mathrm{II}$, as shown in Fig. 13. The existence of heteroclinic orbits as claimed in the theorem now easily follows by standard arguments using Poincaré–Bendixson theory, the hyperbolicity of the three equilibria, and the fact that $\mathcal{P}$, $\mathcal{N}$ contain no nonconstant periodic orbit. Hopf bifurcation of system (65) with $f = 1$ and parameter τ at $(u,v,\varepsilon,\tau) = (0,0,\varepsilon,0)$ is verified in the usual way (see [8]). Since we already know that it is supercritical with asymptotically stable Hopf cycles, we omit the details. Finally, to prove the phenomena claimed when (p,q) crosses L_2, we again rely on the computations (149). This ends the proof of Theorem 2. □

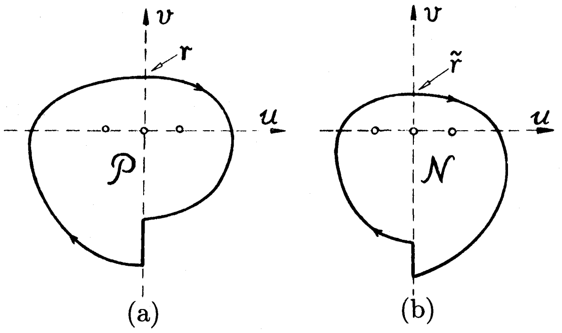


FIG. 13. *Illustrating the proof of Theorem* 2. (a) $(p,q) \in \mathrm{I}$, (b) $(p,q) \in \mathrm{II}$.

13. Concluding remarks. (1) The phenomenon described in Theorem 2, which occurs as (p,q) crosses the p-axis upwards, can be regarded as a counterpart to Hopf bifurcation: Three equilibria first coalesce to a single equilibrium, from which a family of periodic orbits then emanate. As we have seen, the growth of these periodic orbits is not of order $\sqrt{q}$ as in standard Hopf bifurcation, but only of order q.

(2) It is easily verified that, in addition to the heteroclinic orbits addressed in Theorem 2, system (58) has a continuum of heteroclinic orbits connecting the two equilibria $\neq$ (0,0); this is true for any (p,q) sufficiently small ($q \neq 0$).

(3) Returning to the surface $\mathcal{M}$ in (γ,δ,ζ)-space defined by (7), with (α,β) satisfying (54), the results of Theorem 2 and the phase portraits given in Fig. 12 may be carried over to a neighborhood $\mathcal{U}'$ of the cusp point C_0 of $\mathcal{M}$. Looking at $\mathcal{U}'$ parallel to the negative δ-axis, we obtain a subdivision of $\mathcal{U}'$ which is described qualitatively in Fig. 14. Here, the phase portraits in $\mathrm{I}', L_1', \cdots$ are equivalent to those in $\mathrm{I}, L_1, \cdots$. Moreover, Fig. 15 shows a vertical projection of $\mathcal{U}'$ onto the original parameter plane (γ,δ) with corresponding decomposition $\mathrm{I}'', L_1'', \cdots$. It is interesting to note that the restriction $|p| \leq \sqrt{|q|}\,\tau_0$, given in (151), is essential here for $q > 0$, since otherwise, there would exist parameters (p,q), $q > 0$—corresponding in Fig. 15 to the points of III$''$ and V$''$ between C_+ and C_-—for which system (58) has three equilibria, contrary to our previous results.

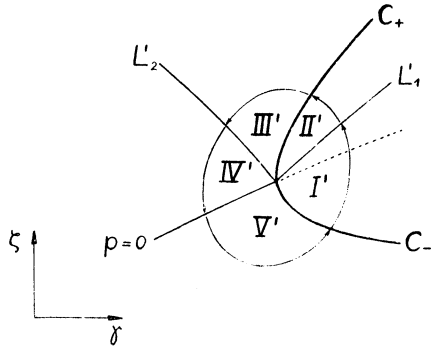


FIG. 14. *The neighborhood* $\mathcal{U}'$ *of* C_0.

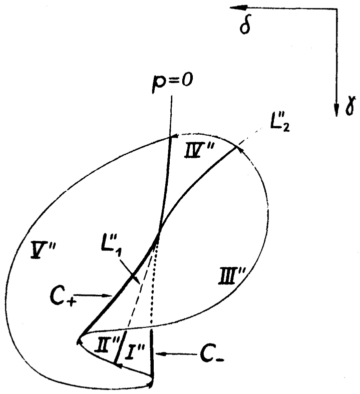


FIG. 15. *Vertical projection of* $\mathcal{U}'$.

(4) The local phase portraits of Fig. 12 could be translated back to the original (x, y)-plane, via the near identity transformation (56) and the orientation reversing transformation (24). Figures 16 and 17 show such phase portraits corresponding to regions II, respectively III. Note that all statements were of local nature such that, e.g., in Fig. 16, only the flow in a common neighborhood of the three inner equilibria is guaranteed.

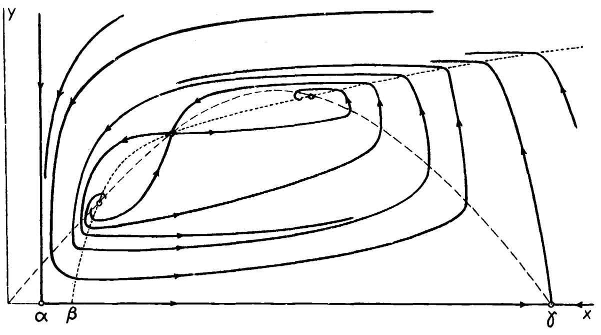


FIG. 16. *Phase portrait for region* II *in the* (x, y)*-plane. The curves* (5) *are indicated, too.*

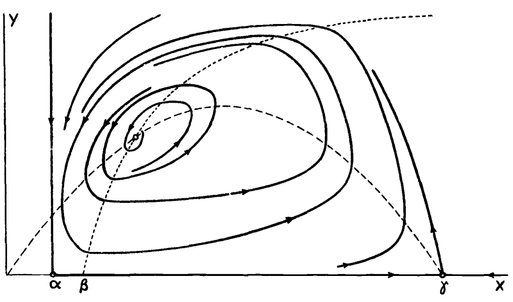


FIG. 17. *Phase portrait for region* III *in the* (x, y)*-plane.*

(5) The parts of this paper concerning the unfolding of the degenerate codimension-2 bifurcation in the cusp point can be treated in a more general setting. This has been done and will be published separately.

REFERENCES

[1] H. AMANN, *Gewöhnliche Differentialgleichungen*, deGruyter, Berlin, New York, 1983.

[2] A. D. BAZYKIN, *Structural and dynamic stability of model predator-prey systems*, Internat. Inst. Appl. Systems Anal., Laxenburg, Austria, 1976.

[3] A. D. BAZYKIN, F. S. BEREZOVSKAYA, G. A. DENISOV, AND YU. A. KUZNETZOV, *The influence of predator saturation effect and competition among predators on predator-prey system dynamics*, Ecological Model., 14 (1981), pp. 39–57.

[4] J. CARR, *Applications of centre manifold theory*, Appl. Math. Sci., 35, Springer-Verlag, New York, Berlin, 1981.

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

[6] W. GRÖBNER AND N. HOFREITER, *Integraltafel* II. *Teil: Bestimmte Integrale*, Springer-Verlag, New York, 1966.

[7] J. GUCKENHEIMER, *Multiple bifurcation problems of codimension two*, SIAM J. Math. Anal., 15 (1984), pp. 1–49.

[8] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, Appl. Math. Sci., 42, Springer-Verlag, New York, Berlin, 1983.

[9] J. HAINZL, *Stability and Hopf bifurcation in a predator-prey system with several parameters*, SIAM J. Appl. Math., 48 (1988), pp. 170–190.

[10] N. KOPELL AND L. N. HOWARD, *Bifurcations and trajectories joining critical points*, Adv. in Math., 18 (1975), pp. 306–358.

[11] J. A. SANDERS AND R. CUSHMAN, *Abelian integrals and global Hopf bifurcations*, in Dynamical systems and bifurcations, Lecture Notes in Math. 1125, Springer-Verlag, New York, 1985, pp. 87–98.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 181–200, January 1992

009

QUALITATIVE ANALYSIS OF ONE- OR TWO-SPECIES NEUTRAL DELAY POPULATION MODELS*

YANG KUANG†

Abstract. In this paper neutral delay models of single population growth, predator-prey, and competition interactions are introduced and investigated. These systems are more general than previous ones by allowing per capita growth rates to be nonlinear and delays to be of the general distributed type. Conditions are given for solutions of these systems to be bounded for proper initial functions. For neutral delay single population models, sufficient conditions for solutions tending to the positive steady states are also presented.

Key words. qualitative analysis, neutral delay equation, predator-prey system, competition system, distributed delay

AMS(MOS) subject classifications. 34K15, 34K20, 92A15

1. Introduction.

The autonomous logistic delay differential equation

$$\dot{x}(t) = rx(t)[1 - x(t-\tau)/K], \tag{1.1}$$

where "$\cdot$"$= d/dt$, r, K, τ are positive constants, has been widely used as a model equation capable of showing oscillations of single-species population sizes in constant environments closed to both immigration and emigration (see Cushing [2], Gopalsamy and Zhang [10], Hale [13], Kuang and Feldstein [19], and Pielou [27]). It has been the object of intensive analysis by numerous authors (see the references cited in [10]). Indeed, it is a natural generalization of the following well-known logistic single-species population equation:

$$\dot{x}(t) = rx(t)[1 - x(t)/K]. \tag{1.2}$$

Here r is called the intrinsic growth rate of the species x, K is interpreted as the environment carrying capacity for x, and $r[1 - x(t)/K]$ is the per capita growth rate of x at time t. Based on his investigation on laboratory populations of Daphnia magna, F. E. Smith [28] argued that a growing population will use food faster than a saturated one; thus the per capita growth rate in (1.2) should be replaced by $r[1 - (x(t) + \rho\dot{x}(t))/K]$ (for details see Pielou [27, pp. 38–40]). This leads to the following equation:

$$\dot{x}(t) = rx(t)[1 - (x(t) + \rho\dot{x}(t))/K]. \tag{1.3}$$

We may think of x as a species grazing upon vegetation, which takes time τ to recover. In this case, it will be even more realistic to incorporate a single discrete delay τ in the per capita growth rate, which results in the following neutral delay logistic equation

$$\dot{x}(t) = rx(t)[1 - (x(t-\tau) + \rho\dot{x}(t-\tau))/K]. \tag{1.4}$$

This equation was first introduced and investigated by Gopalsamy and Zhang [10]. Subsequently, it was studied by Freedman and Kuang [6], and Kuang and Feldstein

*Received by the editors July 25, 1990; accepted for publication (in revised form) April 5, 1991. This research was partially supported by a College of Liberal Arts and Sciences Summer Research Award at Arizona State University.

†Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804.

[19]. The focus of these works was the qualitative behavior of the solutions, such as boundedness, asymptotic stability, and oscillation. In a recent paper Gopalsamy, He, and Wen [9] studied the existence and linear asymptotic stability of periodic solutions of equation (1.4), when r, ρ, and K are replaced by periodic functions of period ω, and $\tau = n\omega$ for some positive integer n.

Assume the population $x(t)$ described by (1.4) is a prey species, and suppose there exists a predatory species $y(t)$ that preys on species $x(t)$; then it is natural to propose the following mathematical model to describe their interaction:

$$\left\{\begin{aligned} \dot{x}(t) &= rx(t)[1 - (x(t-\tau) + \rho\dot{x}(t-\tau))/K] - y(t)p(x(t)), \\ \dot{y}(t) &= y(t)[-\alpha + \beta p(x(t-\sigma))]. \end{aligned}\right. \tag{1.5}$$

Here α, β, and σ are all positive constants, and $p(x)$ is the predator response function for the predator species y with respect to the prey species x. A slightly more general version of (1.5) was introduced and studied in Kuang [16], where the focus of the study was the local stability and oscillation analysis of system (1.5). An even more general version of (1.5) was proposed and studied in [18] where sufficient conditions were obtained for its solutions to be bounded.

Assume $x(t)$ described by (1.4) is the population of a species competing with another species with population $y(t)$ for a shared limited resource—space or a nutrient, for example; then the following system may model their interaction:

$$\left\{\begin{aligned} \dot{x}(t) &= r_1x(t)[1 - k_1x(t) - ax(t-\tau_1) - \beta\dot{x}(t-\tau_0) - c_1y(t-\tau_2)], \\ \dot{y}(t) &= r_2y(t)[1 - c_2x(t-\tau_3) - k_2y(t-\tau_4)]. \end{aligned}\right. \tag{1.6}$$

Here all parameters except β are assumed to be positive constants. We have included $k_1x(t)$ into the per capita growth rate of $x(t)$, which may reflect the possible instantaneous interference within species x. System (1.6) was first introduced and studied in Kuang [17]. Again the focus of that work was the local stability and oscillatory analysis of system (1.6). Sufficient conditions for solutions of (1.6) to be bounded can be found in [18].

Although extensive literature exists on functional differential equations and their applications (cf. [2], [3], [7], [13], [29]), works on global asymptotic stability and boundedness of solutions for nonlinear equations or systems are relatively few. Most of these existing results only apply to systems with special kernel in the distributed delays (cf. the references cited in [1], [2], [5], [20]), or with a strong nondelayed self-crowding effect (e.g., [8], [12], [22]–[25]). However, these requirements are rather artificial and restrictive, and very few real systems may satisfy them. Thus, there is an urgent need to study these questions for more general and realistic models. An effort along this line was documented in [11], [21].

As the reader may already be aware, many real systems are quite sensitive to sudden changes. This fact may suggest that proper mathematical models of the systems should consist of some neutral delay equations. Even though the delay lengths may be short, and the neutral terms relatively small, it is still necessary, for the sake of rigorousness, to justify that the neutral term effects are not important. Indeed, most of the time we may find that neutral term effects can be quite significant. This is largely due to the fact that neutral delay equations are not structurally stable in the sense that the introduction of neutral delay terms may destabilize an asymptotically stable equilibrium. For example, $3\dot{x} = -x$ has a globally asymptotically stable trivial solution, while the same solution in $\dot{x} + 2\dot{x}(t-\tau) = -x$ becomes unstable for any $\tau > 0$. This is because the corresponding characteristic equation of the neutral equation may have roots bifurcating from infinity [6], [13], a phenomenon that cannot occur in retarded equations. This indicates that it is important to deal with neutral delay equations in real mathematical models.

Although neutral delay equations have been studied extensively (see the references cited in Hale [13]), most of the existing works are related to fundamental questions such as existence, uniqueness, local stability analysis, etc. Recently, there seems to be a growing interest in the oscillation theory for neutral delay equations (cf. [7], [10], [16], [17]). The literature on global asymptotic stability for general nonlinear neutral delay equations is scarce. The main difficulty in this kind of analysis is probably the lack of compactness for bounded solutions (indeed, even the boundedness of solutions is usually hard to establish). Frankly speaking, the results to be presented in this paper are rather limited and primitive.

This paper is organized as follows: In the next section the system is described in detail and some preliminary results are presented. Section 3 deals with boundedness of solutions of a scalar neutral equation and §4 contains a discussion of domain of attractivity for the unique positive steady states. Section 5 considers the boundedness problem for the full system. The final section is devoted to discussion and presents some numerical simulations.

2. Preliminaries. In this paper we propose to study the following general neutral delay systems, which may model a two-species interaction (x and y) in a closed environment and include both systems (1.5) and (1.6) as special cases:

$$\begin{aligned}\dot{x}(t) =& x(t)\left[\int_{-\tau_1}^{0} g(x(t+s))d\mu_1(s) - \rho\int_{-\tau_2}^{0} \dot{x}(t+s)d\mu_2(s)\right.\\ &\left. - q(x(t))\int_{-\tau_3}^{0} y(t+s)d\mu_3(s)\right],\\ \dot{y}(t) =& y(t)\left[a + b\int_{-\tau_4}^{0} x(t+s)q(x(t+s))d\mu_4(s) - c\int_{-\tau_5}^{0} y(t+s)d\mu_5(s)\right],\end{aligned} \tag{2.1}$$

where τ_i, $i = 1, \cdots, 5$, ρ and c are nonnegative constants, and a and b are real numbers. We always assume the following:

(H1) $\mu_i(s)$ is nondecreasing and $\int_{-\tau_i}^{0} d\mu_i(s) = 1$, $i = 1, \cdots, 5$;

(H2) $g(x)$ is continuously differentiable such that $g(0) > 0$, $g'(x) < 0$ for $x \geq 0$, and $g(1) = 0$;

(H3) $p(x) = xq(x)$ is continuously differentiable, $p(0) = 0$, $p'(x) > 0$ for $x \geq 0$, and $\lim_{x\to+\infty} p(x) > |a/b|$.

Clearly, when $a < 0$, $b > 0$, $c = 0$, and $g(x) = r(1-x)$, then system (2.1) reduces to a slightly more general form of (1.5). We may refer to this resulting system as Gause-type neutral delay predator-prey system (cf. [16]). When $a > 0$, $b < 0$, we see (2.1) has (1.6) as a special case. System (2.1) is more general than the one considered in [18] in the following two aspects: (i) we allow the per capita growth rate of x to be nonlinear when y is absent, (ii) all delays are of the distributed type.

Let $\tau_0 = \max\{\tau_i : i = 1, 2, \cdots, 5\}$ and $R^+ = \{r : r \geq 0\}$. We always assume that the initial conditions for (2.1) are of the type

$$\begin{aligned} x(s) = \phi_1(s) \geq 0, s \in [-\tau_0, 0], \phi_1(0) > 0 \quad \text{and} \quad \phi_1 \in C^1([-\tau_0, 0], R^+),\\ y(s) = \phi_2(s) \geq 0, s \in [-\tau_0, 0], \phi_2(0) > 0 \quad \text{and} \quad \phi_2 \in C^1([-\tau_0, 0], R^+).\end{aligned} \tag{2.2}$$

We say $(x(t), y(t))$ is a solution of (2.1) on $[-\tau_0, \infty)$ if both $x(t)$ and $y(t)$ are positive continuously differentiable functions and satisfy both the above initial conditions and system (2.1).

We note that the first equation of system (2.1) can be rewritten as (provided $x(t) > 0$)

$$\left[\ln x(t) + \rho \int_{-\tau_2}^{0} x(t+s)d\mu_2(s)\right]' = \int_{-\tau_1}^{0} g(x(t+s))d\mu_1(s) - q(x(t)) \int_{-\tau_3}^{0} y(t+s)d\mu_3(s). \tag{2.3}$$

By letting $u(t) = \ln x(t)$, we can see that system (2.1) falls into the class of neutral systems considered in Hale [13, Chap. 12]. Thus, local existence, uniqueness, and continuous dependence of solutions are guaranteed [13, §12.2]. The following result ensures that the solution $(x(t), y(t))$ of (2.1) and (2.2) is positive and exists for all $t \geq 0$.

PROPOSITION 2.1. *Assume* (H1)–(H3) *in* (2.1). *Then the solution* $(x(t), y(t))$ *of* (2.1) *and* (2.2) *is positive and exists for all* $t > 0$. *Moreover, there exist positive constants* A *and* B (*depending on initial functions* ϕ_1 *and* ϕ_2), *such that* $x(t) < Ae^{Bt}$.

Proof. Suppose the maximal interval of existence for $x(t)$ and $y(t)$ is $[0, \omega)$. The positivity of $x(t)$ and $y(t)$ follows from the fact that the system is of Kolmogorov form. Assume first $\omega < +\infty$. Then equation (2.3) implies that, for $t \in [0, \omega)$,

$$\left[\ln x(t) + \rho \int_{-\tau_2}^{0} x(t+s)d\mu_2(s)\right]' < g(0),$$

which leads to

$$\ln x(t) + \rho \int_{-\tau_2}^{0} x(t+s)d\mu_2(s) < \ln x(0) + \rho \int_{-\tau_2}^{0} \phi_1(s)d\mu_2(s) + g(0)t.$$

Thus, for $t \in [0, \omega)$,

$$\ln x(t) < \ln x(0) + \rho \int_{-\tau_2}^{0} \phi_1(s)d\mu_2(s) + g(0)t. \tag{2.4}$$

Let $A = x(0)\exp\left(\rho \int_{-\tau_2}^{0} \phi_1(s)d\mu_2(s)\right)$, $B = g(0)$, then we see that (2.4) implies that for $t < \omega$,

$$x(t) < Ae^{Bt}. \tag{2.5}$$

Hence, $\lim_{t \to \omega} x(t) \leq Ae^{B\omega} < +\infty$. Clearly, (2.5) holds for all $t \in [0, \omega)$. The second equation of (2.1) implies that, for $t \in [0, \omega)$,

$$(\ln y(t))' \leq |a| + |b|p(Ae^{B\omega}),$$

which leads to

$$\lim_{t \to \omega} y(t) \leq y(0)\exp\left(\omega[|a| + |b|p(Ae^{B\omega})]\right) < +\infty. \tag{2.6}$$

By the well-known continuation theorem (Theorem 12.2.4 in [13]), we conclude that $\omega = +\infty$, proving the proposition. □

In the rest of this paper $\|\phi(s)\| = \max\{\phi(s), s \in [-\tau_0, 0]\}$ is denoted for any continuous function $\phi(s)$ defined on $[-\tau_0, 0]$.

3. Boundedness of $x(t)$. Our main object in this section is to obtain conditions under which $x(t)$ will be bounded. To this end, we will analyze the system independent of y. We need the following lemma. Its proof can be found in [18]. For convenience and completeness, we repeat its proof here.

LEMMA 3.1. *If $0 < \alpha \leq e^{-1}$, then there exists a $\lambda(\alpha)$, $1 < \lambda(\alpha) \leq e$ such that* $\exp(\alpha\lambda(\alpha)) = \lambda(\alpha)$ *and* $\exp(\alpha x) > x$ *for $x < \lambda(\alpha)$.*

Proof. If $\alpha = e^{-1}$, then we easily see that $e^{\alpha x} \geq x$ for all $x \in R$ and $e^{\alpha x} = x$ if and only if $x = e$. Clearly, for $x > 0$, $e^{\alpha x}$ is strictly increasing with respect to α. Thus, we see that if $0 < \alpha < e^{-1}$, $e^{\alpha x}$ will intersect with x at exactly two distinct points, say $x_1(\alpha)$ and $x_2(\alpha)$ and $x_1(\alpha) < x_2(\alpha)$. Then, we must have $1 < x_1(\alpha) < e < x_2(\alpha)$. Let $\lambda(\alpha) = x_1(\alpha)$; then the conclusion of the lemma holds. □

The following theorem is the main result of this section. It generalizes Theorem 3.1 in [18].

THEOREM 3.1. *Assume $\alpha \equiv g(0)\tau_1 + \rho \leq e^{-1}$, and let $\lambda(\alpha)$ be defined as in Lemma 3.1. Assume further that the initial function ϕ_1 for x satisfies $\|\phi_1\| < 1$. Then $x(t) < \lambda(\alpha)$ for $t \geq -\tau_0$.*

Proof. Since $x(t) > 0$, $y(t) > 0$, we have

$$\dot{x}(t) \leq x(t)\left[\int_{-\tau_1}^{0} g(x(t+s))d\mu_1(s) - p\int_{-\tau_2}^{0} \dot{x}(t+s)d\mu_2(s)\right]. \tag{3.1}$$

If $x(t)$ is not bounded by $\lambda(\alpha)$, then there must exist $t^* > t_0 > 0$ such that $x(t^*) = \lambda(\alpha)$, $x(t_0) = 1$, $1 < x(t) < \lambda(\alpha)$ for $t \in (t_0, t^*)$ and $x(t) < \lambda(\alpha)$ for $t \in [-\tau_0, t^*)$. It is easy to see that (3.1) implies that, for $t \geq t_0$,

$$\begin{aligned} x(t) \leq x(t_0)&\exp\left\{\int_{t_0}^{t}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\right\} \\ &\cdot\exp\left\{-\rho\int_{t_0}^{t}\left(\int_{-\tau_2}^{0} \dot{x}(\tau+s)d\mu_2(s)\right)d\tau\right\}. \end{aligned} \tag{3.2}$$

Since $x(t) > 0$, for $t \geq -\tau_0$, we have

$$\begin{aligned} -\rho\int_{t_0}^{t^*}\left(\int_{-\tau_2}^{0} \dot{x}(\tau+s)d\mu_2(s)\right)d\tau &= -\rho\int_{-\tau_2}^{0}\left(\int_{t_0}^{t^*} \dot{x}(\tau+s)d\tau\right)d\mu_2(s) \\ &= -\rho\int_{-\tau_2}^{0}[x(t^*+s) - x(t_0+s)]d\mu_2(s) < \rho\int_{-\tau_2}^{0} x(t_0+s)d\mu_2(s) \\ &< \rho\lambda(\alpha). \end{aligned} \tag{3.3}$$

If $t^* \leq t_0 + \tau_1$, then

$$\int_{t_0}^{t^*}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau < \int_{t_0}^{t^*} g(0)d\tau \leq g(0)\tau_1. \tag{3.4}$$

If $t^* > t_0 + \tau_1$, then $x(\tau+s) \geq 1$ for $s \in [-\tau_1, 0]$, $\tau \in [t_0 + \tau_1, t^*]$, which implies $g(x(\tau+s)) \leq 0$ by (H2). Thus

$$\int_{t_0+\tau_1}^{t^*}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau \leq 0. \tag{3.5}$$

Hence, we have

$$
\begin{aligned}
&\int_{t_0}^{t^*}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\\
&\quad=\int_{t_0}^{t_0+\tau_1}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\\
&\qquad+\int_{t_0+\tau_1}^{t^*}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\\
&\quad<\int_{t_0}^{t_0+\tau_1} g(0)d\tau = g(0)\tau_1.
\end{aligned}
\tag{3.6}
$$

That is, in both cases, we have

$$
\int_{t_0}^{t^*}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau < g(0)\tau_1. \tag{3.7}
$$

Therefore, (3.2) yields

$$
\lambda(\alpha) = x(t^*) < \exp(g(0)\tau_1 + \rho\lambda(\alpha)) < \exp(\alpha\lambda(\alpha)). \tag{3.8}
$$

Clearly, this is a contradiction to the definition of $\lambda(\alpha)$. Therefore, $x(t^*)$ must be less than $\lambda(\alpha)$, and the theorem is proved. □

We call a function $x(t)$ (defined on $[0, +\infty)$) *oscillatory about* x^* (see also [7], [10], [16], [17]) if there exists a sequence $\{t_n\} \to +\infty$ as $n \to +\infty$, such that $x(t_n) = x^*$, $n = 1, \cdots$. Otherwise, we call it *nonoscillatory about* x^*.

If $x(t)$ is unbounded, then the following theorem may roughly characterize its behavior.

THEOREM 3.2. *In system* (2.1), *if $x(t)$ is unbounded, then $x(t)$ is oscillatory about* 1.

Proof. Assume that $x(t)$ is unbounded and not oscillatory about 1, then there must exist a $t_0 > 0$ such that, for $t \geq t_0$, $x(t) > 1$. Let $t_1 = t_0 + \tau_1$; then we have for $t > t_1$,

$$
\int_{t_1}^{t}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau < 0,
$$

and

$$
\begin{aligned}
&-\rho\int_{t_1}^{t}\left(\int_{-\tau_2}^{0} \dot{x}(\tau+s)d\mu_2(s)\right)d\tau\\
&\quad=-\rho\int_{-\tau_2}^{0}[x(t+s)-x(t_1+s)]d\mu_2(s)\\
&\quad<\rho\int_{-\tau_2}^{0} x(t_1+s)d\mu_2(s).
\end{aligned}
$$

Hence, from (3.2), we see that, for $t > t_1$,

$$
x(t) \leq x(t_1)\exp\left\{\rho\int_{-\tau_2}^{0} x(t_1+s)d\mu_2(s)\right\} < +\infty,
$$

which implies that $x(t)$ is bounded. This is a contradiction to our assumption, and the theorem is thus proved. □

4. Attractivity of the positive steady states in single population models. In this section we restrict our attention to the following neutral delay single population model, which results from (2.1) by taking $y(t) \equiv 0$:

$$\dot{x}(t) = x(t)\left[\int_{-\tau_1}^{0} g(x(t+s))d\mu_1(s) - \rho\int_{-\tau_2}^{0} \dot{x}(t+s)d\mu_2(s)\right]. \tag{4.1}$$

We always assume that the initial function of (4.1) satisfies the requirements stated in the first part of (2.2). Clearly, (4.1) has exactly two steady states. They are $x(t) \equiv 0$ and $x(t) \equiv 1$. The variational equation of (4.1) about $x(t) \equiv 0$ is $\dot{x}(t) = g(0)x(t)$. Thus, we see $x(t) \equiv 0$ is always unstable. In fact, we have the following stronger result.

PROPOSITION 4.1. *Let $x(t)$ be the solution of* (4.1); *then*

$$\limsup_{t\to+\infty} x(t) \geq 1.$$

Proof. Otherwise, there is $0 < \epsilon < 1$, $T > 0$, such that, for $t \geq T$, $x(t) \leq 1 - \epsilon$. Then for $t > t_0 \geq T + \tau_1 + \tau_2$,

$$\int_{t_0}^{t}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau \geq \int_{t_0}^{t} g(1-\epsilon)d\tau = g(1-\epsilon)(t-t_0), \tag{4.2}$$

and

$$\begin{aligned} -\rho\int_{t_0}^{t}\left(\int_{-\tau_2}^{0} \dot{x}(\tau+s)d\mu_2(s)\right)d\tau &= -\rho\int_{-\tau_2}^{0}[x(t+s) - x(t_0+s)]d\mu_2(s) \\ &> -\rho\int_{-\tau_2}^{0} x(t+s)d\mu_2(s) \geq -(1-\epsilon)\rho. \end{aligned} \tag{4.3}$$

Therefore

$$\begin{aligned} x(t) &= x(t_0)\exp\left\{\int_{t_0}^{t}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\right\} \\ &\quad\cdot\exp\left\{-\rho\int_{t_0}^{t}\left(\int_{-\tau_2}^{0} \dot{x}(\tau+s)d\mu_2(s)\right)d\tau\right\} \\ &\geq x(t_0)\exp\{g(1-\epsilon)(t-t_0) - (1-\epsilon)\rho\}, \end{aligned}$$

which implies that

$$\lim_{t\to+\infty} x(t) = +\infty,$$

a contradiction to the assumption that $x(t) \leq 1 - \epsilon$ for $t \geq T$. This proves the proposition. □

In order to prove our global stability result we need the following simple lemma.

LEMMA 4.1. *For $0 < r < 1$, there is a strictly decreasing function $h(r)$, such that $h(r) = \exp\{r(h(r)-1)\}$, $\lim_{r\to 0^+} h(r) = +\infty$, $\lim_{r\to 1^-} h(r) = 1$, and $x > \exp\{r(x-1)\}$ for $x \in (1, h(r))$.*

Proof. Clearly, for $0 < r < 1$, $e^{r(x-1)}$ always intersects with x at $x = 1$. $(d/dx)(e^{r(x-1)})|_{x=1} = r < (dx/dx)|_{x=1} = 1$, we see that $e^{r(x-1)}$ will intersect with x at another point, say $h(r)$. Clearly, this $h(r)$ has all those properties described in the lemma. □

As in Theorem 3.1, we denote

$$\alpha = g(0)\tau_1 + \rho,$$

and define $\lambda(\alpha)$ as in Lemma 3.1. Assuming $\alpha \leq e^{-1}$, we define

$$G(\alpha) = \max\{|g'(x)| : x \in [0, \lambda(\alpha)]\}.$$

We denote

$$\bar{\tau} = \max\{\tau_1, \tau_2\}.$$

Now we are ready to state and prove the main result of this section.

THEOREM 4.1. *In system* (4.1), *assume* $0 < \alpha < e^{-1}$, *and*

(i) $\lambda(\alpha)\rho < 1$,

(ii) $G(\alpha)\tau_1 + 2\rho \leq h^{-1}(\lambda(\alpha))$,

where $h^{-1}(\cdot)$ *is the inverse function of* $h(r)$ *defined in Lemma* 4.1. *Then we have that* $\lim_{t\to+\infty} x(t,\phi) = 1$, *provided that its initial function* $\phi(s)$ *satisfies that* $\phi(s) \in C^1([-\bar{\tau}, 0], R^+)$, $\phi(0) > 0$, *and* $\|\phi\| = \max\{|\phi(s)| : s \in [-\bar{\tau}, 0]\} < 1$.

Proof. By Theorem 3.1, we know that $x(t) < \lambda(\alpha)$, where $1 < \lambda(\alpha) \leq e$, and $\alpha \equiv g(0)\tau_1 + \rho \leq e^{-1}$. Denote

$$v = \limsup_{t\to+\infty} |x(t) - 1|; \tag{4.4}$$

then $0 \leq v \leq e - 1$. In the following, we assume $v > 0$. We claim that $x(t)$ must be oscillatory about 1. Otherwise, there is a $T > 0$ such that, for $t \geq T$, $x(t) > 1$ or $x(t) < 1$. We assume first that $x(t) > 1$ for $t \geq T$. Let $\bar{t} \geq T + \tau_1 + \tau_2$, such that

$$|\dot{x}(\bar{t})| = \max\{|\dot{x}(\bar{t} + s)| : s \in [-\tau_2, 0]\}.$$

If $\dot{x}(\bar{t}) \geq 0$, we have

$$\dot{x}(\bar{t}) + \rho x(\bar{t}) \int_{-\tau_2}^{0} \dot{x}(\bar{t} + s) d\mu_2(s) \geq (1 - \lambda(\alpha)\rho)\dot{x}(\bar{t}),$$

and

$$x(\bar{t}) \int_{-\tau_1}^{0} g(x(\bar{t} + s)) d\mu_1(s) < \lambda(\alpha) g(0),$$

which implies that (since $\lambda(\alpha)\rho < 1$)

$$\dot{x}(\bar{t}) < \frac{\lambda(\alpha)g(0)}{1 - \lambda(\alpha)\rho}.$$

If $\dot{x}(\bar{t}) < 0$, we have

$$\dot{x}(\bar{t}) + \rho x(\bar{t}) \int_{-\tau_2}^{0} \dot{x}(\bar{t} + s) d\mu_2(s) \leq (1 - \lambda(\alpha)\rho)\dot{x}(\bar{t}),$$

and

$$x(\bar{t}) \int_{-\tau_1}^{0} g(x(\bar{t} + s)) d\mu_1(s) > \lambda(\alpha) g(\lambda(\alpha)),$$

which implies that

$$\dot{x}(\bar{t}) > \frac{\lambda(\alpha)g(\lambda(\alpha))}{1 - \lambda(\alpha)\rho}.$$

Hence, for all $t \geq T + \tau_1 + \tau_2$,

$$|\dot{x}(t)| \leq \max\left\{\frac{\lambda(\alpha)g(0)}{1 - \lambda(\alpha)\rho}, \frac{-\lambda(\alpha)g(\lambda(\alpha))}{1 - \lambda(\alpha)\rho}\right\}. \tag{4.5}$$

This, together with the assumptions $v > 0$ and $x(t) > 1$ for $t \geq T$, implies that

$$\lim_{t\to+\infty} \max\left\{\int_{t_0}^{t} g(x(\tau + s)) d\tau : s \subset n[-\tau_1, 0]\right\} = -\infty.$$

Thus

$$\lim_{t\to+\infty}\int_{t_0}^{t}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau$$
$$=\lim_{t\to+\infty}\int_{-\tau_1}^{0}\left(\int_{t_0}^{t} g(x(\tau+s))d\tau\right)d\mu_1(s)=-\infty.$$

Therefore, we have, for any $t_0\geq T$,

$$\lim_{t\to+\infty}x(t)=\lim_{t\to+\infty}x(t_0)\exp\left\{\int_{t_0}^{t}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\right\}$$
$$\cdot\exp\left\{-\rho\int_{t_0}^{t}\left(\int_{-\tau_2}^{0}\dot{x}(\tau+s)d\mu_2(s)\right)d\tau\right\}=0,$$

since

$$\exp\left\{-\rho\int_{t_0}^{t}\left(\int_{-\tau_2}^{0}\dot{x}(\tau+s)d\mu_2(s)\right)d\tau\right\}$$
$$=\exp\left\{-\rho\int_{-\tau_2}^{0}[x(t+s)-x(t_0+s)]d\mu_2(s)\right\}$$
$$<\exp\{\rho\lambda(\alpha)\}<+\infty.$$

This clearly contradicts the assumption that $x(t)>1$ for $t\geq T$. The case of $x(t)<1$ for $t\geq T$ can be dealt with similarly. This proves the claim.

Now, since $x(t)$ is oscillatory about 1, we see that there is a sequence $\{T_i\}$, $i=1,\cdots$, such that $0<T_1<\cdots<T_i<T_{i+1}<\cdots$, $\lim_{i\to+\infty}T_i=+\infty$, and $x(T_i)=1$. Denote

$$u=\limsup_{t\to+\infty}x(t),$$
$$w=\liminf_{t\to+\infty}x(t).$$

Then, either $u=1+v$ or $w=1-v$. Assume first that $u=1+v$. Then, for any $0<\epsilon<v$, there is an $i(\epsilon)>1$ such that, for $t\geq T_i$, $1-v-\epsilon<x(t)<u+\epsilon$. Clearly, there is a $t^*=t^*(\epsilon)\geq T_i+\bar{\tau}$ such that $x(t^*)>u-\epsilon$, and $x(t^*)$ is the maximum in $[T_j,T_{j+1}]$ for some $j\geq i$. Without loss of generality, we may assume that $1<x(t)<x(t^*)$ for $t\in(T_j,t^*)$. We have

$$-\rho\int_{T_j}^{t^*}\left(\int_{-\tau_2}^{0}\dot{x}(\tau+s)d\mu_2(s)\right)d\tau=-\rho\int_{-\tau_2}^{0}[x(t^*+s)-x(T_j+s)]d\mu_2(s)$$
$$\leq\rho\int_{-\tau_2}^{0}(|x(t^*+s)-1|+|1-x(T_j+s)|)d\mu_2(s)$$
$$<2\rho(u+\epsilon-1).$$

If $t^*-T_j\leq\tau_1$, then

$$\int_{T_j}^{t^*}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\leq\int_{T_j}^{t^*}\left(\int_{-\tau_1}^{0} G(\alpha)(u+\epsilon-1)d\mu_1(s)\right)d\tau$$
$$\leq G(\alpha)\tau_1(u+\epsilon-1),$$

since

$$|g(x(\tau+s))|=|g(x(\tau+s))-g(1)|$$
$$=|g'(\theta x(\tau+s)+1-\theta)(x(\tau+s)-1)|$$
$$\leq G(\alpha)(v+\epsilon)=G(\alpha)(u+\epsilon-1),$$

where $\theta \in [0,1]$. If $t^* - T_j > \tau_1$, then

$$\int_{T_j+\tau_1}^{t^*} \left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s) \right) d\tau < 0,$$

and

$$\int_{T_j}^{T_j+\tau_1} \left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s) \right) d\tau \le G(\alpha)\tau_1(u+\epsilon-1).$$

Thus, in both cases we have

$$\begin{aligned} u-\epsilon < x(t^*) &= x(T_j)\exp\left\{ \int_{T_j}^{t^*} \left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s) \right) d\tau \right\} \\ &\quad \cdot \exp\left\{ -\rho \int_{T_j}^{t^*} \left(\int_{-\tau_2}^{0} \dot{x}(\tau+s)d\mu_2(s) \right) d\tau \right\} \\ &< \exp\{[G(\alpha)\tau_1 + 2\rho](u+\epsilon-1)\}. \end{aligned} \tag{4.6}$$

Since (4.6) is true for all $0 < \epsilon < v$, by letting $\epsilon \to 0$ we obtain

$$u \le \exp\{[G(\alpha)\tau_1 + 2\rho](u-1)\}. \tag{4.7}$$

This is a contradiction to assumption (ii), since if $G(\alpha)\tau_1 + 2\rho \le h^{-1}(\lambda(\alpha))$, then for $1 < x \le \lambda(\alpha)$,

$$x > \exp\{[G(\alpha)\tau_1 + 2\rho](x-1)\}.$$

This indicates that $u \ne 1+v$ if $v > 0$.

Now we assume $w = 1-v$. Then a similar argument as above yields that, for any $0 < \epsilon < v$,

$$1 - v + \epsilon > \exp\{-[G(\alpha)\tau_1 + 2\rho](v+\epsilon)\},$$

which implies that

$$1 - v \ge \exp\{-[G(\alpha)\tau_1 + 2\rho]v\}.$$

Thus,

$$w \ge \exp\{[G(\alpha)\tau_1 + 2\rho](w-1)\}.$$

This is impossible for $w < 1$ and $G(\alpha)\tau_1 + 2\rho \le h^{-1}(\lambda(\alpha))$, since $h^{-1}(\lambda(\alpha)) < 1$ by Lemma 4.1. This proves that $w \ne 1-v$ if $v > 0$. Hence, v must be zero, proving the theorem. □

In particular, for the equation

$$\dot{x}(t) = x(t)\left[\int_{-\tau_1}^{0} r[1 - x(t+s)]d\mu_1(s) - \rho \int_{-\tau_2}^{0} \dot{x}(t+s)d\mu_2(s) \right], \tag{4.8}$$

where $r > 0$, we have the following corollary.

COROLLARY 4.1. *In* (4.8) *let* $\alpha = r\tau_1 + \rho$. *Assume* $\alpha < e^{-1}$, $\lambda(\alpha)\rho < 1$ *and* $r\tau_1 + 2\rho \le h^{-1}(\lambda(\alpha))$. *Then the conclusion of Theorem* 4.1 *is valid for* (4.8).

Proof. We note in this case $g(x) = r(1-x)$. Thus $g(0) = r$ and $G(\alpha) = r$. The rest follows from Theorem 4.1. □

Remark 4.1. It should be mentioned here that assumption (ii) in Theorem 4.1 can be replaced by

(ii′) $G(\alpha)\tau_1 + 2\rho < (e-1)^{-1}$,

which is more restrictive, but easy to verify. This is because $1 < \lambda(\alpha) < e$ for $0 < \alpha < e^{-1}$; thus by Lemma 4.1 we have

$$h^{-1}(\lambda(\alpha)) > h^{-1}(e) = (e-1)^{-1}.$$

Remark 4.2. As mentioned in the Introduction, our results are rather limited and primitive. This can be seen by taking $\rho = 0$, i.e., when (4.1) is reduced to

$$\dot{x}(t) = x(t) \int_{-\tau_1}^{0} g(x(t+s))d\mu_1(s). \tag{4.9}$$

For (4.9) Theorem 3.2 in [15] implies that if $G\tau_1 \leq 1$, where G is any upper bound of $|g'(x)|$, $x \geq 0$, then $x \equiv 1$ is globally asymptotically stable with respect to initial function ϕ, $\phi \in C([-\bar{\tau}, 0], R^+)$, $\phi(0) > 0$. This is clearly much sharper than the conclusion implied by Theorem 4.1.

5. Boundedness results for system (2.1). We consider first the case when $c > 0$ in system (2.1). This will include the neutral competition system (1.6) as a special case. When $a < 0$, $b > 0$, system (2.1) can be used to model the predatory-prey interaction with self-crowding effect on predator. The following result generalizes Theorem 4.1 in [18].

THEOREM 5.1. *Assume $c > 0$, $\alpha = g(0)\tau_1 + \rho \leq e^{-1}$, $\lambda(\alpha)$ is defined as in Lemma 3.1, and $\|\phi_1\| < 1$. Then solutions of (2.1) are bounded. Moreover, if $x(t) < \lambda(\alpha)$, for $t \geq -\tau_0$, and*

(i) *If $a < 0$, $b > 0$, and $\beta = \max\{0, bp(\lambda(\alpha)) + a\}$, then*

$$\limsup_{t \to +\infty} y(t) \leq c^{-1}\beta e^{\beta\tau_5};$$

(ii) *If $a > 0$, $b < 0$, then*

$$\limsup_{t \to +\infty} y(t) \leq c^{-1}ae^{a\tau_5}.$$

Proof. The assertion on $x(t)$ follows from Theorem 3.1. Thus, in case (i) we have

$$\dot{y}(t) \leq y(t)\left[bp(\lambda(\alpha)) + a - c\int_{-\tau_5}^{0} y(t+s)d\mu_5(s)\right]. \tag{5.1}$$

It is thus clear that if $bp(\lambda(\alpha)) + a \leq 0$, then $\lim_{t \to +\infty} y(t) = 0$. Assume $bp(\lambda(\alpha)) + a = \beta > 0$, then (5.1) implies that $\dot{y}(t) \leq \beta y(t)$. Hence, for $t \geq t_0$, $y(t) \leq y(t_0)e^{\beta(t-t_0)}$, which leads to

$$y(t_0) \geq y(t)e^{-\beta(t-t_0)}.$$

Therefore,

$$\begin{aligned} -\int_{-\tau_5}^{0} y(t+s)d\mu_5(s) &\leq -\int_{-\tau_5}^{0} y(t)e^{\beta s}d\mu_5(s) \\ &\leq -e^{-\beta\tau_5}\int_{-\tau_5}^{0} y(t)d\mu_5(s) = -e^{-\beta\tau_5}y(t), \end{aligned} \tag{5.2}$$

since $y(t) > 0$ for $t \geq 0$. A substitution of (5.2) into (5.1) yields

$$\dot{y}(t) \leq y(t)(\beta - ce^{-\beta\tau_5}y(t)). \tag{5.3}$$

Clearly, solutions of

$$\dot{y}(t) = \beta y(t)(1 - c\beta^{-1}e^{-\beta\tau_5}y(t)) \tag{5.4}$$

satisfy

$$\lim_{t \to +\infty} y(t) = c^{-1}\beta e^{\beta\tau_5}. \tag{5.5}$$

Therefore, solutions of (5.3) must satisfy

$$\limsup_{t\to+\infty} y(t) \le c^{-1}\beta e^{\beta\tau_5}.$$

This proves case (i). For case (ii), we have

$$\dot{y}(t) \le y(t)\left[a - c\int_{-\tau_5}^{0} y(t+s)d\mu_2(s)\right]. \tag{5.6}$$

Thus $\dot{y}(t) \le ay(t)$, which implies that $y(t+s) \ge y(t)e^{as}$, for $s \le 0$. Therefore

$$\dot{y}(t) \le y(t)(a - ce^{-a\tau_5}y(t)).$$

Hence, by repeating the above argument, we can show that

$$\limsup_{t\to+\infty} y(t) \le c^{-1}ae^{a\tau_5}.$$

This completes the proof. □

In the rest of this section we assume $c = 0$, $a = -\delta < 0$, $b > 0$. System (2.1) thus reduces to

$$\begin{aligned}\dot{x}(t) = x(t)\Bigg[\int_{-\tau_1}^{0} & g(x(t+s))d\mu_1(s) - \rho\int_{-\tau_2}^{0}\dot{x}(t+s)d\mu_2(s)\\ & - q(x(t))\int_{-\tau_3}^{0} y(t+s)d\mu_3(s)\Bigg],\\ \dot{y}(t) = y(t)\Bigg[-\delta + b & \int_{-\tau_4}^{0} p(x(t+s))d\mu_4(s)\Bigg].\end{aligned} \tag{5.7}$$

When delays are absent from the above system, it reduces to the so-called Gause-type predator-prey system (cf., [4], [16]). For (5.7) we have the following boundedness result which generalizes Theorem 4.2 in [18].

THEOREM 5.2. *In* (5.7), *assume* $\alpha = g(0)\tau_1 + \rho \le e^{-1}$, $\lambda(\alpha)$ *is defined as in Lemma* 3.1, *and* $\|\phi_1\| < 1$. *Let* $\bar{x} > 0$ *be the unique solution of* $bp(x) = \delta$, $\gamma = \min\{q(x) : x \in [0, \lambda(\alpha)]\}$. *Denote* $\bar{y} = \gamma^{-1}[g(0) + \rho\lambda(\alpha) + \ln(\lambda(\alpha)/\bar{x})]$, $\tilde{y} = \max\{y(0), \bar{y}\}$, $\Delta = \tilde{y}\exp\{[bp(\lambda(\alpha)) - \delta](\tau_3 + \tau_4 + 1)\}$. *Then, for* $t \ge 0$, $x(t) < \lambda(\alpha)$, $y(t) < \Delta$. *Moreover, if* $\lambda(\alpha) < \bar{x}$, *then* $\limsup_{t\to\infty} y(t) = 0$; *and if* $\lambda(\alpha) \ge \bar{x}$, *then*

$$\limsup_{t\to+\infty} y(t) \le \bar{y}\exp\{[bp(\lambda(\alpha)) - \delta](\tau_3 + \tau_4 + 1)\}.$$

Proof. Again, the assertion on $x(t)$ follows from Theorem 3.1. In the following we assume $y(t)$ is not bounded by Δ. Clearly, in this case $\bar{x}$ must be less than $\lambda(\alpha)$, since if $\bar{x} > \lambda(\alpha)$, then $\dot{y}(t) < 0$; which implies that $y(t) \le y(0) < \Delta$.

The first equation in (5.7) gives us, for $t \ge t_0$,

$$\begin{aligned}x(t) = & x(t_0)\exp\left\{\int_{t_0}^{t}\left(\int_{-\tau_1}^{0} g(x(\tau+s))d\mu_1(s)\right)d\tau\right.\\ & \left. - \rho\int_{t_0}^{t}\left(\int_{-\tau_2}^{0}\dot{x}(\tau+s)d\mu_2(s)\right)d\tau\right\}\\ & \cdot\exp\left\{-\int_{t_0}^{t}\left(q(x(\tau))\int_{-\tau_3}^{0} y(\tau+s)d\mu_3(s)\right)d\tau\right\},\end{aligned}$$

which leads to

$$
\begin{aligned}
x(t) \le& \lambda(\alpha)\exp\{g(0)(t-t_0)+\rho\lambda(\alpha)\} \\
&\cdot\exp\left\{-\gamma\int_{-\tau_3}^{0}\left(\int_{t_0}^{t} y(\tau+s)d\tau\right)d\mu_3(s)\right\}.
\end{aligned} \tag{5.8}
$$

The second equation of (5.7) implies that

$$\dot{y}(t) < [bp(\lambda(\alpha))-\delta]y(t),$$

which leads to

$$y(t) < y(t_0)\exp\{[bp(\lambda(\alpha))-\delta](t-t_0)\}, \qquad t \ge t_0. \tag{5.9}$$

Since $y(t)$ is not bounded by Δ, there must be $t_2 > t_1 \ge 0$, such that $y(t_1)=\tilde{y}$, $y(t_2)=\Delta$, and $y(t)\ge\tilde{y}$ for $t\in[t_1,t_2]$. From (5.9) we see that $t_2-t_1>\tau_3+\tau_4+1$. In (5.8) we let $t_0=t_1+\tau_3$ and $t_0+1\le t\le t_2$, then $y(\tau+s)\ge\tilde{y}$ for $\tau\in[t_0,t]$, $s\in[-\tau_3,0]$. Thus,

$$
\begin{aligned}
x(t) \le&\lambda(\alpha)\exp\{g(0)(t-t_0)+\rho\lambda(\alpha)\}\exp\{-\gamma\tilde{y}(t-t_0)\} \\
\le&\lambda(\alpha)\exp\{g(0)(t-t_0)+\rho\lambda(\alpha)\} \\
&\cdot\exp\{-[g(0)+\rho\lambda(\alpha)+\ln(\lambda(\alpha)/\bar{x})](t-t_0)\} \\
\le&\lambda(\alpha)\exp\{-\ln(\lambda(\alpha)/\bar{x})\}\exp\{-[\rho\lambda(\alpha)+\ln(\lambda(\alpha)/\bar{x})](t-t_0-1)\} \\
=&\bar{x}\exp\{-[\rho\lambda(\alpha)+\ln(\lambda(\alpha)/\bar{x})](t-t_0-1)\}.
\end{aligned} \tag{5.10}
$$

Hence, for $t\in[t_0+1,t_2]$, $x(t)\le\bar{x}$. From the second equation of (5.7), we have

$$y(t)=y(t_0)\exp\left\{\int_{t_0}^{t}\left[-\delta+b\int_{-\tau_4}^{0}p(x(\tau+s))d\mu_4(s)\right]d\tau\right\}. \tag{5.11}$$

Let $t_0=t_1$ in (5.11), then for $t_1+\tau_3+\tau_4+1\le t\le t_2$,

$$
\begin{aligned}
y(t)=\tilde{y}&\exp\left\{\int_{t_1}^{t_1+\tau_3+\tau_4+1}\left[-\delta+b\int_{-\tau_4}^{0}p(x(\tau+s))d\mu_4(s)\right]d\tau\right\} \\
&\cdot\exp\left\{\int_{t_1+\tau_3+\tau_4+1}^{t}\left[-\delta+b\int_{-\tau_4}^{0}p(x(\tau+s))d\mu_4(s)\right]d\tau\right\}.
\end{aligned}
$$

Clearly $-\delta+b\int_{-\tau_4}^{0}p(x(t+s))d\mu_4(s)<bp(\lambda(\alpha))-\delta$. For $t_1+\tau_3+\tau_4+1\le t\le t_2$, $-\delta+b\int_{-\tau_4}^{0}p(x(t+s))d\mu_4(s)\le 0$. Therefore, for $t_1+\tau_3+\tau_4+1\le t\le t_2$,

$$y(t)<\tilde{y}\exp\{[bp(\lambda(\alpha))-\delta](\tau_3+\tau_4+1)\}=\Delta,$$

a contradiction to the assumption that $y(t_2)=\Delta$. This proves that $y(t)<\Delta$.

Next we prove the second part of the theorem.

Clearly, if $\lambda(\alpha)<\bar{x}$, then $\dot{y}(t)<0$, and $\limsup_{t\to+\infty}y(t)=0$. Assume now that $\lambda(\alpha)\ge\bar{x}$, and there is a $y(t)$, such that $\limsup_{t\to+\infty}y(t)>\bar{y}\exp\{[bp(\lambda(\alpha))-\delta](\tau_3+\tau_4+1)\}$. We claim that there exists a $\tilde{t}>0$ such that $y(\tilde{t})=\bar{y}$. Otherwise, for large t, $y(t)>\bar{y}$. From (5.10) we see $x(t)<\bar{x}$ for large t; indeed, in this case, $\lim_{t\to+\infty}x(t)=0$. This clearly implies that $\lim_{t\to+\infty}y(t)=0$, a contradiction. Obviously, for any $T>0$, there is a $\hat{t}\ge T$ such that $y(\hat{t})\ge\bar{y}\exp\{[bp(\lambda(\alpha))-\delta](\tau_3+\tau_4+1)\}$. Thus, for such a large $\hat{t}$, we can find a $\tilde{t}$, $\tilde{t}<\hat{t}$ such that $y(\tilde{t})=\bar{y}$ and $y(t)\ge\bar{y}$ for $t\in[\tilde{t},\hat{t}]$. Now we can repeat the previous argument (by letting $\tilde{t}=t_1$, $\hat{t}=t_2$) to derive a contradiction. The proof is thus complete. □

6. Discussion. In order to facilitate our discussion, we would like to present some numerical simulations of the neutral delay systems studied in the previous sections. Consider first the single-species neutral delay equation

$$\dot{x} = rx[1 - x(t-1) - c\dot{x}(t-1)]. \tag{6.1}$$

Figures 1a, 1b, and 1c depict three solutions of (6.1) when $r = 0.1$, $c = 5$. Clearly, we see that solutions of (6.1) seem to be very sensitive to initial functions. The solution with initial function as constant 0.2 seems to tend to a periodic solution with three peaks, while the solutions with initial function as $1.2 + 0.2t$ and 0.1 seem to approach the steady state $x(t) \equiv 1$ monotonically (for $t \geq 10$). This perhaps can explain (roughly) why our main results require initial functions to satisfy some conditions. However, such a drastic change of behaviors cannot happen if $c = 0$. (Indeed, all solutions tend to $x(t) \equiv 1$. See [30].)

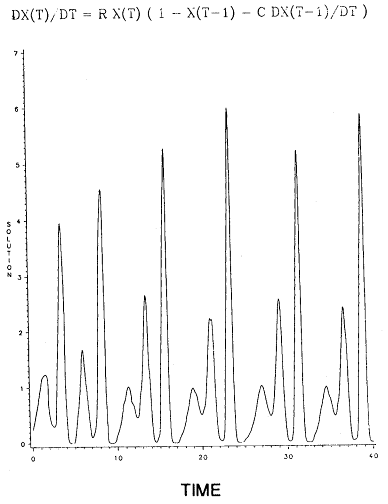


FIG. 1a. *Numerical solution of* (6.1) *with* $r = 0.1$, $c = 5$, *and* $x(t) = 0.2$, $t \leq 0$.

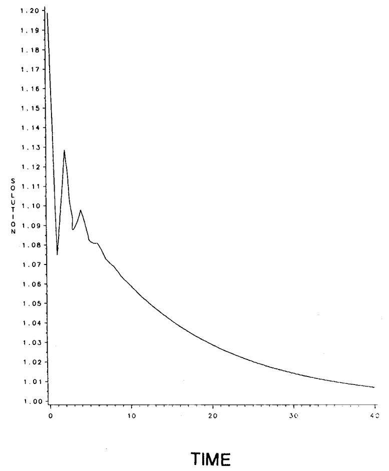


FIG. 1b. *Numerical solution of* (6.1) *with* $r = 0.1$, $c = 5$, *and* $x(t) = 1.2 + 0.2t$, $t \leq 0$.

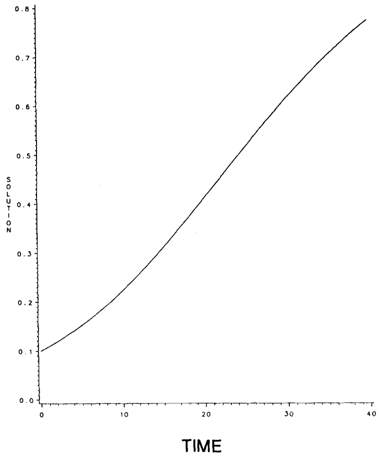


FIG. 1c. *Numerical solution of* (6.1) *with* $r = 0.1$, $c = 5$, *and* $x(t) = 0.1$, $t \leq 0$.

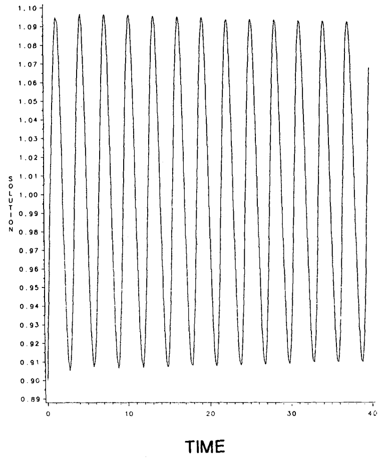


FIG. 2a. *Numerical solution of* (6.1) *with* $r = \pi/\sqrt{3}$, $c = \sqrt{3}/2\pi$, *and* $x(t) = 0.9$, $t \leq 0$.

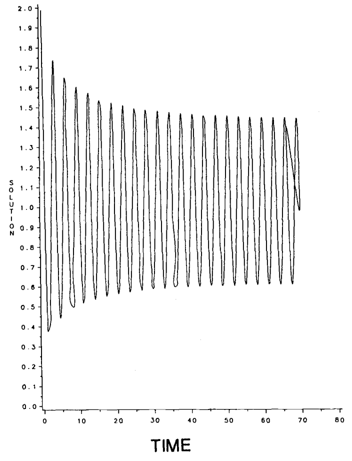


FIG. 2b. *Numerical solution of* (6.1) *with* $r = 0.05 + \pi/\sqrt{3} \approx 1.8638$, $c = \sqrt{3}/2\pi - 0.04$, $x(t) = 2.0 + t$, $t \leq 0$. $l(r, c) \approx 1.8197 < r$.

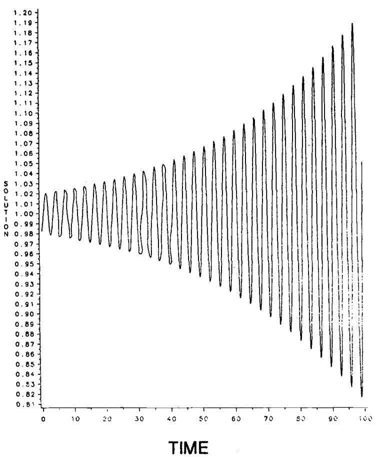


FIG. 2c. *Numerical solution of* (6.1) *with* $r = \pi/\sqrt{3} - 0.06$, $c = \sqrt{3}/2\pi + 0.02$, $x(t) = 0.98$, $t \leq 0$. $l(r,c) \approx 1.8103 > r \approx 1.7538$.

Denote

$$l(r,c) = \sqrt{1 - r^2c^2}\,\mathrm{arc}\cot[-rc(1 - r^2c^2)^{-1/2}]. \tag{6.2}$$

The following statements are proved in [6]:

(i) If $l(r,c) = r$, then the characteristic equation of (6.1) about $x(t) \equiv 1$ has roots $\pm i\omega$, $\omega > 0$.

(ii) If $l(r,c) < r$, then $x(t) \equiv 1$ is locally asymptotically stable.

(iii) If $l(r,c) > r$, then $x(t) \equiv 1$ is unstable.

Figures 2a, 2b, and 2c confirm these results. Clearly, these figures may suggest that a Hopf bifurcation takes place when the local stability of $x(t) \equiv 1$ changes from stable to unstable.

We consider next the following neutral delay predator-prey system:

$$\begin{cases} \dot{x}(t) = x(t)[1 - x(t-\tau) - \rho\dot{x}(t-\tau)] - y(t)\dfrac{x^2(t)}{x^2(t)+1}, \\ \dot{y}(t) = y(t)\left[-0.1 + \dfrac{x^2(t)}{x^2(t)+1}\right]. \end{cases} \tag{6.3}$$

When $\rho = 0$ (i.e., no neutral term), Figs. 3a and 3b indicate that solutions tend to the positive steady state $(\frac{1}{3}, \frac{20}{9})$ monotonically. Figures 4a and 4b seem to show that when ρ increased to 2.9, the solution $(x(t), y(t))$ with initial functions $x(t) = 0.35$, $y(t) = 2.21$, $t \leq 0$, tends to a periodic solution surrounding the positive steady state. Indeed, by a local stability analysis as presented in [16], we can show that the positive steady state is unstable when $\rho = 2.9$, $\tau = 1$. This may suggest that the existence of the neutral term resulted in a Hopf bifurcation, thus producing a stable periodic solution. Other numerical simulations of system (6.3) strongly suggest that solutions of (6.3) are oscillatory when $\rho \neq 0$, in contrast to the monotone behavior depicted in Figs. 3a and 3b when $\rho = 0$.

As pointed out in the beginning of this paper, one of the most useful features of delay equation in population dynamics is its ability to show oscillations of population sizes observed in real systems. Our numerical simulations seem to suggest that neutral delay models may serve better for this purpose.

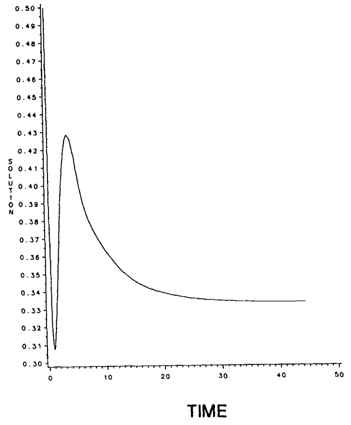


FIG. 3a. *Numerical solution of* $x(t)$ *of* (6.3) *with* $\rho = 0$, $\tau = 1$, *and* $(x(t), y(t)) = (0.5 - t, 1.5)$, $t \leq 0$.

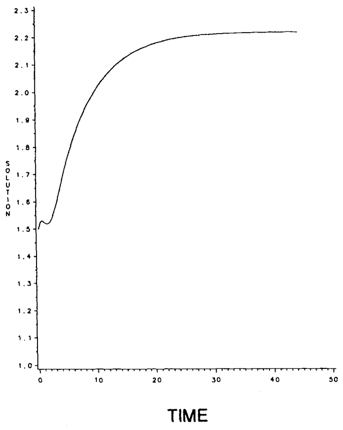


FIG. 3b. *Numerical solution of* $y(t)$ *of* (6.3) *with* $\rho = 0$, $\tau = 1$, *and* $(x(t), y(t)) = (0.5 - t, 1.5)$, $t \leq 0$.

It is well known that the dynamics of the simple-looking Wright's equation $\dot{x}(t) = rx(t)(1 - x(t - 1))$ is already very complicated (cf., [13], [14], [26], [30]). With the addition of a neutral term, we can only expect the dynamics to become even more complex. Numerical simulations depicted in Figs. 2a–2c, and Figs. 4a and 4b seem to suggest that periodic solutions exist and persist. These periodic solutions appear to be the outcome of a sequence of Hopf bifurcations as in the case of Wright's equation. Unfortunately, there seems to be no general Hopf bifurcation theory for nonlinear neutral systems to confirm this observation theoretically. A similar phenomenon is expected for system (2.1). This is an open problem left untouched in this paper. Another important question left unresolved is the domain of attractivity analysis for system (2.1). Indeed, this question is open even for system (2.1) when $\rho = 0$. These questions will be pursued in the future.

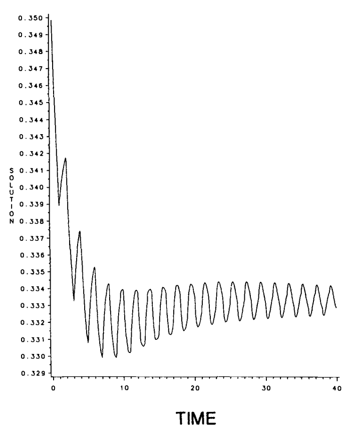


FIG. 4a. *Numerical solution of* $x(t)$ *of* (6.3) *with* $\rho = 2.9$, $\tau = 1$ *and* $(x(t), y(t)) = (0.35, 2.21)$, $t \leq 0$.

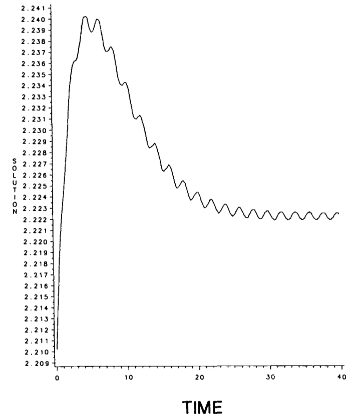


FIG. 4b. *Numerical solution of* $y(t)$ *of* (6.3) *with* $\rho = 2.9$, $\tau = 1$, *and* $(x(t), y(t)) = (0.35, 2.21)$, $t \leq 0$.

Our boundedness result in §3 indicates that if initial population is less than the carrying capacity of the environment (in our case, it is 1), and both the delay τ_1 and the neutral coefficient ρ are small, then the population stays bounded by $\lambda(\alpha)$, where $\lambda(\alpha)$ is between 1 and e, as defined in Lemma 3.1. Our global stability result in §4 suggests that, under slightly more restrictive conditions, the population approaches the environment carrying capacity as time goes by, a phenomenon observed for Wright's equation (for small delay) and logistic equations. This somehow partially justifies that the neutral delay effect can be ignored, provided that the delay length and neutral coefficient are expected to be small.

The results in §5 imply, to some extent, that the dissipativities of the considered systems are maintained for small delay τ_1 and small neutral coefficient ρ. They are certainly not surprising. However, for large τ_1 and ρ, the local stability analysis of [16] and [17] indicated that the neutral delay terms can be destabilizing.

Acknowledgments. The author is grateful to E. Lo for providing the numerical simulations displayed in the figures presented here, using the numerical schemes developed in [31].

The author also thanks the referees for their many valuable suggestions that improved the original manuscript.

REFERENCES

[1] E. Beretta and F. Solimano, *A generalization of Volterra models with continuous time delay in population dynamics: Boundedness and global asymptotic stability*, SIAM J. Appl. Math., 48, (1988), pp. 607–626.

[2] J. M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomath. 20, Springer-Verlag, New York, 1977.

[3] G. Dunkel, *Single species model for population growth depending on past history*, in Seminar on Differential Equations and Dynamical Systems, Lecture Notes in Math. 60, Springer-Verlag, New York, 1968, pp. 92–99.

[4] H. I. Freedman, *Deterministic Mathematical Models in Population Ecology*, Marcel Dekker, New York, 1988.

[5] H. I. Freedman and K. Gopalsamy, *Global stability in time-delayed single-species dynamics*, Bull. Math. Biol., 48 (1986), pp. 485–492.

[6] H. I. Freedman and Y. Kuang, *Stability switches in linear scalar neutral delay equations*, Funkcial. Ekvac., 34 (1991), pp. 187–209.

[7] K. Gopalsamy, *Equations Mathematical Ecology; Part* 1, *Autonomous Systems*, preliminary version, preprint.

[8] ———, *Time lags and global stability in two-species competition*, Bull. Math. Biol., 42 (1980), pp. 729–737.

[9] K. Gopalsamy, X.-Z. He, and L. Wen, *On a periodic neutral logistic equation*, Glasgow Math. J., to appear.

[10] K. Gopalsamy and B. G. Zhang, *On a neutral delay logistic equation*, Dynamics Stability Systems, 2 (1988), pp. 183–195.

[11] J. R. Haddock and Y. Kuang, *Asymptotic theory for a class of nonautonomous delay differential equations*, J. Math. Anal. Appl., to appear.

[12] J. R. Haddock and J. Terjéki, *Liapunov–Razumikhin functions and an invariance principle for functional differential equations*, J. Differential Equations, 48 (1983), pp. 95–122.

[13] J. K. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[14] G. S. Jones, *The existence of periodic solutions of* $f'(x) = -\alpha f(x-1)\{1+f(x)\}$, J. Math. Anal. Appl., 5 (1962), pp. 435–450.

[15] Y. Kuang, *Global stability for a class of nonlinear nonautonomous delay equations*, Nonlinear Anal., to appear.

[16] ———, *On neutral delay logistic Gause-type predator-prey systems*, Dynamics Stability Systems, 6 (1991), pp. 173–189.

[17] ———, *On neutral delay two-species Lotka–Volterra competitive systems*, J. Australian Math. Soc. Ser. B, 32 (1991), pp. 311–326.

[18] ———, *Boundedness of solutions in neutral delay predator-prey and competition systems*, Proc. of the Claremont Conference on Differential Equations and Applications to Biology and Population Dynamics, to appear.

[19] Y. Kuang and A. Feldstein, *Boundedness of solutions of a nonlinear nonautonomous neutral delay equation*, J. Math. Anal. Appl., 156 (1991), pp. 193–204.

[20] Y. Kuang and H. L. Smith, *Global stability in diffusive delay Lotka–Volterra systems*, Differential and Integral Equations, 4 (1991), pp. 117–128.

[21] ———, *Global stability for infinite delay Lotka–Volterra type systems*, J. Differential Equations, to appear.

[22] Y. Kuang, R. H. Martin, and H. L. Smith, *Global stability for infinite delay, dispersive Lotka–Volterra systems: Weakly interacting populations in nearly identical patches*, J. Dynamics and Differential Equations, 3 (1991), 339–360.

[23] S. M. Lenhart and C. C. Travis, *Global stability of a biological model with time delay*, Proc. Amer. Math. Soc., 96 (1986), pp. 75–78.

[24] R. H. Martin and H. L. Smith, *Convergence in Lotka–Volterra systems with diffusion and delay*, to appear.

[25] R. K. Miller, *On Volterra's population equation*, SIAM J. Appl. Math., 14 (1966), pp. 446–452.

[26] R. D. Nussbaum, *Periodic solutions of some nonlinear autonomous functional differential equations*, Ann. Math. Pura. Appl., 10 (1974), pp. 263–306.

[27] E. C. Pielou, *Mathematical Ecology*, Wiley Interscience, New York, 1977.

[28] F. E. Smith, *Population dynamics in Daphnia magna*, Ecology, 44 (1963), pp. 651–653.

[29] H. O. Walther, *Existence of a non-constant periodic solution of a nonlinear autonomous functional differential equation representing the growth of a single species population*, J. Math. Biol., 1 (1975), pp. 227–240.

[30] E. M. Wright, *A non-linear difference-differential equation*, J. Reine Angew. Math., 494 (1955), pp. 66–87.

[31] Z. Jackiewicz and E. Lo, *The numerical solution of neutral functional differential equations by Adams predictor corrector methods*, Tech. Report #118, Arizona State University, Tempe, AZ, 1988.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 201-208, January 1992

010

A BOUNDARY VALUE PROBLEM WITH MULTIPLE SOLUTIONS FROM THE THEORY OF LAMINAR FLOW*

S. P. HASTINGS†, C. LU‡, AND A. D. MACGILLIVRAY§

Abstract. The equation considered is

$$f^{iv}+R(ff'''-f'f'')=0,$$

with boundary conditions

$$f(0)=f''(0)=0, \quad f(1)=1, \quad f'(1)=0.$$

It is shown that for all R there is at least one solution, and for sufficiently large R there are at least three solutions. The asymptotic behavior as $R \to -\infty$ is also studied. This equation arises in studying laminar flow in a porous channel.

Key words. boundary value problems, laminar flow

AMS(MOS) subject classification. 34B15

1. Introduction. In 1953 Berman [1] introduced and studied a mathematical model that describes the laminar flow of a viscous fluid between two porous walls through which fluid is injected or removed. This model arises in a number of applications, including transpiration cooling and the separation of U_{235} from U_{238} by gaseous diffusion.

Berman reduced the boundary value problem involving the steady-state Navier-Stokes equations to a boundary value problem involving a fourth-order nonlinear ordinary differential equation

$$f^{iv}+R(f'f''-ff''')=0, \tag{1}$$

with boundary conditions

$$f(0)=f''(0)=0, \tag{2}$$

$$f'(1)=0, \qquad f(1)=1, \tag{3}$$

where $f(\eta)$ is an unknown function related to the stream function, and η is the normalized transverse coordinate ($\eta=\pm 1$ are the walls).

$R=Vd/\mu$ is a Reynolds number based upon V, the normal outward velocity, μ, the viscosity, and d, where $2d$ is the distance between the walls. $V>0$ corresponds to suction (as in the process of isotope separation of uranium) and $V<0$ corresponds to injection (as in the process of transpiration cooling.) There have been many numerical studies and asymptotic analyses published; see [8] and the references listed there. R. M. Terrill studied the formal asymptotics of some solutions in [5] and [6]. In 1978 rigorous analytical results were obtained by Skalak and Wang [4], who showed that any solution must have one of three possible types of behavior and found each of these three types numerically. In 1987 Shih [3] proved the existence of solutions for $R<0$ using the Leray-Schauder fixed point theorem. Finally, in a preprint received after this paper was accepted for publication, Wang and Hwang gave a different proof of the existence of at least three solutions for large R [7], and they also proved the uniqueness of solutions for negative R.

* Received by the editors May 23, 1990; accepted for publication (in revised form) April 29, 1991.

† University of Pittsburgh, Department of Mathematics and Statistics, Pittsburgh, Pennsylvania 15260.

‡ University of S. Illinois at Edwardsville, Department of Mathematics, Edwardsville, Illinois 62026.

§ State University of New York at Buffalo, Department of Mathematics, Buffalo, New York 14214-3093.

The plan of this paper is as follows. We find it convenient to introduce $g = Rf$. First, in §§ 2 and 3, we give a simple proof of the existence of a concave down solution for $R > 0$. The method is a topological shooting argument based on a technique of Serrin and McLeod [2]. The uniqueness of this solution remains open. These solutions are increasing. In § 3 also we obtain a second family of increasing solutions, but these change concavity. They exist only for sufficiently large R. In § 4 we consider solutions such that g is initially decreasing. First we obtain a family of solutions with g decreasing on $[0, 1]$, so that $R < 0$. This result is not new, having been obtained by Shih [3], but we give a simpler proof. Finally, in § 5, we find a family of nonmonotone solutions, which also exist only for sufficiently large R. In the course of the proofs we obtain information about the asymptotic form of the solutions for large negative R. We will strengthen this result and discuss asymptotic behavior as $R \to +\infty$ in a second paper.

2. Statement of main result and first part of proof. Our main result in this paper is the following theorem.

THEOREM 1. *For each R, the boundary value problem* (1)-(3) *has at least one solution. For sufficiently large positive R,* (1)-(3) *has at least three solutions.*

Proof. With $g = Rf$, (1), (2), (3) become

$$g^{iv} + g'g'' - gg''' = 0, \tag{4}$$

$$g(0) = g''(0) = 0, \tag{5}$$

$$g'(1) = 0, \qquad g(1) = R. \tag{6}$$

Consider the initial value problem consisting of (4)-(5) and

$$g'(0) = \lambda, \qquad g'''(0) = \mu, \tag{7}$$

where λ and μ are to be found so that the solution $g = g_{\lambda,\mu}$ of (4), (5), (7) also satisfies (6). The dependence of g on (λ, μ) will usually be suppressed in our notation.

In the (λ, μ) plane define sets A and B as follows:

$$A = \{(\lambda, \mu) | g'(1) < 0 \text{ or } g(\eta) \text{ blows up before } \eta = 1\},$$

$$B = \{(\lambda, \mu) | g'(1) > 0\}.$$

It was shown in [4] that $g^{iv} < 0$ if $\mu \neq 0$ (see below), so solutions can only blow up by tending to $-\infty$. Since solutions of (4) depend continuously on initial conditions, A and B are open subsets of R^2. If a point (λ, μ) is not in A or B, then $g_{\lambda,\mu}$ solves (4)-(6) for some R. We shall first show how to find some continua in the (λ, μ) plane which lie in the complement of $A \cup B$, and then discuss the range of values of $g_{\lambda,\mu}(1)$ for points (λ, μ) in these continua.

Since $g_{\lambda,0}(\eta) \equiv \lambda\eta$, it is clear that the positive λ axis ($\lambda > 0$, $\mu = 0$) is contained in B, while the negative λ axis is contained in A.

3. Existence of solutions with $g'(0) \geqq 0$.

LEMMA 1. *For each $\lambda > 0$ there is a $\mu_-(\lambda) < 0$ such that if $\mu \leqq \mu_-(\lambda)$, then $(\lambda, \mu) \in A$.*

Proof. From (4), (5), and (7) it follows that $g^{iv}(0) = g^{vi}(0) = 0$, and

$$g^{v} = -g''^2 + gg^{iv}. \tag{8}$$

From this we have $g^{iv}(\eta) = -\int_0^\eta \exp \int_s^\eta g(r)\, dr\, g''(s)^2\, ds$, so that if $\mu \neq 0$, then $g^{iv} < 0$ as long as the solution is defined.

Hence g''' is decreasing, and if $\mu < 0$ and $|\mu|$ is sufficiently large (depending on λ), then $g'(\eta_0) < 0$ for some η_0 in $(0, 1)$, and $g'(\eta) < 0$ so long as the solution exists to the right of η_0. This proves Lemma 1.

Since A is open, $\mu_{-}(\lambda)$ can be assumed to be continuous in λ for $\lambda > 0$. For any $\lambda_2 > \lambda_1 > 0$, let

$$\Omega = \Omega_{\lambda_1, \lambda_2} = \{(\lambda, \mu) | \lambda_1 \leqq \lambda \leqq \lambda_2, \mu_{-}(\lambda) \leqq \mu \leqq 0\}.$$

Then B contains the top boundary ($\mu = 0$, $\lambda_1 \leqq \lambda \leqq \lambda_2$) of Ω and A contains the bottom boundary ($\mu = \mu_{-}(\lambda)$, $\lambda_1 \leqq \lambda \leqq \lambda_2$). By a result in [2], there is a continuum $\gamma_{\lambda_1, \lambda_2} \in \Omega_{\lambda_1, \lambda_2}$ connecting the left and right sides of Ω such that $\gamma_{\lambda_1, \lambda_2} \cap (A \cup B)$ is empty.

Since $g_{\lambda, \mu}$ is a continuous function of (λ, μ) on $\gamma_{\lambda_1, \lambda_2}$, the existence of a solution for any $R > 0$ is a consequence of showing that for sufficiently large λ_2 and small λ_1, with (λ_1, μ_1) and (λ_2, μ_2) in $\gamma_{\lambda_1, \lambda_2}$, λ_1 and λ_2 can be chosen so that $g_{\lambda_1, \mu_1}(1) < R$ and $g_{\lambda_2, \mu_2}(1) > R$.

Since we are dealing with solutions such that $g' > 0$ and $g'' < 0$, we have $g_{\lambda, \mu}(1) < \lambda$, so we simply choose λ_1 in $(0, R)$.

To choose λ_2 we integrate the inequality $g''' < \mu$ twice to show that for a solution of (4)-(7), $0 = g'(1) < \lambda + \mu/2$. Also, $g(1) > g(\frac{1}{4}) > \frac{1}{4} g'(\frac{1}{4}) > \frac{1}{4}(\lambda + \mu/4) > \frac{1}{8}\lambda$. Hence, to obtain $g_{\lambda_2, \mu_2}(1) > R$, we just pick $\lambda_2 > 8R$.

To obtain further solutions, we consider $\lambda \geqq 0$, $\mu > 0$.

LEMMA 2. *For each $\lambda \geqq 0$, there is a $\mu_{+}(\lambda) > 0$ such that if $\mu \geqq \mu_{+}(\lambda)$, then $(\lambda, \mu) \in A$.*

Proof. We first show that for sufficiently large μ (depending on λ, which for the moment is fixed), there is an η_μ in $(0, 1)$ with $g'''(\eta_\mu) = \mu/2$. Also, $\lim_{\mu \to \infty} \eta_\mu = 0$. To see this, use (7) and (8) to show that as long as $g''' \geqq \mu/2$,

$$g''' \leqq \mu - \frac{\mu^2 \eta^4}{48}.$$

Hence, $\eta_\mu \leqq (24/\mu)^{1/4}$. Furthermore, from (8), $g^v < 0$ in $(0, \eta_\mu)$, so g^{iv} is decreasing. By the mean value theorem applied to g''',

$$g^{iv}(\eta_\mu) \leqq -C\mu^{5/4}$$

for some $C > 0$ independent of μ.

Also, g^{iv} continues to decrease as long as $g > 0$, so beyond η_μ, as long as $g' > 0$, we have

$$g''' \leqq \frac{\mu}{2} - C\mu^{5/4}(\eta - \eta_\mu). \tag{9}$$

Integrating this twice and using the initial conditions shows that for sufficiently large μ, either $g(1)$ is not defined, or $g'(1) < 0$. In either case, $(\lambda, \mu) \in A$. This proves Lemma 2.

We now see that if $\lambda_2 > 0$, then there is a continuum δ_{0, λ_2} connecting $\lambda = 0$ with $\lambda = \lambda_2$ in the region $\mu \geqq 0$, and δ_{0, λ_2} is in the complement of $A \cup B$. In fact, since the positive λ-axis is contained in B, the continuum δ_{0, λ_2} lies in $\mu > 0$ except possibly at $\lambda = 0$. But just below, in Lemma 3, we shall show that in δ_{0, λ_2}, $\mu > 0$ when $\lambda = 0$. The corresponding solutions of (4)-(6) satisfy $g' > 0$ on $(0, 1)$, $g'' > 0$ on an initial interval $(0, \alpha)$, and $g'' < 0$ on $(\alpha, 1)$.

Also, $g(1)$ varies continuously on δ_{0, λ_2}. However it will be shown that $g(1)$ does not take on all positive values R.

LEMMA 3. $\inf \{g_{\lambda, \mu}(1) | \lambda \geqq 0, \mu > 0,$ *and* $g'_{(\lambda, \mu)}(1) = 0\} > 0$.

Proof of Lemma 3. Expand $g_{\lambda, \mu}(\eta)$ in a Taylor series in (λ, μ) around $(0, 0)$. The lowest order terms are

$$\phi(\eta)\mu + \psi(\eta)\lambda,$$

where ϕ and ψ satisfy

$$\phi'''=1,\ \phi(0)=\phi'(0)=\phi''(0)=0,$$
$$\psi'''=0,\ \psi(0)=\psi''(0)=0,\ \psi'(0)=1.$$

Here $\phi=\partial g/\partial\mu|_{(\lambda,\mu)=(0,0)}$, $\psi=\partial g/\partial\lambda|_{(\lambda,\mu)=(0,0)}$. Since $\phi'(1)>0$ and $\psi'(1)>0$, it follows that there is some $\rho>0$ such that, if $\mu\geqq 0$, $\lambda\geqq 0$, and $\rho>\mu^2+\lambda^2>0$, then $g'_{\lambda,\mu}(1)>0$ and $(\lambda,\mu)\in B$. Also, $g_{(\lambda,\mu)}(1)>0$ if $\lambda>0$, $\mu>0$ and $g'>0$ on $[0,1)$, while $g'_{(\lambda,\mu)}(1)<0$ for large μ. It follows that

$$\inf\{g_{(\lambda,\mu)}(1)|0\leqq\lambda\leqq\rho,\ \mu>0,\ g'_{(\lambda,\mu)}(1)=0\}>0.$$

We shall show later that for large λ, $g_{(\lambda,\mu)}(1)$ is large if $g'_{(\lambda,\mu)}(1)=0$, which will prove Lemma 3.

We now show that there is a solution of this type for any sufficiently large R. It is convenient to introduce another scaling. For $\lambda\neq 0$, let

$$h(\eta)=\frac{g(\eta)}{|\lambda|}.$$

Then

$$\varepsilon h^{iv}=-h'h''+hh''', \tag{10}$$

where $\varepsilon=1/|\lambda|$. Also,

$$h(0)=h''(0)=0,\quad h'(0)=1,\quad h'''(0)=\varepsilon\mu. \tag{11}$$

We have shown that for each $\varepsilon>0$ there are at least two values of μ, one negative and one positive, such that

$$h'(1)=0. \tag{12}$$

Let h_ε denote some solution of (10)-(12) with $\mu>0$. We must show that $\lim_{\varepsilon\to 0^+}(h_\varepsilon(1))/\varepsilon=\infty$. In fact, we show more.

LEMMA 4. $\liminf_{\varepsilon\to 0^+} h_\varepsilon(1)\geqq 1$.

Proof. Suppose, for some fixed $r<1$, that $h_{\varepsilon_j}(1)<r$ for some sequence of $\varepsilon_j\to 0$. Since $h'_\varepsilon>0$ on $[0,1)$, and h''' is decreasing, it follows that $h'_{\varepsilon_j}(r)<1$. Hence there is an $\eta_j<r$ such that $h''_{\varepsilon_j}(\eta_j)=0$, $h'_{\varepsilon_j}(\eta_j)>1$, and h''_{ε_j} decreases on $[\eta_j,1]$. We distinguish between two cases. Either $h''_{\varepsilon_j}((1+r)/2)\to 0$ or, for some subsequence if necessary, and some $\delta>0$, $h''_{\varepsilon_j}((1+r)/2)\leqq-\delta$. In the first case, $h''_{\varepsilon_j}\to 0$ uniformly on $[\eta_j,(1+r)/2]$. But this implies that $h_{\varepsilon_j}((1+r)/2)>r$ for large j, contradicting an earlier assumption. In the second case, $h''_{\varepsilon_j}\leqq-\delta$ on $[(1+r)/2,1]$, and since $h'_{\varepsilon_j}(r)<1$, we obtain a contradiction using (8). This proves Lemma 4 and establishes the existence of a second solution for large R.

4. Monotone solutions with $\lambda<0$. To obtain further solutions we now consider $\lambda<0$. If $\mu<0$, then g, g', g'', and g''' are all negative and decreasing on $(0,1)$. If $\mu=0$ then $g(\eta)=\lambda\eta$. For $\mu>0$, the following lemmas are useful.

LEMMA 5. *If $\lambda<0$, $\mu>0$, then $g'''>0$ as long as $g'\leqq 0$.*

Proof. We have seen that $g^{iv}<0$ if $\mu\neq 0$. But from (4) it is apparent that at the first zero of g''', where $g''>0$, sign $(g^{iv})=-\text{sign}(g')$. Hence $g'>0$ at this point.

LEMMA 6. *Let*

$$\phi(\eta)=\frac{\partial g_{\lambda,\mu}(\eta)}{\partial\mu}.$$

Then ϕ, ϕ', ϕ'' are all positive as long as $g'\leqq 0$.

Proof. Integrating (4) shows that

$$g''' = \lambda^2 + \mu - g'^2 + gg''. \tag{13}$$

Differentiating with respect to μ shows that

$$\phi''' = 1 - 2\phi' g' + \phi g'' + g\phi'',$$

with

$$\phi(0) = \phi'(0) = \phi''(0) = 0, \qquad \phi'''(0) = 1.$$

At the first zero, if any, of ϕ'', $\phi''' > 0$ if $g' \leqq 0$, a contradiction. This proves Lemma 6.

Still taking a fixed $\lambda < 0$, we consider the solutions for $\mu > 0$. If μ is small, then $g'(1) < 0$, and, by Lemma 5, $g'' > 0$ on $[0, 1]$. Lemma 6 then implies that $\phi'(1) > 0$.

LEMMA 7. *For sufficiently large μ, $g'(\eta) > 0$ for some η in $(0, 1)$.*

Proof. On an initial interval $g' < 0$, and as long as this inequality persists, we have

$$\lambda < g' < 0, \quad \lambda\eta \leqq g(\eta) < 0, \quad g'' > 0, \quad \text{and} \quad g^{iv} \geqq \lambda g'''.$$

Hence $g''' \geqq g'''(0)\, e^{\lambda\eta}$. For fixed λ, integrating two times shows that g' becomes positive at some $\eta < 1$, if $\mu = g'''(0)$ is sufficiently large, completing the proof.

Recall that $g'(1) < 0$ if $\mu = 0$, $\lambda < 0$. Lemmas 6 and 7 show that there is a unique $\mu = \mu_i(\lambda) > 0$ such that $g' < 0$ on $[0, 1]$ and $g'(1) = 0$. (Here "i" stands for "intermediate." The solution is unique because Lemma 6 shows that $g'(1)$ is strictly increasing in μ whenever $g' < 0$ on $[0, 1)$ and $g'(1) = 0$. Hence $g'(1)$ can only increase from negative to positive as μ increases, so long as g is negative.) The region $0 \leqq \mu < \mu_i(\lambda)$ is contained in A. Since $\partial g'_{\lambda,\mu}(1)/\partial\mu|_{\mu=\mu_i(\lambda)} > 0$ at $\eta = 1$, B contains an open set of the (λ, μ) plane with $\mu = \mu_i(\lambda)$ as its lower boundary for $\lambda < 0$.

An additional simple result clarifies the picture further.

LEMMA 8. *For $\mu > \mu_i(\lambda)$, g' has exactly one zero in the initial interval where $g < 0$.*

Proof. If not, then there is a point in $(0, 1)$ where

$$g' = 0, \quad g'' < 0, \quad g''' < 0, \quad g < 0.$$

However, from (13), g''' is positive if $g' = 0$, $gg'' > 0$. This proves Lemma 8.

The curve $\mu = \mu_i(\lambda)$, $\lambda < 0$, gives solutions for at least some negative values of R. Note that this curve is continuous, from the implicit function theorem, because $\mu_i(\lambda)$ can be viewed locally as a solution to the equation

$$g'_{\lambda,\mu}(1) = 0.$$

and from Lemma 6 we know that $\partial g'_{\lambda,\mu}(1)/\partial\mu \neq 0$. The existence of solutions for negative λ is not new, as it was shown in [3] that there is a solution to (1)-(3) for any $R < 0$. To obtain this from our arguments, we must show that if $\mu = \mu_i(\lambda)$, then $g(1) \to 0$ as $\lambda \to 0^-$, while $g(1) \to \infty$ as $\lambda \to -\infty$.

The first of these is trivial, since $g(1) \geqq \lambda$. For the second we again let $h(\eta) = g(\eta)/|\lambda|$, so that h satisfies (10), but with

$$h(0) = h''(0) = 0, \quad h'(0) = -1, \quad h'''(0) = \varepsilon\mu. \tag{14}$$

We now wish to use h_ε to denote the solution of (10), (12), (14) with $h' < 0$ on $[0, 1)$. The asymptotic form of h_ε is given by the following theorem.

THEOREM 2.

$$\lim_{\varepsilon \to 0^+} h_\varepsilon(\eta) = -\frac{2}{\pi} \sin\left(\frac{\pi}{2}\eta\right),$$

uniformly on $0 \leqq \eta \leqq 1$.

LEMMA 9. $\lim_{\varepsilon\to 0}\varepsilon^2\mu=0$.

Proof. Suppose, on the contrary, that there is a $\delta>0$ and a sequence of ε's tending to zero such that $\varepsilon^2\mu>\delta$, where $\mu=\mu_i(-1/\varepsilon)$. Thus $h_\varepsilon'''(0)=\varepsilon\mu\to\infty$. We have seen that for the solution under discussion, on $[0,1]$,

$$h'''>0,\quad h''>0,\quad -1<h'<0,$$

so

$$\varepsilon h'''\geqq\delta+hh''.$$

On some initial interval $[0,\alpha]$, $h''\leqq 2$. In this interval, $h'''\geqq(\delta+2h)/\varepsilon$. It follows easily that for sufficiently small ε, either $h'=0$ or $h''=2$, before $\eta=\frac{1}{2}$. But we are assuming that $h'<0$ on $[0,1)$. If $h''(\alpha)=2$, then $h''>2$ on $(\alpha,1]$, and $h'(1)>0$, which again is impossible. This proves Lemma 9.

Hence, there is a $\delta(\varepsilon)>0$ such that $\delta(\varepsilon)\to 0$ as $\varepsilon\to 0$ and $0<\varepsilon h'''\leqq\delta(\varepsilon)=\varepsilon^2\mu=\varepsilon h'''(0)$ on $[0,1]$. The equivalent of (13) for h is

$$\varepsilon h'''=1+\varepsilon^2\mu-h'^2+hh'', \tag{15}$$

so that on $[0,1]$,

$$-\delta(\varepsilon)\leqq 1-h'^2+hh''\leqq 0.$$

Since $h'<0$ on $[0,1)$, h' can be expressed as a function of h:

$$h'(\eta)=Q(h(\eta)).$$

Then $Q(0)=-1$, $-1<Q<0$ on $(h(1),0)$, and

$$Q^2-1-\delta\leqq hQQ'\leqq Q^2-1<0,$$

or

$$\frac{QQ'}{Q^2-1}\leqq\frac{1}{h}\leqq\frac{QQ'}{Q^2-1-\delta(\varepsilon)}.$$

These inequalities can be integrated from $h(1)$ to $h(\eta)$ (the resulting improper integral is convergent) giving, first,

$$1-\frac{h(\eta)^2}{h(1)^2}\leqq h'(\eta)^2\leqq 1+\delta-\frac{h(\eta)^2}{h(1)^2},$$

and then, integrating from zero to η,

$$-h(1)\sin\left(\frac{\sqrt{1+\delta}\,\eta}{h(1)}\right)\leqq h(\eta)\leqq -h(1)\sin\left(\frac{\eta}{h(1)}\right).$$

Since $h(\eta)$ is decreasing exactly on $[0,1]$, with $h'(1)=0$, the only possibility as $\varepsilon\to 0$ is that $h(1)\to -2/\pi$, and Theorem 2 follows.

5. Nonmonotone solutions for $\lambda<0$. As pointed out earlier, a region above $\mu=\mu_i(\lambda)$, $\lambda<0$ lies in the set B. Also, a neighborhood of the segment $\{\lambda=0, 0<\mu<\rho\lambda=0, 0\leqq\mu<\rho\}$, lies in B. We wish now to show that these regions overlap.

First we observe that $\lim_{\lambda\to 0^-}\mu_i(\lambda)=0$. This is apparent because $g'(1,0,\mu)>0$ if μ is positive and sufficiently small.

Again, let $\phi(\eta,\lambda,\mu)=(\partial g/\partial\mu)(\eta,\lambda,\mu)$. We have seen that $\phi'''(\eta,0,0)\equiv 1$, so $\phi'(1,0,0)=\frac{1}{2}$. Hence, $\phi'(1,\lambda,\mu)\geqq\frac{1}{4}$, for (λ,μ) in some neighborhood of $(0,0)$. Therefore there are $\mu_0>0$, $\lambda_0<0$ such that $\phi'(1,\lambda,\mu)>0$ for $0\leqq\mu\leqq\mu_0$, $\lambda_0\leqq\lambda\leqq 0$. Since

$g'(1, \lambda, \mu_i(\lambda)) = 0$, we see that $g'(1, \lambda, \mu) > 0$ if $|\lambda_0|$ is small enough to ensure that $\mu_i(\lambda_0) < \mu_0$, and $\lambda_0 \leqq \lambda < 0$, $\mu_i(\lambda) < \mu < \mu_0$. This proves that B contains an open set bounded below by $\mu = \mu_i(\lambda)$ for $\lambda \leqq 0$ and by $\mu = 0$ for $\lambda > 0$.

To obtain our final set of solutions we need Lemma 10.

LEMMA 10. *For each $\lambda < 0$ there is a $\mu_+(\lambda) > 0$ such that if $\mu \geqq \mu_+(\lambda)$, then $(\lambda, \mu) \in A$.*

Remark. After proving Lemma 10 we will have defined $\mu_+(\lambda)$ for all λ. The component of A containing the region $\mu \geqq \mu_+(\lambda)$ must be separated from B by a continuum γ which extends over the entire λ-axis. We have seen that in γ, $g(1) \to +\infty$ as $\lambda \to +\infty$. After proving Lemma 10, we will show that in γ, $g(1) \to +\infty$ as $\lambda \to -\infty$, as well. This will give three solutions for any sufficiently large positive R.

Proof. For large μ, as long as $g'' > 0$ and $g' \leqq \sqrt{\mu/2}$ we have, from (13),

$$\mu\eta \geqq g'' > 0, \quad g' > \lambda, \quad g > \lambda\eta,$$

and

$$g''' \geqq \frac{\mu}{2} + \lambda\mu\eta^2.$$

If $\eta < 1/(2\sqrt{|\lambda|})$, then $g''' \geqq \mu/4$, and integrating this three times, using the initial conditions on g, enables us to conclude that before $\eta = \sqrt{24|\lambda|/\mu}$, either $g = 0$ or $g' = \sqrt{\mu/2}$.

On the other hand, $g'' \leqq \mu\eta$, $g' \leqq \lambda + \mu\eta^2/2$, so $g'(\sqrt{24|\lambda|/\mu}) \leqq 11|\lambda| < \sqrt{\mu/2}$ for μ large. Hence we have shown that for large μ, $g = 0$ before $\eta = \sqrt{24|\lambda|/\mu}$. Furthermore, at $g = 0$, $g''' \geqq \lambda^2 - \mu/2 + \mu \geqq \mu/2$ and $0 \leqq g' \leqq 11|\lambda|$. From here the proof that $g'(1) < 0$ proceeds with estimates which are very similar to those in Lemma 2, and we omit the details. This proves Lemma 10.

The proof of Theorem 1 will be completed by showing that

$$\lim_{\substack{\lambda \to -\infty \\ g'_{(\lambda,\mu)}(0)=0}} g_{\lambda,\mu}(1) = \infty.$$

Once again, let $h = g/|\lambda|$, and consider (10), (12), (14). For small $\varepsilon > 0$, we assume $\mu > 0$ has been chosen so that $h'(\eta_1) = 0$ for some η_1 in $(0, 1)$ and also $h'(1) = 0$. We must show that $h_\varepsilon(1)/\varepsilon \to \infty$ (at least for some sequence of ε's tending to zero). Suppose this is not the case and there is an $M > 0$ such that $h_\varepsilon(1) \leqq M\varepsilon$ for all small ε. Equivalently, $h_\varepsilon(\eta) \leqq M\varepsilon$ for all η and small ε. For each ε the solution $h = h_\varepsilon$ has the following qualitative behavior.

$$\begin{aligned} h^{iv} &< 0 \quad \text{on } (0, 1], \\ h' &< 0 \quad \text{on some interval } [0, \eta_1),\ h' > 0 \quad \text{on } (\eta_1, 1], \\ h'' &> 0 \quad \text{on } [0, \eta_2],\ h'' < 0 \quad \text{on } (\eta_2, 1], \\ h''' &> 0 \quad \text{on } [0, \eta_3),\ h''' < 0 \quad \text{on } (\eta_3, 1], \\ h &< 0 \quad \text{on } (0, \eta_4),\ h > 0 \quad \text{on } (\eta_4, 1], \end{aligned}$$

where $0 < \eta_1 < \eta_3 < \eta_2$.

LEMMA 11. $\eta_2 > \eta_4$.

Proof. This is obvious if $\eta_3 \geqq \eta_4$. If $\eta_3 < \eta_4$, then

$$0 > \varepsilon h^{iv} = -h'h'' + hh''' \geqq -h'h''$$

on $[\eta_3, \eta_4]$, which implies that $h'' > 0$ on this interval, as desired.

LEMMA 12. $h'(\eta_2) > 1$.

Proof. Since $h'''(\eta_2) < 0$ and $h''(\eta_2) = 0$, the result follows from (15).

Continuing with the proof of Theorem 1, there is an $\eta_5 > \eta_2$ with $h'(\eta_5) = 1$. Also, $h'' < 0$ on $(\eta_2, \eta_5]$, so it is clear that $h(1) \geqq h(\eta_5) \geqq \eta_5 - \eta_2$, and thus $\eta_5 - \eta_2 \leqq M\varepsilon$. Also, h''' is decreasing, so

$$h'''(\eta_5) \leqq \frac{h''(\eta_5)}{(\eta_5 - \eta_2)} \leqq \frac{h''(\eta_5)}{M\varepsilon}.$$

However, since $h'(\eta_5) = 1$, we have from (15)

$$h'''(\eta_5) = \varepsilon\mu + \frac{h(\eta_5)}{\varepsilon} h''(\eta_5) \geqq Mh''(\eta_5).$$

These inequalities are inconsistent if $\varepsilon < 1/M^2$. We were using the assumption that $h(1)/\varepsilon \leqq M$, which therefore is impossible for small ε. This completes the proof of Theorem 1.

REFERENCES

[1] A. S. BERMAN, *Laminar flow in channels with porous walls*, J. Appl. Phys., 24 (1953), pp. 1232-1235.

[2] J. B. MCLEOD and J. SERRIN, *The existence of similar solutions for some laminar boundary layer problems*, Arch. Rational Mech. Anal., 31 (1968), pp. 288-303.

[3] K.-G. SHIH, *On the existence of solutions of an equation arising in the theory of laminar flow in a uniformly porous channel*, SIAM J. Appl. Math., 47 (1987), pp. 526-533.

[4] F. M. SKALAK and C. Y. WANG *On the non-unique solutions of laminar flow through a porous tube or channel*, SIAM J. Appl. Math., 34 (1978), pp. 535-544.

[5] R. M. TERRILL, *Laminar flow in a uniformly porous channel*, Aeronautical Quart., 15 (1964), pp. 299-310.

[6] ———, *Laminar flow in a uniformly porous channel with large injection*, Aeronautical Quart., 16 (1965), pp. 323-332.

[7] C. A. WANG AND T.-W. HWANG, *On multiple solutions of Berman's equation*, Proc. Roy. Acad. Edinburgh, to appear.

[8] S. W. YUAN, AND A. B. FINKELSTEIN, *Laminar pipe flow with injection and suction through a porous wall*, Trans. ASME Ser. E, J. Appl. Mech., 78 (1956), pp. 719-724.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 209-221, January 1992

A NONLINEAR DIFFERENTIAL OPERATOR SERIES THAT COMMUTES WITH ANY FUNCTION*

PETER J. OLVER†

Abstract. A natural differential operator series is one that commutes with every function. The only linear examples are the formal series operators $e^{\alpha z D}$ representing translations. This paper discusses a surprising natural *nonlinear* "normally ordered" differential operator series, arising from the Lagrange inversion formula. The operator provides a wide range of new higher-order derivative identities and identities among Bell polynomials. These identities specialize to a large variety of interesting identities among binomial coefficients and classical orthogonal polynomials, a number of which are new.

Key words. natural differential operator series, normal ordering, Lagrange inversion, Bell polynomial, orthogonal polynomial

AMS(MOS) subject classifications. 26A24, 05A15, 05A19

1. Introduction. An operator $\mathcal{D}$ is called *natural* if it commutes with arbitrary functions, i.e.,

$$\Phi(\mathcal{D}u) = \mathcal{D}\Phi(u) \tag{1}$$

for all scalar functions Φ. In this paper we will take $u(t)$ to be a formal power series in the variable t, and $\mathcal{D}$ to be a formal series of differential operators. A simple example of a natural operator in this context is the exponential operator e^{zD}, where $D = d/dt$, which, by Taylor's theorem, coincides with the translation operator $e^{zD}u(t) = u(t+z)$. The proof that e^{zD} is natural is then elementary:

$$\Phi(e^{zD}u(t)) = \Phi(u(t+z)) = e^{zD}\Phi(u(t)). \tag{2}$$

In fact, it is not hard to show that the translation operators $e^{\alpha z D}$ are essentially the only linear natural differential operator series. It is therefore rather surprising that there exist nonlinear natural differential operator series! The main result of this paper is that the series operator

$$D^{-1} :e^{zDu}: D = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!} D^{n-1} \cdot u^n \cdot D \tag{3}$$

is natural, i.e., for any analytic function $\Phi(u)$, and any formal power series $u(t)$,

$$D^{-1} :e^{zDu}: D\Phi(u) = \Phi(D^{-1} :e^{zDu}: Du). \tag{4}$$

In (3) the colons mean that the operator is "normally ordered," meaning that all the multiplication terms appear after all the differentiations. This is reminiscent of the Wick ordering in quantum mechanics [4], although not quite the same. I do not know if the identity (4) has any bearing on this subject.

Two proofs of this identity will be discussed. The first is an application of the classical Lagrange inversion theorem [3], [7]. In fact, it will follow that the operator (3) formally represents the implicitly defined variable translation $x = t + zu(x)$, which explains its naturality. The second proof uses techniques from the Frobenius theory of partial differential equations, a method of independent interest, and appears in an

* Received by the editors January 3, 1990; accepted for publication (in revised form) February 21, 1991. This research was supported in part by National Science Foundation grant DMS 89-01600.

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

appendix. By choosing different elementary functions Φ in the identity (4), we are led to a large class of interesting new identities involving higher-order derivatives of scalar functions. Moreover, specializing the resulting derivative identities to various elementary types of functions $u(x)$, leads to, among others, the Hagen–Rothe binomial coefficient identity [5], the Abel identity [3], and a number of interesting identities among classical orthogonal polynomials, including Hermite, Legendre, and Jacobi polynomials, that I have not been able to find in the literature. In another direction, using the standard connection between higher-order derivatives of compositions of functions and the Bell polynomials [3], [10], these derivative identities are easily shown to be equivalent to a large collection of apparently new identities for Bell polynomials.

This work arose from an ongoing investigation into the canonical forms for bi-Hamiltonian systems [9], and applications of these results to the precise integrability of canonical bi-Hamiltonian systems can be found there.

2. Differential operators and normal ordering. We will be concerned with formal power series whose coefficients are differential operators. These in turn can be applied to analytic functions or formal power series, leading in turn to further formal series.

Let t be a scalar variable. We use D to denote the derivative operator d/dt. Let

$$F(z)=\sum_{n=0}^{\infty} c_n z^n \tag{5}$$

be a formal power series in the scalar variable z. We can form the operator series

$$F(zD)=\sum_{n=0}^{\infty} c_n z^n D^n$$

which, when applied to any analytic function $f(t)$ results in a formal power series

$$F(zD)f(t)=\sum_{n=0}^{\infty} c_n z^n f^{(n)}(t) \tag{6}$$

in the derivatives of $f^{(n)}=D^nf=d^nf/dt^n$ of f. For example, by Taylor's theorem, the exponential operator

$$e^{zD}f(t)=\sum_{n=0}^{\infty}\frac{1}{n!}z^nf^{(n)}(t)=f(t+z) \tag{7}$$

coincides with the operator of translation in z. If

$$f(t)=\sum_{i=0}^{\infty} f_i t^i$$

is itself a formal power series, then (6) is a formal power series in both z and t whose coefficients depend on the coefficients f_i of f. In particular, evaluating this identity at $t=0$ leads to the formal series

$$F(zD)f(t)|_{t=0}=\sum_{n=0}^{\infty} n!\,c_nf_nz^n. \tag{8}$$

Note that, under the natural identification of the coefficients of f with the derivatives of f at $t=0$, which is $f_n=(n!)^{-1}f^{(n)}(0)$, we recover the original equality (6). In fact, we can replace zero by any other value of t; hence we can use (8) to evaluate the series (6). This remark will be of use later on.

We now wish to extend our range of operators to certain types of nonlinear operators. By "nonlinear" we mean that the operator itself depends on an analytic function or a formal power series $u(t)$, so that the operator will, in general, be a nonlinear function of u. However, it still acts linearly when applied to other power series. The most elementary operators associated with a formal power series (5) are the nonlinear operators

$$F(zuD)=\sum_{n=0}^{\infty} c_n z^n (uD)^n \quad\text{and}\quad F(zDu)=\sum_{n=0}^{\infty} c_n z^n (D\cdot u)^n. \tag{9}$$

Note that since the operators of differentiation D and multiplication by u do not commute, these two operators are not the same; their commutator

$$[D,u]=D\cdot u-u\cdot D=u' \tag{10}$$

is the operator of multiplication by $u'=u^{(1)}$. For example,

$$\begin{aligned}e^{zuD}v&=\sum_{n=0}^{\infty}\frac{z^n}{n!}(uD)^n v\\&=v+zuv'+\frac{1}{2}z^2(u^2v''+uu'v')\\&\quad+\frac{1}{6}z^3(u^3v'''+3u^2u'v''+uu'^2v'+u^2u''v')+\cdots,\\e^{zDu}v&=\sum_{n=0}^{\infty}\frac{z^n}{n!}(D\cdot u)^n v\\&=v+z(uv'+u'v)+\frac{1}{2}z^2(u^2v''+3uu'v'+(uu''+u'^2)v)\\&\quad+\frac{1}{6}z^3(u^3v'''+6u^2u'v''+(4u^2u''+7uu'^2)v'\\&\qquad+(u^2u'''+4uu'u''+u'^3)v)+\cdots.\end{aligned}$$

A further type of operator is found by ordering the factors in the series in yet another way.

DEFINITION 1. Given a formal power series (5), the *normally ordered operator series* is defined to be

$$:F(zDu):=\sum_{n=0}^{\infty}c_nz^nD^n\cdot u^n. \tag{11}$$

Thus the action of $:F(zDu):$ on a function $f(t)$ is given by

$$:F(zDu):f=\sum_{n=0}^{\infty}c_nz^nD^n\{u^nf\}.$$

The colons in the notation (11) are to distinguish this operator from the more standard operator series (9). For example

$$\begin{aligned}:e^{zDu}:\,v&=\sum_{n=0}^{\infty}\frac{1}{n!}z^nD^n(u^nv)\\&=v+z(uv'+u'v)+\frac{1}{2}z^2(u^2v''+3uu'v'+2(uu''+u'^2)v)\\&\quad+\frac{1}{6}z^3(u^3v'''+9u^2u'v''+(9u^2u''+18uu'^2)v'\\&\qquad+(3u^2u'''+18uu'u''+6u'^3)v)+\cdots.\end{aligned}$$

The colon notation is borrowed from quantum mechanics. Indeed, these operators remind us of the Wick ordering used in quantum field theory [4], in which all the creation operators appear to the left of all the annihilation operators. (Indeed, the commutation relations (10) are also reminiscent of the standard commutation relations, but only coincide when $u = t$.) However, this is not really the ordering adopted here, since in the harmonic oscillator, the creation and annihilation operators are certain combinations of derivative and multiplication operators.

3. Natural operators. Certain formal series differential operators play a distinguished role, in that they commute with functional evaluation. We make the following definition.

DEFINITION 2. A series differential operator $\mathcal{D}$ is called *natural* if it commutes with all functions, i.e.,

$$\Phi(\mathcal{D}u) = \mathcal{D}\Phi(u) \tag{12}$$

for all scalar functions Φ and all formal series u.

A simple example is provided by the translation operator e^{zD}, as shown in the introduction. It is not hard to show that the following is essentially the only linear example.

PROPOSITION 3. *The operators $e^{\alpha zD}$, α a constant, are the only natural formal series linear differential operators of the form $\mathcal{D} = F(zD)$.*

Proof. Let

$$\mathcal{D} = F(zD) = \sum_{n=0}^{\infty} c_n z^n D^n.$$

First set $z = 0$ in (12), which gives $\Phi(c_0 u) = c_0 \Phi(u)$. For this to hold for all Φ, the leading term of $\mathcal{D}$ must be $c_0 = 1$. Now, assume by induction that we have shown that $c_j = \alpha^j/j!$, for $j \leqq n-1$, where $\alpha = c_1$, and $n \geqq 2$. Let $\Phi(u) = u^2$. Then the coefficient of z^n in (9) is

$$2c_0 c_n u u^{(n)} + 2c_1 c_{n-1} u' u^{(n-1)} + \cdots = c_n D^n(u^2).$$

This readily implies that $c_n = \alpha^n/n!$, completing the induction. (Note that in fact we only needed to check (12) for quadratic functions Φ to prove this result.)

The main result of this paper is the following example of a nonlinear natural differential operator.

THEOREM 4. *Let $u(t)$ be a formal power series and let $D = d/dt$. Then the series differential operator*

$$D^{-1} :e^{zDu}: D = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!} D^{n-1} \cdot u^n \cdot D \tag{13}$$

is natural, i.e., for any analytic function $\Phi(u)$,

$$D^{-1} :e^{zDu}: D\Phi(u) = \Phi(D^{-1} :e^{zDu}: Du). \tag{14}$$

Proof. This result follows as a direct consequence of the famous Lagrange inversion formula: cf. [3, p. 150] or [7, pp. 113, 114]. According to equation (5) of Melzak [7, p. 113], if $u(t)$ is any analytic function (or formal power series), and we define $x = \xi(z, t)$ implicitly by the formula

$$x = t + zu(x), \tag{15}$$

then, for any analytic function $f(t)$, we have the classical *Lagrange inversion formula*

$$(16)\qquad f(x)=f(t)+\sum_{n=1}^{\infty}\frac{z^n}{n!}\{D^{n-1}\cdot u(t)^n\cdot D\}f(t)=D^{-1}:e^{zDu}:Df(t).$$

Now set $f(t)=\Phi(u(t))$, so that (16) becomes

$$(17)\qquad \Phi(u(x))=D^{-1}:e^{zDu}:D\Phi(u(t)).$$

On the other hand, according to the formula at the bottom of page 144 of [7], for any analytic function $g(x)$, evaluated at (15),

$$\frac{\partial^n}{\partial z^n}g(\xi(z,t))|_{z=0}=D^{n-1}(u(t)^nDg(t)),\qquad n\geqq 1.$$

Therefore, taking $g=u$, we find the expansion

$$\begin{aligned}u(x)&=\sum_{n=0}^{\infty}\frac{z^n}{n!}\frac{\partial^n}{\partial z^n}u(\xi(z,t))|_{z=0}\\&=u(t)+\sum_{n=0}^{\infty}\frac{z^n}{n!}D^n(u(t)^nDu(t))\\&=D^{-1}:e^{zDu}:Du(t).\end{aligned}$$

Substituting this into (17) completes the proof of (13).

In view of the proof, then, it is no longer surprising that the operator series (13) is natural, since it corresponds to the variable translation (15) via Lagrange inversion. More generally, we can introduce the translation

$$(18)\qquad x=t+\Psi(z,u(x),u'(x),\cdots,u^{(n)}(x)),$$

which has a corresponding differential operator series, which will also clearly be natural. For example, the operators

$$D^{-1}:e^{zD\Psi(u,u',\cdots)}:D,$$

where Ψ is any analytic function of u and its derivatives, are also natural. An interesting problem that I have not tried to investigate is whether there exist other classes of natural differential operators, although it seems reasonable to conjecture that only the operators associated with such translations are natural.

4. Derivative identities. Just as generating function identities leads to combinatorial identities, so any natural differential operator leads to a large class of derivative identities, obtained by considering different functions Φ in the basic condition (14). Here we present some of the more elementary derivative identities to be found as a consequence of the main theorem. We first compute the basic formula

$$(19)\qquad \begin{aligned}\zeta(u)=D^{-1}:e^{zDu}:Du&=u+\sum_{n=1}^{\infty}\frac{z^n}{n!}D^{n-1}\{u^nu'\}\\&=\sum_{n=0}^{\infty}\frac{z^n}{(n+1)!}D^n(u^{n+1}).\end{aligned}$$

More generally, we find that, for $\Phi(u)=u^k$,

$$(20)\qquad D^{-1}:e^{zDu}:Du^k=\sum_{n=0}^{\infty}\frac{k}{n+k}\frac{z^n}{n!}D^n(u^{n+k}).$$

As long as k is not a negative integer, (20) is valid as it stands. It also remains correct when $k=-j$ is a negative integer, *provided* we interpret the term corresponding to $n=j$ in the summation according to the general "rule"

$$(21)\qquad \lim_{m\to 0}\frac{1}{m}D^n(u^m)=\lim_{m\to 0}D^{n-1}(u^{m-1}u')=D^n(\log u),\qquad n\geqq 1.$$

Note that if $n=0$, the term $(k/(n+k))D^n(u^{n+k})=u^k$ is not a problem. Now, according to Theorem 4, the series (20) is the kth power of the series (19). This implies certain unusual identities among higher-order derivatives of powers of u. For instance, taking the case $k=2$, the series identity

$$\sum_{n=0}^{\infty}\frac{2}{n+2}\frac{z^n}{n!}D^n(u^{n+2})=\left(\sum_{n=0}^{\infty}\frac{z^n}{(n+1)!}D^n(u^{n+1})\right)^2$$

implies the following derivative identities:

$$D^n(u^{n+2})=\sum_{i=0}^{n}\frac{n+2}{2(i+1)(n-i+1)}\binom{n}{i}D^i(u^{i+1})\cdot D^{n-i}(u^{n-i+1}).$$

More generally, if we apply the identities corresponding to $\Phi(u)$ being u^{k+l}, u^k, and u^l, then the series identity

$$\sum_{n=0}^{\infty}\frac{k+l}{n+k+l}\frac{z^n}{n!}D^n(u^{n+k+l})=\left(\sum_{n=0}^{\infty}\frac{k}{n+k}\frac{z^n}{n!}D^n(u^{n+k})\right)\left(\sum_{n=0}^{\infty}\frac{l}{n+l}\frac{z^n}{n!}D^n(u^{n+l})\right)$$

implies the additional derivative identities

$$(22)\qquad \begin{aligned}&\frac{k+l}{n+k+l}D^n(u^{n+k+l})\\&=\sum_{i=0}^{n}\frac{kl}{(i+k)(n-i+l)}\binom{n}{i}D^i(u^{i+k})\cdot D^{n-i}(u^{n-i+l}).\end{aligned}$$

These identities are valid for arbitrary (positive and negative) values of k, l, provided we use the rule (21) if either $n+k+l=0$, or any of the summation terms $i+k=0$ or $n-i+l=0$. For example, if we take $k=-1$, $l=1$, we find that the series

$$\eta(u)=\frac{1}{u}-z\frac{u'}{u}+\sum_{n=2}^{\infty}\frac{z^n}{(1-n)\cdot n!}D^n(u^{n-1})$$

is the series inverse for (19), i.e., $\eta(u)=1/\zeta(u)$, and hence we have the series identity

$$1=\left(\sum_{n=0}^{\infty}\frac{z^n}{(n+1)!}D^n(u^{n+1})\right)\cdot\left(\frac{1}{u}-z\frac{u'}{u}+\sum_{n=2}^{\infty}\frac{z^n}{(1-n)\cdot n!}D^n(u^{n-1})\right).$$

Rearranging the terms of degree n in z in this formula results in the derivative identity

$$(23)\quad D^n(u^{n+1})=(n+1)u'D^{n-1}(u^n)+\sum_{i=2}^{n}\frac{1}{i-1}\binom{n+1}{i}uD^i(u^{i-1})D^{n-i}(u^{n-i+1}),$$

valid for $n\geqq 1$. The identities (22), (23) appear to be related to, but interestingly not the same as, some derivative identities appearing in Adams and Hippisley [1, § 7.37, p. 160]. Many more examples can be deduced by choosing other types of elementary functions for $\Phi(u)$ in (14).

5. Binomial and orthogonal polynomial identities. We now specialize the above derivative identities for particular functions $u(t)$, and find that they reduce to a wide range of identities among binomial coefficients and orthogonal polynomials. Some are known, but the orthogonal polynomial identities are apparently new. (However, I have not attempted a completely exhaustive search of the literature.)

1. First consider the case

$$u(t)=t^{\alpha}, \quad \text{so } \frac{1}{n!}D^n u^m=\binom{m\alpha}{n}t^{m\alpha-n}.$$

Then (22) reduces to the identity

$$\frac{k+l}{n+k+l}\binom{(n+k+l)\alpha}{n}=\sum_{i=0}^{n}\frac{kl}{(i+k)(n-i+l)}\binom{(i+k)\alpha}{i}\binom{(n-i+l)\alpha}{n-i}. \tag{24}$$

This is equivalent to the Hagen-Rothe identity [5], [6, Eqn. 3.142], which generalizes the classical Vandermonde convolution identity for binomial coefficients,

$$\binom{r+s}{n}=\sum_{i=0}^{n}\binom{r}{i}\binom{s}{n-i},$$

which follows from (24) in the limit $\alpha\to 0$, $k\alpha\to r$, $l\alpha\to s$. In (24), we use the definition

$$\binom{\beta}{n}=\frac{\beta(\beta-1)\cdots(\beta-n+1)}{n!}$$

for the general binomial coefficients, so that

$$\frac{1}{\beta}\binom{\beta}{n}=\frac{(\beta-1)\cdots(\beta-n+1)}{n!}$$

is well defined even for $\beta=0$. As another example, the formula (23) in this case reduces to the identity

$$\frac{1}{n+1}\binom{(n+1)\alpha}{n}=\alpha\binom{n\alpha}{n-1}+\sum_{i=2}^{n}\frac{1}{(i-1)(n-i+1)}\binom{(i-1)\alpha}{i}\binom{(n-i+1)\alpha}{n-i},$$

which is similar to the Van der Corput identity; cf. [6, Eqn. 3.147].

2. Let

$$u=e^{\alpha t}, \quad \text{so } D^n u^m=m^n\alpha^n e^{\alpha t}.$$

Then (22) reduces to the identity

$$(n+k+l)^{n-1}=\sum_{i=0}^{n}\frac{kl}{k+l}\binom{n}{i}(i+k)^{i-1}(n-i+l)^{n-i-1}. \tag{25}$$

If we set $k=-x/z$, $l=-n-(y/z)$, we deduce

$$(x+y)^{n-1}=\sum_{i=0}^{n}\frac{x(y+nz)}{x+y+nz}\binom{n}{i}(x-iz)^{i-1}(y+iz)^{n-i-1}, \tag{26}$$

which is very similar to the Abel identity [3, p. 128],

$$(x+y)^n=\sum_{i=0}^{n}\binom{n}{i}x(x-iz)^{i-1}(y+iz)^{n-i}. \tag{27}$$

Indeed, they are essentially equivalent identities, since if we denote (27) by $A_n(x, y, z)$, and (26) by $B_{n-1}(x, y, z)$, then we easily verify the relation

$$A_n(x, y, z) + nzA_{n-1}(x, y, z) = B_n(x, y, z).$$

Consequently, we can use Theorem 4 to give yet another proof of the Abel identity.

3. Let

$$u = e^{-t^2}, \quad \text{so } D^n u^m = (-1)^n m^{n/2} H_n(\sqrt{m}\, t)\, e^{-mt^2},$$

where H_n denotes the usual Hermite polynomial [2, § 10.13]. In this case, (22) reduces to the identity

$$\begin{aligned}&(n+k+l)^{(n/2)-1}H_n(\sqrt{n+k+l}\, t)\\ &\quad = \sum_{i=0}^{n} \frac{kl}{k+l}\binom{n}{i}(i+k)^{(i/2)-1}(n-i+l)^{(n-i)/2-1}H_i(\sqrt{i+k}\, t)H_{n-i}(\sqrt{n-i+l}\, t),\end{aligned} \tag{28}$$

which we can interpret as an Abel-type identity for Hermite polynomials. It is not the same as the usual addition formula, since the arguments of the Hermite polynomials appearing in the summation depend on the summation index i. If either $n+k+l=0$, $i+k=0$, or $n-i+l=0$, then we view the corresponding term in (28) according to the rule

$$\lim_{m\to 0}(-1)^n m^{(n/2)-1}H_n(\sqrt{m}\, t) = \begin{cases} -2t, & n=1,\\ -2, & n=2,\\ 0, & n \geqq 3,\end{cases}$$

stemming from the rule (21).

4. Let

$$u = t^\alpha e^{-t}, \quad \text{so} \quad \frac{1}{n!}D^n u^m = t^{m\alpha-n}\, e^{-mt} L_n^{m\alpha-n}(mt),$$

where L_n^α are the generalized Laguerre polynomials [2, § 10.12]. Again (22) reduces to an Abel-type identity

$$\begin{aligned}&\frac{k+l}{n+k+l}L_n^{(n+k+l)\alpha-n}((n+k+l)t)\\ &\quad = \sum_{i=0}^{n}\frac{kl}{(i+k)(n-i+l)}L_i^{(i+k)\alpha-i}((i+k)t)L_{n-i}^{(n-i+l)\alpha-n+i}((n-i+l)t)\end{aligned} \tag{29}$$

for Laguerre polynomials. As in the previous example, we make the convention

$$\lim_{m\to 0}\frac{1}{m}L_n^{m\alpha-n}(mt) = \begin{cases}\alpha - t, & n=1,\\ [(-1)^{n-1}/n]\alpha, & n \geqq 2,\end{cases}$$

stemming from the rule (21), for any exceptional terms in (29).

5. Finally, consider the case

$$u = (1-t)^\alpha(1+t)^\beta.$$

Then

$$\frac{1}{n!}D^n u^m=(-2)^n(1-t)^{m\alpha-n}(1+t)^{m\beta-n}P_n^{(m\alpha-n,m\beta-n)}(t),$$

where $P_n^{(\alpha,\beta)}$ are the Jacobi polynomials [2, § 10.8]. In this case (22) reduces to the Hagen–Rothe type formula

$$\begin{aligned}&\frac{k+l}{n+k+l}P_n^{((n+k+l)\alpha-n,(n+k+l)\beta-n)}(t)\\ &=\sum_{i=0}^{n}\frac{kl}{(i+k)(n-i+l)}P_i^{((i+k)\alpha-i,(i+k)\beta-i)}(t)P_{n-i}^{((n-i+l)\alpha-n+i,(n-i+l)\beta-n+i)}(t),\end{aligned}\tag{30}$$

which again does not appear in the standard literature on Jacobi polynomials. Again, we need a rule

$$\lim_{m\to 0}\frac{1}{m}P_n^{(m\alpha-n,m\beta-n)}(t)=\frac{[\alpha(t+1)^n-\beta(t-1)^n]}{[n(-2)^n]},\qquad n\geqq 1,$$

for any exceptional terms in (30).

6. Bell polynomial identities. The Bell polynomials arise in the formula for the nth derivative of the composition of two functions [3], [10]. Specifically, we have

$$\frac{d^n}{dt^n}f\circ g=\sum_{i=1}^{n}(f^{(i)}\circ g)\cdot B_i^n(g),\tag{31}$$

where the B_i^n are polynomials in the derivatives $g^{(k)}$ of g. Thus, the above derivative identities can be rewritten as identities involving Bell polynomials. Surprisingly, these identities have not appeared in the literature.

First, according to (31) (see [10, Ex. 22, p. 46]),

$$\frac{d^n}{dt^n}u^m=\sum_{i=1}^{\min\{m,n\}}\frac{m!}{(m-i)!}u^{m-i}B_i^n(u).\tag{32}$$

Therefore, (19) can be rewritten as

$$\zeta(u)=u+\sum_{n=1}^{\infty}\sum_{i=1}^{n}\frac{z^n u^{n+1-i}}{(n+1-i)!}B_i^n(u).\tag{33}$$

Furthermore, according to (14), (19), for k a positive integer,

$$\zeta(u)^k=u^k+\sum_{n=1}^{\infty}\sum_{i=1}^{n}\frac{k(n+k-1)!}{n!(n+k-i)!}z^n u^{n+k-i}B_i^n(u).\tag{34}$$

For example, if $k=2$, then

$$\zeta(u)^2=u^2+\sum_{n=1}^{\infty}\sum_{i=1}^{n}\frac{2(n+1)}{(n+2-i)!}z^n u^{n+k-i}B_i^n(u).$$

If we compare this with the square of the series (33), we deduce that

$$\frac{2(n+1)}{(n+2-i)!}B_i^n(u)=\frac{2}{(n+1-i)!}B_i^n(u)+\sum_{\substack{p+q=n\\ r+s=i}}\frac{1}{(p+1-r)!(q+1-s)!}B_r^p(u)B_s^q(u),$$

which is equivalent to the identity

$$2(i-1)B_i^n(u)=\sum_{p,r}\binom{n+2-i}{p+1-r}B_r^p(u)B_{i-r}^{n-p}(u).\tag{35}$$

By way of contrast, consider the Bell polynomial identity coming from squaring the standard exponential series

$$e^{zu}=1+\sum_{n=1}^{\infty}\sum_{i=1}^{n}\frac{z^i t^n}{n!}B_i^n(u). \tag{36}$$

Equating $e^{2zu}=(e^{zu})^2$ and rearranging terms, we deduce

$$2(2^{i-1}-1)B_i^n(u)=\sum_{p,r}\binom{n}{p}B_r^p(u)B_{i-r}^{n-p}(u). \tag{37}$$

which is quite similar to (23), but, except in very special cases, a different identity. More generally, the equation $e^{(a+b)zu}=e^{azu}\,e^{bzu}$ leads to the further standard identities

$$((a+b)^i-a^i-b^i)B_i^n(u)=\sum_{p,r}\binom{n}{p}a^r b^{i-r}B_r^p(u)B_{i-r}^{n-p}(u), \tag{38}$$

valid for any a, b. On the other hand, the identity $\zeta(u)^{a+b}=\zeta(u)^a\zeta(u)^b$ leads to yet more complicated identities

$$\left[\frac{(a+b)(n+a+b-1)}{ab}-\frac{(n+a-1)\binom{n+a+b-i}{b}}{b\binom{n+a+b-i}{b}}-\frac{(n+b-1)\binom{n+a+b-i}{a}}{a\binom{n+a+b-i}{a}}\right]B_i^n(u)$$
$$=\sum_{p,r}\frac{\binom{n}{p}\binom{n+a+b-i}{p+a-r}}{\binom{n+a+b-2}{p+a-1}}B_r^p(u)B_{i-r}^{n-p}(u). \tag{39}$$

More identities can be constructed by using different functions $\Phi(u)$ in the fundamental theorem. The number of different identities satisfied by the Bell polynomials is remarkable!

Appendix: Alternative proof of the main theorem. An alternative proof of Theorem 4 is based on the properties of certain first-order partial differential operators or vector fields; cf. [8]. As such, this proof may be adapted to give an alternative proof of the Lagrange inversion formula. The mathematical methods have not been used in the subject before, and thus are of some interest, possibly being of use in other problems. Let

$$u=\sum_{n=0}^{\infty}u_n t^n$$

be a formal power series. We will also work with the associated formal series

$$v=\frac{1}{u}=\sum_{n=0}^{\infty}v_n t^n$$

and

$$w=\log\frac{u}{u_0}=\sum_{n=1}^{\infty}w_n t^n.$$

We will regard the coefficients of u and v, i.e., $u_0, u_1, u_2, \cdots$, and $v_0, v_1, v_2, \cdots$, as providing different local coordinates on the space of formal power series. They are connected by formulae of the form

$$u_n = R_n(v_0, v_1, \cdots, v_n), \qquad v_n = R_n(u_0, u_1, \cdots, u_n),$$

where the rational functions R_n can be explicitly expressed using determinants; cf. [10, Ex. 20, p. 45]. We can also use $u_0, w_1, w_2, \cdots$, as yet another set of coordinates, connected by formulae of the form

$$\text{(A1)} \qquad u_n = u_0 B_n(w_1, w_2, \cdots, w_n),$$

where B_n is a Bell polynomial coming from the relation $u = u_0\, e^w$. We now write out the basic series

$$\text{(A2)} \qquad \Psi(u) = D^{-1} :e^{zDu}: D\Phi(u) = \sum_{n=0}^{\infty} z^n \Psi_n,$$

where, according to (13),

$$\text{(A3)} \qquad \Psi_n = \frac{1}{n!} D^n \Xi_n(u)\bigg|_{t=0}, \qquad \Xi_n(u) = \int_0^u \tilde{u}^n \Phi'(\tilde{u})\, d\tilde{u}.$$

Note that we are using the identification between series coefficients and derivatives given in (8) in this formula. In particular, the coefficient Ψ_n depends on the first n coefficients $u_0, u_1, \cdots, u_n$ of u. We can also re-express Ψ_n in terms of the coefficients $v_0, v_1, \cdots, v_n$ of $v = 1/u$, or, alternatively, in terms of $u_0, w_1, w_2, \cdots, w_n$ using (A1). It will be very convenient to permit such changes of coordinates during the course of the proof.

The elementary first-order partial differential operators

$$\mathbf{u}_j = \frac{\partial}{\partial u_j}, \quad \mathbf{v}_j = \frac{\partial}{\partial v_j}, \quad \mathbf{w}_j = \frac{\partial}{\partial w_j},$$

will be regarded as vector fields acting on the functions of the coefficients of the formal power series u, and its associated power series $v = 1/u$, $w = \log(u/u_0)$. As such, they can all be re-expressed in any of our three coordinate systems. Note that

$$\text{(A4)} \qquad \mathbf{v}_j(u) = -\mathbf{v}_j\left(\frac{1}{v}\right) = -\frac{t^j}{v^2} = -t^j u^2, \qquad j = 0, 1, 2, \cdots,$$

$$\text{(A5)} \qquad \mathbf{w}_j(u) = \mathbf{w}_j(u_0\, e^w) = t^j u_0\, e^w = t^j u, \qquad j = 1, 2, 3, \cdots.$$

Using these and similar formulae, it is not too difficult to verify the following change of variables formulae for these vector fields:

$$\text{(A6)} \qquad \mathbf{v}_j = -\sum_{m,n=0}^{\infty} u_m u_n \frac{\partial}{\partial u_{m+n+j}} = -\sum_{n=0}^{\infty} u_n \frac{\partial}{\partial w_{n+j}}, \qquad j = 0, 1, 2, \cdots.$$

(In the second summation, we use (A1) to re-express the u's in terms of the w's.) Also

$$\text{(A7)} \qquad \mathbf{w}_j = \sum_{n=0}^{\infty} u_n \frac{\partial}{\partial u_{n+j}} = -\sum_{n=0}^{\infty} v_n \frac{\partial}{\partial v_{n+j}}, \qquad j = 1, 2, 3, \cdots.$$

Finally, we define the vector fields

$$\text{(A8)} \qquad \mathbf{y}_j = \mathbf{w}_j + z\mathbf{v}_{j-1}, \qquad j = 1, 2, 3, \cdots,$$

where z is a scalar parameter.

LEMMA A1. *The vector fields* $\mathbf{y}_j$ *all mutually commute*:

$$[\mathbf{y}_j, \mathbf{y}_k] = 0 \quad \textit{for all } j, k = 1, 2, 3, \cdots . \tag{A9}$$

The proof is simplest in the v coordinates. We just re-express $\mathbf{w}_j$ using the second formula in (A7), and do a simple direct computation.

LEMMA A2. *Let* $\Psi(u) = D^{-1} :e^{zDu}: D\Phi(u)$. *Then*

$$\mathbf{y}_j[\Psi(u)] = 0 \quad \textit{for all } j = 1, 2, 3, \cdots . \tag{A10}$$

Proof. Note that since we are now working with formal power series, the vector fields $\mathbf{v}_j, \mathbf{w}_j$ commute with the derivative operator $D = d/dt$. Using (A3), (A4), we compute

$$\begin{aligned}\mathbf{v}_j[\Psi(u)] &= \sum_{n=0}^{\infty} \frac{z^n}{n!} D^n \left\{ \frac{\partial \Xi_n(u)}{\partial u} \mathbf{v}_j(u) \right\} \bigg|_{t=0} \\ &= -\sum_{n=0}^{\infty} \frac{z^n}{n!} D^n \{ t^j u^{n+2} \Phi'(u) \} |_{t=0},\end{aligned}$$

whereas, using (A3), (A5), we find

$$\begin{aligned}\mathbf{w}_j[\Psi(u)] &= \sum_{n=0}^{\infty} \frac{z^n}{n!} D^n \left\{ \frac{\partial \Xi_n(u)}{\partial u} \mathbf{w}_j(u) \right\} \bigg|_{t=0} \\ &= \sum_{n=0}^{\infty} \frac{z^n}{n!} D^n \{ t^j u^{n+1} \Phi'(u) \} |_{t=0}.\end{aligned}$$

Note that

$$[D^n, t] = D^n \cdot t - t \cdot D^n = nD^{n-1};$$

hence, upon evaluation at $t = 0$,

$$D^n \cdot t|_{t=0} = nD^{n-1}|_{t=0}$$

Substituting this into the previous sum, we find (since $j \geqq 1$)

$$\begin{aligned}\mathbf{w}_j[\Psi(u)] &= \sum_{n=1}^{\infty} \frac{z^n}{(n-1)!} D^{n-1} \{ t^{j-1} u^{n+1} \Phi'(u) \} |_{t=0} \\ &= \sum_{n=0}^{\infty} \frac{z^{n+1}}{n!} D^n \{ t^{j-1} u^{n+2} \Phi'(u) \} |_{t=0}.\end{aligned}$$

Comparing with the previous summation, we deduce that

$$\mathbf{w}_j[\Psi(u)] = -z\mathbf{v}_{j-1}[\Psi(u)], \qquad j = 1, 2, 3, \cdots, \tag{A11}$$

which clearly implies the lemma.

If we write out the coefficient of z^n in the previous formula (A11), we find that

$$\mathbf{w}_j[\Psi_n(u)] = -\mathbf{v}_{j-1}[\Psi_{n-1}(u)], \qquad j, n = 1, 2, 3, \cdots . \tag{A12}$$

Since Ψ_n only depends on $u_0, u_1, \cdots, u_n$, only the first n of the equations in (A12) are nontrivial. We now regard (A12) as a system of first-order partial differential equations for the coefficients Ψ_n of the series (A2). The commutativity Lemma A1 will imply that the system is in involution in the sense of Frobenius [8], and hence can be uniquely solved using suitable initial data. In fact, since the w coordinates straighten out the vector fields $\mathbf{w}_j$, we can explicitly solve the system.

LEMMA A3. *Let Ψ_n, $n=0,1,2,\cdots$, be functions depending on the coefficients w_j of the formal series w which satisfy*

(i) $\Psi_n(u_0, w_1, w_2, \cdots, w_n)$ *depends only on the first n coefficients of w.*

(ii) $\Psi_n(u_0, 0, 0, \cdots, 0)=0$ *for $n>0$.*

(iii) $$\frac{\partial \Psi_n}{\partial w_j}=\mathbf{v}_{j-1}[\Psi_{n-1}], \qquad j=1,\cdots,n.$$

Then Ψ_n are uniquely determined by the function $\Psi_0(u_0)=\Phi(u_0)$.

Proof. According to Lemma A1, the integrability conditions for the elementary system of partial differential equations (iii) are satisfied. Therefore, the value of Ψ_n is uniquely determined by its noncharacteristic Cauchy data prescribed by condition (ii). This completes the proof.

Now, to complete the proof of Theorem 4, it suffices to notice that $\zeta(u)$, as defined by (A1), being a particular case of (A2), satisfies the three conditions of Lemma A3. But then the series $\Phi(\zeta(u))$ also satisfies them since, for example,

$$\mathbf{y}_j\Phi(\zeta(u))=\Phi'(\zeta(u))\mathbf{y}_j(\zeta(u))=0.$$

Also, the leading-order term of $\Phi(\zeta(u))$ is $\Phi(u_0)$, which agrees with that of $\Psi(u)$ as given by (A2). According to the uniqueness result in Lemma A3, the series must agree, i.e., $\Phi(\zeta(u))=\Psi(u)$. This completes the proof of the main theorem.

Acknowledgments. I would like to express my gratitude to two anonymous referees, one for observing how the main theorem easily follows from Lagrange inversion, and the other for encouraging me to investigate the consequences of the new derivative identities (22), (23) which are discussed in § 5.

REFERENCES

[1] E. P. ADAMS AND R. L. HIPPISLEY, *Smithsonian Mathematical Formulae and Tables of Elliptic Functions*, Smithsonian Miscellaneous Collections, Vol. 74, No. 1, Smithsonian Inst., Washington, D.C., 1922.

[2] H. BATEMAN, *Higher Transcendental Functions*, Vol. 2, A. Erdélyi, ed., McGraw-Hill, New York, 1953.

[3] L. COMTET, *Advanced Combinatorics: The Art of Finite and Infinite Expansions*, D. Reidel, Boston, MA, 1974.

[4] J. GLIMM AND A. JAFFE, *Quantum Physics*, Second Ed., Springer-Verlag, New York, 1987.

[5] H. W. GOULD, *Some generalizations of Vandermonde's convolution*, Amer. Math. Monthly, 63 (1956), pp. 84–91.

[6] ———, *Combinatorial Identities*, Morgantown, WV, 1972.

[7] Z. A. MELZAK, *Companion to Concrete Mathematics*, Wiley-Interscience, New York, 1973.

[8] P. J. OLVER, *Applications of Lie Groups to Differential Equations*, Graduate Texts in Math. 107, Springer-Verlag, New York, 1986.

[9] ———, *Canonical forms and integrability of biHamiltonian systems*, Phys. Lett., 148A (1990), pp. 177–187.

[10] J. RIORDAN, *An Introduction to Combinatorial Analysis*, Princeton University Press, Princeton, NJ, 1980.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 222–243, January 1992

012

AN ALGEBRAIC THEORY OF WAVELETS. I. OPERATIONAL CALCULUS AND COMPLEX STRUCTURE*

GERALD KAISER†

Abstract. In wavelet analysis, a function f is split into two parts at each iteration. The first part, Hf, represents a smoothed version of f, sampled half as frequently, while the second part, Gf, represents the detail lost by filtering through H. Although the operators H and G have very different interpretations, they exhibit a remarkable symmetry in their algebraic properties. We examine this symmetry by developing an effective, basis-independent operational calculus for wavelets and use it to show that the symmetry is due to the existence of a *complex structure*, i.e., a map J such that $J^2 = -I$ where I is the identity. This implies that the space V_α of (real) functions at the scale 2^α ($\alpha \in \mathbb{Z}$) may be regarded as a *complexification* $V_{\alpha+1}^c$ of the space $V_{\alpha+1}$ of functions at the next (coarser) scale. Roughly, the low-frequency parts Hf of the functions span the real part of $V_{\alpha+1}^c$ while their high-frequency parts Gf span the imaginary part. The map J mediates between the two and relates the corresponding operators H and G. Furthermore, at the scale $\alpha = -1$, J transforms the fundamental function ϕ associated with H into the "fundamental wavelet" ψ associated with G.

Key words. wavelets, multiscale analysis, complex structure, operational calculus

AMS(MOS) subject classifications. 16, 20, 41, 42

1. Operational calculus. In wavelet analysis (see Daubechies [1988], Mallat [1989], Meyer [1990], Strang [1990], and the references therein), one deals with the representation of a function ("signal") at different scales. We begin with a single real-valued function ϕ of one real variable which we take, for simplicity, to be continuous with compact support. It is assumed that for some $T > 0$, the translates $\phi_n(t) \equiv \phi(t - nT)$, $n \in \mathbb{Z}$, form an orthonormal set in $L^2(\mathbb{R})$ (such functions can be easily constructed). The closure of the span of the vectors ϕ_n in $L^2(\mathbb{R})$ forms a subspace V which can be identified with $\ell^2(\mathbb{Z})$, since for a real sequence $u \equiv \{u_n\}$ we have

$$\left\| \sum_n u_n \phi_n \right\|^2 = \sum_n u_n^2. \tag{1}$$

We introduce the shift operator

$$(Sf)(t) \equiv f(t - T), \tag{2}$$

which leaves V invariant and is an orthogonal operator on $L^2(\mathbb{R})$ (we shall be dealing with *real* spaces, unless otherwise stated). A general element of V can be written uniquely as

$$\sum_n u_n S^n \phi \equiv u(S)\phi, \tag{3}$$

*Received by the editors May 7, 1990; accepted for publication (in revised form) January 28, 1991.

†Mathematics Department, University of Massachusetts at Lowell, Lowell, Massachusetts 01854.

where $u(e^{i\xi T})$ is the square-integrable function on the unit circle ($|\xi| \leq \pi/T$) having $\{u_n\}$ as its Fourier coefficients and $u(S)$ is defined as an operator on "nice" functions (e.g., Schwartz test functions) $f(t)$ through the Fourier transform, i.e.,

$$[u(S)f]\,\hat{}(\xi) \equiv u(e^{i\xi T})\hat{f}(\xi). \tag{4}$$

For the purpose of developing our operational calculus, we shall consider operators $u(S)$ which are *polynomials* in S and S^{-1}. These form an abelian algebra $\mathcal{P}$ of operators on V. Moreover, it will suffice to restrict our attention to the dense subspace of finite combinations in V, i.e., to $\mathcal{P}\phi$, since our goal here is to produce an L^2 theory and this can be achieved by developing the algebraic (finite) theory and then completing in the L^2 norm. Note that the independence of the vectors ϕ_n means that $u(S)\phi = 0$ implies $u(S) = 0$. Our results could actually be extended to operators $u(S)$ with $\{u_n\} \in \ell^1(\mathbb{Z}) \subset \ell^2(\mathbb{Z})$, which also form an algebra since the product $u(S^{-1})w(S)$ corresponds to the convolution of the sequences $\{u_n\}$ and $\{w_n\}$. We resist the temptation.

Let us stop for a moment to discuss the "signal-processing" interpretation of $u(S)\phi$, since that is one of the motivations behind wavelet theory. It is natural to think of $u(S)\phi$ as an approximation to a function ("signal") $f(t)$ obtained by sampling f only at $t_n = nT$. Let f_0 denote the *band-limited* function obtained from f by cutting off all frequencies ξ with $|\xi| > \pi/T$. That is, $\hat{f}_0$ coincides with $\hat{f}$ for $|\xi| \leq \pi/T$ but vanishes outside this interval. The value of f_0 at t_n is then

$$f_0(nT) = \frac{1}{2\pi}\int_{-\pi/T}^{\pi/T} d\xi e^{-i\xi nT}\hat{f}(\xi), \tag{5}$$

which is just the Fourier coefficient of the periodic function

$$\hat{F}_0(\xi) \equiv \sum_n T f_0(nT)e^{i\xi nT} \tag{6}$$

obtained from $\hat{f}_0(\xi)$ by identifying $\xi + 2\pi/T$ with ξ. In the time domain,

$$F_0(t) = \sum_n T f_0(nT)\,\delta(t - nT). \tag{7}$$

This has the same form as $u(S)\phi$, if we set $u_n = Tf_0(nT)$ and $\phi(t) = \delta(t)$ where δ is the Dirac distribution. Hence the usual sampling theory may be regarded as the singular case $\phi = \delta$, and then $u(S)\phi$ characterizes the band-limited approximation f_0 of f. For square-integrable ϕ, the samples u_n are no longer the values at the sharp times t_n but are *smeared* over ϕ_n, since $u_n = \langle \phi_n, u(S)\phi \rangle$. In fact, ϕ acts as a *filter,* i.e., as a convolution operator, since $(u(S)\phi)\hat{}(\xi) = u(e^{i\xi T})\hat{\phi}(\xi)$. Roughly speaking, we may think of ϕ as giving the shape of a *pixel.*

Next, a scaled family of spaces $V_\alpha, \alpha \in \mathbb{Z}$, is constructed from V as follows. The dilation operator D, defined by

$$(Df)(t) = 2^{-1/2}f(t/2), \tag{8}$$

is orthogonal on $L^2(\mathbb{R})$. It stretches a function by a factor of 2 without altering its norm and is related to S by the commutation rules

$$(9) \qquad DS = S^2 D, \qquad D^{-1}S^2 = SD^{-1}.$$

Hence D "squares" S while D^{-1} takes its "square root." A repeated application of the above gives

$$(10) \qquad D^\alpha S = S^{2^\alpha} D^\alpha, \qquad \alpha \in \mathbb{Z}.$$

Define the spaces

$$(11) \qquad V_\alpha = D^\alpha V,$$

which are closed in $L^2(\mathbb{R})$ ($V_0 \equiv V$). An orthonormal basis for V_α is given by

$$(12) \qquad \phi_n^\alpha(t) \equiv D^\alpha S^n \phi(t) = 2^{-\alpha/2}\phi\left(2^{-\alpha}t - nT\right),$$

and V_α can also be identified with $\ell^2(\mathbb{Z})$. The motivation is that V_α will consist of functions containing detail only up to the scale of 2^α, which correspond to sequences $\{u_n^\alpha\}$ in $\ell^2(\mathbb{Z})$ representing samples at $t_n = 2^\alpha nT$. For this to work, we must have $V_{\alpha+1} \subset V_\alpha$ for all α. A *necessary* condition for this is that ϕ must satisfy a functional equation (taking $\alpha = -1$) of the form

$$(13) \qquad \phi = \sum_n h_n \phi_n^{-1} = D^{-1}\sum_n h_n S^n \phi \equiv D^{-1}h(S)\phi$$

for some (unique) set of coefficients h_n. Since we assume that ϕ has compact support, it follows that all but a finite number of the coefficients h_n vanish, so $h(S)$ is a polynomial in S and S^{-1}, i.e., $h(S) \in \mathcal{P}$. This operator *averages,* while D^{-1} *compresses.* Hence ϕ is a fixed point of this dual action of spreading and compression. $D\phi = h(S)\phi$, called a *dilation equation,* states that the dilated pixel $D\phi$ is a linear combination of undilated pixels ϕ_n. An integration with respect to t leads to

$$(14) \qquad \sum_n h_n = \sqrt{2} \quad \text{or} \quad h(I) = \sqrt{2}I,$$

giving a constraint on the coefficiants h_n (other constraints will emerge). Note that the singular case $\phi = \delta$ associated with sharp sampling satisfies the dilation equation with $h(S) = \sqrt{2}I$, where I denotes the identity operator on V. Since δ is not square-integrable, this solution does not fit into the L^2 theory considered here.

On the other hand, the coefficients h_n uniquely determine ϕ, up to a sign. For if we iterate $D^{-1}h(S) = h(S^{1/2})D^{-1}$, we obtain

$$(15) \qquad \phi = \left[D^{-1}h(S)\right]^N \phi = \prod_{\alpha=1}^{N} h\left(S^{2^{-\alpha}}\right) D^{-N}\phi.$$

Since the Fourier transform of $D^{-N}\phi$ satisfies

$$2^{N/2}(D^{-N}\phi)^{\wedge}(\xi) = \hat{\phi}(2^{-N}\xi) \to \hat{\phi}(0) \text{ as } N \to \infty, \tag{16}$$

we formally obtain

$$\phi = \hat{\phi}(0) \lim_{N\to\infty} \prod_{\alpha=1}^{N} \left[2^{-1/2} h\left(S^{2^{-\alpha}}\right)\right] \delta. \tag{17}$$

The normalization is determined up to a sign by $\|\phi\| = 1$. In general, the function ϕ determined by $h(S)$ is highly irregular. Daubechies [1988] has classified all $h(S) \in \mathcal{P}$ that give functions ϕ possessing some regularity. The simplest (and least regular) of these is related to the classical Haar basis. It will be used throughout the paper to illustrate the various operators as they are introduced. For the general case, the actions of these operators on bases are given in the appendix.

Example 1 (The Haar system). Let ϕ be the indicator function $\chi_{[0,1)}$ for the interval $[0, 1)$. Then

$$\begin{aligned} D\phi &= \frac{1}{\sqrt{2}}\chi_{[0,2)} = \frac{1}{\sqrt{2}}\left(\chi_{[0,1)} + \chi_{[1,2)}\right) \\ &= \frac{1}{\sqrt{2}}(I+S)\,\phi; \end{aligned} \tag{18}$$

hence ϕ satisfies the dilation equation with $h(S) = (I + S)/\sqrt{2}$.

The next step is to introduce a "multiscale analysis" based on the sequence of spaces V_α. We shall do this in a basis-independent fashion. Since shifts and dilations are related by $DS = S^2 D$, we have

$$D^{\alpha+1}u(S)\phi = D^{\alpha}u(S^2)D\phi = D^{\alpha}u(S^2)h(S)\phi. \tag{19}$$

This defines a map H_α^*: $V_{\alpha+1} \to V_\alpha$, given by

$$H_\alpha^* D^{\alpha+1}u(S)\phi = D^{\alpha}h(S)\,u(S^2)\phi. \tag{20}$$

Since the two sides of this equation are actually *identical* as functions or elements of $L^2(\mathbb{R})$, H_α^* is simply the *inclusion map* which establishes $V_{\alpha+1} \subset V_\alpha$. This shows that the relation $D\phi = h(S)\phi$ is not only necessary but also *sufficient* for $V_{\alpha+1} \subset V_\alpha$. Although a vector in $V_{\alpha+1}$ is identical with its image under H_α^* as an element of $L^2(\mathbb{R})$, it is useful to distinguish between them since this permits us to use operator theory to define other useful maps, such as the adjoint H_α: $V_\alpha \to V_{\alpha+1}$ of H_α^*. Since the norm on V_α is that of $L^2(\mathbb{R})$ and H_α^* is an inclusion, it follows that $H_\alpha H_\alpha^* = I_{\alpha+1}$, the identity on $V_{\alpha+1}$. In particular, H_α is *onto*; it is just the orthogonal projection from V_α to $V_{\alpha+1}$. H_α^* is interpreted as an operator that *interpolates* a vector in $V_{\alpha+1}$, representing it as the vector in V_α obtained by replacing the "pixel" ϕ with the linear combination of compressed pixels $D^{-1}h(S)\phi$. The adjoint H_α is sometimes called a "low-pass filter" because it *smooths out* the signal and resamples it at half the sampling rate, thus cutting the freqency range in half. However, it is *not* a filter in the traditional sense since it is not a convolution operator, as will be seen below. The kernel of H_α is denoted by $W_{\alpha+1}$. It is the orthogonal complement of the image of H_α^*, i.e., of $V_{\alpha+1}$, in V_α:

$$W_{\alpha+1} \equiv \ker H_\alpha = V_\alpha \ominus H_\alpha^* V_{\alpha+1} = V_\alpha \ominus V_{\alpha+1}. \tag{21}$$

Note that H_α^* is "natural" with respect to the scale gradation, i.e.,

$$H_\alpha^* D^{\alpha+1} = D^\alpha H_0^* D. \tag{22}$$

Our "home space" will be V. All our operators will enjoy the above naturality with respect to scale. Because of this property, it will generally be sufficient to work in V. Define the operator H^*: $V \to V$ by

$$H^* = H_0^* D. \tag{23}$$

We will refer to H^* as the "home version" of H_α^*. Home versions of operators will generally be denoted without subscripts. Note that while H_α^* preserves the scale (it is an inclusion map!), H^* involves a change in scale. It consists of a dilation (which spreads the sample points apart to a distance $2T$) followed by an interpolation (which restores the sampling interval to its original value T). Thus H^* is a *zoom-in operator*! Its adjoint

$$H = D^{-1} H_0 \tag{24}$$

consists of a "filtration" by H_0 (which cuts the density of sample points by a factor of 2 without changing the scale) followed by a compression (which restores it to its previous value). H is, therefore, a *zoom-out operator.* It is related to H_α by

$$H_\alpha D^\alpha = D^{\alpha+1} H. \tag{25}$$

The filtration performed by H (which will be detailed below) represents the (possible) loss of information due to zooming out.

The operators H and H^* are essentially identical with those used by Daubechies, except for the fact that hers act on the sequences $\{u_n\}$ rather than the functions $u(S)\phi$. They are especially useful when considering iterated decomposition and reconstruction algorithms (§3).

To find the action of H_α, it suffices to find the action of H. Note that $H^* u(S)\phi = h(S)u(S^2)\phi$, where $u(S^2)$ is *even* in S. This will be an important observation in what follows, hence we first study the decomposition of V into its even and odd subspaces.

An arbitary polynomial $u(S)$ in S, S^{-1} can be written uniquely as the sum of its even and odd parts,

$$\begin{aligned} u(S) &= \sum_n u_{2n} S^{2n} + \sum_n u_{2n+1} S^{2n+1} \\ &\equiv u_+(S^2) + S u_-(S^2). \end{aligned} \tag{26}$$

Define the operator E^* (for *even*) on V by

$$E^* S = S^2 E^*, \qquad E^* \phi = \phi. \tag{27}$$

Then

$$E^*u(S)\phi = u(S^2)\phi = \sum_n u_n\phi_{2n}. \tag{28}$$

Also define the operator O^* (for *odd*) by $O^* = SE^*$, so that

$$O^*u(S)\phi = Su(S^2)\phi = \sum_n u_n\phi_{2n+1}. \tag{29}$$

H^* is related to E^* by $H^* = h(S)E^*$. Hence to obtain H it suffices to find the adjoint E of E^*.

LEMMA 1. *Let $v(S) \in \mathcal{P}$ and denote the adjoints of E^* and O^* by E and O. Then*

(a)

$$\begin{aligned} EE^* &= OO^* = I, \\ OE^* &= EO^* = 0, \end{aligned} \tag{30}$$

(b)

$$\begin{aligned} Ev(S)\phi &= v_+(S)\phi = \sum_n v_{2n}\phi_n \\ Ov(S)\phi &= v_-(S)\phi = \sum_n v_{2n+1}\phi_n, \end{aligned} \tag{31}$$

(c)

$$\begin{aligned} Ev(S)E^* &= v_+(S) = \tfrac{1}{2}D^{-1}\left[v(S) + v(-S)\right] D, \\ Ov(S)O^* &= v_+(S), \\ Ov(S)E^* &= v_-(S) = \tfrac{1}{2}D^{-1}S^{-1}\left[v(S) - v(-S)\right] D \\ Ev(S)O^* &= Sv_-(S) \end{aligned} \tag{32}$$

(*note that* (a) *is a special case with* $v(S) = I$), *and*

(d)

$$E^*E + O^*O = I. \tag{33}$$

Proof. For $u(S), v(S) \in \mathcal{P}$, we have

$$\begin{aligned} \langle EE^*u(S)\phi, v(S)\phi \rangle &= \langle E^*u(S)\phi, E^*v(S)\phi \rangle \\ &= \langle u(S^2)\phi, v(S^2)\phi \rangle = \langle u(S)\phi, v(S)\phi \rangle, \end{aligned} \tag{34}$$

where the last equality follows from the invariance of the inner product under $S \mapsto S^2$, i.e.,

$$\langle S^{2n}\phi, S^{2m}\phi \rangle = \delta_{nm} = \langle S^n\phi, S^m\phi \rangle. \tag{35}$$

Hence $EE^* = I$, so $OO^* = ES^{-1}SE^* = I$. $EO^* = OE^* = 0$ follows from the orthogonality of even and odd functions of S (applied to ϕ). This proves (a). To show (b), note that due to the orthogonality of even and odd functions,

$$\begin{aligned} \langle E^*u(S)\phi, v(S)\phi \rangle &= \langle u(S^2)\phi, v(S)\phi \rangle = \langle u(S^2)\phi, v_+(S^2)\phi \rangle \\ &= \langle E^*u(S)\phi, E^*v_+(S)\phi \rangle = \langle u(S)\phi, v_+(S)\phi \rangle, \end{aligned} \tag{36}$$

where we have used (a). This proves the first equation in (b). The second follows from $O = ES^{-1}$ and $S^{-1}v(S) = v_-(S^2) + S^{-1}v_+(S^2)$. To prove (c), note that $u(S^2)E^* = E^*u(S)$ and $Su(S^2)E^* = O^*u(S)$; hence

$$\begin{aligned} Ev(S)E^* &= E(v_+(S^2) + Sv_-(S^2))E^* \\ &= EE^*v_+(S) + EO^*v_-(S) = v_+(S), \\ Ov(S)O^* &= ES^{-1}v(S)SE^* = v_+(S), \\ Ov(S)E^* &= O(v_+(S^2) + Sv_-(S^2))E^* \\ &= OE^*v_+(S) + EE^*v_-(S) = v_-(S), \\ Ev(S)O^* &= E(v_+(S^2) + Sv_-(S^2))SE^* \\ &= EO^*v_+(S) + EE^*Sv_-(S) = Sv_-(S). \end{aligned} \tag{37}$$

Lastly, (d) follows from

$$\begin{aligned} E^*Ev(S)\phi + O^*Ov(S)\phi &= E^*v_+(S)\phi + O^*v_-(S)\phi \\ &= v_+(S^2)\phi + Sv_-(S^2)\phi = v(S)\phi. \end{aligned} \tag{38}$$ □

Remark. The algebraic structure above is characteristic of orthogonal decompositions and will be met again in our discussion of low- and high-frequency filters. E^*E and O^*O are the projection operators to the subspaces of even and odd functions of S (applied to ϕ),

$$V^e \equiv \overline{\{v(S^2)\phi \mid v(S) \in \mathcal{P}\}}, \qquad V^o \equiv \overline{\{Sv(S^2)\phi \mid v(S) \in \mathcal{P}\}}, \tag{39}$$

and

$$V = V^e \oplus V^o. \tag{40}$$

This decomposition will play an important role in the sequel.

PROPOSITION 2. *The maps* $H\colon V \to V$ *and* $H_\alpha\colon V_\alpha \to V_{\alpha+1}$ *are given by*

$$\begin{aligned} Hu(S)\phi &= Eh(S^{-1})u(S)\phi \\ &= [h_+(S^{-1})u_+(S) + h_-(S^{-1})u_-(S)]\,\phi, \\ H_\alpha D^\alpha u(S)\phi &= D^{\alpha+1}E\,h(S^{-1})u(S)\phi \\ &= D^{\alpha+1}\,[h_+(S^{-1})u_+(S) + h_-(S^{-1})u_-(S)]\,\phi. \end{aligned} \tag{41}$$

Proof. Since $H^* = h(S)\,E^*$, it follows that $H = E\,h(S^{-1})$ and $H_\alpha D^\alpha = D^{\alpha+1}H = D^{\alpha+1}Eh(S^{-1})$. □

Example 2 (The Haar system, continued). For $h(S) = (I + S)/\sqrt{2}$ and $\phi = \chi_{[0,1)}$, we have $h_+(S) = h_-(S) = 1/\sqrt{2}$. Hence

$$\begin{aligned} H^*\phi_n &= h(S)\,\phi_{2n} = \frac{1}{\sqrt{2}}\,(\phi_{2n} + \phi_{2n+1}) \\ H\phi_n &= \frac{1}{\sqrt{2}}E\,(I + S^{-1})\;S^n\phi = \frac{1}{\sqrt{2}}E\,(S^n + S^{n-1})\,\phi \\ &= \frac{1}{\sqrt{2}}\begin{cases} S^{n/2}\,\phi, & n \text{ even}, \\ S^{(n-1)/2}\,\phi, & n \text{ odd} \end{cases} \\ &= \frac{1}{\sqrt{2}}\,\phi_{[n/2]}. \end{aligned} \tag{42}$$

2. Complex structure. Up to this point, it could be argued, nothing extraordinary has happened. We have a filter which, when applied repeatedly, gives rise to a nested sequence of subspaces V_α. However, the next step is quite surprising and underlies much of the interest wavelets have generated. It is desirable to record the information lost at each stage of filtering, i.e., that part of the signal residing in the orthogonal complement $W_{\alpha+1}$ of $V_{\alpha+1}$ in V_α. The orthogonal decomposition $V_\alpha = V_{\alpha+1} \oplus W_{\alpha+1}$ is described by filters H_α and G_α, where H_α is as above and G_α extracts high-frequency information. For this reason, H_α and G_α obey a set of algebraic relations similar to those satisfied by E and O above. What is quite remarkable is that there exists a vector ψ in V_{-1} that is related to the spaces W_α and the maps G_α in a way almost totally symmetric to the way ϕ is related to V_α and H_α. This is *not* merely a consequence of the orthogonal decomposition but is somehow related to the fact that $V_{\alpha+1}$ is "half" of V_α, due to the doubling of the sampling interval upon dilation, as expressed by the commutation relation $DS = S^2D$. However, the precise reason for this symmetry has not been entirely clear. The usual constructions are somewhat involved and do not appear to shed much light on this question. It was this puzzle which motivated the present work. As an answer, we propose the following new construction. Begin by defining a *complex structure* on V, i.e., a map $J\colon V \to V$ such that $J^2 = -I$. (To illustrate this concept, consider the complex plane $\mathbb{C}$ as the real space $\mathbb{R}^2$. Then multiplication by the unit imaginary i is represented by a real 2×2 matrix whose square is $-I$.) J is defined by giving its commutation rule with respect to the shift and its action on ϕ:

$$JS = -S^{-1}J, \qquad J\phi = \varepsilon(S)\,\phi, \tag{43}$$

where $\varepsilon(S)$ is an as yet undetermined function. It follows that for $u(S) \in \mathcal{P}$,

$$Ju(S)\phi = \varepsilon(S)u(-S^{-1})\phi. \tag{44}$$

We further require that J preserve the inner product, i.e., that $J^*J = I$. Combined with $J^2 = -I$, this gives $J^* = -J$. That is, J will behave like multiplication by i also with respect to the inner product, giving it an interpretation as a *Hermitian* inner product.

In order to study J, we first define two simpler operators C and M as follows.

$$\begin{aligned} CS &= S^{-1}C, \qquad & C\phi &= \phi, \\ MS &= -SM, \qquad & M\phi &= \phi. \end{aligned} \tag{45}$$

Note that $CM = MC$ and that $C^* = C$ and $M^* = M$, since

$$\begin{aligned} \langle Cu(S)\phi, v(S)\phi \rangle &= \langle u(S^{-1})\phi, v(S)\phi \rangle = \langle v(S)\phi, u(S^{-1})\phi \rangle \\ &= \langle u(S)\phi, v(S^{-1})\phi \rangle = \langle u(S)\phi, Cv(S)\phi \rangle, \end{aligned} \tag{46}$$

where we have used the symmetry of the inner product and $u(S)^* = u(S^{-1})$ (both of which depend on the reality of V), and

$$\begin{aligned} \langle Mu(S)\phi, v(S)\phi \rangle &= \langle u(-S)\phi, v(S)\phi \rangle = \langle u(S)\phi, v(-S)\phi \rangle \\ &= \langle u(S)\phi, Mv(S)\phi \rangle, \end{aligned} \tag{47}$$

where the invariance of the inner product under $S \mapsto -S$ was used. ($C^* = C$ could have been proved more simply by noting that the inner product is invariant under

$S \mapsto S^{-1}$, which follows from the orthogonality of the ϕ_n's; however, this orthogonality does not appear to be a fundamental feature of the theory, so we avoid it whenever possible.) Since C and M are also involutions, i.e.,

$$C^2 = M^2 = I, \tag{48}$$

it follows that they are orthogonal operators. Hence they represent *symmetries,* which makes them important in themselves, especially in the abstract context where we begin with an algebra and constructs a representation (see the remark at the end of §3). In fact, the orthogonal decomposition $V = V^e \oplus V^o$ is nothing but the spectral decomposition associated with M, since V^e and V^o are the eigenspaces of M with eigenvalues 1 and -1, respectively. C has a simple interpretation as a *conjugation operator,* since for $u(S) \in \mathcal{P}$,

$$Cu(S)C = u(S^{-1}) = u(S)^*. \tag{49}$$

In terms of C and M,

$$J = \varepsilon(S)CM. \tag{40}$$

PROPOSITION 3. *The conditions $J^* = -J$ and $J^2 = -I$ hold if and only if $\varepsilon(S)$ satisfies*

$$\varepsilon(-S) = -\varepsilon(S), \qquad \varepsilon(S^{-1})\varepsilon(S) = 1. \tag{51}$$

Proof. We have

$$J^* = MC\varepsilon(S^{-1}) = M\varepsilon(S)C = \varepsilon(-S)MC = \varepsilon(-S)CM; \tag{52}$$

hence $J^* = -J$ if and only if $\varepsilon(-S) = -\varepsilon(S)$. Assume this to be the case. Then

$$J^2 = \varepsilon(S)CM\varepsilon(S)CM = \varepsilon(S)\varepsilon(-S^{-1}) = -\varepsilon(S)\varepsilon(S^{-1}); \tag{53}$$

hence $J^2 = -I$ if and only if $\varepsilon(S^{-1})\varepsilon(S) = I$. □

Remarks. (1) J is determined only up to the orthogonal mapping $\varepsilon(S)$. This corresponds to a similar freedom in the standard approach to wavelet theory, where a factor $e^{i\lambda(\xi)}$ in Fourier space relates the functions $H(\xi)$ and $G(\xi)$ associated with the operators H and G (Daubechies [1988, p. 943], where $T = 1$). The relation between $\varepsilon(S)$ and $\lambda(\xi)$ is given in the Appendix.

(2) The simplest examples of a complex structure are given by choosing

$$\varepsilon(S) = \pm S^{2p+1}, \qquad p \in \mathbb{Z}. \tag{54}$$

More interesting examples can be obtained by enlarging $\mathcal{P}$ to a topological algebra, for example, allowing $u(S)$ with $\{u_n\} \in \ell^1(\mathbb{Z})$.

(3) The above proof used the symmetry of the inner product. Later we shall complexify our spaces and the inner product becomes Hermitian. However, this proof easily extends to the complex case (when transposing, also take the complex conjugate). C then becomes $\mathbb{C}$-*antilinear* and is interpreted as Hermitian conjugation.

At an arbitrary scale α, define maps $J_\alpha: V_\alpha \to V_\alpha$ by naturality, i.e.,

$$J_\alpha D^\alpha = D^\alpha J, \tag{55}$$

which implies that $J_\alpha^2 = -I_\alpha$ and $J_\alpha^* = -J_\alpha$. J_α is related to S by

$$\begin{aligned} S^{2^\alpha} J_\alpha D^\alpha &= S^{2^\alpha} D^\alpha J = D^\alpha SJ = -D^\alpha JS^{-1} \\ &= -J_\alpha D^\alpha S^{-1} = -J_\alpha S^{-2^\alpha} D^\alpha, \end{aligned} \tag{56}$$

showing that

$$S^{2^\alpha} J_\alpha = -J_\alpha S^{-2^\alpha}. \tag{57}$$

In particular, note that $S^{1/2}J_{-1} = -J_{-1}S^{-1/2}$; hence

$$SJ_{-1} = +J_{-1}S^{-1}. \tag{58}$$

We are now in a position to construct the basic wavelet ψ, the spaces W_α and an appropriate set of high-frequency filters in a way that will make the symmetry with ϕ, V_α, and H_α quite clear. Consider the *restriction* of J_α to the subspace $V_{\alpha+1}$ of V_α, i.e., the map K_α^*: $V_{\alpha+1} \to V_\alpha$ defined by

$$K_\alpha^* = J_\alpha H_\alpha^*. \tag{59}$$

K_α^* is natural with respect to the scale gradation, and its home version will, as usual, be denoted by $K^* \equiv K_0^* D = JH^*$. It will turn out that its adjoint K is essentially equivalent to the usual filter G (to be introduced below) but is more natural from the point of view of the complex structure. Define the vector $\psi \in V_{-1}$ by

$$\psi = K_{-1}^*\phi = J_{-1}D^{-1}h(S)\phi = D^{-1}Jh(S)\phi \equiv D^{-1}g(S)\phi, \tag{60}$$

where the function

$$g(S) = \varepsilon(S)h(-S^{-1}) \tag{61}$$

will play a similar role for the high-frequency components as does $h(S)$ for the low-frequency components. Namely, $g(S)$ is a "differencing operator," just as $h(S)$ is an averaging operator. (We will see below that for the Haar system, $g(S) = (I - S)/\sqrt{2}$.) For $w(S) \in \mathcal{P}$, we have

$$\begin{aligned} K^*w(S)\phi &= JH^*w(S)\phi = Jh(S)w(S^2)\phi \\ &= g(S)w(S^{-2})\phi = g(S)E^*Cw(S)\phi; \end{aligned} \tag{62}$$

hence

$$K^* = g(S)E^*C = g(S)CE^*. \tag{63}$$

PROPOSITION 4. *The adjoints of K^* and K_α^* are given by*

$$\begin{aligned} Kv(S)\phi &= ECg(S^{-1})v(S)\phi = Ev(S^{-1})D\psi, \\ K_\alpha D^\alpha v(S)\phi &= D^{\alpha+1}ECg(S^{-1})v(S)\phi = D^{\alpha+1}Ev(S^{-1})D\psi. \end{aligned} \tag{64}$$

Proof. Since $K = ECg(S^{-1})$ and $K_\alpha D^\alpha = D^{\alpha+1}K$, we have

$$Kv(S)\phi = ECg(S^{-1})v(S)\phi = Eg(S)v(S^{-1})\phi = Ev(S^{-1})D\psi \tag{65}$$

and

$$K_\alpha D^\alpha v(S)\phi = D^{\alpha+1}Eg(S^{-1})v(S)\phi = D^{\alpha+1}Ev(S^{-1})D\psi. \quad \square \tag{66}$$

Example 3 (The Haar system, continued). Returning to $h(S) = (I+S)/\sqrt{2}$ and choosing $\varepsilon(S) = -S$, we have

$$\begin{aligned}
J\,\phi_n &= -S\,(-S^{-1})^n\,\phi = (-1)^{1-n}\,\phi_{1-n},\\
g(S) &= -S\,\frac{1}{\sqrt{2}}\,(I - S^{-1}) = \frac{1}{\sqrt{2}}\,(I - S),\\
\psi &= D^{-1}\frac{1}{\sqrt{2}}(I - S)\,\phi = \frac{1}{\sqrt{2}}\left(\chi_{[0,1/2)} - \chi_{[1/2,1)}\right),\\
K^*\,\phi_n &= JH^*\,\phi_n = \frac{1}{\sqrt{2}}\,J\,(\phi_{2n} + \phi_{2n+1})\\
&= \frac{1}{\sqrt{2}}\,(\phi_{-2n} - \phi_{1-2n}),\\
K\,\phi_n &= E\,S^{-n}\,D\psi = \frac{1}{\sqrt{2}}\,E\,S^{-n}\,(I - S)\,\phi\\
&= \frac{1}{\sqrt{2}}\,E\,(S^{-n} - S^{1-n})\,\phi\\
&= \frac{1}{\sqrt{2}}\begin{cases} S^{-n/2}\,\phi, & n \text{ even},\\ S^{(1-n)/2}\,\phi, & n \text{ odd}\end{cases}\\
&= \frac{(-1)^n}{\sqrt{2}}\,\phi_{-[n/2]}.
\end{aligned} \tag{67}$$

It can be easily checked that the functions $\psi_n^\alpha \equiv D^\alpha S^n \psi$ $(\alpha, n \in \mathbb{Z})$ are mutually orthogonal. They form the *Haar basis* of $L^2(\mathbb{R})$.

Note. The "time-reversal" associated with K and K^* (i.e., $n \to -n$) is due to the presence of C in J. It is harmless, since ultimately it is only KK^* and K^*K that count. However, it can be removed by replacing K_α^* and K^* by

$$\begin{aligned}
K_\alpha'^* &\equiv J_\alpha H_\alpha^* C = K_\alpha^* C,\\
K'^* &\equiv K_0'^* D = K^* C.
\end{aligned} \tag{68}$$

For then

$$\begin{aligned}
K'K'^* &= I, &\quad K'^*K' &= K^*K,\\
K'H^* &= 0, &\quad HK'^* &= 0,
\end{aligned} \tag{69}$$

with similar formulas for H_α and K_α'. Hence H and K' lead to an orthogonal decomposition of V equivalent to that given by H and K. Furthermore, K' does not reverse time. For the Haar system,

$$\begin{aligned}
K'^*\phi_n &= K^*C\phi_n = K^*\phi_{-n} = \frac{1}{\sqrt{2}}\,(\phi_{2n} - \phi_{2n+1}),\\
K'\phi_n &= CK\phi_n = \frac{(-1)^n}{\sqrt{2}} C\,\phi_{-[n/2]}\\
&= \frac{(-1)^n}{\sqrt{2}}\,\phi_{[n/2]}.
\end{aligned} \tag{70}$$

We prefer K_α and K because they are "cleaner" with respect to the complex structure. For example, the complex decomposition and reconstruction algorithm given in §3 is less natural if K'_α and K' are used.

PROPOSITION 5. *The pairs of operators* $\{H, K\}$ *and* $\{H_\alpha, K_\alpha\}$ *satisfy*

$$\begin{aligned} HH^* &= Eh(S^{-1})h(S)E^* = I, \\ KK^* &= Eg(S^{-1})g(S)E^* = I, \\ HK^* &= Eh(S^{-1})g(S)E^*C = 0, \\ KH^* &= CEg(S^{-1})h(S)E^* = 0. \end{aligned} \tag{71}$$

and

$$H_\alpha H_\alpha^* = K_\alpha K_\alpha^* = I_{\alpha+1}, \qquad H_\alpha K_\alpha^* = K_\alpha H_\alpha^* = 0. \tag{72}$$

Proof. (Note that $H_\alpha H_\alpha^* = I_{\alpha+1}$ has already been shown; it is included here for completeness, since it belongs with the other identities.) The first equation follows from $H^* = H_0^* D$ and $H_0^* H_0 = I$. The second follows from the first and $K^* = JH^*$, since $J^*J = I$. The last two equations follow from Lemma 1, since $h(S^{-1})g(S)$ and $g(S^{-1})h(S)$ are odd functions hence their even parts vanish. This proves the identities for H and K. The other identities follow by naturality. □

PROPOSITION 6. *The pairs* $\{H, K\}$ *and* $\{H_\alpha, K_\alpha\}$ *give orthogonal decompositions of* V *and* V_α. *We have*

(a)

$$\begin{gathered} HV = KV = V, \\ H^*V = V_1, \qquad K^*V = W_1, \\ H^*H + K^*K = I, \end{gathered} \tag{73}$$

(b)

$$\begin{gathered} H_\alpha V_\alpha = K_\alpha V_\alpha = V_{\alpha+1}, \\ H_\alpha^* V_{\alpha+1} = V_{\alpha+1}, \qquad K_\alpha^* V_{\alpha+1} = W_{\alpha+1}, \\ H_\alpha^* H_\alpha + K_\alpha^* K_\alpha = I_\alpha. \end{gathered} \tag{74}$$

Proof. We have

$$HV = D^{-1}H_0 V = D^{-1}V_1 = V, \tag{75}$$

and since $K = -HJ$, it follows that $KV = HJV = HV = V$. Also

$$H^*V = H_0^* DV = H_0^* V_1 = V_1. \tag{76}$$

$KK^* = I$ and $HK^* = 0$ (Proposition 4) imply that K^* is injective and its range is orthogonal to that of H^*, i.e., to V_1. Hence $K^*V \subset W_1$. To show that $K^*V = W_1$, let $u(S) \in \mathcal{P}$. We need to find $v(S), w(S) \in \mathcal{P}$ such that

$$u(S)\phi = H^*v(S)\phi + K^*w(S)\phi = h(S)v(S^2)\phi + g(S)w(S^{-2})\phi \tag{77}$$

or, equivalently, dropping ϕ,

$$u(S) = h(S)v(S^2) + g(S)w(S^{-2}). \tag{78}$$

Use Lemma 1 to decompose this equation into its even and odd parts:

$$(79)\qquad \begin{aligned} u_+(S) &= Eu(S)E^* = E\left[h(S)v(S^2) + g(S)w(S^{-2})\right]E^* \\ &= h_+(S)v(S) + g_+(S)w(S^{-1}), \\ u_-(S) &= Ou(S)E^* = O\left[h(S)v(S^2) + g(S)w(S^{-2})\right]E^* \\ &= h_-(S)v(S) + g_-(S)w(S^{-1}), \end{aligned}$$

which can be written in matrix form as

$$(80)\qquad \begin{bmatrix} u_+(S) \\ u_-(S) \end{bmatrix} = \begin{bmatrix} h_+(S) & g_+(S) \\ h_-(S) & g_-(S) \end{bmatrix} \begin{bmatrix} v(S) \\ w(S^{-1}) \end{bmatrix} \equiv U(S) \begin{bmatrix} v(S) \\ w(S^{-1}) \end{bmatrix}.$$

But Proposition 5 is precisely the statement that the matrix $U(S)$ is *unitary*, i.e., $U(S)^*U(S) = I$. Multiplying by $U(S)^*$, we obtain the unique solution

$$(81)\qquad \begin{bmatrix} v(S) \\ w(S^{-1}) \end{bmatrix} = \begin{bmatrix} h_+(S^{-1}) & h_-(S^{-1}) \\ g_+(S^{-1}) & g_-(S^{-1}) \end{bmatrix} \begin{bmatrix} u_+(S) \\ u_-(S) \end{bmatrix},$$

which shows that $V \equiv V_1 \oplus W_1 = H^*V \oplus K^*V$. Applying H and K to (77) gives

$$(82)\qquad v(S)\phi = Hu(S)\phi, \qquad w(S)\phi = Ku(S)\phi,$$

which proves that $H^*H + K^*K = I$ as claimed. For the range of K_α^* we have $K_\alpha^* V_{\alpha+1} = K_\alpha^* D^{\alpha+1} V = D^\alpha K^* V = D^\alpha W_1 = W_{\alpha+1}$. Finally,

$$(83)\qquad H_\alpha^* H_\alpha + K_\alpha^* K_\alpha = D^\alpha(H^*H + K^*K)D^{-\alpha} = I_\alpha. \qquad \square$$

We now construct the usual "high-frequency filters" G_α. First note that elements in $W_1 \equiv K^*V$ can be written in the form

$$(84)\qquad K^*w(S)\phi = g(S)w(S^{-2})\phi = w(S^{-2})D\psi = Dw(S^{-1})\psi.$$

It follows that ψ is a "basic wavelet," i.e., that

$$(85)\qquad W_\alpha = \overline{D^\alpha \mathcal{P}\psi},$$

and the vectors

$$(86)\qquad \psi_n^\alpha \equiv D^\alpha S^n \psi = D^{\alpha-1}K^*S^{-n}\phi = D^{-1}K_\alpha^* \phi_{-n}^{\alpha+1}$$

form an orthonormal basis for W_α.

Since $K_0^*\colon V_1 \to V$ is injective and its range is W_1, it can be factored uniquely as

$$(87)\qquad K_0^* = G_0^* R_1^*,$$

where $R_1^*\colon V_1 \to W_1$ is an isomorphism and $G_0^*\colon W_1 \to V$ is the inclusion map. From

$$(88)\qquad K_0^* Dw(S)\phi = Dw(S^{-1})\psi = g(S)w(S^{-2})\phi$$

we read off

$$(89)\qquad \begin{aligned} R_1^* Dw(S)\phi &= Dw(S^{-1})\psi, \\ G_0^* Dw(S^{-1})\psi &= g(S)w(S^{-2})\phi. \end{aligned}$$

For the home versions, we have

$$K^* = G_0^* R_1^* D = G_0^* D R^* = G^* R^*, \tag{90}$$

where

$$R^*\colon V \equiv V_0 \to W \equiv W_0 \tag{91}$$

is an isomorphism and $G^*\colon W \to V$, though injective, is not an inclusion map. (This is the price for working with the home versions, which do not preserve the scale.) Note that $R^*R = I_W$ and $RR^* = I_V$; hence $K = RG$ implies that $G = R^*K$ and

$$\begin{aligned} GG^* &= R^*KK^*R = I_W, \\ GH^* &= R^*KH^* = 0, \\ G^*G &= K^*RR^*K = K^*K \end{aligned} \tag{92}$$

(hence $H^*H + G^*G = I$). Therefore H and G give an orthogonal decomposition of V which is equivalent to that given by H and K.

Defining $R_\alpha^*\colon V_\alpha \to W_\alpha$ and $G_\alpha^*\colon W_{\alpha+1} \to V_\alpha$ by $R_\alpha^* D^\alpha = D^\alpha R^*$ and $G_\alpha^* D^{\alpha+1} = D^\alpha G^*$, we get a graded family of filters G_α related to K_α by

$$K_\alpha^* = G_\alpha^* R_{\alpha+1}^*. \tag{93}$$

The operators G_α are, in fact, the usual high-frequency filters, for the latter are defined analogously to H_α, namely, by substituting $g(S)\phi$ for the dilated wavelet $D\psi$:

$$\begin{aligned} G_\alpha^* D^{\alpha+1} u(S)\psi &= G_\alpha^* D^\alpha u(S^2) D\psi \\ &= D^\alpha g(S) u(S^2)\phi = D^\alpha G^* u(S)\psi. \end{aligned} \tag{94}$$

The orthogonal decomposition of V given by H and G induces an orthogonal decomposition of V_α by H_α and G_α which is, in fact, the usual wavelet decomposition and is equivalent to the one given by H_α and K_α in Proposition 6.

Example 4 (The Haar system, final visit). For the Haar system, (87) and (90) give

$$\begin{aligned} R^*\,\phi_n &= D^{-1} R_1^* D\,\phi_n = S^{-n}\,\psi = \psi_{-n}, \\ G^*\,\psi_n &= G_0^* D\,\psi_n = \frac{1}{\sqrt{2}}(I - S)\,\phi_{2n} = \frac{1}{\sqrt{2}}\,(\phi_{2n} - \phi_{2n+1}), \\ G\,\phi_n &= R^* K\,\phi_n = \frac{(-1)^n}{\sqrt{2}}\,R^*\,\phi_{-[n/2]8} \\ &= \frac{(-1)^n}{\sqrt{2}}\,\psi_{[n/2]}. \end{aligned} \tag{95}$$

Remark. We have stated that K_α is more natural than G_α from the point of view of the symmetry associated with J_α. The reason is that both H_α and K_α map V_α to $V_{\alpha+1}$, whereas G_α maps V_α to $W_{\alpha+1}$. This symmetry is reflected by the simple relation $K_\alpha^* = J_\alpha H_\alpha^*$, whereas the relation

$$G_\alpha^* = J_\alpha H_\alpha^* R_{\alpha+1} \tag{96}$$

is somewhat more complicated. A more concrete divident of this symmetry will appear in the next section, where the complex combinations $H_\alpha^* \pm iK_\alpha^*$ will be considered. The corresponding combinations $H_\alpha^* \pm iG_\alpha^*$ do not make sense, as the two operators have different domains.

We now prove an identity which will be useful later.

PROPOSITION 7.

$$\begin{aligned} K^*H - H^*K &= J, \\ K_\alpha^* H_\alpha - H_\alpha^* K_\alpha &= J_\alpha. \end{aligned} \tag{97}$$

Proof. For $u(S)\phi = H^*v(S)\phi + K^*w(S)\phi \in V$, we have

$$\begin{aligned} (K^*H - H^*K)\,u(S)\phi &= K^*v(S)\phi - H^*w(S)\phi \\ &= JH^*v(S)\phi + JK^*w(S)\phi = Ju(S)\phi. \end{aligned} \tag{98}$$

The identity at arbitrary scale follows from naturality:

$$(K_\alpha^* H_\alpha - H_\alpha^* K_\alpha)D^\alpha = D^\alpha(K^*H - H^*K) = D^\alpha J = J_\alpha D^\alpha. \tag{99}$$ □

It is natural to wonder whether the complex structures J_α extend to define a "global" complex structure on $L^2(\mathbb{R})$. We now show that this is not the case.

PROPOSITION 8. *The complex structure J_α on V_α is not an extension of $J_{\alpha+1}$.*

Proof. The statement that J_α is an extension of $J_{\alpha+1}$ means that $J_{\alpha+1}$ is the restriction of J_α to $V_{\alpha+1}$, i.e.,

$$K_\alpha^* \equiv J_\alpha H_\alpha^* = H_\alpha^* J_{\alpha+1}. \tag{100}$$

If this were true, then left-multiplication by H_α would imply

$$0 = H_\alpha K_\alpha^* = H_\alpha H_\alpha^* J_{\alpha+1} = J_{\alpha+1}, \tag{101}$$

which contradicts $J_{\alpha+1}^2 = -I_{\alpha+1}$. □

3.1. Complex decomposition and reconstruction. The decomposition/reconstruction algorithm of the last section can be iterated, and when repeated indefinitely gives a unique representation of any function $f \in L^2(\mathbb{R})$ as an L^2-convergent infinite orthogonal sum of "detail" functions at finer and finer scales. We develop the algorithm formally in the home version, which will be seen to be much more convenient. For a rigorous treatment, see Daubechies [1988b]. Given any $D^\alpha v^\alpha(S)\phi \in V_\alpha$, write

$$v^\alpha \equiv v^\alpha(S)\phi \in V \tag{102}$$

for brevity. Since

$$\begin{aligned} I &= K^*K + H^*H = K^*K + H^*(K^*K + H^*H)H = \cdots \\ &= \sum_{\beta=1}^{N} (H^*)^{\beta-1} K^*KH^{\beta-1} + (H^*)^N H^N, \end{aligned} \tag{103}$$

we have

$$
\begin{aligned}
v^\alpha &= \sum_{\beta=1}^{N} (H^*)^{\beta-1} K^* K H^{\beta-1} v^\alpha + (H^*)^N H^N v^\alpha \\
&= \sum_{\beta=1}^{N} (H^*)^{\beta-1} K^* w^{\alpha+\beta} + (H^*)^N v^{\alpha+N},
\end{aligned} \tag{104}
$$

where

$$
\begin{aligned}
w^{\alpha+\beta} &\equiv K H^{\beta-1} v^\alpha, \\
v^{\alpha+N} &\equiv H^N v^\alpha.
\end{aligned} \tag{105}
$$

The vector $v^{\alpha+N}$ represents an N-fold smoothed version of v^α, while $w^{\alpha+\beta}$ represents the detail filtered out at the βth iteration. The terms in the above sum are orthogonal, since for $\gamma > \beta$ we have

$$
\begin{aligned}
&\langle (H^*)^{\beta-1} K^* w^{\alpha+\beta}, (H^*)^{\gamma-1} K^* w^{\alpha+\gamma} \rangle \\
&\qquad = \langle w^{\alpha+\beta}, K (H^*)^{\gamma-\beta} K^* w^{\alpha+\beta} \rangle = 0.
\end{aligned} \tag{106}
$$

To see what this expansion means in terms of the scale-preserving filters, apply D^α and use naturality (see Appendix):

$$
\begin{aligned}
D^\alpha v^\alpha = {} & \sum_{\beta=1}^{N} H^*_\alpha H^*_{\alpha+1} \cdots H^*_{\alpha+\beta-2} K^*_{\alpha+\beta-1} D^{\alpha+\beta} w^{\alpha+\beta} \\
& + H^*_\alpha H^*_{\alpha+1} \cdots H^*_{\alpha+N-1} D^{\alpha+N} v^{\alpha+N}.
\end{aligned} \tag{107}
$$

This formula shows the advantage of the home versions of the filters, which "zoom" in and out to get the detail at any desirable scale and can therefore be used repeatedly without changing operators.

It can be shown that $v^{\alpha+N} \to 0$ in $L^2(\mathbb{R})$ as $N \to \infty$, which gives the orthogonal decomposition

$$
D^\alpha v^\alpha = D^\alpha \sum_{\beta=1}^{\infty} (H^*)^{\beta-1} K^* w^{\alpha+\beta}. \tag{108}
$$

Since

$$
L^2(\mathbb{R}) = \overline{\bigcup_{\alpha \in \mathbb{Z}} V_\alpha}, \tag{109}
$$

any $L^2(\mathbb{R})$ function can be approximated as accurately as desired in the form of (108). We need only choose a "cut-off" scale α, which means that all detail finer than 2^α will be ignored. Moreover, since

$$
W_{\alpha+1} \subset V_\alpha \perp W_\alpha, \tag{110}
$$

the above gives the orthogonal decomposition

$$
L^2(\mathbb{R}) = \overline{\bigoplus_{\alpha \in \mathbb{Z}}} W_\alpha. \tag{111}
$$

Let us now look at the reconstruction and decomposition formulas in light of the complex structure. A single iteration gives

$$v^{\alpha} = H^{*}v^{\alpha+1} + K^{*}w^{\alpha+1} = H^{*}v^{\alpha+1} + JH^{*}w^{\alpha+1}, \tag{112}$$

where

$$v^{\alpha+1} = Hv^{\alpha}, \qquad w^{\alpha+1} = Kv^{\alpha} = -HJv^{\alpha}. \tag{113}$$

Since J is like multiplication by $\pm i$, let us define the complex conjugate pairs of vectors

$$\begin{aligned} \zeta^{\alpha} &= \frac{v^{\alpha} + iw^{\alpha}}{\sqrt{2}}, \\ \bar{\zeta}^{\alpha} &= \frac{v^{\alpha} - iw^{\alpha}}{\sqrt{2}}, \end{aligned} \tag{114}$$

where the $\sqrt{2}$ in the denominators is for later convenience. ζ^{α} and $\bar{\zeta}^{\alpha}$ belong to the *complexification* of V, i.e., to

$$V^{c} = V \oplus iV = \mathbb{C} \otimes V. \tag{115}$$

We endow V^{c} with the Hermitian inner product obtained from V by extending $\mathbb{C}$—linearly in the *second* factor and antilinearly in the first factor (this is the convention used in the physics literature). Note that under this inner product, iV is not orthogonal to V, and hence the direct sum $V \oplus iV$ is *not* an orthogonal decomposition. We now extend all operators on V to V^{c} by $\mathbb{C}$-linearity, denoting the extensions by the same symbols. Substituting

$$v^{\alpha+1} = \frac{\zeta^{\alpha+1} + \bar{\zeta}^{\alpha+1}}{\sqrt{2}}, \qquad w^{\alpha+1} = \frac{\zeta^{\alpha+1} - \bar{\zeta}^{\alpha+1}}{\sqrt{2}i} \tag{116}$$

into the reconstruction formula, we obtain

$$v^{\alpha} = Z^{*}\zeta^{\alpha+1} + \bar{Z}^{*}\bar{\zeta}^{\alpha+1} = 2\,\Re\left(Z^{*}\zeta^{\alpha+1}\right), \tag{117}$$

where the operators $Z^{*}, \bar{Z}^{*}$: $V^{c} \to V^{c}$ are defined by

$$Z^{*} = \frac{H^{*} - iK^{*}}{\sqrt{2}}, \qquad \bar{Z}^{*} = \frac{H^{*} + iK^{*}}{\sqrt{2}}. \tag{118}$$

Therefore

$$Z = \frac{H + iK}{\sqrt{2}} = \frac{H - iHJ}{\sqrt{2}} = H\left(\frac{I - iJ}{\sqrt{2}}\right). \tag{119}$$

The operators

$$P^{\pm} = \frac{I \mp iJ}{2} \tag{120}$$

are the orthogonal projections in V^{c} to the eigenspaces of J with eigenvalues $\pm i$, since

$$\begin{aligned} JP^{\pm} &= \pm iP^{\pm}, \\ (P^{\pm})^{*} &= P^{\pm} = (P^{\pm})^{2}. \end{aligned} \tag{121}$$

We have the orthogonal decomposition

$$V^c = V^+ \oplus V^-, \qquad V^\pm \equiv P^\pm V. \tag{122}$$

The above shows that the operator $Z\colon V^c \to V^c$ and its adjoint satisfy

$$Z = \sqrt{2}\,HP^+, \qquad Z^* = \sqrt{2}P^+H^*. \tag{123}$$

Since H^* is injective, it follows that the range of Z^* is V^+ and the kernel of Z is V^-. Similarly,

$$\bar{Z} = \sqrt{2}\,HP^-, \qquad \bar{Z}^* = \sqrt{2}P^-H^*, \tag{124}$$

so the range of $\bar{Z}^*$ is V^- and the kernel of $\bar{Z}$ is V^+.

PROPOSITION 9. *The operators Z and $\bar{Z}$ satisfy*

$$\begin{aligned} ZZ^* &= I, & \bar{Z}\bar{Z}^* &= I,\\ Z\bar{Z}^* &= 0, & \bar{Z}Z^* &= 0,\\ Z^*Z &= P^+, & \bar{Z}^*\bar{Z} &= P^-. \end{aligned} \tag{125}$$

Proof. By Proposition 5, we have

$$\begin{aligned} ZZ^* &= 2HP^+P^+H^* = 2HP^+H^*\\ &= H(I - iJ)H^* = HH^* - iHK^* = I, \end{aligned} \tag{126}$$

and

$$Z\bar{Z}^* = 2HP^+P^-H^* = 0. \tag{127}$$

Furthermore, by Propositions 6 and 7,

$$\begin{aligned} 2Z^*Z &= (H^* - iK^*)(H + iK)\\ &= (H^*H + K^*K) + i(H^*K - K^*H)\\ &= I - iJ = 2P^+. \end{aligned} \tag{128}$$

The other equalities follow from complex conjugation, which exchanges Z and $\bar{Z}$. □

Proposition 9 gives a new, *complex* decomposition and reconstruction algorithm. Given $\zeta^\alpha \in V^c$, define $\zeta^{\alpha+1}, \chi^{\alpha+1} \in V^c$ by

$$\begin{aligned} \zeta^{\alpha+1} &= Z\zeta^\alpha = \sqrt{2}HP^+\zeta^\alpha,\\ \chi^{\alpha+1} &= \bar{Z}\zeta^\alpha = \sqrt{2}HP^-\zeta^\alpha. \end{aligned} \tag{129}$$

Then ζ^α can be reconstructed using

$$\zeta^\alpha = Z^*\zeta^{\alpha+1} + \bar{Z}^*\chi^{\alpha+1}. \tag{130}$$

Like the real algorithm, this can be iterated. By Proposition 9,

$$\begin{aligned} I &= \bar{Z}^*\bar{Z} + Z^*Z = \bar{Z}^*\bar{Z} + Z^*(\bar{Z}^*\bar{Z} + Z^*Z)Z = \cdots \\ &= \sum_{\alpha=1}^{N} (Z^*)^{\alpha-1}\bar{Z}^*\bar{Z}Z^{\alpha-1} + (Z^*)^N Z^N. \end{aligned} \tag{131}$$

Hence for any $\zeta^0 \in V^c$ and $N \in \mathbb{N}$,

$$\begin{aligned} \zeta^0 &= \sum_{\alpha=1}^{N} (Z^*)^{\alpha-1}\bar{Z}^*\bar{Z}Z^{\alpha-1}\zeta^0 + (Z^*)^N Z^N \zeta^0 \\ &= \sum_{\alpha=1}^{N} (Z^*)^{\alpha-1}\bar{Z}^*\chi^\alpha + (Z^*)^N \zeta^N, \end{aligned} \tag{132}$$

where

$$\chi^\alpha \equiv \bar{Z}Z^{\alpha-1}\zeta^0, \qquad \zeta^N \equiv Z^N\zeta^0. \tag{133}$$

The convergence and significance (in relation to the usual frequency decomposition) of this algorithm will be studied elsewhere. Here we merely note that the terms in the above sum are mutually orthogonal, since for $\beta > \alpha$ Proposition 9 implies

$$\langle (Z^*)^{\alpha-1}\bar{Z}^*\chi^\alpha, (Z^*)^{\beta-1}\bar{Z}^*\chi^\beta \rangle = \langle \chi^\alpha, \bar{Z}(Z^*)^{\beta-\alpha}\bar{Z}^*\chi^\alpha \rangle = 0. \tag{134}$$

Just as the filters H and K have associated averaging- and "differencing" operators $h(S)$ and $g(S)$, there are operators associated with Z and $\bar{Z}$. If

$$\begin{aligned} Z^*\zeta(S)\phi &= 2^{-1/2}\left(h(S)\zeta(S^2) - ig(S)\zeta(S^{-2})\right)\phi \\ &= \left(\frac{h(S) - ig(S)C}{\sqrt{2}}\right) E^*\zeta(S)\phi; \end{aligned} \tag{135}$$

then $Z^* = z^*E^*$ and, similarly, $\bar{Z}^* = \bar{z}^*E^*$, where

$$z^* = \frac{h(S) - ig(S)C}{\sqrt{2}}, \qquad \bar{z}^* = \frac{h(S) + ig(S)C}{\sqrt{2}}. \tag{136}$$

Note, however, that these operators do not commute with S, since they contain C.

PROPOSITION 10. *The operators z and $\bar{z}$ satisfy*

$$zz^* = z^*z = \bar{z}\bar{z}^* = \bar{z}^*\bar{z} = I. \tag{137}$$

Proof. A straight computation gives

$$\begin{aligned} z^*z &= \frac{1}{2}\left(h(S) - ig(S)C\right)\left(h(S^{-1}) + iCg(S^{-1})\right) \\ &= \frac{1}{2}\left(h(S)h(S^{-1}) + g(S)g(S^{-1})\right) + \frac{i}{2}\left(h(S)g(S) - g(S)h(S)\right)C \\ &= \frac{1}{2}\left(h(S)h(S^{-1}) + h(-S)h(-S^{-1})\right) \\ &= DEh(S)h(S^{-1})E^*D^{-1} = I \end{aligned} \tag{138}$$

by Proposition 5, and the other identities are proved similarly. □

Furthermore, there is a natural complex function which stands in the same relation to z as do ϕ and ψ stand to $h(S)$ and $g(S)$, respectively. Consider the function $\bar{\chi} \in V_{-1}^{+}$ defined by

$$D\bar{\chi} \equiv Z^* \phi = z^* \phi = \frac{h(S) - ig(S)}{\sqrt{2}} \phi. \tag{139}$$

Using $D\phi = h(S)\phi$ and $D\psi = g(S)\phi$, we get

$$\bar{\chi} = \frac{\phi - i\psi}{\sqrt{2}}. \tag{140}$$

Similarly, we define $\chi \in V_{-1}^{-}$ by

$$\chi \equiv D^{-1} \bar{Z}^* \phi = \frac{\phi + i\psi}{\sqrt{2}}. \tag{141}$$

The functions ϕ and ψ are somewhat reminiscent of the cosine- and sine-function. By that analogy, χ and $\bar{\chi}$ correspond to *the complex exponentials*! It may be that χ and $\bar{\chi}$, which combine averaging and "differencing" in a complex way, play a similar role in wavelet analysis as do the complex exponentials in Fourier analysis.

Finally, note that the Hermitian inner product in V^c has the decomposition

$$\langle \zeta, \zeta' \rangle = \langle \zeta, P^+ \zeta \rangle + \langle \zeta, P^- \zeta' \rangle, \tag{142}$$

which is the sum of the inner products in V^+ and V^-. Now the restriction of P^+ to $V \subset V^c$ maps V one-to-one onto V^+, although it cannot preserve the inner product since V is real while V^+ is complex. In fact, for $\zeta \equiv P^+ u$ and $\zeta' \equiv P^+ u'$ in V^+, we have

$$\begin{aligned} \langle \zeta, \zeta' \rangle = \langle u, P^+ u' \rangle &= \frac{1}{2} \langle u, u' \rangle - \frac{i}{2} \langle u, Ju' \rangle \\ &\equiv \frac{1}{2} \langle u, u' \rangle - i\omega(u, u'), \end{aligned} \tag{143}$$

which states that imaginary part of the Hermitian inner product in V^+ is the skew-symmetric bilinear form ω. This form is known as the *symplectic structure* determined by the complex structure J together with the inner product on V. It plays a fundamental role in a broad variety of subjects, e.g., group representation theory, classical mechanics, and quantum mechanics (see Marsden [1981]). I do not know whether it has any significance for wavelet analysis, but that seems to me a question worth exploring.

Some final remarks. (1) Although no attempt has been made here to work in an *abstract* setting, the algebraic approach clearly lends itself to such generalizations. We have made use of just a few facts about our initial set-up, for example, that the inner product is invariant under S and D and also under C and M. This suggests that it is necessary to *begin* with an algebra whose generators include S and D and other operators such as C and M, subject to certain relations, and to look for *representations* of this algebra, i.e., for a vector space on which the elements of the algebra act as operators. In our case, the vector ϕ provides a representation on $L^2(\mathbb{R})$. ϕ is a *cyclic* vector because its orbit under the algebra spans $L^2(\mathbb{R})$, and the representation is *irreducible.* The representation is also *orthogonal* in the sense that the generators are represented by orthogonal operators. Solving the dilation equation $D\phi = h(S)\phi$ therefore amounts to constructing the entire representation!

(2) The complex combinations $H \pm iK$ are reminiscent of certain operators that occur in quantum mechanics in connection with *coherent states* and that are also associated with time-frequency localization. See Kaiser [1990], Kaiser and Streater [1991].

4. Appendix. Here we assemble various information for the reader's convenience. We begin with a summary of the operational calculus.

4.1. Operational calculus and basis representations. We list the actions of various operators in their "home versions," both intrinsically and on the orthonormal bases $\{\phi_n\}$ of $V \equiv V_0$ and $\{\psi_n\}$ of $W \equiv W_0$. (See Daubechies [1988b].) Since $\varepsilon(S)$ is odd, it may be written in the form $\varepsilon(S) = \sum_k \varepsilon_k \, S^{2k+1}$.

$$\begin{aligned}
H^*u(S)\phi &= h(S)u(S^2)\phi, & H^*\phi_k &= \sum_n h_{n-2k}\phi_n, \\
Hu(S)\phi &= Eh(S^{-1})u(S)\phi, & H\phi_n &= \sum_k h_{n-2k}\phi_k, \\
K^*u(S)\phi &= g(S)u(S^{-2})\phi, & K^*\phi_k &= \sum_n g_{n+2k}\phi_n, \\
Ku(S)\phi &= Eg(S)u(S^{-1})\phi, & K\phi_n &= \sum_k g_{n+2k}\phi_k, \\
G^*u(S)\psi &= g(S)u(S^2)\phi, & G^*\psi_k &= \sum_n g_{n-2k}\phi_n, \\
Gu(S)\phi &= Eg(S^{-1})u(S)\psi, & G\phi_n &= \sum_k g_{n-2k}\psi_k, \\
Ju(S)\phi &= \varepsilon(S)u(-S^{-1})\,\phi, & J\phi_n &= (-1)^n \sum_k \varepsilon_k \, \phi_{2k+1-n}.
\end{aligned}$$

4.2. Naturality. Naturality relates the scale-preserving filters $H_\alpha, K_\alpha, G_\alpha$, and Z_α to one another and to their home versions, which involve changes of scale. We give the relations for Z_α; the others are similar. The last two equations show the advantage of the home versions for iterated decomposition and reconstruction.

$$\begin{aligned}
&Z_\alpha D^\alpha = D^\alpha Z_0 = D^{\alpha+1} Z, \\
&Z_\alpha^* D^{\alpha+1} = D^\alpha Z_0^* D = D^\alpha Z^*, \\
&Z_{\alpha+N} Z_{\alpha+N-1} \cdots Z_\alpha D^\alpha = D^{\alpha+N+1} Z^{\alpha+N+1}, \\
&Z_\alpha^* Z_{\alpha+1}^* \cdots Z_{\alpha+N}^* D^{\alpha+N+1} = D^\alpha (Z^*)^{N+1}.
\end{aligned}$$

4.3. Relation to Daubechies' notation. Upon taking the Fourier transform, our operators $u(S)$ become functions $u(e^{i\xi T})$ on the unit circle. Their connections with those occuring in Daubechies' paper, where the sampling interval $T = 1$, are as follows: Write $\varepsilon(S) = S\varepsilon_-(S^2)$. Then

$$\begin{aligned} h(e^{i\xi}) &= H(\xi), & g(e^{i\xi}) &= G(\xi), \\ h_+(e^{i\xi}) &= \alpha(\xi), & g_+(e^{i\xi}) &= \gamma(\xi), \\ h_-(e^{i\xi}) &= \beta(\xi), & g_-(e^{i\xi}) &= \delta(\xi), \\ \varepsilon_-(e^{i\xi}) &= -e^{i\lambda(\xi)}, & U(e^{i\xi}) &= M(\xi). \end{aligned}$$

4.4. Analogy with complex numbers. We give a "pocket dictionary" of the correspondence between wavelet operations and operations on complex numbers. This is done for the relation $V_0 \approx V_1 \oplus iV_1$, although the same analogy holds at every scale.

$$\begin{aligned} V_1 &\longleftrightarrow \mathbb{R} \\ W_1 &\longleftrightarrow i\mathbb{R} \\ V_0 = V_1 \oplus JV_1 &\longleftrightarrow \mathbb{C} = \mathbb{R} \oplus i\mathbb{R} \\ u(S)\phi \mapsto H_0 u(S)\phi &\longleftrightarrow x + iy \mapsto \Re(x+iy) \\ u(S)\phi \mapsto K_0 u(S)\phi &\longleftrightarrow x + iy \mapsto \Im(x+iy) \\ u(S)\phi \mapsto Ju(S)\phi &\longleftrightarrow z \mapsto iz \\ Dv(S)\phi \mapsto H_0^* Dv(S)\phi &\longleftrightarrow x \mapsto x + i0 \\ Dw(S)\phi \mapsto K_0^* Dw(S)\phi &\longleftrightarrow y \mapsto 0 + iy \\ u(S)\phi = H_0^* Dv(S)\phi + JH_0^* Dw(S)\phi &\longleftrightarrow z = x + iy \\ Dw(S)\phi \mapsto R_1^* Dw(S)\phi &\longleftrightarrow y \in \mathbb{R} \mapsto iy \in i\mathbb{R}. \end{aligned}$$

Acknowledgment. The author thanks Yuli Makovoz for helpful discussions.

REFERENCES

I. Daubechies [1988], *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41, pp. 909–996.

G. Kaiser [1990], *Quantum Physics, Relativity, and Complex Spacetime: Towards a New Synthesis*, North-Holland, Amsterdam.

G. Kaiser and R. F. Streater [1991], *Windowed Radon transforms, analytic signals and the wave equations*, in Wavelets—A Tutorial, C. K. Chui, ed., Academic Press, New York, to appear.

S. Mallat [1989] *Multiresolution approximation and wavelet orthonormal bases in $L^2(IR)$*, Trans. Amer. Math. Soc., 315, pp. 69–87.

J. E. Marsden [1981], *Lectures on Geometric Methods in Mathematical Physics*, Notes from a Conference held at the University of Lowell, CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 37, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Y. Meyer [1990], *Ondelettes et Opérateurs* I, Hermann, Paris.

G. Strang [1990], *Wavelets and dilation equations*, SIAM Rev., 31, pp. 614–627.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 244-261, January 1992

ITERATIVE RECONSTRUCTION OF MULTIVARIATE BAND-LIMITED FUNCTIONS FROM IRREGULAR SAMPLING VALUES*

HANS G. FEICHTINGER† AND KARLHEINZ GRÖCHENIG‡

Abstract. This paper describes a real analysis approach to the problem of complete reconstruction of a band-limited, multivariate function f from irregularly spaced sampling values $(f(x_i))_{i\in I}$. The required sampling density of the set $X=(x_i)_{i\in I}$ depends only on the spectrum Ω of f. The proposed reconstruction methods are iterative and stable and converge for a given function f with respect to any weighted L^p-norm $1\leqq p\leqq\infty$, for which f belongs to the corresponding Banach space $L^p_v(\mathbb{R}^m)$. It is also shown that any band-limited function f can be represented as a series of translates $L_{y_j}g$ (with complex coefficients) for a given integrable, band-limited function g if the Fourier transform satisfies $\hat{g}(t)\neq 0$ over Ω and the family $Y=(y_j)_{j\in J}$ is sufficiently dense. Moreover, the behavior of the coefficients (such as weighted p-summability) corresponds precisely to the global behavior of f (i.e., membership in the corresponding weighted L^p-space). The proofs are based on a careful analysis of convolution relations, spline approximation operators, and discretization operators (approximation of functions by discrete measures). In contrast to Hilbert space methods, the techniques used here yield pointwise estimates. Special cases of the algorithms presented provide a theoretical basis for methods suggested recently in the engineering literature. Numerical experiments have demonstrated the efficiency of these methods convincingly.

Key words. irregular sampling, iterative reconstruction, approximation of convolutions, multivariate band-limited functions, function spaces

AMS(MOS) subject classifications. 10G99, 65D99, 42A65, 42B99

1. Introduction.

According to the classical sampling theorem attributed to Whittaker, Shannon, Kotel'nikov, and several others, a band-limited function can be reconstructed from sampling values over any sufficiently fine lattice. Because of its great importance in information theory, electrical engineering, signal processing, and other applications, a lot of work has been carried out on improvements and extensions, e.g., to lattices in higher dimensions [PM], [Me], [DM]. Reviews and extensive references on these investigations are given in [Bu], [Je], [Hi2], [BERS], [BSS], and [Ma] (some of which mention the irregular sampling problem).

The regular sampling theorem is based on Fourier series and Poisson's formula. This limits the possible sampling sets to lattices, in other words, to discrete sets that arise from the standard lattice $\mathbb{Z}^m\subseteq\mathbb{R}^m$ through application of an invertible real $n\times n$ matrix (cf. [Co]). However, in many applications, e.g., optics, tomography, synthetic aperture radar, computer graphics, signal processing, meteorology, and geophysics it is necessary to deal with situations where sampling values are *not* available on a regular grid [SA], [Ce], [PK], [St], [So].

In this paper a new *iterative approach* to the irregular sampling problem is described which represents a *constructive* solution to this question. The proposed algorithms use only standard operations which are also well suited for numerical implementation.

1.1. The requirements for an irregular sampling theory.

As a simple model for sampling theory let us recall the regular sampling theorem: Given a band-limited function (of finite energy) $f\in L^2(\mathbb{R})$ with spectrum in $[-\pi W, \pi W]$, f can be represented as a cardinal series

$$f(t)=\sum_{k\in\mathbb{Z}} f(k/W)\,\mathrm{sinc}\,(\pi(Wt-k))$$

* Received by the editors March 29, 1990; accepted for publication (in revised form) April 29, 1991.

† Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria.

‡ Department of Mathematics, U-9, University of Connecticut, Storrs, Connecticut 06269.

where sinc $(x) := x^{-1} \sin(x)$ for $x \neq 0$ and sinc $(0) = 1$; cf. [Bu], [BSS], [Pa]. The series is convergent in the L^2-sense and uniformly. In the case of oversampling, i.e., if sampling values $(f(\alpha k)_{k \in \mathbb{Z}})$ for some $\alpha < 1/W$ are known, the nonintegrable sinc-function can be replaced by other "windows" (cf. [Sch], [BERS]) with a more rapid decay. Then the series converges with respect to weighted L^p-norms as well, whenever f belongs to such a weighted L^p-space (cf. [F4, Thm. 2]). Such alternative windows with rapid decay are also required if band-limited tempered distributions must be reconstructed from their regular sampling values (cf. [Ca], [BERS], [Se]).

The regular sampling theorem combines *two aspects*:

(a) Any band-limited function can be *completely reconstructed* from its sampled values over a *sufficiently fine* lattice by means of a simple series expansion with the sampled values as coefficients.

(b) Any band-limited function can be *expanded into a series* with *translates of a single function* g as building blocks.

For the irregular sampling problem we will discuss these two aspects separately; see Theorems 3.1 and 3.2. Practical considerations impose the following requirements on an irregular sampling theory. They are satisfied in the case of regular sampling, at least for fast decaying kernels. The theory should be

(1) *Constructive*, i.e., a possibly iterative algorithm should allow numerical reconstruction;

(2) *Multidimensional*, so that it can be used in signal and image processing or for the interpolation of sequences of images;

(3) *Local*, so that the value of a band-limited function at a point is essentially determined by the adjacent sampling values, and more distant sampling values have no influence. Estimates with respect to weighted L^p-norms are a suitable tool to describe decay conditions and the locality properties of the reconstruction operators.

(4) *Stable*, so that small perturbations of the parameters cause only small errors in the reconstruction.

1.2. The real analysis approach of this paper. To achieve these objectives we choose a real variable approach. We start with the observation that a band-limited function satisfies a convolution equation of the form $f = f * g$, and then analyze this convolution equation carefully. The main tools are *pointwise estimates* (a) for spline type approximations of smooth functions and (b) for the approximation of a convolution by a weighted sum of translates, see Lemmas 4.1-4.4.

Iterative algorithms arise through repeated application of these approximation operators to the remainder term. The resulting sequence converges to the original signal at a geometric rate. These reconstruction methods satisfy all the requirements stated above.

In contrast to many previous papers on irregular sampling, we do not use methods from analytic function theory. This is one of the reasons why the results extend easily to multivariate irregular sampling. Our techniques allow for a treatment of irregular sampling in a very general class of Banach spaces.

Discretization operators which are dual to the spline operators lead to nonorthogonal series expansions of band-limited signals in terms of translates of a single function. The coefficients depend in a linear way on the expanded function, and the coefficient mapping is continuous with respect to any of the norms under consideration. Similar techniques can be used on nonabelian locally compact groups to derive very general results on atomic decompositions for function spaces [FG1], [FG2], [Gr1].

1.3. Review of the literature. Most papers concerning the irregular sampling theorem deal with signals of finite energy, i.e., work in the *Hilbert space* $L^2(\mathbb{R})$ [DS], [Be1,2], [Wi]. In general, such arguments are not applicable to the important class of almost periodic functions or even trigonometric polynomials. Unweighted L^p-spaces are treated in [Go] for regular sampling and in the important but nonconstructive paper by Beurling [B]. Only in a few cases [SA], [BH2], [PM] are two-dimensional problems discussed, however, with additional conditions on the sampling set.

In the mathematical literature very strong uniqueness theorems are found for band-limited functions, both in one dimension [BM], [Wa] and in higher dimensions [La], [B]. These results, which use complex analysis and operator theory, are of high theoretical interest, but they have had no practical implications because they are not constructive.

For small deviations from regular sampling the perturbation theory of orthonormal bases in Hilbert spaces yields irregular sampling theorems [Hi1], [Ye], [Ka], [Yo], [Ra1,2], [BH1,2]. In principle, these methods are constructive, but they seem to be too difficult to use for numerical computations. For instance, a reconstruction through a Lagrange type interpolation involves functions which are given as infinite products. A reconstruction by means of the biorthogonal system requires that these functions be computed first. Besides the computational complexity of this task, it is also known that the biorthogonal system depends on the sampling set in an unstable way, so that a small change of a single sampling point affects all functions of the biorthogonal system in an unpredictable way [Sp]. Moreover, sampling theorems of this kind are restricted to the Hilbert space $L^2(\mathbb{R}^m)$, and they cannot treat sampling sets with strong variations of the local density.

The results in [CPL], [So] are based on the idea of transforming the irregular sampling set into a regular one. They give exact reconstruction, but only for certain classes of functions which are not band-limited and which depend heavily on the sampling geometry in a nontransparent way.

For numerical computations *iterative methods* are most useful for recovering a band-limited function from irregularly sampled values [Wi], [SA], [MA]. They are all derived from the fundamental paper of Duffin and Shaeffer [DS] on nonharmonic Fourier series and a theorem in [Sa]. The convergence of these methods is known only for the Hilbert space L^2. The description of convergence therefore lacks the much desired locality.

The numerical implementation of [SA] is quite successful in image restoration, although for their third method no proof of convergence is known.

1.4. A short overview. The plan of this paper is as follows: Section 2 begins with a description of a family of Banach spaces of functions and measures on $\mathbb{R}^m$, which will allow us to describe convergence of the iterative methods with respect to a variety of norms. Together with these spaces, suitable operators such as spline type approximations and approximations through discrete measures are introduced. They are not well defined for L^p-spaces, but bounded on the auxiliary spaces introduced in this section. These concepts are also crucial in order to adapt the approach described in [F3] for L^1-spaces to $L^2(\mathbb{R}^m)$ or to weighted L^p-spaces. The main results of this paper are stated in § 3. In contrast to known results, we need not assume a positive minimal distance of the sampling points and can thus treat local variations of the density. For numerical applications this means that all information in regions of high sampling density can be used. For L^p-spaces and with positive minimal distance between the sampling points, the results allow a much simpler and more accessible formulation. Therefore we state them explicitly as corollaries. Section 4 contains the technical parts and the proof of the results. The underlying estimates are formulated in a series of lemmas.

2. Function spaces and operators. First let us fix some notation. We denote the space of all Radon measures (regular Borel measures on $\mathbb{R}^m$) by $R(\mathbb{R}^m)$. By the Riesz representation theorem we identify it with the topological dual of $\mathcal{K}(\mathbb{R}^m)$: $\mathcal{K}(\mathbb{R}^m):=\{k|k$ continuous, complex-valued, with compact support$\}$. A point measure δ_x is characterized by $\delta_x(f)=\int f(y)\,d\delta_x(y):=f(x)$. For the uniform norm we use the following symbol: $\|f\|_\infty:=\sup_{z\in\mathbb{R}^m}|f(z)|$.

Submultiplicative weight functions, i.e., continuous functions satisfying $w(x)\geqq 1$ for all $x\in\mathbb{R}^m$ and $w(x+y)\leqq w(x)w(y)$ for all $x,y\in\mathbb{R}^m$, are important because the weighted L^1-spaces $L^1_w(\mathbb{R}^m):=\{f|fw\in L^1(\mathbb{R}^m)\}$, with the norm $\|f\|_{1,w}:=\int_{\mathbb{R}^m}|f(x)|w(x)\,dx$, are Banach algebras under convolution (cf. [Rei], [F1] for details)

$$(2.1)\qquad f*g(x):=\int_{\mathbb{R}^m}f(x-y)g(y)\,dy \quad \text{for } f,g\in L^1_w(\mathbb{R}^m).$$

If $w=w_a:y\to(1+|y|)^a$, for some $a\geqq 0$ we write $L^1_a(\mathbb{R}^m)$ instead of $L^1_w(\mathbb{R}^m)$.

We shall describe our approach in the setting of solid BF-spaces (or Banach lattices). Formally we make the following general assumptions:

(B1) $(B,\|\ \|_B)\hookrightarrow L^1_{\text{loc}}(\mathbb{R}^m)$ is a Banach space of locally integrable functions on $\mathbb{R}^m$, and for any compact set $Q\subseteq\mathbb{R}^m$ there exists $C_Q>0$ such that

$$\int_Q|f(x)|\,dx\leqq C_Q\|f\|_B \quad \text{for all } f\in B.$$

(B2) $(B,\|\ \|_B)$ is a Banach module over $(C^0(\mathbb{R}^m),\|\ \|_\infty)$ with respect to pointwise multiplication, i.e., $hf\in B$ and $\|hf\|_B\leqq\|h\|_\infty\|f\|_B$ for $h\in C^0$ and $f\in B$.

(B3) $(B,\|\ \|_B)$ is translation invariant in the following sense: The translation operators L_y, $y\in\mathbb{R}^m$, given by $L_yf(x):=f(x-y)$, map B into itself and for some $a\geqq 0$ we have $\|L_yf\|_B\leqq C_B(1+|y|)^a\|f\|_B$ for all $f\in B$.

(B3′) $(B,\|\ \|_B)$ is a Banach convolution module over $L^1_a(\mathbb{R}^m)$, i.e., we assume that, for $f\in B$ and $g\in L^1_a$, $g*f\in B$ and $\|g*f\|_B\leqq C_B\|g\|_{1,a}\|f\|_B$.

Remark 2.1. It follows from (B3) that B is a space of tempered distributions, i.e., $(B,\|\ \|_B)\hookrightarrow\mathcal{S}'(\mathbb{R}^m)$. Consequently, the Fourier transform $\hat{f}$ and the spectrum $\operatorname{spec} f:=\operatorname{supp}\hat{f}$ (equals the support of $\hat{f}$) are well defined for $f\in B$. For any closed set $\Omega\subseteq\mathbb{R}^m$ the set

$$B^\Omega:=\{f\in B, \operatorname{spec} f\subseteq\Omega\}$$

is a closed subspace of B.

Examples. The most natural examples satisfying (B1)–(B3′) are the spaces

$$L^p_v:=\{f|fv\in L^p(\mathbb{R}^m)\},\quad \text{with norm } \|f\|_{v,p}:=\left(\int_{\mathbb{R}^m}|f(x)v(x)|^p\,dx\right)^{1/p}$$

for $1\leqq p<\infty$ (and a sup norm for $p=\infty$); cf. [F1]. The function v is assumed to be a continuous and positive function which is *moderate* with respect to the *weight function* w_a, i.e., it satisfies $v(x+y)\leqq C_1w_a(x)v(y)$ for all $x,y\in\mathbb{R}^m$. Note that obviously w_b is a moderate function (with respect to w_a) for any $b\in[-a,a]$. Our general approach also includes (weighted) mixed norm spaces in the sense of Benedek–Panzone, or weighted variants of rearrangement invariant Banach spaces such as Lorentz or Orlicz spaces [LT] or spaces of bounded p-mean [F5].

In order to extend the results stated in [F3] to L^p-spaces (with $p>1$) the Wiener type spaces $W(M, B)$ and $W(C^0, B)$ are an important tool [F2]. For $B=L^p$ they coincide with amalgam spaces in the sense of [FSt]. We shall describe them briefly, using the symbols MB and CB, and give a new simple proof of a convolution theorem.

DEFINITION 2.1. For a fixed open, bounded subset $Q\subseteq \mathbb{R}^m$ we define the *local maximal function* $x \mapsto f^{\#}(x)$ by $f^{\#}(x):=\sup_{z\in Q+x}|f(z)|$. Then

$$CB := \{f \mid f \text{ continuous}, f^{\#}\in B\}. \tag{2.2}$$

defines a Banach space with the norm

$$\|f\|_{CB} := \|f^{\#}\|_B. \tag{2.3}$$

We shall denote the space $CL^1_a(\mathbb{R}^m)$ by $C^1_a(\mathbb{R}^m)$. Note that the Schwartz space $\mathcal{S}(\mathbb{R}^m)$ is continuously embedded into these spaces: $\mathcal{S}(\mathbb{R}^m)\hookrightarrow C^1_a(\mathbb{R}^m)$ for any $a\in\mathbb{R}$. The second space associated with B in a natural way allows us to deal with discrete measures with a certain global behavior.

DEFINITION 2.2. For Q as above we set

$$MB = \{\mu\in R(\mathbb{R}^m), \text{with } q_\mu : x\mapsto |\mu|(x+Q)\in B\}, \tag{2.4}$$

$$\|\mu\|_{MB} := \|q_\mu\|_B. \tag{2.5}$$

For a discrete set $X=(x_i)_{i\in I}$ in $\mathbb{R}^m$ we shall write MB_X for the closed subspace $\{\mu=\sum_{i\in I}\lambda_i\delta_{x_i}, \mu\in MB\}$ of measures in MB supported on X.

In the last two definitions different bounded sets Q_1, Q_2 generate the same space and equivalent norms; hence $g\in CB$ if and only if $g^{\#\#}\in B$. For any positive $k\in\mathcal{K}(\mathbb{R}^m)$, $\||\mu|*k\|_B$ is also an equivalent norm for MB. These facts are used in the sequel (e.g., for (2.9) below) without notice.

Next we describe basic properties of these spaces.

THEOREM 2.1. *Let* $(B, \| \|_B)$ *satisfy the general assumptions* (B1)-(B3′). *Then the following hold*:

(i) *CB and MB are Banach spaces satisfying properties* (B1)-(B3′), *with the following continuous embeddings*:

$$CB \hookrightarrow B \hookrightarrow MB. \tag{2.6}$$

(ii) $MB * C^1_a \subseteq CB$, *and* $\|\mu*g\|_{CB}\leqq C_0\|\mu\|_{MB}\cdot\|g\|_{C^1_a}$ *for all* $\mu\in MB$ *and* $g\in C^1_a$.

(iii) *The spaces* MB^Ω, B^Ω, *and* CB^Ω *coincide, and the respective norms are equivalent for any compact subset* $\Omega\subseteq\mathbb{R}^m$, *i.e., any* $\mu\in MB^\Omega$ *is represented by a function* $f\in CB$, *and there exists a constant* $C_\Omega>0$ *such that* $\|f\|_{CB}\leqq C_\Omega\|f\|_B\leqq C_\Omega\|f\|_{MB}$ *for all* $f\in MB^\Omega$.

(iv) *If* $\mathcal{K}(\mathbb{R}^m)$ *is dense in B, then it is dense in CB.*

Remark 2.2. In view of the examples above we assume that (2.6) holds in the form

$$\|f\|_{MB}\leqq\|f\|_B\leqq\|f\|_{CB} \quad \text{for all } f\in CB.$$

Proof of Theorem 2.1. The verification of (i) and (iv) is left to the interested reader; cf. [F2]. For (ii) we have to verify $(\mu*f)^{\#}\in B$. By direct computation the following two pointwise estimates can be obtained:

$$(\mu*f)^{\#}\leqq|\mu|*f^{\#}, \tag{2.7}$$

and for any $k\in\mathcal{K}(\mathbb{R}^m)$, $k\geqq0$, with supp $k\subseteq Q=-Q$ and $\int_{\mathbb{R}^m}k(y)\,dy=1$,

$$h^{\#}*k\geqq|h|. \tag{2.8}$$

Combining these two estimates we obtain

$$(2.9) \qquad (\mu * g)^{\#} \leqq |\mu| * g^{\#} \leqq |\mu| * (k * g^{\#\#}) = (|\mu| * k) * g^{\#\#} \in B * L^1_a \subseteq B$$

whenever $\mu \in MB$ and $g^{\#\#} \in L^1_a$, i.e., $g \in C^1_a$, and the proof of (ii) is complete.

Since spec $\mu \subseteq \Omega$, $\mu = \mu * g$ holds for any $g \in C^1_a$ with $\hat{g}(t) \equiv 1$ on Ω. Consequently, $\mu = \mu * g \in MB * C^1_a \subseteq CB$ by (ii). This also gives the required estimate (cf. Remark 2.2)

$$(2.10) \quad \|\mu\|_{MB} \leqq \|\mu\|_B \leqq \|\mu\|_{CB} = \|\mu * g\|_{CB} \leqq \|(\mu * g)^{\#}\|_B \leqq C_B \|\mu\|_{MB} \|g^{\#}\|_{1,a}. \qquad \square$$

Besides these *convolution relations* two special types of *operators* on *CB* and *MB* are needed. Both involve so-called δ-PUs.

Definition 2.3. We call a family of nonnegative (measurable) functions $\Psi = (\psi_i)_{i \in I}$ a *δ-partition of unity* (δ-PU for short) if the following is satisfied: $\sum_{i \in I} \psi_i(x) \equiv 1$ on $\mathbb{R}^m$, and supp $\psi_i \subseteq K_\delta(x_i)$ for $i \in I$ for some discrete family $X = (x_i)_{i \in I}$ in $\mathbb{R}^m$. Here $K_\delta(x)$ denotes the open ball of radius δ centered at x. We shall use the symbol $|\Psi|$ for the infimum over all numbers δ such that Ψ is a δ-PU.

Note that X has to be *δ-dense* in $\mathbb{R}^m$, i.e., $\mathbb{R}^m = \bigcup_{i \in I} K_\delta(x_i)$ for any δ-PU Ψ.

Examples of δ-partitions of unity. (a) If X is any δ-dense family and $(P_i)_{i \in I}$ is a partition of $\mathbb{R}^m$ such that $P_i \subseteq K_\delta(x_i)$, then the family of indicator functions $(c_{P_i})_{i \in I}$ is a δ-PU. The use of Voronoi regions is most natural for our task (cf. [FG3]).

(b) If $\psi_0(x) = 1 - |x|/\delta$ for $|x| \leqq \delta$ and $\psi_0(x) = 0$ for $|x| > \delta$, then $L_{\delta n}\psi_0$, $n \in \mathbb{Z}$ is a continuous δ-partition for $\mathbb{R}$.

(c) For irregular sequences in $\mathbb{R}$ we can take triangular functions with $\psi_i(x_i) = 1$ and supported by $[x_{i-1}, x_{i+1}]$. The natural analogue for $\mathbb{R}^2$ can be described as follows. Starting with a triangulation induced from the set X, choose the functions to satisfy $\psi_i(x_i) = 1$ and to be piecewise linear over the triangles having x_i as a vertex; cf. [SA].

(d) For any δ-dense family X in $\mathbb{R}^m$ we can find smooth δ-PUs. In the regular case smooth PUs in $\mathbb{R}^m$ of the form $L_{\delta n}\psi_0$, $n \in \mathbb{Z}^m$, can be obtained using B-splines.

Remark 2.3. In many cases it is convenient to work with families $X = (x_i)_{i \in I}$ which are *well spread* in the following sense: X is δ-dense and *relatively separated* in $\mathbb{R}^m$, i.e., X is a *finite union* of subfamilies, such that $|x_i - x_j| \geqq \delta_0 > 0$ for all $i \neq j$ in the same subfamily, for some $\delta_0 > 0$. In this case the covering through the balls $K_\delta(x_i)$ is of finite height. Of course a sequence is relatively separated in $\mathbb{R}$ if $|x_n - n| \leqq C$ for all $n \in \mathbb{N}$ (cf. [BH2] for a two-dimensional version).

Remark 2.4. If the family Y is well spread and v is a moderate function, then $\mu = \sum_{j \in J} c_j \delta_{y_j} \in (ML^p_v)_Y$ if and only if $(\sum_{j \in J} |c_j|^p v(y_j)^p)^{1/p} < \infty$. For well-spread families it is also easy to check that $\sum_{j \in J} |f(y_j)|^p v(y_j)^p \leqq C^p_Y \|f\|^p_{CL^p_v}$ for all $f \in CL^p_v$. The estimate given in Theorem 2.1(iii) is thus a generalization of Nikolskij's inequality to general p and m dimensions (cf. [Ni, pp. 123–125], or [BH2, Thm. 1] for special cases with $p = 2$).

Using δ-PUs we define the following operators.

Definition 2.4. Given a δ-PU Ψ associated with a family X we denote by Sp_Ψ (actually we should write $Sp_{\Psi,X}$) the operator defined for continuous f:

$$(2.11) \qquad Sp_\Psi(f) := \sum_{i \in I} f(x_i)\psi_i.$$

We also use the same symbol Sp_Ψ in order to describe a related operator which maps sequences $\Lambda = (\lambda_i)_{i \in I}$ into functions on $\mathbb{R}^m$:

$$Sp_\Psi(\Lambda) := \sum_{i \in I} \lambda_i \psi_i.$$

Remark 2.5. Note that $\|Sp_\Psi(f)\|_\infty \leqq \|f\|_\infty$ for $f \in C^b(\mathbb{R}^m)$, and $Sp_\Psi(f) \in C^b(\mathbb{R}^m)$ if in addition Ψ consists of continuous functions. If Ψ consists of piecewise polynomials of fixed order, then $Sp_\Psi(f)$ is a spline approximation (or quasi interpolant) of f, which explains our notation. If Ψ is a system of triangular functions on $\mathbb{R}$, then $Sp_\Psi f$ is the piecewise linear interpolation of the sampling values of f at X. Basic properties of Sp_Ψ are collected in the following proposition.

Proposition 2.2. *Assume that* $(B, \| \|_B)$ *satisfies* (B1) *and* (B2).

(i) *There is a constant* $C_S > 0$ *such that* $\|Sp_\Psi f\|_B \leqq C_S \|f\|_{CB}$ *for all* $f \in CB$; *The family* Sp_Ψ *of all spline operators with continuous* Ψ *and* $|\Psi| \leqq 1$ *acts uniformly bounded on any space CB.*

(ii) supp $(Sp_\Psi f) \subseteq$ supp $(f) + K_1(0)$; *hence* $Sp_\Psi(\mathcal{K}(\mathbb{R}^m)) \subseteq \mathcal{K}(\mathbb{R}^m)$ *for* Ψ *in* $\mathcal{K}(\mathbb{R}^m)$.

(iii) $Sp_\Psi f \to f$ *with respect to* $\| \|_\infty$ *for* $|\Psi| \to 0$, *for any* $f \in \mathcal{K}(\mathbb{R}^m)$, *and consequently,* $Sp_\Psi f \to f$ *in CB for any* $f \in CB$, *whenever* $\mathcal{K}(\mathbb{R}^m)$ *is dense in B.*

(iv) *For fixed X the operator* Sp_Ψ *may be considered as a bounded operator from* MB_X *into B, i.e., we have*

$$\|Sp_\Psi(\Lambda)\|_B \leqq C \cdot \left\| \sum_{i \in I} \lambda_i \delta_{x_i} \right\|_{MB}.$$

Proof. (i) It is easily verified that $\sup_{z \in x+Q} |Sp_\Psi f(z)| \leqq \sup_{z \in x+Q_1} |f(z)|$ for $Q_1 := Q + K_1(0)$, because supp $\psi_i \subseteq K_1(x_i)$ for all $i \in I$. Taking the B-norm with respect to x on both sides proves both statements of (i). The simple proof of (ii) and (iii) is left to the interested reader. (iv) follows from

$$|Sp_\Psi(\Lambda)(x)| \leqq \sum_{i \in I} |\lambda_i| \psi_i(x) \leqq \sum_{\psi_i(x) \neq 0} |\lambda_i| \leqq q_\mu(x)$$

with $\mu = \sum_{i \in I} \lambda_i \delta_{x_i}$ and $Q := K_1(0)$, if $|\Psi| \leqq 1$, by taking the B-norm on both sides.

Next we introduce an operator which replaces a given function by a discrete measure. In order to avoid confusion with the PU Ψ used above we now write $\Phi = (\varphi_j)_{j \in J}$ for a PU associated with $Y = (y_j)_{j \in J}$ in $\mathbb{R}^m$.

DEFINITION 2.5. The discrete measure obtained from $f \in L^1_{\text{loc}}(\mathbb{R}^m)$ through concentration of mass by means of Φ is denoted by

$$D_\Phi f := \sum_{j \in J} \langle f, \varphi_j \rangle \delta_{y_j},$$

with $\langle f, \varphi_j \rangle = \int_{\mathbb{R}^m} \varphi_j(x) f(x)\, dx$.

The following properties of these operators are of interest to us.

PROPOSITION 2.3. *Assume that* $(B, \| \|_B)$ *satisfies* (B1)-(B3′).

(i) *There is a constant* $C_D > 0$ *such that for all* $f \in B$ *and all* Φ *with* $|\Phi| \leqq 1$,

$$\|D_\Phi f\|_{MB} \leqq C_D \|f\|_B;$$

(ii) *Assume that* $\mathcal{K}(\mathbb{R}^m)$ *is dense in B. Then* $D_\Phi f * h \to f * h$ *in CB, hence in B and uniformly over compact sets, for* $h \in C^1_a$, *as* $|\Phi| \to 0$.

Proof. In order to determine the norm of $D_\Phi f$ in MB we observe that

$$|D_\Phi f|(x+Q) = \sum_{\{j \mid y_j \in x+Q\}} |\langle f, \varphi_j \rangle|$$

$$\leqq \int_{\mathbb{R}^m} \left(\sum_{\{j \mid y_j \in x+Q\}} |f(y) \varphi_j(y)| \right) \leqq |f| * c_{Q_1}(x)$$

where c_{Q_1} is the indicator function of $Q_1 := Q + K_1(0)$. Since $|f| * c_{Q_1} \in B * L^1_a \subseteq B$ by (B3′) we obtain $\|D_\Phi f\|_{MB} \leqq C_1 \|\mu\|_{MB}$. The proof of (ii) is left to the interested reader.

Remark 2.6. Note that both Sp_Ψ and D_Φ are not bounded on B itself, and that the auxiliary spaces CB and MB are essential for our approach.

3. The main results. With the notations of § 2 we now describe the results about reconstruction and series expansions of irregularly sampled band-limited functions. The main theorems will contain two parts. The first part asserts the existence of a reconstruction operator. This formulation reveals the stability of the reconstruction and the correct norm estimates. The second part realizes the reconstruction operator as an iterative procedure. This form emphasizes the algorithmic aspect of the reconstruction. The corollaries make clear that for well-spread families the results can be described using natural Banach spaces of sequences instead of MB.

The first theorem deals with the complete reconstruction of band-limited functions from their sampling values.

THEOREM 3.1 (General sampling theorem for band-limited functions). *Let Ω be a compact subset of $\mathbb{R}^m$. Then there exist $\delta = \delta(\Omega) > 0$ and $C = C(\delta, \Omega) > 0$ such that for any δ-PU Ψ on $\mathbb{R}^m$ there is a bounded operator $\mathbf{B}: B \to B^\Omega$ with*

$$\|\mathbf{B}(\phi)\|_B \leqq C \cdot \|\phi\|_B, \tag{3.1}$$

which inverts Sp_Ψ on B^Ω, i.e., $f = \mathbf{B}(Sp_\Psi f)$ for all $f \in B^\Omega$. Consequently, f can be completely recovered from the sampling values $(f(x_i))_{i\in I}$.

The operator $\mathbf{B}$ can be realized by the following iterative algorithm: Fix a pair of band-limited functions $g, h \in L_a^1$ such that $\hat{g}(t) \equiv 1$ on Ω and $\hat{h}(t) \equiv 1$ on spec g. *Set*

$$\phi_0 := \phi \quad \textit{and} \quad \phi_{k+1} := \phi_k * h - Sp_\Psi(\phi_k * h); \tag{3.2}$$

Then

$$\mathbf{B}(\phi) = \left(\sum_{k=0}^{\infty} \phi_k\right) * g. \tag{3.3}$$

Formula (3.1) expresses the stability of the reconstruction. The algorithm satisfies all the other natural requirements discussed in the introduction. For more precise statements about the locality, see Theorem 3.4 and [FG4].

The operations involved can be easily implemented and our first numerical experiments have shown that the algorithm works efficiently. As has been shown in [Gr2], the required sampling density (at least for the one-dimensional case and a natural choice of Ψ) is just the Nyquist rate.

If $\Omega = [-1, 1]$, $g = h = \text{sinc}$, and Ψ is a system of triangular (or pyramid) type functions on $\mathbb{R}$ or $\mathbb{R}^2$, then the algorithm can be shown to be equivalent to method (2) suggested in [SA] (without proof for the irregular case there). Since $\text{sinc} \notin L^1$, our argument for the convergence of the algorithm cannot be applied directly, but the arguments given in § 4 are easily adjusted to this case.

COROLLARY A. *Given a compact set $\Omega \subseteq \mathbb{R}^m$ there exists $\delta = \delta(\Omega) > 0$ such that for* any *δ-dense, well-spread family X in $\mathbb{R}^m$ the following is true: There exists a bounded linear operator $\mathbf{R}$ from the sequence space $l_v^p := \{\Lambda | (\sum_{i\in I} |\lambda_i|^p v(x_i)^p)^{1/p} < \infty\}$ into $L_v^p(\mathbb{R}^m)$ such that $\mathbf{R}$ provides a complete reconstruction of f from its sampling values, i.e., $\mathbf{R}(f(x_i)_{i\in I}) = f$ for all $f \in (L_v^p)^\Omega$. For given Ω, X the same $\mathbf{R}$ and δ work for all a-moderate weights v and $1 \leqq p \leqq \infty$.*

A "dual" variant of Theorem 3.1 yields series expansions in terms of translates of a single function.

THEOREM 3.2 (Series expansions for band-limited functions). *For any $g \in L_a^1$ with $\hat{g}(s) \neq 0$ on a compact set $\Omega \subseteq \mathbb{R}^m$ there is a positive number $\gamma = \gamma(\Omega, g) > 0$ such that* *for any γ-dense family $Y=(y_j)_{j\in J}$ there is a bounded linear operator* $\mathbf{D}$ *from B^Ω into MB_Y satisfying*

$$f=D(f)*g \quad \text{for all } f\in B^\Omega.$$

Writing $\mathbf{D}(f)=\sum_{j\in J}c_j\delta_{y_j}$, *this means*

$$f=\sum_{j\in J}c_jL_{y_j}g. \tag{3.4}$$

The coefficients are obtained as $c_j:=\sum_{k=0}^\infty\langle\varphi_j,f_k\rangle$, where the sequence $(f_k)_{k=0}^\infty$ is given iteratively (using an auxiliary function g_1) by

$$f_0:=f*g_1 \quad \text{and} \quad f_{k+1}:=(f_k-D_\Phi f_k)*h, \tag{3.5}$$

Remark 3.1. Since the sampling sets in $\mathbb{R}^m$ do not have any natural order, we understand convergence of a series $h=\sum_{i\in I}h_i$ in the following sense: for any exhausting sequence $F_n\subseteq I$ of finite subsets of I, i.e., $F_n\subseteq F_{n+1}$ and $\bigcup_{n=1}^\infty F_n=I$, the sequence of partial sums $\sum_{i\in F_n}h_i$ converges to h. Consequently, these series converge unconditionally, i.e., they converge for any fixed enumeration of I. If $I=\mathbb{Z}^m$, the interpretation as a multiple iterated sum is also admissible.

COROLLARY B. *Under the assumptions of Theorem* 3.2 *for any η-dense family Y in $\mathbb{R}^m$ there is a bounded linear operator $\mathbf{C}:(L_v^p)^\Omega\to l_v^p$ such that $f(x)=\sum_{j\in J}(\mathbf{C}f)_jg(x-y_j)$ holds for every $f\in(L_v^p)^\Omega$. The series converges uniformly over compact sets and for $1\leqq p<\infty$ also in the norm of L^p.*

The next theorem offers an alternative reconstruction algorithm which is easier to implement numerically and computationally less intensive, but the required sampling density for this algorithm may be higher.

THEOREM 3.3 (Method of adaptive weights). *Given $\Omega\subseteq\mathbb{R}^m$ compact, $g\in L_a^1$ with $\hat g(t)\equiv1$ on Ω and $h\in L_a^1$ with $\hat h(t)\equiv1$ on* spec(g), *there exists $\eta=\eta(\Omega,g)>0$ such that $f\in B^\Omega$ can be reconstructed from its sampled values $(f(x_i)_{i\in I})$ on any η-dense family X by the following algorithm: Set $w_i=\int\psi_i(x)\,dx$ and*

$$\begin{aligned}\phi_0&=\sum_{i\in I}f(x_i)\cdot w_i\cdot\delta_{x_i}\in MB_X,\\ \phi_{k+1}&=\phi_k*h-\sum_{i\in I}\phi_k(x_i)\cdot w_i\cdot L_{x_i}h.\end{aligned}\tag{3.6}$$

Then $f=\sum_{n=0}^\infty\phi_n*g$ and the right side depends only on the sampling values $(f(x_i)_{i\in I})$.

Remark 3.2. The proof will show that the partial sums

$$f^{(n)}:=\sum_{i\in I}w_i\sum_{k=0}^n\phi_k(x_i)\cdot L_{x_i}g \tag{3.7}$$

are convergent to f at a geometric rate.

The following variant of Theorem 3.1 is of interest if many functions with the same spectrum are to be reconstructed from samples taken over the same family X or if X and Ω are given in advance.

THEOREM 3.4. *Under the conditions of Theorem* 3.1 *there is a family $(e_i)_{i\in I}$ in C_a^1 such that $f\in B^\Omega$ can be written as $f=\sum_{i\in I}f(x_i)e_i$. The series converges uniformly over compact sets and, if $\mathcal{K}(\mathbb{R}^m)$ is a-dense in B, in the norm of CB.*

Theorem 3.4 expresses the locality of the reconstruction. In contrast to the sinc-functions in the classical cardinal series the e_i's have much better decay properties. In applications the collection of functions $(e_i)_{i\in I}$ may be calculated in advance, using only the knowledge of Ω and the sampling set X. Given the sampling values $(f(x_i))_{i\in I}$, the reconstruction of f is then obtained quickly by ordinary summation.

Remark 3.3. The representation of f as a series in $(e_i)_{i\in I}$ is not unique. In contrast in Kadec's $\frac{1}{4}$-theorem the functions $(e_i)_{i\in I}$ are not linearly independent in general and thus do not constitute a basis. On the other hand, these series expansions work simultaneously for all p, $1\leqq p<\infty$. The nonuniqueness of the expansion is closely related to numerical stability.

In our last theorem we combine the two aspects of the regular sampling theorem and show how for a given family Y a suitable sequence of coefficients for a series expansion of the form (3.4) can be computed directly from the sampled values of f alone.

THEOREM 3.5 (Combined sampling and expansion). *Given $g\in L^1_a$ with $\hat{g}(x)\equiv 1$ on a compact set $\Omega\subseteq\mathbb{R}^m$, there exist two constants $\delta=\delta(\Omega, g)>0$ and $\gamma=\gamma(\Omega, g)>0$ such that for any two families $(x_i)_{i\in I}$ and $(y_j)_{j\in J}$ which are δ and γ-dense, respectively, there is a linear mapping* $\mathbf{M}$ *from the space of sampling values $\{(f(x_i)_{i\in I}), f\in B^{\Omega}\}$ into MB_Y satisfying*

$$f=\mathbf{M}(f(x_i))*g=\sum_{j\in J} c_jL_{y_j}g. \tag{3.8}$$

The coefficients can be obtained from a sequence defined iteratively by

$$f_0:=f, f_{k+1}:=f_k*h-(D_{\Phi}Sp_{\Psi}f_k)*h, \tag{3.9}$$

through

$$c_j:=\sum_{k=0}^{\infty}\langle\phi_j, Sp_{\Psi}f_k\rangle.$$

COROLLARY C. *If in the situation of Theorem 3.5 the sets X and Y are well spread, there exists a bounded linear operator* $\mathbf{N}$ *from $l^p_v(I)$ into $l^p_v(J)$ such that for $f\in(L^p_v)^{\Omega}$,*

$$f(x)=\sum_{j\in J}\mathbf{N}(f(x_i)_{i\in I})_j g(x-y_j),$$

with convergence in L^p_v for $1\leqq p<\infty$.

4. Proofs.

LEMMA 4.1. *For any compact subset $\Xi\subseteq\mathbb{R}^m$ there exists some constant $C^1_{\Xi}=C(\Xi, a)>0$ such that uniformly for all δ-PUs Ψ and spaces B satisfying* (B1)–(B3′)

$$\|f-Sp_{\Psi}f\|_B\leqq C^1_{\Xi}\cdot\|f\|_B\cdot\delta \quad \textit{for all } f\in B^{\Xi}.$$

Proof. We discuss the one-dimensional case first and check that the condition $\operatorname{spec} f\subseteq\Xi$ implies $f'\in CB$. Since $\hat{f'}(t)=2\pi it\hat{f}(t)$ for all $t\in\mathbb{R}$ it is sufficient to choose some $u\in\mathcal{S}(\mathbb{R})$ such that $\hat{u}$ is in $\mathcal{D}$ (infinitely differentiable with compact support) and satisfies $\hat{u}(t)=2\pi it$ on Ξ. Then $f'=f*u$, and since $\mathcal{S}(\mathbb{R}^m)\subseteq C^1_a$ this implies $f'\in CB$, and by Theorem 2.1(ii),

$$\|f'\|_{CB}\leqq C_0\|f\|_B\|u\|_{C^1_a}. \tag{4.1}$$

The mean value theorem implies

$$|f(x)-f(y)|\leqq\sup_{z\in K_\delta(x)}|f'(z)|\cdot\delta \quad \text{for } x, y\in\mathbb{R}^m \quad \text{with } |x-y|\leqq\delta.$$

It follows therefore that (note that $\operatorname{supp}\psi_i\subseteq K_\delta(x_i)$)

$$|(f(x)-f(x_i))\psi_i(x)|\leqq\delta\cdot\sup_{z\in K_\delta(x)}|f'(z)|\cdot\Psi_i(x). \tag{4.2}$$

By summation over $i \in I$, using the properties of a δ-PU we obtain

$$|(f - Sp_\Psi f)(x)| \leqq \sum_{i \in I} |(f(x) - f(x_i))\psi_i(x)| \leqq \delta \cdot \sup_{z \in K_\delta(x)} |f'(z)|,$$

and, after taking B-norms on both sides,

$$\|Sp_\Psi f - f\|_B \leqq \delta \cdot \|f'\|_{CB} \leqq \delta \cdot C^1_\Xi \|f\|_B, \tag{4.3}$$

for $C^1_\Xi := C_0 \cdot \|u\|_{C^1_a}$. This completes the proof in the one-dimensional case.

In the case $m \geqq 2$ we use that the partial derivatives on B^Ξ can be represented as convolution operators, i.e., $\partial f/\partial x_k = f * u_k$ for all $f \in B^\Xi$ if $u_k \in \mathscr{S}(\mathbb{R}^m)$ satisfies $\hat{u}_k(t) = 2\pi i t_k$ on Ξ. Thus

$$|\operatorname{grad} f(x)| = \left(\sum_{k=1}^m (\partial f/\partial x_k(x))^2 \right)^{1/2} \leqq \sum_{k=1}^m |\partial f/\partial x_k(x)| = \sum_{k=1}^m |f * u_k(x)|.$$

Setting $u(x) := \sum_{k=1}^m |u_k(x)|$, we have $u \in C^1_a$ and

$$|\operatorname{grad} f(x)| \leqq (|f| * u)(x) \quad \text{for all } x \in \mathbb{R}^m. \tag{4.4}$$

Invoking the mean value theorem we obtain for any $y \in B_\delta(x)$

$$|f(x) - f(y)| \leqq |x - y| \sup_{z \in K_\delta(x)} |\operatorname{grad} f(z)| \leqq \delta \cdot \sup_{z \in K_\delta(x)} (|f| * u)(z).$$

By Theorem 2.1(ii) we have $|f| * u \in CB$ and the same arguments as above apply (with f' being replaced by $|f| * u$). □

Remark 4.1. Observe that in the above situation we have the inequality

$$|\operatorname{grad} f|^\# \leqq |f| * u^\#. \tag{4.5}$$

Using (2.7) and (4.4) we therefore derive the estimate

$$\|\operatorname{grad} f\|_{CB} \leqq C^2_\Xi \cdot \|f\|_B \quad \text{for} f \in B^\Xi. \tag{4.6}$$

This result is a generalization of Bernstein's inequality from L^p to general function spaces B.

Proof of Theorem 3.1. Given $\Omega \subseteq \mathbb{R}^m$ compact, we fix a pair of band-limited functions $g, h \in L^1_a(\mathbb{R}^n)$ such that $\hat{g}(t) = 1$ on Ω, and $\hat{h}(t) = 1$ on spec g (any pair of Schwartz functions $\hat{g}, \hat{h} \in \mathscr{S}$ satisfying the condition will do). Consequently,

$$h * g = g, \qquad f * g = f = f * g * h \quad \text{for all } f \in B^\Omega. \tag{4.7}$$

In order to define the operator $\mathbf{B}$ on the Banach space $(B, \| \ \|_B)$ we define for $\phi \in B$ a sequence, starting with $\phi_0 = \phi$:

$$\phi_{k+1} := \phi_k * h - Sp_\Psi(\phi_k * h). \tag{4.8}$$

Since h is band-limited the functions $\phi_k * h$ are band-limited and Lemma 4.1 is applicable with $\Xi :=$ spec h. Hence (using (B3'))

$$\|\phi_{k+1}\|_B \leqq \delta \cdot C^1_\Xi \|\phi_k * h\|_B \leqq \delta \cdot C^1_\Xi \cdot C_B \|h\|_{1,a} \|\phi_k\|_B. \tag{4.9}$$

If $|\Psi| \leqq \delta$, such that $\delta \cdot C^1_\Xi C_B \|h\|_{1,a} =: \gamma < 1$, then

$$\|\phi_{k+1}\|_B \leqq \gamma \cdot \|\phi_k\|_B \leqq \gamma^{k+1} \|\phi\|_B \quad \text{for } k \geqq 0. \tag{4.10}$$

It follows that the operator $\mathbf{B}$ given as

$$\mathbf{B}(\phi) = \left(\sum_{n=0}^\infty \phi_n \right) * g \tag{4.11}$$

is well defined on B, with values in B^{Ξ}, and bounded due to the estimate

$$\begin{aligned}\|\mathbf{B}(\phi)\|_B &\leqq C_B\cdot\left\|\sum_{n=0}^{\infty}\phi_n\right\|_B\|g\|_{1,a}\\ &\leqq C_B\cdot\sum_{n=0}^{\infty}\|\phi_n\|_B\|g\|_{1,a}\leqq(1-\gamma)^{-1}C_B\|\phi\|_B\|g\|_{1,a}.\end{aligned}\tag{4.12}$$

In order to verify that $\mathbf{B}$ inverts the spline operator Sp_Ψ over B^Ω we have to consider the sequence $(\phi_k)_{k=0}^{\infty}$, now starting with $\phi_0:=Sp_\Psi f$, for $f\in B^\Omega$. We also use the sequence $(f_k)_{k=0}^{\infty}$, given by recursion (4.8), and starting with $f_0:=f$, because it satisfies the following identity:

$$f=f_{k+1}*g+\left(\sum_{n=0}^{k}Sp_\Psi(f_n*h)\right)*g.\tag{4.13}$$

Equation (4.13) is clear for $k=0$ by (4.8) and follows immediately by induction for $k>0$. Moreover, since $\|f_{k+1}*g\|_B\to 0$ at a geometric rate, (4.13) yields the following representation of f as an absolutely convergent series in B:

$$f=\left(\sum_{n=0}^{\infty}Sp_\Psi(f_n*h)\right)*g.\tag{4.14}$$

It will now be sufficient to verify that

$$Sp_\Psi(f_k*h)*g=\phi_k*g,\tag{4.15}$$

because then (4.14) implies the required identity

$$\mathbf{B}(Sp_\Psi f)=\left(\sum_{k=0}^{\infty}\phi_k\right)*g=\sum_{k=0}^{\infty}Sp_\Psi(f_k*h)*g=f\quad\text{for} f\in B^\Omega.\tag{4.16}$$

In order to verify (4.15) we show first that

$$f_{k+1}=f_k-\phi_k\quad\text{for } k\geqq 0.\tag{4.17}$$

This is clear for $k=0$ by (4.8) and follows for general $k\geqq 1$ by induction:

$$\begin{aligned}f_k-\phi_k&=f_{k-1}*h-Sp_\Psi(f_{k-1}*h)-\phi_{k-1}*h+Sp_\Psi(\phi_{k-1}*h)\\&=(f_{k-1}-\phi_{k-1})*h-Sp_\Psi((f_{k-1}-\phi_{k-1})*h)\\&=f_k*h-Sp_\Psi(f_k*h)=f_{k+1}.\end{aligned}$$

Equation (4.15) is true for $k=0$ and follows from (4.17) by induction for $k\geqq 1$, using (4.8):

$$\begin{aligned}Sp_\Psi(f_{k+1}*h)*g&=Sp_\Psi(f_k*h)*g-Sp_\Psi(\phi_k*h)*g\\&=\phi_k*h*g-Sp_\Psi(\phi_k*h)*g=\phi_{k+1}*g.\end{aligned}$$

The proof of Theorem 3.1 is thus complete. □

Remark 4.2. Note that both Lemma 4.1 and the proof of Theorem 3.1 work simultaneously for all function spaces B that are convolution modules over the same Beurling algebra L_a^1. Thus the same constants arise for all such spaces having the same constant C_B in (B3). This will be important for Theorem 3.4.

Proof of Corollary A. We define the reconstruction operator $\mathbf{R}$ from l_v^p into L_v^p as follows: For $\Lambda=(\lambda_i)_{i\in I}\in l_v^p$ we set $\mathbf{R}(\Lambda):=\mathbf{B}(Sp_\Psi(\Lambda))$, i.e., we form iteratively

$$\phi_0:=\sum_{i\in I}\lambda_i\psi_i\quad\text{and}\quad\phi_{k+1}:=\phi_k*h-Sp_\Psi(\phi_k*h)\quad\text{for } k\geqq 1.\tag{4.18}$$

Then $\mathbf{R}(\Lambda) = (\sum_{k=0}^{\infty} \phi_k) * g$, the series is well defined in L_v^p and $\operatorname{spec} \mathbf{R}(\Lambda) \subseteq \operatorname{spec} g$. A combination of Theorem 3.1, Remark 2.4, and Proposition 2.2(iv) then yields the boundedness of $\mathbf{R}$ as an operator from l_v^p into L_v^p. Evidently, $\mathbf{R}$ applied to the sequence $\Lambda = (f(x_i))_{i\in I}$ yields $\mathbf{R}(\Lambda) = \mathbf{B}(Sp_\Psi f) = f$. Thus $\mathbf{R}$ is indeed a reconstruction operator. □

The following result will play the same role in the proof of Theorem 3.2 as Lemma 4.1 in the proof of Theorem 3.1. The operator D_Φ is as in Proposition 2.3.

LEMMA 4.2 (Discretization of convolution). *For any compact subset $\Xi \subseteq \mathbb{R}^m$ there exists some constant $C_\Xi^3 = C(\Xi, a) > 0$ such that uniformly for all η-PUs Φ and all spaces B satisfying* (B1)-(B3′),

$$\|(f - D_\Phi f) * h\|_{CB} \leqq \eta \cdot C_\Xi^3 \cdot \|f\|_B \cdot \|h\|_{1,a} \tag{4.19}$$

for any band-limited $h \in L_a^1$ with $\operatorname{spec} h \subseteq \Xi$ and $f \in B$.

Proof. Given an η-PU $\Phi = (\varphi_j)_{j\in J}$ we consider for fixed $j \in J$

$$|(f\varphi_j - \langle f, \varphi_j\rangle \delta_{y_j}) * h| \leqq \int_{K_\eta(y_j)} |h(x-y) - h(x-y_j)|\,|f\varphi_j|(y)\,dy =: I. \tag{4.20}$$

Next we observe that the mean value theorem implies for fixed x, y, y_j:

$$|h(x-y) - h(x-y_j)| \leqq \eta \cdot |\operatorname{grad} h(\xi)|,$$

with ξ between $x-y$ and $x-y_j$; hence $\xi \in K_\eta(x-y)$, and we may continue the estimate by

$$I \leqq \int_{K_\eta(y_j)} \eta \cdot |\operatorname{grad} h|^{\#}(x-y)|f\varphi_j|(y)\,dy \leqq \eta \cdot |f\varphi_j| * (\operatorname{grad} h)^{\#}(x).$$

By summation over $j \in J$ we obtain from this the pointwise estimate

$$|(f - D_\Phi f) * h| \leqq \eta \cdot (|f| * (\operatorname{grad} h)^{\#}). \tag{4.21}$$

Since $\operatorname{spec}(f - D_\Phi f) * h \subseteq \operatorname{spec} h = \Xi$, the CB-norm is equivalent to its B-norm by Theorem 2.1(iii) for these functions. Using (B3′) we obtain as in (4.4) (cf. Remark 4.1):

$$\begin{aligned} \|(f - D_\Phi f) * h\|_{CB} &\leqq \|(f - D_\Phi f) * h\|_B \leqq \eta \|\,|f| * |\operatorname{grad} h|^{\#}\|_B \\ &\leqq \eta \cdot C_B \|f\|_B \,\|\,|\operatorname{grad} h|^{\#}\|_{1,a} \leqq \eta \cdot C_B C_\Xi^2 \cdot \|h\|_{1,a} \cdot \|f\|_B. \end{aligned} \tag{4.22}$$

Thus $C_\Xi^3 := C_B \cdot C_\Xi^2$ is an appropriate choice. □

Proof of Theorem 3.2. By the theorem of Wiener-Levy (cf. [Rei, Chaps. 1, 6.5]) there exists a band-limited function $g_1 \in L_a^1$ such that $\hat{g}_1(t) = 1/\hat{g}(t)$ on Ω. Next we choose a band-limited function $h \in \mathcal{S}(\mathbb{R}^m) \subseteq L_a^1$ with $\hat{h}(t) \equiv 1$ on $\operatorname{spec} g \cup \operatorname{spec} g_1$. Then

$$\begin{gathered} g * h = g, \qquad g_1 * h = g_1, \\ f * h = f * g_1 * g = f \quad \text{for all } f \in B^\Omega. \end{gathered} \tag{4.23}$$

We want to show that in the identity $f = (f * g_1) * g$ the factor $f * g_1$ can be replaced by a discrete measure. To this end we start an iteration procedure similar to that of Theorem 3.1, but with a different approximation of the convolution $f \mapsto f * h$. We define

$$f_0 := f * g_1, \qquad f_{k+1} := (f_k - D_\Phi f_k) * h, \tag{4.24}$$

with $D_\Phi\colon B \to MB$ as in Lemma 4.2. It follows by induction that

$$f = f_{k+1} * g + \left(\sum_{n=0}^{k} D_\Phi f_n\right) * g. \tag{4.25}$$

Substituting (4.24) and (4.23) into (4.25), we obtain for $k=0$

$$f_1 * g + (D_\Phi f_0) * g = f * g_1 * h * g - D_\Phi(f * g_1) * h * g$$
$$+ D_\Phi(f * g_1) * g = f.$$

The step from k to $k+1$ is also clear, if we again use $g * h = g$. If we now choose $\eta = \eta(\Omega, g) < 1/(C_\Xi^3 \cdot \|h\|_{1,a})$, or $\gamma := \eta \cdot C_\Xi^3 \|h\|_{1,a} < 1$, then Lemma 4.2 implies for any η-PU Φ

$$\|f_{k+1}\|_B \leqq \gamma \cdot \|f_k\|_B \leqq \gamma^{k+1} \|f_0\|_B \quad \text{for } k \geqq 0. \tag{4.26}$$

Taking the limit as $k \to \infty$ in (4.25) the series representation for f follows:

$$f = \left(\sum_{n=0}^{\infty} D_\Phi(f_n) \right) * g = D_\Phi\left(\sum_{n=0}^{\infty} f_n \right) * g. \tag{4.27}$$

The interchange of brackets is justified by the fact that the series $F := \sum_{n=0}^{\infty} f_n$ is absolutely convergent in B, by the continuity of D_Φ from B into MB, and by the continuity of convolution by $g \in C_a^1$ (Theorem 2.1(ii)). Thus

$$f = (D_\Phi F) * g = \sum_{j \in J} \langle \varphi_j, F \rangle L_{y_j} g \tag{4.28}$$

yields the desired expansion of f in terms of translates of g. The continuous dependence of the measure $D_\Phi F$ follows from Proposition 2.3(i) and (4.26):

$$\begin{aligned} \|D_\Phi F\|_{MB} &\leqq C_D \|F\|_{MB} \leqq C_D \|F\|_B \leqq C_D \cdot \sum_{k=0}^{\infty} \|f_k\|_B \\ &\leqq C_D \left(\sum_{k=0}^{\infty} \gamma^k \right) \|f_0\|_B \leqq C_D C_B (1-\gamma)^{-1} \|g_1\|_{1,a} \|f\|_B. \end{aligned} \tag{4.29}$$

Thus the operator $\mathbf{D}$ is given as $\mathbf{D}(f) := D_\Phi F$. □

Remark 4.3. If $\hat{g}(t) \equiv 1$ on Ω, the auxiliary function g_1 is not necessary and the reconstruction of f is much simpler. The iteration is then $f_0 = f$

$$f_{k+1} = f_k * h - (D_\Phi f_k) * h \tag{4.30}$$

where $\hat{h}(t) \equiv 1$ on spec g. The rest of the argument is the same.

Proof of Corollary B. It is no loss of generality to assume that X is well spread (by selecting a δ-dense and separated subfamily, and setting the coefficient equal to zero for the omitted points). Thus all we have to show is that in (4.29) $\|D_\Psi F\|_{ML_v^p}$ is equivalent to $(\sum_{i \in I} |\langle \psi_i, F \rangle|^p v(x_i))^{1/p}$. This follows from Remark 2.4. □

For the proof of Theorem 3.3 we have to use a different discretization operator (only valid on B^Ω but not defined on all of B), which will take the same role as D_Φ in the proof of Theorem 3.2. For $f \in CB$ we set

$$D_\Psi^+(f) := \sum_{i \in I} \left(\int \psi_i(y)\, dy \right) f(x_i) \delta_{x_i}. \tag{4.31}$$

These operators D_Ψ^+ combines the features of Sp_Ψ and D_Ψ. Since D_Ψ^+ maps CB into MB_X and uses sampling values of $f \in CB$ it can be used as an approximation operator in both Theorems 3.1 and 3.2. Lemma 4.3 provides the necessary estimate.

LEMMA 4.3. *Let* $(B, \| \|_B)$ *satisfy* (B1)-(B3) (*for some* $a > 0$). *Then there exists* $C_{D+} > 0$ (*depending only on* a) *such that for any* δ-PU Ψ,

$$\|D_\Psi(f) - D_\Psi^+(f)\|_{MB} \leqq \delta \cdot C_{D+} \|f\|_B \quad \textit{for all } f \in B^\Omega. \tag{4.32}$$

In particular, for any $\varepsilon > 0$ *there exists* $\delta_0 > 0$ *such that*

$$\|(D_\Psi^+ f) * h - f * h\|_{CB} \leqq \varepsilon \cdot \|f\|_B \tag{4.33}$$

for all $f \in B^\Omega$ *and all* δ-PU Ψ *with* $\delta \leqq \delta_0$.

In the proof we shall use the following a priori estimate.

Remark 4.4. Given complex-valued sequences $(c_i)_{i\in I}$ and $(d_i)_{i\in I}$ satisfying $|c_i|\leqq d_i$ for all $i\in I$ we have

$$\mu=\sum_{i\in I} d_i\delta_{x_i}\in MB \text{ implies } \nu:=\sum_{i\in I} c_i d_{x_i}\in MB \text{ and } \|\nu\|_{MB}\leqq\|\mu\|_{MB}. \tag{4.34}$$

Proof of Lemma 4.3. The mean-value theorem implies for $i\in I$:

$$\int |f(y)-f(x_i)|\psi_i(y)\,dy\leqq\delta\cdot\int |\operatorname{grad} f|^{\#}(y)\psi_i(y)\,dy. \tag{4.35}$$

After summation over $i\in I$ Remark 4.4 implies

$$\begin{aligned}\|D_\Psi f-D_\Psi^+ f\|_{MB}&\leqq\left\|\sum_{i\in I}\left(\int |f(y)-f(x_i)|\psi_i(y)\,dy\right)\delta_{x_i}\right\|_{MB}\\ &\leqq\delta\cdot\|D_\Psi|\operatorname{grad} f|^{\#}\|_{MB}.\end{aligned} \tag{4.36}$$

Applying Proposition 2.3(i) and Remark 4.1 we obtain

$$\begin{aligned}\|D_\Psi|\operatorname{grad} f|^{\#}\|_{MB}&\leqq C_D\cdot\|\,|\operatorname{grad} f|^{\#}\|_{MB}\leqq C_D\|\,|\operatorname{grad} f|^{\#}\|_B\\ &=C_D\|\operatorname{grad} f\|_{CB}\leqq(C_D C_0\|u\|_{C_a^1})\|f\|_B.\end{aligned} \tag{4.37}$$

Thus $\|D_\Psi f-D_\Psi^+ f\|_{MB}\leqq\delta\cdot C_3\|f\|_B$ for all $f\in B^\Omega$. Combining this fact with Lemma 4.2, the proof is complete if we choose $\delta_0\leqq\varepsilon\cdot\min(\frac{1}{2},\frac{1}{2}C_\Xi^2)$. □

Proof of Theorem 3.3. We refer to the proof of Theorem 3.2 for the iterative procedure, indicating only that D_Φ has to be replaced by D_Ψ^+. Furthermore, g_1 may be chosen to be identical to g (cf. Remark 4.3). By using Lemma 4.3 instead of Lemma 4.2 geometric convergence of the sequence $(f_n)_{n=0}^\infty$ can be verified for sufficiently small δ. In analogy to (4.27) we obtain

$$f=D_\Psi^+\left(\sum_{n=0}^\infty f_n\right)*g. \tag{4.38}$$

Proof of Theorem 3.4. We use the operator **B** from Theorem 3.1 in order to define

$$e_i:=\mathbf{B}(\psi_i). \tag{4.39}$$

Since $\psi_i\in L_a^1$, Theorem 3.1 implies $e_i\in L_a^1(\mathbb{R}^n)$ and $\operatorname{spec} e_i\subseteq\Xi$, thus $e_i\in C_a^1$ for $i\in I$ by Theorem 2.1(iii). If $\mathcal{K}(\mathbb{R}^m)$ is dense in B, then obviously

$$\lim_{n\to\infty}\sum_{i\in F_n} f(x_i)\psi_i=\sum_{i\in I} f(x_i)\psi_i$$

in B for any increasing sequence of finite subsets $F_n\subseteq I$, exhausting I. Therefore by the boundedness of **B**

$$\begin{aligned}f=\mathbf{B}(Sp_\Psi f)&=\mathbf{B}\left(\lim_{n\to\infty}\sum_{i\in F_n} f(x_i)\psi_i\right)\\ &=\lim_{n\to\infty}\sum_{i\in F_n} f(x_i)e_i=\sum_{i\in I} f(x_i)e_i,\end{aligned} \tag{4.40}$$

i.e., the series converges in B. Since the family $(e_i)_{i\in I}$ has joint spectrum we may apply Theorem 2.1(iii) to derive convergence in CB, hence uniform convergence over compact sets in $\mathbb{R}^m$. □

The direct method of obtaining suitable coefficients from the sampling values as described in Theorem 3.5 requires the following technical lemma.

LEMMA 4.4 (From sampling values to coefficients). *Given $\Xi\subseteq\mathbb{R}^m$, compact and $\rho>0$ there exist $\delta=\delta(\Xi)$ and $\gamma=\gamma(\Xi,h)>0$ such that*

$$\|f*h-(D_\Phi Sp_\Psi f)*h\|_B\leqq\rho\cdot\|f\|_B \quad \textit{for all } f\in B^\Xi \tag{4.41}$$

and uniformly for all δ-PUs Φ and all γ-PUs Φ.

Proof. We combine Lemmas 4.1 and 4.2 and obtain for $f \in B^{\Xi}$

$$\begin{aligned} &\|f * h - (D_\Phi Sp_\Psi f) * h\|_B \\ &\quad \leqq \|(f - D_\Phi f) * h\|_B + \|D_\Phi(f - Sp_\Psi f) * h\|_B && \text{(by Lemma 4.2)} \\ &\quad \leqq \gamma C_\Xi^3 \|h\|_{1,a} \|f\|_B + C_0 \|D_\Phi(f - Sp_\Psi f)\|_{MB} \cdot \|h\|_{C_a^1} && \text{(by Theorem 2.1(ii))} \\ &\quad \leqq \gamma C_\Xi^3 \|h\|_{1,a} \|f\|_B + C_D C_0 \|f - Sp_\Psi f\|_B \|h\|_{C_a^1} \\ &&& \text{(by Proposition 2.3 and Lemma 4.1)} \\ &\quad \leqq (\gamma C_\Xi^3 \|h\|_{1,a} + C_D C_0 C_\Xi^1 \delta \|h\|_{C_a^1}) \cdot \|f\|_B. \end{aligned}$$

Now choose γ, δ small enough so that the coefficient of $\|f\|_B$ is less than ρ. □

Proof of Theorem 3.5. The proof is similar to those of Theorems 3.1. and 3.2. We choose $h \in L_a^1$ band-limited with $\hat{h}(t) = 1$ on spec g and define

$$f_0 := f, \qquad f_{k+1} := f_k * h - (D_\Phi Sp_\Psi f_k) * h. \tag{4.42}$$

Then $f_k \in B^{\Xi}$ for all k, and if δ and γ are small enough, by Lemma 4.4

$$\|f_{k+1}\|_B \leqq \rho \cdot \|f_k\|_B \tag{4.43}$$

with $\rho < 1$. Since $f = f * g = f * h * g$ and $g = g * h$ we have for $n \geqq 0$

$$f = f * g = f_{n+1} * g + \sum_{k=0}^{n} D_\Phi(Sp_\Psi f_k) * g. \tag{4.44}$$

Since $\rho < 1$ it follows

$$f = \left(\sum_{k=0}^{\infty} D_\Phi(Sp_\Psi f_k) \right) * g. \tag{4.45}$$

Since the series of discrete measures $\sum_{k=0}^{\infty} D_\Phi(Sp_\Psi f_k) = D_\Phi(\sum_{k=0}^{\infty} Sp_\Psi f_k)$ is supported on $Y = (y_j)_{j \in J}$ and absolutely convergent in MB, the result can be written as $\sum_{j \in J} c_j \delta_{y_j}$, with $c_j := \langle \varphi_j, \sum_{k=0}^{\infty} Sp_\Psi f_k \rangle$. This sum is unconditionally convergent in MB, and norm convergent, if $\mathscr{K}(\mathbb{R}^m)$ is dense in B. It follows therefore that $f = \sum_{j \in J} c_j L_{y_j} g$, as was required.

As in (4.29) we verify that the coefficient mapping (it can be described by an infinite matrix) $\mathbf{M}$: $f \mapsto C = (c_j)_{j \in J}$, is continuous. Finally, we show as in the proof of Theorem 3.1 that the f_k's and therefore $(c_j)_{j \in J}$ depend only on the sampling values $(f(x_i))_{i \in I}$. □

Corollary C follows from Theorem 3.3 in the same way that Corollary B follows from Theorem 3.2.

Note added in proof. Since the submission of this manuscript we have obtained further results on the algorithms of Theorems 3.1–3.3: (a) [FG4] contains a detailed error analysis of these algorithms and shows their stability, again in the general function space setting. (b) The results of numerical experiments have been very convincing; see [FGH], [FCH], [FCS]. Particularly, the adaptive weights method described in Theorem 3.5 using the D_Ψ^+-operator has turned to be simple and extremely effective.

Acknowledgment. We thank the referees for their valuable comments which helped to improve this paper.

REFERENCES

[B] A. BEURLING, *Local harmonic analysis with some applications to differential operators*, in Some Recent Advances in the Basic Sciences, Vol. 1, Belfer Graduate School of Science, Annual Science Conference Proc., A. Gelbart, ed., (1962)-(1964), pp. 109-125.

[BM] A. BEURLING AND P. MALLIAVIN, *On the closure of characters and zeros of entire functions*, Acta Math., 118 (1967), pp. 79-95.

[Be1] F. E. Beutler, *Sampling theorems and bases in Hilbert space*, Inform. Control 4 (1961), pp. 97-117.

[Be2] ———, *Error-free recovery of signals from irregularly spaced samples*, SIAM Rev., 8 (1966), pp. 328-335.

[Bu] P. L. Butzer, *A survey of the Whittaker-Shannon sampling theorem and some of its extensions.* J. Math. Res. Exposition, 3 (1983), pp. 185-212.

[BERS] P. L. Butzer, W. Engels, S. Ries, and R. L. Stens, *The Shannon sampling series and the reconstruction of signals in terms of linear, quadratic and cubic splines*, SIAM J. Appl. Math., 46 (1986), pp. 299-323.

[BH1] P. L. Butzer and L. Hinsen, *Reconstruction of bounded signals from pseudoperiodic irregularly spaced samples*, Signal Proc., 17 (1988), pp. 1-17.

[BH2] ———, *Two-dimensional nonuniform sampling expansions—An iterative approach.* I, II, Appl. Anal. 32 (1989), pp. 53-68, 69-85.

[BSS] P. L. Butzer, W. Splettstösser, and R. L. Stens, *The sampling theorem and linear prediction in signal analysis*, Jahresber. Dt. Math.-Verein., 90 (1988), pp. 1-70.

[Ca] L. L. Campbell, *Sampling theorem for the Fourier transform of a distribution with bounded support*, SIAM J. Appl. Math., 16 (1968), pp. 626-636.

[Ce] A. E. Cetin, *Reconstruction of signals from Fourier transform samples*, Signal Processing, 16 (1989), pp. 129-148.

[CPL] J. J. Clark, M. R. Palmer, and P. D. Lawrence, *A Transformation method for reconstruction of functions from nonuniformly spaced samples*, IEEE Trans. ASSP, 33 (1985), pp. 1151-1165.

[Co] A. Cordoba, *La formule sommatoire de Poisson*, C.R. Acad. Sci. Paris, 306 (1988), pp. 373-376.

[DM] D. E. Dudgeon and R. M. Mersereau, *Multidimensional Signal Processing*, Prentice-Hall, Englewood-Cliffs, NJ, 1984.

[DS] R. Duffin and A. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341-366.

[F1] H. G. Feichtinger, *Gewichtsfunktionen auf lokalkompakten Gruppen*, Sitzungsber. Österr. Akad. Wiss., 188 (1979), pp. 451-471.

[F2] ———, *Banach convolution algebras of Wiener's type*, Proc. Conf. "Functions, Series, Operators," Budapest, August 1980, Colloquia Math. Soc. J. Bolyai, North Holland, Amsterdam; Oxford University Press, New York, 1983, pp. 509-524.

[F3] ———, *Discretization of convolution and reconstruction of band-limited functions from irregular sampling*, Progress in Approximation Theory, J. Approx. Theory (1991), pp. 333-345.

[F4] ———, *New results on regular and irregular sampling based on Wiener amalgams*, Proc. Conf. "Function Spaces," Edwardsville, IL, April 1990.

[F5] ———, *An elementary approach to Wiener's third Tauberian theorem on the Euclidean n-space*; (Proc. Conf. Cortona 1984), Symposia Mathematica, Analisi Armonica, XXIX (1988), pp. 267-301.

[FCH] H. G. Feichtinger, C. Cenker, and H. Hermann, *Iterative methods in irregular sampling: A first comparison of methods*, ICCCP-91, March 1991, Phoenix, AZ, IEEE Comp. Soc. Press, pp. 483-489.

[FCS] H. G. Feichtinger, C. Cenker, and H. Steier, *Fast iterative and non-iterative reconstruction of band-limited functions from irregular sampling values*, ICASSP-91, Toronto, May 1991.

[FG1] H. G. Feichtinger and K. H. Gröchenig, *A unified approach to atomic characterizations via integrable group representations*, Proc. Conf. Lund June 1986, Lecture Notes in Math. 1302 (1988), pp. 52-73.

[FG2] ———, *Banach spaces related to integrable group representations and their atomic decompositions*, I, J. Funct. Anal., 86 (1989), pp. 307-340.

[FG3] ———, *Multidimensional sampling of band-limited functions in L^p-spaces*, Proc. Conf. Oberwolfach, Feb. 1989, ISNM 90, Birkhäuser, 1989, pp. 135-142.

[FG4] ———, *Error analysis in regular and irregular sampling theory*, Appl. Math., to appear.

[FGH] H. G. Feichtinger, K. H. Gröchenig, and M. Hermann, *Iterative methods in irregular sampling theory, numerical results.* 7. Aachener Symposium für Signaltheorie. ASST 90, Sept. 1990, Informatik Fachber. 253, Springer-Verlag, Berlin, 1990, pp. 160-166.

[FSt] J. J. Fournier and J. Stewart, *Amalgams*, Bull. Amer. Math. Soc., 13 (1987), pp. 1-21.

[Go] R. P. Gosselin, *On the L^p-theory of cardinal series*, Ann. Math., 78 (1963), pp. 567-581.

[Gr1] K. Gröchenig, *Describing functions: atomic decompositions versus frames*, Monatsh. Math., 112 (1991), pp. 1-42.

[Gr2] ———, *Reconstruction algorithms in irregular sampling theory*, Math. Comp. (1992), to appear.

[Hi1] J. R. Higgins, *A sampling theorem for irregularly spaced sample points*, IEEE Trans. Inform. Theory, 22 (1976), pp. 621-622.

[Hi2] ———, *Five short stories about the cardinal series*, Bull. Amer. Math. Soc., 12 (1985), pp. 45-89.

[Je] A. J. JERRI, *The Shannon sampling theorem—its various extensions and applications, a tutorial review.* Proc. IEEE, 65 (1977), pp. 1565-1596.

[Ka] M. I. KADEC, *The exact value of the Paley-Wiener constant*, Soviet Math. Dokl., 5 (1964), pp. 559-561.

[La] H. J. LANDAU, *Necessary density conditions for sampling and interpolation of certain entire functions*, Acta Math., 117 (1967), pp. 37-52.

[LT] J. LINDENSTRAUSS AND L. TZAFRIRI, *Classical Banach spaces*, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

[Ma] F. MARVASTI, *A unified approach to zero-crossings and non-uniform sampling of single and multidimensional signals and systems*, Nonuniform, Oak Park, IL, 1987.

[MA] F. MARVASTI AND M. ANALOUI, *Recovery of signals from nonuniform samples using iterative methods*, Proc. Internat. Symposium Circuits Systems, Portland, OR, May 1989.

[Me] R. M. MERSEREAU, *The processing of hexagonally sampled two-dimensional signals*, Proc. IEEE, 67 (1979), pp. 930-949.

[Ni] S. M. NIKOL'SKIJ, *Approximation of Functions of Several Variables and Embedding Theorems*, Springer-Verlag, Berlin, New York, Heidelberg, 1975.

[PK] S. X. PAN AND A. C. KAK, *A computational study of reconstruction algorithms for diffraction tomography: Interpolation versus filtered backpropagation*, Trans. IEEE, Acoust. Speech Signal Process. 31 (1983), pp. 1262-1275.

[Pa] A. PAPOULIS, *Signal Analysis*, McGraw-Hill, New York, 1984.

[PM] D. P. PETERSEN AND D. MIDDLETON, *Sampling and reconstruction of wave number-limited functions in N-dimensional Euclidean space*, Inform. and Control, 5 (1962), pp. 279-323.

[Ra1] M. D. RAWN, *On nonuniform sampling expansions, using entire interpolating functions, and on the stability of Bessel-type sampling expansions*, IEEE Trans. Inform. Theory, 35 (1989), pp. 549-557.

[Ra2] M. D. RAWN, *A stable nonuniform sampling expansion involving derivatives*, IEEE Trans. Inform. Theory, 35, (1989), pp. 1223-1227.

[Rei] H. REITER, *Classical Harmonic Analysis and Locally Compact Groups*, Oxford University Press, London, 1968.

[Sa] L. W. SANDBERG, *On the properties of some systems that distort signal*, I, Bell. Systems. Tech. J., 42 (1963), pp. 2033-2047.

[SA] J. P. ALLEBACH AND K. D. SAUER, *Iterative reconstruction of band-limited images from nonuniformly spaced samples*, IEEE Trans. Circuits and Systems, 34 (1987), pp. 1497-1506.

[Sch] I. J. SCHÖNBERG, *Contribution to the problem of approximation of equidistant data by analytic functions. Part* A: *On the problem of smoothing or graduation. A first class of analytic approximation formulae*, Quart. Appl. Math., 4 (1946), pp. 746-756.

[Se] K. SEIP, *An irregular sampling theorem for functions bandlimited in a generalized sense*, SIAM J. Appl. Math., 47 (1987), pp. 1112-1116.

[So] M. SOUMEKH, *Band-limited interpolation from unevenly spaced sampled data*, IEEE Trans. Acoust. Speech Signal Process, 36 (1988), pp. 110-122.

[Sp] W. SPLETTSTÖSSER, *Unregelmäßige Abtastungen determinierter und zufälliger Signale*, in Kolloqium DFG-Schwerpunktprogramm Digitale Signalverarbeitung, H. G. Zimmer and V. Neuhoff, eds., Göttingen, Germany, 1981, pp. 1-4.

[St] H. STARK, ED., *Image Recovery, Theory and Applications*, Academic Press, New York, 1987.

[Wa] W. J. WALKER, *The separation of zeros for entire functions of exponential type*, J. Math. Anal. Appl., 122 (1987), 257-259.

[Wi] R. G. WILEY, *Recovery of bandlimited signals from unequally spaced samples*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 694-702.

[Ye] J. L. YEN, *On the non-uniform sampling of band-width limited signals*, IRE Trans. Circuit Theory, 3 (1956), pp. 251-257.

[Yo] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 23, No. 1, pp. 262–285, January 1992

ASYMPTOTICS OF THE SWALLOWTAIL INTEGRAL NEAR THE CUSP OF THE CAUSTIC*

D. KAMINSKI†

Abstract. The asymptotic behaviour of the "swallowtail integral" is examined, a generalized Airy function, given by

$$S(x,y,z) = \int_{-\infty}^{+\infty} \exp\{i(t^5/5 + xt^3/3 + yt^2/2 + zt)\}dt,$$

for large values of its parameters. In particular, its caustic is briefly discussed, and asymptotic expansions of S are obtained which are uniformly valid near cusps of the caustic as $|x|+|y|+|z| \to +\infty$.

In obtaining the asymptotics of S, the quartic transformation $f(t) = z^4/4 - \zeta z^2/2 + \eta z + \theta$ is used. Exact expressions for the parameters in this transformation are obtained, displaying ζ, η, and θ in terms of the known function f at its critical points.

Key words. uniform asymptotic expansions, swallowtail integral, diffraction integrals, caustics

AMS(MOS) subject classifications. 41A60, 30E15, 33A70

1. Introduction.

The swallowtail integral is a function of three real variables defined by

$$S(x,y,z) = \int_{-\infty}^{+\infty} e^{i(t^5/5+xt^3/3+yt^2/2+zt)}dt. \tag{1.1}$$

By deforming the path of integration into the complex t-plane so that it begins at $\infty e^{9\pi i/10}$ and ends at $\infty e^{\pi i/10}$, we see that S can be extended to an entire function in $\mathbf{C}^3$. The swallowtail integral occupies important niches in several fields of mathematics and physics—in physics, S appears in both geometric optics and applications of catastrophe theory, paralleling the role played by Airy's integral (see [Gil]) and the Pearcey integral (see [Bri] and [Pea]).

The swallowtail integral arises when considering the large λ asymptotics of integrals of the form

$$G(\lambda;\alpha) = \int_{-\infty}^{+\infty} g(z;\alpha)e^{i\lambda f(z;\alpha)}dz, \tag{1.2}$$

where g and f are typically analytic, and $\alpha = (\alpha_1, \cdots, \alpha_n)$ is a collection of auxiliary parameters varying in some set A. If the saddle points of the integral (i.e., critical points of the phase function f) are all simple, then the asymptotics of (1.2) is given as a sum of terms each of order $\lambda^{-1/2}$. In the event that two simple saddles undergo confluence as $\alpha \to \alpha_0$, then the uniform asymptotic behaviour of (1.2) contains terms involving the Airy function and its derivative multiplied by powers of $\lambda^{-1/3}$. If three simple saddles coalesce as $\alpha \to \alpha_0$, then the uniform asymptotic behaviour of (1.2) can be described by terms containing the Pearcey function and its first-order derivatives,

* Received by the editors September 11, 1990; accepted for publication (in revised form) June 19, 1991. This work was partially funded by a Natural Sciences and Engineering Research Council of Canada operating grant, and by a Natural Sciences and Engineering Research Council of Canada postgraduate scholarship while the author was at the University of Manitoba.

† Department of Mathematics and Computer Science, University of Lethbridge, 4401 University Drive, Lethbridge, Alberta, Canada T1K 3M4 (kaminski@hg.uleth.ca).

each multiplied by powers of $\lambda^{-1/4}$, where the Pearcey function is given by

$$P(x,y) = \int_{-\infty}^{+\infty} e^{i(t^4/4+xt^2/2+yt)}dt. \tag{1.3}$$

The swallowtail integral enters the picture when four simple saddles of (1.2) undergo confluence as $\alpha \to \alpha_0$. In this case, the uniform asymptotic behaviour of (1.2) will be governed by terms involving S, S_x, S_y, and S_z, each multiplied by powers of $\lambda^{-1/5}$. Ursell, in [Urs], discussed the uniform asymptotic theory for integrals with several coalescing saddle points using an integral closely related to (1.1).

One factor serving to limit the utility of integrals of the type in (1.1) and (1.3) is the shortage of tables of values of these functions and a comparative lack of information regarding the large parameter behaviour of (1.1) and (1.2). The past few years have seen a growing body of work dedicated to resolving this difficulty. On the numerical side, Connor and his associates [Con1, Con2, Con3, Con4] have done significant work in developing tables for both P and S, but correspondingly little has been done in examining the large parameter behaviour of S (the Pearcey function has been analyzed in [Kam2, Par, Sta]).

In this work, we will develop asymptotic expansions of $S(x,y,z)$ which remain valid as three saddle points coalesce. The plan of the paper first involves a number of preliminary steps, collected in § 2, where a description of the so-called "caustic" associated with the swallowtail integral is presented. The type of asymptotic behaviour to be found for S near the caustic is briefly discussed, as is a detailed description of the behaviour of the saddle points of S.

The interesting case of large negative x behaviour of S, which serves as the focus of this paper, is taken up in § 3, with an examination of a quartic change of variables first used by Ursell in 1972; see [Urs]. In the process of developing the asymptotics of $S(x,y,z)$, we will provide the first "concrete" application of the Pearcey integral to uniform asymptotic theory.

Section 4 briefly examines the conformal mapping determined by the quartic transformation and shows why the attention we pay to the use of full steepest descent contours is important.

Section 5 examines the limiting forms of the uniform expansion obtained in § 3 and determines the values of the first three coefficients in the uniform expansion, at the caustic.

In the final portion of the paper, we present our results in the form of a theorem and show why restrictions imposed in earlier sections can be relaxed. A brief examination of the termwise differentiation of the uniform expansion is also undertaken.

2. Preliminaries.

2.1. Caustics. Let $F(t) = t^5/5+xt^3/3+yt^2/2+zt$ be the phase function of the integral (1.1). It is well known from stationary phase philosophy that the asymptotic behaviour of (1.1) is governed by the number and order of critical points of F, and in particular, by the number and order of those critical points that are real (if F has only complex critical points, S is of exponentially small order as $|x|+|y|+|z| \to \infty$).

For t_0 to be a zero of F' of order two or more requires that $F'(t) = t^4+xt^2+yt+z$ and $F''(t) = 4t^3+2xt+y$ both vanish simultaneously at $t = t_0$. If t_0 is a zero of order three, then we additionally must have $F'''(t) = 12t^2+2x$ equal to zero when $t = t_0$. Finally, we observe that zeros of F' of order four can occur only when $t_0 = 0$.

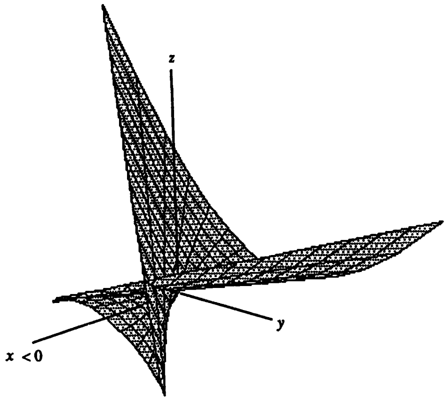


FIG. 1. *Locus of points (x, y, z) for which F' and F'' have simultaneous real zeros.*

From $F'''(t_0) = 0, t_0 = 0$ implies that x must be zero, and resubstituting these two values in $F'' = 0$ and $F' = 0$ shows that F' has a zero of order four only when $x = y = z = 0$. Therefore, in our problem of determining the large $|x| + |y| + |z|$ behaviour of S, we see that we never encounter a stationary point of order four.

Proceeding similarly for zeros of F' of order three shows that such zeros (for real t_0) can occur only when $x \leq 0$, as $F'''(t_0) = 12t_0^2 + 2x$ never vanishes for positive x. If x is positive, then S can have stationary points only of order ≤ 2. In the latter setting, the method of Chester, Friedman and Ursell [CFU] can be applied; this is briefly discussed in the last part of this section.

Interesting behaviour can therefore be expected for S when $x \leq 0$, and we shall find it convenient to characterize the saddle point structure of S after the fashion of Gilmore [Gil]. We accomplish this by plotting those parameter values of (x, y, z) for which F' has (real) zeros of order two or more. The result of this exercise is frequently referred to as a *caustic*, and is displayed in Fig. 1. We will examine this self-intersecting surface by taking plane slices $x =$ constant for $x > 0$, $x = 0$, and $x < 0$.

For $x > 0$, say $x = 1$, we have $F' = t^4 + t^2 + yt + z$ and $F'' = 4t^3 + 2t + y$ whence, if F' and F'' vanish simultaneously, $(y, z) = (-2t - 4t^3, t^2 + 3t^4)$. The resulting parametric curve in the yz-plane is the set of y and z values ($x = 1$) for which F has coalesced real critical points. Notice also that replacing t by $-t$ in the parametric forms for y and z results in only a change in sign of y.

The plane slice $x = 0$ results in a similarly shaped curve.

The negative $x =$ constant plane slice reveals a more complicated structure. For the purpose of illustration, we take $x = -1$. In this case, $F''' = 12t^2 - 2$, so that critical points of order three exist precisely when $t = \pm 1/\sqrt{6}$. The parametric equation of the curve in the yz-plane along which F has critical points of order two (or higher) is $(y, z) = (2t - 4t^3, -t^2 + 3t^4)$; see Fig. 2. Note the presence of two *cusps*, which correspond to those values of y and z for which F (with $x = -1$) has critical points of order three. Other points of the curve are associated with critical points of order two.

At points off the surface in Fig. 1, F has at most simple real zeros. The reader interested in a more detailed discussion of these plane sections is directed to [Gil, p. 62].

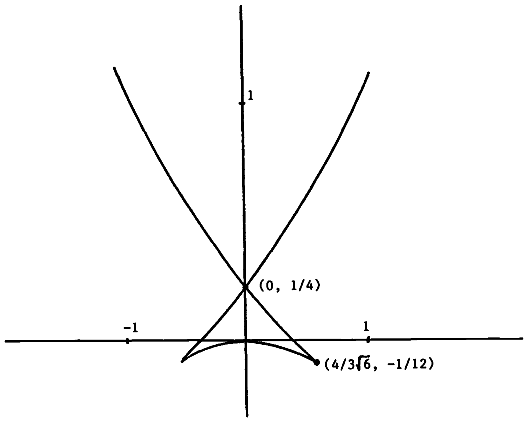


FIG. 2. *The plane slice $x = -1$ of Fig. 1. The bounded region contains those values (y, z) for which F' has exactly four distinct real zeros. Confluence of zeros of F' occurs only for (y, z) on the curve. Elsewhere, F' has either two distinct real zeros, or only complex conjugate pairs of zeros. This illustration is referenced later in the text, with y and z replaced by b and c, respectively.*

2.2. Reduction of S. Since the critical points of the phase function F of S can undergo a confluence only for points on the caustic, and since all such confluences involve only a pair of critical points except at the cusps, it suffices to restrict our attention only to the case of negative x. Elsewhere on the caustic, the method of Chester, Friedman, and Ursell (cf. [CFU]) can be applied to obtain the uniform asymptotic behaviour of S, as has been done in [Kam1].

Furthermore, we may, without loss of generality, assume that $y \geq 0$ in our treatment, as the case of $y < 0$ follows upon conjugation, since $S(x, -y, z)$ and $S(x, y, z)$ are complex conjugates. For notational convenience, we replace x by $-x$ and consider $S(-x, y, z)$ with $x > 0$.

If we introduce the change of variables $t = x^{1/2}u$, then $S(-x, y, z)$ has the integral representation

$$S(-x, y, z) = x^{1/2} \int_{\infty e^{9\pi i/10}}^{\infty e^{\pi i/10}} e^{ix^{5/2} f(u; yx^{-3/2}, zx^{-2})} du, \tag{2.1}$$

where

$$f(u; b, c) = \frac{u^5}{5} - \frac{u^3}{3} + b\frac{u^2}{2} + cu. \tag{2.2}$$

The uniform asymptotic behaviour of $S(-x, y, z)$ for large positive x can therefore be obtained from the uniform asymptotic behaviour of

$$I(\lambda) = \int_{\infty e^{9\pi i/10}}^{\infty e^{\pi i/10}} \exp\left[i\lambda f(t; b, c)\right] dt, \tag{2.3}$$

with $\lambda \to +\infty$. We shall work exclusively with this integral, recovering the desired behaviour of S at the end through the use of (2.1).

To proceed further requires detailed knowledge concerning the behaviour of the saddle points of $I(\lambda)$.

2.3. Zeros of quartic polynomials. It is well known that a solution of quartics by radicals is available, and several works in the theory of equations present formulae for this purpose. Regrettably, the form the zeros take is often unwieldy, due to the nested radicals that appear.

An elegant solution to the problem of extracting the roots of quartic polynomials can be found in the work of A. Greenhill [Gre1, Gre2], published in the late 19th century. Greenhill's approach involves the use of Weierstrass elliptic functions, and the expressions he obtains for the zeros of quartic polynomials reduce in short order to the more familiar forms provided by classical techniques, such as the use of Lagrange resolvents.

Let

$$U = x^4 + 6Cx^2 + 4Dx + E,$$

and set $g_2 = E + 3C^2$, $g_3 = CE - D^2 - C^3$. (Note that every quartic can be given the form possessed by U through the use of a linear change of variables.) Denote by $\wp(z; g_2, g_3)$ the Weierstrass elliptic function formed with the invariants g_2 and g_3. Let α (in the fundamental period parallelogram) be that number for which $\wp(2\alpha; g_2, g_3) = -C$, $\wp'(2\alpha; g_2, g_3) = -D$; for the existence of such α, see either [Gre1, p. 271–272] or [Gre2, p. 152–153]. Put

$$\mathcal{S} = 4s^3 - g_2 s - g_3$$

and let the discriminant of this cubic (called the *discriminating cubic* of the quartic U) be

$$\delta = g_2^3 - 27g_3^2.$$

Denote the zeros of $\mathcal{S}$ by $e_i, i = 1, 2, 3$, and those of U by $x_j, j = 0, 1, 2, 3$. Greenhill found that

$$\begin{aligned} x_0 &= \sqrt{\wp(2\alpha) - e_1} + \sqrt{\wp(2\alpha) - e_2} + \sqrt{\wp(2\alpha) - e_3}, \\ x_1 &= \sqrt{\wp(2\alpha) - e_1} - \sqrt{\wp(2\alpha) - e_2} - \sqrt{\wp(2\alpha) - e_3}, \\ x_2 &= -\sqrt{\wp(2\alpha) - e_1} + \sqrt{\wp(2\alpha) - e_2} - \sqrt{\wp(2\alpha) - e_3}, \\ x_3 &= -\sqrt{\wp(2\alpha) - e_1} - \sqrt{\wp(2\alpha) - e_2} + \sqrt{\wp(2\alpha) - e_3}. \end{aligned}$$

We note here that there is no ambiguity in the previous set of equations, as the square roots are chosen by

$$\sqrt{\wp(z) - e_i} = \frac{\sigma(z + \omega_i)}{\sigma(z) \cdot \sigma(\omega_i)} e^{-\eta_i z}.$$

In this equation, ω_i is an irreducible half-period, and the numbers η_i are determined by $\eta_i = \zeta(\omega_i)$. The functions ζ and σ are, respectively, the Weierstrass zeta and sigma functions. A number of well-known properties of $\wp$ and relations involving the constants ω_i, η_i, and e_i can be found in [Cop]. Of interest to us is the fact the square root determined in the previous equation is always ± the principal branch; see [Cop, p. 367].

This latter piece of information, together with the equation $\wp(2\alpha) = -C, \wp'(2\alpha) = -D$, give us the more familiar Lagrange form for the zeros of the quartic U, only with additional information regarding the branches of the square roots that appear. From this, we find that the zeros of $f'(t) = t^4 - t^2 + bt + c$ are given by

$$\begin{aligned} t_0 &= \sqrt{1/6 - e_1} + \sqrt{1/6 - e_2} - \sqrt{1/6 - e_3}, \\ t_1 &= \sqrt{1/6 - e_1} - \sqrt{1/6 - e_2} + \sqrt{1/6 - e_3}, \\ t_2 &= -\sqrt{1/6 - e_1} + \sqrt{1/6 - e_2} + \sqrt{1/6 - e_3}, \\ t_3 &= -\sqrt{1/6 - e_1} - \sqrt{1/6 - e_2} - \sqrt{1/6 - e_3}. \end{aligned} \tag{2.4}$$

Here, the e_i's are the roots of the discriminating polynomial, given by

$$\mathcal{S} = 4s^3 - \left(c + \frac{1}{12}\right)s + \left(\frac{c}{6} + \frac{b^2}{16} - \frac{1}{216}\right); \tag{2.5}$$

use of the trigonometric solution of cubics allows us to express the e_i's as

(2.6)

$$e_1 = \sqrt{\tfrac{c}{3} + \tfrac{1}{36}}\sin(\pi/3 - \psi), \quad e_2 = \sqrt{\tfrac{c}{3} + \tfrac{1}{36}}\sin\psi, \quad e_3 = -\sqrt{\tfrac{c}{3} + \tfrac{1}{36}}\sin(\pi/3 + \psi),$$

with the angle ψ given by

$$\sin 3\psi = \frac{c/6 + b^2/16 - 1/216}{(c/3 + 1/36)^{3/2}}. \tag{2.7}$$

Before proceeding with our discussion of the asymptotics of I, a few observations regarding the e_i should be recorded.

Along the caustic, we have

$$\left(\frac{c}{6} + \frac{b^2}{16} - \frac{1}{216}\right)^2 = \left(\frac{c}{3} + \frac{1}{36}\right)^3, \tag{2.8}$$

since the discriminating cubic must have repeated zeros when two or more t_i's coincide; thus, this equation gives the caustic in the bc-plane. This is precisely the $x = -1$ caustic of § 2.1, with y and z replaced by b and c (recall Fig. 2). The cusp points are easily found to have coordinates $(\pm 4/3\sqrt{6}, -1/12)$. As we pointed out in § 2.1, because of the symmetry in the caustic with respect to the vertical axis (in either y- or c- coordinates), we need only consider those values (b, c) with nonnegative ordinate.

Along the caustic, (2.7) must become $\sin 3\psi = \pm 1$ or $\psi = \pm\pi/6$. At the point $(b, c) = (0, 0)$, we find $\psi = -\pi/6$, and at $(b, c) = (0, 1/4)$, we have $\psi = \pi/6$. Thus, on the lower arch of the caustic (that segment joining $(0, 0)$ with $(4/3\sqrt{6}, -1/12)$), ψ must be $-\pi/6$, and on the middle segment of the caustic (that portion joining $(4/3\sqrt{6}, -1/12)$ to $(0, \frac{1}{4})$), we have $\psi = \pi/6$; see Fig. 2. By similar reasoning, we find that all level curves $\sin 3\psi = \tau$, with $-1 < \tau < 1$, must also pass through the cusp.

At the point $(b, c) = (0, 0)$, we find $t_0 = t_1 = 0$, $t_2 = 1$, and $t_3 = -1$ since $\psi = -\pi/6$, $e_1 = 1/6$, and $e_2 = e_3 = -1/12$. For $(b, c) = (0, \frac{1}{4})$, we have $\psi = \pi/6, e_1 = e_2 = 1/6$, and $e_3 = -1/3$ whence we find $t_0 = t_3 = -1/\sqrt{2}$, and $t_1 = t_2 = 1/\sqrt{2}$. Finally, at the cusp itself, where $(b, c) = (4/3\sqrt{6}, -1/12)$, all e_i's vanish since the discriminating cubic (2.5) reduces to $\mathcal{S} = 4s^3$. In this case, we have $t_0 = t_1 = t_2 = 1/\sqrt{6}$, and $t_3 = -3/\sqrt{6}$.

Piecing together these special values for the roots of $f' = 0$ shows that

$$t_3 \leq t_0 \leq t_1 \leq t_2. \tag{2.9}$$

Equality can occur only when (b, c) lies on the caustic. Hence, (2.9) is, in fact, strict inside the caustic. If we restrict (b, c) further so that $b \geq b_0 > 0$, then we find that the zero t_3 is isolated from the other three zeros.

With this restriction on the parameters b, c, we note the following. Inside the caustic, all four zeros of f' are real and distinct. On the lower arch, t_0 and t_1 coalesce and then separate into a complex conjugate pair as we pass below the caustic. On the upper arch (between the cusp and the point $(0, \frac{1}{4})$), t_1 and t_2 coincide, and then separate into a complex conjugate pair as we rise out of the caustic. t_3 remains real throughout this range of values of b and c.

The (nonuniform) asymptotic behaviour of I is readily available. For (b, c) inside the caustic, we have

$$I(\lambda; b, c) \sim \sum_{j=0}^{3} e^{\lambda i f(t_j; b,c) + \frac{\pi i}{4} \operatorname{sgn}(f''(t_j; b,c))} \sqrt{\frac{2\pi}{\lambda |f''(t_j; b, c)|}} \tag{2.10}$$

as $\lambda \to +\infty$, with (b, c) fixed. On the caustic (but not at the cusp), we have either $t_0 = t_1$ or $t_1 = t_2$. For the purpose of illustration, assume that $t_0 = t_1$. Then, for large λ and fixed (b, c), we have

$$\begin{aligned} I(\lambda; b, c) \sim & \sum_{j=2}^{3} e^{\lambda i f(t_j; b,c) + \frac{\pi i}{4} \operatorname{sgn}(f''(t_j; b,c))} \sqrt{\frac{2\pi}{\lambda |f''(t_j; b, c)|}} \\ & + \frac{2^{1/3} \Gamma(1/3) e^{\lambda i f(t_0; b,c)}}{3^{1/6} (\lambda f'''(t_0; b, c))^{1/3}}. \end{aligned} \tag{2.11}$$

In (2.11), notice that $f'''(t_0) \neq 0$ since we have fixed $(b, c) \neq (4/3\sqrt{6}, -1/12)$.

At the cusp, we have the triple zero $t_0 = t_1 = t_2 = 1/\sqrt{6}$ and the simple zero $t_3 = -3/\sqrt{6}$. Hence,

$$\begin{aligned} I(\lambda; \frac{4}{3\sqrt{6}}, \frac{-1}{12}) \sim & \frac{3^{1/8} \Gamma(1/4) e^{\pi i/8 - \lambda i/45\sqrt{6}}}{2^{7/8} \lambda^{1/4}} + \frac{3^{3/4}}{2^{7/4}} \sqrt{\frac{\pi}{\lambda}} e^{7\lambda i/5\sqrt{6} - \pi i/4} \\ & + \frac{2^{3/8} 3^{3/8} 63 \cdot \Gamma(3/4)}{32 \cdot 25 \lambda^{3/4}} e^{3\pi i/8 - \lambda i/45\sqrt{6}} \end{aligned} \tag{2.12}$$

as $\lambda \to +\infty$.

If (b, c) lies outside the caustic, then the asymptotic behaviour of $I(\lambda; b, c)$ is given by two terms from (2.10). One of these two terms arises from the real saddle t_3—the other stems from the remaining real saddle. The other two saddles, which now form a complex conjugate pair, provide only an exponentially negligible contribution to the value of I due to considerations of the topography of the saddles.

3. Uniform expansion of I.

3.1. The quartic transformation. Because the zeros t_0, t_1, and t_2 undergo various confluences, with all three zeros coinciding at the cusp, we introduce the quartic transformation first examined in [Urs]:

$$f(t; b, c) = \frac{z^4}{4} - \zeta \frac{z^2}{2} + \eta z + \theta. \tag{3.1}$$

This requires the determination of the parameters ζ, η, and θ as functions of b and c which provide for a (local) uniformly analytic, one-to-one transformation from t to z.

In particular, we shall require that the saddles of f correspond with the saddles of the right-hand side of (3.1). For notational convenience, we will denote the right-hand side of (3.1) by $g(z;\zeta,\eta,\theta)$; frequently, we will suppress the parameters ζ,η, and θ.

From the trigonometric solution of cubic equations, we have, as zeros of $g'(z;\zeta,\eta,\theta) = z^3-\zeta z+\eta$,

$$(3.2)\quad z_1 = 2\sqrt{\frac{\zeta}{3}}\sin(\pi/3-\phi),\quad z_2 = 2\sqrt{\frac{\zeta}{3}}\sin\phi,\quad z_3 = -2\sqrt{\frac{\zeta}{3}}\sin(\pi/3+\phi),$$

with the angle ϕ given by

$$(3.3)\qquad \sin 3\phi = \frac{3^{3/2}\eta}{2\zeta^{3/2}}.$$

An examination of the functions $\sin(\pi/3-\phi), \sin\phi$, and $-\sin(\pi/3+\phi)$ for $\phi \in\ ]-\pi/6,\pi/6[$ reveals that when the z_i are real, they satisfy the inequality

$$(3.4)\qquad z_3 < z_2 < z_1.$$

Note, too, that the quartic g must have $\zeta > 0$ in order to have three real saddles, and that these saddles z_i, for $i = 1,2,3$, coincide for $(\zeta,\eta) = (0,0)$.

The inequalities (2.9) and (3.4), together with the observation that t_3 remains isolated from t_0, t_1, and t_2 for $b \geq b_0 > 0$, strongly suggests that the correspondence

$$t_2 \leftrightarrow z_1,\quad t_1 \leftrightarrow z_2,\quad t_0 \leftrightarrow z_3$$

be made by the uniformly analytic, one-to-one solution of (3.1). How shall the parameters ζ,η, and θ be determined? The straightforward substitution of the t_i's and their associated z_j's into (3.1) rapidly leads to an unappealing system of nonlinear equations. A more elegant approach, adopted here, entails the use of some classical theory of equations.

For a polynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ with zeros $x_1, x_2, \cdots, x_n$, it is well known that if we put $s_1 = \sum_{j=1}^n x_j$, $s_2 = \sum_{i<j} x_ix_j$, $s_3 = \sum_{i<j<k} x_ix_jx_k$, $\cdots$, $s_n = \prod_{j=1}^n x_j$, then the s_i's, the elementary symmetric functions of the roots of $p(x) = 0$, are related to the coefficients $a_0, a_1, \cdots, a_{n-1}$ via the formulae

$$(3.5)\quad a_0 = (-1)^n s_n,\quad a_1 = (-1)^{n+1}s_{n-1},\quad a_2 = (-1)^{n+2}s_{n-2},\quad \cdots,\quad a_{n-1} = -s_1.$$

We note that the saddles z_1, z_2, and z_3 are zeros of the polynomial $z^3 - \zeta z + \eta$. Hence, a straightforward application of (3.5) yields

$$(3.6)\qquad z_1+z_2+z_3 = 0,\quad z_1z_2+z_1z_3+z_2z_3 = -\zeta,\quad z_1z_2z_3 = -\eta.$$

If we form the functions

$$(3.7)\qquad \begin{aligned} \sigma_1 &= f(t_0)+f(t_1)+f(t_2),\\ \sigma_2 &= f(t_0)f(t_1)+f(t_0)f(t_2)+f(t_1)f(t_2),\\ \sigma_3 &= f(t_0)f(t_1)f(t_2), \end{aligned}$$

then use of the correspondence between the t_i's and the z_j's and equation (3.1) permits us to express the functions (3.7) as functions of the z_j's. Subsequent use of (3.6) then provides us with expressions for the σ_i's as functions of the parameters ζ,η, and θ.

These last expressions will prove solvable and will permit the computation of ζ, η, and θ in terms of the known roots t_0, t_1, and t_2.

Before proceeding with the computations, we note that this approach was implemented by Connor and his colleagues in [Con2]. However, the equations they used did not yield ζ, η, and θ as functions of the t_i's—the parameters were instead obtained by applying a numerical scheme to what will, in our treatment, appear as intermediate results. Furthermore, because of the extensive computation involved in solving for ζ, η, and θ as functions of the t_i's, we shall provide details only for one of the three equations we require; additional details regarding the equations we obtain can be found in the appendix of [Kam1].

We have, from (3.1) and the first equation of (3.7),

$$\sigma_1 = f(t_0) + f(t_1) + f(t_2) = \frac{1}{4}\sum_{i=1}^{3} z_i^4 - \frac{\zeta}{2}\sum_{i=1}^{3} z_i^2 + \eta \sum_{i=1}^{3} z_i + 3\theta.$$

By the first equation of (3.6), this reduces to

$$\sigma_1 = \frac{1}{4}(z_1^4 + z_2^4 + z_3^4) - \frac{\zeta}{2}(z_1^2 + z_2^2 + z_3^2) + 3\theta. \tag{3.8}$$

To calculate the sums of squares and of fourth powers, we proceed in the following fashion: squaring the first equation of (3.6) gives $(z_1 + z_2 + z_3)^2 = 0$ which reduces to $z_1^2 + z_2^2 + z_3^2 = 2\zeta$ in view of the second equation of (3.6). Equation (3.8) now becomes

$$\sigma_1 = \frac{1}{4}(z_1^4 + z_2^4 + z_3^4) - \zeta^2 + 3\theta. \tag{3.9}$$

Upon squaring the expression for the sum of squares, we find that $4\zeta^2 = (z_1^2 + z_2^2 + z_3^2)^2 = z_1^4 + z_2^4 + z_3^4 + 2(z_1^2 z_2^2 + z_1^2 z_3^2 + z_2^2 z_3^2)$. The latter term can be evaluated by squaring the second equation in (3.6). Thus, we have $\zeta^2 = (z_1 z_2 + z_1 z_3 + z_2 z_3)^2 = z_1^2 z_2^2 + z_1^2 z_3^2 + z_2^2 z_3^2$ since $z_1 + z_2 + z_3 = 0$. Therefore, $z_1^2 z_2^2 + z_1^2 z_3^2 + z_2^2 z_3^2 = \zeta^2$ so the expression for the sum of fourth powers becomes $4\zeta^2 = z_1^4 + z_2^4 + z_3^4 + 2\zeta^2$. Use of this in (3.9) gives

$$\sigma_1 = -\frac{\zeta^2}{2} + 3\theta. \tag{3.10}$$

Proceeding in a similar fashion gives

$$\sigma_2 = \frac{\zeta^4}{16} - \frac{9\eta^2\zeta}{8} - \theta\zeta^2 + 3\theta^2 \tag{3.11}$$

and

$$\sigma_3 = -\frac{27\eta^4}{64} + \frac{\eta^2\zeta^3}{32} + \frac{\theta\zeta^4}{16} - \frac{9\eta^2\theta\zeta}{8} - \frac{\theta^2\zeta^2}{2} + \theta^3. \tag{3.12}$$

It is with equations (3.10)–(3.12) that we shall obtain expressions for ζ, η, and θ.

From (3.11), we see that $\sigma_2\theta - 3\theta^3 + \theta^2\zeta^2/2 = \theta\zeta^4/16 - 9\eta^2\theta\zeta/8 - \theta^2\zeta^2/2$. Use of this in (3.12) gives

$$\sigma_3 = -\frac{27\eta^4}{64} + \frac{\eta^2\zeta^3}{32} + \sigma_2\theta + \frac{\theta^2\zeta^2}{2} - 2\theta^3. \tag{3.13}$$

From (3.10) we have

$$\theta = \frac{\sigma_1}{3} + \frac{\zeta^2}{6}, \tag{3.14}$$

and this in (3.11) gives

$$\eta^2 = -\frac{\zeta^3}{54} + \frac{8(\sigma_1^2 - 3\sigma_2)}{27\zeta}. \tag{3.15}$$

Equation (3.14) can be applied to (3.13) to eliminate all occurrences of θ. The result of doing this is:

$$\sigma_3 = -\frac{27\eta^4}{64} + \frac{\eta^2\zeta^3}{32} + \frac{1}{3}(\sigma_1\sigma_2 - \frac{2\sigma_1^3}{9}) - \frac{\zeta^2}{18}(\sigma_1^2 - 3\sigma_2) + \frac{\zeta^6}{216}. \tag{3.16}$$

If we replace every instance of η^2 by the right-hand side of (3.15) and multiply the resulting equation by $256\zeta^2$, we obtain

(3.17)
$$\zeta^8 - \frac{32}{3}(\sigma_1^2 - 3\sigma_2)\zeta^4 + \frac{256}{27}(9\sigma_1\sigma_2 - 2\sigma_1^3 - 27\sigma_3)\zeta^2 - \frac{256}{27}(\sigma_1^2 - 3\sigma_2)^2 = 0.$$

Note that

$$9\sigma_1\sigma_2 - 2\sigma_1^3 - 27\sigma_3 = \sigma_1^3 - 3\sigma_1(\sigma_1^2 - 3\sigma_2) - 27\sigma_3. \tag{3.18}$$

Equation (3.18) will prove to be of value in later discussion.

Put $Z = \zeta^2$ in the octic (3.17). The resulting quartic,

(3.19)
$$Z^4 - \frac{32}{3}(\sigma_1^2 - 3\sigma_2)Z^2 + \frac{256}{27}(9\sigma_1\sigma_2 - 2\sigma_1^3 - 27\sigma_3)Z - \frac{256}{27}(\sigma_1^2 - 3\sigma_2)^2 = 0,$$

can be solved using Greenhill's formulae (cf. § 2.3). To that end, we calculate the invariants of the required elliptic function.

From $g_2 = E + 3C^2$ and $g_3 = CE - D^2 - C^3$ in § 2.3, we have

$$\begin{aligned} g_2 &= -\tfrac{256}{27}(\sigma_1^2 - 3\sigma_2)^2 + 3[-\tfrac{16}{9}(\sigma_1^2 - 3\sigma_2)]^2 = 0, \\ g_3 &= \tfrac{4\cdot 16^3}{9^3}(\sigma_1^2 - 3\sigma_2)^3 - \tfrac{64^2}{27^2}(9\sigma_1\sigma_2 - 2\sigma_1^3 - 27\sigma_3)^2. \end{aligned} \tag{3.20}$$

The first of the two preceding equations represents a happy state of affairs, as the discriminating cubic takes the form $\mathcal{S} = 4s^3 - g_3$. Hence, the zeros of $\mathcal{S}$ are given by $\omega^j(g_3/4)^{1/3}, j = 0, 1, 2$, where ω is the cube root of unity $\omega = e^{2\pi i/3}$, and $(g_3/4)^{1/3}$ is taken with its principal value. Use of (3.20) permits us to write

$$\left(\frac{g_3}{4}\right)^{1/3} = \frac{16}{9}(\sigma_1^2 - 3\sigma_2)[1 - \mathcal{N}/\mathcal{D}]^{1/3}, \tag{3.21}$$

where we have set

$$\begin{aligned} \mathcal{N} &= (9\sigma_1\sigma_2 - 2\sigma_1^3 - 27\sigma_3)^2, \\ \mathcal{D} &= 4(\sigma_1^2 - 3\sigma_2)^3. \end{aligned} \tag{3.22}$$

Also, from $\wp(2\alpha) = -C$ (cf. § 2.3), we have $\wp(2\alpha) = \frac{16}{9}(\sigma_1^2 - 3\sigma_2)$.

For the sake of notational ease, let

$$\chi = \left[1 - \frac{\mathcal{N}}{\mathcal{D}}\right]^{1/3}; \tag{3.23}$$

upon applying Greenhill's formulae with the e_j's replaced by the $\omega^j(g_3/4)^{1/3}$, we obtain

$$\begin{aligned} Z_0 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\textstyle\sum_{j=0}^2 \sqrt{1-\omega^j\chi}, \\ Z_1 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\left\{\sqrt{1-\chi} - \textstyle\sum_{j=1}^2 \sqrt{1-\omega^j\chi}\right\}, \\ Z_2 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\left\{\sqrt{1-\omega\chi} - \textstyle\sum_{j=0,2} \sqrt{1-\omega^j\chi}\right\}, \\ Z_3 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\left\{\sqrt{1-\omega^2\chi} - \textstyle\sum_{j=0}^1 \sqrt{1-\omega^j\chi}\right\}, \end{aligned} \tag{3.24}$$

where the first surd, $\sqrt{\sigma_1^2 - 3\sigma_2}$, is taken to have its principal value. The other surds must have their branches chosen with care although, from our discussion in § 2.3, we know the remaining square roots are ± the principal branch. To make the appropriate selection of these branches, we examine the Z_i's for values of (b, c) on the caustic.

On the caustic, the discriminating cubic associated with the quartic in Z has repeated roots. Hence, at least two t_i's coincide (whence at least two $f(t_i)$'s coincide) and from the discriminant $\delta = g_2^3 - 27g_3^2$ (recall § 2.3), we get $\delta = 0$ and

$$(\sigma_1^3 - 3\sigma_1(\sigma_1^2 - 3\sigma_2) - 27\sigma_3)^2 = 4(\sigma_1^2 - 3\sigma_2)^3; \tag{3.25}$$

for the sake of argument, assume that we have $t_0 = t_1 \neq t_2$, so that $f(t_0) = f(t_1) \neq f(t_2)$. Equation (3.7), together with some arithmetic, provides us with $9\sigma_1\sigma_2 - 2\sigma_1^3 - 27\sigma_3 = 2[f(t_0) - f(t_2)]^3$ and $\sigma_1^2 - 3\sigma_2 = [f(t_0) - f(t_2)]^2$. Use of these two equations, together with (3.18), establishes (3.25). Similarly, we can obtain (3.25) in the case $t_0 \neq t_1 = t_2$.

Equation (3.25) leads to the conclusion that $\mathcal{N} = \mathcal{D}$ on the caustic, whence the expressions for the Z_i's reduce to

$$\begin{aligned} Z_0 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}(\pm 1 \pm 1 \pm 1), & Z_1 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}(\pm 1 \mp 1 \mp 1), \\ Z_2 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}(\mp 1 \pm 1 \mp 1), & Z_3 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}(\mp 1 \mp 1 \pm 1), \end{aligned} \tag{3.26}$$

where in only the first row are the signs independently chosen, the choice of signs subsequently determined according to the pattern used for the x_j in terms of $\sqrt{\wp(2\alpha) - e_k}$ in § 2.3. With (3.25) and (3.19), the quartic in Z reduces to

$$Z^4 - \frac{32}{3}(\sigma_1^2 - 3\sigma_2)Z^2 + \frac{512}{27}(\sigma_1^2 - 3\sigma_2)^{3/2}Z - \frac{256}{27}(\sigma_1^2 - 3\sigma_2) = 0. \tag{3.27}$$

The form of the Z_i's in (3.26) strongly suggests putting $Z_i = \frac{4}{3}(\sigma_1^2 - 3\sigma_2)^{1/2}\epsilon_i$. In this event, (3.27) yields the reduced equation

$$\frac{\epsilon_i^4}{3} - 2\epsilon_i^2 + \frac{8\epsilon_i}{3} - 1 = 0. \tag{3.28}$$

Experiment is now easy and shows, for example, that if we choose the square roots appearing in the first equation of (3.24) to be all principal branch choices, then $\epsilon_0 = 3$ does not satisfy (3.28), but if we choose the square roots so the first two have their principal branch, the last one being the negative of the principal branch, then we have $\epsilon_0 = \epsilon_1 = \epsilon_2 = 1$ and $\epsilon_3 = -3$. This latter choice provides ϵ_i's, all of which satisfy (3.28). Hence, we make this latter selection of branches for the surds appearing in (3.24).

All other choices of branches for the surds in (3.24) amount to a reordering of the Z_i's under either of the two choices of branches discussed above. With our choice of branches, (3.24) becomes

$$\begin{aligned}
Z_0 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\left[\sqrt{1-\chi} + \sqrt{1-\omega\chi} - \sqrt{1-\omega^2\chi}\right], \\
Z_1 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\left[\sqrt{1-\chi} - \sqrt{1-\omega\chi} + \sqrt{1-\omega^2\chi}\right], \\
Z_2 &= \tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\left[-\sqrt{1-\chi} + \sqrt{1-\omega\chi} + \sqrt{1-\omega^2\chi}\right], \\
Z_3 &= -\tfrac{4}{3}\sqrt{\sigma_1^2 - 3\sigma_2}\left[\sqrt{1-\chi} + \sqrt{1-\omega\chi} + \sqrt{1-\omega^2\chi}\right],
\end{aligned} \tag{3.29}$$

where all surds are taken with their principal branch.

For our change of variables (3.1), we require $\zeta > 0$ (inside the caustic); recall that $Z = \zeta^2$. Hence, we must determine which of the Z_i's is real and nonnegative. With all square roots chosen to be the principal branch, we know that $\sqrt{\bar{z}}$ is the conjugate of $\sqrt{z}$, whence the observation that ω and ω^2 are complex conjugates implies that $\sqrt{1-\omega\chi}$ and $\sqrt{1-\omega^2\chi}$ are conjugates. Thus, of the Z_i's in (3.29), we see that only Z_2 and Z_3 can be real.

To determine which of Z_2 or Z_3 gives rise to real square roots, we examine the limiting behaviour of Z_2 and Z_3 as we approach the caustic. From (3.25), we know that for δ near zero, the ratio $\mathcal{N}/\mathcal{D}$ is approximately 1, in which case χ is tending to zero. Applying the binomial theorem to the expressions for Z_2 and Z_3 gives us

$$\begin{aligned}
-\sqrt{1-\chi} + \sqrt{1-\omega\chi} + \sqrt{1-\omega^2\chi} &= 1 + \chi + O(\chi^2), \\
\sqrt{1-\chi} + \sqrt{1-\omega\chi} + \sqrt{1-\omega^2\chi} &= 3 + O(\chi^2).
\end{aligned}$$

Thus, Z_2 is positive, Z_3 is negative, and so

$$\zeta = \pm\sqrt{\frac{4}{3}}\sqrt[4]{\sigma_1^2 - 3\sigma_2}\sqrt{-\sqrt{1-\chi} + \sqrt{1-\omega\chi} + \sqrt{1-\omega^2\chi}}.$$

Use of the fact that ζ must be positive inside the caustic yields

$$\zeta = \sqrt{\frac{4}{3}}\sqrt[4]{\sigma_1^2 - 3\sigma_2}\sqrt{-\sqrt{1-\chi} + \sqrt{1-\omega\chi} + \sqrt{1-\omega^2\chi}}. \tag{3.30}$$

This, together with (3.14) and (3.15), give computable expressions for the parameters ζ, η, and θ.

Before continuing with the development of the uniform expansion of I, we note that the term $\sigma_1^2 - 3\sigma_2 \geq 0$ inside and on the caustic, and is real outside the caustic (although it may be negative). Furthermore, in (3.15), when taking the square root to obtain η, it is chosen so that η is negative on the arc of the caustic joining $(0, 0)$ to

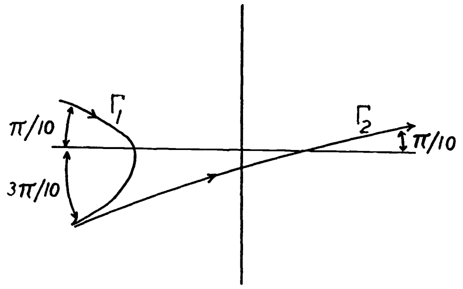


FIG. 3. *t-plane integration contours for the integrals $I_j(\lambda; b, c)$.*

$(4/3\sqrt{6}, -1/12)$, and positive on the arc joining $(4/3\sqrt{6}, -1/12)$ to $(0, \frac{1}{4})$. Additional details may be found in [Kam1, p. 84–85].

It is interesting to note that we have obtained closed form expressions for the parameters η, ζ and θ appearing in (3.1). Ursell, in examining the quartic transformation (3.1) remarks that these parameters can, in principle, be constructed from convergent power series (a consequence of using Levinson's theorem; see [Lev]), and that the power series approach is not practical computationally. He further states that the parameters can be obtained without explicit reference to the uniformly analytic one-to-one solution to (3.1), but does not go on to provide expressions for the parameters; cf. [Urs, p. 64–65]. This has been accomplished here.

With our transformation (3.1) completely determined, we can proceed to the uniform expansion of I.

3.2. The expansion of I. Since t_3 remains isolated from t_0, t_1, t_2 for $b > 0$, we rewrite $I(\lambda)$ as the sum of two path integrals:

$$I(\lambda; b, c) = \int_{\Gamma_1} e^{i\lambda f(t;b,c)} dt + \int_{\Gamma_2} e^{i\lambda f(t;b,c)} dt \equiv I_1(\lambda; b, c) + I_2(\lambda; b, c), \tag{3.31}$$

where $I_j(\lambda; b, c)$ denotes the integral of $e^{i\lambda f(t;b,c)}$ over the contour Γ_j, with the Γ_j as depicted in Fig. 3.

Γ_1 may be taken to be the steepest descent curve through t_3 beginning at $\infty e^{9\pi i/10}$ and ending at $\infty e^{13\pi i/10}$. Thus, I_1 has an asymptotic expansion of the form

$$I_1(\lambda; b, c) \sim e^{i\lambda f(t_3;b,c) - \pi i/4} \sum_{j=1}^{\infty} \frac{a_j(b,c)}{\lambda^{j/2}} \tag{3.32}$$

as $\lambda \to +\infty$. The leading term has the coefficient $a_1(b,c) = \sqrt{2\pi/(-f''(t_3; b, c))}$, which is well behaved for $b \geq b_0 > 0$. Note that $f''(t_3) < 0$ since t_3 gives a local maximum of the quintic f.

For the integral I_2, we invoke the quartic transformation (3.1) and introduce the function sequences $\{p_n\}, \{q_n\}, \{r_n\}, \{g_n\}$, and $\{h_n\}$, defined by

$$\begin{aligned} g_0(z; \zeta, \eta) &= p_0 + q_0 z + r_0 z^2 + (z^3 - \zeta z + \eta) h_0(z; \zeta, \eta), \\ \frac{\partial}{\partial z} h_k(z; \zeta, \eta) &= g_{k+1}(z; \zeta, \eta) \\ &= p_{k+1} + q_{k+1} z + r_{k+1} z^2 + (z^3 - \zeta z + \eta) h_{k+1}(z; \zeta, \eta), \end{aligned} \tag{3.33}$$

where $g_0(z;\zeta,\eta)=dt/dz$ (cf. eqn. (3.1)), and $k=0,1,\cdots$. The coefficients p_n, q_n and r_n are functions of ζ and η which, in view of our expressions for ζ,η, and θ in terms of b and c, in turn can be regarded as functions of b and c.

Through repeated substitution and partial integration, we obtain the expansion

$$(3.34)\quad I_2(\lambda;b,c)\sim e^{i\lambda\theta}\sum_{j=0}^{\infty}\left(\frac{i}{\lambda}\right)^j [p_jF_0(\lambda;\zeta,\eta)+q_jF_1(\lambda;\zeta,\eta)+r_jF_2(\lambda;\zeta,\eta)]$$

as $\lambda\to+\infty$, uniformly valid for (ζ,η) "near" the caustic $27\eta^2-4\zeta^3=0$; i.e., uniformly valid near the caustic in the bc-plane with $b\geq b_0>0$, and (b,c) within a band of fixed distance from the caustic.

The functions F_k appearing in the previous equation are given by the integrals

$$F_k=\int_C e^{i\lambda(z^4/4-\zeta z^2/2+\eta z)}z^k dz,$$

with $k=0,1,2$. C is a contour in the z-plane beginning at $\infty e^{9\pi i/8}$ and ending at $\infty e^{\pi i/8}$. With the change of variables $u=z/\lambda^{1/4}$, we find

$$\begin{aligned}F_0(\lambda;\zeta,\eta)&=\lambda^{-1/4}P(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta),\\ F_1(\lambda;\zeta,\eta)&=-i\lambda^{-1/2}P_y(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta),\\ F_2(\lambda;\zeta,\eta)&=-2i\lambda^{-3/4}P_x(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta),\end{aligned}$$

where the function $P(x,y)$ is the Pearcey function in (1.3), and the functions P_x and P_y are the first-order partial derivatives of $P(x,y)$ with respect to x and y, respectively. Use of the above expressions for the F_k in the expansion (3.34) yields the expansion

$$(3.35)\quad \begin{aligned}I_2(\lambda;b,c)\sim e^{i\lambda\theta}\sum_{j=0}^{\infty}\left(\frac{i}{\lambda}\right)^j&\left[\frac{p_j}{\lambda^{1/4}}P(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta)\right.\\ &\left.-\frac{iq_j}{\lambda^{1/2}}P_y(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta)-\frac{2ir_j}{\lambda^{3/4}}P_x(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta)\right]\end{aligned}$$

as $\lambda\to+\infty$, uniformly valid for (b,c) in a band of fixed distance from the caustic with $b\geq b_0>0$ for some fixed b_0.

The expansion of I is therefore given by the sum of (3.32) and (3.35). In particular, we have the uniform approximation

$$(3.36)\quad \begin{aligned}I(\lambda;b,c)=&\,e^{i\lambda f(t_3;b,c)-\pi i/4}\sqrt{\frac{2\pi}{-\lambda f''(t_3;b,c)}}\left[1+O\left(\frac{1}{\sqrt{\lambda}}\right)\right]\\ &+e^{i\lambda\theta}\left[\frac{p_0}{\lambda^{1/4}}P(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta)-\frac{iq_0}{\lambda^{1/2}}P_y(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta)\right.\\ &\left.-\frac{2ir_0}{\lambda^{3/4}}P_x(-\lambda^{1/2}\zeta,\lambda^{3/4}\eta)\right]\cdot\left[1+O\left(\frac{1}{\lambda}\right)\right],\end{aligned}$$

where the O-symbols are independent of (b,c).

The expansion of $S(-x,y,z)$ now follows directly from $S(-x,y,z)=x^{1/2}\cdot I(x^{5/2}; yx^{-3/2},zx^{-2})$.

4. Conformal mapping. We shall devote the next few pages to an analysis of the conformal mapping determined by the "uniform change of variables" (3.1), for primarily two reasons: first, (3.1) is not as well known as the cubic transformation of [CFU]; second, we have been performing our computations under the assumption that (3.1) gives a solution that is uniformly analytic and one-to-one over an integration contour for I (more accurately, over the contour Γ_2 used in (3.31)).

The latter point is important, for if we use only the local properties of the transformation (3.1), we can still obtain the asymptotics of S in the case where three saddles coalesce, but at the cost of replacing the "=" signs in (3.32), (3.34), and (3.35) by "$\sim$" signs. This in turn means that if a theory of error bounds should emerge for uniform expansions developed through the use of (3.1), we can directly compute the errors in our asymptotic approximations. Without the use of the full steepest descent contour Γ_2, it is unlikely that we could obtain full precision in our expansions. Additional discussion on why full contours should be used in asymptotics can be found in [Olv].

A full treatment of the conformal mapping (3.1) would involve an analysis for each of the cases where there was no confluence of the t_i, two of the t_i coalesced, or three of the t_i coalesced (recall: at most three t_i can coalesce for the large parameter behaviour of S; see the discussion in § 2.1). For the purpose of illustration, we shall examine only one case; the reader will find a treatment of other saddle point configurations in [Kaml, pp. 90–100].

We begin by supposing that a point (b, c) lies inside the caustic (recall Fig. 2). Then there are four real saddles for the integral I, $t_3 < t_0 < t_1 < t_2$ (cf. (2.9)), of which only t_0, t_1, and t_2 are involved with confluence since we have taken $b \geq b_0 > 0$. We proceed by first determining the curves for which $\operatorname{Im} f(t; b, c) = 0$. It will prove to be convenient to introduce an intermediate variable Z in (3.1):

$$f(t; b, c) = Z = \frac{z^4}{4} - \zeta\frac{z^2}{2} + \eta z + \theta. \tag{4.1}$$

Clearly, the real t-axis is mapped to the real Z-axis in the following fashion:

$$\begin{aligned}]-\infty, t_3] &\rightarrow\,]-\infty, f(t_3)]\,, & [t_3, t_0] &\rightarrow [f(t_0), f(t_3)]\,, \\ [t_0, t_1] &\rightarrow [f(t_0), f(t_1)]\,, & [t_1, t_2] &\rightarrow [f(t_2), f(t_1)]\,, \\ [t_2, +\infty[&\rightarrow [f(t_2), +\infty[\,; \end{aligned}$$

we recall that f is a quintic with four relative extrema. The remaining curves in the t-plane sent by f to the real Z-axis can be found by solving the equation

$$\operatorname{Im} f(t) = 0 = \operatorname{Im}[f(t) - f(t_k)], \qquad (k = 0, 1, 2, 3), \tag{4.2}$$

since, in the present case, each of the t_k's real implies that $f(t_k)$ is real. We note that since the left-hand side of the (4.2) is independent of k, the solution curves arising from the right-hand side are the same for each k (in other words, we can select a convenient t_k without affecting the solution).

For each k, we develop f into its Taylor expansion centered at t_k; the right-hand side of (4.2) then has no constant terms. By use of changes of variables of the form $t - t_k = \sigma + i\tau$, where we have suppressed the dependence of σ on k, we obtain easily solved equations which express σ as a function of τ, or vice versa. Plotting these curves results in Fig. 4.

These curves, which include the real axis, partition the t-plane into several disjoint regions. We shall consider only those labelled $R_1, \cdots, R_4$, as our integration contour

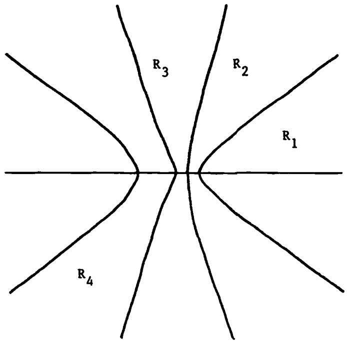


FIG. 4. *Curves in the complex t-plane for which* $\operatorname{Im} f(t; b, c) = 0$. *The real axis is included as such a curve.*

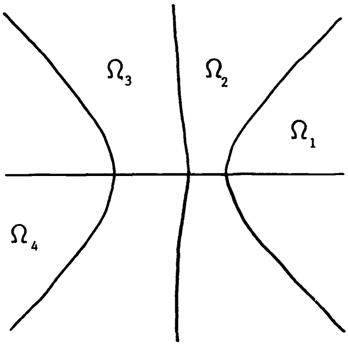


FIG. 5. *Curves in the complex z-plane for which* $\operatorname{Im} g(z) = 0$. *As in Fig.* 4, *the real axis is such a curve.*

for I_2 of equation (3.31), Γ_2, can be chosen to lie entirely within $R_1 \cup R_2 \cup R_3 \cup R_4$, or this union's closure.

We will see that the (local) uniformly analytic, one-to-one solution of (3.1) is indeed one-to-one on this union.

Since the saddles of $g(z; \zeta, \eta, \theta)$ under the tranformation (3.1) correspond with the t_i's in the fashion $t_2 \leftrightarrow z_1, t_1 \leftrightarrow z_2, t_0 \leftrightarrow z_3$, we have $z_3 < z_2 < z_1$ for (b, c) inside the caustic (this can also be seen from (3.2)). Hence, the z-plane curves for which $\operatorname{Im} g(z) = 0$ partition the z-plane into the disjoint regions displayed in Fig. 5. Again, the real axis is a curve for which $\operatorname{Im} g = 0$.

We will be concerned with the regions labelled $\Omega_1, \cdots, \Omega_4$.

Consider R_1 in the t-plane as depicted in Fig. 6. The arc BD is the steepest descent curve of if from t_2 to $\infty e^{\pi i/10}$. The arc BC is the upper extent of the region R_1. Under (4.1), the image of R_1 is the upper half of the Z-plane depicted in Fig. 6.

The map $z \to Z$ on Ω_1, defined by (4.1), produces a similar effect also displayed in Fig. 6. In the illustration, the curve BD is the steepest descent curve of $ig(z)$ beginning at z_1 and ending at $\infty e^{\pi i/8}$. BC is the upper extent of Ω_1. Under (4.1), Ω_1 maps to the half-plane depicted in Fig. 6.

Thus, under (4.1), the region R_1 maps to Ω_1.

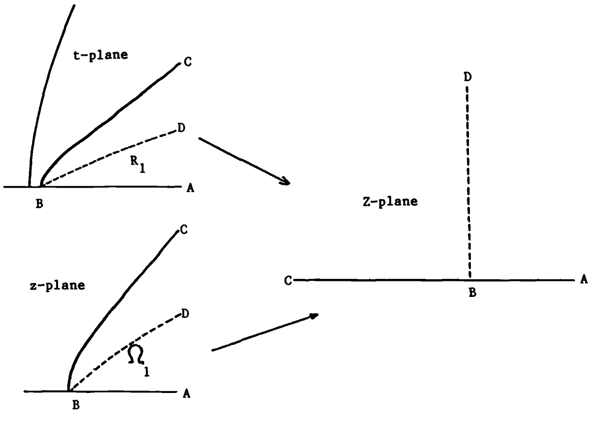


FIG. 6. *Effect of the mappings $t \longrightarrow Z$, $z \longrightarrow Z$ on the regions R_1 and Ω_1, respectively.*

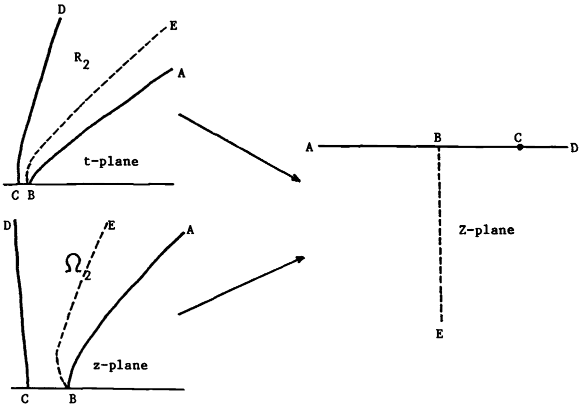


FIG. 7. *Effect of the mappings $t \longrightarrow Z$, $z \longrightarrow Z$ on the regions R_2 and Ω_2, respectively.*

We now direct our attention to R_2. In Fig. 7, the arcs CD and BA bound the region R_2; these are curves for which $\operatorname{Im} f = 0$. The arc BE is the steepest ascent curve for if from t_2 to $\infty e^{3\pi i/10}$. Under (4.1), the region maps to the half Z-plane as illustrated in Fig. 7.

It is easy to see that B, C reverse ordering upon application of the map $t \longrightarrow Z$ since t_1 is a local maximum of f, and t_2 is an adjacent local minimum.

Consider $z \longrightarrow Z$ acting on Ω_2; refer to Fig. 7. CD and BA are curves for which $\operatorname{Im} g(z) = 0$; these two arcs bound Ω_2 on two sides. BE is the steepest ascent curve of $ig(z)$ from z_1 to $\infty e^{3\pi i/8}$. Under (4.1), Ω_2 maps to the half Z-plane depicted in Fig. 7.

As was the case for R_1 and Ω_1, we see that $f(t) = g(z)$ provides us with a one-to-one map from R_2 onto Ω_2.

Similar analysis applies to the mapping from R_3 to Ω_3 (see Fig. 8).

Finally, we turn to $R_4 \longrightarrow \Omega_4$. In Fig. 9, CE is the steepest descent curve of if from t_0 to $\infty e^{13\pi i/10}$, and the corresponding curve in the z-plane is the steepest descent curve of ig from z_3 to $\infty e^{9\pi i/8}$.

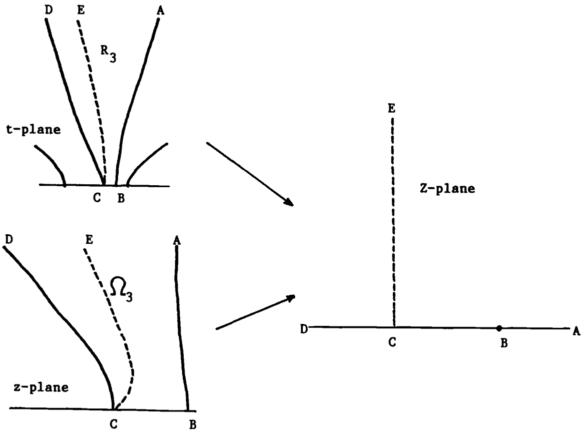


FIG. 8. *Effect of the mappings* $t \to Z$, $z \to Z$ *on the regions* R_3 *and* Ω_3, *respectively.*

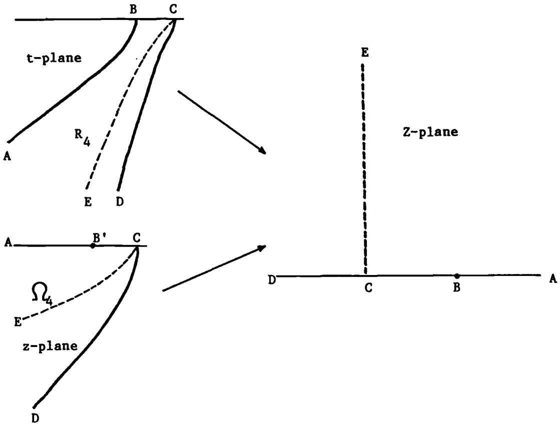


FIG. 9. *Effect of the mappings* $t \to Z$, $z \to Z$ *on the regions* R_4 *and* Ω_4, *respectively.*

We note here that f decreases steadily from $+\infty$ to $-\infty$ as t moves from $\mathrm{A} = \infty e^{6\pi i/5}$ to $\mathrm{B} = t_3$, to $\mathrm{C} = t_0$ and thence to $\mathrm{D} = \infty e^{7\pi i/5}$. Similarly, g decreases steadily from $+\infty$ to $-\infty$ as z moves from $\mathrm{A} = +\infty$ to $\mathrm{C} = z_3$ to $\mathrm{D} = \infty e^{5\pi i/4}$. Since $g(z_3) = f(t_0)$, this implies the existence of a point B' in the interval $]-\infty, z_3[$ such that $g(\mathrm{B}') = f(t_3)$. This is indicated in Fig. 9.

Thus, $R_4 \to \Omega_4$ in a one-to-one fashion. To obtain the mappings of the regions conjugate to the R_i to regions conjugate to the Ω_i, we merely "flip" all illustrations about the real t, z, and Z-axes, and replace "ascent" by "descent" and vice versa.

This concludes our look at the mapping (3.1) in the case where all t_i's are real and separated.

5. A limiting case. As an example of the use of the quartic transformation formulae developed in § 3, and as a check on the validity of our results, we determine the limiting form of the coefficients p_0, q_0, and r_0 in the approximation (3.36) when the parameters (b, c) tend to the cusp $(4/3\sqrt{6}, -1/12)$. This calculation showcases the determination of the parameters ζ, η, and θ, and makes explicit use of the formula for ζ presented as (3.30).

We begin by taking (b, c) to lie on the caustic. From the discussion leading to (2.9), it is readily seen that the $\psi = -\pi/6$ level curve in the bc-plane contains the lower arch of the caustic joining $(0,0)$ to $(4/3\sqrt{6}, -1/12)$, and the $\psi = \pi/6$ level curve contains the upper arch of the caustic joining $(4/3\sqrt{6}, -1/12)$ to $(0, \frac{1}{4})$; see Fig. 6. We will take (b, c) to lie on that part of the caustic joining $(0,0)$ to $(4/3\sqrt{6}, -1/12)$ so that $\psi = -\pi/6$ with $c \neq -1/12$ ($c = -1/12$ would place us at the cusp). With these choices for ψ and c, it is easy to see that (2.7) becomes

$$c/6 + b^2/16 - 1/216 = -(c/3 + 1/36)^{3/2},$$

a quadratic equation in b. We set $c = \Delta c - 1/12$. Use of this in the previous equation implies that $b^2 = \frac{8}{27}\left[1 - 9\Delta c - 54(\Delta c/3)^{3/2}\right]$, so if b is positive, we may write

$$b = \frac{4}{3\sqrt{6}}\left[1 - 9\Delta c - 54(\Delta c/3)^{3/2}\right]^{1/2} \tag{5.1}$$

for a point (b, c) on the arch of the caustic under examination. We will use (5.1) extensively to develop $\Delta c \to 0^+$ limits for a variety of quantities needed in the computation of p_0, q_0, and r_0 in (3.36). Before proceeding, we shall determine expressions for p_0, q_0, and r_0 on the caustic.

First, on the $\psi = -\pi/6$ portion of the caustic, we have $t_0 = t_1 < t_2$ so that $z_3 = z_2 < z_1$ in view of the correspondence $t_2 \leftrightarrow z_1, t_1 \leftrightarrow z_2, t_0 \leftrightarrow z_3$. Furthermore, from $\phi = -\pi/6$, we have $\eta = -2(\zeta/3)^{3/2}$ (cf. eqn. (3.3)), $z_1 = 2\sqrt{\zeta/3}$, and $z_2 = z_3 = -\sqrt{\zeta/3}$.

From the first equation of (3.33), we get

$$g_0(z) = \frac{dt}{dz} = p_0 + q_0 z + r_0 z^2 + (z^3 - \zeta z - 2(\zeta/3)^{3/2})h_0(z), \tag{5.2}$$

which implies

$$\begin{aligned} g_0(z_1) &= p_0 + 2\sqrt{\zeta/3}q_0 + 4\zeta r_0/3, \\ g_0(z_2) &= p_0 - \sqrt{\zeta/3}q_0 + \zeta r_0/3, \end{aligned} \tag{5.3}$$

in view of the fact that $(z^3 - \zeta z - 2(\zeta/3)^{3/2})h_0(z)$ vanishes at z_1 and z_2. As (5.3) forms a pair of linear equations in the three unknowns p_0, q_0, and r_0, (5.3) is insufficient for determining p_0, q_0, and r_0 uniquely. A third (linearly) independent equation can be found by differentiating (5.2) with respect to z, and then evaluating the result to obtain

$$g_0'(z_2) = q_0 - 2\sqrt{\frac{\zeta}{3}}r_0. \tag{5.4}$$

Thus, (5.3) and (5.4) determine p_0, q_0, and r_0. Upon forming the difference of the equations in (5.3), we find $g_0(z_1) - g_0(z_2) = 3\sqrt{\zeta/3}q_0 + \zeta r_0$, so that subsequent use of (5.4) yields

$$r_0 = \frac{1}{3\zeta}\left[g_0(z_1) - g_0(z_2) - 3\sqrt{\frac{\zeta}{3}}g_0'(z_2)\right]. \tag{5.5}$$

Use of this in (5.4) then provides us with

$$q_0 = \frac{g_0'(z_2)}{3} + \frac{2}{\sqrt{27\zeta}}[g_0(z_1) - g_0(z_2)]. \tag{5.6}$$

Finally, (5.5) and (5.6) together in (5.3) gives us

$$p_0 = \frac{g_0(z_1)}{9} + \frac{8g_0(z_2)}{9} + \frac{2}{3}\sqrt{\frac{\zeta}{3}}g_0'(z_2). \tag{5.7}$$

Thus, once we have determined $g_0(z_1), g_0(z_2)$, and $g_0'(z_2)$, equations (5.5)–(5.7) will provide us with values for our coefficients.

To calculate the $g_0(z_i)$ and $g_0'(z_2)$, we return to the mapping (3.1). If we differentiate (3.1) twice with respect to z (bearing in mind that $g_0(z) = dt/dz$), we obtain $f''(t)g_0^2(z) + f'(t)g_0'(z) = 3z^2 - \zeta$. Evaluation at $t = t_0 \leftrightarrow z = z_2$ yields no information (since $f'(t_0) = f''(t_0) = 0$ at the order two saddle t_0), while evaluation at $t = t_2 \leftrightarrow z = z_1$ yields $f''(t_2)g_0^2(z_1) = 3z_1^2 - \zeta = 3\zeta$, since $z_1 = 2\sqrt{\zeta/3}$, or $g_0(z_1) = \pm\sqrt{3\zeta/f''(t_2)}$. Now, $f''(t_2) > 0$ since t_2 is a local minimum of f; hence, the ratio inside the square root is positive. Because z must increase with t, the positive square root must be extracted and so we have

$$g_0(z_1) = \sqrt{\frac{3\zeta}{f''(t_2)}}. \tag{5.8}$$

To continue, we differentiate $f''(t)g_0^2(z) + f'(t)g_0'(z) = 3z^2 - \zeta$ again with respect to z and evaluate the result at $z = z_2 \leftrightarrow t = t_0$ to obtain

$$g_0(z_2) = \sqrt[3]{\frac{-6\sqrt{\zeta/3}}{f'''(t_0)}}. \tag{5.9}$$

If we repeat this process again, we find, after some arithmetic, that

$$g_0'(z_2) = \frac{6 - f''''(t_0)g_0^4(z_2)}{6f'''(t_0)g_0^2(z_2)}. \tag{5.10}$$

It is clear that $g_0(z_1), g_0(z_2)$, and $g_0'(z_2)$ can be expressed in terms of f, its derivatives, and ζ.

We turn to the calculation of these latter quantities. To begin, we apply the binomial theorem to (5.1) to find

$$\begin{aligned} b \ &= \ \frac{4}{3\sqrt{6}}\left[1 - \frac{9}{2}\Delta c - \frac{9}{\sqrt{3}}(\Delta c)^{3/2} - \frac{81}{8}(\Delta c)^2\right. \\ &\qquad \left. - \frac{81}{2\sqrt{3}}(\Delta c)^{5/2} - \frac{27\cdot 35}{16}(\Delta c)^3 + O(\Delta c)^{7/2}\right]. \end{aligned} \tag{5.11}$$

For the limiting forms of the saddles t_0 and t_2, we note that $\psi = -\pi/6$ implies, from (2.7), that $e_1 = \sqrt{\Delta c/3}$, $e_2 = e_3 = -\frac{1}{2}\sqrt{\Delta c/3}$; recall that we have set $c = \Delta c - 1/12$.
Use of these observations in (2.4) provides us with

$$t_0 = \frac{1}{\sqrt{6}}\sqrt{1 - 6\sqrt{\Delta c/3}}.$$

If we apply the binomial theorem to this expression for t_0, we find that

$$\begin{aligned} t_0 - \frac{1}{\sqrt{6}} &= \frac{-1}{\sqrt{6}} \Bigg[\sqrt{3}(\Delta c)^{1/2} + \frac{3}{2}(\Delta c) + \frac{9}{2\sqrt{3}}(\Delta c)^{3/2} + \frac{45}{8}(\Delta c)^2 \\ &\qquad + \frac{7 \cdot 27}{8\sqrt{3}}(\Delta c)^{5/2} + \frac{21 \cdot 27}{16}(\Delta c)^3 + O(\Delta c)^{7/2} \Bigg] \end{aligned}$$

as $\Delta c \to 0$, and in similar fashion for small Δc,

$$\begin{aligned} t_2 - \frac{1}{\sqrt{6}} &= \frac{1}{\sqrt{6}} \Bigg[2\sqrt{3}(\Delta c)^{1/2} + \frac{3}{4}(\Delta c) + \frac{15\sqrt{3}}{8}(\Delta c)^{3/2} + \frac{7 \cdot 45}{64}(\Delta c)^2 \\ &\qquad + \frac{17 \cdot 63\sqrt{3}}{128}(\Delta c)^{5/2} + \frac{21 \cdot 27 \cdot 31}{512}(\Delta c)^3 + O(\Delta c)^{7/2} \Bigg] . \end{aligned}$$

We are now in a position to calculate $f(t_0) - f(t_2)$, which we shall see is required in the calculation of ζ. We accomplish this by first writing f as its Taylor expansion about $t = 1/\sqrt{6}$. We replace b in the result of this computation by (5.11) and compute $f(1/\sqrt{6}), f'(1/\sqrt{6})$, and $f''(1/\sqrt{6})$; this results in approximations in terms of powers of $(\Delta c)^{1/2}$. Use of these approximations for f and its derivatives at $1/\sqrt{6}$, together with the preceding small Δc approximations for t_0 and t_2, in the Taylor series for f gives us small Δc approximations for $f(t_0)$ and $f(t_2)$. Upon taking the outcome of these two (involved) calculations and forming their difference, we find that

$$(5.12) \quad f(t_0) - f(t_2) = \frac{27}{4\sqrt{6}}(\Delta c)^2 \cdot \left[1 + \frac{6\sqrt{3}}{5}(\Delta c)^{1/2} + \frac{45}{8}(\Delta c) + O(\Delta c)^{3/2} \right] .$$

Since $\sigma_1^2 - 3\sigma_2 = (f(t_0) - f(t_2))^2$ on this portion of the caustic (see the discussion following (3.25)), we have $(\sigma_1^2 - 3\sigma_2)^{1/2} = f(t_0) - f(t_2)$.

We observe that (3.25) is equivalent to the expression $\mathcal{N} = \mathcal{D}$, where $\mathcal{N}$ and $\mathcal{D}$ are displayed in (3.22). This in turn implies that χ, in (3.23), is zero. Hence, (5.12) together with $\chi = 0$ and (3.30) yields

$$(5.13) \qquad \zeta = \frac{3^{3/4}}{2^{1/4}}(\Delta c) \left[1 + \frac{3\sqrt{3}}{5}(\Delta c)^{1/2} + \frac{9 \cdot 101}{400}(\Delta c) + O(\Delta c)^{3/2} \right] .$$

With ζ in hand, we can proceed to the calculation of $g_0(z_1), g_0(z_2)$, and $g_0'(z_2)$. We obtain $g_0(z_1)$ first. From the Taylor expansion of f centered at $t = 1/\sqrt{6}$, the small Δc approximation for t_2 and (5.11), we obtain

$$(5.14) \quad f''(t_2) = \frac{18}{\sqrt{6}} \left[(\Delta c) + \sqrt{3}(\Delta c)^{3/2} + \frac{45}{16}(\Delta c)^2 + \frac{33\sqrt{3}}{8}(\Delta c)^{5/2} + O(\Delta c)^3 \right] ,$$

and dividing this into 3ζ gives $3\zeta/f''(t_2) = (3^{1/4}/2^{3/4})[1 - \frac{2}{5}\sqrt{3}(\Delta c)^{1/2} + \frac{33}{50}(\Delta c) + O(\Delta c)^{3/2}]$, from which (5.8) yields

$$(5.15) \qquad g_0(z_1) = \frac{3^{1/8}}{2^{3/8}} \left[1 - \frac{\sqrt{3}}{5}(\Delta c)^{1/2} + \frac{27}{100}(\Delta c) + O(\Delta c)^{3/2} \right] .$$

To get $g_0(z_2)$ requires the limiting behaviour of $f'''(t_0)$ as $\Delta c \to 0$. The Taylor series for f centered at $1/\sqrt{6}$, the small Δc approximation of t_0 and (5.11) yield $f'''(t_0) = -4\sqrt{3}(\Delta c)^{1/2} + O(\Delta c)^{7/2}$; from this and (5.13), we find

$$g_0(z_2) = \frac{3^{1/8}}{2^{3/8}}\left[1 + \frac{\sqrt{3}}{10}(\Delta c)^{1/2} + \frac{3\cdot 81}{800}(\Delta c) + O(\Delta c)^{3/2}\right]. \tag{5.16}$$

To obtain $g_0'(z_2)$, we use the Taylor expansion for $f^{(iv)}$ centered at $t = 1/\sqrt{6}$ and put $t = t_0$. With the small Δc approximation for t_0, we have $f^{(iv)}(t_0) = (24/\sqrt{6})\left[1 - \sqrt{3}(\Delta c)^{1/2} - 3(\Delta c)/2 + O(\Delta c)^{3/2}\right]$. This and (5.10) gives us

$$g_0'(z_2) = -\frac{3^{3/4}}{2^{5/4}\cdot 5}\left[1 + \frac{21\sqrt{3}}{40}(\Delta c)^{1/2} + O(\Delta c)\right]. \tag{5.17}$$

With $g_0(z_1), g_0(z_2)$, and $g_0'(z_2)$ at our disposal, we are in a position to calculate the coefficients p_0, q_0, and r_0. Upon assembling the preceding approximations, and substituting into equations (5.5), (5.6), and (5.7), we find

$$r_0 = \frac{3^{3/8}63}{2^{1/8}32\cdot 25}[1 + O(\Delta c)^{1/2}], \tag{5.18}$$

$$q_0 = -\frac{3^{3/4}}{2^{5/4}5}[1 + O(\Delta c)^{1/2}], \tag{5.19}$$

and

$$p_0 = \frac{3^{1/8}}{2^{3/8}}[1 + O(\Delta c)^{1/2}]. \tag{5.20}$$

To compare the limit, as $\Delta c \to 0$, of (3.36) with the classically obtained result (2.12) requires the calculation of $P(0,0), P_x(0,0)$, and $P_y(0,0)$, where P is the Pearcey integral given in equation (1.3). Standard techniques give

$$P(0,0) = \frac{\Gamma(1/4)}{\sqrt{2}}e^{\pi i/8}; \quad P_x(0,0) = \frac{i\Gamma(3/4)}{\sqrt{2}}e^{3\pi i/8}; \quad P_y(0,0) = 0,$$

so that use of these values for the Pearcey function and its derivatives, along with (5.18)–(5.20), in (3.36) gives us

$$\begin{aligned}&\left[\frac{p_0(4/3\sqrt{6}, -1/12)}{\lambda^{1/4}}P(0,0) - \frac{i\cdot q_0(4/3\sqrt{6}, -1/12)}{\lambda^{1/2}}P_y(0,0)\right.\\ &\left.\quad - \frac{2i\cdot r_0(4/3\sqrt{6}, -1/12)}{\lambda^{3/4}}P_x(0,0)\right] e^{i\lambda f(1/\sqrt{6})}\\ &= \frac{3^{1/8}\Gamma(1/4)}{2^{7/8}\lambda^{1/4}}e^{\pi i/8 - i\lambda/(45\sqrt{6})} + \frac{2^{3/8}3^{3/8}63\Gamma(3/4)}{32\cdot 25\lambda^{3/4}}e^{3\pi i/8 - i\lambda/(45\sqrt{6})}.\end{aligned} \tag{5.21}$$

Equation (5.21), together with $f(-3/\sqrt{6}) = 7/5\sqrt{6}$ and $f''(-3/\sqrt{6}) = -32/3\sqrt{6}$ ($t_3 = -3/\sqrt{6}$ at the cusp), shows that (3.36) agrees with the classically obtained approximation given in (2.12).

6. Summary and closing remarks. Collecting the work of §§2–5, we have established the result:

Let S be the swallowtail integral defined in (1.1). *Then, for positive y and large positive x,*

$$\begin{aligned}\frac{S(-x,y,z)}{\sqrt{x}} &= e^{ix^{5/2}f(t_3;yx^{-3/2},zx^{-2})-\pi i/4}\sqrt{\frac{-2\pi}{x^{5/2}f''(t_3;yx^{-3/2},zx^{-2})}}\\ &\quad\cdot\left[1+O(x^{-5/4})\right]\\ &\quad+e^{ix^{5/2}\theta}\left[p_0P(-x^{5/4}\zeta,x^{15/8}\eta)x^{-5/8}-iq_0P_y(-x^{5/4}\zeta,x^{15/8}\eta)x^{-5/4}\right.\\ &\quad\left.-2ir_0P_x(-x^{5/4}\zeta,x^{15/8}\eta)x^{-15/8}\right]\cdot\left[1+O(x^{-5/2})\right],\end{aligned}$$

where $f(t;b,c)=t^5/5-t^3/3+bt^2/2+ct$, t_3 is the negative root of $t^4-t^2+yx^{-3/2}t+zx^{-2}=0$, and the functions ζ, η, and θ are given by equations (3.30), (3.15) *and* (3.14) *respectively* (*information about the branches that must be used in these formulae can be found in the discussion following equations* (3.29) *and* (3.30)). *The coefficients p_0, q_0, and r_0 are defined by equations* (3.33) *and satisfy, at the cusp of the caustic $4x^{3/2}=3\sqrt{6}\,y$, $z=-x^2/12$,*

$$p_0=3^{1/8}/2^{3/8};\quad q_0=-3^{3/4}/(2^{5/4}5);\quad r_0=3^{3/8}63/(2^{1/8}32\cdot 25)\ .$$

The function P is the Pearcey integral given as equation (1.3). *This asymptotic approximation of $S(-x,y,z)$ remains uniformly valid for large positive x, for y and z in a neighborhood of the cusp of the caustic, containing a disk of the form $(yx^{-3/2}-4/3\sqrt{6})^2+(zx^{-2}+1/12)^2\le\epsilon^2$ for some $\epsilon>0$.*

For negative y, the asymptotic behaviour can be obtained by forming the complex conjugate of the approximation for $S(-x,-y,z)$.

Uniform asymptotic expansions of the derivatives of $S(x,y,z)$ are also readily obtained from the expansions developed in this work. Because the integral I (cf. § 2.2) is analytic in *all* of its arguments (indeed, it is entire in the parameters b and c), the coefficients in the expansion of I are analytic in their parameters (b,c for the p_n, q_n and r_n). Thus, we need only differentiate our expansion termwise to obtain expansions of the derivatives of S; see [Urs, p. 52]. However, this requires the introduction of no new techniques, and so has been excluded from our (already lengthy) discussion.

We close this work by noting that, although the use of Greenhill's work was not necessary for the development of the expansion of I, it proved convenient at times. Further, it appears that elliptic functions provide a means of approaching the problem of expressing the zeros of higher degree polynomials (unsolvable by radicals for degree ≥ 5) as analytic functions of a polynomial's coefficients; see [Kie]. This consideration is of paramount importance for the development of uniform expansions of integrals such as

$$\int_{-\infty}^{+\infty}e^{i(t^6/6+wt^4/4+xt^3/3+yt^2/2+zt)}dt,$$

the next "canonical diffraction integral" in the suite of generalized Airy functions (termed the "butterfly" integral).

Acknowledgments. The author is grateful for the numerous improvements to [Kam1] suggested by Professors R. Wong, J. P. McClure, and F. W. J. Olver. The author is particularly indebted to Professor Wong for his guidance throughout this work.

REFERENCES

[Bri] L. BRILLOUIN, *Sur une méthode de calcul approchée de certaines intégrales, dite méthode de col*, Ann. Éc. Norm, (3), XXXIII (1916), pp. 17–69.

[CFU] C. CHESTER, B. FRIEDMAN, AND F. URSELL, *An extension of the method of steepest descents.* Proc. Cambridge Phil. Soc., 53 (1957), pp. 599–611.

[Cop] E. T. COPSON, *An Introduction to the Theory of Functions of a Complex Variable*, Clarendon Press, Oxford, 1975.

[Con1] J. N. L. CONNOR AND D. FARRELLY, *Molecular collisions and cusp catastrophes: Three methods for the calculation of Pearcey's integral and its derivatives*, Chem. Phys. Lett., 81 (1981), pp. 306–310.

[Con2] J. N. L. CONNOR, P. R. CURTIS AND D. FARRELLY, *A differential equation method for the numerical evaluation of the Airy, Pearcey and swallowtail canonical integrals and their derivatives*, Molecular Phys., 48 (1983), pp. 1305–1330.

[Con3] J. N. L. CONNOR, P. R. CURTIS AND D. FARRELLY, *The uniform swallowtail approximation: practical methods for oscillating integrals with four coalescing saddle points*, J. Phys. A: Math. Gen., 17 (1984), pp. 283–310.

[Con4] J. N. L. CONNOR, *Practical methods for the uniform asymptotic evaluation of oscillating integrals with several coalescing saddle points*, in Asymptotic and Computational Analysis, Lecture Notes in Pure and Applied Math. Series 124, R. Wong, ed., Marcel–Dekker, New York, 1990, pp. 137–173.

[Gil] R. GILMORE, *Catastrophe Theory for Scientists and Engineers*, John Wiley and Sons, New York, 1981.

[Gre1] A. G. GREENHILL, *Solution of the cubic and quartic equations by means of Weierstrass's elliptic functions*, Proc. London Math. Soc., 18 (1886), pp. 262–287.

[Gre2] A. G. GREENHILL, *The Applications of Elliptic Functions*, MacMillan and Co., London, 1892.

[Kam1] D. KAMINSKI, *Asymptotic Expansions of Some Canonical Diffraction Integrals*, Ph.D thesis, University of Manitoba, Winnipeg, Manitoba, 1987.

[Kam2] ______, *Asymptotic expansion of the Pearcey integral near the caustic*, Siam J. Math. Anal., 20 (1989), pp. 987–1005.

[Kie] L. KIEPERT, *Auflösung der Gleichungen fünften Grades.* J. Reine Angew. Math., 87 (1879), pp. 114–133.

[Lev] N. LEVINSON, *Transformation of an analytic function of several variables to a canonical form*, Duke Math. J., 28 (1961), pp. 345–353.

[Olv] F. W. J. OLVER, *Why steepest descents?* SIAM Rev., 12 (1970), pp. 228–247.

[Par] R. B. PARIS, *The Asymptotics of Pearcey's integral for complex variables*, in Asymptotic and Computational Analysis, Lecture Notes in Pure and Applied Math., Series 124, R. Wong, ed., Marcel–Dekker, New York (1990), pp. 653–667.

[Pea] T. PEARCEY, *The structure of an electromagnetic field in the neighborhood of a cusp of a caustic*, Lond. Edinb. Dubl. Phil. Mag., 37 (1946), pp. 311–317.

[Sta] J. J. STAMNES AND B. SPJELKAVIK, *Evaluation of the field near a cusp of a caustic*, Optica Acta, 30 (1983), pp. 1331–1358.

[Urs] F. URSELL, *Integrals with a large parameter. Several nearly coincident saddle points*, Proc. Cambridge Phil. Soc., 72 (1972), pp. 49–65.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 287–325, March 1992

001

ON THE OPTIMAL DESIGN OF COLUMNS AGAINST BUCKLING*

STEVEN J. COX† AND MICHAEL L. OVERTON‡

Abstract. The authors establish existence, derive necessary conditions, infer regularity, and construct and test an algorithm for the maximization of a column's Euler buckling load under a variety of boundary conditions over a general class of admissible designs. It is proven that symmetric clamped-clamped columns possess a positive first eigenfunction and a symmetric rearrangement is introduced that does not decrease the column's buckling load. The necessary conditions, expressed in the language of Clarke's generalized gradient [10], subsume those proposed by Olhoff and Rasmussen [25], Masur [22], and Seiranian [34]. The work of [25], [22], and [34] sought to correct the necessary conditions of Tadjbakhsh and Keller [37], who had not foreseen the presence of a multiple least eigenvalue. This remedy has been hampered by Tadjbakhsh and Keller's miscalculation of the buckling loads of their clamped-clamped and clamped-hinged columns. This issue is resolved in the appendix.

In the numerical treatment of the associated finite-dimensional optimization problem the authors build on the work of Overton [26] in devising an efficient means of extracting an ascent direction from the column's least eigenvalue. Owing to its possible multiplicity, this is indeed a nonsmooth problem and again the ideas of Clarke [10] are exploited.

Key words. eigenvalue, generalized gradient

AMS(MOS) subject classifications. 34, 49, 65, 73

1. Introduction. We recall Pearson's formulation [38, p. 66] of the following problem of Lagrange,

> To find the curve which by its revolution about an axis in its plane determines the column of greatest efficiency.

For columns of unit length and volume, efficiency here denotes the structure's resistance to buckling under axial compression. When λ is the magnitude of the axial load and u the resulting transverse displacement, we postulate the potential energy

$$\int_0^1 EI|u''|^2\,dx - \lambda\int_0^1 |u'|^2\,dx$$

with the two terms measuring bending and elongation respectively. Here I is the second moment of area of the column's cross section and E is its Young's modulus. For sufficiently small λ, the minimum of this potential energy, over all admissible displacements, is zero. The (Euler) buckling load of the column is the greatest λ, call it λ_1, for which this minimum is zero. That is,

$$\lambda_1 = \inf_{u\in V}\frac{\int_0^1 EI|u''|^2\,dx}{\int_0^1 |u'|^2\,dx}, \tag{1.1}$$

where V is a closed subspace of H^2, the space of L^2 functions on the interval $(0,1)$ with first and second distributional derivatives in L^2, distinguished by the choice of boundary conditions. The choice that has generated the greatest interest is the clamped-clamped condition $u(0) = u'(0) = u(1) = u'(1) = 0$. With the corresponding

* Received by the editors June 14, 1990; accepted for publication (in revised form) January 30, 1991. This work was supported by Defense Advanced Research Projects Agency contract F49620-87-C-0065, Air Force Office of Scientific Research grant 89-0363, and National Science Foundation grant CCR-9107640.

† Department of Mathematical Sciences, Rice University, P.O. Box 1872, Houston, Texas 77251.

‡ Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, New York 10012.

V denoted by H_0^2, it is not difficult to show that the infimum in (1.1) is attained at some $u_1 \in H_0^2$. First order necessary conditions then require that u_1 satisfy

$$\int_0^1 EIu_1''v''\,dx = \lambda_1 \int_0^1 u_1'v'\,dx \quad \forall v \in H_0^2. \tag{1.2}$$

When I and E are smooth it follows from (1.2) that

$$-(EIu_1'')'' = \lambda_1 u_1'', \qquad u_1(0) = u_1'(0) = u_1(1) = u_1'(1) = 0. \tag{1.3}$$

With this we recognize (1.1) as Rayleigh's principle for the least eigenvalue of (1.3) and u_1 as an associated first eigenfunction. For the problem of Lagrange, the Young's modulus is assumed constant and, as the column is a solid of revolution, each cross section's second moment of area is simply a constant multiple of the square of its area, A, i.e., $I(x) = cA^2(x)$. Fixing our attention on columns of unit volume, we require

$$\int_0^1 A\,dx = 1. \tag{1.4}$$

We have reduced the problem of Lagrange to the search for that A which, subject to (1.4), maximizes the λ_1 of (1.1). This problem, with clamped-clamped boundary conditions, was first attacked in 1962 by Tadjbakhsh and Keller [37] in the continuation of work Keller [20] had begun at the suggestion of Clifford Truesdell. The work of [37] hinges on the necessary condition that the best A, and its corresponding eigenfunction u, satisfy

$$A^4|u''|^2 = A^3 \tag{1.5}$$

along the entire column. This was obtained on formally differentiating a second-order analogue (see eqn. (2.5)) of (1.3) with respect to A, subject to the integral constraint. Upon reconciling (1.3) and (1.5), Tadjbakhsh and Keller arrived at the representation

$$\begin{aligned} A(x) &= \tfrac{4}{3}\sin^2\theta(x), \qquad -\pi/2 \le \theta \le 3\pi/2, \\ \theta(x) - \tfrac{1}{2}\sin 2\theta(x) + \pi/2 &= 2\pi x, \qquad 0 \le x \le 1. \end{aligned} \tag{1.6}$$

The most striking aspect of this claim is that it requires the cross-sectional area to vanish at $\frac{1}{4}$ and $\frac{3}{4}$. This result should however come as no surprise, for implicit in (1.5) is the assumption that the optimal buckling load is simple, i.e., that the corresponding space of buckled configurations is one-dimensional. This requires the optimal column to buckle in much the same way as the uniform column ($A \equiv 1$), the first eigenfunction of which is $U(x) = 1 - \cos(2\pi x)$. The fact that A vanishes at the inflection points of U agrees then with the heuristic (suggested by (1.1)) that the optimal column need be thick only in regions where it bends, i.e., where the magnitude of the linearized curvature $|u''|$ is large.

Tadjbakhsh and Keller claimed $16\pi^2/3$ as the buckling load of the resulting column. It was not until 1977 that Olhoff and Rasmussen [25], observing that (1.3) does not exclude multiple eigenvalues, noted that as the least eigenvalue does not vary smoothly with A at points where its multiplicity exceeds one, the formal differentiation in [37] would be hard to justify. As evidence that Tadjbakhsh and Keller had indeed taken the wrong course, Olhoff and Rasmussen claimed, on the basis of numerical work, 30.51 for the buckling load of the column constructed according to (1.6). Unfortunately, they neglected to describe the means by which this value was arrived at. Indeed, the fact that A vanishes at $\frac{1}{4}$ and $\frac{3}{4}$ introduces computational difficulties. Although they did go on to suggest how $16\pi^2/3$ was incorrectly obtained, a number of workers have remained unconvinced, e.g., Myers and Spillers [24] and Barnes [4]. Upon fleshing out the relevant remarks of Olhoff and Rasmussen we shall see, in work relegated to an appendix, that the buckling load for the column proposed by Tadjbakhsh and Keller does not exceed $\pi^2/3$. These same arguments will serve to demonstrate that Tadjbakhsh and Keller's best clamped-hinged column also has a much lower buckling load than thought previously.

Having concluded that differentiating (1.3) would lead to less than optimal columns, Olhoff and Rasmussen presented a "bimodal formulation" of the problem of Lagrange, i.e., one that would accommodate double eigenvalues. Their subsequent necessary condition paired the best A with two corresponding linearly independent eigenfunctions u, v and a scalar $t \in [0,1]$ so that

$$A\left(t|u''|^2 + (1-t)|v''|^2\right) = 1 \tag{1.7}$$

along the entire column. On implementing an algorithm that enforced this optimality condition, Olhoff and Rasmussen arrived at a column whose cross sectional area was positive throughout and which could withstand loads up to 52.3563. Their methods were, however, no more rigorous than those of Tadjbakhsh and Keller, and moreover, solely on the basis of claims, the latter still had the stronger column, for $52.3563 < 16\pi^2/3$. Those persuaded by Olhoff and Rasmussen's criticism of the work of Tadjbakhsh and Keller then set out to put (1.7) on a solid foundation. Actually, they joined the discussion of the more general problem: What conditions are necessary for a multiple eigenvalue to attain its extremum? The greatest advances on this question have come in finite dimensions and lie in the apparently little-known work of Bratus and Seiranian [7]. These conditions, later discovered independently in a more general form by Overton [26], will be discussed in detail in §5. For now, we note that Bratus and Seiranian, upon applying their finite-dimensional arguments to the problem of Lagrange, arrived at the conclusion that the best A must, with two corresponding orthogonal eigenfunctions u, v, satisfy

$$A(\delta_1|u''|^2 + \delta_2|v''|^2 + \delta_3 u''v'') = 1, \quad \text{where } \delta_1\delta_2 \geq \frac{\delta_3^2}{4}. \tag{1.8}$$

This condition was also proposed by Masur [22] who, like Seiranian [34], went on to represent the best A via a system of transcendental equations. Their approximate solutions to these systems are in good agreement, with respective buckling loads of 52.3564 and 52.3565, with that proposed by Olhoff and Rasmussen [25]. Note that (1.7) and (1.8), with the introduction of a second buckling mode, possess mechanisms which, at least in principle, rule out the possibility of columns with vanishing cross sectional area.

Our main contribution to the problem of Lagrange is essentially twofold. We employ the generalized gradient of Clarke in (i) a rigorous derivation of the necessary conditions (1.8) and (ii) the construction of an efficient algorithm to solve the associated finite-dimensional optimization problem. Our initial focus on the clamped-clamped case will be extended in §5 to each of the boundary conditions considered by Tadjbakhsh and Keller.

In our discussion of the various optimality criteria something has been conspicuously lacking: the literature contains no proof of the existence of a best A for the problem of Lagrange. Before filling this gap we establish a number of preliminary results and look to a more general problem formulation.

2. The optimal design problem. The moment I is more precisely the second moment of area of the cross section about a line through its centroid normal to the plane of buckling. That is, denoting the cross section by $\Omega(x)$ with centroid at the origin, if η is a unit normal to the plane of buckling then

$$I(x) = \int_{\Omega(x)} |\eta^T y|^2 \, dy. \tag{2.1}$$

When Ω is a circle, in fact, when Ω is a regular polygon, this integral does not depend on η, and we find that I varies as the square of the cross-sectional area, A. On considering so called thin-walled columns we shall now see that I varies as an affine function of A. On the lateral surface of a cylinder with circular cross section of constant radius R we add a layer of variable thickness $\rho(x)$ with $\rho(x) \le \varepsilon R$, $\varepsilon \ll 1$. Neglecting powers of ρ greater than one we find $I(x) = \pi R^3 \rho(x) + \pi R^4$ and $A(x) = 2\pi R \rho(x) + \pi R^2$. Taking $\tilde{A}(x) = A(x) - \pi R^2/2$ for our design variable we find $I(x) = (R^2/2)\tilde{A}(x)$. The effect of this choice on the integral constraint is trivial. Of greater interest is that $\tilde{A}$, by construction, must satisfy the pointwise bounds

$$\frac{\pi R^2}{2} \le \tilde{A}(x) \le \frac{\pi R^2}{2} + 2\varepsilon\pi R^2. \tag{2.2}$$

It is not difficult to continue this line of reasoning and collect a number of examples where I varies as some power of A. We proceed then to consider the case where $EI = \sigma^p$ for some $p > 0$. Compelled by our examination of the previous special cases, we admit those σ in

$$ad = \left\{ \sigma \in L^\infty : 0 < \alpha \le \sigma(x) \le \beta, \ \int_0^1 \sigma(x)\, dx = 1 \right\}.$$

The weak formulation of the buckled column equation for $\sigma \in ad$ is

$$\int_0^1 \sigma^p u'' v'' \, dx = \lambda \int_0^1 u' v' \, dx \quad \forall v \in H_0^2. \tag{2.3}$$

As $\sigma \in L^\infty$ and $\alpha > 0$, (2.3) possesses the sequence of eigenvalues

$$0 < \lambda_1(\sigma) \le \lambda_2(\sigma) \le \cdots \uparrow \infty,$$

repeated according to their finite multiplicities and a corresponding sequence of eigenfunctions $\{u_k(\sigma)\}_{k=1}^\infty \subset H_0^2$, orthonormal in terms of the bilinear form associated with the right side of (2.3). As $H_0^2(0,1) \subset C^1([0,1])$ we find $u_k \in C^1([0,1])$. Upon integrating by parts on the right side of (2.3) we find that $\sigma^p u_k''$ differs from $-\lambda_k(\sigma) u_k$ by an affine function of x. Hence, $\sigma^p u_k'' \in C^1([0,1])$, and, in fact,

$$(\sigma^p u_k'')(x) = (\sigma^p u_k'')'(0)x + (\sigma^p u_k'')(0) - \lambda_k(\sigma) u_k(x). \tag{2.4}$$

We collect those eigenfunctions corresponding to $\lambda_1(\sigma)$ in

$$\mathcal{E}(\sigma) = \operatorname{span}\{u_k(\sigma); \ \lambda_k(\sigma) = \lambda_1(\sigma)\},$$

a subspace of H_0^2 with dimension equal to the multiplicity of $\lambda_1(\sigma)$. Implicit in Olhoff and Rasmussen's bimodal formulation is the assumption that this multiplicity is at most two. Seiranian [34], has confirmed this through Kamke's analysis of the second-order problem with nonseparated boundary conditions

$$w'' + \lambda\sigma^{-p} w = 0, \quad w(1) = w(0) + w'(0), \quad w'(1) = w'(0). \tag{2.5}$$

This is the strong version of (2.3) with $w = \sigma^p u''$ and was first considered in our context by Tadjbakhsh and Keller. Kamke, in [19, §4], proves that the multiplicity of each eigenvalue of (2.5) is no greater than two. Equation (2.4), however, suggests an approach that applies directly to the weak formulation.

If, corresponding to $\lambda_k(\sigma)$, there existed three linearly independent eigenfunctions u_1, u_2, u_3 then it would be possible to choose scalars a, b, c not all zero such that $v = au_1 + bu_2 + cu_3$ satisfies, in addition to $v(0) = v'(0) = v(1) = v'(1) = 0$, the two conditions $(\sigma^p v'')'(0) = 0$, and $(\sigma^p v'')(0) = 0$. From (2.4) we conclude that v satisfies the homogeneous linear second order equation with zero initial conditions

$$\sigma^p(x)v''(x) + \lambda_k(\sigma)v(x) = 0, \qquad v(0) = v'(0) = 0.$$

As the only solution to this equation is the identically zero function, we have established

LEMMA 2.1. *If $\sigma \in ad$ then the multiplicity of $\lambda_k(\sigma)$ is at most two.*

As the least eigenvalue of the uniform column is $4\pi^2$, we find, as a consequence of the monotonicity of the Rayleigh quotient, that

$$4\pi^2\alpha^p \le \lambda_1(\sigma) \le 4\pi^2\beta^p \quad \forall\, \sigma \in ad. \tag{2.6}$$

Corresponding to the least eigenvalue $\lambda_1(\sigma)$, a positive eigenfunction is expected. Indeed, this is the *only* type that Tadjbakhsh and Keller expected. To our knowledge, however, there is no proof that a positive first eigenfunction need exist. We remark that on this point the oscillation theory of Kamke is insufficient, for it concludes only that eigenfunctions corresponding to the least nonzero eigenvalue of (2.5) possess either three *or* two zeros. This translates into either one *or* no zero(s) for eigenfunctions corresponding to $\lambda_1(\sigma)$. We now improve on this situation in the case where σ is even (about 1/2), i.e., $\sigma(x) = \sigma(1-x)$.

THEOREM 2.2. *If $\sigma \in L^\infty$ is even and admits a positive lower bound, then there exists a positive even eigenfunction corresponding to $\lambda_1(\sigma)$.*

Proof 2.2. We exploit the essential idea in inverse iteration, a popular technique for computing the least eigenvalue and eigenvector of a symmetric matrix. In our context this idea amounts to approximating the least eigenfunction by the solution of a related nonhomogeneous boundary value problem. Given $v_0 \in H_0^2$ we consider its expansion in the complete set of eigenfunctions $\{u_k(\sigma)\}$,

$$v_0(x) = \overline{v}(x) + \sum_{k=m+1}^{\infty} a_k u_k(x),$$

where m is the least integer for which $\lambda_m(\sigma) < \lambda_{m+1}(\sigma)$ and $\overline{v}$ is an eigenfunction corresponding to $\lambda_1(\sigma)$. From v_0 we construct the sequence $\{v_j\} \subset H_0^2$ according to

$$\int_0^1 \sigma^p v_j'' \phi''\, dx = \lambda_1(\sigma) \int_0^1 v_{j-1}' \phi'\, dx \quad \forall\, \phi \in H_0^2.$$

On expanding v_j in $\{u_k(\sigma)\}$ we find

$$v_j(x) = \overline{v}(x) + \sum_{k=m+1}^{\infty} a_k \left(\frac{\lambda_1(\sigma)}{\lambda_k(\sigma)}\right)^j u_k(x).$$

As $\lambda_1(\sigma) < \lambda_k(\sigma)$ for all $k > m$, we find that v_j converges pointwise to $\overline{v}$ as $j \to \infty$. It remains then to produce a v_0 whose corresponding $\overline{v}$ is even and positive.

Our choice for v_0 is the first eigenfunction of the uniform column, i.e., $1-\cos(2\pi x)$, a positive even function with exactly two inflection points.

LEMMA 2.3. *Let f be an even member of L^∞ with a positive lower bound and v be a positive, even member of H_0^2 with precisely two inflection points. If $u \in H_0^2$ satisfies*

$$\int_0^1 f u'' \phi''\, dx = \int_0^1 v' \phi'\, dx \quad \forall \phi \in H_0^2, \tag{2.7}$$

then u is positive, even, and possesses precisely two inflection points.

Proof 2.3. Upon integrating by parts on the right of (2.7) we find that fu'' differs from v by an affine function. Dividing by f and integrating twice gives

$$u(x) = \int_0^x (x-y)(ay+b-v(y))g(y)\, dy, \tag{2.8}$$

where $g = 1/f$ and a and b are determined by $u(1) = 0$ and $u'(1) = 0$, i.e., by

$$a\int_0^1 xg(x)\, dx + b\int_0^1 g(x)\, dx = \int_0^1 v(x)g(x)\, dx, \tag{2.9}$$

$$a\int_0^1 x^2 g(x)\, dx + b\int_0^1 xg(x)\, dx = \int_0^1 xv(x)g(x)\, dx. \tag{2.10}$$

That these equations uniquely determine a and b follows from Hölder's inequality

$$\left(\int_0^1 xg(x)\, dx\right)^2 < \int_0^1 g(x)\, dx \int_0^1 x^2 g(x)\, dx.$$

Our hypotheses, in fact, allow us to conclude that

$$a = 0, \qquad b = \frac{\int_0^1 v(x)g(x)\, dx}{\int_0^1 g(x)\, dx}.$$

This obviously satisfies (2.9). Regarding (2.10), recall that every even function satisfies $\int_0^1 \phi(x)\, dx = 2\int_0^1 x\phi(x)\, dx$. Consequently,

$$b = \frac{\int_0^1 v(x)g(x)\, dx}{\int_0^1 g(x)\, dx} = \frac{2\int_0^1 xv(x)g(x)\, dx}{2\int_0^1 xg(x)\, dx} = \frac{\int_0^1 xv(x)g(x)\, dx}{\int_0^1 xg(x)\, dx}$$

satisfies (2.10) as well. Labeling $s(x) = (b - v(x))g(x)$, equation (2.8), $u(1) = 0$, and $u'(1) = 0$ take the form

$$u(x) = \int_0^x (x-y)s(y)\, dy, \qquad \int_0^1 s(y)\, dy = 0, \qquad \int_0^1 ys(y) = 0.$$

With this and the fact that s is even, we find

$$\begin{aligned} u(1-x) &= \int_0^{1-x} (1-x-y)s(y)\,dy \\ &= \int_0^1 (1-x-y)s(y)\,dy - \int_{1-x}^1 (1-x-y)s(y)\,dy \\ &= \int_0^x (x-y)s(y)\,dy = u(x). \end{aligned}$$

Regarding the convexity/concavity of u, we observe that $f(x)u''(x) = b - v(x)$. That $b - v(x)$ has at least two zeros follows from $b > 0$, $v(0) = v(1) = 0$, and $0 < b < \|v\|_\infty$. For $b - v(x)$ to possess more than two zeros v must admit a local minimum, a condition that requires of v no less then four inflection points. These zeros, say x_0 and $1 - x_0$, are the inflection points of u. As u vanishes at zero and is convex on $(0, x_0)$, it must be positive there, and, by symmetry, positive on $(1 - x_0, 1)$ as well. As u is positive at x_0 and $1 - x_0$ while concave between these points it must be positive on this interval as well. □

Proof 2.2. It now follows that $\{v_j\}$ is a sequence of positive even functions. Because the convergence of v_j to $\overline{v}$ is pointwise, we conclude that $\overline{v}$ is itself a positive even function. □

When σ is even and $\lambda_1(\sigma)$ is simple, we now have, up to a scalar multiple, a unique positive even first eigenfunction; call it u_1. When $\lambda_1(\sigma)$ is double, in addition to u_1, there exists a first eigenfunction u_2 for which $\int_0^1 u_1' u_2'\,dx = 0$. Consequently, u_2 is not even, and as u_1 and u_2 span $\mathcal{E}(\sigma)$ we may conclude that when σ is even, there exists, up to a scalar multiple, a unique positive even first eigenfunction.

Though Theorem 2.2 applies only to even functions, we shall see in the next result that this suffices for our purposes. Note that Lemma 2.3 states that the operator $(d^2/dx^2(f\,d^2/dx^2))^{-1}(-d^2/dx^2)$ leaves a subcone of the positive H_0^2 functions invariant when f is even. This cone is, however, too "thin" to allow one to invoke Krein–Rutman arguments. Regarding possible improvements of Lemma 2.3, we note that even the constant coefficient operator $(d^4/dx^4)^{-1}(-d^2/dx^2)$ does *not* leave the positive H_0^2 functions invariant. To see this, we solve for b in (2.9)–(2.10) with $g = 1$,

$$b = 4\int_0^1 v\,dx - 6\int_0^1 xv\,dx.$$

Taking for v any smooth positive function supported in $(\frac{2}{3}, 1)$ produces $b < 0$. As $u(0) = u'(0) = 0$ and $u''(0) = b$, we conclude that u is not positive.

THEOREM 2.4. *Given $\sigma \in ad$ there exists an even $\sigma_* \in ad$ for which $\lambda_1(\sigma) \leq \lambda_1(\sigma_*)$.*

Proof. There is a very simple argument when $0 < p \leq 1$. Given a function ϕ on $(0, 1)$, we denote its even part by $\phi_s(x) = \frac{1}{2}(\phi(x) + \phi(1-x))$. Consider the even function $\tilde{\sigma} \equiv ((\sigma^p)_s)^{1/p}$ and its corresponding even first eigenfunction $\tilde{u}$. With the normalization $\|\tilde{u}'\| = 1$, we find

$$\lambda_1(\sigma) \leq \int_0^1 \sigma^p |\tilde{u}''|^2\,dx = \int_0^1 (\sigma^p)_s |\tilde{u}''|^2\,dx = \int_0^1 \tilde{\sigma}^p |\tilde{u}''|^2\,dx = \lambda_1(\tilde{\sigma}). \tag{2.11}$$

As $t \mapsto t^{1/p}$ is convex, we observe that

$$\tilde{\sigma}(x) = \left(\frac{\sigma^p(x)}{2} + \frac{\sigma^p(1-x)}{2} \right)^{1/p} \leq \frac{\sigma(x)}{2} + \frac{\sigma(1-x)}{2} = \sigma_s(x). \tag{2.12}$$

Now $\sigma_s \in ad$ and (2.11)–(2.12) imply that $\lambda_1(\sigma) \leq \lambda_1(\sigma_s)$. Our attempts to argue in a similar fashion for $p > 1$ with $\tilde{\sigma} \equiv ((\sigma^{-p})_s)^{-1/p}$ and (2.5) have been thwarted by the fact that $\lambda_1(\sigma)$ corresponds to the *third* eigenvalue of (2.5). What is needed is a rearrangement of σ that echoes the curvature of its corresponding first eigenfunction. To make this precise, we first need the following extension of Lemma 2.3.

Recall that a function ϕ is said to be odd about the point $(x_0, \phi(x_0))$ on some interval containing x_0 when

$$\phi(x_0) - \phi(x_0 - x) = \phi(x_0 - x) - \phi(x_0)$$

for each x on the given interval. If, in addition to the original hypotheses of Lemma 2.3, we assume that f and v, when restricted to $(0, \frac{1}{2})$ are even about $\frac{1}{4}$ and odd about $(\frac{1}{4}, v(\frac{1}{4}))$, respectively, we conclude that u, when restricted to $(0, \frac{1}{2})$, is odd about $(\frac{1}{4}, u(\frac{1}{4}))$.

To see this we recall that $\int_0^1 u''\,dx = 0$ and u'' is even about $\frac{1}{2}$, hence $\int_0^{1/2} u''\,dx = 0$. For the remainder of this paragraph all functions will be restricted to $(0, \frac{1}{2})$. Recall as well that $u'' = (b - v)/f$, the quotient of a function odd about $(\frac{1}{4}, b - v(\frac{1}{4}))$ and a function even about $\frac{1}{4}$. Hence u'' is odd about $(\frac{1}{4}, b - v(\frac{1}{4}))$. The condition that $\int_0^{1/2} u''\,dx = 0$ now forces $b = v(\frac{1}{4})$. As u'' is now odd about $(\frac{1}{4}, 0)$ and $u(0) = u'(0) = 0$, we easily conclude that u is indeed odd about $(\frac{1}{4}, u(\frac{1}{4}))$.

If σ_* is now even about $\frac{1}{2}$ and even about $\frac{1}{4}$ when restricted to $(0, \frac{1}{2})$, then beginning the iteration of Theorem 2.2 with a v_0 that is even about $\frac{1}{2}$ and odd about $(\frac{1}{4}, v_0(\frac{1}{4}))$, e.g., $1 - \cos(2\pi x)$ will produce u_*, a positive eigenfunction corresponding to $\lambda_1(\sigma_*)$ that is even about $\frac{1}{2}$ and odd about $(\frac{1}{4}, u_*(\frac{1}{4}))$ on $(0, \frac{1}{2})$. We immediately note that σ_*^p and $|u_*''|^2$ are similarly ordered, i.e.,

$$(\sigma_*^p(x_1) - \sigma_*^p(x_2))(|u_*''(x_1)|^2 - |u_*''(x_2)|^2) \geq 0 \quad \forall x_1, x_2 \in (0,1). \tag{2.13}$$

Given $\sigma \in ad$ we now define its rearrangement σ_*.

$$\ell_c = \{x \in (0,1) : \sigma(x) \geq c\},$$

$$\ell_c^* = \begin{cases} \{x \in R : |x - \frac{1}{2}| \leq \frac{1}{4}|\ell_c|\} & \text{if } \ell_c \neq \emptyset, \\ \emptyset & \text{otherwise} \end{cases}$$

$$\sigma_*(x) = \begin{cases} \sigma_*(\frac{1}{2} - x), & \text{if } 0 \leq x \leq \frac{1}{4} \\ \sup\{c \in R : x \in \ell_c^*\}, & \text{if } \frac{1}{4} \leq x \leq \frac{3}{4} \\ \sigma_*(1 - x), & \text{if } \frac{3}{4} \leq x \leq 1. \end{cases}$$

In essence, this distributes half of σ's mass in a symmetrically decreasing fashion about $\frac{1}{2}$ on $(\frac{1}{4}, \frac{3}{4})$, completing the rest via symmetry. By construction, these two functions are equimeasurable, i.e.,

$$|\{x \in (0,1) : \sigma(x) \geq c\}| = |\{x \in (0,1) : \sigma_*(x) \geq c\}| \quad \forall c \in R,$$

and consequently, $\sigma_* \in ad$. We are now in position to apply the following result of Hardy, Littlewood, and Pólya; see Pólya and Szegö [31, p. 153].

If f and f_1 are equimeasurable, g and g_1 are equimeasurable, $f \in L^q$, $g \in L^{q'}$, and f_1 and g_1 are similarly ordered, then

$$\int_0^1 fg\,dx \leq \int_0^1 f_1 g_1\,dx. \tag{2.14}$$

Given $\sigma \in ad$ we now rearrange it as above into σ_* and consider its corresponding $u_* \in \mathcal{E}(\sigma_*)$. Upon normalizing $\|u_*'\| = 1$ we find

$$\lambda_1(\sigma) \leq \int_0^1 \sigma^p |u_*''|^2\,dx \leq \int_0^1 \sigma_*^p |u_*''|^2\,dx = \lambda_1(\sigma_*).$$

The first inequality is a consequence of Rayleigh's principle, the second, of (2.13)–(2.14). □

The stage now set, we address, in the next two sections, existence and necessary conditions for the generalized problem of Lagrange

$$\sup_{\sigma\in ad} \lambda_1(\sigma). \tag{2.15}$$

3. Existence. We adopt the direct method of the calculus of variations and neglect to relabel subsequences. Denote by $\hat{\lambda}_1$ the value of (2.15) and by $\{\sigma_n\} \subset ad$ an associated maximizing sequence, i.e., $\lambda_1(\sigma_n) \uparrow \hat{\lambda}_1$. Without loss, we may assume that each σ_n is even about 1/2. We abbreviate $\lambda_1(\sigma_n)$ to $\lambda_{1,n}$ and denote by $u_{1,n}$ a corresponding positive eigenfunction for which $\|u_{1,n}'\| = 1$ and $\int_0^1 \sigma_n^p |u_{1,n}''|^2\,dx = \lambda_{1,n}$, where $\|\cdot\|$ denotes the L^2 norm. These normalizations, in light of (2.6), impose a uniform H^2 bound on the sequence $\{u_{1,n}\}$. As a result, there exists a subsequence with weak H^2 limit $\overline{u} \in H_0^2$. The imbedding of H^2 in H^1 being compact, we find $\|\overline{u}'\| = 1$, and so $\overline{u}$ is not identically zero. The natural question is whether $\hat{\lambda}_1$ and $\overline{u}$ are indeed an eigenvalue and eigenfunction for some column with corresponding $\overline{\sigma} \in ad$. If so, then $\overline{\sigma}$ is necessarily the desired optimal design. This question was first addressed by Senatorov [35] in the context of a second-order problem. He discovered that weak convergence of the reciprocals of the coefficients of the highest order term must be considered. This observation continues to hold for fourth order problems, the details of which we now sketch.

Consider the weak formulation

$$\int_0^1 \sigma_n^p u_{1,n}'' v''\,dx = \lambda_{1,n} \int_0^1 u_{1,n}' v'\,dx \quad \forall v \in H_0^2. \tag{3.1}$$

Our previous remarks reveal that the right-hand side converges to $\hat{\lambda}_1 \int_0^1 \overline{u}'v'\,dx$ for each such v. Regarding the left side we define $\xi_n = \sigma_n^p u_{1,n}''$, and, as in (2.4), deduce from (3.1)

$$\xi_n(x) = (\sigma_n^p u_{1,n}'')(0) - \lambda_{1,n} u_{1,n}(x).$$

As the sequences $\{\xi_n\}$ and $\{u_{1,n}\}$ are uniformly bounded in L^2, so too must $\{(\sigma_n^p u_{1,n}'')(0)\}$. Consequently, ξ_n converges strongly in L^2 to some function $\overline{\xi}$. The left side of (3.1) therefore converges to $\int_0^1 \overline{\xi} v''\,dx$. It remains to characterize this $\overline{\xi}$. Recalling the pointwise bounds on the σ_n we may assume that σ_n^{-p} converges in the weak* topology of L^∞ to some function μ. Thus, $\xi_n \sigma_n^{-p}$ converges weakly in L^2 to $\mu\overline{\xi}$. But $\xi_n \sigma_n^{-p} = u_{1,n}''$, whose weak L^2 limit is $\overline{u}''$. Hence, $\overline{\xi} = \overline{u}''\mu^{-1}$. Defining $\overline{\sigma} = \mu^{-1/p}$, we may pass to the limit in (3.1) and obtain

$$\int_0^1 \overline{\sigma}^p \overline{u}'' v''\,dx = \hat{\lambda}_1 \int_0^1 \overline{u}' v'\,dx \quad \forall, v \in H_0^2. \tag{3.2}$$

As symmetry is preserved under weak $*$ convergence, we find $\overline{\sigma}$ to be even. In addition, the pointwise convergence of $u_{1,n}$ to $\overline{u}$ leaves $\overline{u}$ positive and even. Now (3.2) implies that $\hat{\lambda}_1 = \lambda_j(\overline{\sigma})$ for some j. That $j = 1$ follows from the existence of a positive even first eigenfunction for $\overline{\sigma}$ and the fact that $\overline{u}$ is itself positive and even. We need only determine whether $\overline{\sigma} \in ad$. We may verify the pointwise bounds without trouble. With respect to the integral constraint, we consider the convex function $f : R \to R$, $f(t) = t^{-1/p}$. The integral functional $\varphi \mapsto \int_0^1 f(\varphi(x))\,dx$ is then weak* lower semicontinuous on L^∞, see, e.g., Dacorogna [12, Thm. 1.1]. As $1/\sigma_n^p$ converges weak* to $1/\overline{\sigma}^p$, this allows us to conclude that

$$\int_0^1 \overline{\sigma}\,dx = \int_0^1 f(1/\overline{\sigma}^p)\,dx \le \liminf \int_0^1 f(1/\sigma_n^p)\,dx = \lim \int_0^1 \sigma_n\,dx = 1. \tag{3.3}$$

If indeed equality does not hold in (3.3), then there exists an even $\hat{\sigma} \in ad$ such that $\hat{\sigma}(x) \ge \overline{\sigma}(x)$ for almost every $x \in (0,1)$. From Rayleigh's principle we then easily deduce $\lambda_1(\hat{\sigma}) \ge \lambda_1(\overline{\sigma})$. We have now proven the following.

THEOREM 3.1. *There exists an even $\hat{\sigma} \in ad$ for which $\lambda_1(\sigma) \le \lambda_1(\hat{\sigma})$ for every $\sigma \in ad$.*

Our choice of *ad* was motivated by our interest in the "shape" of the strongest column. Theorem 3.1, however, may also be applied in the search for the "composition" of the strongest column. For example, consider the design problem where one must combine two materials in fixed proportion so as to maximize the buckling load of the resulting column. The set of admissible designs is then

$$\begin{aligned} ad_E = \{&\alpha\chi(x) + \beta(1-\chi(x)) : \chi \text{ is the characteristic} \\ &\text{function of a subset of (0,1) with measure } \gamma\}, \end{aligned}$$

where α and β are the Youngs moduli of the respective materials with γ the volume fraction of the first. In this context, Theorem 3.1 states that λ_1 attains its maximum on the weak* closure of ad_E, i.e., on

$$ad_E^* = \left\{ \alpha\theta(x) + \beta(1-\theta(x)) : 0 \le \theta(x) \le 1, \int_0^1 \theta\,dx = \gamma \right\}.$$

4. Necessary conditions.

We search now for a characterization of our optimal design, $\hat{\sigma}$. Typical of many such problems, two distinct approaches are possible. Taking advantage of the variational structure, the so called direct approach attempts to swap the order of the limits in

$$\hat{\lambda}_1 = \lambda_1(\hat{\sigma}) = \sup_{\sigma\in ad} \inf_{u\in H_0^2} \frac{\int_0^1 \sigma^p |u''|^2\,dx}{\int_0^1 |u'|^2\,dx},$$

inferring necessary conditions from the resulting saddle point. The indirect approach strives to determine $\hat{\sigma}$ through knowledge of the tangents to the graph of $\sigma \mapsto \lambda_1(\sigma)$

and the normals to *ad*. Our implementation of these two approaches intersect in their reliance on (i) recent work of Auchmuty [2] on dual variational principles and (ii) a lopsided minimax principle.

PROPOSITION 4.1.

$$\lambda_1^{-1}(\sigma) = \sup_{u\in H_0^2} \mathcal{A}(\sigma,u), \;\; \mathcal{A}(\sigma,u) = \sqrt{2}\|u'\| - \tfrac{1}{2}\int_0^1 \sigma^p |u''|^2\,dx.$$

$u \mapsto \mathcal{A}(\sigma,u)$ attains its maximum only on those $u \in \mathcal{E}(\sigma)$ for which $\|u'\| = \sqrt{2}\lambda_1^{-1}(\sigma)$.

Proof. In addition to being bounded above by $2\lambda_1^{-1}(\sigma)$, the map $u \mapsto \mathcal{A}(\sigma,u)$ is coercive and weakly upper semicontinuous on H_0^2 and therefore attains its maximum at some $\overline{u} \in H_0^2$. Necessarily, $\mathcal{D}_2\mathcal{A}(\sigma,\overline{u})$, the Gâteaux derivative of $u \mapsto \mathcal{A}(\sigma,u)$ at $\overline{u}$, must vanish. That is,

$$\int_0^1 \sigma^p \overline{u}'' v''\,dx = \sqrt{2}\|\overline{u}'\|^{-1}\int_0^1 \overline{u}' v'\,dx \quad \forall\, v \in H_0^2.$$

As a result, $\overline{u}$ is an eigenfunction corresponding to the eigenvalue $\sqrt{2}\|\overline{u}'\|^{-1}$. As $\overline{u}$ maximizes $u \mapsto \mathcal{A}(\sigma,u)$, this must be the least eigenvalue, $\lambda_1(\sigma)$. □

PROPOSITION 4.2. *Consider $F : X \times Y \to R$, where X and Y are topological vector spaces and assume that $x \mapsto F(x,y)$ is concave and upper semicontinuous while $y \mapsto F(x,y)$ is convex and lower semicontinuous. If there exists a $y_o \in Y$ and $c_o \in R$ such that $\{x \in X;\; F(x,y_o) \geq c_o\}$ is compact and $c_o < \inf_{y\in Y}\sup_{x\in X} F(x,y)$, then*

$$\sup_{x\in X}\inf_{y\in Y} F(x,y) = \inf_{y\in Y}\sup_{x\in X} F(x,y).$$

Proof. This is a weakening of Theorem 3.7, Chapter 2, in Barbu and Precupanu [3]. □

It is with the indirect approach that we shall meet with the greatest success. For prior attempts in this context see Haug and Rousselet [18] and Choi and Haug [9]. Our principal tool is the generalized gradient of Clarke [10].

For a real valued Lipschitz function F on a Banach space X we consider the generalized directional derivative of F at x in the direction v,

$$F^o(x;v) \equiv \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y+tv) - F(y)}{t}.$$

Denoting the dual of X by X^* and $x^*(x)$ by $\langle x^*, x\rangle$ when $x^* \in X^*$ and $x \in X$, Clarke's generalized gradient of F at x is the nonempty, convex, weak* compact set

$$\partial F(x) \equiv \{\xi \in X^*;\; F^o(x;v) \geq \langle \xi, v\rangle \;\forall\, v \in X\}.$$

We demonstrate that $\sigma \mapsto \lambda_1^{-1}(\sigma)$ is Lipschitz on $\Sigma = \{\sigma \in L^\infty;\; \|\hat{\sigma} - \sigma\|_\infty < \alpha/2\}$. Choose $\sigma_1, \sigma_2 \in \Sigma$ such that $\lambda_1^{-1}(\sigma_1) > \lambda_1^{-1}(\sigma_2)$ and note that for $u_1 \in \operatorname{Argmax} \mathcal{A}(\sigma_1,\cdot)$, the set on which $u \mapsto \mathcal{A}(\sigma_1,u)$ attains its maximum,

$$\begin{aligned} |\lambda_1^{-1}(\sigma_1) - \lambda_1^{-1}(\sigma_2)| &\leq |\mathcal{A}(\sigma_1,u_1) - \mathcal{A}(\sigma_2,u_1)| \\ &\leq \int_0^1 |\sigma_1^p - \sigma_2^p||u_1''|^2\,dx \\ &\leq \|u_1''\|^2\|\sigma_1^p - \sigma_2^p\|_\infty \leq \alpha^{-2p}p|2\beta|^{p-1}\|\sigma_1 - \sigma_2\|_\infty. \end{aligned}$$

Without loss we assume that $\lambda_1(\hat{\sigma})$ is a double eigenvalue. Then $\mathcal{E}(\hat{\sigma})$ is two-dimensional and $\operatorname{Argmax} \mathcal{A}(\hat{\sigma}, \cdot)$ is the intersection of $\mathcal{E}(\hat{\sigma})$ with the sphere $\|u'\| = \sqrt{2}\hat{\lambda}_1^{-1}$. It will be convenient to choose a basis $\{\hat{u}_1, \hat{u}_2\}$ for $\mathcal{E}(\hat{\sigma})$ for which $\int_0^1 \hat{u}_i'\hat{u}_j'\, dx = 2\delta_{ij}\hat{\lambda}_1^{-2}$. For then,

$$\text{(4.1)} \qquad \operatorname{Argmax} \mathcal{A}(\hat{\sigma}, \cdot) = \{a\hat{u}_1 + b\hat{u}_2;\ a^2 + b^2 = 1\}.$$

Regarding the Gâteaux derivative of $\sigma \mapsto \mathcal{A}(\sigma, u)$ at $\hat{\sigma}$ in the direction η we have

$$\text{(4.2)} \qquad \langle \mathcal{D}_1 \mathcal{A}(\hat{\sigma}, u), \eta \rangle = -\tfrac{p}{2} \int_0^1 \eta \hat{\sigma}^{p-1} |u''|^2 \, dx.$$

Denoting convex hull by 'co,' the sense in which the gradient of a maximum is the maximum of the gradients is the following.

THEOREM 4.3. $\partial \lambda_1^{-1}(\hat{\sigma}) = \operatorname{co}\{-\tfrac{p}{2}\hat{\sigma}^{p-1}(a\hat{u}_1'' + b\hat{u}_2'')^2;\ a^2 + b^2 = 1\}$.

Proof. From (4.1) and (4.2) this set is precisely

$$\text{(4.3)} \qquad \operatorname{co}\{\mathcal{D}_1 \mathcal{A}(\hat{\sigma}, u);\ u \in \operatorname{Argmax} \mathcal{A}(\hat{\sigma}, \cdot)\}.$$

Our claim does not fit neatly into Clarke's result [10, Thm. 2.8.2] on the generalized gradient of a pointwise maximum. The contortions involved in fitting our problem to Clarke's hypotheses are no less difficult, and far less instructive, than an independent proof.

Let us denote the set in (4.3) by Ξ. We show that $\Xi \subset \partial \lambda_1^{-1}(\hat{\sigma})$. For $\xi \in \Xi$ and $\eta \in L^\infty$

$$\begin{aligned} \langle \xi, \eta \rangle &= \sum_{i=1}^n \mu_i \langle \mathcal{D}_1 \mathcal{A}(\hat{\sigma}, u_i), \eta \rangle \\ &= \sum_{i=1}^n \mu_i \lim_{t \downarrow 0} \frac{\mathcal{A}(\hat{\sigma} + t\eta, u_i) - \mathcal{A}(\hat{\sigma}, u_i)}{t} \\ &\le \sum_{i=1}^n \mu_i \limsup_{t \downarrow 0} \frac{\lambda_1^{-1}(\hat{\sigma} + t\eta) - \lambda_1^{-1}(\hat{\sigma})}{t} \\ &\le (\lambda_1^{-1})^o(\hat{\sigma}; \eta), \end{aligned}$$

hence $\xi \in \partial \lambda_1^{-1}(\hat{\sigma})$.

Regarding the reverse inclusion we define

$$g(\sigma; \eta) = \max_{\xi \in \Xi} \langle \xi, \eta \rangle, \qquad \sigma \in \Sigma,\ \eta \in L^\infty$$

and prove

$$(\lambda_1^{-1})^o(\sigma; \eta) \le g(\sigma; \eta).$$

Select $\sigma_n \to \sigma$ in L^∞ and $t_n \downarrow 0$ in R such that

$$q_n \equiv \frac{\lambda_1^{-1}(\sigma_n + t_n\eta) - \lambda_1^{-1}(\sigma_n)}{t_n}$$

converges to $(\lambda_1^{-1})^o(\sigma;\eta)$. Select $u_n \in$ Argmax $\mathcal{A}(\sigma_n + t_n\eta, \cdot)$ and note that

$$q_n \leq \frac{\mathcal{A}(\sigma_n + t_n\eta, u_n) - \mathcal{A}(\sigma_n, u_n)}{t_n}$$

with the right side equal to $\langle \mathcal{D}_1\mathcal{A}(\sigma_n + \bar{t}_n\eta, u_n), \eta\rangle$, for some $\bar{t}_n \in (0, t_n)$, by the Mean Value Theorem. As $\sigma_n + t_n\eta \to \sigma$ in L^∞ and $u_n \in$ Argmax $\mathcal{A}(\sigma_n + t_n\eta, \cdot)$, we recall from our work in Theorem 3.1 that $u_n \rightharpoonup u \in$ Argmax $\mathcal{A}(\sigma, \cdot)$ in H^2 and $(\sigma_n + t_n\eta)^p u_n'' \to \sigma^p u''$ in L^2, and hence $u_n'' \to u''$ in L^2, i.e., $u_n \to u$ in H^2. Recalling (4.2) this establishes

$$\langle \mathcal{D}_1\mathcal{A}(\sigma_n + \bar{t}_n\eta, u_n), \eta\rangle \to \langle \mathcal{D}_1\mathcal{A}(\sigma, u), \eta\rangle$$

with $u \in$ Argmax $\mathcal{A}(\sigma, \cdot)$. As a result,

$$(\lambda_1^{-1})^o(\sigma;\eta) = \lim_{n\to\infty} q_n \leq \langle \mathcal{D}_1\mathcal{A}(\sigma, u), \eta\rangle \leq g(\sigma;\eta).$$

If ζ is now an element of $\partial\lambda_1^{-1}(\hat{\sigma})$, then $g(\hat{\sigma};\eta) \geq \langle\zeta, \eta\rangle$ for each $\eta \in L^\infty$. Consequently,

$$0 = \min_{\eta\in L^\infty} \max_{\xi\in\Xi} \langle \xi - \zeta, \eta\rangle.$$

Noting that Ξ is closed and bounded in L^1 and finite dimensional (it lies in the span of $\{|\hat{u}_1''|^2, \hat{u}_1''\hat{u}_2'', |\hat{u}_2''|^2\}$), we find it compact in $(L^\infty)^*$. Invoking Proposition 4.2 yields a $\hat{\xi} \in \Xi$ for which

$$\langle \hat{\xi} - \zeta, \eta\rangle = 0 \quad \forall \eta \in L^\infty.$$

It follows that $\zeta = \hat{\xi}$ and so $\partial\lambda_1^{-1}(\hat{\sigma}) \subset \Xi$. □

This proof, though identical in outline to Clarke [10, Thm. 2.8.2], has exploited additional properties of λ_1 and $\mathcal{A}$ to make up for the missing hypotheses. Observe that when $\lambda_1(\hat{\sigma})$ is simple the generalized gradient reduces to the singleton

$$\partial\lambda_1^{-1}(\hat{\sigma}) = \{\mathcal{D}_1\mathcal{A}(\hat{\sigma}, \hat{u}_1)\}.$$

As zero is not a tangent direction to λ_1^{-1} at $\hat{\sigma}$, i.e., $0 \neq \partial\lambda_1^{-1}(\hat{\sigma})$, we are compelled to investigate the constraint set *ad*. Separating the equality from the inequality constraints brings

$$C \equiv \{\sigma \in L^\infty;\ \alpha \leq \sigma(x) \leq \beta\} \quad \text{and} \quad V(\sigma) \equiv \int_0^1 \sigma\, dx.$$

As $\hat{\sigma}$ minimizes $\sigma \mapsto \lambda_1^{-1}(\sigma)$ subject to $\sigma \in C$ and $V(\sigma) = 1$, we deduce from the Lagrange Multiplier Rule, [10, Thm. 6.1.1], that a nontrivial linear combination of elements in $\partial\lambda_1^{-1}(\hat{\sigma})$ and $\partial V(\hat{\sigma})$ is normal to C at $\hat{\sigma}$. In particular,

$$\left(\nu_1\partial\lambda_1^{-1}(\hat{\sigma}) + \nu_2\partial V(\hat{\sigma})\right) \cap N_C(\hat{\sigma}) \neq \emptyset,$$

where $\nu_1 \leq 0$, $\nu_1^2 + \nu_2^2 > 0$, and

$$N_C(\hat{\sigma}) = \left\{\zeta \in (L^\infty)^*;\ \int_0^1 (\hat{\sigma} - \sigma)\, d\zeta \geq 0,\ \forall \sigma \in C\right\}$$

is the cone of normals to C at $\hat{\sigma}$. In light of our previous calculations, and the fact that $\partial V(\hat{\sigma}) = 1 \in L^1$, there exists a $\hat{\xi} \in \partial\lambda_1^{-1}(\hat{\sigma})$ for which

$$\int_0^1 (\hat{\sigma} - \sigma)(\nu_1\hat{\xi} + \nu_2)\,dx \geq 0 \quad \forall\sigma \in C. \tag{4.4}$$

Observing that $\nu_1\hat{\xi} \geq 0$, we find that $\nu_2 \geq 0$ requires, through (4.4), that $\hat{\sigma} \equiv \beta$. This is an impossibility. Likewise, should $\nu_1 = 0$, (4.4) would require $\hat{\sigma} \equiv \alpha$ (since $\nu_2 < 0$). Taking $\ell^2 = \nu_2/\nu_1$, we arrive at

$$\int_0^1 (\hat{\sigma} - \sigma)(\hat{\xi} + \ell^2)\,dx \leq 0 \quad \forall\sigma \in C.$$

The subsequent reduction to pointwise optimality conditions follows a well-known course, see, e.g., Cea and Malanowski [8]. In particular,

$$\hat{\sigma}(x) = \alpha \quad\Rightarrow\quad -\hat{\xi}(x) \leq \ell^2, \tag{4.5}$$

$$\alpha < \hat{\sigma}(x) < \beta \quad\Rightarrow\quad -\hat{\xi}(x) = \ell^2, \tag{4.6}$$

$$\hat{\sigma}(x) = \beta \quad\Rightarrow\quad -\hat{\xi}(x) \geq \ell^2 \tag{4.7}$$

for almost every $x \in (0,1)$. To appreciate this result, we must recall that $\hat{\xi} \in \partial\lambda_1^{-1}(\hat{\sigma})$ means

$$-\hat{\xi}(x) = \frac{p}{2}\sum_{i=1}^{n} t_i\hat{\sigma}^{p-1}(x)(a_i\hat{u}_1''(x) + b_i\hat{u}_2''(x))^2,$$

where

$$t_i \geq 0, \quad \sum_{i=1}^{n} t_i = 1, \quad \text{and} \quad a_i^2 + b_i^2 = 1.$$

On expanding this sum of squares, (4.6) becomes

$$\hat{\sigma}^{p-1}\left(\delta_1|\hat{u}_1''|^2 + \delta_2|\hat{u}_2''|^2 + \delta_3\hat{u}_1''\hat{u}_2''\right) = 1, \tag{4.8}$$

where

$$\delta_1 = \frac{p}{2}\sum_{i=1}^{n} t_i a_i^2/\ell^2, \quad \delta_2 = \frac{p}{2}\sum_{i=1}^{n} t_i b_i^2/\ell^2, \quad \delta_3 = p\sum_{i=1}^{n} t_i a_i b_i/\ell^2.$$

Observing that $\delta_1\delta_2$ indeed dominates $\delta_3^2/4$, we have recovered (1.8), the necessary condition of Bratus and Seiranian [7] and Masur [22]. If in fact $\delta_1\delta_2 = \delta_3^2/4$, then for $\hat{u} \equiv \sqrt{\delta_1}\hat{u}_1 + \sqrt{\delta_2}\hat{u}_2$, equation (4.8) yields $\hat{\sigma}^{p-1}|\hat{u}''|^2 = 1$, the optimality condition of Tadjbakhsh and Keller. Since $\hat{u}$ is an eigenfunction and, therefore, admits at least two inflection points, the pointwise bounds must become active, i.e., $\hat{\sigma}^{p-1}|\hat{u}''|^2 = 1$ cannot hold on the entire interval. Ignoring any bound constraints, Masur [22] and Seiranian [34] found a σ and two orthogonal elements of $\mathcal{E}(\sigma)$ for which (4.8) holds with $p = 2$. This appears to be the design obtained by Olhoff and Rasmussen [25] and, by all indications, the one preferred by our algorithm as well (see §7, Fig. 1). It appears likely that in this case the bound constraints are inactive due to the fact that where $\hat{\sigma}$ is less than one, $\hat{\sigma}^2$ is much less than one. Since $\hat{\sigma}^2$ is the quantity that appears in the Rayleigh quotient, we expect it to be as large as possible. This suggests that $\hat{\sigma}$ is bounded away from zero, independent of α. This lower bound with the integral constraint supports the conjecture that $\hat{\sigma}$ is in fact bounded above as well. Hence, when α and β are, respectively, chosen below and above these 'natural' bounds, condition (4.8) is free to stand on its own. Clearly these natural bounds must depend on p. In fact, we shall provide numerical evidence in §7 in favor of the argument that the natural lower (upper) bound is an increasing (decreasing) function of p for $p > 1$.

Unfortunately, it is not known whether (4.8) is a sufficient condition for optimality. The proof of sufficiency offered by Tadjbakhsh and Keller [37] is incorrect. They proceed as if $\lambda_1(\sigma)$ corresponds to the least eigenvalue of (2.5) and, accordingly, admit all functions that satisfy the boundary conditions as test functions in a Rayleigh principle argument. In fact, (2.5) possesses a double zero eigenvalue, hence only those functions that are orthogonal to the first two eigenfunctions can be admitted. We remark that Ramm's claim [32], that Tadjbakhsh and Keller mistakenly applied Hölders inequality in their sufficiency proof, is incorrect, though [37, §6 (25)] is only valid for $n < 0$.

Though (4.8) need not hold over the entire length of the column, we now show that where it does hold it requires that $\hat{\sigma}$ be smooth.

THEOREM 4.4. *If* $\alpha < \hat{\sigma}(x) < \beta$ *for each* $x \in (a,b) \subset (0,1)$, *then* $\hat{\sigma} \in C^\infty(a,b)$.

Proof. We observed in (2.4) that

$$\hat{\sigma}^p \hat{u}_i'' = l_i - \hat{\lambda}_1 \hat{u}_i, \tag{4.9}$$

where l_i is an affine function of x. Now multiply (4.8) by $\hat{\sigma}^{p+1}$,

$$\delta_1(\hat{\sigma}^p \hat{u}_1'')^2 + \delta_2(\hat{\sigma}^p \hat{u}_2'')^2 + \delta_3(\hat{\sigma}^p \hat{u}_1'')(\hat{\sigma}^p \hat{u}_2'') = \hat{\sigma}^{p+1}. \tag{4.10}$$

From (4.9) we find, on recalling $H_0^2 \subset C^1$, that each term on the left of (4.10) is C^1, and hence, that $\hat{\sigma} \in C^1$. Writing (4.9) in the form

$$\hat{u}_i'' = \frac{l_i - \hat{\lambda}_1 \hat{u}_i}{\hat{\sigma}^p},$$

we conclude $\hat{u}_i'' \in C^1$, that is, $\hat{u}_i \in C^3$. Repeating this exact argument leads to $\hat{\sigma} \in C^3$ and $\hat{u}_i \in C^5$. The result then follows from continued repetitions. □

Having succeeded in pursuing the indirect approach, we now look to the possibility (and the implications) of exchanging the limits in the characterization

$$\hat{\lambda}_1^{-1} = \mathcal{A}(\hat{\sigma}, \hat{u}) = \inf_{\sigma \in ad} \sup_{u \in H_0^2} \mathcal{A}(\sigma, u).$$

Recalling Proposition 4.2, this will require convexity and lower semicontinuity of $\sigma \mapsto \mathcal{A}(\sigma, u)$, and concavity and upper semicontinuity of $u \mapsto \mathcal{A}(\sigma, u)$, as well as compactness of one of its upper level sets.

Remark 4.5. We noted the weak H^2 upper semicontinuity of $u \mapsto \mathcal{A}(\sigma, u)$ in Proposition 4.1. As $\{u \in H_0^2;\ \mathcal{A}(\sigma, u) \geq c\}$ is bounded it is also weakly compact (independent of $c \in R$ and $\sigma \in ad$). Convexity of $\sigma \mapsto \mathcal{A}(\sigma, u)$ follows on restricting $p \leq 1$.

The two remaining properties require more work. Note that $u \in$ Argmax $\mathcal{A}(\sigma, \cdot)$ implies $\mathcal{A}(\sigma, u) = \mathcal{A}(\sigma, -u) = \lambda_1^{-1}(\sigma)$ while $\mathcal{A}(\sigma, 0) = 0$. Hence, $u \mapsto \mathcal{A}(\sigma, u)$ is not concave on any set that contains Argmax $\mathcal{A}(\sigma,\cdot)$. This suggests that we examine the half-spaces exterior to $\{u \in H_0^2;\ \|u'\| \le \sqrt{2}\lambda_1^{-1}(\sigma)\}$. Unfortunately, this ball, and hence its support planes, depend on σ. Consequently, if we expect these half-spaces to vary continuously with σ, we must be careful in our choosing. This choice is greatly facilitated by the assumption that σ lies in ad_s, those functions in ad that are even about $\frac{1}{2}$. In this case, we may speak unambiguously of u_1, the positive even eigenfunction corresponding to $\lambda_1(\sigma)$. We normalize $\|u_1'\| = \sqrt{2}\lambda_1^{-1}(\sigma)$ and consider the associated half-space

$$\Pi_\sigma = \left\{ v \in H_0^2;\ \lambda_1^2(\sigma) \int_0^1 u_1' v'\, dx > 2 \right\}.$$

PROPOSITION 4.6. *For $\sigma \in ad_s$, $u \mapsto \mathcal{A}(\sigma, u)$ is concave on Π_σ.*

Proof. The quadratic form associated with the second Gâteaux derivative of $u \mapsto \mathcal{A}(\sigma,u)$ at $\overline{u} \in \Pi_\sigma$ satisfies

$$\begin{aligned}\langle \mathcal{D}_2^2 \mathcal{A}(\sigma,\overline{u})v,v\rangle &= \sqrt{2}\|\overline{u}'\|^{-1} \int_0^1 |v'|^2\,dx - \int_0^1 \sigma^p |v''|^2\,dx - \sqrt{2}^{-1}\|\overline{u}'\|^{-3}\left(\int_0^1 \overline{u}'v'\,dx\right)^2 \\ &\le (\sqrt{2}\|\overline{u}'\|^{-1} - \lambda_1(\sigma)) \int_0^1 |v'|^2\,dx \\ &\le 0 \quad \forall v \in H_0^2.\end{aligned}$$

□

This suggests that we penalize $\mathcal{A}$ with the indicator function of Π_σ,

$$\pi(\sigma,u) = \begin{cases} 0 & \text{if } u \in \Pi_\sigma; \\ \infty, & \text{otherwise.}\end{cases}$$

This not only guarantees concavity but also respects lower semicontinuity.

PROPOSITION 4.7. *$\sigma \mapsto \mathcal{A}(\sigma,u) - \pi(\sigma,u)$ is lower semicontinuous for the strong L^∞ topology on ad_s.*

Proof. Now $\sigma_n \to \sigma$ in L^∞ clearly implies $\mathcal{A}(\sigma_n,u) \to \mathcal{A}(\sigma,u)$ for each $u \in H_0^2$. Regarding $\limsup \pi(\sigma_n,u) \le \pi(\sigma,u)$, it suffices to show that

$$\pi(\sigma_n,u) \to 0 \quad \forall u \in \Pi_\sigma.$$

From the proof of Theorem 3.1 it is clear that $\lambda_1(\sigma_n) \to \lambda_1(\sigma)$ and $u_1(\sigma_n) \rightharpoonup u_1(\sigma)$ in H_0^2. Hence,

$$\lambda_1^2(\sigma_n) \int_0^1 u_1'(\sigma_n)u'\,dx \to \lambda_1^2(\sigma)\int_0^1 u_1(\sigma)u'\,dx > 2.$$

From this we conclude that u is eventually in each Π_{σ_n}, i.e., $\pi(\sigma_n,u) = 0$. □

We may now modify Proposition 4.2 (see Cox and McLaughlin [11, §7]) and conclude the following.

THEOREM 4.8. *If $p \le 1$, then $(\hat{\sigma}, u_1(\hat{\sigma}))$ is a saddle point for $\mathcal{A}$ over $ad_s \times H_0^2$. That is, denoting $u_1(\hat{\sigma})$ by $\hat{u}$,*

$$\mathcal{A}(\hat{\sigma},u) \le \mathcal{A}(\hat{\sigma},\hat{u}) \le \mathcal{A}(\sigma,\hat{u}) \quad \forall\,(\sigma,u) \in ad_s \times H_0^2.$$

The latter inequality yields the following maximum principle

$$\int_0^1 \sigma^p |\hat{u}''|^2 \, dx \leq \int_0^1 \hat{\sigma}^p |\hat{u}''|^2 \, dx \quad \forall \, \sigma \in ad_s.$$

The subsequent pointwise conditions call for an $\ell^2 > 0$ such that

$$\hat{\sigma}(x) = \alpha \;\Rightarrow\; \hat{\sigma}^{p-1}(x)|\hat{u}''(x)|^2 \leq \ell^2, \tag{4.11}$$

$$\alpha < \hat{\sigma}(x) < \beta \;\Rightarrow\; \hat{\sigma}^{p-1}(x)|\hat{u}''(x)|^2 = \ell^2, \tag{4.12}$$

$$\hat{\sigma}(x) = \beta \;\Rightarrow\; \hat{\sigma}^{p-1}(x)|\hat{u}''(x)|^2 \geq \ell^2 \tag{4.13}$$

for almost every $x \in (0,1)$. As $\sigma \mapsto \lambda_1^{-1}(\sigma)$ is convex when $p \leq 1$, these conditions are also sufficient. Hence, we see that where the direct method applies it gives more information. In particular, the necessary conditions (4.11)–(4.13) involve only a single buckling mode. Comparing these to the more general conditions in (4.5)–(4.7) suggests that $\lambda_1(\hat{\sigma})$ is indeed a simple eigenvalue when $p \leq 1$. We shall see numerical evidence of this in §7. The critical case, $p = 1$, where the optimal buckling load changes multiplicity, has received considerable attention. In this case, the right side of (4.11)–(4.13) is independent of $\hat{\sigma}$. In particular, a number of workers have claimed that

$$|\hat{u}''(x)| = \ell. \tag{4.14}$$

We remark, however, that in the absence of a second buckling mode the bound constraints must become active near the inflection points of $\hat{u}$, making (4.11) and (4.13) indeed necessary. Nonetheless, Seiranian [34], who deduced (4.14) from (1.8), proceeded to solve (4.14) in conjunction with (2.3), yielding

$$\hat{\sigma}(x) = \begin{cases} 3/2(1 - 16x^2) & \text{if } 0 \leq x \leq \frac{1}{4} \\ 3/2(16x - 16x^2 - 3) & \text{if } \frac{1}{4} \leq x \leq \frac{3}{4} \\ 3/2(32x - 16x^2 - 15) & \text{if } \frac{3}{4} \leq x \leq 1. \end{cases} \tag{4.15}$$

On evaluating the Rayleigh quotient with this $\hat{\sigma}$ and a specific C^1 test function, Seiranian arrived at a buckling load of 48. This design, like that of Tadjbakhsh and Keller for $p = 2$, vanishes at $\frac{1}{4}$ and $\frac{3}{4}$. Unlike the design of Tadjbakhsh and Keller, however, we are not able to show it to be suboptimal. We can only stress that, lacking an existence proof for $\alpha = 0$, $p = 1$, there is no reason to believe that (4.14) is a necessary condition for optimality.

5. Other boundary conditions. Intent on a clean exposition, we have to this point concentrated solely on the clamped-clamped boundary conditions $u(0) = u'(0) = u(1) = u'(1) = 0$. We now apply the work of the previous sections to the other standard sets of boundary conditions, in particular, hinged and free. A column is said to be free at a point when no conditions are prescribed, while it is hinged, or simply supported, when its displacement is required to vanish there. As a matter of notation, the weak formulation of the buckled column equation will read

$$\int_0^1 \sigma^p u'' v'' \, dx = \mu \int_0^1 u'v' \, dx \quad \forall \, v \in V_{i,j}, \tag{5.1}$$

where $V_{i,j}$ is a subspace of H^2, with i and j chosen from $\{0,1,2\}$ according to whether the respective end is either free, hinged, or clamped. For example,

$$V_{1,2} = \{u \in H^2;\ u(0) = 0,\ u(1) = u'(1) = 0\}$$

specifies the hinged-clamped column. We denote the least eigenvalue of (5.1) by $\mu_{i,j}(\sigma)$, and the corresponding space of eigenfunctions by $\mathcal{E}_{i,j}(\sigma)$. As before, $u \in \mathcal{E}_{i,j}(\sigma)$ implies that both u and $\sigma^p u''$ are elements of $C^1([0,1])$. In addition, such functions satisfy so called natural boundary conditions. In particular, if $i = 1$ then, in addition to $u(0) = 0$, we find

$$\sigma^p u''(0) = 0, \tag{5.2}$$

while if $i = 0$ we have, in addition to (5.2),

$$(\sigma^p u'')'(0) + \mu_{0,j}(\sigma)u'(0) = 0. \tag{5.3}$$

We shall consider only those $\mu_{i,j}(\sigma)$ for which $i+j \geq 2$, as otherwise $\mu_{i,j}(\sigma) = 0$. For comparison purposes, we record these eigenvalues in the case of the uniform column.

$$\mu_{0,2}(1) = \pi^2/4, \quad \mu_{1,1}(1) = \pi^2, \quad \mu_{1,2}(1) \approx 2.046\pi^2, \quad \mu_{2,2}(1) = 4\pi^2. \tag{5.4}$$

Clearly, $\mu_{i,j}(1) = \mu_{j,i}(1)$. Analogous to (2.5), for $i+j \geq 2$, (5.4) gives the uniform bounds

$$\pi^2\alpha^p/4 \leq \mu_{i,j}(\sigma) \leq 4\pi^2\beta^p \quad \forall\, \sigma \in ad. \tag{5.5}$$

As in §2, we address the multiplicity of $\mu_{i,j}(\sigma)$ and the presence of positive eigenfunctions.

LEMMA 5.1. *For $\sigma \in ad$,*
(a) *If $2 \leq i+j < 4$, then $\mu_{i,j}(\sigma)$ is simple and there exists a corresponding positive eigenfunction.*
(b) *$\mu_{0,2}(\sigma) < \mu_{1,2}(\sigma)$ and $\mu_{1,1}(\sigma) < \mu_{1,2}(\sigma) < \mu_{2,2}(\sigma)$.*

Proof. (a) Seiranian noted for these boundary conditions that (5.1) is equivalent, except for the presence of a simple zero eigenvalue when the product ij equals 2, to a second-order problem with *separated* boundary conditions. It now follows from the oscillation theory of Stürm, see, e.g., Atkinson [1], that each $\mu_{i,j}(\sigma)$ is simple and that for $ij \neq 2$, $\sigma^p u''$ is of one sign for each $u \in \mathcal{E}_{i,j}(\sigma)$. In case ij equals zero or 1, this yields, respectively, a positive convex or concave element of $\mathcal{E}_{i,j}(\sigma)$. When $ij = 2$ we find that $\sigma^p u''$ vanishes exactly once on $(0,1)$ for each $u \in \mathcal{E}_{i,j}(\sigma)$. Here we find an eigenfunction that is convex on $(0, x_0)$ and concave on $(x_0, 1)$ for some x_0. As this function must vanish at zero and 1, we conclude that it must be positive on $(0,1)$.

(b) As $V_{i+1,j} \subset V_{i,j}$ we find $\mu_{i,j}(\sigma) \leq \mu_{i+1,j}$. Should equality hold, we conclude $\mathcal{E}_{i+1,j}(\sigma) \subset \mathcal{E}_{i,j}(\sigma)$. As in (2.4), for $u \in \mathcal{E}_{i,j}(\sigma)$ we deduce from (5.1) that

$$(\sigma^p u'')(x) = ((\sigma^p u'')'(0) + \mu_{i,j}(\sigma)u'(0))\,x + \sigma^p u''(0) + \mu_{i,j}(\sigma)u(0) - \mu_{i,j}(\sigma)u(x). \tag{5.6}$$

If $\mu_{0,2}(\sigma) = \mu_{1,2}(\sigma)$, then for each $u \in \mathcal{E}_{1,2}(\sigma) \subset \mathcal{E}_{0,2}(\sigma)$ equation (5.6), in view of (5.3), reads

$$(\sigma^p u'')(x) = -\mu_{i,j}u(x). \tag{5.7}$$

On recalling that $u(1) = u'(1) = 0$, we see that u satisfies a linear homogeneous equation with zero terminal data, and hence, $u \equiv 0$.

If $\mu_{1,1}(\sigma) = \mu_{1,2}(\sigma)$, then for each $u \in \mathcal{E}_{1,2}(\sigma) \subset \mathcal{E}_{1,1}(\sigma)$ (5.6), in view of (5.2), reads

$$(\sigma^p u'')(x) = ((\sigma^p u'')'(0) + \mu_{i,j}(\sigma)u'(0))\,x - \mu_{i,j}(\sigma)u(x). \tag{5.8}$$

So $(\sigma^p u'')(1) = (\sigma^p u'')'(0) + \mu_{i,j}(\sigma)u'(0)$. But $(\sigma^p u'')(1) = 0$ so (5.8) reduces to (5.7) and again the clamped conditions at 1 imply that $u \equiv 0$.

If $\mu_{1,2}(\sigma) = \mu_{2,2}(\sigma)$, then for each $u \in \mathcal{E}_{2,2}(\sigma) \subset \mathcal{E}_{1,2}(\sigma)$ equation (5.6), in view of (5.2), reads

$$(\sigma^p u'')(x) = (\sigma^p u'')'(0)x - \mu_{i,j}(\sigma)u(x).$$

Hence, $(\sigma^p u'')(1) = (\sigma^p u'')'(0)$, from which we conclude that u is either identically zero or not of one sign. This excludes the positive element of $\mathcal{E}_{1,2}(\sigma)$ established in part (a). □

Thanks to the presence of positive first eigenfunctions, the existence theory of §3 applies directly to the problem of Lagrange

$$\sup_{\sigma \in ad} \mu_{i,j}(\sigma), \qquad 2 \le i + j \le 4. \tag{5.9}$$

We note that only for symmetric boundary conditions, i.e., $i = j$, should we expect an even optimal design. As $\mu_{i,j}(\sigma)$ is simple when $i+j < 4$, we deduce from Theorem 4.3 and conditions (4.5–4.7) that

$$\hat{\sigma}_{i,j}(x) = \alpha \quad \Rightarrow \quad \hat{\sigma}_{i,j}^{p-1}(x)|\hat{u}''(x)|^2 \le \ell^2, \tag{5.10}$$

$$\alpha < \hat{\sigma}_{i,j}(x) < \beta \quad \Rightarrow \quad \hat{\sigma}_{i,j}^{p-1}(x)|\hat{u}''(x)|^2 = \ell^2, \tag{5.11}$$

$$\hat{\sigma}_{i,j}(x) = \beta \quad \Rightarrow \quad \hat{\sigma}_{i,j}^{p-1}(x)|\hat{u}''(x)|^2 \ge \ell^2 \tag{5.12}$$

for almost every $x \in (0, 1)$, where $\hat{u} \in \mathcal{E}_{i,j}(\hat{\sigma}_{i,j})$. As before, $\hat{\sigma}_{i,j}$ is smooth where (5.11) holds.

The right side of (5.11) is the sole necessary condition offered by Keller [20] and Tadjbakhsh and Keller [37]. We now investigate the extent to which their claim is valid. Recall that their analysis of the clamped-clamped column erred in neglecting (a) double eigenvalues and (b) bounds on σ. As the previous lemma precludes the former phenomenon, we need only consider the latter. The observation to be made is that (5.10) and (5.12) are only needed near the zeros of $\hat{u}''$. As noted above, members of $\mathcal{E}_{1,1}(\sigma)$ and $\mathcal{E}_{2,0}(\sigma)$ have second derivatives of one sign. As such, in these cases, (5.11) stands on its own (with the minor adjustment that σ be allowed to vanish at zero and/or 1). In addition, as the related second-order problems are fully equivalent, i.e., there are no spurious eigenvalues, Tadjbakhsh and Keller's sufficiency proof is correct. In summary, Keller [20] has the correct necessary condition for the hinged-hinged column, Tadjbakhsh and Keller [37] have the correct necessary condition for the clamped-free column, and the proof of sufficiency in [37] holds for both. We now recall their analytical solutions to these problems.

Keller, in [20], with $p = 2$ and $i = j = 1$ reconciled (5.11) and (5.1) and found

$$\begin{aligned} \hat{\sigma}_{1,1}(x) &= \tfrac{4}{3}\sin^2\theta(x), \qquad 0 \le \theta \le \pi, \\ \theta(x) - \tfrac{1}{2}\sin 2\theta(x) &= \pi x, \qquad 0 \le x \le 1. \end{aligned} \tag{5.13}$$

We have observed that this is a shortened cycloid with parametrization

$$\begin{aligned} x(t) &= \tfrac{3}{4\pi}\left(\tfrac{2}{3}(t-\sin t)\right) \\ y(t) &= \tfrac{2}{3}(1-\cos t) \end{aligned} \qquad 0 \le t \le 2\pi.$$

This column buckles under an axial load of $4\pi^2/3$. In [37], Tadjbakhsh and Keller with $p = 2$ and $i = 2$, $j = 0$ reconciled (5.11) and (5.1) and found

$$\hat{\sigma}_{2,0}(x) = \tfrac{4}{3}\sin^2\theta(x), \qquad -\pi/2 \le \theta \le 0,$$

$$\theta(x) - \tfrac{1}{2}\sin 2\theta(x) + \pi/2 = \pi x/2, \qquad 0 \le x \le 1,$$

our parametrization being,

$$\begin{aligned} x(t) &= \tfrac{3}{2\pi}\left(\tfrac{2}{3}(t-\sin t)\right)+1 \\ y(t) &= \tfrac{2}{3}(1-\cos t) \end{aligned} \qquad -\pi \le t \le 0.$$

This column buckles under an axial load of $\pi^2/3$. Having argued in favor of the existing solutions to the clamped-free and hinged-hinged problems, we now turn to the clamped-hinged problem.

We saw in Lemma 5.1 that the second derivative of each function in $\mathcal{E}_{2,1}(\sigma)$ must change sign. The effect of this is that (5.11) forces $\hat{\sigma}_{2,1}$ to vanish at an interior point. In particular, when Tadjbakhsh and Keller reconciled (5.1) and (5.11) they found

$$\hat{\sigma}_{2,1}(x) = \frac{4\sin^2\theta(x)}{3\sin^2\theta(0)}, \qquad \theta(0) \le \theta \le \pi, \tag{5.14}$$

$$\theta(x) - \tfrac{1}{2}\sin 2\theta(x) + \tfrac{1}{2}\sin 2\theta(0) - \theta(0) = x(\pi + \tfrac{1}{2}\sin 2\theta(0) - \theta(0)), \qquad 0 \le x \le 1,$$

$$\tfrac{1}{2}\sin 2\theta(0) - \theta(0) = -\tfrac{2}{3}\sin^3\theta(0)\cos^{-1}\theta(0) - \pi.$$

Taking $a = \frac{1}{2}\sin 2\theta(0) - \theta(0)$, note that this $\hat{\sigma}_{2,1}(x)$ vanishes at $a/(\pi+a)$. Tadjbakhsh and Keller assert that the column built according to (5.14) will not buckle under loads less than approximately 27.22 in magnitude. We show in the appendix that this column cannot withstand loads exceeding $\pi^2/3$ – and so, in fact, is much weaker than the uniform column. In addition, as $\mu_{2,1}(\sigma)$ corresponds to the *second* eigenvalue of its associated second-order problem, the sufficiency proof of [37] is invalid. Hence, (5.14) is not an optimal design. In summary, (5.11) cannot stand alone in the clamped-hinged case, $\sigma(x) \ge \alpha$ is indeed an active constraint and (5.10) absolutely necessary. We suspect that there exists no solution to (5.9) when $ij = 2$ and $\alpha = 0$.

6. The finite-dimensional problem. We discretize the interval $[0, h, 2h, \cdots, (N-1)h = 1]$ and approximate $V_{i,j}$ by the finite-dimensional space $V_{i,j}^h$, the subspace of $V_{i,j}$ whose elements, when restricted to $[kh, (k+1)h]$, are cubic polynomials (see Strang and Fix [36]). As each member of $V_{i,j}^h$ is completely determined by the value of it and its derivative at each of the N mesh points, we identify $V_{i,j}^h$ with R^{2N-i-j}. We next approximate *ad* with the class of piecewise constant functions

$$ad^h \equiv \left\{ \sigma \in R^{N-1} : \alpha \le \sigma_k \le \beta, \sum_{k=1}^{N-1} \sigma_k = N-1 \right\}.$$

We have refrained from labeling elements of ad^h by σ^h to avoid confusion with powers of σ. In this context, the infinite-dimensional eigenvalue problem of (5.1) is now approximated by

$$B_h(\sigma)q_h = \mu K_h q_h, \quad \sigma \in ad^h, \quad q_h \in R^{2N-i-j}. \tag{6.1}$$

$B_h(\sigma)$ and K_h, the so-called bending and stiffness matrices, are each real, $(2N - i - j) \times (2N - i - j)$, symmetric, positive definite, and banded with half bandwidth of four. Our interest is, of course, in $\mu^h_{i,j}(\sigma)$, the least eigenvalue of (6.1). For, as $h \to 0$, one finds, e.g., in [36], that $\mu^h_{i,j}(\sigma) \to \mu_{i,j}(\sigma)$. The connection between the finite and infinite-dimensional problems now understood, we concentrate solely on (6.1). It should cause no confusion if, in our presentation of the finite-dimensional optimization problem, we suppress most dependence on h, i, and j. With this, (6.1) becomes

$$B(\sigma)q = \lambda K q, \quad \sigma \in ad^h, \quad q \in R^n, \tag{6.2}$$

and we denote its least eigenvalue by $\lambda_1(\sigma)$. Our finite-dimensional problem of Lagrange is now

$$\max_{\sigma \in ad^h} \lambda_1(\sigma). \tag{6.3}$$

The care that was taken in differentiating $\sigma \mapsto \lambda_1(\sigma)$ in §4 must also be exercised here. The occurrence of multiple eigenvalues is still possible. Clarke [10, Prop. 2.8.8] specifies the generalized gradient of the largest eigenvalue of a symmetric matrix in terms of a convex hull; see also [28] and [29]. Though such a characterization may suffice for an analytical description, as in §4, for computational purposes we have found it more useful to specify first-order conditions in terms of, less well-known, "dual matrices." We state the result in general terms. We shall need $\mathcal{S}^n$, the class of $n \times n$ real symmetric matrices, and the Frobenius matrix inner product, $\langle A, B \rangle = \operatorname{tr} A^T B$.

THEOREM 6.1. *Let $B : R^{N-1} \to \mathcal{S}^n$ be continuously Fréchet differentiable with $B_k(\sigma) = \partial B(\sigma)/\partial \sigma_k$ and let K be a fixed symmetric positive semidefinite matrix of the same order n. Assume $\sigma \in ad^h$ is such that $\lambda_1(\sigma)$ has multiplicity t, with corresponding eigenvectors given by the columns of a matrix $Q_1 \in R^{n \times t}$, normalized so that $Q_1^T K Q_1 = I$. Then a necessary condition for σ to solve* (6.3) *is that there exist a symmetric positive semidefinite matrix U of order t, with trace equal to 1, and Lagrange multipliers ν and γ_k, $k = 1, \cdots, N-1$, such that*

$$\langle U, Q_1^T B_k(\sigma) Q_1 \rangle = \nu + \gamma_k, \tag{6.4}$$

and

$$\sigma_k = \alpha \quad \Rightarrow \quad \gamma_k \le 0, \tag{6.5}$$

$$\alpha < \sigma_k < \beta \quad \Rightarrow \quad \gamma_k = 0, \tag{6.6}$$

$$\sigma_k = \beta \quad \Rightarrow \quad \gamma_k \ge 0 \tag{6.7}$$

for each k. Furthermore, this condition is also sufficient for optimality in the case that $\sigma \mapsto B(\sigma)$ is affine.

Proof. In the following we use the notation $U \ge 0$ to mean that a symmetric matrix U is positive semidefinite. Regarding $\lambda_1 : \mathcal{S}^n \to R$, we invoke Rayleigh's principle in

$$\lambda_1 = \min\{\langle q, Bq\rangle;\ q \in R^n,\ \langle q, Kq\rangle = 1\}$$
$$= \min\{\langle qq^T, B\rangle;\ q \in R^n,\ \langle q, Kq\rangle = 1\}.$$

Let $Q \in R^{n\times n}$ be any matrix satisfying

$$Q^T K Q = I. \tag{6.8}$$

It is easily shown that

$$\text{co}\{qq^T;\ q \in R^n,\ \langle q, Kq\rangle = 1\} = \{Q\hat{U}Q^T;\ \hat{U} \in \mathcal{S}^n,\ \text{tr}\,\hat{U} = 1,\ \hat{U} \geq 0\},$$

by using the spectral decomposition of $\hat{U}$, which by assumption has nonnegative eigenvalues adding to one, to obtain the requisite convex combination showing that the second set is contained in the first. It follows that

$$\lambda_1 = \min\{\langle Q\hat{U}Q^T, B\rangle;\ \hat{U} \in \mathcal{S}^n,\ \text{tr}\,\hat{U} = 1,\ \hat{U} \geq 0\}. \tag{6.9}$$

Now take Q to be a matrix whose columns are eigenvectors of (6.2), normalized so that (6.8) holds. The first t columns of Q are the columns of Q_1 and

$$Q^T B Q = \ \text{Diag}(\lambda_i),$$

where $\lambda_1 \leq \lambda_2 \leq \ldots$ are the eigenvalues of (6.2), repeated according to multiplicity. Therefore, the matrices achieving the minimization in (6.9) are those defined by

$$\hat{U} = Q\begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix} Q^T = Q_1 U Q_1^T,$$

where $U \in \mathcal{S}^t$, with $\text{tr}\,U = 1$ and $U \geq 0$. Consequently, the generalized gradient of $B \mapsto -\lambda_1(B)$ is the set of such matrices $\hat{U}$ (see Rockafellar [33, pp. 29 and 35] or Clarke [10, §2.8]), no convex hull operation being required since the set of such $\hat{U}$ is convex.

With $\lambda_1(\sigma) = (\lambda_1 \circ B)(\sigma)$, the desired necessary conditions now follow from (i) the chain rule for generalized gradients [10, Thm. 2.3.10], (ii) the standard Lagrange multiplier rule [10, Thm. 6.1.1 and Thm. 6.4.4], and (iii) properties of the inner product. In particular,

$$\partial(-\lambda_1(\sigma)) = \{v \in R^{N-1};\ v_k = \langle U, Q_1^T B_k(\sigma) Q_1\rangle,\ U \in \mathcal{S}^t,\ U \geq 0, \text{tr}\,U = 1\}. \tag{6.10}$$

These necessary conditions are also sufficient in the case that $\sigma \mapsto B(\sigma)$ is affine because the composition of a concave function with an affine function is concave. □

Our attention to $\partial(-\lambda_1)$ in (6.10) and $\partial\lambda_1^{-1}$ in Theorem 4.3, rather than simply $\partial\lambda_1$, is merely an artifact of Clarke's concern with functions defined as pointwise maxima rather than minima. Here, it was convenient to characterize $-\lambda_1$ as the maximum of a Rayleigh quotient where, in §4, we found it more advantageous to maximize a functional of Auchmuty, and hence, to consider λ_1^{-1}.

A simpler version of this theorem (for the unconstrained case, with $K = I$ and $B(\sigma)$ an affine function) was given by Overton [26], following work of Fletcher [14]. The $n \times n$ matrix $\hat{U}$ is known as a "dual matrix" by analogy with "dual variables"

(Lagrange multipliers) familiar from mathematical programming. The $t \times t$ matrix U may be called a "reduced dual matrix," but since it is the one we shall need as a computational tool we shall also refer to it as simply the dual matrix. The distinction between $\hat{U}$ and U is analogous to the notational question of whether inactive constraints in a nonlinear program should be assigned zero Lagrange multipliers.

In the case that $t = 1$ and the bound constraints are inactive, the necessary condition reduces to the requirement that the gradient of $\lambda_1(\sigma)$, whose elements are $q_1^T B_k(\sigma) q_1$, has the constant value ν. (Here q_1 is the only column in Q_1, and U is the scalar one.) In the case that $t = 2$, let

$$U = \begin{pmatrix} \delta_1 & \delta_3/2 \\ \delta_3/2 & \delta_2 \end{pmatrix}. \tag{6.11}$$

Let the two columns of Q_1 be q_1 and q_2 and again assume that all bound constraints are inactive. The necessary condition then becomes

$$\delta_1 q_1^T B_k(\sigma) q_1 + \delta_2 q_1^T B_k(\sigma) q_2 + \delta_3 q_1^T B_k(\sigma) q_2 = \nu,$$

together with the trace and positive semidefinite constraints on U. Without loss of generality, δ_1 and δ_2 may be taken to have nonnegative sign, and the normalizing trace condition may be replaced by the assumption that $\nu = 1$. The positive semidefinite constraint is then simply

$$\delta_1 \delta_2 \geq \frac{\delta_3^2}{4}.$$

This is the same necessary condition given by Bratus and Seiranian [7] and Masur [22]. We note that the derivation given here not only applies for $t > 2$, but is much simpler than that given by [7] and [22] for the case $t = 2$. In a footnote, Masur conjectured that the positive semidefinite condition on U would also be the correct necessary condition for $t > 2$. Bratus [6] gave a discussion of necessary and sufficient conditions for general multiplicity t, but the given necessary condition concerns the necessary sign of the directional derivative of λ_1 for all feasible directions; the positive semidefinite condition on U was apparently not obtained.

Before discussing the algorithm that springs from Theorem 6.1, we will investigate the extent to which it suggests a new tack on the infinite-dimensional problem. Regarding the variational principle of (6.9), we consider $K^+(X)$, the space of positive compact linear operators on a real separable Hilbert space X. Each $T \in K^+(X)$ possesses a countable sequence of eigenvalues $\lambda_1(T) \geq \lambda_2(T) \geq \cdots \downarrow 0$ repeated according to multiplicity and a (possibly infinite) trace $\operatorname{tr} T = \sum_{i=1}^{\infty} \lambda_i$. In this context it is not difficult to show for symmetric $T \in K^+(X)$ that

$$\lambda_1(T) = \max \{\operatorname{tr} TU;\ U \in K^+(X),\ \operatorname{tr} U = 1\}.$$

Recall that $u \mapsto (\sigma^p u'')''$ and $u \mapsto -u''$ are positive symmetric isomorphisms of H_0^2 onto H^{-2} and H_0^1 onto H^{-1}, respectively. We denote these maps by A and B, remark that $B^{1/2}$ is a positive isomorphism of H_0^1 onto L^2, and denote by I the compact imbedding of H_0^2 in H_0^1. With $\phi = B^{1/2} J u$, and $*$ denoting adjoint, the buckled column equation receives the formulation

$$(1/\lambda)\phi = B^{1/2} I A^{-1} (B^{1/2} I)^* \phi.$$

By construction, $B^{1/2}IA^{-1}(B^{1/2}I)^*$ is a symmetric member of $K^+(L^2)$. Although we may now proceed to compute $\partial\lambda_1^{-1}$, as in the previous theorem, this representation suffers from its dependence on the unknown A^{-1} and $B^{1/2}$ in contrast to Theorem 4.3 which works directly with A and B.

We now turn to the question of how to solve the finite-dimensional optimization problem. Although there is a substantial literature on the generalization of gradient methods to nonsmooth problems (see the survey [28] by Polak), little attention has been given to applying nonsmooth optimization techniques to (6.3). One exception is [29], which describes an algorithm for maximizing the least eigenvalue of a variety of important structures, accounting for the presence of multiple eigenvalues. It focuses however on the clamped vibrating column, $(\sigma^p u'')'' = \lambda u$, $u \in H_0^2$, a problem which has long been known to admit only simple eigenvalues (see, e.g., Leighton and Nehari [21, Lemma 4.1]). Our algorithm differs from that of [29] in our attention to the added structure of the generalized gradient of λ_1 as revealed in the theorem above. We use an algorithm specifically designed to exploit this structure, which is based on Overton [26], but modified to be far more efficient for moderate to large mesh size N.

Given $\sigma \in ad^h$ with $\lambda_1(\sigma)$ and $\lambda_2(\sigma)$ the two least eigenvalues of (6.2), we normalize the corresponding eigenvectors q_1 and q_2 so that $Q_1 = [q_1\; q_2]$ satisfies $Q_1^T K Q_1 = I$, the 2×2 identity matrix. These eigenvalues and eigenvectors are computed by subspace iteration with a block size of two, with the necessary linear systems solved directly using the Cholesky factors of $B(\sigma)$ (see Bathe and Wilson [5] for details).

Define the approximate multiplicity t of λ_1 by $t = 2$ if

$$\lambda_2(\sigma) - \lambda_1(\sigma) \leq \tau\lambda_1(\sigma)$$

and $t = 1$ otherwise. Here τ is a positive tolerance which may be adjusted during the optimization process. A multiplicity higher than two is excluded (for sufficiently small h) by Lemmas 2.1 and 5.1. Now consider the following linear program (LP):

(LP0) $$\max_{d\in R^N} d_N$$

subject to

(LP1) $$Ed = e,$$

(LP2) $$Fd \leq f,$$

(LP3) $$d_k = 0, \qquad k \in J,$$

(LP4) $$\alpha - \sigma_k \leq d_k \leq \beta - \sigma_k, \qquad k = 1, \cdots N-1,$$

(LP5) $$|d_k| \leq \rho, \qquad k = 1, \cdots, N,$$

where

(i) The first $N-1$ components of d represent proposed changes to σ, while the last component approximates the corresponding change in $\lambda_1(\sigma)$. Let us write $d = [\eta^T \omega]^T$, with $\eta \in R^{N-1}$, $\omega \in R$;

(ii) ρ is a positive scalar, whose purpose is to ensure $\|d\|$ is not too large;

(iii) J is an index set, which effectively removes the corresponding variables from the linear program;

(iv) The first row of the matrix E is $[1, \cdots, 1, 0]$, and the first element of the right-hand side vector e is zero. This ensures that the changes to σ respect the integral constraint;

(v) The second row of E is

$$[-q_1^T B_1(\sigma) q_1, \cdots, -q_1^T B_{N-1}(\sigma) q_1, 1]$$

and the corresponding element of e is zero. If $t = 2$, then E contains an additional two rows,

$$[-q_2^T B_1(\sigma) q_2, \cdots, -q_2^T B_{N-1}(\sigma) q_2, 1] \quad \text{and} \quad [-q_1^T B_1(\sigma) q_2, \cdots, -q_1^T B_{N-1}(\sigma) q_2, 0],$$

with corresponding elements of e set to $\lambda_2(\sigma) - \lambda_1(\sigma)$ and zero, respectively;

(vi) If $t = 1$, F contains the single row

$$[-q_2^T B_1(\sigma) q_2, \cdots, -q_2^T B_{N-1}(\sigma) q_2, 1]$$

and f is the scalar $\lambda_2(\sigma) - \lambda_1(\sigma)$. If $t = 2$, F and f are empty, i.e., (LP2) may be removed.

We note that the given rows may be computed very efficiently, since the derivative matrices $B_k(\sigma)$ are extremely sparse.

The justification for (v) and (vi) is as follows. If $t = 1$, the second row of E imposes a linearization of the nonlinear equation $\lambda_1(\sigma + \eta) = \lambda_1(\sigma) + \omega$, while the only row in F imposes a linearization of the inequality $\lambda_2(\sigma + \eta) \geq \lambda_1(\sigma) + \omega$. Since ω is being maximized, the solution to the linear program yields the steepest ascent direction for $\lambda_1(\sigma)$, projected to satisfy the integral and bound constraints, with steplength required to be short enough that the linearized value for $\lambda_2(\sigma + \eta)$ does not drop below that for $\lambda_1(\sigma + \eta)$, and that the various bounds are satisfied. If $t = 2$, the second through fourth rows of E give a linearization of the appropriate set of three nonlinear equations, imposing the coalescence of $\lambda_1(\sigma + \eta)$ and $\lambda_2(\sigma + \eta)$, see [16]. The common linearized value, $\lambda_1(\sigma) + \omega$, is maximized, subject to the given constraints.

THEOREM 6.2. *Suppose that $\tau = 0$ so that the multiplicity estimate t is exact, and suppose that $\rho > 0$ and J is the empty set. Then $d = 0$ is a (nonunique) solution to the linear program given above if and only if* (6.4) *holds for some $U \in S^t$ with* $\operatorname{tr} U = 1$, *some $\nu \in R$ and some $\gamma \in R^{N-1}$ satisfying* (6.5)–(6.7).

Proof. By the usual Lagrange multiplier rule, the linear program admits the solution $d = 0$ if and only if there exist multipliers $\nu \in R$, $\delta \in R^{t(t+1)/2}$ and $\gamma \in R^{N-1}$ satisfying

$$E^T \begin{pmatrix} \nu \\ \delta \end{pmatrix} + \begin{pmatrix} \gamma \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

with γ subject to the standard sign condition. Setting $U = \delta_1 (= 1)$ if $t = 1$ and defining U by (6.11) if $t = 2$, we have (6.4)–(6.7) with $\operatorname{tr} U = 1$. The same argument holds in the reverse direction. □

Note two points: there is no positive semidefinite condition obtained on U, and the solution $d = 0$ is generally not at a vertex of the feasible region, so is not unique.

Our algorithm for solving (6.3) generates a sequence in ad^h. Each successive approximation is obtained from the previous one by consideration of a linear program

(LP) of the form given above. We first define a simple version of the algorithm, but one which is too costly for practical use. In this version, we obtain $\sigma+\eta$, a candidate replacement for σ, by solving the LP for $d=[\eta^T\omega]^T$. If $\lambda_1(\sigma+\eta)>\lambda_1(\sigma)$, σ is replaced by $\sigma+\eta$ and the process repeated. Otherwise, σ remains unchanged, the trust region radius ρ is decreased by a factor of two, and the revised LP is considered. This kind of trust region approach can be made very effective by modifying the size of ρ according to how well the actual increase in $\lambda_1(\sigma)$ agrees with the linear prediction ω, as discussed in Fletcher [14] in the context of general nonlinear programming. As recommended by Fletcher, we double ρ if the ratio of actual to predicted increase exceeds 0.75 and halve ρ if the ratio is less than 0.25. The process is terminated when $\|d\|\le\epsilon$, a convergence tolerance.

However, the expense of obtaining the optimal solution of each linear program is not justified. Although the "limit" LP defined by σ equal to a solution of (6.2)–(6.4) has an optimal solution which is not a vertex, generically, any LP solved during the successive approximation process can be expected to have a unique solution which must be at a vertex. Since only a few constraints involve all the variables, most of the constraints defining a vertex are simple bounds, and most of these may be trust radius bounds of the form (LP5). Since the only purpose of the trust radius bounds is to avoid taking steps too large for the linearizations to be accurate, there is little to be gained by finding the exact set of active bounds. We, therefore, partially solve the LP as follows. We first attempt to obtain a feasible point for the LP by setting d to the least norm solution of (LP1, LP3), contracting this step if necessary to satisfy the various inequalities. This contraction effectively scales the right-hand side of the only possible inhomogeneous equality constraint in the LP, that corresponding to the third row of E in the case $t=2$. The rationale here is that if the least norm step satisfying the equality constraints is not feasible, the underlying approximations are probably not good enough to justify the solution of the unmodified LP. We then start the LP solution process as in a projected gradient method, augmenting d by projected gradient steps with steplengths determined by the inequality constraints and bounds. The gradient being projected is that of the LP objective, i.e., the vector $[0,\cdots,0,1]$, while the constraints determining the projection are the equality constraints and any inequality constraint and bounds encountered during the process. This continues until either (a) a trust radius bound of the form (LP5) is encountered, or (b) the norm of the current projected gradient increment drops below the tolerance ϵ (unlikely to happen first). At this point the process of partially solving the LP is terminated. Anywhere from zero to many active bounds of the form (LP4) may be encountered by this process, as well as, possibly, the inequality (LP2) (in the case $t=1$). By adding any active bounds encountered to the set J, we avoid having to process these bounds again during the (partial) solution of subsequent LP's. However, the signs of the associated bound multipliers must be checked after the partial LP solution and bounds with the wrong sign removed from J if necessary. The entire process is very efficient, requiring QR factorizations of matrices with only two to four columns, with rows removed corresponding to active bounds. For complete details of the process, see [27].

In the case $t=2$, the LP partial solution process generates four multipliers corresponding to the rows of E, namely ν,δ_1,δ_2, and δ_3. If the corresponding dual matrix U, defined by (6.11), is not positive semidefinite, this is a clear indication that the multiplicity estimate t is incorrect, and so τ is reduced by a factor of ten. In principle, it might be necessary to use a more sophisticated technique to recover from a multiplicity estimate which is too large. For example, if the algorithm was started at a point where all the optimality conditions except $U \geq 0$ were satisfied, it would be necessary to split the multiple eigenvalue to obtain an ascent direction; this is explained further in [26] and [27]. However, this technique has not been required in our computational experiments for the Lagrange problem.

This completes our outline of the algorithm used to generate the numerical results given in the next section. For more algorithmic details, see [27]. We do not have any proof that the given algorithm will converge to a solution of (6.3), but given any approximate solution, we may verify the required signs of γ and the eigenvalues of the dual matrix U and compute the residual of the approximate equation (6.4). We have found the algorithm to be very effective in practice as the numerical results attest.

7. Computational results. The algorithm outlined in the previous section has been implemented in Fortran 77 and tested extensively. Subroutines from the Linpack [13] and Eispack [17] libraries were used to (a) perform the QR factorizations required during the partial LP solution process (for matrices with at most four columns), (b) obtain the Cholesky factorizations of $B(\sigma)$ needed for subspace iteration (these matrices have only seven nonzero diagonals), and (c) solve the reduced generalized eigenvalue problems required for subspace iteration (these have order two). Parameters were set as follows: τ, the relative multiplicity tolerance, was initialized to 0.1; ρ, the trust radius, was initialized to 5.0; ϵ, the convergence tolerance, was set to 10^{-3}. The initial σ was set to the constant one, corresponding to the uniform column. Runs were made for various values of N, the number of mesh points; p, the power of σ in the differential equation; α and β, the lower and upper bounds on σ, and the various homogeneous boundary conditions: clamped-clamped, clamped-hinged, clamped-free, and hinged-hinged.

The algorithm was found to be very efficient, typically invoking subspace iteration, in the computation of the two least eigenvalues of (6.2), about 50 times prior to reaching the convergence tolerance. At the final iterate, the residual of the approximate equation (6.4) was typically found to have norm about 10^{-3}. Smaller choices of ϵ required significantly more computation but did not produce a qualitatively improved solution. Other initial choices of ρ affected only the initial efficiency of the algorithm. The results were relatively insensitive to the initial choice of the multiplicity tolerance τ, although smaller choices delayed identification of the final multiplicity. There was usually no difficulty in determining the correct final multiplicity t, with corresponding positive semidefinite dual matrix U. In the cases where the final multiplicity t was two, the gap between λ_1 and λ_2 was typically reduced to 10^{-6}. The subspace iteration was itself very efficient, requiring only one or two steps on all but the first few steps of the optimization, reflecting the good separation of λ_2 from λ_3 and the availability of an excellent initial two-dimensional subspace, namely the span of the eigenvectors q_1 and q_2 computed at the previous optimization step. (The first subspace iteration was initialized using the first two columns of the identity matrix.) The initial σ in each case was that of the uniform column, $\sigma \equiv 1$. Symmetry was not imposed on the algorithm's subsequent choices of σ. A typical run for $N = 513$ required 1.5 hours on a Sparc station.

We begin our summary of the results with $p = 2$. Under the assumption that each transverse cross section of the column is circular, we recall that σ is proportional to the square of the cross section's radius. Plotting both $\pm\sqrt{\sigma}$ then gives a lengthwise cross section of the associated column. With this representation we may then view the corresponding buckling mode(s) simultaneously. Our figures, generated by Matlab

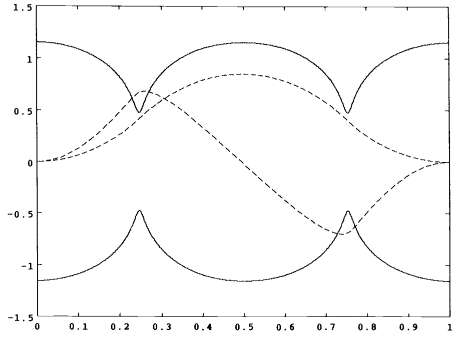


FIG. 1. *Strongest clamped-clamped column and first two buckling modes.* $p = 2$, $\alpha = 0$.

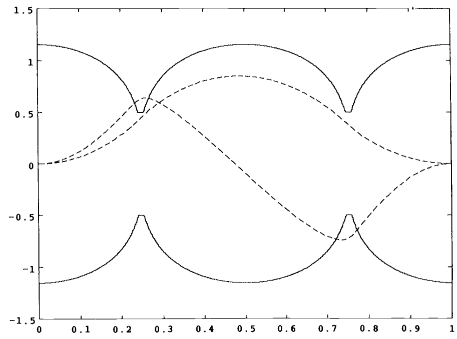


FIG. 2. *Strongest clamped-clamped column and first two buckling modes.* $p = 2$, $\alpha = .25$.

[23], portray the column in the piecewise fashion produced by the algorithm of §6, while using dashed curves(s) to indicate the buckling mode(s). We remark that for those optimal designs with double buckling loads the corresponding buckling modes depend on our initial choice of subspace, in subspace iteration, and σ.

Figure 1 gives our strongest clamped-clamped column and its first two buckling modes. Here $p = 2$, $\alpha = 0$, and $\beta = 10$, with a double buckling load of 52.3533. This value agrees to four figures with that obtained by Olhoff and Rasmussen [25], Masur [22], and Seiranian [34].

On increasing α or decreasing β, these bound(s) will eventually become active. Figure 2 gives our strongest clamped-clamped column and its first two buckling modes when $p = 2$, $\alpha = 0.25$, and $\beta = 10$. The least eigenvalue is still double, though now reduced to 52.3467.

As the uniform column has a simple first eigenvalue, we would expect that a sufficient increase in α would produce an optimal design with a simple first eigenvalue. Figure 3 gives our best strongest clamped-clamped column and its first buckling mode when $p = 2$, $\alpha = 0.5$, and $\beta = 10$. In this case the first two eigenvalues are 51.07086 and 62.3479.

Clearly there must exist (at least) one value of α between $\frac{1}{4}$ and $\frac{1}{2}$ at which the optimal buckling load switches multiplicity. Olhoff and Rasmussen [25] declare 0.28 to be the only such value. Our algorithm also indicates the presence of such a critical α in the vicinity of 0.28. We note that in addition to being able to approach 0.28 from above—proceeding until the gap between the least two eigenvalues vanishes—we

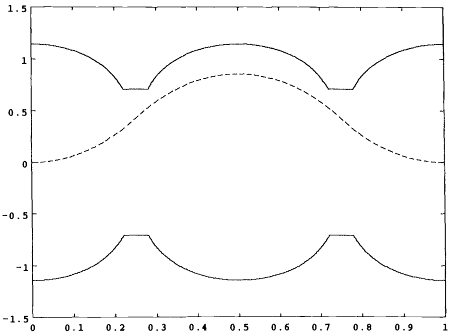


FIG. 3. *Strongest clamped-clamped column and first buckling mode. $p = 2$, $\alpha = .5$.*

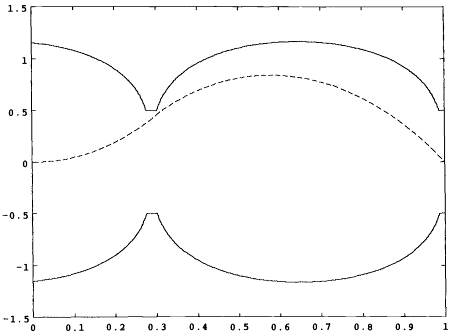


FIG. 4. *Strongest clamped-hinged column and first buckling mode. $p = 2$, $\alpha = .25$.*

have also approached from below, in this case proceeding until the least eigenvalue of the corresponding 2 by 2 dual matrix vanishes.

Figure 4 gives our strongest clamped-hinged column and its first buckling mode when $p = 2$, $\alpha = 0.25$, and $\beta = 10$. The buckling load of 27.0762 is, as expected, simple. Although decreasing α increases the buckling load, our designs converge, as $\alpha \to 0$, to the Tadjbakhsh and Keller solution, (5.13). As shown in the appendix, this column buckles at $\pi^2/3$, and so cannot possibly be optimal. The convergence of our algorithm to (5.13) only strengthens our belief that the problem, as stated by Tadjbakhsh and Keller, is without a solution. That is, $\sigma \mapsto \mu_{2,1}(\sigma)$ with $p = 2$, does *not* attain its maximum on *ad* when $\alpha = 0$.

Our numerical results also agree with Tadjbakhsh and Keller in the cases for which we have argued that they are correct. In particular, Fig. 5 gives our strongest clamped-free column and first buckling mode when $p = 2$, $\alpha = 0$, and $\beta = 10$. The buckling load, again simple, is 3.2897. Fig. 6 gives our strongest hinged-hinged column and first buckling mode when $p = 2$, $\alpha = 0$, and $\beta = 10$. Its simple buckling load is 13.1579.

We return to the clamped-clamped case and consider its dependence on p. Our analysis of (4.5–4.7) led us to believe that, for $p > 1$, the minimum (maximum) of the optimal design increases (decreases) with p. This is reinforced by Fig. 7, whose lower (upper) curve traces the minimum (maximum) of the optimal design as a function of p. As the buckling load is double for each of these designs, there must exist a curve, between the lower one and the curve that is constantly one, across which the

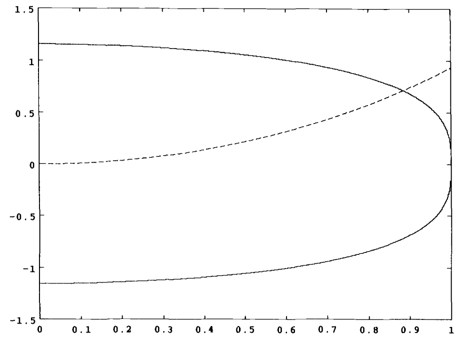


FIG. 5. *Strongest clamped-free column and first buckling mode.* $p = 2, \ \alpha = 0$.

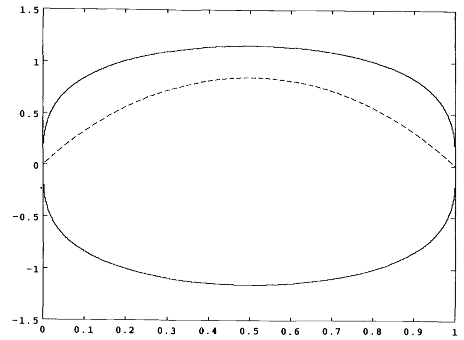


FIG. 6. *Strongest hinged-hinged column and first buckling mode.* $p = 2, \ \alpha = 0$.

optimal buckling load changes multiplicity. We have seen that $(2, 0.28)$ lies near such a curve. With respect to the range of p considered in Fig. 7, we have found that both the optimal buckling load and the least eigenvalue of its corresponding dual matrix increase with p. Regarding the behavior as p tends to 1 from above, we have found that the minimum of the optimal design tends to zero, and, though the optimal buckling load remains double, the least eigenvalue of the corresponding dual matrix tends to zero. Below $p = 1$ we found optimal designs with simple buckling loads regardless of our choice of α. Figure 8 gives our strongest clamped-clamped column when $p = 1, \ \alpha = 0$, and $\beta = 10$. The buckling load of 47.9898 and the design itself are very close to the analytical result of (4.15). Refining the mesh in neighborhoods of $\frac{1}{4}$ and $\frac{3}{4}$, and perhaps using piecewise linear elements for σ, would presumably bring us even closer to (4.15). We have not pursued this for two reasons. First, we argued in §4 that in the absence of an existence proof we cannot fully trust (4.15), and second, in both of the physical contexts where $p = 1$ has arisen, there is an a priori strictly positive lower bound on the admissible σ. Regarding the latter, we present in Fig. 9 our strongest clamped-clamped column when $p = 1, \ \alpha = 0.8$, and $\beta = 1.2$. Its simple buckling load is 43.4921.

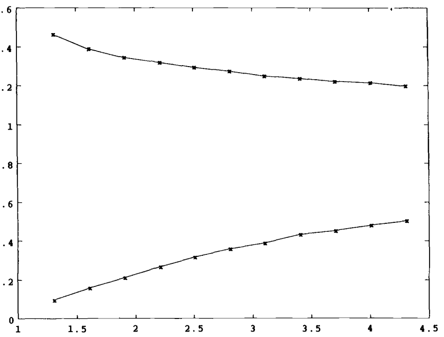


FIG. 7. *Maximum and minimum of* $\hat{\sigma}_{2,2}$ *vs. p.*

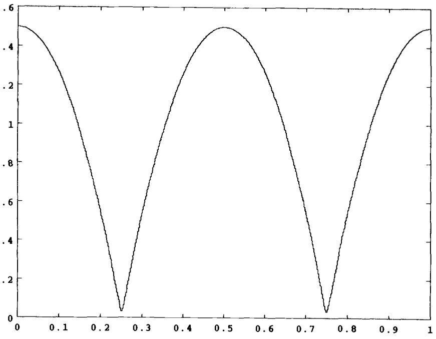


FIG. 8. $\hat{\sigma}_{2,2}$ *vs.* x. $p = 1$, $\alpha = 0$.

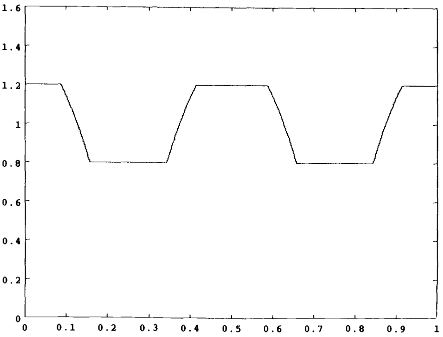


FIG. 9. $\hat{\sigma}_{2,2}$ *vs.* x. $p = 1$, $\alpha = 0.8$, $\beta = 1.2$.

Having addressed dependence on p and α at a particular level of discretization, we now fix $p = 2$, $\alpha = 0$, and $\beta = 10$ and with clamped-clamped boundary conditions demonstrate the convergence of several relevant parameters as N, the number of mesh points, becomes large. In particular, Table 1 lists $\hat{\mu}_{2,2}$ (the optimal buckling load), the least eigenvalue of the associated dual matrix U, and $\|\hat{\sigma}_N - \hat{\sigma}_{1025}\|_\infty$ (the greatest difference between the optimal design on a mesh of N points and the optimal design on a mesh of 1025 points) for values of N from 65 to 1025.

We close our study with a glance at the numerical range of the buckling load over

TABLE 1

N	$\hat{\mu}_{2,2}$	min ev(U)	$\|\hat{\sigma}_N - \hat{\sigma}_{1025}\|_\infty$
65	52.14944	0.023859	0.1066
129	52.31027	0.043317	0.0415
257	52.33615	0.047034	0.0424
513	52.35332	0.046435	0.0059
1025	52.35548	0.046607	0.0000

ad. To this point we have concentrated on its maximization, and, though we may compare this value to that of the associated uniform column, it would be of interest to weigh it against the minimum buckling load. Clearly, α must now be strictly positive, for a buckling load of zero could be produced by prescribing that σ vanish on some interval. Regarding the existence of a minimizer for $\sigma \mapsto \mu_{i,j}(\sigma)$ over *ad*, we note that Theorem 3.1 is insufficient. Recall in (3.3) that the limit of the maximizing sequence integrated to less than one. This was not an obstacle, for adding mass could only increase the buckling load. As our goal now is to minimize this load, it appears that we must either relax the cost functional or construct the so called G-closure of *ad* to obtain a minimizing design. Instead of embarking on this, we modified the strongest column algorithm to minimize instead of maximize $\sigma \mapsto \mu_{i,j}(\sigma)$.

The modification to the algorithm is very simple, namely changing the sign of (LP0) and requiring a decrease instead of an increase in the smallest eigenvalue. The modified algorithm generated plausible weakest designs in *ad*, and, though we lack a proof of optimality, we shall content ourselves with a discussion of these numerical results. In all cases (independent of p and α) the minimum buckling load was simple; this is to be expected since the minimization should tend to separate the least eigenvalue from the remainder of the spectrum. In addition, we find that the generated designs have their mass concentrated near the inflection points of their associated positive buckling mode. This, too, is to be expected, making the opposite argument to that made in §2.

Figure 10 gives our weakest clamped-clamped column with $p = 2$, $\alpha = 0.25$, and $\beta = 10$. This column buckles under a simple axial load of 2.5658. The buckling load of our strongest column in this class (see Fig. 2) is 52.3467.

Figure 11 gives our weakest clamped-clamped column with $p = 1$, $\alpha = 0.8$, and $\beta = 1.2$. This column buckles under the simple axial load of 33.5631. The buckling load of our strongest column in this class (see Fig. 9) is 43.4921.

8. Concluding remarks. Having thoroughly discussed the work of Tadjbakhsh and Keller [37], we now take up two related issues, that, though broached by Keller in [20], have received little rigorous attention since.

The first involves the optimal design of cylindrical columns. Here, given again a fixed amount of material to be distributed over a column of fixed length, we seek the shape of the cross-section that, when used to generate a cylinder, produces a column with the greatest buckling load. We are not allowed to "taper" the column as we have in past sections. Keller quickly reduced this problem to the search for that planar domain of fixed area with the greatest least second moment of area. Recall that the

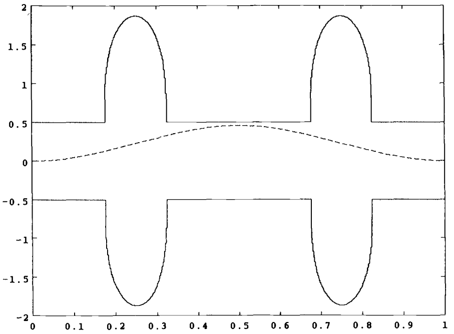


FIG. 10. *Weakest clamped-clamped column and first buckling mode.* $p = 2$, $\alpha = .25$.

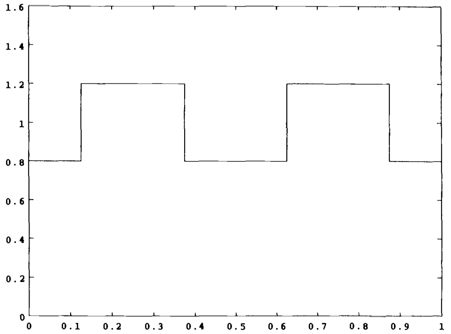


FIG. 11. *Weakest clamped-clamped design.* $p = 1$, $\alpha = 0.8$, $\beta = 1.2$.

second moment of area of the domain Ω with centroid at the origin in the direction η (with $|\eta| = 1$) is

$$I(\Omega, \eta) \equiv \int_{\Omega} |\eta^T y|^2 \, dy. \tag{8.1}$$

Denoting the unit circle by S, Keller's problem takes the form

$$\sup_{|\Omega|=A} \inf_{\eta \in S} I(\Omega, \eta). \tag{8.2}$$

Keller noted the existence of Ω for which this value is infinite. To exclude such Ω he restricted himself to convex domains. Within this smaller class he then argued, without proof, that the equilateral triangle possesses the greatest least second moment of area. Citing Pólya, Truesdell later observed that Keller's conjecture was indeed true. In particular, Pólya in [30] found that among convex sets the maximum of $4I_1 I_2 |\Omega|^{-4}$ occurs for triangles. Here I_1 and I_2 are the respective principal second moments of area, i.e., the min and max of (8.1). Keller's result follows, noting that only sets for which $I_1 = I_2$ need be considered, for if $I_1 < I_2$ we can simply redistribute the mass in such a way that I_1 increases while I_2 decreases.

This reduction to convex domains is too severe. It is not the lack of convexity that allows (8.2) to grow without bound but the possibility that Ω itself may be unbounded, though of finite area. To exclude this behavior, we may simply bound

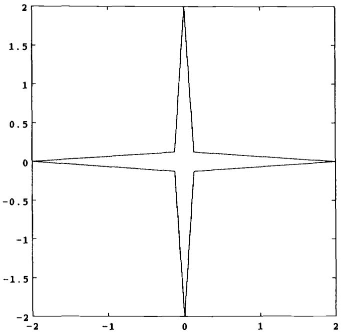


FIG. 12. *A domain whose second moment of area exceeds that of the equilateral triangle of the same area.*

$|\partial\Omega|$, the length of Ω's boundary. The fact that this does indeed bound $I(\Omega,\eta)$ follows from the isoperimetric inequality

$$2I_p(\Omega)/\pi \leq (|\partial\Omega|/2\pi)^4$$

of Pólya and Szegö [31, §1.5], where $I_p(\Omega)$ is the polar moment of inertia. As the second moment of area will be independent of η for the best Ω, its value will be one half that of its polar moment of inertia. We must now consider,

$$\sup_{\substack{|\Omega|=A \\ |\partial\Omega|\leq L}} \inf_{\eta\in S} I(\Omega,\eta). \tag{8.3}$$

For fixed A the value of (8.3) will grow as L is increased, suggesting that for sufficiently large L one may produce a domain whose second moment of area exceeds that of the equilateral triangle of the same area. We shall accomplish this with $A = 1$ and $L = 16$. To produce large values in (8.3), we need only consider domains that are symmetric about the coordinate axes, as well as the two diagonals that stretch out towards infinity. The symmetry will render $\eta \mapsto I(\Omega,\eta)$ constant, while the latter condition will ensure us that $I(\Omega,\eta)$ is large. The domain of Fig. 12 has an area of one, a boundary whose length does not exceed 16, and a second moment of area of $\frac{16}{45}+\frac{71}{8^3 90}$. This value is more than three times greater than $1/6\sqrt{3}$, the second moment of area of the equilateral triangle of the same area. Though we have not solved (8.2), this example demonstrates that (8.2) produces, through the designer's choice of L, columns with arbitrarily large buckling load.

The other issue we wish to resurrect from Keller [20] is that of post-buckling. Analysis of the buckled column must proceed from the nonlinear model. For example, in the hinged-hinged case we consider

$$(\sigma^p\phi')' + \lambda\sin\phi = 0, \quad \int_0^1 \sin\phi\, ds = 0, \quad \phi'(0) = \phi'(1) = 0, \tag{8.4}$$

where ϕ measures the angle between the column and a fixed axis in its plane of buckling. Equation (5.1) arises from linearizing $\sin\phi$ to ϕ, identifying $u' = \phi$, and differentiating the differential equation in (8.3). The least eigenvalue of (5.1), $\mu_{1,1}$, is indeed a bifurcation point for (8.3). Moreover, Keller showed that the direction of the solution branch emanating from $\mu_{1,1}$ was indeed rightward, i.e., that (8.3) admits only the trivial solution for $\lambda < \mu_{1,1}$. It remains to determine the nature of the nontrivial solution branch(es) for the other sets of boundary conditions. Here we would also like to understand the role of imperfections in the model and/or design. In particular, the splitting of the double eigenvalue in the optimal clamped-clamped column via an unfolding of the ideal bifurcation problem in a parameter that introduces material asymmetry would be of interest.

A.1. Appendix. We have argued throughout that the necessary and sufficient conditions proposed by Tadjbakhsh and Keller are incorrect. This does not in itself, however, invalidate their designs. Indeed, we argued that their solutions to the hinged-hinged and clamped-free problems are correct. This appendix serves to demonstrate that these are their only correct designs.

In particular, we show that Tadjbakhsh and Keller incorrectly calculated the buckling loads of their proposed solutions to the clamped-clamped and clamped-hinged problems. Recall their solution of the former,

$$\begin{aligned} A(x) &= \tfrac{4}{3}\sin^2\theta(x), \qquad -\pi/2 \le \theta \le 3\pi/2, \\ \theta(x) - \tfrac{1}{2}\sin 2\theta(x) + \pi/2 &= 2\pi x, \qquad 0 \le x \le 1. \end{aligned} \tag{A.1}$$

As Olhoff and Rasmussen [25] observed in their numerical work, the column constructed according to (A.1) tends to deform (under axial compression) on $(0, \frac{1}{4})$ and $(\frac{3}{4}, 1)$, with the center of the column experiencing only a rigid motion. To make this precise we recall from [37] that this A, when restricted to $(0, \frac{1}{4})$, is actually the optimal design of the clamped-free problem there, with a corresponding buckling load of $\pi^2/3$ (as the volume of this piece is also $1/4$) and first eigenfunction $u_{2,0}$. We normalize $u_{2,0}$ so that

$$\int_0^{1/4} |u_{2,0}'|^2\, dx = 1,$$

and use it to construct a sequence $\{\varphi_n\} \subset H_0^2$ on which

$$\frac{\int_0^1 A^2|\varphi_n''|^2\, dx}{\int_0^1 |\varphi_n'|^2\, dx} \to \frac{\pi^2}{3} \quad \text{as } n \to \infty. \tag{A.2}$$

First we build the continuous displacement

$$\tilde{u}(x) = \begin{cases} u_{2,0}(x), & \text{if } 0 \le x \le \frac{1}{4} \\ u_{2,0}(\frac{1}{4}), & \text{if } \frac{1}{4} \le x \le \frac{3}{4} \\ u_{2,0}(1-x), & \text{if } \frac{3}{4} \le x \le 1. \end{cases}$$

This function is not a member of H_0^2, but we shall see that it suffices to introduce the cubic perturbation

$$p_n(x) = a(x - 2nx^2 + n^2x^3), \quad \text{where } a = u_{2,0}'(\tfrac{1}{4}),$$

near its singularities. As $p_n(0) = p_n(1/n) = p_n'(1/n) = 0$, the function

$$\varphi_n(1-x) = \varphi_n(x) = \begin{cases} \tilde{u}(x), & \text{if } 0 \le x \le \frac{1}{4} \\ \tilde{u}(\frac{1}{4}) + p_n(x - \frac{1}{4}), & \text{if } \frac{1}{4} \le x \le \frac{1}{4} + \frac{1}{n} \\ \tilde{u}(\frac{1}{4}), & \text{if } \frac{1}{4} + \frac{1}{n} \le x \le \frac{1}{2} \end{cases}$$

possesses a continuous derivative. It remains to show that $\varphi_n'' \in L^2$. The only possible obstacle is the behavior of $u_{2,0}''$ near $\frac{1}{4}$. Returning to Tadjbakhsh and Keller [37,§3, (25)] we find

$$u_{2,0}''(x) = \frac{\sqrt{3}}{2 \sin \theta(x)}.$$

And, as (A.1) implies that $\theta(x) = O(|x - 1/4|^{1/3})$ as $x \to \frac{1}{4}$, we find $u_{2,0}''(x) = 0(|x - \frac{1}{4}|^{-1/3})$. This singularity is clearly square integrable, hence $\varphi_n \in H_0^2$, and we can consider the Rayleigh quotient

$$\text{(A.3)} \qquad \frac{\int_0^1 A^2 |\varphi_n''|^2\,dx}{\int_0^1 |\varphi_n'|^2\,dx} = \frac{2\int_0^{1/4} A^2 |u_{2,0}''|^2\,dx + 2\int_0^{1/n} A^2(x - 1/4)|p_n''|^2\,dx}{2\int_0^{1/4} |u_{2,0}'|^2\,dx + 2\int_0^{1/n} |p_n'|^2\,dx}.$$

By construction,

$$\int_0^{1/4} A^2 |u_{2,0}''|^2\,dx = \pi^2/3 \quad \text{and} \quad \int_0^{1/4} |u_{2,0}'|^2\,dx = 1.$$

It remains to show that the remaining terms in (A.3) tend to zero with increasing n. Beginning with the numerator, from $A(x) = O(|x - \frac{1}{4}|^{2/3})$ comes

$$\int_0^{1/n} A^2(x - 1/4)|p_n''(x)|^2\,dx = \int_0^{1/n} O(x^{4/3})|2an(3nx - 2)|^2\,dx = O(1/n^{1/3}),$$

while in the denominator

$$\int_0^{1/n} |p_n'(x)|^2\,dx = \int_0^{1/n} |a(1 - 4nx + 3n^2x^2)|^2\,dx = O(1/n).$$

As we have constructed a sequence of admissible displacements whose Rayleigh quotients tend to $\pi^2/3$, we conclude that the clamped-clamped column built according to (A.1) buckles at a load not exceeding $\pi^2/3$.

Now recall Tadjbakhsh and Keller's solution to the clamped-hinged problem,

$$\text{(A.4)} \qquad \hat{\sigma}_{2,1}(x) = \frac{4 \sin^2 \theta(x)}{3 \sin^2 \theta(0)}, \qquad \theta(0) \le \theta \le \pi,$$

$$\theta(x) - \tfrac{1}{2}\sin 2\theta(x) + \tfrac{1}{2}\sin 2\theta(0) - \theta(0) = x(\pi + \tfrac{1}{2}\sin 2\theta(0) - \theta(0)), \qquad 0 \le x \le 1,$$

$$\tfrac{1}{2}\sin 2\theta(0) - \theta(0) = -\tfrac{2}{3}\sin^3\theta(0)\cos^{-1}\theta(0) - \pi,$$

and the fact that it vanishes at $x_0 = y/(\pi+y)$, where $y = \frac{1}{2}\sin 2\theta(0) - \theta(0)$. Analogous to the above, this design is optimal for the clamped-free column on $(0, x_0)$. With the volume of this piece being x_0 as well, we find that it buckles at $\pi^2/3$. Denoting by $u_{2,0}$ the corresponding clamped-free eigenfunction on $(0, x_0)$ whose derivative has L^2 norm 1, we define,

$$\tilde{u}(x) = \begin{cases} u_{2,0}(x), & \text{if } 0 \le x \le x_0 \\ \frac{u_{2,0}(x_0)}{1-x_0}(1-x), & \text{if } x_0 \le x \le 1. \end{cases}$$

Here we introduce the perturbation

$$p_n(x) = bx - 2n(b+c)x^2 + n^2(b+c)x^3, \quad \text{where} \quad b = u_{2,0}(x_0),\ c = u_{2,0}'(x_0),$$

and the corresponding regularization

$$\varphi_n(x) = \begin{cases} \tilde{u}(x), & \text{if } 0 \le x \le x_0 \\ \tilde{u}(x_0) + p_n(x - x_0), & \text{if } x_0 \le x \le x_0 + 1/n \\ \tilde{u}(x), & \text{if } x_0 + 1/n \le x \le 1. \end{cases}$$

By construction, $\varphi_n \in C^1$, while, as in the clamped-clamped case, $u_{2,0}''$ behaves like $|x - x_0|^{-1/3}$ near x_0, and so $\varphi_n \in H_0^2$. Moreover, $\hat{\sigma}_{2,1}(x) = O(|x - x_0|^{2/3})$ implies, as above, that

$$\int_0^{1/n} \hat{\sigma}_{2,1}^2(x - x_0)|p_n''(x)|^2\,dx = O(1/n^{1/3}), \quad \text{and} \quad \int_0^{1/n} |p_n'|^2\,dx = O(1/n).$$

Consequently,

$$\frac{\int_0^1 \hat{\sigma}_{2,1}^2 |\varphi_n''|^2\,dx}{\int_0^1 |\varphi_n'|^2\,dx} \to \frac{\pi^2/3}{1 + u_{2,0}^2(x_0)/(1 - x_0)} \quad \text{as } n \to \infty,$$

i.e., the clamped-hinged column built according to (A.4) is even weaker than the clamped-clamped column of (A.1).

A.2. Acknowledgments. This work began during the first author's visit to the Courant Institute. The efforts of Bob Kohn and Cathleen Morawetz in this regard are greatly appreciated. We thank Andre Cherkaev for introducing us to the work of Bratus and Seiranian.

REFERENCES

[1] F. V. Atkinson, *Discrete and Continuous Boundary Problems*, Academic Press, New York, 1964.

[2] G. Auchmuty, *Dual variational principles for eigenvalue problems*, in Nonlinear Functional Analysis and Its Applications, F. Browder, ed., American Mathematical Society, Providence, 1986, pp. 55–71.

[3] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Reidel, Boston, MA, 1986.

[4] D. Barnes, *The shape of the strongest column is arbitrarily close to the shape of the weakest column*, Quart. Appl. Math., XLVI (1988), pp. 605–609.

[5] K. Bathe and E. Wilson, *Numerical Methods in Finite Element Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1976.

[6] A. BRATUS, *Multiple eigenvalues in problems of optimizing the spectral properties of systems with a finite number of degrees of freedom*, Zh. Vychisl. Mat. i Mat. Fiz., 26 (1986), pp. 645–654; USSR Comput. Math. and Math. Phys., 26 (1986), pp. 1–7.

[7] A. BRATUS AND A. SEIRANIAN, *Bimodal solutions in eigenvalue optimization problems*, Prikl. Mat. Mekh., 47 (1983), pp. 546–554; Appl. Math. Mech., 47 (1983), pp. 451–457.

[8] J. CEA AND K. MALANOWSKI, *An example of a max-min problem in partial differential equations*, SIAM J. Control, 8 (1970), pp. 305–316.

[9] K. K. CHOI AND E. J. HAUG, *Optimization of structures with repeated eigenvalues*, in Optimal Design of Distributed Parameter Structures, E. J. Haug and J. Cea, eds., Sijthoff–Noordhoff, Leyden, 1981, pp. 219–277.

[10] F. CLARKE, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics 5, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[11] S. J. COX AND J. R. MCLAUGHLIN, *Extremal eigenvalue problems for composite membranes*, II, Appl. Math. Optim., 22 (1990), pp. 169–187.

[12] B. DACOROGNA, *Weak Continuity and Weak Lower Semicontinuity of Non-Linear Functionals*, Lecture Notes in Math. 922, Springer–Verlag, New York, 1982.

[13] J. J. DONGARRA, JAMES R. BUNCH, CLEVE B. MOLER, AND G. W. STEWART, *LINPACK Users Guide*, SIAM, Philadelphia, 1978.

[14] R. FLETCHER, *Semidefinite constraints in optimization*, SIAM J. Control Optim., 23 (1985), pp. 493–513.

[15] ———, *Practical Methods of Optimization*, Second ed., John Wiley, New York, 1987.

[16] S. FRIEDLAND, J. NOCEDAL, AND M. L. OVERTON, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal., 24 (1987), pp. 634–667.

[17] B. S. GARBOW, B. T. SMITH, J. M. BOYLE, J. J. DONGARRA, Y. IKEBE, V. C. KLEMN, AND C. B. MOLER, *Matrix Eigensystem Routines:* EISPACK *Guide Extension*, Lecture Notes in Computer Science 51, Springer–Verlag, New York, 1977.

[18] E. J. HAUG AND B. ROUSSELET, *Design sensitivity analysis in structural mechanics.* II. *Eigenvalue variations*, J. Struct. Mech., 8 (1980), pp. 161–186.

[19] E. KAMKE, *Neue Herleitung der Oszillationssätze für die linearen selbstadjungierten Randwetaufgaben zweiter Ordnung*, Math. Z., 44 (1938), pp. 635–658.

[20] J. KELLER, *The shape of the strongest column*, Arch. Rational Mech. Anal., 5 (1960), pp. 275–285.

[21] W. LEIGHTON AND Z. NEHARI, *On the oscillation of solutions of self–adjoint linear differential equations of the fourth order*, Trans. Amer. Math. Soc., 89 (1958), pp. 325–378.

[22] E. MASUR, *Optimal structural design under multiple eigenvalue constraints*, Internat. J. Solids and Structures, 20 (1984), pp. 211–231.

[23] C. MOLER, J. LITTLE, AND S. BANGERT, *Pro–Matlab Users Guide*, MathWorks, Sherborn, MA, 1987.

[24] M. MYERS AND W. SPILLERS, *A note on the strongest fixed-fixed column*, Quart. Appl. Math., XLIV (1986), pp. 583–588.

[25] N. OLHOFF AND S. RASMUSSEN, *On single and bimodal optimum buckling loads of clamped columns*, Internat. J. Solids and Structures, 13 (1977), pp. 605–614.

[26] M. L. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.

[27] ———, *Large-scale optimization of eigenvalues*, SIAM J. Optim., 2 (1992), pp. 88–120.

[28] E. POLAK, *On the mathematical foundation of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–89.

[29] E. POLAK AND Y. WARDI, *A nondifferentiable optimization algorithm for structural problems with eigenvalue inequality constraints*, J. Struct. Mech., 11 (1983), pp. 561–577.

[30] G. PÓLYA, *More isoperimetric inequalities conjectured and proved*, Comment. Math. Helv., 29 (1955), pp. 112–119.

[31] G. PÓLYA AND G. SZEGÖ, *Isoperimetric Inequalities in Mathematical Physics*, Annals of Math. Stud., No. 27, Princeton, 1951.

[32] A. RAMM, *Queries*, Notices Amer. Math. Soc., 29 (1982), pp. 326–329.

[33] R. T. ROCKAFELLAR, *The Theory of Subgradients and its Applications to Problems of Optimization: Convex and Nonconvex Functions*, Series on Research and Education in Mathematics 1, Heldermann Verlag, Berlin, 1981.

[34] A. SEIRANIAN, *On a problem of Lagrange*, Inzh. Zh., Mekhanika Tverdogo Tela, 19 (1984), pp. 101–111; Mech. Solids, 19 (1984), pp. 100–111.

[35] P. Senatorov, *The stability of the eigenvalues and eigenfunctions of a Sturm–Liouville problem*, Differentsial'nye Uravneniia, 7 (1971), pp. 1667–1671; Differential Equations, 7 (1971), pp. 1266–1269.

[36] G. Strang and G.J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[37] I. Tadjbakhsh and J. Keller, *Strongest columns and isoperimetric inequalities for eigenvalues*, J. Appl. Mech., 29 (1962), pp. 159–164.

[38] I. Todhunter and K. Pearson, *A History of the Theory of Elasticity and of the Strength of Materials*, I, University Press, Cambridge, 1886.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 326-333, March 1992

L^∞ ESTIMATE FOR SOME NONLINEAR ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS AND APPLICATION TO AN EXISTENCE RESULT*

L. BOCCARDO†, F. MURAT‡, AND J.-P. PUEL§

Abstract. Consider the nonlinear elliptic equation (E): $\mathcal{A}(u)+H(x,u,Du)=f(x)-\operatorname{div} g(x)$ where $\mathcal{A}(u)=-\operatorname{div}(a(x,u,Du))+a_0(x,u,Du)$ is a Leray-Lions operator defined on $W_0^{1,p}(\Omega)$ with $a_0(x,s,\xi)s\geqq \alpha_0|s|^p$, $\alpha_0>0$, and where H is a first-order term satisfying $|H(x,s,\xi)|\leqq C_0+C_1|\xi|^p$. The main goal of this paper is to prove an L^∞ estimate for the *bounded* solutions of (E) when f belongs to $L^q(\Omega)$ and g belongs to $(L^r(\Omega))^N$ with $r=p'q$ and $\max(1,N/p)<q\leqq+\infty$. In view of the method and results developed in the author's previous work, this implies the existence of a solution for equation (E).

Résumé. Considérons l'équation elliptique non linéaire (E): $\mathcal{A}(u)+H(x,u,Du)=f(x)-\operatorname{div} g(x)$ où $\mathcal{A}(u)=-\operatorname{div}(a(x,u,Du))+a_0(x,u,Du)$ est un opérateur de Leray-Lions défini sur $W_0^{1,p}(\Omega)$ avec $a_0(a(x,u,\xi)s\geqq\alpha_0|s|^p$, $\alpha_0>0$, et où H est un terme du premier ordre qui vérifie $|H(x,s,\xi)|\leqq C_0+C_1|\xi|^p$. Le résultat principal de ce travail est une estimation L^∞ pour les solutions *bornées* de (E) lorsque f appartient à $L^q(\Omega)$ et g appartient à $(L^r(\Omega))^N$ avec $r=p'q$ et $\max(1,N/p)<q\leqq+\infty$. A la lumière de la méthode et des résultats présentés dans nos précédents travaux, ceci implique l'existence d'une solution pour l'équation (E).

Key words. L^∞ estimate, nonlinear elliptic pde, natural growth with respect to Du

AMS(MOS) subject classifications. 35E, 35G

1. Introduction. Consider on a bounded open set $\Omega\subset R^N$ the nonlinear elliptic equation

$$\mathcal{A}(u)+H(x,u,Du)=f(x)-\operatorname{div} g(x) \quad \text{in } \Omega; \qquad u=0 \quad \text{on } \partial\Omega, \tag{1.1}$$

where

$$\mathcal{A}(u)=-\operatorname{div}(a(x,u,Du))+a_0(x,u,Du)$$

is a Leray-Lions operator from $W_0^{1,p}(\Omega)$ $(1<p<+\infty)$ into its dual $W^{-1,p'}(\Omega)$, $(1/p+1/p'=1)$, such that

$$\exists\alpha>0,\quad \text{for a.e. } x\in\Omega,\quad \forall s\in R,\quad \forall\xi\in R^N,\quad a(x,s,\xi)\xi\geqq\alpha|\xi|^p, \tag{1.2}$$

$$\exists\alpha_0>0,\quad \text{for a.e. } x\in\Omega,\quad \forall s\in R,\quad \forall\xi\in R^N,\quad a_0(x,s,\xi)s\geqq\alpha_0|s|^p, \tag{1.3}$$

and where H is a nonlinear first-order term defined through a Carathéodory function satisfying the "natural" growth condition with respect to ξ, i.e.,

$$\begin{gathered}\exists C_0>0,\quad \exists C_1>0,\quad \text{for a.e. } x\in\Omega,\quad \forall s\in R,\quad \forall\xi\in R^N,\\ |H(x,s,\xi)|\leqq C_0+C_1|\xi|^p.\end{gathered} \tag{1.4}$$

In [BMP1], [BMP2], [BMP3], and [BMP4] we have shown the existence of solutions for nonlinear elliptic equations analogous to (1.1) (in some cases without assuming

* Received by the editors March 2, 1990; accepted for publication (in revised form) October 30, 1990.

† Dipartimento di Matematica, Università di Roma I, Piazzale A. Moro, 00185 Roma, Italy. The research of this author was partially supported by INRIA, GNAFA-CNR and MPI (40%+60%).

‡ Laboratoire d'Analyse Numérique, Tour 55-65, 5 ème étage, Université Paris VI, 75252 Paris Cedex 05, France.

§ Département de Mathématiques et d'Informatique, Université d'Orléans, et Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau Cedex, France.

(1.3) but assuming existence of a sub- and a supersolution). We have also developed a method which essentially reduces the existence proof to the proof of an a priori L^∞ estimate for the solutions of a family of approximate equations.

The main goal of this paper is to prove an L^∞ estimate for the *bounded* solutions of (1.1): we assume that $f \in L^q(\Omega)$ and $g \in (L^r(\Omega))^N$ with $\max(1, N/p) < q \leqq \pm\infty$ and $r = p'q$ and we prove that any $u \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ that solves (1.1) satisfies

$$\|u\|_{L^\infty(\Omega)} \leqq \gamma, \tag{1.5}$$

where γ depends only on the data, i.e., Ω, N, p, q, α, α_0, C_0, C_1, $\|f\|_{L^q(\Omega)}$, $\|g\|_{(L^r(\Omega))^N}$. The main thrust of the result is thus to turn the *qualitative* information $u \in L^\infty(\Omega)$ into the *quantitative* estimate (1.5).

The a priori assumption that u belongs to $L^\infty(\Omega)$ is crucial to obtain (1.5) since there are examples satisfying all the above hypotheses where (1.1) possesses not only solutions in $W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ but also solutions in $W_0^{1,p}(\Omega)$ which do not belong to $L^\infty(\Omega)$ (see Remark 2.1 below).

The L^∞ estimate (1.5) generalizes to the case of nonlinear elliptic equations the results of De Giorgi [DG], Moser [M], and Stampacchia [S1], [S2] for the linear case. These are known to be optimal, in the sense that (for $p = 2$) the hypotheses $f \in L^q(\Omega)$, $q > N/2$ and $g \in (L^r(\Omega))^N$, $r = 2q$ are optimal. This indicates that our result for the nonlinear case is also optimal.

Our proof is an adaptation of the nonlinear context of (1.1) of Stampacchia's method, in the spirit of our previous work. Indeed the first step of Stampacchia's proof (see [S1], [S2]) consists of using in the linear equation the test function $G_k(u)$, where $G_k(s)$ is the function defined by $G_k(s) = 0$ if $|s| \leqq k$ and $G_k'(s) = 1$ if $|s| > k$ (see (2.9)). We adapt this first step to the nonlinear equation (1.1) by using here the nonlinear (with respect to $G_k(u)$) test function $\varphi(G_k(u))$, where $\varphi(s)$ is the odd function defined by $\varphi(s) = e^{\lambda s} - 1$ if $s \geqq 0$ (see (2.8)). Nonlinear test functions $\varphi(u)$, where $\varphi(s)$ has an exponential behaviour, are known to be well adapted to the study of (1.1) since they allow the absorption of the nonlinearity H by the coerciveness of the operator $\mathscr{A}$ (see, e.g., [BMP1], [BMP2], [BMP3], [BMP4]).

We mention that L^∞ bounds for similar problems have been recently proved in [MS] and [ALT] by different methods based on rearrangement techniques; see also the references added in proof.

Once the L^∞ estimate is obtained, the proof of existence of a solution for (1.1) that belongs to $W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ follows the method presented, for example, in [BMP4]. This will be done in § 3, where we also collect some comments about regularity of the solutions.

2. L^∞ estimate. We consider the nonlinear elliptic equation (1.1), and as we only look for bounded solutions, we rewrite it in the more precise way

$$\mathscr{A}(u) + H(x, u, Du) = f(x) - \operatorname{div} g(x) \text{ in } \mathscr{D}'(\Omega); \qquad u \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega). \tag{2.1}$$

On the operator

$$\mathscr{A}(u) = -\operatorname{div}(a(x, u, Du)) + a_0(x, u, Du)$$

we assume that a and a_0 are Carathéodory functions defined on $\Omega \times R \times R^N$ with values in R^N and R, respectively, which satisfy (1.2), (1.3) and

$$\begin{aligned} &\exists \psi \in L^{p'}(\Omega), \quad \exists \beta > 0, \quad \text{for a.e. in } x \in \Omega, \quad \forall s \in R, \quad \forall \xi \in R^N, \\ &|a(x, s, \xi)| \leqq |\psi(x)| + \beta[|s|^{p-1} + |\xi|^{p-1}], \\ &|a_0(x, s, \xi)| \leqq |\psi(x)| + \beta[|s|^{p-1} + |\xi|^{p-1}]; \end{aligned} \tag{2.2}$$

$$\text{for a.e. } x\in\Omega,\quad \forall s\in R,\quad \forall\xi,\xi^*\in R^N,\quad \xi\neq\xi^*, \tag{2.3}$$
$$[a(x,s,\xi)-a(x,s,\xi^*)][\xi-\xi^*]>0.$$

These hypotheses ensure that $\mathscr{A}$ is a bounded, continuous, coercive, and pseudomonotone operator of Leray-Lions type from $W_0^{1,p}(\Omega)$ into its dual $W^{-1,p'}(\Omega)$ $(1/p+1/p'=1)$ (see [LL], [L]).

Furthermore, H is assumed to be a Carathéodory function defined on $\Omega\times R\times R^N$ with values in R, satisfying (1.4). Note that no extra requirement such as a Lipschitz continuity or a one-sided condition is imposed on H.

The operator $u\to\mathscr{A}(u)+H(x,u,Du)$ is then well defined, continuous, and bounded from $W_0^{1,p}(\Omega)$ into $W^{-1,p'}(\Omega)+L^1(\Omega)$, but it is neither pseudomonotone nor coercive.

We can now state the main result of this paper.

THEOREM 1. *Let us assume that* (1.2)-(1.4), (2.2) *hold true and that*

$$f\in L^q(\Omega),\ g\in(L^r(\Omega))^N \quad \textit{with } \max\left(1,\frac{N}{p}\right)<q\leqq+\infty,\quad r=p'q. \tag{2.4}$$

Then any solution u of (2.1) *satisfies the estimate*

$$\|u\|_{L^\infty(\Omega)}\leqq\gamma, \tag{2.5}$$

where γ is a constant which depends only on Ω, N, p, q, α, α_0, C_0, C_1, $\|f\|_{L^q(\Omega)}$, $\|g\|_{(L^r(\Omega))^N}$.

Remark 2.1. As mentioned in the Introduction, Theorem 1 provides an "explicit" estimate in $L^\infty(\Omega)$ for the solutions of (2.1) which are already known to belong to $W_0^{1,p}(\Omega)\cap L^\infty(\Omega)$. This does not mean that every solution of (1.1) that belongs to $W_0^{1,p}(\Omega)$ also belongs to $L^\infty(\Omega)$. In fact, this last assertion is false, as is shown in Contre-exemple 3.2, § 3.7 of [BMP1].

Remark 2.2. Estimate (2.5) continues to hold if the right-hand side of (2.1) has the more general form

$$F(x,u,Du)-\operatorname{div}(G(x,u,Du)), \tag{2.6}$$

where F and G are Carathéodory functions from $\Omega\times R\times R^N$ into R and R^N, respectively, satisfying

$$\text{for a.e. } x\in\Omega,\quad \forall s\in R,\quad \forall\xi\in R^N,\quad |F(x,s,\xi)|\leqq|f(x)|,\quad |G(x,s,\xi)|\leqq|g(x)|,$$

where f and g satisfy (2.4). There is essentially no change in the proof of Theorem 1 given below, but the proof of the existence theorem given in § 3 does not work in this context. It would require that G be independent of Du.

Note also that (2.3) is not required to obtain Theorem 1. However, this hypothesis is needed for the existence result (see Theorem 2 below).

Proof of Theorem 1. Set

$$\lambda=\frac{p'C_1}{\alpha}+p', \tag{2.7}$$

and define for $k\in R^+$ the real functions $\varphi\in C^1(R)$ and $G_k\in W^{1,\infty}(R)$ by

$$\varphi(s)=\begin{cases}e^{\lambda s}-1 & \text{if } s\geqq0\\ -e^{-\lambda s}+1 & \text{if } s\leqq0,\end{cases} \tag{2.8}$$

$$G_k(s)=\begin{cases}s-k & \text{if } s\geqq k\\ 0 & \text{if } -k\leqq s\leqq k,\\ s+k & \text{if } s\leqq -k.\end{cases} \tag{2.9}$$

Let $A(k)$ be the set

$$A(k)=\{x\in\Omega, |u(x)|>k\}. \tag{2.10}$$

The functions $G_k(u)$ have been used in [S1] and [S2] as test functions to obtain the L^∞ estimate for linear equations. Adapting Stampacchia's method to the nonlinear case, we will use the test functions

$$v=\varphi(G_k(u)). \tag{2.11}$$

Since u belongs to $W_0^{1,p}(\Omega)\cap L^\infty(\Omega)$, so does v, and it is easy to show that

$$v=\varphi((|u|-k)^+)\chi_{A(k)}\operatorname{sign}(u), \tag{2.12}$$

$$Dv=\varphi'((|u|-k)^+)\chi_{A(k)}Du, \tag{2.13}$$

where $\chi_{A(k)}$ is the characteristic function of the set $A(k)$. The function v is then an admissible test function for (2.1). Using v in the weak formulation of (2.1) as well as (1.2)-(1.4) and Young's inequality, we obtain

$$\begin{aligned}\alpha\int_{A(k)}&|Du|^p\varphi'((|u|-k)^+)\,dx+\alpha_0\int_{A(k)}|u|^{p-1}\varphi((|u|-k)^+)\,dx\\ &\leqq\int_{A(k)}(C_0+C_1|Du|^p)\varphi((|u|-k)^+)\,dx\\ &+\int_{A(k)}|f|\varphi((|u|-k)^+)\,dx\\ &+\int_{A(k)}\varphi'((|u|-k)^+)\left(\frac{\alpha}{p}|Du|^p+\frac{1}{p'\alpha^{1/(p-1)}}|g|^{p'}\right)dx.\end{aligned} \tag{2.14}$$

From the choice (2.7) of λ we have for $s\geqq 0$,

$$\frac{\alpha}{p'}\varphi'(s)-C_1\varphi(s)\geqq\alpha\,e^{\lambda s}=\frac{\alpha}{\lambda^p}\left(\varphi'\left(\frac{s}{p}\right)\right)^p.$$

This implies

$$\begin{aligned}\frac{\alpha}{\lambda^p}\int_{A(k)}&\left|\varphi'\left(\frac{(|u|-k)^+}{p}\right)Du\right|^p dx+\alpha_0k^{p-1}\int_{A(k)}\varphi((|u|-k)^+)\,dx\\ &\leqq\int_{A(k)}(C_0+|f|)\varphi((|u|-k)^+)\,dx\\ &+\int_{A(k)}\frac{1}{p'\alpha^{1/(p-1)}}|g|^{p'}\varphi'((|u|-k)^+)\,dx.\end{aligned} \tag{2.15}$$

Let us define the function w_k by

$$w_k=\varphi\left(\frac{(|u|-k)^+}{p}\right). \tag{2.16}$$

Then w_k belongs to $W_0^{1,p}(\Omega)$ and

$$Dw_k=\frac{1}{p}\varphi'\left(\frac{(|u|-k)^+}{p}\right)\chi_{A(k)}\operatorname{sign}(u)\,Du.$$

One can easily show that

$$\forall s \geqq 0, \quad e^{\lambda s} - 1 \geqq (e^{\lambda s/p} - 1)^p,$$

$$\begin{cases} \exists C_2 > 0 \text{ (depending only on } \lambda \text{ and } p), \quad \forall s \geqq 1, \\ e^{\lambda s} - 1 \leqq C_2(e^{\lambda s/p} - 1)^p, \qquad \lambda\, e^{\lambda s} \leqq C_2 \lambda (e^{\lambda s/p} - 1)^p. \end{cases}$$

This implies

$$\varphi((|u| - k)^+) \geqq |w_k|^p \quad \text{a.e. on } \Omega,$$

$$\varphi((|u| - k)^+) \leqq C_2 |w_k|^p \quad \text{a.e. on } A(k+1),$$

$$\varphi'((|u| - k)^+) \leqq C_2 \lambda |w_k|^p \quad \text{a.e. on } A(k+1).$$

Combining these remarks with (2.15) yields

$$\begin{aligned} &\frac{\alpha p^p}{\lambda^p} \int_\Omega |Dw_k|^p \, dx + \alpha_0 k^{p-1} \int_\Omega |w_k|^p \, dx \\ &\leqq \int_{A(k+1)} C_2 \left((C_0 + |f|) + \frac{\lambda}{p' \alpha^{1/(p-1)}} |g|^{p'} \right) |w_k|^p \, dx \\ &\quad + \int_{A(k) - A(k+1)} \left((C_0 + |f|) \varphi((|u| - k)^+) \right. \\ &\qquad \left. + \frac{1}{p' \alpha^{1/(p-1)}} |g|^{p'} \varphi'((|u| - k)^+) \right) dx. \end{aligned} \tag{2.17}$$

Define

$$h = (C_0 + |f|) + \frac{\lambda}{p' \alpha^{1/(p-1)}} |g|^{p'}.$$

Since h belongs to $L^q(\Omega)$ and since

$$\varphi((|u| - k)^+) \leqq e^\lambda - 1, \qquad \varphi'((|u| - k)^+) \leqq \lambda\, e^\lambda \quad \text{on } A(k) - A(k+1),$$

we obtain

$$\begin{aligned} &\frac{\alpha p^p}{\lambda^p} \int_\Omega |Dw_k|^p \, dx + \alpha_0 k^{p-1} \int_\Omega |w_k|^p \, dx \\ &\leqq \int_{A(k+1)} C_2 h |w_k|^p \, dx + e^\lambda \int_{A(k) - A(k+1)} h \, dx. \end{aligned} \tag{2.18}$$

Sobolev's inequality asserts that for $p*$ defined by

$$\begin{cases} p^* = \dfrac{Np}{N-p} \quad \text{if } 1 < p < N, \\ p^* \text{ is any fixed real number such that } pq' < p^* < +\infty \quad \text{if } N \leqq p < +\infty, \end{cases} \tag{2.19}$$

there exists C^* which depends only on Ω, N, p, and $p*$ such that

$$\forall w \in W_0^{1,p}(\Omega), \quad C^* \|w\|_{L^{p^*}(\Omega)} \leqq \|Dw\|_{(L^p(\Omega))^N}. \tag{2.20}$$

Therefore, there exist strictly positive constants C_3 and C_4, depending only on Ω, N, p, q, $p*$, C_1, and α such that

$$\tag{2.21} \begin{aligned} C_3\left(\int_\Omega |w_k|^{p*}\,dx\right)^{p/p*} + C_4\alpha_0 k^{p-1}\int_\Omega |w_k|^p\,dx \\ \leqq \int_\Omega h|w_k|^p\,dx + \int_{A(k)} h\,dx. \end{aligned}$$

Because of (2.4) and (2.19) we have $p<pq'<p*$. From Hölder's inequality and an interpolation inequality, we get

$$\int_\Omega h|w_k|^p\,dx \leqq \|h\|_{L^q(\Omega)}\|w_k\|^p_{L^{pq'}(\Omega)} \leqq \|h\|_{L^q(\Omega)}\|w_k\|^{\theta p}_{L^p(\Omega)}\|w_k\|^{(1-\theta)p}_{L^{p*}(\Omega)},$$

where $\theta\in\,]0,1[$ is defined by

$$\frac{1}{pq'}=\frac{\theta}{p}+\frac{(1-\theta)}{p*}.$$

Using Young's inequality, we obtain for any $\eta, 0<\eta<+\infty$,

$$\int_\Omega h|w_k|^p\,dx \leqq (1-\theta)\eta^{1/(1-\theta)}\|w_k\|^p_{L^{p*}(\Omega)}+\theta\eta^{-1/\theta}\|h\|^{1/\theta}_{L^q(\Omega)}\|w_k\|^p_{L^p(\Omega)}.$$

Now choose η such that

$$(1-\theta)\eta^{1/(1-\theta)}=\frac{C_3}{2},$$

and then k_0 such that

$$C_4\alpha_0 k_0^{p-1}=\theta\eta^{-1/\theta}\|h\|^{1/\theta}_{L^q(\Omega)},$$

(this is the only place where the hypothesis $\alpha_0>0$ is used). We obtain from (2.21)

$$\forall k \geqq k_0,\quad \frac{C_3}{2}\left(\int_\Omega |w_k|^{p*}\,dx\right)^{p/p*} \leqq \int_{A(k)} h\,dx \leqq \|h\|_{L^q(\Omega)}|A(k)|^{1/q'},$$

where $|A(k)|$ denotes the Lebesgue measure of the set $A(k)$. We rewrite this inequality as

$$\tag{2.22} \forall k\geqq k_0,\quad \int_\Omega |w_k|^{p*}\,dx \leqq \left(\frac{2}{C_3}\|h\|_{L^q(\Omega)}\right)^{p*/p}|A(k)|^{p*/pq'}=C_5|A(k)|^{p*/pq'}.$$

Take $l>k\geqq k_0$ and observe that

$$|A(l)|\left(\lambda\left(\frac{l-k}{p}\right)\right)^{p*} \leqq |A(l)|\left|\varphi\left(\frac{l-k}{p}\right)\right|^{p*} \leqq \int_{A(l)}|w_k|^{p*}\,dx \leqq \int_\Omega |w_k|^{p*}\,dx.$$

Then

$$\tag{2.23} \forall l, k \text{ with } l>k\geqq k_0,\quad (l-k)^{p*}|A(l)| \leqq C_6|A(k)|^{p*/pq'},$$

where C_6 depends only on Ω, N, α, $\alpha_0, p, q, p*, C_0, C_1, \|f\|_{L^q(\Omega)}, \|g\|_{(L^r(\Omega))^N}$. Since $p*>0$ and $p*/pq'>1$, we can use a result of Stampacchia (cf. [S1, Lemma 4.1] or

[S2, Lemma 4.1]) which implies that there exists γ that depends only on C_6, k_0, p^* and p^*/pq' such that

$$A(k)=0, \quad \forall k \geqq \gamma.$$

This finishes the proof of Theorem 1.

3. Application to an existence result. In this section we will apply the method and results presented, for example, in [BMP4] and the L^∞ estimate of the previous section to a family of approximate equations for (2.1) in order to prove the existence of at least one solution of (2.1).

THEOREM 2. *Let us assume that* (1.2)-(1.4), *and* (2.2)-(2.4) *hold true. Then there exists at least one solution u of* (2.1).

Proof of Theorem 2. Let us define for $\varepsilon>0$ the approximation

$$H^\varepsilon(x, s, \xi)=\frac{H(x, s, \xi)}{1+\varepsilon|H(x, s, \xi)|}. \tag{3.1}$$

Note that H^ε satisfies (1.4). We consider the approximate problem

$$\begin{cases} -\operatorname{div}(a(x, u^\varepsilon, Du^\varepsilon))+a_0(x, u^\varepsilon, Du^\varepsilon) \\ \quad +H^\varepsilon(x, u^\varepsilon, Du^\varepsilon)=f(x)-\operatorname{div} g(x) \quad \text{in } \mathscr{D}'(\Omega), \\ u^\varepsilon \in W_0^{1,p}(\Omega). \end{cases} \tag{3.2}$$

For fixed $\varepsilon>0$, H^ε is uniformly bounded by $1/\varepsilon$. Therefore there exists at least one solution u^ε to (3.2) (see [LL], [L]). Using Stampacchia's method (see, e.g., [BG]) it can be proved that any solution u^ε of (3.2) belongs to $L^\infty(\Omega)$ (for fixed $\varepsilon>0$). We can now apply Theorem 1 which yields

$$\|u^\varepsilon\|_{L^\infty(\Omega)} \leqq \gamma,$$

where γ does not depend on ε.

Once the L^∞ estimate is obtained, we can apply Theorem 2.1 of [BMP4] (with a slight modification due to the term div g), or the remark at the end of § 3 of [BMP3], and obtain the relative compactness of the family (u^ε) in the strong topology of $W_0^{1,p}(\Omega)$. Then, by extracting a subsequence $(u^{\varepsilon'})$ which strongly converges in $W_0^{1,p}(\Omega)$ it is easy to pass to the limit when ε' tends to zero and to obtain Theorem 2.

Remark 3.1. (Regularity). We make the same hypotheses as in Theorem 1. Then, according to the results of [LU, Chap. 4, Thm. 1.1, p. 251], any solution of (2.1) actually belongs to $C^{0,\delta}(\Omega)$ for some δ, $0<\delta<1$, whenever $\partial\Omega$ is sufficiently smooth (otherwise one only obtains $C^{0,\delta}_{\text{loc}}(\Omega)$).

Moreover, if in (2.2) we assume that $\psi \in L^s(\Omega)$, with $s>p'$, a slight modification of Proposition 3.8 of [BMP2] (due to the term div g) (see also [GM], p. 156] when $p=2$) proves that u belongs to $W_0^{1,p+\delta}(\Omega)$ for some $\delta>0$ if $\partial\Omega$ is sufficiently smooth (otherwise one has only the local result).

Acknowledgments. A previous version of this work was presented by the first author at the Conference on Nonlinear Evolution Equations and Their Applications (Visegrad, Hungary, 1988). The authors are indebted to the organizers of this meeting for this invitation.

REFERENCES

[ALT] A. Alvino, P.-L. Lions, and G. Trombetti, *Comparison results for elliptic and parabolic equations via Schwarz symmetrisation*, Ann. Inst. Henri Poincaré, Analyse non linéaire, 7 (1990), pp. 37-65.

[BG] L. Boccardo and D. Giachetti, *Alcune osservazioni sulla regolarità delle soluzioni di alcuni problemi fortemente nonlineari e applicazioni*, Ricerche Mat., 34 (1985), pp. 309-323.

[BMP1] L. Boccardo, F. Murat, and J.-P. Puel, *Existence de solutions faibles pour des équations elliptiques quasi linéaires à croissance quadratique*, in Nonlinear partial differential equations and their applications, Collège de France Seminar, Vol. IV, (H. Brezis and J. L. Lions, eds.), Res. Notes in Math. 84, Pitman, London, 1983, pp. 19-73.

[BMP2] ———, *Résultats d'existence pour certains problèmes quasi linéaires*, Ann. Sc. Norm. Sup. Pisa, 11 (1984), pp. 213-235.

[BMP3] ———, *Existence of bounded solutions for nonlinear elliptic unilateral problems*, Ann. Mat. Pura Appl., 152 (1988), pp. 183-196.

[BMP4] ———, *Quelques propriétés des opérateurs elliptiques quasi linéaires*, C. R. Acad. Sci. Paris, Sér. I, 307 (1988), pp. 749-752.

[DG] E. De Giorgi, *Sulla differenziabilità e l'analiticità delle estremali degli integrali multipli regolari*, Mem. Accad. Sci. Torino, Cl. Sci. Fis. Mat. Natur., 3 (1957), pp. 25-43.

[FP] V. Ferone and M. R. Posteraro, *On a class of quasilinear elliptic equations with quadratic growth in the gradient*, Preprint 33 (1991), Dipartimento di Matematica ed Applicazioni dell'Università di Napoli, Naples, Italy.

[GM] M. Giaquinta and G. Modica, *Regularity results for some classes of higher order nonlinear elliptic systems*, J. Reine Angew. Math., 311-312 (1979), pp. 145-169.

[L] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod et Gauthier Villars, Paris, 1969.

[LL] J. Leray and J.-L. Lions, *Quelques résultats de Visik sur les problèmes elliptiques non linéaires par les méthodes de Minty-Browder*, Bull. Soc. Math. France, 93 (1965), pp. 97-107.

[LU] O. A. Ladyzhenskaya and N. N. Uralt'seva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[MPS] C. Maderna, C. Pagani, and S. Salsa, *Quasilinear elliptic equations with quadratic growth in the gradient*, J. Differential Equations, to appear.

[MS] C. Maderna and S. Salsa, *Dirichlet problem for elliptic equations with nonlinear first order term: A comparison result*, Ann. Mat. Pura Appl., 148 (1987), pp. 277-288.

[M] J. Moser, *A new proof of De Giorgi's theorem concerning the regularity problem for elliptic differential equations*, Comm. Pure Appl. Math., 13 (1960), pp. 457-468.

[S1] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier, 15 (1965), pp. 189-258.

[S2] ———, *Equations elliptiques du second ordre à coefficients discontinus*, Séminaires de Mathématiques Supérieures, 16, Les Presses de l'Université de Montréal, Montréal, 1966.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 334-345, March 1992

THE SURFACE EVAPORATION PROBLEM WITH SIGNORINI BOUNDARY CONDITION*

ZHIDA HUANG[†] AND MARIO PRIMICERIO[‡]

Abstract. The one-dimensional filtration of an incompressible liquid in a homogeneous, isotropic, rigid porous medium is considered. The bottom of the layer is impermeable, whereas on the top surface a Signorini-type boundary condition is imposed. Existence and uniqueness of the weak solution are proved under general conditions. Then some qualitative properties of the solution and its asymptotic behaviour are analyzed. In particular, the characterization of the set $D \equiv \{t: u(0, t)=0\}$ is discussed.

Key words. filtration, porous media, unilateral boundary conditions, evaporation, free boundary problem

AMS(MOS) subject classifications. 35K65, 35K85, 76S05, 35R35, 35B40

1. Introduction. In this paper we consider the flow of an incompressible liquid, say water, through a homogeneous isotropic and rigid porous medium. If θ represents water content, i.e., the mass of water per unit volume of the medium, and $\underset{\sim}{q}$ the discharge, the mass balance is expressed by

$$\theta_t + \operatorname{div} \rho \underset{\sim}{q} = 0, \tag{1.1}$$

where ρ is the density of water.

Equation (1.1) is completed by prescribing the dependence of θ upon the capillary pressure ψ (state equation) and a relationship between $\underset{\sim}{q}$ and the pressure gradient (law of motion).

It is generally agreed that when ψ exceeds a saturation value (say $\psi = 0$), θ is constant and equal to $\theta_* = \varepsilon\rho$,where ε is the porosity. In such situation Darcy's law holds

$$\underset{\sim}{q} = -K \text{ grad } [\psi/\rho g - x], \tag{1.2}$$

where g is the gravity acceleration, x is a vertical coordinate pointing downwards, and K is the hydraulic conductivity of the medium.

Law (1.2) is usually assumed to hold also in the unsaturated region ($\theta < \theta_*$ and $\psi < 0$) where the conductivity K is supposed to be experimentally determined as a function of θ or of ψ (see, e.g., [11], [12]).

When situations with low water content are to be dealt with, the situation becomes much more complicated; the notion of a "residual" water content θ_0 is introduced such that it cannot be further reduced by means of applied pressure gradients. It is reasonable to assume that, as θ approaches θ_0, the hydraulic conductivity falls to zero: actually, for $\theta = \theta_0$, the liquid phase is no longer connected and it is clearly understood that no pressure can be transmitted from one "pendular ring" of water to the others [12], [13].

Accordingly, we will assume shapes of $\theta - \psi$ and $K - \psi$ curves shown in Fig. 1.1 (see also [1], [14], [15]).

We will consider a one-dimensional case as a first approximation to practical problems in which a much more complicated situation can appear: essentially three-dimensional filtration (see [7], [17]), anisotropy, hysteresis phenomena, etc. This approximation is commonly adopted in most of the papers devoted to problems of

* Received by the editors November 13, 1989; accepted for publication (in revised form) April 29, 1991.

[†] Department of Mathematics, South China Normal University, Guangzhou, China.
[‡] Universita degli Studi Istituto Matematico, Firenze, Italy.

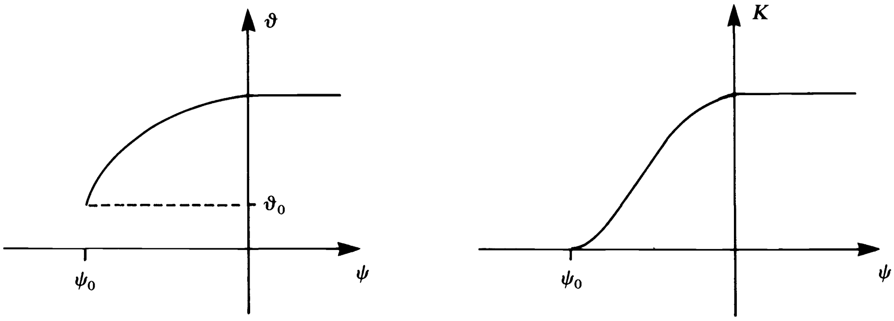


FIG. 1.1

this class and, from a mathematical point of view, makes it possible to use the Kirchhoff transformation

$$u(x, t) = \int_{\psi_0}^{\psi(x, t)} K(s)\, ds.$$

Denote

$$u_s = \int_{\psi_0}^{0} K(s)\, ds.$$

which means that the medium is saturated when $u \geqq u_s$. Then, when $u > 0$, $u(x, t)$ satisfies the equation

$$\theta(u)_t = u_{xx} - k(u)_x. \tag{1.3}$$

A sketch of $\theta(u)$ and $k(u)$ is given in Fig. 1.2.
In general, it is assumed in the literature (see [9], [10]) that $\theta'(u) \to +\infty$ as $u \to 0^+$ and $\theta'(u) \to 0$ as $u \to u_s$. Hence, (1.3) is degenerate when $u = 0$ and $u \geqq u_s$.

We will study (1.3) in a slab $0 < x < 1$, with initial condition

$$u(x, 0) = u_0(x), \qquad 0 < x < 1,$$

where u_0 has values in $(0, 1)$. We assume that the bottom of the slab is impervious, i.e.,

$$u_x(1, t) - k(u(1, t)) = 0,$$

and that evaporation takes place at the top surface $x = 0$. In previous papers (e.g., [5], [10], [16]) evaporation is modeled by prescribing a constant boundary flux; in particular, in [10] it is supposed that the medium can become dry not only on the surface

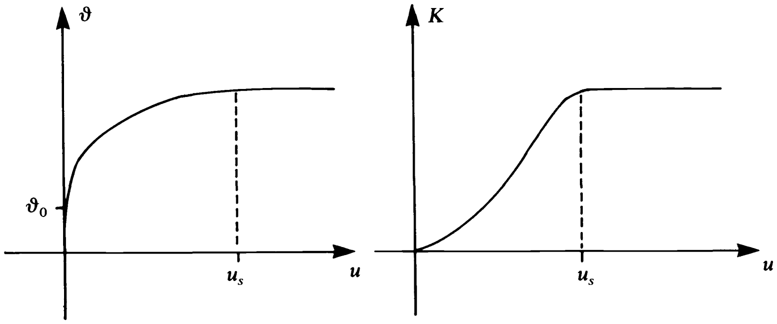


FIG. 1.2

but also below it and the rate of evaporation is a given function of the thickness of the dry part.

On the other hand, it seems reasonable and consistent with the model to assume that the flux becomes zero when the capillary piezometric head reaches the critical value ψ_0. In [8] a filtration problem is considered in which the boundary condition is $u_x(0,t)-k(u(0,t))=f(u(0,t))$ where $f(0)=0$ and f is a smooth function. Here we want to investigate the delicate case in which a discontinuity appears for $u=0$. Hence we will assume that the rate of evaporation is a given constant q if $u(0,t)>0$, whereas the flux is between 0 and q when $u(0,t)=0$. From a mathematical point of view, this is a unilateral or "Signorini type" boundary condition (see also [7]), which has the form

$$\begin{aligned} &u_x(0,t)-k(u(0,t))=q && \text{if } u(0,t)>0,\\ &0\leqq u_x(0,t)\leqq q && \text{if } u(0,t)=0,\end{aligned}$$

i.e.,

$$u_x(0,t)-k(u(0,t))\in qH(u(0,t)),$$

where

$$H(z)=\begin{cases}0, & z<0,\\ [0,1], & z=0,\\ 1, & z>0.\end{cases}$$

Summing up, the evaporation problem we will study is the following:

$$(\mathbb{P})\qquad \begin{aligned} &\theta(u)_t=u_{xx}-k(u)_x, \qquad (x,t)\in H_T=(0,1)x(0,T],\\ &u(x,0)=u_0(x), \qquad x\in[0,1],\\ &(u_x-k(u))|_{x=0}\in qH(u(0,t)), \qquad t\in(0,T],\\ &(u_x-k(u))|_{x=1}=0, \qquad t\in(0,T].\end{aligned}$$

The plan of this paper is as follows. In § 2, some assumptions on the problem ($\mathbb{P}$) and the definition of the weak solution of problem ($\mathbb{P}$) will be given. In § 3, using the parabolic regularization process, we prove the existence of a weak solution of ($\mathbb{P}$). In § 4, the uniqueness of weak solutions is proved. In § 5, we discuss the properties of the weak solution. We will see that $u(x,t)>0$ for any $x>0$ and finite t. Moreover we will consider the set $D=\{t\colon u(0,t)=0\}$, proving that there exists a T_0 such that $[T_0,+\infty)\subset D$ (but possibly $D\neq[T_0,+\infty)$); in other words, the surface will become eventually dry. Finally we prove that the weak solution tends to zero uniformly as $t\to+\infty$.

In § 6, we first consider the special case in which the initial condition satisfies

$$(1.4)\qquad 0\leqq u_0'(x)-k(u_0(x))\leqq q,$$

and we prove that the set D is connected, i.e., $D=[T_0,+\infty)$. Then, we pass to estimate the number of the "switching times," i.e., the number of times at which the Signorini boundary condition "switches" from a Neumann-type condition ($u_x-k(u)=q$) to a Dirichlet-type condition ($u=0$). We prove that this number is less than the number of zeros of the function $u_0'(x)-k(u_0(x))-q$.

2. Assumptions and weak solution. We assume that the functions $\theta(u)$, $k(u)$ and $u_0(x)$ satisfy the following conditions:

(A1) $\theta(u)$ is Lipschitz continuous in any closed subset of $\mathbb{R}^+$. $\theta(0)=\theta_0>0$; $\theta(u)\equiv\theta_s$ for $u\geqq u_s$; $\theta'(u)>0$ if $u\in[0, u_s)$; $\theta'(u)\in L^1(0, M_0)$, where M_0 is the upper bound of $u_0(x)$. It is permitted that $\theta'(u)\to\infty$ as $u\to 0^+$.

(A2) $k(u)\in C^{0+1}(\mathbb{R})$; $k(u)\equiv 0$ if $u\in(-\infty, 0]$; $k(u)\equiv K_s$ if $u\in[u_s, +\infty)$; $k'(u)>0$ when $u\in(0, u_s)$; there exists a constant $\sigma_0\in(0, u_s)$ such that $k(u)$ has a continuous first derivative in the interval $(-\infty, \sigma_0)$.

(A3) $\exists C>0$ s.t. $(k'(u))^2/\theta'(u)\leqq C$ in $[0, u_s]$.

(A4) $u_0(x)\in C^{0+1}([0, 1])$; there exists $x_0\in(0, 1)$ such that $m_0\leqq u_0(x)<u_s$ if $x\in[0, x_0)$ and $u_s\leqq u_0(x)\leqq M_0$ if $x\in[x_0, 1]$, where $m_0\in(0, u_s)$; $u_0'(0)-k(u_0(0))=q$; $u_0'(1)-k(u_0(1))=0$.

We define

$$C(u)=\int_0^u k^p(r)\theta'(r)\,dr, \tag{2.1}$$

where $p\geqq 1$ is a constant which can always be determined, as in [9], in such a way that

(C1) $C(r)\in C^{0+1}(\mathbb{R})$; $C(r)\equiv 0$ if $r\in(-\infty, 0]$; $C(r)\equiv C_s$ if $r\in[u_s, \infty)$; $C'(r)>0$ if $r\in(0, u_s)$; $C(r)$ has a continuous second derivative in $(-\infty, \sigma_0)$.

$$(k'(r))^2=0\left(\frac{C'(r)}{k^p(r)}\right) \quad\text{if } r\in(0, M_0);\quad \frac{C'(r)}{k^p(r)}\in L^1(0, M_0). \tag{C2}$$

Let

$$A(u)=\theta(u)-\theta_0 \tag{2.2}$$

and note that

$$A(u)=\int_0^u \frac{c'(s)}{k^p(s)}\,ds. \tag{2.2$'$}$$

Using transformations (2.1) and (2.2), equation (1.3) becomes

$$A(u)_t=u_{xx}-(k(u))_x. \tag{2.3}$$

Let G be a simply connected open region of the (x, t) plane. We will say that G belongs to the class U if $\exists t_0\leqq T$ such that $G\subset H_{t_0}$ and if $I=\partial G\cap\{t=t_0\}$ and $I_0=\partial G\cap\{x=0\}\equiv\{x=0, t_1\leqq t\leqq t_2\leqq t_0\}$ are nonempty. Moreover, the remainder of ∂G is piecewise smooth.

We will use class U in the definition of a weak solution where we will denote by $I_1=\partial G\cap\{x=1\}$ and by $I_c=\partial G\backslash(I\cup I_0\cup I_1)$.

Definition 2.1. A real, bounded, nonnegative, and measurable function $u(x, t)$ defined on $\bar{H}_T$ is called a weak solution to problem $(\mathbb{P})$, if $C(u)$ is continuous in $\bar{H}_T$ and smooth in $\{(x, t)\in H_T: 0<u(x, t)<u_s\}$, $u>0$ in H_T and if

$$\begin{aligned}\iint_G \{(u_x-k(u))\phi_x-A(u)\phi_t\}\,dx\,dt+q\int_{t_1}^{t_2} H(u(0, t))\phi(0, t)\,dt\\ =\int_{\mathrm{l}\overset{+}{\mathrm{c}}} A(u)\phi\,dx+\int_{\mathrm{l}\overset{+}{\mathrm{c}}}(u_x-k(u))\phi\,dt\end{aligned} \tag{2.4}$$

for any $G\in U$, where $\mathrm{l}\overset{+}{\mathrm{c}}$ means that the line integral is counterclockwise, and $\phi\in C^1(\bar{G})$ is an arbitrary test function which vanishes when $(x, t)\in I$. In addition, the mass conservation law is satisfied:

$$\int_0^1\{A(u(x, t))-A(u_0(x))\}\,dx+\int_0^t qH(u(0, t))\,dt=0 \quad\text{for any } t\in[0, T]. \tag{2.5}$$

3. The existence of a weak solution. As in [9], we construct function sequences $\{C_n(s)\}$, $\{k_n(s)\}$, $\{H_n(s)\}$, and $\{u_{0n}(x)\}$, which satisfy the following conditions:

(C_n): $C_n(s) \in C^2(\mathbb{R})$; $C_n(s) = (1/n)s + C(s)$ for $s \in (-\infty, \sigma_1)$, where $\sigma_1 \in (0, \sigma)$; $C_n(s) \to C(s)$ uniformly on any bounded subset of $\mathbb{R}$; $C_n'(s) \to C'(s)$ uniformly on any closed interval in $[0, u_s)$; $1/n \leqq C_n'(s) \leqq C_0$ for $n \geqq 1$ and $s \in \mathbb{R}$, where C_0 is a constant.

(k_n): $k_n(s) \in C^2(\mathbb{R})$; $k_n(s) = k(s)$ for $s \in (-\infty, \sigma_1)$; $k_n(s) \to k(s)$ uniformly on any bounded subset of $\mathbb{R}$; $0 \leqq k_n(s) \leqq k_0$, $0 \leqq k_n'(s) \leqq k_{01}$ for $n \geqq 1$ and $s \in \mathbb{R}$, where k_0 and k_{01} are constants.

(H_n): $H_n(s) \in C^1(\mathbb{R})$; $H_n(s) \equiv 0$ for $s \in (-\infty, 0]$; $H_n(s) \equiv 1$ for $s \in [1/n, \infty)$; $H_n'(s) \geqq 0$.

(u_{0n}): $u_{0n}(x) \in C^2([0, 1])$; $u_{0n}(x) \to u_0(x)$ uniformly on $[0, 1]$; $m_0 - (1/n) \leqq u_{0n}(x) < M_0 + (1/n)$. $|u_{0n}'(x)| \leqq M_{01}$.

At $x = 0$ or $x = 1$, $u_{0n}(x)$ satisfies the compatibility condition

$$u_{0n}''' = (A_n'(u_{0n}))^{-1} A_n''(u_{0n}) u_{0n}' u_{0n}'' + 2k_n'(u_{0n}) u_{0n}'' \\ + \{k_n''(u_{0n}) - (A_n'(u_{0n}))^{-1} A_n''(u_{0n}) k_n'(u_{0n})\}(u_{0n}')^2 - (k_n'(u_{0n}))^2 u_{0n}',$$

where $u_{0n} = k_n(u_{0n}) + qH_n(u_{0n})$ when $x = 0$ and $u_{0n} = k_n(u_{0n}) + k_n(\frac{1}{n})$ when $x = 1$, the function $A_n(s)$ is defined as

$$A_n(s) = \frac{1}{n} + \int_0^s \frac{C_n'(r)}{k_n^p(r) + (1/n)} \, dr.$$

Then, on the basis of [3], the problem

$$\begin{aligned} &(3.1) \qquad (A_n(u_n))_t = u_{nxx} - (k_n(u_n))_x, \qquad (x, t) \in H_T, \\ &(3.2) \qquad u_n(x, 0) = u_{0n}(x), \qquad 0 \leqq x \leqq 1, \\ &(3.3) \qquad u_{nx}(0, t) - k_n(u_n(0, t)) = qH_n(u_n(0, t)), \qquad 0 < t \leqq T, \\ &(3.4) \qquad u_{nx}(1, t) - k_n(u_n(1, t)) = k_n\left(\frac{1}{n}\right), \qquad 0 < t \leqq T, \end{aligned} \qquad (I_n)$$

has a unique solution $u_n(x, t) \in C^{2+\alpha, 1+(\alpha/2)}(\bar{H}_T)$, where $\alpha \in (0, 1)$.

The following lemmas can be proved by essentially standard methods.

LEMMA 3.1. *Let $u_n(x, t)$ be the solution of (I_n). Then there exists a constant M such that*

$$(3.5) \qquad 0 < u_n(x, t) \leqq M$$

for large n.

Proof. Let $U_0 = M_0 + (1/n)$ and define a function $U_n(x)$ by

$$x = \int_{U_0}^{U_n} \frac{ds}{k_n(s) + k_n(2/n)}.$$

Using the maximum principle it is easily proved that

$$0 < u_n(x, t) < U_n(x) \leqq M.$$

LEMMA 3.2. *There exists a constant $M_1 > 0$ such that*

$$(3.6) \qquad |u_{nx}(x, t)| \leqq M_1, \qquad (x, t) \in \bar{H}_T.$$

Proof. Let $V(x, t) = u_{nx}(x, t) - K_n(u_n)$. Then $V(x, t)$ satisfies

$$A_n'(u_n) V_t = V_{xx} - [(A_n'(u_n))^{-1} A_n''(u_n) u_{nx} + k_n(u_n)] V_x.$$

The maximum principle applied to V proves (3.6).

LEMMA 3.3. *There exists a constant $N>0$ such that $C_n(u_n(x,t))$ is uniformly Lipschitz continuous with respect to x and uniformly Hölder continuous (exponent $\frac{1}{2}$) with respect to t in $\bar{H}_T$ when $n>N$.*

Proof. The proof is obtained using condition (C_n), estimate (3.6) and Proposition 1 of [6].

Using the lemmas above, as in [9] we can select a subsequence from $\{u_n(x,t)\}$, which converges weakly in $L^2(H_T)$ to the weak solution $u(x,t)$ of $(\mathbb{P})$. The positivity of u in H_T will be proved in Theorem 5.1. Thus we have the following theorem.

THEOREM 3.1. *Problem $(\mathbb{P})$ has at least one weak solution.*

4. Uniqueness. To prove uniqueness we need the following lemmas.

LEMMA 4.1. *For any fixed constant $M>0$, there exists a constant $F_0>0$ such that*

$$[k(s_1)-k(s_2)]^2 \leqq F_0(s_1-s_2)[A(s_1)-A(s_2)] \tag{4.1}$$

for any $s_1, s_2 \in [0, M]$ if $(k'(s))^2 = 0(A'(s))$ in $(0, M]$.

Proof. See Lemma 1 of [4].

LEMMA 4.2. *Suppose $G \in H_T$ is a simply connected open region with continuous parabolic boundary $\partial_p G$, and u_1 and u_2 are weak solutions of (2.3) with the same boundary value: $u_1(x,t)=u_2(x,t)=f(x,t)$ or u_{1x} $(x,t)-k(u_1(x,t))=u_{2x}(x,t)-k(u_2(x,t))=h(x,t)$ on $\partial_p G$, where the definition of weak solution is usual (e.g., see [6], [9]), then $u_1(x,t)\equiv u_2(x,t)$ in $\bar{G}$.*

Proof. The proof is similar to that of Theorem 2 in [6] or Theorem 2 in [4] and we omit it.

LEMMA 4.3. *Suppose that $G \in U$ (defined in § 2), u_1 and u_2 are two weak solutions of $(\mathbb{P})$. If $C(u_1) \geqq C(u_2)$ in G and $C(u_1)=C(u_2)$ on I_c, then $u_1=u_2$ in G.*

Proof. Let $u_2(x,t)=u_1(x,t)$ when $(x,t)\in \bar{H}_T \backslash \bar{G}$, then the remainder of the proof is similar to that of Theorem 2 in [6] or Theorem 2 in [4] and we omit it.

THEOREM 4.1. *Problem $(\mathbb{P})$ has at most one weak solution.*

Proof. Suppose that u_1 and u_2 are two weak solutions of $(\mathbb{P})$. Then there will be a point $(x_0,t_0)\in H_T$ such that, e.g., $C(u_1(x_0,t_0))>C(u_2(x_0,t_0))$. Thus, let G_0 be the largest simply connected open subset of H_T such that $(x_0,t_0)\in G_0$ and $C(u_1)>C(u_2)$ when $(x,t)\in G_0$. Let G_1 be the intersection of the saturated regions of u_1 and u_2 (G_1 may be empty);

$$t^* = \inf\{t \mid (x,t)\in \partial G_0 \cap \partial G_1\};$$

$$G_2 = G_1 \cap \{(x,t) \mid t^* \leqq t \leqq t_0\};$$

$$I_c = \{(x,t)\in \partial(G_0 \cup G_2) \mid t=\inf\{t \mid (x,t)\in G_0\cup G_2\}\};$$

$$G = \{(x,t)\in H_T \mid \inf\{t \mid (x,t)\in G_0 \cup G_2\} < t < t_0\}.$$

Then, $G\in U$ (defined in § 2). By Lemmas 4.2 and 4.3, $C(u_1)\geqq C(u_2)$ when $(x,t)\in G$. Using the result of Lemma 4.3, we know that $u_1=u_2$ in G. This contradicts $C(u_1)>C(u_2)$ in G_0, and the proof is complete.

5. Some properties of the weak solution. In this section we first prove $u(x,t)>0$ in $\bar{H}_T \backslash \{(x,t) \mid x=0\}$ under the following assumption:

(A5) $\theta'(r)>C_0$ when $r\in[0,\frac{1}{2}u_s]$, where $C_0>0$ is a constant.

Then we prove that there exists a constant $T_0>0$ so that $u(0,t)\equiv 0$ when $t\geqq T_0$. Finally we prove $u(x,t)\to 0$ uniformly as $t\to\infty$.

THEOREM 5.1. *Let $u(x,t)$ be the weak solution of $(\mathbb{P})$. If assumption (A5) holds, then $u(x,t)>0$ in $\bar{H}_T\backslash\{(x,t)\mid x=0\}$.*

Proof. Suppose $u_n(x, t)$ is the solution of problem (I_n). With no loss of generality, we assume $0 \leqq u_n(x, t) \leqq \frac{1}{2}u_s$ so that $A'_n(u_n) \geqq C_0 > 0$ (by assumption (A5)).

For any fixed $x_0 \in (0, 1)$, let $B \in (0, 1)$ such that $\{x | (x - x_0) \leqq B\} \subset (0, 1)$. Then we construct a function

$$V(x, t) = \begin{cases} D\, e^{1/[(x-x_0)^2 - B^2] - \alpha t}, & |x - x_0| < B, \\ 0, & |x - x_0| \geqq B, \end{cases}$$

where D and α are positive constants to be determined later. Let

$$w(x, t) = u_n(x, t) - V(x, t).$$

Take D sufficiently small such that

$$u_{0n}(x) \geqq V(x, 0) \quad \text{for large } n.$$

We have

$$\begin{aligned} Lw &= w_{xx} - k'_n(u_n) w_x - A'_n(u_n) w_t \\ &= -V_{xx} + k'_n(u_n) V_x + A'_n(u_n) V_t \\ &= D\, e^{1/[(x-x_0)^2 - B^2] - \alpha t} \left\{ -\frac{6(x-x_0)^4 + 4(1-B^2)(x-x_0)^2 - 2B^4}{((x-x_0)^2 - B^2)^4} \right. \\ &\qquad \left. \cdot\, k'_n(u_n) \frac{2(x-x_0)}{((x-x_0)^2 - B^2)^2} - \alpha A'_n(u_n) \right\} \end{aligned}$$

in the region $Q = \{(x, t) |\, |x - x_0| < B, 0 < t < T\}$. Consider the sum of the first two terms in the above braces

$$\begin{aligned} S &= -\frac{6(x-x_0)^4 + 4(1-B^2)(x-x_0)^2 - 2B^4}{((x-x_0)^2 - B^2)^4} - \frac{2k'_n(u_n)(x-x_0)}{((x-x_0)^2 - B^2)^2} \\ &= -\frac{6(x-x_0)^4 + 4(1-B^2)(x-x_0)^2 - 2B^4 + 2k'_n(u_n)(x-x_0)((x-x_0)^2 - B^2)^2}{((x-x_0)^2 - B^2)^4}. \end{aligned}$$

Because $0 \leqq k'_n(u_n) \leqq k_{01}$ and

$$6(x-x_0)^4 + 4(1-B^2)(x-x_0)^2 - 2B^4 \to 4B^2 \quad \text{as } |x - x_0| \to B,$$

$$((x-x_0)^2 - B^2)^2 \to 0 \quad \text{as } |x - x_0| \to B,$$

we can choose an appropriate $B_0 \in (0, B)$ such that

$$S < 0 \quad \text{if } B_0 < |x - x_0| < B.$$

Thus we have

(5.1) $$Lw < 0 \quad \text{when } B_0 < |x - x_0| < B.$$

Since $A'_n(u_n) \geqq C_0 > 0$, we can select an appropriate α such that

(5.2) $$Lw < 0 \quad \text{if } 0 < |x - x_0| \leqq B_0.$$

Therefore, $w(x, t)$ cannot take its minimum in Q. Since $w(x, t) \geqq 0$ when $t = 0$ or $|x - x_0| = B$, it follows that

(5.3) $$w(x, t) = u_n(x, t) - V(x, t) \geqq 0 \quad \text{in } \bar{Q}.$$

By (5.3) and the arbitrariness of x_0, we conclude that $u(x, t) > 0$ in H_T.

It remains to prove $u(1, t) > 0$. For this purpose we choose constants B and x_0 such that

$$0 < 1 - x_0 < B < x_0 - B < 2x_0 - 1 < x_0 < 1.$$

Then, we construct a function

$$V(x, t) = \begin{cases} D\, e^{1/[(x-x_0)^2 - B^2] - \alpha t}, & x \in (x_0 - B, 1], \\ 0, & x \in [0, x_0 - B], \end{cases}$$

where D is sufficiently small such that

$$V(x, 0) \leqq u_{0n}(x) \quad \text{for large } n.$$

Let

$$w(x, t) = u_n(x, t) - V(x, t),$$
$$Q = \{(x, t) | x_0 - B < x < 1, t \in (0, T]\},$$
$$Lw = w_{xx} - k'_n(u_n) w_x - A'_n(u_n) w_t.$$

As we did above, we can take α large enough so that

$$Lw \leqq 0 \quad \text{in } Q.$$

Therefore $w(x, t)$ must take its minimum on $\partial_p Q$. Because $w \geqq 0$ when $t = 0$ or $x = x_0 - B$ and $w(1, 0) > 0$, $w(x, t) \geqq 0$ as long as $w(1, t) > 0$.

If there exists $t_1 = \min\{t | w(1, t) = 0\}$, we have $w(x, t) \geqq 0$ in $\bar{Q}_{t1} = \{[x, t) | x_0 - B \leqq x \leqq 1, 0 \leqq t \leqq t_1\}$. Thus,

$$u_{nx}(1, t_1) \leqq V_x(1, t_1). \tag{5.4}$$

But

$$V_x(1, t_1) = -\frac{2(1 - x_0)}{((1 - x_0)^2 - B^2)^2} V(1, t_1)$$
$$< 0 < k_n(u_n(1, t)) + k_n\left(\frac{1}{n}\right) = u_{nx}(1, t),$$

which contradicts (5.4). Thus,

$$w(1, t) > 0 \quad \text{for any } t \in [0, T].$$

This means $u(1, t) > 0$ for any $t \in [0, T]$ and Theorem 5.1 is proved.

THEOREM 5.2. *For the weak solution $u(x, t)$ of* ($\mathbb{P}$), *there exists a constant $T_0 > 0$ such that $u(0, t) \equiv 0$ when $t \geqq T_0$ (i.e., the surface will eventually become dry).*

Proof. Suppose (2.3) has a traveling wave solution $U(X) = U(x\text{-}mt)$, where m is a positive constant. Then $U(X)$ can be defined by

$$X = \int_{U_0}^{U(X)} \frac{ds}{k(s) - (mA(s) + B)}, \tag{5.5}$$

where U_0 and B are constants.

(i) We first prove that there exists a positive constant m_{01} which is small enough such that $q/2m_{01} > \max\{M_0, A(u_s)\}$, where $M_0 \geqq u_0(x)$, and

$$\int_{M_0}^{q/2m_{01}} \frac{ds}{k(s) - (m_{01}A(s) - (q/2))} > 1. \tag{5.6}$$

In fact, if we take

$$m_{01} \in \left(0, \min\left\{\frac{q}{2M_0}, \frac{q}{2A(u_s)}, \frac{q}{2k_s+q+2M_0}\right\}\right), \tag{5.7}$$

then

$$\int_{M_0}^{q/2m_{01}} \frac{ds}{k(s)-(m_{01}A(s)-(q/2))} > \int_{M_0}^{q/2m_{01}} \frac{ds}{k_s+(q/2)} = \left(\frac{q}{2m_{01}} - M_0\right)\left(\frac{1}{k_s+(q/2)}\right) > 1.$$

(ii) In (5.5), take $B = -q/2$, $m = m_{01}$ which satisfies (5.7). Thus, there exists a constant $U_0 \in (M_0, q/2m_{01})$ such that

$$X = \int_{M_0}^{q/2m_{01}} \frac{ds}{k(s)-(m_{01}A(s)-(q/2))} = 1. \tag{5.8}$$

The function $U(X)$, determined by

$$X = \int_{U_0}^{U(X)} \frac{ds}{k(s)-(m_{01}A(s)-(q/2))} \tag{5.9}$$

(where m_{01} and U_0 satisfy (5.7) and (5.8)), has the following properties:

(a) There exists a constant $a < 0$ such that $U(a) = 0$ and $U(\underset{\sim}{x})$ is monotonically increasing for $a \leqq X \leqq 1$.

(b) $U(0) = U_0 > M_0$, $U(1) = q/2M_{01}$.

Because $U(x - m_{01}t) > u(x, t)$ when $t = 0$, by Theorem 3.1 of [5], the first point $P_1 \equiv (x_1, t_1)$ such that $u(P_1) = U(P_1)$ and $u(x, t) < U(x - m_{01}t)$ for $t < t_1$, must belong to the boundaries $x = 0$ or $x = 1$.

If $x_1 = 1$, then $u_x(P_1) \geqq U_x(P_1)$. However,

$$u_x(P_1) = k(u(P_1)) < k(U(P_1)) - (m_{01}A(U(P_1)) - (q/2)) = U_x(P_1).$$

This is a contradiction.

If $x_1 = 0$, we have

$$u_x(P_1) \leqq U_x(P_1). \tag{5.10}$$

In this case, if $U(P_1) > 0$, then

$$\begin{aligned} U_x(P_1) &= k(U(P_1)) - (m_{01}A(U(P_1)) - (q/2)) \\ &= k(U(P_1)) + q - (m_{01}A(U(P_1)) + (q/2)) \\ &< k(U(P_1)) + q \\ &= u_x(P_1). \end{aligned}$$

This contradicts (5.10). Therefore, $U(x - m_{01}t)$ cannot take the same value of $u(x, t)$ before the time at which $U(0\text{-}m_{01}t) = 0$.

Because $U(x - m_{01}t)$ moves rightwards with the constant velocity m_{01}, the zero point of $U(X)$ must arrive at the boundary $x = 0$. In fact, we have $U(x - mt) = u(x, t) = 0$ when $t = -a/m_{01} \equiv T_0$ and $x = 0$.

It is quite evident that for any $\alpha > 0$, the function $U(x - m_{01}t + \alpha)$ possesses the same properties of $U(x - m_{01}t)$. This means that $u(0, t) \equiv 0$ when $t \geqq T_0$, as we had to prove.

THEOREM 5.3. *Suppose $u(x, t)$ is the weak solution of* ($\mathbb{P}$). *Then*

$$\lim_{t\to\infty} \sup_{x\in[0,1]} u(x, t) = 0. \tag{5.11}$$

Proof. On the basis of Theorem 5.2, there exists a constant t_0 such that $u(0, t) \equiv 0$ when $t \geqq t_0$. In the region $H_{t_0}^{+} = \{(x, t) | 0 < x < 1, t_0 < t < \infty\}$, we construct a dominant function $v(x, t)$ which is the usual weak solution of the problem

$$\begin{aligned} &A(V)_t = V_{xx} - k(V)_x, && (x, t) \in H_{t_0}^{+}, \\ &V(x, 0) = V_0(x), && 0 \leqq x \leqq 1, \\ &V(0, t) = h(t), && t_0 < t < \infty, \\ &V_x(1, t) - k(V(1, t)) = 0, && t_0 < t < \infty, \end{aligned}$$

where $V_0(x)$ is constructed by the relation

$$x = \int_{V(0)}^{V(x)} \frac{ds}{k(s)}$$

such that $V_0(x) > u(x, t_0)$, $h(t) \in C^1([t_0, \infty)) \cap C^{0+1}([0, \infty))$, $h'(t) \leqq 0$ and $\lim_{t \to \infty} h(t) = 0$.

There is no difficulty (see [1]) proving

$$\begin{aligned} &0 \leqq u(x, t) \leqq V(x, t) \quad \text{in } \bar{H}_{t_0}^{+}, \\ &V(x, t_2) \leqq V(x, t_1) \quad \text{if } t_2 > t_1, \\ &|V_x(x, t)| \leqq \text{constant}. \end{aligned} \tag{5.12}$$

Then, as in [1], we can prove

$$\lim_{t \to \infty} \sup_{x \in [0, 1]} V(x, t) = 0. \tag{5.13}$$

Consequently, (5.11) follows directly from (5.12) and (5.13).

6. The number of switching. In this section we consider the characterization of the set $D \equiv \{t: u(0, t) = 0\}$. This is an interesting question. For instance, we could look for conditions such that D is a (finite) union of intervals, so that solving the problem with the Signorini-type condition is equivalent to solving a sequence of problems in which the condition on $x = 0$ is $u_x - k(u) = q$ and $u = 0$, alternatively. First, we give here a sufficient condition that guarantees that D is connected, i.e., that once the surface becomes dry, it will not become wet again. We assume

$$0 \leqq u_0'(x) - k(u_0(x)) \leqq q, \tag{6.1}$$

and prove the following theorem.

THEOREM 6.1. *Suppose $u(x, t)$ is the weak solution of* ($\mathbb{P}$) *with assumption* (6.1) *and $t_0 = \inf\{t | u(0, t) = 0\}$, then $u(0, t) \equiv 0$ when $t \geqq t_0$.*

Proof. By (6.1), using the same method we used to prove Lemma 3.2, we can get

$$0 \leqq u_x(x, t) - k(u(x, t)) \leqq q \quad \text{in } \bar{H}_{t0}. \tag{6.2}$$

In the region $H_{t0T} = \{(x, t) | 0 < x < 1, t_0 < t \leqq T\}$ we construct a problem

$$\begin{aligned} &A(V)_t = V_{xx} - k(V)_x, && (x, t) \in H_{t0T}, \\ &V(x, t_0) = u(x, t_0), && 0 \leqq x \leqq 1, \\ &V(0, t) = 0, && t_0 < t \leqq T, \\ &V_x(1, t) - k(V(1, t)) = 0, && t_0 < t \leqq T. \end{aligned} \tag{6.3} \qquad (P_{t0})$$

Then, the problem (P_{t0}) has unique and usual weak solution $V(x, t)$. Using comparison techniques we can prove that

$$0 \leqq V_x(0, t) - k(V(0, t)) \leqq q. \tag{6.4}$$

Then, let

$$u(x, t) = \begin{cases} u(x, t), & (x, t) \in \bar{H}_{t0}, \\ V(x, t), & (x, t) \in \bar{H}_{t0T}. \end{cases} \tag{6.5}$$

It is easy to verify that the $u(x, t)$ is really a weak solution of $(\mathbb{P})$. Then, Theorem 6.1 follows from the uniqueness of the weak solution.

DEFINITION 6.1. We will say that $\bar{t}$ is a switching time for the Signorini condition if $u(0, \bar{t}) = 0$ and $\exists a > 0$ such that $u(0, t) > 0$ in $(\bar{t} - a, \bar{t})$ or in $(\bar{t}, \bar{t} + a)$.

Our aim is to estimate the number of switching times. We will show that it is controlled by the function

$$V_0(x) = u_0'(x) - k(u_0(x)) - q. \tag{6.6}$$

According to our assumptions it is

$$V_0(0) = 0, \qquad V_0(1) = -q. \tag{6.7}$$

To simplify the analysis below we will assume that V_0 has a finite number of zeros and (just to be specific) that $V_0(x) < 0$ in a neighborhood of $x = 0$.

Let u be the weak solution of $(\mathbb{P})$ and set

$$V(x, t) = u_x(x, t) - k(u(x, t)) - q \tag{6.8}$$

and note that $V(1, t) = -q$. We have the following lemma.

LEMMA 6.1. *Let $m \geqq 0$ be the (even) number of zeros of $V_0(x)$ in $(0, 1)$. For any $\tilde{t}$, $V(x, \tilde{t})$ has at most m zeros.*

Proof. The function $V(x, t)$ satisfies the degenerated parabolic equation

$$A'(u) V_t = V_{xx} - [(A'(u))^{-1} A''(u) u_x - k(u)] V_x$$

in H_T. From the maximum principle we have that the level curves $V = 0$ can only originate on $t = 0$. Of course, they can merge or they can end on $x = 0$ or on $t = T$ (the words "originate" and "end" have an obvious meaning in connection with the orientation of the t axis). This concludes the proof.

Next we have Lemma 6.2.

LEMMA 6.2. *Let $\varepsilon, x_1 > 0$ and $u(x, t)$ be the weak solution of the problem*

$$\begin{aligned} &A(u)_t = u_{xx} - k(u)_x, && 0 < x < x_1, \quad 0 < t < T, \\ &u(x, 0) = v_0(x), && 0 \leqq x \leqq x_1, \\ &u_x - k(u)|_{x=0} = q, && 0 < t < T, \\ &u_x - k(u)|_{x=x_1} = q + \varepsilon, && 0 < t < T. \end{aligned}$$

Assume that $v_0(x)$ satisfies assumptions (A4) where, apart from trivial modifications we substitute the last condition by $v_0' - k(v_0)|_{x=x_1} = q + \varepsilon$ and we add the condition $v_0' - k(v_0) > q$ in $(0, x_1]$. Then $u(x, t) > 0$ in $[0, x_1]x[0, T]$.

Proof. Define $U(x)$ by

$$x = \int_{v_0(0)}^{U(x)} \frac{ds}{k(s) + q}.$$

It is clear that $U(x)$ is a classical solution of (1.2). Using a comparison technique, we obtain

$$u(x, t) \geqq U(x) \geqq m_0 > 0,$$

where m_0 appears in the assumption (A4).

THEOREM 6.2. *Let $u(x, t)$ be the weak solution of* ($\mathbb{P}$). *Then the number of switching times cannot exceed $m+1$ where m is the number of zeros of $V_0(x)$ in* $(0, 1)$.

Proof. If the positivity set of V (call it Pos (V)) does not reach $x=0$ in finite time for $t>t_0$, where t_0 is the first switching time, then the argument of Theorem 6.1 still applies and no more switching times can appear.

After t_0 a new switching time t_1 can exist only if a level curve $V=0$ hits $x=0$ at $t=t_1$ and a segment $\{x=0, t \in (t_1, t_1+a)\}$ belongs to ∂ Pos (V). Afterwards, the argument of Lemma 6.2 applies unless a new line $V=0$ hits $x=0$, separating Pos (V) from the t axis. This concludes our proof.

REFERENCES

[1] D. KRÖNER AND J. F. RODRIGUES, *Global behavior for solutions of a porous media equation of elliptic-parabolic type*, J. Math. Pures Appl., 64 (1985), pp. 105-120.

[2] A. FASANO AND M. PRIMICERIO, *Liquid flow in partially saturated porous media*, J. Inst. Math. Appl., 23 (1979), pp. 503-517.

[3] T. I. FOKINA, *On a boundary value problem for parabolic equations with strong nonlinearities*, Vestinik Moskov. Univ. Math. Mech., 30 (1975), pp. 22-27.

[4] B. H. GILDING, *A nonlinear degenerate parabolic equation*, Ann. Scuola Norm. Sup. Pisa, 4 (1977), pp. 393-432.

[5] Z. HUANG, *Asymptotic behavior of the generalized solution of infiltration problem with constant surface flux*, Acta Math. Sinica, 26 (1983), pp. 677-698.

[6] C. J. VAN DUYN AND L. A. PELETIER, *Nonstationary filtration in partially saturated porous media*, Arch. Rational Mech. Anal, 78 (1982), pp. 173-198.

[7] U. HORNUNG, *A parabolic-elliptic variational inequality*, Manuscripta Math., 39 (1982), pp. 155-172.

[8] Z. HUANG, *A filtration problem with surface evaporation*, Boll. Un. Mat. Ital. A (7), 4 (1990), pp. 253-261.

[9] S. XIAO, Z. HUANG, AND C.-Z. ZHOU, *The infiltration problem with constant rate in partially saturated porous media*, Acta Math. Appl. Sinica, 1 (1984), pp. 108-126.

[10] L.-P. YANG, *The free boundary problem of water evaporation*, Doctoral dissertation, Tsinghua Univ., Beijing, 1988.

[11] M. MUSKAT, *The Flow of Homogeneous Fluids Through Porous Media*, McGraw-Hill, New York, 1937.

[12] J. BEAR, *Dynamics of fluids in porous media*, Elsevier, New York, 1972.

[13] J. BEAR AND A. VERRUIJT, *Modeling Groundwater Flow and Pollution*, D. Reidel, Dordrecht, the Netherlands, 1987.

[14] D. KRÖNER, *Parabolic regularization and behaviour of the free boundary for unsaturated flow in a porous medium*, J. Reine Angew. Math., 348 (1984), pp. 180-196.

[15] M. PRIMICERIO AND R. GIANNI, *La filtracion en medios porosos*, Cuadernos Instituto B. Levi, Rosario, 18 (1989), pp. 1-89.

[16] H. M. YIN, *A singular degenerate free boundary problem arising from the moisture evaporation in partially saturated porous media*, preprint, McMaster University, Hamilton, Ontario, 1989.

[17] H. W. ALT AND S. LUCKHAUS, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 184 (1985), pp. 311-341.

[18] J. HULSHOF, *Elliptic-parabolic problems: The interface*, Doctoral dissertation, University of Leiden, Leiden, the Netherlands, 1986.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 346–363, March 1992

004

TRANSIENT STIMULATED RAMAN SCATTERING*

CURTIS R. MENYUK[†] [‡] AND THOMAS I. SEIDMAN[†] [§]

Abstract. The system: $u_\xi = -zv$, $v_\xi = \bar{z}u$, $z_\tau = u\bar{v} - \gamma z$ with $z \to 0$ at $\tau \to -\infty$ and initial data for $\mathbf{u} = (u, v)$ at $\xi = 0$ are considered. Well posedness results are obtained for this and also for a version discretized in τ. Stability is considered as $\xi \to \infty$.

Key words. Raman scattering, well posed, partial differential equations, system, stability

AMS(MOS) subject classifications. 35B40, 35Q60, 78A60

1. Introduction. The Raman effect has played a conspicuous role in physics since its discovery in the 1920s [14], [11]. Specifically, the system of partial differential equations

$$\begin{cases} \text{(i)} \quad \partial u/\partial \xi = -zv, \\ \text{(ii)} \quad \partial v/\partial \xi = \bar{z}u, \\ \text{(iii)} \quad \partial z/\partial \tau = u\bar{v} - \gamma z, \end{cases} \tag{1.1}$$

was first derived to model the interaction of two laser beams with gases when the frequency difference between the beams corresponds to a resonance of the gas molecules [16], [1]. Here, u and v are unknown $\mathbb{C}$-valued functions on $\mathbb{R}_+ \times \mathbb{R}$ (i.e., functions of (ξ, τ) with $0 < \xi < \infty$; $-\infty < \tau < \infty$) which represent the two laser beams, usually referred to as the pump beam and the Stokes beam, respectively. Then the function $-iz$ corresponds to the off-diagonal density matrix element which describes the quantum mechanical state of the gas. The real parameter $\gamma \geq 0$ represents a de-excitation rate due to molecular collisions.

In recent years, these equations have been the focus of intense activity, both experimental and theoretical; some references to the relevant physical literature are provided in our bibliography. It has long been known that (1.1) has a Lax pair when $\gamma = 0$ and so has soliton solutions [2]. On the other hand, we note that it is physically reasonable to require that

$$z(\tau) \to 0 \quad \text{as } \tau \to -\infty \tag{1.2}$$

so that z should be "independent of the infinite past"; yet it was later shown that this physical boundary condition leads to special difficulties which require modification of the standard inverse scattering approach [8], [15], [9]. These modifications seriously complicate the theory, leading to results which are difficult to interpret [10]. We note that soliton-like pulses have been observed in experiments with laser beams whose durations are long compared to the collisional de-excitation time [3], [17] but,

* Received by the editors June 18, 1990; accepted for publication (in revised form) June 6, 1991.

[†] Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21228.

[‡] Also: Department of Electrical Engineering, University of Maryland. The work of this author has been partially supported by the Naval Research Laboratory. (e-mail: menyuk@umbcl.umbc.edu or menyuk@umbc.bitnet.)

[§] The work of this author has been partially supported by U.S. Air Force Office of Scientific Research grant AFOSR-87-0190 and by National Science Foundation grant ECS-8814788. (e-mail: seidman@umbcl.umbc.edu or seidman@umbc.bitnet.)

surprisingly, soliton-like pulses are not observed in experiments with laser beams whose durations are short compared to the molecular de-excitation time [4], [5] — even though setting $\gamma = 0$ is presumably "more legitimate" for the latter case. Indeed, numerical experiments indicate that both u and z tend toward zero almost everywhere as $\xi \to \infty$ for a fairly broad set of initial data [13], [7]; compare §5. Following the somewhat less formal argument of [12], we show here that $z \to 0$ as $\xi \to \infty$. The more detailed asymptotic behavior of u and v as $\xi \to \infty$ remains an open question and, as will become evident in the course of this paper, a somewhat delicate one.

Clearly, there is a need for careful mathematical work. Remarkably, despite the importance of (1.1) in the physics literature, no one until now has even shown that these equations are well posed! The goal of this paper is to place the study of (1.1) on a firm mathematical foundation by demonstrating well posedness, obtaining a number of other simple results relating to the asymptotic behavior of these equations as $\xi \to \infty$, and outlining the remaining difficulties and some open problems. The key insights[1] will be that solutions satisfy the identities

$$(1.4) \qquad |\mathbf{u}(\xi,\tau)|^2 = |\mathbf{u}_0|^2 \quad \text{a.e. } \tau \in \mathbb{R} \quad \text{for all } \xi \geq 0$$

(where $|\mathbf{u}_0|^2 = |u_0(\tau)|^2 + |v_0(\tau)|^2$; see (2.1)) and, also pointwise in τ,

$$(1.5) \qquad \frac{d}{d\xi}\left(\int_{-\infty}^{\tau} |e^{-\gamma(\tau-\tilde{\tau})}u|^2\right) = -\frac{d}{d\xi}\left(\int_{-\infty}^{\tau} |e^{-\gamma(\tau-\tilde{\tau})}v|^2\right) = -|z(\tau)|^2.$$

2. Formulation.

We will use a subscript $_\xi$ (or simply $'$) for (partial) differentiation with respect to ξ and subscript $_\tau$ for differentiation with respect to τ, etc. We will consistently use the notation

$$\mathbf{u} := (u,v), \quad X := \begin{pmatrix} 0 & -z \\ \bar{z} & 0 \end{pmatrix}, \quad \mathbf{u}_0 := (u_0, v_0),$$

$$(2.1) \qquad K^2(\tau) := |\mathbf{u}_0(\tau)|^2 := [|u_0(\tau)|^2 + |v_0(\tau)|^2],$$

$$\sigma = \sigma(\tau) := \int_{-\infty}^{\tau} K^2(\tilde{\tau})\,d\tilde{\tau} \quad \text{so } K^2 d\tau =: d\sigma.$$

We will assume the initial data $\mathbf{u}(0) = \mathbf{u}_0$ is to be in $\mathcal{H}$ so

$$\kappa^2 := \|K\|^2 := \int_{-\infty}^{\infty} K^2\,d\tau < \infty.$$

We can solve (1.1.iii) as an ordinary differential equation in τ, temporarily ignoring the ξ dependence, to obtain

$$z(\tau) = e^{-\gamma(\tau-\tau_*)}z(\tau_*) + \int_{\tau_*}^{\tau} e^{-\gamma(\tau-\tilde{\tau})}u(\tilde{\tau})\overline{v(\tilde{\tau})}\,d\tilde{\tau}$$

[1] We note also that the system (1.1) is invariant under the action of the group $\mathcal{G} = \{g_\vartheta : \vartheta : \mathbb{R} \to \mathbb{R}\}$ of transformations

$$(1.3) \qquad g = g_\vartheta : (u,v) \mapsto (e^{i\vartheta}u, e^{i\vartheta}v)$$

for "arbitrary" real $\vartheta = \vartheta(\tau)$, independent of ξ. So far, however, we have not been able to exploit this insight effectively.

for arbitrary real τ_*, τ. Imposing (1.2), the first term on the right can be omitted "at $-\infty$" so the differential equation (1.1.iii) can replaced by

$$z(\tau) := \int_{-\infty}^{\tau} e^{-\gamma(\tau-\tilde{\tau})} u(\tilde{\tau})\overline{v(\tilde{\tau})}\, d\tilde{\tau} \tag{2.2}$$

as a *definition.* We note that a principal point of difference between our present treatment and most previous work is precisely this imposition of the boundary condition (1.2); compare Remark 4.4 below.

For (2.2) to be meaningful, we need $u\bar{v}$ to be integrable and we will therefore seek solutions in the L^2 space[2]

$$\mathcal{H} := \{\mathbf{u} = (u,v) : \mathbb{R} \to \mathbb{C}^2 : \|\mathbf{u}\|^2 := \int_{-\infty}^{\infty} |\mathbf{u}(\tau)|^2\, d\tau < \infty\}.$$

Along with $\mathcal{H}$, we introduce the spaces

$$\begin{aligned} \mathcal{H}_K &:= \{\mathbf{u} = (u,v) \in \mathcal{H} : |\mathbf{u}(\tau)| = K(\tau) \text{ a.e. } \tau \in \mathbb{R}\}, \\ \mathcal{Z} &:= \{z \in C((-\infty,\infty] \to \mathbb{C}) : z(-\infty) = 0\}, \\ \mathcal{X} &:= \left\{ X = \begin{pmatrix} 0 & -z \\ \overline{z} & 0 \end{pmatrix} : z \in \mathcal{Z} \right\}. \end{aligned}$$

Note that $\sup\{|z(\cdot)|\}$ just gives the norm of $X(\cdot) \in \mathcal{X}$ as an operator on $\mathcal{H}$ (or on any $\mathcal{H}_K \subset \mathcal{H}$), acting by pointwise multiplication. We will also introduce the linear space $\boldsymbol{\mathcal{U}}$ of functions $\mathbf{u}(\cdot) \in C(\mathbb{R}_+ \to \mathcal{H})$ for which the exponentially weighted norm

$$\|\mathbf{u}\|_\kappa := \sup_{\xi \ge 0}\{e^{-2\kappa^2\xi}\|\mathbf{u}(\xi,\cdot)\|_{\mathcal{H}}\} \tag{2.3}$$

is finite for some κ. Convergence in $\boldsymbol{\mathcal{U}}$ is given by convergence in $\|\cdot\|_\kappa$ for every large enough κ; $\boldsymbol{\mathcal{U}}$ is then metrizable and complete. We finally let $\boldsymbol{\mathcal{U}}_K$ be the subset of $\mathbf{u} \in \boldsymbol{\mathcal{U}}$ taking values almost everywhere in $\mathcal{H}_K$—topologized through $\|\cdot\|_\kappa$ with $\kappa := \|K\|$.

Finally, we introduce the map

$$\mathbf{X} : \mathbf{u} \mapsto X := \begin{pmatrix} 0 & -z \\ \overline{z} & 0 \end{pmatrix} \quad \text{for } \mathbf{u} = (u,v) \in \mathcal{H} \tag{2.4}$$

with $z = z(\cdot)$ defined by (2.2). Then (1.1.i, ii) can be written as an abstract ordinary differential equation with respect to ξ in the more succinct form

$$\mathbf{u}' = \mathbf{X}(\mathbf{u})\mathbf{u}, \qquad \mathbf{u}(0) = \mathbf{u}_0 \in \mathcal{H}. \tag{2.5}$$

[2] Clearly it would be sufficient to have this integrability only "to the left," i.e., on each semi-infinite interval $(-\infty, \tau]$ with τ finite, imposing no growth condition as $\tilde{\tau} \to +\infty$. The present formulation permits us to work with a Hilbert space formulation for $\mathbf{u} = (u,v)$ and the greater generality can be recovered—observing that, by setting everything equal to zero for $\tau > \tau^*$, we can always restrict the problem to $(-\infty, \tau^*]$ for each (arbitrary) finite τ^* to make the present formulation appropriate.

Note that although we use a $\mathbb{C}^2$ notation, thinking of $\mathcal{H}$ as a complex Hilbert space, it will later be convenient (cf., Thm. 3.4) to treat it also as a real Hilbert space, effectively identifying $\mathbb{C}^2$ with $\mathbb{R}^4$.

3. Well posedness. Our principal concern here is to show that the problem (2.5) has a unique solution, but we begin with a lemma about the map $\mathbf{u} \mapsto \mathbf{X}(\mathbf{u})$.

LEMMA 3.1. *The map* $\mathbf{X}$ *is well defined by* (2.2) *and* (2.4), *and is continuous from* $\mathcal{H}$ *to* $\mathcal{X}$. *For each* $\tilde{\mathcal{H}}_K := \cup\{\mathcal{H}_{\tilde{K}} : \tilde{K} \leq K\}$, *the set of functions* z *defined by* (2.2) *with* $\mathbf{u} \in \tilde{\mathcal{H}}_K$ *is precompact in* $\mathcal{Z}$ *and the map* $\mathbf{X}$ *is uniformly Lipschitzian on* $\tilde{\mathcal{H}}_K$ *with Lipschitz constant* $\kappa := \|K\|$.

Proof. Suppose $\mathbf{u} \in \mathcal{H}$ with $|\mathbf{u}(\tau)| \leq K(\tau)$ almost everywhere and obtain z from $\mathbf{u}$ as in (2.2). Noting that $e^{-\gamma(\tau-\tilde{\tau})} \leq 1$ and that $2|u\overline{v}| \leq K^2$, we then clearly have

$$|z(\tau) - z(\hat{\tau})| = \left|\int_{\hat{\tau}}^{\tau} e^{-\gamma(\tau-\tilde{\tau})} u\overline{v}\right| \leq \tfrac{1}{2}\int_{\hat{\tau}}^{\tau} K^2$$

which shows the continuity of z and, indeed, equicontinuity on $\tilde{\mathcal{H}}_K$; essentially the same computation shows that $\{z(\tau_n)\}$ is (uniformly on $\tilde{\mathcal{H}}_K$) always Cauchy as $\tau_n \to \infty$ so $z(\tau)$ always has a limit as $\tau \to \infty$. Similarly, there is a uniform bound: $|z(\tau)| \leq \frac{1}{2}\kappa^2$. By the Arzela–Ascoli Theorem, it follows that the relevant $\{z\}$ will be in a compact subset of $\mathcal{Z}$. Now let z_1, z_2 be obtained from $\mathbf{u}_1, \mathbf{u}_2$ and set $z := z_1 - z_2$, $\mathbf{u} := \mathbf{u}_1 - \mathbf{u}_2$ so, pointwise in τ, we have

$$\begin{aligned} &u_1\overline{v}_1 - u_2\overline{v}_2 = u_1\overline{v} + u\overline{v}_2 = u\overline{v}_1 + u_2\overline{v}, \\ &|u_1\overline{v}_1 - u_2\overline{v}_2| \leq \min\{[|u_1|^2 + |v_2|^2], [|v_1|^2 + |u_2|^2]\}^{1/2}|\mathbf{u}|. \end{aligned}$$

Note that this minimum is bounded by the average—which is bounded by $K(\tau)$ for $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{H}_K$. Thus,

$$|z(\tau)| \leq \int_{-\infty}^{\tau} K|\mathbf{u}| \leq \|K\|\|\mathbf{u}\|.$$

Since we are using $\sup\{|z(\cdot)|\}$ as our $\mathcal{X}$-norm, we then get for $\mathbf{X}(\cdot)$ the desired Lipschitz condition with constant $\kappa := \|K\|$. □

THEOREM 3.2. *Let* $\mathbf{u}_0 = (u_0, v_0)$ *be given in* $\mathcal{H}$. *Then there is a unique function* $\mathbf{u} = (u, v) : \mathbb{R}_+ \to \mathcal{H}$ *in* $\boldsymbol{\mathcal{U}}_K$ *satisfying the nonlinear equation* (2.5) *with the notation of* (2.1) *and* (2.4).

Before beginning the proof, we remark that an essentially identical argument works for the problem with ξ reversed (here and also in Theorem 3.3) so, in particular, we have *backward uniqueness* for the solution as well. We also remark that our definition of $\boldsymbol{\mathcal{U}}_K$ means that finding $\mathbf{u} \in \boldsymbol{\mathcal{U}}_K$ implicitly includes the assertion of (1.4).

Proof. Fix $\mathbf{u}_0 \in \mathcal{H}$, thus fixing $K := |\mathbf{u}| \in L^2_+(\mathbb{R})$ and the spaces $\mathcal{H}_K, \boldsymbol{\mathcal{U}}_K$ as above. There is no difficulty in defining a map $\mathbf{F} : \tilde{\mathbf{u}} \mapsto \mathbf{u}$ for $\tilde{\mathbf{u}} \in \boldsymbol{\mathcal{U}}_K$ by solving the linear ordinary differential equation

$$\mathbf{u}' = \mathbf{X}(\tilde{\mathbf{u}})\mathbf{u}, \qquad \mathbf{u}(0) = \mathbf{u}_0. \tag{3.1}$$

Indeed, we note from Lemma 3.1 that $\mathbf{X}(\tilde{\mathbf{u}})\mathbf{u}$ is continuous on $\mathbb{R}_+ \times \mathbb{R}$ so (3.1) can be interpreted pointwise in τ as a (finite-dimensional) ordinary differential equation in ξ—with an adequately defined initial condition for almost every τ. Note that, since $X := \mathbf{X}(\tilde{\mathbf{u}})$ is skew-adjoint, we have from (3.1) that

$$(|\mathbf{u}|^2)' = 2\langle \mathbf{u}, \mathbf{u}' \rangle = 2\langle \mathbf{u}, X\mathbf{u} \rangle \equiv 0$$

whence $|\mathbf{u}(\cdot,\tau)|$ is constant and we have (1.4) for solutions of (3.1), i.e., we have $\mathbf{u}(\cdot) \in \boldsymbol{\mathcal{U}}_K$. (Indeed, we need not even have $\tilde{\mathbf{u}} \in \boldsymbol{\mathcal{U}}_K$ to have $\mathbf{u} =: \mathbf{F}(\tilde{\mathbf{u}}) \in \boldsymbol{\mathcal{U}}_K$ and $\|\mathbf{u}(\xi,\cdot)\| \equiv \kappa$.)

A fixed point for $\mathbf{F}$ is a solution of (2.5) so it will be sufficient to show that $\mathbf{F}$ is a uniformly strict contraction from $\boldsymbol{\mathcal{U}}_K$ to itself. Given $\mathbf{u}_j := \mathbf{F}(\tilde{\mathbf{u}}_j)$ for $j = 1, 2$, set $\mathbf{u} := \mathbf{u}_1 - \mathbf{u}_2$, $\tilde{\mathbf{u}} := \tilde{\mathbf{u}}_1 - \tilde{\mathbf{u}}_2$ and let $\tilde{\nu}$ be the $\boldsymbol{\mathcal{U}}$-norm of $\tilde{\mathbf{u}}$. Then

$$\begin{aligned}(e^{-2\kappa^2\xi}\|\mathbf{u}\|^2)' &= e^{-2\kappa^2\xi}\left(-2\kappa^2\|\mathbf{u}\|^2 + 2\langle \mathbf{u}, [\mathbf{X}(\tilde{\mathbf{u}}_1) - \mathbf{X}(\tilde{\mathbf{u}}_2)]\,\mathbf{u}_2\rangle\right)\\ &\le e^{-2\kappa^2\xi}\left(-2\kappa^2\|\mathbf{u}\|^2 + 2\kappa\|\tilde{\mathbf{u}}\|\|\mathbf{u}_2\|\right)\\ &\le (\kappa^2/2)\, e^{-2\kappa^2\xi}\|\tilde{\mathbf{u}}\|^2.\end{aligned}$$

Since $\|\tilde{\mathbf{u}}(\xi)\|^2 \le \exp[4\kappa^2\xi]\tilde{\nu}^2$ and $\mathbf{u}(0) = 0$, integrating gives

$$e^{-2\kappa^2\xi}\|\mathbf{u}\|^2 \le \frac{\kappa^2}{2}\int_0^\xi e^{2\kappa^2\bar{\xi}} \le \tfrac{1}{4}e^{2\kappa^2\xi}\tilde{\nu}^2$$

and then

$$\left(e^{-2\kappa^2\xi}\|\mathbf{u}\|\right)^2 \le \tilde{\nu}^2/4$$

which shows that $\mathbf{F}$ is uniformly Lipschitzian on $\boldsymbol{\mathcal{U}}_K$ with Lipschitz constant $\frac{1}{2}$. The result then follows by the Contraction Mapping Theorem. □

We complete our treatment of well posedness by considering the continuous dependence of the solution on the initial data $\mathbf{u}_0$. It is clear that the estimate (3.2) gives continuous dependence of solutions on initial data in the sense of uniform convergence (with respect to the $\mathcal{H}$-norm) on bounded ξ-intervals but the estimate grows exponentially in ξ. Since we know that the solutions themselves are bounded uniformly in ξ, it might seem plausible that this could be improved to have convergence uniform on $\mathbb{R}_+$. That, however, is false; see Remark 6.4.

THEOREM 3.3. *Let* $\mathbf{u}_{j0} = (u_{j0}, v_{j0})$ *for* $j = 1, 2$ *be given in* $\mathcal{H}$ *with corresponding solutions* $\mathbf{u}_j : \mathbb{R}_+ \to \mathcal{H}$ *satisfying* (2.5). *Then*

$$\|\mathbf{u}_1(\xi,\cdot) - \mathbf{u}_2(\xi,\cdot)\| \le e^{(\kappa_+\kappa_-)\xi}\|\mathbf{u}_{10} - \mathbf{u}_{20}\| \tag{3.2}$$

where $\kappa_\pm := \|K_\pm\|$ *with* $K_\pm(\tau) := \max, \min\{|\tilde{\mathbf{u}}_0(\tau)|, |\hat{\mathbf{u}}_0(\tau)|\}$.

Proof. The argument is standard. Set $\mathbf{v} := \mathbf{u}_1 - \mathbf{u}_2$ and $X := X_1 - X_2$ with $X_j := \mathbf{X}(\mathbf{u}_j)$, etc. Pointwise in τ, we have $|\mathbf{v}|^2 = \langle \mathbf{v}, \mathbf{v}\rangle$ so

$$\begin{aligned}\partial|\mathbf{v}|^2/\partial\xi &= 2\mathrm{Re}[\langle \mathbf{v}, X_1\mathbf{v}\rangle + \langle \mathbf{v}, X\mathbf{u}_2\rangle]\\ &= 2\mathrm{Re}[\langle \mathbf{v}, X_2\mathbf{v}\rangle + \langle \mathbf{v}, X\mathbf{u}_1\rangle]\\ &= 2\mathrm{Re}\langle \mathbf{v}, X\mathbf{u}_2\rangle = 2\mathrm{Re}\langle \mathbf{v}, X\mathbf{u}_1\rangle\end{aligned}$$

by the skew symmetry of X_1, X_2. By Lemma 3.1, we have $|X(\tau)| \le \|K_+\|\,\|\mathbf{v}\|$ and (1.4) gives, pointwise in τ, $\min\{|\mathbf{u}_1|, |\mathbf{u}_2|\} \equiv K_-$. Thus,

$$\begin{aligned}\partial|\mathbf{v}|^2/\partial\xi &\le 2|\mathbf{v}(\tau)|\|K_+\|\|\mathbf{v}\|K_-(\tau),\\ d\|\mathbf{v}\|^2/d\xi &\le 2\|\mathbf{v}\|^2\|K_+\|\|K_-\|.\end{aligned}$$

The result now follows on applying the Gronwall inequality. □

Extending Theorem 3.3, we next wish to consider linearization of the system, i.e., differentiability of the dependence on initial data.

THEOREM 3.4. *The solution map* $\mathbf{S} : \mathcal{H} \to \boldsymbol{\mathcal{U}}$ *for* (2.5), *is "Fréchet differentiable" (in a sense to be made precise below). At each* $\mathbf{u}_0^* \in \mathcal{H}$ *and corresponding* $\mathbf{u}^* := \mathbf{S}(\mathbf{u}_0^*)$ *the derivative is the linear map:* $\mathcal{H} \to \boldsymbol{\mathcal{U}} : \mathbf{u}_0 = (u_0, v_0) \mapsto \mathbf{u}$ *defined by the linearized system:*

$$(3.3) \qquad \begin{cases} u' &= -z^*v - v^*z, \\ v' &= \overline{z}^*u + u^*\overline{z}, \end{cases} \qquad z := \int_{-\infty}^{\tau} e^{-\gamma(\tau-\tilde{\tau})}[\overline{v}^*u + u^*\overline{v}]\, d\tilde{\tau}$$

with $\mathbf{u}(0) = \mathbf{u}_0$. *In particular, for any* $\mathbf{u}_0^* \in \mathcal{H}$ *yielding* $\mathbf{u}^*$ *by* (2.5) *the equation* (3.3) *has a unique solution* $\mathbf{u}$ *for each initial* $\mathbf{u}_0 \in \mathcal{H}$ *and* $\mathbf{u}(\xi, \cdot)$ *will be bounded in* $\mathcal{H}$ *uniformly on bounded* ξ*-intervals.*

A word of caution is in order here since we have been working with complex spaces: *the solution map is not differentiable when complex differentiation is considered* since (2.5) involves conjugations. Instead, as noted earlier, although we have made no alteration in the notation, we are here considering $\mathcal{H} = L^2(\mathbb{R} \to \mathbb{C}^2)$ as isometrically equivalent to the *real* Hilbert space $L^2(\mathbb{R} \to \mathbb{R}^4)$, etc.

Proof. Now consider Theorem 3.3 with $\mathbf{u}_{20} = \mathbf{u}_0^*$ and, for $s \neq 0$, $\mathbf{u}_{10} = \mathbf{u}_0^* + s\mathbf{u}_0$; set $\mathbf{v} = \mathbf{v}(\xi, \tau; s) := [\mathbf{u}_1 - \mathbf{u}^*]/s$. Then (3.2) gives the uniform estimate $\|\mathbf{v}\| \leq e^{(\kappa_+ \kappa_-)\xi}\|\mathbf{u}_0\|$, where $\kappa_\pm = \kappa_\pm(s) \to \|\mathbf{u}_0^*\|^2$ as $s \to 0$. A standard argument then shows that $\mathbf{v}(\cdot; s)$ satisfies a system whose right-hand side tends to that of (3.3) as $s \to 0$ ($\mathcal{O}(s)$ difference in the coefficients) so $\|\mathbf{v}(\cdot; s) - \mathbf{u}\|_\kappa \to 0$ for any $\kappa > \|\mathbf{u}_0^*\|$. Temporarily fixing any such κ, we may treat (the relevant subspace of) $\boldsymbol{\mathcal{U}}$ as a Banach space with the norm $\|\cdot\|_\kappa$ and $\mathbf{u}$, given by (3.3), is the Gâteaux differential of $\mathbf{S}(\cdot)$ at $\mathbf{u}^*$ in the direction of $\mathbf{u}_0$. This is clearly linear in $\mathbf{u}_0$, so this gives a Gâteaux derivative. It is continuous in $\mathbf{u}_0^*$ (as long as we stay close enough to the original $\mathbf{u}_0^*$ so as not to disturb the choice of κ), so this is necessarily a Fréchet derivative, working with this $\|\cdot\|_\kappa$. [We do note that when considering bounded ξ-intervals, the choice of κ is irrelevant; in any case, (3.2) gives control of errors in the norm with $\kappa = \|\mathbf{u}_0^*\|$.] □

4. Some remarks.

Our first concern here is to verify (1.5).

LEMMA 4.1. *Let* $\mathbf{u}$ *be any solution of the system* (2.5) *with* $\mathbf{u}_0$ *in* L^2. *Then,*

$$(4.1) \qquad \begin{cases} \text{(i)} & \left(\int_{-\infty}^{\tau} |e^{-\gamma(\tau-\tilde{\tau})}u|^2\right)' = -|z|^2, \\ \text{(ii)} & \left(\int_{-\infty}^{\tau} |e^{-\gamma(\tau-\tilde{\tau})}v|^2\right)' = |z|^2, \\ \text{(iii)} & z'(\cdot,\tau) = \int_{-\infty}^{\tau} e^{-\gamma(\tau-\tilde{\tau})}\left(|u|^2 - |v|^2\right) z \end{cases}$$

for all $\xi > 0$ *and all* $\tau \in \mathbb{R}$.

Proof. From (1.1) we have $[e^{\gamma\tau}z]_\tau{}' = e^{\gamma\tau}\left(|u|^2 - |v|^2\right) z$ and, using (4.5), integrating this gives (4.1.iii). Similarly, we have

$$\overline{e^{\gamma\tau}z}(e^{\gamma\tau}z)_\tau = e^{2\gamma\tau}\overline{z}(u\overline{v}) = e^{2\gamma\tau}\overline{v}v_\xi = -e^{2\gamma\tau}u\overline{u_\xi}$$

and integrating this gives (4.1.i) and (4.1.ii). That these identities hold pointwise for *all* $\tau \in [0, 1]$ follows from the known continuity in τ of z and continuity of the indefinite integral in the third identity. □

COROLLARY 4.2. *If z is real (alternatively, if z is pure imaginary or if z vanishes identically) for $\tau \le \tau_*$ at $\xi = \xi_0$, then this holds for all $\xi \ge 0$. For the case in which $z \equiv 0$ on $\mathbb{R}_+ \times (-\infty, \tau_*)$, we have $\mathbf{u}$ stationary (independent of ξ) there and conversely.*

Proof. The first assertion follows immediately from (4.1.iii), viewing $e^{-\gamma(\tau-\tilde{\tau})}(|u|^2 - |v|^2)$ simply as an integrable real function and integrating this ODE forward or, as in the remark following the statement of Theorem 3.2, backward in ξ. The case of $z \equiv 0$ is obvious with the converse following by, e.g., (4.1.i). □

Remark 4.3. While the system (1.1) cannot give analyticity in its dependence on the initial data, we observe that we could consider the analytic system

$$(4.2) \qquad \begin{cases} u_1' = -z_1 v_1, \\ v_1' = \ \ z_2 u_1, \\ u_2' = -z_2 v_2, \\ v_2' = \ \ z_1 u_2, \end{cases} \qquad \begin{cases} z_1 := \int_{-\infty}^{\tau} e^{-\gamma(\tau-\tilde{\tau})} u_1 v_2, \\ z_2 := \int_{-\infty}^{\tau} e^{-\gamma(\tau-\tilde{\tau})} u_2 v_1, \end{cases}$$

and have

$$(4.3) \qquad [u_1, v_1, z_1] \equiv [u, v, z] \qquad [u_2, v_2, z_2] \equiv [\overline{u}, \overline{v}, \overline{z}]$$

for all real $\xi > 0$ if this holds initially, at $\xi = 0$.

Without (4.3), we do not have the estimate (1.4) and so it is not clear when solutions for (4.2) will exist globally. On any finite ξ-interval, however, we can get existence for initial data almost satisfying (4.3) and thus, analytic linearization, subject to (4.3), with (3.3) suitably extended to a complex neighborhood of $\mathbb{R}_+$. In particular, this shows that we can also obtain higher derivatives of the solution map with respect to the initial data.

Remark 4.4. For nonstationary solutions, it is interesting to consider the case in which $\gamma = 0$ and z is real on $\mathbb{R}$ — initially, at $\xi = 0$, and so for all $\xi \ge 0$ by Corollary 4.2.

For this we will first reduce the problem to a more convenient form, modifying our notation somewhat, without (at first) taking $\gamma = 0$. If we set $\mathbf{u} =: K\tilde{\mathbf{u}}$ pointwise in τ, then we have the identity

$$(4.4) \qquad |\tilde{\mathbf{u}}|^2 := |\tilde{u}|^2 + |\tilde{v}|^2 \equiv 1 \quad \text{pointwise in } \xi, \tau$$

in view of Theorem 3.2. It is easily seen that $\tilde{\mathbf{u}}$ also satisfies (2.5), provided we modify the definition of the operator $\mathbf{X}(\cdot)$ by replacing (2.2) with

$$(4.5) \qquad z(\tau) := \int_{-\infty}^{\tau} e^{-\gamma(\tau-\tilde{\tau})} u(\tilde{\tau}) \overline{v(\tilde{\tau})} K^2(\tilde{\tau})\, d\tilde{\tau}.$$

Of course, the initial data now must satisfy: $|\tilde{\mathbf{u}}_0| \equiv 1$, pointwise in τ.

That much reduction is available for all γ, but when $\gamma = 0$ we can conveniently use the variable σ of (2.1) to view $\tilde{\mathbf{u}}$ as a function of (ξ, σ), rather than of (ξ, τ). This further reduction will actually (σ—almost everywhere by Sard's Theorem) avoid any difficulty with the definition of $\tilde{u}$ when $K(\tau) = 0$. We note that it is possible to view z also as a function of (ξ, σ) since, while the function $\sigma(\cdot)$ may not be injective, this can happen only if K vanishes on some subinterval — in which case $\mathbf{u}$ (whence $u\overline{v} = z_\tau$)

also vanishes on this subinterval so z is constant there: the value of $z(\xi,\tau)$ depends only on (ξ,σ). In this case, the domain of σ is $[0,\kappa^2]$ and, henceforth omitting the $\tilde{}$, (4.5) becomes simply

$$z(\cdot,\sigma) := \int_0^\sigma u\bar{v}\,d\tilde{\sigma}. \tag{4.6}$$

Apart from the name of the variable ($\sigma \leftrightarrow \tau$), we observe that this is identical to the *original* problem for $\gamma = 0$ with initial data giving

$$K(\tau) = \{1 \text{ for } 0 \le \tau \le \kappa^2; 0 \text{ else}\}.$$

Note that use of (4.6) means that we have the boundary condition

$$z = 0 \quad \text{at } \sigma = 0, \tag{4.7}$$

corresponding to (1.2).

In view of (4.4), $\mathbf{u}$ must have the form

$$u = e^{i\vartheta}\cos\varphi, \qquad v = e^{i\vartheta}\sin\varphi \tag{4.8}$$

with ϑ,φ real. Assume ϑ is independent of ξ — first as an *ansatz*, but then confirmed by our subsequent calculations. We then see that (2.5), (4.6) are equivalent to the requirement that $z = \varphi_\xi$ and

$$\varphi_{\xi\sigma} = u\bar{v} = \tfrac{1}{2}\sin 2\varphi. \tag{4.9}$$

If we set $t := \sigma + \xi$ and $x := \sigma - \xi$, then (4.9) becomes

$$2\varphi_{tt} - 2\varphi_{xx} = \sin 2\varphi, \tag{4.10}$$

i.e., 2φ satisfies the sine-Gordon equation. Conversely, if $2\varphi = 2\varphi(t,x)$ is any real solution of the sine-Gordon equation, then, for arbitrary real $\vartheta(\sigma)$,

$$\begin{aligned} u(\xi,\sigma) &:= e^{i\vartheta(\sigma)}\cos\varphi(\sigma+\xi,\sigma-\xi),\\ v(\xi,\sigma) &:= e^{i\vartheta(\sigma)}\sin\varphi(\sigma+\xi,\sigma-\xi),\\ z &:= [\varphi_t - \varphi_x](\sigma+\xi,\sigma-\xi) \end{aligned}$$

gives a solution of (1.1) for $\gamma = 0$ with $z \equiv \varphi_\xi$ real.

For this to be consistent with the boundary conditions (4.7) we are imposing on z, it is necessary that 2φ be a solution of the sine-Gordon equation satisfying

$$\varphi_t(t,x) \equiv \varphi_x(t,x) \quad \text{along the line: } t + x = 0. \tag{4.11}$$

While the sine-Gordon equation has nontrivial traveling wave solutions, we emphasize that those are all excluded by this constraint (4.11)—corresponding to (4.7) and so to our original boundary condition (1.2) at $\tau \to -\infty$.

5. Discrete approximation. In this section we consider the "obvious" discretizations (with respect to τ) of the system (2.5). For convenience, we restrict our attention to the case $\gamma = 0$ and take the system reduced as in Remark 4.4 so that $|u_0|^2 + |v_0|^2 \equiv 1$ for $0 \le \sigma \le \kappa^2$ and

$$\begin{cases} u' &= -zv \\ v' &= \bar{z}u \end{cases} \qquad \text{with } z(\cdot,\sigma) := \int_0^\sigma u\bar{v}\,d\tilde{\sigma}. \tag{5.1}$$

While much of our analysis here directly parallels that for the partial differential equation system, for this finite-dimensional approximation we have the advantage of local compactness and will be able to obtain a more complete description of the asymptotic behavior of solutions as $\xi \to \infty$. This may be viewed both as a theoretical complement to the results observed in computational simulation and for comparison with our less complete asymptotic analysis in the next section, as an indication of goals and conjectures for future work.

For the remainder of this section, we adopt the notation that

$$\begin{aligned}
\mathcal{H} &:= (\mathbb{C}^2)^J,\\
\mathbf{u} &:= (\mathbf{u}_1, \cdots, \mathbf{u}_J) \in \mathcal{H}, \qquad [\mathbf{u}_j := (u_j, v_j) \in \mathbb{C}^2],\\
\mathcal{H}_1 &:= \{\mathbf{u} \in \mathcal{H} : |\mathbf{u}_j|^2 := |u_j|^2 + |v_j|^2 = 1\},\\
\mathcal{S} &:= \{\mathbf{u}^0 \in \mathcal{H}_1 : u_j^0 = 0 \text{ or } v_j^0 = 0 \quad \text{for each } j = 1, \cdots, J,\\
U_j &:= \delta\Sigma_1^j |u_k|^2,\\
\nu &:= \max\{U_j/j\delta : j = 1, \cdots, J\}.
\end{aligned}$$

In general, we have $U_j = U_j(\cdot)$ for some particular solution $\mathbf{u}(\cdot)$ of (5.3), below, and ν will similarly relate to this; we could write explicitly $\nu := \nu(\mathbf{u})$ for $\mathbf{u} \in \mathcal{H}_1$ or $\nu := \nu(\xi; \mathbf{u}^0) := \nu(\mathbf{u}(\xi))$, with $\mathbf{u}(\cdot)$ satisfying (5.3) with initial data $\mathbf{u}^0$.

Assuming the nodes are equally spaced with respect to σ, we can then define $z_j \approx z(j\delta)$ by a discretized approximation to the integral

$$z_j := \delta\Sigma_1^j u_k \overline{v_k} = \delta u_j \overline{v_j} + z_{j-1}, \qquad z_0 := 0, \tag{5.2}$$

for $j = 1, \cdots, J$. Thus, we consider here the system of ordinary differential equations

$$\begin{aligned}
u_j' &= -z_j v_j\\
v_j' &= \overline{z_j} u_j\\
\text{with}\, u_j(0) &= u_j^0, \quad v_j(0) = v_j^0
\end{aligned} \tag{5.3}$$

for each $j = 1, \cdots, J$, using (5.2). For the initial conditions we assume, as earlier, that

$$|u_j^0|^2 + |v_j^0|^2 = 1 \quad \text{for } j = 1, \cdots, J.$$

The factor δ could be removed by rescaling ξ, but we will retain $\delta := \kappa^2/J$ here to remind us of the correspondence: $u_j(\xi) \approx u(\xi, j\delta)$, etc. We note that (5.3) is equivalent to

$$\begin{aligned}
u_j' &= -\delta |v_j|^2 u_j - \delta\zeta_j v_j,\\
v_j' &= \delta |u_j|^2 v_j + \delta\zeta_j u_j
\end{aligned} \tag{5.4}$$

with $\delta\zeta_j := z_{j-1}$.

We begin by asserting the set of "background" results.

LEMMA 5.1. *For each* $\mathbf{u}^0 \in \mathcal{H}$ *there is a unique global solution* $\mathbf{u}(\cdot) = \mathbf{u}(\cdot; \mathbf{u}^0)$ *of* (5.3)*, depending on* $\mathbf{u}^0$ *uniformly on bounded* ξ*-intervals. The functions* $u_j, \overline{u}_j, \cdots, \overline{z}_j$ *are all real-analytic functions of* ξ. *For* $\mathbf{u}^0 \in \mathcal{H}_1$*, we have* $\mathbf{u}(\xi) \in \mathcal{H}_1$ *for all* $\xi \geq 0$.

Proof. The arguments are straightforward parallels of those given above for Theorems 3.2, 3.4, and Remark 4.3. Details are left to the reader. □

LEMMA 5.2. *For $j = 1, \cdots, J$ we have*

$$(5.5) \qquad \begin{cases} \text{(i)} & [u_j \bar{v}_j]' = z_j \left[|u_j|^2 - |v_j|^2 \right], \\ \text{(ii)} & z_j' = \delta \Sigma_1^j z_k \left[|u_k|^2 - |v_k|^2 \right], \\ \text{(iii)} & -U_j' = |z_j|^2 + \delta^2 \Sigma_1^j |u_k \bar{v}_k|^2. \end{cases}$$

If, for $k = 1, \cdots, j$, we have $z_k(0) = 0$, then this persists for all ξ, and u_k, v_k are then constant (independent of ξ).

Proof. The formulas (5.5.i, ii) are direct from (5.3) and the final assertion follows; compare Corollary 4.2. Also from (5.3), for each k we have

$$\begin{aligned} -\delta(|u_k|^2)' &= \delta[\bar{u}_k(z_k v_k) + u_k \overline{z_k v_k}] \\ &= \overline{(z_k - z_{k-1})} z_k + \bar{z}_k (z_k - z_{k-1}) \\ &= |z_k|^2 - |z_{k-1}|^2 + |z_k - z_{k-1}|^2, \end{aligned}$$

since $z_k - z_{k-1} = \delta u_k \bar{v}_k$. Summing over $k = 1, \cdots, j$ gives (5.5.iii). □

LEMMA 5.3. *If (for some k) z_k is not identically zero, then (for each $j \geq k$) z_j can vanish at most on a discrete set (with no finite limit points) and U_j is strictly decreasing on every ξ-interval.*

Proof. By the real-analyticity noted in Lemma 5.1, it is only possible for z_j to vanish on a set with a finite limit point if $z_j \equiv 0$; else $|z_j|^2 > 0$ almost everywhere and (5.5) implies U_j strictly decreasing. To have $z_j \equiv 0$ we must either have $u_j \equiv 0$ or $v_j \equiv 0$; suppose the former, so $|v_j| \equiv 1$. Since (5.4) would then give $0 \equiv u_j' = -z_{j-1} v_j$, this is only possible if also $z_{j-1} \equiv 0$. Similarly, $v_j \equiv 0$ would also require $z_{j-1} \equiv 0$. Induction on the index completes the proof. □

LEMMA 5.4. *For arbitrary initial data in $\mathcal{H}_1$, we have each $z_j \to 0$ and $\mathbf{u} \to \mathcal{S}$ as $\xi \to \infty$.*

Proof. Since U_J is nonincreasing by (5.5) and is obviously bounded below by zero, we must have $[|z_j|^2 + \delta^2 \Sigma_1^j |u_k \bar{v}_k|^2]$ integrable. As we know that u_j, v_j, z_j are bounded, this has a bounded derivative and so must go pointwise to zero. In particular, this immediately gives $z_j \to 0$. Each component of $\mathbf{u}$ lives in the (compact) unit sphere of $\mathbb{C}^2$ so, setting $\varphi(u, v) := |u\bar{v}|$, convergence $\varphi(u, v) \to 0$ implies that $(u, v) \to \varphi^{-1}(\{0\}) = \{(u, v) \in \mathcal{H}_1 : u = 0 \text{ or } v = 0\}$. Thus, $\mathbf{u} \to \mathcal{S} = \{\text{each } u_j = 0 \text{ or } v_j = 0\}$ as asserted. □

LEMMA 5.5. *Let $\mathbf{u}(\cdot)$ be a solution of (5.3) such that $U_J(\xi_*) < \delta$ for some $\xi_* \geq 0$. Then $\mathbf{u}(\tilde{\xi}) \in \mathcal{I}_0$ for each $\tilde{\xi} \geq 0$, where*

$$(5.6) \qquad \mathcal{I}_0 := \{\mathbf{u}^0 \in \mathcal{H}_1 : \text{ each } u_j \to 0 \text{ as } \xi \to \infty\}.$$

The set $\mathcal{I}_0$ is open in $\mathcal{H}_1$.

Proof. For each j, the assumption precludes having $|v_j| \to 0$ since U_J is nonincreasing and $|u_j| = 1$ implies $U_J \geq \delta$; by Lemma 5.4, this ensures $u_j \to 0$. By the definition, this gives $\mathbf{u}^0 \in \mathcal{I}_0$ and, of course, each $\mathbf{u}(\tilde{\xi} \in \mathcal{I}_0$. To see that $\mathcal{I}_0$ is open, consider any $\tilde{\mathbf{u}}^0 \in \mathcal{I}_0$ so each $\tilde{u}_j \to 0$. We may then find ξ_* such that $\tilde{U}_J(\xi_*) < \delta/2$. By Lemma 5.1 we have uniform continuity on bounded ξ-intervals for the dependence of solutions of (5.3) on the initial data, so there is a neighborhood of $\tilde{\mathbf{u}}^0$ giving $U_J(\xi_*) < \delta$. The first part of this proof then shows this neighborhood is in $\mathcal{I}_0$. □

THEOREM 5.6. *For arbitrary initial data in $\mathcal{H}_1$ we have convergence $\mathbf{u} \to \mathbf{u}^*$ (at an asymptotically exponential rate) as $\xi \to \infty$ for some steady state solution $\mathbf{u}^* \in \mathcal{S}$. Thus, for each $j = 1, \cdots, J$ we have either* Case 1: *$u_j \to 0$ and $v_j \to v_j^*$ (with $|v_j^*| = 1$) or* Case 2: *$v_j \to 0$ and $u_j \to u_j^*$ (with $|u_j^*| = 1$); in particular, we always have* Case 1 *for $j = 1$.*

Proof. We will proceed inductively in j, taking the system in the form (5.4) with the index j suppressed so

$$(5.7) \qquad \begin{cases} \text{(i)} \quad u' = -\delta(|v|^2 u + \zeta v), \\ \text{(ii)} \quad v' = \ \ \delta(|u|^2 v + \bar{\zeta} u), \end{cases}$$

and with the inductive assumption that we know

$$(5.8) \qquad |\zeta(\xi)| \le C e^{-\mu\delta\xi}$$

for $C = C_\mu$ and arbitrary $0 < \mu < 1$. By Lemma 5.4 we know that $u\bar{v} \to 0$ as $\xi \to \infty$, so we must be in one of the two possible cases—which we then consider separately to show the exponential decay rate for the appropriate component. With (4.4) and the inductive hypothesis on $\zeta = \zeta_j$, this completes the induction by giving the corresponding exponential decay for ζ_{j+1}. Returning to (5.3) with knowledge of exponential decay of z_j, integrability of the derivative gives existence of a specific limit for the nonvanishing component as well.

Case 1 $\boxed{u \to 0}$. Fix $\mu < 1$. Set $y := |u|^2$ so $y' = 2\,\mathrm{Re}\,\bar{u}u' = -2\delta[|v|^2 y + \mathrm{Re}\,\zeta\bar{u}v$, using (5.7.i) and note that $\mathrm{Re}\,\zeta\bar{u}v \le |\zeta||v|\sqrt{y} \le \varepsilon|v|^2 y + |\zeta|^2/4\varepsilon$. Now choose $0 < \varepsilon < 1 - \mu$ and use (5.8) with μ replaced by $\tilde{\mu} := \mu + \varepsilon < 1$. By (4.4), if $u \to 0$, then $|v| \to 1$ so (noting that $\mu/(1-\varepsilon) < 1$) there exists $\tilde{\xi}$ such that $(1-\varepsilon)|v(\xi)|^2 \ge \mu$ for $\xi \ge \tilde{\xi}$. Thus, we have

$$y'(\xi) \le -2\delta\mu y + (\delta C^2/2\varepsilon)e^{-2(\mu+\varepsilon)\delta\xi}$$

for $\xi \ge \tilde{\xi}$. The Gronwall inequality and some simple manipulation then give the desired estimate for $y = |u|^2$:

$$\begin{aligned} |u(\xi)|^2 &\le e^{-2\mu\delta(\xi-\tilde{\xi})}|u(\tilde{\xi})|^2 + \frac{\delta C^2}{2\varepsilon}\int_{\tilde{\xi}}^{\xi} e^{-2\mu\delta(\xi-\hat{\xi})}e^{-2(\mu+\varepsilon)\hat{\xi}}\,d\hat{\xi} \\ &\le \left[e^{2\mu\delta\tilde{\xi}}|u(\tilde{\xi})|^2 + \frac{\delta C^2}{2\varepsilon}\int_{\tilde{\xi}}^{\infty} e^{-2\varepsilon\hat{\xi}}\,d\hat{\xi}\right]e^{-2\mu\delta\xi} \\ &=: \tilde{C}^2 e^{-2\mu\delta\xi} \end{aligned}$$

for $\xi \ge \tilde{\xi}$; this also applies to all $\xi \ge 0$ with a modification of the $\tilde{C}$.

Case 2 $\boxed{v \to 0}$. Again, fix $\mu < 1$ and now set $y := |v|^2$ and choose $0 < 2\varepsilon < 1 - \mu$. Much as above, we get

$$y'(\xi) \ge 2\delta(\mu - \varepsilon)y - (\delta C^2/2\varepsilon)e^{-2\mu\delta\xi}$$

for ξ large enough ($\xi \ge \tilde{\xi}_0$) that $(1-\varepsilon)|u(\xi)|^2 \ge \mu - \varepsilon$. Applying the (reversed)

Gronwall inequality we then obtain, for any $\xi > \tilde{\xi} \geq \tilde{\xi}_0$,

$$\begin{aligned} y(\xi) &\geq e^{2(\mu-\varepsilon)\delta(\xi-\tilde{\xi})} y(\tilde{\xi}) - \frac{\delta C^2}{2\varepsilon} \int_{\tilde{\xi}}^{\xi} e^{2(\mu-\varepsilon)\delta(\xi-\hat{\xi})} e^{-2\mu\hat{\xi}}\, d\hat{\xi} \\ &= e^{2(\mu-\varepsilon)\delta(\xi-\tilde{\xi})} \left[y(\tilde{\xi}) - \left(\frac{\delta C^2}{2\varepsilon} \int_{\tilde{\xi}}^{\xi} e^{-2\varepsilon\delta(\hat{\xi}-\tilde{\xi})}\, d\hat{\xi} \right) e^{-2\mu\delta\tilde{\xi}} \right] \\ &\geq e^{2(\mu-\varepsilon)\delta(\xi-\tilde{\xi})} \left[y(\tilde{\xi}) - \frac{C^2}{4\varepsilon} e^{-2\mu\delta\tilde{\xi}} \right]. \end{aligned}$$

It follows that we must have $|v(\tilde{\xi})| \leq (C/2\varepsilon)e^{-\mu\delta\tilde{\xi}}$ for all $\tilde{\xi} \geq \tilde{\xi}_0$ (hence, for all $\tilde{\xi} \geq 0$ with a modified coefficient) or the Gronwall estimate would give $y(\xi) \to \infty$, contradicting the case that $v \to 0$.

In view of Lemmas 5.3, 5.4, and 5.5 and the experience with computational experimentation with a variety of sets of initial data, it is plausible to conjecture that one always has $u_j \to 0$ as $\xi \to \infty$ unless $z_k \equiv 0$ for each $k = 1, \cdots, j$. This is false!

THEOREM 5.7. *Let* $\mathbf{u}^*$ *be any stationary solution such that* $u_j = 0$ *for some* $j < J$. *Then there is a nonstationary solution* $\mathbf{u}$ *such that* $\mathbf{u}(\xi) \to \mathbf{u}^*$ *as* $\xi \to \infty$.

Proof. For each j we have either:

Case 1 $\boxed{u_j^* = 0 \text{ so } |v_j^*| = 1}$. Set

$$u_j = v_j^* \sin\alpha_j, \quad v_j = v_j^* \cos\alpha_j, \quad \chi_j := -1$$

so $(v_j^* \cos\alpha_j)\alpha_j' = u_j' = -z_j v_j^* \cos\alpha_j$ or:

Case 2 $\boxed{v_j^* = 0 \text{ so } |u_j^*| = 1}$. Set

$$u_j = u_j^* \cos\alpha_j, \quad v_j = u_j^* \sin\alpha_j, \quad \chi_j := +1$$

so $(u_j^* \cos\alpha_j)\alpha_j' = v_j' = \overline{z_j} u_j^* \cos\alpha_j$.

In either case, we have $u_j\overline{v_j} = \frac{1}{2}\sin 2\alpha_j$ whence, after scaling ξ for convenience to permit the omission of a factor δ from our definition of z_j, we must have

$$\text{(5.9)} \qquad \alpha_j' = \chi_j z_j, \qquad z_j = \Sigma_{k=1}^{j} \tfrac{1}{2} \sin 2\alpha_k$$

(provided $\cos 2\alpha_j \neq 0$) for $j = 1, \cdots, J$. Note that by restricting our attention here to the real case, we have isolated stationary solutions: each $\alpha_j = \pm\nu\pi$. Our notation ensures that any nonstationary solution $\alpha = [\alpha_1, \cdots, \alpha_J]$ of (5.9) for which $\alpha(\xi) \to 0$ corresponds to a nonstationary solution $\mathbf{u}$ of (5.4) for which $\mathbf{u} \to \mathbf{u}^*$ as $\xi \to \infty$. The linearization of (5.9) around $\alpha = 0$ is just

$$\text{(5.10)} \qquad \alpha' = \mathbf{A}\alpha \quad \text{with } \mathbf{A} := \begin{pmatrix} \chi_1 & 0 & 0 & \cdots \\ \chi_2 & \chi_2 & 0 & \cdots \\ \chi_3 & \chi_3 & \chi_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Since $\mathbf{A}$ is lower triangular, we see that its eigenvalues (with multiplicity) are the diagonal elements $\{\chi_1, \cdots, \chi_J\}$ and if any of these is -1, we have a corresponding eigenvector giving an exponentially decaying solution. Since there are no purely imaginary eigenvalues, the linearized problem (5.10) gives the local splitting into stable and unstable manifolds whence the nonlinear problem (5.9) must also have an exponentially decaying solution near zero—asymptotically behaving precisely like the solution of (5.10). □

Remark 5.8. While we have carried through the analysis for the discretization corresponding to (5.2), we could equally well have considered a trapezoidal rule approximation to (4.6):

$$z_j := \delta\Sigma_1^j(u_{k-1}\overline{v_{k-1}} + u_k\overline{v_k})/2. \tag{5.11}$$

With trivial modification, we would then have obtained for that setting the same results obtained above. Indeed, (5.11) gives the *identical* system (5.4) if we were to scale ξ by 2 and set $\zeta_j := 2\sum_{k=1}^{j-1} u_k\overline{v_k}$, instead.

6. Stationary solutions. We now introduce the set $\mathcal{S}$ of all stationary solutions. Note that $\mathbf{u} \in \mathcal{S}$ means $\mathbf{u}' \equiv 0$ so $z \equiv 0$ on $\mathbb{R}$. By (2.2), this corresponds to having $u\overline{v} = 0$ almost everywhere; note from this that $\mathcal{S}$ is entirely independent of γ. It is easy to see that $\mathcal{S}$, $\mathcal{S}_0 := \{\mathbf{u} \in \mathcal{S} : u = u_0 \equiv 0\}$ and $\mathcal{S}\backslash\mathcal{S}_0 := \{\mathbf{u} \in \mathcal{S} : u_0 \not\equiv 0\}$ are each uncountable arc-wise connected sets in $\mathcal{H}_1$—even if we were to restrict attention to the "purely real" case, taking $\mathcal{H} := L^2(\mathbb{R} \to \mathbb{R}^2)$, or to factor out the action of the group $\mathcal{G} := \{g_\vartheta : \mathbf{u} \mapsto e^{i\vartheta(\tau)}\mathbf{u}\}$.

From the characterization $u\overline{v} \equiv 0$ we observe that, for each $\mathbf{u} \in \mathcal{S}$, we can partition[3] $\mathbb{R}$, independently of ξ in view of (1.4), as a disjoint union $\mathcal{A} \cup \mathcal{B}$ such that

$$\begin{cases} |u| = K,\ v = 0 & \text{on } \mathcal{A}, \\ u = 0,\ |v| = K & \text{on } \mathcal{B}. \end{cases} \tag{6.1}$$

We now wish to consider the linearization around a stationary solution $\mathbf{u}^* \in \mathcal{S}$ so that $z^* \equiv 0$ in (3.3). We fix $\mathcal{A}, \mathcal{B}$, and $K(\cdot)$ corresponding to $\mathbf{u}^*$ and note that (3.3) now gives

$$u' = -v^*z = \begin{cases} 0 & \text{on } \mathcal{A}, \\ -v^*z & \text{on } \mathcal{B}, \end{cases} \qquad \overline{v}' = \overline{u^*}z = \begin{cases} \overline{u^*}z & \text{on } \mathcal{A}, \\ 0 & \text{on } \mathcal{B}, \end{cases} \tag{6.2}$$

$$e^{\gamma\tau}z = \int_{-\infty}^{\tau} e^{\gamma\tilde{\tau}} \left\{ \begin{matrix} u^*\overline{v} & \text{on } \mathcal{A} \\ \overline{v^*}u & \text{on } \mathcal{B} \end{matrix} \right\} d\tilde{\tau}.$$

Now introduce

$$K^2 w := \begin{cases} -e^{\gamma\tau}u^*\overline{v} \quad (\text{so } v = -e^{-\gamma\tau}u^*\overline{w}) & \text{on } \mathcal{A}, \\ e^{\gamma\tau}\overline{v^*}u \quad (\text{so } u = e^{-\gamma\tau}\overline{v^*}w) & \text{on } \mathcal{B} \end{cases} \tag{6.3}$$

with w itself irrelevant, where $K = 0$; e.g., we may set $w := 0$ there. Then, noting that $\mathbf{u}^*$ is independent of ξ with $|v^*| = K$ on $\mathcal{A}$ and $|u^*| = K$ on $\mathcal{B}$, (6.2) gives (where $K \neq 0$)

[3] Any τ for which $K(\tau) = 0$ so $u = v = 0$ may arbitrarily be assigned either to $\mathcal{A}$ or to $\mathcal{B}$; the partition is unique to within $d\sigma$-nullsets.

$$(6.4)\qquad w' = \frac{1}{K^2}\left\{\begin{matrix}(-e^{\gamma\tau}u^*)(\overline{u^*}z) & \text{on } \mathcal{A}\\ (e^{\gamma\tau}\overline{v^*})(-v^*z) & \text{on } \mathcal{B}\end{matrix}\right\}$$

$$= -e^{\gamma\tau}z = \int_{-\infty}^{\tau}\left\{\begin{matrix}-e^{\gamma\tau}u^*\overline{v} & \text{on } \mathcal{A}\\ -e^{\gamma\tau}\overline{v^*}u & \text{on } \mathcal{B}\end{matrix}\right\}d\tilde{\tau}.$$

If we define $\chi : \mathbb{R} \to \{\pm 1\}$ by

$$(6.5)\qquad \chi := \{+1 \text{ on } \mathcal{A};\ -1 \text{ on } \mathcal{B}\}$$

and change variables to think of w as a function of (ξ, σ), using (2.1) so $K^2 d\tilde{\tau} = d\tilde{\sigma}$, then (6.4) takes the simple form

$$(6.6)\qquad w' = \int_0^{\sigma} \chi w \, d\tilde{\sigma}$$

or, equivalently,

$$(6.7)\qquad w_{\xi\sigma} = \chi w \quad \left(= \left[-e^{-\gamma\tau}z\right]_{\sigma}\right) \quad w(0,\cdot) = w_0 \in L^1_{d\sigma}.$$

Note that this formulation omits the irrelevant values of w for the set of τ where $K(\tau) = 0$, which disappears when we write things in terms of σ. Since the operator: $w \mapsto \int_0^{\sigma} \chi w \, d\tilde{\sigma}$ is certainly bounded, the solution operator $\mathbf{S}_{\xi} : w_0 \mapsto w(\xi, \cdot)$ for (6.6) forms a group on L^1. Alternatively, we might note that Theorem 3.4 ensures integrability of $K^2 w$ with respect to τ and so integrability of w with respect to σ.

We now restate the results of this discussion as a lemma without further proof.

LEMMA 6.1. *Let $\mathbf{u}^* \equiv \mathbf{u}_0^*$ be given in $\mathcal{S}$, determining $K(\cdot)$, $\sigma(\cdot)$ as in* (2.1) *and χ as in* (6.5); *let $\mathbf{u}$ be a linearized perturbation, obtained from* (3.3), *corresponding to a perturbation $\mathbf{u}_0 \in \mathcal{H}$ of the initial conditions $\mathbf{u}_0^*$. Then,*

$$(6.8)\qquad u = \left\{\begin{matrix}u_0,\\ e^{-\gamma\tau}v^*w,\end{matrix}\right. \qquad v = \left\{\begin{matrix}-e^{-\gamma\tau}u^*\overline{w} & \textit{on } \mathcal{A},\\ v_0 & \textit{on } \mathcal{B},\end{matrix}\right.$$

where w, viewed as a function of (ξ, σ), satisfies (6.7) *and has initial data w_0 at $\xi = 0$ given by*

$$K^2 w_0 := \left\{\begin{matrix}-e^{\gamma\tau}u^*\overline{v_0} & \textit{on } \mathcal{A},\\ e^{\gamma\tau}\overline{v^*}u_0 & \textit{on } \mathcal{B}.\end{matrix}\right.$$

Note that w_0 is necessarily integrable with respect to σ.

LEMMA 6.2. *When χ is constant $(\chi \equiv \pm 1)$, the solution of* (6.7) *is given explicitly by*

$$(6.9)\qquad w(\xi, \sigma) := w_0(\sigma) + \xi \int_0^{\sigma} w_0(\tilde{\sigma})\Psi'(r)\, d\tilde{\sigma},$$

where we set $r := \xi[\sigma - \tilde{\sigma}]$ and have

$$(6.10)\qquad \Psi(r) := \left\{\begin{matrix}J_0(2\sqrt{r}) & \textit{for } \chi \equiv -1 \ (\mathbf{u}^* \in \mathcal{S}_0),\\ I_0(2\sqrt{r}) & \textit{for } \chi \equiv +1,\end{matrix}\right.$$

where J_0 is the usual Bessel function of first kind and I_0 is the modified Bessel function: $I_0(s) := J_0(is)$.

Proof. We begin with the *ansatz* that $e^{-\gamma\tau}z$ can be obtained from w_0 by a convolution with respect to σ:

$$-e^{-\gamma\tau}z = \chi w_0 * \Psi(\xi\cdot) := \chi\int_0^\sigma w_0(\tilde\sigma)\Psi(r)\,d\tilde\sigma$$

with $r := \xi[\sigma - \tilde\sigma]$. Differentiating with respect to σ then gives[4] (6.9)—provided we require $\Psi(0) = 1$ so as to have the correct condition at $\xi = 0$. Next, differentiating first with respect to ξ and then with respect to σ gives

$$\begin{aligned} w_\xi &= \int_0^\sigma w_0(\tilde\sigma)\Psi'(r)\,d\tilde\sigma + \int_0^\sigma w_0(\tilde\sigma)[r\Psi''(r)]\,d\tilde\sigma \\ &= \int_0^\sigma w_0(\tilde\sigma)[r\Psi'(r)]'\,d\tilde\sigma, \\ w_{\xi\sigma} &= w_0(\sigma)[r\Psi'(r)]'\,|_{r=0} + \xi\int_0^\sigma w_0(\tilde\sigma)[r\Psi'(r)]''\,d\tilde\sigma \\ &= \chi\left[w_0(\sigma) + \xi\int_0^\sigma w_0(\tilde\sigma)\Psi'(r)\,d\tilde\sigma\right] \end{aligned}$$

with the final equality coming from (6.7). We obtain this last, provided that

$$[r\Psi'(r)]'\,|_{r=0} = \chi, \qquad [r\Psi'(r)]'' = \chi\Psi'(r).$$

The differential equation gives $([r\Psi'(r)]' - \chi\Psi(r)) =$ constant and evaluating at $r = 0$ shows this constant must be zero. If we now set $\Psi(r) =: \Phi(s)$ with $s := 2\sqrt{r}$, this gives

$$s^2\Phi'' + s\Phi' - \chi s^2\Phi = 0, \qquad \Phi(0) = 1.$$

For $\chi = -1$, this is Bessel's equation with parameter $a = 0$ and the normalization gives $\Phi(s) = J_0(s)$; for $\chi = +1$, we then get $\Phi(s) = J_0(is) =: I_0(s)$.

Note that $d\Psi(r)/dr = \Phi'(2\sqrt{r})/\sqrt{r}$ so, since

$$J_0'(s) = -J_1(s) \qquad I_0'(s) = iJ_0'(is) = -iJ_1(is) =: I_1(s),$$

we have

$$\Psi'(r) = \begin{cases} -J_1(2\sqrt{r})/\sqrt{r} & \text{for } \chi = -1, \\ I_1(2\sqrt{r})/\sqrt{r} & \text{for } \chi = +1 \end{cases} \tag{6.11}$$

for use in (6.9). Since $J_0(z)$ is an *even* analytic function of z, it is also analytic in $\sqrt{z}$ so Ψ is analytic (a fortiori bounded) near zero and so on any bounded interval. Thus, (6.9) makes sense for all $d\sigma$-integrable w_0. □

Remark 6.3. What information can we draw from this in the case of $\mathbf{u}^* \in \mathcal{S}_0$ (so $\chi \equiv -1$)? We observe, first, that the integral term in (6.9) is a convolution so, using L^1 norms,

$$\|w(\xi,\cdot)\| \le \|w_0\|\,[1 + \|\xi\Psi'\|]\,.$$

[4] It is at this point already that we need the assumption: $\chi \equiv$ constant.

Here we have, setting $s^2 = \xi\sigma$,

$$\|\xi\Psi'\| := \int_0^{\kappa^2} |\Psi'(\xi\sigma)|\,\xi d\sigma = 2\int_0^{\kappa\sqrt{\xi}} |J_1(2s)|\,ds,$$

and since $J_1(s)$ decays like $1/\sqrt{s}$, we see that we have $\|\xi\Psi'\| = \mathcal{O}(\xi^{1/4})$ so, at worst, we always have

$$\|w(\xi,\cdot)\|_{L^1} = \mathcal{O}(\xi^{1/4}) \quad \text{as } \xi \to \infty. \tag{6.12}$$

On the other hand, if we are considering perturbations for which $w_0 \in BV$ (bounded variation) so that it is justifiable to integrate by parts in (6.9), then we obtain

$$w(\xi,\sigma) = w_0(0)J_0(2\sqrt{\xi\sigma}) + \int_0^{\sigma} J_0(2\sqrt{\xi(\sigma-\tilde{\sigma})})\,dw_0(\tilde{\sigma}). \tag{6.13}$$

The integral term on the right can be estimated in the same way as for (6.12) to give $\mathcal{O}(\xi^{-1/4})$ decay. The first term goes to zero pointwise in σ as $\xi \to 0$ at a rate $\mathcal{O}(\xi^{-1/4})$. This is not uniform, but can certainly be integrated in σ to give a decay rate

$$\|w(\xi,\cdot)\|_{L^1} = \mathcal{O}(\xi^{-1/4}) \quad \text{as } \xi \to \infty \tag{6.14}$$

in this case.

Clearly, even as "linearized stability," this is far weaker than the results we have for the discretized setting, corresponding to Lemma 5.5 and Theorem 5.6. Nevertheless, it is as strong a result as we have been able to obtain here. That the result is weaker is certainly related to the fact that $\mathcal{S}_0$ is not isolated from other stationary solutions, but we might still hope to improve this. Indeed, if (6.12) could be improved to give boundedness as $\xi \to \infty$ for each initial $w_0 \in L^1(0,\kappa^2)$, then a simple argument would show decay for all w_0. So far we do not know whether this is true and, further, note that this linearized stability by itself would not show $\mathbf{u} \to \mathcal{S}_0$ (locally) for the nonlinear problem.

Remark 6.4. We next consider the case of $\mathbf{u}^* \in \mathcal{S}\backslash\mathcal{S}_0$ so $\mathcal{A}$ is nonempty. We are seeking here to demonstrate instability, so it is only necessary to construct special examples. Suppose $\mathcal{A}$ contains an interval $[\sigma_-,\sigma_+]$ and we take the perturbation $\mathbf{u}_0$ so $w_0 \equiv 1$ on this interval and vanishes otherwise. This effectively lets us take $\sigma_- = 0$ with no loss of generality. Thus, at least for σ in the interval, we are considering (6.7) with $\chi \equiv +1$ and Lemma 6.2 applies. Integrating by parts as for (6.13), we then have

$$w(\xi,\sigma) = I_0(2\sqrt{\xi\sigma}) \quad \text{for } 0 < \sigma < \sigma_+$$

and, since I_0 grows exponentially, we have instability: growth of $w(\xi,\cdot)$ which is exponential in $\xi^{1/2}$. We remark that this is consistent with having $\|w\| = \mathcal{O}(e^{\varepsilon\xi})$ for arbitrarily small $\varepsilon > 0$, as is suggested by the fact that the Volterra operator on the right of (6.6) has spectrum $\{0\}$.

What does this tell us for the nonlinear problem? Using the fact that, as observed in Remark 4.3, the solution map has a locally bounded second derivative we can show that *for arbitrarily large M and arbitrarily small $\varepsilon > 0$ there exist solutions of* (2.5) *which initially differ from $\mathbf{u}^*$ by less than ε but at a later time differ by more than $M\varepsilon$, assuming $M\varepsilon$ is not too big.*

Again, as in Remark 6.3, this is not very much. It certainly is enough, however, to guarantee that the exponential factor in (3.2) cannot be omitted to give the Lipschitz continuity uniform in ξ.

7. Asymptotic behavior. In this section we consider the asymptotic behavior of solutions of the system (2.5) as $\xi \to \infty$. We will assume that the reduction of §4 has been made, if necessary, so in (2.1) we have $K(\tau) \equiv 1$ for $\tau \in [0,1]$ with everything vanishing for $\tau \notin [0,1]$ whence we have (4.5) and (4.4). While we might conjecture for this context essentially the same results which we obtained for the discretized system in the previous section, we have so far not been able to carry out this program completely. Our results here are primarily the consequences of (4.1).

Remark 7.1.

DEFINITION. A function $\mathbf{v} : \mathbb{R}_+ \to \mathcal{X}$ will be called **recurrent** (for some ξ_0) if there is a sequence $\xi_n \to \infty$ for which $\|\mathbf{v}(\xi_n) - \mathbf{v}(\xi_0)\| \to 0$.

Clearly, every periodic function is recurrent for arbitrary ξ_0. Following Bohr, $\mathbf{v}$ is *almost periodic* on $\mathbb{R}_+$ if, for any $\varepsilon > 0$, there exists $\ell(\varepsilon)$ such that (with $\tilde{\xi}$ arbitrary) $\|\mathbf{v}(\xi) - \mathbf{v}(\xi_\varepsilon)\| \le \varepsilon$ for some $\tilde{\xi} < \xi_\varepsilon < \tilde{\xi} + \ell(\varepsilon)$ and all $\xi \in \mathbb{R}_+$; clearly, this also implies recurrence for arbitrary ξ_0. This would include sums of (incommensurately) periodic functions: if $\mathbf{v} := \sum_j \mathbf{v}_j$ (where each $\mathbf{v}_j$ is continuous with period ξ_j and $\sum_j \sup_\xi |\mathbf{v}_j|$ convergent); then it is easily seen to be almost periodic, using the number-theoretic result that we can always find positive integers q and $\{n_j\}$, making $|q - n_j\xi_j|$ arbitrarily small simultaneously for $j = 1, \cdots, J$ [6, Thm. 201] every "positive ray" $q[1/\xi_1, \cdots, 1/\xi_J]$ passes arbitrarily close to integer lattice points in $\mathbb{R}_+^J$ for infinitely many integers q].

Finally, we note that, for $\mathbf{v}$ satisfying an autononomous ODE, if we were to have $\|\mathbf{v}(\xi_n) - \mathbf{v}(\xi_0)\| \to 0$ for any sequence $\{\xi_n\}$ bounded away from ξ_0, then we would have recurrence at ξ_0. To see this when $\{\xi_n\}$ is bounded, extract a subsequence converging to some $\xi_1 \ne \xi_0$ and observe that continuity gives $\mathbf{v}(\xi_1) = \mathbf{v}(\xi_0)$ whence, assuming uniqueness for the ODE, $\mathbf{v}$ would necessarily be periodic with period $|\xi_1 - \xi_0|$.

THEOREM 7.2. *Under the hypotheses of Lemma* 4.1 *and taking the system reduced as in the previous section, if* $\mathbf{u}$ *is recurrent (for some* ξ_0*), then it is stationary:* $z \equiv 0$ *so* $\mathbf{u}$ *independent of* ξ.

Proof. By (4.1.ii), for each τ and for $\xi_n - \xi_0 \ge \delta > 0$ we have

$$\int_{\xi_0}^{\xi_0+\delta} |z(\cdot,\tau)|^2 \le \int_0^1 |v(\xi_n,\cdot)|^2 - \int_0^1 |v(\xi_0,\cdot)|^2 \le \sqrt{2\tau}\|\mathbf{u}(\xi_n,\cdot) - \mathbf{u}(\xi_0,\cdot)\| \to 0,$$

which gives $z \equiv 0$ on $[\xi_0, \xi_0 + \delta] \times [0,1]$ — and so everywhere, as in Corollary 4.2, giving stationarity of $\mathbf{u}$ as asserted. □

THEOREM 7.3. *Under the hypotheses of Lemma* 4.1 *and taking the system reduced as in the previous section, we always have* $z \in L^2(\mathbb{R}_+ \times [0,1])$ *and uniform convergence:* $z(\xi,\cdot) \to 0$ *as* $\xi \to \infty$.

Proof. It is convenient to set

$$U(\cdot,\tau) := \int_0^\tau |u|^2, \qquad V(\cdot,\tau) := \int_0^\tau |v|^2. \tag{7.1}$$

From (4.4) we see that $0 \le U, V \le \tau$ and, also using (4.1), we see that each is uniformly Lipschitzian (jointly in ξ, τ) with $U' = -|z|^2$, $V' = |z|^2$ so U is nonincreasing in ξ and V nondecreasing. For each (fixed) τ, we have

$$\begin{aligned}\int_\xi^{\tilde{\xi}} |z(\cdot,\tau)|^2 &= U(\xi,\tau) - U(\tilde{\xi},\tau) \le U(\xi,\tau) \le \tau \\ &= V(\tilde{\xi},\tau) - V(\xi,\tau) \le V(\tilde{\xi},\tau) \le \tau.\end{aligned}$$

So, taking $\xi = 0$ and letting $\tilde{\xi} \to \infty$, we see that we have $\|z(\cdot,\tau)\| \le \sqrt{\tau}$ (this is the $L^2(\mathbb{R}_+)$-norm) whence $\|z\| \le 1/\sqrt{2}$. (Here, this is the $L^2(\mathbb{R}_+ \times [0,1])$-norm.) Using the bounds $0 \le |u|^2, |v|^2 \le 1$ in (4.1.iii), we have $|z'| \le \sqrt{\tau}\|z\|$ so $\|z'\| \le \|z\|/\sqrt{2}$; thus

$$\left|\left(\|z\|^2\right)'\right| = 2\left|\operatorname{Re}\int_0^1 \bar{z}z'\right| \le \sqrt{2}\,\|z\|^2.$$

Then $(\|z\|^2)'$ is integrable on $\mathbb{R}_+$ whence $\|z\|^2$ has a limit as $\xi \to \infty$, necessarily zero almost everywhere in τ since $z \in L^2(\mathbb{R}_+ \times [0,1])$. This gives $z(\xi,\cdot) \to 0$ in $L^2(0,1)$-norm. Since $z(\xi,\cdot)$ stays in a compact subset of $C[0,1]$, as we have noted in the previous section, this is actually uniform convergence. □

Remark 7.4. Since the set $\mathcal{S}$ of stationary solutions is characterized by having $z \equiv 0$, it would be tempting to conclude, from the uniform convergence above, that it is a global attractor, i.e., that we must necessarily have $\mathbf{u} \to \mathcal{S}$; indeed, computation suggests the stronger conjecture that we have $\mathbf{u} \to \mathcal{S}_0$ $(u \to 0)$ as $\xi \to \infty$ for all $\mathbf{u}_0 \in \mathcal{U}_1$ — except, of course, $\mathcal{S}\backslash\mathcal{S}_0$. We do note, however, that this cannot follow simply from Theorem 7.3 since, e.g., there exist sequences like $\mathbf{u}_j(\tau) := (\cos j\pi\tau, \sin j\pi\tau)$ for which we have $|z_j(\tau)| \le 1/4\pi j \to 0$, but $\|\mathbf{u}_j - \mathcal{S}\| = 2\sqrt{\pi - 2} \not\to 0$. (Of course, this example has *not* been constructed by taking $\mathbf{u}_j := \mathbf{u}(\xi_j)$ with $\xi_j \to \infty$ for a solution $\mathbf{u}$ of (2.5).)

REFERENCES

[1] R. L. Carman, F. Shimizu, C. S. Wang, and N. Bloembergen, *Theory of Stokes pulse shapes in transient stimulated Raman scattering*, Phys. Rev. A, 2 (1970), pp. 60–72.

[2] F. Y. F. Chu and A. C. Scott, *Inverse scattering transform for wave-wave scattering*, Phys. Rev. A, 12 (1975), pp. 2060–2064.

[3] K. Drühl, R. G. Wenzel, and J. L. Carlsten, *Observation of solitons in stimulated Raman scattering*, Phys. Rev. Lett., 51 (1983), pp. 1171–1174.

[4] M. D. Duncan, R. Mahon, L. L. Tankersley, and J. Reintjes, *Transient stimulated Raman amplification in hydrogen*, J. Opt. Soc. Amer. B, 5 (1988), pp. 37–52.

[5] M. D. Duncan, R. Mahon, L. L. Tankersley, G. Hilfer, and J. Reintjes, *Phase pulling in transient Raman amplifiers*, J. Opt. Soc. Amer. B, 7 (1990), pp. 202–210.

[6] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Third ed., Clarendon Press, Oxford, 1954.

[7] G. Hilfer and C. R. Menyuk, *Stimulated Raman scattering in the transient limit*, J. Opt. Soc. Amer. B, 7 (1990), pp. 739–749.

[8] D. J. Kaup, *The method of solution for stimulated Raman scattering and two-photon propagation*, Phys. D, 6 (1983), pp. 143–154.

[9] ———, *Creation of a soliton out of dissipation*, Phys. D, 19 (1986), pp. 125–134.

[10] D. J. Kaup and C. R. Menyuk, *A model initial value problem in stimulated Raman scattering*, Phys. Rev. A, 42 (1990), pp. 1712–1717.

[11] G. Landsberg and L. Mandelstam, *Eine neue Erscheinung bei der Lichtzerstreuung in Krystallen*, Die Naturwissenshaften, 28 (1928), pp. 557–558.

[12] C. R. Menyuk, *Transient solitons in stimulated Raman scattering*, Phys. Rev. Lett., 62 (1989), pp. 2937–2940.

[13] C. R. Menyuk and G. Hilfer, *Asymptotic evolution of transient pulses undergoing stimulated Raman scattering*, Opt. Lett., 14 (1989), pp. 227–229.

[14] C. V. Raman and K. S. Krishnan, *A new type of secondary radiation*, Nature, 121 (1928), pp. 501–502.

[15] H. Steudel, *Solitons in stimulated Raman scattering and resonant two-photon propagation*, Phys. D, 6 (1983), pp. 155–178.

[16] C.-S. Wang, *Theory of stimulated Raman scattering*, Phys. Rev., 182 (1969), pp. 482–494.

[17] R. G. Wenzel, J. L. Carlsten, and K. J. Drühl, *Soliton experiments in stimulated Raman scattering*, J. Statist. Phys., 39 (1985), pp. 621–632.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 364–386, March 1992

BUNSEN FLAMES AS STEADY SOLUTIONS OF THE KURAMOTO–SIVASHINSKY EQUATION*

DANIEL MICHELSON†

Abstract. Rotationally invariant steady solutions of the Kuramoto–Sivashinsky equation in two space dimensions are studied. Specifically, conical solutions, i.e., solutions which tend to a constant slope at infinity, are sought after. These solutions in combustion theory have a physical interpretation as the Bunsen flames. The technique of rigorous estimates by computer is used to prove the existence of such solutions for all physically reasonable values of the slopes.

Key words. Bunsen flames, Kuramoto–Sivashinsky equation, interval arithmetic, computer proofs

AMS(MOS) subject classifications. 34A34, 34A45, 35A50

1. Introduction.

The well-known Kuramoto–Sivashinsky equation is

$$u_t + \nabla^4 u + \nabla^2 u + \tfrac{1}{2}\,|\nabla u|^2 = 0, \qquad u = u(x,t). \tag{1.1}$$

In the context of the combustion theory, the function $u(x,t)$ represents the perturbation of the plane flame front that propagates in a fuel-oxygen mixture (e.g., see [6]).

In [5] we studied the steady solutions of this equation in one space dimension, i.e., solutions of the form

$$u(x,t) = -c^2 t + v(x), \qquad x \in R^1. \tag{1.2}$$

It was proved analytically that this equation has for small c periodic and quasi-periodic solutions and for large c a unique conical solution. For intermediate values of c it was shown numerically that besides the conical and periodic solutions there is, probably, a Cantor set of chaotic solutions. In this paper we consider equation (1.1) in the plane $x \in R^2$ and look for steady solutions of the form (1.2) with rotational symmetry, i.e., independent of the angle φ. The corresponding function $v(r)$ of the radius r satisfies the equation

$$\left(D^2 + \frac{1}{r}D\right)^2 v + \left(D^2 + \frac{1}{r}D\right) v = c^2 - \frac{1}{2}\left(Dv\right)^2, \qquad 0 \le r < \infty, \tag{1.3}$$

where D is the operator of differentiation with respect to r. With Dv denoted by y, the equation is reduced to a third-order nonautonomous ordinary differential equation (O.D.E.)

$$\left(D^2 + \frac{1}{r}D\right)\left(D + \frac{1}{r}\right) y + \left(D + \frac{1}{r}\right) y = c^2 - \frac{y^2}{2}, \qquad 0 \le r < \infty. \tag{1.4}$$

As $r \to \infty$ this equation approaches the one-dimensional limit

$$y''' + y' = c^2 - \frac{y^2}{2}, \quad y = y(x), \quad -\infty < x < +\infty. \tag{1.5}$$

* Received by the editors April 17, 1989; accepted for publication (in revised form) November 7, 1990.

† Weizmann Institute, Department of Applied Mathematics and Computer Science, Rehovot, Israel. (email: mtmiche@weizmann.weizmann.ac.il)

Our ultimate goal is to study the set of all bounded solutions of (1.4). In this paper, however, we restrict ourselves only to so-called Bunsen flame profiles; namely, (1.1) with the right-hand side replaced by c^2 models a flame front generated by a circular gas burner. The constant c is proportional to the slope of the flame cone and is determined by the gas influx. A stationary radial solution of such a problem will obviously satisfy (1.4) with boundary condition

$$(1.6) \qquad y(0) = 0, \qquad \lim_{r \to +\infty} y(r) = -c\sqrt{2}.$$

The condition $y(0) = 0$ follows from the request that u is a weak solution of (1.1). Our numerical experiments with the partial differential equation (P.D.E.) (1.1) showed that the above stationary solutions of (1.1) are stable only for large values of c, namely, $c > 3$. For large c it is convenient to switch to the variables

$$(1.7) \qquad y_{\text{new}} = \frac{y_{\text{old}}}{(c\sqrt{2})}, \qquad r_{\text{new}} = \frac{r_{\text{old}} \cdot c^{1/3}}{2^{1/6}},$$

so that the new function $y(r)$ satisfies

$$(1.8) \qquad \left(D^2 + \frac{1}{r}D\right)\left(D + \frac{1}{r}\right)y + \alpha\left(D + \frac{1}{r}\right)y + y^2 - 1 = 0, \qquad 0 < r < \infty,$$

where

$$(1.9) \qquad \alpha = (\sqrt{2}/c)^{2/3}.$$

Our problem is now stated as follows: for small α prove the existence of bounded solutions of (1.8) that satisfy

$$(1.10) \qquad \lim_{r \to \infty} y(r) = -1.$$

On Fig. 1(a)–(c) one can see the graph of the solution $y(r)$ for $\alpha = 0$ in the planes yy', ry, and rv respectively. These graphs were obtained by a nonrigorous numerical experiment.

Unlike the autonomous equation (see [5]) the theoretical study of (1.8) runs into considerable difficulties. Because of the singularity at $r = 0$ the flow defined by (1.8) cannot be studied by topological methods. In addition, (1.8) does not possess a Lyapunov function even when $\alpha = 0$.

Lately a new approach of computer-assisted proofs in mathematics has come into being (e.g., see [1]–[4]). The idea is to do computations in interval arithmetic so that computer produces rigorous bounds of the true result. This idea could be used in solving O.D.E.'s on a bounded interval (e.g., see [3]). When the interval is unbounded, an asymptotic expansion should be constructed at infinity and bounds of the truncation error should be analytically derived. The main analytical difficulty is to maximize the domain of validity of the asymptotic expansion in order to reduce the finite interval on which the O.D.E. is solved numerically in interval arithmetic. Another problem is to stabilize the algorithm of numerical solution to prevent rapid exponential growth of the error.

Rewrite (1.8) as an autonomous system in R^4:

$$(1.11) \qquad d\bar{y}/dr = f(\bar{y}) = (y_2, y_3, 1 - y_1^2 - \alpha y_2 - 2y_3y_4 + y_2y_4^2 - y_1y_4^3, -y_4^2),$$

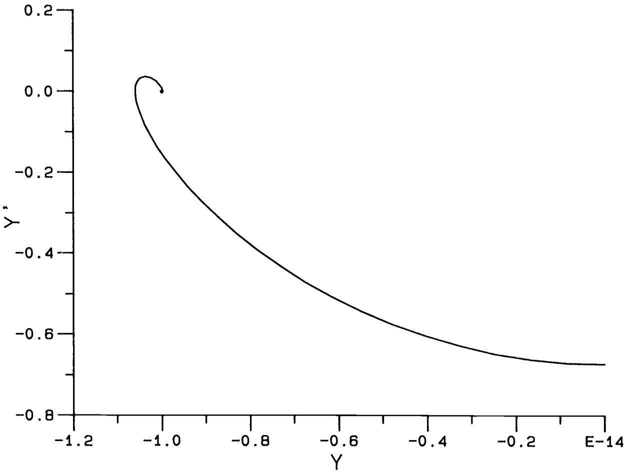


FIG. 1a. *The solution of* (1.8), (1.10) *with* $\alpha = 0$ *in the* yy' *plane.*

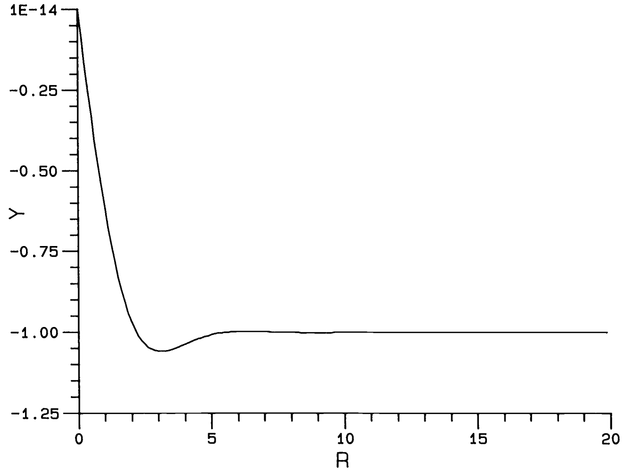


FIG. 1b. *The same curve in the ry plane.*

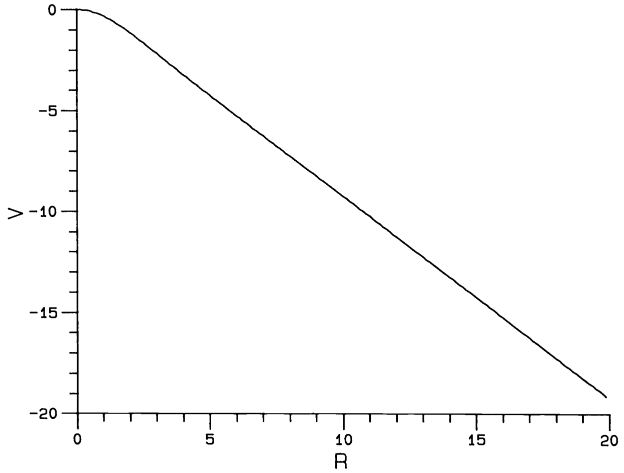


FIG. 1c. *The same curve in the rv plane.*

where

$$\bar{y} = (y_1, y_2, y_3, y_4) = (y, y', y'', r^{-1}). \tag{1.12}$$

The point $\bar{y}_0 = (-1, 0, 0, 0)$ is a critical point of the flow, and the differential $df(\bar{y}_0)$ besides the zero eigenvalue corresponding to the equation $y_4' = -y_4^2$ has a positive eigenvalue λ_1 and two eigenvalues $\lambda_2 = \bar{\lambda}_3$ with a negative real part. These eigenvalues are the roots of the characteristic equation

$$\lambda^3 + \alpha\lambda - 2 = 0. \tag{1.13}$$

Hence there exists a local three-dimensional central stable manifold M_{cs} of the flow that passes through the point $\bar{y}_0$. Since $y_4 = r^{-1} \to 0$ as $r \to \infty$, each point on M_{cs} that is sufficiently close to $\bar{y}_0$ is attracted by the flow to $\bar{y}_0$. Thus the surface M_{cs} replaces the asymptotic formula for large r. In a neighborhood of $r = 0$ the solution $y(r)$ of (1.8) could be expanded into converging power series

$$y(r) = \sum_{i=1}^{\infty} a_i r^{2i-1}, \tag{1.14}$$

where the slope

$$a_1 = y'(0) = s \tag{1.15}$$

is a free parameter which defines uniquely the rest of the coefficients a_i. We use this expansion to compute $\bar{y}(r_0)$, $r_0 = 1$. From that point, (1.8) is solved by the Taylor method with a small step $h = .125$ until the trajectory $\bar{y}(r)$ reaches the domain of existence of the manifold M_{cs}, say at $r = r_1$. Our algorithm described in §§2 and 4 takes into account the truncation and round-off errors. Thus, as we start with a point value of s, the final result $\bar{y}(r_1, s)$ is a small box in R^4. Suppose that for all $s \in I_{\Delta s} = [s_0 - \Delta s, s_0 + \Delta s]$:

$$\begin{aligned} &\text{(i) the boxes } \bar{y}(r_1, s) \text{ lie in the domain of existence of } M_{cs}, \\ &\text{(ii) the boxes } \bar{y}(r_1, s_0 - \Delta s) \text{ and } \bar{y}(r_1, s_0 + \Delta s) \text{ are separated by } M_{cs}. \end{aligned} \tag{1.16}$$

It follows then that there exists s in the interval $I_{\Delta s}$ such that $\bar{y}(r_1, s)$ belongs to M_{cs}, and hence $\bar{y}(r, s) \to \bar{y}_0$ as $r \to \infty$. The computation of $\bar{y}(r_1, s)$ for $s \in I_{\Delta s}$ is done by a single run of the program with s defined as the interval $I_{\Delta s}$. Since we want to prove the result for a continuum of α, we define α to be an interval $\alpha = I_{\Delta\alpha} = [\alpha_0 - \Delta\alpha, \alpha_0 + \Delta\alpha]$. The properties in (1.16) should then hold uniformly for α as above. Thus the idea of the proof is quite simple. However, its technical fulfillment is not trivial at all. The usual crude estimates of the domain of attraction for M_{cs} show that r_1 in the case where $\alpha = 0$ should be about 15. The roots of (1.13) are of absolute value $2^{1/3}$ so that the initial errors at $r = 0$ are amplified by the factor $\exp(15 \cdot 2^{1/3}) \approx 1.6 \cdot 10^8$. The domain of M_{cs} is of the order 10^{-2}. Thus, to satisfy (1.16) (i), Δs has to be of the order 10^{-10}. But then $\bar{y}(r_1, s_0 - \Delta s), \bar{y}(r_1, s_0 + \Delta s)$ are so close that they are not separated by M_{cs}. Besides, $\Delta\alpha$ also should be of the order 10^{-10}. Each run of the program takes about 10 seconds of computer time on Cyber 180/855. Thus, to cover an interval $\alpha \in [0, 1]$, we would need $\approx 2 \cdot 10^8$ hours of computer time. Instead, by a careful estimate of M_{cs} we reduced r_1 to about 5. The amplification factor thus was ≈ 500. Next, (1.8) was solved in characteristic variables. As a result, the solution $\bar{y}$ increased only in the direction normal to M_{cs}. Hence, smallness of $\bar{y}(r_n)-\bar{y}_0$ in the direction tangential to M_{cs} did not contradict the separability condition (1.16)(ii). Finally, the dependence of $\bar{y}(r)$ on the parameters s and α was expressed by a Taylor expansion of order 2. Thus, $\bar{y}(r)$ and the first derivatives with respect to s and α were computed at the central point (s_0, α_0) while only the second derivatives were computed in the domain $I_{\Delta s} \times I_{\Delta\alpha}$. As a result, we run the program successfully with $\Delta s = .0015$ and $\Delta\alpha = .005$ from $\alpha = 0$ through $\alpha = 2.39$ at the expense of 1 hour of CPU on Cyber 180/855. We could continue the proof for larger α with a smaller Δs, but for our purposes the interval $\alpha \in [0, 2.39]$ is more than enough.

Recall that the empirical stability domain of Bunsen flames extends only to $c = 3$, i.e., $\alpha = .31$. Our nonrigorous numerical experiments suggest that for $\alpha > 1.1$ problem (1.8), (1.10) has infinitely many solutions. The total picture is similar to that obtained in [5] for the antonomous (1.5). The branch followed by our program corresponds to the minimal slope $s = y'(0)$ of all the bounded solutions for given α. The values of s_0 corresponding to α_0 and the derivatives $ds_0/d\alpha_0$ were obtained by a nonrigorous preliminary program. In addition to conditions in (1.16) our interval arithmetic program verified the transversality of the intersection of the curve $\bar{y}(r_1, s)$ for $s \in I_{\Delta s}$ and the manifold M_{cs} and the negativeness of y. We summarize the results of all computations in the following theorem.

THEOREM 1. *Problem* (1.8), (1.10) *has a negative solution for all* $\alpha \in [0, 2.39]$, *i.e.,* $0.383 < c < \infty$. *This solution is formed by a transversal intersection of the curve* $\bar{y}(r_1, s), s \in {}^{I}\!_{\Delta s}$ *and the central stable manifold* M_{cs} *of the critical point* $\bar{y}_0 = (-1, 0, 0, 0)$. *The values of the slopes* $s = y'(0)$ *and the corresponding values of* α *appear in Table* 1.

Remark 1. Because of the lack of space we display in this article only the results for $\alpha = n \cdot 0.1$, $0 \le n \le 23$ and $\alpha = 2.39$. The complete Table 1 may be obtained from the author upon request.

The negativeness of $y(r)$ for all $r > 0$ means that the Bunsen-flame cone $v(r)$ is monotonically decreasing. We conjecture that the negativeness of $y(r)$ singles it out as the only bounded solution of (1.8) with $y(0) = 0$.

2. The power series expansion at $r = 0$. Let us substitute the formal expansion (1.14) into (1.8). We obtain the following recurrent relation:

$$a_{i+1} = -\left(\alpha a_i + \left(\sum_{j=1}^{i-1} a_j a_{i-j}\right)/(2i)\right)/(4i(i+1)), \qquad i \ge 2 \tag{2.1}$$

and

$$a_2 = \frac{1 - 2\alpha a_1}{16}, \tag{2.2}$$

where a_1 is a free parameter. For appropriate ρ_0 and ρ_1 the coefficients a_i are bounded by

$$|a_i| < \rho_1 \rho_0^{i-1}. \tag{2.3}$$

Indeed, for that purpose, ρ_1 and ρ_0 should satisfy

$$\rho_1 \ge |a_1|, \qquad \rho_1\rho_0 \ge \frac{|1 - 2\alpha a_1|}{16}. \tag{2.4}$$

From (2.1) we obtain

$$|a_{i+1}| \le \frac{\alpha\rho_1\rho_0^{i-1}}{4i(i+1)} + \frac{(i-1)\rho_1^2\rho_0^{i-2}}{8i^2(i+1)}. \tag{2.5}$$

Thus, to assure the bounds in (2.3) for $i \ge 2$, we impose

$$\frac{\alpha\rho_0}{4i(i+1)} + \frac{(i-1)\rho_1}{8i^2(i+1)} \le \rho_0^2, \qquad i \ge 2. \tag{2.6}$$

In particular,

$$\frac{\alpha\rho_0}{24} + \frac{\rho_1}{96} \le \rho_0^2. \tag{2.7}$$

The last inequality clearly implies all inequalities in (2.6). Since $\rho_0 > 0$, by (2.7)

$$\rho_0 \ge \frac{1}{48}\,(\alpha + (\alpha^2 + 24\rho_1)^{1/2}) = F_1(\rho_1), \tag{2.8}$$

while by (2.4)

$$\rho_0 \ge \frac{|1-2\alpha a_1|}{16\rho_1} = F_2(\rho_1). \tag{2.9}$$

Since $F_1(\rho)$ is increasing and $F_2(\rho)$ is a decreasing function there exists a unique ρ_1^* such that $F_1(\rho_1^*) = F_2(\rho_1^*)$. It is easy to see that this ρ_1^* satisfies the cubic equation

$$\rho_1^3 + \frac{\alpha(1-2\alpha a_1)}{4}\,\rho_1 - \frac{3}{8}\,(1-2\alpha a_1)^2 = 0. \tag{2.10}$$

Thus, the pair (ρ_0, ρ_1) with the smallest ρ_0 is given by the formula

$$\rho_1 = \max(\rho_1^*, |a_1|), \qquad \rho_0 = F_1(\rho_1). \tag{2.11}$$

The radius of convergence of the series is bounded from below by $R \ge \rho_0^{-1/2}$. For example, for $\alpha = 0$ the corresponding slope $a_1 = s = -.6739$. Since $F_1(|a_1|) = .0838$ is smaller than $F_2(|a_1|) = .0927$ it follows that $\rho_1 = \rho_1^* = .7211$ and $\rho_0 = .0867$, while $R > 3.39$. In computations we employ instead of y a finite sum

$$y_N = \sum_{i=1}^{N} a_i r^{2i-1}. \tag{2.12}$$

The remainder $E_N(r)$ is bounded by

$$|E_N(r)| \le \sum_{i=N+1}^{\infty} |a_i| r^{2i-1} \le \rho_1\rho_0^N r^{2N+1}(1-\rho_0 r^2)^{-1}. \tag{2.13}$$

In particular, for ρ_0, ρ_1 as above, and

$$N = 80, \qquad r = r_0 = 1, \tag{2.14}$$

we obtain $|E_N(r)| < 10^{-85}$. The derivatives $y'(r)$ and $y''(r)$ are approximated by the sums

$$y'_N = \sum_{i=1}^{N}(2i-1)a_i r^{2i-2}, \qquad y''_N = \sum_{i=1}^{N}(2i-1)(2i-2)a_i r^{2i-3}. \tag{2.15}$$

By (2.5)

$$(2i-1)(2i-2)|a_i| \le \alpha\rho_1\rho_0^{i-1} + \frac{\rho_1^2\rho_0^{i-2}}{2}, \tag{2.16}$$

and hence the remainder E''_N of y'' is bounded by

$$|E''_N(r)| \le \left(\alpha + \frac{\rho_1}{2\rho_0}\right)\rho_1\rho_0^N r^{2N-1}(1-\rho_0 r^2)^{-1}. \tag{2.17}$$

Clearly, for $r < 2N$ the terms in the sum for E'_N are smaller than the corresponding terms in $E''_N(r)$, and hence also $|E'_N(r)|$ is bounded by the right-hand side of (2.17). In addition to the function $y(r)$ we will have to compute its partial derivatives up to order 2 with respect to the parameters $y'(0) = s$ and α. More precisely, when differentiating with respect to α we assume that

$$s = s_0 + s_{\alpha_0}\cdot(\alpha-\alpha_0), \tag{2.18}$$

where s_0 and s_{α_0} are constants. In our computer program these constants are denoted by $S0$ and $SA0$. The coefficients of the expansion

$$\partial_p y(r) = \sum_{i=0}^{\infty}\partial_p a_i r^{2i-1}, \qquad p = s \text{ or } p = \alpha \tag{2.19}$$

satisfy for $i \ge 1$

$$\partial_p a_{i+1} = \left(\alpha\partial_p a_i + \partial_p\alpha\cdot a_i + 2\left(\sum_{j=1}^{i-1}\partial_p a_{i-j}\cdot a_j\right)/(2i)\right)/(4i(i+1)) \tag{2.20}$$

with the first coefficient

$$\partial_s a_1 = 1, \qquad \partial_\alpha a_1 = s_{\alpha_0}. \tag{2.21}$$

The corresponding formulas for the second derivatives are

$$\begin{aligned}\partial_{p_1}\partial_{p_2}a_{i+1} = \Bigg(&\alpha\partial_{p_1}\partial_{p_2}a_i + \partial_{p_1}\alpha\partial_{p_2}a_i + \partial_{p_2}\alpha\partial_{p_1}a_i\\ &+2\sum_{j=1}^{i-1}(\partial_{p_1}\partial_{p_2}a_{i-j}\cdot a_j + \partial_{p_1}a_{i-j}\cdot\partial_{p_2}a_j)/(2i)\Bigg)/(4i(i+1)),\end{aligned} \tag{2.22}$$

where p_1, p_2 are s or α. Clearly, only $\partial_\alpha\alpha = 1$ is nonzero. The first coefficient for all p_1, p_2 is

$$\partial_{p_1}\partial_{p_2}a_1 = 0. \tag{2.23}$$

We are looking for a common estimate

$$|\partial_s^{j_1}\partial_\alpha^{j_2}a_i| \le \rho_3\rho_2^{i-1}, \quad \rho_2 \ge \rho_0, \quad \rho_3 \ge \rho_1, \quad 0 \le j_1 + j_2 \le 2. \tag{2.24}$$

Then (2.20) implies

$$|\partial_p a_{i+1}| \le \frac{\alpha\rho_3\rho_2^{i-1} + \rho_1\rho_2^{i-1}}{4i(i+1)} + \frac{(i-1)\rho_3\rho_1\rho_2^{i-2}}{4i^2(i+1)} \tag{2.25}$$

and by (2.21) and (2.24)

$$\rho_3 = \max(1, |s_{\alpha_0}|, \rho_1). \tag{2.26}$$

From (2.22) we obtain the upper bound (in the case where $p_1 = p_2 = \alpha$)

$$\begin{aligned} |\partial_{p_1}\partial_{p_2}a_{i+1}| &\leq \frac{\alpha\rho_3\rho_2^{i-1} + 2\rho_3\rho_2^{i-1}}{4i(i+1)} + \frac{(i-1)(\rho_3\rho_1\rho_2^{i-2} + \rho_3^2\rho_2^{i-2})}{4i^2(i+1)} \\ &\leq \frac{(\alpha+2)\rho_3\rho_2^{i-1}}{4i(i+1)} + \frac{(i-1)\rho_3^2\rho_2^{i-2}}{2i^2(i+1)}. \end{aligned} \tag{2.27}$$

Clearly, the last bound is higher than that in (2.25). In order to satisfy (2.24) we have to request

$$\rho_2^2 \geq \frac{(\alpha+2)\rho_2}{4i(i+1)} + \frac{(i-1)\rho_3}{2i^2(i+1)}. \tag{2.28}$$

The strongest inequalities are attained in the cases $i = 1$, $i = 2$:

$$\rho_2^2 \geq \frac{(\alpha+2)\rho_2}{24} + \frac{\rho_3}{24}, \qquad \rho_2 \geq \frac{\alpha+2}{8}. \tag{2.29}$$

Hence we can define

$$\rho_2 = \max\left(\frac{1}{48}\,(\alpha+2+((\alpha+2)^2+96\rho_3)^{1/2}),\ \frac{\alpha+2}{8}\right). \tag{2.30}$$

Obviously, ρ_2 is greater than ρ_0 defined in (2.11). For example, for $\alpha = 0$ we obtain $\rho_3 = 1$ and $\rho_2 = .25$. The truncation error $\partial_s^{j_1}\partial_\alpha^{j_2}E_N$, $j_1 + j_2 \leq 2$ is estimated as in (2.13) with ρ_0, ρ_1 replaced by ρ_2 and ρ_3, respectively. For the second derivative $\partial_s^{j_1}\partial_\alpha^{j_2}E_N''$, instead of (2.16) we obtain

$$(2i-1)(2i-2)|\partial_s^{j_1}\partial_\alpha^{j_2}a_i| \leq (\alpha+2)\rho_3\rho_2^{i-1} + 2\rho_3^2\rho_2^{i-2}. \tag{2.31}$$

Hence, the common upper bound for the derivatives of E_N is

$$\begin{gathered} |\partial_s^{j_1}\partial_\alpha^{j_2}E_N^{(k)}| \leq (\alpha+2+2\rho_3/\rho_2)\rho_3\rho_2^N r^{2N-1}(1-\rho_2 r^2)^{-1}, \\ i_1 + i_2 \leq 2, \qquad k \leq 2. \end{gathered} \tag{2.32}$$

The above bound is denoted in the subroutine SERIES (see the Appendix) by $ERROR(2) = -ERROR(1)$. In the worst case of $\alpha = 2.39$, $s = -2.989$, and $s_{\alpha_0} = -1.55$ we obtain $\rho_1^* = 3.78 = \rho_1$, $\rho_0 = .254$, $\rho_3 = \rho_1$, $\rho_2 = \max(.50, .55) = .55$, and hence the bound in (2.32) for r and N as in (2.14) is $2.6 \cdot 10^{-19}$. This number, relative to the computed values of $\partial_s^{j_1}\partial_\alpha^{j_2}y_N^{(k)}(r)$, seems to be far below the maximal roundoff error of our computer. Still, to be sure, we add the interval ($ERROR(1)$, $ERROR(2)$) to the above derivatives of y_N. In subroutine SERIES we compute the derivatives $\partial_s^{j_1}\partial_\alpha^{j_2}y^{(k)}(r)$, $k \leq 2$ for $0 \leq j_1 + j_2 \leq 1$ at the central point s_0, α_0 and for $j_1 + j_2 = 2$ at the intervals $I_{\Delta s} = (s_0 - \Delta s, s_0 + \Delta s)$ and $I_{\Delta\alpha} = (\alpha_0 - \Delta\alpha, \alpha_0 + \Delta\alpha)$. In order to simplify notation we employ in our computer program the following general convention. For an interval variable f which depends on the parameters s and α we store the endpoints of the interval $\partial_s^{j_1}\partial_\alpha^{j_2}f$ at the vector

$$\partial_s^{j_1}\partial_\alpha^{j_2}f \to (f(j), f(j+1)), \qquad j = (j_1+j_2)(j_1+j_2+1) + 2j_2 + 1. \tag{2.33}$$

Thus, the index j varies from 1 through 11. The above derivatives for $j_1+j_2 \le 1$ are routinely computed at the central point (s_0, α_0), while the second-order derivatives are evaluated for the whole intervals $I_{\Delta s}$ and $I_{\Delta\alpha}$. In intermediate computations we sometimes need the values of f and its first derivatives at the above intervals. To distinguish between the central and interval functions, in the former case we modify the name f as $f0$. Note that the second derivatives are never evaluated at the central point. To allow a uniform treatment of the derivatives of f in majority of loops, we have abused the notation and denoted the second derivatives at the intervals of α and s by $f0$ instead of f. Thus, $f0(11), f0(12)$ is the derivative $\partial_\alpha^2 f$ at $I_{\Delta s} \times I_{\Delta s}$ while $f0(5) = \partial_\alpha f$ at (s_0, α_0). The values of f and the derivatives $\partial_s f$, $\partial_\alpha f$ at $I_{\Delta s} \times I_{\Delta\alpha}$ are computed by Taylor's formulas

$$\begin{aligned} f &= f0 + \partial_s f0 \cdot (s-s_0) + \partial_\alpha f0 \cdot (\alpha - \alpha_0) \\ &\quad + \tfrac{1}{2}\partial_s^2 f0 \cdot (s-s_0)^2 + \partial_s\partial_\alpha f0 \cdot (s-s_0)\cdot(\alpha-\alpha_0) + \tfrac{1}{2}\partial_\alpha^2 f0 \cdot (\alpha-\alpha_0)^2, \end{aligned} \tag{2.34}$$

$$\partial_s f = \partial_s f0 + \partial_s^2 f0 \cdot (s-s_0) + \partial_s\partial_\alpha f0 \cdot (\alpha-\alpha_0), \tag{2.35}$$

and similar expression for $\partial_\alpha f$. In actual computations we replace $(s-s_0)$, $(\alpha-\alpha_0)$, and their products by the intervals

$$\begin{aligned} &DP(3), DP(4) = -\Delta s, \Delta s; \qquad DP(5), DP(6) = -\Delta\alpha, \Delta\alpha, \\ &DP(7), DP(8) = -(\Delta s)^2, (\Delta s)^2; \qquad DP(9), DP(10) = -\Delta s\cdot\Delta\alpha,\ \Delta s \cdot \Delta\alpha, \\ &DP(11), DP(12) = -(\Delta\alpha)^2, (\Delta\alpha)^2. \end{aligned} \tag{2.36}$$

In our Fortran program formula (2.34) is written as

$$\begin{aligned} \text{I}\quad F(1) &= F0(1) + F0(3)^*DP(3) + F0(5)^*DP(5) \\ &\quad + 0.5^*F0(7)^*DP(7) + F0(9)^*DP(9) + 0.5^*F0(11)^*DP(11). \end{aligned} \tag{2.37}$$

Thus, interval variables are represented by the address of their left endpoints. The letter I in the first column of the instruction indicates that this is an interval statement. Our preprocessor COMPINT then translates this instruction into a sequence of CALL statements. This preprocessor and its output are described in more detail in §5.

Following the above convention we store the derivatives $\partial_s^{j_1}\partial_\alpha^{j_2} a_i$ in case $(s, \alpha) = I_{\Delta s} \times I_{\Delta\alpha}$ and $j_1 + j_2 \le 1$ at the locations $A(j,i)$, $A(j+1,i)$, where j is defined as in (2.31), while in all remaining cases they are stored in $A0(j,i)$, $A0(j+1,i)$. Formulas (2.1), (2.2) are used to compute $A0(1,i)$, $A0(2,i)$, formulas (2.20), (2.21) for $A0(3,i)$, $A0(4,i)$ and $A0(5,i)$, $A0(6,i)$, and formula (2.22) for $A0(j,i)$, $A0(j+1,i)$, where $j = 7, 9, 11$. In (2.22) the lower-order derivatives a_i, $\partial_s a_i$, and $\partial_\alpha a_i$ are replaced by the corresponding intervals $A(j,i)$, $A(j+1,i)$. The latter ones in turn are computed as in the example (2.37). The final series $\sum_{i=1}^N \partial_s^{j_1}\partial_\alpha^{j_2} a_i r^{2i-1}$ are summed up using the Horner method.

3. The central stable manifold. As mentioned in the Introduction, there exists a local three-dimensional central stable manifold M_{cs} of the flow (1.11) which passes through the point $\bar{y}_0 = (-1, 0, 0, 0)$. The neutral direction corresponds to the fourth equation $\dot{y}_4 = -y_4^2$. Hence the rate of convergence of $\bar{y}(r)$ to $\bar{y}_0$ is not exponential but of order $O(r^{-1})$. Note that (1.8) has a formal asymptotic solution

$$y = \sum_{i=0}^{\infty} b_i r^{-1}, \qquad b_0 = -1, \tag{3.1}$$

where the remaining coefficients are uniquely determined by b_0. Although the series in (3.1) is a diverging one, we will see below that the partial sums

$$y_N = \sum_{i=0}^{N} b_i r^{-i} \tag{3.2}$$

constitute the leading part of all solutions $y(r)$ that tend to the critical point $y = -1$ as $r \to \infty$. Let us pass to the variable

$$y_{\text{new}} = y_{\text{old}} - y_N. \tag{3.3}$$

In the new variable $y = y_{\text{new}}$ (1.8) becomes

$$\begin{aligned} y''' + \alpha y' - 2y &= f(y, r) \\ &= -2r^{-1}y'' + y'r^{-2} - yr^{-3} - y^2 - 2\Delta y_N \cdot y - p(r^{-1}), \end{aligned} \tag{3.4}$$

where

$$\Delta y_N = \sum_{i=2}^{N} b_i r^{-i} = O(r^{-2}) \tag{3.5}$$

and $p(r^{-1})$ is the result of substitution of y_N into (1.8):

$$p(r^{-1}) = \sum_{i=N+1}^{2N} p_i r^{-i} = O(r^{-N-1}). \tag{3.6}$$

Since $f(0, r) = O(r^{-N-1})$ and the roots of (1.13) are not imaginary, it follows easily that all solutions of (3.4) that tend to zero as $r \to \infty$ are of the order $O(r^{-N-1})$. We will need, however, an exact bound of y. For that purpose, let us pass to the characteristic variables $u = (u_1, u_2, u_3)$:

$$u_1 = \frac{(D - \lambda_2)(D - \lambda_3)y}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)}, \quad u_2 = \frac{(D - \lambda_1)(D - \lambda_3)y}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)}, \quad u_3 = \bar{u}_2, \tag{3.7}$$

where $\lambda_1, \lambda_2 = \bar{\lambda}_3$ are the roots of (1.13) and $\lambda_1 > 0$. Equation (3.4) becomes

$$u_1' - \lambda_1 u_1 = \frac{f(u, r)}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)}, \qquad u_2' - \lambda_2 u_2 = \frac{f(u, r)}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)}. \tag{3.8}$$

Assume for a moment that $f = f(r)$ is a given bounded function of r. Then solutions of (3.8) that are bounded on the interval (r_1, ∞) satisfy the estimate

$$\|u_1\|_\infty \le c_1 \|f\|_\infty, \qquad \|u_2\|_\infty \le c_2 \|f\|_\infty + |u_2(r_1)|, \tag{3.9}$$

where

$$c_1 = \frac{1}{\lambda_1|\lambda_1 - \lambda_2|^2}, \qquad c_2 = \frac{1}{|Re\ \lambda_2| \cdot |\lambda_2 - \lambda_1| \cdot |\lambda_2 - \lambda_3|} = \frac{1}{\lambda_1 \cdot |\lambda_2 - \lambda_1| \cdot |\mathrm{Im}\ \lambda_2|} \tag{3.10}$$

while $\|u\|_\infty = \sup_{r \geq r_1} |u(r)|$ is the usual maximum norm. The function y and its derivatives in terms of u_1, u_2 are

$$y = u_1 + 2\mathrm{Re}\ u_2, \quad y' = \lambda_1 u_1 + 2\mathrm{Re}(\lambda_2 u_2), \quad y'' = \lambda_1^2 u_1 + 2\mathrm{Re}(\lambda_2^2 u_2). \tag{3.11}$$

Now we recall that f depends on y, and with the aid of (3.9) we estimate

$$\begin{aligned}\|f\|_\infty &\leq 2r_1^{-1}(\lambda_1^2\|u_1\|_\infty + 2|\lambda_2|^2\|u_2\|_\infty) + r_1^{-2}(\lambda_1\|u_1\|_\infty + 2|\lambda_2|\ \|u_2\|_\infty) \\ &\quad + (r_1^{-3} + 2\|\Delta y_N\|_\infty)(\|u_1\|_\infty + 2\|u_2\|_\infty) \\ &\quad + (\|u_1\|_\infty + 2\|u_2\|_\infty)^2 + \|p\|_\infty \\ &\leq c_3\|f\|_\infty + c_4\|f\|_\infty^2 + c_5,\end{aligned} \tag{3.12}$$

where

$$\begin{aligned}c_3 &= 2r_1^{-1}(c_1\lambda_1^2 + 2c_2|\lambda_2|^2) + r_1^{-2}(c_1\lambda_1 + 2c_2|\lambda_2|) \\ &\quad + (r_1^{-3} + 2\|\Delta y_N\|_\infty + 4|u_2(r_1)|)(c_1 + 2c_2),\end{aligned} \tag{3.13}$$

$$c_4 = (c_1 + 2c_1)^2, \tag{3.14}$$

and

$$\begin{aligned}c_5 &= 2(2r_1^{-1}|\lambda_2|^2 + r_1^{-2}|\lambda_2| + r_1^{-3} + 2\|\Delta y_N\|_\infty)|u_2(r_1)| \\ &\quad + 4|u_2(r_1)|^2 + \sum_{i=N+1}^{2N} |p_i| r_1^{-i}.\end{aligned} \tag{3.15}$$

Given $u_2(r_1)$, system (3.8) is solved by iterations where $f = f_n = f(u^{(n)})$ is evaluated at $u^{(n)}$ and the corresponding bounded solution of the initial value problem with $u_2^{(n+1)}(r_1) = u_2(r_1)$ is denoted by $u^{(n+1)}$. The zero approximation $u^{(0)}$ is set to be zero. Now, inequality (3.12) is replaced by

$$\|f_{n+1}\|_\infty \leq c_3\|f_n\|_\infty + c_4\|f_n\|_\infty^2 + c_5. \tag{3.16}$$

Suppose that

$$c_3 < 1 \quad \text{and} \quad c_6 = (1 - c_3)^2 - 4c_4c_5 > 0. \tag{3.17}$$

Then (3.16) implies

$$\|f_{n+1}\|_\infty \leq c_7 = (1 - c_3 - \sqrt{c_6})/(2c_4), \tag{3.18}$$

since $\|f_0\|_\infty = \|p\|_\infty \leq c_5 \leq c_7$. Thus, there exists a subsequence $u^{(n_i)}$ that converges to the solution of the original nonlinear equation and has the prescribed value of

$u_2(r_1)$. The conditions in (3.17) restrict the size of r_1^{-1} and of $|u_2(r_1)|$. But once these conditions are met, the correspondence

$$r_1^{-1}, u_2(r_1) \rightarrow u_1(r_1) \tag{3.19}$$

defines the local central-stable manifold M_{cs}. The value of $u_1(r_1)$ is then bounded by

$$|u_1(r_1)| \leq c_1 c_7. \tag{3.20}$$

In our computation we set $N = 2$. Then

$$y_N = -1 + b_1 r^{-1} + b_2 r^{-2}, \qquad b_1 = \frac{-\alpha}{2}, \quad b_2 = \frac{\alpha^2}{8}, \tag{3.21}$$

$$\Delta y_N = b_2 r^{-2}, \tag{3.22}$$

and

$$\begin{aligned} p(r) &= p_3 r^{-3} + p_4 r^{-4} + p_5 r^{-5} \\ &= -(1 + \frac{\alpha^3}{4})r^{-3} + \frac{\alpha^4}{64} r^{-4} - \frac{9}{8}\alpha^2 r^{-4}. \end{aligned} \tag{3.23}$$

Since α varies in the interval $[\alpha_0 - \Delta\alpha, \alpha_0 + \Delta\alpha]$, we freeze the coefficient α in (3.4) at α_0 and instead modify f by

$$f \rightarrow f - (\alpha - \alpha_0)y'. \tag{3.24}$$

This effects the coefficients c_3 and c_5

$$c_3 \rightarrow c_3 + |\Delta\alpha|(c_1\lambda_1 + 2c_2|\lambda_2|), \qquad c_5 \rightarrow c_5 + 2|\Delta\alpha| \cdot |\lambda_2|\,|u_2(r_1)|. \tag{3.25}$$

The eigenvalues λ_i are roots of the equation

$$\lambda^3 + \alpha_0\lambda - 2 = 0. \tag{3.26}$$

Let us now look at Table 1. The numbers in the columns from left to right are as follows: $\alpha_0, s_0, s_{\alpha_0}, r_1$, the upper bound of $|z_2(r_1)|$ for $s, \alpha \in I_{\Delta s} \times I_{\Delta\alpha}$, the interval value of $z_1(r_1)$ for $s = s_0 - \Delta s$, $\alpha \in I_{\Delta\alpha}$, and the interval value of $z_1(r_1)$ for $s = s_0 + \Delta s$, $\alpha \in I_{\Delta\alpha}$. The increments $\Delta\alpha$ and Δs are fixed $\Delta\alpha = .005$, $\Delta s = .0015$. Because of the lack of space we display in this article only the results for $\alpha_1 = n \cdot 0.1$, $0 \leq n \leq 23$, and $\alpha = 2.39$. The complete Table 1 is on deposit with the author and may be obtained by the interested reader upon request. Let us check, for example, that for $\alpha_0 = 0$ the values r_1, $z_2(r_1)$ lie in the domain of existence of M_{cs} and $z_1(r_1)$ for $s = s_0 + \Delta s$ and $s_0 - \Delta s$ are separated by M_{cs}. The eigenvalues λ_i are

$$\lambda_1 = 2^{1/3} \approx 1.26, \qquad \lambda_2 = \frac{-\lambda_1 + i(3\lambda_1^2 + 4\alpha)^{1/2}}{2}. \tag{3.27}$$

Hence

$$\begin{aligned} &|\lambda_2| = (\lambda_1^2 + \alpha)^{1/2} \approx 1.26, \qquad |\lambda_1 - \lambda_2| = (3\lambda_1^2 + \alpha)^{1/2} \approx 2.18, \\ &|\text{Im } \lambda_2| = (3\lambda_1^2 + 4\alpha)^{1/2}/2 \approx 1.09. \end{aligned} \tag{3.28}$$

TABLE 1

The output of the program BUNSEN. *The increments* DALF *and* DS *of* ALF *and* S *are always* .005 *and* .0015, *respectively. The order of the Taylor method* NORDER *is* 8 *for* S0 *less than* 1.71 *and* 10 *otherwise.*

```
-------------------------------------------------------------------------------
 ALFO      S0      SA0    RMAX    |Z2(S)|        Z1(S0-DS)           Z1(S0+DS)
-------------------------------------------------------------------------------

  .00  -.67385   -.44   5.375   .12E-01   -.12E+00  -.12E+00   .11E+00  .11E+00
  .10  -.71981   -.48   5.375   .12E-01   -.11E+00  -.11E+00   .10E+00  .10E+00
  .20  -.76932   -.51   5.375   .11E-01   -.10E+00  -.97E-01   .91E-01  .93E-01
  .30  -.82256   -.55   5.375   .11E-01   -.91E-01  -.88E-01   .82E-01  .85E-01
  .40  -.87971   -.59   5.500   .11E-01   -.94E-01  -.91E-01   .85E-01  .89E-01
  .50  -.94093   -.63   5.500   .11E-01   -.86E-01  -.82E-01   .77E-01  .81E-01
  .60 -1.00639   -.68   5.625   .11E-01   -.89E-01  -.85E-01   .80E-01  .84E-01
  .70 -1.07622   -.72   5.625   .11E-01   -.81E-01  -.76E-01   .72E-01  .77E-01
  .80 -1.15056   -.77   5.750   .10E-01   -.84E-01  -.77E-01   .73E-01  .80E-01
  .90 -1.2[illegible]952   -.81   5.750   .11E-01   -.76E-01  -.70E-01   .66E-01  .73E-01
 1.00 -1.31320   -.86   5.875   .11E-01   -.79E-01  -.70E-01   .66E-01  .75E-01
 1.10 -1.40168   -.91   6.000   .11E-01   -.82E-01  -.70E-01   .66E-01  .78E-01
 1.20 -1.49504   -.96   6.125   .12E-01   -.85E-01  -.69E-01   .65E-01  .80E-01
 1.30 -1.59333  -1.01   6.250   .12E-01   -.89E-01  -.68E-01   .62E-01  .83E-01
 1.40 -1.69659  -1.06   6.375   .13E-01   -.93E-01  -.65E-01   .58E-01  .86E-01
 1.50 -1.80486  -1.11   6.500   .14E-01   -.99E-01  -.61E-01   .53E-01  .91E-01
 1.60 -1.91814  -1.16   6.750   .15E-01   -.12E+00  -.60E-01   .50E-01  .11E+00
 1.70 -2.03646  -1.21   7.250   .17E-01   -.18E+00  -.61E-01   .45E-01  .16E+00
 1.80 -2.15981  -1.26   6.875   .13E-01   -.87E-01  -.74E-01   .62E-01  .76E-01
 1.90 -2.28819  -1.31   7.000   .14E-01   -.89E-01  -.71E-01   .59E-01  .76E-01
 2.00 -2.42157  -1.36   7.125   .14E-01   -.89E-01  -.69E-01   .56E-01  .76E-01
 2.10 -2.55993  -1.41   7.250   .14E-01   -.90E-01  -.65E-01   .52E-01  .77E-01
 2.20 -2.70323  -1.46   7.375   .15E-01   -.90E-01  -.60E-01   .48E-01  .78E-01
 2.30 -2.85146  -1.51   7.625   .15E-01   -.99E-01  -.57E-01   .47E-01  .89E-01
 2.39 -2.98904  -1.55   8.000   .16E-01   -.12E+00  -.57E-01   .47E-01  .11E+00
```

The constants

$$(3.29)\qquad \begin{aligned} c_1 &= \frac{1}{\lambda_1(3\lambda_1^2+\alpha)} = \frac{1}{6}, \\ c_2 &= \frac{2}{\lambda_1(3\lambda_1^2+\alpha)^{1/2}\cdot(3\lambda_1^2+4\alpha)^{1/2}} = \frac{1}{3}. \end{aligned}$$

For $r_1 = 5.375$ and $|u_2(r_1)| \le .012$ we have

$$(3.30)\qquad \|\Delta y_N(r)\|_\infty = |\alpha|^2/8\cdot r_1^{-2} \le |\Delta\alpha|^2/8\cdot r_1^{-2} = .11\cdot 10^{-6},$$

$$(3.31)\qquad \|p(r)\|_\infty \le \left(1+\frac{|\Delta\alpha|^3}{4}\right) r_1^{-3} + \frac{|\Delta\alpha|^4}{4} r_1^{-4} + \frac{9}{8}|\Delta\alpha|^2 r_1^{-4} \approx .6\cdot 10^{-2},$$

$$(3.32)\qquad c_3 \approx .58,\quad c_4 \approx .70,\quad c_5 \approx .022,\quad c_6 \approx .12,\quad c_7 \approx .058\ .$$

Thus r_1, $u_2(r_1)$ lie in the domain of M_{cs} and $|u(r_1)| \le c_1c_7 < .01$. The values $-.12$ and $.11$ of $u_1(r_1)$ for $s = s_0 \pm \Delta s$ are clearly separated by the strip $[-.01, .01]$. To verify the separability condition we wrote a computer program which appears as the SUBROUTINE ESTIM in the computer code in the Appendix. This subroutine is written also in interval arithmetic and the input values of all parameters are obtained from the main program. The trajectory $\bar{y}(r)$ is advanced in r with steps $h = .125$ until ESTIM gives a positive result, which constitutes a formal proof of the existence of solution. The final values of r_1 and $u(r_1)$ are printed only to make the claim more convincing.

In addition we checked the transversality of the intersection of the curve $\bar{y}(r_1; s)$ $s \in I_{\Delta s}$ with the manifold M_{cs}. Besides the values of $\bar{y}(r_1; s)$ and $u(r_1; s)$ our program computes the interval values of $\partial_s u(r_1; s)$ for $(s, \alpha) \in I_{\Delta s} \times I_{\Delta\alpha}$. Suppose that for some $s_1 \in I_{\Delta s}$, $u(r_1; s)$ lies in M_{cs} and $\partial_s u(r_1; s_1)$ is tangent to M_{cs}. Then the ratio $\partial_s u_1(r_1; s_1)/\partial_s u_2(r_1; s)$ should tend to zero as $r \to \infty$ since at the fixed point $\bar{y}_0$ the tangent plane to M_{cs} is just $u_1 = 0$. The values of $\partial_s u(r_1; s)$ computed by the program were such that $|\partial_s u_1(r_1; s)| \gg |\partial_s u_2(r_1; s)|$. We wish to prove that for these values of $\partial_s u(r_1; s_1)$ there exists a constant $c > 0$ such that

$$c|\partial_s u_1(r; s_1)|^2 - |\partial_s u_2(r; s_1)|^2 > 0 \quad \text{for all } r > r_1. \tag{3.33}$$

This clearly implies the transversality of the intersection. The function $u_s(r) = \partial_s u(r; s_1)$ satisfies the system

$$\begin{aligned} u'_{s1} &= \lambda_1 u_{s1} + \frac{1}{|\lambda_1 - \lambda_2|^2}\, df \cdot u_s, \\ u'_{s2} &= \lambda_2 u_{s2} + \frac{1}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)}\, df \cdot u_s, \qquad u_{s3} = \bar{u}_{s2}, \end{aligned} \tag{3.34}$$

where df is the differential of f with respect to u. From the explicit formulas of f in (3.4) and (3.24) and from the relations in (3.11) it follows that

$$|df \cdot u_s| \le a_1|u_{s1}| + a_2|u_{s2}|, \tag{3.35}$$

where

$$\begin{aligned} a_1 &= |\Delta\alpha|\lambda_1 + 2r_1^{-1}\lambda_1^2 + 2|b_2|r_1^{-2} + r_1^{-2}\lambda_1 + r_1^{-3} + 2\|y\|_\infty, \\ a_2 &= 2(|\Delta\alpha| \cdot |\lambda_2| + 2r_1^{-1}|\lambda_2|^2 + 2|b_2|r_1^{-2} + r_1^{-2}|\lambda_2| + r_1^{-3} + 2\|y\|_\infty), \end{aligned} \tag{3.36}$$

and

$$\|y\|_\infty \le \|u_1\|_\infty + 2\|u_2\|_\infty \le (c_1 + 2c_2)\|f\|_\infty + 2|u_2(r_1)|. \tag{3.37}$$

Multiplying the system in (3.34) by the vector $(cu_{s1}, -\bar{u}_{s2})$ and taking real parts yields

$$\begin{aligned} \frac{1}{2}(c|u_{s1}|^2 - |u_{s2}|^2)' &= c\lambda_1|u_{s1}|^2 - \operatorname{Re}\lambda_2|u_{s2}|^2 \\ &\quad + \frac{c}{|\lambda_1 - \lambda_2|^2}(df \cdot u_s)u_{s1} - \operatorname{Re}\left(\frac{(df \cdot u_s)\bar{u}_{s2}}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)}\right). \end{aligned} \tag{3.38}$$

We wish to show that the right-hand side of (3.38) is nonnegative. Recall that $-\operatorname{Re}\lambda_2 = \lambda_1/2 > 0$. In view of (3.35) it is enough to show that for some $c > 0$

$$c|u_{s1}|^2 + \frac{1}{2}|u_{s2}|^2 \ge (a_1|u_{s1}| + a_2|u_{s2}|)\left(cc_1|u_{s1}| + \frac{c_2}{2}|u_{s2}|\right), \tag{3.39}$$

where c_1 and c_2 are defined by (3.10). The last inequality of quadratic forms is equivalent to the inequalities of coefficients

$$a_1c_1 \le 1, \tag{3.40}$$

$$2d_1c \ge (d_2c + d_3)^2, \tag{3.41}$$

where

$$d_1 = (1 - a_1c_1)(1 - a_2c_2), \quad d_2 = c_1a_2, \quad \text{and} \quad d_3 = \frac{c_2a_1}{2}. \tag{3.42}$$

To satisfy (3.41) c should be

$$\begin{aligned} &\frac{d_4 - d_6}{d_2^2} \le c \le d_7 = \frac{d_4 + d_6}{d_2^2}, \qquad d_4 = d_1 - d_2d_3, \quad d_6 = \sqrt{d_5}, \\ &d_5 = d_4^2 - d_2^2d_3^2 = d_1(d_1 - 2d_2d_3). \end{aligned} \tag{3.43}$$

Finally, if the computed interval vector $\partial_s u(r_1; s), (s, \alpha) \in I_{\Delta s} \times I_{\Delta\alpha}$ satisfies

$$\frac{|\partial_s u_2(r_1; s)|^2}{|\partial_s u_1(r_1; s)|^2} < d_7, \tag{3.44}$$

we take $c = d_7$. This implies that $c|u_{s1}|^2 - |u_{s2}|^2 > 0$ for all $r > r_1$ and proves the tranvsersality. Our subroutine ESTIM also computed the above coefficients d_i and a_i and verified the inequalities (3.40) and (3.44). If transversality failed, the trajectory $y(r)$ was advanced in r until both the conditions of separation and of transversality were met. Thus the printout in Table 1 testifies both to the existence of the solution and the transversality of the intersection.

Let $M_{cs}(r_1^{-1}, \delta)$ be the restriction of M_{cs} to the domain $y_4 = r_1^{-1}$, $|u_2| < \delta$ with the parameter $\alpha \in I_{\Delta\alpha}$, where δ is the bound on $|u_2(r_1)|$ as it appears in the fourth column of Table 1. If du_1, du_2 is a tangent vector to $M_{cs}(r_1^{-1}, \delta)$ then our previous analysis shows that

$$\frac{|du_2|^2}{d_7} > |du_1|^2. \tag{3.45}$$

In view of (3.44) it follows that the curve $\bar{u}(r_1, s)$, $s \in I_{\Delta s}$ for each $\alpha \in I_{\Delta\alpha}$ has a unique intersection with $M_{cs}(r_1^{-1}, \delta)$. Note also that $M_{cs}(r_1^{-1}, \delta)$ is uniquely determined by the conditions

$$\|u_1\|_\infty \le c_1c_7, \quad \|u_2\|_\infty \le c_2c_7 + \delta, \quad \lim_{r \to \infty} u(r) = 0. \tag{3.46}$$

(Here the constants c_7 and δ replace the terms $\|f\|_\infty$ and $|u_2(r_1)|$ in (3.9).) Indeed, let $u(r), \tilde{u}(r)$ be two different solutions of (3.8) satisfying (3.46) such that $u_2(r_1) = \tilde{u}_2(r_1)$. In our analysis of the system (3.34) we can replace the function $u_s(r)$ by $\Delta u(r) = \tilde{u}(r) - u(r)$. Estimates of all constants a_i, d_i remain valid since the norm $\|y\|_\infty$ in (3.37) is replaced by the same bound

$$\tfrac{1}{2} \, \|y + \tilde{y}\|_\infty \le (c_1 + 2c_2)c_7 + 2\delta. \tag{3.47}$$

Clearly, $|\Delta u_2(r_1)|^2/|\Delta u_1(r_1)|^2 = 0 < d_7$, and hence for all $r > r_1$

$$|\Delta u_1(r)| > |\Delta u_2(r)|/d_7^{1/2}. \tag{3.48}$$

But for large r, $u(r)$ and $\tilde{u}(r)$ belong to a small neighborhood of zero where the central-stable manifold M_{cs} is unique and close to the tangent plane $u_1 = 0$. The last is contradicted by (3.48). Thus, the output in Table 1 shows that for each $\alpha \in I_{\Delta\alpha}$ there exists a unique $s^* \in I_{\Delta s}$ such that the corresponding solution of (1.8) in u coordinates satisfies conditions (3.46). This, however, does not mean that there are no other solutions with slopes s arbitrarily close to s^* that escape the neighborhood defined by (3.46) but ultimately converge to the point $\bar{y}_0$. In effect we demonstrated numerically in [5] that for $c \approx .835 \div 86$ and $c \approx .4845 - .4982$ equation (1.4) has homoclinic solutions, i.e., solutions which have the limits $y(\pm\infty) = -c\sqrt{2}$. Thus for corresponding values of α we may expect that (1.8) has infinitely many solutions with slopes $s_n = y'(0)$, $s_n \to s_0$ such that $\lim_{r\to\infty} y(r, s_n) \to -1$ and $y(r, s_0)$ is the negative solution established by Theorem 1.

4. Numerical solution of the O.D.E. in interval arithmetic. The power series expansion of §2 provides the box value of $\bar{y}$ and its partial derivatives with respect to s and α at $r = r_0 = 1$. On the other hand, the invariant manifold M_{cs} was estimated by us in §3 for $r \geq r_1 \approx 5 \div 7$. In order to connect between $r = r_0$ and $r = r_1$ equation (1.8) should be solved in the interval $[r_0, r_1]$ numerically. An O.D.E. could be solved in integral form as suggested by [2] or by an explicit Taylor method as in [3]. We prefer the second approach. Consider a general initial value problem for a system of differential equations

$$(4.1) \qquad y' = f(y), \quad y(r_0) = y_0, \quad y \in R^d.$$

Given $y(r)$ we compute $y(r+h)$ by the Taylor formula

$$(4.2) \qquad y(r+h) = y(r) + y'(r)h + \cdots + \frac{y^{(n)}(r)h^n}{n!} + E_n$$

with the remainder

$$(4.3) \qquad E_n = (E_{n1}, \cdots, E_{nd}), \qquad E_{ni} = \frac{y_i^{(n+1)}(r+\theta_i h)h^{n+1}}{(n+1)!},$$
$$0 < \theta_i < 1, \qquad 1 \leq i \leq d.$$

The coefficients $y^{(k)}(r)$, $1 \leq k \leq n$ are uniquely defined by $y(r)$ from (4.1). However, to obtain a rigorous estimate of $y(r+h)$ we must estimate the derivative $y^{(n+1)}$ in the interval $(r, r+h)$. Suppose we have a subroutine DERIV whose input consists of an interval value of r and a d-dimensional box value of $y(r)$, while the output consists of the box values of the derivatives $y^{(k)}(r)$, $1 \leq k \leq n+1$. Our algorithm below uses two calls to DERIV to compute $y(r+h)$. The algorithm could be split into the following 5 steps:

(1) Call DERIV with the input $r, y(r)$ and obtain the output $y^{(k)}(r), 1 \leq k \leq n$.

(2) Define the box

$$(4.4) \qquad y_{h0} = y(r) + FACT \cdot \sum_{k=1}^{m} y^{(k)}(r) \cdot [0, h]^k / k!, \qquad m \leq n,$$

where $FACT$ is a "safety" factor. We took $FACT = 1.5$, $m = 5$, and $n = 8$ or $n = 10$. The box y_{h0} is about to include all possible values of the vector $\{y_i(r_i)\}_{i=1}^d$, $r_i \in [r, r+h]$ corresponding to the true solution y.

(3) Call DERIV with the input $[r, r+h]$ instead of r and the box y_{h0} instead of $y(r)$. The resulting box values of the derivatives are denoted by $y_{h0}^{(k)}$, $1 \leq k \leq n+1$.

(4) Compute the box

$$y_h = \sum_{k=0}^{n} y^{(k)}(r)[0,h]^k/k! + y_{h0}^{(n+1)}[0,h]^{n+1}/(n+1)! \tag{4.5}$$

and check whether y_h lies in y_{h0}. If not, the program fails and the computation is stopped. If yes, go to Step 5.

(5) Compute $y(r+h)$ by

$$y(r+h) = \sum_{k=0}^{n} y^{(k)}(r)h^k/k! + y_{h0}^{(n+1)}h^{n+1}/(n+1)!. \tag{4.6}$$

The test in (4) proves that the remainder E_n in (4.3) is contained in the remainder of (4.6).

The main problem with interval computations is their exponential instability. Consider a linear system of O.D.E.'s with constant coefficients

$$y' = Ay. \tag{4.7}$$

If y is a d-dimensional symmetric box $[-\delta,\delta]^d$, then $Ay =$ abs$(A)y$, where abs$(A) = \{|a_{ij}|\}$. Hence the solutions of (4.7) when computed in interval arithmetic will grow exponentially even when the original matrix A has no positive eigenvalues. In the case of (3.4), if we disregard the term $f(y,r) = O(|\bar{y}|)\cdot O(r^{-1}+|\bar{y}|)$ then the resulting constant coefficients O.D.E. has positive eigenvalue λ_1 and complex eigenvalues λ_2, λ_3. If we turn to the characteristic variables u, the first component u_1 in (3.8) will grow with the rate $e^{\lambda_1 r}$, also in interval arithmetic, while the second component u_2 will grow as $e^{|\lambda_2|r}$, where really it is decreasing as $e^{\text{Re }\lambda_2 r} = e^{-\lambda_1 r/2}$. To prevent this exponential growth we switch to the real variables $z = (z_1, z_2, z_3)$,

$$z_1 = u_1, \qquad z_2 + iz_3 = e^{-i\text{Im }\lambda_2 r}u_2. \tag{4.8}$$

Then, instead of (3.8), variables z_2, z_3 will satisfy O.D.E.'s with the exponent Re λ_2. The precise transformation formulas from $\bar{y}$ to z are

$$u_1 = \sum_{j=0}^{2} c_{1j}y^{(j)}, \quad \text{Re } u_2 = \sum_{j=0}^{2} c_{2j}y^{(j)}, \quad \text{Im } u_2 = \sum_{j=0}^{2} c_{3j}y^{(j)}, \tag{4.9}$$

$$\begin{aligned} z_2 &= \cos\ \text{Im}\ \lambda_2 r \cdot \text{Re}\ u_2 + \sin\ \text{Im}\ \lambda_2 r \cdot \text{Im}\ u_2, \\ z_3 &= -\sin\ \text{Im}\ \lambda_2 r \cdot \text{Re}\ u_2 + \cos\ \text{Im}\ \lambda_2 r \cdot \text{Im}\ u_2 \end{aligned} \tag{4.10}$$

The variable y in (4.9) is, of course, the y_{new} of (3.4). In variable z equation (3.8) becomes

$$\begin{aligned} z_1' &= \lambda_1 z_1 + c_{13}f, \\ z_2' &= \text{Re}\ \lambda_2 \cdot z_2 + (c_{23}\cos\ \text{Im}\ \lambda_2 r + c_{33}\sin\ \text{Im}\ \lambda_2 r)f, \\ z_3' &= \text{Re}\ \lambda_2 \cdot z_3 + (c_{33}\cos\ \text{Im}\ \lambda_2 r - c_{23}\sin\ \text{Im}\ \lambda_2 r)f. \end{aligned} \tag{4.11}$$

The constants c_{ij} as follow from (3.7) are

$$
\begin{aligned}
&c_{11} = \frac{|\lambda_2|^2}{|\lambda_1 - \lambda_2|^2} = \frac{\lambda_1^2 + \alpha}{3\lambda_1^2 + \alpha}, \qquad c_{12} = \frac{-\lambda_2 - \lambda_3}{|\lambda_1 - \lambda_2|^2} = \frac{\lambda_1}{3\lambda_1^2 + \alpha},\\
&c_{13} = \frac{1}{3\lambda_1^2 + \alpha}, \qquad c_{21} = \operatorname{Re} \frac{\lambda_1 \lambda_3}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} = \frac{\lambda_1^2}{3\lambda_1^2 + \alpha},\\
&c_{22} = -\operatorname{Re} \frac{\lambda_1 + \lambda_3}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} = -\frac{\lambda_1}{2(3\lambda_1^2 + \alpha)},\\
(4.12)\quad &c_{23} = \operatorname{Re} \frac{1}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} = -\frac{1}{2(3\lambda_1^2 + \alpha)},\\
&c_{31} = \operatorname{Im} \frac{\lambda_1 \lambda_3}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} = \frac{\lambda_1 \alpha}{2(3\lambda_1^2 + \alpha)\operatorname{Im} \lambda_2},\\
&c_{32} = -\operatorname{Im} \frac{\lambda_1 + \lambda_3}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} = -\frac{(3\lambda_1^2/2 + \alpha)}{2(\lambda_1^2 + \alpha) \operatorname{Im} \lambda_2},\\
&c_{33} = \operatorname{Im} \frac{1}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} = \frac{3\lambda_1/2}{2(\lambda_1^2 + \alpha)\operatorname{Im} \lambda_2}.
\end{aligned}
$$

The higher-order derivatives of z are computed by (4.11):

$$
\begin{aligned}
&z_1^{(i+1)} = \lambda_1 z_1^{(i)} + c_{13} f^{(i)},\\
(4.13)\quad &z_2^{(i+1)} = \operatorname{Re} \lambda_2 \cdot z_2^{(i)} + \sum_{j=0}^{i} \binom{i}{j} (c_{23} \cos^{(i-j)} \operatorname{Im} \lambda_2 r\\
&\qquad\qquad + c_{33} \sin^{(i-j)} \operatorname{Im} \lambda_2 r) f^{(j)}
\end{aligned}
$$

and similarly for z_3. It was more convenient to compute the derivatives of f by differentiating (3.4) instead of expressing f in terms of z. The differentiation formulas appear in the listing of the computer program in the subroutine DERIV. For the convenience of the reader we present them here in mathematical notation:

$$
(4.14)\qquad (y^2)^{(i)} = \sum_{j=0}^{i} \binom{i}{j} y^{(j)} y^{(i-j)},
$$

$$
\begin{aligned}
f^{(i)} &= -\sum_{j=0}^{i} \binom{i}{j} (2y^{(j+2)} (r^{-1})^{(i-j)} - 2b_2 y^{(j)} (r^{-1})^{(i-j+1)}\\
(4.15)\quad &+ y^{(j+1)} (r^{-1})^{(i+j+1)} + y^{(j)} (r^{-1})^{(i+j+2)}/2) - (y^2)^{(i)} - (\alpha - \alpha_0) y^{(i+1)}\\
&- q_3 (r^{-1})^{(i+2)} - q_4 (r^{-1})^{(i+3)} - q_5 (r^{-1})^{(i+4)},
\end{aligned}
$$

where $b_2 = \alpha^2/8$ as in (3.21) and q_i are related to p_i in (3.23) as

$$
(4.16)\qquad \begin{aligned}
&q_3 = p_3/2 = -(1 + \alpha^3/4)/2, \qquad q_4 = -p_4/6 = -\alpha^4/384,\\
&q_5 = p_5/24 = -3\alpha^2/64.
\end{aligned}
$$

Finally, by (3.4)

$$y^{(i+3)} = -\alpha_0 y^{(i+1)} + 2y^{(i)} + f^{(i)}. \tag{4.17}$$

Recall that the term $-(\alpha - \alpha_0)y^{(i+1)}$ is absorbed in $f^{(i)}$. Thus we compute first the derivatives $f^{(i)}$ and $y^{(i+3)}$ for $i = 1, \cdots, n$ and then the derivatives $z^{(i)}$, $i = 1, \cdots, n+1$. The value of $z(r+h)$ is then computed by formula (4.6) with y replaced by z. Recall that Re $\lambda_2 = -\lambda_1/2 < 0$. If $z_2 + hz_2'$ were added in interval arithmetic as

$$z_2 + h \text{ Re } \lambda_2 z_2 + O(f) \tag{4.18}$$

then, for a symmetric interval $z_2 = (-\delta, \delta)$, the result would be $z_2(1+h|\text{Re } \lambda_2|)+O(f)$. Instead we add it as

$$z_2 + hz_2' = z_2(1 + h \text{ Re } \lambda_2) + (c_{23} \cos \text{ Im } \lambda_2 r + c_{33} \sin \text{ Im } \lambda_2 r)f. \tag{4.19}$$

The remaining part of the Taylor formula is added in a straightforward manner by Horner's method. As r increases and z becomes small, $z_2(1 + h \text{ Re } \lambda_2)$ becomes a leading term in the Taylor formula. As a result z_2 decreases exponentially. The component z_1 grows, of course, but with the correct rate $e^{\lambda_1 r}$. Once $z(r+h)$ has been computed, $y(r+h)$ is expressed in terms of $z(r+h)$. The explicit inverse transformation formulas are

$$\begin{aligned} y &= u_1 + 2\text{Re } u_2, \\ y' &= \lambda_1 u_1 + 2(\text{Re } \lambda_2 \cdot \text{ Re } u_2 - \text{ Im } \lambda_2 \cdot \text{ Im } u_2), \\ y'' &= \lambda_1^2 u_1 + 2(\text{Re } \lambda_2^2 \cdot \text{ Re } u_2 - \text{ Im } \lambda_2^2 \cdot \text{ Im } u_2), \\ u_1 &= z_1, \qquad \text{Re } u_2 = z_2 \cos \text{ Im } \lambda_2 r - z_3 \sin \text{ Im } \lambda_2 r, \\ \text{Im } u_2 &= z_2 \sin \text{ Im } \lambda_2 r + z_3 \cos \text{ Im } \lambda_2 r. \end{aligned} \tag{4.20}$$

Despite the above stabilization procedure, the computations with reasonable intervals $\Delta\alpha = .005$ and $\Delta s = .0015$ exploded before reaching $r_1 \approx 6$. Therefore, instead of solving the differential equations (3.4), (4.11) with $\alpha = I_{\Delta\alpha}$, $s = I_{\Delta s}$ we evaluated the function $y(r,s,\alpha)$ and $z(r,s,\alpha)$ by the Taylor formula of order 2:

$$\begin{aligned} y(r,s,\alpha) &= y(r,s_0,\alpha_0) + \partial_s y(r,s_0,\alpha_0)(s-s_0) + \partial_\alpha y(r,s_0,\alpha_0)(\alpha-\alpha_0) \\ &\quad + \sum_{j_1+j_2=2} \frac{\partial_s^{j_i}\partial_\alpha^{j_2}}{j_1!j_2!} y(r,s,\alpha)(s-s_0)^{j_i}(\alpha-\alpha_0)^{j_2} \end{aligned} \tag{4.21}$$

and similarly for z. Hence $y(r,s_0,\alpha_0)$ and its first derivatives were computed for the point values of $s = s_0, \alpha = \alpha_0$ and only the second derivatives $\partial_s^{j_1}\partial_\alpha^{j_2}y$, $j_1 + j_2 = 2$ were computed for the original interval values of s and α. With Δs and $\Delta\alpha$, as above, $(s-s_0)^{j_1}(\alpha-\alpha_0)^{j_2} \approx 10^{-6}$, so we could afford second-order derivatives of a magnitude $10^3 - 10^4$. The equations for the derivatives with respect to s and α are obtained by differentiating (4.13)–(4.16). The eigenvalues λ_i and the coefficients c_{ij} in (4.13) are constant since they are computed at $\alpha = \alpha_0$; however, $b_2 = \alpha^2/8$ and q_i at (4.16) depend on α. When differentiating the nonlinear term y^2 in (4.14) we obtain

$$(\partial_s^2 y^2)^{(i)} = 2\sum_{j=0}^{i} \binom{i}{j} (\partial_s^2 y^{(j)} y^{(i-j)} + \partial_s y^{(j)} \partial_s y^{(i-j)}) \tag{4.22}$$

and similarly for the other second derivatives. Here we have to know the values of $y^{(i)}$ and $\partial_s y^{(i)}$ for $(s, \alpha) \in I_{\Delta s} \times I_{\Delta\alpha}$. The former ones are computed by (4.21) and the latter ones by

$$\begin{aligned}\partial_s y^{(i)}(r, s, \alpha) = \partial_s y^{(i)}(r, s_0, \alpha_0) + \partial_s^2 y^{(i)}(r, s, \alpha)(s - s_0) \\ + \partial_s \partial_\alpha y^{(i)}(r, s, \alpha)(\alpha - \alpha_0)\end{aligned} \tag{4.23}$$

and similarly for $\partial_\alpha y^{(i)}$. These interval values also appear when $b_2 y^{(j)}$ and $(\alpha - \alpha_0) y^{(i+1)}$ in (4.15) are differentiated twice with respect to α or in the mixed derivative $\partial_s \partial_\alpha$. The precise differentiation formulas appear in the listing of our computer code in the Appendix, in the subroutine DERIV. As already mentioned in §2, the derivatives $\partial_s^{j_1} \partial_\alpha^{j_2} f$ of a generic function f are stored in our program as shown in (2.33). The derivatives $\partial_s^{j_1} \partial_\alpha^{j_2}, 0 \le j_1 + j_2 \le 1$ computed at the central point (s_0, α_0) and the second-order derivatives in $I_{\Delta s} \times I_{\Delta\alpha}$ carry the name $f0$, while the derivatives up to order 1 in $I_{\Delta s} \times I_{\Delta\alpha}$ carry the original name f. More detailed comments that explain the purpose of each variable appear in the listing of the program in the Appendix. Now, with the Taylor expansion in $(s - s_0)$ and $(\alpha - \alpha_0)$, our program could advance the solution up to $r_1 \approx 6 \div 8$ (depending on α) with $\Delta\alpha$ and Δs as above. The step size h in r was constant $h = \frac{1}{8}$. The order n of the Taylor method was 8 for $0 \le \alpha \le 1.7$ and 10 for $1.7 < \alpha \le 2.4$. The slope s_0 corresponding to α_0 was computed by a preliminary nonrigorous program so that

$$z_1(\tilde{r}_1, s_0, \alpha_0) = 0, \tag{4.24}$$

where $\tilde{r}_1$ is the expected value of r_1 at which the subroutine ESTIM should succeed in proving the transversality of the intersection. As mentioned in (2.18), we assumed that when differentiating with respect to α, the slope s is a linear function of α. The derivative s_α in (2.18) is set by the preliminary program so that

$$\partial_\alpha z_1(\tilde{r}_1, s_0, \alpha_0) = 0. \tag{4.25}$$

As a result,

$$z_1(r_1, s, \alpha) \approx \partial_s z_1(r_1, s_0, \alpha_0) \cdot (s - s_0), \tag{4.26}$$

so that

$$z_1(r_1, s_0 \pm \Delta s, \alpha) = \pm\, \partial_s z_1(r_1, s_0, \alpha_0) \cdot \Delta s. \tag{4.27}$$

Because of the exponential growth, $\partial_s z_1(r_1, s_0, \alpha_0)$ is quite large, of the order $40 \div 70$. The interval values of $z_1(r_1, s_0 - \Delta s, \alpha)$ and $z_1(r_1, s_0 + \Delta s, \alpha)$ appear in the last four columns of Table 1. Indeed they are almost opposite to each other and well separated for all α by the plane $z_1 = 0$. The subroutine ESTIM testifies that they are also separated by the manifold M_{cs}.

5. The computer program. Below we will explain in general the structure of our computer code BUNSEN. The listing of the program appears in the Appendix. The program is written in Fortran and consists of the main part and of nine subroutines. The main part reads the input prepared by the preliminary program. This input consists of the values s_0, α_0, and s_{α_0} denoted correspondingly by $S0$, $ALF0$, and $SA0$. It also defines the vector DP as explained in (2.36), computes the coefficients $b_2 = B2$ and $q_i = Q(i)$ and their derivatives with respect to s and α, and computes the derivatives $RD(i) = (r^{-1})^{(i)}$ and other constants. It also calls various subroutines in the following order:

(1) Subroutine CONST, which computes the coefficients $COEF(i,j) = c_{ij}$ as defined in (4.12), the eigenvalues $LAM(1) = \lambda_1$, $LAM(2) = \operatorname{Re}\lambda_2$, $LAM(3) = \operatorname{Im}\lambda_2$, $LAM2(1) = \lambda_1^2$, $LAM2(2) = \operatorname{Re}(\lambda_2^2)$, $LAM2(3) = \operatorname{Im}(\lambda_2^2)$, $LAM2(4) = (\operatorname{Im}\lambda_2)^2$.

(2) Subroutine SERIES, which returns the vector $Y0$ defined as $(Y0(j,k),\ Y0(j+1,k)) = \partial_s^{j_1}\partial_\alpha^{j_2}y^{(k)}(r_0)$, with $0 \le k \le 2$, $j_1+j_2 \le 2$, and j as in (2.33). For $j_1+j_2 = 2$ the corresponding values are computed in $I_{\Delta s} \times I_{\Delta\alpha}$, otherwise at the central point (s_0, α_0).

(3) Next, the main program transforms the vector $Y0$ from the original y variables to the new variable $y_{\text{new}} = y_{\text{old}} - y_N$, where y_N is defined by (3.21). After that the subroutine TRANS computes the corresponding vector $z0$ as defined by (4.9), (4.10).

(4) The five step algorithm presented in §4 calls twice the subroutine DERIV and computes the derivatives $y^{(i)}(r)$, $z^{(i)}(r)$, and the truncation error E_n of the Taylor series. The derivatives $\partial_s^{j_1}\partial_\alpha^{j_2}y^{(i)}(r)$ are stored in the vector $YD0(j,i)$ and similarly for z. In the second call to DERIV the corresponding values for the interval $[r, r+h]$ are stored in the vectors $YDH0(j,i)$ and $ZDH0(j,i)$. Since for E_n we need to know only the derivatives of order $i = n+1$, in the second call to DERIV the derivatives $ZDH0(j, n+1)$ are computed from $YDH0(j,i)$, $i \le n+1$ by formulas (4.9)–(4.10) and not by (4.11). The last parameter $IFLAG$ in the calling sequence of DERIV controls the above-mentioned option of computing $ZDH0(j, n+1)$. The transformation from $YDH0$ to $ZDH0$ is carried out by the subroutine TRANSD. After the second call to DERIV the main program verifies that $y_h \subset y_{h_0}$ (see (4.5)) and computes the vector $Z0$ for $r = r+h$. The inverse transformation from z to y as defined by (4.20) is carried out by the subroutine TRANSIN.

(5) $z(r)$ for $\alpha = I_{\Delta\alpha}$ and $s = I_{\Delta s}$ or $s = s_0 \pm \Delta s$ is computed. The results are stored in vectors z and $z1$ correspondingly. Then subroutine ESTIM is called. It checks whether z lies in the domain of M_{cs} and whether $z1$ is separated by M_{cs}. The subroutine ESTIM also verifies the transversality of the intersection. Step (5) is executed if r becomes greater than $RMAX1 = 5$. If the result of ESTIM is positive, the program reads the new input for a higher value of α and repeats the whole procedure. If not, the trajectory is advanced in r. If r exceeds a fixed value $RMAX = 8$ the program stops.

As already mentioned in §2 our program is processed by a compiler which we called COMPINT. All statements indicated by the letter I in the first column are considered to be interval statements. All variables declared by the INTERVAL type statements preceded by the label ID in the first two columns are considered to be interval variables. Likewise, all variables assigned their values by the I statements are treated by COMPINT as interval variables. The interval variables are always pairs of real machine numbers $x(1), x(2), x(1) \le x(2)$. They are stored in the successive order and named by the address of the first number $x(1)$. Thus, all interval variables are defined in our program by the DIMENSION statement as arrays of length 2, and in case of vector variables we add the first dimension of length 2. The exception are the partial derivatives $\partial_s^{j_1}\partial_\alpha^{j_2}y$ and $\partial_s^{j_1}\partial_\alpha^{j_2}z$, which are stored as explained in (2.33). This was done in order to reduce the number of dimensions to 3, which is the maximal dimension allowed in Fortran.

The compiler COMPINT translates all I statements into a sequence of elementary operations +, −, *, and / and replaces them by a call to corresponding subroutines SUM, DIF, MUL, and DIV. The names of the subroutines carry two additional digits

1 or 2. For example, MUL21(A, B, C) computes the product of an interval variable A with a real scalar variable B. Thus the digit 1 stands for scalar variable and 2 for the interval one. In all cases the result C is an interval variable. When encountering the expression $C = A * B$ the compiler checks the type of the variables A and B and then sets in the corresponding name of the subroutine. The listing of the most involved elementary subroutine MUL22(A, B, MUL) appears in the Appendix. The resulting interval MUL satisfies

$$(5.1) \qquad D(1)(1-\varepsilon) < MUL(1) < D(1), \qquad D(2) < MUL(2) < D(2)(1+\varepsilon),$$

where

$$(5.2) \qquad D(1) = \min(A(i), B(j)), \quad D(2) = \max(A(i), B(j)), \quad i, j = 1, 2$$

and

$$(5.3) \qquad \varepsilon = 2^{-47} + 2^{-48}.$$

When computing $D(1)$ and $D(2)$, because of symmetric rounding we obtain instead numbers $C1$ and $C2$. Then $MUL(1)$ and $MUL(2)$ are defined by

$$(5.4) \qquad MUL(1) = C1 - 2^{e_1 - 48}, \qquad MUL(2) = C2 + 2^{e_2 - 48},$$

where e_1 and e_2 are the exponents of $C1$ and $C2$ in the binary floating point representation. The number 48 here stands for 48-digit binary mantissa in our computer. The rounding in (5.4) may cause a relative error up to 2^{-47} with an additional maximal relative error of 2^{-48} in the symmetric rounding. The constants $2^{e_1 - 48}$, $i = 1, 2$ are formed by logical operation at the level of binary words $C1$ and $C2$. To adjust this and other subroutines to a different word length we must change only two logical constants. Besides the above four operations our library named INTAR (interval arithmetic) includes two subroutines ISIN22(X, Y), ICOS22(X, Y), which for an interval X compute the interval Y such that $Y \supset \sin X$ and $Y \supset \cos X$, correspondingly. In order to give a correct result the length $x(2) - x(1)$ of X should in the worst case be less than 0.1. In our computation these programs were called with intervals $X = \operatorname{Re} \lambda_2 \cdot r$ of negligible length. The content of the library INTAR as well as the compiler COMPINT are not listed in this publication and can be obtained from the author by request. Two more interval arithmetic subroutines ISQRT22 and CUBEQ appear in the listing of our program BUNSEN. The first finds a square root of an interval and the second solves a cubic equation with interval coefficients. We need the latter to solve (1.13) and (2.10).

The parts of the program and the program as a whole were extensively checked. As a practical and most convincing test we should mention that (1.8) was solved independently by Taylor's method in the original variables y. In this case the algorithm is very simple and short. The results, however, agreed completely with the box values obtained by the program BUNSEN.

6. Conclusion. The Bunsen flame solutions are a particular though physically important case of two-dimensional structures described by the Kuramoto–Sivashinsky equation. In this paper we managed to prove the existence of these solutions for the physically relevant values of parameters. There are also other types of radial solutions which tend to periodic, quasi-periodic, or chaotic solutions of the O.D.E (1.5) as $r \to \infty$. A rigorous computer assisted proof of existence of such solutions seems to be a hard task even with modern supercomputers. Laboratory experiments have shown that Bunsen flames may rotate with a constant speed. The stationary solution we found is closely related to the leading term in the Fourier series expansion of the rotating flame with respect to the polar angle φ. It now seems possible to justify the existence of such rotating solutions, at least on the physical level of rigor.

7. Appendix. The computer program BUNSEN is not published here due to its length of 700 lines. It can be obtained from the author by electronic mail.

REFERENCES

[1] J.-P. Eckmann and P. Wittwer, *A complete proof of the Feigenbaum conjectures*, J. Stat. Phys., 46 (1987), pp. 455–475.

[2] E. W. Kaucher and W. L. Miranker, *Self-validating Numerics for Function Space Problems*, Academic Press, New York, 1984.

[3] R. de la Llave and C. Fefferman, *Relativistic stability of matter*, I, Rev. Math. Iberoamericana, 1–2 (1986), pp. 119–213.

[4] O. E. Lanford III, *A computer-assisted proof of the Feigenbaum conjectures*, Bull. Amer. Math. Soc. (NS), 6 (1982), pp. 427–434.

[5] D. Michelson, *Steady solutions of the Kuramoto–Sivashinsky equation*, Phys., 19D (1986), pp. 89–111.

[6] G. Sivashinsky, *Nonlinear analysis of hydrodynamic instability in laminar flames*, Acta Astronautica, 4 (1977), pp. 1117–1206.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 387–411, March 1992

006

LARGE-TIME BEHAVIOR OF A TIME-PERIODIC COOPERATIVE SYSTEM OF REACTION-DIFFUSION EQUATIONS DEPENDING ON PARAMETERS*

PETER TAKÁČ†

Abstract. A kind of *structural stability* with respect to a parameter $\theta \in \Theta$ for a generic strongly monotone discrete-time dynamical system $\{T_\theta^n : X \longrightarrow X;\ n \in \mathbb{Z}_+\}$ is studied. Here, X and Θ are strongly ordered spaces, and the mapping $(x, \theta) \longmapsto T_\theta x$ from $\mathcal{X}$ into X is assumed to be continuous, strongly monotone and satisfying a compactness hypothesis. A classification of structurally stable points in $\mathcal{X}$ is introduced; the set of all such points is denoted by $\mathcal{S}$. *No* hyperbolicity hypothesis is assumed. If $\mathcal{X}$ is an open subset of a strongly ordered separable Banach space $\mathcal{V}$, it is proved that (1) $\mu(\mathcal{X} \setminus \mathcal{S}) = 0$ for every Gaussian measure μ on $\mathcal{V}$; (2) $(x, \theta) \in \mathcal{S}$ implies $\omega_\theta(x) \times \{\theta\} \subset \mathcal{S}$, where $\omega_\theta(x)$ denotes the ω-limit set of $x \in X$ under the semigroup $\{T_\theta^n : n \in \mathbb{Z}_+\}$; and (3) $\omega_\theta(x)$ is a "quasi cycle" for T_θ whenever $(x, \theta) \in \mathcal{S}$. These results are applied to a very general strictly cooperative time-periodic system of weakly coupled reaction-diffusion equations with (space- and/or time-dependent) parameters in both the reaction functions and Robin's boundary conditions. Here T_θ is the period map.

Key words. reaction-diffusion equation, strictly cooperative system, strongly monotone mapping, period map, structural stability, quasi cycle, ω-limit set

AMS(MOS) subject classifications. 35B30, 35B40, 35K55, 47H07

Introduction. This work deals with the large-time asymptotic behavior of a very general time-periodic system of reaction-diffusion equations which depend upon certain parameters. We restrict ourselves to those systems which have the following *monotonicity* property:

(M) The solution $u(t)$ (where $t \geq 0$ is the time variable) depends monotonically upon its initial value $u(0) = u_0$ and the parameter(s) θ.

More precisely, we assume that the solution $u(t) \equiv u(t, u_0, \theta)$ belongs to an ordered phase space $(X, \leq)$ and the parameter θ belongs to an ordered parameter space $(\Theta, \leq)$. Typically, X and Θ are suitable open subsets of ordered Banach spaces whose positive cone has nonempty interior (cf. Schaefer [23]). Then, given any $t \geq 0$, the mapping $u(t, \cdot, \cdot) \colon X \times \Theta \longrightarrow X$ is called *monotone* if

$$u_0 \leq u_0' \text{ in } X \quad \text{and} \quad \theta \leq \theta' \text{ in } \Theta \Longrightarrow u(t, u_0, \theta) \leq u(t, u_0', \theta') \quad \text{in } X.$$

We are interested in how the asymptotic behavior as $t \longrightarrow \infty$ of $u(t, u_0, \theta)$ depends upon u_0 and θ. In particular, we focus our study on (some sort of) *structural stability* of our monotone dynamical system with respect to the parameter θ. This problem can also be viewed as an *input-output* problem where (u_0, θ) represents the input, and $u(t, u_0, \theta)$ is the output, with the output depending monotonically and continuously upon the input, for each $t \geq 0$. Monotone dynamical systems without parameters have been studied recently by a number of authors, cf. Alikakos, Hess, and Matano [1], Chen and Matano [7], Hess [10], Hirsch [11]–[15], Matano and Mimura [17], Poláčik [21], Smith [25], Smith and Thieme [26], [27], Takáč [28]–[30], and others. Generically, autonomous monotone dynamical systems show convergence to an equilibrium (or a set of equilibria) for "almost every" relatively compact semiorbit. Under some additional restrictions the corresponding conclusion also holds for time-periodic

* Received by the editors March 5, 1990; accepted for publication (in revised form) April 29, 1991. This research was supported in part by National Science Foundation grant DMS-8802646.

† Mathematics Department, Vanderbilt University, Nashville, Tennessee 37240.

or discrete-time systems. However, until now the only work with parameters and monotonicity playing a crucial rôle is that of Thron [31], [32] who has studied simple input-output problems in pharmacokinetics by numerical simulations. In [31] a compartmental model deals with patients on a regular (time-independent) drug dosage schedule where drug dosage represents the input, and drug concentration (in a cell) is the output. It is conjectured that in such autonomous dynamical systems monotonicity with respect to certain parameters should prevent oscillatory (time-periodic) or chaotic large-time behavior in drug concentrations. The problem of stability of equilibria (steady states in drug concentrations) upon changes in drug dosage is mentioned as well. Our present article is the very beginning of rigorous analytical treatment of monotone dynamical systems depending (monotonically) upon parameters, with *no* hyperbolicity assumption of any kind.

The main objective of this article is to show that, in a very general time-periodic monotone dynamical system with parameters, "almost every" input $(u_0, \theta) \in X \times \Theta$ has the following two properties:

(1) Uniformly in time $t \in [0, \infty)$, the output $u(t, u_0, \theta)$ depends continuously upon the input varying near $(u_0, \theta) \in X \times \Theta$ (structural stability).

(2) The output $u(t, u_0, \theta)$ can be approximated by nearby periodic orbits, uniformly in time $t \in [t_0, \infty)$, where $t_0 \in [0, \infty)$ is sufficiently large (near periodic large-time behavior).

The kinds of applications of our results we have in mind include (a) chemical and biochemical reactions among diffusing substances in a vessel (e.g., a biochemical control circuit for enzymes metabolizing a drug in a cell) (cf. Othmer [19] and Thron [31]), (b) cooperative migrating populations with diffusive migration within a common region as studied in population biology, epidemiology, and ecology (cf. Fife [8] and Hirsch [12], [14]), and (c) two competing populations with diffusive migration within a region where they compete for resources, with competition, diffusion, and resource availability observing time-periodic seasonal oscillations (cf. Matano and Mimura [17]). Mathematical formulation of such models typically results in a time-periodic cooperative system of reaction-diffusion equations which may depend on certain parameters. A special case of such a system is a spatially homogeneous cooperative system of ordinary differential equations, a frequently used approximation. Below we present two problems, (**P**) and (**P′**), on which we can explain applications of our results.

Problem (**P**). We consider the following time-periodic, strictly cooperative system of reaction-diffusion equations with (spatially and/or temporally dependent) parameters in both the reaction-diffusion equations and the boundary conditions:

$$(\mathbf{P})\qquad \begin{aligned} &\frac{\partial u_k}{\partial t} - d_k(x,t)\Delta u_k = f_k(x,t,u_1,\cdots,u_n) + \sum_{\ell=1}^{n} \gamma_{k\ell}(x,t)u_\ell \quad \text{in } \Omega \times (0,\infty);\\ &\frac{\partial u_k}{\partial n} + \vartheta_k(x,t)u_k = 0 \qquad \text{on } \partial\Omega \times (0,\infty);\\ &\qquad u_k(x,0) = u_{k,0}(x) \qquad \text{in } \Omega. \end{aligned}$$

Here, $n \in \mathbb{N}$ is the number of equations indexed by $k = 1, 2, \cdots, n$, which are considered in an open bounded domain $\Omega \subset \mathbb{R}^N$ with the boundary $\partial\Omega$ of class C^3. The diffusion phenomenon is modelled by the N-dimensional Laplacian Δ with the diffusion coefficient $d_k(x,t) \geq d_k^* > 0$, $(x,t) \in \Omega \times (0,\infty)$. The systems of reaction functions f_k and $\gamma_{k\ell}$, $1 \leq k$, $\ell \leq n$, are assumed to be *cooperative* and *strictly* *cooperative*, respectively:

$$\frac{\partial f_k}{\partial u_l} \geq 0 \quad \text{and} \quad \gamma_{k\ell} > 0 \quad \text{whenever} \quad k \neq \ell.$$

We consider Robin's boundary conditions with $\vartheta_k(x,t) \geq \vartheta_k^* > 0$, $(x,t) \in \partial\Omega \times (0,\infty)$. All functions d_k, f_k, $\gamma_{k\ell}$, and ϑ_k are assumed to be *τ-periodic* in time $t \in \mathbb{R}^1_+$, where $\tau \in (0,\infty)$, and of class C^2 in their respective domains:

$$d_k : \overline{\Omega} \times \mathbb{R}^1_+ \longrightarrow [d_k^*, \infty), \qquad f_k : \overline{\Omega} \times \mathbb{R}^1_+ \times \mathbb{R}^n \longrightarrow \mathbb{R}^1,$$
$$\gamma_{k\ell} : \overline{\Omega} \times \mathbb{R}^1_+ \longrightarrow \mathbb{R}^1, \qquad \text{and} \qquad \vartheta_k : \partial\Omega \times \mathbb{R}^1_+ \longrightarrow [\vartheta_k^*, \infty).$$

The functions d_k and f_k are assumed to be fixed, whereas $\gamma \equiv (\gamma_{k\ell})_{k,\ell=1}^n$ and $\vartheta \equiv (\vartheta_k)_{k=1}^n$ are *parameters* varying in suitable open subsets of the Banach spaces

$$\tilde{V}_\gamma = C^2(\overline{\Omega} \times (\mathbb{R}^1/\tau\mathbb{Z}) \longrightarrow \mathbb{R}^{n\times n}) \quad \text{and} \quad \tilde{V}_\vartheta = C^2(\partial\Omega \times (\mathbb{R}^1/\tau\mathbb{Z}) \longrightarrow \mathbb{R}^n),$$

respectively. We denote all these parameters by $\theta = (\gamma, \vartheta) \in \Theta \subset \tilde{V}_\Theta = \tilde{V}_\gamma \times \tilde{V}_\vartheta$, where Θ is a suitable open subset of the strongly ordered Banach space $\tilde{V}_\Theta$. Of course, we consider the natural *pointwise* and *coordinatewise* ordering in all our function spaces. In particular, $\theta \leq \theta'$ in $\tilde{V}_\Theta$ is equivalent to: $\gamma_{k\ell} \leq \gamma'_{k\ell}$ in $\overline{\Omega} \times \mathbb{R}^1_+$ and $\vartheta_k \leq \vartheta'_k$ in $\partial\Omega \times \mathbb{R}^1_+$, where $\theta = (\gamma, \vartheta)$ and $\theta' = (\gamma', \vartheta')$ have coordinates indexed by k, $\ell = 1, \cdots, n$. The positive cone $(\tilde{V}_\Theta)_+ = \{\theta \in \tilde{V}_\Theta : \theta \geq 0\}$ in $\tilde{V}_\Theta$ has nonempty interior $\operatorname{Int}((\tilde{V}_\Theta)_+)$ which is equivalent to saying that $\tilde{V}_\Theta$ is *strongly ordered.*

Problem (**P**$'$). A particularly interesting example, which we owe to C. Dennis Thron [31], [32], of a (spatially homogeneous) time-periodic, strictly cooperative system of ordinary differential equations with (temporally dependent) parameters arises in pharmacology as a dynamical problem in drug metabolism:

$$(\mathbf{P}') \qquad \begin{aligned} &\frac{du_k}{dt} = f_k(t, u_1, \cdots, u_n, v_1, \cdots, v_{n'}) + \gamma_k(t) && \text{in } (0,\infty), \\ &\frac{dv_{k'}}{dt} = g_{k'}(t, u_1, \cdots, u_n, v_1, \cdots, v_{n'}) && \text{in } (0,\infty); \\ &u_k(0) = u_{k,0}, \\ &v_{k'}(0) = v_{k',0}. \end{aligned}$$

Here, $u_k : [0,\infty) \longrightarrow [0,\infty)$ denotes the (spatially homogeneous) concentration of the kth drug $(1 \leq k \leq n)$, $v_{k'} : [0,\infty) \longrightarrow [0,\infty)$ denotes the concentration of the k'th enzyme $(1 \leq k' \leq n')$, and $\gamma_k : [0,\infty) \longrightarrow [0,\infty)$ is the infusion rate of the kth drug. We are dealing with a compartmental model of a metabolic system of $(n + n')$ interacting drugs and enzymes occupying a single cell. Their interaction (e.g., induction of enzymes which metabolize drugs, synthesis of enzymes, etc.) is described by the system of (smooth) reaction functions $f_k : [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^{n'} \longrightarrow \mathbb{R}$ and $g_{k'} : [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^{n'} \longrightarrow \mathbb{R}$, which is assumed to be strictly cooperative by the biological nature of the interaction, i.e., all off-diagonal entries in the corresponding Jacobian matrix of the mapping $\left((f_k(t, \cdots))_{k=1}^n, (g_{k'}(t, \cdots))_{k'=1}^{n'}\right) : \mathbb{R}^n \times \mathbb{R}^{n'} \longrightarrow \mathbb{R}^n \times \mathbb{R}^{n'}$ in (**P**$'$) are > 0, for each $t \in [0,\infty)$. Particular forms of these reaction functions are obtained from intercompartmental fluxes and their dependence upon concentrations of drugs and enzymes; cf. Thron [31] for biological arguments. The drugs are injected in equal time-periods $\tau > 0$, and also all reaction functions are assumed to be τ-periodic in time t. For instance, we can consider the following simple *enzyme-substrate* system ($n = n' = 1$):

$$\begin{aligned} \frac{du}{dt} &= -c_1 u(E-v) + c_2 v + \gamma(t) \quad \text{for } t \in (0,\infty), \\ \frac{dv}{dt} &= c_1 u(E-v) - (c_2+c_3)v \quad \text{for } t \in (0,\infty). \end{aligned}$$

Here u, E, and v denote the (nonnegative) concentrations of substrate, total enzyme, and enzyme-substrate complex, respectively, and $\gamma(t)$ is the substrate influx rate which is considered to be τ-periodic. All constants E, c_1, c_2, and c_3 are assumed to be positive. (Of course, our results apply also to the case when E, c_1, c_2, and c_3 are known continuous τ-periodic functions of time t.) This system is strictly cooperative at any time $t \in [0,\infty)$ when its solution $\big(u(t), v(t)\big)$ satisfies $0 \le u(t) < \infty$ and $0 \le v(t) < E$.

Problem ($\mathbf{P}'$) poses an input-output problem in pharmacokinetics, where the infusion rates $\gamma \equiv (\gamma_k)_{k=1}^n$ are parameters representing time-periodic input, and the concentrations $(u,v) \equiv \big((u_k)_{k=1}^n, (v_k)_{k'=1}^{n'}\big)$ are solutions of ($\mathbf{P}'$) representing output. Of interest is the dependence of the output upon the input. Continuity of this dependence is closely related to structural stability of the dynamical system ($\mathbf{P}'$), with respect to the parameter γ varying in a suitable open subset Θ of the Banach space $\tilde{V}_\Theta = C(\mathbb{R}^1/\tau\mathbb{Z} \longrightarrow \mathbb{R}^n)$.

We now return to our first problem ($\mathbf{P}$); problem ($\mathbf{P}'$) can be treated analogously. We denote by $u \equiv (u_k)_{k=1}^n : \Omega \times \mathbb{R}_+^1 \longrightarrow \mathbb{R}^n$ the (unique) L_p-*solution* of our problem ($\mathbf{P}$) with the initial value $u_0 \equiv (u_{k,0})_{k=1}^n : \Omega \longrightarrow \mathbb{R}^n$, for a fixed $p \in (2N,\infty)$, cf. Amann [2], [3], [5]. The underlying space X for the solution $u(\,\cdot\,,t) \in X$, $t \in \mathbb{R}_+^1$, of ($\mathbf{P}$) is a suitably chosen open subset $X \subset \mathrm{Int}_V(V_+)$ of the interior of the positive cone V_+ in the strongly ordered Banach space $V = W_p^1(\Omega \longrightarrow \mathbb{R}^n)$, cf. Triebel [33]. This choice of X is justified by our applications, namely, u_k is the density of the kth component in chemical and biochemical reactions, population biology, epidemiology, ecology, etc., cf. Fife [8], Hirsch [12], [14], Matano and Mimura [17], and Othmer [19], and numerous references therein. If both X and Θ are suitably matched with the given system of reaction functions f_k, $1 \le k \le n$, then the solution $u(\,\cdot\,,t) \in X$ exists globally in time, i.e., for all $t \in \mathbb{R}_+^1$, whenever $u_0 \in X$ and $\theta \in \Theta$ are given. We set $\mathcal{X} = X \times \Theta$ and want to investigate the asymptotic behavior as $t \longrightarrow \infty$ of $u(\,\cdot\,,t) \equiv u(t) \in X$ depending upon $(u_0,\theta) \in \mathcal{X}$. Therefore, we write $u = u(t,u_0,\theta)$ to indicate this dependence.

We define the corresponding *period map* $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ by $\mathcal{T}(u_0,\theta) = \big(u(\tau,u_0,\theta),\theta\big)$ for all $(u_0,\theta) \in \mathcal{X}$. We will study the large-time asymptotic behavior of the solution u in terms of the asymptotic behavior of the iterates $\mathcal{T}^n$, $n \in \mathbb{Z}_+$, of $\mathcal{T}$ as $n \longrightarrow \infty$ applied to the points $(u_0,\theta) \in \mathcal{X}$. More precisely, we are interested in the ω-*limit sets* for the sequences $\mathcal{T}^n(u_0,\theta)$, $n \in \mathbb{Z}_+$, which are defined for every $(u_0,\theta) \in \mathcal{X}$ by

$$\begin{aligned} \omega(u_0,\theta) = \{(w,\theta) \in \mathcal{X} :\ & u(n_k\tau,u_0,\theta) \longrightarrow w \text{ in } X \ (k \longrightarrow \infty) \\ & \text{for some sequence } n_k \longrightarrow \infty \text{ in } \mathbb{Z}_+\}. \end{aligned}$$

To guarantee all the usual interesting properties of these ω-limit sets, e.g., $\omega(u_0,\theta)$ is nonempty, compact, and totally invariant under $\mathcal{T}$, we make use of the following two properties of the mapping $\mathcal{T}$ which can be derived from a number of results of Amann [5, Thm. 7.3] under appropriate hypotheses on f_k, X, and Θ:

(a) $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ is continuous;

(b) If $u_0 \in X$ and $\Sigma \subset \Theta$ is bounded in $\tilde{V}_\Theta$, then the closure in V of the set $u(\tau\mathbb{Z}_+, u_0, \Sigma) \equiv \{u(n\tau, u_0, \theta) : (n, \theta) \in \mathbb{Z}_+ \times \Sigma\}$ is compact in X.

In Example 4.1 we state the precise hypotheses under which both (a) and (b) are valid for the case of a single reaction-diffusion equation ($n = 1$). The reader can easily generalize these hypotheses to a system of n equations using Example 4.6.

To investigate the large-time asymptotic behavior of the solution $u(t, u_0, \theta)$ of $(\mathbf{P})$ we take advantage of the *strong monotonicity* of $u(t, \cdot, \cdot)$ with respect to $(u_0, \theta) \in \mathcal{X}$, for any fixed $t \in (0, \infty)$, which is a direct consequence of our choice of $X \subset \operatorname{Int}_V(V_+)$, $V \subset C(\overline{\Omega} \longrightarrow \mathbb{R}^n)$, and the strong maximum and boundary point principles for strictly cooperative (weakly coupled) parabolic systems, cf. Protter and Weinberger [22, Chap. 3, §8].

(c) If $u_0 \le u_0'$ in X and $\theta \le \theta'$ in Θ satisfy $(u_0, \theta) \ne (u_0', \theta')$, then

$$u(t, u_0', \theta') - u(t, u_0, \theta) \in \operatorname{Int}_V(V_+) \quad \text{holds for every} \quad t \in (0, \infty).$$

For v, $v' \in V$ we write $v \ll v'$ if and only if $v' - v \in \operatorname{Int}_V(V_+)$, and call the relation "$\ll$" the *strong ordering* in V. We use the same notation for the strong orderings in V_Θ and $\mathcal{V} = V \times V_\Theta$. Observe that each of the spaces V, V_Θ, and $\mathcal{V}$ is continuously imbedded into the space $C(K)$ of all continuous functions $\varphi : K \longrightarrow \mathbb{R}^1$ over a compact Hausdorff space K, endowed with the pointwise ordering "$\le$," and hence, for φ, $\varphi' \in C(K)$ we have $\varphi \ll \varphi'$ if and only if $\min_K(\varphi' - \varphi) > 0$.

Using properties (a), (b), and (c) of the mapping $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$, in §2 we define very essential *lower* and *upper stability* notions for an arbitrary point $(u_0, \theta) \in \mathcal{X}$ under the discrete-time semigroup $\{\mathcal{T}^n : n \in \mathbb{Z}_+\}$ acting on $\mathcal{X}$. In Lemma 2.3 we show that our stability notion is, in fact, equivalent to the classical notion of *Lyapunov stability*. This stability incorporates also continuous dependence of our discrete-time dynamical system $\{\mathcal{T}^n : \mathcal{X} \longrightarrow \mathcal{X};\ n \in \mathbb{Z}_+\}$ upon the parameter $\theta \in \Theta$, which reflects the *structural stability* of our system $(\mathbf{P})$ with respect to the parameter $\theta \in \Theta$. The set $\mathcal{S}$ of all stable points in $\mathcal{X}$ can be quite complicated even if we consider the dynamical system generated by a single autonomous ordinary differential equation with a scalar parameter, cf. Example 4.10. We would like to point out once again that *no* hyperbolicity of any kind is assumed in this article.

The most important parts of our main results, Theorems 3.3 and 3.4, can be stated for the problem $(\mathbf{P})$ as follows.

THEOREM 0.1. *Let $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ be the period τ map for the problem* $(\mathbf{P})$, *and assume it satisfies* (a), (b), *and* (c). *Then $\mathcal{U} \equiv \mathcal{X} \setminus \mathcal{S}$ is a Borel subset of $\mathcal{V} = V \times V_\Theta$, a strongly ordered separable Banach space, and $\mu(\mathcal{U}) = 0$ for every Gaussian measure μ on $\mathcal{V}$.*

If $(u_0, \theta) \in \mathcal{S}$ then also $\omega(u_0, \theta) \subset \mathcal{S}$, and $\omega(u_0, \theta)$ is a "quasi cycle" for $\mathcal{T}$ which can be approximated in $\mathcal{X}$ from below and from above, respectively, by monotone sequences of (*true*) *cycles $\omega(v_n, \theta)$ and $\omega(w_n, \theta)$ for $\mathcal{T}$, as $n \longrightarrow \infty$ in $\mathbb{N}$:*

$$v_n \le v_{n+1} \le w_{n+1} \le w_n \ \textit{in}\ X;$$

$$\mathcal{T}^{k_n}(v_n, \theta) = (v_n, \theta) \ \textit{and}\ \mathcal{T}^{\ell_n}(w_n, \theta) = (w_n, \theta) \quad \textit{for some}\ k_n,\ \ell_n \in \mathbb{N};$$

and

$$\omega(u_0,\theta) = \bigcap_{m=1}^{\infty} \mathrm{Cl} \bigcup_{n=m}^{\infty} \omega(v_n,\theta) = \bigcap_{m=1}^{\infty} \mathrm{Cl} \bigcup_{n=m}^{\infty} \omega(w_n,\theta)$$

with "Cl" *denoting the* (*compact*) *closures in* $\mathcal{X}$.

Loosely speaking, this theorem states that the ω-limit set $\omega(u_0,\theta)$ of almost every point $(u_0,\theta) \in \mathcal{X}$ is a stable *quasi cycle* for the period map $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$, meaning that $\omega(u_0,\theta)$ can be approximated by nearby ω-limit sets which are true cycles; see Definition 3.1. We prove this and other results by developing a suitable new theory for the large-time asymptotic behavior of very general discrete-time dynamical systems of the form $\{\mathcal{T}^n : \mathcal{X} \longrightarrow \mathcal{X};\ n \in \mathbb{Z}_+\}$, where $\mathcal{X} = X \times \Theta$, X and Θ are strongly ordered spaces, $\mathcal{T}(x,\theta) \equiv (T_\theta x, \theta)$ for $(x,\theta) \in \mathcal{X}$, and the mapping $(x,\theta) \longmapsto T_\theta x$ from $\mathcal{X}$ into X is continuous and strongly increasing. This theory generalizes a number of earlier results of the author [28], [30].

Without parameters, analogous results have been obtained for strongly increasing continuous-time semiflows by Hirsch [12], [14], Poláčik [21], Smith and Thieme [26], [27], and Takáč [30], and for time-periodic semilinear heat equations in one space dimension by Chen and Matano [7]. Structural stability of systems of weakly coupled reaction-diffusion equations with a single scalar parameter $\theta \in [0,\infty)$ in Robin's boundary conditions has been investigated by Hale and Rocha [9] who assumed large diffusivities d_k and hyperbolicity of all equilibrium solutions, but no cooperativeness hypothesis. They studied autonomous systems by showing the upper semicontinuity of their attractors with respect to d_k and θ, and then applying bifurcation methods. Although the importance of structural stability with respect to certain parameters is well known, we would like to emphasize another application of Theorem 0.1 and several other results in this article, namely, simple numerical simulations of the large-time asymptotic behavior of strongly increasing dynamical systems. Since roundoff computer errors may actually change the prescribed parameters, it is necessary to look for some kind of continuous dependence of the ω-limit sets or attractors upon these parameters to make sure that the computed results remain close to the true (precise) asymptotic behavior of the given system for *all* times.

This article is organized as follows. In §1 we generalize a number of basic preliminary results from Takáč [28], [30]. In §2 we introduce the lower and upper ω-limit sets which, in turn, play the key rôle in our definitions of lower and upper ω-stable points; these points form the sets $\mathcal{S}_-$ and $\mathcal{S}_+$, respectively. The equivalence of Ljapunov stability with ours is shown in Lemma 2.3. We describe some elementary properties of the sets $\mathcal{S}_-$, $\mathcal{S}_+$ and $\mathcal{U}_- = \mathcal{X} \setminus \mathcal{S}_-$, $\mathcal{U}_+ = \mathcal{X} \setminus \mathcal{S}_+$ in Theorems 2.4 and 2.5, respectively. In §3 we first define quasi cycles (Definition 3.1) and then state and prove our main results, Theorems 3.3, 3.4, and 3.6. Finally, in §4 we present three examples, Examples 4.1, 4.6, and 4.10, to which we apply our main results. The reader may find numerous additional applications of our results, for instance, to systems of delay equations or other functional differential equations; cf. Smith [25].

1. Families of semigroups: basic results. This section is devoted primarily to a generalization of the basic results from Takáč [28, §§1 and 2] and Takáč [30, §§1 and 2]. In these articles the author investigated a single strongly increasing continuous mapping $T : X \longrightarrow X$ in a strongly ordered space X, whereas in the present article a strongly increasing continuous family $\mathcal{T} \equiv \{T_\theta : \theta \in \Theta\}$ of such mappings is considered. Here Θ is a strongly ordered space of parameters. In particular, we are interested in the asymptotic behavior as $n \longrightarrow \infty$ of the discrete-time semigroups

$\{T_\theta^n : n \in \mathbb{Z}_+\}$ depending upon $\theta \in \Theta$, where $\mathbb{Z}_+ = \{0, 1, 2, \cdots\}$. Throughout the entire paper we assume the following hypotheses (X), (V), (Θ), (V_Θ), and $(\mathcal{T})$.

(X): X is a *strongly ordered space*, i.e., X is a metrizable topological space with a closed partial order relation "$\leq$" satisfying the following axiom, for every open subset U of X;

(SO1) If $x \in U$ then $a \ll x \ll b$ for some a, $b \in U$.

Here $x \ll y$ for x, $y \in X$ means that (x, y) belongs to the interior of the order relation in $X \times X$. We write $x < y$ if $x \leq y$, $x \neq y$. It is easy to see that for every open subset U of X, (SO1) implies the following.

(SO2) If a, $b \in U$ and $a \ll b$ then $a \ll x \ll b$ for some $x \in U$.

(V): V is a *strongly ordered vector space*, i.e., V is a metrizable topological vector space whose order relation is defined by a closed cone $V_+ = \{x \in V : x \geq 0\}$ with nonempty interior denoted by $\mathrm{Int}(V_+)$. (In some of our results we will assume that X is a nonempty open subset of V with closure $\mathrm{Cl}(X)$.)

(Θ): Θ is another strongly ordered space whose ordering is denoted by $\leq$ again.

(V_Θ): V_Θ is another strongly ordered vector space with the positive cone $(V_\Theta)_+ = \{\theta \in V_\Theta : \theta \geq 0\}$. (Again, $\Theta \subset V_\Theta$ nonempty and open will be assumed when needed.)

Observe that also the product space $\mathcal{X} = X \times \Theta$ is strongly ordered by $(x, \theta) \leq (y, \sigma) \overset{\text{def}}{\Longleftrightarrow} x \leq y$ (in X) and $\theta \leq \sigma$ (in Θ). Then $(x, \theta) \ll (y, \sigma)$ in $\mathcal{X}$ if and only if $x \ll y$ and $\theta \ll \sigma$. A similar statement is valid for $\mathcal{V} = V \times V_\Theta$. Our last hypothesis is as follows.

$(\mathcal{T})$: $\mathcal{T}_X$ is a continuous, strongly increasing mapping from $\mathcal{X}$ into X, i.e., $(x, \theta) < (y, \sigma)$ in $\mathcal{X}$ implies $\mathcal{T}_X(x, \theta) \ll \mathcal{T}_X(y, \sigma)$. Given any $\theta \in \Theta$, we define the mapping $T_\theta : X \longrightarrow X$ by $T_\theta x = \mathcal{T}_X(x, \theta)$, $x \in X$, and set $\mathcal{T}(x, \theta) \equiv (T_\theta x, \theta)$.

Clearly, $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ is continuous and increasing ($x \leq y$ in $\mathcal{X} \Longrightarrow \mathcal{T}x \leq \mathcal{T}y$), and it preserves also the strong ordering "$\ll$" in $\mathcal{X}$ ($x \ll y$ in $\mathcal{X} \Longrightarrow \mathcal{T}x \ll \mathcal{T}y$). We identify $\mathcal{T} \equiv \{T_\theta : \theta \in \Theta\}$ in a natural way.

For a fixed $\theta \in \Theta$, the *positive semiorbit under* T_θ (shortly, *θ-orbit*) of any $x \in X$ is defined by $\mathcal{O}_\theta^+(x) = \{T_\theta^n x : n \in \mathbb{Z}_+\}$, and the corresponding *ω-limit set* of x is defined by $\omega_\theta(x) = \{y \in X : T_\theta^{n_k} x \longrightarrow y \ (k \longrightarrow \infty)$ for some sequence $n_k \longrightarrow \infty$ in $\mathbb{Z}_+\}$. Obviously, $\omega_\theta(x) \neq \emptyset$ provided $\mathcal{O}_\theta^+(x)$ is relatively compact in X. A subset Y of X is called *positively invariant under* T_θ (shortly, *θ-invariant*) if $T_\theta(Y) \subset Y$, and *totally θ-invariant* if $T_\theta(Y) = Y$. For instance, every $\mathcal{O}_\theta^+(x)$ is θ-invariant, and if $\mathcal{O}_\theta^+(x)$ is relatively compact then $\omega_\theta(x)$ is totally θ-invariant. For the mapping $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ in place of T_θ we define analogous concepts and drop the subscript θ from the notation. Observe that $\mathcal{O}^+(x, \theta) = \mathcal{O}_\theta^+(x) \times \{\theta\}$ and $\omega(x, \theta) = \omega_\theta(x) \times \{\theta\}$. Clearly, a set $\mathcal{Y} \subset \mathcal{X}$ is invariant (totally invariant) if and only if, for each $\theta \in \Theta$, the set $Y_\theta = \{x \in X : (x, \theta) \in \mathcal{Y}\}$ is θ-invariant (totally θ-invariant).

Given a, $b \in X$, the set $[a, b] = \{x \in X : a \leq x \leq b\}$ is called a *closed order interval*, and $[[a, b]] = \{x \in X : a \ll x \ll b\}$ is called an *open order interval* in X. We write $[a, \infty]] = \{x \in X : x \geq a\}$, and proceed similarly for $[[-\infty, b]$, etc. A subset Y of X is called *order-convex* in X if $[a, b] \subset Y$ whenever a, $b \in Y$ and $a < b$; *lower closed* if $[[-\infty, b] \subset Y$ whenever $b \in Y$; and *upper closed* if $[a, \infty]] \subset Y$ whenever $a \in Y$. We denote closed order intervals in V by $[a, b]_V = \{x \in V : a \leq x \leq b\}$, and similarly, all other concepts in V will be marked by the subscript V in case confusion might arise.

Analogous concepts and notation as in X and V are introduced in Θ and V_Θ, and also in $\mathcal{X} = X \times \Theta$ and $\mathcal{V} = V \times V_\Theta$.

Now we are ready to introduce one of our basic concepts.

DEFINITION 1.1. A pair (A, B) of subsets A, B of $\mathcal{X}$ is called an *order decomposition* of $\mathcal{X}$ if it has the following five properties: (i) $A \neq \emptyset$ and $B \neq \emptyset$, (ii) A and B are closed, (iii) A is lower closed and B is upper closed, (iv) $A \cup B = \mathcal{X}$, and (v) $\mathrm{Int}(A \cap B) = \emptyset$.

An order decomposition (A, B) of $\mathcal{X}$ is called *invariant* if $\mathcal{T}(A) \subset A$ and $\mathcal{T}(B) \subset B$. The set $H = A \cap B$ (possibly empty) is called the *boundary* of the order decomposition (A, B) of $\mathcal{X}$. A *d-hypersurface* is any nonempty subset H of $\mathcal{X}$ such that $H = A \cap B$ for some order decomposition (A, B) of $\mathcal{X}$.

Notice that the boundary H of an order decomposition (A, B) of $\mathcal{X}$ satisfies $H = \partial A = \partial B$, where "$\partial$" is the boundary symbol in $\mathcal{X}$, and H is invariant whenever (A, B) is invariant. It is also easy to see that a d-hypersurface H never contains two strongly ordered points x, y (with $x \ll y$). Consequently, if H is invariant, then it contains no pair of points (x, θ), $(y, \sigma) \in \mathcal{X}$ satisfying $x \leq y$ and $\theta \ll \sigma$.

If X is a strongly ordered space, it turns out to be very useful to work with the *order topology* on X whose neighborhood base is generated by all open order intervals $[[a, b]]$ with $a \ll b$. If $Y \subset X$, we denote by $\hat{Y}$ the set Y endowed with the induced order topology. A subset Y of X is called *order open* (*order closed*, respectively) if it is open (closed, respectively) in $\hat{X}$. Notice that the identity mapping $\hat{\imath} : X \longrightarrow \hat{X}$ is continuous, but in general not homeomorphic. It is proved in Hirsch [13], [14] that if $f : X_1 \longrightarrow X_2$ is a continuous, increasing mapping between two strongly ordered spaces (i.e., $x \leq_{X_1} y$ implies $f(x) \leq_{X_2} f(y)$), then f is continuous also in the order topologies; that is, the induced mapping $\hat{f} : \hat{X}_1 \longrightarrow \hat{X}_2$ is continuous. It is easy to see that the order topology on V is induced by any *ordered norm* $|\cdot|_e$ on V defined by

$$|x|_e = \inf\{\lambda \in \mathbb{R}^1_+ : \ -\lambda e \leq x \leq \lambda e\}$$

for some $e \in \mathrm{Int}(V_+)$, where $\mathbb{R}^1_+ = [0, \infty)$.

Our first result deals with the existence of invariant d-hypersurfaces. We give a proof of existence which is much simpler than the original one in Takáč [28, Prop. 1.1] or [30, Prop. 1.2], even though our present hypotheses are slightly weaker.

PROPOSITION 1.2. *Let $\mathcal{X}$ be a strongly ordered space, and let $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ be continuous and preserving the strong ordering $\ll$ in $\mathcal{X}$ ($x \ll y$ in $\mathcal{X} \Longrightarrow \mathcal{T}x \ll \mathcal{T}y$). Assume that $G \subset \mathcal{X}$ is nonempty and invariant (under $\mathcal{T}$), and contains no pair of strongly ordered points x, y (with $x \ll y$). Then there exists an invariant order decomposition (A, B) of $\mathcal{X}$ such that $G \subset H = A \cap B$. In particular, we can define (A, B) in either of the following two ways.*

(i) *$A = \mathrm{Cl}(A^\circ)$ and $B = \mathcal{X} \setminus A^\circ$ where $A^\circ = \{x \in \mathcal{X} : \ \mathcal{T}^n x \ll y$ for some $n \in \mathbb{Z}_+$ and $y \in G\}$.*

(ii) *$A = \mathcal{X} \setminus B^\circ$ and $B = \mathrm{Cl}(B^\circ)$ where $B^\circ = \{x \in \mathcal{X} : \ \mathcal{T}^n x \gg y$ for some $n \in \mathbb{Z}_+$ and $y \in G\}$.*

For instance, if $\mathcal{X} = X \times \Theta$ and $\mathcal{T}$ satisfy (X), (Θ), and $(\mathcal{T})$, we may take $G = \omega(x, \theta)$ for any relatively compact $\mathcal{O}^+(x, \theta)$; cf. Proposition 1.5.

Proof. We prove only (i), the proof of (ii) being analogous. So let (A, B) be defined by (i). Clearly, A° is open since $\mathcal{T}$ is continuous, and A° is lower closed since $\mathcal{T}$ is increasing. It is easy to see that also $A = \mathrm{Cl}(A^\circ)$ is lower closed, whereas $B = \mathcal{X} \setminus A^\circ$ is upper closed, cf. Takáč [28, Lemma 2.1] or [30, Lemma 1.4]. Since G is invariant and contains no pair of strongly ordered points $x \ll y$, we have $A^\circ \cap G = \emptyset$. On the other hand, the fact that $\mathcal{X}$ is strongly ordered implies $G \subset \mathrm{Cl}\big(\cup_{y \in G}[[-\infty, y]]\big)$

and in particular $G \subset \mathrm{Cl}(A^\circ)$. Consequently, $G \subset \partial A^\circ \overset{\mathrm{def}}{=} \mathrm{Cl}(A^\circ) \setminus \mathrm{Int}(A^\circ) = A \cap B$ as desired. It is now obvious that the pair (A, B) satisfies properties (i)–(iv) in Definition 1.1. Suppose that (v) is false, i.e., there exists $b \in \mathrm{Int}(A \cap B) \neq \emptyset$. Then also $z \ll b \ll w$ for some $z, w \in \mathrm{Int}(A \cap B)$, since $\mathcal{X}$ is strongly ordered, and, consequently, $z' \ll b \ll w'$ for some $z' \in B$ and $w' \in A^\circ$. Hence $b \in [[z', w']] \subset A^\circ \cap B$ since A° is lower closed and B is upper closed. But then $A^\circ \cap B \neq \emptyset$ is a contradiction. We conclude that (A, B) is an order decomposition of $\mathcal{X}$.

Finally, making use of the fact that $\mathcal{T}$ preserves the strong ordering "$\ll$," we arrive at $\mathcal{T}(A^\circ) \subset A^\circ$ and $\mathcal{T}(\mathcal{X} \setminus A^\circ) \subset \mathcal{X} \setminus A^\circ$ which proves that (A, B) is invariant. □

In case $\mathcal{X} \subset \mathcal{V}$, our second result describes d-hypersurfaces as Lipschitz hypersurfaces in $\mathcal{V}$, cf. Takáč [28, Prop. 1.2] or [30, Prop. 1.3] for a proof. The first version of Part (i) was proved by Hirsch [15, Prop. 2.6] for the case $\dim(\mathcal{V}) < \infty$, the dimension of $\mathcal{V}$.

We recall that an everywhere defined linear mapping $L : V_1 \longrightarrow V_2$ between two ordered vector spaces is called *positive* (*strongly positive*, respectively) if $x < y$ in V_1 implies $Lx \leq Ly$ ($Lx \ll Ly$, respectively) in V_2. We set $I =$ identity mapping on $\mathcal{V}$, and $\mathbb{R}^1 = (-\infty, \infty)$.

PROPOSITION 1.3. *Let $\mathcal{X}$ be a nonempty open subset of $\mathcal{V}$, and let (A, B) be an order decomposition of $\mathcal{X}$ with the boundary $H = A \cap B$. Fix any vector $v \in \mathrm{Int}(\mathcal{V}_+)$, and denote by $R = \mathrm{lin}\{v\}$ the linear subspace of $\mathcal{V}$ spanned by v. Let Q be a positive continuous projection of $\mathcal{V}$ onto R, which always exists, and set $P = I - Q$ with $W = P(\mathcal{V})$, the range of P, so that $\mathcal{V} = W \oplus R$ is the direct algebraic and topological sum of W and R. Then we have the following statements.*

(i) *The restriction $P\big|_H$ of P to H is one-to-one, and both $P\big|_H$ and its inverse $\pi = (P\big|_H)^{-1} : P(H) \longrightarrow H$ are Lipschitz continuous in the ordered norm $|\cdot|_v$ with a common Lipschitz constant* 2.

(ii) *$P\big|_H$ is a homeomorphism of H onto $P(H)$ in the topologies induced by that on $\mathcal{V}$.*

(iii) *Furthermore, set*

$$H \oplus R = \{x \in \mathcal{V} : \ x = x_0 + \tau v \text{ for some } x_0 \in H \text{ and } \tau \in \mathbb{R}^1\},$$

where x_0 and τ are uniquely determined by $Px = Px_0$, and define a mapping $h : H \oplus R \longrightarrow \mathcal{V}$ by

$$h(x) = Px_0 + \tau v, \qquad x = x_0 + \tau v \in H \oplus R,$$

and similarly for $P(H) \oplus R$. Then also h and its inverse $h^{-1} : P(H) \oplus R \longrightarrow H \oplus R$ are Lipschitz continuous in the ordered norm $|\cdot|_v$ with a common Lipschitz constant 7, and h is a homeomorphism of $H \oplus R$ onto $P(H) \oplus R$ in the topologies induced by that on $\mathcal{V}$.

(iv) *If, in addition, $\mathcal{X}$ is order open in $\mathcal{V}$ (i.e., open in $\hat{\mathcal{V}}$), then $P(H)$ is order open in W, and $P(H) \oplus R$ is order open in $\mathcal{V}$.*

Let $\mathcal{X} = X \times \Theta$ and $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ satisfy (X), (Θ), and $(\mathcal{T})$. For a fixed $\theta \in \Theta$, we denote by $\mathcal{E}_\theta = \{x \in X : T_\theta x = x\}$ the set of all *equilibria* (i.e., *fixed points*) of T_θ. If $k \in \mathbb{N}$, the elements of $\mathcal{E}_\theta^k = \{x \in X : \ T_\theta^k x = x\}$ are called *k-periodic points* of T_θ, and their θ-orbits $\mathcal{O}_\theta^+(x)$ are called *k-cycles*. The following two elementary results were proved in Hirsch [13, Lemma 3.1] and Takáč [28, Lemma 2.2].

PROPOSITION 1.4. (Convergence criterion for strongly monotone semigroups.) *Assume that $(x, \theta) \in X \times \Theta$, $\mathcal{O}_\theta^+(x)$ is relatively compact, and either $T_\theta^k x > x$ or*

$T^k_\theta x < x$ *for some* $k \in \mathbb{N}$. *Then* $T^{nk+\ell}_\theta x \longrightarrow T^\ell_\theta p$ *as* $n \longrightarrow \infty$, $\ell = 0, 1, \cdots, k-1$, *for some* $p \in \mathcal{E}^k_\theta$, *and either* $p \gg x$ *or* $p \ll x$, *respectively. Moreover,* $\omega_\theta(x)$ *is a* k*-cycle.*

A set $Y \subset X$ is called *unordered* if no pair of points x, $y \in Y$ satisfies $x < y$.

PROPOSITION 1.5. (Nonordering of limit sets.) *Assume that* $(x, \theta) \in X \times \Theta$ *and* $\mathcal{O}^+_\theta(x)$ *is relatively compact. Then* $\omega_\theta(x)$ *is nonempty and unordered. If* $\mathrm{Cl}(\mathcal{O}^+_\theta(x))$ *is not unordered, then* $\omega_\theta(x)$ *is a cycle.*

From now on it is convenient to introduce the following ordering "$\preceq$" of unordered subsets of X: If F, $G \subset X$ are unordered, we write $F \preceq G$ if and only if

$$F \subset G_- = \cup\{[[-\infty, x] : x \in G\} \quad \text{and} \quad G \subset F_+ = \cup\{[x, \infty]] : x \in F\}.$$

We write $F \prec\!\!\prec G$ if and only if $F \subset \mathrm{Int}(G_-)$ and $G \subset \mathrm{Int}(F_+)$, while $F \prec G$ means $F \preceq G$, $F \neq G$.

We conclude this section with the following result.

PROPOSITION 1.6. *Let* x, $y \in X$ *and* σ, $\theta \in \Theta$ *be such that* $y \le x$ *and* $\sigma < \theta$. *Assume that both* $\mathcal{O}^+_\sigma(y)$ *and* $\mathcal{O}^+_\theta(x)$ *are relatively compact. Then* $\omega_\sigma(y) \prec\!\!\prec \omega_\theta(x)$. *If also* $w \in \omega_\theta(x)$ *is such that* $\mathcal{O}^+_\sigma(w)$ *is relatively compact, then* $\omega_\sigma(w)$ *is a cycle satisfying*

$$\omega_\sigma(y) \preceq \omega_\sigma(w) \prec\!\!\prec \omega_\theta(x).$$

Proof. By $(\mathcal{T})$ we have $T^n_\sigma y \ll T^n_\theta x$ for $n = 1, 2, \cdots$. Since $\mathcal{O}^+_\sigma(y)$ and $\mathcal{O}^+_\theta(x)$ are relatively compact, we obtain also $\omega_\sigma(y) \preceq \omega_\theta(x)$. Now take any $u \in \omega_\theta(x)$. Then $u = T_\theta u'$ for some $u' \in \omega_\theta(x)$, $v' \le u'$ for some $v' \in \omega_\sigma(y)$, and $v = T_\sigma v' \in \omega_\sigma(y)$ satisfies $v = T_\sigma v' \ll T_\theta u' = u$. Analogously, given any $v \in \omega_\sigma(y)$, we can find $u \in \omega_\theta(x)$ satisfying $v \ll u$. We have proved $\omega_\sigma(y) \prec\!\!\prec \omega_\theta(x)$.

Now fix any $w \in \omega_\theta(x)$ with $\mathcal{O}^+_\sigma(w)$ relatively compact. Then $\omega_\sigma(y) \prec\!\!\prec \omega_\theta(x)$ shows that $T^m_\sigma y \ll w$ for some $m \in \mathbb{Z}_+$ which, in turn, implies $\omega_\sigma(y) \preceq \omega_\sigma(w)$. Finally, from $T_\sigma w \ll T_\theta w \in \omega_\theta(x)$ we obtain $T_\sigma w \ll T^m_\theta x$ for some $m \in \mathbb{Z}_+$, and, consequently, $\omega_\sigma(w) = \omega_\sigma(T_\sigma w) \prec\!\!\prec \omega_\theta(T^m_\theta x) = \omega_\theta(x)$ as desired. In particular, $\omega_\sigma(w)$ is a cycle for T_σ, by Proposition 1.5., since $\mathcal{O}^+_\sigma(w)$ is not unordered. □

Remark. Observe that the cycle $C = \omega_\sigma(w)$ in Proposition 1.6 does *not* depend on a particular choice of $w \in \omega_\theta(x)$ with $\mathcal{O}^+_\sigma(w)$ relatively compact. Namely, if $\tilde{w} \in \omega_\theta(x)$ with $\mathcal{O}^+_\sigma(\tilde{w})$ relatively compact, then $\omega_\sigma(w) \prec\!\!\prec \omega_\theta(x)$ entails $\omega_\sigma(w) \preceq \omega_\sigma(\tilde{w})$, and, similarly, $\omega_\sigma(\tilde{w}) \preceq \omega_\sigma(w)$. We conclude that $\omega_\sigma(\tilde{w}) = \omega_\sigma(w)$ as desired.

2. Lower and upper ω-limit sets. In this section we introduce important stability concepts with respect to the varying parameter $\theta \in \Theta$. They are motivated by a number of results from Takáč [30, §3]. Throughout this entire section we assume that $\mathcal{X} = X \times \Theta$ and $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ satisfy (X), (Θ) and $(\mathcal{T})$. We say that the mapping $\mathcal{T}$ is *ω-compact* in a subset $\mathcal{Y}$ of $\mathcal{X}$ if $\mathcal{O}^+(x)$ is relatively compact for each $x \in \mathcal{Y}$, and also $\cup_{x \in \mathcal{Y}} \omega(x)$ is relatively compact in $\mathcal{X}$. Now let $x \in X$ and $\Sigma \subset \Theta$ be such that $\mathcal{T}$ is ω-compact in $\{x\} \times \Sigma$. We define the *lower* and *upper ω-limit sets* of $(x, \theta) \in \{x\} \times \Sigma$ *relative to* Σ by

$$\omega^\Sigma_{\theta-}(x) = \bigcap_{\substack{\sigma \in \Theta \\ \sigma \ll \theta}} \mathrm{Cl} \bigcup_{\substack{\varrho \in \Sigma \\ \sigma \le \varrho \ll \theta}} \omega_\sigma(x) \quad \text{and} \quad \omega^\Sigma_{\theta+}(x) = \bigcap_{\substack{\sigma \in \Theta \\ \sigma \gg \theta}} \mathrm{Cl} \bigcup_{\substack{\varrho \in \Sigma \\ \sigma \ge \varrho \gg \theta}} \omega_\sigma(x),$$

respectively. Observe that if there exists a sequence $\varrho_n \in \Sigma$, $\varrho_n \ll \theta$ ($\varrho_n \gg \theta$, respectively) and $\varrho_n \longrightarrow \theta$, then $\omega^\Sigma_{\theta-}(x) \neq \emptyset$ ($\omega^\Sigma_{\theta+}(x) \neq \emptyset$), by the ω-compactness of $\mathcal{T}$.

The lower and upper ω-limit sets satisfy the following analogue of Proposition 1.5.

PROPOSITION 2.1. (Nonordering of lower and upper limit sets.) *Assume that $x \in X$ and $\Sigma \subset \Theta$ are such that $\mathcal{T}$ is ω-compact in $\{x\} \times \Sigma$. Let $\theta \in \Sigma$. Then both $\omega^{\Sigma}_{\theta-}(x)$ and $\omega^{\Sigma}_{\theta+}(x)$ are compact, unordered, and totally θ-invariant.*

Proof. We consider only $\omega^{\Sigma}_{\theta-}(x)$. It is compact because $\mathcal{T}$ is ω-compact. Suppose it is not unordered, i.e., $a < b$ for some $a,\, b \in \omega^{\Sigma}_{\theta-}(x)$. Hence $T_\theta a \ll T_\theta b$, and there exist $u,\, v \in X$, and $\lambda \in \Theta$ such that $a \ll u$, $v \ll b$, $\lambda \ll \theta$, and $T_\sigma u \ll T_\sigma v$ whenever $\lambda \le \sigma \le \theta$. Next we find $\alpha,\, \beta \in \Sigma$ with $\lambda \le \beta \ll \alpha \ll \theta$ and $a' \in \omega_\alpha(x)$ and $b' \in \omega_\beta(x)$ so close to a and b, respectively, that $a' \ll u$ and $v \ll b'$. Hence, $T_\sigma a' \ll T_\sigma b'$ for all $\sigma \in \Theta$, $\lambda \le \sigma \le \theta$. By Proposition 1.6, we have $\omega_\beta(x) \ll \omega_\alpha(x)$. So $b' \ll a''$ for some $a'' \in \omega_\alpha(x)$. We obtain $T_\sigma a' \ll T_\sigma b' \ll T_\sigma a''$ for $\lambda \le \sigma \le \theta$. Taking $\sigma = \alpha$ we arrive at $T_\alpha a' \ll T_\alpha a''$ in $\omega_\alpha(x)$, a contradiction to $\omega_\alpha(x)$ is unordered. We have verified that $\omega^{\Sigma}_{\theta-}(x)$ is unordered.

To show that $\omega^{\Sigma}_{\theta-}(x)$ is totally θ-invariant, we may assume it is nonempty. If $\sigma \in \Sigma$, $\sigma \ll \theta$, then $\omega_\sigma(x) \preceq \omega^{\Sigma}_{\theta-}(x)$ and $\omega_\sigma(x) = T_\sigma \omega_\sigma(x) \preceq T_\theta \omega^{\Sigma}_{\theta-}(x)$, which entails $\omega^{\Sigma}_{\theta-}(x) \preceq T_\theta \omega^{\Sigma}_{\theta-}(x)$. On the other hand, if we fix $\lambda \in \Theta$, $\lambda \ll \theta$, and choose any $\sigma \in \Sigma$, $\lambda \le \sigma \ll \theta$, then $T_\lambda \omega_\sigma(x) \preceq T_\sigma \omega_\sigma(x) = \omega_\sigma(x) \preceq \omega^{\Sigma}_{\theta-}(x)$, and so $T_\lambda \omega^{\Sigma}_{\theta-}(x) \preceq \omega^{\Sigma}_{\theta-}(x)$. Using the compactness of $\omega^{\Sigma}_{\theta-}(x)$, we let $\lambda \longrightarrow \theta$ in Θ, thus obtaining $T_\theta \omega^{\Sigma}_{\theta-}(x) \preceq \omega^{\Sigma}_{\theta-}(x)$. Finally, $\omega^{\Sigma}_{\theta-}(x)$ is unordered forces $T_\theta \omega^{\Sigma}_{\theta-}(x) = \omega^{\Sigma}_{\theta-}(x)$ as desired. □

COROLLARY 2.2. *Let all hypotheses of* Proposition 2.1 *be satisfied. If $\varrho_n \longrightarrow \theta$ for some sequence $\varrho_n \in \Sigma$, $\varrho_n \ll \theta$, then*

$$\omega^{\Sigma}_{\theta-}(x) = \bigcap_{k=1}^{\infty} \operatorname{Cl} \bigcup_{n=k}^{\infty} \omega_{\varrho_n}(x)$$

and $\omega_\varrho(x) \prec\prec \omega^{\Sigma}_{\theta-}(x) \preceq \omega_\theta(x)$ for each $\varrho \in \Theta$, $\varrho \ll \theta$, and $\mathcal{O}^{+}_{\varrho}(x)$ relatively compact. A corresponding result holds for $\omega^{\Sigma}_{\theta+}(x)$.

Proof. Since $\{\varrho_n\}_{n=1}^{\infty}$ contains a strongly increasing subsequence (ordered by $\ll$), we may assume $\varrho_1 \ll \varrho_2 \ll \cdots \ll \theta$. Set $\Omega = \{\theta, \varrho_1, \varrho_2, \cdots\}$. If $\sigma \in \Sigma$, $\sigma \ll \theta$, and $m \in \mathbb{N}$ is so large that $\sigma \ll \varrho_m$, then $\omega_\sigma(x) \prec\prec \omega_{\varrho_m}(x) \preceq \omega^{\Omega}_{\theta-}(x)$. Hence, $\omega^{\Sigma}_{\theta-}(x) \preceq \omega^{\Omega}_{\theta-}(x)$. On the other hand, if $m \in \mathbb{N}$ is arbitrary and $\sigma_m \in \Sigma$, $\varrho_m \ll \sigma_m \ll \theta$, then $\omega_{\varrho_m}(x) \prec\prec \omega_{\sigma_m}(x) \preceq \omega^{\Sigma}_{\theta-}(x)$. Hence, $\omega^{\Omega}_{\theta-}(x) \preceq \omega^{\Sigma}_{\theta-}(x)$. But both $\omega^{\Sigma}_{\theta-}(x)$ and $\omega^{\Omega}_{\theta-}(x)$ are unordered, and, therefore,

$$\omega^{\Sigma}_{\theta-}(x) = \omega^{\Omega}_{\theta-}(x) = \bigcap_{k=1}^{\infty} \operatorname{Cl} \bigcup_{n=k}^{\infty} \omega_{\varrho_n}(x).$$

The remaining statements are obvious. □

Remark. It is clear from Corollary 2.2 that the set $\omega^{\Sigma}_{\theta-}(x)$ ($\omega^{\Sigma}_{\theta+}(x)$, respectively) is independent from the choice of $\Sigma \subset \Theta$ such that $\theta \in \Sigma$, $\varrho_n \longrightarrow \theta$ for some sequence $\varrho_n \in \Sigma$, $\varrho_n \ll \theta$ ($\varrho_n \gg \theta$), and $\mathcal{T}$ is ω-compact in $\{x\} \times \Sigma$. Therefore, we say that a point $(x, \theta) \in X \times \Theta$ is *lower* (*upper*, respectively) *approximable* if there exists a sequence $\varrho_n \in \Theta$, $\varrho_n \ll \theta$ ($\varrho_n \gg \theta$), satisfying $\varrho_n \longrightarrow \theta$ and $\mathcal{T}$ is ω-compact in $\{x\} \times \Omega$, where $\Omega = \{\theta, \varrho_1, \varrho_2, \cdots\}$. We define the *lower* (*upper*) *ω-limit set* of such (x, θ) by $\omega_{\theta-}(x) = \omega^{\Omega}_{\theta-}(x)$ ($\omega_{\theta+}(x) = \omega^{\Omega}_{\theta+}(x)$).

From this Remark we introduce the following stability classification of a lower (upper, respectively) approximable point $(x, \theta) \in X \times \Theta$.

We say that $(x, \theta) \in X \times \Theta$ is *lower* (*upper*, respectively) *ω-stable* if $\omega_{\theta-}(x) = \omega_\theta(x)$ ($\omega_{\theta+}(x) = \omega_\theta(x)$); otherwise, (x, θ) is *lower* (*upper*) *ω-unstable*. The set of all lower (upper) ω-stable points (x,θ) is denoted by $\mathcal{S}_-$ ($\mathcal{S}_+$), and the set of all lower (upper) ω-unstable points by $\mathcal{U}_-$ ($\mathcal{U}_+$). We denote $\mathcal{S}_{\theta-} = \{x \in X : (x,\theta) \in \mathcal{S}_-\}$ and $\mathcal{U}_{\theta-} = \{x \in X : (x,\theta) \in \mathcal{U}_-\}$ and, analogously, $\mathcal{S}_{\theta+}$ and $\mathcal{U}_{\theta+}$.

Observe that our stability notions are equivalent to the continuity properties of the set-valued mapping $\theta\ (\in \Theta) \longmapsto \omega_\theta(x)\ (\subset X)$, for $x \in X$ fixed.

Remark. Assume that $\hat{d}(x,y)$ is an *ordered metric* for $\hat{X}$, the space X with the order topology, i.e., $u \le a \le b \le v$ in X implies $\hat{d}(a,b) \le \hat{d}(u,v)$. Let $\hat{\delta}(\theta,\sigma)$ be an ordered metric for $\hat{\Theta}$, and define an ordered product metric

$$\hat{\mathcal{D}}\big((x,\theta),(y,\sigma)\big) = \max\{\hat{d}(x,y), \hat{\delta}(\theta,\sigma)\}$$

for $\hat{\mathcal{X}} = \hat{X} \times \hat{\Theta}$. For instance, if V is a strongly ordered vector space and $e \in \operatorname{Int}(V_+)$, then the ordered norm $|\cdot|_e$ on $\hat{V}$ defines an ordered metric $\hat{d}(x,y) = |x-y|_e$ for $\hat{V}$. The following lemma shows that our stability concept uses the metric $\hat{\mathcal{D}}$.

LEMMA 2.3. *Let $(x,\theta) \in X \times \Theta$ be lower approximable. Then $(x,\theta) \in \mathcal{S}_-$ if and only if the following statement holds.*

($*$) *For every $\epsilon > 0$ there exists $\delta > 0$ such that, for each $(y,\sigma) \in X \times \Theta$ with $(y,\sigma) \le (x,\theta)$ and $\mathcal{O}_\sigma^+(y)$ relatively compact, we have*

$$\hat{\mathcal{D}}\big((x,\theta),(y,\sigma)\big) \le \delta \Longrightarrow \hat{d}(T_\theta^n x, T_\sigma^n y) \le \epsilon \quad \textit{for all } n \in \mathbb{Z}_+.$$

Corresponding statements hold for (x,θ) upper approximable and $\mathcal{S}_+$.

Proof. We first deduce from $(\mathcal{T})$ that for every $\varrho \in \Theta$, $\varrho < \theta$, there exists $\delta > 0$ such that, for each $(y,\sigma) \in X \times \Theta$ with $(y,\sigma) < (x,\theta)$, we have

$$\hat{\mathcal{D}}\big((x,\theta),(y,\sigma)\big) \le \delta \Longrightarrow T_\varrho x \le T_\sigma y \ll T_\theta x.$$

We conclude that ($*$) is equivalent to the following.

$$(**) \qquad\qquad \text{Take } (y,\sigma) = (x,\varrho) \text{ in } (*).$$

Now assume $(x,\theta) \in \mathcal{S}_-$. Suppose ($**$) is not valid. Then, since $\hat{\mathcal{D}}$ and $\hat{d}$ are ordered metrics, there exist $\epsilon_0 > 0$ and sequences $\varrho_1 \ll \varrho_2 \ll \cdots \ll \theta$ in Θ and $n_1, n_2, \cdots$ in $\mathbb{Z}_+$ such that $\varrho_k \longrightarrow \theta$ in $\hat{\Theta}$ as $k \longrightarrow \infty$, $\mathcal{T}$ is ω-compact in $\{x\} \times \Omega$ where $\Omega = \{\theta, \varrho_1, \varrho_2, \cdots\}$, and

$$\hat{d}(T_\theta^{n_k} x, T_{\varrho_k}^{n_k} x) \ge \epsilon_0 \quad \text{for all } k \in \mathbb{N}.$$

The continuity of $\mathcal{T}$ forces $\{n_k\}$ to be unbounded. Passing to a subsequence, we may assume $n_1 < n_2 < \cdots$. Observe that $T_{\varrho_m}^{n_k} x \le T_{\varrho_k}^{n_k} x \le T_\theta^{n_k} x$ whenever $k \ge m \in \mathbb{N}$. Hence $\hat{d}(T_\theta^{n_k} x, T_{\varrho_m}^{n_k} x) \ge \epsilon_0$ for $k \ge m$, and letting $k \longrightarrow \infty$ we get $w \in \omega_\theta(x)$ and $u_m \in \omega_{\varrho_m}(x)$, satisfying $u_m \le w$ and $\hat{d}(w, u_m) \ge \epsilon_0$. Using the ω-compactness of $\mathcal{T}$ in $\{x\} \times \Omega$, we find a subsequence of $\{u_m\}$ convergent in X to some $u \in \omega_{\theta-}(x) = \omega_\theta(x)$. Consequently $u \le w$ and $\hat{d}(w,u) \ge \epsilon_0$, which means $u < w$ in an unordered set $\omega_\theta(x)$, a contradiction. We have proved that $(x,\theta) \in \mathcal{S}_- \Longrightarrow (**)$.

To prove the converse statement, we assume ($**$) is valid. We apply Corollary 2.2: Let $\varrho \ll \varrho_2 \ll \cdots \ll \theta$ be any sequence in Θ such that $\varrho_k \longrightarrow \theta$ and $\mathcal{T}$ is ω-compact in $\{x\} \times \Omega$ where $\Omega = \{\theta, \varrho_1, \varrho_2, \cdots\}$. Suppose $\omega_{\theta-}(x) \ne \omega_\theta(x)$. Then $u < v$ for some $u \in \omega_{\theta-}(x)$ and $v \in \omega_\theta(x)$. Making use of $\omega_{\varrho_1}(x) \mathrel{-\!\!\!\ll} \omega_{\varrho_2}(x) \mathrel{-\!\!\!\ll} \cdots \mathrel{-\!\!\!\ll} \omega_{\theta-}(x)$, we find a sequence $u_k \in \omega_{\varrho_k}(x)$ such that $u_k \ll u$ for all $k \in \mathbb{N}$. Consequently, there exists another sequence $n_1 < n_2 < \cdots$ in $\mathbb{Z}_+$ such that $T_{\varrho_k}^{n_k} x \le u$ and $T_\theta^{n_k} x \longrightarrow w \in \omega_\theta(x)$.

From $(**)$ we obtain $\hat{d}(T_\theta^{n_k}x, T_{\varrho_k}^{n_k}x) \longrightarrow 0$ as $k \longrightarrow \infty$, and, therefore, $w \leq u$. But then $w < v$ in $\omega_\theta(x)$ contradicts the fact that $\omega_\theta(x)$ is unordered. We have also verified $(**) \Longrightarrow (x,\theta) \in \mathcal{S}_-$. □

The structure of the ω-limit sets $\omega_\sigma(x)$ for $\sigma \ll \theta$ near θ, where $(x,\theta) \in X \times \Theta$ is lower ω-unstable, is very simple; and, similarly, for upper ω-unstable points we have the following.

THEOREM 2.4. *Let $(x,\theta) \in \mathcal{U}_-$. Assume there exists $\lambda_0 \in \Theta$, $\lambda_0 \ll \theta$, such that $\mathcal{O}_\sigma^+(y)$ is relatively compact for every $(y,\sigma) \in X \times \Theta$, satisfying $\lambda_0 \leq \sigma \leq \theta$ and $u \leq y \leq w$ for some $u \in \omega_{\lambda_0}(x)$ and $w \in \omega_\theta(x)$.*

Then there exists $\lambda \in \Theta$, $\lambda_0 \leq \lambda \ll \theta$, such that for every $\sigma \in \Theta$, $\lambda \leq \sigma \ll \theta$, the set $\omega_\sigma(x)$ is a k_σ-cycle for T_σ, satisfying $\omega_\sigma(w) = \omega_\sigma(x)$ for each $w \in \omega_\theta(x)$. In particular, if $u \in \omega_\sigma(x)$ then also $\omega_\theta(u)$ is a k_σ-cycle for T_θ, satisfying $\omega_\sigma(x) \prec\!\!\prec \omega_\theta(u) \preceq \omega_{\theta-}(x) \prec\!\!\prec \omega_\theta(x)$.

A corresponding result holds for $(x,\theta) \in \mathcal{U}_+$.

Proof. Since $\omega_{\theta-}(x) \prec \omega_\theta(x)$ and both $\omega_{\theta-}(x)$ and $\omega_\theta(x)$ are totally θ-invariant, there exist $\tilde{v} \in \omega_{\theta-}(x)$ and $\tilde{w} \in \omega_\theta(x)$ with $\tilde{v} \ll \tilde{w}$. Then $\tilde{v} \ll T_\theta^m x$ for some $m \in \mathbb{Z}_+$. Fix any $\lambda \in \Theta$, $\lambda_0 \leq \lambda \ll \theta$, so close to θ that $\tilde{v} \ll T_\lambda^m x$ holds. Now let $\sigma \in \Theta$, $\lambda \leq \sigma \ll \theta$. Then $v_\sigma \ll \tilde{v}$ for some $v_\sigma \in \omega_\sigma(x) \prec\!\!\prec \omega_{\theta-}(x)$, and, consequently, $T_\sigma^{m_\sigma}x \ll \tilde{v}$ for some $m_\sigma > m$ in $\mathbb{Z}_+$. Applying Proposition 1.4 to $T_\sigma^{m_\sigma}x \ll \tilde{v} \ll T_\lambda^m x \leq T_\sigma^m x$, we conclude that $\omega_\sigma(x)$ is a cycle. If $w \in \omega_\theta(x)$ then $T_\sigma w \ll T_\theta w \in \omega_\theta(x)$, and so $T_\sigma w \ll T_\theta^k x$ for some $k \in \mathbb{Z}_+$. We find $\varrho \in \Theta$, $\sigma \leq \varrho \ll \theta$, such that $T_\sigma w \ll T_\varrho^k x$, and, consequently, $\omega_\sigma(w) \preceq \omega_\varrho(x) \prec\!\!\prec \omega_{\theta-}(x)$. Hence, $T_\sigma^{n_\sigma}w \ll \tilde{v} \ll T_\sigma^m x$ for some $n_\sigma > m$ in $\mathbb{Z}_+$. It follows that $\omega_\sigma(w) \preceq \omega_\sigma(x)$. On the other hand, $\omega_\sigma(x) \prec\!\!\prec \omega_\theta(x)$ and $w \in \omega_\theta(x)$ force $\omega_\sigma(x) \preceq \omega_\sigma(w)$. Thus $\omega_\sigma(w) = \omega_\sigma(x)$ as desired. If $u \in \omega_\sigma(x)$ then $\omega_\sigma(x) \prec\!\!\prec \omega_\theta(u) \preceq \omega_{\theta-}(x)$ by Proposition 1.6, and if k_σ denotes the cardinality of the cycle $\omega_\sigma(x)$, then $u = T_\sigma^{k_\sigma}u \ll T_\theta^{k_\sigma}u$ entails that $\omega_\theta(u)$ is a k_σ-cycle by Proposition 1.4.

In order to verify $\omega_{\theta-}(x) \prec\!\!\prec \omega_\theta(x)$, we first recall that $\omega_{\theta-}(x) \prec \omega_\theta(x)$ and $\tilde{v} \ll \tilde{w}$ for some $\tilde{v} \in \omega_{\theta-}(x)$ and $\tilde{w} \in \omega_\theta(x)$. Since both $\omega_{\theta-}(x)$ and $\omega_\theta(x)$ are totally θ-invariant, it suffices to show $\omega_{\theta-}(x) \cap \omega_\theta(x) = \emptyset$. Then T_θ strongly increasing entails $\omega_{\theta-}(x) = T_\theta\omega_{\theta-}(x) \prec\!\!\prec T_\theta\omega_\theta(x) = \omega_\theta(x)$ as desired. Now suppose there exists $w^* \in \omega_{\theta-}(x) \cap \omega_\theta(x)$; hence, $\omega_\theta(w^*) \subset \omega_{\theta-}(x) \cap \omega_\theta(x)$. We have shown above that $\omega_\sigma(x) = \omega_\sigma(w^*) \prec\!\!\prec \omega_\theta(w^*)$ for all $\sigma \in \Theta$, $\lambda \leq \sigma \ll \theta$. From Corollary 2.2 we deduce $\omega_{\theta-}(x) \preceq \omega_\theta(w^*)$. But then $\omega_\theta(w^*) \subset \omega_{\theta-}(x)$ and $\omega_{\theta-}(x)$ unordered force $\omega_\theta(w^*) = \omega_{\theta-}(x)$. Similarly, we use $\omega_{\theta-}(x) \preceq \omega_\theta(x)$, $\omega_{\theta-}(x) \subset \omega_\theta(x)$, and $\omega_\theta(x)$ unordered to obtain $\omega_{\theta-}(x) = \omega_\theta(x)$, a contradiction to $(x,\theta) \in \mathcal{U}$. □

The last result in this section describes the structure of the ω-limit sets $\omega_\sigma(x)$ for $\sigma \ll \theta$ near θ, where $(x,\theta) \in X \times \Theta$ is lower ω-stable, and for upper ω-stable points as well.

THEOREM 2.5. *Let $(x,\theta) \in \mathcal{S}_-$. Assume there exists $\lambda_0 \in \Theta$, $\lambda_0 \ll \theta$, such that $\mathcal{O}_\sigma^+(y)$ is relatively compact for every $(y,\sigma) \in X \times \Theta$, satisfying $\lambda_0 \leq \sigma \leq \theta$ and $u \leq y \leq w$ for some $u \in \omega_{\lambda_0}(x)$ and $w \in \omega_\theta(x)$.*

Then for all $\sigma \in \Theta$, $\lambda_0 \leq \sigma \ll \theta$, and $w \in \omega_\theta(x)$, the set $\omega_\sigma(w)$ is a k_σ-cycle for T_σ independent from w and satisfies $\omega_\sigma(x) \preceq \omega_\sigma(w) \prec\!\!\prec \omega_\varrho(x)$ for some $\varrho \in \Theta$, $\sigma < \varrho \ll \theta$. In particular, if $u \in \omega_\sigma(w)$ then also $\omega_\theta(u)$ is a k_σ-cycle for T_θ satisfying $\omega_\sigma(w) \prec\!\!\prec \omega_\theta(u) \preceq \omega_\theta(x)$.

Finally, given arbitrary $v \in \omega_\theta(x)$ and $a \in X$, $a \ll v$, there exists $z \in X$, $a \ll z \ll v$, such that $z \ll T_\theta^k z$ for some $k \in \mathbb{N}$. In particular, for $\hat{U}_i = \omega_\theta(x) \cap [[T_\theta^i z, \infty]]$, $i \in \mathbb{Z}_+$, we have $T_\theta^k(\hat{U}_i) \subset \hat{U}_i$ and $\cup_{i=j}^{j+k-1}\hat{U}_i = \omega_\theta(x)$ for each $j \in \mathbb{Z}_+$. Also $\mathcal{O}_\theta^+(v)$ is dense in $\omega_\theta(x)$.

A corresponding result holds for $(x, \theta) \in \mathcal{S}_+$.

Proof. Let $\sigma \in \Theta$, $\lambda_0 \leq \sigma \ll \theta$, and $w \in \omega_\theta(x)$. Then $\omega_\sigma(w)$ is a k_σ-cycle for T_σ satisfying $\omega_\sigma(x) \preceq \omega_\sigma(w) \mathrel{-\!\!\!\prec\!\!\prec} \omega_\theta(x)$, by Proposition 1.6, which is independent from w, by the Remark after Proposition 1.6. Now observe $T_\sigma^m w \ll w$ for some $m \in \mathbb{N}$, and hence $T_\sigma^m w \ll T_\theta^\ell x$ for some $\ell \in \mathbb{N}$. Using the continuity of $\mathcal{T}$, we obtain $T_\sigma^m w \ll T_\varrho^\ell x$ for some $\varrho \in \Theta$, $\sigma < \varrho \ll \theta$, whence $\omega_\sigma(w) \mathrel{-\!\!\!\prec\!\!\prec} \omega_\varrho(x)$ because $\sigma < \varrho$. If $u \in \omega_\sigma(w)$ then $\omega_\sigma(w) \mathrel{-\!\!\!\prec\!\!\prec} \omega_\theta(u) \preceq \omega_\theta(x)$ by Proposition 1.6, and $u = T_\sigma^{k_\sigma} u \ll T_\theta^{k_\sigma} u$ shows that $\omega_\theta(u)$ is a k_σ-cycle by Proposition 1.4.

Fix arbitrary $v \in \omega_\theta(x)$ and $a \in X$, $a \ll v$. Then $\omega_\sigma(x) \preceq \omega_\sigma(v) \mathrel{-\!\!\!\prec\!\!\prec} \omega_\theta(x)$ whenever $\lambda_0 \leq \sigma \ll \theta$. Recalling $(x, \theta) \in \mathcal{S}_-$ and Corollary 2.2, we can find $\sigma \in \Theta$ so close to θ that $a \ll z \ll v$ for some $z \in \omega_\sigma(v)$. Hence, $z = T_\sigma^{k_\sigma} z \ll T_\theta^{k_\sigma} z$. Set $k = k_\sigma$ and $\hat{U}_i = \omega_\theta(x) \cap [[T_\theta^i z, \infty]]$, $i \in \mathbb{Z}_+$. Clearly, $T_\theta^i z \ll T_\theta^{k+i} z$ entails $T_\theta^k(\hat{U}_i) \subset \hat{U}_i$. From $\omega_\theta(z) \preceq \omega_\theta(x)$, we deduce $\cup_{i=j}^{j+k-1} \hat{U}_i = \omega_\theta(x)$ for each $j \in \mathbb{Z}_+$.

To show that $\mathcal{O}_\theta^+(v)$ is dense in $\omega_\theta(x)$ we first observe that the topologies from X and $\hat{X}$ coincide on $\omega_\theta(x)$, by compactness. Consequently, choose any $v' \in \omega_\theta(x)$ and $a' \in X$, $a' \ll v'$. Now take $\sigma \in \Theta$ above so close to θ that also $a' \ll z' \ll v'$ for some $z' \in \omega_\sigma(v) = \omega_\sigma(v')$. Hence, $z' = T_\sigma^\ell z$ for some $\ell \in \mathbb{Z}_+$, $0 \leq \ell \leq k_\sigma - 1$. We obtain $a' \ll z' = T_\sigma^\ell z \ll T_\theta^\ell v$. Letting $a' \longrightarrow v'$ in X we arrive at $v' \leq v^*$ for some $v^* \in \mathrm{Cl}(\mathcal{O}_\theta^+(v)) \subset \omega_\theta(x)$. Finally, $\omega_\theta(x)$ unordered forces $v' = v^*$. The proof is complete. □

3. Main results. We have seen in Theorems 2.4 and 2.5 that the set $\omega_{\theta-}(x)$ can be approximated from below by a sequence of cycles for T_θ, namely,

$$\omega_\theta(u_1) \preceq \omega_\theta(u_2) \preceq \cdots \preceq \omega_{\theta-}(x),$$

where $u_k \in \omega_{\sigma_k}(w)$ for any fixed $w \in \omega_\theta(x)$ and $\lambda_0 \leq \sigma_1 \ll \sigma_2 \ll \cdots \ll \theta$ in Θ with $\sigma_k \longrightarrow \theta$ as $k \longrightarrow \infty$. In particular, $\omega_{\theta-}(x)$ is approximated also by the sequence of cycles

$$\omega_{\sigma_1}(w) \mathrel{-\!\!\!\prec\!\!\prec} \omega_{\sigma_2}(w) \mathrel{-\!\!\!\prec\!\!\prec} \cdots \mathrel{-\!\!\!\prec\!\!\prec} \omega_{\sigma-}(x).$$

Similar approximation holds for $\omega_{\theta+}(x)$.

In our next theorem we will show that $\omega_{\theta-}(x)$ must be a quasi cycle for T_θ:

DEFINITION 3.1. *Given $\theta \in \Theta$, we say that a set $C \subset X$ is a quasi cycle for T_θ if C is nonempty, compact, and totally θ-invariant, and every open cover of C (by its relatively open subsets) possesses a refinement forming another finite open cover of C by the sets U_0, U_1, $\cdots$, U_{k-1} such that $T_\theta^j(U_i) \subset U_{i+j}$ for all i, $j \in \mathbb{Z}_+$, where $U_{\ell+k} \equiv U_\ell$, $\ell \in \mathbb{Z}_+$.*

We recall that if $\mathcal{C}$ and $\mathcal{C}'$ are two open covers of a topological space K, then $\mathcal{C}'$ is a *refinement* of $\mathcal{C}$ if for every $V \in \mathcal{C}'$ there exists $U \in \mathcal{C}$ such that $V \subset U$.

We start with the following technical lemma.

LEMMA 3.2. *Let G be a compact and unordered subset of X. Then the sets $U_a = G \cap [[a, \infty]]$, for all $a \in X$, form a base for a Hausdorff topology on G which is coarser than the topology from X restricted to G (and hence, both these topologies are identical).*

The same result is valid for the sets $U^b = G \cap [[-\infty, b]]$, $b \in X$.

Proof. Clearly every U_a, $a \in X$, is relatively open in $G \subset X$. If $x \in G$ and $x \in U_a \cap U_b$ for some a, $b \in X$, then there exists $c \in [[a, \infty]] \cap [[b, \infty]]$ with $c \ll x$;

hence, $x \in U_c \subset U_a \cap U_b$. It follows that $\{U_a : a \in X\}$ is a base for a topology on G which is coarser than the X-topology. Suppose the former one is not Hausdorff, i.e., there exist x, $y \in G$, $x \neq y$, such that for all a, $b \in X$ we have: $x \in U_a$, $y \in U_b$ $\Longrightarrow U_a \cap U_b \neq \emptyset$. So let $z_{ab} \in U_a \cap U_b$. Letting $(a,b) \longrightarrow (x,y)$ in $X \times X$, $a \ll x$, $b \ll y$, we obtain $x \leq z$ and $y \leq z$ for any limit point z of z_{ab} in $G \subset X$. But G unordered forces $x = z = y$, a contradiction. Thus, $\{U_a : a \in X\}$ is a base for a Hausdorff topology on G. □

Now we are ready to prove the following theorem.

THEOREM 3.3. *Let $(x,\theta) \in X \times \Theta$ be lower approximable. Assume there exists $\lambda_0 \in \Theta$, $\lambda_0 \ll \theta$, such that $\mathcal{O}_\sigma^+(y)$ is relatively compact for every $(y,\sigma) \in X \times \Theta$, satisfying $\lambda_0 \leq \sigma \leq \theta$ and $u \leq y \leq w$ for some $u \in \omega_{\lambda_0}(x)$ and $w \in \omega_\theta(x)$.*

Then $\omega_{\theta-}(x) \subset \mathcal{S}_{\theta-}$, and $\omega_{\theta-}(x)$ is a quasi cycle for T_θ. In particular $\omega_{\theta-}(x) = \mathrm{Cl}(\mathcal{O}_\theta^+(w))$ for every $w \in \omega_{\theta-}(x)$.

A corresponding result holds if $(x,\theta) \in X \times \Theta$ is upper approximable.

Proof. Set $C = \omega_{\theta-}(x)$. It follows from Theorems 2.4 and 2.5 that $(w,\theta) \in \mathcal{S}_-$ for every $w \in C$, i.e., $C \subset \mathcal{S}_{\theta-}$. To show that C is a quasi cycle for T_θ, we consider an arbitrary open cover $\mathcal{C}$ of C. Combining Lemma 3.2 with the compactness of C, we can find a finite set $M \subset X$ such that the sets $U_a = C \cap [[a,\infty]] \neq \emptyset$, $a \in M$, form a refinement of the cover $\mathcal{C}$ of C. Furthermore, observe that $\hat{N} = \cup_{a \in M}[[a,\infty]]$ is an order-open set in X with $C \subset \hat{N}$. Fix any $w \in C$. Thus, combining $\omega_\sigma(x) \preceq \omega_\sigma(w) \prec \prec C$ with Corollary 2.2 (and its proof) we can choose $\sigma \in \Theta$, $\lambda_0 \leq \sigma \ll \theta$, so close to θ that $\omega_\sigma(w) \subset \hat{N}$. Consequently, M can be replaced by $\omega_\sigma(w)$ which is a k-cycle for T_σ, and so $z \ll T_\theta^k z$ for $z \in \omega_\sigma(w)$. Fix any $z \in \omega_\sigma(w)$ and define $\hat{U}_i = C \cap [[T_\theta^i z, \infty]]$, $0 \leq i \leq k-1$, and $\hat{U}_{\ell+k} \equiv \hat{U}_\ell$, $\ell \in \mathbb{Z}_+$. Then $\hat{U}_i \subset U_a$ for some $a = a(i) \in M$ depending upon $i \in \mathbb{Z}$, $0 \leq i \leq k-1$. We conclude that $\{\hat{U}_i : 0 \leq i \leq k-1\}$ is a refinement of $\mathcal{C}$ having all properties required in Definition 3.1. So C is a quasi cycle. Finally, letting $\sigma \longrightarrow \theta$, we obtain $C = \mathrm{Cl}(\mathcal{O}_\theta^+(w))$ for $w \in C$. □

This theorem implies that the dynamics of T_θ on $\omega_{\theta-}(x)$ is not "very complicated"; it is very much cycle-like and stable in the sense of Definition 3.1. On the other hand, an example due to Smale [24] suggests that for $(x,\theta) \in \mathcal{U}_-$ the dynamics of T_θ on $\omega_\theta(x)$ can be rather "arbitrary." However, if $\mathcal{X} = X \times \Theta \subset V \times V_\Theta = \mathcal{V}$ with $\mathcal{V}$ separable, this case is very improbable because the set $\mathcal{U}_-$ has zero Gaussian measure as we state it more precisely in Theorem 3.4. This theorem is closely related to results in Hirsch [14, §7] and Takáč [30, Prop. 5.5 and Cor. 5.6]. A similar statement holds for $\mathcal{U}_+$. But first we need some additional notation.

Given $(x,\theta) \in X \times \Theta$, we set

$$\mathcal{X}_{(x,\theta)-} = \{(y,\sigma) \in X \times \Theta : \ \sigma \ll \theta \ \text{and} \ T_\sigma^n y \ll T_\theta^m x \ \text{for some} \ m,n \in \mathbb{Z}_+\}$$

and $\mathcal{X}'_{(x,\theta)-} = \partial\mathcal{X}_{(x,\theta)-} \cap \{(y,\sigma) \in X \times \Theta : \ \sigma \ll \theta\}$, where "$\partial$" stands for the boundary symbol in $\mathcal{X}$. The sets $\mathcal{X}_{(x,\theta)+}$ and $\mathcal{X}'_{(x,\theta)+}$ are defined analogously with the reversed ordering. Observe that if $\mathcal{X} = \mathcal{S}_- \cup \mathcal{U}_-$ then

$$\mathcal{X}_{(x,\theta)-} = \{(y,\sigma) \in \mathcal{X} : \ \sigma \ll \theta \ \text{and} \ \omega_\sigma(y) \prec\!\prec \omega_{\theta-}(x)\}.$$

We say that a subset $\mathcal{J}$ of $\mathcal{X}$ is *simply ordered* (*simply strongly ordered*, respectively) if x, $y \in \mathcal{J}$ and $x \neq y$ imply either $x < y$ or $x > y$ (either $x \ll y$ or $x \gg y$). The reader is referred to H-H. Kuo [16] for general facts about *Gaussian measures* in Banach spaces, and to Aronszajn [6] and Phelps [20] for descriptions of their null sets. Some additional details about null sets can be found in Hirsch [14, Lemma 7.7].

THEOREM 3.4. *Let X and Θ be nonempty open subsets of separable strongly ordered Banach spaces V and V_Θ, respectively. Assume that every point in $\mathcal{X} = X \times \Theta$ is lower approximable, i.e., $\mathcal{X} = \mathcal{S}_- \cup \mathcal{U}_-$. Then we have the following statements.*

(a) *If $\mathcal{J}$ is a simply strongly ordered subset of $\mathcal{X}$, then $\mathcal{U}_- \cap \mathcal{J}$ is at most countable.*

(b) *If μ is a Gaussian measure on $\mathcal{V} = V \times V_\Theta$, then $\mu(\mathcal{U}_-) = 0$. In particular, $\mathcal{U}_-$ is a Borel set in $\mathcal{V}$.*

(c) *If $(x,\theta) \in \mathcal{X}$, then the sets $A = \mathrm{Cl}(\mathcal{X}_{(x,\theta)-})$ and $B = \mathcal{X} \setminus \mathcal{X}_{(x,\theta)-}$ form an invariant order decomposition of $\mathcal{X}$. Its boundary $H = A \cap B = \partial\mathcal{X}_{(x,\theta)-}$ is an invariant Lipschitz hypersurface in $\hat{\mathcal{V}}$ as described in Proposition* 1.3, $(x,\theta) \in H$ *and* $\mathcal{X}'_{(x,\theta)-} \subset \mathcal{U}_- \cap H$.

Corresponding statements hold for $\mathcal{S}_+$ and $\mathcal{U}_+$.

Proof. (a) Let $\mathcal{J} \in \mathcal{X}$ be simply strongly ordered. Given $(x,\theta) \in \mathcal{U}_-$, we define

$$U_{\theta-}(x) = \cup\{[[v,w]] : \ v \in \omega_{\theta-}(x) \text{ and } w \in \omega_\theta(x)\}$$

which is a nonempty open subset of X. If $(x,\theta) \ll (y,\sigma)$ in $\mathcal{U}_- \cap \mathcal{J}$, it is obvious that $\omega_\theta(x) \mathrel{-\!\!\!\ll} \omega_{\sigma-}(y)$ whence $U_{\theta-}(x) \cap U_{\sigma-}(y) = \emptyset$. Since X is separable, we conclude that $\mathcal{U}_- \cap \mathcal{J}$ must be at most countable.

(b) First we show that $\mathcal{U}_-$ is a Borel set. We employ Lemma 2.3. If $e \in \mathrm{Int}(V_+)$ and $\eta \in \mathrm{Int}((V_\Theta)_+)$ are fixed, then $\hat{d}(x,y) = |x-y|_e$ and $\hat{\delta}(\theta,\sigma) = |\theta - \sigma|_\eta$ define ordered metrics in $\hat{V}$ and $\hat{V}_\Theta$, respectively. Given $\epsilon > 0$, $\delta > 0$, $n \in \mathbb{Z}_+$, and $(y,\sigma) \in \mathcal{X}$, we denote by $\mathcal{U}_-^{\epsilon,\delta,n}(y,\sigma)$ the set of all $(x,\theta) \in \mathcal{X}$ such that

$$(y,\sigma) \ll (x,\theta), \quad \hat{\mathcal{D}}\big((x,\theta),(y,\sigma)\big) < \delta \quad \text{and} \quad \hat{d}(T_\theta^n x, T_\sigma^n y) > \epsilon.$$

Obviously $\mathcal{U}_-^{\epsilon,\delta,n}(y,\sigma)$ is open in $\mathcal{V}$. From Lemma 2.3 we obtain

$$\mathcal{U}_- = \bigcup_{\epsilon}\bigcap_{\delta}\bigcup_{n}\bigcup_{(y,\sigma)} \mathcal{U}_-^{\epsilon,\delta,n}(y,\sigma),$$

where we take ϵ and δ rational and positive, $n \in \mathbb{Z}_+$, and (y,σ) from a dense countable subset of $\mathcal{X}$. Consequently, $\mathcal{U}_-$ is a Borel set in $\mathcal{V}$.

Now let μ be a Gaussian measure on $\mathcal{V}$. Fix any $v \in \mathrm{Int}(\mathcal{V}_+)$. Define the line $x + \mathbb{R}^1 v = \{x + rv \in \mathcal{V} : r \in \mathbb{R}^1\}$ for every $x \in \mathcal{V}$. By Part (a) the set $\mathcal{U}_- \cap (x + \mathbb{R}^1 v)$ is at most countable, and hence, it has Lebesgue measure zero. Since $\mathrm{Int}(\mathcal{V}_+) \neq \emptyset$, it is a straightforward matter to verify that $\mathcal{U}_-$ is an *exceptional set* in $\mathcal{V}$ in the sense of Aronszajn [6, Def. I.1.2]. Applying a result of Phelps [20, Prop. 5], we conclude that $\mu(\mathcal{U}_-) = 0$.

(c) Let $(x,\theta) \in \mathcal{X}$. Taking $G = \mathcal{O}^+(x,\theta)$ in Proposition 1.2(i) we observe that the sets $A = \mathrm{Cl}(\mathcal{X}_{(x,\theta)-})$ and $B = \mathcal{X} \setminus \mathcal{X}_{(x,\theta)-}$ form an invariant order decomposition (A,B) of $\mathcal{X}$ whose boundary $H = A \cap B = \partial\mathcal{X}_{(x,\theta)-}$ is described in Proposition 1.3. Obviously, $(x,\theta) \in G \subset H$.

It remains to show $\mathcal{X}'_{(x,\theta)-} \subset \mathcal{U}_-$. We take any $(z,\varrho) \in H$ with $\varrho \ll \theta$. Since A is lower closed in a strongly ordered space $\mathcal{X}$, we can find a sequence $z_1 \ll z_2 \ll \cdots \ll z$ in X with $z_k \longrightarrow z$, and another sequence $\varrho_1 \ll \varrho_2 \ll \cdots \ll \varrho$ in Θ with $\varrho_k \longrightarrow \varrho$. From our definition of $\mathcal{X}_{(x,\theta)-}$ and A we obtain $(z_k, \varrho_k) \in \mathcal{X}_{(x,\theta)-}$, whence

$$\omega_{\varrho_k}(z_k) \mathrel{-\!\!\!\ll} \omega_{\theta-}(x) \quad \text{for all } k \in \mathbb{N}. \tag{1}$$

We claim also

$$\omega_{\varrho-}(z) = \bigcap_{n=1}^{\infty} \mathrm{Cl} \bigcup_{k=n}^{\infty} \omega_{\varrho_k}(z_k). \tag{2}$$

Namely, it is easy to deduce from $(\mathcal{T})$ that for every $m \in \mathbb{N}$ there exists $k > m$ such that $T_{\varrho_m} z \ll T_{\varrho_k} z_k$, and hence, $\omega_{\varrho_m}(z_m) \preceq \omega_{\varrho_m}(z) \prec\!\prec \omega_{\varrho_k}(z_k) \preceq \omega_{\varrho_k}(z)$ by Proposition 1.6. Applying Corollary 2.2 to these inequalities, we arrive at (2). Next, we combine (1) and (2) with $\varrho \ll \theta$ to obtain

$$\omega_{\varrho-}(z) \prec\!\prec \omega_{\theta-}(x) \preceq \omega_\theta(x). \tag{3}$$

On the other hand, $(x,\theta) \in H$ and $(z,\varrho) \in H$ imply $\omega(x,\theta) = (\omega_\theta(x),\theta) \subset H$ and $\omega(z,\varrho) = (\omega_\varrho(z),\varrho) \subset H$.

Finally, suppose $(z,\varrho) \in \mathcal{S}_-$. Then $\omega_\varrho(z) = \omega_{\varrho-}(z) \prec\!\prec \omega_\theta(x)$ by (3), and $\varrho \ll \theta$ as well, and, consequently, H must contain a pair of points (a,ϱ) and (b,θ) satisfying $a \ll b$. But $(a,\varrho) \ll (b,\theta)$ is impossible by Definition 1.1. We have proved also $(z,\varrho) \in \mathcal{U}_-$. This completes our proof. □

In contrast with the results in Takáč [30, Props. 5.5 and 5.6] for $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ strongly increasing we are unable to obtain a similar complete characterization of the sets $\mathcal{U}_-$ and $\mathcal{U}_+$. We can prove only the following consequence of Theorem 3.4(c). Here we write $(X,\sigma) = X \times \{\sigma\} \subset \mathcal{X}$ whenever $\sigma \in \Theta$.

COROLLARY 3.5. *Let all hypotheses of Theorem* 3.4 *be satisfied, including* $\mathcal{X} = \mathcal{S}_- \cup \mathcal{U}_-$. *In addition, assume* X *is connected. Let* $(x,\theta) \in X \times \Theta$, $(z,\varrho) \in \mathcal{X}'_{(x,\theta)-}$ *and* $\sigma \in \Theta$ *with* $\varrho \le \sigma \ll \theta$. *Then* $(X,\sigma) \cap \mathcal{X}'_{(x,\theta)-} \neq \emptyset$.

Finally, set $A^{\sigma,0} = \{y \in X : (y,\sigma) \in \mathcal{X}_{(x,\theta)-}\}$, $A^\sigma = \mathrm{Cl}(A^{\sigma,0})$, $B^\sigma = X \setminus A^{\sigma,0}$ *and* $X^\sigma = \{y \in X : (y,\sigma) \in \mathcal{X}'_{(x,\theta)-}\}$ *for* $\sigma \in \Theta$, $\sigma \ll \theta$, *and denote by* Λ *the set of all such* σ *for which the interior of* X^σ *in* X *is nonempty. Then the following statements are valid.*

(i) *If* $B^\sigma \neq \emptyset$ *then* (A^σ, B^σ) *is a* σ*-invariant order decomposition of* X, *and* $A^\sigma \cup X^\sigma \subset A^{\lambda,0} = A^\lambda \setminus X^\lambda$ *whenever* $\lambda \ll \sigma \ll \theta$ *in* Θ. *If* $\lambda \ll \theta$ *then* $A^{\lambda,0} = \cup\{A^{\sigma,0} : \lambda \ll \sigma \ll \theta\}$.

(ii) *If* $\sigma \notin \Lambda$ *then* $X^\sigma = A^\sigma \cap B^\sigma$, *and, in particular, if also* $\varrho \le \sigma \ll \theta$ *then* X^σ *is a nonempty* σ*-invariant Lipschitz d-hypersurface in* $\hat{V}$.

(iii) *Every simply strongly ordered subset of* Λ *is at most countable, and if* μ *is a Gaussian measure on* V_Θ *then* $\mu(\Lambda) = 0$. *In particular,* Λ *is a Borel set in* V_Θ.

Analogous results hold for $\mathcal{X}_{(x,\theta)+}$ *and* $X'_{(x,\theta)+}$.

Proof. Suppose there exists $\sigma \in \Theta$, $\varrho \le \sigma \ll \theta$, such that $(X,\sigma) \cap \mathcal{X}'_{(x,\theta)-} = \emptyset$. Set $A^\circ = \mathcal{X}_{(x,\theta)-}$ (which is open and lower closed in $\mathcal{X}$), $A = \mathrm{Cl}(A^\circ)$, $B = \mathcal{X} \setminus \mathcal{X}_{(x,\theta)-}$, and $B^\circ = \mathrm{Int}(B)$ (which is open and upper closed in $\mathcal{X}$). By Theorem 3.4(c), (A,B) is an invariant order decomposition of $\mathcal{X}$ with the boundary $H = A \cap B$. Since X is connected and $(X,\sigma) \cap H = \emptyset$, we have either $(X,\sigma) \subset A^\circ$ or else $(X,\sigma) \subset B^\circ$. The former case forces $(X,\varrho) \subset A^\circ$ by $\varrho \le \sigma$, thus contradicting $(z,\varrho) \in H$. The latter one forces $(X,\theta) \subset B^\circ$ by $\sigma \ll \theta$, thus contradicting $(x,\theta) \in H$. Hence, we have proved $(X,\sigma) \cap \mathcal{X}'_{(x,\theta)-} \neq \emptyset$ as desired.

Proof of (i). Set $\tilde{B}^{\sigma,0} = \{y \in X : (y,\sigma) \in B^\circ\}$ for $\sigma \in \Theta$, $\sigma \ll \theta$, and observe that $X = A^{\sigma,0} \cup X^\sigma \cup \tilde{B}^{\sigma,0}$ is a disjoint union of σ-invariant sets, where $A^{\sigma,0}$ is open and lower closed and $\tilde{B}^{\sigma,0}$ is open and upper closed in X. Consequently, $B^\sigma = X^\sigma \cup \tilde{B}^{\sigma,0}$, and it is easy to see that (A^σ, B^σ) is a σ-invariant order decomposition of X provided $B^\sigma \neq \emptyset$. Now let $\lambda \ll \sigma \ll \theta$ in Θ. Obviously, $A^{\sigma,0} \subset A^{\lambda,0} = A^\lambda \setminus X^\lambda$. To show also $X^\sigma \subset A^{\lambda,0}$, we fix any $y \in X^\sigma$. Then $(y,\sigma) \in H$ and $\lambda \ll \theta$ imply $\mathcal{T}(y,\lambda) \ll \mathcal{T}(y,\sigma) \in H$, whence $(y,\lambda) \in A^\circ$, and we have shown $X^\sigma \subset A^{\lambda,0}$. We conclude that $A^\sigma \cup X^\sigma = A^{\sigma,0} \cup X^\sigma \subset A^{\lambda,0}$.

Finally, let $\lambda \ll \theta$ in Θ. We have already proved $\cup\{A^\sigma : \sigma \in [[\lambda,\theta]]\} \subset A^{\lambda,0}$. On the other hand, if $y \in A^{\lambda,0}$ then the fact that $(y,\lambda) \in A^\circ$ and A° is open in a strongly ordered space $\mathcal{X} = X \times \Theta$ shows that $(y, \sigma) \in A^\circ$ for any $\sigma \in \Theta$ close enough to λ, whence $y \in A^{\sigma,0}$ for some $\sigma \in [[\lambda, \theta]]$. It follows that $A^{\lambda,0} = \cup\{A^{\sigma,0} : \sigma \in [[\lambda, \theta]]\}$ as desired.

Proof of (ii). It follows immediately from the proof of (i) that $B^\sigma = \mathrm{Cl}(\tilde{B}^{\sigma,0})$ and $X^\sigma = A^\sigma \cap B^\sigma$ for every $\sigma \notin \Lambda$. If also $\varrho \le \sigma \ll \theta$ then X^σ is nonempty by $(X, \sigma) \cap \mathcal{X}'_{(x,\theta)-} \neq \emptyset$, and X^σ is a Lipschitz d-hypersurface in $\hat{V}$ by Proposition 1.3. The invariance of X^σ under T_σ has been proved in (i).

Proof of (iii). From (i) we obtain $X^\lambda \cap X^\sigma = \emptyset$ whenever $\lambda \ll \sigma \ll \theta$. Since X is separable, every simply strongly ordered subset of Λ can be only finite or countable. To show that Λ is a Borel set in V_Θ we fix any $\eta \in \mathrm{Int}((V_\Theta)_+)$. Given $u \ll v$ in X and $\delta \in (0, \infty)$, we define $U^{u,v,\delta}$ to be the union of all order intervals $[[\alpha, \beta]]$ in Θ such that $(v, \alpha) \in A^\circ$, $(u, \beta) \in B^\circ$ and $0 \ll \beta - \alpha \ll \delta\eta$. Clearly $\sigma \in \cap_{\delta>0} U^{u,v,\delta}$ if and only if $[[u, v]] \subset X^\sigma$ for $\sigma \in \Theta$, $\sigma \ll \theta$. Hence $\Lambda = \cup_{u \ll v} \cap_{\delta>0} U^{u,v,\delta}$ which remains valid even if we take u and v from a countable dense subset of X, $\delta > 0$ rational, and α and β from a countable dense subset of Θ. We conclude that Λ is a Borel set in V_Θ. If μ is a Gaussian measure on V_Θ, then $\mu(\Lambda) = 0$ by the same arguments as in the proof of Theorem 3.4(b). The proof is now complete. □

Remark. Loosely speaking, Corollary 3.5 contains the following results for a given $(x, \theta) \in X \times \Theta$: The set $\mathcal{X}'_{(x,\theta)-}$ forms the boundary of the set $\mathcal{X}_{(x,\theta)-}$ in $X \times [[-\infty, \theta]]$ of all $(y, \sigma) \in \mathcal{X}$ such that $\sigma \ll \theta$ and $\omega_\sigma(y) \prec\!\!\prec \omega_{\theta-}(x)$. This boundary is an invariant Lipschitz hypersurface in $\hat{\mathcal{V}}$ which can be expressed as the union of the sets (X^σ, σ) for $\sigma \ll \theta$. Except for a set Λ of Gaussian measure zero in Θ, every X^σ, $\sigma \notin \Lambda$, is a σ-invariant Lipschitz hypersurface in $\hat{V}$ which is strictly decreasing in σ, i.e., $A^\sigma \cup X^\sigma \subset A^\lambda \setminus X^\lambda$ provided $\lambda \ll \sigma \ll \theta$. If X is an open order interval in V, and λ, $\sigma \notin \Lambda$ are such that $\lambda \ll \sigma \ll \theta$ and $X^\lambda \neq \emptyset$, then $X^\sigma \prec\!\!\prec X^\lambda$.

Combining Theorems 3.3 and 3.4(b), we obtain the following analogue of well-known results due to Hirsch [14, Thm. 7.8 and 8.10(d)] for strongly increasing continuous-time semiflows.

THEOREM 3.6. *Let all hypotheses of Theorem* 3.4 *be satisfied, together with* $\mathcal{X} = \mathcal{S}_- \cup \mathcal{U}_- = \mathcal{S}_+ \cup \mathcal{U}_+$. *Let* $(x, \theta) \in \mathcal{X} = X \times \Theta$, *and let* μ *be any Gaussian measure on* $\mathcal{V} = V \times V_\Theta$. *If* $(x, \theta) \in \mathcal{S} \equiv \mathcal{S}_- \cap \mathcal{S}_+$ *then* $\omega_\theta(x)$ *is a quasi cycle for* T_θ *contained in* $\mathcal{S}_\theta \equiv \mathcal{S}_{\theta-} \cap \mathcal{S}_{\theta+}$; *otherwise,* $(x, \theta) \in \mathcal{U} \equiv \mathcal{U}_- \cup \mathcal{U}_+$ *where* $\mu(\mathcal{U}) = 0$.

4. Examples. In this section we present three examples to which our results from §3 can be applied. Our first, Example 4.1, is a single reaction-diffusion equation depending upon parameters in both the reaction function and Robin's boundary conditions. Our second, Example 4.6, is a strictly cooperative system of ordinary differential equations depending upon parameters in the reaction functions. It is easy to see that these two examples can be combined into a strictly cooperative system of weakly coupled reaction-diffusion equations or into a system of two reaction-diffusion equations for two competing species, cf. Matano and Mimura [17]. Our last example, Example 4.10, illustrates and clarifies our results for the sets $\mathcal{U}_-$ and $\mathcal{U}_+$ obtained in §3.

Example 4.1. We consider the initial-boundary value problem (IBVP) for the following reaction-diffusion equation:

$$\frac{\partial u}{\partial t} + A(x,t)u = f(x,t,u) + \gamma(x,t)u \quad \text{in } \Omega \times (0,\infty); \tag{1}$$

$$\frac{\partial u}{\partial n} + \vartheta(x,t)u = 0 \quad \text{on } \partial\Omega \times (0,\infty); \tag{1_b}$$

$$(1_i) \qquad u(x,0) = u_0(x) \quad \text{in } \Omega.$$

Here, Ω is a bounded open domain in $\mathbb{R}^N$ with the boundary $\partial\Omega$ of class C^3 and, with the outer unit normal $n = n(x)$ to $\partial\Omega$ at $x \in \partial\Omega$, $A(x,t)$ is a uniformly strongly elliptic operator of the form

$$A(x,t) \equiv -\sum_{i,j=1}^{N} a_{ij}(x,t)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{N} a_i(x,t)\frac{\partial}{\partial x_i}$$

whose coefficients satisfy a_{ij}, $a_i \in C^2(\overline{\Omega} \times \mathbb{R}^1_+)$, the reaction function f satisfies $f \in C^2(\overline{\Omega} \times \mathbb{R}^1_+ \times \mathbb{R}^1)$, and $\gamma \in C^2(\overline{\Omega} \times \mathbb{R}^1_+)$. In Robin's boundary conditions we assume $0 \le \vartheta \in C^2(\partial\Omega \times \mathbb{R}^1_+)$. Finally, $u : \Omega \times \mathbb{R}^1_+ \longrightarrow \mathbb{R}^1$ is the unknown function of (x,t) with a prescribed initial distribution $u_0 : \Omega \longrightarrow \mathbb{R}^1$. All functions a_{ij}, a_i, f, γ, and ϑ entering (IBVP) are assumed to be *τ-periodic* in time $t \in \mathbb{R}^1_+$ with a period $\tau \in (0,\infty)$. We will express the τ-periodicity in time of a function $\varphi : \mathbb{R}^1_+ \longrightarrow \mathbb{R}^1$ by writing $\varphi : \mathbb{R}^1/\tau\mathbb{Z} \longrightarrow \mathbb{R}^1$.

In order to guarantee global existence in time of an *L_p-solution* u (cf. Amann [2], [3], [5]) to (IBVP) for every sufficiently smooth initial distribution u_0 such that $0 \le u_0(x) \le M$ for all $x \in \Omega$, we make the following two hypotheses.

(f): $f(x,t,0) \ge 0$ for all $(x,t) \in \Omega \times \mathbb{R}^1_+$.

(γ): There exists a constant $M \in (0,\infty)$ such that $\gamma(x,t) \le -f(x,t,M)/M$ for all $(x,t) \in \Omega \times \mathbb{R}^1_+$.

In particular, then $\underline{u}(x,t) \equiv 0$ is a *subsolution* and $\overline{u}(x,t) \equiv M$ is a *supersolution* of (IBVP). Consequently, if u is a *classical solution* to (IBVP) in $\Omega \times [0,t_0)$, for some $t_0 \in (0,\infty]$, i.e., $u \in C(\overline{\Omega}\times[0,t_0)) \cap C^{1,0}(\overline{\Omega}\times(0,t_0)) \cap C^{2,1}(\Omega\times(0,t_0))$ satisfies (IBVP) in $\Omega \times [0,t_0)$, then we may apply the maximum and boundary point principles for parabolic equations (cf. Protter and Weinberger [22]) to conclude that

$$(2) \qquad 0 \le u(x,t) \le M \quad \text{for all } (x,t) \in \overline{\Omega} \times [0,t_0).$$

Now we are ready to employ existence, uniqueness, and regularity results due to Amann [2], [3], [5] to construct our mapping $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$.

We denote by $W^s_p = W^s_p(\Omega)$the Sobolev–Slobodeckii interpolation spaces for $1 < p < \infty$ and $0 \le s < \infty$, and, by $C^\mu = C^\mu(\overline{\Omega})$, the Hölder spaces for $0 \le \mu \le \infty$, cf. Triebel [33, Chap. 4]. If $\mu < s - (N/p)$ then $W^s_p \hookrightarrow C^\mu$ is a compact imbedding, and both W^s_p and C^μ are strongly ordered Banach spaces with the *pointwise ordering.* If u, $v : \Omega \longrightarrow \mathbb{R}^1$ are Lebesgue measurable functions, we write $u \le v$ if and only if $u(x) \le v(x)$ for almost everywhere $x \in \Omega$.

Next, we fix real numbers p and μ such that

$$(\mathrm{H}) \qquad 2 \le p < \infty \quad \text{and} \quad 0 < \mu < \frac{1}{2} - \frac{N}{p}.$$

Finally, we combine results of Amann [3, Thm. 2.1(i, ii) and Cor. 2.2] with our a priori estimate (2) to conclude that, given any $u_0 \in [0,M]_{W^1_p}$, where

$$[0,M]_{W^1_p} = \{\varphi \in W^1_p : \ 0 \le \varphi \le M\},$$

there exists a unique global *L_p-solution* $u : \mathbb{R}^1_+ \longrightarrow W^1_p$ of our (IBVP), cf. Amann [2, §15] or [3, §2] or [5, §7] for its definition. Moreover, u is also a classical solution of (IBVP) satisfying

$$(3) \qquad u \in C(\mathbb{R}^1_+ \longrightarrow C^\mu) \cap C^1((0,\infty) \longrightarrow C^\mu) \cap C((0,\infty) \longrightarrow C^{\mu+2}).$$

We set $V = W_p^1$, $\tilde{V}_\gamma = C^2(\overline{\Omega} \times (\mathbb{R}^1/\tau\mathbb{Z}))$, $\tilde{V}_\vartheta = C^2(\partial\Omega \times (\mathbb{R}^1/\tau\mathbb{Z}))$ and $\tilde{V}_\Theta = \tilde{V}_\gamma \times \tilde{V}_\vartheta$. From now on we assume that all a_{ij}, a_i, and f in (1) are fixed, and we also fix an arbitrary constant $M \in (0,\infty)$. We denote $\overline{X} = [0, M]_V$;

$\tilde{\Theta}_\gamma$ is the set of all $\gamma \in \tilde{V}_\gamma$ such that the hypothesis (γ) is satisfied for this M, i.e., $\gamma \le -f(\,\cdot\,, \cdot\,, M)/M$; and $\tilde{\Theta}_\vartheta = (\tilde{V}_\vartheta)_+$, the positive cone in $\tilde{V}_\vartheta$. Then $\overline{X}$ is a closed subset of V, and $\tilde{\Theta} = \tilde{\Theta}_\gamma \times \tilde{\Theta}_\vartheta$ is a closed subset of $\tilde{V}_\Theta$ whose elements are denoted by $\theta = (\gamma, \vartheta)$. All these spaces are endowed with the pointwise ordering.

Next we apply the strong maximum and boundary point principles (cf. Protter and Weinberger [22]) in a way very similar to Hirsch [14, Proof of Thm. 4.1] to derive the following strongly monotone dependence of the solution $u(t, u_0, \theta)$ to (IBVP) upon $u_0 \in \overline{X}$ and $\theta \in \tilde{\Theta}$ for each fixed $t \in (0,\infty)$.

PROPOSITION 4.2. (a) *If $u_0 < u_0'$ in $\overline{X}$, $\theta \in \tilde{\Theta}$ and $t \in (0,\infty)$, then we have*

$$u(t, u_0, \theta) \ll u(t, u_0', \theta) \quad \text{in } \overline{X}.$$

(b) *Let $0 \ll u_0 \le u_0'$ in $\overline{X}$, $\theta \le \theta'$ in $\tilde{\Theta}$, and $t \in (0,\infty)$, and assume $(u_0, \theta) \ne (u_0', \theta')$. Then we have*

$$0 \ll u(t, u_0, \theta) \ll u(t, u_0', \theta') \quad \text{in } \overline{X}.$$

Observe that $\varphi \ll \varphi'$ in $\overline{X}$ means $\min_{\overline{\Omega}}(\varphi' - \varphi) > 0$ for φ, $\varphi' \in \overline{X}$.

Amann [5, Thm. 7.3(i)] has shown that, when regarded as a mapping of $(t, u_0, \theta) \in \mathbb{R}_+^1 \times \overline{X} \times \tilde{\Theta} \subset \mathbb{R}_+^1 \times W_p^1 \times C^2(\partial\Omega \times (\mathbb{R}^1/\tau\mathbb{Z}))$ into $\overline{X} \subset W_p^1$, the solution $u(\,\cdot\,, \cdot\,, \cdot\,)$ of our (IBVP) is continuous and also Lipschitz continuous with respect to both u_0 and θ locally uniformly in $\mathbb{R}_+^1 \times \overline{X} \times \tilde{\Theta}$, i.e., the following result is valid.

PROPOSITION 4.3. *We have*

$$u \in C^{0,1-,1-}(\mathbb{R}_+^1 \times \overline{X} \times \tilde{\Theta} \longrightarrow \overline{X}).$$

From Amann [5, Thm. 7.3(v)] and (3) we easily obtain the following compactness properties of the mapping $u(\tau, \cdot\,, \cdot\,) : \overline{X} \times \tilde{\Theta} \subset W_p^1 \times \tilde{V}_\Theta \longrightarrow \overline{X} \subset W_p^1$:

PROPOSITION 4.4. *Let Σ be any subset of $\tilde{\Theta}$ which is bounded in $\tilde{V}_\Theta$. Then $u(\tau, \overline{X}, \Sigma)$, the image of $\overline{X} \times \Sigma$ under the mapping $u(\tau, \cdot\,, \cdot\,)$, is a subset of $\overline{X} \cap C^{\mu+2}$ which is bounded in $C^{\mu+2}$, and, in particular, it is relatively compact in W_p^1.*

Definition of $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$. We set $X = [[0, M]]_V$ in $V = W_p^1$, i.e., $X = \operatorname{Int}_V(\overline{X})$. We choose V_γ to be any closed linear subspace of $\tilde{V}_\gamma$ such that $V_\gamma \cap \operatorname{Int}((\tilde{V}_\gamma)_+) \ne \emptyset$ in $\tilde{V}_\gamma$, and V_ϑ to be any closed linear subspace of $\tilde{V}_\vartheta$ such that $V_\vartheta \cap \operatorname{Int}((\tilde{V}_\vartheta)_+) \ne \emptyset$ in $\tilde{V}_\vartheta$. Hence, $V_\Theta = V_\gamma \times V_\vartheta$ is a strongly ordered Banach space. The following three examples present the basic choices for V_γ, and similarly for V_ϑ: (i) $V_\gamma = \tilde{V}_\gamma = C^2(\overline{\Omega} \times (\mathbb{R}^1/\tau\mathbb{Z}))$, (ii) $V_\gamma = C^2(\overline{\Omega}) \subset \tilde{V}_\gamma$ (temporally independent functions), or (iii) $V_\gamma = \mathbb{R}^1 \subset \tilde{V}_\gamma$ (constant functions).

We may also choose one of the spaces V_γ or V_ϑ to be trivial, thus eliminating one of the parameters γ or ϑ. Next we choose Θ to be the set of all $\theta = (\gamma, \vartheta) \in V_\Theta$ such that $\gamma \ll -f(\,\cdot\,, \cdot\,, M)/M$ in V_γ and $\vartheta \gg 0$ in V_ϑ, i.e., $\Theta = \operatorname{Int}_{V_\Theta}(\tilde{\Theta} \cap V_\Theta)$ in V_Θ. Finally, we set $\mathcal{X} = X \times \Theta$ and $\mathcal{V} = V \times V_\Theta$, and define $\mathcal{T}$ by

$$\mathcal{T}(u_0, \theta) = \big(u(\tau, u_0, \theta), \theta\big) \quad \text{for all } (u_0, \theta) \in \mathcal{X}.$$

It follows from Propositions 4.2(b) and 4.3 that $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ satisfies hypothesis ($\mathcal{T}$) in §1. By Proposition 4.4, given any $\theta \in \Theta$, the θ-orbit $\mathcal{O}_\theta^+(u_0)$ of every $u_0 \in X$ is relatively compact in $\overline{X} \subset V$, whence $\omega_\theta(u_0) \subset \overline{X}$, and, more precisely, we have either $\omega_\theta(u_0) = \{0\}$ or $\omega_\theta(u_0) \subset X$ or $\omega_\theta(u_0) = \{M\}$ by Propositions 4.2(a) and 1.5. The case $\omega_\theta(u_0) = \{M\}$ cannot occur because $u(t, M, \theta) \ll M$ for $t > 0$, by the strong maximum and boundary point principles. The case $\omega_\theta(u_0) = \{0\}$ is easily excluded by strengthening hypothesis (f) to

$$(\mathrm{f}') : f(\,\cdot\,, \cdot\,, 0) \geq 0 \quad \text{in } \overline{\Omega} \times \mathbb{R}^1_+ \text{ and } f(x', t', 0) > 0 \quad \text{for some } (x', t') \in \overline{\Omega} \times \mathbb{R}^1_+.$$

From (f′) we derive $u(t, 0, \theta) \gg 0$ for $t > 0$. We conclude from Proposition 4.4 that every point in $\mathcal{X}$ is both lower and upper approximable, and, therefore, we can apply Theorems 2.5 and 3.6 to obtain the following theorem for our (IBVP).

THEOREM 4.5. *Assume that a_{ij}, a_i, and f satisfy the C^2-smoothness hypotheses stated above, together with the uniform strong ellipticity of $A(x, t)$ and with* (f'). *Then $\mathcal{U} = \mathcal{U}_- \cup \mathcal{U}_+$, the set of all ω-unstable points for $\mathcal{T}$, has zero Gaussian measure in $\mathcal{V}$.*

If $(u_0, \theta) \in \mathcal{S} = \mathcal{X} \setminus \mathcal{U}$ then $\omega_\theta(u_0)$ is a quasi cycle in $\mathcal{S}_\theta$ with the following three properties.

(i) *Let $w_0 \in \omega_\theta(u_0)$, and let $\theta_1, \theta_2 \in \Theta$ be such that $\theta_1 \ll \theta \ll \theta_2$. Then $\omega_{\theta_i}(w_0)$ is a k_i-cycle for T_{θ_i} independent from the choice of $w_0 \in \omega_\theta(u_0)$ for $i = 1, 2$. In particular, given any $w_i \in \omega_{\theta_i}(w_0)$ with $w_1 \leq w_2$, the function $u(\,\cdot\,, w_i, \theta_i) : \mathbb{R}^1_+ \longrightarrow W^1_p$ is a $k_i\tau$-periodic solution of* (IBVP) *with the parameter $\theta_i = (\gamma_i, \vartheta_i)$ and the initial distribution w_i.*

(ii) *Let w_0, θ_i, k_i, and w_i be as in* Part (i). *Then there exists $r \in \mathbb{R}^1_+$ such that*

$$u(s, w_1, \theta_1) \ll u(r + s, u_0, \theta) \ll u(s, w_2, \theta_2) \quad \textit{in } W^1_p \quad \textit{for all } s \in \mathbb{R}^1_+.$$

(iii) *If $\epsilon \in (0, \infty)$, one can choose $\theta_i \in \Theta$ in* Part (i) *so close to θ that for any choice of $w_0 \in \omega_\theta(u_0)$ and $w_i \in \omega_{\theta_i}(w_0)$ with $w_1 \leq w_2$ we have*

$$0 \ll u(\,\cdot\,, w_2, \theta_2) - u(\,\cdot\,, w_1, \theta_1) \leq \epsilon \quad \textit{in} \quad C(\overline{\Omega} \times \mathbb{R}^1_+).$$

Moreover, we can even choose θ_i so close to θ that there also exists $r_\epsilon \in \mathbb{R}^1_+$ such that

$$\|u(r_\epsilon + s, u_0, \theta) - u(s, w_i, \theta_i)\|_{W^1_p} \leq \epsilon \quad \textit{for all } s \in \mathbb{R}^1_+ \textit{ and } i = 1, 2.$$

Example 4.6. We consider the initial value problem (IVP) for the following system of n ordinary differential equations indexed by $k = 1, 2, \cdots, n$:

$$(4) \qquad \frac{du_k}{dt} = f_k(t, u_1, \cdots, u_n) + \sum_{\ell=1}^{n} \gamma_{k\ell}(t) u_\ell \quad \text{in } (0, \infty);$$

$$(4_i) \qquad u_k(0) = u_{k,0}.$$

Here, $u \equiv (u_1, \cdots, u_n) : \mathbb{R}^1_+ \longrightarrow \mathbb{R}^n$ denotes the unknown vector-valued function of time t with a prescribed initial vector $u_0 \equiv (u_{k,0}) \in \mathbb{R}^n$. We assume that the reaction functions f_k satisfy f_k, $\partial f_k/\partial u_\ell \in C(\mathbb{R}^1_+ \times \mathbb{R}^n)$, and $\gamma_{k\ell} \in C(\mathbb{R}^1_+)$. Let $\gamma = (\gamma_{k\ell})$ denote the matrix with the entries $\gamma_{k\ell}$, $1 \leq k$, $\ell \leq n$, and let $\gamma_k = (\gamma_{k\ell})_k$ denote its kth row. We assume that, for each $t \in \mathbb{R}^1_+$, the Jacobian matrix $((\partial f_k/\partial u_\ell) + \gamma_{k\ell})$ of the vector field $f(t,\cdot)+\gamma(t)=\big(f_k(t,\cdot)+\gamma_k(t)\big):\mathbb{R}^n\longrightarrow\mathbb{R}^n$ is *strictly cooperative* in $\mathbb{R}^n$.

(SC): If $k\neq\ell$ and $(t,u)\in\mathbb{R}^1_+\times\mathbb{R}^n$, then

$$\gamma_{k\ell}(t)>-\frac{\partial f_k}{\partial u_\ell}(t,u).$$

In order to guarantee global existence in time of a *classical solution* $u\in C^1(\mathbb{R}^1_+\longrightarrow\mathbb{R}^n)$ to (IVP) for every initial vector $u_0\in\mathbb{R}^n$ such that $0\le u_{k,0}\le M_k$ for $1\le k\le n$, we make the following two hypotheses.

(f): $f_k(t,0,\cdots,0)\ge 0$ for all $t\in\mathbb{R}^1_+$ and $k=1,\cdots,n$.

(γ): There exist constants $M_k\in(0,\infty)$, $1\le k\le n$, such that

$$\sum_{\ell=1}^{n}\gamma_{k\ell}(t)M_\ell\le -f_k(t,M_1,\cdots,M_k)\quad\text{for all } t\in\mathbb{R}^1_+.$$

In particular, then $0\le u_k(t)\le M_k$ for all $t\in\mathbb{R}^1_+$ provided $0\le u_{k,0}\le M_k$. All functions f_k and $\gamma_{k\ell}$ entering (IVP) are assumed to be *τ-periodic* in time $t\in\mathbb{R}^1_+$ with a period $\tau\in\mathbb{R}^1_+\setminus\{0\}$. The Euclidean space $\mathbb{R}^n$ is strongly ordered by the *coordinatewise ordering*: If $u=(u_k)\in\mathbb{R}^n$ and $v=(v_k)\in\mathbb{R}^n$, we write $u\le v$ if and only if $u_k\le v_k$ for $k=1,2,\cdots,n$.

We set $V=\mathbb{R}^n$ and $\tilde{V}_\Theta=C(\mathbb{R}^1/\tau\mathbb{Z}\longrightarrow\mathbb{R}^{n\times n})$. From now on we assume that all f_k in (4) are fixed, and we also fix an arbitrary vector $M=(M_k)\in\mathbb{R}^n$ with $M_k>0$ for $k=1,2,\cdots,n$. We denote $\overline{X}=[0,M]_V$; and $\tilde{\Theta}$ is the set of all $\gamma\in\tilde{V}_\Theta$ such that the hypotheses (SC) and (γ) are satisfied. The space $\tilde{V}_\Theta$ is endowed with the pointwise and coordinatewise ordering.

Next we apply the Müller–Kamke theorem (cf. Müller [18]) to derive the following strongly monotone dependence of the solution $u(t,u_0,\gamma)$ to (IVP) upon $u_0\in\overline{X}$ and $\gamma\in\tilde{\Theta}$, for each $t\in(0,\infty)$:

PROPOSITION 4.7. *All statements in Proposition* 4.2 *are valid also for* (IVP) *if we write $\theta=\gamma$.*

It is easy to prove also the following continuity properties of the solution u of our (IVP) with respect to t, u_0, and γ.

PROPOSITION 4.8. *We have*

$$u\in C^{1,1,1-}(\mathbb{R}^1_+\times\overline{X}\times\tilde{\Theta}\longrightarrow\overline{X}).$$

Definition of $\mathcal{T}:\mathcal{X}\longrightarrow\mathcal{X}$. We set $X=[[0,M]]_V$ in $V=\mathbb{R}^n$, i.e., $X=\mathrm{Int}_V(\overline{X})$. We choose V_Θ to be any closed linear subspace of $\tilde{V}_\Theta$ such that $V_\Theta\cap\mathrm{Int}((\tilde{V}_\Theta)_+)\neq\emptyset$ in $\tilde{V}_\Theta$. The following two examples present the basic choices for V_Θ: (i) $V_\Theta=\tilde{V}_\Theta=C(\mathbb{R}^1/\tau\mathbb{Z}\longrightarrow\mathbb{R}^{n\times n})$, or (ii) $V_\Theta=\mathbb{R}^{n\times n}\subset\tilde{V}_\Theta$ (matrices with constant entries).

Next, we choose Θ to be the set of all $\gamma\in V_\Theta$ such that

$$\gamma_{k\ell}(t)>-\inf_{u\in\mathbb{R}^n}\frac{\partial f_k}{\partial u_\ell}(t,u)\quad\text{for all } k\neq\ell \text{ and } t\in\mathbb{R}^1/\tau\mathbb{Z}$$

and

$$\gamma_{kk}(t)<-M_k^{-1}\left[f_k(t,M_1,\cdots,M_n)+\sum_{\substack{\ell=1\\ \ell\neq k}}^{n}\gamma_{k\ell}(t)M_\ell\right]\quad\text{for all } k \text{ and } t\in\mathbb{R}^1/\tau\mathbb{Z}.$$

Equivalently, $\Theta = \mathrm{Int}_{V_\Theta}(\tilde{\Theta} \cap V_\Theta)$ in V_Θ. Finally, we set $\mathcal{X} = X \times \Theta$ and $\mathcal{V} = V \times V_\Theta$ and define $\mathcal{T}$ by

$$\mathcal{T}(u_0, \gamma) = \big(u(\tau, u_0, \gamma), \gamma\big) \quad \text{for all } (u_0, \gamma) \in \mathcal{X}.$$

It follows from Propositions 4.7 and 4.8 that $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ satisfies hypothesis $(\mathcal{T})$ in §1. Since $\overline{X} = [0, M]_{\mathbb{R}^n}$ is compact, every point in $\mathcal{X}$ is both lower and upper approximable provided $\omega_\theta(u_0) \subset X$ for all $(u_0, \theta) \in \mathcal{X}$. Similarly, as in Example 4.1, we can achieve $\omega_\theta(u_0) \subset X$ by strengthening hypothesis (f) to

(f'): $f_k(t, 0, \cdots, 0) \geq 0$ for all $t \in \mathbb{R}^1_+$ and $k = 1, \cdots, n$, and $f_{k'}(t', 0, \cdots, 0) > 0$ for some $t' \in \mathbb{R}^1_+$ and k'.

Finally, we can apply Theorems 2.5 and 3.6 to obtain the following analogue of Theorem 4.5 for our (IVP).

THEOREM 4.9. *Assume that f_k satisfy $f_k, \partial f_k/\partial u_\ell \in C((\mathbb{R}^1/\tau\mathbb{Z}) \times \mathbb{R}^n)$, and* (f'). *Then $\mathcal{U} = \mathcal{U}_- \cup \mathcal{U}_+$ has zero Gaussian measure in $\mathcal{V}$. In particular, if* $\dim(V_\Theta) \equiv n_\theta < \infty$ *then also* $\dim(\mathcal{V}) = n + n_\Theta < \infty$*, and the $(n + n_\Theta)$-dimensional Lebesgue measure of $\mathcal{U}$ is zero.*

If $(u_0, \theta) \in \mathcal{S} = \mathcal{X} \setminus \mathcal{U}$ *then $\omega_\theta(u_0)$ is a quasi cycle in $\mathcal{S}_\theta$ having all properties* (i), (ii), *and* (iii) *from Theorem* 4.5 *with $\mathbb{R}^n$ in place of W^1_p.*

Remark. Problem $(\mathbf{P}')$ from the Introduction can be treated in the same way as Example 4.6. More generally, in both problems $(\mathbf{P})$ and $(\mathbf{P}')$, in place of $\sum_{\ell=1}^n \gamma_{k\ell}(x, t) u_\ell$ and $\gamma_k(t)$, respectively, we could consider any polynomial $p_k(u_1, \cdots, u_n)$ in the variables $u_1, \cdots, u_n \in [0, \infty)$ whose coefficients would play the rôle of (possibly space- and/or time-dependent) parameters. Naturally, condition (γ) from Examples 4.1 and 4.6 might then become more complicated.

Example 4.10. We consider the following single autonomous ordinary differential equation:

$$\frac{du}{dt} = \big(\theta - \varphi(u)\big) u \quad \text{for } t \in (0, \infty); \tag{5}$$

$$u(0) = u_0. \tag{5_i}$$

Here, $\varphi \in C^{1-}(\mathbb{R}^1_+)$, i.e., $\varphi : \mathbb{R}^1_+ \longrightarrow \mathbb{R}^1$ is locally Lipschitz continuous, and $\theta \in \mathbb{R}^1_+$ is a parameter. We assume that $\varphi(0) = 0$ and $\varphi(u) \longrightarrow \infty$ as $u \longrightarrow \infty$. The unknown function $u \in C^1(\mathbb{R}^1_+)$ exists globally in time for every initial value $u_0 \in \mathbb{R}^1_+$. Moreover, the limit $u(\infty) \equiv \lim_{t \to \infty} u(t)$ exists in $\mathbb{R}^1_+$ and satisfies

$$u(\infty) \in \{0\} \cup \varphi^{-1}(\theta),$$

where $\varphi^{-1}(\theta) = \{v \in \mathbb{R}^1_+ : \varphi(v) = \theta\}$ is closed in $\mathbb{R}^1_+$. Clearly, $u(\infty) > 0$ provided $u_0 > 0$ and $\theta > 0$. Again, we write $u(t, u_0, \theta)$ to indicate the dependence of the solution u of (5) upon t, u_0, and θ. We have

$$u \in C^{1,1-,1-}(\mathbb{R}^1_+ \times \mathbb{R}^1_+ \times \mathbb{R}^1_+ \longrightarrow \mathbb{R}^1_+),$$

and if $t \in (0, \infty)$, $u_0 \leq u_0'$, and $\theta \leq \theta'$ in $(0, \infty)$, and $(u_0, \theta) \neq (u_0', \theta')$, then also

$$0 < u(t, u_0, \theta) < u(t, u_0', \theta').$$

We set $V = V_\Theta = \mathbb{R}^1$ and $X = \Theta = (0, \infty)$ and fix any period $\tau \in (0, \infty)$. We define $\mathcal{T} : \mathcal{X} \longrightarrow \mathcal{X}$ by $\mathcal{T}(u_0, \theta) = \big(u(\tau, u_0, \theta), \theta\big)$ for $(u_0, \theta) \in \mathcal{X} = (0, \infty)^2$. Then $\mathcal{T}$

satisfies hypothesis $(\mathcal{T})$, and every point in $\mathcal{X}$ is both lower and upper approximable. It is evident from Lemma 2.3 that the sets $\mathcal{S}_-$ and $\mathcal{S}_+$, and thus $\mathcal{U}_- = \mathcal{X} \setminus \mathcal{S}_-$ and $\mathcal{U}_+ = \mathcal{X} \setminus \mathcal{S}_+$ as well, are independent from our choice of $\tau \in (0, \infty)$. They can be determined from the graph of φ as follows.

LEMMA 4.11. *Let $\theta \in (0,\infty)$ be fixed. Then $(0,\infty) \setminus \varphi^{-1}(\theta) = \cup_{n \in K}(a_n, b_n)$ is the union of at most countably many pairwise disjoint, nonempty open intervals (a_n, b_n), for $n \in K \subset \mathbb{N}$. Furthermore, given any $n \in K$ and $c_n \in (a_n, b_n)$, we have the following two alternatives.*

(i) *$\varphi(c_n) < \theta$ in which case $\varphi(x) < \theta$ and $u(\infty, x, \theta) = b_n \in \mathcal{S}_{\theta -}$ for all $x \in (a_n, b_n)$, and $a_n \in \mathcal{U}_{\theta +}$; in particular, we have $(a_n, b_n] \subset \mathcal{S}_{\theta -}$. Furthermore, either $(a_n, b_n] \subset \mathcal{S}_{\theta +}$ or else $(a_n, b_n] \subset \mathcal{U}_{\theta +}$.*

(ii) *$\varphi(c_n) > \theta$ in which case $\varphi(x) > \theta$ and $u(\infty, x, \theta) = a_n \in \mathcal{S}_{\theta +}$ for all $x \in (a_n, b_n)$, and $b_n \in \mathcal{U}_{\theta -}$; in particular, we have $[a_n, b_n) \subset \mathcal{S}_{\theta +}$. Furthermore, either $[a_n, b_n) \subset \mathcal{S}_{\theta -}$ or else $[a_n, b_n) \subset \mathcal{U}_{\theta -}$.*

LEMMA 4.12. *Let $\theta \in (0, \infty)$ and $x \in \varphi^{-1}(\theta)$ be fixed. Then $(x, \theta) \in \mathcal{S}_-$ if and only if there exists a sequence $x_1 < x_2 < \cdots < x$ in $(0,\infty)$ such that $x_n \longrightarrow x$ and $\varphi(x_n) < \theta$ for all $n \in \mathbb{N}$.*

An analogous statement holds for $\mathcal{S}_+$.

It follows from these two lemmas (whose proofs are simple exercises) that in order to determine the sets $\mathcal{S}_-$ and $\mathcal{S}_+$ we must first compute their intersections with the graph of φ by Lemma 4.12 and then $\mathcal{S}_-$ and $\mathcal{S}_+$ by Lemma 4.11. Roughly speaking, they can be as complicated as the graph of φ.

Acknowledgment. The author expresses his appreciation to C. Dennis Thron, M.D. (Dartmouth Medical School), for helpful discussions concerning applications of the results reported in this article to mathematical problems in pharmacology.

REFERENCES

[1] 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.

[2] H. AMANN, *Existence and regularity for semilinear parabolic evolution equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 11 (1984), pp. 593–676.

[3] ———, *Global existence for semilinear parabolic systems*, J. Reine Angew. Math., 360 (1985), pp. 47–83.

[4] ———, *Dynamic theory of quasilinear parabolic equations-* I. *Abstract evolution equations*, Nonlinear Anal., 12 (1988), pp. 895–919.

[5] ———, *Dynamic theory of quasilinear parabolic equations-* II. *Reaction-diffusion systems*, Differential Integral Equations, 3 (1990), pp. 13–75.

[6] N. ARONSZAJN, *Differentiability of Lipschitzian mappings between Banach spaces*, Studia Math., 57 (1976), pp. 147–190.

[7] X-Y. CHEN AND H. MATANO, *Convergence, asymptotic periodicity, and finite-point blow-up in one-dimensional semilinear heat equations*, J. Differential Equations, 78 (1989), pp. 160–190.

[8] P. C. FIFE, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer-Verlag, Berlin, New York, 1979.

[9] J. K. HALE AND C. ROCHA, *Interaction of diffusion and boundary conditions*, Nonlinear Anal., 11 (1987), pp. 633–649.

[10] P. HESS, *On stabilization of discrete strongly order-preserving semigroups and dynamical processes*, Proc. Trends in Semigroup Theory and Applications, Trieste, 1987; Lecture Notes in Pure and Applied Math., Marcel Dekker, New York, 1988.

[11] M. W. HIRSCH, *Differential equations and convergence almost everywhere in strongly monotone semiflows*, Contemp. Math., 17 (1983), pp. 267–285.

[12] ———, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc., 11 (1984), pp. 1–64.

[13] ———, *Attractors for discrete-time monotone dynamical systems in strongly ordered spaces*, in Geometry and Topology, Lecture Notes in Math. 1167, Springer-Verlag, Berlin, New York, 1985, pp. 141–153.

[14] ———, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1–53.

[15] ———, *Systems of differential equations that are competitive or cooperative:* III, *Competing species*, Nonlinearity, 1 (1988), pp. 51–71.

[16] H-H. KUO, *Gaussian Measures in Banach Spaces*. Lecture Notes in Math., 463, Springer-Verlag, Berlin, New York, 1975.

[17] H. MATANO AND M. MIMURA, *Patern formation in competition-diffusion systems in nonconvex domains*, Publ. Res. Inst. Math. Sci., 19 (1983), pp. 1049–1079.

[18] M. MÜLLER, *Über das Fundamentaltheorem in der Theorie der gewöhnlichen Differentialgleichungen*, Math. Z., 26 (1926), pp. 619–645.

[19] H. OTHMER, *The qualitative dynamics of a class of biochemical control circuits*, J. Math. Biol., 3 (1976), pp. 53–78.

[20] R. R. PHELPS, *Gaussian null sets and differentiability of Lipschitz map on Banach spaces*, Pacific J. Math., 77 (1978), pp. 523–531.

[21] P. POLÁČIK, *Convergence in smooth strongly monotone flows defined by semilinear parabolic equations*, J. Differential Equations, 79 (1989), pp. 89–110.

[22] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[23] H. H. SCHAEFER, *Topological Vector Spaces*, Springer-Verlag, Berlin, New York, 1971.

[24] S. SMALE, *On the differential equations of species in competition*, J. Math. Biol., 3 (1976), pp. 5–7.

[25] H. L. SMITH, *Monotone semiflows generated by functional differential equations*, J. Differential Equations, 66 (1987), pp. 420–442.

[26] H. L. SMITH AND H. R. THIEME, *Quasiconvergence and stability for strongly order-preserving semiflows*, SIAM J. Math. Anal., 21 (1990), pp. 673–692.

[27] ———, *Convergence for strongly order-preserving semiflows*, SIAM J. Math. Anal., 22 (1991), pp. 1081–1101.

[28] P. TAKÁČ, *Convergence to equilibrium on invariant d-hypersurfaces for strongly increasing discrete-time semigroups*, J. Math. Anal. Appl., 148 (1990), pp. 223–244.

[29] ———, *Asymptotic behavior of discrete-time semigroups of sublinear, strongly increasing mappings with applications in biology*, Nonlinear Anal., 14 (1990), pp. 35–42.

[30] ———, *Domains of attraction of generic ω-limit sets for strongly monotone discrete-time semigroups*, J. Reine Angew. Math., to appear.

[31] C. D. THRON, *Pharmacokinetic Instability*, in Mathematical Biology: Proceedings of the International Conference on Biomathematics, Chen Lansun, G.P. Patil, You Zhaoyong, Li Dianmo, and Ma Zhien, eds., Xi'an, China, 1988, Xi'an Jiaotong University Press, 1988.

[32] ———, Private communication (1990).

[33] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland Publishing, Amsterdam, 1978.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 412–448, March 1992

VECTOR-VALUED TAUBERIAN THEOREMS AND ASYMPTOTIC BEHAVIOR OF LINEAR VOLTERRA EQUATIONS*

WOLFGANG ARENDT† AND JAN PRÜSS‡

Abstract. The asymptotic behavior of the solutions of linear Volterra equations in a Banach space X of the form

$$(*)\qquad u(t) = f(t) + \int_0^t a(t-\tau)Au(\tau)d(\tau), \quad t \geq 0$$

is studied, in particular that of the resolvent $S(t)$ for $(*)$; here $a \in L^1_{\text{loc}}(\mathbb{R}_+)$ and A is a closed linear operator in X with dense domain. A complete characterization of the existence of $\lim_{t\to 0} S(t)x = Px$ for all $x \in X$ in the sense of Abel is obtained, and the nature of the ergodic limit P is studied. By means of vector-valued Tauberian theorems for the Laplace transform, a general result on convergence of $S(t)$ in the strong sense is derived. Several examples are given which illustrate this result, and also an application to the theory of linear viscoelasticity is presented.

Key words. Volterra equations, resolvents, asymptotic behavior, Laplace transform, ergodic limit, Abel-limit, Cesaro-limit, Tauberian theorems, C_0-semigroups, cosine families, viscoelasticity

AMS(MOS) subject classifications. primary 45N05, 45K05, 44A10; secondary 47D05, 47G05, 76A10

1. Introduction. Let X be a Banach space, $a \in L^1_{\text{loc}}(\mathbb{R}_+)$, A a closed linear operator in X with dense domain $D(A)$, and consider the abstract linear Volterra equation in X

$$(1.1)\qquad u(t) = f(t) + \int_0^t a(t-\tau)Au(\tau)d\tau, \qquad t \geq 0,$$

where $f : \mathbb{R}_+ \to X$ is continuous, $\mathbb{R}_+ = [0, \infty)$. X_A denotes the Banach space $D(A)$ equipped with the graph norm $|.|_A$ of A. A function $u \in C(\mathbb{R}_+; X_A)$ satisfying (1.1) on $\mathbb{R}_+$ is called a *strong solution* of (1.1), while $u \in C(\mathbb{R}_+; X)$ is a *mild solution* of (1.1) if $a * u \in C(\mathbb{R}_+; X_A)$ holds and

$$(1.2)\qquad u(t) = f(t) + A\int_0^t a(t-\tau)u(\tau)d\tau, \qquad t \geq 0,$$

is satisfied on $\mathbb{R}_+$. A family $\{S(t)\}_{t\geq 0} \subset \mathcal{B}(X)$ of bounded linear operators in X is called a *resolvent* for (1.1) if $S(t)$ commutes with A and satisfies the *resolvent equation*

$$(1.3)\qquad S(t)x = x + \int_0^t a(t-\tau)AS(\tau)x d\tau, \qquad t \geq 0, \quad x \in D(A).$$

Once a resolvent $S(t)$ for (1.1) is known to exist, it is unique, and the solution of (1.1) is represented by the *variation of parameters formula*

$$(1.4)\qquad u(t) = \frac{d}{dt}\int_0^t S(t-\tau)f(\tau)d\tau, \qquad t \geq 0,$$

* Received by the editors July 3, 1990; accepted for publication (in revised form) June 14, 1991. This research was partially supported by Deutsche Forschungsgemeinschaft, Bonn and by the Université de Franche-Comté, Besançon.

† Equipe de Mathématiques, URA 741 Centre National de la Recherche Scientifique, Université de Franche-Comté, 25030 Besançon Cedex, France.

‡ FB 17 Mathematik, Universität Paderborn, Warburger Strasse 100, 4790 Paderborn, Germany.

whenever u is a mild solution of (1.1), then $S * f \in C^1(\mathbb{R}_+; X)$ and u is represented by (1.4).

By now the question of existence of a resolvent for (1.1) has been settled for many different classes of pairs (a, A); for a general exposition of the theory, we refer to Prüss [33]. Here we always assume the existence of a resolvent $S(t)$ for (1.1) which is in addition of subexponential growth, i.e., which satisfies

$$\overline{\lim_{t\to\infty}} \frac{1}{t} \log |S(t)| \leq 0. \tag{1.5}$$

It is the purpose of this paper to study the asymptotic behavior of the solutions of (1.1), in particular that of the resolvent $S(t)$ itself. More precisely, the existence of the limits $\lim_{t\to\infty} u(t) = u(\infty)$ and $\lim_{t\to\infty} S(t) = P$ in various senses are investigated, and the nature of the limits $u(\infty)$ and P are discussed.

Our approach is based on the theory of vector-valued Laplace transforms. A well-known Abelian theorem shows that if $\lim_{t\to\infty} S(t)x = Px$ for all $x \in X$, then

$$H(\lambda) = \hat{S}(\lambda) = \int_0^\infty S(t)e^{-\lambda t}dt, \qquad \text{Re}\, \lambda > 0, \tag{1.6}$$

satisfies

$$A - \lim_{t\to\infty} S(t)x := \lim_{\lambda\to 0+} \lambda H(\lambda)x = Px \quad \text{for all } x \in X. \tag{1.7}$$

Therefore it is natural to study first the existence of the Abelian limit P of $S(t)$ as well as its properties. This will be done in §4, where we also apply some elementary vector-valued Tauberian theorems to deduce the convergence of $S(t)$ in the ordinary sense from existence of the ergodic limit $P \in \mathcal{B}(X)$; for that, several strong assumptions on $S(t)$ are needed. Once the Abelian limit $P \in \mathcal{B}(X)$ of $S(t)$ is known to exist, it follows easily that $A - \lim_{t\to\infty} u(t) = u(\infty)$ also exists whenever $f(t)$ admits an Abelian limit $f(\infty)$ and then $u(\infty) = Pf(\infty)$ holds.

The main result of this paper, the General Convergence Theorem stated and proved in §5, gives sufficient conditions for the strong convergence of $S(t)$ to its ergodic limit P as $t \to \infty$. For the special case $a(t) \equiv 1$ and A the generator of a bounded C_0-semigroup $T(t)$ in X, we have $S(t) = T(t)$ and the result reduces to a stability theorem for C_0-semigroups obtained recently by Arendt and Batty [2] and independently by Lyubich and Phong [28]; cf. also §7. The proof of the General Convergence Theorem relies on the complex Tauberian theory for the vector-valued Laplace transform. In fact, it is very much inspired by the proof of Arendt and Batty [2] for the semigroup case. However, due to the more complicated structure of the Laplace transform $H(\lambda)$ of the resolvent $S(t)$ for (1.1), i.e.,

$$H(\lambda) = \frac{1}{\lambda}(I - \hat{a}(\lambda)A)^{-1}, \qquad \text{Re}\, \lambda > 0, \tag{1.8}$$

the Tauberian arguments involved are more delicate and differ from those employed in the proof of Arendt and Batty [2].

Since Abelian and Tauberian theorems for the vector-valued Laplace transform are at the heart of our approach, and since there is no coherent presentation of this material available in the literature, we have included two sections on this matter. Section 2 contains the basic Abelian theorem, as well as the vector-valued extension of the classical real Tauberian theorems due to Hardy–Littlewood, Wiener, Pitt, and

Karamata; cf. Doetsch [15] and Widder [42] for their classical statements. Complex Tauberian theorems are presented in §3. Here a condition on the Laplace transform $\hat{f}$ of f is given which implies convergence of $f(t)$ $(t \to \infty)$. A first result of this type had been given in 1938 by Ingham [20], but recently a simple new technique of proof due to Newman [30] has led to considerable extensions; see Korevaar [25], Allan, O'Farrell, and Ransford [1], Arendt and Batty [2], Ransford [34], and Batty [3].

Section 6 is devoted to an elaboration of several examples and special cases of the theory developed in §§3–5. In particular, several classes of kernels are presented, for which the assumptions of the General Convergence Theorem reduce to boundedness of $S(t)$ (which is necessary for the existence of the strong limit of $S(t)$ anyway) and to a spectral condition that cannot be relaxed (and to some extent is also necessary). In §7, we apply our results to the theory of linear viscoelasticity. Here we show that if A generates a uniformly bounded cosine family and $a(t)$ is of the form

$$(1.9) \qquad a(t) = a_0 + a_\infty t + \int_0^t a_1(\tau)d\tau, \qquad t \geq 0$$

with $a_0, a_\infty \geq 0$, $a_1(t) \geq 0$ nonincreasing, $\log a_1(t)$ convex, and $\lim_{t\to\infty} a_1(t) = 0$, then $S(t)$ converges strongly as $t \to \infty$ if in addition $N(A)^\perp \cap N(A') = \{0\}$ and $a(t) \not\equiv a_\infty t$ hold. This result shows that any viscoelastic fluid in a smooth domain $\Omega \subset \mathbb{R}^n$ with compact boundary is asymptotically stable in the strong sense, whether Ω is bounded or not. It has been shown in Prüss [32] that viscoelastic fluids are uniformly asymptotically stable if and only if $A = P\Delta$ is invertible. This is always true for bounded domains Ω, but it is in general not the case for unbounded domains; cf. §7 for these concepts and further discussion.

2. Abelian and real Tauberian theorems. Throughout this section, $(X, |\ \ |)$ is a Banach space and $f \in L^1_{\text{loc}}([0,\infty), X)$ is such that

$$\hat{f}(\lambda) = \int_0^\infty e^{-\lambda t} f(t)dt := \lim_{b\to\infty} \int_0^b e^{-\lambda t} f(t)dt$$

exists for Re $\lambda > 0$ (this is equivalent to $\sup_{t\geq 0} e^{-\lambda t}|\int_0^t f(s)ds| < \infty$ for all $\lambda > 0$).

DEFINITION 2.1. Let $f_\infty \in X$. The function f converges to f_∞ in the sense of Cesaro $(t \to \infty)$ if $C - \lim_{t\to\infty} f(t) := \lim_{t\to\infty}(1/t)\int_0^t f(s)ds = f_\infty$, and f converges to f_∞ in the sense of Abel $(t \to \infty)$ if $A - \lim_{t\to\infty} f(t) := \lim_{\lambda\to 0+} \lambda\hat{f}(\lambda) = f_\infty$.

The following *Abelian theorem* is easy to prove (see [19, Thm. 18.2.1]). It will be convenient to introduce $F(t) = \int_0^t f(s)ds$ as an auxiliary function.

THEOREM 2.2. *Let* f_∞, $F_\infty \in X$.

(a) *If* $\lim_{t\to\infty} f(t) = f_\infty$, *then* $C - \lim_{t\to\infty} f(t) = f_\infty$.

(b) *If* $C - \lim_{t\to\infty} f(t) = f_\infty$, *then* $A - \lim_{t\to\infty} f(t) = f_\infty$.

(c) *If* $\lim_{t\to\infty} F(t) = F_\infty$, *then* $\lim_{\lambda\to 0+} \hat{f}(\lambda) = F_\infty$.

Note that (c) is a special case of (b) since $\hat{f}(\lambda) = (F')^\wedge(\lambda) = \lambda\hat{F}(\lambda) \quad (\lambda > 0)$.

A result if called a *Tauberian theorem* if a condition on f is given under which the converse implications of (a), (b), or (c) are valid. Such theorems are presented in sections A, B, C, D, and E respectively. Most of these results are well known at least in the numerical case. We include proofs here for the sake of completeness.

Our main objective is to find conditions under which Abelian convergence implies convergence (see §D).

A. Conditions under which $C-\lim_{t\to\infty} f(t) = f_\infty$ **implies** $\lim_{t\to\infty} f(t) = f_\infty$. A vector-valued function f is called *feebly oscillating* (when $t \to \infty$) if

$$\lim_{\substack{t,s\to\infty \\ t/s\to 1}} |f(t) - f(s)| = 0$$

(cf. [19, Def. 18.3.1], [43, Def. 8.4]).

Example 2.3. Assume that $t|f(t)| \leq M$ for $t \geq \tau$, where $\tau \geq 0$. Then F is feebly oscillating. In fact, $|F(t) - F(s)| \leq \int_s^t r|f(r)|(dr/r) \leq M\log(t/s)$ for $t \geq s \geq \tau$.

THEOREM 2.4. *Assume that f is feebly oscillating and let $f_\infty \in X$. If $C - \lim_{t\to\infty} f(t) = f_\infty$, then $\lim_{t\to\infty} f(t) = f_\infty$.*

Proof. We can suppose that $f_\infty = 0$. Let $\epsilon > 0$. There exist $\delta > 0$, $t_0 > 0$ such that $|f(s) - f(t)| < \epsilon$ whenever $s, t > t_0$, $s \in [t - \delta t, t + \delta t]$. Hence $|f(t) - (1/2\delta t)\int_{t(1-\delta)}^{t(1+\delta)} f(s)ds| = |(1/2\delta t)\int_{t(1-\delta)}^{t(1+\delta)} (f(t) - f(s))ds| \leq \epsilon$ if $t \geq t_0$. Since

$$\lim_{t\to\infty} \frac{1}{2\delta t}\int_{t(1-\delta)}^{t(1+\delta)} f(s)ds = 0,$$

we conclude $\lim_{t\to\infty} f(t) = 0$. □

B. Conditions under which $A-\lim_{t\to\infty} f(t) = f_\infty$ **implies** $C-\lim_{t\to\infty} f(t) = f_\infty$. The following result is a particular case of [19, Thms. 18.3.3, 18.3.2].

THEOREM 2.5. *Let $f_\infty \in X$. Assume that $f \in L^\infty([\tau,\infty); X)$ for some $\tau \geq 0$. If $A - \lim_{t\to\infty} f(t) = f_\infty$ then $C - \lim_{t\to\infty} f(t) = f_\infty$.*

Proof. 1. We first assume that $\tau = 0$. For $\beta > 0$ let $e_\beta(t) = \beta e^{-\beta t}(t > 0)$. Then span $\{e_\beta : \beta > 0\}$ is dense in $L^1[0,\infty)$ (in fact, if $g \in L^\infty[0,\infty)$ such that $0 = \langle e_\beta, g\rangle = \beta\hat{g}(\beta)$ for all $\beta > 0$, then $g = 0$ almost everywhere by uniqueness theorem for Laplace transforms). By hypothesis $\lim_{\alpha\to\infty}\int_0^\infty e^{-s}f(\alpha s)ds = \lim_{\lambda\to 0+}\int_0^\infty e^{-s}f(s/\lambda)ds = \lim_{\lambda\to 0+}\lambda\int_0^\infty e^{-\lambda s}f(s)ds = f_\infty$. Hence

$$\lim_{\alpha\to\infty}\langle e_\beta, f(\alpha\cdot)\rangle = \lim_{\alpha\to\infty}\beta\int_0^\infty e^{-s\beta}f(\alpha s)ds = \lim_{\alpha\to\infty}\int_0^\infty e^{-s}f(\frac{\alpha}{\beta}s)ds = f_\infty = \langle e_\beta, f_\infty\rangle$$

for all $\beta > 0$. It follows that $\lim_{\alpha\to\infty}\langle h, f(\alpha\cdot)\rangle = f_\infty\int_0^\infty h(t)dt$ for all $h \in L^1[0,\infty)$. Letting $h = X_{[0,1]}$ we obtain $\lim_{\alpha\to\infty} 1/\alpha\int_0^\alpha f(s)ds = \lim_{\alpha\to\infty}\int_0^1 f(\alpha s)ds = \lim_{\alpha\to\infty}\langle h, f(\alpha\cdot)\rangle = f_\infty$.

2. If $\tau > 0$ the result follows by applying 1. to $g(t) = f(t + \tau)$. □

Another result of this type involves an order condition. We assume in the following theorem that X is an ordered Banach space with normal cone X_+ (i.e., X_+ is a closed convex cone such that $X_+ \cap (-X_+) = \{0\}$ and $X'_+ - X'_+ = X'$ where X'_+ denotes the dual cone; see [5] for details). For example, X may be a Banach lattice.

THEOREM 2.6. *Assume that $f(t) \geq 0$ (i.e., $f(t) \in X_+$) for $t \geq 0$. If $A - \lim_{t\to\infty} f(t) = f_\infty$, then $C - \lim_{t\to\infty} f(t) = f_\infty$.*

Karamata's proof of this result (see [43, Thm. 8.5.3]) goes through in the vector-valued case described above. A very short and elegant proof in the scalar case is given by König [24].

C. Conditions under which $\lim_{\lambda\to 0+}\hat{f}(\lambda) = F_\infty$ **implies** $\lim_{t\to\infty} F(t) = F_\infty$. The following theorem is due to Hardy and Littlewood in the numerical case; see [43, Thm. 8.4.3].

THEOREM 2.7. *Let $F_\infty \in X$. Assume that for some $\tau \geq 0$*

$$M = \sup_{t\geq\tau} t|f(t)| < \infty. \tag{2.1}$$

If $\lim_{\lambda\to 0+}\hat{f}(\lambda) = F_\infty$, *then* $\lim_{t\to\infty} F(t) = F_\infty$.

Proof. 1. We first assume that $\tau = 0$. For $t > 0$ we have

$$\begin{aligned}|F(t) - \hat{f}(\frac{1}{t})| &= |\int_0^t f(s)[1 - e^{-s/t}]ds - \int_t^\infty f(s)e^{-s/t}ds| \\ &\leq M(\int_0^t [1 - e^{-s/t}]/s^{-1}ds + \int_t^\infty s^{-1}e^{-s/t}ds) \\ &\leq M[\sup_{0<s<t} t[1 - e^{-s/t}]/s + \int_1^\infty e^{-r}r^{-1}dr] \\ &\leq M[\sup_{0<x\leq 1}(1 - e^{-x})/x + \int_1^\infty e^{-r}r^{-1}dr] < \infty.\end{aligned}$$

Since $\lim_{t\to\infty}\hat{f}(1/t) = F_\infty$, it follows that F is bounded. But $A - \lim_{t\to\infty} F(t) = \lim_{\lambda\to 0+}\hat{f}(\lambda) = F_\infty$. So it follows from Theorem 2.5 that $C - \lim_{t\to\infty} F(t) = F_\infty$. The function F is slowly oscillating (see Example 2.3). Hence $\lim_{t\to\infty} F(t) = F_\infty$ by Theorem 2.4.

2. If $\tau > 0$ the result follows from 1. by considering $f(t+\tau)$ instead of $f(t)$. □

D. Conditions under which $A - \lim_{t\to\infty} f(t) = f_\infty$ implies $\lim_{t\to\infty} f(t) = f_\infty$. Since $\lambda\hat{F}(\lambda) = \hat{f}(\lambda)$, any Tauberian theorem of type D yields one of type C. Conversely, if $f \in C^1([\tau,\infty), X)$ we can apply a result of type C to the function $f'(t+\tau)$ and obtain a Tauberian theorem of type D.

Following an idea of Batty [3] we apply instead Tauberian theorems of type C to the function f_δ defined by

$$f_\delta(t) = (f(t+\delta) - f(t))/\delta \qquad (t \geq 0) \tag{2.2}$$

for some $\delta > 0$. The following implications hold.

LEMMA 2.8. *Let* $f_\infty \in X$, $\delta > 0$.

$$\begin{array}{ll}\text{(i)} & \lim_{t\to\infty} f(t) = f_\infty \\ & \Downarrow \\ \text{(ii)} & \lim_{t\to\infty}\int_t^{t+\delta} f(s)ds = f_\infty\delta \\ & \Updownarrow \\ \text{(iii)} & \lim_{t\to\infty}\int_0^t f_\delta(s)ds = f_\infty - \frac{1}{\delta}\int_0^\delta f(s)ds \\ & \Downarrow \\ \text{(iv)} & \lim_{\lambda\to 0+}\hat{f}_\delta(\lambda) = f_\infty - \frac{1}{\delta}\int_0^\delta f(s)ds \\ & \Updownarrow \\ \text{(v)} & A - \lim_{t\to\infty} f(t) = f_\infty.\end{array}$$

Proof. Since $|(1/\delta)\int_t^{t+\delta} f(s)ds - f_\infty| = |(1/\delta)\int_t^{t+\delta}(f(s) - f_\infty)ds| \leq \sup_{s\geq t}|f(s) - f_\infty|$, (i) implies (ii), and (ii) is equivalent to (iii) since $\int_0^t f_\delta(s)ds = (1/\delta)\int_t^{t+\delta} f(s)ds - (1/\delta)\int_0^\delta f(s)ds$.

By Theorem 2.2(c), (iii) implies (iv). Since

$$\hat{f}_\delta(\lambda) = \frac{1}{\delta\lambda}(e^{\lambda\delta} - 1)\lambda\hat{f}(\lambda) - \frac{e^{\lambda\delta}}{\delta}\int_0^\delta e^{-\lambda s}f(s)ds, \tag{2.3}$$

(iv) is equivalent to (v). □

Let $f_\infty \in X$. We say, f is *B-convergent* to f_∞, or simply write $B - \lim_{t\to\infty} f(t) = f_\infty$, if (ii) of Lemma 2.8 holds for all $\delta > 0$. A vector-valued function f is called *slowly oscillating* (when $t \to \infty$) if

$$\lim_{\substack{t,s\to\infty \\ t-s\to 0}} |f(t) - f(s)| = 0.$$

PROPOSITION 2.9. *Let $f_\infty \in X$.*

(a) *If $B - \lim_{t\to\infty} f(t) = f_\infty$, then $A - \lim_{t\to\infty} f(t) = f_\infty$.*

(b) *If $\lim_{t\to\infty} f(t) = f_\infty$, then $B - \lim_{t\to\infty} f(t) = f_\infty$.*

(c) *If f is slowly oscillating then $B - \lim_{t\to\infty} f(t) = f_\infty$ implies $\lim_{t\to\infty} f(t) = f_\infty$.*

Proof. (a) and (b) follow from Lemma 2.8. Assume that $B - \lim_{t\to\infty} f(t) = f_\infty$. Then

$$\begin{aligned}\overline{\lim_{t\to\infty}}|f(t) - f_\infty| &\le \overline{\lim_{t\to\infty}}|f(t) - \frac{1}{\delta}\int_t^{t+\delta} f(s)ds| \\ &= \overline{\lim_{t\to\infty}}|\frac{1}{\delta}\int_t^{t+\delta}(f(t) - f(s))ds| \le \overline{\lim_{t\to\infty}} \sup_{t\le s\le t+\delta}|f(t) - f(s)|.\end{aligned}$$

Hence if f is slowly oscillating, we obtain $\overline{\lim}_{t\to\infty}|f(t) - f_\infty| = 0$ by letting $\delta \downarrow 0$. □

Every feebly oscillating function is slowly oscillating (this is obvious from the definitions); moreover, f is slowly oscillating whenever there exists $\tau \ge 0$ such that $f = g + h$, where $g \in UC([\tau,\infty); X)$ (the space of all uniformly continuous functions on $[\tau,\infty)$ with values in X), and $h \in L^\infty([\tau,\infty); X)$ converges to zero as $t \to \infty$.

Remark. In order that $B - \lim_{t\to\infty} f(t) = f_\infty$ it suffices that (ii) holds for all $\delta \in (0,\delta_0)$ for some $\delta_0 > 0$. In fact, if (ii) holds for $\delta > 0$ and $\eta > 0$ it does so for $\delta + \eta$.

Now we are able to deduce from Theorem 2.7 the following Tauberian theorem of type D.

THEOREM 2.10. *Let $f_\infty \in X$. Assume that for some $\delta > 0$*

$$\overline{\lim_{t\to\infty}} \sup_{t\le s\le t+\delta} t|f(t) - f(s)| < \infty. \tag{2.4}$$

If $A - \lim_{t\to\infty} f(t) = f_\infty$ then $\lim_{t\to\infty} f(t) = f_\infty$.

Proof. Assumption (2.4) implies that f_δ satisfies (2.1) for $\delta > 0$ small enough and also that f is slowly oscillating. Since $A - \lim_{t\to\infty} f(t) = f_\infty$, it follows that (iv) of Lemma 2.8 is satisfied. We conclude (iii) from Theorem 2.7 so that $B - \lim_{t\to\infty} f(t) = f_\infty$. Hence $\lim_{t\to\infty} f(t) = f_\infty$ by Proposition 2.9. □

Note that (2.4) is satisfied whenever $f \in C^1([\tau,\infty), X)$ and $|tf'(t)| \le M$ for $t \ge \tau$; in fact, $|f(t) - f(s)| = |\int_t^s f'(r)dr| \le M\log(s/t) \le M(s-t)/t$ for $t < s$.

Applying Theorem 2.10 to F we obtain an improvement of Theorem 2.7.

COROLLARY 2.11. *Let $F_\infty \in X$. Assume that for some $\delta > \infty$*

$$\overline{\lim_{t\to\infty}} \int_t^{t+\delta} r|f(r)|dr < \infty. \tag{2.5}$$

If $\lim_{\lambda\to 0+} \hat{f}(\lambda) = F_\infty$, *then* $\lim_{t\to\infty} F(t) = F_\infty$.

Proof. We have

$$\begin{aligned}\overline{\lim_{t\to\infty}} \sup_{t\le s\le t+\delta} t|F(t) - F(s)| &\le \overline{\lim_{t\to\infty}} t \int_t^{t+\delta} |f(s)|ds \\ &\le \overline{\lim_{t\to\infty}} \int_t^{t+\delta} s|f(s)|ds < \infty.\end{aligned}$$

Hence F satisfies (2.4) and the conclusion follows from Theorem 2.10. □

E. Power series. Let $p(z) = \sum_{n=0}^\infty a_n z^n$ be a power series, where $a_n \in X$, which converges for $|z| < 1$. Defining $f \in L^1_{\mathrm{loc}}([0,\infty); X)$ by

$$f(t) = a_n \quad \text{if } t \in [n, n+1) \tag{2.6}$$

the preceding results yield Tauberian theorems for p. In fact,

$$\hat{f}(\lambda) = \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{n=0}^\infty a_n e^{-\lambda n} \qquad (\mathrm{Re}\,\lambda > 0). \tag{2.7}$$

From Theorem 2.7 Hardy's theorem can be obtained.

THEOREM 2.12. *Assume that* $\sup\{n|a_n| : n \in \mathbb{N}_0\} < \infty$, *and let* $b_\infty \in X$. *If* $\lim_{z\uparrow 1} p(z) = b_\infty$, *then* $\sum_{n=0}^\infty a_n = b_\infty$.

The special case when $\lim_{n\to\infty} na_n = 0$ had been proven by Tauber [38] (in the scalar case) and was the starting point of Tauberian theory.

In the case of power series, theorems of type C and D are equivalent. In fact, let $b_n = \sum_{k=0}^n a_k$, or equivalently, $a_0 = b_0$, $a_n = b_n - b_{n-1}$ $(n = 1, 2\cdots)$. Then $q(z) = \sum_{n=0}^\infty b_n z^n$ has the same radius of convergence as $p(z)$. The formula for the Cauchy product yields

$$\frac{1}{z-1} \sum_{k=0}^\infty a_k z^k = \sum_{k=0}^\infty z^k \cdot \sum_{k=0}^\infty a_k z^k = \sum_{k=0}^\infty b_k z^k \qquad (|z| < 1);$$

that is,

$$\sum_{k=0}^\infty a_k z^k = (1-z) \sum_{k=0}^\infty b_k z^k \qquad (|z| < 1).$$

Thus $A - \lim_{n\to\infty} b_n := \lim_{z\uparrow 1} (1-z) \sum_{k=0}^\infty b_k z^k = \lim_{z\uparrow 1} \sum_{k=0}^\infty a_k z^k$ whenever one of the limits exists. So we obtain the following.

COROLLARY 2.13. *Let $b_n \in X$ be such that* $\sup\{n|b_n - b_{n-1}| : n \in \mathbb{N}\} < \infty$. *If* $A - \lim_{n\to\infty} b_n = b_\infty$, *then* $\lim_{n\to\infty} b_n = b_\infty$.

3. Complex Tauberian theorems. We assume throughout this section that $f \in L^1_{\mathrm{loc}}([0,\infty); X)$ is such that

$$\hat{f}(\lambda) = \lim_{b\to\infty} \int_0^b e^{-\lambda t} f(t)dt =: \int_0^\infty e^{-\lambda t} f(t)dt$$

exists for Re $\lambda > 0$. In this section we consider conditions on $\hat{f}$ (rather than on f) in order to establish Tauberian theorems. The following theorem (of type C, see §2) is a variant of [2, Thm. 4.1].

THEOREM 3.1. *Let $f_\infty \in X$. Assume that $f \in L^\infty([\tau,\infty);X)$ for some $\tau \geq 0$ and that $(\hat{f}(\lambda) - F_\infty)/\lambda$ has a continuous extension to $\bar{\mathbb{C}}_+\backslash iE$, where $E \subset \mathbb{R}$ is a closed null set and $0 \notin E$. If for all $R > 0$*

$$M(R) := \sup_{\substack{\eta\in E \\ |\eta|\leq R}} \sup_{t\geq 0} \left| \int_0^t \exp(-i\eta s) f(s) ds \right| < \infty, \tag{3.1}$$

then $\lim_{t\to\infty} F(t) = F_\infty$.

Here and in the sequel we let $F(t) = \int_0^t f(s)ds$, $\mathbb{C}_+ = \{\lambda \in \mathbb{C} : \text{Re } \lambda > 0\}$, and $\bar{\mathbb{C}}_+$ the closure of $\mathbb{C}_+$. Note that the hypothesis implies that $\hat{f}(\lambda)$ has a continuous extension to $\bar{\mathbb{C}}_+\backslash iE$ and that $\hat{f}(0) = F_\infty$. In particular, $A - \lim_{t\to\infty} F(t) = F_\infty$.

The proof of [2, Thm. 4.1] works for Theorem 3.1 as well if the basic estimate Lemma 5.2 which will be proved in §5 is used instead of [2, Lemma 3.1].

For $E = \emptyset$ Theorem 3.1 is a version of a theorem due to Ingham [20]. A very short and elegant proof based on an ingenious contour argument due to Newman [30] is given by Korevaar [25]. In [2] the technique of Newman and Korevaar has been extended in order to treat singularities in $i\mathbb{R}$. Whereas in [2] it is assumed that $\hat{f}$ has a holomorphic extension to $\bar{\mathbb{C}}_+\backslash iE$, our slightly more general version is more natural in view of the applications to Volterra equations we have in mind (see §5).

We give several comments on Theorem 3.1, starting with the case when $E = \emptyset$.

Remark 3.2. *Quantitative estimates.* Korevaar's argument actually yields the following more precise result. Assume that $f \in L^\infty([\tau,\infty);X)$ for some $\tau > 0$ and that $F_\infty \in X$ such that $(\hat{f}(\lambda) - F_\infty)/\lambda$ has a continuous extension to $\mathbb{C}_+ \cup i[-R,R]$ where $R > 0$. Then

$$\overline{\lim_{t\to\infty}} |F(t) - F_\infty| \leq \frac{2}{R} \overline{\lim_{t\to\infty}} |f(t)|. \tag{3.2}$$

Proof. In fact, Korevaar shows (a special case of Lemma 5.2 below)

$$\overline{\lim_{t\to\infty}} |F(t) - F_\infty| \leq \frac{2}{R} \sup_{t\geq 0} |f(t)|. \tag{3.3}$$

Applying this to $g(t) = f(t+s)$ with $s \geq \tau$, we have $\hat{g}(\lambda) = e^{\lambda s}[\hat{f}(\lambda) - \int_0^s e^{-\lambda r} f(r)dr]$ (Re $\lambda > 0$) so that $(\hat{g}(\lambda) - G_\infty)/\lambda$ has a continuous extension to $\mathbb{C}_+ \cup i[-R,R]$ with $G_\infty = F_\infty - \int_0^s f(r)dr$. Hence, by (3.3)

$$\overline{\lim_{t\to\infty}} |F(t) - F_\infty| = \overline{\lim_{t\to\infty}} \left| \int_0^{t+s} f(r)dr - F_\infty \right| = \overline{\lim_{t\to\infty}} |G(t) - G_\infty| \leq \sup_{t\geq s} \frac{2}{R} |f(t)|.$$

Letting $s \to \infty$ yields (3.2). □

Quantitative estimates in the case $E \neq \emptyset$ are given in Batty [3].

Remark 3.3. *Convergence of the Laplace integral at regular points.* Assume that $\lim_{t\to\infty} |f(t)| = 0$. If $\hat{f}$ has a holomorphic extension to $\mathbb{C}_+ \cup U$, where U is a neighborhood of $i\eta \in i\mathbb{R}$, then

$$\hat{f}(i\eta) = \int_0^\infty e^{-i\eta s} f(s)ds = \lim_{t\to\infty} \int_0^t e^{-i\eta s} f(s)ds. \tag{3.4}$$

To see this, it suffices to replace $f(t)$ in (3.2) by $e^{-i\eta t}f(t)$.

Remark 3.4. *Riesz's theorem on power series* [40, Thm. 7.3]. Let $a_n \in X$ be such that $\lim_{n\to\infty}|a_n| = 0$ and let $p(z) = \sum_{n=0}^{\infty} a_n z^n$ ($|z| < 1$). If p has a holomorphic extension to $D \cup U$ ($D = \{z \in \mathbb{C} : |z| < 1\}$) where U is an open neighborhood of $z_0 \in \Gamma := \{z \in \mathbb{C} : |z| = 1\}$, then $p(z_0) = \lim_{N\to\infty}\sum_{n=0}^{N} a_n z_0^n$. This is obtained by applying (3.4) to the function f defined by (2.6).

Next we establish a complex Tauberian theorem of type D. The following is a variant of [3, Cor. 2.6].

THEOREM 3.5. *Assume that f is slowly oscillating, let $f_\infty \in X$, and suppose that $\hat{f}(\lambda) - (f_\infty/\lambda)$ has a continuous extension to $\bar{\mathbb{C}}_+\backslash iE$, where $E \subset \mathbb{R}$ is a closed null set such that $0 \notin E$. If for all $R > 0$,*

$$M(R) := \sup_{\eta\in E\cap[-R,R]}\sup_{t\geq 0}\left|\int_0^t \exp(-i\eta s)f(s)ds\right| < \infty, \tag{3.5}$$

then $\lim_{t\to\infty} f(t) = f_\infty$.

Remark. The assumption implies that $A - \lim_{t\to\infty} f(t) = f_\infty$.

Proof. The function f_δ defined by (2.2) is eventually bounded for $\delta > 0$ sufficiently small. Let $c := f_\infty - (1/\delta)\int_0^\delta f(s)ds$. Then by (2.3),

$$\begin{aligned}(\hat{f}_\delta(\lambda) - c)/\lambda &= \frac{1}{\delta\lambda}(e^{\lambda\delta} - 1)\hat{f}(\lambda) - \frac{e^{\lambda\delta}}{\lambda\delta}\int_0^\delta e^{\lambda s}f(s)ds - \frac{c}{\lambda}\\ &= \frac{1}{\delta\lambda}(e^{\lambda\delta} - 1)(\hat{f}(\lambda) - \frac{f_\infty}{\lambda}) + \left[\frac{1}{\delta\lambda}(e^{\lambda\delta} - 1) - 1\right]f_\infty/\lambda\\ &\qquad - \frac{e^{\lambda\delta}}{\delta}\int_0^\delta \frac{e^{-\lambda s} - 1}{\lambda}f(s)ds - \frac{e^{\lambda\delta} - 1}{\lambda\delta}\int_0^\delta f(s)ds.\end{aligned}$$

Since the functions $(1/\delta\lambda)(e^{\lambda\delta} - 1)$, $[(1/\delta\lambda)(e^{\lambda\delta} - 1) - 1]/\lambda$, and $(e^{-\lambda s} - 1)/\lambda$ are entire, it follows that $(\hat{f}_\delta(\lambda) - c)/\lambda$ has a continuous extension to $\bar{\mathbb{C}}_+\backslash iE$. Moreover,

$$\begin{aligned}\left|\int_0^t \exp(-i\eta s)f_\delta(s)ds\right| &= \left|\delta^{-1}\int_0^t \exp(-i\eta s)(f(s+\delta) - f(s)ds\right|\\ &= \delta^{-1}\left|\int_\delta^{\delta+t}\exp i\eta(\delta - s)f(s)ds - \int_0^t \exp(-i\eta s)f(s)ds\right|\\ &\leq 3\delta^{-1}M(R) \quad \text{for all } \eta \in E\cap[-R,R].\end{aligned}$$

It follows from Theorem 3.1 that $\lim_{t\to\infty}\int_0^t f_\delta(s)ds = c$. Hence $B-\lim_{t\to\infty} f(t) = f_\infty$ by Lemma 2.8. It follows from Proposition 2.9(c) that $\lim_{t\to\infty} f(t) = f_\infty$. □

Remark 3.6. (a) If in Theorem 3.5, instead of f slowly oscillating, we merely assume that $f_\delta \in L^\infty([\tau,\infty);X)$ for all $\delta > 0$, then we obtain $B - \lim_{t\to\infty} f(t) = f_\infty$

(b) However, if f is not slowly oscillating, then f does not converge in general, even if f is bounded. An example is the function $f(t) = T(t)y$ from [2, proof of Ex. 2.5]. The function f is bounded and $\hat{f}$ has a holomorphic extension to $\bar{\mathbb{C}}_+$. However, $f(t)$ does not converge for $t \to \infty$.

Next we consider the case where $0 \in E$. For simplicity we assume $f_\infty = 0$. We let $\mathrm{Lip}([\tau,\infty),X) = \{f : [\tau,\infty) \to X : f$ is Lipschitz continuous $\}$.

THEOREM 3.7. *Assume that* $f \in \mathrm{Lip}([\tau,\infty);X)$, *for some* $\tau \geq 0$. *Suppose that* $\hat{f}(\lambda)$ *has a continuous extension to* $\bar{\mathbb{C}}_+\backslash iE$, *where* E *is a closed null set,* $0 \in E$, *and that for each* $R \geq 0$

$$M(R) := \sup_{\eta\in E\cap[-R,R]} \sup_{t\geq 0} \left| \int_0^t \exp(-i\eta s) f(s)ds \right| < \infty. \tag{3.6}$$

Then $\lim_{t\to\infty} f(t) = 0$.

Remark. Since $0 \in E$, condition (3.6) implies that $C - \lim_{t\to\infty} f(t) = 0$.

Proof. We first show that $f \in L^\infty([\tau,\infty);X)$. There exists $L \geq 0$ such that $|f(t) - f(s)| \leq L|t-s|$ for all $s,t \geq \tau$. Let $\varphi \in X'$, $|\varphi| \leq 1$. Then, by the Taylor expansion for $F(t) = \int_0^t f(r)dr$ in $s \geq \tau$, we have

$$\langle F(s+1),\varphi\rangle = \langle F(s),\varphi\rangle + \langle f(s),\varphi\rangle + \int_s^{s+1} (s+1-r)\frac{d}{dr}\langle f(r),\varphi\rangle dr.$$

Hence

$$\begin{aligned}|\langle f(s),\varphi\rangle| &\leq |\langle F(s+1),\varphi\rangle| + |\langle F(s),\varphi\rangle| + \int_s^{s+1} (s+1-r)|\frac{d}{dr}\langle f(r),\varphi\rangle|dr \\ &\leq 2M(0) + L\int_s^{s+1}(s+1-r)dr \leq 2M(0) + \frac{L}{2}.\end{aligned}$$

Fix $\mu \in \mathbb{R}\backslash E$ and define $g(t) = e^{i\mu t}f(t)$. Then $g \in \mathrm{Lip}([\tau,\infty);X)$ and $\hat{g}(\lambda) = \hat{f}(\lambda - i\mu)(\mathrm{Re}\ \lambda > 0)$. Hence $\hat{g}(\lambda)$ has a continuous extension to $\bar{\mathbb{C}}_+\backslash iE'$ where $E' = E + \mu$. Moreover, for $\eta' = \eta + \mu \in E' \cap [-R,R]$ we have $|\int_0^t \exp(-i\eta' s)g(s)ds| = |\int_0^t \exp(-i\eta s)f(s)ds| \leq M(R + |\mu|)$ for all $t \geq 0$. Since $0 \notin E'$, the assertion follows from Theorem 3.5. □

Applying Theorem 3.5 to power series we obtain a variant of a result due to Allan, O'Farrell, and Ransford [1]. We let $\bar{D} = \{z \in \mathbb{C} : |z| \leq 1\}$, and $\Gamma = \{z \in \mathbb{C} : |z| = 1\}$.

THEOREM 3.8. *Let* $b_n \in X$ *be such that* $\sup\{|b_n| : n \in \mathbb{N}_0\} < \infty$ *and set* $p(z) = \sum_{n=0}^\infty b_n z^n$ *for* $|z| < 1$. *Assume that* p *has a continuous extension to* $\bar{D}\backslash F$, *where* $F \subset \Gamma$ *is a closed null set.*

If $\sup_{z\in F}\sup_{N\in\mathbb{N}} |\sum_{n=0}^N b_n z^n| < \infty$, *then* $\lim_{n\to\infty} b_n = 0$.

Remark. The hypothesis of the theorem directly implies $A - \lim_{n\to\infty} b_n = 0$ if $1 \notin F$ and $C - \lim_{n\to\infty} b_n = \lim_{n\to\infty} 1/n \sum_{k=0}^{n-1} b_k = 0$ if $1 \in F$.

Proof. Replacing b_n by $b_n w^{-n}$ for some $w \in \Gamma\backslash F$ if necessary, we may assume that $1 \notin F$. Let $f(t) = b_n$ for $t \in [n, n+1)$. Then $\hat{f}(\lambda) = [(1-e^{-\lambda})/\lambda]\sum_{n=0}^\infty b_n e^{-\lambda n}$ has a continuous extension to $\mathbb{C}_+\backslash iE$ where $E = \{\eta \in \mathbb{R} : e^{-i\eta} \in F\}$. Moreover, for $t \in [n, n+1)$ we have $\int_0^t \exp(-i\eta s)f(s)ds = \sum_{m=0}^n b_m \exp(-i\eta m)(1-\exp(-i\eta))/i\eta + b_n\exp(-i\eta n)(1-\exp(-i\eta(t-n)))/i\eta$, so that (3.5) is satisfied (since $0 \notin E$). It follows from Theorem 3.5 and Remark 3.6 that $\lim_{n\to\infty} b_n = \lim_{n\to\infty}\int_n^{n+1} f(s)ds = B - \lim_{t\to\infty} f(t) = 0$. □

It is implied by Riesz's theorem (Remark 3.4) that in the situation of Theorem 3.8 we have $p(z) = \sum_{n=0}^\infty b_n z^n$ for all regular $z \in \Gamma$ (and this is precisely what is shown in [1], assuming that p has a holomorphic extension to $\bar{D}\backslash F$).

An immediate consequence of Theorem 3.8 is the Katznelson–Tzafriri theorem (which actually was the motivation of the work by Allan, O'Farrell, and Ransford [1]).

THEOREM 3.9. (*Katznelson–Tzafrifi* [22].) *Let* $T \in \mathcal{L}(X)$ *such that* $\sup_{n\geq 0} |T^n| < \infty$ *and* $\sigma(T) \cap \Gamma \subset \{1\}$. *Then* $\lim_{n\to\infty} |(T-I)T^n| = 0$.

Proof. Let $p(z) = \sum_{n=0}^{\infty}(T-I)T^n z^n = (T-I)(I-zT)^{-1}, |z| < 1$. Since $\sup_{n\geq 0} |\sum_{n=0}^{N}(I-T)T^n| = \sup_{N\geq 0}|I - T^{N+1}| < \infty$, the hypotheses of Theorem 3.8 are satisfied for $b_n = (I-T)T^n$ and $F = \{1\}$. □

We are going to prove a continuous version of the Katznelson–Tzafriri theorem. Formally, it is expected that $(T^n)_{n\geq 0}$ has to be replaced by $(T(t))_{t\geq 0}$ and $T - I = (T(1) - I)/1$ by A, the generator of $(T(t))_{t\geq 0}$. We make this more precise. A C_0-semigroup $(T(t))_{t\geq 0}$ on X is called *eventually differentiable* if there exists $\tau > 0$ such that $T(\tau)X \subset D(A)$. Note that then $T(t)X \subset D(A)$ and $AT(t) \in \mathcal{L}(X)$ for all $t \geq \tau$.

THEOREM 3.10. *Let* $(T(t))_{t\geq 0}$ *be a bounded, eventually differentiable semigroup with generator* A. *The following are equivalent.*

(i) $\lim_{t\to\infty} |AT(t)| = 0$;

(ii) $\sigma(A) \cap i\mathbb{R} \subset \{0\}$.

Proof. Let $M = \sup_{t\geq 0} |T(t)|$. Assume that (ii) holds and that $\tau > 0$ such that $T(\tau)X \subset D(A)$. Then $T(t)X \subset D(A^2)$ for all $t \geq 2\tau$. Let $f : [0,\infty) \to \mathcal{L}(X)$ be given by $F(t) = AT(t + 2\tau)$. Then

$$\begin{aligned} |f(t) - f(s)| &= \left|\int_t^s \frac{d}{dr} f(r)dr\right| = \left|\int_t^s A^2 T(r + 2\tau)dr\right| \\ &= \left|\int_t^s T(r)A^2T(2\tau)dr\right| \leq M|A^2T(2\tau)||s-t|, \qquad s,t \geq 0, \end{aligned}$$

so that f is Lipschitz continuous. Moreover, $\hat{f}(\lambda) = R(\lambda, A)AT(2\tau)$ has a continuous extension to $\overline{\mathbb{C}}_+\backslash\{0\}$. Since $|\int_0^t f(s)ds| = |T(t+2\tau) - T(t)| \leq 2M \quad (t \geq 0)$, it follows from Theorem 3.7 that $\lim_{t\to\infty} |AT(t)| = \lim_{t\to\infty} |f(t)| = 0$.

Conversely, assume that (i) holds. (a) We show that $\lambda e^{t\lambda} \in \sigma(AT(t))$ for all $t \geq \tau$ whenever $\lambda \in \sigma(A) \cap i\mathbb{R}$. In fact, let $\lambda \in \sigma(A) \cap i\mathbb{R}$; then $\lambda \in \sigma_p(A) \cup \sigma_c(A)$ since λ is a boundary point of $\rho(A)$. Hence, there exist $x_n \in D(A)$, $|x_n| = 1$ such that $\lim_{t\to\infty} |(\lambda - A)x_n| = 0$. Consequently,

$$\begin{aligned} (\lambda e^{t\lambda} - AT(t))x_n &= \lambda(e^{\lambda t} - T(t))x_n + T(t)(\lambda - A)x_n \\ &= \lambda e^{\lambda t} \int_0^t e^{-\lambda s} T(s)(\lambda - A)x_n ds + T(t)(\lambda - A)x_n \to 0, \end{aligned}$$

as $n \to \infty$. Thus $\lambda e^{\lambda t} \in \sigma(AT(t))$ for $t \geq \tau$.

(b) Let $\eta \in \mathbb{R}$ be such that $i\eta \in \sigma(A)$. Then by (a), $i\eta e^{i\eta t} \in \sigma(AT(t))$ for $t \geq \tau$. Consequently, $|\eta| = |i\eta e^{i\eta t}| \leq |AT(t)| \to 0$ as $t \to \infty$, i.e., $\eta = 0$. □

As another application of Theorem 3.7 we obtain the following result which in some sense is complementary to Theorem 3.10.

THEOREM 3.11. *Let* $U(t)$ *be* C_0*-group with generator* A *and suppose that* $\sup_{t\geq 0} |U(t)x| < \infty$ *for all* $x \in D_\infty := \cap_{n\geq 0} D(A^n)$. *If* $\sigma(A) \cap i\mathbb{R} \subset \{0\}$, *then* $U(t) = I$ *for all* $t \in \mathbb{R}$.

Proof. Let $x \in D_\infty$ and $f(t) = AU(t)x = U(t)Ax$. Then f is Lipschitz continuous since $|f(t) - f(s)| = |\int_s^t (d/dr)U(r)Axdr| = |\int_s^t U(r)A^2xdr| \leq |t-s| \sup_{r\geq 0} |U(r)A^2x|$, and $|\int_0^t f(s)dr| = |U(t)x - x|$ is bounded for $t \geq 0$. For Re $\lambda > 0$ we have $\hat{f}(\lambda) \in D(A)$ and $(\lambda - A)\hat{f}(\lambda) = Ax$. Hence $\hat{f}(\lambda) = (\lambda - A)^{-1}Ax$ whenever $\lambda \in \varrho(A)$, Re $\lambda > 0$. This shows that $\hat{f}(\lambda)$ has a continuous extension to $\overline{\mathbb{C}}_+\backslash\{0\}$. It follows from Theorem 3.7 that $\lim_{t\to\infty} U(t)Ax = \lim_{t\to\infty} f(t) = 0$.

So far we have shown that $\lim_{t\to\infty} U(t)Ax = 0$ for all $x \in D_\infty$. We will deduce from this that $Ax = 0$ for all $x \in D_\infty$ and hence $A = 0$ since D_∞ is a core. In fact, D_∞ is a Fréchet space for the topology defined by the norms $p_n(x) = |x|+|Ax|+\cdots+|A^n x|$, $n \in \mathbb{N} \cup \{0\}$. We show that there exists $k \in \mathbb{N}$ such that

$$|U(t)x| \leq kp_k(x) \tag{3.7}$$

for all $x \in D_\infty$, $t \in \mathbb{R}$. If this is false, there exist $x_m \in D_\infty$, $t_m \in \mathbb{R}$ such that $p_m(x_m) = 1$ and $|U(t_m)x_m| \geq m$, $m \in \mathbb{N}$. Let $Y_k = \{x \in D_\infty : |U(t)x| \leq kp_k(x)$ for all $t \in \mathbb{R}\}$. Then Y_k is closed in D_∞ and $\cup_{k\geq 0} Y_k = D_\infty$. So by Baire's theorem there exists $k \in \mathbb{N}$ such that Y_k has a nonempty interior; i.e., we find $a \in D_\infty$, $\epsilon > 0$, $\ell \geq k$ such that $p_\ell(a-x) \leq \epsilon$ implies $|U(t)x| \leq kp_k(x)$ for all $t \in \mathbb{R}$. Consequently,

$$\begin{aligned} m\cdot \epsilon/p_\ell(x_m) - |U(t_m)a| &\leq \epsilon/p_\ell(x_m)|U(t_m)x_m| - |U(t_m)a| \leq |U(t_m)(a-\epsilon/p_\ell(x_m)x_m| \\ &\leq kp_k(a-(\epsilon/p_\ell(x_m))x_m) \leq k(p_k(a)+\epsilon p_k(x_m)/p_\ell(x_m)) \\ &\leq k(p_k(a)+\epsilon) \quad \text{since } \ell \geq k, \end{aligned}$$

hence,

$$\begin{aligned} \epsilon m &\leq p_\ell(x_m)[k(p_k(a)+\epsilon)+|U(t_m)a|] \leq p_m(x_m)[k(p_k(a)+\epsilon)+|U(t_m)a|] \\ &= [k(p_k(a)+\epsilon)+|U(t_m)a|] \quad \text{for all } m \geq \ell. \end{aligned}$$

But $(U(t_m)a)_{m\geq 0}$ is bounded in X, a contradiction. So (3.7) is proved.

Let $x \in D_\infty$. Then by (3.7)

$$\begin{aligned} |Ax| &= |U(-t)U(t)Ax| \leq kp_k(U(t)Ax) \\ &= k\{|U(t)Ax| + |U(t)A^2x| + \cdots + |U(t)A^{k+1}x|\} \to 0 \quad \text{as } t \to \infty. \end{aligned}$$

Hence $Ax = 0$ for all $x \in D_\infty$. □

4. Real ergodic theorems for Volterra equations. Throughout the remainder of the paper, we make the assumptions of the Introduction. In particular, $(S(t))_{t\geq 0}$ denotes the resolvent governing (1.1). Recall that we assume

$$\sup_{t\geq 0} |e^{-\lambda t}S(t)| < \infty$$

for all $\lambda > 0$. By $\hat{S}(\lambda) = \int_0^\infty e^{-\lambda t}S(t)dt$, Re $\lambda > 0$, we denote the Laplace transform of $S(t)$. In addition, we assume that the (complex-valued) kernel $a \in L^1_{\text{loc}}(\mathbb{R}_+)$ is Laplace transformable, i.e., there exists $\alpha \geq 0$ such that $\int_0^\infty e^{-\alpha t}|a(t)|dt < \infty$. We let $\hat{a}(\lambda) = \int_0^\infty e^{-\lambda t}a(t)dt$ (Re $\lambda \geq \alpha$). If (1.3) holds, then the closedness of A implies

$$\int_0^t a(t-s)S(s)x\,ds \in D(A) \quad \text{and} \quad S(t)x = x + A\int_0^t a(t-s)S(s)x\,ds \tag{4.1}$$

for all $x \in X$. Moreover, due to the assumptions above we have the following proposition.

PROPOSITION 4.1. (a) *$\hat{a}(\lambda)$ has a meromorphic extension to $\mathbb{C}_+$.*

(b) *$\hat{a}(\lambda) \neq 0$ on $\mathbb{C}_+$ if A is unbounded.*

(c) *$\lambda\hat{S}(\lambda) = (I - \hat{a}(\lambda)A)^{-1}$ for all $\lambda \in \mathbb{C}_+$ such that λ is not a pole of $\hat{a}$.*

We refer to [33] for the proof of (4.1) and Proposition 4.1.

COROLLARY 4.2. *If $\hat{a}$ has a pole in $\mathbb{C}_+$, then $R(A)$ is closed and $X = N(A) \oplus R(A)$.*

Remark. Here $R(A) := \{Ax : x \in D(A)\}$ denotes the *range* and $N(A) := \{x \in D(A) : Ax = 0\}$ the *kernel* of A.

Proof. Assume that $\lambda_0 \in \mathbb{C}_+$ is a pole of $\hat{a}$ of order n; then $1/\hat{a}(\lambda)$ maps a neighborhood of λ_0 onto a neighborhood of zero. It follows from Proposition 4.1(c) that there exists $\epsilon > 0$ such that $V := \{z \in \mathbb{C} : 0 < |z| < \epsilon\} \subset \varrho(A)$. Moreover, $|((1/\hat{a}(\lambda)) - A)^{-1}| = |\hat{a}(\lambda)\lambda\hat{S}(\lambda)| \leq$ const $|\hat{a}(\lambda)|$ near λ_0. Hence $|(z - A)^{-1}| \leq$ const$/|z|$ $(z \in V)$. Thus zero is at most a pole of order 1 of $(z - A)^{-1}$. Now the claim follows from [41, Chap. VIII.8]. □

In order to study the asymptotic behavior of the resolvent, we use the following terminology.

DEFINITION 4.3. The resolvent S is called (a) *uniformly* (*strongly, weakly*) *Abel-ergodic* if $\lim_{\lambda \to 0+} \lambda\hat{S}(\lambda) = P$ exists in the uniform (respectively, strong, weak) operator topology;

(b) *uniformly* (*strongly, weakly*) *Cesaro-ergodic* if $\lim_{t \to \infty} 1/t \int_0^t S(s)ds = P$ exists uniformly (respectively, strongly, weakly);

(c) *uniformly* (*strongly, weakly*) *ergodic* if $\lim_{t \to \infty} S(t) = P$ exists uniformly (respectively, strongly, weakly).

Notation. We shall use the abbreviation (i,J)-ergodic where i runs through the symbols u, s, w with obvious meaning, and J runs through A, C, E. Then the following implication scheme holds.

$$\begin{tikzcd}
(u,A) \arrow[d, Rightarrow] & (u,C) \arrow[l, Rightarrow] \arrow[d, Rightarrow] & (u,E) \arrow[l, Rightarrow] \arrow[d, Rightarrow] \\
(s,A) \arrow[d, Rightarrow] & (s,C) \arrow[l, Rightarrow] \arrow[d, Rightarrow] & (s,E) \arrow[l, Rightarrow] \arrow[d, Rightarrow] \\
(w,A) & (w,C) \arrow[l, Rightarrow] & (w,E) \arrow[l, Rightarrow]
\end{tikzcd}$$

Our goal is to characterize (i, J) ergodicity of $S(t)$ in terms of the operator A and the kernel a; or, at least, to find sufficient conditions. We need the following.

PROPOSITION 4.4. *Let B be a densely defined linear operator on X, $\mu_n \in \mathbb{C}$ such that* $\lim_{n \to \infty} |\mu_n| = \infty$, $1/\mu_n \in \varrho(B)$ *and* $\sup_{n \geq 0} |(I - \mu_n B)^{-1}| < \infty$. *Then*

(a) $N(B) \cap \overline{R(B)} = \{0\}$.

(b) *The following are equivalent.*

(i) $\lim_{n \to \infty}(I - \mu_n B)^{-1} = P$ *exists strongly*;

(ii) $\lim_{n \to \infty}(I - \mu_n B)^{-1} = P$ *exists weakly*;

(iii) $N(B) \oplus \overline{R(B)} = X$;

(iv) $N(B)^\perp \cap N(B') = \{0\}$.

If this is the case, then P is the projection onto $N(B)$ along $\overline{R(B)}$.

(c) *If X is reflexive, the equivalent conditions of* (b) *are automatically satisfied.*

(d) *Assume that the equivalent conditions of* (b) *hold. Then the following are equivalent.*

(i) $R(B)$ *is closed*;

(ii) $\lim_{n \to \infty}(I - \mu_n B)^{-1} = P$ *in* $\mathcal{L}(X)$;

(iii) $\lim_{n \to \infty}(I - \mu_n B)^{-2} = P$ *in* $\mathcal{L}(X)$.

This result is well known; we refer to [41, Chap. VIII.4] and [19, Chap. XVIII]. We add the analogous properties of $(I - \mu_n B)^{-1}$ at zero.

PROPOSITION 4.5. *Let B be an operator on X, $0 \neq \mu_n \in \mathbb{C}$ such that $1/\mu_n \in \varrho(B)$, $\lim_{n\to\infty} \mu_n = 0$ and $\sup_{n\geq 0} |(I - \mu_n B)^{-1}| < \infty$.*

(a) *The following are equivalent.*

(i) *$D(B)$ is dense in X;*

(ii) *$\lim_{n\to\infty}(I - \mu_n B)^{-1} = I$ strongly;*

(iii) *$\lim_{n\to\infty}(I - \mu_n B)^{-1} = I$ weakly.*

(b) *The following are equivalent.*

(i) *$D(B) = X$;*

(ii) *$\lim_{n\to\infty}(I - \mu_n B)^{-1} = I$ in $\mathcal{L}(X)$;*

(iii) *$\lim_{n\to\infty}(I - \mu_n B)^{-2} = I$ in $\mathcal{L}(X)$.*

For the proof we refer to [19, Chap. XVIII].

Strong and weak Abel ergodicity of the resolvent $S(T)$ of (1.1) are characterized as follows.

THEOREM 4.6. *The following are equivalent.*

(i) *$S(t)$ is strongly Abel ergodic.*

(ii) *$S(t)$ is weakly Abel ergodic.*

(iii) (a) *$|\lambda\hat{S}(\lambda)|$ is bounded on $(0,1]$;*

(b) *$\lim_{\lambda\to 0+} \hat{a}(\lambda) =: \hat{a}(0)$ exists in $\mathbb{C} \cup \{\infty\}$;*

(c) *$N(A)^\perp \cap N(A') = \{0\}$ if $\hat{a}(0) = \infty$.*

Moreover, if these equivalent conditions are satisfied, then $\lim_{\lambda\to 0} \lambda\hat{S}(\lambda) = (I - \hat{a}(0)A)^{-1}$ in $\mathcal{L}(X)$ if $0 \neq \hat{a}(0) \in \mathbb{C}$, $\lim_{\lambda\to 0+} \lambda\hat{S}(\lambda) = I$ strongly if $\hat{a}(0) = 0$, and $\lim_{\lambda\to 0} \lambda\hat{S}(\lambda) = P$ strongly if $\hat{a}(0) = \infty$, where P denotes the projection onto $N(A)$ along $\overline{R(A)}$. If X is reflexive, then (c) in (iii) can be omitted.

Proof. (ii) $\Rightarrow$ (iii). Assume that $w - \lim_{\lambda\to 0+} \lambda\hat{S}(\lambda)x = w - \lim_{\lambda\to 0+}(I - \hat{a}(\lambda)A)^{-1}x = Px$ for all $x \in X$. Then

$$\sup_{\lambda\in(0,1]} |(I - \hat{a}(\lambda)A)^{-1}| < \infty. \tag{4.2}$$

Choose a sequence $\lambda_n \to 0$ such that $\mu_n = \hat{a}(\lambda_n) \to \mu_\infty \in \mathbb{C} \cup \{\infty\}$. We distinguish three cases.

Case 1. $0 < |\mu_\infty| < \infty$.

Then, by (4.2), $\mu_\infty^{-1} \in \varrho(A)$ and $P = (I - \mu_\infty A)^{-1} = \lim_{n\to\infty}(I - \mu_n A)^{-1}$ in $\mathcal{L}(X)$.

Case 2. $\mu_\infty = 0$.

Then $\lim_{n\to\infty}(I - \mu_n A)^{-1} = I$ strongly by Proposition 4.5.

Case 3. $\mu_\infty = \infty$.

It follows from Proposition 4.4 and 4.5, that $(I - \mu_n A)^{-1} \to P$ strongly, where P is the projection onto $N(A)$ along $\overline{R(A)}$.

Now suppose that there exists another sequence $\lambda_n' \to 0$ such that $\hat{a}(\lambda_n') \to \mu_\infty' \neq \mu_\infty$, $\mu_\infty' \in \mathbb{C} \cup \{\infty\}$. Since $A \neq 0$, the limit operators P and P' are different. But this is impossible since $P = \lim_{\lambda\to 0+} \lambda\hat{S}(\lambda) = P'$. This shows that $\hat{a}(0) := \lim_{\lambda\to 0+} \hat{a}(\lambda)$ exists in $\mathbb{C} \cup \{\infty\}$. We have proved (iii). It follows from Propositions 4.4 and 4.5 that (iii) implies (i). □

From the preceding proof, we also obtain the following characterization of uniform Abel ergodicity.

THEOREM 4.7. *$S(t)$ is uniformly Abel ergodic if and only if the following four conditions hold.*

(a) *$|\lambda\hat{S}(\lambda)|$ is bounded on $(0,1]$;*

(b) *$\lim_{\lambda\to 0+} \hat{a}(\lambda) =: \hat{a}(0)$ exists in $\mathbb{C} \cup \{\infty\}$;*

(c) *if $\hat{a}(0) = \infty$ then $R(A)$ is closed and $X = N(A) \oplus R(A)$;*

(d) $\hat{a}(0) \neq 0$ *if* A *is unbounded.*

COROLLARY 4.8. *Suppose that* $(\lambda_0 - A)^{-1}$ *is compact for some* $\lambda_0 \in \varrho(A)$. *We assume that* $\hat{a}(0) \neq 0$ *if* A *is unbounded. If* $S(t)$ *is* (w, A)*-ergodic, then* $S(t)$ *is* (u, A)*-ergodic.*

Proof. This follows from Theorem 4.7 and Theorem 4.6 since $R(A)$ is closed because of the compactness of $(\lambda_0 - A)^{-1}$. □

It is instructive to classify Abel ergodicity by the limits of $\hat{a}(\lambda)$ as $\lambda \to 0+$. Assume that $\hat{a}(0) = \lim_{\lambda \to 0+} \hat{a}(\lambda) \in \mathbb{C} \cup \{\infty\}$ exists.

Case 1. $\hat{a}(0) = 0$. Then

(a) $S(t)$ is (u, A)-ergodic iff A is bounded; and

(b) $S(t)$ is (s, A)-ergodic iff $(I - \hat{a}(\lambda)A)^{-1}$ is bounded for $\lambda \to 0+$.

The ergodic limit then is $P = I$.

Case 2. $\hat{a}(0) \neq 0, \infty$. Then $S(t)$ is (u, A)-ergodic iff it is (s, A)-ergodic iff $(I - \hat{a}(\lambda)A)^{-1}$ is bounded for $\lambda \to 0+$ iff $\hat{a}(0)^{-1} \in \rho(A)$.

The ergodic limit then is $P = (I - \hat{a}(0)A)^{-1}$.

Case 3. $\hat{a}(0) = \infty$. Then

(a) $S(t)$ is (u, A)-ergodic iff $\overline{\lim}_{\lambda \to 0+} |(I - \hat{a}(\lambda)A)^{-1}| < \infty$, $N(A)^{\perp} \cap N(A') = \{0\}$ and $R(A)$ is closed;

(b) $S(t)$ is (s, A)-ergodic iff $\overline{\lim}_{\lambda \to 0+} |(I - \hat{a}(\lambda)A)^{-1}| < \infty$, and $N(A)^{\perp} \cap N(A') = \{0\}$.

The ergodic limit P is then the projection onto $N(A)$ along $\overline{R(A)}$.

In particular, we obtain the following necessary conditions.

COROLLARY 4.9. *If* $A - \lim_{t \to \infty} S(t) = 0$ *strongly, then* $\lim_{\lambda \to 0+} \hat{a}(\lambda) = \infty$ *and* $0 \notin \sigma_p(A) \cup \sigma_p(A')$.

Proof. For the second assertion observe that $Ax = 0$ implies $S(t)x = x$ $(t \geq 0)$ and so $x = 0$. This shows $N(A) = 0$. Hence $N(A') = N(A)^{\perp} \cap N(A') = \{0\}$. Thus $0 \notin \sigma_p(A) \cup \sigma_p(A')$. □

Next, we consider Cesaro ergodicity.

THEOREM 4.10. (a) *If* $S(t)$ *is bounded and* (w, A)*-ergodic, then* $S(t)$ *is* (s, C)*-ergodic.*

(b) *Suppose that* X *is an ordered Banach space with normal and generating cone. If* $S(t) \geq 0$ $(t \geq 0)$ *and* $S(t)$ *is* (w, A)*-ergodic, then* $S(t)$ *is* (s, C)*-ergodic.*

Proof. This follows from Theorem 2.5 and 2.6. □

Finally, we consider ergodicity of $S(t)$. We say that $S(t)$ is a *bounded analytic resolvent* if there exists a bounded, analytic extension of S to a sector $\Sigma(\theta) = \{z : |\arg z| < \theta\}$ for some $\theta \in (0, \pi/2)$.

Remark 4.11. Equation (1.1) is governed by a bounded analytic resolvent if and only if the following conditions are satisfied for some $\theta \in (0, \pi/2)$.

(a) $\hat{a}$ admits a meromorphic extension to $\Sigma(\theta + \pi/2)$.

(b) $\hat{a}(\lambda) \neq 0$ if A is unbounded and $1/\hat{a}(\lambda) \in \varrho(A)$ for all $\lambda \in \Sigma(\theta + \pi/2)$ with $\hat{a}(\lambda) \neq 0$.

(c) $|(I - \hat{a}(\lambda)A)^{-1}|$ is bounded on $\Sigma(\theta + \pi/2)$.

We refer to [33] for a proof.

In the semigroup case $(a(t) \equiv 1)$ the notion of bounded analytic resolvent coincides with that of a bounded analytic semigroup.

PROPOSITION 4.12. *Assume that* S *is a bounded analytic resolvent. Then there exists* $M \geq 0$ *such that*

$$t|S'(t)| \leq M \quad \textit{for all } t > 0$$

(see [33, Cor. 2.1] for a proof).

THEOREM 4.13. *Assume that $S(t)$ is a bounded analytic resolvent. If $S(t)$ is weakly Abel ergodic, then $S(t)$ is strongly ergodic. Moreover, $S(t)$ is even uniformly ergodic, if in addition $(\lambda_0 - A)^{-1}$ is compact for some $\lambda_0 \in \varrho(A)$.*

Proof. It follows from Theorem 4.6 that $S(t)$ is (s, A)-ergodic, and by Corollary 4.8 that S is (u, A)-ergodic if $(\lambda_0 - A)^{-1}$ is compact for some $\lambda_0 \in \varrho(A)$. Let $f(t) = S(t)$ $(t \geq 0)$. Then $f : (0, \infty) \to \mathcal{L}(X)$ is analytic and bounded, hence $f \in L^\infty([0, \infty); \mathcal{L}(X))$. Moreover, $\overline{\lim}_{t\to\infty} t|f'(t)| < \infty$ (by Proposition 4.12). So the claim follows from Theorem 2.10. □

Example 4.14. Consider the kernel $a(t) = t^{\alpha-1}/\Gamma(\alpha)$ where $\alpha \in (0, 2]$ and assume that (1.1) is well posed. For $\alpha = 1$ this means that A generates a C_0-semigroup, for $\alpha = 2$, that A generates a cosine function. We assume again that $\sup_{t\geq 0} |e^{-\lambda t} S(t)| < \infty$ for all $\lambda > 0$. Since $\hat{a}(\lambda) = \lambda^{-\alpha}$, it follows that $\Sigma(\alpha\frac{\pi}{2}) \subset \varrho(A)$. Moreover, $\lim_{\lambda\to 0+} \hat{a}(\lambda) = \infty$ and $\lambda\hat{S}(\lambda) = (I - \lambda^{-\alpha} A)^{-1} = \lambda^\alpha(\lambda^\alpha - A)^{-1}$. Thus Abel ergodicity is the same for all $\alpha \in (0, 2]$:

(a) $S(t)$ is (s, A)-ergodic iff $\sup_{\mu\in(0,1]} |\mu(\mu - A)^{-1}| < \infty$ and $N(A') \cap N(A)^\perp = \{0\}$.

(b) $S(t)$ is (u, A)-ergodic iff (a) holds and $R(A)$ is closed.

In order to characterize strong ergodicity we assume $\alpha < 2$ and $\Sigma(\theta) \subset \varrho(A)$, $|\mu(\mu - A)^{-1}| \leq M$ on $\Sigma(\theta)$ for some $\theta \in (\alpha, \pi/2, \pi)$. Then (1.1) is governed by a bounded analytic resolvent (Remark 4.11). If $N(A') \cap N(A)^\perp = \{0\}$, it follows from Theorem 4.13 that $\lim_{t\to\infty} S(t) = P$ strongly, where P is the projection onto $N(A)$ along $\overline{R(A)}$.

Finally, we consider Volterra equations on $L^\infty = L^\infty(\Omega, \Sigma, \mu)$, where (Ω, Σ, μ) denotes a positive measure space; this Banach space plays an exceptional role.

THEOREM 4.15. *If $X = L^\infty$, then the well-posedness of* (1.1) *implies that A is bounded.*

Remark. Conversely, if A is bounded, then (1.1) is well posed for every kernel.

Theorem 4.15 is due to Lotz [26] in the case $a(t) = 1$, where A is the generator of a C_0-semigroup (see also [29, A-II.3]); for the special case of contraction semigroups it was obtained independently by Coulhon [11]; and for positive semigroups it is due to Kishimoto and Robinson [23].

The reasons for the phenomenon expressed in Theorem 4.15 are two properties of $X = L^\infty$, namely,

$$\text{(DP)} \qquad x_n \to 0 \quad \text{in } (X, \sigma(X, X')) \quad \text{and} \quad x_n' \to 0 \quad \text{in } (X', \sigma(X', X'')) \\ \text{imply } \langle x_n, x_n' \rangle \to 0$$

and

$$\text{(G)} \qquad x_n' \to 0 \quad \text{in } (X', \sigma(X', X)) \quad \text{implies } x_n' \to 0 \quad \text{in } (X', \sigma(X', X''))$$

(see [36, Chaps. II.9.7 and II.10.4]). The first property is called the *Dunford–Pettis property*; a space satisfying the second is called a *Grothendieck space*. For further details on the background in geometry of Banach spaces, we refer to [26] (see also [27], [12]).

The key of the proof of Theorem 4.5 is the following result due to Lotz [26, Thm. 2].

LEMMA 4.16. *Let X satisfy* (G) *and* (DP). *Suppose $T_n \in \mathcal{L}(X)$ is such that $\lim_{n\to\infty} T_n = 0$ strongly and $\lim_{n\to\infty} T_n' = 0$ strongly. Then $\lim_{n\to\infty} |T_n^2| = 0$.*

Using this lemma we obtain the following general result which contains Theorem 4.15 as a special case.

THEOREM 4.17. *Assume that X satisfies* (G) *and* (DP). *Let B be an operator on X and (μ_n) be a sequence in $\mathbb{C}\backslash\{0\}$ such that $1/\mu_n \in \varrho(B)$, $\sup_{n\geq 0} |(I-\mu_n B)^{-1}| < \infty$ and $\lim_{n\to\infty} |\mu_n| = 0$. If $\overline{D(B)} = X$, then B is bounded.*

Proof. Let $J_n = (1-\mu_n B)^{-1}$. Then $\lim_{n\to\infty} J_n = I$ strongly by Proposition 4.5. Hence, $\sigma(X', X) - \lim_{n\to\infty} J_n' x' = x'$ and so by (G), $\sigma(X', X'') - \lim_{n\to\infty} J_n' x' = x'$ for all $x' \in X'$. It follows from Proposition 4.5 that $\lim_{n\to\infty} J_n' = I$ strongly. Now we deduce from Proposition 4.16 that $\lim_{n\to\infty} |(J_n - I)^2| = 0$ which implies $D(B) = X$ by Proposition 4.5. □

Next we consider ergodicity of (1.1) in L^∞.

THEOREM 4.18. *If $X = L^\infty$ and $S(t)$ is weakly Abel ergodic, then $S(t)$ is uniformly Abel ergodic.*

Remark. Since by our general assumption (1.1) is well posed, A is bounded in the situation of Theorem 4.18 (by Theorem 4.15).

We first show the following.

THEOREM 4.19. *Assume that X satisfies* (G) *and* (DP). *Let B be an operator on X such that $1/\mu_n \in \varrho(B)$ for a sequence $(\mu_n) \subset \mathbb{C}$ such that $\lim_{n\to\infty} |\mu_n| = \infty$. If $\lim_{n\to\infty}(I - \mu_n B)^{-1} = P$ weakly, then $\lim_{n\to\infty}(I - \mu_n B)^{-1} = P$ in $\mathcal{L}(X)$.*

Proof. By Proposition 4.4 we have $X = N(B) \oplus \overline{R(B)}$. We can assume $N(B) = 0$ and $P = 0$. Moreover, since $J_n := (I-\mu_n B)^{-1} \to 0$ strongly, it follows that $J_n' x' \to 0$ for $\sigma(X', X)$ and so by (G) for $\sigma(X', X'')$ for all $x' \in X'$. It follows from Proposition 4.4 that $\lim_{n\to\infty} J_n' = I$ strongly. Thus $\lim_{n\to\infty} |J_n^2| = 0$ by Lemma 4.16. This implies $R(B) = X$ by Proposition 4.4(d). □

Proof of Theorem 4.18. Assume that $S(t)$ is (w, A)-ergodic on L^∞. If $\hat{a}(0) \in \mathbb{C}$, then $S(t)$ is (u, A)-ergodic by Theorem 4.7 (note that A is bounded). If $\hat{a}(0) = \infty$, then S is (u, A)-ergodic by Theorem 4.19. □

Remark. Lotz [26] investigates ergodic properties of discrete semigroups $(T^n)_{n\geq 0}$ where T is a bounded linear operator on L^∞.

5. A general convergence theorem for Volterra equations.

This section contains the main theorem which is based on the complex methods introduced in §3.

We assume throughout that a is a kernel as described in §4, A is a linear closed densely defined operator and that the Volterra equation (1.1) is well posed and governed by the resolvent $S(t)$, which is bounded.

Then we know in particular that $\hat{a}$ has a meromorphic extension to $\mathbb{C}_+$. For later purposes (§§6 and 7) we set

$$\varrho(a) := \{i\mu : \mu \in \mathbb{R}, \hat{a} \text{ has a continuous extension to } \mathbb{C}_+ \cup i[\mu-\epsilon, \mu+\epsilon] \text{ with values in } \mathbb{C} \cup \{\infty\} \text{ for some } \epsilon > 0\} \tag{5.1}$$

and still denote by $\hat{a}$ the continuous extension of $\hat{a}$ to $\mathbb{C}_+ \cup \varrho(a)$.

In this section, though, we assume throughout that

$$\varrho(a) = i\mathbb{R}. \tag{5.2}$$

Moreover, we assume that $S(t)$ is strongly Abel ergodic, and set

$$\lim_{\lambda\to 0+} \lambda \hat{S}(\lambda) = Q \tag{5.3}$$

Remark. Since by assumption $S(t)$ is bounded, this is automatically satisfied if X is reflexive (see Thm. 4.4).

From Theorem 4.6, we know the following. If $\hat{a}(0) \in \mathbb{C}$, then $Q = (I - \hat{a}(0)A)^{-1}$; if $\hat{a}(0) = \infty$, then $X = N(A) \oplus \overline{R(A)}$ and $Q = P$, the projection onto $N(A)$ along $\overline{R(A)}$. The following "resolvent set $\varrho(a, A)$ of (a, A)" plays an important role.

$$\varrho(a, A) := \Big\{ i\eta \in i\mathbb{R} : \text{ there exists } \epsilon > 0 \text{ such that } \tfrac{1}{\lambda}[(1 - \hat{a}(\lambda)A)^{-1} - Q] \quad \text{has a strongly continuous extension to } \mathbb{C}_+ \cup i[\eta - \epsilon, \eta + \epsilon] \Big\}. \tag{5.4}$$

Now we are able to formulate the General Convergence Theorem. It is valid for arbitrary kernels (satisfying (5.2)). In the forthcoming sections it will be shown that, for many interesting classes of kernels, hypotheses (H2) and (H3) are automatically satisfied so that (H1) remains to be verified in order to conclude that $S(t)$ is strongly ergodic. Note that in the reflexive case (H1) reduces to a condition on the spectral behavior of (a, A) on $i\mathbb{R}$: the singular set iE has to be countable and $1/\hat{a}(i\eta) \notin \sigma_p(A')$ whenever $\eta \in E$ such that $\hat{a}(i\eta) \neq 0, \infty$ (by $\sigma_p(A')$ we denote the point spectrum of the adjoint A' of A).

THEOREM 5.1. *Assume* (5.2), (5.3), *and suppose the following three hypotheses are satisfied.*

(H1) *The singular set* $iE := i\mathbb{R} \backslash \varrho(a, A)$ *is countable and* $\mu \in E \backslash \{0\}$, $\hat{a}(i\mu) \neq 0, \infty$ *implies* $R(I - \hat{a}(i\mu)A) = X$; $\mu \in E \backslash \{0\}$, $\hat{a}(i\mu) = \infty$ *implies* $X = N(A) \oplus \overline{R(A)}$.

(H2) *For all* $\mu \in E$ *there exists* $C(\mu) \geq 1$ *such that* $|\int_0^t e^{-i\mu s}(a * S(s) - \hat{a}(i\mu)S(s))Ax\,ds| \leq C(\mu)|x|_A$ *for all* $x \in D(A)$ *if* $\hat{a}(i\mu) \in \mathbb{C}$, *and* $|\int_0^t e^{-i\mu s} S(s)Ax\,ds| \leq C(\mu)|x|_A$ *for all* $x \in D(A)$ *if* $\hat{a}(i\mu) = \infty$.

(H3) *There exist* $\tau \geq 0$, $M \geq 0$ *such that* $|S'(t)x| \leq M|x|_A$ $(x \in D(A),\ t \geq \tau)$, *and* $|S(t)| \leq M$ $(t \geq 0)$.

Then $\lim_{t\to\infty} S(t)x = Qx$ *for all* $x \in X$, *where* $Q = (I - \hat{a}(0)A)^{-1}$ *if* $\hat{a}(0) \in \mathbb{C}$, *and* Q *is the projection onto* $N(A)$ *along* $\overline{R(A)}$ *if* $\hat{a}(0) = \infty$.

We start with the following estimate which is a variant of [2, Lemma 3.1].

LEMMA 5.2. *Let* $f : [0, \infty) \to X$ *be measurable,* $|f(t)| \leq M_0$ $(t \geq 0)$. *Let* $R > 0$. *Assume that* $\hat{f}(\lambda)/\lambda$ (*which is defined for* Re $\lambda > 0$) *has a continuous extension to* $\mathbb{C}_+ \cup i([-R, R] \backslash \bigcup_{j=1}^n (\xi_j - \epsilon_j, \xi_j + \epsilon_j))$ *where* $\xi_j \in \mathbb{R}$, $\epsilon_j > 0$ *such that the intervals* $(\xi_j - \epsilon_j, \xi_j + \epsilon_j)$ $(j = 1 \cdots n)$ *are pairwise disjoint and* $0 \notin \bigcup_{j=1}^n [\xi_j - \epsilon_j, \xi_j + \epsilon_j] \subset (-R, R)$. *Furthermore, suppose that for* $j = 1, \cdots, n$ *there exist* $\eta_j \in (\xi_j - \epsilon_j, \xi_j + \epsilon_j)$ *such that*

$$M_j = \sup_{t \geq 0} \Big| \int_0^t \exp(-i\eta_j s) f(s) ds \Big| < \infty \qquad (j = 1 \cdots n).$$

Then,

$$\overline{\lim_{t\to\infty}} \Big| \int_0^t f(s) ds \Big| \leq \frac{2M_0}{R} \prod_{j=1}^n a_j + 12 \sum_{j=1}^n M_j \delta_j \prod_{\substack{k=1 \\ k \neq j}}^n b_{jk}, \tag{5.5}$$

where

$$
\begin{aligned}
&a_j = (1+\epsilon_j^2(R-|\xi_j|)^{-2})\xi_j^2(\xi_j^2-\epsilon_j^2)^{-1}; \\
(5.6) \qquad &b_{jk} = (1+\epsilon_k^2(|\xi_j-\xi_k|-\epsilon_j)^{-2})\xi_k^2(\xi_k^2-\epsilon_k^2)^{-1} \qquad (k \neq j); \\
&\delta_j = \epsilon_j\xi_j^2(|\xi_j|-\epsilon_j)^{-1}(\xi_j^2-\epsilon_j^2)^{-1}.
\end{aligned}
$$

Proof. We modify the proof of [2, Lemma 3.1] in the following way, keeping the notation used there (cf. also [25, 2.2]). The paths γ_j are replaced by straight lines on the imaginary axis ($j = 0, \cdots, n$). Applying (a slight extension of) Cauchy's theorem to $g(\lambda) = \hat{f}(\lambda)$, we have $0 = -(1/2\pi i)\int_\gamma h(z)(g(z)/z)e^{tz}dz$. Moreover, g_t being entire implies

$$\int_0^t f(s)ds = g_t(0) = \frac{1}{2\pi i}\int_{|z|=R} h(z)g_t(z)e^{tz}\frac{dz}{z}$$

and

$$0 = \sum_{j=1}^n \frac{1}{2\pi i}\int_{|z-i\eta_j|=\epsilon_j} h(z)g_t(z)e^{tz}\frac{dz}{z}.$$

Summing up, we obtain

$$
\begin{aligned}
\int_0^t f(s)ds = \frac{1}{2\pi i} &\int_{\substack{|z|=R \\ \mathrm{Re}z>0}} h(z)(g_t(z)-g(z))e^{tz}\frac{dz}{z} \\
&+\sum_{j=1}^n \frac{1}{2\pi i}\int_{\substack{|z-i\eta_j|=\epsilon_j \\ \mathrm{Re}z>0}} h(z)(g_t(z)-g(z))e^{tz}\frac{dz}{z} \\
&-\sum_{j=1}^n \frac{1}{2\pi i}\int_{\gamma_j} h(z)g(z)e^{tz}\frac{dz}{z} + \frac{1}{2\pi i}\int_{\substack{|z|=R \\ \mathrm{Re}z<0}} h(z)g_t(z)e^{tz}\frac{dz}{z} \\
&+\sum_{j=1}^n \frac{1}{2\pi i}\int_{\substack{\mathrm{Re}z<0 \\ |z-i\eta_j|=\epsilon_j}} h(z)g_t(z)e^{tz}\frac{dz}{z}
\end{aligned}
$$

Now the third term converges to zero ($t \to \infty$) by the Riemann–Lebesgue lemma; the other estimates are given in [2, Lemma 3.1]. □

We put Lemma 5.2 in a different form (corresponding to Tauberian theorems of type D) keeping the definition (5.6) throughout this section.

LEMMA 5.3. *Let $\varphi \in L^1_{\mathrm{loc}}([0,\infty), X) \cap C^1([\tau,\infty), X)$ where $\tau \geq 0$. Assume that $\hat{\varphi}(\lambda)$ has a continuous extension to $K := \mathbb{C}_+ \cup i([-R,R]\setminus\bigcup_{j=1}^n(\eta_j-\epsilon_j,\eta_j+\epsilon_j))$ where $\eta_j \in \mathbb{R}$, $\epsilon_j > 0$ such that the intervals $(\eta_j-\epsilon_j,\eta_j+\epsilon_j)$ are pairwise disjoint ($j = 1\cdots n$) and $0 \notin \bigcup_{j=1}^n[\eta_j-\epsilon_j,\eta_j+\epsilon_j] \subset (-R,R)$. Suppose that*

$$N_0 := \sup_{t\geq\tau}|\varphi'(t)| + |\varphi(\tau)| < \infty$$

and

$$N_j := \sup_{t\geq\tau}\left|\int_\tau^t e^{-i\eta_j s}\varphi'(s)ds\right| + |\varphi(\tau)| < \infty, \qquad \textit{for } j = 1,\cdots,n.$$

Then,

$$\overline{\lim_{t\to\infty}}|\varphi(t)| \le \frac{2N_0}{R}\prod_{j=1}^{n} a_j + 12\sum_{j=1}^{n} N_j\delta_j \prod_{\substack{k=1\\k\ne j}}^{n} b_{jk}.$$

Proof. (a) We assume that $\tau = 0$. Let $f(t) = \varphi'(t) + \varphi(0)\exp(-t)$. Then $\hat{f}(\lambda)/\lambda = \hat{\varphi}(\lambda) - \varphi(0)/(1+\lambda)$ has a continuous extension to K. Moreover, $|f(t)| \le |\varphi'(t)| + |\varphi(0)| \le N_0$ $(t \ge 0)$ and $|\int_0^t \exp(-i\eta_j s)f(s)ds| = |\int_0^t \exp(-i\eta_j s)\varphi'(s)ds + \varphi(0)\int_0^t \exp(-i\eta_j s)\exp(-s)ds| \le N_j$ $(t \ge 0)$, $j = 1\cdots n$. Since $\int_0^t f(s)ds = \varphi(t) - \varphi(0)\exp(-t)$ one has $\overline{\lim}_{t\to\infty}|\varphi(t)| = \overline{\lim}_{t\to\infty}|\int_0^t f(s)ds|$. So the claim follows from Lemma 5.2.

(b) If $\tau \ge 0$ is arbitrary we apply (a) to $\psi(t) = \varphi(t+\tau)$. □

Proof of Theorem 5.1. Since $S(t)x = x$ on $N(A)$ we can assume that $P = Q = 0$ in the case when $\hat{a}(0) = \infty$. Choose $\nu_0 \in \varrho(A)$ and let $L = (\nu_0 - A)^{-1}$. Let $0 \le \mu_0 \in \mathbb{R}\backslash E$ be fixed. Let $R > \mu_0$ such that $\pm R \notin E$. We set $E_0 = E \cap [\mu_0 - R, \mu_0 + R]$. For every ordinal α, we define inductively subsets E_α of E in the following way. Suppose that E_β has been defined for all $\beta < \alpha$. We let E_α be the set of all cluster points of $E_{\alpha-1}$, if α has a predecessor $\alpha - 1$, and $E_\alpha = \bigcap_{\beta<\alpha} E_\beta$ if not.

For $\mu \in E$, $\mu \ne 0$ we define

$$B(\mu) = \begin{cases} \dfrac{1-\hat{a}(i\mu)A}{C(\mu)} i\mu & \text{if } |\hat{a}(i\mu)| \le 1 \\ \left(\dfrac{1}{\hat{a}(i\mu)} - A\right) i\mu/C(\mu) & \text{if } |\hat{a}(i\mu)| > 1, \end{cases}$$

where $C(\mu)$ is the constant from hypothesis (H2), and

$$B(0) = \begin{cases} A & \text{if } \hat{a}(0) = \infty \\ 1 - \hat{a}(0)A & \text{if } \hat{a}(0) \in \mathbb{C}. \end{cases}$$

We shall prove the following.

Inductive statement. If $\mu_j \in E_\alpha$, $\epsilon_j > 0$ $(j = 1, \cdots, n)$ such that $(\mu_j - \epsilon_j, \mu_j + \epsilon_j)$ are pairwise disjoint,

$$E_\alpha \subset \bigcup_{j=1}^{n}(\mu_j - \epsilon_j, \mu_j + \epsilon_j) \quad \text{and } \mu_0 \notin \bigcup_{j=1}^{n}[\mu_j - \epsilon_j, \mu_j + \epsilon_j] \subset (\mu_0 - R, \mu_0 + R),$$

then

$$\overline{\lim_{t\to\infty}}|(S(t) - Q)LUx| \le \frac{2N_0|Ux|}{R}\prod_{j=1}^{n} a_j + 12\sum_{j=1}^{n} C_0|U_jx|\delta_j \prod_{\substack{k=1\\k\ne j}} b_{jk} \tag{5.7}$$

for all $x \in D(A^n)$, where

$$U = \prod_{j=1}^{n} B_j; \quad U_j = \prod_{\substack{k=1\\k\ne j}}^{n} B_k, \quad B_j = B(\mu_j);$$

the constants a_j, δ_j, b_{jk} are given by (5.6), and C_0, N_0 are constants which will be defined below and do not depend on μ_j, ϵ_j.

It is part of the inductive statement that

$$\overline{\lim_{t\to\infty}}|(S(t)-Q)Lx| \le \frac{2N_0}{R}|x| \tag{5.8}$$

for all $x \in X$ if $E_\alpha = \emptyset$ (which is (5.7) with the convention that the empty product is 1 and the empty sum zero).

Once the inductive statement has been established, the theorem is proved as follows. Since E_α is compact and countable, E_α is either empty or contains isolated points, so that $E_\alpha = \emptyset$ or $E_{\alpha+1} \neq E_\alpha$. Thus it follows that for some α (at most ω_1), $E_\alpha = \emptyset$. Hence, (5.8) holds. We can choose $0 < R \notin E \cup -E$ arbitrarily large. Thus $\lim_{t\to\infty}|(S(t)-Q)Lx| = 0$ for all $x \in X$. Since $R(L) = D(A)$ is dense in X and $S(t)$ is bounded the claim follows.

It remains to prove the inductive statement.

(1) $\alpha = 0$. Let $\mu_j \in E_0$, $\epsilon_j > 0$ such that

$$E_0 \subset \bigcup_{j=1}^{n}(\mu_j-\epsilon_j, \mu_j+\epsilon_j) \text{ and } \mu_0 \notin \bigcup_{j=1}^{n}[\mu_j-\epsilon_j, \mu_j+\epsilon_j] \subset (\mu_0-R, \mu_0+R),$$

according to the statement. Let $y \in X$ and set $\varphi(t) = e^{-i\mu_0 t}(S(t)-Q)Ly$ $(t \ge 0)$. We verify that φ satisfies the hypotheses of Lemma 5.3 (after specification of y).

For $\operatorname{Re}\lambda > 0$ we have

$$\hat{\varphi}(\lambda) = \hat{S}(\lambda+i\mu_0)Ly - \frac{1}{i\mu_0+\lambda}QLy = \frac{1}{\lambda+i\mu_0}((1-\hat{a}(\lambda+i\mu_0)A)^{-1}-Q)Ly.$$

Set $\eta_j = \mu_j - \mu_0$ $(j = 1, \cdots, n)$. Then $0 \notin \cup_{j=1}^{n}[\eta_j-\epsilon_j, \eta_j+\epsilon_j] \subset (-R, R)$ and $\hat{\varphi}(\lambda)$ has a continuous extension to $\mathbb{C}_+ \cup i([-R,R]\setminus\bigcup_{j=1}^{n}(\eta_j-\epsilon_j, \eta_j+\epsilon_j))$.

Setting $C_L = |L| + |AL|$ we have $|Ly|_A \le C_L|y|$. We have

$$\varphi'(t) = -i\mu_0\exp(-i\mu_0 t)(S(t)-Q)Ly + \exp(-i\mu_0 t)S'(t)Ly.$$

Using (H3), we obtain

$$|\varphi'(t)| + |\varphi(\tau)| \le N_0|y| \qquad (t \ge \tau) \tag{5.9}$$

with $N_0 = \mu_0(M+|Q|)|L| + MC_L + (M+|Q|)|L|$.

Now let $y = Ux = B_jU_jx$ where $x \in D(A^n)$, and observe that $|\eta_j| = |\mu_j-\mu_0| \le R$, hence $|\mu_j| \le R+\mu_0$. Moreover, since $C(\mu) \ge 1$, it follows from the definition of $B(\mu)$ that

$$|B(\mu)L| \le |\mu|C_L \qquad (0 \neq \mu \in E); \tag{5.10}$$

in particular,

$$|B_jL| \le (R+\mu_0)C_L \quad \text{if } \mu_j \neq 0. \tag{5.11}$$

Due to (1.1), hypothesis (H2) implies

$$\left|\int_0^t e^{-i\mu s}S(s)(1-\hat{a}(i\mu)A)x\,ds\right| \le C(\mu)|x|_A \qquad (x \in D(A))$$

if $\mu \in E\backslash\{0\}$ and $\hat{a}(i\mu) \in \mathbb{C}$. Consequently, it follows from (H2) that

$$\left|\int_0^t e^{-i\mu s} S(s)B(\mu)y\,ds\right| \leq |\mu||y|_A \quad (y \in D(A), \quad \mu \in E\backslash\{0\}). \tag{5.12}$$

We estimate $\int_\tau^t e^{-i\mu_j s}\varphi'(s)ds$.

Case 1. $\mu_j \neq 0$, i.e., $\eta_j = \mu_j - \mu_0 \neq -\mu_0$.

$$\begin{aligned}
&\int_\tau^t \exp(-i\eta_j s)\varphi'(s)ds \\
&\quad = -i\mu_0 \int_\tau^t \exp(-i\mu_j s)(S(s) - Q)Ly\,ds + \int_\tau^t \exp(-i\mu_j s)S'(s)Ly\,ds \\
&\quad = -i\mu_0 \int_\tau^t \exp(-i\mu_j s)(S(s) - Q)Ly\,ds + \exp(-i\mu_j t)S(t)Ly \\
&\qquad\qquad - \exp(-i\mu_j \tau)S(\tau)Ly + i\mu_j \int_\tau^t \exp(-i\mu_j s)S(s)Ly\,ds \\
&\quad = i(\mu_j - \mu_0) \int_0^t \exp(-i\mu_j s)S(s)Ly\,ds + \frac{\mu_0}{\mu_j}(\exp(-i\mu_j\tau) - \exp(-i\mu_j t))QLy \\
&\qquad\qquad + \exp(-i\mu_0 t)S(t)Ly - \exp(-i\mu_j\tau)S(\tau)Ly \\
&\qquad\qquad -i(\mu_j - \mu_0) \int_0^\tau \exp(-i\mu_j s)S(s)Ly\,ds.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\left|\int_\tau^t \exp(-i\eta_j s)\varphi'(s)ds\right| \\
&\quad \leq R\left|\int_0^t \exp(-i\mu_j s)S(s)B_j LU_j x\,ds\right| \\
&\qquad\qquad +2\frac{\mu_0}{|\mu_j|}|Q||B_j LU_j x| + 2M|B_j LU_j x| + R\tau M|B_j LU_j x| \\
&\quad \leq R|\mu_j|.|LU_j x|_A + 2\mu_0|Q|C_L|U_j x| + 2M(R + \mu_0)C_L|U_j x| \\
&\qquad\qquad +R\tau M(R + \mu_0)C_L|U_j x|,
\end{aligned}$$

by (5.12), (5.10), and (5.11). Setting

$$C_1 := R(R + \mu_0)C_L + 2\mu_0|Q|C_L + 2M(R + \mu_0)C_L + R\tau M(R + \mu_0)C_L,$$

we obtain

$$\left|\int_\tau^t \exp(-i\eta_j s)\varphi'(s)ds\right| \leq C_1|U_j x|. \tag{5.13}$$

Case 2. $\mu_j = 0$; that is, $\eta_j = -\mu_0$. Then,

$$\begin{aligned}
\int_\tau^t e^{-i\eta_j s}\varphi'(s)ds &= -i\mu_0 \int_\tau^t (S(s) - Q)Ly\,ds + \int_\tau^t S'(s)Ly\,ds \\
&= -i\mu_0 \int_0^t (S(s) - Q)y\,ds + S(t)Ly - S(\tau)Ly + i\mu_0 \int_0^\tau (S(s) - Q)Ly\,ds.
\end{aligned}$$

We must distinguish two cases. (a) If $\hat{a}(0)=\infty$, then $Q=0$ and $B_j=A$, $y=AU_jx$. Then,

$$\begin{aligned}\left|\int_\tau^t \exp(-i\eta_j s)\varphi'(s)ds\right| &\le \mu_0\left|\int_0^t S(s)ALU_jx\,ds\right| + 2M|ALU_jx| + \mu_0\tau M|ALU_jx| \\ &\le \mu_0 C(0)|LU_jx|_A + 2MC_L|U_jx| + \mu_0\tau MC_L|U_jx|\end{aligned}$$

by (H2). Hence, $|\int_\tau^t \exp(-i\eta_j s)\varphi'(s)ds| \le C_2|U_jx|$ if we set $C_2 := (\mu_0 C(0)C_L + 2MC_L + \mu_0\tau MC_L)$ if $0\in E$ and $\hat{a}(0)=\infty$.

(b) If $\hat{a}(0)\in\mathbb{C}$, then $Q=(I-\hat{a}(0)A)^{-1}$, $B_j=Q^{-1}$, $y=Q^{-1}U_jx$ and so $(S(s)-Q)Ly=(S(s)Q^{-1}-I)LU_jx=(S(s)(I-\hat{a}(0)A)-I)LU_jx=S(s)LU_jx-LU_jx-\hat{a}(0)S(s)ALU_jx=A(a*S)(s)LU_jx-\hat{a}(0)S(s)ALU_jx$ by (4.1). Thus,

$$\begin{aligned}\left|\int_\tau^t \exp(-i\eta_j s)\varphi'(s)ds\right| &\le \mu_0\left|\int_0^t ((a*S)(s)-\hat{a}(0)S(s))ALU_jx\,ds\right| + 2M|Q^{-1}LU_jx| \\ &\qquad + \mu_0\tau(M+|Q|)|Q^{-1}LU_jx|\end{aligned}$$

$$\le \mu_0 C(0)|LU_jx|_A + 2M|(I-\hat{a}(0)A)L||U_jx| + \mu_0\tau(M+|Q|)|(I-\hat{a}(0)A)L||U_jx|$$

by (H2). Hence, $|\int_\tau^t \exp(-i\eta_j s)\varphi'(s)ds| \le C_2|U_jx|$ $(t\ge\tau)$ if we set

$$C_2=\mu_0C(0)C_L+2M|(I-\hat{a}(0)A)L|+\mu_0\tau(M+|Q|)+\mu_0\tau(M+|Q|)|(I-\hat{a}(0)A)L|$$

in the case $0\in E$, $\hat{a}(0)\in\mathbb{C}$. So far, we have proved that

$$\left|\int_\tau^t \varphi'(s)\exp(-i\eta_j s)ds\right| \le C_3|U_jx|, \quad j=1,\cdots,n, \quad t\ge\tau$$

if we put $C_2:=0$ in the case where $0\notin E$ and $C_3=\max\{C_1,C_2\}$ (see (5.13)). Finally, we let

$$\begin{aligned} C_4 &= \begin{cases} MC_L & \text{if } 0\in E, \hat{a}(0)=\infty \\ (M+|Q|)(I-\hat{a}(0)A)L| & \text{if } 0\in E, \hat{a}(0)\in\mathbb{C} \\ 0 & \text{if } 0\notin E \end{cases} \\ C_5 &= (M+|Q|)(R+\mu_0)C_L \\ C_6 &= \max\{C_4,C_5\}. \end{aligned}$$

Then $|\varphi(\tau)|\le C_6|U_jx|$ $(j=1\cdots n)$. In fact, if $\mu_j\ne 0$, then

$$|\varphi(\tau)|\le(M+|Q|)|B_jLU_jx|\le C_5|U_jx|;$$

if $\mu_j=0$ and $\hat{a}(0)=\infty$, then $Q=0$, $B_j=A$, $y=AU_jx$, and so $|\varphi(\tau)|\le M|LAU_jx|\le MC_L|U_jx|=C_4|U_jx|$; if $\mu_j=0$ and $\hat{a}(0)\in\mathbb{C}$, then $Q=(I-\hat{a}(0)A)^{-1}$, $y=Q^{-1}U_jx$ and so

$$|\varphi(\tau)|\le(M+|Q|)|(I-\hat{a}(0)A)L||U_jx|=C_4|U_jx|.$$

Letting $C_0:=\max\{C_3,C_6\}$, we finally have

$$\left|\int_\tau^t \varphi'(s)\exp(-i\eta_j s)ds\right| + |\varphi(\tau)| \le C_0|U_jx|$$

$(j = 1, \cdots, n)$. In view of (5.9), now the claim (5.7) follows from Lemma 5.3. This proves the inductive statement for $\alpha = 0$.

(2) Let α be an ordinal greater than zero and assume that the inductive statement holds for all ordinals $\beta < \alpha$. We show the statement to hold for α. Let $(\mu_j - \epsilon_j, \mu_j + \epsilon_j)$ $(j = 1, \cdots, n)$ be disjoint intervals such that $\mu_0 \notin \cup_{j=1}^{n}[\mu_j - \epsilon_j, \mu_j + \epsilon_j] \subset (\mu_0 - R, \mu_0 + R)$ and $E_\alpha \subset \Omega := \cup_{j=1}^{n}(\mu_j - \epsilon_j, \mu_j + \epsilon_j)$.

Case 1. $\alpha - 1$ does not exist. Then $E_\alpha = \cap_{\beta<\alpha} E_\beta$. Since Ω is open and E_0 compact, it follows that $E_\beta \subset \Omega$ for some $\beta < \alpha$. So (5.7) follows trivially from the inductive hypothesis.

Case 2. $\alpha - 1$ exists. Since E_α is the set of all accumulation points of $E_{\alpha-1}$, $E_{\alpha-1} \backslash E_\alpha$ is finite, say $E_{\alpha-1} \backslash E_\alpha = \{\mu_{n+1}, \cdots, \mu_{n+p}\}$. Let $\epsilon_j > 0$, $j = n+1, \cdots, n+p$ be small enough so that $\mu_0 \notin \cup_{j=1}^{n+p}[\mu_j - \epsilon_j, \mu_j + \epsilon_j] \subset (\mu_0 - R, \mu_0 + R)$. Since $E_{\alpha-1} \subset \cup_{j=1}^{n+p}(\mu_j - \epsilon_j, \mu_j + \epsilon_j)$, we conclude from the inductive hypothesis for $\alpha - 1$ that

$$\overline{\lim_{t\to\infty}}|S(t) - Q)LVy| \le \frac{2N_0|Vy|}{R} \prod_{j=1}^{n+p} a_j + 12 \sum_{j=1}^{n+p} C_0 |V_j y| \delta_j \prod_{\substack{k=1 \\ k\neq j}}^{n+p} b_{jk}$$

for all $y \in D(A^{n+p})$, where

$$V = \prod_{j=1}^{n+p} B_j, \quad V_j = \prod_{\substack{k=1 \\ k\neq j}}^{n+p} B_k, \quad B_j = B(\mu_j) \quad (j = 1 \cdots n+p).$$

Letting $\epsilon_j \downarrow 0$ for $j = n+1, \cdots, n+p$, we obtain

$$\overline{\lim_{t\to\infty}}|(S(t) - Q)LVy| \le 2(N_0/R)|Vy| \prod_{j=1}^{n} a_j + 12 \sum_{j=1}^{n} C_0 |V_j y| \delta_j \prod_{\substack{k=1 \\ k\neq j}}^{n} b_{jk}. \tag{5.14}$$

Letting $W = \prod_{j=n+1}^{n+p} B_j$, $U = \prod_{j=1}^{n} B_j$, $U_j = \prod_{\substack{k=1 \\ k\neq j}}^{n} B_k$, we can rewrite (5.14) as

$$\begin{aligned} \overline{\lim_{t\to\infty}}|(S(t) - Q)LUWy| &\le 2(N_0/R)|UWy| \prod_{j=1}^{n} a_j \\ &\quad + 12 \sum_{j=1}^{n} C_0 |U_j W y| \delta_j \prod_{\substack{k=1 \\ k\neq j}}^{n} b_{jk} \qquad (y \in D(A^{n+p})). \end{aligned} \tag{5.15}$$

Now the operators $B_j (j = n+1, \cdots, n+p)$ commute and have dense range by (H1). This implies that $WD(A^{n+p})$ is dense in $(D(A^n); |\ |_{A^n})$, where $|x|_{A^n} := |x| + |Ax| + \cdots + |A^n x|$ for $x \in D(A^n)$. Thus, given $x \in D(A^n)$ we find $y_m \in D(A^{n+p})$ such that $\lim_{m\to\infty} |Wy_m - x|_{A^n} = 0$, hence $\lim_{m\to\infty} UWy_m = Ux$ in $(D(A), |\ |_A)$. Setting $y = y_m$ in (5.15), we obtain (5.7) by letting $m \to \infty$. This completes the proof of Theorem 5.1. □

6. Some examples and illustrations. In this section we want to discuss several examples of kernels $a(t)$ and operators A to which the General Convergence Theorem applies and also to present conditions on the kernel $a(t)$ such that assumptions (H2) and (H3) of Theorem 6.1 are satisfied.

We begin with the semigroup case $a(t) \equiv 1$, $t \geq 0$. Then the resolvent $S(t)$ satisfying

$$S(t) = I + A\int_0^t a(\tau)S(t-\tau)d\tau, \qquad t \geq 0, \tag{6.1}$$

is the semigroup generated by A, i.e., $S(t) = e^{At}$. Therefore the relation $S'(t)x = S(t)Ax$ shows that (H3) is trivially satisfied whenever the semigroup is bounded. To verify (H2), observe that $\hat{a}(\lambda) = 1/\lambda$; hence, $\hat{a}(0) = \infty$ and $\hat{a}(i\mu) \in \mathbb{C}$ otherwise. For $\mu = 0$ we obtain

$$\int_0^t S(\tau)Ax\,d\tau = \int_0^t S'(\tau)x\,d\tau = S(t)x - x, \qquad t > 0,$$

and so (H2) is valid for $\mu = 0$. If $\mu \neq 0$ we get, via an integration by parts,

$$\begin{aligned}\int_0^t e^{-i\mu\tau}((a*S)(\tau) - \hat{a}(i\mu)S(\tau))Ax\,d\tau &= \int_0^t e^{-i\mu\tau}\left(\int_0^\tau S(s)ds - \frac{1}{i\mu}S(\tau)\right)Ax\,d\tau \\ &= \frac{1}{i\mu}e^{-i\mu t}(x - S(t)x);\end{aligned}$$

hence, (H2) is valid for all $\mu \in \mathbb{R}$. Since $E = \sigma(A) \cap i\mathbb{R}$, (H1) becomes $(\sigma_p(A) \cup \sigma_p(A')) \cap i\mathbb{R} \subset \{0\}$ and $N(A)^\perp \cap N(A') = \{0\}$. Thus, the General Convergence Theorem reduces for the case $a(t) \equiv 1$ to the following version of the stability theorem of Arendt and Batty [2], and Lyubich and Phong [28].

COROLLARY 6.1. *Suppose A generates a bounded C_0-semigroup in X, let $\sigma(A) \cap i\mathbb{R}$ be countable, $\sigma_p(A') \cap i\mathbb{R} \subset \{0\}$, and assume $N(A)^\perp \cap N(A') = \{0\}$. Then $\lim_{t\to\infty} S(t)x = Px$ for each $x \in X$, where P denotes the projection onto $N(A)$ along $\overline{R(A)}$.*

Next we show that condition (H2) for $\mu \neq 0$ is satisfied for a large class of kernels, provided the resolvent $S(t)$ is known to be bounded. We denote by $BV(\mathbb{R}_+)$ the space of all functions $a : \mathbb{R} \to \mathbb{R}$ of bounded variation, which are left-continuous and such that $a(t) = 0$ for $t \leq 0$.

PROPOSITION 6.2. *Suppose that the resolvent $S(t)$ of* (1.1) *is bounded, let $a(t)$ be of the form*

$$a(t) = \sum_{k=0}^n a_k(t), \qquad t > 0, \tag{6.2}$$

where

$$a_0, ta_0 \in L^1(\mathbb{R}_+) \quad \textit{and} \quad \textit{for } k = 1, \cdots, n,$$

$$a_k \in W^{k-1,1}_{\text{loc}}(\mathbb{R}_+), \quad a_k^{(k-1)} \in BV(\mathbb{R}_+), \quad \int_0^\infty t|da_k^{(k-1)}(t)| < \infty. \tag{6.3}$$

Then $\varrho(a) \supset i\mathbb{R}\backslash\{0\}$, $\hat{a}(i\mu) \in \mathbb{C}$ for all $\mu \in \mathbb{R}\backslash\{0\}$, and for each $\mu \in \mathbb{R}\backslash\{0\}$ there is constant $c(\mu)$, such that

$$\left|\int_0^t e^{-i\mu s}[(a*S)(s) - \hat{a}(i\mu)S(s)]Ax\,dx\right| \leq c(\mu)|x|_A, \qquad x \in D(A). \tag{6.4}$$

If $n = 0$ the assertions also hold for $\mu = 0$.

Proof. Adding suitable constants to the functions of bounded variation $b_k(t) = a_k^{(k-1)}(t)$, we may assume $a_k^{(i)}(0) = 0$ for all $0 \le i \le k-2 \le n-2$. Let $b_0(t) = \int_0^t a_0(\tau)d\tau$. The familiar formula

$$\widehat{db}_k(\lambda) = (da_k^{(k-1)})^\wedge(\lambda) = \lambda^k \hat{a}_k(\lambda), \quad \operatorname{Re}\lambda > 0, \quad k = 1, \cdots, n,$$

by (6.2) yields the representation

$$\hat{a}(\lambda) = \sum_{k=0}^{n} \widehat{db}_k(\lambda)\lambda^{-k}, \qquad \operatorname{Re}\lambda > 0. \tag{6.5}$$

Since $b_k \in BV(\mathbb{R}_+)$, $k = 0, \cdots, n$, (6.5) shows that $\hat{a}(\lambda)$ admits a continuous extension at least to $\overline{\mathbb{C}}_+\backslash\{0\}$; hence, we obtain $\varrho(a) \supset i\mathbb{R}\backslash\{0\}$ and $\hat{a}(\lambda) \in \mathbb{C}$ for all $\lambda \in \overline{\mathbb{C}}_+\backslash\{0\}$. Integrating by parts k times leads to

$$\int_0^t e^{-i\mu\tau}(a_k * S)(\tau)d\tau = (i\mu)^{-k}\int_0^t e^{-i\mu\tau}(db_k * S)(\tau)d\tau - e^{-i\mu t}\sum_{j=0}^{k-1}(a_k^{(j)} * S)(t)(i\mu)^{-j-1};$$

hence, summation over k gives

$$\begin{aligned}
&\int_0^t e^{-i\mu\tau}[(a * S)(\tau) - \hat{a}(i\mu)S(\tau)]Ax\,d\tau \\
&= \sum_{k=0}^{n}(i\mu)^{-k}\int_0^t e^{-i\mu\tau}[(db_k * S)(\tau) - \widehat{db}_k(i\mu)S(\tau))]Ax\,d\tau \\
&\quad - e^{-i\mu t}\sum_{k=1}^{n}\sum_{j=0}^{k-1}(a_k^{(j)} * S)(t)(i\mu)^{-j-1}Ax \\
&= \sum_{k=0}^{n}(i\mu)^{-k}(T_k(t) - e^{-i\mu t}R_k(t))Ax,
\end{aligned} \tag{6.6}$$

where

$$T_k(t) = \int_0^t e^{-i\mu\tau}[(db_k * S)(\tau) - \widehat{db}_k(i\mu)S(\tau)]d\tau \tag{6.7}$$

and

$$R_k(t) = \sum_{j=k}^{n}(a_j^{(k-1)} * S)(t), \qquad R_0(t) = 0. \tag{6.8}$$

To estimate $T_k(t)$, we write

$$\begin{aligned}
T_k(t) &= \int_0^t e^{-i\mu\tau}\int_0^\tau db_k(\tau - s)S(s)d\tau - \int_0^t e^{-i\mu\tau}\widehat{db}_k(i\mu)S(\tau)d\tau \\
&= \int_0^t S(s)e^{-i\mu s}\left(\int_s^t db_k(\tau - s)e^{-i\mu(\tau - s)} - \widehat{db}_k(i\mu)\right)ds, \\
&= -\int_0^t S(s)e^{-i\mu s}\left(\int_{t-s}^\infty db_k(\tau)e^{-i\mu\tau}\right)ds.
\end{aligned}$$

Hence,

$$|T_k(t)| \le M \int_0^t \int_{t-s}^{\infty} |db_k(\tau)| ds = M \left(\int_t^{\infty} \int_0^t ds |db_k(\tau)| + \int_0^t \int_{t-\tau}^t ds |db_k(\tau)| \right)$$
$$= M \left(t \int_t^{\infty} |db_k(\tau)| + \int_0^t \tau |db_k(\tau)| \right) \le M \int_0^{\infty} \tau |db_k(\tau)| = M_k < \infty,$$

where $M = \sup_{\tau \ge 0} |S(\tau)|$. To derive a bound on the $R_k(t)$, we expand $(a_k * S)(t)$ into a Taylor series up to order k,

$$(a_k * S)(t+h) - (a_k * S)(t) = \sum_{j=0}^{k-1} (a_k^{(j)} * S)(t) \frac{h^j}{j!} + \int_t^{t+h} (db_k * S)(\tau) \frac{(t+h-\tau)^{k-1}}{(k-1)!} d\tau.$$

Summing over k, we obtain with (6.1)

$$S(t+h)x - S(t)x = \sum_{k=1}^{n} \sum_{j=1}^{k} (a_k^{(j-1)} * SAx)(t) \frac{h^{j-1}}{(j-1)!}$$
$$+ \sum_{k=0}^{n} \int_t^{t+h} (db_k * SAx)(\tau) \frac{(t+h-\tau)^{k-1}}{(k-1)!} d\tau.$$

Since $S(t)$ is bounded and $b_k \in BV(\mathbb{R}_+)$ the polynomials

$$P_t(h)x = \sum_{j=1}^{n} \frac{h^{j-1}}{(j-1)!} \left(\sum_{k=j}^{n} a_k^{(j-1)} * SAx \right) = \sum_{j=1}^{n} R_{j-1}(t) Ax \frac{h^{j-1}}{(j-1)!}$$

are bounded, uniformly for $0 \le h \le 1$, $t \ge 0$; but this implies the existence of a constant $C > 0$ such that

$$(6.9) \qquad |R_k(t)Ax| \le C|x|_A, \quad x \in D(A), \quad k = 1, \cdots, n.$$

The proof is now complete. □

A special case of Proposition 6.2 will be used in §7, namely, the following.

COROLLARY 6.3. *Suppose that the resolvent $S(t)$ for* (1.1) *is bounded; let $a(t)$ be of the form*

$$(6.10) \qquad a(t) = b_0 + b_\infty t + \int_0^t b_1(s) ds, \qquad t > 0,$$

where $b_0, b_\infty \ge 0$ are constants and $b_1 \in L^1_{\text{loc}}(\mathbb{R}_+)$ is nonnegative, nonincreasing, and convex. Then $\varrho(a) \supset i\mathbb{R}\backslash\{0\}$, $\hat{a}(i\mu) \in \mathbb{C}$ for all $\mu \in \mathbb{R}\backslash\{0\}$ and (6.4) *holds for each $\mu \in \mathbb{R}\backslash\{0\}$.*

Proof. We may assume $\lim_{t\to\infty} b_1(t) = 0$, changing b_∞ otherwise. Let $t_0 > 0$ and

$$c_1(t) = \begin{cases} b_1(t) - b_1(t_0) & \text{for } t \le t_0, \\ 0 & \text{for } t \ge t_0, \end{cases}$$
$$c_3(t) = \begin{cases} 0 & \text{for } t \le t_0, \\ b_1(t_0) - b_1(t) & \text{for } t > t_0, \end{cases}$$

and define $a_0(t) = 0$,

$$a_1(t) = b_0 + \int_0^t c_1(\tau)d\tau, \qquad t > 0,$$

$$a_2(t) = (b_\infty + b_1(t_0))t, \qquad t > 0,$$

$$a_3(t) = -\int_0^t c_3(\tau)d\tau, \qquad t > 0.$$

Obviously, $a(t) = a_1(t) + a_2(t) + a_3(t)$, $a_1 \in BV(\mathbb{R}_+)$ and $da_1 = b_0\delta + c_1(t)dt$ has all moments since its support is compact; $a_2 \in W^{1,1}_{\text{loc}}(\mathbb{R}_+)$, $\dot{a}_2 = b_\infty + b_1(t_0) \in BV(\mathbb{R}_+)$ and $d\dot{a}_2 = (b_\infty + b_1(t_0))\delta$ also has all moments. a_3 belongs to $W^{2,1}_{\text{loc}}(\mathbb{R}_+)$ since b_1 is nonincreasing and convex, and $\ddot{a}_3(t) = -\dot{c}_3(t)$ for $t > t_0$, $\ddot{a}_3(t) = 0$, for $t < t_0$; moreover, by convexity, $-\dot{c}_3(t)$ is nonincreasing for $t > t_0$ and nonnegative, hence $\ddot{a}_3 \in BV(\mathbb{R}_+) \cap L^1(\mathbb{R}_+)$ and in particular $d\ddot{a}_3$ admits a finite first moment, since $\ddot{a}_3(t)$ is nondecreasing, as integration by parts shows. □

The argument at the end of the proof of Proposition 6.2 also yields (H3), i.e., boundness of $S'(t)x$, whenever $S(t)$ is bounded and $a(t)$ is of the form (6.2), (6.3) with $a_0 = 0$. More precisely, we have the following.

PROPOSITION 6.4. *Suppose the resolvent $S(t)$ for* (1.1) *is bounded; let $a(t)$ be of the form*

$$a(t) = \sum_{k=1}^{n} a_k(t), \qquad t > 0, \tag{6.11}$$

where

$$a_k \in W^{k-1,1}_{\text{loc}}(\mathbb{R}_+), \quad a_k^{(k-1)} \in BV(\mathbb{R}_+), \quad k = 1, \cdots, n. \tag{6.12}$$

Then there is a constant $C > 0$, such that

$$|S'(t)x| \le C|x|_A \quad \textit{for all } x \in D(A), \quad t > 0. \tag{6.13}$$

Proof. Equation (6.1) yields for $x \in D(A)$

$$S'(t)x = \sum_{k=2}^{n} (\dot{a}_k * SAx)(t) + (da_1 * SAx)(t) = R_2(t)Ax + (da_1 * SAx)(t), \qquad t > 0,$$

where $R_2(t)$ is given by (6.8). Since $a_1 \in BV(\mathbb{R}_+)$, estimate (6.9) yields the assertion. Observe that for the proof of (6.9) no moment condition was used. □

For the applications in §7 we shall need the following special case of Proposition 6.4.

COROLLARY 6.5. *Suppose the resolvent $S(t)$ for* (6.1) *is bounded; let $a(t)$ be of the form*

$$a(t) = b_0 + b_\infty t + \int_0^t b_1(\tau)d\tau, \qquad t > 0, \tag{6.14}$$

where $b_0, b_\infty \ge 0$ and $b_1 \in L^1_{\text{loc}}(\mathbb{R}_+)$ is nonnegative and nonincreasing. Then there is a constant $C > 0$ such that

$$|S'(t)x| \le C|x|_A \quad \textit{for all } x \in D(A), \quad t > 0. \tag{6.15}$$

Proof. We may assume $\lim_{t\to\infty} b_1(t) = 0$. Define

$$a_1(t) = b_0 + \int_0^t c_1(\tau)d\tau, \qquad a_2(t) = (b_\infty + b_1(t_0))t - \int_0^t c_3(\tau)d\tau,$$

where $c_1(t)$ and $c_3(t)$ are defined as in the proof of Corollary 6.3. Then $a_1 \in BV(\mathbb{R}_+)$ and $\dot{a}_2 = b_\infty + b_1(t_0) - c_3(t) \in BV(\mathbb{R}_+)$; hence Proposition 6.4 applies and yields (6.15). □

Observe that in Proposition 6.4 we have to assume that the non-BV part of $a_0(t)$ of $a(t)$ in decomposition (6.2) is absent. This clearly restricts its applicability; however, in case $a_0 \neq 0$, Estimate (6.13) cannot be expected. In general, $S(t)x$ need not be differentiable at all. In this case, we must use the structure of $a_0(t)$ and A directly to obtain a bound on $S'(t)$.

The verification of (H2) for $\mu = 0$ is more difficult. If $n = 0$ in Proposition 6.2 then $\varrho(a) = i\mathbb{R}$, $\hat{a}(i\mu) \in \mathbb{C}$ for all $\mu \in \mathbb{R}$ and (6.4) remains valid for $\mu = 0$ as the proof given there shows (in fact, no integration by parts is needed). On the other hand, if $n \geq 1$ then generically $\hat{a}(0) = \infty$ as (6.5) shows (only one of the $\widehat{db_k}(0) = b_k(\infty)$, $k = 1, \cdots, n$ must be nonzero for $\hat{a}(0) = \infty$); then we have to prove that

$$U(t)Ax = \int_0^t S(\tau)Ax\, d\tau, \qquad t > 0, \tag{6.16}$$

is bounded by the graph norm $|x|_A$ of x. Since by (6.1) we obtain the relations

$$\hat{U}(\lambda)Ax = \frac{1}{\lambda\hat{a}(\lambda)}((S-I)x)^\wedge(\lambda) = \frac{1}{\lambda^2\hat{a}(\lambda)}(\dot{S})^\wedge(\lambda)x$$

for the Laplace transform of $U(t)Ax$, we see that $U(t)Ax$ will be bounded if there is $k \in BV(\mathbb{R}_+)$, such that $\widehat{dk}(\lambda) = (\lambda\hat{a}(\lambda))^{-1}$, or if there is $\ell \in BV(\mathbb{R}_+)$ such that $\widehat{d\ell}(\lambda) = (\lambda^2\hat{a}(\lambda))^{-1}$ and $S'(t)x$ is bounded. It should be clear that more information on the kernel $a(t)$ must be available in order to achieve this, rather than just an expansion of the form (6.2) and (6.3). In §7 it will be shown how this can be done. Let us summarize.

PROPOSITION 6.6. *Suppose that the resolvent $S(t)$ for* (1.1) *is bounded, and assume either of the following.* (a) *There is $k \in BV(\mathbb{R}_+)$ such that $(\lambda\hat{a}(\lambda))^{-1} = \widehat{dk}(\lambda)$, $\lambda > 0$, i.e.,*

$$(k * a)(t) \equiv t, \qquad t > 0.$$

(b) *There is $\ell \in BV(\mathbb{R}_+)$ such that $(\lambda^2\hat{a}(\lambda))^{-1} = \widehat{d\ell}(\lambda)$, $\lambda > 0$, i.e.,*

$$(\ell * a)(t) = t^2/2, \qquad t > 0,$$

and, in addition, suppose that (6.13) *holds.*

Then $0 \in \varrho(a)$, $\hat{a}(0) = \infty$, and there is a constant $C > 0$ such that

$$\left|\int_0^t S(\tau)Ax\, d\tau\right| \leq C|x|_A \quad \textit{for all } x \in D(A), \quad t \geq 0. \tag{6.17}$$

Consider now the cosine case, i.e., $a(t) \equiv t$ and A generating a bounded strongly continuous cosine family $C(t)$. Then we have $S(t) = C(t)$, $t \geq 0$, $a(t)$ is of the form (6.2), (6.3) and also of the form (6.11), (6.12). Since $\lambda^2\hat{a}(\lambda) = 1 = \widehat{d\ell}(\lambda)$ with $\ell(\lambda) = 1$ for $t > 0$, Propositions 6.2, 6.4, and 6.6 imply that (H2) and (H3) of the General Convergence Theorem are satisfied. Since $\sigma(A) \subset (-\infty, 0]$ (H1) becomes $\sigma(A)$ countable and $\sigma_p(A') \subset \{0\}$, $N(A)^\perp \cap N(A') = \{0\}$. Thus we have the following.

COROLLARY 6.7. *Suppose A generates a bounded, strongly continuous cosine family $C(t)$ in X, assume $\sigma(A)$ is at most countable, $\sigma_p(A') \subset \{0\}$ and $N(A)^\perp \cap N(A') = \{0\}$. Then $\lim_{t\to\infty} C(t)x = Px$ for all $x \in X$, where P denotes the projection onto $N(A)$ along $R(A)$.*

We conclude this section with an example which is such that none of the results of this section can be applied, although (H1), (H2), and (H3) hold, and so the General Convergence Theorem can still be used.

Example 6.8. Let X be a Hilbert space, A a dissipative operator in X such that $\varrho(A) \supset i\mathbb{R}$, and let $a(t) = \cos(t)$, $t > 0$. We claim that the resolvent $S(t)$ of (1.1) satisfies

$$\lim_{t\to\infty} S(t)x = x \quad \text{for all } x \in X. \tag{6.18}$$

To prove this we will apply the General Convergence Theorem of §5. Observe first that $\hat{a}(\lambda) = \lambda(\lambda^2+1)^{-1}$; hence $a(t)$ is not of the form (6.2), (6.3) in view of the poles $\lambda = \pm i$ of $\hat{a}(\lambda)$. For the Laplace transform of $S(t)$, we obtain

$$\lambda\hat{S}(\lambda) = (\lambda + 1/\lambda)(\lambda + 1/\lambda - A)^{-1}, \quad \operatorname{Re}\lambda \geq 0, \quad \lambda \neq 0, \tag{6.19}$$

which exists on $\overline{\mathbb{C}}_+\backslash\{0\}$, since A is dissipative and $\varrho(A) \supset i\mathbb{R}$, and the function $\varphi(\lambda) = \lambda + 1/\lambda$ maps $\overline{\mathbb{C}}_+\backslash\{0\}$ onto $\bar{\mathbb{C}}_+$. Furthermore, (6.19) yields

$$\lim_{\lambda\to 0+} \lambda\hat{S}(\lambda)x = \lim_{r\to\infty} r(r-A)^{-1}x = x \quad \text{for all } x \in X. \tag{6.20}$$

The set of singularities E of (a, A) consists only of the point zero and so we only have to prove that $S(t)$, $S'(t)A^{-1}$, and $V(t) = 1 * (a * S)(t)$ are bounded (existence of $S(t)$ follows, e.g., from the paper of Grimmer and Prüss [18] since $a(0+) > 0$ and $a(t)$ is smooth). Let $x \in D(A)$ and put $u(t) = V(t)x$; then it is easy to see that $u(t)$ satisfies

$$u'' = Au' - u + x, \qquad u(0) = u'(0) = 0. \tag{6.21}$$

Take the inner product of (6.21) with $u'(t)$ and integrate to the result

$$|u'(t)|^2 + |u(t) - x|^2 \leq |x|^2 + 2\int_0^t (Au'(s), u'(s))ds \leq |x|^2,$$

since A is dissipative and $u(0) = u'(0) = 0$. But this means

$$|(a * S)(t)x|^2 + |V(t)x - x|^2 \leq |x|^2, \qquad t > 0, \tag{6.22}$$

i.e., $V(t)$ and $(a*S)(t)$ are both bounded. Similarly, $u(t) = S(t)x$, $x \in D(A^2)$ satisfies (6.21) with initial values $u(0) = x$ and $u'(0) = Ax$; therefore, the same argument yields

$$|S'(t)x|^2 + |S(t)x - x|^2 \leq |Ax|^2, \qquad t > 0, \tag{6.23}$$

i.e., $S'(t)A^{-1}$ is bounded by 1 and so $S(t) = S'(t)A^{-1} + V(t)$ is bounded as well.

7. Applications to viscoelasticity. Let $\Omega \subset \mathbb{R}^n$ be a domain with compact and smooth boundary $\partial\Omega$ that is occupied by a linear incompressible viscoelastic fluid. Assuming the fluid at rest for $t \geq 0$, its velocity field $u(t,x)$ is governed for $t > 0$ by the following problem

$$(7.1)\quad \begin{cases} u_t(t,x) = \int_0^t \Delta u(t-\tau,x)da(\tau) - \nabla p(t,x) + g(t,x) & \\ (\nabla \circ u)(t,x) = 0 & \text{for } x \in \Omega, \quad t > 0, \\ u(t,x) = 0 & \text{for } x \in \partial\Omega, \quad t > 0, \\ u(0,x) = u_0(x) & \text{for } x \in \Omega. \end{cases}$$

Here $p(t,x)$ denotes the (also unknown) hydrostatic pressure; $g(t,x)$ a (given) external force field, $u_0(x)$ the (given) initial velocity field (induced by a δ-perturbation at time $t = 0$); $\Delta, \nabla, \nabla\circ$ designate the Laplacian, gradient, divergence with respect to the x-variables, respectively. The *stress relaxation modulus* $da(t)$ of a linear viscoelastic material is of the general form

$$(7.2)\qquad a(t) = a_0 + a_\infty t + \int_0^t a_1(s)ds, \qquad t \geq 0,$$

where $a_0, a_\infty \geq 0$ are constants, and $a_1(t) \geq 0$ is nonincreasing and of positive type, $\lim_{t\to\infty} a_1(t) = 0$; for a viscoelastic fluid we even have $a_\infty = 0$ and $a_1 \in L^1(\mathbb{R}_+)$.

For the derivation of (7.1), the properties of the kernel $da(t)$, and more on the physical background of viscoelasticity, we refer to the monographs of Christensen [9], Renardy, Hrusa, and Nohel [35], and Pipkin [31].

Equation (7.1) can be rewritten as an abstract Volterra equation in a Banach space X of the form (1.1), i.e.,

$$(7.3)\qquad u(t) = \int_0^t a(t-\tau)Au(\tau)d\tau + f(t), \qquad t \geq 0,$$

where A denotes a closed linear operator in X with dense domain $D(A)$ and $f \in C(\mathbb{R}_+, X)$. In fact, we may choose $X = L_0^2(\Omega;\mathbb{R}^n)$, the space of all divergence-free L^2-vector fields, $A = P\Delta$, the Stokes operator with $D(A) = W^{2,2}(\Omega;\mathbb{R}^n) \cap W_0^{1,2}(\Omega;\mathbb{R}^n) \cap X$ (P denotes the Helmholtz projection in $L^2(\Omega;\mathbb{R}^n)$) and $f : \mathbb{R}_+ \to X$ is defined by $f(t) = u_0 + \int_0^t g(s)ds$. It is well known that the Stokes operator is self-adjoint and negative semidefinite, and hence gives rise to a bounded cosine family in X.

For the Helmholtz projection and the properties of the Stokes operator mentioned above, as well as others, we refer to the paper by Giga and Sohr [17], and to the monograph of Temam [39].

Existence of the resolvent in the general case relevant for the theory of viscoelasticity was first obtained in a Hilbert space setting by Carr and Hannsgen [7].

PROPOSITION 7.1. *Let X be a Hilbert space, A self-adjoint and negative semidefinite and let $a(t)$ be of the form* (7.2) *with $a_0, a_\infty \geq 0$, $a_1(t) \geq 0$ nonincreasing and of positive type with $a_1 \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ and $\lim_{t\to\infty} a_1(t) = 0$. Then* (7.3) *admits a resolvent $S(t)$ such that $|S(t)| \leq 1$ on $\mathbb{R}_+$.*

Actually, Carr and Hannsgen assumed in addition that $a_1(t)$ is convex; however, for existence this is not needed. The proof of Proposition 7.1 relies on the spectral decomposition of self-adjoint operators in Hilbert spaces and estimates on the solutions

$s(t;\mu)$ of the scalar equations

$$s(t)+\mu\int_0^t a(t-\tau)s(\tau)d\tau = 1, \quad t\geq 0, \quad \mu \geq 0. \tag{7.4}$$

A different approach was introduced in Prüss [32].

PROPOSITION 7.2. *Let X be a Banach space, A the generator of a bounded cosine family $C(t)$ in X, $a(t)$ of the form* (7.2) *with $a_0, a_\infty \geq 0$, $a_1 \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$, $a_1(t)\geq 0$ nonincreasing and $\log a_1(t)$ convex, $\lim_{t\to\infty} a_1(t) = 0$. Then* (7.3) *is governed by a bounded resolvent $S(t)$.*

The proof of this result is based on the complete monotonicity of the functions $h(\lambda,\tau) = \exp(-\tau/\hat{a}(\lambda)^{1/2})/(\lambda\hat{a}(\lambda)^{1/2})$ with respect to $\lambda > 0$, for each fixed $\tau \geq 0$, on the representation formula

$$\hat{S}(\lambda) = \int_0^\infty C(\tau)h(\lambda,\tau)d\tau, \qquad \lambda > 0, \tag{7.5}$$

and on the generation theorem for resolvents due to Da Prato and Iannelli [13] and Grimmer and Prüss [18].

Here we are interested in the asymptotic behavior of the resolvent $S(t)$. Before we quote some known results, let us introduce the following definition.

DEFINITION 7.3. Suppose (7.3) admits a resolvent $S(t)$.

(i) Equation (7.3) is called *uniformly asymptotically stable* if there is $\varphi \in L^1(\mathbb{R}_+)\cap C_0(\mathbb{R}_+)$ such that $|S(t)| \leq \varphi(t)$ on $\mathbb{R}_+$.

(ii) Equation (7.3) is called *asymptotically stable* if $S(t)x \to 0$ as $t\to\infty$ for each $x\in X$.

Carr and Hannsgen [7] obtained the following result.

THEOREM 7.4. *Let the assumptions of Proposition* 7.1 *be satisfied, and assume in addition that $a_1 \in C^1(0,\infty)$ and that $-\dot{a}_1(t)$ is nonincreasing and convex. If A is invertible and $a(t) \not\equiv a_\infty t$, then* (7.3) *is uniformly asymptotically stable.*

Observe that A must necessarily be invertible if (7.3) is uniformly asymptotically stable. In fact, if $0\in\sigma(A)$, then $|\mu(\mu-A)^{-1}| \geq 1$ for each $\mu\in\varrho(A)$; on the other hand, $S(\cdot)x \in L^1(\mathbb{R}_+, X)$ for each $x\in X$ implies that $\hat{S}(\lambda) = (1/\lambda)(I-\hat{a}(\lambda)A)^{-1}$ is uniformly bounded for $\operatorname{Re}\lambda \geq 0$, i.e., $M \geq |\hat{S}(\lambda)| \geq 1/|\lambda|$ which is impossible. Also $a(t)\not\equiv a_\infty t$ is necessary for uniform asymptotic stability, since otherwise $S(t) = C(\sqrt{a_\infty}t)$ where $C(t)$ denotes the cosine family generated by A; but cosine families are never integrable.

There is a similar result for the situation of Proposition 7.2; see Prüss [32].

THEOREM 7.5. *Let the assumptions of Proposition* 7.2 *be satisfied. Then* (7.3) *is uniformly asymptotically stable if and only if $a(t)\not\equiv a_\infty t$ and A is invertible.*

In the case $A = P\Delta$, the Stokes operator in $L^2_0(\Omega;\mathbb{R}^n)$, A is invertible if the domain $\Omega\subset\mathbb{R}^n$ is bounded; thus Theorems 7.4 and 7.5 show that viscoelastic fluids with (sufficiently) convex stress relaxation moduli are always uniformly asymptotically stable, unless they are purely elastic, i.e., $da(t) = a_\infty dt$.

However, for unbounded domains Ω, the operator $A = P\Delta$ will in general not be invertible and so (7.3) is not uniformly asymptotically stable. As a consequence of our General Convergence Theorem and the results in §7, in this situation (7.3) will still be asymptotically stable, as the next theorems show.

THEOREM 7.6. *Let the assumption of Proposition* 7.2 *be satisfied, and assume in addition $a(t)\not\equiv a_\infty t$ and $N(A)^\perp \cap N(A') = \{0\}$. Then $\lim_{t\to\infty} S(t)x = Px$ for each $x\in X$, where P denotes the projection onto $N(A)$ along $\overline{R(A)}$.*

Proof. (a) By Proposition 7.2 there is a resolvent $S(t)$ for (7.3) which is bounded on $\mathbb{R}_+$. Since log-convex functions are convex, we see that Corollaries 6.3 and 6.5 apply; hence, $\varrho(a) \supset i\mathbb{R}\backslash\{0\}$ and (H2) holds for $\mu \neq 0$, and also (H3) is satisfied.

(b) We next compute the set E of singularities of $\hat{S}(\lambda)$. Since A generates a bounded cosine family, we have $\sigma(A) \subset (-\infty, 0]$. Convex functions are of positive type; hence, $\hat{a}(\lambda) \notin (-\infty, 0]$ for Re $\lambda > 0$. Therefore, $E \subset \{0\}$ will follow if we show that Im $\hat{a}(i\mu) \neq 0$ for $\mu \in \mathbb{R}$, $\mu \neq 0$. Since $ta_1(t) \leq \int_0^t a_1(\tau)d\tau \to 0$ as $t \to 0$, via an integration by parts, we obtain with $\mu > 0$,

$$(7.6) \qquad -\mu \mathrm{Im}\ \hat{a}(i\mu) = a_0 + \ \mathrm{Re}\ \hat{a}_1(i\mu) = a_0 + \mu^{-1}\int_0^\infty (-\dot{a}_1(t))\sin(\mu t)dt \geq 0,$$

since $a_1(t)$ is nondecreasing and convex (hence also absolutely continuous on $(0,\infty)$). Equality in (7.6) can only hold in case $a_0 = 0$, and $-\dot{a}_1(t)$ is constant on each of the intervals $(2k\pi\mu^{-1}; 2(k+1)\pi\mu^{-1})$; but this cannot happen since $a_1(t)$ is log-convex by assumption and is nontrivial, for otherwise, $a(t) \equiv a_\infty t$. Thus, $E \subset \{0\}$ holds.

(c) We next show $0 \in \varrho(a)$ and (H2) for $\mu = 0$. This will be done with the help of the following result.

LEMMA 7.7. *Let $a(t)$ satisfy the assumptions of Proposition* 7.2 *and define* $g(\lambda) = \hat{a}(\lambda)^{-1/2}$ *for* Re $\lambda \geq 0$. *Then there are* k, $\ell \in BV(\mathbb{R}_+)$ *such that*

$$(7.7) \qquad \widehat{dk}(\lambda) = \frac{g(\lambda)}{1+g(\lambda)}, \quad \widehat{d\ell}(\lambda) = \widehat{dk}(\lambda)g(\lambda)/\lambda, \quad \mathrm{Re}\ \lambda \geq 0.$$

The proof of Lemma 7.7 is based on Bernstein's theorem and the Wiener–Levy theorem; see Prüss [32, pp. 341–342].

Observe that Lemma 7.7 yields $0 \in \varrho(a)$ and $\hat{a}(0) = \infty$, since $\widehat{dk}(\lambda)$ is continuous on $\overline{\mathbb{C}}_+$, $\hat{a}(\lambda) = (1/\widehat{dk}(\lambda) - 1)^2$ and $\lim_{\lambda \to 0+} \hat{a}(\lambda) = \hat{a}(0) = \infty$.

Now let $U(t) = \int_0^t S(\tau)d\tau$; then for $x \in D(A)$, we have

$$(UAx)^\wedge(\lambda) = \lambda^{-1}\hat{S}(\lambda)Ax = \lambda^{-2}(I - \hat{a}(\lambda)A)^{-1}Ax = \lambda^{-2}g(\lambda)^2(g(\lambda)^2 - A)^{-1}Ax$$

and

$$(S-I)^\wedge(\lambda)x = \hat{a}(\lambda)A\lambda^{-1}(I - \hat{a}(\lambda)A)^{-1} = \lambda^{-1}A(g(\lambda)^2 - A)^{-1}x,$$

as well as

$$(S')^\wedge(\lambda)x = \lambda\hat{S}(\lambda)x - x = \hat{a}(\lambda)A(I - \hat{a}(\lambda)A)^{-1}x = A(g(\lambda)^2 - A)^{-1}x.$$

These relations and the identity

$$\lambda^{-2}g(\lambda)^2 = \lambda^{-1}\widehat{d\ell}(\lambda)(1 + \widehat{dk}(\lambda)) + \widehat{d\ell}(\lambda)^2, \qquad \mathrm{Re}\ \lambda \geq 0,$$

yield

$$\hat{U}(\lambda)Ax = \widehat{d\ell}(\lambda)^2(\hat{S}')^\wedge(\lambda)x + \widehat{d\ell}(\lambda)(1 + \widehat{dk}(\lambda))(S-I)^\wedge(\lambda)x;$$

hence,

$$U(t)Ax = (d\ell * d\ell * S')(t)x + (d\ell + d\ell * dk) * (S-I)(t)x.$$

Since the measures $d\ell$ and dk are bounded, we obtain from the boundedness of $S(t)$ and $S'(t)$ on $D(A)$ the desired bound on $U(t)Ax$, i.e., (H2) holds.

(d) Finally, the assumption $N(A)^{\perp} \cap N(A') = \{0\}$ implies $\lambda \hat{S}(\lambda) \to P$ strongly as $\lambda \to 0+$ by Theorem 4.6; hence, the General Convergence Theorem applies and the proof is complete. □

The proof of Theorem 7.6 shows that boundedness of $U(t)Ax$ is the difficult thing to prove. This turns out to be even more difficult in the situation of Proposition 7.1, where the assumptions on $a_1(t)$ are weaker so that, in general, Bernstein's theorem can no longer be employed. We want to discuss this case now in some detail. So suppose that X is a Hilbert space, A negative semidefinite, and let $a(t)$ be of the form (7.2) with $a_0, a_\infty \geq 0$, $a_1 \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ nonnegative, nonincreasing of positive type, and $\lim_{t\to\infty} a_1(t) = 0$; let us exclude the cosine case $a(t) \equiv a_\infty t$ which has already been discussed in §6.

Proposition 7.1 shows the existence and boundedness of the resolvent $S(t)$, Corollary 6.5 yields the boundedness of $S'(t)$ on $D(A)$. That is, (H3) holds, and since $\lim_{\lambda\to 0+} \hat{a}(\lambda) = \hat{a}(0) = \infty$, we obtain $\lim_{\lambda \to 0+} \lambda \hat{S}(\lambda)x = Px$ for all $x \in X$ where P denotes the orthogonal projection onto $N(A)$. By means of the decomposition $a_1(t) = a_2(t) + a_3(t)$, where

$$(7.8) \qquad a_2(t) = (a_1(t) - a_1(t_0))_+, \quad \text{and} \quad a_3(t) = \min(a_1(t), a_1(t_0))$$

for $t > 0$, $a_2(t) = a_3(t) = 0$ for $t \leq 0$ as before, we obtain

$$(7.9) \qquad \hat{a}(\lambda) = a_0/\lambda + a_\infty/\lambda^2 + \hat{a}_2(\lambda)/\lambda + \widehat{da_3}(\lambda)/\lambda^2, \qquad \mathrm{Re}\,\lambda \geq 0,$$

and therefore, $\varrho(a) \supset i\mathbb{R}\backslash\{0\}$, $\hat{a}(i\mu) \in \mathbb{C}$ for all $\mu \in \mathbb{R}$, $\mu \neq 0$.

Since $a_1(t)$ is nonincreasing, it follows that for $\mu \neq 0$

$$\mu^2 \mathrm{Re}\ \hat{a}(i\mu) = -a_\infty + \mu \mathrm{Im}\ \hat{a}_1(i\mu) \leq 0$$

and even strictly if $a_1(t)$ is also continuous on $(0, \infty)$ or in case $a_\infty > 0$. On the other hand, we have for $\mu \neq 0$

$$-\mu \mathrm{Im}\ \hat{a}(i\mu) = a_0 + \mathrm{Re}\ \hat{a}_1(i\mu) \geq 0,$$

since $a_1(t)$ is of positive type and even strictly if $a_0 > 0$. Therefore, we have

$$E_0 = E\backslash\{0\} = \{\mu \in \mathbb{R}\backslash\{0\} : \ \mathrm{Re}\ \hat{a}_1(i\mu) = -a_0, \hat{a}(i\mu)^{-1} \in \sigma(A) \text{ or } \hat{a}(i\mu) = 0\}$$

Thus, the spectral assumption (H1) reduces to

$$(7.10) \qquad E_0 \text{ is at most countable, and } \mu \in E_0 \text{ implies } \hat{a}(i\mu)^{-1} \notin \sigma_p(A).$$

Observe that $E_0 = \emptyset$ if $a_0 > 0$ or if $\mathrm{Re}\ \hat{a}_1(i\mu) \neq 0$ for all $\mu \neq 0$.

By Proposition 6.2 and Corollary 6.3, it is also not difficult to verify (H2) for $\mu \in E_0$; in fact, either $-\int_0^\infty t da_1(t) < \infty$ or a_1 convex will be sufficient; note, however, that the former is equivalent to $a_1 \in L^1(\mathbb{R}_+)$ since $a_1(t) \geq 0$ is nonincreasing.

We turn now to the question whether $0 \in \varrho(a)$ and whether $U(t)Ax = \int_0^t S(\tau)Ax\,d\tau$ is bounded on $D(A)$. The first two cases will be a consequence of Proposition 6.6.

Case 1. $a_\infty > 0$ (a "solid"). This one is easy. In fact, if $a_\infty > 0$, then $g_1(\lambda) = (\lambda^2 \hat{a}(\lambda))^{-1}$ is bounded and completely monotonic for $\lambda > 0$, since $a_1(t)$ is nonincreasing. Therefore, by Bernstein's theorem there is a function $\ell \in BV(\mathbb{R}_+)$

such that $g_1(\lambda) = \widehat{d\ell}(\lambda)$ for $t > 0$. Proposition 6.6 then implies $0 \in \varrho(a)$, $\hat{a}(0) = \infty$ and boundedness of $U(t)Ax$ on $D(A)$.

Case 2. $a_\infty = 0$, $a_0 > 0$, $a_1 \in L^1(\mathbb{R}_+)$ ("viscous fluid"). Here we use

$$g_2(\lambda) = \frac{1}{\lambda\hat{a}(\lambda)} = \frac{1}{a_0 + \hat{a}_1(\lambda)} = a_0^{-1}\left(1 - \frac{\hat{a}_1(\lambda)}{a_0 + \hat{a}_1(\lambda)}\right), \qquad \text{Re}\,\lambda \geq 0.$$

Since $a_1 \in L^1(\mathbb{R}_+)$ is of positive type, Re $\hat{a}_1(\lambda) \geq 0$ for Re $\lambda \geq 0$, and so $a_0 + \hat{a}_1(\lambda)$ does not vanish on $\overline{\mathbb{C}}_+$. By the Paley–Wiener theorem there is a function $r \in L^1(\mathbb{R}_+)$ such that

$$g_2(\lambda) = a_0^{-1}(1 - \hat{r}(\lambda)), \qquad \text{Re}\,\lambda \geq 0,$$

and so assumption (a) of Proposition 6.6 is satisfied; therefore, $0 \in \varrho(a)$, $\hat{a}(0) = \infty$, and $U(t)Ax$ is bounded on $D(A)$.

Case 3. $a_\infty = a_0 = 0$, $a_1 \in L^1(\mathbb{R}_+)$ (a "rigid fluid"). We assume in addition that a_1 is absolutely continuous on $(0, \infty)$ in this case. As before, decompose $a_1(t) = a_2(t) + a_3(t)$, where a_2, a_3 are as in (7.8) and $t_0 > 0$ is small enough for $\alpha = a_3(0+) = a_1(t_0) > 0$. Since

$$S(t)x - x = (a_1 * UAx)(t)$$

and

$$S'(t)x - (a_2 * SAx)(t) = (da_3 * UAx)(t) = \alpha U(t)Ax + (\dot{a}_3 * UAx)(t),$$

we obtain

$$\hat{U}(\lambda)Ax = (\alpha + \hat{a}_1(\lambda) + (\dot{a}_3)^\wedge(\lambda))^{-1}(S(t)x - x + S'(t)x - (a_2 * SAx)(t))^\wedge(\lambda).$$

By boundedness of $S(t)x$, $S'(t)x$, and $S(t)Ax$ on $D(A)$, and since $a_2 \in L^1(\mathbb{R}_+)$, it is sufficient to show that

$$g_3(\lambda) = (\alpha + \hat{a}_1(\lambda) + (\dot{a}_3)^\wedge(\lambda))^{-1} = (\alpha + \hat{b}(\lambda))^{-1} = \alpha^{-1}\left(1 - \frac{\hat{b}(\lambda)}{\alpha + \hat{b}(\lambda)}\right)$$

is the Laplace transform of a bounded measure.

By assumption, Re $\hat{a}_1(\lambda) \geq 0$ for Re $\lambda > 0$ and $\hat{a}_1(0) = \int_0^\infty a_1(t)dt > 0$; on the other hand, $|(\dot{a}_3)^\wedge(\lambda)| \leq a_3(0+) = \alpha$ since $\dot{a}_3(t) \leq 0$ on $(0, \infty)$, and equality only holds for $\lambda = 0$. Therefore, $\hat{b}(\lambda) = \hat{a}_1(\lambda) + (\dot{a}_3)^\wedge(\lambda) \neq -\alpha$ for Re $\lambda \geq 0$, and so by the Paley–Wiener theorem there is a function $r \in L^1(\mathbb{R}_+)$, such that

$$g_3(\lambda) = \alpha^{-1}(1 - \hat{r}(\lambda))) = \widehat{dk}(\lambda), \qquad \text{Re}\,\lambda \geq 0,$$

where $k(t) = \alpha^{-1}(1 - \int_0^t r(\tau)d\tau)$ belongs to $BV(\mathbb{R}_+)$. Thus, $U(t)Ax$ is bounded on $D(A)$, $0 \in \varrho(a)$, and $\hat{a}(0) = \infty$ follow in this case easily from (7.9).

Case 4. $a_\infty = 0$, $a_1 \notin L^1(\mathbb{R}_+)$. If a_1 is not integrable then we cannot apply the Paley–Wiener theorem directly to show that the functions $g_j(\lambda)$ in Case 2 and Case 3 above are Laplace transforms of bounded measures. However, it is enough to know that for every $\alpha > 0$ there is $r_\alpha \in L^1(\mathbb{R}_+)$ such that

$$\hat{r}_\alpha(\lambda) = \frac{\hat{a}_1(\lambda)}{\alpha + \hat{a}_1(\lambda)}, \qquad \text{Re}\,\lambda \geq 0,$$

holds. Obviously, this is enough in case $a_0 > 0$; put $\alpha = a_0$ to see this. If $a_0 = 0$, rewrite $g_3(\lambda)$ as

$$g_3(\lambda) = \alpha^{-1}\left(1 - \frac{\hat{r}_\alpha(\lambda) + (\dot{a}_3)^\wedge(\lambda)\hat{dk}_\alpha(\lambda)}{1 + (\dot{a}_3)^\wedge(\lambda)\hat{dk}_\alpha(\lambda)}\right)$$

where $\widehat{dk}_\alpha(\lambda) = (\alpha + \hat{a}_1(\lambda))^{-1} = \alpha^{-1}(1 - \hat{r}_\alpha(\lambda))$, and apply Paley–Wiener to this representation.

Shea and Wainger [37] have shown that if in addition $a_1(t)$ is convex such $r_\alpha \in L^1(\mathbb{R}_+)$ exist; see also Jordan, Staffans, and Wheeler [21]. It is clear that then we also have $0 \in \varrho(a)$.

We summarize this in the following theorem.

THEOREM 7.8. *Let the assumptions of Proposition* 7.1 *be satisfied. In addition we assume that one of the following conditions is satisfied:* (a) $a_\infty > 0$*;*

(b) $a_1 \in L^1(\mathbb{R}_+)$*, and either* a_1 *is absolutely continuous on* $(0, \infty)$ *or* $a_0 > 0$*;*

(c) $a_1(t)$ *is convex on* $(0, \infty)$*.*

Moreover, suppose that the spectral condition (7.10) *is satisfied. Then* $\lim_{t\to\infty} S(t)x = Px$ *for each* $x \in X$*, where* P *denotes the orthogonal projection onto* $N(A)$*.*

Finally, we want to mention that for a_1 convex, Re $\hat{a}_1(i\mu) = 0$ if $-\dot{a}_1(t)$ is constant on each of the intervals $(2k\pi\mu^{-1}, 2(k+1)\pi\mu^{-1})$; in particular, $E_0 = \emptyset$ if $-\dot{a}_1(t)$ is nonincreasing and continuous on $(0, \infty)$.

8. Acknowledgment. It is a pleasure for the second author to thank the Equipe de Mathématiques de Besançon for its kind hospitality during his visits.

REFERENCES

[1] G. R. Allan, A. G. O'Farrell, and T. J. Ransford, *A Tauberian theorem arising in operator theory*, Bull. London Math. Soc., 19 (1987), pp. 537–545.

[2] W. Arendt and C. J. K. Batty, *Tauberian theorems and stability of one-parameter semigroups*, Trans. Amer. Math. Soc., 306 (1988), pp. 837–852.

[3] C. J. K. Batty, *Tauberian theorems for the Laplace–Stieltjes transform*, Trans. Amer. Math. Soc., 322 (1990), pp. 783–804.

[4] C. J. K. Batty and V. Q. Phong, *Stability of individual elements under one-parameter semigroups*, Trans. Amer. Math. Soc., 322 (1990), pp. 805–818.

[5] C. J. K. Batty and D. Robinson, *Positive one-parameter semi-groups on ordered spaces*, Acta Appl. Math., 2 (1984), pp. 221–296.

[6] W. Borchers and H. Sohr, *On the semigroup of the Stokes operator for exterior domains*, Math. Z., 196 (1987), pp. 415–425.

[7] R. W. Carr and K. B. Hannsgen, *A nonhomogeneous integrodifferential equation in Hilbert space*, SIAM J. Math. Anal., 10 (1979), pp. 961–984.

[8] S. D. Chatterji, *Tauber's theorem—a few historical remarks*, Jahrb. Überblicke Math., (1984), pp. 167–175.

[9] R. M. Christensen, *Theory of Viscoelasticity*, Academic Press, New York, 1971.

[10] Ph. Clément and J. A. Nohel, *Asymptotic behavior of solutions of Volterra equations with completely positive kernels*, SIAM J. Math. Anal., 12 (1981), pp. 514–535.

[11] T. Coulhon, *Suites d'opérateurs sur un espace $C(K)$ de Grothendieck*, C. R. Acad. Sci. Paris Sér. I Math. 298, (1984), pp. 13–15.

[12] T. Coulhon, *Semi-groupes d'opérateurs et suites de contractions sur les espaces L^∞ et $C(K)$*, Thèse, Université de Paris 6, Paris (1984).

[13] G. Da Prato and M. Iannelli, *Linear integrodifferential equations in Banach spaces*, Rend. Sem. Math. Univ. Padova, C2 (1980), pp. 207–219.

[14] R. Derndinger, R. Nagel, and G. Palm, *Ergodic theory in the perspective of functional analysis*, Lecture Notes, Tübingen, Germany, 1987.

[15] G. Doetsch, *Handbuch der Laplace-Transformation, Band* 1, Birkhäuser, Basel, 1971.

[16] J. Esterle, E. Strouse, and F. Zouakia, *Stabilité asymptotique de certains semigroupes d'opérateurs*, J. Operator Theory, to appear.

[17] Y. Giga and H. Sohr, *On the Stokes operator in exterior domains*, J. Fac. Sci. Sec. IA, 36 (1989), pp. 103–130.

[18] R. Grimmer and J. Prüss, *On linear Volterra equations in Banach spaces*, Comput. Math. Appl., 11 (1985), pp. 189–205.

[19] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, Amer. Math. Soc. Colloq. Publ. 31, Providence, RI, 1957.

[20] A. E. Ingham, *On Wiener's method in Tauberian theorems*, Proc. London Math. Soc. (2), 38 (1933), pp. 458–480.

[21] C. G. Jordan, O. J. Staffans, and R. L. Wheeler, *Local analyticity on weighted spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749–782.

[22] Y. Katznelson and L. Tzafriri, *On power bounded operators*, J. Funct. Anal., 86 (1986), pp. 313–328.

[23] A. Kishimoto and D. W. Robinson, *Subordinate semigroups and order properties*, J. Austral. Math. Soc. Ser. A, 31 (1981), pp. 59–76.

[24] H. König, *Neuer Beweis eines klassischen Tauber–Satzes*, Arch. Math., XI (1960), pp. 278–279.

[25] J. Korevaar, *On Newman's quick way to the prime number theorem*, Math. Intelligencer, 4 (1982), pp. 108–115.

[26] H. P. Lotz, *Uniform convergence of operators on L^∞ and similar spaces*, Math. Z., 190 (1985), pp. 207–220.

[27] H. P. Lotz, *Tauberian theorems for operators on L^∞ and similar spaces. Functional analysis. Functional analysis, surveys and recent results* III, K. D. Bierstedt and B. Fuchssteiner, eds., North Holland, Amsterdam, 1984.

[28] Y. I. Lyubich and V. Q. Phong, *Asymptotic stability of linear differential equations in Banach spaces*, Studia Math., 88 (1988), pp. 37–42.

[29] R. Nagel, *One-parameter Semigroups of Positive Operators.* Lecture Notes in Math. 1184, Springer–Verlag, Berlin, 1986.

[30] D. J. Newman, *Simple analytic proof of the prime number theorem*, Amer. Math. Monthly, 87 (1980), pp. 693–696.

[31] J. Pipkin, *Lectures on Viscoelasticity Theory*, Springer–Verlag, Berlin, 1972.

[32] J. Prüss, *Positivity and regularity of hyperbolic Volterra equations in Banach spaces*, Math. Ann., 279 (1987), pp. 317–344.

[33] J. Prüss, *Linear Evolutionary Integral Equations in Banach Spaces and Applications*, Birkhäuser Verlag, Basel, to be published.

[34] T. J. Ransford, *Some quantitative Tauberian theorems for power series.* Bull. London Math. Soc., 20 (1988), pp. 37–44.

[35] M. Renardy, W. J. Hrusa, and J. A. Nohel, *Mathematical Problems in Viscoelasticity*, Longman, Harlow, Essex, 1987.

[36] M. M. Schaefer, *Banach Lattices and Positive Operators*, Springer–Verlag, Berlin, 1974.

[37] D. F. Shea and S. Wainger, *Variants of the Wiener–Levy theorem, with applications to problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312–343.

[38] A. Tauber, *Ein satz aus der Theorie der unendlichen Reihen*, Monatsh. Math., 8 (1897), pp. 273–277.

[39] R. Temam, *The Navier–Stokes Equation*, North Holland, Amsterdam, 1975.

[40] E. C. Titchmarsh, *The Theory of Functions*, Oxford University Press, Oxford, 1932.

[41] K. Yosida, *Functional Analysis*, Springer–Verlag, Berlin, 1980.

[42] W. von Wahl, *The Equations of Navier–Stokes and Abstract Parabolic Equations*, Vieweg, Braunschweig, Germany, 1985.

[43] D. V. Widder, *An Introduction to Transform Theory*, Academic Press, New York, 1971.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 449-481, March 1992

008

RICCATI DIFFERENTIAL EQUATIONS WITH UNBOUNDED COEFFICIENTS AND NONSMOOTH TERMINAL CONDITION—THE CASE OF ANALYTIC SEMIGROUPS*

I. LASIECKA† AND R. TRIGGIANI†

Abstract. This paper provides, constructively, an explicit solution to the (operator) Differential Riccati Equation in Hilbert space with unbounded coefficients on a fixed time interval $[0, T)$, $T<\infty$, which arises in the optimal control problem with nonsmoothing terminal condition at $t=T$ for an abstract dynamics modeled by an analytic semigroup. The results are sharp as illustrated by counterexamples. Regularity properties of all the quantities involved are also given. Uniqueness of the solution is asserted under some additional assumptions on the terminal condition. Applications include parabolic equations with Dirichlet, or Neumann (Robin) boundary control, or else with point control, as well as plate-like equations with a high degree of damping, etc.

Key words. Riccati Differential Equations, analytic semigroups

AMS(MOS) subject classifications. 47, 35, 93

1. Introduction: statement of main results; literature. Consider the following abstract differential equation,

$$\dot{y}=Ay+Bu \quad \text{on, say, } [\mathscr{D}(A^*)]', \quad y(0)=y_0\in Y, \tag{1.1}$$

subject to the following assumptions to be maintained throughout the article.

(i) A is the infinitesimal generator of a strongly continuous analytic semigroup, denoted by e^{At} on the Hilbert space Y. Without loss of generality for the problem considered here, where the dynamics (1.1) is studied over a finite interval $[0, T]$, $T<\infty$, we may assume that A is boundedly invertible, i.e., $A^{-1}\in\mathscr{L}(Y)$. Then the fractional powers $(-A)^\theta$, $0<\theta<1$ are well defined.

(ii) B is a linear (generally unbounded from U to Y) but continuous operator: $U=\mathscr{D}(B)\to[\mathscr{D}(A^*)]'$, where U is another Hilbert space, such that

$$A^{-\gamma}B\in\mathscr{L}(U;Y) \quad \text{or} \quad \|A^{-\gamma}B\|_{\mathscr{L}(U;Y)}=\|B^*A^{*-\gamma}\|_{\mathscr{L}(Y;U)}\leqq c_\gamma \tag{1.2}$$
$$\text{for some fixed } \gamma, 0\leqq\gamma<1.$$

Generally, dependence on γ will not necessarily be explicitly noted in the sequel. In (1.1), A^* is the Y-adjoint of A, and $[\mathscr{D}(A^*)]'$ denotes the dual space of $\mathscr{D}(A^*)$ with respect to the Y-inner product, so that $|y|_{[\mathscr{D}(A^*)]'}=|A^{-1}y|_Y$.

Remark 1.1. It was pointed out in [L], [T], on the basis of preliminary work in [B1], [B2], [W], that in the case of a second-order parabolic equation defined on a bounded domain $\Omega\subset R^n$ the relevant values of the constant γ are as follows: $\gamma=\frac{3}{4}+\varepsilon$, $\forall\varepsilon>0$, for Dirichlet boundary control with $U=L_2(\Gamma)$, $Y=L_2(\Omega)$; or else, for Neumann boundary control with $U=L_2(\Gamma)$, $Y=H^1(\Omega)$; while $\gamma=\frac{1}{4}+\varepsilon$, for the Neumann boundary control with $U=L_2(\Gamma)$ and $Y=L_2(\Omega)$. For a more detailed discussion of these and other examples of partial differential equations with boundary-point control which fit into the present abstract theory, we refer to [L-T4].

With the dynamics (1.1), we associate the following quadratic functional cost over a preassigned fixed time interval $[0, T]$, $0<T<\infty$:

$$J(u,y)\equiv\int_0^T[|Ry(t)|_W^2+|u(t)|_U^2]\,dt+|Gy(T)|_Z^2, \tag{1.3}$$

* Received by the editors May 9, 1990; accepted for publication (in revised form) January 30, 1991. This research was partially supported by National Science Foundation grant NSF-DMS-8902811 and by Air Force Office of Scientific Research grant AFOSR-87-0321.

† Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia 22903.

where in (1.3), $y(t) = y(t; y_0)$, $R \in \mathscr{L}(Y; W)$, $G \in \mathscr{L}(Y; Z)$, and W, Z are other Hilbert spaces.

The corresponding optimal control problem is:

$$\begin{aligned} &\text{minimize } J(u, y) \text{ over all } u \in L_2(0, T; U), \\ &\text{where } y \text{ is the solution of (1.1) due to } u. \end{aligned} \tag{1.4}$$

The solution to (1.1) is

$$y(t) = e^{At}y_0 + (Lu)(t) \tag{1.5}$$

$$(Lu)(t) = \int_0^t e^{A(t-\tau)}Bu(\tau)\,d\tau \tag{1.6a}$$

$$: \text{continuous } L_2(0, T; U) \to L_2(0, T; \mathscr{D}((-A)^{1-\gamma})). \tag{1.6b}$$

The adjoint L^* of L: $(Lu, f)_{L_2(0,T;Y)} = (u, L^*f)_{L_2(0,T;U)}$ is given by

$$(L^*f)(t) = \int_t^T B^*\, e^{A^*(\tau-t)}f(\tau)\,d\tau \tag{1.7a}$$

$$: \text{continuous } L_2(0, T; [\mathscr{D}((-A)^{1-\gamma})]') \to L_2(0, T; U). \tag{1.7b}$$

Complementing (1.6*a*), we shall let L_T be the (unbounded) operator

$$L_T u = \int_0^Y e^{A(T-t)}Bu(t)\,dt \tag{1.8}$$

with densely defined domain $\mathscr{D}(L_T) = \{u \in L_2(0, T; U)\colon L_T u \in Y\}$, which describes the map (1.5) from the input u to the solution $y(T)$ of (1.1) at time $t = T$, with $y_0 = 0$. Its adjoint L_T^*, $(L_T u, y)_Y = (u, L_T^* y)_{L_2(0,T;Y)}$ is the closed operator

$$\{L_T^* y\}(t) = B^*\, e^{A^*(T-t)}y, \qquad 0 \leqq t \leqq T, \quad y \in Y. \tag{1.9}$$

1.1. Nonsmoothing case. Our main result is the following theorem.

THEOREM 1.1. *Let the* (*densely defined*) *operator GL_T be closed* (*or closeable*), *as an operator $L_2(0, T; U) \supset \mathscr{D}(GL_T) \to Z$. Then there exists a unique optimal pair $\{u^0(t, 0; y_0), y^0(t, 0; y_0)\}$ of problem* (1.1), (1.3), *with $T < \infty$, explicitly given by*

$$-u^0(t, 0; x) = \{\Lambda_{0T}^{-1}[L_T^* G^*\, e^{AT}x + L^*R^*R(e^{A\cdot}x)]\}(t), \tag{1.10}$$

$$y^0(t, 0; x) = e^{At}x + (Lu^0)(t), \tag{1.11}$$

$$\Lambda_{0T} = I + L^*R^*RL + L_T^* G^* GL_T, \tag{1.12}$$

with L, L^*, defined in (1.6), (1.7), *and L_T, L_T^* defined by* (1.8) *and* (1.9). *Moreover, there exists a nonnegative, self-adjoint operator $P(t) = P^*(t) \geqq 0$* (*see* (3.31) *of Proposition* 3.8), *defined explicitly in terms of the data in* (vii) = (1.19) *below, such that*

$$\text{(i)} \qquad P(\cdot) \in \mathscr{L}(Y; C([0, T]; Y)) \tag{1.13}$$

(*see* (2.27) *of Proposition* 2.2 *and comments in the proof thereof*);

(ii) *In fact, even more, for* $0 \leqq \theta < 1$,

$$|(-A^*)^\theta P(t)|_{\mathscr{L}(Y)} \leqq \frac{C_{T\gamma\theta}}{(T-t)^\theta}, \tag{1.14}$$

(*see* (3.26) *of Corollary* 3.7);

(iii) *For all* $0 < \varepsilon \leqq T$,

$$(-A^*)^\theta P(t) \in \mathscr{L}(Y; C([0, T-\varepsilon]; Y)), \qquad 0 \leqq \theta < 1, \tag{1.15}$$

(*see* (4.31) *of Proposition* 4.6);

$$\text{(iv)} \qquad |B^*P(t)|_{\mathscr{L}(Y;U)} \leqq \frac{C_{T\gamma}}{(T-t)^\gamma}, \qquad 0 \leqq t < T \tag{1.16}$$

(*see* (3.27) *of Corollary* 3.7);

(v) *For any* $0 < \varepsilon \leqq T$,

$$B^*P(\,\cdot\,) \in \mathscr{L}(Y; C([0, T-\varepsilon]; Y)) \tag{1.17}$$

(*see* (4.32) *of Proposition* 4.6).

(vi) *For each* $y_0 \in Y$, *the optimal control* $u^0(t, 0; y_0)$ *is given in pointwise feedback form by*

$$u^0(t, 0; y_0) = -B^*P(t)y^0(t, 0; y_0), \qquad 0 \leqq t < T \tag{1.18}$$

(*see* (2.28) *of Proposition* 2.2).

(vii) *The operator* $P(t)$ *is given* (*constructively*) *by*

$$P(t)x = \int_t^T e^{A^*(\tau - t)} R^*Ry^0(\tau, t; x)\,d\tau + e^{A^*(T-t)}G^*Gy^0(T, t; x). \tag{1.19}$$

(viii) *The optimal cost of the optimal control problem on* $[t, T]$ *initiating at time* t *at point* $x \in Y$ *is*

$$J(u^0(\,\cdot\,, t; x), \quad y^0(\,\cdot\,, t; x)) = (P(t)x, x)_Y \tag{1.20}$$

(*see* (3.32) *of Proposition* 3.8);

(ix) *For* $0 < t < T$, $P(t)$ *satisfies the following Differential Riccati Equation* (DRE), *for all* $x, y \in \mathscr{D}((-A)^\varepsilon)$, *for all* $\varepsilon > 0$ (*see Theorem* 4.5),

$$\begin{aligned}(\dot{P}(t)x, y)_Y = &-(R^*Rx, y)_Y - (P(t)x, Ay)_Y - (P(t)Ax, y)_Y \\ &+ (B^*P(t)x, B^*P(t)y)_U.\end{aligned} \tag{1.21}$$

(x) *The following regularity properties hold true for the optimal pair*

$$\|u^0(\,\cdot\,, s; x)\|_{L_2(s,T;U)} + \|y^0(\,\cdot\,, s; x)\|_{L_2(s,T;U)} \leqq c_T\|x\|_Y \tag{1.22}$$

(*see* (2.19), (2.20) *of Proposition* 2.1)

$$\|Gy^0(T, s; x)\|_Z \leqq c_T\|x\|_Y; \tag{1.23}$$

(*see* (2.21) *of Proposition* 2.1)

$$\|u^0(\,\cdot\,, s; x)\|_{C_\gamma([s,T];U)} \leqq c_{T\gamma}\|x\|_Y \tag{1.24}$$

(*see* (3.23) *of Theorem* 3.6)

$$\|y^0(\,\cdot\,, s; x)\|_{C([s,T];Y)} \leqq c_{T\gamma}\|x\|_Y \qquad \text{if } 0 \leqq \gamma < \tfrac{1}{2}; \tag{1.25a}$$

$$\|y^0(\,\cdot\,, s; x)\|_{C_{2\gamma-1+\varepsilon}([s,T];Y)} \leqq c_{T\gamma}\|x\|_Y \qquad \text{if } \tfrac{1}{2} \leqq \gamma < 1. \tag{1.25b}$$

(*see* (3.24) *of Theorem* 3.6).

In (1.24) *and* (1.25), *if* X *is a Hilbert space and* r *any real number*, $C_r([s, T]; X)$ *denotes the Banach space defined by*

$$\begin{aligned} C_r([s, T]; X) = \Big\{ & f(t) \in C([s, T); X): \\ & \|f\|_{C_r([s,T];X)} = \sup_{s \leqq t < T} (T-t)^r \|f(t)\|_X < \infty \Big\}. \end{aligned} \tag{1.26}$$

Moreover (*see Theorem* 4.1), *for* $x \in Y$ *and for each s fixed*, $0 \leqq s < T$, *the optimal control* $u^0(t, s; x)$ *and the optimal solution* $y^0(t, s; x)$ *are, respectively, U-valued and Y-valued functions that are differentiable in* $t \in (s, T)$ *with* $(\partial u^0/\partial t)(t, s; x) \in U$, $(\partial y^0/\partial t)(t, s; x) \in Y$. *In fact, these U-valued and Y-valued functions* $u^0(t, s; x)$ *and* $y^0(t, s; x)$ *are analytic in* $t \in (s, T)$ *if the operator A has compact resolvent in Y, or* $(-A)^p B$ *is compact, for some* $0 < p < 1$.

(xi) *If we define the evolution operator*

$$\Phi(t, \tau)x = y^0(t, \tau; x), \tag{1.27}$$

the following weak convergence results hold true (*see* (5.3) *and* (5.4) *of Proposition* 5.2):

$$\lim_{t \uparrow T} (G\Phi(T, t)x, z)_Z = (Gx, z) \quad \forall x \in X, \forall z \in Z; \tag{1.28}$$

$$\lim_{t \uparrow T} (P(t)x, y)_Y = (G^*Gx, y) \quad \forall x, y \in Y. \tag{1.29}$$

Remark 1.2. With reference to the assumption of Theorem 1.1, we have

(1.30) closed operator $(GL_T)^*$, written as $L_T^* G^*$, be *densely defined* as an operator $Z \supset \mathscr{D}((GL_T)^*) \rightarrow L_2(0, T; U)$ $\Leftrightarrow$ densely defined operator GL_T be *closeable* as an operator $L_2(0, T; U) \supset \mathscr{D}(GL_T) \rightarrow Z$

$$\Uparrow$$

(1.31) $(-A^*)^{\beta/2{}^*_G}$ be *densely defined* as an operator $Z \supset \mathscr{D}((-A^*)^{\beta/2} G^*) \rightarrow Y$ for some $\beta > 2\gamma - 1$

The equivalence is a standard result [K1, p. 168]. To see the sufficient condition, we compute from (1.9),

$$\begin{aligned} \{L_T^* G^* z\}(t) &= B^* e^{A^*(T-t)} G^* z \\ &= B^*(-A^*)^{-\gamma}(-A^*)^{\gamma-\beta/2} e^{A^*(T-t)}(-A^*)^{\beta/2} G^* z, \end{aligned} \tag{1.32}$$

use (1.2), and notice that $(-A^*)^{\gamma-\beta/2} e^{A^*(T-t)} \in \mathscr{L}(Y; L_2(0, T; Y))$ for $2\gamma - \beta < 1$.

We emphasize that condition (1.31) on $(-A^*)^{\beta/2} G^*$, which does not involve B, is *only sufficient* for the ultimate requirement that GL_T be closeable, which instead involves B. This will be seen in one example in § 7.2 below.

Remark 1.3. An example in § 7.1 will show that the assumption that GL_T be closed (closeable) cannot be dispensed with, for otherwise the optimal control may not exist.

1.2. The smoothing case. Next we shall assume that G is a smoothing operator in the sense that

$$(-A^*)^\beta G^* G \in \mathscr{L}(Y) \quad \text{for some } \beta > 2\gamma - 1 \tag{1.33}$$

(which is automatically satisfied with $\beta = 0$ if $0 \leqq \gamma < \frac{1}{2}$). Then, equation (1.33) implies that GL_T is, in fact, bounded: $GL_T \in \mathscr{L}(L_2(0, T; U), Z)$ (so that a fortiori the assumption of Theorem 1.1 is satisfied). To justify this claim, a useful result from [F1, Lemma 3.1] is invoked (which will be recalled in its entirety at the beginning of § 6), which says, in particular, that the operator $(-A^*)^{\beta/2} G^* G(-A)^{\beta/2-\varepsilon}$ admits a bounded extension in $\mathscr{L}(Y)$, $\varepsilon > 0$. Thus, $(-A^*)^{\beta/2-\varepsilon} G^* G(-A)^{\beta/2-\varepsilon}$ is self-adjoint and in $\mathscr{L}(Y)$. Then, $(-A^*)^{\beta/2-\varepsilon} G^* \in \mathscr{L}(Z, Y)$. Returning to (1.32) and the line below it, we see that $L_T^* G^*$ is bounded: $L_T^* G^* \in \mathscr{L}(Z, L_2(0, T; U))$), as desired, since $(-A^*)^{\gamma-\beta/2+\epsilon} e^{A^*(T-t)} \in \mathscr{L}(Y; L_2(0, T; Y))$.

Thus, if condition (1.33) is assumed, then, accordingly, stronger results follow; in particular, y^0 becomes continuous. Moreover, the solution of the DRE (1.21) given explicitly by (1.19), is unique and the limits (1.28) and (1.29) as $t \uparrow T$ of Theorem 1.1(xi) are strong. Thus, we recover results of [D-I], which were obtained via the "direct" method.

THEOREM 1.2 *Assume* (1.33). *Then*; *we have the following.*

(i) (*Regularity of optimal pair*). *For* $x \in Y$ *and any* $\varepsilon > 0$,

$$|u^0(\cdot, s; x)|_{C_{1-\gamma-\varepsilon}([s,T];U)} + |y^0(\cdot, s; x)|_{C([s,T];Y)} \leqq c_{T\gamma}|x|_Y, \tag{1.34}$$

$$y^0(T, \cdot\,; x) = \Phi(T, \cdot)x \in C([s, T]; Y), \tag{1.35}$$

(*see* (6.30) *and* (6.31) *of Corollary* 6.3) *from which, in particular,* (*see* (6.32))

$$\lim_{t\uparrow T} \Phi(T, t)x = x, \qquad x \in Y; \tag{1.36}$$

(ii) *For any* $0 \leqq \theta < 1$, $\varepsilon > 0$, $x \in Y$, *we have* $(-A^*)^\theta P(\cdot)x \in C_{\theta+1-2\gamma+\varepsilon}([0, T]; Y)$ (*see* (6.33) *of Corollary* 6.3)

$$|(-A^*)^\theta P(t)|_{\mathscr{L}(Y)} \leqq \frac{C_{T\gamma}}{1-\theta} \frac{1}{(T-t)^{\theta+1-2\gamma+\varepsilon}}, \tag{1.37}$$

(iii) $$B^*P(\cdot) \in \mathscr{L}(Y; C_{1-\gamma-\varepsilon}([0, T]; U));$$

that is,

$$|B^*P(t)x|_U \leqq \frac{C_T}{1-\gamma} \frac{1}{(T-t)^{1-\gamma-\varepsilon}} |x|_Y \tag{1.38}$$

(*see* (6.34) *of Corollary* 6.3);

(1.39) (iv) $$\lim_{t\uparrow T} P(t)x = G^*Gx, \qquad x \in Y$$

(*see* (6.35) *of Corollary* 6.3);

(v) (*uniqueness, see Theorem* 6.4) *the solution* $P(t)$, *given constructively by* (1.19), *of the DRE* (1.21) *and of the terminal condition* (1.39) *is unique within the class of self-adjoint operators* $\bar{P}(t)$ *such that*

$$\begin{array}{ll}(1.40) & \\ (1.41) & \end{array} \quad B^*\bar{P}(t)x \in \begin{cases} C_\gamma([0, T]; U) & \text{if } 0 \leqq \gamma < \frac{1}{2}, \quad \text{where } \beta = 0, \\ C_{1-\gamma-\varepsilon}([0, T]; U) & \text{if } \frac{1}{2} \leqq \gamma < 1, \end{cases}$$

If we assume further smoothing properties on G, we obtain, accordingly, more regular results. In particular, now u^0 becomes continuous as well. We recover results of [F.1], which were obtained via the "direct" approach.

THEOREM 1.3. *Under the assumption* (*which* a fortiori *implies that* $GL_T \in \mathscr{L}(L_2(0, T; U), Z)$)

$$(-A^*)^\gamma G^*G \in \mathscr{L}(Y), \tag{1.42}$$

which is stronger than assumption (1.33) *since* $2\gamma - 1 < \gamma$, *additional regularity results hold true, namely* (*see* (6.54))

(1.43) (i) $$|u^0(\cdot, s; x)|_{C([s,T];U)} \leqq c_T|x|_Y, \qquad x \in Y;$$

(ii) *For any* $0 \leqq \theta < 1$, $A^*P(\cdot) \in \mathscr{L}(Y; C_{\theta-\gamma}([0, T]; Y))$ (*see* (6.55)),

$$|(-A^*)^\theta P(t)|_{\mathscr{L}(Y)} \leqq \frac{C_T}{1-\theta} \frac{1}{(T-t)^{\theta-\gamma}}; \tag{1.44}$$

(1.45) (iii) $$B^*P(\cdot) \in \mathscr{L}(Y; C[0, T]; U) \qquad (\textit{see } (6.56)).$$

1.3. Literature. We shall concentrate here only on the case where e^{At} is analytic and where B is genuinely unbounded, particularly in the mathematically more demanding and physically more interesting range $\frac{1}{2}<\gamma<1$. (In the case $0<\gamma<\frac{1}{2}$, or even $\gamma=\frac{1}{2}$ if the generator A is, say, self-adjoint or normal, the analysis drastically simplifies and it technically reduces to the case $\gamma=0$ where B is then bounded, $B\in\mathscr{L}(U, Y)$: this is so, since then the operator L in (1.6) is continuous $L_2(0, T; U)\to C([0, T]; Y)$[DS], [L1, Appendix A]. This mildly unbounded case $\gamma<\frac{1}{2}$ where, moreover, G is very smoothing, essentially $(A^*)^{-1}$, is treated in [P−S]. A preliminary study via semigroup theory for a parabolic problem with $G=0$ was carried out in [B2]. The presence of the penalization operator G in (1.3) introduces additional genuine difficulties. Qualitatively, the analyticity of e^{At} tends "to compensate" the effects of the unboundedness of B on any interval of the type $[0, T-\varepsilon]$, $\forall\varepsilon>0$ small. Instead, the presence of a *non-smoothing* operator G produces a singularity at $t=T$ for $\{L_T^*G^*G\,e^{AT}x\}(t)=B^*\,e^{A^*(T-t)}G^*Ge^{AT}x$, which occurs in the explicit formula (1.10) for the optimal $u^0(t, 0; x)$. In [L-T1], the optimal control problem (1.4) was studied in several directions (from the regularity of the optimal pair to the synthesis thereof via a Riccati operator) in the case of a general second-order parabolic equation defined on a bounded domain Ω of R^n with Dirichlet boundary control, where the constant γ in (1.2) is $\gamma=\frac{3}{4}+\varepsilon$, $\varepsilon>0$[T], [L]. Moreover, in [L-T1], the operator G was taken to be the identity $G=I$ (with $Z=Y$), certainly a nonsmoothing case. The variational approach (from the optimal control problem to the Riccati equation) introduced in [L-T1] is explicit and constructive in the sense that: first, the optimal pair u^0, y^0 is characterized solely in terms of the data of the problem (see (1.10), (1.11)); next, an operator $P(t)$ is constructed (see (1.19)) in terms of original and optimal evolution, hence, ultimately in terms of the original data of the problem; finally, the operator $P(t)$ is shown to satisfy the DRE and its limiting condition as $t\uparrow T$.

Another approach, in a sense the converse of the first, so-called "direct" (as it proceeds in reverse from a direct study of the well-posedness of the Riccati Equation to the optimal control problem via dynamic programming) is proposed in [F1], [D-I], following [D]. Here, the operator G is taken to be "smoothing" for both the purposes of asserting a unique solution of the Riccati Equation (by local contraction argument and global a priori bound), as well as for the limiting condition as $t\uparrow T$. In these references, smoothing assumptions on G are (1.33) for [D-I] and (1.42) for [F1], in which cases existence and uniqueness of the solution to the DRE is asserted. In a more recent work [F2], the direct study of the Riccati Equation for existence (not for uniqueness) is carried out in the nonsmoothing case $G=$ Identity under the crucial assumption that A be dissipative, which then yields strong convergence of $P(t)$ to G^*G as $t\uparrow T$. An even more recent and almost contemporaneous work on the direct study of the Riccati Equation for existence (not for uniqueness) in the nonsmoothing case for G is [F4]. Instead of assuming that GL_T is closed (closeable)—a natural hypothesis on G in the variational approach of the present paper—[F4] makes the following assumption on G, say in the case $G\in\mathscr{L}(Y, Z)$:

There exists a sequence $G_n\in\mathscr{L}(Y, Z)$ of operators such that (a) there exists $\beta>2\gamma-1$ such that each G_n satisfies the assumption $G_n(-A)^{\beta/2}\in\mathscr{L}(Y, Z)$;

(b) $\{G_n^*G_n\}$ is a nondecreasing family of self-adjoint operators which converges monotonically to G^*G in the sense that as $n\to\infty$:

$$(G_n^*G_nx, x)_Y=\|G_nx\|^2\uparrow(G^*Gx, x)_Y=\|Gx\|_Z^2,\qquad\forall x\in Y. \tag{1.46}$$

Under this assumption (1.46), [F4, Thm. 3.2] shows existence of a solution $P(t)$ of the DRE (1.21) (with $d/dt(P(t)x, y)_Y$ on the left side), which satisfies the regularity properties (1.13) and (1.14), among others, as in Theorem 1.1 above. In addition, [F4] obtains also the strong convergence of $P(t) \to G^*G$ as $t \uparrow T$ (versus weak convergence in (1.29)), because of the postulated monotonic approximation property (b). We shall see in § 7.3 that assumption (1.46) in [F4]—which does not invoke B, only A and G—is stronger than the assumption GL_T closeable of our Theorem 1.1, which involves B.

The question was also raised as to whether the approach of [L-T1] could be extended to include a general nonsmoothing G, not just the identity. This paper is in response to this query.

The present contribution offers the most general treatment to date of the optimal control problem (1.4) for the dynamics (1.1) in the analytic case. This, in particular, includes the following.

(i) The relaxation of regularity assumptions on the (say, bounded) operator G of terminal state penalization (see also Remark 7.3 pointing out the applicability of our treatment—as well as that of [F4]—also to the case where G is unbounded, say $G \in \mathscr{L}(\mathscr{D}((-A)^\rho), Z)$, $\rho > 0$);

(ii) Sharpness of estimates on the behavior of the various relevant quantities u^0, y^0, $B^*P(t)$, $(-A^*)^\theta P(t)$, etc., in the neighborhood of T.

Moreover, the results of our Theorem 1.1—which are based only on the assumption that GL_T be closeable—are sharp: this is confirmed by the classes of counterexamples of § 7.3, where in fact, GL_T is not closeable and the optimal control does not exist.

In the process of restudying the optimal control problem (1.4) for the abstract equation (1.1) with a general G, we also dispense altogether with the assumption that the resolvent of A be compact (which was automatically satisfied and used in [L-T1], and likewise also in [F2]), thus incorporating in the treatment, in particular, Kelvin-Voigt models of plates, where the resolvent is not compact. Moreover, we recover through our variational approach the smoothing cases (1.33) and (1.42) of [D-I] and [F1], which were originally obtained by the direct approach.

We generally follow the variational approach of [L-T1] and incorporate an idea of [D-I] to quantitatively describe the singularity of the various quantities at $t = T$ via the Banach spaces defined in (1.26).

2. Preliminaries.

2.1. Explicit representation formulas for the optimal pair $\{u^0, y^0\}$. Following [L-T1] we shall collect here some preliminary results, culminating with the representation formulas of the optimal pair $\{u^0, y^0\}$. The solution to (1.1) with initial datum $y_s \in Y$ at initial time s, $0 \leqq s \leqq t \leqq T$, is given by (see (1.5), (1.6) for $s = 0$, where $L = L_0$)

$$(2.1)\qquad y(t, s; y_s) = e^{A(t-s)}y_s + (L_s u)(t),$$

$$(2.2)\qquad (L_s u)(t) = \int_s^t e^{A(t-\tau)}Bu(\tau)\, d\tau$$

(2.3a) : continuous $L_2(s, T; U) \to L_2(s, T; \mathscr{D}((-A)^{1-\gamma}))$ with operator norm uniform with respect to $s, 0 \leqq s \leqq T$;

that is

$$(2.3b)\qquad |L_s u|_{L_2(s,T;\mathscr{D}((-A)^{1-\gamma}))} \leqq K_{T\gamma}|u|_{L_2(s,T;U)},$$

where $K_{T\gamma}$ does not depend on s, $0 \leqq s \leqq T$. The continuity expressed by (2.3a, b) is a consequence of the basic assumption (1.2) on B, as well as of the standard result:

$$f \to \int_0^t A\, e^{A(t-\tau)}f(\tau)\, d\tau : L_2(0, T; Y) \to L_2(0, T; Y)$$

[DS], [L; Appendix A]. The adjoint operator $L_s^*:(L_s u, f)_{L_2(s,T;Y)}=(u, L_s^* f)_{L_2(s,T;U)}$ is given by (see (1.7) for $s=0$, where $L^*=L_0^*$)

$$(L_s^* f)(t)=\int_t^T B^* e^{A^*(\tau-t)} f(\tau)\, d\tau, \qquad s \leqq t \leqq T \tag{2.4}$$

(2.5a) : continuous $L_2(s, T; [\mathscr{D}((-A)^{1-\gamma})]') \to L_2(s, T; U)$ with operator norm uniform in $s: 0 \leqq s \leqq T$

(2.5b) : continuous $L_\infty(s, T; [\mathscr{D}((-A)^{\theta})]') \to C([s, T]; U)$ for $\gamma+\theta<1$, with operator norm uniform with respect to $s, 0 \leqq s \leqq T$;

that is,

$$|L_s^* f|_{C([s, T]; U)} \leqq K_{T\gamma} |f|_{L_\infty(s,T;[\mathscr{D}((-A)^\theta)]')}. \tag{2.5c}$$

Other regularity results for L_s and L_s^* will be given in Theorem 3.3 below. We shall also need the (unbounded) operator L_{sT} (see (1.8) for $s=0$, where $L_T=L_{0T}$)

$$L_{sT} u=(L_s u)(T)=\int_s^T e^{A(T-t)} B u(t)\, dt \tag{2.6}$$

with domain $\mathscr{D}(L_{sT})=\{u \in L_2(s, T; U): L_{sT} u \in Y\}$, and its adjoint $L_{sT}^*:(L_{sT} u, y)_Y=(u, L_{sT}^* y)_{L_2(s,T;U)}$ given by (see (1.9) for $s=0: L_T^*=L_{0T}^*$)

$$\{L_{sT}^* y\}(t)=B^* e^{A^*(T-t)} y, \qquad s \leqq t \leqq T, \tag{2.7}$$

which is unbounded from $Y \supset \mathscr{D}(L_{sT}^*)$ into $L_2(s, T; U)$. We note that $H^1(s, T; U) \subset \mathscr{D}(L_{sT})$, so that $\mathscr{D}(L_{sT})$ is dense in $L_2(s, T; U)$ and that L_{sT} is a closed operator. Next, by using the assumption that GL_T is closed (closeable), we shall convert $\mathscr{D}(GL_{sT})$ into a Hilbert space $V(s, T; U)$ equipped with the following inner product

$$(u, v)_{V(s,T;U)}=(u, v)_{L_2(s,T;U)}+(GL_{sT} u, GL_{sT} v)_Z \tag{2.8}$$

for $u, v \in \mathscr{D}(GL_{sT})$. Let $[V(s, T; U)]'$ denote the dual space of $V(s, T; U)$ with respect to $L_2(s, T; U)$ as pivot space:

$$V(s, T; U) \subset L_2(s, T; U) \subset [V(s, T; U)]' \tag{2.9a}$$

with continuous injections (from (2.8)),

$$|u|_{[V(s,T;U)]'} \leqq |u|_{L_2(s,T;U)} \leqq |u|_{V(s,T;U)} \tag{2.9b}$$

We now introduce the (unbounded) operator

$$\Lambda_{sT} \equiv I_s+L_s^* R^* R L_s+L_{sT}^* G^* G L_{sT} \tag{2.10a}$$

(2.10b) : continuous $L_2(s, T; U) \supset \mathscr{D}(\Lambda_{sT}) \to L_2(s, T; Y)$.

Using (2.3a) and (2.10), we readily verify via (2.8) that

$$|(\Lambda_{sT} u, v)_{L_2(s,T;U)}| \leqq M_{T\gamma} |u|_{V(s,T;U)} |v|_{V(s,T;U)}, \tag{2.11}$$

$$(\Lambda_{sT} u, u)_{L_2(s,T;U)} \geqq |u|_{V(s,T;U)}^2, \tag{2.12}$$

with constant $M_{T\gamma}$ independent on $s, 0 \leqq s \leqq T$, by (2.3a, b). Then, by (2.11) and (2.12), the Lax-Milgram theorem applies, and we can extend Λ_{sT} as

(2.13) Λ_{sT}: isomorphism $V(s, T; U)$ onto $[V(s, T; U)]'$

so that in particular,

$$|\Lambda_{sT}^{-1} v|_{V(s,T;U)} \leqq C_T |v|_{[V(s,T;U)]'} \tag{2.14}$$

with constant C_T independent of s, $0 \leqq s \leqq T$. From (2.8), we obtain

$$|GL_{sT}|_{\mathscr{L}(V(s,T;U);Z)} = |L_{sT}^* G^*|_{\mathscr{L}(Z;[V(s,T;U)]')} \leqq 1. \tag{2.15}$$

We now return to the optimal control problem (1.3), except that we shall consider it over the time interval $[s, T]$, with initial time $t = s$, rather than $t = 0$; $0 \leqq s < T$, and initial datum x. We shall call $\{u^0(\cdot, s; x)$ and $y^0(\cdot, s; x)\}$ the corresponding unique optimal pair. By standard minimization approaches (e.g., Lagrange multipliers, or direct computation, etc.), the following explicit characterization of the optimal pair can be derived [L-T1]:

$$-u^0(\cdot, s; x) = L_s^* R^* R y^0(\cdot, s; x) + L_{sT}^* G^* G y^0(T, s; x), \tag{2.16}$$

$$-u^0(\cdot, s; x) = \Lambda_{sT}^{-1}[L_{sT}^* G^* G\, e^{A(T-s)}x + L_s^* R^* R\, e^{A(\cdot - s)}x] \in V(s, T; U), \tag{2.17}$$

where we note that the element in the square bracket in (2.17) belongs, say, to $[V(s, T; U)]'$ so that (2.17) is well defined by (2.13). In going from (2.16) to (2.17) we have used the optimal dynamics

$$y^0(t, s; x) = e^{A(t-s)}x + \{L_s u^0(\cdot, s; x)\}(t) \tag{2.18}$$

for both $y^0(\cdot, s; x)$ and $y^0(T, s; x)$ in (2.16). Note that u^0 in (2.17), and hence y^0 in (2.18), are given explicitly in terms of the data of the problem.

2.2. L_2-estimates for $\{u^0, y^0\}$ and Z-estimate for $Gy^0(T; \cdot\,; x)$.

Proposition 2.1. *With reference to the optimal pair* $\{u^0(\cdot, s; x), y^0(\cdot, s; x)\}$, *we have*

$$\text{(i)} \quad |u^0(\cdot, s; x)|_{L_2(s,T;U)} \leqq |u^0(\cdot, s; x)|_{V(s,T;U)} \leqq C_T |x|_Y; \tag{2.19}$$

$$\text{(ii)} \quad |y^0(\cdot, s; x)|_{L_2(s,T;Y)} \leqq C_T |x|_Y; \tag{2.20}$$

$$\text{(iii)} \quad |Gy^0(T, s; x)|_Z \leqq C_T |x|_Y \tag{2.21}$$

with C_T a generic constant independent of s, $0 \leqq s \leqq T$.

Proof. (i) From (2.17) we compute via (2.9b), (2.14), (2.15), and (2.5a),

$$\begin{aligned}
|u^0(\cdot, s; x)|_{L_2(s,T;U)} &\leqq |u^0(\cdot, s; x)|_{V(s,T;U)} \\
&\leqq C_T |L_{sT}^* G^* G\, e^{A(T-s)}x + L_s^* R^* R\, e^{A(\cdot - s)}x|_{[V(s,T;U)]'} \qquad \text{(by (2.14))} \\
&\leqq C_T\{|G\, e^{A(T-s)}x|_Z + |L_s^* R^* R\, e^{A(\cdot - s)}x|_{L_2(s,T;U)}\} \qquad \text{(by (2.15), (2.9b))} \\
&\leqq C_T |x|_Y \qquad \text{(by (2.5a))}.
\end{aligned} \tag{2.22}$$

(ii) Inequality (2.20) follows then from the optimal dynamics (2.18) via (2.3) and inequality (2.19) established in (i).

(iii) From (2.18) we have for $t = T$ via (2.6),

$$Gy^0(T. s; x) = G\, e^{A(T-s)}x + GL_{sT}u^0(\cdot, s; x). \tag{2.23}$$

Then, inequality (2.21) follows from (2.23) via (2.15) (left) and (2.19) (right), with u^0 in $V(s, T; U)$. □

2.3. Definition of operator $P(t)$ and first properties. Following [L-T1], we next define the operator $P(t) \in \mathscr{L}(Y)$, $0 \leqq t < T$, by

$$P(t)x = \int_t^T e^{A^*(\tau - t)} R^* R y^0(\tau, t; x)\, d\tau + e^{A^*(T-t)} G^* G y^0(T, t; x). \tag{2.24}$$

$P(t)$ in (2.24) is defined explicitly in terms of the data of the problem, since y^0 is also, as remarked below (2.18).

It is convenient to introduce the (evolution) operator $\Phi(t, s)$,

$$\Phi(t, s)x = y^0(t, s; x), \quad x \in Y, \quad 0 \leqq s \leqq t \leqq T, \tag{2.25a}$$

which plainly satisfies

$$\Phi(t, t) = \text{Identity:} \quad \Phi(t, \tau)\Phi(\tau, s) = \Phi(t, s), \quad 0 \leqq s \leqq \tau \leqq t \leqq T. \tag{2.25b}$$

Further properties of $\Phi(\cdot, \cdot)$ will be collected in Lemma 4.3. We rewrite (2.24) via (2.25a) as

$$P(t)x = \int_t^T e^{A^*(\tau - t)} R^* R \Phi(\tau, t)x \, d\tau + e^{A^*(T-t)} G^* G \Phi(T, t)x. \tag{2.26}$$

PROPOSITION 2.2. *With reference to* (2.24), *we have*

$$\text{(i)} \qquad P(t) \in \mathscr{L}(Y; L_\infty(0, T; Y)) \tag{2.27}$$

$$\text{(ii)} \qquad -u^0(t, 0; x) = B^* P(t) y^0(t, 0; x), \quad 0 \leqq t < T; \quad x \in Y. \tag{2.28}$$

Proof. (i) Property (2.27) follows from (2.24) via the regularity properties (2.20) and (2.21) of Proposition 2.1. It may be boosted to $P(t) \in \mathscr{L}(Y; C([0, T]; Y))$, using the properties of Lemma 4.3(iii) (as done in the proof of Proposition 4.6).

(ii) As usual [L-T1], we rewrite (2.16) explicitly by virtue of (2.4), (2.7), and (2.25a),

$$\begin{aligned} -u^0(t, s; x) = \int_t^T & B^* \, e^{A^*(\tau - t)} R^* R \Phi(\tau, s)x \, d\tau \\ & + B^* \, e^{A^*(T-t)} G^* G \Phi(T, s)x. \end{aligned} \tag{2.29}$$

Choosing initial time s equal to t with corresponding initial datum $y^0(t, 0; x) = \Phi(t, 0)x$, we obtain (2.28) from (2.29) via (2.26) and (2.25b). □

We note that, by virtue of (2.28), the optimal dynamics (2.18) can be explicitly rewritten as

$$y^0(t, s; x) = e^{A(t-s)}x - \int_s^t e^{A(t-\tau)} BB^* y^0(\tau, s; x) \, d\tau; \tag{2.30a}$$

$$\Phi(t, s)x = e^{A(t-s)}x - \int_s^t e^{A(t-\tau)} BB^* \Phi(\tau, s)x \, d\tau. \tag{2.30b}$$

3. Pointwise estimates for $u^0(t, s; x)$, $y^0(t, s; x)$, and $P(t)$. Let X be a Hilbert space and let r be a real number. Following [D-I], we introduce the Banach space $C_r([s, T]; X)$ defined by

$$\begin{aligned} C_r([s, T]; X) = \Big\{ & f(t) \in C([s, T); X): \\ & |f|_{C_r([s, T]; X)} = \sup_{s \leqq t < T} (T - t)^r |f(t)|_X < \infty \Big\}. \end{aligned} \tag{3.1}$$

In the interesting case $r > 0$, r measures the singularity of $f(t)$ at $t \to T$. Note that $C_r([s, T]; X) \subset L_q(s, T; X)$ for $rq < 1$.

PROPOSITION 3.1. *With reference to the operators L_s and L_s^* of* (2.2) *and* (2.4), *we have*

(i) *For $r+\gamma<1$, L_s: continuous $C_r([s, T]; U) \to C([s, T]; Y)$:*

$$|L_s u|_{C([s,T];Y)} \leqq \frac{C_{T,\gamma}}{1-(\gamma+r)}(T-s)^{1-(\gamma+r)}|u|_{C_r([s,T];U)}, \tag{3.2}$$

so that the bound in (3.2) *may be made independent of s, $0 \leqq s \leqq T$;*

(ii) *For $r+\gamma \geqq 1$ and $\varepsilon>0$ arbitrary,*

$$L_s\text{: continuous } C_r([s, T]; U) \to C_{r+\gamma-1+\varepsilon}([s, T]; Y);$$

$$|L_s u|_{C_{r+\gamma-1+\varepsilon}([s,T];Y)} \leqq \frac{C_{T\gamma}}{\varepsilon}(T-s)^{\varepsilon}|u|_{C_r([s,T];U)}, \tag{3.3}$$

so that the bound in (3.3) *may be made independent of s, $0 \leqq s \leqq T$;*

(iii) *For $0 \leqq r < 1$,*

$$L_s^*\text{: continuous } C_r([s, T]; Y) \to C_{r+\gamma-1}([s, T]; U),$$

$$|L_s^* f|_{C_{r+\gamma-1}([s,T];U)} \leqq C_{T\gamma} 2^{r+\gamma-1} \max\left\{\frac{1}{1-r}, \frac{1}{1-\gamma}\right\}|f|_{C_r([s,T];Y)}, \tag{3.4}$$

so that the uniform norm is independent of s, $0 \leqq s \leqq T$.

Proof. We first note that $(L_s u)(t) \in C([s, T); Y)$ and $(L_s^* f)(t) \in C([s, T); U)$.

(i), (ii). By (2.2), assumption (1.2) on B, and the analyticity of e^{At},

$$\begin{aligned} |(L_s u)(t)|_Y &= \left|\int_s^t (-A)^\gamma e^{A(t-\tau)}(-A)^{-\gamma} B u(\tau)\, d\tau\right|_Y \\ &\leqq C_{T\gamma} \int_s^t \frac{|u(\tau)|_U (T-\tau)^r}{(t-\tau)^\gamma (T-\tau)^r}\, d\tau \\ &\leqq C_{T\gamma} |u|_{C_r([s,T];U)} \int_s^t \frac{d\tau}{(t-\tau)^\gamma (T-\tau)^r} \qquad \text{(by (3.1)).} \end{aligned} \tag{3.5}$$

For any $r \geqq 0$, we use $(T-\tau)^r \geqq (t-\tau)^r$ in the last integral in (3.5) so that

$$|(L_s u)(t)|_Y \leqq C_{T\gamma}|u|_{C_r([s,T];U)} \frac{(t-s)^{1-(\gamma+r)}}{1-(\gamma+r)}, \tag{3.6}$$

for $r+\gamma<1$ as assumed, and (3.6) yields (3.2) as desired.

For $r+\gamma \geqq 1$, we write using $(T-\tau)^{1-\gamma-\varepsilon} \geqq (t-\tau)^{1-\gamma-\varepsilon}$ and $(T-\tau)^{r+\gamma-1+\varepsilon} \geqq (T-t)^{r+\gamma-1+\varepsilon}$:

$$\begin{aligned} \int_s^t \frac{d\tau}{(t-\tau)^\gamma (T-\tau)^r} &= \int_s^t \frac{d\tau}{(t-\tau)^\gamma (T-\tau)^{1-\gamma-\varepsilon} (T-\tau)^{r+\gamma-1+\varepsilon}} \\ &\leqq \frac{1}{(T-t)^{r+\gamma-1+\varepsilon}} \int_s^t \frac{d\tau}{(t-\tau)^{1-\varepsilon}} \\ &= \frac{(t-s)^\varepsilon}{\varepsilon (T-t)^{r+\gamma-1+\varepsilon}}. \end{aligned} \tag{3.7}$$

Then (3.7) used in (3.5) yields (3.3) as desired, via (3.1).

(iii) By (2.4), assumption (1.2) on B, and analyticity,

$$|(L_s^*f)(t)|_U = \left| \int_t^T B^*(-A^*)^{-\gamma}(-A^*)^{\gamma} e^{A^*(\tau-t)} f(\tau)\, d\tau \right|_U$$

(by (3.1))
$$\leqq C_{T\gamma} \int_t^T \frac{|f(\tau)|_Y (T-\tau)^r}{(\tau-t)^{\gamma}(T-\tau)^r}\, d\tau$$

$$\leqq C_{T\gamma} |f|_{C_r([s,T];Y)}$$
$$\cdot \left\{ \int_t^{t+(T-t)/2} \frac{d\tau}{(\tau-t)^{\gamma}(T-\tau)^r} + \int_{t+(T-t)/2}^T \frac{d\tau}{(\tau-t)^{\gamma}(T-\tau)^r} \right\}$$

(using $T-\tau \geqq (T-t)/2$ in the first integral, and $(\tau-t) \geqq (T-t)/2$ in the second integral with $0 \leqq r < 1$),

$$\leqq C_{T\gamma} |f|_{C_r([s,T];Y)}$$
$$\cdot \left\{ \left(\frac{2}{T-t}\right)^r \left(\frac{T-t}{2}\right)^{1-\gamma} \frac{1}{1-\gamma} + \left(\frac{2}{T-t}\right)^{\gamma} \left(\frac{T-t}{2}\right)^{1-r} \frac{1}{1-r} \right\}$$

$$|(L_s^*f)(t)|_U \leqq C_{T\gamma} 2^{r+\gamma-1} \max\left\{\frac{1}{1-\gamma}, \frac{1}{1-r}\right\} \frac{1}{(T-t)^{r+\gamma-1}} |f|_{C_r([s,T];Y)}, \tag{3.8}$$

and (3.8) yields (3.4) via (3.1), as desired. □

Proposition 3.1 says that L_s and L_s^* are smoothing operators: L_s (respectively, L_s^*) reduces the order of the singularity from r to zero if $r+\gamma<1$, and from r to $r+(\gamma-1+\varepsilon)$ if $r+\gamma \geqq 1$ (respectively, to $r+(\gamma-1)$).

COROLLARY 3.2. *Given* $0<\gamma<1$, *there exists a positive integer* $n_0 = n_0(\gamma)$ *such that for all positive integers* $n \geqq n_0(\gamma)$ *we have*

$$(L_s^* R^* R L_s)^n\text{: continuous } C_{\gamma}([s, T]; U) \to C([s, T]; U) \tag{3.9}$$

$$(L_s^* R^* R L_s)^n v|_{C([s,T];U)} \leqq C_T |v|_{C_{\gamma}([s,T];U)}, \tag{3.10}$$

with uniform norm bound which may be taken independent of s, $0 \leqq s \leqq T$.

Proof. The results of Proposition 3.1 are applied, recursively, with R^* and R bounded. After n_0-iterations, a space $C_r([s,T];U)$ is obtained with $r \leqq 0$. Details are omitted. □

THEOREM 3.3. *With reference to the operators* L_s *and* L_s^* *in* (2.2), (2.4) *we have*:

(3.11) (i) L_s: *continuous* $L_2(s, T; U) \to L_r(s, T; Y)$,

where r is an arbitrary positive number satisfying $r<2/(2\gamma-1)$, *where* $2/(2\gamma-1)>2$ *for* $\frac{1}{2}<\gamma<1$; *for* $0 \leqq \gamma \leqq \frac{1}{2}$ *we may take* $r=\infty$;

(3.12) (ii) L_s^*: *continuous* $L_r(s, T; Y) \to L_{r'}(s, T; U)$,

where r is as in (i), *and* r' *is any positive number satisfying* $r'<2/(4\gamma-3)$, *where* $2/(4\gamma-3)>r$ *for* $\frac{3}{4}<\gamma<1$; *for* $0<\gamma \leqq \frac{3}{4}$, *we may take* $r'=\infty$;

(iii) *For* $p>1/(1-\gamma)$,

$$L_s\text{: continuous } L_p(s, T; U) \to C([s, T]; Y). \tag{3.13}$$

(iv) *Thus, a fortiori, there exists a positive integer* $n_1 = n_1(\gamma)$ *such that*

$$(L_s^* R^* R L_s)^{n_1}\text{: continuous } L_2(s, T; U) \to C([s, T]; U). \tag{3.14}$$

We may take $n_1=1$ *if* $\gamma \leqq \frac{3}{4}$; *and* $n_1=2$ *if* $\frac{3}{4}<\gamma<\frac{5}{6}$; *etc.*

In all cases the operator norm has a bound which may be taken not to depend on s, $0 \leqq s \leqq T$.

Proof. The proof uses Young's inequality. For a similar (but not identical) situation, see, e.g., [L-T2, Lemma 4.2] (or [L-T3, Thm. 2.5]). Details are omitted. □

THEOREM 3.4. *The operator* $[I_s+L_s^*R^*RL_s]$ *is boundedly invertible on the space* $C_\gamma([s, T; U)$ *defined by* (3.1):

$$[I_s+L_s^*R^*RL_s]^{-1}\in\mathcal{L}(C_\gamma([s,T];U)) \tag{3.15}$$

with uniform norm bound which depends on T and γ, *but may be taken not to depend on* s, $0\leqq s\leqq T$.

Proof. Let $h\in C_\gamma([s,T];U)$. We seek a unique $g\in C_\gamma([s,T];U)$ such that

$$g+L_s^*R^*RL_sg=h. \tag{3.16}$$

To simplify the notation, we may take $R=I$ in the argument below.

Step 1. Given such h, if $n_0=n_0(\gamma)$ is the positive integer of Corollary 3.2, equation (3.9), then there exists a unique $v\in L_2(s,T;U)$ such that

$$v+L_s^*L_sv=(L_s^*L_s)^{n_0}h\in C([s,T];U)\subset L_2(s,T;U), \tag{3.17}$$

since $I_s+L_s^*L_s$ is boundedly invertible on $L_2(s,T;U)$.

Step 2. We shall show that, in fact,

$$v\in C([s,T];U). \tag{3.18}$$

In fact, if $n_1=1$ (i.e., $0\leqq\gamma\leqq\frac{3}{4}$) in (3.14) of Theorem 3.3, then $L_s^*L_sv\in C([s,T];U)$ and (3.17) yields (3.18). In general, we write from (3.17), with $r=0,1,\cdots,n_1-1$:

$$(L_s^*L_s)^rv+(L_s^*L_s)^{r+1}v=(L_s^*L_s)^{n_0+r}h\in C([s,T];U), \tag{3.19}$$

where the regularity on the right of (3.19) is a consequence of (3.13) via (3.17) (right). Starting from $r=n_1-1$ in (3.19), we first obtain $(L_s^*L_s)^{r+1}v=(L_s^*L_s)^{n_1}v\in C([s,T];U)$ by (3.14); hence $(L_s^*L_s)^{n_1-1}v\in C([s,T];U)$ by (3.19). Next, using this latter information in (3.19), this time with $r=n_1-2$, leads to $(L_s^*L_s)^{n_1-2}v\in C([s,T];U)$. By repeating this procedure a finite number of times, we arrive at (3.18), as desired.

Step 3. Starting from the given h and the v obtained in (3.17), we shall finally define a finite sequence of vectors called $g_{n_0-1}, g_{n_0-2}, g_{n_0-3},\cdots,g_1, g$, whose last element g will be precisely the sought-after unique solution of (3.16). We define recursively

$$\begin{aligned}
&(3.20_{n_0-1}) &\quad g_{n_0-1}&=(L_s^*L_s)^{n_0-1}h-v &\quad &\in C_\gamma([s,T];U)\\
&(3.20_{n_0-2}) &\quad g_{n_0-2}&=(L_s^*L_s)^{n_0-2}h-g_{n_0-1} &\quad &\in C_\gamma([s,T];U)\\
& & &\ \ \vdots & &\\
& &\quad g_1&=(L_s^*L_s)h-g_2 &\quad &\in C_\gamma([s,T];U)\\
&(3.20_0) &\quad g&=h-g_1 &\quad &\in C_\gamma([s,T];U).
\end{aligned}$$

The regularity noted on the right of (3.20_{n_0-1}) is a consequence of Proposition 3.1 applied to h and of (3.18). Then, recursively, the other regularity statements follow. In particular, $g\in C_\gamma([s,T];U)$. It is now an easy matter to show that such g is the unique sought-after solution of (3.16). Moreover, the bound on the uniform norm in (3.15) may be taken not to depend on s, $0\leqq s\leqq T$, since this is the case for L_s and L_s^* in Proposition 3.1 and Theorem 3.3. Theorem 3.4 is proved. □

The above results are now used to obtain pointwise estimates.

LEMMA 3.5. *With reference to the last term in* (2.16), *we have*

$$|L_{sT}^*G^*Gy^0(T,s;x)|_{C_\gamma([s,T];U)}\leqq C_{T\gamma}|x|_Y, \tag{3.21}$$

where the constant $C_{T\gamma}$ *does not depend on* s, $0\leqq s\leqq T$.

Proof. From the definition (2.7) of L_{sT}^* and the assumption (1.2) on B, we readily obtain

$$
\begin{aligned}
|L_{sT}^* G^* G y^0(T, s; x)|_U &= |B^*(-A^*)^{-\gamma}(-A^*)^{\gamma} e^{A^*(T-t)} G^* G y^0(T, s; x)|_U \\
&\leqq \frac{C_{T,\gamma}}{(T-t)^{\gamma}} |Gy^0(T, s; x)|_Z \leqq \frac{C_{T,\gamma}}{(T-t)^{\gamma}} |x|_Y,
\end{aligned} \tag{3.22}
$$

where in the last step we have used (2.21). Then (3.22) yields (3.21) via the definition (3.1). □

The major result of this section is the following.

THEOREM 3.6. *For the optimal pair* $\{u^0, y^0\}$, *we have with constant* $C_{T\gamma}$ *independent of* s, $0 \leqq s \leqq T$:

(i)

$$|u^0(\cdot, s; x)|_{C_\gamma([s,T]; U)} \leqq C_{T\gamma} |x|_Y; \tag{3.23}$$

(ii) *If* $0 \leqq \gamma < \frac{1}{2}$,

$$|y^0(\cdot, s; x)|_{C([s,T]; Y)} \leqq C_{T\gamma} |x|_Y; \tag{3.24a}$$

(iii) *If* $\frac{1}{2} \leqq \gamma < 1$,

$$|y^0(\cdot, s; x)|_{C_{2\gamma-1+\varepsilon}([s,T]; Y)} \leqq C_{T\gamma} |x|_Y, \quad \forall \varepsilon > 0. \tag{3.24b}$$

Proof. (i) We return to (2.16) where we now substitute for the optimal $y^0(\cdot, s; x)$ from (2.18), thus obtaining

$$-u^0(\cdot, s; x) = [I_s + L_s^* R^* R L_s]^{-1} \{L_s^* R^* R\, e^{A(\cdot - s)} x + L_{sT}^* G^* G y^0(T, s; x)\}. \tag{3.25}$$

In fact, the expression in the bracket { } in (3.25) is a well-defined element of $C_\gamma([s, T]; U)$ for $x \in Y$: this is so by (3.21) for its second term and is plainly true (a fortiori) from (2.5) for its first term. Moreover, the inverse in (3.25) is well defined in $C_\gamma([s, T]; U)$ by Theorem 3.4. Indeed, these results collectively yield the bound (3.23), independent of s.

(ii) Estimates (3.24a, b) now follow from estimate (3.23), via the optimal dynamics (2.18), where we use property (3.2) with $r = \gamma$ and $r + \gamma < 1$ for (3.24a), and property (3.3) with $r = \gamma$ and $r + \gamma \geqq 1$ for (3.24b). □

COROLLARY 3.7. *With reference to the operator* $P(t)$ *introduced in* (2.24), *we have the following pointwise estimates for* $s \leqq t < T$;

(i) For $0 \leqq \theta < 1$,

$$|(-A^*)^{\theta} P(t) x| \leqq \frac{C_{T\gamma\theta}}{(T-t)^{\theta}} |x|_Y, \tag{3.26}$$

(ii)

$$|B^* P(t) x| \leqq \frac{C_{T\gamma}}{(T-t)^{\gamma}} |x|_Y, \tag{3.27}$$

where the constant C_T does not depend on s, $0 \leqq s \leqq T$.

Proof. (i) Recalling (2.24) we compute by analyticity

$$
\begin{aligned}
|(-A^*)^{\theta} P(t) x|_Y &= \left| \int_t^T (-A^*)^{\theta} e^{A^*(\tau - t)} R^* R y^0(\tau, t; x)\, d\tau \right. \\
&\qquad \left. + (-A^*)^{\theta} e^{A^*(T-t)} G^* G y^0(T, t; x) \right|_Y \\
&\leqq C_T \int_t^T \frac{1}{(\tau - t)^{\theta}} |y^0(\tau, t; x)|_Y\, d\tau + \frac{C_T}{(T-t)^{\theta}} |Gy^0(T, t; x)|_Z
\end{aligned} \tag{3.28}
$$

$$\leqq C_T\left\{\int_t^T \frac{1}{(\tau-t)^\theta}|y^0(\tau,t;x)|_Y\,d\tau+\frac{1}{(T-t)^\theta}|x|_Y\right\},$$

where in the last step we have used estimate (2.21) for the second term. We now distinguish two cases according to (3.24a, b). If $\gamma<\frac{1}{2}$, then (3.24a) applies in (3.28), and (3.28) yields (3.26). If, instead, $\frac{1}{2}\leqq\gamma<1$, then (3.24b) applies in (3.28), and we obtain for the right-hand side (R.H.S.) of (3.28),

$$\text{(3.29)}\qquad \text{R.H.S. of (3.28)}\leqq C_T\left\{\int_t^T \frac{d\tau}{(\tau-t)^\theta(T-\tau)^{2\gamma-1+\varepsilon}}+\frac{1}{(T-t)^\theta}\right\}|x|_Y.$$

But the integral in (3.29) is the same as the one that occurs when estimating L_s^* in the proof of Proposition 3.1(iii), which culminates with the bound in (3.8), with γ, r there replaced by θ, $2\gamma-1+\varepsilon$ now. Thus, we obtain by (3.28), (3.29), via (3.8),

$$\begin{aligned}&|(-A^*)^\theta P(t)x|_Y\\ \text{(3.29)}\qquad &\leqq C_{T\gamma\theta}\left\{\max\left\{\frac{1}{1-\theta},\frac{1}{1-2\gamma-\varepsilon}\right\}\frac{1}{(T-t)^{2(\gamma-1)+\theta+\varepsilon}}+\frac{1}{(T-t)^\theta}\right\}|x|_Y,\end{aligned}$$

from which (3.26) follows, since $\gamma-1<0$.

(ii) Estimate (3.27) now follows from estimate (3.26) with $\theta=\gamma>\theta+\gamma-1$, via hypothesis (1.2) on B. □

Further properties of $P(t)$ are obtained next, as a consequence of (3.27).

PROPOSITION 3.8.

(i) *For $0\leqq t<T$, the following identity, symmetric in $x, y\in Y$, holds,*

$$\begin{aligned}(P(t)x,y)_Y&=\int_t^T (R\Phi(\tau,t)x,R\Phi(\tau,t)y)_W\,d\tau+(G\Phi(T,t)x,G\Phi(T,t)y)_Z\\ \text{(3.30)}\qquad&\quad+\int_t^T (B^*P(\tau)\Phi(\tau,t)x,B^*P(\tau)\Phi(\tau,t)y)_U\,d\tau.\end{aligned}$$

(ii) *As a consequence,*

$$\text{(3.31)}\qquad P(t)=P^*(t)\geqq 0.$$

(iii) *The optimal cost of the optimal control problem on $[t,T]$ initiating at the point $x\in Y$ at the initial time t is*

$$\begin{aligned}J^0&=J(u^0(\cdot,t;x),y^0(\cdot,t;x))\\ &=\int_t^T |R\Phi(\tau,t)x|_W^2+|B^*P(\tau)\Phi(\tau,t)x|_U^2\,d\tau+|G\Phi(T,t)x|_Z^2\\ \text{(3.32)}\qquad&=(P(t)x,x)_Y.\end{aligned}$$

Proof. (i) As in [L-T1, Prop. 3.3(iii)] we substitute for $e^{A(\tau-t)}y$ and $e^{A(T-t)}y$ occurring in

$$(P(t)x,y)_Y=\int_t^T (R\Phi(\tau,t)x,R\,e^{A(\tau-t)}y)_W\,d\tau+(G\Phi(T,t)x,G\,e^{A(T-t)}y)_Z,$$

the expression obtained from (2.30). Interchanging the order of integration and using that $B^*P(t)$ is well defined for $t<T$, by (3.27), yields (3.30). Details are omitted, see, e.g., [L-T1, Prop. 3.3]. Then (ii) and (iii) follow at once. □

4. Derivation of the Riccati equation (1.13). In this section, our main goal is to show that the operator $P(t) \in \mathscr{L}(Y; L_\infty(0, T; Y))$ explicitly defined by (2.24) in terms of the data of the problem, is a Riccati operator; i.e., it satisfies the Riccati equation (1.13). To accomplish this, we need part (i) of the next theorem; part (ii) will not be invoked in the sequel and is listed here for completeness.

THEOREM 4.1 [L-T1]. (i) *For $x \in Y$ and for each s fixed, $0 \leqq s < T$, the optimal control $u^0(t, s; x)$ and the optimal solution $y^0(t, s; x)$ are respectively U-valued and Y-valued functions which are differentiable in $t \in (s, T)$ with $(\partial u^0/\partial t)(t, s; x) \in U$, $(\partial y^0/\partial t)(t, s; x) \in Y$; see* [L-T5] *for sharp regularity results.*

(ii) *In fact, these U-valued and Y-valued functions $u^0(t, s; x)$ and $y^0(t, s; x)$ are analytic in $t \in (s, T)$ if the operator A has compact resolvent in Y.*

Remark 4.1. Although paper [L-T1] deals specifically with the case of a parabolic equation with Dirichlet boundary control and with $G = I$, its proof of the result in Theorem 4.1 is general. To assert analyticity as in part (ii), the proof uses the analyticity of e^{At}, the representation formula (3.25) for u^0, and the compactness of the resolvent operator of A. The idea (similar to the idea behind the proof of Theorem 3.4) is to assert the invertibility of the operator $[I_s + L_s^* R^* R L_s]$ on the space of U-valued functions ($\mathscr{A}(\mathscr{F}, L_2(\Gamma))$ in the notation of [L-T1]), which are (i) analytic (holomorphic) on a suitable set $\mathscr{F}$ of the complex variable z which contains the interval $(0, T)$, and (ii) continuous on $\bar{\mathscr{F}}$. The proof in [L-T1] works essentially verbatim for the general case.

We next introduce the operator

$$A_P(t) = A - BB^*P(t): \qquad Y \supset \mathscr{D}(A_P(t)) \to Y, \tag{4.1}$$

where $P(t)$ is defined by (2.24) and show that

LEMMA 4.2. *With reference to* (4.1), *we have, for $s \leqq t < T$,*

$$\mathscr{D}(A_P(t)) \subset \mathscr{D}((-A)^{1-\gamma}). \tag{4.2}$$

Proof. Let $x \in \mathscr{D}(A_P(t))$, i.e., $z(t) = A_P(t)x = [A - BB^*P(t)]x \in Y$,

$$-(-A)^\gamma[(-A)^{1-\gamma} + (-A)^{-\gamma}BB^*P(t)]x = z(t) \in Y. \tag{4.3}$$

But $(-A)^{-\gamma}BB^*P(t)x \in Y$ for $t < T$ by assumption (1.2) on B and by property (3.27) of Corollary 3.7. From here and (4.3), we then obtain that

$$(-A)^{1-\gamma}x = -(-A)^{-\gamma}z(t) - (-A)^{-\gamma}BB^*P(t)x \in Y, \tag{4.4}$$

which means that $x \in \mathscr{D}((-A)^{1-\gamma})$, as desired. □

We next recall the operator $\Phi(t, s) \in \mathscr{L}(Y)$, $0 \leqq s \leqq t < T$,

$$\Phi(t, s)x = y^0(t, s; x), \qquad x \in Y, \tag{4.5}$$

which is defined via (2.18), (2.17), solely in terms of the data of the problem. All the preceding results on $y^0(t, s; x)$ can be expressed in terms of $\Phi(t, s)$ via (4.5). In particular, we recall that the optimal dynamics is rewritten (see (2.30)) as

$$\Phi(t, s)x = e^{A(t-s)}x + \int_s^t e^{A(t-\tau)}Bu^0(\tau, s; x)\, d\tau \tag{4.6a}$$

$$= e^{A(t-s)}x + \int_s^t e^{A(t-\tau)}BB^*P(\tau)\Phi(\tau, s)x\, d\tau. \tag{4.6b}$$

In the next two lemmas, we collect some properties of Φ, which in the case of the heat equation were proved in [L-T1].

LEMMA 4.3 [L-T1, Prop. 3.2].

(i) $\Phi(t, t) =$ *identity, and* $\Phi(t, \tau)\Phi(\tau, s) = \Phi(t, s)$ *for* $0 \leqq s \leqq \tau \leqq t < T$ (*transitivity*).

(ii) *For s fixed, the map* $t \to \Phi(t, s)x$ *is continuous in* Y, $s \leqq t < T$, $x \in Y$.

(iii) *For $t < T$ fixed, the map* $s \to \Phi(t, s)x$ *is continuous in* Y, $0 \leqq s \leqq t < T$, $x \in Y$.

(iv) *The map* $s \to G\Phi(T, s)x$ *is continuous in* Z, $0 \leqq s < T$, $x \in Y$.

(v) *For* $0 \leqq s < t < \tau < T$, *the following identity holds for* $x \in Y$:

$$\frac{\partial \Phi(\tau, t)}{\partial t} \Phi(t, s)x = -\Phi(\tau, t) \frac{\partial \Phi}{\partial t}(t, s)x \in Y. \tag{4.7}$$

(vi) *For* $0 \leqq s < t < T$, *the following identity holds for* $x \in Y$

$$\frac{\partial G(T, t)}{\partial t} \Phi(t, s)x = -G\Phi(T, t) \frac{\partial \Phi}{\partial t}(t, s)x \in Y. \tag{4.8}$$

(We note that (4.7), and (4.8) reduce the derivative of Φ and $G\Phi$ in the second argument, computed along the optimal trajectory, $\Phi(t, s)x$, in terms of the derivative of Φ in the first argument.)

(vii) *For* $0 \leqq s < t < T$ *and* $x \in Y$ *we have* $\Phi(t, s)x \in \mathscr{D}((-A)^\theta)$ *with* $\theta < 1 - \gamma$; *moreover, for* $x \in \mathscr{D}((-A)^\theta)$ *we have*

$$\lim_{s \uparrow t} (-A)^\theta \Phi(t, s)x = (-A)^\theta x. \tag{4.10}$$

Proof. (ii) This is a restatement of the result contained in (3.24) of Theorem 3.6 and a fortiori in Theorem 4.1(i).

(iii) For right continuity, we choose $h > 0$ such that $s < s + h \leqq t < T$;

$$\begin{aligned} |\Phi(t, s+h)x - \Phi(t, s)x| &= |\Phi(t, s+h)[x - \Phi(s+h, s)x]| \\ &\leqq \frac{C_T}{(T-t)^r} |\Phi(s+h, s)x - x| \to 0, \end{aligned} \tag{4.11}$$

where in the last step we have used (3.24) via (4.5) with $r = 2\gamma - 1 + \varepsilon$. Then the right-hand side of (4.11) goes to zero as $h \downarrow 0$ by (ii). As to the left continuity, we compute for $h > 0$, again by (3.24) via (4.5):

$$\begin{aligned} |\Phi(t, s-h)x - \Phi(t, s)x| &= |\Phi(t, s)[\Phi(s, s-h)x - x]| \\ &\leqq \frac{C_T}{(T-t)^r} |\Phi(s, s-h)x - x| \to 0, \end{aligned} \tag{4.12}$$

where the right-hand side of (4.12) goes to zero as $h \downarrow 0$ by

$$|\Phi(s, s-h)x - x| \leqq |e^{Ah}x - x| + \int_{s-h}^{s} |e^{A(s-\tau)}(-A)^\gamma A^{-\gamma} Bu^0(\tau, s-h; x)| \, d\tau, \tag{4.13}$$

which follows from (4.6) where (3.23) of Theorem 3.6 is used in (4.13).

(iv) The case $t = T$ is reduced to case (iii); for $|h|$ sufficiently small and $s + h < t < T$:

$$\begin{aligned} |G(\Phi(T, s+h)x - G\Phi(T, s)x| &= |G\Phi(T, t)[\Phi(t, s+h)x - \Phi(t, s)x]| \\ &\leqq C_T |\Phi(t, s+h)x - \Phi(t, s)x| \to 0, \end{aligned} \tag{4.14}$$

where we have used (2.21) of Proposition 2.1, and the right-hand side of (4.14) goes to zero as $h \to 0$ by part (iii).

(v) For $|h|$ sufficiently small so that $0 \leqq s < t+h < \tau < T$ and $s < t < \tau$, we compute for $x \in Y$,

$$\tag{4.15} \begin{aligned} &\frac{1}{h}[\Phi(\tau, t+h)\Phi(t, s)x - \Phi(\tau, t)\Phi(t, s)x] \\ &\qquad = \Phi(\tau, t+h)\frac{1}{h}[\Phi(t, s)x - \Phi(t+h, s)x]. \end{aligned}$$

But since $\partial\Phi(t, s)x/\partial t$ exists in Y by Theorem 4.1(i), we have for the right-hand side of (4.15),

$$\tag{4.16} \begin{aligned} &\Phi(\tau, t+h)\frac{1}{h}[\Phi(t+h, s)x - \Phi(t, s)x] - \Phi(\tau, t)\frac{\partial\Phi}{\partial t}(t, s)x \\ &\quad = \Phi(\tau, t+h)\left\{\frac{1}{h}[\Phi(t+h, s)x - \Phi(t, s)x] - \frac{\partial\Phi}{\partial t}(t, s)x\right\} \\ &\qquad + [\Phi(\tau, t+h) - \Phi(\tau, t)]\frac{\partial\Phi}{\partial t}(t, s)x \to 0 \quad \text{as } h \to 0. \end{aligned}$$

But $\Phi(\tau, t+h)$ is strongly continuous in h (property (iii)) and hence uniformly bounded in h in the $\mathscr{L}(Y)$-norm by the Principle of Uniform Boundedness: then the right-hand side of (4.16) goes to zero as $h \to 0$, and (4.7) is proved via (4.15).

(vi) The proof of (4.8) is similar to the one in (4.15) and (4.16) and uses the fact that $G\Phi(T, t+h)$ is strongly continuous in h for $t < T$ (property (iv)) and hence uniformly bounded in h in the $\mathscr{L}(Z)$-norm.

(vii) From (4.6) we have for $s < t < T$,

$$\tag{4.17} \begin{aligned} (-A)^\theta\Phi(t, s)x &= (-A)^\theta e^{A(t-s)}x \\ &\quad + \int_s^t (-A)^\theta e^{A(t-\tau)}(-A)^\gamma(-A)^{-\gamma}Bu^0(\tau, s; x)\, d\tau, \end{aligned}$$

and the integral in (4.17) is bounded in norm by the expression

$$\frac{C_T}{(T-t)^\gamma}\int_s^t \frac{d\tau}{(t-\tau)^{\theta+\gamma}}$$

(by virtue of property (1.2) on B and of property (3.23) of Theorem 3.6), which is well defined and converges to zero as $s \uparrow t$ if $\theta + \gamma < 1$. Equation (4.17) is well defined in Y, if $x \in Y$, $s < t$; or if $x \in \mathscr{D}((-A)^\theta)$, $s = t$. □

LEMMA 4.4.

(i) *For any $x \in Y$ and any t, $s < t < T$, we have with reference to the operator $A_P(t)$ in* (4.1),

$$\tag{4.18} \frac{\partial\Phi(t, s)x}{\partial t} = A_P(t)\Phi(t, s)x \in Y.$$

(ii) *For $0 \leqq s < t < \tau < T$ and $x \in Y$,*

$$\tag{4.19} \frac{\partial\Phi(\tau, t)}{\partial t}\Phi(t, s)x = -\Phi(\tau, t)A_P(t)\Phi(t, s)x \in Y.$$

(iii) *For $0 \leqq s < t < T$ and $x \in Y$,*

$$\tag{4.20} \frac{\partial G\Phi(T, t)}{\partial t}\Phi(t, s)x = -G\Phi(T, t)A_P(t)\Phi(t, s)x \in Y.$$

Proof. (i) As in [L-T1, Prop. 3.4], we start from (4.6b), differentiate in t (as guaranteed by Theorem 4.1(i)) after taking the Y-inner product with $y \in \mathscr{D}(A^*)$ and obtain for $s < t < T$,

$$\left(\frac{\partial \Phi(t,s)}{\partial t} x, y\right)_Y = (e^{A(t-s)}x, A^*y) - (A^{-1}BB^*P(t)\Phi(t,s)x, A^*y)$$

$$- \left(\int_s^t e^{A(t-\tau)}BB^*P(\tau)\Phi(\tau,s)x\,d\tau, A^*y\right)_Y$$

$$\text{(4.21)} \qquad \text{(by (4.6))} \qquad = ([I - A^{-1}BB^*P(t)]\Phi(t,s)x, A^*y), \qquad y \in \mathscr{D}(A^*),$$

where all the terms are well defined by property (1.2) on B and property (3.27) on $B^*P(t)$, $t < T$. By Theorem 4.1(i), the left-hand side of (4.21) is a well-defined Y-inner product $\forall x, y \in Y$, therefore, so is the right-hand side extended as a duality pairing. Thus, A^* can be moved to the left and (4.18) follows.

Properties (ii) and (iii) are a direct consequence of (4.7) and (4.8) of Lemma 4.3 via (4.18). □

The main result of this section is the following.

THEOREM 4.5. *The operator* $P(t)$ *defined by* (2.24) *satisfies the following Riccati equation for* $0 \leqq t < T$ *and* $x, y \in \mathscr{D}(A)$, *indeed for* $x, y \in \mathscr{D}((-A)^{\varepsilon})$, *for all* ε *with* $0 < \varepsilon < 1 - \gamma$,

$$\text{(4.22)} \qquad \begin{aligned}(\dot{P}(t)x, y)_Y = -(R^*Rx, y)_Y - (P(t)x, Ay)_Y - (P(t)Ax, y)_Y \\ + (B^*P(t)x, B^*P(t)y)_U,\end{aligned}$$

where $\dot{P}(t)$ *is a closed operator and* $(-A^*)^{-\varepsilon}\dot{P}(t)(-A)^{-\varepsilon}$ *can be extended to a bounded operator in* $\mathscr{L}(Y)$.

Proof. With, say, $x \in Y$, $y \in \mathscr{D}(A)$, and $0 \leqq s < t < T$, we differentiate in t (2.24), rewritten now via (4.5) as

$$\text{(4.23)} \qquad (P(t)x, y)_Y = \left(\int_t^T e^{A^*(\tau-t)}R^*R\Phi(\tau,t)x\,d\tau, y\right)_Y + (e^{A^*(T-t)}G^*G\Phi(T,t)x, y)_Y$$

to obtain after replacing x with $\Phi(t,s)x$ (all inner products are in Y),

$$(\dot{P}(t)\Phi(t,s)x, y) = -(R^*R\Phi(t,s)x, y)_Y - \left(\int_t^T e^{A^*(\tau-t)}R^*R\Phi(\tau,t)\Phi(t,s)x\,d\tau, Ay\right)$$

$$+ \left(\int_t^T e^{A^*(\tau-t)}R^*R\frac{\partial \Phi(\tau,t)}{\partial t}\Phi(t,s)x\,d\tau, y\right)$$

$$- (e^{A^*(T-t)}G^*G\Phi(T,t)\Phi(t,s)x, Ay)$$

$$\text{(4.24)} \qquad + (e^{A^*(T-t)}G^*\frac{\partial G\Phi(T,t)}{\partial t}\Phi(t,s)x, y\Big)$$

(using (4.23) with x replaced by $\Phi(t,s)x$ for the second and fourth terms in (4.24))

$$= -(R^*R\Phi(t,s)x, y)_Y - (P(t)\Phi(t,s)x, Ay)$$

$$+ \left(\int_t^T e^{A^*(\tau-t)}R^*R\frac{\partial \Phi(\tau,t)}{\partial t}\Phi(t,s)x\,d\tau, y\right)$$

$$+ \left(e^{A^*(T-t)}G^*\frac{\partial G\Phi(T,t)}{\partial t}\Phi(t,s)x\,d\tau, y\right)$$

(using identities (4.19) and (4.20))

$$= -(R^*R\Phi(t,s)x, y) - (P(t)\Phi(t,s)x, Ay)$$
$$-\left(\int_t^T e^{A^*(\tau-t)}R^*R\Phi(\tau,t)A_P(t)\Phi(t,s)x\,d\tau, y\right)$$
$$-(e^{A^*(T-t)}G^*G\Phi(T,t)A_P(t)\Phi(t,s)x, y)$$

(using again (4.23) with x replaced by $\Phi(t,s)x$)

$$= -(R^*R\Phi(t,s)x, y) - (P(t)\Phi(t,s)x, Ay) - (P(t)A_P(t)\Phi(t,s)x, y).$$

From here, recalling the definition of $A_P(t)$ in (4.1), we obtain

$$(4.25)\qquad (\dot{P}(t)\Phi(t,s)x, y) = -(R^*R\Phi(t,s)x, y) - (P(t)\Phi(t,s)x, Ay)$$
$$-(P(t)[A - BB^*P(t)]\Phi(t,s)x, y).$$

We now verify the following claim: The right-hand side of (4.25) is well defined with $s < t < T$ for all $x \in Y$ and all $y \in \mathscr{D}((-A)^\varepsilon)$, for all $\varepsilon > 0$.

In fact, we first note that from (3.27) of Corollary 3.7, we have for $t < T$,

$$(4.26)\qquad B^*P(t) \in \mathscr{L}(Y, U), \quad \text{and hence } P(t)BB^*P(t) \in \mathscr{L}(Y),$$

since $P(t)$ is self-adjoint (see (3.31) in Proposition 3.8(ii)). Moreover, for $t < T$, $(-A^*)^{1-\varepsilon}P(t) \in \mathscr{L}(Y)$, for all $\varepsilon > 0$, by (3.26) of Corollary 3.7, and likewise, since $P(t)$ is self-adjoint, $P(t)(-A)^{1-\varepsilon}$ can be extended to an operator in $\mathscr{L}(Y)$. Thus, we decompose the operator in the last term of (4.25) as

$$(4.27)\qquad P(t)[A - BB^*P(t)] = -P(t)(-A)^{1-\varepsilon}(-A)^\varepsilon - P(t)BB^*P(t),$$

so that (4.27) is well defined on $\mathscr{D}((-A)^\varepsilon)$, for all $\varepsilon > 0$, by (4.26). But, recalling Lemma 4.3(vii), we then see that $\Phi(t,s)x \in \mathscr{D}((-A)^\varepsilon)$ for all $0 < \varepsilon < 1 - \gamma$, $x \in Y$, $s < t < T$, so that the corresponding term $P(t)[A - BB^*P(t)]\Phi(t,s)x$ is well defined in Y. Thus, our claim has been verified.

We now restrict to $x \in \mathscr{D}((-A)^\varepsilon)$ and obtain from (4.10) with any $\varepsilon < 1 - \gamma$ and from Lemma 4.3(iii), $t < T$, recalling (4.26),

$$\lim_{s\uparrow t} P(t)[A - BB^*P(t)]\Phi(t,s)x = -P(t)(-A)^{1-\varepsilon}\lim_{s\uparrow t}(-A)^\varepsilon\Phi(t,s)x$$
$$-P(t)BB^*P(t)\lim_{s\uparrow t}\Phi(t,s)x$$
$$(4.28)\qquad = P(t)[A - BB^*P(t)]x, \qquad x \in \mathscr{D}((-A)^\varepsilon).$$

Thus, taking the limit of the right-hand side of (4.25) as $s \uparrow t$, $t < T$, and using (4.28) and Lemma 4.3(iii), we obtain the right-hand side of (4.22), well defined for all $x, y \in \mathscr{D}((-A)^\varepsilon)$, for all $\varepsilon > 0$, as desired.

As to the left-hand side of (4.25), we may consider the operator $\dot{P}(t)$ to be well-defined at least on the set $\mathscr{M}_t$ defined by

$$(4.29)\qquad \mathscr{M}_t \equiv \bigcup_{0 \leqq s < t} \mathscr{M}_{ts} \subset \mathscr{D}((-A)^\theta), \quad \text{where } \mathscr{M}_{ts} = \Phi(t,s)Y \subset \mathscr{D}((-A)^\theta)$$

for θ fixed, $\theta < 1 - \gamma$ (by Lemma 4.3(vii)) with $s < t < T$. By using the transitivity property of $\Phi(\cdot,\cdot)$, one then sees that $s_1 < s_2$ implies $\mathscr{M}_{s_1t} \subset \mathscr{M}_{s_2t}$ and that $\mathscr{M}_t$ is actually a subspace of Y, which is dense in Y by Lemma 4.3(ii), (iii). We next show that $\dot{P}(t)$ with domain $\mathscr{M}_t$ is *closeable*. In fact, let $\mathscr{M}_t \ni x_n = \Phi(t, s_n)y_n \to 0$, $s_n < t < T$, $y_n \in Y$, and let $\dot{P}(t)x_n \to v$ in Y. Then, $v = 0$. In fact, identity (4.25) with $\Phi(t,s)x$ replaced now by

$\Phi(t, s_n)y_n$ implies as $n\to\infty$ that $(v, y)=0$ for all $y\in\mathscr{D}((-A)^\varepsilon)$, hence for all $y\in Y$, and then $v=0$. We denote the *closure* of $\dot{P}(t)$ (smallest closed extension) still by $\dot{P}(t)$. Moreover, $\dot{P}(t)$ is also self-adjoint and thus, recalling Lemma 4.3(ii), we have $t<T$,

$$\lim_{s\uparrow t}(\dot{P}(t)\Phi(t, s)x, y)_Y=\lim_{s\uparrow t}(\Phi(t, s)x, \dot{P}(t)y)_Y \tag{4.30}$$
$$=(x, \dot{P}(t)y)_Y=(\dot{P}(t)x, y)_Y,$$

at least $\forall x\in Y$ and $\forall y\in\mathscr{M}_t$. Thus, at this point, we have that the DRE (4.22) holds true for all $x\in\mathscr{D}((-A)^\varepsilon)$, for all ε with $0<\varepsilon<1-\gamma$, and for all $y\in\mathscr{M}_t$. But, as seen above, the right-hand side of (4.22) is well defined also for all $y\in\mathscr{D}((-A)^\varepsilon)$. Moreover, $\mathscr{M}_t$ is dense in $\mathscr{D}((-A)^\varepsilon)$ in the $\mathscr{D}((-A)^\varepsilon)$-topology, as it follows a fortiori from (4.10) of Lemma 4.3. Then, the left-hand side of (4.22) can be extended likewise to all $y\in\mathscr{D}((-A)^\varepsilon)$. This implies that $(-A^*)^{-\varepsilon}\dot{P}(t)(-A)^{-\varepsilon}$ can be extended to a bounded operator in $\mathscr{L}(Y)$. Thus, the DRE (4.22) holds true for all $x, y\in\mathscr{D}((-A)^\varepsilon)$, as desired. The proof of Theorem 4.5 is complete. □

A final property of $P(t)$ which complements property (2.27) and property (3.26) is the following.

PROPOSITION 4.6. *For any $\varepsilon>0$ small we have*

$$(-A^*)^\theta P(t)\in\mathscr{L}(Y; C([0, T-\varepsilon]; Y)), \qquad 0\leqq\theta<1; \tag{4.31}$$

$$B^*P(t)\in\mathscr{L}(Y; C([0, T-\varepsilon]; Y)). \tag{4.32}$$

Proof. (i) From (2.26) with $x\in Y$, we have

$$(-A^*)^\theta P(t)x=\int_t^T(-A^*)^\theta e^{A^*(\tau-t)}R^*R\Phi(\tau, t)x\,d\tau \tag{4.33}$$
$$+(-A^*)^\theta e^{A^*(T-t)}G^*G\Phi(T, t)x,$$

and conclusion (4.31) follows from (4.33) using the properties of Lemma 4.3(ii) and (iv) with $t\leqq T-\varepsilon$.

(ii) Then (4.32) follows from (4.31) via assumption (1.2) on B. □

5. The issue of the limit of $P(t)$ as $t\uparrow T$. We rewrite (2.26) for convenience as

$$P(t)x=\int_t^T e^{A^*(\tau-t)}R^*R\Phi(\tau, t)x\,d\tau+e^{A^*(T-t)}G^*G\Phi(T, t)x, \tag{5.1}$$

and see that its first term satisfies the following result.

LEMMA 5.1.

$$\lim_{t\uparrow T}\int_t^T e^{A^*(\tau-t)}R^*R\Phi(\tau, t)x\,d\tau=0, \qquad x\in Y. \tag{5.2}$$

Proof. Conclusion (5.2) follows just by invoking the L_2-estimate (2.20) for $y^0(\tau, t; x)=\Phi(\tau, t)x$ and using the Schwarz inequality, or else by invoking the sharper estimate in (3.24). □

Lemma 5.1 reduces the strong (weak) convergence of $P(t)$ in (5.1) to the strong (weak) convergence of $G^*G\Phi(T, t)$. We begin with weak convergence results.

PROPOSITION 5.2. *Assume the standing hypothesis that $GL_T=GL_{0T}$ be closeable. Then*

$$\text{(i)}\qquad \lim_{t\uparrow T}(G\Phi(T, t)x, z)_Z=(Gx, z)_Z \qquad \forall x\in Y\quad \forall z\in Z; \tag{5.3}$$

$$\text{(ii)}\qquad \lim_{t\uparrow T}(P(t)x, y)_Y=(G^*Gx, y)_Y \quad \forall x, y\in Y. \tag{5.4}$$

Proof. (i) From the optimal dynamics (4.6a), we have

$$G\Phi(T,t)x = G\,e^{A(T-t)}x + \int_t^T G\,e^{A(T-\tau)}Bu^0(\tau,t;x)\,d\tau, \tag{5.5}$$

and thus (5.3) follows, as soon as we show that

$$\lim_{t\uparrow T}\left(\int_t^T G\,e^{A(T-\tau)}Bu^0(\tau,t;x)\,d\tau, z\right)_Z = 0 \quad \forall x\in Y \quad \forall x \in Z. \tag{5.6}$$

But the following uniform bound in $0\leqq t\leqq T$ holds true,

$$\begin{aligned}\left|\int_t^T G\,e^{A(T-\tau)}Bu^0(\tau,t;x)\,d\tau\right|_Z &= |G\Phi(T,t)x - G\,e^{A(T-t)}x|_Z \\ &\leqq C_T|x|_Y \quad \forall\, 0\leqq t\leqq T\end{aligned} \tag{5.7}$$

from (5.5), recalling (2.21) of Proposition 2.1 and (4.5). Thus, in view of (5.7), we see that the desired limit in (5.6) holds true, as soon as we prove that for all $x\in Y$, for all z in a dense set of Z

$$\lim_{t\uparrow T}\left(\int_t^T G\,e^{A(T-\tau)}Bu^0(\tau,t;x)\,d\tau, z\right)_Z = 0. \tag{5.8}$$

To prove (5.8) we choose $z\in\mathscr{D}((GL_T)^*)$, which is dense in Z by assumption via the equivalence of (1.30). We write $(GL_T)^*$ as usual as $L_T^*G^*$ (to denote, in effect, the extension) and note from (2.7) that $\mathscr{D}(L_{tT}^*G^*)\equiv\mathscr{D}(L_{0T}^*G^*)$, constant in t. Then, (5.8) follows from

$$\begin{aligned}&\left|\left(\int_t^T G\,e^{A(T-\tau)}Bu^0(\tau,t;x)\,d\tau, z\right)_Z\right| \\ &= \left|\int_t^T (u^0(\tau,t;x), B^*\,e^{A^*(T-\tau)}G^*z)_U\,d\tau\right| \\ &\leqq |u^0(\cdot,t;x)|_{L_2(t,T;U)}|L_{tT}^*G^*z|_{L_2(t,T;U)}\to 0 \quad \text{as } t\uparrow T. \quad \square\end{aligned} \tag{5.9}$$

COROLLARY 5.3. *Assume, in addition to GL_{0T} closeable, that G is compact. Then, the following strong convergence results hold*:

$$\text{(i)} \qquad \lim_{t\uparrow T} G^*G\Phi(T,t)x = G^*Gx, \qquad x\in Y; \tag{5.10}$$

$$\text{(ii)} \qquad \lim_{t\uparrow T} P(t)x = G^*Gx, \qquad x\in Y. \tag{5.11}$$

Proof. (i) Since G^* is compact, the weak convergence of $G\Phi(T,t)$ as in (5.3) becomes strong convergence as in (5.10).

(ii) Returning to (5.1) and recalling (5.2), we obtain (5.11) by (5.10). □

6. The smoothing case $(-A^*)^\beta G^*G\in\mathscr{L}(Y)$, $\beta>2\gamma-1$. In this section, we point out a more regular theory, which becomes available under the stronger assumption that

$$(-A^*)^\beta G^*G\in\mathscr{L}(Y), \qquad \beta>2\gamma-1, \tag{6.1}$$

i.e., when G^*G maps all of Y into $\mathscr{D}((-A^*)^\beta)$, so that (6.1) holds by the closed graph theorem. We first recall from [F1, Lemma 3.1] that as a consequence of (6.1), since $G^*G\in\mathscr{L}(Y)$ is self-adjoint, then the operator $(-A^*)^{\beta-\theta}G^*G(-A)^{\theta-\rho}$ admits a bounded extension in $L(Y)$, a condition which we write simply as

$$(-A^*)^{\beta-\theta}G^*G(-A)^{\theta-\rho}\in\mathscr{L}(Y) \quad \forall\, 0<\rho<\theta<\beta. \tag{6.2}$$

Next, we return to (2.17), rewritten here for convenience for $x \in Y$ as

$$-u^0(\cdot, s; x) = \Lambda_{sT}^{-1}[L_{sT}^* G^* G\, e^{A(T-s)}x + L_s^* R^* R\, e^{A(\cdot - s)}x], \tag{6.3a}$$

$$\Lambda_{sT} = I_s + L_s^* R^* R L_s + L_{sT}^* G^* G L_{sT}, \tag{6.3b}$$

which provides $u^0(\cdot, s, x)$ in terms of the data of the problem.

LEMMA 6.1. *Assume hypothesis* (6.1). *Then, for any $x \in Y$, the term in the square bracket of* (6.3a) *satisfies* (*recall* (3.1)),

$$L_{sT}^* G^* G\, e^{A(T-s)}x + L_s^* R^* R\, e^{A(\cdot - s)}x \in C_{\gamma-\beta}([s, T]; U). \tag{6.4}$$

Proof. Since $L_s^* R^* R\, e^{A(\cdot - s)}x \in C([s, T]; U)$ from (2.5a), and since $L_{sT}^* G^* G\, e^{A(T-s)}x \in C([s, T]; U)$ from (2.7), it remains to show that

$$|(L_{sT}^* G^* G\, e^{A(T-s)}x)(t)|_U \leqq \frac{C_{T\gamma}}{(T-t)^{\gamma-\beta}} |x|_Y. \tag{6.5}$$

But (6.5) follows readily via (2.7), (6.1), (1.2), and analyticity from

$$\begin{aligned} |(L_{sT}^* G^* G\, e^{A(T-s)}x)(t)|_U = |B^*(-A^*)^{-\gamma}\, e^{A^*(T-t)}(-A^*)^{\gamma-\beta} \\ \cdot (-A^*)^{\beta} G^* G\, e^{A(T-s)}x|_U. \end{aligned} \tag{6.6}$$ □

The crucial result of this section is the following theorem concerning Λ_{sT} in (6.3b).

THEOREM 6.2. *Assume hypothesis* (6.1). *Then the operator Λ_{sT} satisfies*

$$\Lambda_{sT}^{-1} = [I_s + L_s^* R^* R L_s + L_{sT}^* G^* G L_{sT}]^{-1} \in \mathscr{L}(C_{\gamma-\beta}([s, T]; U)) \tag{6.7}$$

with uniform bound which may be taken independent of s, $s \leqq t \leqq T$.

Proof. The proof will resemble the arguments of § 3 leading to Theorem 3.4. □

Step 1. *Claim.* Let $v_0 \in L_2(s, T; U)$; then

$$|(-A)^{-\sigma_0} L_{sT} v_0|_Y \leqq c_{T,\varepsilon_0}(T-s)^{\varepsilon_0} |v_0|_{L_2(s,T;U)}; \tag{6.8}$$

$$\forall \sigma_0 > \gamma - \tfrac{1}{2}, \quad \text{so that } \sigma_0 = \gamma - \tfrac{1}{2} + \varepsilon_0. \tag{6.9}$$

In fact, we compute from (2.6) via (1.2) and analyticity,

$$\begin{aligned} |(-A)^{-\sigma_0} L_{sT} v_0|_Y &= \left| (-A)^{-\sigma_0} \int_s^T (-A)^{\gamma} e^{A(T-\tau)}(-A)^{-\gamma} B v_0(\tau)\, d\tau \right|_Y \\ &\leqq C_T \int_s^T \frac{|v_0(\tau)|_U\, d\tau}{(T-\tau)^{\gamma-\sigma_0}} \\ &\leqq C_T \left\{ \int_s^T \frac{d\tau}{(T-\tau)^{2(\gamma-\sigma_0)}} \right\}^{1/2} |v_0|_{L_2(s,T;U)}, \end{aligned} \tag{6.10}$$

and for $2(\gamma - \sigma_0) < 1$, we obtain (6.8), as desired.

Step 2. *Claim.* Setting

$$v_1(t) = (L_{sT}^* G^* G L_{sT} v_0)(t), \tag{6.11}$$

we have, with $\beta > 2\sigma_0 = 2\gamma - 1 + 2\varepsilon_0$,

$$|v_1(t)|_U = |(L_{sT}^* G^* G L_{sT} v_0)(t)|_U \leqq \frac{C_T}{\sqrt{\varepsilon_0}} \frac{(T-s)^{\varepsilon_0}}{(T-t)^{1/2-\varepsilon_0}} |v_0|_{L_2(s,T;U)}, \tag{6.12}$$

so that a fortiori $L_{sT}^* G^* G L_{sT}$ is a smoothing operator:

$$L_{sT}^* G^* G L_{sT}: \text{continuous } L_2(s, T; U) \to L_{r_1}(s, T; U), \tag{6.13a}$$

$$2 < r_1 < \frac{1}{\gamma - \sigma_0} = \frac{2}{1 - 2\varepsilon_0}; \qquad r_1 - 2 < \frac{4\varepsilon_0}{1 - 2\varepsilon_0}. \tag{6.13b}$$

In fact, we compute from (2.7) and (2.6), via (1.2) and (6.2),

$$\begin{aligned}|(L_{sT}^*G^*GL_{sT}v_0)(t)|_U &= |B^*(-A^*)^{-\gamma}\, e^{A^*(T-t)}(-A^*)^{\gamma-\sigma_0} \\ &\quad \cdot (-A^*)^{\sigma_0}G^*G(-A)^{\sigma_0}(-A)^{-\sigma_0}L_{sT}v_0|_U \\ &\leqq C_T\frac{1}{(T-t)^{\gamma-\sigma_0}}|(-A)^{-\sigma_0}L_{sT}v_0|_Y,\end{aligned} \tag{6.14}$$

where in the last step we have used (6.2) for $(-A^*)^{\sigma_0}G^*G(-A)^{\sigma_0} \in \mathscr{L}(Y)$, which is legal since $\beta > 2\sigma_0 > 2\gamma - 1$ from (6.1) and (6.9). Then, (6.12) readily follows from (6.14) via (6.8), with $\gamma - \sigma_0 = \frac{1}{2} - \varepsilon_0$ from (6.9). In turn, (6.12) implies that $v_1 \in L_r(s, T; U)$ for all r such that $(\frac{1}{2} - \varepsilon_0)r = (\gamma - \sigma_0)r < 1$, i.e., in particular, for r_1 as in (6.13b).

Step 3. We reiterate the procedure of Steps 1-2 above. Since v_1 is more regular than the original v_0, we have the following claim.

Claim. We have

$$|(-A)^{-\sigma_1}L_{sT}v_1|_Y \leqq C_T(T-s_0)^{\varepsilon_0}|v_0|_{L_2(s,T;U)}; \tag{6.15}$$

$$\sigma_1 = \sigma_0 - \varepsilon_0 = \gamma - \tfrac{1}{2} < \sigma_0. \tag{6.16}$$

In fact, from (2.6) we compute via (1.2) and (6.12),

$$\begin{aligned}|(-A)^{-\sigma_1}L_{sT}v_1|_Y &= \left|(-A)^{-\sigma_1}\int_s^T (-A)^{\gamma}\, e^{A(T-\tau)}(-A)^{-\gamma}Bv_1(\tau)\, d\tau\right|_Y \\ &\leqq C_T\int_s^T\frac{|v_1(\tau)|_U\, d\tau}{(T-\tau)^{\gamma-\sigma_1}} \\ &\leqq C_T\left(\int_s^T\frac{d\tau}{(T-\tau)^{\gamma-\sigma_1+\gamma-\sigma_0}}\right)|v_0|_{L_2(s,T;U)},\end{aligned} \tag{6.17}$$

and (6.15) follows since $\gamma - \sigma_1 + \gamma - \sigma_0 = 1 - \varepsilon_0 < 1$ from (6.16) and (6.9).

Step 4. *Claim.* Setting

$$v_2(t) = (L_{sT}^*G^*GL_{sT}v_1)(t), \tag{6.18}$$

we have that

$$|v_2(t)|_U = |L_{sT}^*G^*GL_{sT}v_1)(t)|_U \leqq C_T\frac{(T-s)^{\varepsilon_0}}{(T-t)^{1/2-2\varepsilon_0}}|v_0|_{L_2(s,T;U)}, \tag{6.19}$$

so that a fortiori,

$$L_{sT}^*G^*GL_{sT}\text{: continuous } L_{r_1}(s, T; U) \to L_{r_2}(s, T; U); \tag{6.20a}$$

$$2 < r_1 < r_2 < \frac{1}{\gamma - \sigma_0 - \varepsilon_0} = \frac{2}{1 - 4\varepsilon_0}; \qquad r_2 - r_1 < r_2 - 2 < \frac{8\varepsilon_0}{1 - 4\varepsilon_0}. \tag{6.20a}$$

In fact, since $\sigma_1 = \sigma_0 - \varepsilon_0 < \sigma_0$, we rewrite the counterpart of (6.14) as

$$\begin{aligned}|(L_{sT}^*G^*GL_{sT}v_1)(t)|_U &= |B^*(-A^*)^{-\gamma}\, e^{A^*(T-t)}(-A^*)^{\gamma-(\sigma_0+\varepsilon_0)} \\ &\quad \cdot (-A^*)^{\sigma_0+\varepsilon_0}G^*G(-A)^{\sigma_0-\varepsilon_0}(-A)^{-\sigma_1}L_{sT}v_1|_U \\ &\leqq C_T\frac{1}{(T-t)^{\gamma-(\sigma_0+\varepsilon_0)}}|(-A)^{-\sigma_1}L_{sT}v_1|_Y,\end{aligned} \tag{6.21}$$

where in the last step we have used (6.2), for $(-A^*)^{\sigma_0+\varepsilon_0}G^*G(-A)^{\sigma_0-\varepsilon_0}\in\mathscr{L}(Y)$, which is legal since $\beta>(\sigma_0+\varepsilon_0)+(\sigma_0-\varepsilon_0)=2\sigma_0>2\gamma-1$ by (6.1) and (6.16). Then, (6.19) readily follows from (6.21) via (6.15) with $\gamma-(\sigma_0+\varepsilon_0)=\frac{1}{2}-2\varepsilon_0$ from (6.9). In turn, (6.19) implies that $v_2\in L_r(s, T; U)$ for all r such that $(\frac{1}{2}-2\varepsilon_0)r=[\gamma-(\sigma_0+\varepsilon_0)]r<1$; i.e., in particular, for r_2 as in (6.20b).

Step 5. The above procedure can be iterated a finite number of times yielding the following result: Let

$$v_n(t)=(L^*_{sT}G^*GL_{sT}v_{n-1})(t). \tag{6.22}$$

Then

$$|v_n(t)|_U=|(L^*_{sT}G^*GL_{sT}v_{n-1})(t)|_U\leqq C_T\frac{(T-s)^{\varepsilon_0}}{(T-t)^{1/2-n\varepsilon_0}}|v_0|_{L_2(s,T;U)}, \tag{6.23}$$

so that a fortiori

$$L^*_{sT}G^*GL_{sT}\text{: continuous } L_{r_{n-1}}(s, T; U)\to L_{r_n}(s, T; U); \tag{6.24a}$$

$$2<r_1<r_2<\cdots<r_n=\frac{1}{\gamma-\sigma_0-(n-1)\varepsilon_0}=\frac{2}{1-2n\varepsilon_0}; \tag{6.24b}$$

for $1-2n\varepsilon_0>0$.

Step 6. *Claim.* If $n>[\frac{1}{2}-(\gamma-\beta)]/\varepsilon_0$, then $v_n(t)$ in (6.22) satisfies

$$v_n(t)=(L^*_{sT}G^*GL_{sT}v_{n-1})(t)\in C_{\gamma-\beta}([s, T]; U), \tag{6.25}$$

as it readily follows from Step 5.

Step 7. Having obtained the results in (6.4) and (6.25), we return to identity (6.39) and apply a bootstrap argument (similar to the one carried out in Theorem 3.4). There is, to begin with, $u^0=u^0(\cdot, s; x)\in L_2(s, T; U)$ such that by (6.4) we have

$$u^0+L^*_sR^*RL_su^0+L^*_{sT}G^*GL_{sT}u^0=w\in C_{\gamma-\beta}([s, T]; U), \tag{6.26}$$

with $w=L^*_{sT}G^*G\,e^{A(T-s)}x+L^*_sR^*R\,e^{A(\cdot-s)}x$. Starting from (6.26) and proceeding as in Step 2 in the proof of Theorem 3.4, we then apply to (6.26) the operator $L^*_{sT}G^*GL_{sT}$ $n-1$ times, consecutively, thus obtaining $n-1$ additional identities, the last of which is

$$(L^*_{sT}G^*GL_{sT})^{n-1}u^0+(L^*_{sT}G^*GL_{sT})^nu^0=f_n\in C_{\gamma-\beta}([s, T]; U), \tag{6.27}$$

where $f_n=(L^*_{sT}G^*GL_{sT})^{n-1}[w-L^*_sR^*RL_su^0]$. Now, if n is chosen sufficiently large as in the claim of Step 6, then we obtain

$$(L^*_{sT}G^*GL_{sT})^nu^0\in C_{\gamma-\beta}([s, T]; U) \tag{6.28}$$

by (6.25). Then (6.28) and (6.27) imply that

$$(L^*_{sT}G^*GL_{sT})^{n-1}u^0\in C_{\gamma-\beta}([s, T]; U).$$

Proceeding backward along the remaining $n-2$ identities, we eventually obtain

$$u^0(\cdot, s; x)\in C_{\gamma-\beta}([s, T]; U), \tag{6.29}$$

and Theorem 6.2 is proved.

COROLLARY 6.3. *Assume hypothesis* (6.1) (*which is empty if* $0\leqq\gamma<\frac{1}{2}$), *where we set* $\beta=2\gamma-1+\varepsilon$. *Then, for* $x\in Y$,

(i)

$$u^0(\cdot, s; x)\in C_{1-\gamma-\varepsilon}([s, T]; U), \tag{6.30a}$$

and

$$\text{(6.30b)}\qquad |u^0(\cdot, s; x)|_{C_{1-\gamma-\varepsilon}([s,T];U)} \leqq C_{T\gamma}|x|_Y, \qquad \gamma-\beta = 1-\gamma-\varepsilon,$$

with bound which may be taken independent of s;

(ii)

$$\text{(6.31a)}\qquad y^0(\cdot, s; x) = \Phi(\cdot, s)x \in C([s, T]; Y),$$

and

$$\text{(6.31b)}\qquad |y^0(\cdot, s; x)|_{C([s,T];Y)} \leqq C_{T\gamma}|x|_Y,$$

with bound which may be taken independent of s;

(iii)

$$\text{(6.32a)}\qquad y^0(T, \cdot\,; x) = \Phi(T, \cdot)x \in C([s, T]; Y),$$

in particular.

$$\text{(6.32b)}\qquad \lim_{t\uparrow T} \Phi(T, t)x = x.$$

(iv) *For any* $0 \leqq \theta < 1$, *and* $\theta-\beta = \theta+1-2\gamma-\varepsilon$,

$$\text{(6.33a)}\qquad (-A^*)^\theta P(t)x \in C_{\theta-\beta}([0, T]; Y),$$

and

$$\text{(6.33b)}\qquad |(-A^*)^\theta P(t)x|_Y \leqq \frac{C_{T\gamma}}{1-\theta}\frac{1}{(T-t)^{\theta-\beta}}|x|_Y.$$

(v) *With* $\gamma-\beta = 1-\gamma-\varepsilon$,

$$\text{(6.34a)}\qquad B^*P(t)x \in C_{\gamma-\beta}([0, T]; U),$$

and

$$\text{(6.34b)}\qquad |B^*P(t)x|_U \leqq \frac{C_T}{1-\gamma}\frac{1}{(T-t)^{\gamma-\beta}}|x|_Y;$$

(vi)

$$\text{(6.35)}\qquad \lim_{t\uparrow T} P(t)x = \lim_{t\uparrow T} e^{A^*(T-t)}G^*G\Phi(T, t)x = G^*Gx.$$

Proof. (i) Conclusion (6.30) is a rewriting of (6.29); i.e., of Theorem 6.1.

(ii) Conclusion (6.31) follows from (6.30) via the optimal dynamics (2.18) and property (3.2) for the operator L_s with $r = 1-\gamma-\varepsilon$ so that $r+\gamma = 1-\varepsilon < 1$, as required.

(iii) Conclusion (6.32) follows from (2.18) with $t = T$, i.e., from (see (2.23)),

$$\text{(6.36)}\qquad y^0(T, t; x) = \Phi(T, t)x = e^{A(T-t)}x + L_{tT}u^0(\cdot, t; x),$$

where, by virtue of (2.6), (6.30), and (1.2), we obtain as desired,

$$\begin{aligned} |L_{tT}u^0(\cdot, t; x)|_Y &= \left|\int_t^T (-A)^\gamma e^{A(T-\tau)}(-A)^{-\gamma}Bu^0(\tau, t; x)\,d\tau\right|_Y \\ \text{(6.37)}\qquad &\leqq C_{T\gamma}\int_t^T \frac{d\tau}{(T-\tau)^\gamma(T-\tau)^{1-\gamma-\varepsilon}}|x|_Y \\ &= C_{T\gamma}(T-t)^\varepsilon|x|_Y. \end{aligned}$$

(iv) Recalling (2.24), we compute via (6.31), (6.32), and (6.1),

$$
\begin{aligned}
|(-A^*)^\theta P(t)x|_Y &= \bigg| \int_t^T (-A^*)^\theta e^{A^*(\tau-t)} R^*Ry^0(\tau, t; x)\, d\tau \\
&\qquad + (-A^*)^{\theta-\beta} e^{A^*(T-t)}(-A^*)^\beta G^*Gy^0(T, t; x)\bigg|_Y \\
&\leqq C_{T\gamma}\left\{\int_t^T \frac{d\tau}{(\tau-t)^\theta} + \frac{C_T}{(T-t)^{\theta-\beta}}\right\}|x|_Y, \qquad (6.38)
\end{aligned}
$$

and (6.33b) follows.

(v) By (6.33) with $\theta=\gamma$ and (1.2),

$$
|B^*P(t)x|_U = |B^*(-A^*)^{-\gamma}(-A^*)^\gamma P(t)x|_U = \frac{C_{T\gamma}}{1-\gamma}\frac{1}{(T-t)^{\gamma-\beta}}|x|_Y, \qquad (6.39)
$$

where $\gamma-\beta = 1-\gamma-\varepsilon$, as desired.

(vi) Conclusion (6.35) follows from (6.32b) via (5.1) and (5.2). □

THEOREM 6.4. (*Uniqueness of Riccati operator.*) *The operator $P(t)$ defined constructively in* (2.24) *in terms of the data of the problem is the unique solution to the* DRE (1.13) = (4.22) *and its terminal condition* (6.35) *within the class of self-adjoint operators $\bar{P}(t) \in \mathscr{L}(Y)$ such that,*

$$
\text{(i)} \qquad B^*\bar{P}(t)x \in C_\gamma([0, T]; Y), \qquad x \in Y, \qquad (6.40)
$$

if $0 \leqq \gamma < \frac{1}{2}$ (and no other assumption except the standing hypotheses specified below (1.1));

$$
\text{(ii)} \ B^*\bar{P}(t)x \in C_{\gamma-\beta}([0, T]; Y), \qquad x \in Y, \quad \gamma-\beta = 1-\gamma-\varepsilon, \qquad (6.41)
$$

if $\frac{1}{2} \leqq \gamma < 1$, provided that assumption (6.1) holds with $\beta = 2\gamma - 1 + \varepsilon$.

Proof. It suffices to show uniqueness within the specified class for the corresponding Riccati Integral Equation,

$$
\begin{aligned}
(P(t)x, y)_Y &= (G e^{A(T-t)}x, G e^{A(T-t)}y)_Z + \int_t^T (R e^{A(\tau-t)}x, R e^{A(\tau-t)}y)_W\, d\tau \\
&\quad - \int_t^T (B^*P(\tau) e^{A(\tau-t)}x, B^*P(\tau) e^{A(\tau-t)}y)_U\, d\tau, \qquad (6.42)
\end{aligned}
$$

$x, y \in Y$. Let $P_1(t)$ and $P_2(t)$ be two solutions within the specified class, and let $Q(t) = P_1(t) - P_2(t)$. Then $Q(t)$ satisfies for $x \in Y$,

$$
B^*Q(t)x \in \begin{cases} C_\gamma([0, T]; U) & \text{in case (i)}, \quad (6.43) \\ C_{1-\gamma-\varepsilon}([0, T]; U) & \text{in case (ii)}, \quad (6.44) \end{cases}
$$

as well as, for $0 < t < T$ and $x, y \in Y$,

$$
\begin{aligned}
(Q(t)x, y)_Y &= \int_t^T (B^*P_2(\tau) e^{A(\tau-t)}x, B^*Q(\tau) e^{A(\tau-t)}y)_U\, d\tau \\
&\quad - \int_t^T (B^*Q(\tau) e^{A(\tau-t)}x, B^*P_1(\tau) e^{A(\tau-t)}y)_U\, d\tau. \qquad (6.45)
\end{aligned}
$$

Set

$$
y = Bv, \ v \in U, \quad \text{and } B^*Q(t) = V(t), \qquad (6.46)
$$

so that $V(t)$ solves for $0 \leqq t < T$,

$$
\begin{aligned}
V(t)x = & \int_t^T B^*(-A^*)^{-\gamma}(-A^*)^{\gamma}\, e^{A^*(\tau-t)}\, V^*(\tau)B^*P_2(\tau)\, e^{A(\tau-t)}x\, d\tau \\
& - \int_t^T B^*(-A^*)^{-\gamma}(-A^*)^{\gamma}\, e^{A^*(\tau-t)}(B^*P_1(\tau))^*\, V(\tau)\, e^{A(\tau-t)}x\, d\tau.
\end{aligned} \tag{6.47}
$$

We seek to establish uniqueness of the solution $V(t)$ of (6.47) within the classes specified, respectively, in (6.43) and (6.44) for the two cases. We do this first locally, near T, and extend globally to all of $[0, T]$.

Case (i). $\gamma < \frac{1}{2}$. Multiplying (6.47) across for $(T-t)^{\gamma}$ we obtain after using (1.2), and $|B^*P_i(\tau)y|_U \leqq (C_T/(T-\tau)^{\gamma})|y|_Y$:

$$
\begin{aligned}
(T-t)^{\gamma}|V(t)x|_U & \leqq (T-t)^{\gamma}C_T \int_t^T \frac{(T-\tau)^{\gamma}|V^*(\tau)|\, d\tau}{(\tau-t)^{\gamma}(T-\tau)^{2\gamma}}|x|_Y \\
& \leqq (T-t)^{\gamma}C_T\left\{\int_t^T \frac{d\tau}{(\tau-t)^{\gamma}(T-\tau)^{2\gamma}}\right\}\left\{\sup_{t\leqq\tau\leqq T}(T-\tau)^{\gamma}|V(\tau)|\right\}|x|_Y.
\end{aligned} \tag{6.48}
$$

Since $r = 2\gamma < 1$ in our case, the computations (of case (iii) in the proof of Proposition 3.1) leading to (3.8) can be applied to the integral in (6.48). We thus obtain

$$
\begin{aligned}
(T-t)^{\gamma}|V(t)x|_U & \leqq (T-t)^{\gamma}C_T\frac{1}{(T-t)^{\gamma+2\gamma-1}}\left\{\sup_{t\leqq\tau\leqq T}(T-\tau)^{\gamma}|V(\tau)|\right\}|x|_Y \\
& \leqq C_T(T-t)^{1-2\gamma}\left\{\sup_{t\leqq\tau\leqq T}(T-\tau)^{\gamma}|V(\tau)|\right\}|x|_Y,
\end{aligned} \tag{6.49}
$$

where $1-2\gamma>0$ in our case. Letting $t_0 \leqq t \leqq T$, we obtain from (6.49),

$$
\sup_{t_0\leqq t\leqq T}(T-t)^{\gamma}|V(t)| \leqq C_T(T-t_0)^{1-2\gamma}\left\{\sup_{t\leqq\tau\leqq T}(T-\tau)^{\gamma}|V(\tau)|\right\}, \tag{6.50}
$$

and selecting $T-t_0$ sufficiently small, we obtain $C_T(T-t_0)^{1-2\gamma}<1$ and uniqueness on $[t_0, T]$ is established within the class $C_{\gamma}([t_0, T]; U)$.

Case (ii). $\frac{1}{2} \leqq \gamma < 1$. We now multiply (6.47) by $(T-t)^{1-\gamma-\varepsilon}$ and obtain after using $|B^*P_i(\tau)y|_U \leqq (C_T/(T-\tau)^{1-\gamma-\varepsilon})|y|_Y$,

$$
\begin{aligned}
(T-t)^{1-\gamma-\varepsilon}|V(t)x|_U \leqq & (T-t)^{1-\gamma-\varepsilon}C_T\left\{\int_t^T \frac{d\tau}{(\tau-t)^{\gamma}(T-\tau)^{2(1-\gamma-\varepsilon)}}\right\} \\
& \cdot\left\{\sup_{t\leqq\tau\leqq T}(T-\tau)^{1-\gamma-\varepsilon}|V(\tau)|\right\}|x|_Y
\end{aligned} \tag{6.51}
$$

as a counterpart of (6.48). Since $r = 2(1-\gamma-\varepsilon)<1$ in our case the integral in (6.51) can be estimated again as in (3.8), and we obtain since $\gamma+[2(1-\gamma-\varepsilon)]-1 = 1-\gamma-2\varepsilon$,

$$
\begin{aligned}
(T-t)^{1-\gamma-\varepsilon}|V(t)x|_U & \leqq (T-t)^{1-\gamma-\varepsilon}C_T\frac{1}{(T-t)^{1-\gamma-2\varepsilon}}\left\{\sup_{t\leqq\tau\leqq T}(T-\tau)^{1-\gamma-\varepsilon}|V(\tau)|\right\}|x|_Y \\
& \leqq (T-t)^{\varepsilon}C_T\left\{\sup_{t\leqq\tau\leqq T}(T-\tau)^{1-\gamma-\varepsilon}|V(\tau)|\right\}|x|_Y.
\end{aligned} \tag{6.52}
$$

The desired conclusion of uniqueness of $V(t)$ over $[t_0, T]$ with $T-t_0$ sufficiently small is obtained as in case (i), within the class $C_{1-\gamma-\varepsilon}([t_0, T]; U)$.

Finally, after a finite number of steps we obtain uniqueness of $V(t)$ on all of $[0, T]$, in each case, within its specified class. □

Remark 6.1. Under the assumption (as in [F1]),

$$(-A^*)^\gamma G^*G \in \mathscr{L}(Y), \tag{6.53}$$

which is stronger than assumption (6.1) since $2\gamma-1<\gamma$, additional regularity results hold true, namely,

(i)

$$u^0(\cdot, s; x) \in C([s, T]; U), \qquad x \in Y, \tag{6.54a}$$

$$\max_{s \leqq t \leqq T} |u^0(t, s; x)|_U \leqq C_T |x|_Y, \qquad x \in Y, \tag{6.54b}$$

with C_T independent of s;

(ii) For any $0 \leqq \theta < 1$,

$$(-A^*)^\theta P(t)x \in C_{\theta-\gamma}([0, T]; Y), \tag{6.55a}$$

$$|(-A^*)^\theta P(t)x|_Y \leqq \frac{C_T}{1-\theta} \frac{1}{(T-t)^{\theta-\gamma}} |x|_Y; \tag{6.55b}$$

(iii)

$$B^*P(t)x \in C([0, T]; U), \qquad x \in Y, \tag{6.56a}$$

$$\max_{0 \leqq t \leqq T} |B^*P(t)x| \leqq C_T |x|_Y, \qquad x \in Y. \tag{6.56b}$$

A sketch of the proof of the crucial property (6.54) is as follows. Under assumptions (6.53), we can show using (2.7), (2.6), [F1, Lemma 3.1] (as above in § 6) that

$$L_{sT}^* G^* G L_{sT}\text{: continuous } L_2(s, T; U) \to L_r(s, T; Y); \tag{6.57}$$

$$2 < r < \frac{2}{\gamma + 2\varepsilon} \quad \forall \varepsilon > 0,$$

with norm bounds which may be taken independent of s. Hence, by iteration, there exists a positive integer n_1 such that

$$(L_{sT}^* G^* G L_{sT})^{n_1}\text{: } L_2(s, T; U) \to C([s, T]; Y) \tag{6.58}$$

with norm bound which may be taken independent of s. Thus (6.58) lets us use a bootstrap argument (as in, say, Step 7 in the proof of Theorem 6.2) and obtain (6.54).

7. Counterexamples. It was independently noted in [F3] and in the first version of this paper that suitable one-dimensional range (finite range) operators G furnish examples which illustrate the sharpness or limitations of the theory presented.

7.1. Counterexample to the existence of the optimal control u^0 when GL_T is not closeable. The following example is proposed in [F3], but it is analyzed here from a different viewpoint (the example from [F3] is summarized in Remark 7.1). More precisely, in this example we shall see that the operator GL_T is not closeable and that the optimal control does not exist. This shows that our Theorem 1.1, more generally the treatment of this article, is sharp. Recall from § 2.1 that the assumption that GL_T be closeable is used to obtain a *complete* inner product space (Hilbert) $V(s, T; U)$ defined by (the extension of) $\mathscr{D}(GL_T)$ with respect to the inner product in (2.8), and that, moreover, the optimal control is characterized by (2.17), with Λ_{sT}^{-1} an isomorphism from $[V(s, T; U)]'$ onto $V(s, T; U)$.

The example. Consider, say the heat equation defined on a (smooth) bounded domain $\Omega \subset R^n$ with $L_2(0, T; L_2(\Gamma))$-control in the Dirichlet boundary conditions.

$$y_t = \Delta y \qquad \text{in } Q = (0, T] \times \Omega \tag{7.1a}$$

$$y(0, \cdot) = y_0 \qquad \text{in } \Omega \tag{7.1b}$$

$$y|_\Sigma = u \qquad \text{in } \Sigma = (0, T] \times \Gamma. \tag{7.1c}$$

Here $Y = L_2(\Omega)$, $U = L_2(\Gamma)$. There exists $\phi \in Y$, $|\phi| = 1$ such that

$$\int_0^T |B^* e^{A^*(T-t)}\phi|_U^2 \, dt = \infty; \tag{7.2}$$

for, otherwise, by transposition, the map $u \to y(T)$ (where $y_0 = 0$) would be continuous $L_2(0, T; L_2(\Gamma)) \to L_2(\Omega) = Y$, which is false even in the one-dimensional case, e.g., [Lio1, p. 217]. Following [F3], we consider the associated optimal control problem (1.3) with

$$R = 0; \quad Gy = (y, \phi)_Y \phi; \quad \phi \in Y; \quad G^* = G = G^*G. \tag{7.3}$$

Note that we have by (1.8) and (7.3),

$$GL_T u = \left(\int_0^T e^{A(T-t)} Bu(t) \, dt, \phi \right)_Y \phi = (u, B^* e^{A^*(T-\cdot)}\phi)_{L_2(0,T;U)} \phi \tag{7.4}$$

so that GL_T is finite rank and unbounded by (7.2), hence uncloseable [K1, p. 166].

Claim. There is no optimal control in this case. In fact, following § 2.1, if an optimal control $u^0(\cdot, 0; x) = u^0 \in L_2(0, T; U)$ exists, it satisfies the present version of (2.17) (or (1.10)), i.e., since $R = 0$,

$$-[u^0 + L_T^* G^* G L_T u^0] = L_T^* G^* G \, e^{AT} x = (e^{AT}x, \phi)_Y B^* e^{A^*(T-t)}\phi, \tag{7.5}$$

where we have used (7.3) on G^*G and (1.9) for L_T^*. Moreover, by (7.3) and (7.4),

$$L_T^* G^* G L_T u = L_T^* \{(L_T u, \phi)_Y \phi\} = \left(\int_0^T e^{A(T-t)} Bu(t) \, dt, \phi \right)_Y B^* e^{A^*(T-t)}\phi. \tag{7.6}$$

Using (7.6) in (7.5) yields

$$-u^0 = \{(u^0, B^* e^{A^*(T-\cdot)}\phi)_{L_2(0,T;U)} + (e^{AT}x, \phi)_Y\} B^* e^{A^*(T-t)}\phi. \tag{7.7}$$

Since $B^* e^{A^*(T-t)}\phi \notin L_2(0, T; U)$ by (7.2), then (7.7) yields that $u^0 \notin L_2(0, T; U)$, a contradiction.

Remark 7.1. It is argued in [F3] that, in the present case, it is *not possible* for the corresponding optimal problem (1.3), (7.3) to satisfy the following three desirable properties.

(i) That there exists a unique optimal control u^0;

(ii) That there exists $P(t)$, $0 \leqq t \leqq T$, nonnegative self-adjoint, such that identity (1.20) holds;

(iii) That for every $0 \leqq t < T$ and $x \in \mathscr{D}(A)$, $(P(t)x, x)$ is differentiable, $P(t)x \in \mathscr{D}((-A^*)^\gamma)$ and the DRE (1.21) is satisfied.

The analysis above shows, more fundamentally, that the optimal control does not exist in this case, with no need to involve the DRE.

Remark 7.2. We note that the choice (7.2) for ϕ implies that $\phi \notin \mathscr{D}((-A^*)^{\beta/2})$ for all $\beta > 2\gamma - 1$, and hence [F4, § 3.1], G in (7.3) does *not* satisfy assumption (1.46) of [F4]. In fact, if we had $\phi \in \mathscr{D}((-A^*)^{\beta/2})$ we would obtain that

$$B^* e^{A^*(T-t)}\phi = B^*(-A^*)^{-\gamma}(-A^*)^{\gamma-\beta/2} e^{A^*(T-t)}(-A^*)^{\beta/2}\phi \tag{7.8}$$

would belong to $L_2(0, T; U)$ by (1.3) and analyticity with $2\gamma-\beta<1$, thus contradicting (7.2). We note that in this case we have $\mathscr{D}((-A^*)^{\beta/2}G^*)=\{0\}$ for all $\beta>2\gamma-1$.

7.2. Assumption (1.31) is only sufficient for GL_T to be closed. We shall provide a class of examples where condition (1.31) is *violated*, yet GL_T is *closed*. This is not surprising as condition (1.31)—unlike GL_T—does not involve B. Let the generator A be negative, self-adjoint, say with compact resolvent. (We shall, however, maintain the notation A^*.) Let $\{e_n, n=1,2,\cdots\}$ be the corresponding orthonormal basis of eigenvectors of A on Y with eigenvalues $\{-\mu_n\}$, $\mu_n>0$. Let $\mathscr{S}_i$, $i=1,2$, be two infinite, disjoint sequences of positive integers that exhaust all of the positive integers $\mathbb{Z}$: $\mathscr{S}_1\cup\mathscr{S}_2=\mathbb{Z}$; $\mathscr{S}_1\cap\mathscr{S}_2=\emptyset$. For example, $\mathscr{S}_1=\{n=2,4,6,\cdots\}$, $\mathscr{S}_2=\{n=1,3,5,\cdots\}$. Consider the orthogonal decomposition of Y

$$Y=Y_1+Y_2,\quad Y_i=\overline{\operatorname{span}}\,\{e_n, n\in\mathscr{S}_i\},\quad i=1,2. \tag{7.9}$$

Let Π_i be the orthogonal projection of Y onto Y_i, so that Π_i commutes with A, hence with the semigroup e^{At} and Y_i are invariant under e^{At}. Define a vector $b\in Y_1$ by setting

$$(b, e_n)_Y=\begin{cases}\text{sequence in } n\in\mathscr{S}_1 \text{ such that } \sum\limits_{n\in\mathscr{S}_1}\mu_n^\beta|(b,e_n)_Y|^2=\infty\\ 0,\qquad n\in\mathscr{S}_2\end{cases} \tag{7.10}$$

for all $\beta>2\gamma-1$, so that

$$b\notin\mathscr{D}((-A^*)^{\beta/2})\quad\forall\beta>2\gamma-1. \tag{7.11}$$

Next, with $U=Y=W$, define the bounded operators G^*, G and the unbounded operators B^*, B by

$$G^*y=(y_1,a)_Yb+y_2;\quad Gy=(y_1,b)_Ya+y_2;\quad y_i=\Pi_iy\in Y_i;\quad a\in Y. \tag{7.12}$$

$$\begin{aligned}&By_1=0; &&B^*y_1=0; &&y_1=\Pi_1y\in Y_1;\\ &By_2=(-A)^\gamma y_2; &&B^*y_2=(-A^*)^\gamma y_2; &&y_2=\Pi_2y\in Y_2\cap\mathscr{D}((-A)^\gamma).\end{aligned} \tag{7.13}$$

We readily obtain by (7.12), (7.11) that

$$\begin{aligned}&\mathscr{D}((-A^*)^{\beta/2}G^*)=\mathscr{D}((-A^*)^{\beta/2})\cap Y_2;\\ &(-A^*)^{\beta/2}G^*y=(-A^*)^{\beta/2}y_2,\qquad y\in\mathscr{D}((-A^*)^{\beta/2}G^*).\end{aligned} \tag{7.14}$$

Thus, $\mathscr{D}((-A^*)^{\beta/2}G^*)$ is *not dense* in Y_2, and condition (1.31) is *violated.*

On the other hand, since $B\Pi_1u(t)\equiv0$ and Y_2 is invariant under A and e^{At}, we obtain by (1.8) and (7.12),

$$\begin{aligned}GL_Tu&=G\int_0^T e^{A(T-t)}Bu(t)\,dt\\ &=G\int_0^T e^{A(T-t)}\cancel{B\Pi_1}u(t)\,dt+G\int_0^T e^{A(T-t)}B\Pi_2u(t)\,dt\\ &=G\int_0^T e^{A(T-t)}(-A)^\gamma\Pi_2u(t)\,dt=(-A)^\gamma\int_0^T e^{A(T-t)}\Pi_2u(t)\,dt,\end{aligned} \tag{7.15}$$

where in the last step we have used (7.12) on G, with the integral term in Y_2. Thus, GL_T is a *closed* operator (being the product of a closed, boundedly invertible operator $(-A)^\gamma$ and of a bounded operator [K1, p. 164]). Our claim is proved. Note that, by (1.9), one likewise has $\{L_T^*G^*y\}(t)=(-A^*)^\gamma e^{A^*(T-t)}y_2$, $y_2=\Pi_2y\in Y_2$.

7.3. Variational versus direct approach: assumption (1.46) of the direct approach fails, yet GL_T is closed. We have already noted that assumption (1.46) for the direct approach of [F4; Thm. 3.2] in the case where G is nonsmoothing, involves only the operators A and G, not B. Instead, the assumption of the variational approach of the present paper in Theorem 1.1 that GL_T be closeable involves all the data of the problem: G, A, and B. Thus, not surprisingly, we provide new classes of examples where *assumption* (1.46) *fails*, yet GL_T is *closed*. Thus, [F4; Thm. 3.2] is not applicable, while our present Theorem 1.1 is; and more generally our present §§ 2-5 are applicable. We return to the example of § 7.2 and set

$$(7.16) \qquad G_1 y = (y_1, v)_{Y_1} v, \quad G_1^* = G_1 = G_1^* G_1, \quad v \in Y_1, \quad |v| = 1,$$

where we recall that the subspaces Y_i are invariant under A and e^{At}. It follows from an observation in [F4; § 3.1] that

$$(7.17) \qquad \begin{gathered} G_1 \text{ in (7.16) satisfies assumption (1.46)} \\ \Leftrightarrow v \in \mathscr{D}((-A^*)^{\beta/2}) \cap Y_1 \quad \text{for some } \beta > 2\gamma - 1. \end{gathered}$$

Next, choose $v = b$, with $b \in Y_1$ the (normalized) vector defined in (7.10), satisfying (7.11). Thus, the operator G_1,

$$(7.18) \qquad G_1 y = (y_1, b)_{Y_1} b \text{ does not satisfy (1.46) on } Y_1.$$

Since Y_i are invariant under A, it follows that the operator G,

$$(7.19) \qquad \begin{gathered} G = G_1 + I_2 \text{ does not satisfy (1.46) on } Y; \ G_1 \text{ as in (7.18);} \\ I_2 = \text{identity on } Y_2. \end{gathered}$$

Yet, as seen in § 7.2, GL_T is *closed.*

Remark 7.3. The variational approach of this article (as well as the direct approach of [F4]) can be readily extended to allow G to be unbounded, say $G \in \mathscr{L}(\mathscr{D}(-A)^\rho, Z)$, $\rho > 0$, or $G(-A)^{-\rho} \in \mathscr{L}(Y; Z)$. Recalling the explicit formula (1.10) for the optimal $u^0(t, 0; x)$, we write accordingly,

$$\{L_T^* G^* G e^{AT} x\}(t) = B^*(-A^*)^{-\gamma}(-A^*)^\gamma e^{A^*(T-t)}(-A^*)^\rho G^* G(-A)^{-\rho}(-A)^\rho e^{AT} x$$

with $(-A)^\rho e^{AT} x \in Y$ for $x \in Y$, whereby the original γ deteriorates to $\gamma + \rho$ in terms of singularity. This will not be pursued further here.

Remark 7.4. (a) We have seen in § 7.4 that the operator G in (7.3) with ϕ as in (7.2) neither satisfies assumption (1.46) of the direct approach of [F4] (see Remark 7.2) nor does it make GL_T closeable (see below (7.4)).

(b) Suppose that $G(-A)^{\beta/2}$ is closeable for some $\beta > 2\gamma - 1$ equivalently that $(-A^*)^{\beta/2} G^*$ is densely defined, see (1.31). Then the assumptions of both approaches are satisfied; i.e., (b_1) GL_T is closeable and (b_2) assumption (1.46) holds true.

Statement (b_2) is proved in [F4; § 3.4.1]. Statement (b_1) was already noted in (1.31): since then $GL_T = G(-A)^{\beta/2} V_T$ is the product of a closeable operator and of a *bounded* operator

$$V_T u = \int_0^T (-A)^p e^{A(T-t)}(-A)^{-\gamma} Bu(t)\, dt, \ p = \gamma - \beta/2 < \tfrac{1}{2},$$

so that $V_T \in \mathscr{L}(L_2(0, T; U), Y)$.

REFERENCES

[B1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, Berlin, 1981.

[B2] ———, *Boundary control of parabolic equations. L-Q-R theory*, Proc. V. Internat. Summer School, Control Inst. Math. Mech. Acad. Sci. GDR, Berlin (1977).

[D] G. DA PRATO, *Quelques résultats d'existence, unicité et regularité pour un probleme de la théorie du contrôle*, J. Math. Pures Appl., 52 (1973), pp. 353-375.

[D-I] G. DA PRATO AND A. ICHIKAWA, *Riccati equations with unbounded coefficients*, Ann. Mat. Pura Appl., 140 (1985), pp. 209-221.

[DS] L. DESIMON, *Un applicazione della teoria degli integrali singolari allo studio delle equazioni differenziali lineari astratte del primo ordine*, Rendi. Sem. Mat. Univ. Padova, 34 (1964), pp. 205-223.

[F1] F. FLANDOLI, *Riccati equations arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984), 76-86.

[F2] ———, *A Riccati equation with unbounded coefficients and nonsmooth final data*, 1989, preprint.

[F3] ———, *A counterexample in the boundary control of parabolic systems*, Appl. Math. Lett., 2 (1989), pp. 341-343.

[F4] ———, *On the direct solutions of Riccati equations arising in boundary control theory*, Ann. Mat. Pura Appl., to appear.

[K1] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.

[L] I. LASIECKA, *Unified theory for abstract parabolic boundary problems: A semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287-333.

[L-T1] I. LASIECKA AND R. TRIGGIANI, *Dirichlet boundary control problem for parabolic equations with quadratic cost: Analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1983), pp. 41-67.

[L-T2] ———, *The regulator problem for parabolic equations with Dirichlet boundary control. Part* I, *Riccati's feedback synthesis and regularity of optimal solutions*, *Part* II, *Galerkin approximation*, Appl. Math. Optim., 16 (1987), pp. 147-168, 187-216.

[L-T3] ———, *Numerical approximations of algebraic Riccati equations for abstract systems modelled by analytic semigroups, and applications.* Math. Comp., to appear.

[L-T4] ———, *Algebraic Riccati Equations with application to boundary/point control problems: Continuous and approximation theory*, in Perspectives in Control Theory, Proceedings of the Sielpia Conference, Sielpia, Poland, September 1988; Birkhauser, Basel, 1989, pp. 175-210.

[L-T5] ———, book in progress.

[L-T6] ———, *Differential and Algebraic Riccati Equations with Applications to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes Inform. Control Sci. 164, Springer-Verlag, New York, 1991.

[Lio] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[P-S] A. PRITCHARD AND D. SALAMON, *The linear quadratic control problem for infinite dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., (1987), pp. 121-144.

[T] R. TRIGGIANI, *Boundary feedback stabilizability of parabolic equations*, Appl. Math. Optim., (1980), pp. 201-220.

[W] D. WASHBURN, *A bound on the boundary input map for parabolic equations with application to time optimal control*, SIAM J. Control, 17 (1979), pp. 652-671.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 482–504, March 1992

009

THE RECOVERY OF POTENTIALS FROM FINITE SPECTRAL DATA*

BRUCE D. LOWE[†], MICHAEL PILANT[†‡], AND WILLIAM RUNDELL[†‡]

Abstract. The reconstruction of a Sturm–Liouville potential from finite spectral data is considered. A numerical technique based on a shooting method determines a potential with the given spectral data. Convergence of reconstructed potentials is shown and numerical examples are considered.

Key words. inverse spectral theory, Sturm–Liouville problem

AMS(MOS) subject classification. 34B25

1. Introduction. The inverse Sturm–Liouville problem is one of the most celebrated of the class of inverse problems consisting of the determination of unknown coefficients in differential equations. The problem consists of recovering the potential function $q(x)$ in

$$-y'' + q(x)y = \lambda y, \tag{1.1}$$

$$y'(0) - hy(0) = 0, \tag{1.2}$$

$$y'(1) + Hy(1) = 0 \tag{1.3}$$

from a knowledge of spectral data. There are various versions of the problem depending on the exact nature of this data. However, in each of these, it is assumed that a complete spectrum, $\{\lambda_j\}_{j=1}^{\infty}$ for (1.1)–(1.3) is given. It is well known that this is, in itself, insufficient information for the recovery of q, and thus some additional information must be provided. There are exceptions to this, when, for example, the potential is known to be symmetric about the mid point and identical boundary conditions ($h = H$) are imposed at both ends. This is the so-called symmetric case. Some of the better known versions are described in the survey article [13] and the paper [18]. We shall briefly describe these in §5.

Questions of existence and uniqueness for the inverse spectral problem have been extensively studied over the past forty years and most of the fundamental questions answered. More recently the attention has shifted to the question of obtaining efficient reconstructive methods, and this paper continues the trend. In particular, we are interested in the question of obtaining approximations to $q(x)$ from limited data. For example, consider the symmetric case where the overposed data consists of the eigenvalues $\{\lambda_n\}$. In many physical problems of interest we are able to perform accurate measurements of only a relatively small number of the lowest eigenvalues, and the accuracy obtained degrades quite rapidly with higher modes. What if, instead of a complete set of eigenvalues, we are given only the first N eigenvalues $\{\lambda_n\}_{n=1}^{N}$? The answer to this question is known; uniqueness fails of course, but the difference between any two functions $q_1(x)$ and $q_2(x)$ that have $\{\lambda_n\}_{n=1}^{N}$ as their first N eigenvalues is a sum containing only the eigenfunctions of (1.1)–(1.3), $y(x; q_1, \lambda_n)$, and $y(x; q_2, \lambda_n)$ for $n > N$, [7], [9]. Since these eigenfunctions are quite "close to" the functions $\sin n\pi x$ (for Dirichlet conditions), the difference of q_1 and q_2 consists of functions

* Received by the editors December 6, 1990; accepted for publication June 3, 1991.

[†] Department of Mathematics, Texas A&M University, College Station, Texas 77843.

[‡] This research was supported in part by the Office of Naval Research under grant ONR N00014-89-J-1008.

whose frequency is at least N. This means that from finite spectral data we should be able to recover all the low frequency modes of the potential $q(x)$, but none higher than the number of eigenvalues given. It also suggests that the goal of a recovery method might be to assume that $q(x)$ has a finite trigonometric expansion and attempt to recover the Fourier coefficients from the spectral data. Indeed, this has been a popular approach and is further exploited in this paper. Thus our goal is to obtain information about the function $q(x)$ from limited spectral data—typically less than the first ten eigenvalues, and we shall demonstrate that this is quite feasible for a wide class of potentials.

For certain problems it is often the case that we are able to say that, despite the large number of techniques used in the solution, all of them start from the same basic formula or use variations within a given family of methods. This is certainly not the situation for numerical solutions to the inverse spectral problem (1.1), where there has been a wide diversity of approaches. There are methods that, because they rely on a particular representation, are only applicable to a particular type of data; there are other methods that are applicable to most versions of the inverse Sturm–Liouville family. Some methods recover the value of q at discrete points $\{x_i\}_1^M$ and require M items of spectral data; other methods place no restriction on the number of output points for the potential. We can look at this in another light; some approaches place restrictions on the types of basis expansions allowed in the approximation of $q(x)$. Some algorithms are more naturally solved by a direct method, while others require an iterative solution.

Some of the early work on the reconstruction of the potential relied on replacing the original boundary problem for an ordinary differential equation with a discrete version of the problem. This leads to the question of recovering a matrix with unknown diagonal elements from knowledge of its eigenvalues. For a discussion of this approach, including the historical perspective, see [5]. For the symmetric case with Dirichlet boundary conditions, Hald [8] developed an algorithm based on the Rayleigh–Ritz formula. This reduces to finding a matrix of a certain form, an approximation to the Sturm–Liouville operator, whose eigenvalues are prescribed. There is some overlap in the methodology between this work and the present paper: in the present paper we generate a potential whose first N eigenvalues agree with the prescribed eigenvalues $\{\lambda_j\}_{j=1}^N$. We also show that our scheme is applicable to other inverse spectral problems.

In the original paper by Gel'fand and Levitan [6], a formula is developed that has been used to provide a numerical scheme. This requires solving the integral equation

$$K(x,t) + \int_0^x K(x,s)f(s,t)ds = f(x,t) \qquad 0 \le t \le x \tag{1.4}$$

for the function $K(x,t)$ at each fixed $x \in [0,1]$. Here the function $f(x,t)$ is obtained from data of the spectral function type. The potential is then recovered from $q(x) = 2(d/dx)K(x,x)$. The main difficulty with the method, and one that limits its effectiveness, is the sensitivity of the solution of the integral equation on the data function $f(x,t)$. This is given as the difference of two series in which, although the combination converges, each component is rapidly divergent. This leads to difficulties in accurate computation of f. In addition, the method has a relatively high operational count and, at least in the usual formulation, is restricted to spectral data consisting of eigenvalues and norming constants. Despite these drawbacks, the approach has provided a starting place for other techniques. McLaughlin and Handelman [12] have given a method of solution based on the Gel'fand–Levitan integral equation. Sacks [19] has given an iterative method (actually a quasi-Newton method) that is based on the mapping from $K(x,x)$ to $K_t(x,0)$. This significantly reduces the operational count from the original Gel'fand–Levitan formulation and provides an efficient reconstruction of $q(x)$ from eigenvalue plus norming constant data. Recently, Rundell and Sacks [18] gave an approach based on a representation theorem of Gel'fand and Levitan. This method translated the spectral data into Cauchy data for a certain hyperbolic equation which in turn converted the original Sturm–Liouville problem into an overposed problem for this wave equation, and this could be solved by an iterative procedure. This approach leads to a scheme with considerable flexibility in the type of overposed data it can handle and in the type of basis expansions for $q(x)$. The algorithm is extremely robust and has a very low operational count. Hald [7] showed that an algorithm, based on the work of Hochstadt [9], will always provide a solution of the inverse problem in the symmetric case. More recently, Andersson [1] has extended these techniques to the so called "impedance case," the recovery of the function $p(x)$ in $-(py')' = \lambda py$. There are extremely elegant characterizations of the isospectral sets corresponding to Dirichlet data, that is, the set of potentials $q(x)$ that have a given (Dirichlet) spectrum, [17]. These ideas can be used to provide constructive algorithms. Finally, Paine [16] has shown that a Newton type method can be used in the symmetric case. This relies on finding a solution of the potential to eigenvalue map, and there is some overlap with the ideas in the present paper. In all overposed problems, there is a certain latitude in deciding which items of data are "basic" to the problem and which are "additional." Put another way, the question is: what is considered to be the underlying direct problem; that is, the problem of recovering $y(x)$ in (1.1)–(1.3). It is more natural perhaps to consider the boundary conditions as "basic" and the eigenvalues to be the extra information used in the recovery of q. This was the approach taken by Paine. However, for a given $q(x)$ and eigenvalue sequence we can determine the eigenfunctions completely using only left endpoint data. This leaves the boundary condition imposed at the right endpoint to be used as the additional information for the recovery of q, and is the method of the present paper. At the start of this paragraph we mentioned the diversity of approaches to the problem. There is one thing that seems to be common to all of the above approaches—their restriction of applicability to one space variable. In most cases this is inherent in the method. For example, those methods based on the Gel'fand–Levitan representation are limited due to the fact that this only holds in one space dimension. The method of the present paper can, at least formally, be applied in higher dimensions, and this is one of the motivations for our study.

The plan of the paper is as follows. In the next section we will present our method for the symmetric case with Dirichlet boundary conditions. We shall show that our algorithm is always well defined and the associated iteration scheme converges to a function in a well-defined approximating set. In §3 we consider the two spectrum case for a general nonsymmetric potential. In §4 we consider the problem of the convergence of the approximations. Finally, in §5, we present some numerical results and discuss some of the other inverse Sturm–Liouville problems, indicating the minor modifications required to the algorithm.

2. Single spectrum case for a symmetric potential.

The operator

$$L \equiv -\frac{d^2}{dx^2} + q(x) \tag{2.1}$$

with unknown symmetric potential $q(x)$ on $[0,1]$ can be uniquely determined from the Dirichlet spectrum [3], or equivalently, the zeros of the entire function $u(1;q,\lambda)$, where $u(x;q,\lambda)$ satisfies

$$\begin{aligned} -u''(x;q,\lambda)+q(x)u(x;q,\lambda)&=\lambda u(x;q,\lambda),\\ u(0;q,\lambda)&=0,\\ u'(0;q,\lambda)&=1. \end{aligned} \tag{2.2}$$

Knowledge of the first N Dirichlet eigenvalues $\lambda_1,\cdots,\lambda_N$ will not uniquely determine the operator L. Indeed, expanding $q(x)$ into a Fourier cosine series

$$q(x)=q_0+\sum_{k=1}^{\infty} q_k\cos(2\pi kx),$$

we can at best hope to determine the first N coefficients $\{q_0,\cdots,q_{N-1}\}$. However, for N not too large, this should be adequate to approximate a sufficiently smooth $q(x)$. It is natural to seek an approximation

$$q^N(x)=q_N+\sum_{k=1}^{N-1} q_k\phi_k(x)$$

that lies in the space

$$\mathcal{S}_N\equiv \text{ span }\{1,\phi_k(x)=\cos(2\pi kx),k=1,\cdots,N-1\},$$

and for which $\lambda_1,\cdots,\lambda_N$ are zeros of $u(1;q^N(x),\lambda)$.

Let $\Lambda_N\equiv(\lambda_1,\cdots,\lambda_N)$ and define $F:\mathcal{R}^N\to\mathcal{R}^N$ by

$$F_j(\Lambda_N;q_1,\cdots,q_N)\equiv u_j(1;q_1,\cdots,q_N,\lambda_j)\equiv u\left(1;q_N+\sum_{i=1}^{N-1} q_i\cos(2\pi ix),\lambda_j\right). \tag{2.3}$$

We seek a vector $\vec{q}\equiv(q_1,\cdots,q_N)$ for which $F(\Lambda_N;q)=0$. When no ambiguity can occur, we drop the explicit vector notation. Theorem 1 shows that this is possible for $q(x)$ sufficiently close to zero.

The proof of Theorem 1 relies on the following result.

LEMMA 1. *Let $\Lambda_{N,0}=(\pi^2,4\pi^2,\cdots,N^2\pi^2)$ denote the first N Dirichlet eigenvalues corresponding to $q(x)=0$. The Jacobian $F_q(\Lambda;q)|_{\Lambda=\Lambda_{N,0},q=0}$ is upper triangular with elements*

$$\frac{\partial F_j}{\partial q_j}(\Lambda;q)|_{\Lambda=\Lambda_{N,0},q=0}=\frac{(-1)^j}{4\pi^2 j^2},\qquad j=1,\cdots,N-1,$$

$$\frac{\partial F_j}{\partial q_N}(\Lambda;q)|_{\Lambda=\Lambda_{N,0},q=0}=\frac{(-1)^{j+1}}{2\pi^2 j^2},\qquad j=1,\cdots,N$$

and with all the other elements zero.

As a direct consequence, the eigenvalues of $F_q(\Lambda_{N,0};0)$ are given by $(-1)^j/(4\pi^2j^2)$ for $j=1,\cdots,N-1$ and $(-1)^{N+1}/(2\pi^2N^2)$. The spectral condition number of $F_q(\Lambda_{N,0};0)$ is therefore

$$\mu\equiv\frac{|\lambda_{\max}|}{|\lambda_{\min}|}=\frac{N^2}{2}.$$

Proof. Using (2.2), it is easily verified that $u(x;0,\lambda_j) = \sin\sqrt{\lambda_j}x/\sqrt{\lambda_j}$. Furthermore, $w \equiv (\partial u_j/\partial q_k)(x;0,\lambda_j)$ satisfies the ordinary differential equation

$$\begin{aligned} -w'' &= \lambda_j w - \phi_k(x)u(x;0,\lambda_j), \\ w(0) &= 0 = w'(0) \end{aligned} \tag{2.4}$$

which has the solution

$$w(x) = \int_0^x u(x-t;0,\lambda_j)\phi_k(t)u(t;0,\lambda_j)dt. \tag{2.5}$$

Consequently, the jkth element of the Jacobian matrix is given by

$$\frac{\partial u_j}{\partial q_k}(1;0,\lambda_j) = \int_0^1 \frac{1}{\lambda_j}\sin(\sqrt{\lambda_j}(1-t))\sin(\sqrt{\lambda_j}t)\phi_k(t)dt, \tag{2.6}$$

where $\lambda_j = j^2\pi^2$. The result follows easily from a direct integration. □

The special structure of the Jacobian arises from the particular choice of the Fourier basis.

THEOREM 1. *For fixed N and $q(x) \in \mathcal{S}_N$ sufficiently small, $F(\Lambda_N, q)$ has a unique zero, denoted by $q^N(x)$.*

Proof. The Dirichlet eigenvalues of L depend continuously on L^∞ perturbations of the potential [17]. Consequently, for $q(x)$ near zero, Λ_N is a perturbation of $\Lambda_{N,0}$. Since $F(\Lambda_{N,0},0) = 0$, $F_q(\Lambda_{N,0},0)$ is nonsingular, and the function $F(\Lambda_N,q)$ depends smoothly on Λ_N and q, the result follows from the Implicit Function Theorem. □

Several comments are in order. First, the implicit function theorem is not constructive. Second, the theorem does not provide us with a uniform estimate on $q^N(x)$. We require such an estimate in order to obtain a convergence proof, and we derive this estimate in Lemma 4.

In order to actually determine the basis coefficients of $q^N(x)$, $(q_1^N, q_2^N, \cdots, q_N^N)$, we iterate using Newton's method, which yields the scheme

$$\begin{aligned} F_q(\Lambda_N; q^{(m)})\delta q^{(m)} &= -F(\Lambda_N; q^{(m)}), \\ q^{(m+1)} &= q^{(m)} + \delta q^{(m)}. \end{aligned} \tag{2.7}$$

If $q(x)$ is sufficiently close to zero, then $F_q(\Lambda_N; q_1^N, q_2^N, \cdots, q_N^N)$ is nonsingular and the scheme converges quadratically to $(q_1^N, q_2^N, \cdots, q_N^N)$ for all initial guesses which are sufficiently close. If instead of evaluating the Jacobian at $q^{(m)}$, we evaluate it at $q = 0$, we obtain a quasi-Newton scheme. This is more computationally efficient in practice.

Let us now consider the case where the mean of the potential is known. This can in fact be recovered from the spectral data, see [18]. When $\alpha \equiv \int_0^1 q(x)dx$, the problem reduces to finding $q^N(x) \in$ span $\{\cos(2\pi ix), i = 1, \cdots, N\}$ for which

$$u(1; q^N(x), \lambda_j - \alpha) = 0 \quad \text{for } j = 1, \cdots, N. \tag{2.8}$$

If we define $\bar{F} : \mathcal{R}^N \to \mathcal{R}^N$ by

$$\bar{F}_j(\Lambda_N, q_1, \cdots, q_N) = u\left(1; \sum_{r=1}^N q_r\cos(2\pi rx), \lambda_j\right), \qquad j = 1, \cdots, N$$

with

$$\Lambda_N = (\lambda_1 - \alpha, \cdots, \lambda_N - \alpha),$$

then the following lemma holds.

LEMMA 2.

$$\frac{\partial \bar{F}_j}{\partial q_k} = \frac{\partial u}{\partial q_k}(1; 0, \lambda_j) = \begin{cases} \dfrac{\sin\sqrt{\lambda_j}}{2\sqrt{\lambda_j}(\lambda_j - \pi^2 k^2)} & \textit{if } \dfrac{1}{\pi}\sqrt{\lambda_j} \neq \textit{ integer}, \\ 0 & \textit{if } \dfrac{1}{\pi}\sqrt{\lambda_j} = \textit{ integer } (\neq k), \\ \dfrac{(-1)^k}{4\pi^2 k^2} & \textit{if } \dfrac{1}{\pi}\sqrt{\lambda_j} = k. \end{cases}$$

Using the estimate $\lambda_j = \pi^2 j^2 + c_j$ where $c_j \in l^2(n)$ [17], the following theorem gives a sufficient condition for $\bar{F}_q(\Lambda_N, 0)$ to be nonsingular. This is particularly useful when employing the quasi-Newton method.

THEOREM 2. *If* $\Lambda_N = \{\lambda_j\}_{j=1}^{j=N}$, $\lambda_j = j^2\pi^2 + c_j$ *and* $|c_j| \leq \min\{C(j/\sqrt{\lambda_j}), \pi^2/4\}$, *then for* C *independent of* N *and sufficiently small*

$$\det \bar{F}_q(\Lambda_N, 0) \neq 0.$$

Proof. The proof of Theorem 2 follows from some simple inequalities which imply diagonal dominance.

Setting $a_j \equiv \sqrt{\lambda_j} - j\pi = \sqrt{\pi^2 j^2 + c_j} - \pi j$, a first-order MacLaurin expansion gives

$$|a_j| \leq \frac{c_j}{\sqrt{3}\pi j} \leq \frac{C}{\sqrt{3}\pi\sqrt{\lambda_j}}.$$

Summing the off-diagonal entries of the Jacobian matrix yields

$$\begin{aligned} \sum_{j\neq k} \left|\frac{\partial \bar{F}_j}{\partial q_k}(\Lambda_N, 0)\right| &= \sum_{j\neq k} \frac{|\sin\sqrt{\lambda_j}|}{2\sqrt{\lambda_j}|\lambda_j - \pi^2 k^2|} \\ &= \sum_{j\neq k} \frac{|\sin(a_j)|}{2\sqrt{\lambda_j}|\lambda_j - \pi^2 k^2|} \leq \sum_{j\neq k} \frac{|a_j|}{2\sqrt{\lambda_j}|\lambda_j - \pi^2 k^2|} \\ &\leq \frac{C}{2\sqrt{3}\pi} \sum_{j\neq k} \frac{1}{\lambda_j|\lambda_j - \pi^2 k^2|} \leq \frac{C}{2\sqrt{3}\pi^3 k^2}\left[\sum_{j\neq k} \frac{1}{\lambda_j} + \frac{1}{|\lambda_j - \pi^2 k^2|}\right]. \end{aligned}$$

With the estimates

$$\sum_{j\neq k} \frac{1}{\lambda_j} \leq \sum_{j\neq k} \frac{1}{\pi^2(j^2 - \frac{1}{4})} \leq \frac{1}{\pi^2}\sum_{j=1}^{\infty} \frac{2}{j^2} = \frac{1}{3}$$

and

$$\sum_{j\neq k} \frac{1}{|\pi^2 k^2 - \lambda_j|} \leq \sum_{j\neq k} \frac{1}{\pi^2|j^2 - k^2 - \frac{1}{4}|} \leq \frac{1}{3} + \frac{2(2+\sqrt{3})}{3\pi^2},$$

we obtain

$$\sum_{j\neq k} |\frac{\partial \bar{F}_j}{\partial q_k}(\Lambda_N, 0)| \leq \frac{C}{3\sqrt{3}\pi^3 k^2}\left[1+\frac{2+\sqrt{3}}{\pi^2}\right].$$

By a simple argument, we can show that the diagonal element satisfies

$$|\frac{\partial \bar{F}_i}{\partial q_i}(\Lambda_N, 0)| \geq \frac{1}{4\pi^2 i^2}\,\frac{96-24\sqrt{3}-4\sqrt{3}\pi}{97-12\sqrt{3}},$$

and the matrix will be diagonal dominant if

$$C < \frac{-54\pi^3+72\sqrt{3}\pi^3-9\pi^4}{158+73\sqrt{3}+97\pi^2-12\sqrt{3}\pi^2} \approx 1.26918.$$

We remark that the constant $C \approx 1$ corresponds to extremely large potentials.

3. Two spectrum case for a general potential. In the general case of a potential $q(x) \in L^2[0,1]$, we choose the $2N$ basis functions

$$\{\phi_k(x)\}|_{k=1}^{k=2N} = \{\sin 2\pi x, \cos 2\pi x, \cdots, \sin 2N\pi x, \cos 2N\pi x\}. \tag{3.1}$$

As in the previous section, we have assumed (by suitable renormalization) that the potential has zero mean.

In order to determine $q(x)$ uniquely, we must give two sets of spectra which, for purposes of this section, we suppose are Dirichlet–Dirichlet and Dirichlet–Neumann spectra:

$$\Lambda = \{\lambda_j, \mu_j\}|_{j=1}^{j=\infty},$$

where

$$-u'' + q(x)u = \lambda_j u, \quad u(0) = 0 = u(1), \quad u'(0) = 1$$

and

$$-v'' + q(x)v = \mu_j v, \quad v(0) = 0 = v'(1), \quad v'(0) = 1.$$

As before, we define u_j and v_j, respectively, as solutions of the initial value problems (for a given $q(x)$)

$$-u_j'' + q(x)u_j = \lambda_j u_j, \quad u_j(0) = 0, \quad u_j'(0) = 1$$

and

$$-v_j'' + q(x)v_j = \mu_j v_j, \quad v_j(0) = 0, \quad v_j'(0) = 1.$$

We have

$$\frac{\partial u}{\partial q_k}(1; 0, \lambda_j) = \int_0^1 u(1-t; 0, \lambda_j)\phi_k(t)u(t; 0, \lambda_j)dt,$$

$$\frac{\partial v}{\partial q_k}(1; 0, \lambda_j) = \int_0^1 v(1-t; 0, \lambda_j)\phi_k(t)v(t; 0, \lambda_j)dt,$$

which can be explicitly computed. An argument similar to that of Theorem 2 yields the following.

THEOREM 3. *If the Dirichlet–Dirichlet and Dirichlet–Neumann spectra for a potential of mean zero are sufficiently close to $q = 0$ spectra, in the sense that*

$$|\sqrt{\lambda_j} - j\pi| \leq C/\sqrt{\lambda_j}$$

and

$$|\sqrt{\mu_j} - \left(j - \frac{1}{2}\right)\pi| \leq C/\sqrt{\mu_j}$$

with C sufficiently small, and if ϕ_k is the Fourier basis for $L^2[0,1]$, then there is a unique $q^N \in$ span $\{\phi_k\}_{k=1}^{k=2N}$ with the prescribed spectra.

The coefficients of $q(x)$, with respect to the basis ϕ_k, can be recovered by either the Newton or quasi-Newton procedure given in §2.

4. A convergence theorem. The following heuristic argument justifies, in part, the excellent approximation properties of our algorithm. For a symmetric potential, the actual solution of (1.1)–(1.3) satisfies

$$(4.1) \qquad 0 = u_j(1; q, \Lambda_N), \qquad j = 1, 2, \cdots, N,$$

while the approximation $q^N \in S_N$ satisfies

$$(4.2) \qquad 0 = u_j(1; q^N, \Lambda_N), \qquad j = 1, 2, \cdots, N.$$

Expanding (4.1), we have

$$(4.3) \quad 0 = F(\Lambda_N, q) = F(\Lambda_N, q^N) + \frac{\partial F}{\partial q}(\Lambda_N, q^N) \cdot (q - q^N) + \mathcal{O}(|q - q^N|^2).$$

If $\det(\partial F/\partial q)(\Lambda_N, q^N) \neq 0$ and $0 = F(\Lambda_N, q^N)$, then we have to first order $q - q^N \notin S_N$. If we chose an orthonormal basis for S_N, we therefore have (to leading order) $q - q^N \perp S_N$; that is, q^N is very close to the best L^2 approximation to q in that space. This is in fact shown by the numerical computations given in §5. This also motivates the choice of the Fourier basis.

For the remainder to this section we shall assume that $q(x)$ is a symmetric L^2 potential with mean zero and spectrum $\{\lambda_j\}_{j=1}^{\infty}$, satisfying the following.

Assumption A. For some constants $\alpha, \beta > 0$ the eigenvalues of q satisfy $|\lambda_j - \pi^2 j^2| \leq \beta/(j^{1+\alpha})$.

We show that for β sufficiently small, $q^N(x)$ generated by the shooting method of §2 converges in L^2 to $q(x)$. Assumption A implies that $q(x) \in L^\infty$, [17, p. 39].

Consider the map $F : V_N \to \mathbf{R}^N$ given by

$$F_j(\Lambda_N, q) = 4(-1)^j \lambda_j u(1; q, \lambda_j), \qquad j = 1, \cdots, N,$$

where $\Lambda_N = (\lambda_1, \cdots, \lambda_N)$, $V_N = \{q \in \text{span}(\phi_k, k = 1, \cdots, N), \|\vec{q}\|_{L^1} \leq 1\}$, and $u(x; q, \lambda_j)$ solves (2.2). Here $\phi_k(t) = \cos 2\pi k t$. For β sufficiently small, we show that $q^N(x) \in V_N$ for all N. This readily follows once a Lipschitz bound independent of N is established for the Jacobian of F. This is given in Lemma 3, the proof of which can be found in the appendix. In what follows, C denotes a constant that is independent of k, β, and N.

LEMMA 3. *For $q^{(i)} \in V_N$ and all β sufficiently small*

$$\sum_{j=1}^{N}\left|\frac{\partial F_j}{\partial q_k}(\Lambda_N,q^{(1)}) - \frac{\partial F_j}{\partial q_k}(\Lambda_N,q^{(2)})\right| \le C\|q^{(1)} - q^{(2)}\|_{L^2}.$$

LEMMA 4. *If the eigenvalues λ_n satisfy* (A), *then for all β sufficiently small, $\|q^N\|_{L^2} \le 1$.*

Proof. By Lemma 3, $\|F'(\Lambda_N,q^{(2)}) - F'(\Lambda_N,q^{(1)})\|_{L^1} \le C\|q^{(1)}(\cdot) - q^{(2)}(\cdot)\|_{L^2} \le C\|\vec{q}^{(1)} - \vec{q}^{(2)}\|_{L^1}$ for $q^{(i)} \in V_N$, where C is independent of N. It is easy to show that $F'(\Lambda_N,0) = I - H_N$, where $\|H_N\|_{L^1} < \frac{1}{2}$ for β sufficiently small. Consequently, $\|F'(\Lambda_N,0)^{-1}\|_{L_1} < 2$ for all N and $\|F'(\Lambda_N,0)^{-1}F(\Lambda_N,0)\|_{L_1} \le 2\|F(\Lambda_N,0)\|_{L_1} \le 8\sum_{j=1}^{\infty}|\sqrt{\lambda_j}\sin\sqrt{\lambda_j}| \le C\beta\{\sum_{j=1}^{\infty}j^{-1-\alpha}\}$, which can be made arbitrarily small by choosing β appropriately. Using Newton's iteration, [4, p. 157], $\|\vec{q}^N\|_{L^1} \le 1$, giving $\|q^N\|_{L^\infty} \le 1$ and consequently $\|q^N\|_{L^2} \le 1$ for all N.

The convergence of $q^N(x)$ to $q(x)$ follows from the asymptotics of the eigenvalues of $q^N(x)$ and the continuity of the eigenvalue to potential map.

THEOREM 4. *If the eigenvalues λ_n satisfy* (A), *then for β sufficiently small q^N converges strongly to q in L^2.*

Proof. The eigenvalues $\lambda_k(q^N)$ satisfy the estimate

$$\left|\lambda_k(q^N) - \pi^2k^2 - \int_0^1(\cos 2\pi kx)q^N(x)dx\right| \le \frac{C}{k},$$

where C is independent of N [17, p. 35]. For N sufficiently large, $\lambda_k(q^N) = \lambda_k$ for $k = 1,\cdots,N$, by the counting lemma of [17, p. 28] and for $k \ge N+1$, $\langle q^N, \cos 2\pi kx\rangle_{L^2} = 0$, giving $|\lambda_k(q^N) - \pi^2k^2| \le C/k$. Now

$$\begin{aligned}\sum_{k=1}^{\infty}(\lambda_k(q^N) - \pi^2k^2)^2 &= \sum_{k=1}^{N}(\lambda_k - \pi^2k^2)^2 + \sum_{k=N+1}^{\infty}(\lambda_k(q^N) - \pi^2k^2)^2\\ &\le \sum_{k=1}^{\infty}(\lambda_k - \pi^2k^2)^2 + C\sum_{k=N+1}^{\infty}\frac{1}{k^2}\\ &\le \beta^2\sum_{k=1}^{\infty}\frac{1}{k^{2(1+\alpha)}} + C\sum_{k=N+1}^{\infty}\frac{1}{k^2}.\end{aligned}$$

For β sufficiently small and N sufficiently large, Hald's continuous dependence result [8] is applicable, giving

$$\begin{aligned}\|q^N - q\|_{L^2}^2 &\le C\sum_{k=1}^{\infty}(\lambda_k(q^N) - \lambda_k)^2 = C\sum_{k=N+1}^{\infty}(\lambda_k(q^N) - \lambda_k)^2\\ &\le C\left\{\sum_{k=N+1}^{\infty}(\lambda_k(q^N) - \pi^2k^2)^2 + \sum_{k=N+1}^{\infty}(\lambda_k - \pi^2k^2)^2\right\}\\ &\le C\left\{\sum_{k=N+1}^{\infty}\frac{1}{k^2} + \beta^2\sum_{k=N+1}^{\infty}\frac{1}{k^{2(1+\alpha)}}\right\}\\ &\le \frac{C}{N}.\end{aligned}$$

Clearly, $q^N \to q$ in L^2 with convergence rate of order half.

5. Numerical experiments. In this section we present some numerical experiments that we carried out to illustrate the performance of the algorithm. We will do so in three particular cases of the inverse Sturm–Liouville problem: symmetric $q(x)$; data consisting of two spectra; data consisting of the function $\lambda_j(H)$. We shall be able to point out some of the features of the method and be able to show its strengths as well as its weaknesses.

In order to construct data for the direct problem, that is, for given $q(x)$, the eigenvalues (and eigenfunctions if appropriate) we used the routine SLEIGN which is described in [2].

For the remainder of this section we will adopt the following notations. The function $u_j = u_j(x; q, \lambda)$ is the solution of the initial value problem

$$-u_j''(x) + q(x)u_j(x) = \lambda u_j(x), \qquad u_j(0) = 1, \quad u_j'(0) = h, \tag{5.1}$$

and the Frêchet derivative of u_j, with respect to q in the direction $\phi_k(x)$, we denote by $\hat{u}_{jk}(x)$. This satisfies

$$\begin{gathered} -\hat{u}_{jk}''(x) + q(x)\hat{u}_{jk}(x) = \lambda\hat{u}_{jk}(x) - \phi_k(x)u_j(x; q(x), \lambda), \\ \hat{u}_{jk}(0) = 0, \quad \hat{u}_{jk}'(0) = 0 \end{gathered} \tag{5.2}$$

with $\lambda = \lambda_j$ for $1 \leq j \leq N$ and $\phi_k(x)$ the kth element of the basis set, $1 \leq k \leq B$. If the boundary condition at the left endpoint is of Dirichlet type, then we modify the above in the obvious way, choosing $u_j(0) = 0$ and normalizing by $u_j'(0) = 1$. We denote by A_{jk} the $N \times B$ matrix with elements $\hat{u}_{jk}$. The boundary condition at $x = 1$ we shall denote by

$$\mathcal{B}[u] \equiv u'(1) + Hu(1) = 0 \tag{5.3}$$

and the resolution of the (finite-dimensional) inverse Sturm–Liouville problem requires that there be q such that $\mathcal{B}[u(\cdot; q; \lambda)] = 0$ for $\lambda = \lambda_j$, $1 \leq j \leq N$ for the given eigenvalues $\{\lambda_j\}$. Given the mth iterative approximation for $q(x)$, we update by

$$\mathcal{B}\left[u_j[q^{(m)}] + \frac{\partial u_j}{\partial q_k}[q^{(m)}].\delta q^{(m)}\right] = 0.$$

Since $\mathcal{B}$ is linear, we have

$$\mathcal{B}\left[\hat{u}_{jk}[q^{(m)}]\right].\delta q^{(m)} = -\mathcal{B}\left[u_j[q^{(m)}]\right]$$

as our update strategy. In each of the computations of $u_j(x)$ and $\hat{u}_{jk}(x)$, the current approximation $q^{(m)}$ is used. We used a Runge–Kutta–Fehlberg adaptive routine to perform the integration steps for the solutions of the initial value problems.

For the case of the recovery of a symmetric potential from the sequence of eigenvalues λ_j, $1 \leq j \leq N$, the algorithm is outlined in Fig. 1. Note that in the general symmetric case, symmetry is also required of the boundary data so that $\mathcal{B}[u] = u'(1) + hu(1)$.

Note that this algorithm does not require any separate input of the value of the mean $\int_0^1 q(s)ds$; the assumption of zero mean was only to simplify the analysis.

In all computations to be described we used the quasi-Newton version of the algorithm; in every example we tried, the differences in each iterate between the Newton

```
set $\mathcal{B}[u] = u'(1) + hu(1)$
set $q_0(x)$
do for $1 \le n \le max\ iterations$ {
    do for $1 \le j \le N$ {
        ode compute( $u_j$ )
        $b_j := \mathcal{B}[u_j]$
        do for $1 \le k \le B$ {
            ode compute( $\hat{u}_{jk}$ )
            $A_{jk} := \mathcal{B}[\hat{u}_{jk}]$
        }
    }
    svd matrix $A_{jk}$
    zero out small singular values
    solve $A_{jk} c_k = -b_j$
    $\delta q(x) := \sum c_k \phi_k(x)$
    $q_n(x) = q_{n-1}(x) + \delta q(x)$
}
```

FIG. 1. *Algorithm flowchart.*

and the quasi-Newton was within 1 percent. The difference in program running time was, of course, significant.

An initial approximation $q_0(x)$ is required, and we took this to be the zero function. Provided reasonable values were used, the algorithm was not sensitive to the choice of q_0.

We found that the number of iterations of the procedure before effective convergence of the functions q_n was obtained was always very small; most of the reconstruction coming in the first iteration, and the remainder in the next two.

The reason for the phenomena described in each of the three previous paragraphs is that the mapping from a given potential $q(x)$ to the spectral data is apparently very nearly linear. This observation was made in [18] where similar results were obtained.

The basis functions $\phi_k(x)$, $1 \le k \le B$ should be selected to match the expected functions $q(x)$. We used trigonometric functions, linear splines, and cubic splines. The algorithm allows the number of data items N to be different from the number of basis functions B. Usually one would choose $B = N$ to maximize the amount of information utilized. As the previous sections showed for the case of trigonometric basis functions and λ_j corresponding to the first N eigenvalues for the symmetric potential, we can simply invert the system $Ac = b$ in order to solve for the basis coefficients c_k of the function δq. The matrix remains well conditioned for all reasonable values of N. For other basis functions and spectral problems this is not always the case. In addition, we may have the situation where the spectral data is subject to measurement error and we wish to give additional data, that is, we have $N > B$. This is where the singular value decomposition is employed. We can use this to edit out the small singular values (those values σ_k with $\sigma_k < \sigma_{\max} TOL$) before back substituting to obtain the solution vector c_k. For general values of N and B we are therefore finding a least squares fit to this solution vector from the given data. All the runs shown below were made with TOL in the range $(10^{-6}, 10^{-4})$. In the case of the recovery problem with data consisting of the values of $\lambda_1(H_j)$ for N values of the parameter H, we were never able to recover functions with more than about six significant modes, and the use of the least squares approach becomes an essential feature of the method.

As we shall show, the algorithm outlined above can be extended to those spectral problems that consist entirely of eigenvalue data or eigenvalue data plus some pointwise information about the eigenfunction. For convenience we describe here those problems for which a uniqueness result for the inverse problem is known. In most of these cases the modification required to the algorithm consists simply of using the correct boundary operator $\mathcal{B}$ and perhaps a modification to the basis set.

Two spectrum case. Here we must recover the potential $q(x)$ from a double sequence of eigenvalues $\{\lambda_j^{(1)}, \lambda_j^{(2)}\}$, where we let $\{\lambda_j^{(1)}\}$ denote the eigenvalues corresponding to the boundary condition $u'(1) + H_1 u(1)$ and $\{\lambda_j^{(2)}\}$ denote the eigenvalues corresponding to the boundary condition $u'(1) + H_2 u(1)$. If there are $N/2$ eigenvalues from each set, and B basis functions, then A is again an $N \times B$ matrix whose entries are the values of $\mathcal{B}[\hat{u}_{jk}]$, where $H = H_1$ for $j \le N/2$ and $H = H_2$ for $N/2 < j \le N$.

Partially known q. If the function $q(x)$ is known on the interval $[0, \frac{1}{2}]$ and one complete spectrum is provided, then it is known that $q(x)$ can be uniquely determined over the remainder of the interval, [10]. To modify the basic algorithm we only need modify the basis set to consist of functions that span the interval $[\frac{1}{2}, 1]$.

Endpoint data. Here we let $\{\lambda_j\}$ denote the $N/2$ eigenvalues corresponding to the boundary condition $u'(1) + Hu(1)$. The data also includes the $N/2$ numbers k_j which are the values of the jth eigenfunction at $x = 1$, which, in view of the known boundary condition at the right-hand endpoint, also determines the values of $u_j'(1)$. Thus, in Fig. 1 we set the vector b_j to be

$$b_j = \begin{cases} u_j(1) - k_j & \text{if } j \le N/2, \\ u_j'(1) + Hk_j & \text{if } j > N/2. \end{cases}$$

Variable boundary condition data. In this version of the inverse Sturm–Liouville problem, we are given the value of $\lambda_M(H)$ as a function of H for some *fixed* eigenvalue number M, and where H runs over some interval. The algorithm is, thus, virtually identical to that depicted in Fig. 1. Suppose the data is of the form $\{H_j, \lambda_j\}$, where λ_j denotes the value of (say) the smallest eigenvalue of the problem when $H = H_j$. Then for each j we set $\mathcal{B}[u] = u_j'(1) + H_j u_j(1)$. Of course, the basis set must include both even and odd functions on the interval.

Figure 2a shows a reconstruction of the symmetric potential

$$q(x) = 1 - \exp(-20(x - 0.5)^2)$$

from Dirichlet eigenvalues and $B = N = 3$. The supremum norm difference between the reconstructed q and the actual q is 0.038.

With $B = N = 5$, this error is reduced to 0.004. In this and subsequent such graphs, the actual function is represented by a dashed line.

The asymptotic formula for the (Dirichlet) eigenvalues is given by $\lambda_n = n^2\pi^2 + \int_0^1 q(s)ds + c_n$, where c_n is in ℓ^2. As was shown in [14] and [18], the appropriate measure of accuracy for the data are the numbers c_n; we can expect to recover the potential from a reasonably small error in the c_n, but we can never expect to do so from even a few percent error in the eigenvalues themselves. We added 10 percent random

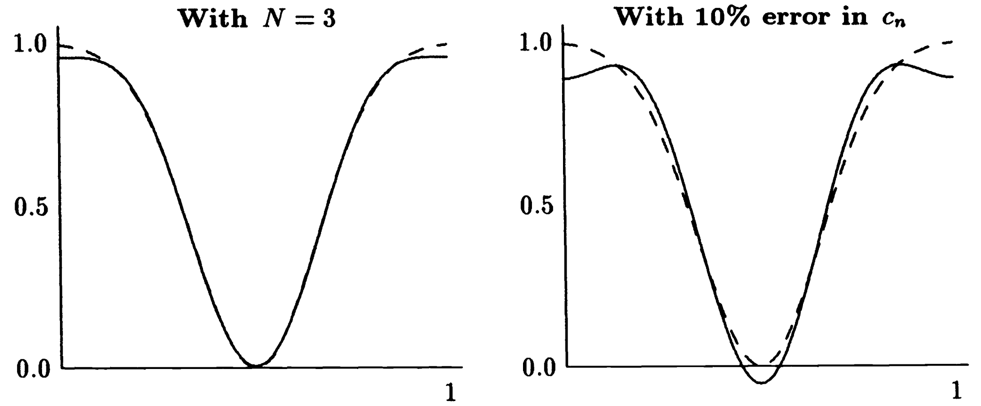


FIG. 2. *Reconstruction of a symmetric $q(x)$ from Dirichlet spectra.*

error to the numbers $\lambda_n - n^2\pi^2 - \int_0^1 q(s)ds$ and used the result in the recovery routine. Figure 2b shows a reconstruction under these conditions using the first five Dirichlet eigenvalues. Notice that in both these figures the reconstruction is much poorer at the endpoints of the interval. This is due to the fact that the actual potential is not a smooth periodic function on the real line, having discontinuities in the derivative at the points $x = 0, 1$, whereas any element in the span of the basis set is smooth. Similar effects can be expected in any method that uses a finite trigonometric basis, [18].

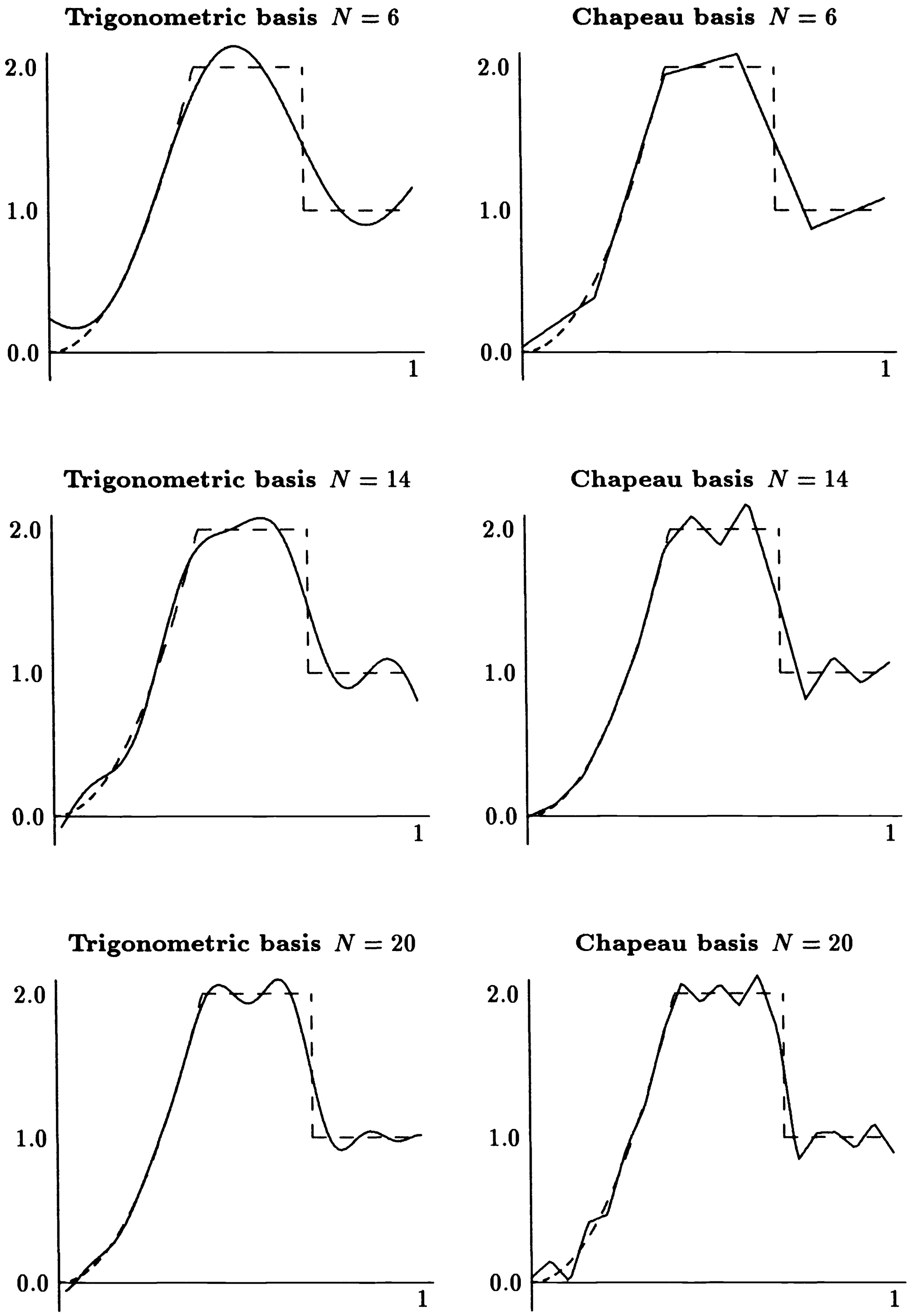
When the data consisted of two spectra we used as a test case the function

$$q(x) = \begin{cases} 12.5x^2 & \text{if } 0 \le x \le 0.4 \\ 2 & \text{if } 0.4 \le x \le 0.7 \\ 1 & \text{if } 0.7 \le x \le 1 \end{cases}$$

and provided both Dirichlet–Dirichlet ($h = H = \infty$) and Dirichlet–Neumann ($h = \infty$, $H = 0$) eigenvalues. Figure 3 shows the reconstructions for $N/2$ eigenvalues from each problem where $N = 6, 14, 20$. Both trigonometric $\{1, \cos(2\pi x), \sin(2\pi x), \cdots\}$ and linear spline (chapeau function) basis were used with $B = N$.

These computations were obtained by zeroing out some of the singular values. As a general rule we found that the further the function deviates from a smooth function when periodically extended, the larger the tolerance required to prevent overshoot at the endpoints. The case of $N = 20$ was run with TOL set to 10^{-4}; with this setting, only the first 14 trigonometric basis functions were retained. A lower value of TOL will result in the algorithm overemphasizing the higher modes in order to compensate for the discontinuity in the periodic function. This is already evident in the case of chapeau basis functions with $N = 20$, where a slightly larger value of TOL would have offered some improvements. This problem is not an issue with trigonometric basis functions in the symmetric case.

The problem of recovering the potential from spectral data consisting of the function $\lambda(H)$ for λ corresponding to some fixed eigenfunction number is a considerably more ill-posed problem than the other usual inverse Sturm–Liouville problems. The easiest proofs of uniqueness rely on analytic continuation, [15] and this has to be considered unreasonable as the basis of a numerical algorithm. The method of [18] was able to handle all the known inverse Sturm–Liouville problems, except this one.



Trigonometric basis $N = 6$

Chapeau basis $N = 6$

Trigonometric basis $N = 14$

Chapeau basis $N = 14$

Trigonometric basis $N = 20$

Chapeau basis $N = 20$

FIG. 3. *Reconstruction of a $q(x)$ from two spectra.*

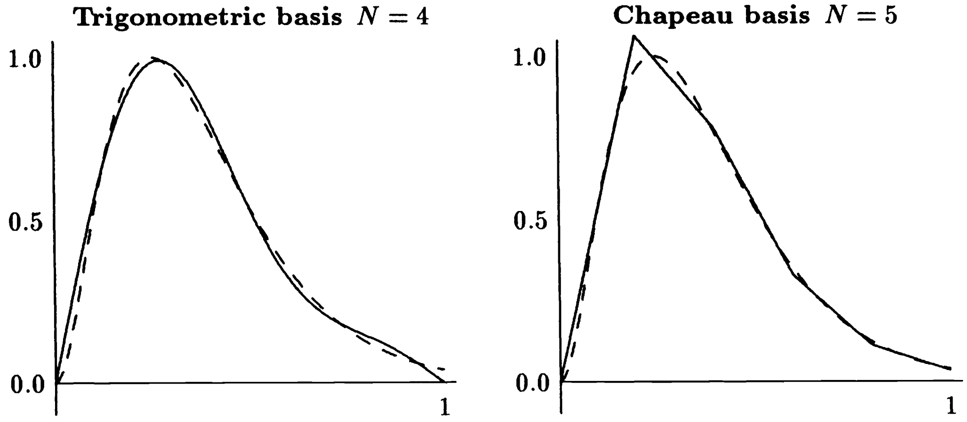


FIG. 4. *Reconstruction of $q(x)$ from $\lambda_1(H)$.*

The uniqueness theorem requires that the set of values H_j should have a finite accumulation point, and it is not known, at least to these authors, whether any finite data set can allow one to recover the lowest frequency modes of q as is the case when eigenvalue data are prescribed. To see the performance of the algorithm, consider the function $q(x) = 16x^2 \exp(-8x)$ which has been normalized to have a maximum value of one. This is a smooth function that has a very small discontinuity in its periodic extension. As such, the algorithm would recover it to within less than 1 percent error with spectral data consisting of 5 eigenvalues from each of two spectra, that is, $N = 10$. For the output of Fig. 4 we used as spectral data the value of $\lambda_1(H)$ for 10 values of H, namely $H = \{0, 1, 2, 3, 4, 5, 10, 20, 50, 100\}$. With the value of TOL set to 10^{-6} we were only able to keep about 5 or 6 modes with any of the basis functions we tried. Here, for good accuracy, we should keep the value of TOL small for best results and then delete any basis functions for which the corresponding singular value is less than the tolerance. The results are shown in Fig. 4.

The use of the singular value decomposition is essential for the functioning of this particular version of the algorithm. The number of values of H at which we computed λ_1 is not so crucial; much the same reconstructions would have resulted if we had used anywhere between 7 and 20 points, provided they were roughly equally spaced. This is due to the behavior of the analytic function $\lambda_1(H)$.

In terms of its numerical performance, how does this algorithm compare with some of the others mentioned in the introduction? The method of Paine is the closest in terms of methodology, although the implementation in [16] is only for the symmetric case, and in fact it is not clear how to implement the algorithm for noneigenvalue data. The mapping used is that between the potential and its eigenvalues. There is an additional disadvantage in this approach; the computation of the Frêchet derivative requires solving N^2 (where N is the number of points at which $q(x)$ is determined) Sturm–Liouville eigenvalue problems, and numerically this is an extremely expensive proposition. In contrast, our method requires $N \times B$ computations of an initial value problem and this requires considerably less computation. In addition, the quasi-Newton implementation means we have to compute this matrix only once. The method of Rundell and Sacks [18] is able to recover virtually the full family of inverse

Sturm–Liouville problems, the exception being the variable boundary condition data. This method leads to an extremely fast algorithm. Indeed, the computational time required to recover a potential from 10 or 20 items of spectral data is significantly less than the time taken to compute even *one* eigenvalue of the direct problem using the program SLEIGN. The method of this paper is unable, at least in the form presented here, to recover $q(x)$ from data consisting of eigenvalues and norming constants, but this is the only exception. In comparison with the Rundell–Sacks method the accuracy obtained is roughly the same, but the amount of computation required is considerably greater. Our Newton-type algorithm performs slightly better under noise in the spectral data, but this might be more a function of the particular implementations. One of the drawbacks to the method of [18] is that it is necessary to have an estimate of the mean of q, $\bar{q} = \int_0^1 q(s)ds$. For smooth functions, an adequate estimate can be obtained from the spectral data, but this ability degrades with rougher potentials. The method used here does not require the estimation of the mean. In addition, and this we believe to be the most important point, the ideas presented here can formally be extended to higher space dimensions, and this we intend to make the subject of future work.

Appendix: proofs of the lemmas. In what follows, C denotes a constant that is independent of k, β, and N.

LEMMA A1. *For $q \in V_N$ and $j \neq k$, $w(t) = \partial u_j(t; q, \lambda_j)/\partial q_k$ satisfies the estimate*

$$\left| w(t) + \frac{1}{4\pi k \lambda_j} \sin 2\pi k t \cos \sqrt{\lambda_j} t \right| \leq C \left(\frac{1}{\lambda_j |\sqrt{\lambda_j} - \pi k|} + \frac{1}{\lambda_j^{3/2}} \right).$$

Proof. w solves

$$\begin{gathered} -w'' + qw = \lambda_j w - \phi_k u_j[q], \\ w(0) = w'(0) = 0 \end{gathered}$$

which has the solution

$$w(t) = \int_0^t (y_1(s)y_2(t) - y_1(t)y_2(s))\phi_k(s)u_j[q](s)ds,$$

where y_1 and y_2 solve

$$\begin{gathered} -y_i'' + qy_i = \lambda_j y_i, \\ y_2(0) = y_1'(0) = 0, \\ y_2'(0) = y_1(0) = 1. \end{gathered}$$

Using the estimates [18, p. 13]

$$\left| y_1(s) - \cos \sqrt{\lambda_j} s \right| \leq \frac{1}{\sqrt{\lambda_j}} e^{\|q\|_{L^2}},$$

$$\left| y_2(s) - \frac{\sin \sqrt{\lambda_j} s}{\sqrt{\lambda_j}} \right| \leq \frac{1}{\lambda_j} e^{\|q\|_{L^2}},$$

we obtain

$$\begin{aligned}
|(y_1(s)y_2(t) - y_1(t)y_2(s)) - \frac{1}{\sqrt{\lambda_j}}\sin\sqrt{\lambda_j}(t-s)| &\le |y_1(s)|\left|y_2(t) - \frac{\sin\sqrt{\lambda_j}t}{\sqrt{\lambda_j}}\right| \\
&+|y_1(s) - \cos\sqrt{\lambda_j}s|\left|\frac{\sin\sqrt{\lambda_j}t}{\sqrt{\lambda_j}}\right| + |y_2(s)||y_1(t) - \cos\sqrt{\lambda_j}t| \\
&+\left|\cos\sqrt{\lambda_j}t\right|\left|y_2(s) - \frac{\sin\sqrt{\lambda_j}s}{\sqrt{\lambda_j}}\right| \\
&\le \frac{e^{\|q\|_{L^2}}}{2\lambda_j}\left(2 + \frac{1}{\sqrt{\lambda_j}}e^{\|q\|_{L^2}}\right).
\end{aligned}$$

Observing that $u_j[q] = y_2$ and using the estimate $|y_2(s) - (\sin\sqrt{\lambda_j}s/\sqrt{\lambda_j})| \le (e^{\|q\|_{L^2}}/\lambda_j)$ gives

$$\begin{aligned}
|w(t) &- \frac{1}{\lambda_j}\int_0^t \sin\sqrt{\lambda_j}(t-s)\phi_k(s)\sin\sqrt{\lambda_j}s\,ds| \\
&\le \left|\int_0^t \left\{y_1(s)y_2(t) - y_1(t)y_2(s) - \frac{1}{\sqrt{\lambda_j}}\sin\sqrt{\lambda_j}(t-s)\right\}\phi_k(s)y_2(s)ds\right| \\
&+\left|\int_0^t \frac{1}{\sqrt{\lambda_j}}\sin\sqrt{\lambda_j}(t-s)\phi_k(s)\left(y_2(s) - \frac{\sin\sqrt{\lambda_j}s}{\sqrt{\lambda_j}}\right)ds\right| \\
&\le \left\{1 + \frac{1}{2}\left(1 + \frac{e^{\|q\|_{L^2}}}{\sqrt{\lambda_j}}\right)\left(2 + \frac{e^{\|q\|_{L^2}}}{\sqrt{\lambda_j}}\right)\right\}\frac{\|\phi_k\|e^{\|q\|_{L^2}}}{\lambda_j^{3/2}}.
\end{aligned}$$

Now

$$\begin{aligned}
\int_0^t &\sin\sqrt{\lambda_j}(t-s)\phi_k(s)\sin\sqrt{\lambda_j}s\,ds \\
&= -\frac{\sin 2\pi kt\cos\sqrt{\lambda_j}t}{4\pi k} + \frac{\cos\sqrt{\lambda_j}t}{2}\int_0^t \cos 2\pi ks\cos 2\sqrt{\lambda_j}s\,ds \\
&\quad+\frac{\sin\sqrt{\lambda_j}t}{2}\int_0^t \sin 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds,
\end{aligned}$$

$$\int_0^t \cos 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds = \frac{\sin 2(\sqrt{\lambda_j} - \pi k)t}{4(\sqrt{\lambda_j} - \pi k)} + \frac{\sin 2(\sqrt{\lambda_j} + \pi k)t}{4(\sqrt{\lambda_j} + \pi k)},$$

$$\int_0^t \sin 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds = \frac{1 - \cos 2(\sqrt{\lambda_j} + \pi k)t}{4(\sqrt{\lambda_j} + \pi k)} + \frac{1 - \cos 2(\sqrt{\lambda_j} - \pi k)t}{4(\sqrt{\lambda_j} - \pi k)},$$

and the result follows using $\|\phi_k\|_{L^2} = 1/\sqrt{2}$ and $\lambda_j \ge 1$ for β sufficiently small.

LEMMA A2. $\left|\int_0^1 \sin\sqrt{\lambda_j}(1-t)\cos\sqrt{\lambda_j}t\sin 2\pi kt(q^{(1)} - q^{(2)})dt\right|$

$$\le \left(\frac{\sqrt{3}\beta}{2\pi}\right)\left(\frac{1}{j^{2+\alpha}}\right) + \frac{1}{2}\left|\int_0^1 \sin 2\pi jt\sin 2\pi kt(q^{(1)} - q^{(2)})dt\right|.$$

Proof. Since $q^{(i)} \in \operatorname{span}\{\cos 2\pi lx\}$, this gives

$$\begin{aligned}\int_0^1 &\sin\sqrt{\lambda_j}(1-t)\cos\sqrt{\lambda_j}t\sin 2\pi kt(q^{(1)}-q^{(2)})dt\\ &= \frac{\sin\sqrt{\lambda_j}}{2}\int_0^1 \cos 2\sqrt{\lambda_j}t\sin 2\pi kt[q^{(1)}-q^{(2)}]dt\\ &\quad -\frac{\cos\sqrt{\lambda_j}}{2}\int_0^1 \sin 2\sqrt{\lambda_j}t\sin 2\pi kt[q^{(1)}-q^{(2)}]dt.\end{aligned}$$

Using the inequalities $|\sin 2\sqrt{\lambda_j}t - \sin 2\pi jt| \le 2|\sqrt{\lambda_j}-\pi j||t| \le 2|t|\beta/\sqrt{3}\pi j^{2+\alpha}$ and $|\sin\sqrt{\lambda_j}| \le \beta/\sqrt{3}\pi j^{2+\alpha}$ gives the desired result.

LEMMA A3.

$$\sum_{j=1}^{\infty}\frac{1}{\sqrt{\lambda_j}}\left|\int_0^1 \sin 2\pi jt(\sin 2\pi kt(q^{(1)}-q^{(2)}))dt\right| \le \frac{1}{\sqrt{2}}\left(\sum_{j=1}^{\infty}\frac{1}{\lambda_j}\right)^{1/2}\|q^{(1)}-q^{(2)}\|_{L^2}.$$

Proof. $\int_0^1 \sin 2\pi jt\sin 2\pi kt[q^{(1)}-q^{(2)}]dt$ is the jth Fourier sine coefficient of $\sin 2\pi kt\cdot(q^{(1)}-q^{(2)})$. The result follows from the Cauchy–Schwarz inequality and Parseval's identity.

LEMMA A4. *For all β sufficiently small, $\sum_{\substack{j=1\\ j\neq k}}^{\infty}(1/\sqrt{\lambda_j}|\sqrt{\lambda_j}-\pi k|) \le C$.*

Proof. It is easily verified that

$$\sum_{\substack{j=1\\ j\neq k}}^{\infty}\frac{1}{\sqrt{\lambda_j}|\sqrt{\lambda_j}-\pi k|} \le \frac{1}{\pi^2}\sum_{\substack{j=1\\ j\neq k}}^{\infty}\frac{|\sqrt{\lambda_j}-\pi j|}{|j-k||\sqrt{\lambda_j}-\pi k|}+\frac{1}{\pi^2}\sum_{\substack{j=1\\ j\neq k}}^{\infty}\frac{|\sqrt{\lambda_j}-\pi j|}{j\sqrt{\lambda_j}}+\frac{1}{\pi^2}\sum_{\substack{j=1\\ j\neq k}}^{\infty}\frac{1}{j|j-k|}.$$

For β sufficiently small $\sqrt{\lambda_j} \ge 1$ and $|\sqrt{\lambda_j}-\pi k| \ge \frac{\pi}{2}$ for all $j \neq k$. The estimate

$$\sum_{\substack{j=1\\ j\neq k}}^{\infty}\frac{1}{j|j-k|} \le \frac{3(1+\ln k)}{k} \le 3$$

and Assumption A gives the desired result.

LEMMA A5. *For $q^{(i)} \in V_N$ and β sufficiently small*

$$\begin{aligned}&\left|(u_j[q^{(1)}]-u_j[q^{(2)}])(t)-\frac{1}{\lambda_j}\int_0^t \sin\sqrt{\lambda_j}(t-s)\sin\sqrt{\lambda_j}s(q^{(1)}-q^{(2)})ds\right|\\ &\quad\le \frac{C\|q^{(1)}-q^{(2)}\|_{L^2}}{\lambda_j^{3/2}}.\end{aligned}$$

Proof. $u := u_j[q^{(1)}]-u_j[q^{(2)}]$ satisfies

$$\begin{gathered}-u'' = \lambda_j u - q^{(1)}u - (q^{(1)}-q^{(2)})u_j[q^{(2)}],\\ u(0)=u'(0)=0\end{gathered}$$

with solution

$$u(t) = \int_0^t \frac{\sin\sqrt{\lambda_j}(t-s)}{\sqrt{\lambda_j}}[q^{(1)}u + (q^{(1)}-q^{(2)})u_j[q^{(2)}]]ds.$$

An application of Gronwall's inequality gives

$$|u(t)| \le \sqrt{\frac{2}{\lambda_j}} e^{\frac{1}{\lambda_j}\|q^{(1)}\|^2_{L^2}} \|u_j[q^{(2)}]\|_{L^2} \|q^{(1)} - q^{(2)}\|_{L^2},$$

and from the estimate

$$\left| u_j[q^{(2)}] - \frac{\sin\sqrt{\lambda_j}s}{\sqrt{\lambda_j}} \right| \le \frac{1}{\lambda_j} e^{\|q^{(2)}\|_{L^2}}$$

we conclude that

$$\begin{aligned}&\left| u(t) - \frac{1}{\lambda_j}\int_0^t \sin\sqrt{\lambda_j}(t-s)\sin\sqrt{\lambda_j}s(q^{(1)} - q^{(2)})ds \right| \\ &\quad \le \left\{ e^{\|q^{(2)}\|_{L^2}} + \left(\sqrt{2}\|q^{(1)}\|_{L^2} e^{\frac{1}{\lambda_j}\|q^{(1)}\|^2_{L^2}} \left(1 + \frac{1}{\sqrt{\lambda_j}} e^{\|q^{(2)}\|_{L^2}} \right)\right)\right\} \frac{\|q^{(1)} - q^{(2)}\|_{L^2}}{\lambda_j^{3/2}}.\end{aligned}$$

The result follows from $\lambda_j \ge 1$ for β sufficiently small.

LEMMA A6. *For $j \ne k$ and all β sufficiently small*

$$\begin{aligned}&\left| \int_0^1 \sin\sqrt{\lambda_j}(1-t)\cos 2\pi kt \int_0^t \sin\sqrt{\lambda_j}(t-s)\sin\sqrt{\lambda_j}s[q^{(1)} - q^{(2)}]ds\,dt \right| \\ &\qquad \le C\left\{ \left(\frac{1}{j^{2+\alpha}} + \frac{1}{\sqrt{\lambda_j}} + \frac{1}{|\sqrt{\lambda_j} - \pi k|} \right) \|q^{(1)} - q^{(2)}\|_{L_2} \right. \\ &\qquad\qquad \left. + \left| \int_0^1 \sin 2\pi jt \sin 2\pi kt[q^{(1)} - q^{(2)}]dt \right| \right\}.\end{aligned}$$

Proof. An integration by parts gives

$$\begin{aligned}&\int_0^1 \sin\sqrt{\lambda_j}(1-t)\cos 2\pi kt \left\{ \int_0^t \sin\sqrt{\lambda_j}(t-s)\sin\sqrt{\lambda_j}s[q^{(1)} - q^{(2)}]ds \right\} dt \\ &= \frac{1}{2}\int_0^1 \left\{ \int_t^1 \sin\sqrt{\lambda_j}(1-s)\sin\sqrt{\lambda_j}s\cos 2\pi ks\,ds \right\} \sin 2\sqrt{\lambda_j}t[q^{(1)} - q^{(2)}]dt \\ &\quad - \int_0^1 \left\{ \int_t^1 \sin\sqrt{\lambda_j}(1-s)\cos\sqrt{\lambda_j}s\cos 2\pi ks\,ds \right\} \sin^2\sqrt{\lambda_j}t(q^{(1)} - q^{(2)})dt,\end{aligned}$$

where

$$\begin{aligned}&\int_t^1 \sin\sqrt{\lambda_j}(1-s)\sin\sqrt{\lambda_j}s\cos 2\pi ks\,ds \\ &\quad = \frac{\cos\sqrt{\lambda_j}\sin 2\pi kt}{4\pi k} + \frac{\sin\sqrt{\lambda_j}}{2}\int_t^1 \sin 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds \\ &\quad + \frac{\cos\sqrt{\lambda_j}}{2}\int_t^1 \cos 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds,\end{aligned}$$

$$\begin{aligned}\int_t^1 &\sin\sqrt{\lambda_j}(1-s)\cos\sqrt{\lambda_j}s\cos 2\pi ks\,ds\\ &= -\frac{\sin\sqrt{\lambda_j}\sin 2\pi kt}{4\pi k}+\frac{\sin\sqrt{\lambda_j}}{2}\int_t^1\cos 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds\\ &\quad -\frac{\cos\sqrt{\lambda_j}}{2}\int_t^1 \sin 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds,\end{aligned}$$

$$\begin{aligned}\int_t^1 &\sin 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds\\ &= \frac{\cos 2(\sqrt{\lambda_j}+\pi k)t-\cos 2(\sqrt{\lambda_j}+\pi k)}{4(\sqrt{\lambda_j}+\pi k)}\\ &\quad+\frac{\cos 2(\sqrt{\lambda_j}-\pi k)t-\cos 2(\sqrt{\lambda_j}-\pi k)}{4(\sqrt{\lambda_j}-\pi k)},\end{aligned}$$

$$\begin{aligned}\int_t^1 \cos 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds &= \frac{\sin 2(\sqrt{\lambda_j}-\pi k)-\sin 2(\sqrt{\lambda_j}-\pi k)t}{4(\sqrt{\lambda_j}-\pi k)}\\ &\quad+\frac{\sin 2(\sqrt{\lambda_j}+\pi k)-\sin 2(\sqrt{\lambda_j}+\pi k)t}{4(\sqrt{\lambda_j}+\pi k)}.\end{aligned}$$

A substitution gives

$$\begin{aligned}\int_0^1 &\sin\sqrt{\lambda_j}(1-t)\cos 2\pi kt\left\{\int_0^t \sin\sqrt{\lambda_j}(t-s)\sin\sqrt{\lambda_j}s[q^{(1)}-q^{(2)}]ds\right\}dt\\ &= \frac{\cos\sqrt{\lambda_j}}{8\pi k}\int_0^1 \sin 2\sqrt{\lambda_j}t\sin 2\pi kt[q^{(1)}-q^{(2)}]dt\\ &\quad+\frac{\cos\sqrt{\lambda_j}}{4}\int_0^1\left\{\int_t^1 \cos 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds\right\}\sin 2\sqrt{\lambda_j}t[q^{(1)}-q^{(2)}]dt\\ &\quad+\frac{\sin\sqrt{\lambda_j}}{4\pi k}\int_0^1 \sin 2\pi kt\sin^2\sqrt{\lambda_j}t[q^{(1)}-q^{(2)}]dt\\ &\quad+\frac{\cos\sqrt{\lambda_j}}{2}\int_0^1\left\{\int_t^1 \sin 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds\right\}\sin^2\sqrt{\lambda_j}t[q^{(1)}-q^{(2)}]dt\\ &\quad+\frac{\sin\sqrt{\lambda_j}}{4}\int_0^1\left\{\int_t^1 \sin 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds\right\}\sin 2\sqrt{\lambda_j}t[q^{(1)}-q^{(2)}]dt\\ &\quad-\frac{\sin\sqrt{\lambda_j}}{2}\int_0^1\left\{\int_t^1 \cos 2\sqrt{\lambda_j}s\cos 2\pi ks\,ds\right\}\sin^2\sqrt{\lambda_j}t[q^{(1)}-q^{(2)}]dt,\end{aligned}$$

and the result follows using $|\sin 2\sqrt{\lambda_j}t-\sin 2\pi jt|\le 2|t|\beta/\sqrt{3}\pi j^{2+\alpha}$, $|\sin\sqrt{\lambda_j}|\le\beta/\sqrt{3}\pi j^{2+\alpha}$, and $k\ge 1$.

LEMMA A7. *Let* $w(x)=w_1(x)-w_2(x)$ *where the functions* $w_1(x)=(\partial u_j/\partial q_k)\cdot(x;q^{(1)},\lambda_j)$ *and* $w_2(x)=(\partial u_j/\partial q_k)(x;q^{(2)},\lambda_j)$. *For* $q^{(i)}\in V_N$ *and all* j *and* k, $w(x)$

satisfies the estimate

$$|w|_{L^\infty} \leq \frac{C}{\lambda_j^{3/2}} \|q^{(1)} - q^{(2)}\|_{L^2}$$

for all β sufficiently small.

Proof. $w(x)$ solves

$$-w'' = \lambda_j w - q^{(1)} w - (q^{(1)} - q^{(2)}) w_2 - \phi_k (u_j[q^{(1)}] - u_j[q^{(2)}]),$$
$$w(0) = w'(0) = 0,$$

and this has the solution

$$w(x) = \int_0^x \frac{\sin\sqrt{\lambda_j}(x-t)}{\sqrt{\lambda_j}} [q^{(1)} w(t) + (q^{(1)} - q^{(2)}) w_2(t) + \phi_k (u_j[q^{(1)}] - u_j[q^{(2)}])] dt.$$

Using Gronwall's inequality, we have

$$|w(x)| \leq \sqrt{\frac{2}{\lambda_j}} \left[\|w_2\|_{L^2} \|q^{(1)} - q^{(2)}\|_{L^2} + \|\phi_k\|_{L^2} \|u_j[q^{(2)}] - u_j[q^{(1)}]\|_{L^2} \right] e^{\frac{1}{\lambda_j}\|q^{(1)}\|_{L^2}},$$

$$\|u_j[q^{(2)}] - u_j[q^{(1)}]\|_{L^2} \leq \sqrt{\frac{2}{\lambda_j}} e^{\frac{1}{\lambda_j}\|q^{(1)}\|_{L^2}^2} \|u_j[q^{(2)}]\|_{L^2} \|q^{(1)} - q^{(2)}\|_{L^2},$$

$$\|w_2\|_{L^2} \leq \sqrt{\frac{2}{\lambda_j}} e^{\frac{1}{\lambda_j}\|q^{(2)}\|_{L^2}^2} \|u_j[q^{(2)}]\|_{L^2} \|\phi_k\|_{L^2}.$$

The estimate now follows using $\left| u_j[q^{(2)}](s) - (\sin\sqrt{\lambda_j} s / \sqrt{\lambda_j}) \right| \leq \frac{1}{\lambda_j} e^{\|q^{(2)}\|_{L^2}}$ and $\lambda_j \geq 1$ for all β sufficiently small.

LEMMA A8. *For $q^{(i)} \in V_N$, $j \neq k$ and all β sufficiently small*

$$\left| \lambda_j \left(\frac{\partial u_j}{\partial q_k}(1; q^{(1)}, \lambda_j) - \frac{\partial u_j}{\partial q_k}(1; q^{(2)}, \lambda_j) \right) \right|$$
$$\leq C \left\{ \left(\frac{1}{\lambda_j} + \frac{1}{j^{1+\alpha}} + \frac{1}{\sqrt{\lambda_j}|\sqrt{\lambda_j} - \pi k|} \right) \|q^{(1)} - q^{(2)}\|_{L^2} \right.$$
$$\left. + \frac{1}{\sqrt{\lambda_j}} \left| \int_0^1 \sin 2\pi j t \sin 2\pi k t [q^{(1)} - q^{(2)}] dt \right| \right\}.$$

Proof. Let $w(x) = w_1(x) - w_2(x)$, where $w_1(x) = (\partial u_j / \partial q_k)(x; q^{(1)}, \lambda_j)$ and $w_2(x) = (\partial u_j / \partial q_k)(x; q^{(2)}, \lambda_j)$. Then

$$w(1) = \int_0^1 \frac{\sin\sqrt{\lambda_j}(1-t)}{\sqrt{\lambda_j}} [q^{(1)} w(t) + (q^{(1)} - q^{(2)}) w_2(t) + \phi_k (u_j[q^{(1)}] - u_j[q^{(2)}])] dt$$
$$= \int_0^1 \frac{\sin\sqrt{\lambda_j}(1-t)}{\sqrt{\lambda_j}} q^{(1)} w(t) dt$$

$$+\int_0^1 \frac{\sin\sqrt{\lambda_j}(1-t)}{\sqrt{\lambda_j}}\left[w_2(t)+\frac{\sin 2\pi kt\cos\sqrt{\lambda_j}t}{4\pi k\lambda_j}\right](q^{(1)}-q^{(2)})dt$$

$$+\frac{1}{4\pi k\lambda_j^{3/2}}\int_0^1 \sin\sqrt{\lambda_j}(1-t)\sin 2\pi kt\cos\sqrt{\lambda_j}t(q^{(1)}-q^{(2)})dt$$

$$+\int_0^1 \frac{\sin\sqrt{\lambda_j}(1-t)}{\sqrt{\lambda_j}}\phi_k\left\{u_j[q^{(1)}]-u_j[q^{(2)}]\right.$$

$$\left.-\frac{1}{\lambda_j}\int_0^t \sin\sqrt{\lambda_j}(t-s)\sin\sqrt{\lambda_j}s[q^{(1)}-q^{(2)}]ds\right\}dt$$

$$+\frac{1}{\lambda_j^{3/2}}\int_0^1 \sin\sqrt{\lambda_j}(1-t)\phi_k\left\{\int_0^t \sin\sqrt{\lambda_j}(t-s)\sin\sqrt{\lambda_j}s[q^{(1)}-q^{(2)}]ds\right\}dt.$$

Now $k \geq 1$ and $\|\phi_k\|_{L^2} = 1/\sqrt{2}$ for all k. The result follows from Lemmas A1, A2, A5, A6, and A7.

Proof of Lemma 3. By Lemma A8,

$$\sum_{\substack{j=1\\ j\neq k}}^{N}\left|\lambda_j\left(\frac{\partial u_j}{\partial q_k}(1;q^{(2)},\lambda_j)-\frac{\partial u_j}{\partial q_k}(1;q^{(1)},\lambda_j)\right)\right|$$

$$\leq C\left\{\sum_{j=1}^{\infty}\left(\frac{1}{\lambda_j}+\frac{1}{j^{1+\alpha}}\right)+\sum_{\substack{j=1\\ j\neq k}}^{\infty}\frac{1}{\sqrt{\lambda_j}|\sqrt{\lambda_j}-\pi k|}\right\}\|q^{(1)}-q^{(2)}\|_{L^2}$$

$$+\sum_{j=1}^{\infty}\frac{1}{\sqrt{\lambda_j}}\left|\int_0^1 \sin 2\pi jt\sin 2\pi kt[q^{(1)}-q^{(2)}]dt\right|.$$

By Lemma A7, when $j = k$, the remaining term is bounded by

$$\left|\lambda_k\left(\frac{\partial u_k}{\partial q_k}(1;q^{(2)},\lambda_k)-\frac{\partial u_k}{\partial q_k}(1;q^{(1)},\lambda_k)\right)\right| \leq \frac{C}{\sqrt{\lambda_k}}\|q^{(1)}-q^{(2)}\|_{L^2} \leq C\|q^{(1)}-q^{(2)}\|_{L^2}.$$

The result follows from Lemmas A3 and A4.

Acknowledgment. The authors thank Roger Knobel for several helpful discussions.

REFERENCES

[1] L. Andersson, *Algorithms for solving inverse eigenvalue problems for Sturm–Liouville equations*, in Inverse Methods in Action, P. Sabatier, ed., Springer–Verlag, New York, Berlin, 1990, pp. 138–145.

[2] P. B. Bailey, B. S. Garbow, H. G. Kaper, and A. Zettl, *Eigenvalue and eigenfunction computations for Sturm–Liouville problems*, to appear.

[3] G. Borg, *Eine Umkehrung der Sturm–Liouville Eigenwertaufgabe*, Acta Math., 76 (1946), pp. 1–96.

[4] K. Deimling, *Nonlinear Functional Analysis*, Springer–Verlag, Berlin, 1985.

[5] G. M. L. Gladwell, *Inverse Problems in Vibration*, Martinus Nijhoff, Dordrecht, the Netherlands, 1986.

[6] I. M. Gel'fand and B. M. Levitan, *On the determination of a differential equation from its spectral function*, Amer. Math. Soc. Trans., 1 (1951), pp. 253–304.

[7] O. H. Hald, *The inverse Sturm–Liouville problem with symmetric potentials*, Acta Math., 141 (1978), pp. 263–291.

[8] O. H. HALD, *The inverse Sturm–Liouville problem and the Rayleigh–Ritz method*, Math. Comp., 32 143 (1978), pp. 687–705.

[9] H. HOCHSTADT, *The inverse Sturm–Liouville problem*, Comm. Pure Appl. Math., 26 (1973), pp. 715–729.

[10] H. HOCHSTADT AND B. LIEBERMAN, *An inverse Sturm–Liouville problem with mixed given data*, SIAM J. Appl. Math., 34 (1976), pp. 676–680.

[11] B. M. LEVITAN, *Inverse Sturm–Liouville Problems*, VNU Science Press, Utrecht, the Netherlands, 1987.

[12] J. R. MCLAUGHLIN AND G. H. HANDELMAN, *Sturm–Liouville inverse eigenvalue problems*, in Mechanics Today, The Reissner Volume, Pergamon Press, New York, 1980, pp. 281–295.

[13] J. R. MCLAUGHLIN, *Analytic methods for recovering coefficients in differential equations from spectral data*, SIAM Rev., 28 (1986), pp. 53–72.

[14] ______, *Stability theorems for two inverse spectral problems*, Inverse Problems, 4 (1988), pp. 529–540.

[15] J. R. MCLAUGHLIN AND W. RUNDELL, *A Uniqueness Theorem for an Inverse Sturm–Liouville Problem*, J. Math. Phys., 28 (1987), pp. 1471–1472.

[16] J. PAINE, *An inverse Sturm–Liouville eigenvalue method based on piecewise constant potentials*, preprint.

[17] J. PÖSCHEL AND E. TRUBOWITZ, *Inverse Spectral Theory*, Academic Press, London, 1987.

[18] W. RUNDELL AND P. E. SACKS, *Reconstruction techniques for classical inverse Sturm–Liouville problems*, Math. Comp., to appear.

[19] P. E. SACKS, *An iterative method for the inverse Dirichlet problem*, Inverse Problems, 4 (1988), pp. 1055–1069.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 505-511, March 1992

ERROR BOUNDS FOR THE ASYMPTOTIC EXPANSION OF THE RATIO OF TWO GAMMA FUNCTIONS WITH COMPLEX ARGUMENT*

C. L. FRENZEN†

Abstract. Error bounds are obtained for an asymptotic expansion of the ratio of two gamma functions $\Gamma(z+a)/\Gamma(z+b)$ when a and b are complex constants and $|z|$ is large. These bounds reduce to earlier bounds for the real case when a, b, and z are real. Properties of completely monotonic functions are used to provide error bounds in the complex case, as in the earlier real case.

Key words. error bound, asymptotic expansion, gamma function

AMS(MOS) subject classifications. primary 41A60; secondary 33A15

1. Introduction.

In [1], Fields showed that for all $n \geqq 1$,

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = \sum_{j=0}^{n-1} \frac{\Gamma(1-2\rho+2j)}{\Gamma(1-2\rho)(2j)!} B_{2j}^{(2\rho)}(\rho) w^{2\rho-1-2j} + O(w^{2\rho-1-2n}) \tag{1.1}$$

as $w \to \infty$ with $|\arg(w+\rho)| < \pi$, where a and b are fixed complex numbers, $2w = 2z+a+b-1$ and $2\rho = a-b+1$. The $B_{2j}^{(2\rho)}(\rho)$ are generalized Bernoulli polynomials, defined by

$$\left(\frac{t}{e^t-1}\right)^{2\rho} e^{\rho t} = \left(\frac{\sinh t/2}{t/2}\right)^{-2\rho} = \sum_{j=0}^{\infty} \frac{t^{2j}}{(2j)!} B_{2j}^{(2\rho)}(\rho), \tag{1.2}$$
$$B_0^{(2\rho)}(\rho) = 1 \qquad (|t| < 2\pi).$$

The first few of these polynomials are

$$B_2^{(2\rho)}(\rho) = -\frac{\rho}{6}, \quad B_4^{(2\rho)}(\rho) = \frac{\rho(5\rho+1)}{60}, \quad B_6^{(2\rho)}(\rho) = -\frac{\rho(35\rho^2+21\rho+4)}{504}. \tag{1.3}$$

A list of the $B_{2j}^{(2\rho)}(\rho)$ for $j = 1, 2, \cdots, 6$ and a recurrence formula can be found in [3].

In [2], it was shown that when a, b, and z are real, $z+a > 0$, $w > 0$, and $0 < 2\rho < 1$, the error made by truncating the asymptotic expansion in (1.1) is numerically less than and has the same sign as the first neglected term. The purpose of this paper is to indicate how, with a little more work, computable error bounds can be obtained for (1.1) when z, a, and b are complex. More specifically, we shall show that when $|\arg w| < 3\pi/4$ and a and b (and therefore ρ) are fixed complex constants such that $0 \leqq \operatorname{Re}(2\rho) < 1$, then

$$\frac{\Gamma(w+\rho)}{\Gamma(w-\rho+1)} = \sum_{j=0}^{n-1} \frac{\Gamma(1-2\rho+2j)}{\Gamma(1-2\rho)(2j)!} B_{2j}^{(2\rho)}(\rho) w^{2\rho-1-2j} + R_n(w,\rho), \tag{1.4}$$

where

$$|R_n(w,\rho)| \leqq \frac{\Gamma(1-\operatorname{Re}(2\rho)+2n)}{|\Gamma(1-2\rho)|(2n)!} |B_{2n}^{(|2\rho|)}(|\rho|)| \cdot (\operatorname{Re}(w e^{-i\beta}) - |\operatorname{Im}\rho||\sin\beta|)^{\operatorname{Re}(2\rho)-1-2n} e^{|\beta||\operatorname{Im}(2\rho)|}. \tag{1.5}$$

* Received by the editors August 13, 1990; accepted for publication (in revised form) June 21, 1991.

† Mathematics Department, Naval Postgraduate School, Monterey, California 93943.

In (1.5) β is any number in the interval $[-\pi/4, \pi/4]$. The bound (1.5) is valid when w lies in the annular sector

$$|\arg(w e^{-i\beta})| < \pi/2, \qquad \operatorname{Re}(w e^{-i\beta}) > |\operatorname{Im} \rho| |\sin \beta|. \tag{1.6}$$

In the conclusion of this paper we shall show that the bound in (1.5) may, for computational purposes, be replaced by

$$\begin{aligned}|R_n(w, \rho)| \leqq |1-2\rho|\, e^{\pi|\operatorname{Im} \rho|} |B_{2n}^{(|2\rho|)}(|\rho|)| (\operatorname{Re}(w e^{-i\beta}) \\ -|\operatorname{Im} \rho| |\sin \beta|)^{\operatorname{Re}(2\rho)-1-2n}\, e^{|\beta||\operatorname{Im}(2\rho)|}.\end{aligned} \tag{1.7}$$

For given values of w and ρ, the bound for $R_n(w, \rho)$ on the right side of (1.5) depends on the value given to β. A further discussion of this point will be given at the conclusion of this article.

2. Derivation of the error bound. As in [2], we shall assume that simple modifications to $\Gamma(w+\rho)/\Gamma(w-\rho+1)$ have been made using the functional equation for the gamma function $(z\Gamma(z) = \Gamma(z+1))$ so that $0 \leqq \operatorname{Re}(2\rho) < 1$. From [4, p. 119], we can write

$$\frac{\Gamma(w+\rho)}{\Gamma(w-\rho+1)} = \frac{1}{\Gamma(1-2\rho)} \int_0^\infty e^{-wt} t^{-2\rho} \left(\frac{\sinh t}{t/2}\right)^{-2\rho} dt, \tag{2.1}$$

assuming that $\operatorname{Re}(w+\rho) > 0$. All powers in (2.1) have their principal values. Note that if $\operatorname{Re} w < 0$, then the reflection formula for the gamma function can be used to write

$$\frac{\Gamma(w+\rho)}{\Gamma(w-\rho+1)} = \frac{-\sin \pi(w-\rho)}{\sin \pi(w+\rho)} \frac{\Gamma(-w+\rho)}{\Gamma(-w-\rho+1)}. \tag{2.2}$$

Without loss of generality, (2.2) allows us to assume $\operatorname{Re} w > 0$.

The function $((\sinh t/2)/t/2)^{-2\rho}$ is holomorphic within the sector $\mathscr{S}$: $|\arg t| \leqq \pi/2 - \delta$, where δ is any positive number satisfying $0 < \delta < \pi/4$. Putting $t = r + is$, it follows that

$$\begin{aligned}\frac{|r+is|^2}{|\sinh(r+is)|^2} &= \frac{r^2+s^2}{(\sinh r)^2 + (\sin s)^2} \\ &\leqq \left(\frac{r}{\sinh r}\right)^2 (1 + (\cot \delta)^2) \leqq (\csc \delta)^2,\end{aligned} \tag{2.3}$$

so that

$$\left|\left(\frac{\sinh t/2}{t/2}\right)^{-2\rho}\right| \leqq (\csc \delta)^{\operatorname{Re}(2\rho)}\, e^{\pi|\operatorname{Im}(2\rho)|} \tag{2.4}$$

within $\mathscr{S}$. Using the estimate in (2.4), we can rotate the path of integration in (2.1) (although $t^{-2\rho}$ is not analytic at $t=0$ the rotation is justified since $\operatorname{Re}(2\rho) < 1$). It is straightforward to show that (2.1) becomes, after rotating the path of integration,

$$\frac{\Gamma(w+\rho)}{\Gamma(w-\rho+1)} = \frac{1}{\Gamma(1-2\rho)} \int_0^{\infty e^{-i\beta}} e^{-wt} t^{-2\rho} \left(\frac{\sinh t/2}{t/2}\right)^{-2\rho} dt, \tag{2.5}$$

where β satisfies $-(\pi/2) < \beta < (\pi/2)$ and $|\arg(w e^{-i\beta})| < (\pi/2)$.

Using (1.2), we write

$$\left(\frac{\sinh t/2}{t/2}\right)^{-2\rho} = \sum_{j=0}^{n-1} \frac{t^{2j}}{(2j)!} B_{2j}^{(2\rho)}(\rho) + r_n(t, \rho). \tag{2.6}$$

Substituting (2.6) into (2.5) and integrating the sum termwise then gives (1.4) with

$$R_n(w,\rho)=\frac{1}{\Gamma(1-2\rho)}\int_0^{\infty e^{-i\beta}} e^{-wt}t^{-2\rho}r_n(t,\rho)\,dt, \tag{2.7}$$

where $|\beta|<\pi/2$ and $|\arg(we^{-i\beta})|<\pi/2$.

By letting $t=u^{1/2}e^{-i\beta}$ (for $u\in[0,\infty)$) and $w=|w|e^{i\alpha}$, (2.7) becomes, using (2.6),

$$\begin{aligned}R_n(w,\rho)=\frac{1}{2\Gamma(1-2\rho)}&\int_0^\infty e^{-u^{1/2}|w|\exp(i(\alpha-\beta))}u^{-\rho-1/2}\\ &\cdot e^{i\beta(2\rho-1)}\left[F(u)-\sum_{j=0}^{n-1}\frac{F^{(j)}(0)}{j!}u^j\right]du.\end{aligned}\tag{2.8}$$

In (2.8) the complex-valued function F of the real variable u is defined by

$$F(u)=(G(u))^{-2\rho} \tag{2.9}$$

where

$$G(u)=\frac{\sinh\sqrt{u\,e^{-2i\beta}/4}}{\sqrt{u\,e^{-2i\beta}/4}} \tag{2.10}$$

and

$$\frac{F^{(j)}(0)}{j!}=\frac{e^{-2i\beta j}}{(2j)!}B_{2j}^{(2\rho)}(\rho). \tag{2.11}$$

From (2.9) and (2.10) we can write

$$F(u)=\exp\left(-2\rho S\left(\frac{u\,e^{-2i\beta}}{4}\right)\right) \tag{2.12}$$

where

$$S(z)=\ln\frac{\sinh\sqrt{z}}{\sqrt{z}}. \tag{2.13}$$

Recall that a function f is completely monotonic on $[0,\infty)$ if $(-1)^nf^{(n)}(x)\geqq 0$ on $[0,\infty)$ for $n=0,1,2,\cdots$. If f is completely monotonic on $[0,\infty)$ then $|f^{(n)}(x)|\leqq|f^{(n)}(0)|$ for $x\in[0,\infty)$ and $n=0,1,2,\cdots$. Since

$$S^{(1)}(z)=\sum_{n=1}^{\infty}\frac{1}{z+(n\pi)^2},$$

it follows that $S^{(1)}$ is a completely monotonic function on $[0,\infty)$. In [2] we showed that the function

$$h_\gamma(x)=e^{-\gamma S(x)}=\left(\frac{\sinh x^{1/2}}{x^{1/2}}\right)^{-\gamma} \tag{2.14}$$

is completely monotonic on $[0,\infty)$ for $\gamma>0$.

To establish the bound for $R_n(w,\rho)$ in (2.8), we note that, by Taylor's theorem for vector-valued functions of a real variable, there exists a ξ between zero and u so that

$$\left|F(u)-\sum_{j=0}^{n-1}\frac{F^{(j)}(0)}{j!}u^j\right|\leqq\frac{|F^{(n)}(\xi)|}{n!}u^n. \tag{2.15}$$

To compute $F^{(n)}$, we apply the Faa di Bruno formula for the derivative of a composite function (see [5]):

$$[f(g(u))]^{(n)}=\sum_{k=1}^{n} f^{(k)}(g(u))A_{nk}\{g^{(1)}(u),\quad g^{(2)}(u),\cdots,g^{(n)}(u)\} \tag{2.16}$$

where

$$\begin{aligned} &A_{nk}\{g^{(1)}(u),g^{(2)}(u),\cdots,g^{(n)}(u)\}\\ &\quad=\sum\frac{n!}{m_1!\,m_2!\cdots m_n!}\left(\frac{g^{(1)}(u)}{1!}\right)^{m_1}\cdots\left(\frac{g^{(n)}(u)}{n!}\right)^{m_n}, \end{aligned} \tag{2.17}$$

and the sum in (2.17), for given values of n and k, ranges over all nonnegative integer solutions of the two equations

$$m_1+m_2+\cdots+m_n=k,\qquad m_1+2m_2+\cdots+nm_n=n. \tag{2.18}$$

From (2.12), we can write $F(u)=f(g(u))$, where $f(u)=e^{-2\rho u}$ and $g(u)=S(u\,e^{-2i\beta}/4)$. Applying (2.16), it follows that

$$F^{(n)}(u)=\left(\frac{e^{-2i\beta}}{4}\right)^{n}\sum_{k=1}^{n}(-2\rho)^k B_{nk}, \tag{2.19}$$

where

$$B_{nk}=\sum\frac{n!}{m_1!m_2!\cdots m_n!}\left(\frac{S^{(1)}(u\,e^{-2i\beta}/4}{1!}\right)^{m_1}\cdots\left(\frac{S^{(n)}(u\,e^{-2i\beta}/4)}{n!}\right)^{m_n}, \tag{2.20}$$

the sum in (2.20), again ranging over all nonnegative integer solutions of (2.18).

Because $S^{(1)}$ is completely monotonic on $[0,\infty)$, it follows that (see [7, p. 158])

$$|S^{(1)}(x+iy)|\leqq|S^{(1)}(x)|\leqq|S^{(1)}(0)|\qquad(0\leqq x<\infty,\ -\infty<y<\infty). \tag{2.21}$$

Indeed, we also have

$$|S^{(j)}(x+iy)|\leqq|S^{(j)}(x)|\leqq|S^{(j)}(0)|\qquad(0\leqq x<\infty,\ -\infty<y<\infty), \tag{2.22}$$

since $(-1)^{j+1}S^{(j)}$ is completely monotonic on $[0,\infty)$ for $j=2,3,\cdots$, when $S^{(1)}$ is. Therefore, when $\operatorname{Re}(u\,e^{-2i\beta}/4)\geqq 0$, i.e., when

$$-\frac{\pi}{4}\leqq\beta\leqq\frac{\pi}{4}. \tag{2.23}$$

equation (2.20), the triangle inequality, and (2.22) together imply

$$|B_{nk}|\leqq C_{nk}, \tag{2.24}$$

where

$$C_{nk}=\sum\frac{n!}{m_1!\,m_2!\cdots m_n!}\left(\frac{|S^{(1)}(0)|}{1!}\right)^{m_1}\cdots\left(\frac{|S^{(n)}(0)|}{n!}\right)^{m_n} \tag{2.25}$$

and the sum again is over all nonnegative integer solutions of (2.18). However, letting

$$h_{|2\rho|}(x)=e^{-|2\rho|S(x)}=\left(\frac{\sinh x^{1/2}}{x^{1/2}}\right)^{-|2\rho|},$$

where S is given in (2.13), it also follows from (1.2) and the complete monotonicity of both $S^{(1)}$ and $h_{|2\rho|}$ on $[0,\infty)$ (see (2.14)) that

$$(-1)^n h_{|2\rho|}^{(n)}(0)=\sum_{k=1}^{n}|2\rho|^k C_{nk}=\frac{n!4^n}{(2n)!}|B_{2n}^{(|2\rho|)}(|\rho|)|, \tag{2.26}$$

with C_{nk} given in (2.25). Consequently, (2.19), together with (2.24) and (2.26), then implies

$$|F^{(n)}(u)| \leqq \frac{n!}{(2n)!} |F(u)| |B_{2n}^{(|2\rho|)}(|\rho|)|. \tag{2.27}$$

To bound $|F(u)|$, from (2.9) and (2.10)

$$|F(u)| = |G(u)|^{-\mathrm{Re}\,(2\rho)}\, e^{\mathrm{Im}\,(2\rho) \arg G(u)} \leqq e^{\mathrm{Im}\,(2\rho) \arg G(u)}, \tag{2.28}$$

where we have used the complete monotonicity of $h_{\mathrm{Re}\,(2\rho)}(x)$ on $[0, \infty)$ (see (2.14), (2.21) and (2.23)). Since $|\arg G(u)| \leqq \pi$, one way to estimate $|F(u)|$ from (2.28) is

$$|F(u)| \leqq e^{\pi |\mathrm{Im}\,(2\rho)|}. \tag{2.29}$$

An alternative way to estimate $|F(u)|$ is to note that since $G(0) = 1$ and $\arg G(0) = 0$, for small positive u (2.10) implies

$$\arg G(u) = \beta + \arg \left(\sinh \left(\frac{u^{1/2}\, e^{-i\beta}}{2} \right) \right). \tag{2.30}$$

Noting that

$$\sinh \left(\frac{u^{1/2}\, e^{-i\beta}}{2} \right) = \exp \left(\frac{u^{1/2}\, e^{-i\beta}}{2} \right) \left(\frac{1 - \exp\,(-u^{1/2}\, e^{-i\beta})}{2} \right), \tag{2.31}$$

equation (2.30) implies, for small positive u,

$$\arg G(u) = \beta - \frac{u^{1/2}}{2} \sin \beta + \arg\,(1 - \exp\,(-u^{1/2}\, e^{-i\beta})). \tag{2.32}$$

The right-hand side of (2.32) represents $\arg G(u)$ for those u for which it is less than or equal to π in absolute value. When $u > 0$, $\mathrm{Re}\,(1 - \exp\,(-u^{1/2}\, e^{-i\beta})) > 0$, so that

$$\arg\,(1 - \exp\,(-u^{1/2}\, e^{-i\beta})) = \tan^{-1} \left(\frac{-e^{-u^{1/2} \cos \beta} \sin\,(u^{1/2} \sin \beta)}{1 - e^{-u^{1/2} \cos \beta} \cos\,(u^{1/2} \sin \beta)} \right). \tag{2.33}$$

Employing the inequalities

$$\frac{e^{-x}}{1 - e^{-x}} < \frac{1}{x} \quad (x > 0); \qquad \frac{\sin x}{x} < 1 \quad (x \neq 0)$$

we have, for $u > 0$,

$$\begin{aligned} \frac{e^{-u^{1/2} \cos \beta} |\sin\,(u^{1/2} \sin \beta)|}{1 - e^{-u^{1/2} \cos \beta} \cos\,(u^{1/2} \sin \beta)} &\leqq \frac{e^{-u^{1/2} \cos \beta} |\sin\,(u^{1/2} \sin \beta)|}{1 - e^{-u^{1/2} \cos \beta}} \\ &\leqq \frac{|\sin\,(u^{1/2} \sin \beta)|}{u^{1/2} \cos \beta} \\ &\leqq |\tan \beta|. \end{aligned} \tag{2.34}$$

Since $\tan^{-1}$ is an odd increasing function and $-(\pi/4) \leq \beta \leqq (\pi/4)$, (2.33) and (2.34) yield

$$|\arg\,(1 - e^{-u^{1/2} e^{-i\beta}})| \leqq |\beta| \qquad (u \geqq 0). \tag{2.35}$$

Equation (2.35) implies that

$$\mathrm{sgn}\,[\beta + \arg\,(1 - e^{-u^{1/2} e^{-i\beta}})] = \mathrm{sgn}\,\beta, \tag{2.36}$$

and, therefore, the absolute value of the right-hand side of (2.32) is less than π for all u satisfying $0<u^{1/2}|\sin\beta|/2<\pi$. For this range of u, the triangle inequality and (2.35) applied to (2.32) then imply

$$|\arg G(u)-\beta|\leqq\frac{u^{1/2}|\sin\beta|}{2}+|\beta|. \tag{2.37}$$

We now note that (2.37) holds for all u, for when $u^{1/2}|\sin\beta|/2\geqq\pi$ the inequality there is still satisfied because $|\arg G(u)-\beta|\leqq\pi+|\beta|$ for all u. Combining (2.15), (2.27), (2.28), and (2.8) then yields

$$\begin{aligned}|R_n(w,\rho)|\leqq|B_{2n}^{(|2\rho|)}(|\rho|)|\,&\frac{1}{2\Gamma(1-2\rho)}\\ &\cdot\int_0^\infty e^{-u^{1/2}(\operatorname{Re}(we^{-i\beta}))}u^{n-\operatorname{Re}\rho-\frac{1}{2}}\,e^{\operatorname{Im}(2\rho)(\arg G(u)-\beta)}\,du.\end{aligned} \tag{2.38}$$

From (2.37), it follows that

$$e^{\operatorname{Im}(2\rho)(\arg G(u)-\beta)}\leqq e^{|\operatorname{Im}(2\rho)|(u^{1/2}|\sin\beta|/2+|\beta|)} \tag{2.39}$$

Assuming now that the second inequality in (1.6) is satisfied, we use the inequality (2.39) in the integrand on the right side of (2.38), and, evaluating the resulting integral, we obtain the error bound in (1.5).

3. Discussion of the error bound and conclusion. For prescribed values of w and ρ the magnitude of the bound on the right-hand side of (1.5) depends on the value assigned to β. With $\arg w=\alpha$, the ratio of the absolute value of the first neglected term in the series (1.4) to the right side in (1.5) is

$$\begin{aligned}&\frac{|\Gamma(1-2\rho+2n)|}{|\Gamma(1-\operatorname{Re}(2\rho)+2n)|}\,\frac{|B_{2n}^{(2\rho)}(\rho)|}{|B_{2n}^{(|2\rho|)}(|\rho|)|}\,e^{-\alpha\operatorname{Im}(2\rho)-|\beta||\operatorname{Im}(2\rho)|}\\ &\quad\cdot\left(\cos(\alpha-\beta)-\frac{|\operatorname{Im}\rho||\sin\beta|}{|w|}\right)^{2n+1-\operatorname{Re}(2\rho)}.\end{aligned} \tag{3.1}$$

For large $|w|$ this ratio is approximately an $O(1)$ function of w,ρ, and β multiplied by $(\cos(\alpha-\beta))^{2n+1-\operatorname{Re}(2\rho)}$. When α lies in the interval $[-\pi/4,\pi/4]$ we can choose $\beta=\alpha$ so that $\cos(\alpha-\beta)=1$. Note that in this case the exponential term in (3.1) disappears when $\alpha\operatorname{Im}(2\rho)<0$. However, when α lies in $(\pi/4,3\pi/4)$ or $(-3\pi/4,-\pi/4)$ β must differ from α. As α approaches $3\pi/4$ or $-3\pi/4$ β must approach $\pi/4$ or $-\pi/4$, respectively, and the bound (1.5) starts to exceed the absolute value of the first neglected term by an increasingly large factor. In practice, however, (2.2) can and should be used to avoid this possibility. We stress that the restriction of β to $[-\pi/4,\pi/4]$ is necessary to obtain the error bound. A similar restriction is required to obtain one of the standard forms of error bound for Stirling's series for $\ln\Gamma(z)$ (see [6, p. 252]).

We can use

$$\begin{aligned}&\frac{\Gamma(1-\operatorname{Re}(2\rho)+2n)}{|\Gamma(1-2\rho)|(2n)!}\\ &\quad=\frac{(2n-\operatorname{Re}(2\rho))(2n-1-\operatorname{Re}(2\rho))\cdots(1-\operatorname{Re}(2\rho))\Gamma(1-\operatorname{Re}2\rho)}{(2n)!|\Gamma(1-2\rho)|},\end{aligned} \tag{3.2}$$

and (see [4, p. 38])

$$\begin{aligned}\frac{\Gamma(x)^2}{|\Gamma(x+iy)|^2}&=\left(1+\frac{y^2}{x^2}\right)\prod_{s=1}^{\infty}\left(1+\frac{y^2}{(x+s)^2}\right)\\&\leqq\left(1+\frac{y^2}{x^2}\right)\prod_{s=1}^{\infty}\left(1+\frac{y^2}{s^2}\right)\qquad(x>0),\end{aligned}\tag{3.3}$$

together with

$$\prod_{s=1}^{\infty}\left(1+\frac{y^2}{s^2}\right)=\frac{\sinh \pi y}{\pi y},\qquad \frac{\sinh \pi y}{\pi y}\leqq e^{\pi|y|}\tag{3.4}$$

to obtain from (3.2) the inequality

$$\frac{\Gamma(1-\operatorname{Re}2\rho+2n)}{|\Gamma(1-2\rho)|(2n)!}\leqq|1-2\rho|\,e^{\pi|\operatorname{Im}\rho|}.\tag{3.5}$$

Using (3.5) in (1.5) then gives the bound in (1.6).

To conclude, note that we could also have chosen the bound (2.29) for $|F(u)|$ instead of the bound used in (2.37). For large $\operatorname{Re}(w\,e^{-i\beta})$, the resulting difference in the error bound (1.5) is a factor of $e^{\pi|\operatorname{Im}(2\rho)|}$ instead of $e^{|\beta||\operatorname{Im}(2\rho)|}$. Since $|\beta|\leqq\pi/4$, the bound obtained from (2.37) is usually better. Finally, observe that the bound on the right side of (1.5) reduces to the absolute value of the first neglected term in the real case, for which $\operatorname{Im}(2\rho)=0$ and $\beta=0$. In fact, by the results in [2], the error in this case is numerically less than, and has the same sign as, the first neglected term.

Acknowledgment. The author acknowledges the support of the Naval Postgraduate School Research Program.

REFERENCES

[1] J. L. FIELDS, *A note on the asymptotic expansion of a ratio of gamma functions*, Proc. Edinburgh Math. Soc., 15 (1966), pp. 43-45.

[2] C. L. FRENZEN, *Error bounds for asymptotic expansions of the ratio of two gamma functions*, SIAM J. Math. Anal., 18 (1987), pp. 890-896.

[3] Y. L. LUKE, *The Special Functions and Their Approximations*, Vol. 1, Academic Press, New York, 1969.

[4] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[5] J. RIORDAN, *An Introduction to Combinatorial Analysis*, Princeton University Press, Princeton, NJ 1980.

[6] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, 1973.

[7] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1941.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 512–524, March 1992

011

FUNCTIONAL INEQUALITIES FOR HYPERGEOMETRIC FUNCTIONS AND COMPLETE ELLIPTIC INTEGRALS*

G. D. ANDERSON[†], M. K. VAMANAMURTHY,[‡] AND M. VUORINEN[§]

Abstract. The authors study monotoneity and convexity properties of the Gaussian hypergeometric function, particularly the special cases of complete elliptic integrals. They also prove functional inequalities for these functions and various combinations of them and present some conjectures about inequalities for these functions.

Key words. elliptic integral, hypergeometric function

AMS(MOS) subject classifications. primary 33E05, 33C05; secondary 33C75

1. Introduction. For real numbers a, b, c such that $c \neq -p, p = 0, 1, 2, 3, \cdots$, the Gaussian *hypergeometric function* is defined by

$$F(a,b;c;x) = \sum_{n=0}^{\infty} \frac{(a,n)(b,n)}{(c,n)} \frac{x^n}{n!} \tag{1.1}$$

for $-1 < x < 1$, where we have used the ascending factorial notation $(a,n) = a(a+1)\cdots(a+n-1)$ for $n = 1, 2, 3, \cdots$ and $(a,0) = 1$. In the exceptional case $c = -p$, $p = 0, 1, 2, \cdots$, $F(a,b;c;x)$ is defined also if $a = -j$ or $b = -j$, where $j = 0, 1, 2, \cdots$ and $j \leq p$. We shall study some properties of this function and of its special cases, namely, the complete elliptic integrals

$$\mathcal{K}(r) = \frac{\pi}{2} F\left(\frac{1}{2}, \frac{1}{2}; 1; r^2\right), \quad \mathcal{E}(r) = \frac{\pi}{2} F\left(-\frac{1}{2}, \frac{1}{2}; 1; r^2\right). \tag{1.2}$$

The basic properties of these functions can be found in [WW]. Numerous identities satisfied by these functions are given, for example, in [AS], [C2], [SO], and [PBM]. For a partial survey and bibliography of the hypergeometric function, see [A].

We now state some results of this paper.

THEOREM 1.1. *For $a \in I\!R$, $c > b > 0$, $x \in (-1, 1)$,*

(1) $F(a,b;c;x) \cdot F(-a,b;c;x) \geq 1$,

(2) $F(a,b;c;x) + F(-a,b;c;x) \geq 2$.

Moreover, (2) *also holds for $a, x \in (0,1)$ and $b, c \in (0, \infty)$. In particular,*

(3) $\mathcal{K}(x)\mathcal{E}(x) \geq \pi^2/4$, $\mathcal{K}(x) + \mathcal{E}(x) \geq \pi$ *for all $x \in [0, 1)$.*

In this paper, we denote $r' = \sqrt{1 - r^2}$ for any $r \in (0, 1)$. In [AVV2, Thm. 2.2(3)] it was shown that $r'\mathcal{K}(r)^2$ is strictly decreasing from $[0, 1)$ onto $(0, \pi^2/4]$. In this paper we extend this result and prove an analogous result for the other elliptic integral $\mathcal{E}$.

* Received by the editors June 20, 1990; accepted for publication (in revised form) July 23, 1991.

[†] Department of Mathematics, Michigan State University, East Lansing, Michigan 48824. E-mail: 21144gda@msu.bitnet. The research of this author was completed while he was a Visiting Professor at Indiana University in the ITM/MUCIA Cooperative Program in Malaysia.

[‡] Department of Mathematics and Statistics, University of Auckland, Auckland, New Zealand. E-mail: vamanamu@mat.aukuni.ac.nz.

[§] Department of Mathematics, University of Helsinki, Hallituskatu 15, Helsinki, Finland. E-mail: vuorinen@convex.csc.fi. The research of this author was supported in part by grants from the Academy of Finland, the University of Auckland, and the New Zealand Mathematical Society. The paper was completed during a visit to the Mittag-Leffler Institute on a grant from the Royal Swedish Academy of Sciences.

THEOREM 1.2. *For each $c \in [\frac{1}{2}, 2]$ the function $f(r) \equiv (r')^c \mathcal{K}(r)$ is decreasing and concave from $[0,1)$ onto $(0, \pi/2]$; the function $\sqrt{f}$ is also decreasing and concave.*

THEOREM 1.3. *The function $f(r) \equiv (r')^c \mathcal{E}(r)$ is increasing on $(0,1)$ if and only if $c \le -\frac{1}{2}$ and decreasing if and only if $c \ge 0$. Further, if $-2 \le c \le -1$, then f is convex.*

The elliptic integrals $\mathcal{K}(r)$ and $\mathcal{E}(r)$ are indispensable in many applications in mathematics as well as in physics and engineering. Some examples are: rectification of curves, quadrature of surfaces, and electromagnetic field computations [B], [L]. Carlson [C2], [C3] has found a unified approach for effective computation of many such integrals (see also [PT1]). Many generic classes of special funtions of mathematical physics, such as Chebyshev polynomials and Legendre functions, are special cases of the Gaussian hypergeometric function (1.1), while others, such as Bessel functions and parabolic cylinder functions, are limiting cases of (1.1). (For an extensive list of special cases see [SO, pp. 157–164], [AS], and [PBM].) Some generalized elliptic-type definite integrals that occur in radiation field problems are also special cases of the hypergeometric function (1.1) [KCH], [KLH].

The complete elliptic integrals also arise in a number of problems of geometric function theory [LV], [AVV1], [Vu]—our work was in part motivated by this fact. We also state some conjectures on the asymptotic behavior of $\mathcal{K}(r)$ near $r = 1$, refining the well-known asymptotic formulas in [BF].

2. The hypergeometric function.

In this section we study some monotoneity and convexity properties of the hypergeometric function and complete elliptic integrals. The evaluation of the hypergeometric function can be based on the numerical solution of the differential equation satisfied by this function (cf. [PT2]). A table of the values of $F(a,b;c;r)$ is given, e.g., in [CK].

THEOREM 2.1. (1) *Let $a, b, c \in (0, \infty)$. Then $f_1(x) \equiv F(a,b;c;x)$ is strictly increasing and convex on $[0,1)$. In particular, $g_1(x) \equiv \mathcal{K}(x)$ is strictly increasing and convex on $[0,1)$.*

(2) *Let $a \in (0,1), b, c \in (0,\infty)$. Then $f_2(x) \equiv F(-a,b;c;x)$ is strictly decreasing and concave on $[0,1)$. In particular, $g_2(x) \equiv \mathcal{E}(x)$ is strictly decreasing and concave on $[0,1)$.*

(3) *For $a \in (-1,1), b, c \in (0,\infty)$ the function $f_3(x) \equiv F(a,b;c;x) + F(-a,b;c;x)$ is strictly increasing and convex on $[0,1)$. In particular, the function $g_3(x) \equiv \mathcal{K}(x) + \mathcal{E}(x)$ is strictly increasing and convex on $[0,1)$.*

(4) *Let $a, b, c \in (0,\infty)$. Then $f_4(x) \equiv \frac{1}{x}(F(a,b;c;x) - 1)$ is strictly increasing and convex on $[0,1)$. In particular, the function $g_4(x) \equiv x^{-2}(\mathcal{K}(x) - \frac{\pi}{2})$ is strictly increasing and convex from $(0,1)$ onto $(\pi/8, \infty)$.*

(5) *Let $a \in (0,1), b, c \in (0,\infty), b < c$. Then $f_5(x) \equiv \frac{1}{x}(1 - F(-a,b;c;x))$ is strictly increasing and convex on $[0,1)$. In particular, $g_5(x) \equiv x^{-2}(\frac{\pi}{2} - \mathcal{E}(x))$ is strictly increasing and convex from $(0,1]$ onto $(\frac{\pi}{8}, \frac{\pi}{2} - 1]$.*

(6) *Let $a \in (0,1), b, c \in (0,\infty)$. Then $f_6(x) \equiv \frac{1}{x}(F(a,b;c;x) - F(-a,b;c;x))$ is strictly increasing and convex on $[0,1)$. In particular, $g_6(x) \equiv x^{-2}(\mathcal{K}(x) - \mathcal{E}(x))$ is strictly increasing and convex from $[0,1)$ onto $[\frac{\pi}{4}, \infty)$.*

(7) *Let $a \in [\frac{1}{2}, 1], b, c \in (0,\infty), a + b \le c$. Then*

$$f_7(x) \equiv \frac{1}{x}(F(-a,b;c;x) - (1-x)F(a,b;c;x))$$

is strictly increasing and convex from $(0,1)$ *onto* $(1-(2ab/c), B(b,a+c-b)/B(b,c-b))$. *In particular,* $g_7(x) \equiv x^{-2}(\mathcal{E}(x)-(1-x^2)\mathcal{K}(x))$ *is strictly increasing and convex from* $(0,1)$ *onto* $(\frac{\pi}{4},1)$.

Proof. The functions $f_1, 1-f_2, f_3, f_4, f_5$, and f_6 are represented by power series with nonnegative coefficients in the specified ranges of a,b,c,x. Moreover, $g_j(x) = (2/\pi)f_j(x^2)$ for $j=1,\cdots,7$, with $a=b=\frac{1}{2}, c=1$.

We shall give a proof for (7), since it is nontrivial. Using the series (1.1) we may write

$$f_7(x) = 1 - \frac{2ab}{c} + \sum_{n=1}^{\infty} C_n x^{2n},$$

where

$$C_n = \frac{(a,n)(b,n)}{(c,n)n!} - \frac{((a,n+1)-(-a,n+1))(b,n+1)}{(c,n+1)(n+1)!}.$$

Now

$$\begin{aligned} C_n &= \frac{(b,n)}{(c,n)(n+1)!} \\ &\quad \times \left[(n+1)(a,n) - \frac{b+n}{c+n}(a(a+1)\cdots(a+n)+a(-a+1)\cdots(-a+n))\right] \\ &> \frac{(b,n)}{(c,n)(n+1)!}[(n+1)(a,n)-(a,n)(a+n)-a(-a+1)\cdots(-a+n)] \\ &= \frac{a(b,n)(-a+1)}{(c,n)(n+1)!}[(a+1)\cdots(a+n-1)-(-a+2)\cdots(-a+n)], \end{aligned}$$

which is nonnegative since $a+m-1 \geq -a+m$ for $m=1,2,\cdots,n$ and $a \in [\frac{1}{2},1]$. Thus $f_7(x)$ is increasing and convex, and $f_7(0+) = 1-(2ab/c)$. The value of $f_7(1-)$ follows from [WW, §14.11, pp. 281–282]. Putting $a=b=\frac{1}{2}, c=1$ in the above theorem we see that $g_7(x)$ is strictly increasing and convex from $(0,1)$ on $(\frac{\pi}{4},1)$. □

Proof of Theorem 1.1. In (1), by a well-known integral representation [R, Thm. 16, p. 47],

$$(2.1) \quad \begin{cases} F(a,b;c;x) = \dfrac{1}{B(b,c-b)} \displaystyle\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-xt)^{-a}dt, \\ F(-a,b;c;x) = \dfrac{1}{B(b,c-b)} \displaystyle\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-xt)^{a}dt, \end{cases}$$

for $c > b > 0$ and $x \in (-1,1)$. Hence by Hölder's inequality [M],

$$F(a,b;c;x)\cdot F(-a,b;c;x) \geq \frac{1}{B(b,c-b)^2}\left(\int_0^1 t^{b-1}(1-t)^{c-b-1}dt\right)^2 = 1.$$

For (2), in the first case it is clear that both summands are positive; hence by (1) and the arithmetic-geometric mean inequality,

$$F(a,b;c;x)+F(-a,b;c;x) \geq 2[F(a,b;c;x)\cdot F(-a,b;c;x)]^{1/2} \geq 2.$$

In the second case the power series has constant term 2 and all coefficients positive, so that (2) follows immediately.

Finally, (3) follows from (1) and (2) if we take the special values $a = b = \frac{1}{2}$, $c = 1$. □

Remark 2.1. Theorem 1.1 (1) can also be derived from Carlson's logarithmic convexity results in [C1], as he has pointed out to us.

LEMMA 2.2. *Given* $a \in (0,1)$, $0 < b < c$, $(1+r)F(a,b;c;r^2) \geq F(a,b;c;r)$ *for* $r \in [0,1)$. *In particular,* $(1+r)\mathcal{K}(r) \geq \mathcal{K}(\sqrt{r})$ *for* $r \in [0,1)$.

Proof. This follows directly from the power series expansion for F. □

LEMMA 2.3. *For* $a,b,c,d > 0$ *and* $0 < r < 1$,

(1) $F(a,b;c;r) \leq F(a,b+d;c+d;r)$ *for* $c \geq b$,

(2) $F(a,b;c;r) \geq F(a,b+d;c+d;r)$ *for* $c \leq b$.

Proof. Since the proofs of the two cases are similar we prove only (1). Because both series in (1) have positive coefficients it is enough to show that the coefficients on the right side are larger, that is, $(a,n)(b,n)/(c,n) \leq (a,n)(b+d,n)/(c+d,n)$. But this holds if and only if $c \geq b$. □

Remark 2.2. The asymptotic relation

$$F(a,b;a+b;r) \sim -\frac{\log(1-r)}{B(a,b)} \quad \text{as } r \to 1 \tag{2.2}$$

is well known [WW, Ex. 18, p. 299], [Ch, pp. 266–7] (for a generalization see [SS] and the references there). See also [H].

We next prove an inequality relating the two functions in (2.2).

THEOREM 2.4. *Denote* $g(x) = (1/x)\log(1-x)$. *Then*

(1) *For* $a,b,x \in (0,1)$ *we have*

$$\frac{-g(x)}{B(a,b)} < F(a,b;a+b;x) < -g(x).$$

The lower estimate is sharp at $x = 1$ *and the upper estimate is sharp at* $x = 0$.

(2) *For* $a,b \in (1,\infty)$ *and* $x \in (0,1)$ *the inequalities in* (1) *are reversed.*

(3) *For* $a = b = 1$ *we have* $F(1,1;2;x) = -g(x)/B(1,1) = -g(x)$.

Proof. Part (3) follows by definition. The first part of (1) is proved by means of the integral representation (2.1). We observe that for $t \in (0,1)$ we have $t^{b-1} > 1$ and $[(1-t)/(1-xt)]^{a-1} > 1$, and the first inequality in (1) follows by simple integration. For the second inequality in (1) we compare the series expansions, noting that it is sufficient to prove that

$$\frac{(a,n)(b,n)}{(a+b,n)} < \frac{n!}{(n+1)} \tag{2.3}$$

for $n = 1,2,3,\cdots$. Inequality (2.3) follows by induction, as we now show. For $n = 1$, it reduces to $2ab < a+b$, which is true for $a,b \in (0,1)$ since $2ab \leq a^2+b^2 < a+b$. Next, assuming (2.3) for n, it is true for $n+1$ if and only if $n(ab-1)+2ab-a-b < 0$, which is true. The proof of (2) is similar. Sharpness at $x = 1$ follows from (2.2), while at $x = 0$ it is obvious. □

3. Complete elliptic integrals. In this section we continue the study of the integrals $\mathcal{K}$ and $\mathcal{E}$ begun in [AV2] and [AVV2]. In particular, we obtain some estimates for these integrals in terms of elementary functions and prove a monotoneity property for a function defined in terms of them. These inequalities supplement similar previously known inequalities that appear in many books on special functions, e.g. [AS, 17.3.33–36], [SO, 61:9:1–8]. Note, however, that our $\mathcal{K}(r)$ and $\mathcal{E}(r)$ defined in (1.2) are $\mathcal{K}(\sqrt{r})$ and $\mathcal{E}(\sqrt{r})$ in the notation of [AS, 17.3.9–10].

The complete elliptic integrals $\mathcal{K}(r)$ and $\mathcal{E}(r)$ of the first and second kind, respectively, are much better known than the hypergeometric function $F(a,b;c;r)$ of which they are special cases (cf. (1.1)). Thus algorithms for the computation of $\mathcal{K}(r)$ and $\mathcal{E}(r)$ are given in many program libraries and they are tabulated in many tables (e.g. [AS], [BF]). An algorithm for the computation of these functions can be based on Gauss' arithmetic-geometric mean iteration [AS, 17.6], which is adequate for most purposes. See also [BB] and [PT1].

For later reference we list two basic identities due to Landen [BF, 163.01, 164.02]; cf. [WW, p. 507]:

$$(3.1) \qquad \mathcal{K}\left(\frac{2\sqrt{r}}{1+r}\right) = (1+r)\mathcal{K}(r), \quad \mathcal{K}\left(\frac{1-r}{1+r}\right) = \frac{1}{2}(1+r)\mathcal{K}'(r).$$

In studying the convexity properties of $\mathcal{K}(r)$ and $\mathcal{E}(r)$ we shall need the following lemma.

LEMMA 3.1. *Let I, J be intervals in $I\!R$, $f : I \longrightarrow J$ be decreasing and concave, and $g : J \longrightarrow I\!R$ be increasing and concave. Then $h \equiv g \circ f : I \longrightarrow I\!R$ is decreasing and concave.*

Proof. For concavity let $a, b \in I$ and $t \in (0,1)$. Then

$$(g \circ f)((1-t)a + tb) \geq g((1-t)f(a) + tf(b)) \geq (1-t)(g \circ f)(a) + t(g \circ f)(b).$$

That $g \circ f$ is decreasing is obvious. □

Proof of Theorem 1.2. By [BF, 710.00], $f'(r) = -(r')^{c-2}g(r)$, where $g(r) \equiv cr\mathcal{K} - (\mathcal{E} - (r')^2\mathcal{K})/r$. For $c \geq 1/2$, differentiation [BF, 710.00, 710.04] gives

$$g'(r) = \frac{c\mathcal{E}}{(r')^2} + \frac{\mathcal{E} - \mathcal{K}}{r^2} \geq \frac{h(r)}{2r^2(r')^2},$$

where $h(r) \equiv (\mathcal{E} - (r')^2\mathcal{K}) - (r')^2(\mathcal{K} - \mathcal{E})$. But $h'(r) = 3r(\mathcal{K} - \mathcal{E}) > 0$, and it follows that $g(r) > g(0+) = 0$ for $0 < r < 1$. Hence f' is negative and decreasing, so that f is decreasing and concave on $(0,1)$. The limit $f(0+) = \pi/2$ is clear, while $f(1-) = 0$ follows from [AVV2, Thm. 2.2(2)]. Since $\sqrt{t}$ is concave and increasing on $(0,\infty)$, it follows from Lemma 3.1 that $\sqrt{f}$ is decreasing and concave on $(0,1)$. □

Proof of Theorem 1.3. By [BF, 710.02]

$$f'(r) = \frac{(r')^{c-2}}{r}[-cr^2\mathcal{E} - (r')^2(\mathcal{K} - \mathcal{E})],$$

which is nonnegative if and only if $-c \geq \sup\{g(r)h(r) : 0 < r < 1\}$, where

$$g(r) \equiv \frac{r'(\mathcal{K} - \mathcal{E})}{r^2}, \qquad h(r) \equiv \frac{r'}{\mathcal{E}}.$$

Now $h'(r) = -(\mathcal{E} - (r')^2\mathcal{K})/(rr'\mathcal{E}^2) < 0$, so that $h(r)$ is decreasing on $(0,1)$. Next, by [BF, 710.05],

$$r^4 g'(r) = \frac{r}{r'}\varphi(r), \qquad \varphi(r) \equiv (\mathcal{E} - (r')^2\mathcal{K}) - (\mathcal{K} - \mathcal{E}).$$

Since $\varphi(0) = 0$ and $\varphi'(r) = (r/(r')^2)((r')^2\mathcal{K} - \mathcal{E}) < 0$ for $0 < r < 1$ by [BF, 710.04, 710.05] and [AVV2, Thm. 2.2(7)], we have $g'(r) < 0$, so that $g(r)$ is decreasing on $(0,1)$. Hence $g(r)h(r)$ is decreasing on $(0,1)$. By l'Hôpital's rule and [BF, 710.05] we have

$$\lim_{r\to 0+} g(r)h(r) = \lim_{r\to 0+} \frac{2(\mathcal{K} - \mathcal{E})}{\pi r^2} = \lim_{r\to 0+} \frac{\mathcal{E}}{\pi(r')^2} = \frac{1}{2}.$$

Thus $\sup\{g(r)h(r) : 0 < r < 1\} = g(0+)h(0+) = 1/2$, so that $f'(r) \geq 0$ if and only if $c \leq -\frac{1}{2}$.

Next, $f'(r) \leq 0$ if and only if $c \geq -\inf\{g(r)h(r) : 0 < r < 1\} = -g(1-)h(1-) = 0$.

Finally, for $-2 \leq c \leq -1$, we write $f'(r)$ as $(r')^{c-2}\psi(r)$, where $\psi(r) \equiv -cr\mathcal{E} - ((r')^2/r)(\mathcal{K} - \mathcal{E})$. Then

$$\psi'(r) = (2 + c + (r')^2 r^{-2})(\mathcal{K} - \mathcal{E}) - (c+1)\mathcal{E} \geq 0 \text{ on } (0,1),$$

so that f is convex. □

THEOREM 3.2. (1) *The function* $f_1(r) \equiv r^{-4}((\pi/(2\sqrt{r'})) - \mathcal{K}(r))$ *is strictly increasing and convex from* $(0,1)$ *onto* $(\pi/128, \infty)$.

(2) *The function* $f_2(r) \equiv (\mathcal{E}(r') - 1)/(r^2 \log(1/r))$ *is increasing and convex from* $(0,1)$ *onto* $(\frac{1}{2}, \infty)$.

(3) *The function* $f_3(r) \equiv r^2\mathcal{K}(r) + c\log r'$ *is increasing if and only if* $c \leq 1$ *and decreasing if and only if* $c \geq \pi$. *In particular, for* $0 < r < 1$,

$$1 < \frac{r^2\mathcal{K}(r)}{\log(1/r')} < \pi.$$

The lower bound is sharp as $r \to 1-$, *and the upper bound is sharp as* $r \to 0+$.

Proof. The properties of f_1 follow if we subtract the series for $\mathcal{K}(r)$ from the series

$$\frac{\pi}{2\sqrt{r'}} = \frac{\pi}{2}(1 - r^2)^{-1/4} = \frac{\pi}{2}\left(1 + \frac{1}{4}r^2 + \frac{1\cdot 5}{4\cdot 8}r^4 + \frac{1\cdot 5\cdot 9}{4\cdot 8\cdot 12}r^6 + \cdots\right).$$

For (2) we use the series expansion

$$f_2(r) = \sum_{n=1}^{\infty} a_n \left(1 + \frac{(\log 4 - 2b_n) + \frac{1}{(2n-1)(2n)}}{\log(1/r)}\right) r^{2n-2},$$

where

$$a_1 = \frac{1}{2}, \qquad a_n = \left(\frac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots(2n-2)}\right)^2 \frac{2n-1}{2n} \quad \text{for } n \geq 2,$$

and

$$b_n = \frac{1}{1\cdot 2} + \frac{1}{3\cdot 4} + \cdots + \frac{1}{(2n-1)(2n)}$$

[BB, (1.3.11)]. Since

$$b_n = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{2n-1} - \frac{1}{2n} < \log 2$$

it follows that $f_2(r)$ is strictly increasing and convex. Clearly $f_2(0+) = a_1 = \frac{1}{2}$ and $f_2(1-) = \infty$.

In part (3) since $\mathcal{E}(r) + (r')^2\mathcal{K}(r)$ is decreasing from $[0,1]$ onto $[1,\pi]$ by [AVV2, Thm. 2.2(3)] it follows that

$$f_3'(r) = \frac{r}{(r')^2}[\mathcal{E}(r) + (r')^2\mathcal{K}(r) - c]$$

is nonnegative if and only if

$$c \leq \inf\{\mathcal{E}(r) + (r')^2\mathcal{K}(r) : 0 < r < 1\} = 1$$

and $f_3'(r) \leq 0$ if and only if

$$c \geq \sup\{\mathcal{E}(r) + (r')^2\mathcal{K}(r) : 0 < r < 1\} = \pi.$$

The bounds for $r^2\mathcal{K}(r)/\log(1/r')$ clearly follow, and they are sharp since $\mathcal{K}(0) = \pi/2$ and $\lim_{r\to 1-} \mathcal{K}(r)/\log(1/r') = 1$ [BF, 112.01]. □

COROLLARY 3.3. *For* $0 < r < 1$,

$$\frac{1+r}{4r}\log\left(\frac{1+r}{1-r}\right) < \mathcal{K}(r) < \pi\frac{1+r}{4r}\log\left(\frac{1+r}{1-r}\right).$$

Proof. Replace r by $2\sqrt{r}/(1+r)$ in the inequalities in Theorem 3.2(3) and use (3.1). □

Remark 3.1. The two lower bounds for $\mathcal{E}(r)$ in Theorem 2.1(5) and 3.2(2) are not comparable.

In the proof of our next theorem, which is close to the conjecture in [AVV2, §6], we need a technical lemma.

LEMMA 3.4. *For any natural number* n,

$$\frac{\pi}{2}\frac{(\frac{1}{2},n)^2}{(n!)^2}\left(1 + (8.5)\frac{(2n+1)^2}{(2n+2)^2}\right) > \frac{9}{2n+2}.$$

Proof. For $n = 1,2,3,4,5,6$, the lemma follows by direct calculation. If $n \geq 7$, by Wallis' inequalities [M, p. 192] it is sufficient to prove that

$$\frac{1}{2n+1}\left(1 + (8.5)\frac{(2n+1)^2}{(2n+2)^2}\right) > \frac{9}{2n+2},$$

which is equivalent to $2n^2 - 12n - 5.5 > 0$, true for $n \geq 7$. □

The upper bound in Theorem 3.5 below also follows from [CG, p. 1072, (1.1)], but our proof here is more direct. In connection with the lower bound see Conjecture 3.1(5) and Remark 3.3 below.

THEOREM 3.5. *For* $0 < r < 1$,

$$\frac{9}{8.5 + r^2} < \frac{\mathcal{K}(r)}{\log(4/r')} < \frac{4}{3 + r^2}.$$

Proof. By Wallis' inequalities [M, p. 192] the first inequality is equivalent to

$$(8.5+r^2)\frac{\pi}{2}\left(1+\sum_{n=1}^{\infty}\frac{(\frac{1}{2},n)^2}{(n!)^2}r^{2n}\right) > 9\log 4+\frac{9}{2}r^2+\frac{9}{2}\sum_{n=1}^{\infty}\frac{r^{2n+2}}{n+1}.$$

Comparing coefficients on both sides, for the constant term we have $(8.5)(\pi/2) > 9\log 4$ and for the coefficients of r^2 we have $(\pi/2)(1+\frac{1^2}{2^2}(8.5)) > 9/2$ by direct calculation, while Lemma 3.4 shows that the coefficient of r^{2n+2} on the left exceeds the corresponding one on the right.

For the second inequality, let

$$f(r)=(3+r^2)\mathcal{K}(r)-4\log\frac{4}{r'}.$$

Then $f(0)=3\pi/2-4\log 4<0$ and

$$f(1)=\lim_{r\to 1}[(3+r^2)(\mathcal{K}(r)-\log(4/r'))-(r')^2\log(4/r')]=0$$

by [BF, 112.01]. Thus it is sufficient to prove that f is increasing on $(0,1)$. Now writing

$$r(r')^2f'(r)=g(r)\equiv(3+r^2)(\mathcal{E}-(r')^2\mathcal{K})+2r^2(r')^2\mathcal{K}-4r^2,$$

we have $g(0)=0=g(1)$. It is sufficient to prove that $g(r)>0$ at each critical point in $(0,1)$. Now

$$g'(r)=rh(r),$$

where $h(r)\equiv 4\mathcal{E}+3(r')^2\mathcal{K}-8$ is strictly decreasing on $(0,1)$. Since $h(0)=7\pi/2-8>0$ and $h(1)=-4<0$, h has a unique zero $r_0\in(0,1)$, and we need only prove that $g(r_0)>0$. By numerical estimates we see that $h(0.74)<0<h(0.72)$, so that $0.72<r_0<0.74$. Then, since $(\mathcal{E}-(r')^2\mathcal{K})/r^2$ is increasing and $(r')^2\mathcal{K}$ is decreasing we find from standard tables (cf. [BF, p. 323]) that $g(r_0)/r_0^2>3.31$. □

For later reference we remark that the double inequalities

$$(3.2)\qquad r'<\frac{\mathcal{E}}{\mathcal{K}}<1-\frac{r^2}{2},\qquad \frac{\pi r^2}{4}<\mathcal{E}-(r')^2\mathcal{K}<r^2$$

for $0<r<1$, relating complete elliptic integrals of the first and second kind, were obtained in [AV1, (6)], [AV2, (1)], respectively. The first two inequalities are sharp at $r=0$, the third is asymptotically sharp as $r\to 0$, and the fourth as $r\to 1$. The inequality giving the upper estimate for $\mathcal{E}/\mathcal{K}$ in (3.2) can be written as

$$(3.3)\qquad \mathcal{E}-(r')^2\mathcal{K}<\frac{r^2\mathcal{K}}{2},\qquad 0<r<1.$$

In our next theorems we compare the quotient and product of the elliptic integrals $\mathcal{E}$ and $\mathcal{K}$ with elementary functions. A power series for $\mathcal{E}/\mathcal{K}$ is given in [AS, 17.3.23]. We let arth denote the inverse hyperbolic tangent function.

THEOREM 3.6. *For* $0<r<1$,

$$\frac{1}{2}\left((r')^2+\frac{r}{\operatorname{arth} r}\right)<\frac{\mathcal{E}(r)}{\mathcal{K}(r)}<\frac{r}{\operatorname{arth} r}.$$

Proof. The first inequality is equivalent to

$$f(r) \equiv \log\frac{1+r}{1-r} - \frac{2r\mathcal{K}}{2\mathcal{E}-(r')^2\mathcal{K}} > 0.$$

Since $f(0) = 0$, it is sufficient to prove that $f'(r) > 0$ for $0 < r < 1$. But by [BF, 710.00, 710.02, 710.04] and algebraic simplification we have

$$f'(r) = \frac{4\mathcal{E}(\mathcal{E}-(r')^2\mathcal{K})}{(r')^2(2\mathcal{E}-(r')^2\mathcal{K})^2} > 0.$$

Similarly, the second inequality is equivalent to

$$g(r) \equiv \log\frac{1+r}{1-r} - \frac{2r\mathcal{K}}{\mathcal{E}} < 0;$$

but $g(0) = 0$ and $g'(r) = 2\mathcal{K}(\mathcal{E}-\mathcal{K})/(r')^2\mathcal{E}^2 < 0$ for $0 < r < 1$. □

THEOREM 3.7. *Let* $f(r) = (\mathcal{E}-r'\mathcal{K})/(1-r')^2$, $0 < r < 1$, $f(0) = \pi/8$, $f(1) = 1$. *Then* f *is strictly increasing on* $[0,1]$. *In particular,*

$$\frac{\pi}{8}(1-r')^2 < \mathcal{E}(r) - r'\mathcal{K}(r) < (1-r')^2$$

for $0 < r < 1$. *The first inequality is sharp at* $r = 0$, *and the second inequality is sharp at both zero and* 1.

Proof. The limit $f(1-) = 1$ is clear, while by l'Hôpital's rule and [BF, 710.04, 710.05] we have

$$f(0+) = \lim_{r\to 0+}\frac{\mathcal{K}-\mathcal{E}}{2r^2} = \lim_{r\to 0+}\frac{\mathcal{E}}{4(r')^2} = \frac{\pi}{8}.$$

Next, for $0 < r < 1$, by differentiation [BF, 710.00, 710.02] and algebraic simplification,

$$rr'(1-r')^2 f'(r) = g(r),$$

where $g(r) \equiv (1-r')(\mathcal{K}-\mathcal{E}) - 2(1+r')(\mathcal{E}-r'\mathcal{K})$. Then $g(0) = 0$ and, for $0 < r < 1$, $(r')^2 g'(r) = r(1+2r')(\mathcal{E}-r'\mathcal{K})$, which is positive by the first inequality in (3.2). □

THEOREM 3.8. *Let* $f(r) \equiv \mathcal{E}(r)\mathcal{K}(r)$, $0 \le r < 1$. *Then* f *is strictly increasing and convex from* $[0,1)$ *onto* $[\pi^2/4, \infty)$ *and* $g(r) \equiv \sqrt{r'}f(r)$ *is strictly decreasing from* $[0,1)$ *onto* $(0, \pi^2/4]$. *In particular, for* $0 < r < 1$,

$$\frac{\pi^2}{4} < \mathcal{E}(r)\mathcal{K}(r) < \frac{\pi^2}{4\sqrt{r'}}.$$

Proof. By differentiation [BF, 710.00, 710.02] and algebraic manipulations,

$$f'(r) = \frac{\mathcal{E}^2-(r')^2\mathcal{K}^2}{r(r')^2} = \frac{\mathcal{E}-r'\mathcal{K}}{(1-r')^2}\cdot\left(\frac{\mathcal{E}}{r'}+\mathcal{K}\right)\cdot\frac{(1-r')^2}{r}\cdot\frac{1}{r'}.$$

By Theorems 3.7 and 1.3 the first two factors on the right are positive increasing functions, and it is easy to see that the last two factors are increasing. Hence f is increasing and convex. Next, by Theorem 1.4 and [AVV2, Thm. 2.2(3)] $g(r)$ is the product of two positive strictly decreasing functions $\sqrt{r'}\mathcal{K}(r)$ and $\mathcal{E}(r)$, hence is strictly decreasing. Finally, the limiting values and inequalities are clear. □

COROLLARY 3.9. *For* $0 < r < 1$,

$$\text{(1)} \qquad \frac{\pi}{2}\left(1 - \frac{r^2}{2}\right)^{-1/2} < \mathcal{K}(r) < \frac{\pi}{2}(r')^{-1/2},$$

$$\text{(2)} \qquad \frac{\pi}{2}\sqrt{r'} < \mathcal{E}(r) < \frac{\pi}{2}.$$

Proof. The lower estimates follow from (3.2) and Theorem 3.8. The upper estimate in (1) is a consequence of Theorem 1.2, and in (2) it follows from the fact that $\mathcal{E}$ is decreasing. □

It follows immediately from the definition (1.2) that $\mathcal{K}(r) \geq \pi/2$. We now obtain another pair of significant estimates for $\mathcal{K}(r)$ in terms of elementary functions.

THEOREM 3.10. *For* $0 < r < 1$, *define the functions* f, g *on* $[0,1)$ *by* $f(r) = r\mathcal{K}(r)^2/\log((1+r)/(1-r))$, $f(0) = \pi^2/8$, *and* $g(r) = f(r)/\mathcal{K}(r)$. *Then* f *is strictly increasing and* g *is strictly decreasing. In particular, for* $0 < r < 1$,

$$\left(\frac{\operatorname{arth} r}{r}\right)^{1/2} < \frac{2\mathcal{K}(r)}{\pi} < \frac{\operatorname{arth} r}{r}.$$

Both inequalities are sharp as $r \to 0$. *The second inequality is of the correct order as* $r \to 1$.

Proof. Since f and g are continuous at zero, it is sufficient to prove the monotoneity properties on $(0,1)$. Now

$$\begin{aligned}\left(\log\frac{1+r}{1-r}\right)^2 f'(r) &= \left(\mathcal{K}^2 + 2\mathcal{K}\frac{\mathcal{E} - (r')^2\mathcal{K}}{(r')^2}\right)\log\frac{1+r}{1-r} - \frac{2r\mathcal{K}^2}{(r')^2} \\ &= \frac{\mathcal{K}}{(r')^2}(2\mathcal{E} - (r')^2\mathcal{K})\left(\log\frac{1+r}{1-r} - \frac{2r\mathcal{K}}{2\mathcal{E} - (r')^2\mathcal{K}}\right),\end{aligned}$$

which is positive by Theorem 3.6.

Similarly,

$$\begin{aligned}\left(\log\frac{1+r}{1-r}\right)^2 g'(r) &= \left(\mathcal{K} + \frac{\mathcal{E} - (r')^2\mathcal{K}}{(r')^2}\right)\log\frac{1+r}{1-r} - \frac{2r\mathcal{K}}{(r')^2} \\ &= \frac{\mathcal{E}}{(r')^2}\left(\log\frac{1+r}{1-r} - \frac{2r\mathcal{K}}{\mathcal{E}}\right),\end{aligned}$$

which is negative by Theorem 3.6.

The lower estimate is sharp by continuity, and the upper estimate is of the correct order as $r \to 1$ since, by [BF, 112.01], $\mathcal{K}(r) \sim \operatorname{arth} r$ as $r \to 1$. □

Remark 3.2. The second inequality in Theorem 3.10 can be rewritten in terms of the hypergeometric function as $F(\frac{1}{2}, \frac{1}{2}; 1; r) \leq F(\frac{1}{2}, 1; \frac{3}{2}; r)$ [C2, p. 15], which is true by virtue of Lemma 2.3.

The next result improves an estimate in [AVV2, (1.8)].

THEOREM 3.11. *For* $0 < r < 1$,

$$1 < \sqrt[4]{1+r}\frac{\mathcal{K}(r)}{\mathcal{K}(\sqrt{r})} < \min\{\sqrt[4]{2}, (r')^{-1/2}\}.$$

Proof. Since $r < \sqrt{r} < 2\sqrt{r}/(1+r)$ it follows from [AVV2, Thm. 2.2(3)] and the first identity in (3.1) that

$$r'\mathcal{K}(r)^2 > \sqrt{1-r}\mathcal{K}(\sqrt{r})^2 > \frac{1-r}{1+r}((1+r)\mathcal{K}(r))^2 = (1-r^2)\mathcal{K}(r)^2.$$

Dividing by $\sqrt{1-r}$ and taking square roots we obtain

$$\sqrt{r'}\sqrt[4]{1+r}\mathcal{K}(r) < \mathcal{K}(\sqrt{r}) < \sqrt[4]{1+r}\mathcal{K}(r).$$

Finally, $\sqrt[4]{1+r}\mathcal{K}(r) < \sqrt[4]{2}\mathcal{K}(\sqrt{r})$ since $\mathcal{K}$ is increasing and $r < \sqrt{r}$. □

It is well known that $\mathcal{K}(r)+\log r'$ is strictly decreasing from $(0,1)$ onto $(\log 4, \pi/2)$ [AVV2, Thm. 2.2(1)]. The next result provides a dual of this fact.

THEOREM 3.12. *The function* $f(r) \equiv \sqrt{r}\mathcal{K}(r) + \log r'$ *is strictly increasing from* $(0,1)$ *onto* $(0, \log 4)$. *In particular, for* $0 < r < 1$,

$$\log\frac{1}{r'} < \sqrt{r}\mathcal{K}(r) < \log\frac{4}{r'}.$$

Proof. The limit $f(0+) = 0$ is obvious, while

$$f(1-) = \lim_{r\to 1-}[\sqrt{r}(\mathcal{K} + \log r') + (1-\sqrt{r})\log r'] = \log 4$$

by [BF, 112.01]. For the monotoneity we write $f'(r) = g(r)/(r^{1/2}(r')^2)$, where $g(r) \equiv \mathcal{E} - \frac{1}{2}(r')^2\mathcal{K} - r^{3/2}$ tends to zero as $r \to 1-$. Thus it is sufficient to prove that $g(r)$ decreases on $(0,1)$. But $g'(r) = h(r)/(2r)$, where $h(r) \equiv \mathcal{E} - (r')^2\mathcal{K} - 3r^{3/2}$, and by [AVV2, Thm. 2.2(7)], $h(r) \le r^2 - 3r^{3/2} < 0$ for $0 < r < 1$. □

V. I. Semenov [Se, (6)] has obtained the inequality

$$\frac{2}{\pi}r^2(1-r^2)\mathcal{K}(r)\mathcal{K}(r') \le \min\{h(r), h(r')\},\quad h(r) = r^2\log\frac{4}{r}. \tag{3.4}$$

The following corollary, which is precisely the second inequality of [AVV2, Thm. 4.2(5)], improves upon (3.4).

COROLLARY 3.13. *For* $0 < r < 1$,

$$\frac{2}{\pi}rr'\mathcal{K}(r)\mathcal{K}'(r) \le \min\left\{r\log\frac{4}{r}, r'\log\frac{4}{r'}\right\}.$$

Proof. By symmetry it is sufficient to prove one of these inequalities. By Theorem 3.12 and [AVV2, Thm. 2.2(3)] we have $\sqrt{r'}\mathcal{K}' \le \log(4/r)$ and $(2/\pi)\sqrt{r'}\mathcal{K} \le 1$. Multiplying, we obtain $(2/\pi)r'\mathcal{K}\mathcal{K}' \le \log(4/r)$, and the first inequality follows. □

Conjectures 3.1. Various generalizations of the classical relation (2.2) were obtained by S. Ramanujan [A] (see also [SS]). We here focus on the particular case

$a = b = \frac{1}{2}$ of (2.2). Our computational work supports the validity of the following conjectures, where $0 < r < 1$, $r' = (1-r^2)^{1/2}$:

(1) $\mathcal{K}(r) < \log(1+\frac{4}{r'}) - (\log 5 - \frac{\pi}{2})(1-r)$.

(2) The function $\sqrt[4]{1+r}\mathcal{K}(r)/\mathcal{K}(\sqrt{r})$ is increasing from $[0,1)$ onto $[1, \sqrt[4]{2})$.

(3) The function $(\mathcal{K}(r) - \log(4/r'))/((r')^2 \log(4/r'))$ is increasing from $(0,1)$ onto $((\pi/\log 16) - 1, \frac{1}{4})$.

(4) The function $(\mathcal{E}(r) - (r')^2\mathcal{K}(r))(\mathcal{E}(r)-1)/(r')^2(\mathcal{K}(r)-\mathcal{E}(r))$ is increasing from $(0,1)$ onto $(\frac{1}{2}, \frac{\pi}{2} - 1)$.

(5) (cf. [AVV2, 6.4(2)]) The inequality in Theorem 3.5 can be replaced by

$$9 < \frac{(8+r^2)\mathcal{K}(r)}{\log(4/r')} < 9.1.$$

(6) $(r')^2 \leq \dfrac{r'\exp(\mathcal{K}(r)) - 4}{\exp\left(\frac{\pi}{2}\right) - 4} \leq \dfrac{2\sqrt{1-r}}{2-r}$.

Remark 3.3. After this manuscript had been completed R. Kühnau kindly informed the authors about his work in [K] related to conjecture 3.1(5).

Acknowledgments. We are grateful to Professor B. C. Carlson and to the referees for their helpful comments.

REFERENCES

[AS] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Dover, New York, 1965.

[AV1] G. D. Anderson and M. K. Vamanamurthy, *Affine mappings and elliptic functions*, Publ. Inst. Math. (Beograd) (N.S.), 11 (1971), pp. 85–87.

[AV2] ———, *Inequalities for elliptic integrals*, Publ. Inst. Math. (Beograd) (N.S.), 37 (1985), pp. 61–63.

[AVV1] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Special functions of quasiconformal theory*, Exposition. Math., 7 (1989), pp. 97–138.

[AVV2] ———, *Functional inequalities for complete elliptic integrals and their ratios*, SIAM J. Math. Anal., 21 (1990), pp. 536–549.

[A] R. Askey, *Ramanujan and hypergeometric and basic hypergeometric series*, Ramanujan International Symposium on Analysis, December 26–28, 1987, N.K. Thakara, ed., Pune, India.

[BB] J. M. Borwein and P. B. Borwein, *Pi and the* AGM, John Wiley & Sons, New York, 1987.

[B] F. Bowman, *Introduction to Elliptic Functions with Applications*, Dover, New York, 1961.

[BF] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Physicists*, Grundlehren Math. Wiss., 57, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1954.

[C1] B. C. Carlson, *A hypergeometric mean value*, Proc. Amer. Math. Soc., 16 (1965), pp. 759–766.

[C2] ———, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[C3] ———, *A table of elliptic integrals of the second kind*, Math. Comp., 49 (1987), pp. 595–606 and S13–S17.

[CG] B. C. Carlson and J. L. Gustafson, *Asymptotic expansion of the first elliptic integral*, SIAM J. Math. Anal., 16 (1985), pp. 1072–1092.

[Ch] T. W. Chaundy, *Elementary Differential Equations*, Clarendon Press, Oxford, 1969.

[CK] S. Conde and S. L. Kalla, *A table of Gauss' hypergeometric function* ${}_2F_1(a,b;c;x)$, Universidad del Zulia, Venezuela, 1979.

[H] E. Hille, *Hypergeometric functions and conformal mapping*, J. Differential Equations, 34 (1979), pp. 147–152.

[KCH] S. L. Kalla, S. Conde, and J. L. Hubbell, *Some results on generalized elliptic-type integrals*, Appl. Anal., 22 (1986), pp. 273–287.

[KLH] S. L. Kalla, C. Leubner, and J. L. Hubbell, *Further results on generalized elliptic-type integrals*, Appl. Anal., 25 (1987), pp. 269–274.

[K] R. Kühnau, *Eine Methode, die Positivität einer Funktion zu prüfen*, submitted.

[L] D. F. Lawden, *Elliptic Functions and Applications*, Applied Mathematical Sciences, 80, Springer-Verlag, 1989.

[LV] O. Lehto and K. I. Virtanen, *Quasiconformal Mappings in the Plane*, Second ed., Die Grundlehren der math. Wiss., 126, Springer-Verlag, New York, Berlin, 1973.

[M] D. S. Mitrinović, *Analytic inequalities*, Grundlehren Math. Wiss., 165, Springer-Verlag, New York, 1970.

[PT1] W. H. Press and S. A. Teukolsky, *Elliptic Integrals*, Comput. Phys., January-February (1990), pp. 92–96.

[PT2] ———, *Hypergeometric functions by direct path integration*, Comput. Phys., May-June (1990), pp. 320–323.

[PBM] A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series, Vol. 3: More Special Functions*, trans. from the Russian by G.G. Gould, Gordon and Breach Science Publishers, New York, 1988.

[R] E. D. Rainville, *Special Functions*, Macmillan, New York, 1960.

[SS] M. Saigo and H. M. Srivastava, *The behavior of the zero-balanced hypergeometric series ${}_pF_{p-1}$ near the boundary of its convergence region*, Proc. Amer. Math. Soc., 110 (1990), pp. 71–76.

[Se] V. I. Semenov, *Certain applications of the quasiconformal and quasiisometric deformations* preprint.

[SO] J. Spanier and K. B. Oldham, *An Atlas of Functions*, Hemisphere (Harper & Row), Washington, New York, London, 1987.

[Vu] M. Vuorinen, *Conformal Geometry and Quasiregular Mappings*, Lecture Notes in Math., Vol. 1319, Springer-Verlag, Berlin, New York, 1988.

[WW] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Fourth ed., Cambridge University Press, Cambridge, 1958.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 525–551, March 1992

012

SOME q-BETA AND MELLIN–BARNES INTEGRALS WITH MANY PARAMETERS ASSOCIATED TO THE CLASSICAL GROUPS*

ROBERT A. GUSTAFSON[†]

Abstract. Multidimensional generalizations of a q-beta integral of Nasrallah–Rahman and Barnes' second lemma are evaluated. These are integral analogues of Jackson, Dougall, and the Pfaff–Saalschütz summation theorem for hypergeometric series. These integrals can also be written as group integrals over the special unitary group, the compact symplectic groups or conjugation invariant integrals over the corresponding Lie algebras.

Key words. multivariate beta integrals, multivariate Mellin–Barnes integrals, q-beta integrals, the classical groups

AMS(MOS) subject classifications. 33A15, 33A30, 33A65, 33A75

1. Introduction. One of the foundations for the classical theory of hypergeometric series is Euler's beta integral and the Mellin–Barnes contour integrals (see Baily [4]). On the one hand, integral representations of hypergeometric series as well as orthogonal polynomials are obtained. On the other hand, an important tool for discovering and proving many important special function identities is found. More recently, multidimensional generalizations of Euler's beta integral and the Mellin–Barnes integral have been discovered in the setting of root systems or simple Lie groups. Some examples of these are Selberg's beta integral [22], the q-Dyson–Zeilberger–Bressoud theorem [1], [23], the Macdonald–Morris conjectures [15], [19], and some generalizations and analogues of the Askey–Wilson integral [3] for various compact Lie groups and Lie algebras [8], [9]. The connection of these integrals to multidimensional special functions is only beginning to be understood (e.g., see [10], [11], [16]). Another aspect of this circle of problems is to obtain group-theoretic interpretations for these fascinating multidimensional integrals and functions. Some progress has been made in one dimension in interpreting some of the basic (or q-analogue) special functions as spherical functions on quantum groups (e.g., see [12], [17]). In higher dimensions, there is the theory of zonal spherical polynomials (see, e.g., [7]), spherical functions for p-adic groups [14], and Macdonald's polynomials [16], which contain all of the above as special cases. It often appears that the discovery of new kinds of special function identities has given the impetus for the discovery of new group-theoretic interpretations of special functions.

Let q be a real number, $0 < q < 1$. For any complex number c, define

$$[c]_\infty = [c;q]_\infty = \prod_{k=0}^{\infty}(1 - cq^k)$$

and also

$$[c]_n = [c]_\infty/[cq^n]_\infty$$

for any integer n.

* Received by the editors August 6, 1990; accepted for publication (in revised form) June 14, 1991. This research was partially supported by National Science Foundation grants INT-8713472 and DMS-9002342.

[†] Department of Mathematics, Texas A & M University, College Station, Texas 77843.

A one variable integral with many parameters generalizing Euler's beta integral was isolated by Rahman [21] (see [2]) and is a special case of an integral of Nasrallah–Rahman [20].

THEOREM 1.1. *Let $a_i \in C$, $1 \leq i \leq 5$, with $|a_i| < 1$, then*

$$\frac{1}{(2\pi i)} \int_T \frac{\left[z \prod_{i=1}^{5} a_i\right]_\infty \left[z^{-1} \prod_{i=1}^{5} a_i\right]_\infty [z^2]_\infty [z^{-2}]_\infty}{\prod_{i=1}^{5} [a_i z]_\infty [a_i z^{-1}]_\infty} \frac{dz}{z} = \frac{2 \prod_{k=1}^{5} \left[\prod_{\substack{i=1 \\ i \neq k}}^{5} a_i\right]_\infty}{[q]_\infty \prod_{1 \leq i < j \leq 5} [a_i a_j]_\infty}, \tag{1.2}$$

where the unit circle T taken in the positive direction.

If we set, for example, $a_1 = 0$, then identity (1.2) reduces to the important Askey–Wilson integral [3]. Identity (1.2) can be viewed as an integral analogue of Jackson's summation theorem for very well poised ${}_8\varphi_7$ hypergeometric series. The integrand in (1.2) is also the weight function for a very general family of biorthogonal rational functions [21], which include the Askey–Wilson polynomials [3] as a limiting case.

The purpose of this paper is to generalize integral (1.2) in the setting of the special unitary groups $SU(n)$ in §2 and the compact symplectic groups $Sp(n)$ in §4. In §5, we also evaluate some Mellin–Barnes integrals associated to the Lie algebras $su(n)$ and $sp(n)$ which generalize a Mellin–Barnes integral analogue of (1.2) due to Rahman [21]. Finally, in §3 we give a $u(n)$ (the Lie algebra of hermitian matrices) generalization of Barnes' second lemma [5], [4]. Barnes' second lemma is a Mellin–Barnes integral analogue of the important Pfaff–Saalschütz summation theorem for ${}_3F_2$ hypergeometric series. The integrals in this paper are given as multiple contour integrals, but can also be written as integrals over the corresponding Lie groups or Lie algebras. For details, see §10 of [9].

In future work [6], we plan to use the integrals evaluated here in order to derive a number of other special function identities, including Jackson-type, summation theorems for hypergeometric series associated to the classical groups and a generalization of the ${}_{10}\varphi_9$ transformation for $SU(n)$. We also hope to show that these integrals will be useful in constructing generating functions and reproducing kernels for multivariate orthogonal polynomials (see Nasrallah and Rahman [20]).

2. A generalization of Nasrallah–Rahman integral for $SU(n)$.

In this section we prove a generalization of the Nasrallah–Rahman integral (1.2) associated to the Lie groups $SU(n)$. The $SU(2)$ case is just the Nasrallah–Rahman integral (1.2). Briefly, the proof goes as follows. We first show that both sides of the integral identity (2.2) satisfy the same difference equation (2.8). This allows the proof of (2.2) to be reduced to a double induction, one parameter n measuring dimension and the other parameter k measuring the distance from an Askey–Wilson type integral for $SU(n)$ that has previously been evaluated [9]. An essential part of the induction proof is a technical multiple contour integral argument.

The main result is the following.

THEOREM 2.1. *For $n \geq 2$, let $a_i \in C$, $1 \leq i \leq n$, and $b_j \in C$, $1 \leq j \leq n+1$, with $|a_i|, |b_j| < 1$. Set $A = \prod_{i=1}^{n} a_i, B = \prod_{j=1}^{n+1} b_j$ and $C = AB$, then*

$$\frac{1}{(2\pi i)^{n-1}}\int_{T^{n-1}} \frac{\prod_{k=1}^{n} [Cz_k]_\infty \prod_{\substack{1\le i,j\le n\\ i\ne j}}[z_iz_j^{-1}]_\infty}{\prod_{k=1}^{n}\left\{\prod_{i=1}^{n} [a_iz_k^{-1}]_\infty \prod_{j=1}^{n+1}[b_jz_k]_\infty\right\}} \frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}} = \frac{n!\prod_{j=1}^{n+1}[b_j^{-1}C]_\infty \prod_{i=1}^{n}[a_iB]_\infty}{[q]_\infty^{n-1}[A]_\infty \prod_{j=1}^{n+1}[b_j^{-1}B]_\infty \prod_{i=1}^{n}\prod_{j=1}^{n+1}[a_ib_j]_\infty},\tag{2.2}$$

where $\prod_{k=1}^{n} z_k = 1$, the integral in each variable $z_1,\cdots,z_{n-1}$ is over the unit circle T taken in the positive direction, and T^{n-1} is the $n-1$ fold direct product of T.

Proof. Observe that if we define the $n=1$ case of the integral on the left-hand side of (2.2) to be the evaluation of this integrand at $z_1 = 1$, then identity (2.2) is valid for $n = 1$. The $n = 2$ case of identity (2.2) is a restatement of integral (1.2). The general proof of Theorem 2.1 given here will include the $n = 2$ case, but for $n = 2$ our proof is related to an earlier proof of Askey [2], though Askey's simpler proof does not seem to generalize for $n > 2$.

To begin our proof of identity (2.2) for $n \ge 2$, we shall require that for any integer $\ell, b_i \ne b_jq^\ell$ for $1 \le i,j \le n+1, i \ne j$, and $b_j \ne Cq^\ell$ for $1 \le j \le n+1$. This restriction will be removed at the end of the proof of (2.2). We also remark that

$$b_k^{-1}C = \prod_{i=1}^{n} a_i \prod_{\substack{j=1\\ j\ne k}}^{n+1} b_j \quad\text{and}\quad b_k^{-1}C = \prod_{\substack{j=1\\ j\ne k}}^{n+1} b_j$$

both make sense when $b_k = 0$.

To prove (2.2), we first show that both sides of (2.2) satisfy the same q-difference equation in the parameters b_j, $1 \le j \le n+1$. We will need the following version of the partial fraction expansion.

LEMMA 2.3. (*Cf.* [18, *Lemma* 7.1].) *Let $\{x_1,\cdots,x_m\}$,, $\{y_1,\cdots,y_{m+1}\}$ and t be indeterminants with the y_i distinct. Then*

$$\sum_{\ell=1}^{m+1}\left\{\prod_{\substack{j=1\\ j\ne\ell}}^{m+1}\left(\frac{t-y_j}{y_\ell-y_j}\right)\prod_{k=1}^{m}(1-y_\ell x_k)\right\} = \prod_{k=1}^{m}(1-tx_k).\tag{2.4}$$

Proof. Identity (2.4) is equivalent to the $r < s$ case of Lemma 7.1 of [18] (see also [13]), which is a statement of the classical partial fractions expansion. To prove identity (2.4), divide both sides by $\prod_{j=1}^{m+1}(t - y_j)$ and obtain the following identity:

$$\sum_{\ell=1}^{m+1}\left\{(t-y_\ell)^{-1}\prod_{\substack{j=1\\ j\ne\ell}}^{m+1}(y_\ell-y_j)^{-1}\prod_{k=1}^{m}(1-y_\ell x_k)\right\} = \prod_{k=1}^{m}(1-tx_k)\prod_{j=1}^{m+1}(t-y_j)^{-1}.\tag{2.5}$$

Identity (2.5) is proved by observing that the left-hand side of (2.5) is simply the partial fraction expansion of the right-hand side of (2.5).

We now show that both sides of (2.2) satisfy the same q-difference equation.

DEFINITION 2.6. If f is a function involving the parameters b_ℓ and C, then define

$$R_\ell f(b_\ell, C) = f(b_\ell q, Cq).$$

LEMMA 2.7. *With notation as in Theorem* 2.1, *let f be either the left-hand side or right-hand side of* (2.2). *Assume that the parameters b_i are all distinct. Then*

$$\sum_{\ell=1}^{n+1} \left\{ \prod_{\substack{j=1\\ j\neq\ell}}^{n+1} \left(\frac{C-b_j}{b_\ell - b_j} \right) \cdot (R_\ell f) \right\} = f. \tag{2.8}$$

Proof. Let I and Q denote, respectively, the left-hand side and right-hand side of equation (2.2). We have

$$R_\ell I = \frac{1}{(2\pi i)^{n-1}} \int_{T^{n-1}} \frac{\prod_{k=1}^{n} [Cz_k]_\infty \prod_{\substack{1\leq i,j\leq n\\ i\neq j}} [z_i z_j^{-1}]_\infty}{\prod_{k=1}^{n} \left\{ \prod_{i=1}^{n} [a_i z_k^{-1}]_\infty \prod_{j=1}^{n+1} [b_j z_k]_\infty \right\}} \tag{2.9}$$
$$\cdot \prod_{k=1}^{n} \left(\frac{1-b_\ell z_k}{1-Cz_k} \right) \frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}}.$$

Substitute the expression on the right-hand side of (2.9) into the left-hand side of (2.8) and apply Lemma 2.3 with $m = n, t = C, x_i = z_i$ and $y_i = b_i$. This proves that I satisfies the q-difference equation (2.8).

Similarly, we have

$$R_\ell Q = \left\{ \prod_{\substack{j=1\\ j\neq\ell}}^{n+1} \left(\frac{1-b_j^{-1}B}{1-b_j^{-1}C} \right) \prod_{i=1}^{n} \left(\frac{1-a_i b_\ell}{1-a_i B} \right) \right\} Q. \tag{2.10}$$

To verify that Q satisfies the q-difference equation (2.8), we need to show that

$$\sum_{\ell=1}^{n+1} \left\{ \prod_{\substack{j=1\\ j\neq\ell}}^{n+1} \left[\frac{(C-b_j)(1-b_j^{-1}B)}{(b_\ell - b_j)(1-b_j^{-1}C)} \right] \prod_{i=1}^{n} \left(\frac{1-a_i b_\ell}{1-a_i B} \right) \right\} = 1. \tag{2.11}$$

Simplifying, we are reduced to showing that

$$\sum_{\ell=1}^{n+1} \left\{ \prod_{\substack{j=1\\ j\neq\ell}}^{n+1} \left(\frac{B-b_j}{b_\ell - b_j} \right) \prod_{i=1}^{n} (1-a_i b_\ell) \right\} = \prod_{i=1}^{n} (1-a_i B). \tag{2.12}$$

Identity (2.12) now follows from Lemma 2.3 upon setting $m = n, t = B, x_i = a_i$, and $y_i = b_i$ in identity (2.4). This completes the proof of Lemma 2.7.

We now prove Theorem 2.1 by induction on n. As mentioned above, the $n = 1$ case of identity (2.2) is trivially true. We will assume from now on that $n \geq 2$ and that the $n - 1$ case of identity (2.2) is valid.

We first prove identity (2.2) in the special case that $a_n = q^{k+1}C^{-1}$ and $b_{n+1} = q^{-k}C$, where k is a positive integer. This will proceed by induction on k, with the cases $k = 1$ and $k > 1$ handled differently.

Assume that $q^{k+1} < |C| < q^k$ and that $|q^kC^{-1}| < r^{-1}$ where r is a real number chosen as close to 1 as necessary. We will also assume that C^{n-1} is not a real number.

Let $R = r^{n-1}$. For $k > 1$, we let

$$a_i < R \quad \text{for } i = 1, \cdots, n-1 \tag{2.13a}$$

and

$$b_j < R^{n-1} \quad \text{for } j = 1, \cdots, n \tag{2.13b}$$

For $k = 1$, let $1 \geq t > 0$ be a real number so that

$$ta_i < R \quad \text{for } i = 1, \cdots, n-1 \tag{2.14a}$$

and

$$tb_j < R^{n-1} \quad \text{for } j = 1, \cdots, n. \tag{2.14b}$$

For $k > 1$, we will always let $t = 1$.

Let N be the contour which is the union of the circle of radius R centered at the origin, traversed in the positive direction, and the circle of radius ε centered at the point q^kC^{-1} traversed in the positive direction, where ε is a sufficiently small positive real number. We will now use the following notation:

$$H(m,t) = \frac{\prod_{\ell=1}^{n} [Cz_\ell]_\infty \prod_{\substack{1\leq i,j\leq n \\ i\neq j}} [z_iz_j^{-1}]_\infty}{\prod_{\ell=1}^{n} \left\{ [q^{m+1}C^{-1}z_\ell^{-1}]_\infty [q^{-m}Cz_\ell]_\infty \prod_{i=1}^{n-1} [a_itz_\ell^{-1}]_\infty \prod_{j=1}^{n} [b_jtz_\ell]_\infty \right\}} \tag{2.15a}$$

and

$$H(m) = H(m,1), \tag{2.15b}$$

where m is a nonnegative integer.

Rewriting the factors $\prod_{\ell=1}^n (1 - q^{-k}Cz_\ell)$ as $(-q^{-k}C)^n \prod_{\ell=1}^n (1 - q^kC^{-1}z_\ell^{-1})$, we find

$$\int\limits_{T^{n-1}} H(k,t)\frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}} = \frac{(-1)^n}{(q^{-k}C)^n} \int\limits_{T^{n-1}} H(k-1,t)\frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}}. \tag{2.16}$$

We also have

$$\begin{aligned} &\int\limits_{T^{n-1}} H(k-1,t)\frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}} \\ &= \int\limits_{T^{n-2}} \int\limits_{T-N} H(k-1,t)\frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}} \end{aligned} \tag{2.17}$$

$$+\int_{T^{n-3}}\int_{T-N}\int_{N} H(k-1,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}$$

$$+\int_{T^{n-4}}\int_{T-N}\int_{N^2} H(k-1,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}$$

$$+\cdots+\int_{N^{n-1}} H(k-1,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}.$$

Using Fubini's theorem and the symmetry of the integrands with respect to permutation of the variables $z_1,\cdots,z_{n-1}$, we find

$$\begin{aligned}\int_{T^{n-1}} H(k-1,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}} &= \sum_{j=0}^{n-2}\int_{T^{n-2-j}}\int_{N^j}\int_{T-N} H(k-1,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}\\ &\quad+\int_{N^{n-1}} H(k-1,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}.\end{aligned}\tag{2.18}$$

With respect to the variable z_1, consider the poles of the integrand $H(k-1,t)\prod_{i=1}^{n-1} z_i^{-1}$ inside the region bounded by the contour $T-N$. The first pole is at $z_1 = C^{-1}q^k$ and the other pole is at $z_1 = Cq^{-k}z_2^{-1}\cdots z_n^{-1}$ (i.e., $z_n = C^{-1}q^k$). Notice that for the second pole to occur either $z_i \in T$ for all $i = 2,\cdots,n-1$ or at least one of the z_i, $i = 2,\cdots,n-1$ must lie on the circle of radius ε centered at q^kC^{-1}.

Let

$$J(k-1,t)=\frac{\prod_{j=2}^{n}\{(1-Cq^{-k}z_j)[Cz_j]_\infty\}}{[q]_{k-1}[q]_\infty\prod_{i=1}^{n-1}[a_itCq^{-k}]_\infty\prod_{j=1}^{n}[b_jtC^{-1}q^k]_\infty}\cdot\frac{\prod_{\substack{2\le i,j\le n\\ i\ne j}}[z_iz_j^{-1}]_\infty}{\prod_{\ell=2}^{n}\left\{\prod_{i=1}^{n-1}[a_itz_\ell^{-1}]_\infty\prod_{j=1}^{n}[b_jtz_\ell]_\infty\right\}},$$

where $\prod_{i=2}^{n} z_i = Cq^{-k}$. From identity (2.18) and the following remarks, it follows that

$$\begin{aligned}&\sum_{j=0}^{n-2}\frac{1}{(2\pi i)^{n-1}}\int_{T^{n-2-j}}\int_{N^j}\int_{T-N} H(k-1,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}\\ &= -\frac{2}{(2\pi i)^{n-2}}\int_{T^{n-2}} J(k-1,t)\frac{dz_2}{z_2}\cdots\frac{dz_{n-1}}{z_{n-1}}\\ &\quad-\sum_{j=1}^{n-2}\frac{1}{(2\pi i)^{n-2}}\int_{T^{n-2-j}}\int_{N^j} J(k-1,t)\frac{dz_2}{z_2}\cdots\frac{dz_{n-1}}{z_{n-1}}+\varepsilon M(t),\end{aligned}\tag{2.19}$$

where $M(t)$ is a number which is bounded independently of ε.

In the integrals on the right-hand side of (2.19) we can deform the contours N to the contour T by attaching the circle of radius R in N to the circle of radius ε by two parallel line segments in opposite directions and then separating the line segments. We can use Fubini's theorem repeatedly, crossing no poles of the integrand during the deformations. It follows that

$$(2.20) \qquad \int\limits_{T^{n-2-j}} \int\limits_{N^j} J(k-1,t)\frac{dz_2}{z_2}\cdots\frac{dz_{n-1}}{z_{n-1}} = \int\limits_{T^{n-2}} J(k-1,t)\frac{dz_2}{z_2}\cdots\frac{dz_{n-1}}{z_{n-1}}$$

for $0 \leq j \leq n-2$.

Now let

$$d = (q^k C^{-1})^{\frac{1}{n-1}}$$

for some choice of $(n-1)$th root of $q^k C^{-1}$. We make the change of variables

$$z_i' = z_i d \quad \text{for } i = 2, \cdots, n-1,$$

$$a_i' = a_i d \quad \text{for } i = 1, \cdots, n,$$

$$b_j' = b_j d^{-1} \quad \text{for } j = 1, \cdots, n,$$

and

$$C' = Cd^{-1}.$$

Note that $C' = q\prod_{i=1}^{n-1} a_i' \prod_{j=1}^{n} b_j'$.

In terms of the new variables we have

$$J(k-1,t) = \frac{\prod\limits_{j=2}^{n}\{(1-d^{-n}z_j')[C'z_j']_\infty\}}{[q]_{k-1}[q]_\infty \prod\limits_{i=1}^{n-1}[a_i'tC'q^{-k}]_\infty \prod\limits_{j=1}^{n}[b_j't(C')^{-1}q^k]_\infty}$$

$$\cdot \frac{\prod\limits_{\substack{2\leq i,j\leq n\\ i\neq j}} [z_i'(z_j')^{-1}]_\infty}{\prod\limits_{\ell=2}^{n}\left\{\prod\limits_{i=1}^{n-1}[a_i't(z_\ell')^{-1}]_\infty \prod\limits_{j=1}^{n}[b_j'tz_\ell']_\infty\right\}},$$

where $\prod_{i=2}^{n} z_i' = 1$. After making this change of variables and moving the contours of integration, we find

$$(2.21) \qquad \int\limits_{T^{n-2}} J(k-1,t)\frac{dz_2}{z_2}\cdots\frac{dz_{n-1}}{z_{n-1}} = \int\limits_{T^{n-2}} J(k-1,t)\frac{dz_2'}{z_2'}\cdots\frac{dz_{n-1}'}{z_{n-1}'}.$$

Consider the case $k=1$. Then

$$H(0,t) = \frac{\prod\limits_{1\leq i,j\leq n}[z_iz_j^{-1}]_\infty}{\prod\limits_{\ell=1}^{n}\left\{[qC^{-1}z_\ell^{-1}]_\infty \prod\limits_{i=1}^{n-1}[a_itz_\ell^{-1}]_\infty \prod\limits_{j=1}^{n}[b_jtz_\ell]_\infty\right\}}$$

and the integral $\int_{N^{n-1}} H(0,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}$ can be evaluated by means of the $SU(n)$ generalization of the Askey–Wilson integral [9, Thm. 6.1]. By a contour deformation argument similar to the above and by shifting the parameters a_it, $1 \le i \le n-1, b_jt$, $1 \le j \le n$, and $q^{-1}C$, it can be verified that the $SU(n)$ Askey–Wilson integral is also valid for the contour N and the parameters a_it, b_jt, and qC^{-1}. We find

$$\int\limits_{N^{n-1}} H(0,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}} = \frac{(2\pi i)^{n-1}n![t^{2n-1}]_\infty}{[q]_\infty^{n-1}\left[qC^{-1}t^{n-1}\prod_{i=1}^{n-1}a_i\right]_\infty\left[t^n\prod_{j=1}^{n}b_j\right]_\infty} \cdot\prod_{j=1}^{n}\left\{[tqC^{-1}b_j]_\infty\prod_{i=1}^{n-1}[t^2a_ib_j]_\infty\right\}^{-1}. \tag{2.22}$$

Considering the limit as ε tends to zero, then from (2.18)–(2.22), it follows that

$$\begin{aligned}\frac{1}{(2\pi i)^{n-1}}\int\limits_{T^{n-1}} H(0,t)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}} &= \frac{n![t^{2n-1}]_\infty}{[q]_\infty^{n-1}\left[qC^{-1}t^{n-1}\prod_{i=1}^{n-1}a_i\right]_\infty\left[t^n\prod_{j=1}^{n}b_j\right]_\infty}\\ &\quad\cdot\prod_{j=1}^{n}\left\{[tqC^{-1}b_j]_\infty\prod_{i=1}^{n-1}[t^2a_ib_j]_\infty\right\}^{-1}\\ &\quad-\frac{n}{(2\pi i)^{n-2}}\int\limits_{T^{n-2}} J(0,t)\frac{dz_2'}{z_2'}\cdots\frac{dz_{n-1}'}{z_{n-1}'}.\end{aligned} \tag{2.23}$$

Both sides of (2.23) are analytic functions of the parameters a_i, b_j, and t. If we choose the a_i and b_j appropriately and deform the contour T on the right-hand side of (2.23), then we can take the limit as $t \to 1$ on both sides of (2.23). We obtain

$$\frac{1}{(2\pi i)^{n-1}}\int\limits_{T^{n-1}} H(0)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}} = -\frac{n}{(2\pi i)^{n-2}}\int\limits_{L^{n-2}} J(0,1)\frac{dz_2'}{z_2'}\cdots\frac{dz_{n-1}'}{z_{n-1}'}, \tag{2.24}$$

where L is an appropriate contour.

We have

$$J(0,1) = \frac{\prod_{j=2}^{n}[d^{-n}z_j']_\infty\prod_{\substack{2\le i,j\le n\\ i\ne j}}[z_i'(z_j')^{-1}]_\infty}{[q]_\infty\prod_{i=1}^{n-1}[a_i'C'q^{-1}]_\infty\prod_{j=1}^{n}[b_j'(C')^{-1}q]_\infty\prod_{\ell=2}^{n}\left\{\prod_{i=1}^{n-1}[a_i'(z_\ell')^{-1}]_\infty\prod_{j=1}^{n}[b_j'z_\ell']_\infty\right\}}.$$

Since $\prod_{i=1}^{n-1}a_i'\prod_{j=1}^{n}b_j' = d^{-n}$, then the integral on the right-hand side of equation (2.24) can be evaluated by the dimension $n-1$ case of identity (2.2), which we assume is valid by the induction hypothesis on the dimension n. From (2.16) and (2.24), we then obtain

$$\frac{1}{(2\pi i)^{n-1}} \int_{T^{n-1}} H(1)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}$$

$$(2.25) \qquad = \frac{(-1)^{n-1}n! \prod_{j=1}^{n} [(b'_j)^{-1}d^{-n}]_\infty \prod_{i=1}^{n-1} \left[a'_i \prod_{j=1}^{n} b'_j\right]_\infty}{(q^{-1}C)^n[q]_\infty^{n-1} \left[\prod_{i=1}^{n-1} a'_i\right]_\infty \prod_{j=1}^{n} \left[b_j'^{-1} \prod_{i=1}^{n} b'_i\right]_\infty}$$

$$\cdot \left\{ \prod_{i=1}^{n-1} \prod_{j=1}^{n} [a'_i b'_j]_\infty \prod_{i=1}^{n-1} [a'_i C' q^{-1}]_\infty \prod_{j=1}^{n} [b'_j (C')^{-1} q]_\infty \right\}^{-1}.$$

Now rewrite identity (2.25) in terms of the parameters a_i, b_j, and C. After some computation and using that $a_n = q^2C^{-1}$ and $b_{n+1} = q^{-1}C$, then identity (2.25) becomes the $k = 1$ dimension n case of identity (2.2). This completes the proof of the $k = 1$, dimension n case of identity (2.2).

We now consider the $k > 1$, dimension n case of identity (2.2). The $k - 1$, dimension n case is assumed to be valid. By a contour deformation argument similar to the above and by shifting the parameters a_i and b_j, it can be verified that the $k-1$ case of identity (2.2) is also valid for the contour N and the parameters a_i and b_j. We find

(2.26)

$$\int_{N^{n-1}} H(k-1)\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}} = \frac{(2\pi i)^{n-1}n! \prod_{j=1}^{n} [b_j^{-1}C]_\infty}{[q]_\infty^n \left[q^k C^{-1} \prod_{i=1}^{n-1} a_i\right]_\infty}$$

$$\cdot \frac{[q^{k-1}]_\infty \left[q \prod_{j=1}^{n} b_j\right]_\infty \prod_{i=1}^{n-1} \left[a_i q^{-k+1} C \prod_{j=1}^{n} b_j\right]_\infty}{\left[\prod_{j=1}^{n} b_j\right]_\infty \prod_{j=1}^{n} \left\{ \left[b_j^{-1} q^{-k+1} C \prod_{i=1}^{n} b_i\right]_\infty [q^k C^{-1} b_j]_\infty \prod_{i=1}^{n-1} [q_i b_j]_\infty \right\}}$$

$$\cdot \prod_{i=1}^{n-1} [q^{-k+1}Ca_i]_\infty^{-1} = \frac{(2\pi i)^{n-1}n! \prod_{j=1}^{n} [(b'_j)^{-1}C']_\infty \prod_{i=1}^{n-1} \left[q a'_i \prod_{j=1}^{n} b'_j\right]_\infty}{[q]_\infty^{n-1}[q]_{k-2} \left[\prod_{i=1}^{n-1} a'_i\right]_\infty \prod_{i=1}^{n-1} [q^{-k+1}C'a'_i]_\infty}$$

$$\cdot \prod_{j=1}^{n} \left\{ \left[q(b'_j)^{-1} \prod_{i=1}^{n} b'_i\right]_\infty [q^k (C')^{-1} b'_j]_\infty \prod_{i=1}^{n-1} [a'_i b'_j]_\infty \right\}^{-1}$$

$$\cdot \left(1 - d^n \prod_{j=1}^{n} b'_j\right)^{-1}.$$

In order to compute the integral of the function $J(k-1,1)$, we will define a related function $I(k-1)$ and a difference equation involving $I(k-1)$ and $J(k-1,1)$. Let

$$I(k-1) = \frac{\prod_{j=2}^{n}[C'z'_j]_\infty \prod_{\substack{2\le i,j\le n\\ i\ne j}} [z'_i(z'_j)^{-1}]_\infty}{\prod_{\ell=2}^{n}\left\{\prod_{i=1}^{n-1}[a'_i(z'_\ell)^{-1}]_\infty \prod_{j=1}^{n}[b'_j z'_\ell]_\infty\right\}}.$$

If $f(b'_\ell, C')$ is a function involving the parameters b'_ℓ and C', then we define the shift operator L_ℓ by

$$L_\ell f(b'_\ell, C') = f(b'_\ell q, C'),$$

i.e., which shifts b'_ℓ, but fixes C'. It then follows from Lemma 2.3 that

$$\begin{aligned} J(k-1,1) = {} & \left\{[q]_{k-1}[q]_\infty \prod_{i=1}^{n-1}[a'_i C' q^{-k}]_\infty \prod_{j=1}^{n}[b'_j(C')^{-1}q^k]_\infty\right\}^{-1} \\ & \cdot \sum_{\ell=1}^{n}\left\{\prod_{\substack{j=1\\ j\ne\ell}}^{n}\left(\frac{d^{-n}-b'_j}{b'_\ell-b'_j}\right) L_\ell I(k-1)\right\}, \end{aligned} \tag{2.27}$$

where we require that the parameters b'_j, $1 \le j \le n$, are distinct.

Identity (2.27) shows that the integral of the function $J(k-1,1)$ can be computed in terms of the integrals of the functions $L_\ell I(k-1)$. Since $C' = q\prod_{i=1}^{n-1} a'_i \prod_{j=1}^{n} b'_j$, then the dimension $n-1$ case of identity (2.2) (which we assume is valid by the induction hypothesis) can be applied to evaluate the integrals of the $L_\ell I(k-1)$ functions. With a little simplification we have

$$\begin{aligned} & \frac{1}{(2\pi i)^{n-2}} \int_{T^{n-2}} J(k-1,1)\frac{dz'_2}{z'_2}\cdots\frac{dz'_n}{z'_n} \\ & = \frac{(n-1)!\prod_{j=1}^{n}[(b'_j)^{-1}C']_\infty \prod_{i=1}^{n-1}\left[qa'_i\prod_{j=1}^{n}b'_j\right]_\infty}{[q]_\infty^{n-1}[q]_{k-1}\prod_{i=1}^{n-1}[q'_i C' q^{-k}]_\infty \prod_{j=1}^{n}[b'_j(C')^{-1}q^k]_\infty} \\ & \cdot\left\{\left[\prod_{i=1}^{n-1}a'_i\right]_\infty \prod_{j=1}^{n}\left[q(b'_j)^{-1}\prod_{i=1}^{n}b'_i\right]_\infty \prod_{i=1}^{n-1}\prod_{j=1}^{n}[a'_i b'_j]_\infty\right\}^{-1} \\ & \cdot\sum_{\ell=1}^{n}\left\{\prod_{\substack{j=1\\ j\ne\ell}}^{n}\left(\frac{d^{-n}-b'_j}{b'_\ell-b'_j}\right)\cdot\frac{(1-q^{-1}(b'_\ell)^{-1}C')\prod_{i=1}^{n-1}(1-a'_i b'_\ell)}{\left(1-\prod_{\substack{i=1\\ i\ne\ell}}^{n}b'_i\right)}\right\}. \end{aligned} \tag{2.28}$$

The expression inside the sum on the right-hand side of (2.28) can be rewritten as

$$\tag{2.29} \frac{(-q^{-1}C')}{\left(d^{-n}-\prod_{i=1}^{n} b'_i\right)} \sum_{\ell=1}^{n} \left\{ \prod_{\substack{j=1\\ j\neq\ell}}^{n} \left(\frac{d^{-n}-b'_j}{b'_\ell - b'_j}\right) \cdot \left(\frac{d^{-n}-\prod_{i=1}^{n} b'_i}{b'_\ell - \prod_{i=1}^{n} b'_i}\right) (1-q(C')^{-1}b'_\ell) \prod_{i=1}^{n-1}(1-a'_i b'_\ell) \right\}.$$

An application of Lemma 2.3 shows that (2.29) equals

$$\tag{2.30} \frac{(q^{-1}C')}{\left(d^{-n}-\prod_{i=1}^{n} b'_i\right)} \left\{ \prod_{j=1}^{n} \left(\frac{d^{-n}-b'_j}{\prod_{i=1}^{n}(b'_i)-b_j}\right) \cdot \left(1-q(C')^{-1}\prod_{i=1}^{n} b'_i\right) \prod_{i=1}^{n-1}\left(1-a'_i \prod_{j=1}^{n} b'_j\right) - (1-q(C')^{-1}d^{-n}) \prod_{i=1}^{n}(1-a'_i d^{-n}) \right\}.$$

Taking the limit as $\varepsilon \to 0$, then from (2.16)–(2.21) and (2.26)–(2.30), it follows that

$$\tag{2.31} \frac{1}{(2\pi i)^{n-1}} \int_{T^{n-1}} H(k) \frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}} = \frac{n! \prod_{j=1}^{n} [(b'_j)^{-1}C']_\infty \prod_{i=1}^{n-1} \left[q a'_i \prod_{j=1}^{n} b'_j\right]_\infty}{[q]_\infty^{n-1}[q]_{k-1} \left[\prod_{i=1}^{n-1} a'_i\right]_\infty \prod_{i=1}^{n-1} [a'_i C' q^{-k}]_\infty} \cdot \prod_{j=1}^{n} \left\{ \left[q(b'_j)^{-1}\prod_{i=1}^{n} b'_i\right]_\infty [q^k b'_j (C')^{-1}]_\infty \prod_{i=1}^{n-1}[a'_i b'_j]_\infty \right\}^{-1} \cdot \frac{(-1)^n}{(q^{-k}C'd)^n} \left\{ \frac{(1-q^{k-1}) \prod_{i=1}^{n-1}(1-q^{-k}C'a'_i)}{\left(1-d^n \prod_{j=1}^{n} b'_j\right)} \right.$$

$$+\frac{(q^{-1}C')}{\left(d^{-n}-\prod_{i=1}^{n} b'_i\right)}(1-q(C')^{-1}d^{-n})\prod_{i=1}^{n-1}(1-a'_i d^{-n})$$

$$-\frac{(q^{-1}C')}{\left(d^{-1}-\prod_{i=1}^{n} b'_i\right)}\prod_{j=1}^{n}\left(\frac{d^{-n}-b'_j}{\prod_{i=1}^{n}(b'_i)-b'_j}\right)$$

$$\left.\cdot\left(1-q(C')^{-1}\prod_{i=1}^{n} b'_i\right)\cdot\prod_{i=1}^{n-1}\left(1-a'_i\prod_{j=1}^{n} b'_j\right)\right\}.$$

Since $d^{-n} = q^{-k}C'$, it follows that the first two terms cancel in the bracketed sum on the right-hand side of (2.31). After some further simplification, then the right-hand side of (2.31) reduces to

$$\begin{aligned}&\frac{n!\prod_{j=1}^{n}[b_j^{-1}C]_\infty\prod_{i=1}^{n-1}\left[a_i q^{-k}C\prod_{j=1}^{n} b_j\right]_\infty}{[q]_\infty^{n-1}[q]_{k-1}\left[q^{k+1}C^{-1}\prod_{i=1}^{n-1} a_i\right]_\infty\prod_{i=1}^{n-1}[a_i q^{-k}C]_\infty}\\ (2.32)\qquad&\cdot\prod_{j=1}^{n}\left\{\left[q^{-k}Cb_j^{-1}\prod_{i=1}^{n} b_i\right]_\infty[q^{k+1}C^{-1}b_j]_\infty\prod_{i=1}^{n-1}[a_i b_j]_\infty\right\}^{-1}\\ &\cdot\left(1-\prod_{i=1}^{n} b_j\right)^{-1}.\end{aligned}$$

Recalling that $a_n = q^{k+1}C^{-1}$ and $b_{n+1} = q^{-k}C$, this completes the proof of the $k > 1$, dimension n case of identity (2.2).

To complete the proof of the full dimension n case of identity (2.2), we begin by choosing an arbitrary complex number a_n such that $|q| < |a_n| < 1$ and setting $b_{n+1} = qa_n^{-1}$. Let $a_1, \cdots, a_{n-1}, b_1, \cdots, b_{n-1} \in C$ be fixed such that $|a_i|, |b_j| < 1$. Then set $b_n = q^k a_n^{-1}\prod_{i=1}^{n-1}(a_i b_i)^{-1}$, with $k > 0$ a sufficiently large integer so that $|b_n| < 1$. From (2.31) and (2.32) it then follows that identity (2.2) is valid for a set of values of b_n which includes the limit point $b_n = 0$. (The case $b_n = 0$ reduces to the $SU(n)$ generalization of the Askey–Wilson integral [9, Thm. 6.1].) By analytic continuation in the parameter b_n, we conclude that identity (2.2) is valid whenever $a_n b_{n+1} = q$.

This last condition can also be eliminated. With notation as in Lemma 2.7, we rewrite the q-difference equation (2.8) as

$$(2.33)\qquad R_{n+1}f = \prod_{j=1}^{n}\left(\frac{b_{n+1}-b_j}{C-b_j}\right)\left[f-\sum_{\ell=1}^{n}\left\{\prod_{\substack{j=1\\ j\neq\ell}}^{n+1}\left(\frac{C-b_j}{b_\ell-b_j}\right)(R_\ell f)\right\}\right],$$

where f can be either the left or right-hand side of identity (2.2). Now suppose that identity (2.2) is valid whenever $b_{n+1} = q^\ell a_n^{-1}$ for some positive integer ℓ. If we choose the parameters on the left-hand side of (2.33) to satisfy $b_{n+1} = q^{\ell+1}a_n^{-1}$, then those on the right-hand side will satisfy $b_{n+1} = q^{\ell}a_n^{-1}$. It follows that identity (2.2) will also be valid for $b_{n+1} = q^{\ell+1}a_n^{-1}$. The set of points $b_{n+1} = q^{\ell}a_n^{-1}$ for $\ell = 1, 2, 3, \cdots$ has the limit point $b_{n+1} = 0$. Again, at $b_{n+1} = 0$, identity (2.2) reduces to the $SU(n)$ Askey–Wilson integral. By analytic continuation of the parameter b_{n+1}, we now conclude that identity (2.2) is valid for arbitrary b_{n+1}. Finally, by analytic continuation we drop the other restrictions on the values of the parameters b_j and C. This completes the proof of the dimension n case of identity (2.2). By induction, this proves Theorem 2.1.

3. A generalization of Barnes' second lemma for $u(n)$. In this section we prove a generalization of Barnes' second lemma associated to the Lie algebra $u(n)$ (of the Lie group $U(n)$). The $u(1)$ case is Barnes' second lemma [5]. The main result, Theorem 3.1, is proved analogously to Theorem 2.1. We again use a double induction on the dimension n and another parameter k.

The main result is the following.

THEOREM 3.1. *For $n \geq 1$, let $\alpha_i \in C$, $1 \leq i \leq n+1$, and $\beta_j \in C$, $1 \leq j \leq n+2$. Set $\gamma = \sum_{i=1}^{n+1} \alpha_i + \sum_{j=1}^{n+2} \beta_j$, then*

$$\begin{aligned}(3.2)\qquad &\frac{1}{(2\pi i)^n} \int_{-i\infty}^{i\infty} \cdots \int_{-i\infty}^{i\infty} \frac{\prod_{j=1}^{n} \left\{ \prod_{i=1}^{n+1} \Gamma(\alpha_i - z_j) \prod_{i=1}^{n+2} \Gamma(\beta_i + z_j) \right\}}{\prod_{j=1}^{n} \Gamma(\gamma + z_j) \prod_{\substack{1 \leq i,j \leq n \\ i \neq j}} \Gamma(z_i - z_j)} dz_1 \cdots, dz_n \\ &= \frac{n! \prod_{i=1}^{n+1} \prod_{j=1}^{n+2} \Gamma(\alpha_i + \beta_j)}{\prod_{i=1}^{n+2} \Gamma(\gamma - \beta_j)},\end{aligned}$$

where the contours of integration are deformed so as to separate the sequences of poles going to the right $\{\alpha_i + k \mid 1 \leq i \leq n+1, k = 0, 1, 2, \cdots\}$ from the sequences of poles going to the left $\{-\beta_i - k \mid 1 \leq i \leq n+2, k = 0, 1, 2, \cdots\}$.

Proof. If we define the $n = 0$ case of the integral on the left-hand side of (3.2) to be 1, then identity (3.2) is valid for $n = 0$. The $n = 1$ case of identity (3.2) is Barnes' second lemma [5].

We begin our proof of identity (3.2) for $n \geq 1$ by requiring that for any integer $\ell, \beta_i \neq \beta_j + \ell$ for $1 \leq i, j \leq n+2, i \neq j$, and $\beta_j \neq \gamma + \ell$ for $1 \leq j \leq n+2$. This restriction will be removed at the end of the proof of (3.2).

We will first show that both sides of (3.2) satisfy the same difference equation in the parameters β_j and γ. We will need the following version of the partial fractions expansion.

LEMMA 3.3. *Let $s \leq m$ and let $\{x_1, \cdots, x_s\}, \{y_1, \cdots, y_{m+1}\}$ and t be indeterminants with the y_i distinct. Then*

$$(3.4)\qquad \sum_{\ell=1}^{m+1} \left\{ \prod_{\substack{j=1 \\ j \neq \ell}}^{m+1} \left(\frac{t - y_j}{y_\ell - y_j} \right) \prod_{k=1}^{s} (y_\ell - x_k) \right\} = \prod_{k=1}^{s} (t - x_k).$$

Proof. The proof is very similar to that of Lemma 2.3. If we divide both sides of identity (3.4) by $\prod_{j=1}^{m+1}(t - y_j)$, then we obtain an obvious partial fractions expansion.

We now show that both sides of (3.2) satisfy the same difference equation.

DEFINITION 3.5. If f is a function involving the parameters β_ℓ and γ, then define

$$\rho_\ell f(\beta_\ell,\gamma) = f(\beta_\ell+1,\gamma+1).$$

LEMMA 3.6. *With notation as in Theorem* 3.1, *let f be either the left-hand side or the right-hand side of* (3.2). *Then*

$$\sum_{\ell=1}^{n+2}\left\{\prod_{\substack{j=1\\ j\neq\ell}}^{n+2}\left(\frac{\gamma-\beta_j}{\beta_\ell-\beta_j}\right)\cdot(\rho_\ell f)\right\} = f. \tag{3.7}$$

Proof. The proof of identity (3.7) is quite similar to that of identity (2.8). A simple application of (3.4) shows that the left-hand side of (3.2) satisfies identity (3.7). Another application of a special case of (3.4) (i.e., equating the t^m term on both sides of (3.4) when $s = m$) shows that the right-hand side of (3.2) also satisfies identity (3.7).

We now prove Theorem 3.1 by induction on n. As mentioned above, by defining the left-hand side of (3.2) to be 1, the $n = 0$ case of identity (3.2) is trivially true. We will assume from now on that $n \geq 1$ and that the $n-1$ case of identity (3.2) is valid.

Similarly to §2, we first prove identity (3.2) in the special case that $\alpha_{n+1} = k+1-\gamma$ and $\beta_{n+2} = \gamma - k$, where k is a positive integer. We will proceed by induction on k, with the cases $k = 1$ and $k > 1$ handled slightly differently.

Assume that $k+1 > \mathrm{Re}(\gamma) > k$ and that $\mathrm{Re}(\gamma - k) < r$, where r is a positive real number chosen as small as necessary. For $k \geq 1$ we let $\mathrm{Re}(\alpha_i) > r$ and $\mathrm{Re}(\beta_j) > r$ for $1 \leq i \leq n$ and $1 \leq j \leq n+1$. Let C be the circle of radius ε centered at the point $\kappa - \gamma$ and traversed in the negative direction, where ε is a sufficiently small positive real number. Let T be the imaginary axis tranversed from $-i\infty$ to $i\infty$, and let N be the union of C and T. We will use the following notation

$$H(m) = \frac{\displaystyle\prod_{\ell=1}^{n}\left\{\Gamma(m+1-\gamma-z_\ell)\Gamma(\gamma-m+z_\ell)\prod_{i=1}^{n}\Gamma(\alpha_i-z_\ell)\prod_{j=1}^{n+1}\Gamma(\beta_j+z_\ell)\right\}}{\displaystyle\prod_{\ell=1}^{n}\Gamma(\gamma+z_\ell)\prod_{\substack{1\leq i,j\leq n\\ i\neq j}}\Gamma(z_i-z_j)},$$

where m is a nonnegative integer.

We have

$$\int_{T^n} H(k)dz_1\cdots dz_n = (-1)^n\int_{T^n} H(k-1)dz_1\cdots dz_n. \tag{3.8}$$

Similarly to §2, we find that

$$\begin{aligned}&\int_{T^n} H(k-1)dz_1\cdots dz_n\\ &\quad= \int_{N^n} H(k-1)dz_1\cdots dz_n - \sum_{j=0}^{n-1}\int_{T^{n-1-j}}\int_{N^j}\int_{C} H(k-1)dz_1\cdots dz_n.\end{aligned} \tag{3.9}$$

Let

$$J(k-1)=\frac{\prod_{i=1}^{n}\Gamma(\alpha_i+\gamma-k)\prod_{j=1}^{n+1}\Gamma(\beta_j+k-\gamma)\prod_{\ell=2}^{n}(z_\ell-k+\gamma)}{(k-1)!\prod_{\ell=2}^{n}\Gamma(\gamma+z_\ell)\prod_{\substack{2\le i,j\le n\\ i\ne j}}\Gamma(z_i-z_j)}$$
$$\cdot\prod_{\ell=2}^{n}\left\{\prod_{i=1}^{n}\Gamma(\alpha_i-z_\ell)\prod_{j=1}^{n+1}\Gamma(\beta_j+z_\ell)\right\}.$$

It follows that

$$\begin{aligned}(3.10)\qquad &\sum_{j=0}^{n-1}\frac{1}{(2\pi i)^n}\int_{T^{n-1-j}}\int_{N^j}\int_{C}H(k-1)dz_1\cdots dz_n\\ &=\sum_{j=0}^{n-1}\frac{1}{(2\pi i)^{n-1}}\int_{T^{n-1-j}}\int_{N^j}J(k-1)dz_2\cdots dz_n\\ &=\sum_{j=0}^{n-1}\frac{1}{(2\pi i)^{n-1}}\int_{T^{n-1}}J(k-1)dz_2\cdots dz_n\\ &=\frac{n}{(2\pi i)^{n-1}}\int_{T^{n-1}}J(k-1)dz_2\cdots dz_n,\end{aligned}$$

since the function $J(k-1)$ in the variable z_j, $2\le j\le n$, has no poles in the disk bounded by the contour C.

Consider the case $k=1$. Then

$$H(0)=\frac{\prod_{\ell=1}^{n}\left\{\Gamma(1-\gamma-z_\ell)\prod_{i=1}^{n}\Gamma(\alpha_i-z_\ell)\prod_{j=1}^{n+1}\Gamma(\beta_j+z_\ell)\right\}}{\prod_{\substack{1\le i,j\le n\\ i\ne j}}\Gamma(z_i-z_j)}$$

and the integral $\int_{N^n}H(0)dz_1\cdots,dz_n$ can be evaluated by means of the $u(n)$ generalization of Barnes' first lemma [9, Thm. 5.1]. We find

$$(3.11)\qquad \int_{N^n}H(0)dz_1\cdots dz_n=0.$$

Applying identities (3.9)–(3.11), we obtain

$$(3.12)\qquad \frac{1}{(2\pi i)^n}\int_{T^n}H(0)dz_1\cdots dz_n=-\frac{n}{(2\pi i)^{n-1}}\int_{T^{n-1}}J(0)dz_2\cdots dz_n.$$

We have

$$J(0)=\prod_{i=1}^{n}\Gamma(\alpha_i+\gamma-1)\prod_{j=1}^{n+1}\Gamma(\beta_j+1-\gamma)\cdot\frac{\prod_{\ell=2}^{n}\left\{\prod_{i=1}^{n}\Gamma(\alpha_i-z_\ell)\prod_{j=1}^{n+1}\Gamma(\beta_j+z_\ell)\right\}}{\prod_{\ell=2}^{n}\Gamma(\gamma-1+z_\ell)\prod_{\substack{2\le i,j\le n\\ i\ne j}}\Gamma(z_i-z_j)}.$$

Therefore, the integral on the right-hand side of equation (3.12) can be evaluated by the dimension $n-1$ case of identity (3.2), which we assume is valid by the induction hypothesis on the dimension n. From (3.8) and (3.12), we obtain, after simplification,

$$\begin{aligned}&\frac{1}{(2\pi i)^n}\int_{T^n} H(1)dz_1\cdots dz_n\\ (3.13)\qquad &= \frac{\prod_{i=1}^{n}\Gamma(\alpha_i+\gamma-1)\prod_{j=1}^{n+1}\Gamma(\beta_j+2-\gamma)\prod_{i=1}^{n}\prod_{j=1}^{n+1}\Gamma(\alpha_i+\beta_j)}{\prod_{i=1}^{n+1}\Gamma(\gamma-\beta_j)}.\end{aligned}$$

Since identity (3.13) is equivalent to the $k=1$ case of identity (3.2), this completes the $k=1$, dimension n case of identity (3.2).

We now consider the $k>1$, dimension n case of identity (3.2). The $k-1$, dimension n case is assumed to be valid. We have

$$\begin{aligned}\int_{N^n} H(k-1)dz_1\cdots dz_n &= \frac{(2\pi i)^n n!\prod_{i=1}^{n}\prod_{j=1}^{n+1}\Gamma(\alpha_i+\beta_j)}{(k-2)!\prod_{j=1}^{n+1}\Gamma(\gamma-\beta_j)}\\ (3.14)\qquad &\quad\cdot\prod_{j=1}^{n+1}\Gamma(k-\gamma+\beta_j)\prod_{i=1}^{n}\Gamma(\alpha_i+\gamma+1-k).\end{aligned}$$

Similarly to §2, to compute the integral of the function $J(k-1)$ we will define a related function $I(k-1)$ and a difference equation involving $I(k-1)$ and $J(k-1)$. Let

$$I(k-1)=\frac{\prod_{\ell=2}^{n}\left\{\prod_{i=1}^{n}\Gamma(\alpha_i-z_\ell)\prod_{j=1}^{n+1}\Gamma(\beta_j+z_\ell)\right\}}{\prod_{\ell=2}^{n}\Gamma(\gamma+z_\ell)\prod_{\substack{2\le i,j\le n\\ i\neq j}}\Gamma(z_i-z_j)}.$$

If $f(\beta_\ell,\gamma)$ is a function involving the parameters β_ℓ and γ, then we define the shift operator Λ_ℓ by

$$\Lambda_\ell f(\beta_\ell,\gamma)=f(\beta_\ell+1,\gamma),$$

i.e., which shifts β_ℓ, but fixes γ. It then follows from Lemma 3.3 that

$$\begin{aligned}J(k-1)&=\frac{\prod_{i=1}^{n}\Gamma(\alpha_i+\gamma-k)\prod_{j=1}^{n+1}\Gamma(\beta_j+k-\gamma)}{(k-1)!}\\ (3.15)\qquad &\quad\cdot\sum_{\ell=1}^{n+1}\left\{\prod_{\substack{j=1\\ j\neq\ell}}^{n+1}\frac{(\gamma-k-\beta_j)}{(\beta_\ell-\beta_j)}\Lambda_\ell I(k-1)\right\}.\end{aligned}$$

Since $\gamma = 1 + \sum_{i=1}^{n} \alpha_i + \sum_{j=1}^{n+1} \beta_j$, then the dimension $n-1$ case of identity (3.2) (which we assume is valid by the induction hypothesis) can be used to evaluate the integrals of the $\Lambda_\ell I(k-1)$ functions. Identity (3.15) then allows us to evaluate the integral of $J(k-1)$. We find

$$\begin{aligned}\frac{1}{(2\pi i)^{n-1}} \int\limits_{T^{n-1}} J(k-1) dz_2 \cdots dz_n \\ (3.16) \qquad = \frac{(n-1)! \prod_{i=1}^{n} \Gamma(\alpha_i + \gamma - k) \prod_{j=1}^{n+1} \Gamma(\beta_j + k - \gamma) \prod_{i=1}^{n} \prod_{j=1}^{n+1} \Gamma(\alpha_i + \beta_j)}{(k-1)! \prod_{j=1}^{n+1} \Gamma(\gamma - \beta_j)} \\ \cdot \sum_{\ell=1}^{n+1} \left\{ \prod_{\substack{j=1 \\ j \neq \ell}}^{n+1} \frac{(\gamma - k - \beta_j)}{(\beta_\ell - \beta_j)} (\gamma - 1 - \beta_\ell) \prod_{i=1}^{n} (\alpha_i + \beta_\ell) \right\}.\end{aligned}$$

A simple extension of Lemma 3.3 (to the case $s = m+1$ in the notation of Lemma 3.3) can be used to evaluate the sum on the right-hand side of (3.16). We have

$$\begin{aligned}-\sum_{\ell=1}^{n+1} \left\{ \prod_{\substack{j=1 \\ j \neq \ell}}^{n+1} \frac{(\gamma - k - \beta_j)}{(\beta_\ell - \beta_j)} (1 - \gamma + \beta_\ell) \prod_{i=1}^{n} (\alpha_i + \beta_\ell) \right\} \\ (3.17) \qquad = \prod_{j=1}^{n+1} (\gamma - k - \beta_j) - (\gamma - k + 1 - \gamma) \prod_{i=1}^{n} (\gamma - k + \alpha_i).\end{aligned}$$

Substituting (3.17) into the right-hand side of (3.16) and applying identities (3.8)–(3.10), (3.14), and (3.16), we find

$$\begin{aligned}(3.18) \qquad \frac{1}{(2\pi i)^n} \int\limits_{T^n} H(k) dz_1 \cdots dz_n = \frac{n! \prod_{i=1}^{n} \Gamma(\alpha_i + \gamma - k)}{\Gamma(k) \prod_{j=1}^{n+1} \Gamma(\gamma - \beta_j)} \\ \cdot \prod_{j=1}^{n+1} \Gamma(\beta_j + 1 + k - \gamma) \prod_{i=1}^{n} \prod_{j=1}^{n+1} \Gamma(\alpha_i + \beta_j).\end{aligned}$$

Recalling that $\alpha_{n+1} = k + 1 - \gamma$ and $\beta_{n+2} = \gamma - k$ completes the proof of the $k > 1$, dimension n case of identity (3.2).

To complete the proof of the full dimension n case of identity (3.2), we begin by choosing an arbitrary complex number α_{n+1} such that $0 < \mathrm{Re}(\alpha_{n+1}) < 1$ and setting $\beta_{n+2} = 1 - \alpha_{n+1}$. Let $\alpha_1, \cdots, \alpha_n, \beta_1, \cdots, \beta_n \in C$ be fixed such that $\mathrm{Re}(\alpha_i)$, $\mathrm{Re}(\beta_j) > 0$. Then set $\beta_{n+1} = 1 + s - \alpha_n - \sum_{i=1}^{n} (\alpha_i + \beta_i)$ with $s \in C$ such that $\mathrm{Re}(s) \geq 0$ and $\mathrm{Re}(\beta_{n+1}) > 0$. From (3.18) it follows that (3.2) is valid for $s = 0, 1, 2, \cdots$. Using Stirling's formula to estimate the growth the left-hand side and right-hand side of (3.2) as a function of s, it can be seen that the growth conditions for Carlson's theorem [4, p. 39] are satisfied. Carlson's theorem now shows that identity (3.2) is valid for all $s \in C$ (subject to the restriction $\mathrm{Re}(\alpha_1), \cdots, \mathrm{Re}(\alpha_{n+1}), \mathrm{Re}(\beta_1), \cdots, \mathrm{Re}(\beta_{n+2}) > 0$). In other words, identity (3.2) is valid whenever $\alpha_{n+1} + \beta_{n+2} = 1$.

This last condition can also be eliminated. With notation as in Lemma 3.7, we rewrite the difference equation (3.8) as

$$\rho_{n+2}f = \prod_{j=1}^{n+1}\left(\frac{\beta_{n+2}-\beta_j}{\gamma-\beta_j}\right)\left[f - \sum_{\ell=1}^{n+1}\left\{\prod_{\substack{j=1\\ j\neq\ell}}^{n+2}\left(\frac{\gamma-\beta_j}{\beta_\ell-\beta_j}\right)(\rho_\ell f)\right\}\right], \tag{3.19}$$

where f can be either the left or right-hand side of identity (3.2). As in §2, suppose now that identity (3.2) is valid when $\beta_{n+2} = \ell - \alpha_{n+1}$ for some positive integer ℓ. If we choose the parameters on the left-hand side of (3.19) to satisfy $\beta_{n+2} = \ell + 1 - \alpha_{n+1}$, then those on the right-hand side will satisfy $\beta_{n+2} = \ell - \alpha_{n+1}$. It follows that identity (3.2) will also be valid for $\beta_{n+2} = \ell + 1 - \alpha_{n+1}$. An application of Carlson's theorem similar to the above now shows that identity (3.2) is valid for arbitrary $\beta_{n+2} \in \mathcal{C}$ with $\mathrm{Re}(\beta_{n+2}) > 0$. Finally, by analytic continuation, we drop the restriction on the values of the parameters α_i, β_j, and γ. This completes the proof of the dimension n case of identity (3.2). By induction, this proves Theorem 3.1.

4. A generalization of the Nasrallah–Rahman integral for the symplectic groups. In this section we prove a generalization of the Nasrallah–Rahman integral (1.2) associated to the compact symplectic groups $Sp(n)$. The $Sp(1) \approx SU(2)$ case is just the Nasrallah–Rahman integral. The proof of the general integral identity is very similar to that discussed in §2. The main result is the following.

THEOREM 4.1. *For $n \geq 1$, let $a_i \in \mathcal{C}$, $1 \leq i \leq 2n+3$, with $|a_i| < 1$. Set $C = \prod_{i=1}^{2n+3} a_i$, then*

$$\begin{aligned}\frac{1}{(2\pi i)^n}\int_{T^n} &\frac{\prod_{1\leq i<j\leq n}[z_iz_j]_\infty[z_i^{-1}z_j^{-1}]_\infty[z_iz_j^{-1}]_\infty[z_i^{-1}z_j]_\infty}{\prod_{i=1}^{2n+3}\prod_{j=1}^{n}[a_iz_j]_\infty[a_iz_j^{-1}]_\infty}\\ &\cdot\prod_{j=1}^{n}\left\{[z_j^2]_\infty[z_j^{-2}]_\infty[Cz_j]_\infty[Cz_j^{-1}]_\infty\frac{dz_j}{z_j}\right\}\\ &= \frac{n!2^n\prod_{j=1}^{2n+3}[Ca_j^{-1}]_\infty}{[q]_\infty^n\prod_{1\leq i<j\leq 2n+3}[a_ia_j]_\infty},\end{aligned}\tag{4.2}$$

where the integral in each variable z_j is over the unit circle T taken in the positive direction.

Proof. Note that if the contour T is allowed to be shifted, then the restriction $|a_i| < 1$, $1 \leq i \leq 2n+3$, may be eliminated. Also observe that the $n = 1$ case of the integral on the left-hand side of (4.2) is just the Nasrallah–Rahman integral (1.2).

To begin the proof of identity (4.2), we require that for any integer $\ell, a_i^{\pm 1} \neq a_jq^\ell$ for $1 \leq i,j \leq 2n+3, i \neq j$. This restriction will be removed at the end of the proof of (4.2).

As in §2, we first show that both sides of (4.2) satisfy the same q-difference equation in the parameters a_i, $1 \leq i \leq 2n+3$. We will need another version of the partial fractions expansion.

LEMMA 4.3. *Let $\{x_1, \cdots, x_m\}, \{y_1, \cdots, y_{m+1}\}$ and t be indeterminants with $y_i \neq y_j$ and $y_i y_j \neq 1$ for $1 \leq i, j \leq s, i \neq j$. Then*

$$\sum_{\ell=1}^{m+1} \left\{ \prod_{\substack{j=1 \\ j\neq\ell}}^{m+1} \frac{(t-y_j)(ty_j-1)}{(y_\ell - y_j)(y_\ell y_j - 1)} \prod_{k=1}^{m} (1-y_\ell x_k)(1-y_\ell x_k^{-1}) \right\} \tag{4.4}$$

$$= \prod_{k=1}^{m} (1-tx_k)(1-tx_k^{-1}).$$

Proof. Divide both sides of (4.4) by $\prod_{j=1}^{m+1}(t-y_j)(ty_j-1)$ and compute residues with respect to the poles y_ℓ and y_ℓ^{-1}, $1 \leq \ell \leq m+1$.

We now show that both sides of (4.2) satisfy the same q-difference equation.

DEFINITION 4.5. If f is a function involving the parameters a_ℓ and C, then define

$$R_\ell f(a_\ell, C) = f(a_\ell q, Cq).$$

LEMMA 4.6. *With notation as in Theorem 4.1, let f be either the left-hand side or right-hand side of (4.2). Then*

$$\sum_{\ell=1}^{n+1} \left\{ \prod_{\substack{j=1 \\ j\neq\ell}}^{n+1} \left(\frac{C-a_j}{a_\ell - a_j}\right)\left(\frac{Ca_j - 1}{a_\ell a_j - 1}\right) \cdot (R_\ell f) \right\} = f. \tag{4.7}$$

Proof. Let I and Q denote, respectively, the left-hand side and right-hand side of (4.2). Using (4.4), it is easily checked that I satisfies identity (4.7). To show that Q satisfies the q-difference equation (4.7) is equivalent to verifying the following identity:

$$\sum_{\ell=1}^{n+1} \left\{ \prod_{\substack{i=1 \\ i\neq\ell}}^{2n+3} a_i \prod_{\substack{j=1 \\ j\neq\ell}}^{n+1} \left(\frac{Ca_j - 1}{a_\ell - a_j}\right) \prod_{k=n+2}^{2n+3} \left(\frac{a_\ell a_k - 1}{C - a_k}\right) \right\} = 1. \tag{4.8}$$

We can rewrite the left-hand side of identity (4.8) as

(4.9)
$$\prod_{k=n+2}^{2n+3} \frac{a_k}{C-a_k} \prod_{j=1}^{n+1}(Ca_j - 1) \sum_{\ell=1}^{n+1} \left\{ \left(\frac{C^{-1}}{a_\ell - C^{-1}}\right) \prod_{j=1}^{n+1} \left(\frac{a_j}{a_\ell - a_j}\right) \prod_{k=n+2}^{2n+3} (a_\ell a_k - 1) \right\}.$$

Letting t be an indeterminant, then we have the following identity (which comes from an extension of (2.4)):

$$\sum_{\ell=1}^{n+1}\left\{\left(\frac{t-C^{-1}}{a_\ell-C^{-1}}\right)\prod_{\substack{j=1\\ j\neq\ell}}^{n+1}\left(\frac{t-a_j}{a_\ell-a_j}\right)\prod_{k=n+2}^{2n+3}(a_\ell a_k-1)\right\}$$

$$(4.10)\qquad =\prod_{k=n+2}^{2n+3}(ta_k-1)-(t-C^{-1})\prod_{j=1}^{n+1}(t-a_j)\prod_{k=n+2}^{2n+3}a_k$$

$$-\prod_{j=1}^{n+1}\left(\frac{t-a_j}{C^{-1}-a_j}\right)\prod_{k=n+2}^{2n+3}(C^{-1}a_k-1).$$

Identity (4.10) can be verified by dividing both sides of (4.10) by $(t-C^{-1})\prod_{j=1}^{n+1}(t-a_j)$ and then doing a partial fractions expansion of the right-hand side of (4.10) with respect to the indeterminant t. Setting $t=0$ in identity (4.10) and substituting into (4.9), we then prove (4.8). This completes the proof of Lemma 4.6.

The proof of Theorem 4.1 proceeds by induction on n, similarly to the proof of Theorem 2.1. If we define the left-hand side of (4.2) to be identically 1 for $n=0$, then the $n=0$ case of identity (4.2) is trivially true. We will assume from now on that $n\geq 1$ and that the $n-1$ case of identity (4.2) is valid. We first prove identity (4.2) in the case that $a_{2n+2}=q^{k+1}C^{-1}$ and $a_{2n+3}=q^{-k}C$, where k is a positive integer. As before, for fixed n we will prove identity (4.2) by induction with respect to k, with the cases $k=1$ and $k>1$ handled differently.

Assume that $q^{k+1}<|C|<q^k$ and that $|q^kC^{-1}|<r^{-1}$, where r is a real number chosen as close to 1 as necessary. We also assume that $|a_i|<r$ for $1\leq i\leq 2n+2$.

Let N be the contour which is the union of the circle of radius r centered at the origin, traversed in the positive direction, and the circle of radius ε centered at the point of q^kC^{-1}, traversed in the positive direction, where ε is a sufficiently small positive real number. We will use the notation

$$H(m)=$$

$$\frac{\prod_{\ell=1}^{n}\{[Cz_\ell]_\infty[Cz_\ell^{-1}]_\infty[z_\ell^2]_\infty[z_\ell^{-2}]_\infty\}\prod_{1\leq i<j\leq n}\{[z_iz_j]_\infty[z_i^{-1}z_j^{-1}]_\infty[z_iz_j^{-1}]_\infty[z_i^{-1}z_j]_\infty\}}{\prod_{\ell=1}^{n}\{[q^{m+1}C^{-1}z_\ell]_\infty[q^{m+1}C^{-1}z_\ell^{-1}]_\infty[q^{-m}Cz_\ell]_\infty[q^{-m}Cz_\ell^{-1}]_\infty\}}$$

$$\cdot\left\{\prod_{\ell=1}^{n}\prod_{i=1}^{2n+1}[a_iz_\ell]_\infty[a_iz_\ell^{-1}]_\infty\right\}^{-1},$$

where m is a nonnegative integer.

As in §2, we find

$$(4.11)\qquad \int_{T^n}H(k)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n}=(q^kC^{-1})^{2n}\int_{T^n}H(k-1)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n}.$$

We also have

$$\begin{aligned}(4.12)\qquad \int_{T^{n-1}} H(k-1)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n} &= \sum_{j=0}^{n-1}\int_{T^{n-1-j}}\int_{N^j}\int_{T-N} H(k-1)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n}\\ &\quad+\int_{N^{n-1}} H(k-1)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n}.\end{aligned}$$

With respect to the variable z_1, the poles of the integrand $H(k-1)\prod_{i=1}^n z_i^{-1}$ inside the region bounded by the contour $T-N$ are at $z_1 = C^{-1}q^k$ and at $z_1 = Cq^{-k}$. It follows that

$$\begin{aligned}(4.13)\qquad &\sum_{j=0}^{n-1}\frac{1}{(2\pi i)^n}\int_{T^{n-1-j}}\int_{N^j}\int_{T-N} H(k-1)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n}\\ &=\sum_{j=0}^{n-1}-\frac{2}{(2\pi i)^{n-1}}\int_{T^{n-1-j}}\int_{N^j} J(k-1)\frac{dz_2}{z_2}\cdots\frac{dz_n}{z_n},\end{aligned}$$

where

$$\begin{aligned}J(k-1) = {}&\frac{(1-q^{-2k}C^2)[C^2q^{-k}]_\infty}{[q]_{k-1}[q]_\infty}\\ &\cdot\frac{\prod_{j=2}^{n}\{(1-Cq^{-k}z_j)(1-Cq^{-k}z_j^{-1})[Cz_j]_\infty[Cz_j^{-1}]_\infty[z_j^2]_\infty[z_j^{-2}]_\infty\}}{\prod_{i=1}^{2n+1}\{[a_iC^{-1}q^k]_\infty[a_iCq^{-k}]_\infty\prod_{j=2}^{n}[a_iz_j]_\infty[a_iz_j^{-1}]_\infty\}}\\ &\cdot\prod_{2\le i<j\le n}\{[z_iz_j]_\infty[z_i^{-1}z_j^{-1}]_\infty[z_iz_j^{-1}]_\infty[z_i^{-1}z_j]_\infty\}.\end{aligned}$$

As in §2, we can deform the contour N into the contour T and show that

$$(4.14)\qquad \int_{T^{n-1-j}}\int_{N^j} J(k-1)\frac{dz_2}{z_2}\cdots\frac{dz_n}{z_n} = \int_{T^{n-2}} J(k-1)\frac{dz_2}{z_2}\cdots\frac{dz_n}{z_n}$$

for $0\le j\le n-1$.

Consider the case $k=1$. Then

$$H(0) = \frac{\prod_{\ell=1}^{n}[z_\ell^2]_\infty[z_\ell^{-2}]_\infty\prod_{1\le i<j\le n}[z_iz_j]_\infty[z_i^{-1}z_j^{-1}]_\infty[z_iz_j^{-1}]_\infty[z_i^{-1}z_j]_\infty}{\prod_{\ell=1}^{n}\left\{[qC^{-1}z_\ell]_\infty[qC^{-1}z_\ell^{-1}]_\infty\prod_{i=1}^{2n+1}[a_iz_\ell]_\infty[a_iz_\ell^{-1}]_\infty\right\}}$$

and the integral $\int_{N^n} H(0)dz_1/z_1\cdots dz_n/z_n$ can be evaluated by means of the $Sp(n)$ generalization of the Askey–Wilson integral [9, Thm. 7.1]. We find

$$(4.15)\qquad \int_{N^n} H(0)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n} = 0.$$

We also have

$$J(0) = \frac{[C^2q^{-2}]_\infty}{[q]_\infty} \cdot \frac{\prod_{j=2}^{n} \{[Cq^{-1}z_j]_\infty [Cq^{-1}z_j^{-1}]_\infty [z_j^2]_\infty [z_j^{-2}]_\infty\}}{\prod_{i=1}^{2n+1} \left\{ [a_iC^{-1}q]_\infty [a_iCq^{-1}]_\infty \prod_{j=2}^{n} [a_iz_j]_\infty [a_iz_j^{-1}]_\infty \right\}}$$
$$\cdot \prod_{2\le i<j\le n} [z_iz_j]_\infty [z_i^{-1}z_j^{-1}]_\infty [z_iz_j^{-1}]_\infty [z_i^{-1}z_j]_\infty .$$

The integral of $J(0)$ can be evaluated by the dimension $n-1$ case of identity (4.2), which we assume is valid by the induction hypothesis on the dimension n. From (4.11)–(4.14) and (4.2), we find

$$\frac{1}{(2\pi i)^n} \int H(1) \frac{dz_1}{z_1} \cdots \frac{dz_n}{z_n}$$
$$= -\frac{2^n n! q^{2n} [C^2q^{-2}]_\infty \prod_{i=1}^{2n+1} [Cq^{-1}a_i^{-1}]_\infty}{C^{2n} [q]_\infty^n \prod_{i=1}^{2n+1} [a_iC^{-1}q]_\infty [a_iCq^{-1}]_\infty} \tag{4.16}$$
$$\cdot \left\{ \prod_{1\le i<j\le 2n+1} [a_ia_j]_\infty \right\}^{-1}$$
$$= \frac{2^n n! [C^2q^{-2}]_\infty \prod_{i=1}^{2n+1} [Ca_i^{-1}]_\infty}{[q]_\infty^n \prod_{i=1}^{2n+1} [a_iC^{-1}q^2]_\infty [a_iCq^{-1}]_\infty \prod_{1\le i<j\le 2n+1} [a_ia_j]_\infty}.$$

This completes the proof of the $k=1$, dimension n case of identity (4.2).

We now consider the $k>1$, dimension n case of identity (4.2). By the induction hypothesis for $k-1$, we have

$$\frac{1}{(2\pi i)^n} \int_{N^n} H(k-1) \frac{dz_1}{z_1} \cdots \frac{dz_n}{z_n}$$
$$= n! 2^n \cdot \frac{[q^{k-1}]_\infty [q^{-k}C^2]_\infty \prod_{i=1}^{2n+1} [Ca_i^{-1}]_\infty}{[q]_\infty^{n+1} \prod_{i=1}^{2n+1} [q^kC^{-1}a_i]_\infty [q^{-k+1}Ca_i]_\infty \prod_{1\le i<j\le 2n+1} [a_ia_j]_\infty}. \tag{4.17}$$

As in §2, we define a function $I(k-1)$, to be used in computing the integral of $J(k-1)$. Let

$$I(k-1)=\frac{\prod_{j=2}^{n}\{[Cz_j]_\infty[Cz_j^{-1}]_\infty[z_j^2]_\infty[z_j^{-2}]_\infty\}}{\prod_{\ell=2}^{n}\prod_{i=1}^{2n+1}[a_iz_\ell]_\infty[a_iz_\ell^{-1}]_\infty}\cdot\prod_{2\leq i<j\leq n}[z_iz_j]_\infty[z_i^{-1}z_j^{-1}]_\infty[z_iz_j^{-1}]_\infty[z_i^{-1}z_j]_\infty .$$

If $f(a_\ell, c)$ is a function involving the parameters a_ℓ and c, then we define the shift operator

$$L_\ell f(b,c)=f(b_\ell q,c).$$

It follows from Lemma 4.3 that

$$\begin{aligned}J(k-1)={}&\frac{(1-q^{-2k}C^2)[C^2q^{-k}]_\infty}{[q]_{k-1}[q]_\infty\prod_{i=1}^{2n+1}\{[a_iC^{-1}q^k]_\infty[a_iCq^{-k}]_\infty\}}\\&\cdot\sum_{\ell=1}^{n}\left\{\prod_{\substack{j=1\\j\neq\ell}}^{n}\frac{(Cq^{-k}-a_j)(Cq^{-k}a_j-1)}{(a_\ell-a_j)(a_\ell a_j-1)}L_\ell I(k-1)\right\}.\end{aligned}\tag{4.18}$$

Identity (4.18) allows us to compute the integral of the function $J(k-1)$ in terms of the integrals of the functions $L_\ell I(k-1)$. We have

$$\begin{aligned}&\frac{1}{(2\pi i)^{n-1}}\int\limits_{T^{n-1}}J(k-1)\frac{dz_2}{z_2}\cdots\frac{dz_n}{z_n}\\&=\frac{(n-1)!2^{n-1}(1-q^{-2k}C^2)[C^2q^{-k}]_\infty\prod_{i=1}^{2n+1}[Ca_i^{-1}]_\infty}{[q]_\infty^n[q]_{k-1}\prod_{1\leq i<j\leq 2n+1}[a_ia_j]_\infty\prod_{i=1}^{2n+1}\{[a_iC^{-1}q^k]_\infty[a_iCq^{-k}]_\infty\}}\\&\quad\cdot\sum_{\ell=1}^{n}\left\{\prod_{\substack{j=1\\j\neq\ell}}^{n}\frac{(Cq^{-k}-a_j)(Cq^{-k}a_j-1)}{(a_\ell-a_j)(a_\ell a_j-1)}\right.\\&\qquad\left.\cdot(1-q^{-1}Ca_\ell^{-1})\prod_{\substack{i=1\\i\neq\ell}}^{2n+1}(1-a_ia_\ell)\right\}.\end{aligned}\tag{4.19}$$

To evaluate the sum on the right-hand side of (4.19), we will use the special case of identity (2.4) where we equate the t^m terms on both sides of (2.4). This special case of identity (2.4) is due to Louck and Biedenharn [13]. We have

$$\begin{aligned}
&\sum_{\ell=1}^{n}\left\{\prod_{\substack{j=1\\ j\neq\ell}}^{n}\frac{(Cq^{-k}-a_j)(Cq^{-k}a_j-1)}{(a_\ell-a_j)(a_\ell a_j-1)}\right.\\
&\qquad\qquad\left.\cdot(1-q^{-1}Ca_\ell^{-1})\prod_{\substack{i=1\\ i\neq\ell}}^{2n+1}(1-a_ia_\ell)\right\}\\
&\quad=q^{k-1}\prod_{j=1}^{n}(q^{-k}C-a_j)(q^{-k}Ca_j-1)\\
&\qquad\qquad\cdot\sum_{\ell=1}^{n}\left\{\frac{(1-qC^{-1}a_\ell)}{a_\ell(a_\ell-q^kC^{-1})}\cdot\frac{\prod_{i=n+1}^{2n+1}(a_ia_\ell-1)}{(a_\ell-q^{-k}C)\prod_{\substack{j=1\\ j\neq\ell}}^{n}(a_\ell-a_j)}\right\}\\
&=q^{k-1}\prod_{j=1}^{n}(q^{-k}C-a_j)(q^{-k}Ca_j-1)\left\{-qC^{-1}\prod_{i=n+1}^{2n+1}a_i-\frac{(-1)^{n+1}}{\prod_{j=1}^{n}(-a_j)}\right.\\
&\qquad\qquad-\frac{(1-qC^{-1}q^kC^{-1})\prod_{i=n+1}^{2n+1}(a_iq^kC^{-1}-1)}{q^kC^{-1}(q^kC^{-1}-q^{-k}C)\prod_{j=1}^{n}(q^kC^{-1}-a_j)}\\
&\qquad\qquad\left.-\frac{(1-qC^{-1}q^{-k}C)\prod_{i=n+1}^{2n+1}(a_iq^{-k}C-1)}{q^{-k}C(q^{-k}C-q^kC^{-1})\prod_{j=1}^{n}(q^{-k}C-a_j)}\right\}\\
&=-\frac{(q^{-k}C)^{2n}(1-q^{-k-1}C^2)\prod_{i=1}^{2n+1}(1-a_iq^kC^{-1})}{(1-q^{-2k}C^2)}\\
&\qquad\qquad+\frac{(1-q^{k-1})\prod_{i=1}^{2n+1}(1-a_iq^{-k}C)}{(1-q^{-2k}C^2)},
\end{aligned}\tag{4.20}$$

where we have used the identity $q^{-1}C=\prod_{j=1}^{2n+1}a_j$. Substituting (4.20) into the right-hand side of (4.19) and using identities (4.11)–(4.14) and (4.17), we find

$$\frac{1}{(2\pi i)^n}\int\limits_{T^n} H(k)\frac{dz_1}{z_1}\cdots\frac{dz_n}{z_n} = \frac{2^n n![C^2q^{-k-1}]_\infty \prod\limits_{i=1}^{2n+1}[Ca_i^{-1}]_\infty}{[q]_\infty^n[q]_{k-1}\prod\limits_{1\le i<j\le 2n+1}[a_ia_j]_\infty \prod\limits_{i=1}^{2n+1}[a_iC^{-1}q^{k+1}]_\infty[a_iCq^{-k}]_\infty}. \tag{4.21}$$

This completes the proof of the $k > 1$, dimension n case of identity (4.2).

The proof of the general dimension n case of identity (4.2) is completed similarly to §2. Let a_{2n+3} be an arbitrary complex number such that $|q| < |a_{2n+3}| < 1$ and set $a_{2n+2} = qa_{2n+3}^{-1}$. Let $a_1, \cdots, a_{2n+1} \in C$ be fixed so that $|a_i| < 1$ for $1 \le i \le 2n+1$. Set $a_{2n+1} = q^k a_{2n+3}^{-1}\prod_{i=1}^{2n} a_i^{-1}$, with $k > 0$ a sufficiently large integer so that $|a_{2n+1}| < 1$. From (4.21) it follows that identity (2.2) is valid for a set of values of a_{2n+1} which includes the limit point $a_{2n+1} = 0$. (The case $a_{2n+1} = 0$ reduces to the $Sp(n)$ generalization of the Askey–Wilson integral [9, Thm. 7.1].) By analytic continuation in the parameters a_{2n+1}, we conclude that identity (4.2) is valid whenever $a_{2n+1}a_{2n+3} = q$.

We use Lemma 4.6 to eliminate this last condition. With notation as in Lemma 4.6, we rewrite the q-difference equation (4.7) as

$$\begin{aligned} R_{2n+2}f = &\prod_{j=n+2}^{2n+1}\left(\frac{a_{2n+2}-a_j}{C-a_j}\right)\left(\frac{a_{2n+2}a_j-1}{Ca_j-1)}\right) \\ &\cdot\left[f-\sum_{\ell=n+2}^{2n+1}\left\{\prod_{\substack{j=n+2\\ j\neq\ell}}^{2n+2}\left(\frac{C-a_j}{a_\ell-a_j}\right)\left(\frac{Ca_j-1}{a_\ell a_j-1}\right)(R_\ell f)\right\}\right], \end{aligned} \tag{4.22}$$

where f can be either the left or right-hand side of identity (4.2). Now suppose that identity (4.2) is valid whenever $a_{2n+2} = q^\ell a_{2n+3}^{-1}$ for some positive integer ℓ. If we choose the parameters on the left-hand side of (4.22) to satisfy $a_{2n+2} = q^{\ell+1}a_{2n+3}^{-1}$, then those on the right-hand side will satisfy $a_{2n+2} = q^\ell a_{2n+3}^{-1}$. It follows that identity (4.2) will also be valid for $a_{2n+2} = q^{\ell+1}a_{2n+3}^{-1}$. Just as above, we conclude by analytic continuation that identity (4.2) is valid for arbitrary a_{2n+2}. Finally, by analytic continuation we also drop the other restrictions on the values of the parameters a_j and C. This completes the proof of the dimension n case of identity (4.2) and the proof of Theorem 4.1.

5. Some multidimensional Mellin–Barnes integrals.

In this section we give analogues of Nasrallah–Rahman integrals associated to the Lie algebras $su(n)$ and $sp(n)$. The integrals below can be viewed as limiting cases as $q \to 1^-$ of Theorems 2.1 and 4.1. The proofs are very similar to that of Theorem 3.1. The $n = 2$ case of Theorem 5.1 and the $n = 1$ case of Theorem 5.3 below (which are equivalent) are due to Rahman [21].

The first result is an analogue of Theorem 2.1.

THEOREM 5.1. *For $n \ge 2$, let $\alpha_i \in C$, $1 \le i \le n$, and $\beta_j \in C$, $1 \le j \le n+1$,* *with* $\mathrm{Re}(\alpha_i)$, $\mathrm{Re}(\beta_j) > 0$. *Set* $A = \sum_{i=1}^{n} \alpha_i, B = \sum_{j=1}^{n+1} \beta_j$ *and* $\gamma = A + B$, *then*

$$\int_{-i\infty}^{i\infty}\cdots\int_{-i\infty}^{i\infty} \frac{\prod_{k=1}^{n} \left\{\prod_{i=1}^{n} \Gamma(\alpha_i - z_k)\prod_{j=1}^{n+1}\Gamma(\beta_j+z_k)\right\} dz_1\cdots dz_{n-1}}{\prod_{k=1}^{n}\Gamma(\gamma+z_k)\prod_{\substack{1\le i,j\le n\\ i\neq j}}\Gamma(z_i-z_j)} = \frac{n!\Gamma(A)\prod_{j=1}^{n+1}\Gamma(B-\beta_j)\prod_{i=1}^{n}\prod_{j=1}^{n+1}\Gamma(\alpha_i+\beta_j)}{\prod_{j=1}^{n+1}\Gamma(\gamma-\beta_j)\prod_{i=1}^{n}\Gamma(\alpha_i+B)}, \tag{5.2}$$

where $\sum_{k=1}^{n} z_k = 0$.

Proof. The proof follows exactly the same lines as the proof of Theorem 3.1. The only modification needed is that in the $k = 1$ case of the induction on k, the parameters α_i and β_j should be modified by adding on an additional parameter t with $\mathrm{Re}(t) > 0$. Thus the $k = 1$ case exactly parallels the proof in §2. Finally, an application of Stirling's formula and Carlson's theorem, just as in §3, allows us to complete the proof of identity (5.2) for general values of the parameters.

We also prove the following analogue of Theorem 4.1 in an entirely similar way.

THEOREM 5.3. *For* $n \geq 1$, *let* $\alpha_i \in C$, $1 \leq i \leq 2n+3$. *Set* $\gamma = \sum_{i=1}^{2n+3} \alpha_i$, *then*

$$\frac{1}{(2\pi i)^n}\int_{-i\infty}^{i\infty}\cdots\int_{-i\infty}^{i\infty} \frac{\prod_{i=1}^{2n+3}\prod_{j=1}^{n}\Gamma(\alpha_i+z_j)\Gamma(\alpha_i-z_j)}{\prod_{1\le i<j\le n}\Gamma(z_i+z_j)\Gamma(-z_i-z_j)\Gamma(z_i-z_j)\Gamma(z_j-z_i)} \cdot\prod_{j=1}^{n}\frac{dz_j}{\Gamma(2z_j)\Gamma(-2z_j)\Gamma(\gamma+z_j)\Gamma(\gamma-z_j)} = \frac{n!2^n\prod_{1\le i<j\le 2n+3}\Gamma(\alpha_i+\alpha_j)}{\prod_{i=1}^{2n+3}\Gamma(\gamma-\alpha_i)}, \tag{5.4}$$

where the contours of integration are deformed so as to separate the sequences of poles going to the right $\{\alpha_i + k \mid 1 \leq i \leq 2n+3, k = 0, 1, 2, \cdots\}$ *from the sequences of poles going to the left* $\{-\alpha_i - k \mid 1 \leq i \leq 2n+3, k = 0, 1, 2, \cdots\}$.

REFERENCES

[1] G. ANDREWS, *Problems and prospects for basic hypergeometric functions*, in The Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 191–224.

[2] R. ASKEY, *Beta integrals in Ramanujan's papers, his unpublished work and further examples*, in Ramanujan Revisited, Proceedings of Centenary Conference, G. Andrews et al., ed., Academic Press, New York, 1988, pp. 561–590.

[3] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985), p. 55.

[4] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge Math. Tract 32, Cambridge University Press, 1935; reprinted; Hafner, New York, 1964.

[5] E. W. BARNES, *A transformation of generalized hypergeometric series*, Quart. J. Math., 41 (1910), pp. 136–140.

[6] R. Y. DENIS AND R. A. GUSTAFSON, *An SU(n) q-beta integral transformation and multiple hypergeometric series identities*, SIAM J. Math. Anal., 23 (1992), pp. 552-561.

[7] K. I. GROSS AND D. ST. P. RICHARDS, *Special functions of matrix argument* I. *Algebraic induction, zonal polynomials and hypergeometric function*, Trans. Amer. Math. Soc., 30 (1987), pp. 781–811.

[8] R. A. GUSTAFSON, *A generalization of Selberg's beta integral*, Bull. Amer. Math. Soc., 22 (1990), pp. 97–105.

[9] ______, *Some q-beta and Mellin–Barnes integrals on compact Lie groups and Lie algebras*, preprint.

[10] G. J. HECKMAN AND E. M. OPDAM, *Root systems and hypergeometric functions* I, *Compositio Mathematica*, 64 (1987), pp. 329–352.

[11] K. W. J. KADELL, *The Selberg–Jack symmetric functions*, preprint.

[12] T. H. KOORNWINDER, *Representation of the twisted $SU(2)$ quantum group and some q-hypergeometric orthogonal polynomials*, Proc. Kon. Nederl. Akad. Wetensch., A 92 (1989), pp. 97–117.

[13] J. D. LOUCK AND L. C. BIEDENHARN, *Canonical unit adjoint tensor products in $U(n)$*, J. Math. Phys., 11 (1970), pp. 2368–2414.

[14] I. G. MACDONALD, *Symmetric Functions and Hall Polynomials*, Oxford University Press, Oxford, 1979.

[15] ______, *Some conjectures for root systems*, SIAM J. Math. Anal., 13 (1982), pp. 988–1007.

[16] ______, *Orthogonal polynomials associated with root systems*, preprint.

[17] T. MASUDA, K. MIMACHI, Y. NAKAGAMI, M. NOUMI, AND K. UENO, *Representation of the quantum group $SU_q(2)$ and the little q-Jacobi polynomials*, J. Funct. Anal., to appear.

[18] S. C. MILNE, *A q-analog of the Gauss summation theorem for hypergeometric series in $U(n)$*, Adv. in Math., 72 (1988), pp. 59–131.

[19] W. G. MORRIS, *Constant term identities for finite and affine root systems: conjectures and theorems*, Ph.D. thesis, University of Wisconsin-Madison, Madison, WI, 1982.

[20] B. NASRALLAH AND M. RAHMAN, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, SIAM J. Math. Anal., 16 (1985), pp. 186–197.

[21] M. RAHMAN, *An integral representation of a ${}_{10}\varphi_9$ and continuous bi-orthogonal ${}_{10}\varphi_9$ rational functions*, Canad. J. Math., 38 (1986), pp. 601–618.

[22] A. SELBERG, *Bemerkinger om et multipelt integral*, Norsk Mat. Tidsskr., 26 (1944), pp. 71–78.

[23] D. ZEILBERGER AND D. M. BRESSOUD, *A proof of Andrews q-Dyson conjecture*, Discrete Math., 54 (1985), pp. 201–224.

SIAM J. MATH. ANAL.
Vol. 23, No. 2, pp. 552–561, March 1992

AN $SU(n)$ q-BETA INTEGRAL TRANSFORMATION AND MULTIPLE HYPERGEOMETRIC SERIES IDENTITIES*

R. Y. DENIS† AND R. A. GUSTAFSON‡

Abstract. A multidimensional integral transformation is proved which is an $SU(n)$ integral analogue of Bailey's classical very well poised ${}_{10}\varphi_9$ hypergeometric series transformation. By applying Cauchy's theorem and specializing parameters, an $SU(n)$ ${}_{10}\varphi_9$ hypergeometric series transformation is then deduced. An $Sp(n)$ generalization of Jackson's very well poised ${}_8\varphi_7$ summation theorem is also proved.

Key words. q-beta integrals, integral transformation, hypergeometric series very well poised on Lie algebras, multiple hypergeometric series, Jackson's theorem

AMS(MOS) subject classifications. 33A15, 33A30, 33A65, 33A75

1. Introduction. In previous papers [4]–[6], the connection between summation theorems for hypergeometric series very well poised on semisimple Lie algebras and corresponding evaluations of multidimensional Mellin–Barnes and q-beta integrals on Lie groups and Lie algebras was developed. In particular, integral analogues of Gauss's sum, the Pfaff–Saalschütz sum, and Dougall's and Jackson's sum on various Lie groups and Lie algebras were found. These integrals generalize important one-dimensional integrals such as Barnes' first and second lemmas [3], the Askey–Wilson integral [1], the Nasrallah–Rahman integral [14], [15], and others. At the most basic level, the proofs of the simplest of these multidimensional integral identities depend on a use of Cauchy's theorem to express the integrals as sums of multiple series and then evaluation of these multiple series using a multidimensional form of the bilateralized Gauss's (${}_1H_1$) sum or a ${}_6\psi_6$ sum. At this point, the series depart from the scene and other methods such as difference equations are used to prove the integral identities. Just as with classical hypergeometric series, there is a hierarchy in the integral identities: a $u(n)$ Barnes' first lemma [5] is used to prove a $u(n)$ Barnes' second lemma [6]; the generalizations of the Askey–Wilson integral [5] are used to prove the corresponding generalization of the Nasrallah–Rahman integral [6].

In this paper we will come full circle, in the sense that integral identities will now be used to prove series identities. In §2 we will prove an integral analogue of the $SU(n)$ ${}_{10}\varphi_9$ transformation and then, in §3, we use this integral to deduce the corresponding $SU(n)$ ${}_{10}\varphi_9$ series identity. In §4 we use an $Sp(n)$ integral analogue of Jackson's ${}_8\varphi_7$ sum found in [6] to prove the corresponding $Sp(n)$ series identity. The methods used in this paper clearly have a wider application and could likely be used to give alternative proofs of the $u(n)$ Pfaff–Saalschütz series identities [7], the $SU(n)$ Jackson series identities [12], and other identities. It should be mentioned that the $SU(n)$ ${}_{10}\varphi_9$ series transformation found in §3 is probably equivalent to results of Milne [13], though the method of proof here is completely different.

One application of the integral and series identities, such as Theorems 2.1, 3.1, and 4.1, would be in the theory of orthogonal polynomials in several variables. It

* Received by the editors August 6, 1990; accepted for publication (in revised form) June 14, 1991.

† Department of Mathematics and Statistics, University of Gorakhpur, Gorakpur-273009, India.

‡ Department of Mathematics, Texas A&M University, College Station, Texas 77843. This research was partially supported by National Science Foundation grants INT-8713472 and DMS-9002342.

is possible that they will be useful in finding generating functions and reproducing kernels for some of Macdonald's polynomials [11]. For example, in the one-dimensional case, see Nasrallah and Rahman [14]. Another possible application is to look for finite field or p-adic analogues of Theorem 2.1 or the previous integral identities in [4]–[6]. For some beautiful one-dimensional finite field and p-adic analogues of Barnes' first lemma see [9]–[10]. These papers also give some clue to the group representation theoretic significance of these integral identities.

2. A multivariate integral transformation. In this section we prove an integral analogue of the $SU(n)$ ${}_{10}\varphi_9$ hypergeometric series transformation. The proof involves a double $SU(n)$ integral. Applying Fubini's theorem and the $SU(n)$ generalization of the Nasrallah–Rahman integral [15, eqn. (2.4)], the double integral can be partially evaluated in two different ways. The resulting identity is the desired transformation.

THEOREM 2.1. *For $n \geq 2$ let $a, b_i, c_j \in \boldsymbol{C}$ for $1 \leq i \leq n+1$ and $1 \leq j \leq n+1$, with $|a|, |b_i|, |c_j| < 1$. Set $B = \prod_{i=1}^{n+1} b_i$ and $C = \prod_{j=1}^{n+1} c_j$, then*

$$(2.2)\qquad \prod_{i=1}^{n+1} \frac{[a^n c_i^{-1} C]_\infty}{[c_i^{-1} C]_\infty} \int_{T^{n-1}} \frac{\prod_{j=1}^{n} \{[aCz_j^{-1}]_\infty [a^n B z_j]_\infty\} \prod_{\substack{1\leq i,j\leq n\\ i\neq j}} [z_i z_j^{-1}]_\infty}{\prod_{i=1}^{n+1} \prod_{j=1}^{n} \{[ac_i z_j^{-1}]_\infty [b_i z_j]_\infty\}} \frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}}$$

$$= \prod_{i=1}^{n+1} \frac{[a^n b_i^{-1} B]_\infty}{[b_i^{-1} B]_\infty} \int_{T^{n-1}} \frac{\prod_{j=1}^{n} \{[aBz_j^{-1}]_\infty [a^n C z_j]_\infty\} \prod_{\substack{1\leq i,j\leq n\\ i\neq j}} [z_i z_j^{-1}]_\infty}{\prod_{i=1}^{n+1} \prod_{j=1}^{n} \{[ab_i z_j^{-1}]_\infty [c_i z_j]_\infty\}} \frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}},$$

where $\prod_{k=1}^{n} z_k = 1$, and the integral in each variable $z_1, \cdots, z_{n-1}$ is over the unit circle T taken in the positive direction.

Proof. Consider the following integral:

$$(2.3)\qquad \frac{1}{(2\pi i)^{n-1}} \int_{T^{n-1}} \int_{T^{n-1}} \frac{\prod_{k=1}^{n} \{[a^n B z_k]_\infty [a^n C u_k^{-1}]_\infty\} \prod_{\substack{1\leq i,j\leq n\\ i\neq j}} \{[z_i z_j^{-1}]_\infty [u_i u_j^{-1}]_\infty\}}{\prod_{k=1}^{n} \prod_{j=1}^{n+1} \{[b_j z_k]_\infty [c_j u_k^{-1}]_\infty\} \prod_{i,k=1}^{n} [a z_k^{-1} u_i]_\infty}$$

$$\cdot \frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}} \frac{du_1}{u_1} \cdots \frac{du_{n-1}}{u_{n-1}},$$

where $\prod_{k=1}^{n} z_k = \prod_{k=1}^{n} u_k = 1$. If we evaluate the integral (2.3) with respect to the variables z_i, $1 \leq i \leq n$, or with respect to the variables u_i, $1 \leq i \leq n$, by means of the $SU(n)$ Nasrallah–Rahman integral [6, Thm. 2.1], we obtain the following identity:

(2.4)
$$\frac{n!\prod_{j=1}^{n+1}[a^nb_j^{-1}B]_\infty}{[q]_\infty^{n-1}[a^n]_\infty\prod_{j=1}^{n+1}[b_j^{-1}B]_\infty}\int_{T^{n-1}}\frac{\prod_{k=1}^{n}\{[aBu_k]_\infty[a^nCu_k^{-1}]_\infty\}}{\prod_{k=1}^{n}\prod_{j=1}^{n+1}\{[ab_ju_k]_\infty[c_ju_k^{-1}]_\infty\}}$$
$$\cdot\prod_{\substack{1\leq i,j\leq n\\ i\neq j}}[u_iu_j^{-1}]_\infty\frac{du_1}{u_1}\cdots\frac{du_{n-1}}{u_{n-1}}$$
$$=\frac{n!\prod_{j=1}^{n+1}[a^nc_j^{-1}C]_\infty}{[q]_\infty^{n-1}[a^n]_\infty\prod_{j=1}^{n+1}[c_j^{-1}C]_\infty}\int_{T^{n-1}}\frac{\prod_{k=1}^{n}\{[aCz_k^{-1}]_\infty[a^nBz_k]_\infty\}}{\prod_{k=1}^{n}\prod_{j=1}^{n+1}\{[ac_jz_k^{-1}]_\infty[b_jz_k]_\infty\}}$$
$$\cdot\prod_{\substack{1\leq i,j\leq n\\ i\neq j}}[z_iz_j^{-1}]_\infty\frac{dz_1}{z_1}\cdots\frac{dz_{n-1}}{z_{n-1}}.$$

Simplifying (2.4) and replacing the variables u_k by z_k^{-1}, $1\leq k\leq n$, on the left-hand side of (2.4), we obtain identity (2.2).

3. An $SU(n)$ ${}_{10}\varphi_9$ transformation. In this section we will prove an $SU(n)$ very well poised hypergeometric series transformation which correspronds to Theorem 2.1. The proof consists of writing both sides of identity (2.2) as sums of multiple series of residues and then specializing the parameters to make the series terminate. The $SU(2)$ case of this series transformation (3.2) is due to Bailey [2] (see Remark 3.7 below) and is one of the most powerful classical basic hypergeometric series identities. At the end of this section we will also state the limit $q\to 1^-$ of identity (3.2). We begin with the following theorem.

THEOREM 3.1. *For $n\geq 2$ let $a, b_j, c_j\in C$ for $1\leq j\leq n+1$, with $ab_ic_i=q^{-m_i}$ for some nonnegative integers m_i for $1\leq i\leq n-1$. Let $\gamma_i=ac_i$ and $w_i=ab_i$ for $1\leq i\leq n-1$, and $\gamma_n=\prod_{i=1}^{n-1}\gamma_i^{-1}$ and $w_n=\prod_{i=1}^{n-1}w_i^{-1}$. Set $B=\prod_{j=1}^{n+1}b_j$ and $C=\prod_{j=1}^{n+1}c_j$, then*

(3.2)
$$\prod_{i=1}^{n+1}\frac{[a^nc_i^{-1}C]_\infty}{[c_i^{-1}C]_\infty}\frac{\prod_{j=1}^{n}\{[aC\gamma_j^{-1}]_\infty[a^nB\gamma_j]_\infty]\}}{\prod_{i=1}^{n+1}\prod_{\substack{j=1\\ i\neq j}}^{n-1}\{[ac_i\gamma_j^{-1}]_\infty[b_i\gamma_j]_\infty\}}$$
$$\cdot\frac{\prod_{\substack{1\leq i,j\leq n\\ i\neq j}}[\gamma_i\gamma_j^{-1}]_\infty}{\prod_{i=1}^{n+1}\{[ac_i\gamma_n^{-1}]_\infty[b_i\gamma_n]_\infty\}}\sum_{\substack{y_1\geq 0,\cdots,y_{n-1}\geq 0\\ y_1+\cdots+y_n=0}}\left\{\prod_{1\leq i<j\leq n}\left(\frac{\gamma_iq^{y_i}-\gamma_jq^{y_j}}{\gamma_i-\gamma_j}\right)\right.$$
$$\left.\cdot\prod_{j=1}^{n}\frac{[q\gamma_j/aC]_{y_j}\prod_{i=1}^{n+1}[b_i\gamma_j]_{y_j}}{[a^nB\gamma_j]_{y_j}\prod_{i=1}^{n+1}[q\gamma_j/ac_i]_{y_j}}\right\}$$

$$= \prod_{i=1}^{n+1} \frac{[a^n b_i^{-1} B]_\infty}{[b_i^{-1} B]_\infty} \frac{\prod\limits_{j=1}^{n} \{[aBw_j^{-1}]_\infty [a^n C w_j]_\infty\}}{\prod\limits_{i=1}^{n+1} \prod\limits_{\substack{j=1 \\ i\neq j}}^{n-1} \{[ab_i w_j^{-1}]_\infty [c_i w_j]_\infty\}}$$

$$\cdot \frac{\prod\limits_{\substack{1\leq i,j\leq n \\ i\neq j}} [w_i w_j^{-1}]_\infty}{\prod\limits_{i=1}^{n+1} \{[ab_i w_n^{-1}]_\infty [c_i w_n]_\infty\}} \sum_{\substack{y_1\geq 0,\cdots,y_{n-1}\geq 0 \\ y_1+\cdots+y_n=0}} \Bigg\{ \prod_{1\leq i<j\leq n} \left(\frac{w_i q^{y_i} - w_j q^{y_j}}{w_i - w_j}\right)$$

$$\cdot \prod_{j=1}^{n} \frac{[qw_j/aB]_{y_j} \prod\limits_{i=1}^{n+1} [c_i w_j]_{y_j}}{[a^n C w_j]_{y_j} \prod\limits_{i=1}^{n+1} [qw_j/ab_i]_{y_j}} \Bigg\}.$$

Proof. Let $a\alpha = \{ac_1, ac_2, \cdots, ac_{n+1}\}$, $\alpha^{-1} = \{c_1^{-1}, c_2^{-1}, \cdots, c_{n+1}^{-1}\}$, $a\beta = \{ab_1, \cdots, ab_{n+1}\}$, and $\beta^{-1} = \{b_1^{-1}, \cdots, b_{n+1}^{-1}\}$. Let Δ be the set of all $\delta = (u_1, \cdots, u_n)$, where for some j, $0 \leq j \leq n-1$,

$$\begin{aligned} &\{u_1, \cdots, u_j\} \subset a\alpha, \\ &\{u_{j+1}, \cdots, u_{n-1}\} \subset \beta^{-1}, \\ &u_n = \prod_{i=1}^{n-1} u_i^{-1}, \end{aligned}$$

and

$$u_i \neq u_k \quad \text{for all} \quad i \neq k,\ 1 \leq i, k \leq n.$$

Similarly, let Δ' be the set of all $\delta' = (v_1, \cdots, v_n)$, where for some j, $0 \leq j \leq n-1$,

$$\begin{aligned} &\{v_1, \cdots, v_j\} \subset a\beta, \\ &\{v_{j+1}, \cdots, v_{n-1}\} \subset \alpha^{-1}, \\ &v_n = \prod_{i=1}^{n-1} v_i^{-1}, \end{aligned}$$

and

$$v_i \neq v_k \quad \text{for all} \quad i \neq k,\ 1 \leq i, k \leq n.$$

We will temporarily assume that $|a|, |b_i|, |c_i| < 1$ for all i, $1 \leq i \leq n+1$. Following the argument in [5, §6], upon expanding identity (2.2) in a series of residues and cancelling various terms, we obtain

(3.3)

$$\sum_{\delta\in\Delta}\left\{\prod_{i=1}^{n+1}\frac{[a^nc_i^{-1}C]_\infty}{[c_i^{-1}C]_\infty}\,\frac{\prod\limits_{j=1}^{n}[aCu_j^{-1}]_\infty[a^nBu_j]_\infty}{\prod\limits_{i=1}^{n+1}\prod\limits_{j=1}^{n}{}'[ac_iu_j^{-1}]_\infty[b_iu_j]_\infty}\right.$$

$$\cdot\prod_{\substack{1\le i,j\le n\\ i\neq j}}[u_iu_j^{-1}]_\infty\sum_{\substack{y_1,\cdots,y_n=-\infty\\ y_1+\cdots+y_n=0}}^{\infty}\left\{\prod_{1\le i<j\le n}\left(\frac{u_iq^{y_i}-u_jq^{y_j}}{u_i-u_j}\right)\right.$$

$$\left.\left.\cdot\prod_{j=1}^{n}\frac{[qu_j/aC]_{y_j}\prod\limits_{i=1}^{n+1}[b_iu_j]_{y_j}}{[a^nBu_j]_{y_j}\prod\limits_{i=1}^{n+1}[qu_j/ac_i]_{y_j}}\right\}\right\}$$

$$=\sum_{\delta'\in\Delta'}\left\{\prod_{i=1}^{n+1}\frac{[a^nb_i^{-1}B]_\infty}{[b_i^{-1}B]_\infty}\,\frac{\prod\limits_{j=1}^{n}[aBv_j^{-1}]_\infty[a^nCv_j]_\infty}{\prod\limits_{i=1}^{n+1}\prod\limits_{j=1}^{n}{}'[ab_iv_j^{-1}]_\infty[c_iv_j]_\infty}\right.$$

$$\cdot\prod_{\substack{1\le i,j\le n\\ i\neq j}}[v_iv_j^{-1}]_\infty\cdot\sum_{\substack{y_1,\cdots,y_n=-\infty\\ y_1+\cdots+y_n=0}}^{\infty}\left\{\prod_{1\le i<j\le n}\left(\frac{v_iq^{y_i}-v_jq^{y_j}}{v_i-v_j}\right)\right.$$

$$\left.\left.\cdot\prod_{j=1}^{n}\frac{[qv_j/aB]_{y_j}\prod\limits_{i=1}^{n+1}[c_iv_j]_{y_j}}{[a^nCv_j]_{y_j}\prod\limits_{i=1}^{n+1}[qv_j/ab_i]_{y_j}}\right\}\right\},$$

where $\prod'$ means the usual product except that if $c=q^{-\ell}$ for some nonnegative integer ℓ, then the factor $[c]_\infty$ in the product is replaced by $[q^{-\ell}]_\ell[q]_\infty$. Now multiply both sides of equation (3.3) by $\prod_{i=1}^{n-1}[ab_ic_i]_\infty$ and set $ab_ic_i=q^{-m_i}$, $1\le i\le n-1$, as in the statement of Theorem 3.1. The only terms not vanishing on the left-hand side of (3.3) will be when $\delta=(u_1,\cdots,u_n)$ satisfies

$$\{u_1,\cdots,u_j\}\subset\{ac_1,\cdots,ac_{n-1}\},$$
$$\{u_{j+1},\cdots,u_{n-1}\}\subset\{b_1^{-1},\cdots,b_{n-1}^{-1}\},$$

and

$$u_iu_j^{-1}\neq ac_\ell b_\ell\quad\text{for } 1\le i,j,\ell\le n-1.$$

Similarly, the only terms not vanishing on the right-hand side of (3.3) will be when $\delta'=(v_1,\cdots,v_n)$ satisfies

$$\{v_1,\cdots v_j\}\subset\{ab_1,\cdots,ab_{n-1}\},$$
$$\{v_{j+1},\cdots,v_{n-1}\}\subset\{c_1^{-1},\cdots,c_{n-1}^{-1}\},$$

and

$$v_iv_j^{-1}\neq ac_\ell b_\ell\quad\text{for } 1\le i,j,\ell\le n-1.$$

We now claim that all of the nonvanishing series on the left-hand side of (3.3) are equal. This can be proved by simply reversing the appropriate summations. For example, consider the series on the left-hand side of (3.3), where $\delta = (u_1, \cdots, u_n)$ and for some j, $0 \leq j \leq n-1$, we have $u_i = ac_i$ for $1 \leq i \leq j$ and $u_i = b_i^{-1}$ for $j < i \leq n-1$. Then we find

$$
\begin{aligned}
(3.4)\qquad & \prod_{i=1}^{n+1} \frac{[a^n c_i^{-1} C]_\infty}{[c_i^{-1} C]_\infty} \sum_{y_1=0}^{m_1} \cdots \sum_{y_j=0}^{m_j} \sum_{y_{j+1}=-m_{j+1}}^{0} \cdots \sum_{y_{n-1}=-m_{n-1}}^{0} \\
& \left\{ \prod_{\ell=1}^{j} \frac{[ab_\ell c_\ell]_{y_\ell}}{[q]_\infty [q^{-y_\ell}]_{y_\ell}} \prod_{\ell=j+1}^{n-1} \frac{[ab_\ell c_\ell]_{-y_\ell}}{[q]_\infty [q^{y_\ell}]_{-y_\ell}} \cdot \prod_{\substack{1 \leq i,\ell \leq n \\ i \neq \ell}} [u_i u_\ell^{-1}]_\infty \right. \\
& \left. \cdot \frac{\prod_{\ell=1}^{n} [aCu_\ell^{-1} q^{-y_\ell}]_\infty [a^n B u_\ell q^{y_\ell}]_\infty}{\prod_{\substack{i=1 \\ i \neq \ell}}^{n+1} \ \prod_{\substack{\ell=1 \\ \text{for } 1 \leq \ell \leq n-1}}^{n} [ac_i u_\ell^{-1} q^{-y_\ell}]_\infty [b_i u_\ell q^{y_\ell}]_\infty} \right\} \\
& = \prod_{i=1}^{n+1} \frac{[a^n c_i^{-1} C]_\infty}{[c_i^{-1} C]_\infty} \sum_{y_1=0}^{m_1} \cdots \sum_{y_{n-1}=0}^{m_{n-1}} \\
& \left\{ \prod_{\ell=1}^{n-1} \frac{[ab_\ell c_\ell]_{y_\ell}}{[q]_\infty [q^{-y_\ell}]_{y_\ell}} \cdot \prod_{\substack{1 \leq i,\ell \leq n \\ i \neq \ell}} [\gamma_i \gamma_\ell^{-1}]_\infty \right. \\
& \left. \cdot \frac{\prod_{\ell=1}^{n} [aC\gamma_\ell q^{-y_\ell}]_\infty [a^n B \gamma_\ell q^{y_\ell}]_\infty}{\prod_{\substack{i=1 \\ i \neq \ell}}^{n+1} \ \prod_{\substack{\ell=1 \\ \text{for } 1 \leq \ell \leq n-1}}^{n} [ac_i \gamma_\ell^{-1} q^{-y_\ell}]_\infty [b_i \gamma_\ell q^{y_\ell}]_\infty} \right\},
\end{aligned}
$$

where $y_1 + \cdots + y_n = 0, \gamma_i = ac_i$ for $1 \leq i \leq n-1$ and $\gamma_n = \prod_{i=1}^{n-1} \gamma_i^{-1}$. Similarly, we show that all of the nonvanishing series on the right-hand side of (3.3) are equal. Putting this together, we obtain identity (3.2). This completes the proof of Theorem 3.1.

Let $q \to 1^-$ in identity (3.2). We then obtain the following.

THEOREM 3.5. *For $n \geq 2$ let $a, b_j, c_j \in C$ for $1 \leq j \leq n+1$, with $a+b_i+c_i = -m_i$ for some nonnegative integer m_i for $1 \leq i \leq n-1$. Let $\gamma_i = a + c_i$ and $w_i = a + b_i$ for $1 \leq i \leq n-1$, and $\gamma_n = -\sum_{i=1}^{n-1} \gamma_i$ and $w_n = -\sum_{i=1}^{n-1} w_i$. Set $B = \sum_{j=1}^{n+1} b_j$ and $C = \sum_{j=1}^{n+1} c_j$, then*

(3.6)

$$\prod_{i=1}^{n+1}\frac{\Gamma(C-c_i)}{\Gamma(C-c_i+na)}\,\frac{\prod_{i=1}^{n+1}\prod_{\substack{j=1\\ i\neq j}}^{n-1}\{\Gamma(a+c_i-\gamma_j)\Gamma(b_i+\gamma_j)\}}{\prod_{j=1}^{n}\{\Gamma(a+C-\gamma_j)\Gamma(na+B+\gamma_j)\}}$$

$$\cdot\frac{\prod_{i=1}^{n+1}\{\Gamma(a+c_i-\gamma_n)\Gamma(b_i+\gamma_n)\}}{\prod_{\substack{1\leq i,j\leq n\\ i\neq j}}\Gamma(\gamma_i-\gamma_j)}$$

$$\cdot\sum_{\substack{y_1\geq 0,\cdots,y_{n-1}\geq 0\\ y_1+\cdots+y_n=0}}\left\{\prod_{1\leq i<j\leq n}\left(\frac{\gamma_i+y_i-\gamma_j-y_j}{\gamma_i-\gamma_j}\right)\right.$$

$$\left.\cdot\prod_{j=1}^{n}\frac{(1+\gamma_j-a-C)_{y_j}}{(na+B+\gamma_j)_{y_j}}\,\frac{\prod_{i=1}^{n+1}(b_i+\gamma_j)_{y_j}}{\prod_{i=1}^{n+1}(1+\gamma_j-a-c_j)_{y_j}}\right\}$$

$$=\prod_{i=1}^{n+1}\frac{\Gamma(B-b_i)}{\Gamma(na+B-b_i)}\,\frac{\prod_{i=1}^{n+1}\prod_{\substack{j=1\\ i\neq j}}^{n-1}\{\Gamma(a+b_i-w_j)\Gamma(c_i+w_j)\}}{\prod_{j=1}^{n}\{\Gamma(a+B-w_j)\Gamma(na+C+w_j)\}}$$

$$\cdot\frac{\prod_{i=1}^{n+1}\{\Gamma(a+b_i-w_n)\Gamma(c_i+w_n)\}}{\prod_{\substack{1\leq i,j\leq n\\ i\neq j}}\Gamma(w_i-w_j)}$$

$$\cdot\sum_{\substack{y_1\geq 0,\cdots,y_{n-1}\geq 0\\ y_1+\cdots+y_n=0}}\left\{\prod_{1\leq i<j\leq n}\left(\frac{w_i+y_i-w_j-y_j}{w_i-w_j}\right)\right.$$

$$\left.\cdot\prod_{j=1}^{n}\frac{(1+w_j-a-B)_{y_j}\prod_{i=1}^{n+1}(c_i+w_j)_{y_j}}{(na+C+w_j)_{y_j}\prod_{i=1}^{n+1}(1+w_j-a-b_i)_{y_j}}\right\}.$$

Remark 3.7 To show that the $n=2$ case of identity (3.2) is equivalent to Bailey's very well poised ${}_{10}\varphi_9$ transformation ([3], p. 68), set $ab_1c_1=q^{-m}$ in (3.2) and reverse the series on the right-hand side of (3.2). Set a in Bailey's notation equal to $a^2c_1^2$ here, and $c=a^2c_2c_1$, $d=a^2c_3c_1$, $e=qc_2^{-1}c_3^{-1}$, $f=ab_1c_1=q^{-m}$, $g=ab_2c_1$, $h=ab_3c_1$, $j=qa^{-1}B^{-1}c_1$, and $k=c_1^2$, where the notation on the left-hand side of each equality is Bailey's and on the right-hand side is that of (3.2) here. From (3.2), we then obtain the following transformation in Bailey's notation:

$$
\begin{aligned}
(3.8)\qquad & {}_{10}\varphi_9\left[\begin{matrix} a, & q\sqrt{a}, & -q\sqrt{a}, & c, & d, & e, \\ & \sqrt{a}, & -\sqrt{a}, & aq/c, & aq/d, & aq/e, \end{matrix}\right. \\
& \qquad \left.\begin{matrix} f, & g, & h, & j; & q \\ aq/f, & aq/g, & aq/h, & aq/j & \end{matrix}\right] \\
& = \frac{[f/a]_\infty, [g/a]_\infty [fgh/a]_\infty [h/a]_\infty}{[a^{-1}]_\infty [fg/a]_\infty [gh/a]_\infty [fh/a]_\infty} \\
& \cdot \frac{[qa/fj]_\infty [qa/gj]_\infty [qa/hj]_\infty [qa/fghj]_\infty}{[qa/j]_\infty [qa/ghj]_\infty [qa/hjf]_\infty [qa/jfg]_\infty} \\
& \cdot {}_{10}\varphi_9\left[\begin{matrix} k, & q\sqrt{k}, & -q\sqrt{k}, & kc/a, & kd/a, & ke/a, \\ & \sqrt{k}, & -\sqrt{k}, & aq/c, & aq/d, & aq/e, \end{matrix}\right. \\
& \qquad \left.\begin{matrix} f, & g, & h, & j; & q \\ kq/f, & kq/g, & kq/h, & kq/j & \end{matrix}\right],
\end{aligned}
$$

where $k = a^2q/cde$, $cdefghj = a^3q^2$, and $f = q^{-m}$ for some nonnegative integer m. Since $f = q^{-m}$, observe that

$$
\begin{aligned}
(3.9)\qquad & \frac{[f/a]_\infty [g/a]_\infty [fgh/a]_\infty [h/a]_\infty}{[a^{-1}]_\infty [fg/a]_\infty [gh/a]_\infty [fh/a]_\infty} \\
& = \frac{[a^{-1}q^{-m}]_m [gha^{-1}q^{-m}]_m}{[ga^{-1}q^{-m}]_m [ha^{-1}q^{-m}]_m} = \frac{[qa]_m [qa/gh]_m}{[qa/g]_m [qa/h]_m} \\
& = \frac{[qa]_\infty [qa/fg]_\infty [aq/fh]_\infty [qa/gh]_\infty}{[qa/f]_\infty [qa/g]_\infty [qa/h]_\infty [qa/fgh]_\infty}.
\end{aligned}
$$

Using identity (3.9) to substitute for the first part of the product on the it right-hand side of (3.8), we obtain Bailey's ${}_{10}\varphi_9$ transformation [3, p. 68].

4. A Jackson summation theorem for hypergeometric series on $Sp(n)$. In this section we generalize Jackson's summation theorem [8] for very well poised ${}_8\varphi_7$ hypergeometric series to the setting of basic hypergeometric series very well poised on $Sp(n)$. The $Sp(1)$ case reduces to the classical Jackson sum. The proof is very similar to that in §3, except that the starting point is the $Sp(n)$ integral identity [6, Thm. 4.1].

THEOREM 4.1. *For $n \geq 1$, let $a_i \in C$ for $1 \leq i \leq 2n+3$, with $a_j a_{n+j} = q^{-m_j}$ for some nonnegative integers m_j for $1 \leq j \leq n$. Set $C = \prod_{i=1}^{2n+3} a_i$. Then*

$$
\begin{aligned}
(4.2)\qquad \sum_{y_1\geq 0,\cdots,y_n\geq 0} \Bigg\{ & q^{-\sum_{i=1}^{n}(n+1-i)y_i} \prod_{j=1}^{n} \frac{(1-a_j^2 q^{2y_j})}{(1-a_j^2)} \\
& \cdot \prod_{1\leq i<j\leq n} \frac{(1-a_i a_j^{-1} q^{y_i-y_j})(1-a_i a_j q^{y_i+y_j})}{(1-a_i a_j^{-1})(1-a_i a_j)}
\end{aligned}
$$

$$\left.\cdot\prod_{j=1}^{n}\frac{\prod\limits_{i=n+2}^{2n+3}[a_ia_j]_{y_j}[a_ia_j^{-1}]_{-y_j}}{[Ca_j]_{y_j}[Ca_j^{-1}]_{-y_j}\prod\limits_{i=1}^{n+1}[qa_i^{-1}a_j]_{y_j}[qa_i^{-1}a_j^{-1}]_{-y_j}}\right\}$$

$$=\frac{\prod\limits_{i=n+1}^{2n+3}\{[Ca_i^{-1}]_\infty\prod\limits_{j=1}^{n}[a_ia_j^{-1}]_\infty\}}{\prod\limits_{n+1\leq i<j\leq 2n+3}[a_ia_j]_\infty\prod\limits_{1\leq i\leq j\leq n}[a_i^{-1}a_j^{-1}]_\infty\prod\limits_{i=1}^{n}[Ca_i]_\infty}.$$

Proof. Let us temporarily assume that $|a_i| < 1$ for $1 \leq i \leq 2n+3$. From [6, Thm. 4.1] we have the following identity:

$$\frac{1}{(2\pi i)^n}\int_{T^n}\frac{\prod\limits_{1\leq i<j\leq n}[z_iz_j]_\infty[z_i^{-1}z_j^{-1}]_\infty[z_iz_j^{-1}]_\infty[z_i^{-1}z_j]_\infty}{\prod\limits_{i=1}^{2n+3}\prod\limits_{j=1}^{n}[a_iz_j]_\infty[a_iz_j^{-1}]_\infty}$$

$$\cdot\prod_{j=1}^{n}\left\{[z_j^2]_\infty[z_j^{-2}]_\infty[Cz_j]_\infty[Cz_j^{-1}]_\infty\frac{dz_j}{z_j}\right\}=\frac{n!2^n\prod\limits_{j=1}^{2n+3}[Ca_j^{-1}]_\infty}{[q]_\infty^n\prod\limits_{1\leq i<j\leq 2n+3}[a_ia_j]_\infty},\tag{4.3}$$

where the integral in each variable z_j is over the unit circle T taken in the positive direction.

Let S be the set of all injective mappings π: $\{1,2,\cdots,n\}\to\{1,2,\cdots,2n+3\}$. Using Cauchy's theorem and following the argument in §7 of [5], we can rewrite (4.3) as

$$\sum_{\pi\in S}\sum_{y_1,\cdots,y_n=0}^{\infty}\left\{\frac{\prod\limits_{1\leq i\leq j\leq n}[a_{\pi(i)}a_{\pi(j)}q^{y_i+y_j}]_\infty[a_{\pi(i)}^{-1}a_{\pi(j)}^{-1}q^{-y_i-y_j}]_\infty}{\prod\limits_{i=1}^{2n+3}\prod\limits_{j=1}^{n}{}'[a_ia_{\pi(j)}q^{y_j}]_\infty[a_ia_{\pi(j)}^{-1}q^{-y_j}]_\infty}\right.$$

$$\left.\cdot\prod_{\substack{1\leq i,j\leq n\\ i\neq j}}[a_{\pi(i)}a_{\pi(j)}^{-1}q^{y_i-y_j}]_\infty\prod_{j=1}^{n}[Ca_{\pi(j)}q^{y_j}]_\infty[Ca_{\pi(j)}^{-1}q^{-y_j}]_\infty\right\}\tag{4.4}$$

$$=\frac{n!2^n\prod\limits_{j=1}^{2n+3}[Ca_j^{-1}]_\infty}{[q]_\infty^n\prod\limits_{1\leq i<j\leq 2n+3}[a_ia_j]_\infty},$$

where $\prod'$ is defined as in (3.3). Now multiply both sides of equation (4.4) by $\prod_{j=1}^{n}[a_ja_{n+j}]_\infty$ and set $a_ja_{n+j}=q^{-m_j}$. The only terms not vanishing on the left-hand side of (4.4) will be when the image of $\pi\subset\{1,\cdots,2n\}$, and if $j\in$ image of π, then $n+j\notin$ image of π for all j, $1\leq j\leq n$. An argument similar to that in §3 shows that on the left-hand side of (4.4), the corresponding series in $y_1,\cdots y_n$ for all such π are equal. Identity (4.2) and Theorem 4.1 now follow from (4.4) after simplification.

We also have the following result when $q \to 1^-$ in (4.2).

THEOREM 4.5. *For $n \geq 1$, let $a_i \in C$ for $1 \leq i \leq 2n+3$, with $a_j + a_{n+j} = -m_j$ for some nonnegative integers m_j for $1 \leq j \leq n$. Set $C = \sum_{i=1}^{2n+3} a_i$. Then*

(4.6)

$$\sum_{y_1\geq 0,\cdots,y_n\geq 0}\left\{\prod_{j=1}^{n}\left(\frac{a_j+y_j}{a_j}\right)\cdot\prod_{1\leq i<j\leq n}\frac{(a_i+y_i-a_j-y_j)(a_i+y_i+a_j+y_j)}{(a_i-a_j)(a_i+a_j)}\right.$$

$$\left.\cdot\prod_{j=1}^{n}\frac{\prod\limits_{i=n+2}^{2n+3}(a_i+a_j)_{y_j}(a_i-a_j)_{-y_j}}{(C+a_j)_{y_j}(C-a_j)_{-y_j}\prod\limits_{i=1}^{n+1}(1-a_i+a_j)_{y_j}(1-a_i-a_j)_{-y_j}}\right\}$$

$$=\frac{\prod\limits_{n+1\leq i<j\leq 2n+3}\Gamma(a_i+a_j)\prod\limits_{1\leq i\leq j\leq n}\Gamma(-a_i-a_j)\prod\limits_{i=1}^{n}\Gamma(C+a_i)}{\prod\limits_{i=n+1}^{2n+3}\left\{\Gamma(C-a_i)\prod\limits_{j=1}^{n}\Gamma(a_i-a_j)\right\}}.$$

REFERENCES

[1] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985), pp. 55.

[2] W. N. BAILEY, *An identity involving Heine's basic hypergeometric series*, Journal London Math. Soc., 4 (1929), pp. 254–257.

[3] ———, *Generalized Hypergeometric Series*, Cambridge Math. Tract, 32, Cambridge University Press, 1935; reprinted; Hafner, New York, 1954.

[4] R. A. GUSTAFSON, *A generalization of Selberg's beta integral*, Bul. Amer. Math. Soc., 22 (1990), pp. 97–105.

[5] ———, *Some q-beta and Mellin–Barnes integrals on compact Lie groups and Lie algebras*, preprint.

[6] ———, *Some q-beta and Mellin–Barnes integrals with many parameters associated to the classical groups*, SIAM J. Math. Anal., this issue (1992), pp. 525–551.

[7] W. J. HOLMAN III, *Summation theorems for hypergeometric series in $U(n)$*, SIAM J. Math. Anal., 11 (1980), pp. 523–532.

[8] F. H. JACKSON, *Summation of q-hypergeometric series*, Messenger of Math., 47 (1917), pp. 101–112.

[9] W.-C. W. LI AND J. SOTO-ANDRADE, *Barnes identities and representations of $GL(2)$, I. Finite field case*, J. Reine Angew. Math., 344 (1983), pp. 171–179.

[10] W.-C. W. LI, *Barnes' identities and representations of $GL(2)$ II. Non-Archimedean local field case*, J. Reine Angew. Math., 345 (1983), pp. 69–92.

[11] I. G. MACDONALD, *Orthogonal polynomials associated with root systems*, preprint.

[12] S. C. MILNE, *Multiple q-series and $U(n)$ generalizations of Ramanujan's ${}_1\psi_1$ sum,* in Ramanujan Revisited, G. Andrews et al., eds., Academic Press, New York, 1988, pp. 473–524.

[13] ———, private communication.

[14] B. NASRALLAH AND M. RAHMAN, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, SIAM J. Math. Anal., 16 (1985), pp. 186–197.

[15] M. RAHMAN, *An integral representation of a ${}_{10}\varphi_9$ and continuous bi-orthogonal ${}_{10}\varphi_9$ rational functions*, Canad. J. Math., 38 (1986), pp. 601–618.

SIAM. J. MATH. ANAL.
Vol. 23, No. 2, pp. 562–569, March 1992

014

SOME FORMULAS OF RAMANUJAN, REVISITED*

EMILIO MONTALDI† AND GIUSEPPE ZUCCHELLI‡

Abstract. An alternative form of a famous result of Ramanujan is given, which is quoted as Entry 29(b) in Chapter 10 of Berndt's [*Ramanujan's Notebooks*, Part II, Springer-Verlag, New York, 1988]. Other results of a similar kind are reconsidered and generalized.

Key words. elliptic integrals, hypergeometric functions

AMS(MOS) subject classifications. 33A25, 33A30

1. In his first letter to Hardy, dated January 16, 1913, Ramanujan [1] communicated the formula

$$\frac{1}{n+1}\,{}_3F_2\left[\begin{matrix}\frac12, & \frac12, n+1\\ 1, & n+2\end{matrix}\right]=\frac{(n!)^2}{\Gamma^2(n+\frac32)}\sum_{r=0}^{n}\frac{(\frac12)_r^2}{(r!)^2} \tag{1.1}$$

which, together with certain generalizations, gave rise to a flurry of papers in the years 1929–1931. This result is quoted as Entry 29(b) in Chapter 10 of Berndt's [2] second book, where the related formulas

$$\frac{\pi^2}{4}\varphi\left(n+\frac12\right)=2G+\sum_{r=0}^{n-1}\frac{(r!)^2}{(\frac32)_r^2} \tag{1.2}$$

and

$$2\varphi\left(n+\frac14\right)=1+\frac{16\Gamma^4(\frac34)}{\pi^3}\sum_{r=0}^{n-1}\frac{(\frac34)_r^2}{(\frac54)_r^2} \tag{1.3}$$

are also discussed. In (1.2) and (1.3), $\varphi(z)$ is defined by

$$\varphi(z)=\frac{\Gamma^2(z+\frac12)}{\Gamma(z)\Gamma(z+1)}\,{}_3F_2\left[\begin{matrix}\frac12, & \frac12, z\\ 1, & z+1\end{matrix}\right] \tag{1.4}$$

for all complex z; moreover,

$$G=\sum_{r=0}^{\infty}\frac{(-1)^r}{(2r+1)^2} \tag{1.5}$$

is Catalan's constant.

As far as we know, the most recent proof of (1.1) and (1.2) is due to Dutka [3], who performed two distinct evaluations of the integral $\int_0^1 x^m\,{}_2F_1(\frac12,\frac12;1;x^2)\,dx$, m being a nonnegative integer.

In this note, we first derive an alternative expression for the finite sum on the right side of (1.1). This is done in § 2. In § 3, we generalize Dutka's procedure and obtain some (perhaps new) formulas for ${}_3F_2$ and ${}_4F_3$.

* Received by the editors August, 1989; accepted for publication September 11, 1990.

† Dipartimento di Fisica, Università di Milano, Istituto Nazionale di Fisica Nucleare, Sezione di Milano, Via Celoria 16, 20133 Milano, Italy.

‡ Centro Consiglio Nazionale delle Ricerche, Dipartimento di Biologia, Università di Milano, Via Celoria 26, 20133 Milano, Italy.

2. Let us consider the integral

$$\begin{aligned}
I_n(\alpha) &\equiv \frac{1}{4}\frac{(\frac{3}{2})_n^2}{(n!)^2}\int_0^1 t^{\alpha+n-(1/2)}(1-t)^{-1/2}\,dt\int_0^1 x^n(1-tx)^{-1/2}\,dx\\
&= \frac{1}{4}\frac{(\frac{3}{2})_n^2}{(n!)^2}\int_0^1 t^{\alpha-(1/2)}(1-t)^{-1/2}\,dt\int_0^1 (1-tx)^{-1/2}\sum_{r=0}^{n}(-1)^r\binom{n}{r}(1-xt)^r\,dx\\
&= \frac{1}{4}\frac{(\frac{3}{2})_n^2}{(n!)^2}\int_0^1 t^{\alpha-(3/2)}\sum_{r=0}^{n}(-1)^r\binom{n}{r}\frac{(1-t)^{-1/2}-(1-t)^r}{r+\frac{1}{2}}\,dt\\
&= \frac{1}{4}\frac{(\frac{3}{2})_n^2}{(n!)^2}\Gamma\left(\alpha-\frac{1}{2}\right)\sum_{r=0}^{n}(-1)^r\frac{\binom{n}{r}}{r+\frac{1}{2}}\left[\frac{\pi^{1/2}}{\Gamma(\alpha)}-\frac{r!}{\Gamma(\alpha+r+\frac{1}{2})}\right].
\end{aligned}\tag{2.1}$$

By putting $\alpha=-n$, we get

$$\frac{\pi}{4}\frac{(\frac{3}{2})_n^2}{(n!)^2}\int_0^1 x^n\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;x\right)dx=\frac{(\frac{3}{2})_n}{n!}\sum_{r=0}^{n}\frac{(\frac{1}{2})_r}{r!}\frac{1}{2(n-r)+1}.\tag{2.2}$$

Now, by recalling Ramanujan's formula (1.1),

$$\begin{aligned}
\frac{\pi}{4}\frac{(\frac{3}{2})_n^2}{(n!)^2}\int_0^1 x^n\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;x\right)dx &= \frac{\pi}{4}\frac{(\frac{3}{2})_n^2}{(n!)^2}\frac{1}{n+1}\,{}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}, n+1\\ 1, & n+2\end{matrix}\right]\\
&= \sum_{r=0}^{n}\frac{(\frac{1}{2})_r^2}{(r!)^2}.
\end{aligned}\tag{2.3}$$

Hence, by comparing (2.2) and (2.3), we obtain the identity

$$\frac{(\frac{3}{2})_n}{n!}\sum_{r=0}^{n}\frac{(\frac{1}{2})_r}{r!}\frac{1}{2(n-r)+1}=\sum_{r=0}^{n}\frac{(\frac{1}{2})_r^2}{(r!)^2}.\tag{2.4}$$

An alternative proof of (2.4) may be of some interest. To this aim, we start from

$$(1-x)^{-1/2}={}_2F_1\left(1,\frac{1}{2};1;x\right)=\frac{1}{\pi}\int_0^1 t^{-1/2}(1-t)^{-1/2}(1-xt)^{-1}\,dt,$$

i.e.,

$$\frac{d}{dx}\frac{1}{2}\log\frac{1+\sqrt{x}}{1-\sqrt{x}}=\frac{1}{\pi}\int_0^1 t^{-1/2}(1-t)^{-1}\frac{\partial}{\partial x}\tan^{-1}\left(\sqrt{\frac{x(1-t)}{1-x}}\right)dt,$$

whence

$$\frac{1}{2}\log\frac{1+\sqrt{x}}{1-\sqrt{x}}=\frac{1}{\pi}\int_0^1 t^{-1/2}(1-t)^{-1}\tan^{-1}\left(\sqrt{\frac{x(1-t)}{1-x}}\right)dt.\tag{2.5}$$

By observing that

$$\begin{aligned}
&\int_0^1 (1-xu)^{-1}(1-u)^{-1/2}(1-xtu)^{-1/2}\,du\\
&\qquad=2[x(1-x)(1-t)]^{-1/2}\tan^{-1}\left(\sqrt{\frac{x(1-t)}{1-x}}\right),
\end{aligned}$$

(2.5) becomes

$$
\begin{aligned}
&x^{-1/2}(1-x)^{-1/2}\log\frac{1+\sqrt{x}}{1-\sqrt{x}}\\
&\quad=\frac{1}{\pi}\int_0^1 t^{-1/2}(1-t)^{-1/2}\,dt\\
&\qquad\cdot\int_0^1 (1-xu)^{-1}(1-u)^{-1/2}(1-xtu)^{-1/2}\,du\\
&\quad=\int_0^1 (1-u)^{-1/2}(1-xu)^{-1}\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;xu\right)du,
\end{aligned}
\tag{2.6}
$$

and, by comparing the coefficients of x^n on both sides, the required identity (2.4) follows at once. Incidentally, this derivation of (2.4) gives, together with (2.2), another proof of Ramanujan identity (1.1).

It may be noted that, with $\delta=x(d/dx)$, (2.6) is equivalent to

$$x^{-1/2}(1-x)^{-1/2}\log\frac{1+\sqrt{x}}{1-\sqrt{x}}=B\left(\delta+1,\frac{1}{2}\right)(1-x)^{-1}\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;x\right). \tag{2.7}$$

This follows from the fact that $u^{x(d/dx)}=u^\delta$ has the operational property $u^\delta f(x)=f(ux)$, $f(x)$ being analytic in a neighborhood of $x=0$. By using

$$\frac{\Gamma(\delta+\frac{3}{2})}{\Gamma(\delta+1)\sqrt{\pi}}=\frac{1}{2\pi}\int_0^1 t^{-1/2}(1-t)^{-3/2}(1-t^{1+\delta})\,dt, \tag{2.8}$$

(2.7) is formally inverted to give (with $u(x)=x^{-1/2}(1-x)^{-1/2}\log(1+\sqrt{x}/1-\sqrt{x})$

$${}_2F_1\left(\frac{1}{2},\frac{1}{2};1;x\right)=\frac{1-x}{\pi}\int_0^1 t^{-1/2}(1-t)^{-3/2}[u(x)-tu(xt)]\,dt. \tag{2.9}$$

This is a rather unusual expression for the complete elliptic integral of the first kind.

Equation (2.9) can also be proven directly from (2.4) and (2.8). Indeed, by observing that

$$
\begin{aligned}
&u(x)=2\sum_{n=0}^{\infty}x^n\sum_{r=0}^{n}\frac{(\frac{1}{2})_r}{r!}\frac{1}{2(n-r)+1}=2\sum_{n=0}^{\infty}x^n\frac{n!}{(\frac{3}{2})_n}c_n,\\
&c_n=\sum_{r=0}^{n}\frac{(\frac{1}{2})_r^2}{(r!)^2},
\end{aligned}
\tag{2.10}
$$

the right side is

$$
\begin{aligned}
&\frac{1-x}{\pi}\sum_{n=0}^{\infty}x^n\frac{n!}{(\frac{3}{2})_n}c_n\int_0^1 dt\,t^{-1/2}(1-t)^{-3/2}(1-t^{n+1})\\
&\quad=(1-x)\sum_{n=0}^{\infty}c_nx^n\\
&\quad=1+\sum_{n=0}^{\infty}(c_{n+1}-c_n)x^{n+1}={}_2F_1\left(\frac{1}{2},\frac{1}{2};1;x\right),
\end{aligned}
$$

since

$$c_{n+1}-c_n=\left(\frac{(\frac{1}{2})_{n+1}}{(n+1)!}\right)^2.$$

This gives another reason why (2.4) is a useful formula.

3. Let us consider

$$A_\alpha \equiv \int_0^1 x^\alpha K(x)\,dx, \tag{3.1}$$

$$B_\alpha \equiv \int_0^1 x^\alpha E(x)\,dx, \tag{3.2}$$

where

$$K(x) = \int_0^1 (1-t^2)^{-1/2}(1-x^2t^2)^{-1/2}\,dt, \tag{3.3}$$

$$E(x) = \int_0^1 (1-t^2)^{-1/2}(1-x^2t^2)^{1/2}\,dt \tag{3.4}$$

are the complete elliptic integrals of the first and second kind, respectively.

From the well-known formulas [4]

$$x(1-x^2)K'(x) = E(x)-(1-x^2)K(x), \tag{3.5}$$

$$xE'(x) = E(x)-K(x), \tag{3.6}$$

we easily obtain

$$\begin{aligned} B_\alpha-(A_\alpha-A_{\alpha+2}) &= \int_0^1 x^{\alpha+1}(1-x^2)K'(x)\,dx \\ &= -\int_0^1 K(x)[(\alpha+1)x^\alpha-(\alpha+3)x^{\alpha+2}]\,dx \\ &= -(\alpha+1)A_\alpha+(\alpha+3)A_{\alpha+2}, \end{aligned}$$

i.e.,

$$(\alpha+2)A_{\alpha+2}-\alpha A_\alpha = B_\alpha \tag{3.7}$$

and

$$\begin{aligned} B_\alpha - A_\alpha &= \int_0^1 x^{\alpha+1}E'(x)\,dt = 1-(\alpha+1)\int_0^1 x^\alpha E(x)\,dx \\ &= 1-(\alpha+1)B_\alpha, \end{aligned}$$

i.e.,

$$(\alpha+2)B_\alpha = 1+A_\alpha. \tag{3.8}$$

Equations (3.7) and (3.8) imply

$$(\alpha+2)^2A_{\alpha+2}-(\alpha+1)^2A_\alpha = 1, \tag{3.9}$$

$$(\alpha+2)(\alpha+4)B_{\alpha+2}-(\alpha+1)^2B_\alpha = 2. \tag{3.10}$$

By introducing

$$C_\alpha \equiv \left[\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+\frac{1}{2})}\right]^2 A_{2\alpha} = \frac{\pi}{4}\frac{\Gamma^2(\alpha+1)}{\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha+\frac{3}{2})}\,{}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}, \alpha+\frac{1}{2} \\ 1, & \alpha+\frac{3}{2}\end{matrix}\right] \tag{3.11}$$

and

$$\begin{aligned} D_\alpha &\equiv \frac{\Gamma(\alpha+1)\Gamma(\alpha+2)}{\Gamma^2(\alpha+\frac{1}{2})} B_{2\alpha} \\ &= \frac{\pi}{4} \frac{\Gamma(\alpha+1)\Gamma(\alpha+2)}{\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha+\frac{3}{2})} {}_3F_2\left[\begin{matrix} -\frac{1}{2}, & \frac{1}{2}, \alpha+\frac{1}{2} \\ 1, & \alpha+\frac{3}{2} \end{matrix}\right], \end{aligned} \tag{3.12}$$

it follows that

$$C_{\alpha+1} - C_\alpha = \frac{1}{4}\left[\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+\frac{3}{2})}\right]^2 \tag{3.13}$$

and

$$D_{\alpha+1} - D_\alpha = \frac{1}{2} \frac{\Gamma(\alpha+1)\Gamma(\alpha+2)}{\Gamma^2(\alpha+\frac{3}{2})}, \tag{3.14}$$

whence ($n = 0, 1, 2, \cdots$)

$$C_{\alpha+n} = C_\alpha + \frac{1}{4}\sum_{r=0}^{n-1}\left[\frac{\Gamma(\alpha+r+1)}{\Gamma(\alpha+r+\frac{3}{2})}\right]^2 \tag{3.15}$$

and

$$D_{\alpha+n} = D_\alpha + \frac{1}{2}\sum_{r=0}^{n-1} \frac{\Gamma(\alpha+r+1)\Gamma(\alpha+r+2)}{\Gamma^2(\alpha+r+\frac{3}{2})}. \tag{3.16}$$

By observing that [5]

$$C_{1/2} = \frac{\pi}{4}, \quad C_0 = \frac{2G}{\pi}, \quad C_{-1/4} = \frac{\pi}{8}, \tag{3.17}$$

it is easy to obtain (1.1), (1.2), and (1.3) from (3.15). Our results (3.15) and (3.16) are an interesting generalization of Ramanujan's formulas.

With $\alpha = 1/2$, (3.16) becomes

$$D_{n+(1/2)} = \frac{\pi}{4}\sum_{r=0}^{n} \frac{(\frac{1}{2})_r(\frac{3}{2})_r}{(r!)^2}. \tag{3.18}$$

On the other hand, we have (cf. the derivation of (2.1))

$$\begin{aligned} D_{n+(1/2)} &= \frac{1}{4} \frac{\Gamma(n+\frac{3}{2})\Gamma(n+\frac{5}{2})}{(n!)^2} \lim_{\alpha\to -n} \int_0^1 t^{\alpha-(1/2)}(1-t)^{-1/2}\,dt \int_0^1 (tx)^n(1-tx)^{1/2}\,dx \\ &= \frac{3\pi}{8} \frac{(\frac{5}{2})_n}{n!} \sum_{r=0}^{n} \frac{1}{(n-r)!} \frac{r+1}{r+\frac{3}{2}} \left(-\frac{1}{2}\right)_{n-r}. \end{aligned} \tag{3.19}$$

Hence, we get the identity

$$\sum_{r=0}^{n} \frac{(\frac{1}{2})_r(\frac{3}{2})_r}{(r!)^2} = \frac{3}{2} \frac{(\frac{5}{2})_n}{n!} \sum_{r=0}^{n} \frac{1}{(n-r)!} \frac{r+1}{r+\frac{3}{2}} \left(-\frac{1}{2}\right)_{n-r}, \tag{3.20}$$

equivalent to the integral formula

$$\begin{aligned} &\frac{1}{x}(1-x)^{-1/2} - \frac{1}{2} x^{-3/2}(1-x)^{1/2} \log\frac{1+\sqrt{x}}{1-\sqrt{x}} \\ &= \int_0^1 (1-t)^{1/2}(1-xt)^{-2} {}_2F_1\left(-\frac{1}{2}, \frac{1}{2}; 1; xt\right) dt \\ &= B\left(\delta+1, \frac{3}{2}\right)(1-x)^{-2} {}_2F_1\left(-\frac{1}{2}, \frac{1}{2}; 1; x\right), \end{aligned} \tag{3.21}$$

which, by using

$$\frac{\Gamma(\delta+\frac{5}{2})}{\Gamma(\delta+1)\Gamma(\frac{3}{2})}=\frac{3}{2\pi}\int_0^1 t^{1/2}(1-t)^{-5/2}\left[t^{1+\delta}-1+(1+\delta)\frac{1-t}{t}\right]dt, \tag{3.22}$$

can be formally inverted to give

$$\begin{aligned}{}_2F_1\left(-\frac{1}{2},\frac{1}{2};1;x\right)&=\frac{3}{2\pi}(1-x)^2\int_0^1 t^{1/2}(1-t)^{-5/2}\\&\quad\cdot\left[tw(xt)-w(x)+\frac{1-t}{t}\left(1+x\frac{d}{dx}\right)w(x)\right]dt,\end{aligned} \tag{3.23}$$

$$\left(w(x)=\frac{1}{x}(1-x)^{-1/2}-\frac{1}{2}x^{-3/2}(1-x)^{1/2}\log\left(\frac{1+\sqrt{x}}{1-\sqrt{x}}\right)\right).$$

Some final remarks are in order. First, we note that

$${}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}, \alpha+n\\ 1, & \alpha+n+1\end{matrix}\right]=\frac{1}{\pi}\int_0^1 x^{-1/2}(1-x)^{-1/2}{}_2F_1\left(\frac{1}{2},\alpha+n;\alpha+n+1;x\right)dx.$$

Now [6, (25)]

$$\begin{aligned}{}_2F_1\left(\frac{1}{2},\alpha+n;\alpha+n+1;x\right)&=\frac{(\alpha+1)_n}{(\alpha+\frac{1}{2})_n}\sum_{r=0}^n(-1)^r\binom{n}{r}\frac{(\frac{1}{2})_r}{(\alpha+1)_r}(1-x)^r\\&\quad\cdot{}_2F_1\left(\alpha+r,\frac{1}{2}+r;\alpha+r+1;x\right),\end{aligned} \tag{3.24}$$

and thus

$$\begin{aligned}{}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}, \alpha+n\\ 1, & \alpha+n+1\end{matrix}\right]&=\frac{(\alpha+1)_n}{(\alpha+\frac{1}{2})_n}\sum_{r=0}^n(-1)^r\binom{n}{r}\frac{(\frac{1}{2})_r}{(\alpha+1)_r r!}\\&\quad\cdot{}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}+r, \alpha+r\\ r+1, & \alpha+r+1\end{matrix}\right].\end{aligned} \tag{3.25}$$

This formula—which, for $\alpha=1$, reduces to (2.4)—enables us to express

$${}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}+n, \alpha+n\\ n+1, & \alpha+n+1\end{matrix}\right]$$

in terms of $C_{\alpha+r-(1/2)}$, $r=0,1,\cdots,n$.

Next, we have

$$\begin{aligned}\int_0^1 x^{\alpha-1}{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;x\right)dx&=\frac{1}{\pi}\int_0^1 t^{(1/2)-\alpha}(1-t)^{-1/2}\,dt\int_0^1(tx)^{\alpha-1}(1-tx)^{-1/2}\,dx\\&=\frac{1}{\pi}\int_0^1 t^{-(1/2)-\alpha}(1-t)^{-1/2}\,dt\int_0^t x^{\alpha-1}(1-x)^{-1/2}\,dx\\&=\frac{1}{\pi}\int_0^1 x^{\alpha-1}(1-x)^{-1/2}\,dx\\&\quad\cdot\left[\int_0^1 t^{-(1/2)-\alpha}(1-t)^{-1/2}\,dt-\int_0^x t^{-(1/2)-\alpha}(1-t)^{-1/2}\,dt\right],\end{aligned}$$

that is, by expanding $(1-t)^{-1/2}$ in a binomial series, and integrating term by term [7, (2)],

$$\frac{1}{\alpha}\,{}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}, \alpha\\ 1, & \alpha+1\end{matrix}\right]=\frac{\Gamma(\alpha)\Gamma(\frac{1}{2}-\alpha)}{\Gamma(1-\alpha)\Gamma(\frac{1}{2}+\alpha)}-\sum_{r=0}^{\infty}\frac{(\frac{1}{2})_r^2}{(r!)^2}\frac{1}{r-\alpha+\frac{1}{2}}. \tag{3.26}$$

Let us put $\alpha=n+(1/2)$, $n=0,1,2,\cdots$. A straightforward calculation shows that (as usual, $\Psi(z)=(d/dz)\log\Gamma(z)$)

$$\lim_{\alpha\to n+1/2}\left[\frac{\Gamma(\alpha)\Gamma(\frac{1}{2}-\alpha)}{\Gamma(1-\alpha)\Gamma(\frac{1}{2}+\alpha)}-\frac{(\frac{1}{2})_n^2}{(n!)^2}\frac{1}{n-\alpha+\frac{1}{2}}\right]=2\frac{(\frac{1}{2})_n^2}{(n!)^2}\left[\Psi(n+1)-\Psi\left(n+\frac{1}{2}\right)\right]$$

and thus

$$\frac{1}{n+\frac{1}{2}}\,{}_3F_2\left[\begin{matrix}\frac{1}{2}, & \frac{1}{2}, n+\frac{1}{2}\\ 1, & n+\frac{3}{2}\end{matrix}\right]=-\sum_{r=0}^{\infty}{}'\,\frac{(\frac{1}{2})_r^2}{(r!)^2}\frac{1}{r-n}+2\frac{(\frac{1}{2})_r^2}{(n!)^2}\left[\Psi(n+1)-\Psi\left(n+\frac{1}{2}\right)\right], \tag{3.27}$$

$\sum'$ having an obvious meaning. Since

$$\sum_{r=0}^{\infty}{}'\,\frac{(\frac{1}{2})_r^2}{(r!)^2}\frac{1}{r-n}=\sum_{r=0}^{n-1}\frac{(\frac{1}{2})_r^2}{(r!)^2}\frac{1}{r-n}+\left[\frac{(\frac{1}{2})_{n+1}}{(n+1)!}\right]^2{}_4F_3\left[\begin{matrix}1, & 1, n+\frac{3}{2}, n+\frac{3}{2}\\ 2, & n+2, n+2\end{matrix}\right], \tag{3.28}$$

(3.27) gives

$${}_4F_3\left[\begin{matrix}1, & 1, n+\frac{3}{2}, n+\frac{3}{2}\\ 2, & n+2, n+2\end{matrix}\right]$$

in terms of C_n.

We have explicitly checked (3.27), for $n=0,1,2,3$, by writing

$$\begin{aligned}\sum_{r=n+1}^{\infty}\frac{(\frac{1}{2})_r^2}{(r!)^2}\frac{1}{r-n}&=\int_0^1 x^{-(n+1)}\left[{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;x\right)-\sum_{r=0}^{n}\frac{(\frac{1}{2})_r^2}{(r!)^2}x^r\right]dx\\ &=\frac{1}{\pi}\int_0^1 t^{-1/2}(1-t)^{-1/2}\,dt\int_0^1 x^{-(n+1)}\\ &\quad\cdot\left[(1-tx)^{-1/2}-\sum_{r=0}^{n}\frac{(\frac{1}{2})_r}{r!}(tx)^r\right]dx,\end{aligned}$$

and by using the formula

$$\begin{aligned}H_n(t)&\equiv\int_0^1 x^{-(n+1)}\left[(1-tx)^{-1/2}-\sum_{r=0}^{n}\frac{(\frac{1}{2})_r}{r!}(tx)^r\right]dx\\ &=-\frac{1}{n}\varphi_{n,0}(t)-\frac{1}{2}\frac{t}{n(n-1)}\varphi_{n-1,1}(t)\\ &\quad-\frac{1\cdot 3}{2^2}\frac{t^2}{n(n-1)(n-2)}\varphi_{n-2,2}(t)\\ &\quad-\cdots+\frac{1\cdot 3\cdots(2n-1)}{2^n n!}t^nL_n(t),\end{aligned} \tag{3.29}$$

where

$$\varphi_{n,r}(t)\equiv(1-t)^{-(r+(1/2))}-\sum_{j=0}^{n}\frac{(r+\frac{1}{2})_j}{j!}t^j \tag{3.30}$$

and

$$L_n(t) \equiv \int_0^1 \frac{1}{x}[(1-tx)^{-(n+(1/2))}-1]\,dx. \tag{3.31}$$

Explicitly, we have

$$L_0(t) = -2\log\left(\frac{1+\sqrt{1-t}}{2}\right); \tag{3.32}$$

furthermore, it is easy to show that

$$L_{n+1}(t) = L_n(t) + \frac{1}{n+\frac{1}{2}}[(1-t)^{-(n+(1/2))}-1]. \tag{3.33}$$

For the sake of completeness, we also quote the recurrence relation

$$M_{n+1} = \frac{2n+1}{2n+2}M_n - \frac{1}{(n+1)(2n+1)} + \frac{\pi}{4}\frac{(\frac{1}{2})_n}{(n+1)(n+1)!}, \tag{3.34}$$

where

$$\begin{aligned} M_n &\equiv \int_0^1 t^{n-(1/2)}(1-t)^{-1/2}\log\left(\frac{1+\sqrt{1-t}}{2}\right) dt \\ &= 2\int_0^{\pi/2} \sin^{2n}\Theta \log\left(\frac{1+\cos\Theta}{2}\right) d\Theta. \end{aligned} \tag{3.35}$$

Last, we recall that

$$M_0 = -2\pi\log 2 + 4G. \tag{3.36}$$

This result follows from (3.35) by using

$$\log\left(\frac{1+\cos\Theta}{2}\right) = 2\sum_{r=0}^{\infty}\frac{(-1)^r}{r+1}\cos(r+1)\Theta - 2\log 2.$$

The proof of (3.29)-(3.34) is left as an exercise to the reader.

Acknowledgment. We express our sincere gratitude to Professor B. C. Berndt for his kind encouragement.

REFERENCES

[1] S. Ramanujan, *Collected Papers*, Chelsea, New York, 1962, p. 351.
[2] B.C. Berndt, *Ramanujan's Notebooks, Part* II, Springer-Verlag, New York, 1988.
[3] J. Dutka, *Two results of Ramanujan*, SIAM J. Math. Anal., 12 (1981), pp. 471-476.
[4] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Fourth ed., Cambridge University Press, Cambridge, UK, 1966, p. 521.
[5] B. C. Berndt, *Ramanujan's Notebooks, Part* II, Entry 35(iv), Springer-Verlag, New York, 1988, p. 45.
[6] A. Erdély, ed., *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953, p. 102.
[7] W. N. Bailey, *Generalized Hypergeometric Series*, Stechert-Hafner, New York, 1964, p. 21.

SIAM J. MATH ANAL.
Vol. 23, No. 3, pp. 571–578, May 1992

ON TWO CONJECTURES CONCERNING THE MULTIPLICITY OF SOLUTIONS OF A DIRICHLET PROBLEM*

HANS G. KAPER† AND MAN KAM KWONG†

Abstract. It is shown that there exist two strictly convex functions, both denoted by f on $(-\infty, \infty)$, the first satisfying the conditions $f'(-\infty) < 1$ and $n^2 < f'(\infty) < (n+1)^2$ and the second satisfying the conditions $(n-1)^2 < f'(-\infty) < n^2 < f'(\infty) < (n+1)^2$, such that the Dirichlet problem for the second-order nonlinear differential equation $u'' + f(u) = h(x)$ on $[0, \pi]$ has at least $2(n+1)$ and five solutions, respectively. This settles in the negative two questions raised by Lazer and McKenna [*SIAM Review,* 32 (1990), pp. 537–578].

Key words. nonlinear boundary value problem, multiplicity of solutions, variation index, topological degree

AMS(MOS) subject classifications. primary 34B15; secondary 35J25

1. The conjectures.

Recently, Lazer and McKenna [1] proposed a modified mathematical model for the onset of large-amplitude oscillations in suspension bridges forced by winds with specific velocities. Their study was motivated by the inadequacy of older theories' explanation of the collapse of the Tacoma Narrows Bridge of Seattle in 1941.

In the Lazer and McKenna model, the motion of the bridge is, as usual, governed by a system of differential equations, the complexity of which depends on the degree of approximation and the simplifications one is willing to accept. One of the new ideas introduced by Lazer and McKenna concerned the asymmetry of the restoring force that a cable exerts under expansion and compression. The authors' basic assumption is that the cable "strongly resists expansion, but does not resist compression." The study leads to boundary value problems of semilinear elliptic equations with Dirichlet conditions or second-order ordinary differential equations with periodic boundary conditions.

The study of semilinear elliptic equations with nonlinear restoring-force terms is a worthwhile subject, still largely unexplored. In their article [1], the authors collected several interesting open problems, some of which have not been answered even in the one-dimensional case, when the elliptic equation reduces to a second-order nonlinear ordinary differential equation.

In this article, we take up two of these open problems, namely problems 4 and 3. The problems concern the Dirichlet boundary value problem

$$u'' + f(u) = h(x), \qquad 0 < x < \pi, \tag{1.1}$$

$$u(0) = u(\pi) = 0, \tag{1.2}$$

where f is a genuinely nonlinear Lipschitz continuous function on $(-\infty, \infty)$ and h is any continuous function on $[0, \pi]$. The main objective is to determine upper and lower

* Received by the editors February 19, 1991; accepted for publication September 3, 1991. This work was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U. S. Department of Energy, under contract W-31-109-Eng-38.

† Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439-4801.

bounds for the number of distinct solutions of (1.1)–(1.2), when f is strictly convex and satisfies one of the following growth conditions:

$$f'(-\infty) < 1, \qquad n^2 < f'(\infty) < (n+1)^2, \tag{1.3}$$

$$(n-1)^2 < f'(-\infty) < n^2 < f'(\infty) < (n+1)^2. \tag{1.4}$$

Strict convexity is imposed to exclude certain degenerate cases. For example, when $f(u) = n^2 u$ for $u \in [-A, A]$, while f' still satisfies the inequalities in (1.3), the choice $h(x) = 0$ gives a continuum of solutions, namely, $u = a\sin(nx)$ for any $a < A$. However, when we construct our counterexamples in §5, we first work under less restrictive conditions on f; we subsequently use a perturbation argument to meet the full requirement.

In the following, we shall refer to a solution of (1.1)–(1.2) as a *D-solution* and reserve the simpler term *solution* for one that satisfies (1.1), but not necessarily the Dirichlet boundary conditions (1.2).

Lazer and McKenna proposed the following two conjectures:

CONJECTURE 1. *If f satisfies the conditions* (1.3), *then the boundary value problem* (1.1)–(1.2) *has at most* $2n$ *solutions for any given* h.

CONJECTURE 2. *If f satisfies the conditions* (1.4), *then the boundary value problem* (1.1)–(1.2) *has at most* 3 *solutions for any given* h.

In this article we show that there exist a constant-valued function h and a function f, which satisfy conditions (1.3), such that the boundary value problem (1.1)–(1.2) has at least $2(n+1)$ solutions. A slight modification of this counterexample will give a function which satisfies the conditions (1.4), such that the corresponding boundary value problem has at least five solutions. These results refute both conjectures of Lazer and McKenna.

Although all it takes to disprove a conjecture is one counterexample, augmented by convincing numerical data, we prefer to discuss in more depth the theoretical reasoning that led us to the discovery of the example. Such analytical considerations give a better understanding of the structure of the solution space and shed light on similar open problems.

We use the familiar shooting method. In §2 we sketch the simple concept of the variation index of a solution and its use in tracking the number of solutions. Although the index does not precisely indicate what must happen, it does show what may happen and, in our case, suggests that something can go wrong if we start with a solution having an undesirable index.

In §3 we present the analytical reasoning that led us to the counterexample. Our discussion pertains only to the particular case $n = 2$; extension to general n is obvious. In §4 we present a serendipitous result about the number of solutions of the Dirichlet problem (1.1)–(1.2) when the function h is even. In §5 we present the counterexamples along with some numerical results.

2. The variation index. Proofs in this section are only sketched, since all are elementary; in fact, some are quite well known. We consider the general inhomogeneous case (1.1) and the Dirichlet conditions (1.2).

The shooting method is intuitive. We replace the boundary value problem with an initial value problem, where, instead of (1.2), we impose the conditions

$$u(0) = 0, \qquad u'(0) = \alpha. \tag{2.1}$$

Here, $\alpha \in (-\infty, \infty)$ serves as a parameter. The solution of the initial value problem is a continuous function of both x and α; we denote its values by $u(x, \alpha)$. If, for some choice of α, $u(\pi, \alpha) = 0$, we have obtained a D-solution. Thus, by keeping tabs on how often $u(\pi, \alpha)$ changes sign as α varies from $-\infty$ to ∞, we can determine the total number of D-solutions. The job is often done by analyzing the upward/downward movement of the "tail end" of the solution trajectory.

Another technique, suitable for $h(x) = 0$, is to carefully follow the zeros of $u(x, \alpha)$, that is, the intersections of the graph of u with the x axis. Since a solution is never tangent to the x axis, the zeros of $u(x, \alpha)$ can only slide along the x axis and appear or disappear through the right endpoint π. More generally, and especially for inhomogeneous equations, we can follow the intersections of $u(x, \alpha)$ with a fixed D-solution, $u(x, \beta)$ say, for a suitable β. By the uniqueness theorem for initial value problems, no two distinct solutions $u(x, \alpha)$ and $u(x, \beta)$ of (1.1) can be tangent to each other at any $x \in [0, \pi]$. Thus, as α varies continuously, the intersection points of $u(x, \alpha)$ with $u(x, \beta)$ can only slide along the graph of the latter. The intersection number of the two solutions is an elementary interpretation of the abstract concept of topological degree. If a decrease in the number of intersections is noticed as α varies from one value to another, the lost intersection point must have slipped away through the right endpoint at some intermediate value of α. That value corresponds to a D-solution.

The intersection number of a solution $u(x, \beta)$, with a neighboring solution $u(x, \alpha)$, $\alpha \approx \beta$, can be approximated by studying the function w,

$$w(x) = w(x, \beta) = \partial u(x, \alpha)/\partial \alpha|_{\alpha=\beta}, \tag{2.2}$$

which satisfies the variational equation

$$w''(x) + f'(u)w(x) = 0, \qquad x > 0, \tag{2.3}$$

with initial conditions

$$w(0) = 0, \qquad w'(0) = 1. \tag{2.4}$$

In (2.3) we substitute the solution $u(x, \beta)$ into the expression $f'(u)$ as if it is a known function, and we regard (2.3) as a linear differential equation of w with the known coefficient $f'(u(x, \beta))$. The number of times that $w(x)$ changes sign in $[0, \pi]$ gives valuable information. The fact that w satisfies a linear differential equation makes it possible to apply the classical Sturm comparison technique.

We say that the *variation index* of $u(x, \beta)$ is k if the corresponding function w has k zeros in $(0, \pi]$. Note that $x = 0$ is always a zero, but it is not counted. If, furthermore, $w(\pi) \neq 0$, we say that the variation index is k^+. For simplicity, we shall use the abbreviated term *index*.

The use of the variational equation to study multiplicity of solutions has a long history, usually associated with bifurcation theory. In particular we note its use by Coffman [3], and subsequently by McLeod and Serrin [4], Ni and Nussbaum [5], and Kwong et al. [6]–[9] in the study of the uniqueness and nonuniqueness of the ground state of semilinear elliptic equations. The concept of the index is used throughout these works, even though it is not identified by name. A definition of the variation index was given recently by Clemons [10].

What is the significance of the sign of $w(x)$ at a given point x? The sign tells us whether $u(x, \alpha)$ increases or decreases as α is slightly increased from β. What about the index itself? If the index is k^+, there is sequence of $k + 1$ points

$$0 < x_1 < x_2 < \cdots < x_{k+1} < \pi, \tag{2.5}$$

such that

$$(2.6)\quad \left.\frac{\partial u(x_1,\alpha)}{\partial\alpha}\right|_{\alpha=\beta} > 0,\ -\left.\frac{\partial u(x_2,\alpha)}{\partial\alpha}\right|_{\alpha=\beta} > 0, \ldots, (-1)^k \left.\frac{\partial u(x_{k+1},\alpha)}{\partial\alpha}\right|_{\alpha=\beta} > 0.$$

Thus, there exists a small enough $\epsilon > 0$ such that for all $\alpha \in (\beta, \beta+\epsilon)$,

$$(2.7)\qquad u(x_1,\alpha) > u(x_1,\beta),\ u(x_2,\alpha) < u(x_2,\beta),\ \cdots$$

It follows that $u(x,\alpha)$ intersects $u(x,\beta)$ at least k times. The same is true for $\alpha \in (\beta-\epsilon,\beta)$. Note that in the degenerate case, when the index is k but not k^+, we can assert only the existence of $k-1$ intersection points.

Let us consider an example to see how the index can be used to deduce multiplicity results. Under the assumption $f'(-\infty) < 1$, it is easy to see that, given any β, we can find an α sufficiently negative that $u(x,\alpha)$ does not intersect $u(x,\beta)$ at all. On the other hand, under the additional assumption that $f'(\infty) > 1$, it can be shown that there exists a δ sufficiently large that $u(x,\delta)$ intersects $u(x,\beta)$ exactly once. Suppose now that $u(x,\beta)$ is actually a D-solution and that its index is known to be k^+. Let us decrease α from β to γ. Initially, when α is sufficiently close to β, $u(x,\alpha)$ has k intersection points with $u(x,\beta)$. As α reaches γ, all of these are lost. We thus conclude that at least k intermediate values of α give rise to D-solutions. Likewise, if we increase α from β to δ, $k-1$ intersection points are lost, adding $k-1$ more D-solutions. Including $u(x,\beta)$, there is a total of $2k$ D-solutions.

In [2], Lazer and McKenna studied the equation (1.1) with $h(x) = s\sin(x) + h_1(x)$ for large s. They constructed a large, almost linear solution, which can be shown to have index n^+. Starting with this D-solution, we can use the above arguments to obtain the main theorem of [2], namely, that there exist at least $2n$ D-solutions. The proof we sketched here is, of course, only a paraphrase of that in [2], with the role of the index underscored.

3. Toward the counterexamples. In this section, we confine ourselves mainly to Conjecture 1 described in §1, assuming that (1.3) holds. Some of the observations below, indicated by the presence of the modifiers *generic* or *generically*, are meant to be intuitive and nonrigorous.

If there exist an infinite number of D-solutions, then the conjecture is obviously false. So we henceforth assume that there are a finite number of D-solutions.

We order the D-solutions according to the magnitude of the corresponding parameter α and refer to the D-solution with the smallest (largest) α as the *first* (*last*) D-solution. We know from the discussion in the preceding section that the index of the first solution is zero or at most 1 (but not 1^+). Intuitively, we know that a degenerate index is nongeneric and occurs rarely. In any case, a small perturbation can bump it into a nondegenerate index. Thus, generically the first solution has index zero.

Using the comparison principle and the convexity assumption of f, it can easily be shown that from the third D-solution onward, the index cannot be zero. The discussion in the preceding section also shows that the index of the last D-solution is at most 2. Hence, if more than two D-solutions exists, the last one generically has index 1^+.

The first solution cannot intersect any other solution in $(0,\pi)$; otherwise, a continuity argument involving decreasing α will give an extra solution before the first one. On the other hand, we claim that any two other solutions must intersect.

We prove the claim by contradiction. Let u_1 denote the first solution and u_2, u_3 two other solutions. If $u_2 \leq u_3$ (they do not intersect), then both $u_2 - u_1$ and $u_3 - u_1$ are nonnegative functions satisfying a second-order differential equation of the form

$$(3.1) \qquad (u_i - u_1)'' + \frac{f(u_i) - f(u_1)}{u_i - u_1}(u_i - u_1) = 0, \qquad i = 2, 3.$$

The convexity of f, however, implies that the coefficient in the equation for $i = 3$ is larger than that in the equation for $i = 2$. Hence, by the Sturm comparison theorem, $u_3 - u_1$ oscillates faster than $u_2 - u_1$, a contradiction.

We now restrict ourselves to the case $n = 2$ and assume that h is symmetric,

$$(3.2) \qquad h(x) = h(\pi - x), \qquad 0 < x < \pi.$$

By (1.3), the coefficient $f'(u)$ in (2.3) is less than 9. By the Sturm comparison theorem, $w(x)$ cannot change sign more than twice. Thus the index of any $u(x, \alpha)$ is at most 2^+. Ignoring the degenerate cases, we therefore have three choices: 0, 1^+, and 2^+. The Sturm comparison theorem also tells us that no two solutions can intersect more than twice in $(0, \pi]$. It follows that no two D-solutions can intersect more than once in the interior $(0, \pi)$.

The first D-solution, denoted by $u(x, \beta)$, has index zero; hence, as we increase α from β, the tail of the graph, $u(\pi, \alpha)$, initially moves upward. Thus $u(\pi, \alpha) > 0$ for α greater than but close to β. In order to produce the next solution, $u(x, \gamma)$, the tail must start to move downward for some value of α before γ and continue to do so until γ is reached. Such movement of the tail can happen only if the index of $u(x, \gamma)$ is odd (again ignoring the degenerate case); thus, the only choice is 1^+. If $u(x, \gamma)$ is not the last D-solution, we can continue this argument to conclude that, generically, adjacent D-solutions have indices of opposite parity. The argument also shows that generically we expect an even number of D-solutions, an expectation that is confirmed in the symmetric case below by Theorem 1. A degeneracy may occur when the tail of the graph moves towards zero and then bounces back in the opposite direction, in which case we have a degenerate index. With all these observations, we conclude that the possible generic configurations of indices are $0, 0$ or $0, 1^+$ for two D-solutions and $0, 1^+, 2^+, 1^+$ for four D-solutions.

What additional properties are imposed by the symmetry of h? The first D-solution must be symmetric; otherwise it intersects its reflection, contradicting our assertion that the first solution does not intersect any other solutions. If more than two D-solutions exists, the last one cannot be symmetric. Indeed, it must intersect the second D-solution exactly once in $(0, \pi)$. Reflecting the second D-solution will then give another D-solution having a larger α than the last D-solution, a contradiction. Now that we know the last D-solution is not symmetric, its reflection must be a different D-solution and must have the same index 1^+. The only way this situation can happen in the four D-solution case is, therefore, that the second and fourth D-solutions are reflections of each other, and this forces the third to be symmetric with index 2^+ (recall that the first D-solution must be symmetric).

Although intuitive, these observations provide us enough clues for the search for a counterexample. Suppose we start with an equation that has a symmetric D-solution with index 1^+. We have just seen that this is incompatible with the case with four D-solutions. Hence, either we have only two D-solutions or the conjecture is false. We pick an example in which the index of the symmetric D-solution is 1^+ but very close to 2. Let the maximum of the D-solution be M. We still have freedom to modify

$f(u)$ for $u > M$. A small decrease in α brings part of the solution above M, and this enables us to manipulate $f(u)$, for $u > M$, so that the solution will bend down fast enough to give one more D-solution—thus excluding the case with two D-solutions.

Once a concrete example has been constructed, it is not difficult to furnish a rigorous proof.

We can argue in a similar way to obtain a counterexample to Conjecture 2. It so happened that almost the same nonlinear function $f(u)$ worked for both conjectures.

4. Interlude. Before proceeding with the actual construction of a counterexample, we digress for a moment to give a serendipitous result on the evenness of the number of D-solutions for symmetric equations, which follows from the arguments used in the preceding section.

THEOREM 1. *Suppose that* (1.3) *and* (3.2) *hold. Then* (1.1)–(1.2) *has either exactly one D-solution or an even number of D-solutions. In the latter case, there is a pair of distinct symmetric D-solutions.*

Proof. As seen before, the first D-solution is symmetric. If there is a nonsymmetric D-solution $u(x,\beta)$, its reflection $u(x,\gamma) = u(\pi - x, \beta)$ is also a D-solution. At the midpoint of the interval, $u'(\pi/2,\beta) = -u'(\pi/2,\gamma) \neq 0$. A shooting argument shows that at some intermediate α between β and γ, $u'(\pi/2,\alpha) = 0$, and this will give a second symmetric D-solution. There cannot be a third symmetric D-solution; otherwise it will intersect the second symmetric D-solution at some point different from $\pi/2$. By symmetry the two D-solutions will have more than two intersection points in $(0,\pi)$, contrary to the upper bound imposed on $f(u)$. The conclusion now follows from the fact that the remaining nonsymmetric D-solutions must occur in pairs. □

5. The counterexamples. Let us, for the time being, overlook the requirement of strict convexity on f and choose for f a non-C^1 function. This will simplify the analysis and the programming for our numerical computation. For the first counterexample, let

$$f(u) = \begin{cases} 8u - 439/100, & u \in [1,\infty), \\ 361u/100, & u \in [-C,1), \\ -361C/100, & u \in (-\infty,-C), \end{cases} \tag{5.1}$$

where C is some suitable large positive number. For the second counterexample, we simply let C be ∞ in the second line of (5.1) and eliminate the third line. We choose a constant forcing term on the right-hand side of (1.1),

$$h(x) = \frac{361}{100}\left(\frac{1+\cos\frac{\pi}{20}}{\cos\frac{\pi}{20}}\right), \tag{5.2}$$

which is determined so that the function

$$u(x,\beta) = \frac{1}{1+\cos\frac{\pi}{20}}\cos\left(\frac{19}{10}\left(x-\frac{\pi}{2}\right)\right) + \frac{1+\cos\frac{\pi}{20}}{\cos\frac{\pi}{20}} \tag{5.3}$$

is a D-solution of (1.1), where $\beta = u'(0,\beta)$.

The existence of at least six D-solutions, for large C, and at least five D-solutions, for $C = \infty$, is confirmed numerically. We start our computation by shooting solutions from $x = 0$ with initial slope between -10 and 10 at increments of 0.1. For the integration we use a fourth/fifth-order Runge–Kutta method from MATLAB, with an error tolerance of 10^{-6}. More accurate computations carried out near the D-solution, by using smaller tolerances and increments of the initial slope, confirm our conclusion.

The data given in Figs. 1 and 2 belong to the choice $C = 1$.

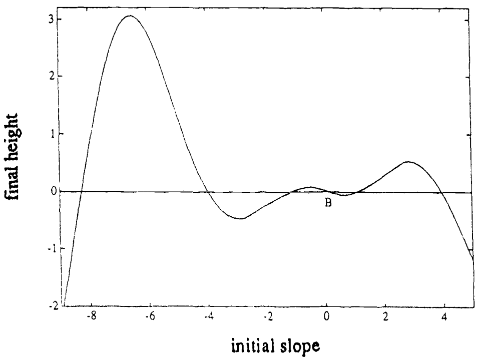


FIG. 1. *Final height $u(\pi, \alpha)$ vs. initial slope α.*

Figure 1 shows the graph of the final height $u(\pi, \alpha)$ as a function of the initial slope $\alpha = u'(0, \alpha)$ for $\alpha \in [-9, 5]$. Each intersection of the graph with the α axis gives a D-solution. Six D-solutions are clearly identifiable. No other solutions turn up when a wider range of α is used. For the second counterexample, with $C = \infty$, the corresponding graph is almost the same, except that the part where the initial slope is less than -5 is entirely above the axis; hence, there are five D-solutions.

Figure 2 shows the graphs of three shooting solutions $x \mapsto u(x, \alpha)$ for different values of the initial slope α.

Note that the graph in Fig. 1 is nowhere tangent to the α axis, indicating that none of the D-solutions is degenerate. It follows that a sufficiently small perturbation will not alter the total number of D-solutions. Thus, if we replace our choice (5.1) of f by a strictly convex smooth function, all six D-solutions will be preserved.

A theoretical proof of the existence of more than four D-solutions can be given along the line of reasoning set forth in §3. The computations, which are straightforward, are omitted.

The D-solution (5.3) is represented by the point B in Fig. 1 and by the graph labeled u(x,b) in Fig. 2. It has index 1^+, since the variational equation is

$$w''(x) + \frac{361}{100} w(x) = 0, \quad w(0) = 0, \quad w'(0) = 1, \tag{5.4}$$

whose solution $w(x) = \sin(19x/10)$ has only one zero in $(0, \pi)$. As shown in §2, this means that if α decreases from β, the final height $u(\pi, \alpha)$ moves upward, at least for a while. This property is confirmed in Fig. 1, as the part of the graph to the immediate left of B is above the α axis. In Fig. 2, the same property is manifested by the fact that the solution labeled u(x,a) intersects u(x,b) once and remains above it afterwards.

We continue to decrease α to γ; the corresponding solution $u(x, \gamma)$ is labeled u(x,c) in Fig. 2. All these solutions can be computed analytically because the equations are linear in each of the two disjoint regions $u < 1$ and $u > 1$. Indeed, the difference $u(x, \beta) - u(x, \gamma)$ is simply a multiple of $\sin(19x/20)$, as long as $u(x, \alpha)$ remains less than one. The portion of the solution above the dotted line $u = 1$ is now subjected to a much larger "restoring force," as the coefficient of the linear term

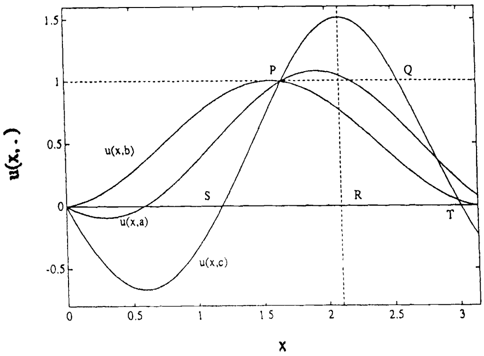


FIG. 2. *Shooting solutions for three different initial slopes.*

has jumped from 361/100 to 8. It is not hard to show that the duration for which the solution can remain above $u = 1$, i.e., the length PQ in Fig. 2, cannot be more than $\pi/\sqrt{8}$. The horizontal distance between the point S, where u(x,c) crosses the x axis, and the point P diminishes as γ decreases and can be arranged to be arbitrarily small if C is chosen sufficiently large. The symmetry of u(x,c) about the point R, halfway between P and Q, implies that the graph must intersect the x axis again at some point T, with $SR = RT$. If S is sufficiently close to P, then T falls within $(0, \pi)$, as shown. The change of sign from $u(\pi, \alpha)$ to $u(\pi, \gamma)$ implies the existence of a D-solution between α and γ. There are two intersection points between $u(x, \beta)$ and $u(x, \gamma)$. As the initial slope is further decreased below γ, two more D-solutions must occur before both of the intersection points are completely lost. Of these three extra solutions, one may be the first D-solution and so is symmetric. The other two will give two more D-solutions upon reflection.

It is easy to see how the construction can be extended to values of $n > 2$. We need to start with a symmetric D-solution having index $(n-1)^+$, which can be perturbed slightly to give a symmetric D-solution with index n^+. We leave it to the readers to supply the details.

REFERENCES

[1] A. C. LAZER AND P. J. MCKENNA, *Large-amplitude periodic oscillations in suspension bridges: Some new connections with nonlinear analysis*, SIAM Rev., 32 (1990), pp. 537–578.

[2] ———, *On a conjecture related to the number of solutions of a nonlinear Dirichlet problem*, Proc. Roy. Soc. Edinburgh, Sect. A, 95 (1983), pp. 275–283.

[3] C. V. COFFMAN, *On the positive solutions of boundary value problems for a class of nonlinear differential equations*, J. Differential Equations, 3 (1967), pp. 92–111.

[4] K. MCLEOD AND J. SERRIN, *Uniqueness of positive radial solutions of $\Delta u + f(u) = 0$ in R^n*, Arch. Rational Mech. Anal., 99 (1987), pp. 115–145.

[5] W. M. NI AND R. NUSSBAUM, *Uniqueness and nonuniqueness for positive radial solutions of* $\Delta u + f(u, r) = 0$, Comm. Pure Appl. Math., 38 (1985), pp. 69–108.

[6] MAN KAM KWONG, *On the Kolodner–Coffman method for the uniqueness problem of Emden–Fowler BVP*, Z. Angew. Math. Phys. (J. Appl. Math. Phys.), 41 (1990), pp. 79–104.

[7] ———, *Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in R^n*, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[8] MAN KAM KWONG AND YI LI, *Uniqueness of radial solutions of semilinear elliptic equations*, Argonne National Laboratory, Argonne, IL, preprint MCS-P156-0590; Trans. Amer. Math. Soc., to appear.

[9] MAN KAM 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.

[10] C. B. CLEMONS, *Uniqueness results for semilinear elliptic equations*, Ph.D. dissertation, University of Maryland, 1990.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 579-608, May 1992

002

VARIATIONAL FORMULATIONS FOR THE DETERMINATION OF RESONANT STATES IN SCATTERING PROBLEMS*

M. LENOIR†, M. VULLIERME-LEDARD†, AND C. HAZARD†

Abstract. Consider the scattering of an acoustic wave by a rigid obstacle. The poles of the analytical continuation of the resolvent operator are called scattering frequencies. On their localization depend the time-decay of the solution and the location of the energy peaks of the steady-state solution.

Two methods are proposed to construct explicitly the analytical continuation of the resolvent: the localized finite element method or the coupling between variational formulation and integral representation, which both rely upon the reduction of the exterior Helmholtz problem to a bounded domain. The determination of the scattering frequencies then amounts to solving a nonlinear eigenvalue problem for a completely continuous operator.

Then, the expansion of the approximate steady-state solution in the vicinity of a scattering frequency is computed. Numerical results for a simple one-dimensional problem are presented.

Key words. scattering frequencies, localized finite element method, integral representation

AMS(MOS) subject classifications. 35B60, 35P25, 45C05, 65N25

1. Introduction.

1.1. Motivation.

One of the most important questions in the study of coupled vibrations of an elastic solid and a compressible fluid is the determination of the energy transfer between the solid and the fluid. The question of determining the values of the excitation frequency of the structure which makes maximum the radiated acoustic pressure is usually denoted a "radiation problem" (see [7]); these frequencies are referred to as "resonant frequencies." In the case of the scattering of a plane monochromatic acoustic wave, they are the frequencies for which the response of the system shows maxima of amplitude. A similar problem arises in naval hydrodynamics in the study of periodic motions of a ship under the influence of a monochromatic swell; the question is now to compute the frequencies inducing motions of maximum amplitude.

The purpose of this work is to describe a *practical method of computation* of these frequencies together with the associated values of the response of the system. The direct determination of the resonant frequencies appears to be difficult; the study of the singularities of the scattering matrix (the so-called "scattering frequencies" which occur for complex values of the frequency) is, however, easier and provides valuable information about the resonant frequencies which stay along the real axis in the vicinity of these scattering frequencies (see Fig. 1). Recall that the scattering matrix is the operator connecting the asymptotic behaviour of the scattered wave to the incident plane monochromatic wave.

For the wave equation in the exterior of a rigid obstacle, Shenk and Thoe [10] have shown that the scattering frequencies are nothing but the poles of the analytical continuation of the resolvent of the Helmholtz operator. The method we propose relies on this result.

1.2. A brief description of the principle.

The method of coupling between variational formulation and integral representation [5] or the localized finite element method [6] allows us to reduce the exterior Helmholtz problem to a problem set in a

* Received by the editors August 22, 1988; accepted for publication (in revised form) October 22, 1990.

† Groupe Hydrodynamique Navale, associé au Centre National de Recherche Scientifique et à l'U.P.M.C. Ecole Nationale Supérieure de Techniques Avancées, Centre de l'Yvette, Chemin de la Hunière, 91120 Palaiseau France.

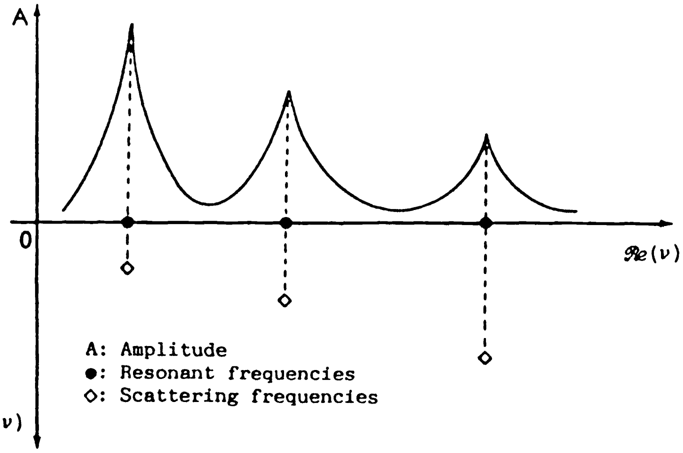


FIG. 1

bounded domain and to exhibit the analytical continuation of the resolvent. The determination of the scattering frequencies reduces then to the solution of a nonlinear eigenvalue problem for a completely continuous operator. This approach is similar to that of Shenk and Thoe [9], which relies upon an integral equation method; it is, however, easier to handle, due to its variational character and due to the fact that no volume potential is needed to remove fictitious singularities of the analytical continuation of the resolvent. The next step consists in expanding the solution in the vicinity of a scattering frequency in order to compute the location of the resonance and the associated amplitude of the solution.

The practical efficiency of our method has been tested in the case of the coupling between a vibrating string and a beam through a density of springs; it is the simplest model for fluid-structure interaction. In this paper we shall only describe the method in the simple case of the wave equation; the application to naval hydrodynamics will be accounted for in a forthcoming paper.

1.3. Some other procedures. Other procedures have been designed for the computation of resonant frequencies. The method of Wei, Majda and Strauss described by Wei [13] relies upon the transient theory of scattering of Lax and Phillips [4]; the solution of the wave equation is actually subject to an asymptotic expansion with respect to time in the following form:

$$u(t)=\sum_{j\in\mathbb{N}} \varphi_j(x)\, e^{-i\nu_j t},$$

where the ν_j are the scattering frequencies. The numerical computation of the longtime solution of the wave equation then allows the identification of the ν_j. The advantage of our procedure is to rely upon the stationary theory of scattering and, therefore, to apply to situations which are not relevant to Lax and Phillips theory, for example, the sea-keeping problem [12]. Also worth mentioning is the work of Ohayon and Sanchez-Palencia [8], which is about the coupling between a structure and a slightly compressible fluid. They first investigate the convergence of the scattering frequencies to the eigenfrequencies of the limit problem associated to an incompressible fluid. Then, they determine the resonant frequencies by expanding the solution with respect to the compressibility in the vicinity of the eigenfrequencies.

2. The resolvent. By Ω we denote the exterior of a compact set in $\mathbb{R}^n$; its boundary is denoted by Γ. The function $\Phi(x, t)$ is the solution of the wave equation

$$\frac{\partial^2\Phi}{\partial t^2}-c^2\Delta\Phi=0 \quad \text{in } \Omega,$$

$$\frac{\partial\Phi}{\partial n}=\operatorname{Re}\,(f(x)\,e^{-i\omega t}) \quad \text{on } \Gamma, f \text{ and } \omega \text{ being given},$$

$$\left(\Phi(x,0), \frac{\partial\Phi}{\partial t}(x,0)\right)=(\Phi_1(x), \Phi_2(x)), \quad \text{given.}$$

The limiting amplitude principle (Eidus [2]) shows that, as $t\to+\infty$, $\frac{1}{2}e^{i\omega t}(\Phi(x,t)+i\Phi(x,t+(\pi/2\omega)))$ tends to $\varphi(x)$ in the energy norm on compact sets, where φ is the solution of the reduced wave equation

$$\Delta\varphi+\nu_0\varphi=0 \quad \text{in } \Omega, \quad \text{with } \nu_0=\frac{\omega^2}{c^2},$$

$$(P_{\nu_0}) \qquad \frac{\partial\varphi}{\partial n}=f(x) \quad \text{on } \Gamma,$$

$$\left(\frac{\partial\varphi}{\partial R}-i\sqrt{\nu_0}\,\varphi\right)=o(R^{-(n-1)/2}) \quad (\textit{outgoing} \text{ radiation condition}).$$

If ν is any complex number of *positive imaginary part*, then problem

$$\Delta\varphi+\nu\varphi=0 \quad \text{in } \Omega,$$

$$(Q_\nu) \qquad \frac{\partial\varphi}{\partial n}=f \quad \text{on } \Gamma$$

has a unique solution in $H^1(\Omega)$. In variational form (Q_ν) reads:

$$\int_\Omega \nabla\varphi\,\nabla\bar\psi\,d\omega-\nu\int_\Omega \varphi\bar\psi\,d\omega=\int_\Gamma f\bar\psi\,ds \quad \forall\psi\in H^1(\Omega), \text{ i.e.,}$$

$$(I+S(\nu))\varphi=F(f) \quad \text{in } H^1(\Omega), \text{ where}$$

$$(S(\nu)\varphi\,|\,\psi)_{H^1(\Omega)}=-(1+\nu)\int_\Omega \varphi\bar\psi\,d\omega \quad \text{and} \quad (F(f)\,|\,\psi)_{H^1(\Omega)}=\int_\Gamma f\bar\psi\,ds.$$

By $R(\nu)$ (resolvent) we shall denote the operator $(I+S(\nu))^{-1}$: $H^1(\Omega)\to H^1(\Omega)$.

As $\nu\to\nu_0\in\mathbb{R}^+$, with $\operatorname{Im}(\nu)>0$, by the *limiting absorption principle* the solution φ of (Q_ν) tends to the solution of (P_{ν_0}) (Wilcox [14]). In the sequel we shall show that $R(\nu)$, defined for values of ν such that $\operatorname{Im}(\nu)>0$, actually extends to $\mathbb{C}\backslash\mathbb{R}^-$ as a meromorphic function of ν; the poles are, of course, located in the half-plane of complex numbers satisfying $\operatorname{Im}(\nu)<0$.

Remark 1. When the same limiting procedure is applied to the solution of problem (Q_ν) with $\operatorname{Im}(\nu)<0$, it no longer leads to the solution of problem (P_{ν_0}), but to a function satisfying the incoming radiation condition

$$\left(\frac{\partial\varphi}{\partial R}+i\sqrt{\nu_0}\,\varphi\right)=o(R^{-(n-1)/2}).$$

3. Coupling between variational formulation and integral representation.

3.1. Reduction to a bounded domain. The main difficulty we face in studying the way $R(\nu)$ depends on ν lies in the fact that the limit of $R(\nu)F(f)$ no longer belongs to $H^1(\Omega)$ when ν tends to the positive real axis. We shall thus be led to give another formulation of (Q_ν) allowing a precise control of the behaviour of $R(\nu)F(f)$ at infinity. For the time being we shall assume that $\text{Im}(\nu)>0$.

By $G_\nu(x)$ we shall denote the fundamental solution of $\Delta+\nu I$; for instance:

$$G_\nu(x)=\frac{1}{4i}H_0^{(1)}(\sqrt{\nu}\,\|x\|) \quad \text{in two dimensions} \tag{1}$$

$$G_\nu(x)=\frac{-1}{4\pi}\frac{e^{i\sqrt{\nu}\|x\|}}{\|x\|} \quad \text{in three dimensions,} \tag{2}$$

where $\sqrt{\nu}=\sqrt{\rho}\,e^{i\theta/2}$ for $\nu=\rho\,e^{i\theta}$ and $\theta\in]-\pi,+\pi]$.

The following representation formula holds for the solution φ of (Q_ν):

$$\varphi(y)=\int_\Gamma\left\{\varphi(x)\frac{\partial}{\partial n_x}G_\nu(x-y)-\frac{\partial\varphi}{\partial n}(x)G_\nu(x-y)\right\}ds_x.$$

It follows that on any boundary Σ surrounding Γ, we have

$$D_\lambda\varphi(y)=\int_\Gamma\left\{\varphi(x)\frac{\partial}{\partial n_x}\gamma_\nu(x,y)-\frac{\partial\varphi}{\partial n}(x)\gamma_\nu(x,y)\right\}ds_x,$$

with $D_\lambda\chi=((\partial\chi/\partial n)+\lambda\chi)_{|\Sigma}$ and $\gamma_\nu(x,y)=-(D_\lambda G_\nu)(x-y)$.

We are thus led to set the following problem in the bounded domain $\tilde{\Omega}$ limited by Γ and Σ (see Fig. 2):

$$(\tilde{Q}_\nu)\qquad \begin{aligned}&\Delta\tilde{\varphi}+\nu\tilde{\varphi}=0 \quad \text{in } \tilde{\Omega},\\ &\frac{\partial\tilde{\varphi}}{\partial n}=f \quad \text{on } \Gamma,\\ &D_\lambda\tilde{\varphi}(\cdot)=\int_\Gamma\left\{\tilde{\varphi}(x)\frac{\partial}{\partial n_x}\gamma_\nu(x,\cdot)-\frac{\partial\tilde{\varphi}}{\partial n}(x)\gamma_\nu(x,\cdot)\right\}ds_x.\end{aligned}$$

By $V(\Sigma,\lambda)$ we shall denote the denumerable set of the eigenvalues of the following associated problem:

$$(Z_\lambda)\qquad \begin{aligned}&\Delta\psi+\nu\psi=0 \quad \text{in } \Omega',\\ &D_\lambda\psi=0 \quad \text{on } \Sigma,\end{aligned}$$

where Ω' is the bounded domain with boundary Σ (see Fig. 3).

Remark 2. When $\text{Im}(\lambda)=0$ (respectively, $\text{Im}(\lambda)<0$, respectively, $\text{Im}(\lambda)>0$), then $V(\Sigma,\lambda)$ is included into $\mathbb{R}$ (respectively, into $\{\nu\,|\,\text{Im}(\nu)<0\}$, respectively, into $\{\nu\,|\,\text{Im}(\nu)>0\}$). □

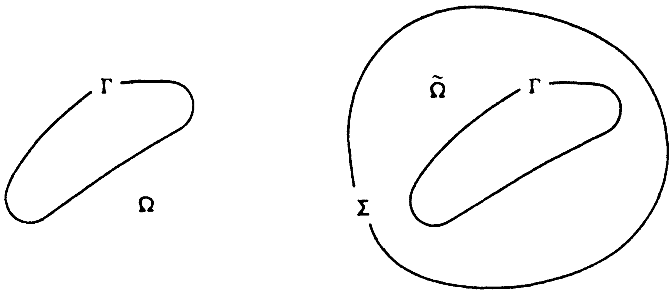


FIG. 2

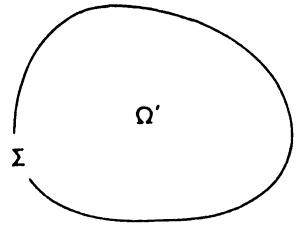


FIG. 3

PROPOSITION 1. *For* $\operatorname{Im}(\nu)>0$ *and* $\nu\notin V(\Sigma,\lambda)$, *problem* $(\tilde{Q}_\nu)$ *is well posed and its solution* $\tilde{\varphi}$ *is nothing but the restriction to* $\tilde{\Omega}$ *of the solution* φ *of* (Q_ν).

Proof. From Green's formula, we have, for $y\in\tilde{\Omega}$,

$$\tilde{\varphi}(y)=\int_{\Gamma\cup\Sigma}\left\{\tilde{\varphi}(x)\frac{\partial}{\partial n_x}G_\nu(x-y)-\frac{\partial\tilde{\varphi}}{\partial n}(x)G_\nu(x-y)\right\}ds_x.$$

Put

$$\psi(y)=\int_{\Sigma}\left\{\tilde{\varphi}(x)\frac{\partial}{\partial n_x}G_\nu(x-y)-\frac{\partial\tilde{\varphi}}{\partial n}(x)G_\nu(x-y)\right\}ds_x;$$

clearly, ψ extends to the whole Ω' as a solution of (Z_λ). If $\nu\notin V(\Sigma,\lambda)$, then

$$\tilde{\varphi}(y)=\int_{\Gamma}\left\{\tilde{\varphi}(x)\frac{\partial}{\partial n_x}G_\nu(x-y)-\frac{\partial\tilde{\varphi}}{\partial n}(x)G_\nu(x-y)\right\}ds_x,$$

from which we deduce that $\tilde{\varphi}$ extends to the whole Ω as a solution of (P_ν). The conclusion follows then by uniqueness of (P_ν). □

Under variational form, problem $(\tilde{Q}_\nu)$ writes as

$$(I+\tilde{S}(\nu))\tilde{\varphi}=\tilde{F}(f,\nu)\quad\text{in } H^1(\tilde{\Omega}),\quad\text{with}$$

$$(3)\qquad \begin{aligned}(\tilde{S}(\nu)\tilde{\varphi}\,|\,\tilde{\psi})_{H^1(\tilde{\Omega})}&=-(1+\nu)\int_{\tilde{\Omega}}\tilde{\varphi}\bar{\tilde{\psi}}\,d\omega+\lambda\int_{\Sigma}\tilde{\varphi}\bar{\tilde{\psi}}\,ds\\&\quad-\int_{\Sigma}\bar{\tilde{\psi}}(y)\int_{\Gamma}\tilde{\varphi}(x)\frac{\partial\gamma_\nu(x,y)}{\partial n_x}\,ds_x\,ds_y,\end{aligned}$$

and

$$(\tilde{F}(f,\nu)\,|\,\tilde{\psi})_{H^1(\tilde{\Omega})}=\int_{\Gamma}f\bar{\tilde{\psi}}\,ds-\int_{\Sigma}\bar{\tilde{\psi}}(y)\int_{\Gamma}f(x)\gamma_\nu(x,y)\,ds_x\,ds_y.$$

3.2. Analytical continuation of the resolvent. Formulas (1) and (2) extend analytically to $\mathbb{C}\backslash\mathbb{R}^-$; the continuation of $\tilde{S}(\nu)$ follows.

PROPOSITION 2. *The operator* $\tilde{S}(\nu)$ *is completely continuous on* $H^1(\tilde{\Omega})$; *moreover, function* $\tilde{S}$: $\nu\mapsto\tilde{S}(\nu)$, $\mathbb{C}\backslash\mathbb{R}^-\to\mathscr{L}(H^1(\tilde{\Omega}),H^1(\tilde{\Omega}))$ *is holomorphic.*

Proof. (i) We have

$$\begin{aligned}\|\tilde{S}(\nu)\tilde{\varphi}\|_{H^1(\tilde{\Omega})}\leqq\operatorname*{Sup}_{\tilde{\psi}\in H^1(\tilde{\Omega})}\frac{1}{\|\tilde{\psi}\|_{H^1(\tilde{\Omega})}}\Bigg(&|1+\nu|\,\|\tilde{\varphi}\|_{L^2(\tilde{\Omega})}\|\tilde{\psi}\|_{L^2(\tilde{\Omega})}+|\lambda|\,\|\tilde{\varphi}\|_{L^2(\Sigma)}\|\tilde{\psi}\|_{L^2(\Sigma)}\\&+\|\tilde{\varphi}\|_{L^2(\Gamma)}\|\tilde{\psi}\|_{L^2(\Sigma)}\int_{\Gamma}\left\|\frac{\partial\gamma_\nu(x,\cdot)}{\partial n_x}\right\|_{L^2(\Sigma)}ds_x\Bigg),\end{aligned}$$

and as a consequence,

$$\|\tilde{S}(\nu)\tilde{\varphi}\|_{H^1(\tilde{\Omega})} \leqq C_1\|\tilde{\varphi}\|_{L^2(\tilde{\Omega})} + C_2\|\tilde{\varphi}\|_{L^2(\Sigma)} + C_3\|\tilde{\varphi}\|_{L^2(\Gamma)};$$

the compactness of $\tilde{S}(\nu)$: $H^1(\tilde{\Omega}) \to H^1(\tilde{\Omega})$ follows.

(ii) The holomorphic dependence of $\tilde{S}(\nu)$ on ν is a straightforward consequence of that of G_ν. □

From now on, we shall denote $(I+\tilde{S}(\nu))^{-1}$ by $\tilde{R}(\nu)$.

COROLLARY 1. *Function $\tilde{R}(\nu)$ extends meromorphically to $\mathbb{C}\backslash\mathbb{R}^-$.*

Proof. From Proposition 1, $I+\tilde{S}(\nu)$ is invertible for $\operatorname{Im}(\nu)>0$ and $\nu \notin V(\Sigma, \lambda)$; the conclusion follows then from Proposition 2 and from Steinberg's theorem [11]. □

Proposition 1 shows that $V(\Sigma, \lambda)$ is included in the set $\tilde{E}$ of the poles of $\tilde{R}(\nu)$. The construction we carried out thus seems not to be intrinsic: it depends on Σ and λ. We shall, however, show below that it provides the analytical continuation of the resolvent $R(\nu)$. Let us recall that $\nu \to \psi(\nu)$ is holomorphic on D if, for each open bounded set U included in Ω, $\nu \mapsto \psi(\nu)_{|U}$ is itself holomorphic on D; notice that the uniqueness of the analytical continuation remains valid for functions with values in $H^1_{\mathrm{loc}}(\Omega)$.

COROLLARY 2. *The poles of the analytical continuation of $\nu \mapsto R(\nu)F(f, \nu)$ and those of $\nu \mapsto \tilde{R}(\nu)\tilde{F}(f, \nu)$ are the same.*

Proof. (i) For fixed λ and Σ, we put $\tilde{\varphi}_\nu = \tilde{R}(\nu)\tilde{F}(f, \nu)$ and by φ_ν we denote the function equal to $\tilde{\varphi}_\nu$ in $\tilde{\Omega}$ and to

$$\int_\Gamma \left\{\tilde{\varphi}_\nu(x)\frac{\partial}{\partial n_x}G_\nu(x,\cdot) - f(x)G_\nu(x,\cdot)\right\} ds_x \quad \text{in } \Omega\backslash\tilde{\Omega}. \tag{4}$$

For $\operatorname{Im}(\nu)>0$ and $\nu \notin V(\Sigma, \lambda)$, Proposition 1 shows that this function is nothing but $R(\nu)F(f)$; formula (4) together with the holomorphic properties of G_ν show that $\nu \mapsto \varphi_\nu$ is holomorphic for $\nu \in \mathbb{C}\backslash(\mathbb{R}^- \cup \tilde{E})$. It follows that φ_ν defines the analytical continuation of $R(\nu)F(f)$.

(ii) If $\tilde{R}(\nu)\tilde{F}(f, \nu)$ is holomorphic in the vicinity of $\hat{\nu}$, then by formula (4), φ_ν is also holomorphic; each pole of the analytical continuation of $R(\nu)F(f)$ is thus a pole of $\tilde{R}(\nu)\tilde{F}(f, \nu)$.

(iii) Assume on the other hand that the analytical continuation φ_ν of $R(\nu)F(f)$ is holomorphic in the vicinity of $\hat{\nu}$; as $\varphi_{\nu|\tilde{\Omega}} = \tilde{R}(\nu)\tilde{F}(f, \nu)$ $\forall \nu \notin \tilde{E}$, then $\tilde{R}(\nu)\tilde{F}(f, \nu)$ is holomorphic in the vicinity of $\hat{\nu}$. It follows that each pole of $\tilde{R}(\nu)\tilde{F}(f, \nu)$ is a pole of the analytical continuation of $R(\nu)F(f)$. □

We shall now show that a relevant choice of parameter λ allows the determination of the poles of $R(\nu)F(f)$ from those of $\tilde{R}(\nu)$.

This result is especially meaningful because it is much easier to compute the poles of $\tilde{R}(\nu)$ than those of $R(\nu)F(f)$. They are the values of ν for which -1 is an eigenvalue of $\tilde{S}(\nu)$; in other words, they are the solutions of the following nonlinear eigenvalue problem:

$$\int_{\tilde{\Omega}} \nabla\tilde{\varphi}\,\nabla\bar{\tilde{\psi}}\,d\omega + \lambda\int_\Sigma \tilde{\varphi}\bar{\tilde{\psi}}\,ds - \int_\Sigma \bar{\tilde{\psi}}(y)\int_\Gamma \tilde{\varphi}(x)\frac{\partial}{\partial n_x}\gamma_\nu(x, y)\,ds_x\,ds_y = \nu\int_{\tilde{\Omega}} \tilde{\varphi}\bar{\tilde{\psi}}\,d\omega. \tag{E_λ}$$

In the sequel, we shall devise a numerical method for its solution.

THEOREM 1. *For* $\operatorname{Im}(\lambda)>0$ *and* $\operatorname{Im}(\nu_*)<0$, *the following assertions are equivalent*:

(i) *ν_* is a pole of $\nu \mapsto \tilde{R}(\nu)$,*

(ii) *$\exists f \in L^2(\Gamma)$ such that ν_* is a pole of $\nu \mapsto R(\nu)F(f)$.*

Proof of (ii)-(i). We shall show that, if ν_* is a pole of $\tilde{R}(\nu)$, then $\tilde{F}(f, \nu_*)$ cannot belong to Im $(I+\tilde{S}(\nu_*))$ for all $f \in L^2(\Gamma)$. The proof is by contradiction: let us assume that

$$\forall f \in L^2(\Gamma),\ \tilde{F}(f, \nu_*) \in \operatorname{Im}(I+\tilde{S}(\nu_*)) = (\operatorname{Ker}(I+\tilde{S}^*(\nu_*))^{\perp}.$$

From formula (3), we have

$$(\tilde{S}^*(\nu_*)\tilde{\varphi} \,|\, \psi)_{H^1(\tilde{\Omega})} = -(1+\overline{\nu_*}) \int_{\tilde{\Omega}} \tilde{\varphi}\bar{\psi}\, d\omega + \bar{\lambda} \int_{\Sigma} \tilde{\varphi}\bar{\psi}\, ds \tag{5}$$
$$- \int_{\Sigma} \tilde{\varphi}(y) \int_{\Gamma} \bar{\psi}(x) \frac{\partial}{\partial n_x} \overline{\gamma_{\nu_*}}(x, y)\, ds_x\, ds_y.$$

Assume that $\tilde{\varphi}$ belongs to Ker $(I+\tilde{S}^*(\nu_*))$, then $\forall f \in L^2(\Gamma)$,

$$\int_{\Gamma} f\bar{\tilde{\varphi}}\, ds - \int_{\Sigma} \bar{\tilde{\varphi}}(y) \int_{\Gamma} f(x)\gamma_{\nu_*}(x, y)\, ds_x\, ds_y = 0, \quad \text{i.e.,}$$

$$\int_{\Gamma} f(x) \left\{ \bar{\tilde{\varphi}}(x) - \int_{\Sigma} \bar{\tilde{\varphi}}(y)\gamma_{\nu_*}(x, y)\, ds_y \right\} ds_x = 0;$$

thus

$$\tilde{\varphi}(x) = \int_{\Sigma} \tilde{\varphi}(y)\overline{\gamma_{\nu_*}}(x, y)\, ds_y \quad \text{on } \Gamma. \tag{6}$$

On the other hand, by formula (5), $\tilde{\varphi}$ is a solution of the following problem:

$$\Delta\tilde{\varphi} + \overline{\nu_*}\tilde{\varphi} = 0 \quad \text{in } \tilde{\Omega},$$
$$\frac{\partial\tilde{\varphi}}{\partial n} = \frac{\partial}{\partial n}\left(\int_{\Sigma} \tilde{\varphi}(y)\overline{\gamma_{\nu_*}}(\cdot, y)\, ds_y \right) \quad \text{on } \Gamma,$$
$$\frac{\partial\tilde{\varphi}}{\partial n} + \bar{\lambda}\tilde{\varphi} = 0 \quad \text{on } \Sigma.$$

Function $\int_{\Sigma} \tilde{\varphi}(y)\overline{\gamma_{\nu_*}}(\cdot, y)\, ds_y$ extends to the whole Ω'; it agrees with $\tilde{\varphi}$ in $\tilde{\Omega}$, since they both are solutions of $\Delta\psi + \overline{\nu_*}\psi = 0$ on $\tilde{\Omega}$, with the same Dirichlet and Neumann boundary conditions on Γ. As a consequence, by formula (6), $\tilde{\varphi}$ extends to Ω' and satisfies:

$$\Delta\tilde{\varphi} + \overline{\nu_*}\tilde{\varphi} = 0 \quad \text{in } \Omega',$$
$$\frac{\partial\tilde{\varphi}}{\partial n} + \bar{\lambda}\tilde{\varphi} = 0 \quad \text{on } \Sigma.$$

By Remark 2, it follows that $\tilde{\varphi} = 0$, and consequently Ker $(I+\tilde{S}^*(\nu_*)) = \{0\}$, which is inconsistent with (i). We thus proved the existence of f in $L^2(\Gamma)$ such that $\tilde{F}(f, \nu_*) \notin$ Im $(I+\tilde{S}(\nu_*))$, therefore, ν_* is a pole of $\tilde{R}(\nu)\tilde{F}(f, \nu)$; as a consequence by Corollary 2, it is a pole of $R(\nu)F(f)$.

(ii)-(i) Conversely, if ν_* is a pole of $R(\nu)F(f)$, then by Corollary 2, ν_* is a pole of $\tilde{R}(\nu)\tilde{F}(f, \nu)$; it is thus a pole of $\tilde{R}(\nu)$, since $\tilde{F}(f, \nu)$ depends holomorphically on ν. □

4. The localized finite element method.

4.1. Reduction to a bounded domain. We shall describe below an alternative way for the reduction of (P_ν) to a bounded domain. For the sake of definiteness, we shall choose the space dimension equal to 2; the method actually applies to any space dimension. For the time being we assume that Im $(\nu) > 0$.

Let Σ be a circle of radius r surrounding Γ, $\tilde{\Omega}$ the domain bounded by Γ and Σ, and $\check{\Omega}$ the domain exterior to Σ (see Fig. 4).

As before, we must find boundary conditions on Σ such that the solution $\tilde{u}$ of the problem set in $\tilde{\Omega}$ is the restriction to $\tilde{\Omega}$ of the solution of (Q_ν).

Let $\chi \in H^{1/2}(\Sigma)$, and consider the following auxiliary problem:

$$\text{(7)} \qquad \begin{aligned} &\Delta\check{u} + \nu\check{u} = 0 \quad \text{in } \check{\Omega}, \\ &\check{u} = \chi \quad \text{on } \Sigma. \end{aligned}$$

By $\mathcal{T}_\nu$ we denote the Calderon operator associated to problem (7), i.e., the continuous function $\mathcal{T}_\nu$:

$$\chi \mapsto \mathcal{T}_\nu(\chi) = \frac{\partial\check{u}}{\partial n}_{|\Sigma} \quad \text{from } H^{1/2}(\Sigma) \quad \text{to } H^{-1/2}(\Sigma).$$

Assume that $\tilde{u}$, defined on $\tilde{\Omega}$, satisfies the reduced wave equation; we can easily prove that condition:

$$\text{(8)} \qquad \frac{\partial\tilde{u}}{\partial n}_{|\Sigma} = -\mathcal{T}_\nu(\tilde{u}_{|\Sigma})$$

implies the analytical matching between $\tilde{u}$ and the solution $\mathcal{S}(\tilde{u}_{|\Sigma})$ of problem (7), with the Dirichlet datum $\chi = \tilde{u}_{|\Sigma}$. We are thus led to the following problem, set in the bounded domain $\tilde{\Omega}$:

$$(\tilde{\Pi}_\nu) \qquad \begin{aligned} &\text{Find } \tilde{u} \in H^1(\tilde{\Omega}), \quad \text{such that} \\ &\Delta\tilde{u} + \nu\tilde{u} = 0 \quad \text{in } \tilde{\Omega}, \\ &\frac{\partial\tilde{u}}{\partial n} = f \quad \text{on } \Gamma, \\ &\frac{\partial\tilde{u}}{\partial n} = -\mathcal{T}_\nu(\tilde{u}_{|\Sigma}) \quad \text{on } \Sigma. \end{aligned}$$

PROPOSITION 3. *Problem $(\tilde{\Pi}_\nu)$ has a unique solution $\tilde{u}$, which is nothing but the restriction to $\tilde{\Omega}$ of the solution u of (Q_ν).*

Proof. The restriction to $\tilde{\Omega}$ of the solution u of (Q_ν) is a solution of $(\tilde{\Pi}_\nu)$. Conversely, if $\tilde{u}$ is a solution of $(\tilde{\Pi}_\nu)$, then the function whose restriction to $\tilde{\Omega}$ is equal to $\tilde{u}$ and whose restriction to $\check{\Omega}$ is equal to $\mathcal{S}(\tilde{u}_{|\Sigma})$ is the solution of (Q_ν). □

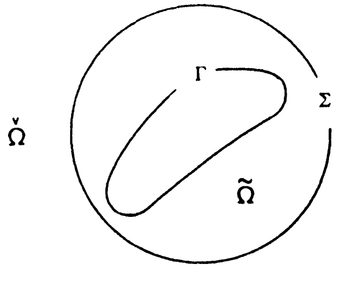


FIG. 4

An explicit formula for $\mathcal{T}_\nu$ can be obtained through diagonalization: we perform a Fourier expansion of the solution in the exterior domain. Let $\Phi_n = 1/\sqrt{2\pi}\, e^{in\theta}$, $n \in \mathbb{Z}$, and τ_n be the orthogonal projection on Φ_n in $L^2(-\pi, +\pi)$:

$$\tau_n(\chi) = \int_{-\pi}^{\pi} \chi \overline{\Phi_n}\, d\theta;$$

we have

$$\chi = \sum_{n\in\mathbb{Z}} \tau_n(\chi)\Phi_n \quad \forall \chi \in L^2(-\pi, +\pi) \quad \text{and}$$

(9)

$$\mathcal{T}_\nu(\chi) = -\sqrt{\nu} \sum_{n\in\mathbb{Z}} \tau_n(\chi) \frac{(H_n^{(1)})'(\sqrt{\nu}\, r)}{H_n^{(1)}(\sqrt{\nu}\, r)} \Phi_n.$$

The definition of the Hankel function $H_n^{(1)}$ is taken from [1]. A variational formulation of problem $(\tilde{\Pi}_\nu)$ then leads to

$$(I + \tilde{T}(\nu))\tilde{u} = \tilde{M}(f),$$

where

$$(\tilde{T}(\nu)\tilde{u}\,|\,v)_{H^1(\tilde{\Omega})} = -(\nu+1) \int_{\tilde{\Omega}} \tilde{u}\bar{v}\, d\omega - \sqrt{\nu} \sum_{n\in\mathbb{Z}} \tau_n(\tilde{u}_{|\Sigma}) \frac{(H_n^{(1)})'(\sqrt{\nu}\, r)}{H_n^{(1)}(\sqrt{\nu}\, r)} \overline{\tau_n(v_{|\Sigma})}, \quad \text{and}$$

(10)

$$(\tilde{M}(f)\,|\,v)_{H^1(\tilde{\Omega})} = \int_\Gamma f\bar{v}\, ds.$$

4.2. Analytical continuation of the resolvent. By Z_n we denote the set of the zeros of $H_n^{(1)}(\sqrt{\nu}\, r)$, and we set $Z = \bigcup_{n\in\mathbb{Z}} Z_n$.

PROPOSITION 4. *Formula* (10) *actually defines the analytical continuation of* $\nu \mapsto T(\nu)$ *to* $\mathbb{C}\backslash\{\mathbb{R}^- \cup Z\}$ *as an holomorphic function with values in* $\mathcal{L}(H^1(\tilde{\Omega}), H^1(\tilde{\Omega}))$.

Proof. The following formula holds:

$$\langle \mathcal{T}_\nu u, \bar{v}\rangle_{H^{-1/2}(\Sigma), H^{1/2}(\Sigma)} = -\sqrt{\nu} \sum_{n\in\mathbb{Z}} \tau_n(u)\overline{\tau_n(v)} \frac{(H_n^{(1)})'(\sqrt{\nu}\, r)}{H_n^{(1)}(\sqrt{\nu}\, r)}.$$

Let $z = \sqrt{\nu}\, r$; we have

$$\frac{(H_n^{(1)})'(z)}{H_n^{(1)}(z)} = \frac{(H_m^{(1)})'(z)}{H_m^{(1)}(z)} = \frac{-H_{m'}^{(1)}(z) + (m/z)H_m^{(1)}(z)}{H_m^{(1)}(z)} = -\frac{H_{m'}^{(1)}(z)}{H_m^{(1)}(z)} + \frac{m}{z},$$

with $m = |n|$, and $m' = |n| + 1$. From [1], we have

$$\frac{H_{m'}^{(1)}(z)}{H_m^{(1)}(z)} \sim 2\frac{m}{z} \quad \text{for } m \to \infty.$$

It follows that

(11)

$$\frac{(H_n^{(1)})'(z)}{H_n^{(1)}(z)} \sim -\frac{|n|}{z}, \quad \text{for } |n| \to \infty,$$

and consequently,

$$\begin{aligned} &|\langle \mathcal{T}_\nu u, \bar{v}\rangle_{H^{-1/2}(\Sigma), H^{1/2}(\Sigma)}| \\ &\leqq \sqrt{|\nu|} \left(\sum_{n\in\mathbb{Z}} |\tau_n(u)|^2 \frac{1}{|n|} \left| \frac{(H_n^{(1)})'(\sqrt{\nu}\, r)}{H_n^{(1)}(\sqrt{\nu}\, r)} \right|^2 \right)^{1/2} \left(\sum_{n\in\mathbb{Z}} |\tau_n(v)|^2 |n| \right)^{1/2}, \\ &\leqq C(r) \|u\|_{H^{1/2}(\Sigma)} \|v\|_{H^{1/2}(\Sigma)}. \end{aligned}$$

As a consequence, the series expansion of $\mathcal{T}_\nu u$ converges uniformly over any compact subset of $\mathbb{C}\backslash\mathbb{R}^-$, and the continuation of $\mathcal{T}_\nu$ that it defines is continuous from $H^{1/2}(\Sigma)$ into $H^{-1/2}(\Sigma)$. The statement of the proposition follows. □

As operator $\tilde{T}(\nu)$ is not completely continuous, Steinberg's theorem no longer applies; the following results hold, however.

PROPOSITION 5. *$I+\tilde{T}(\nu)$ is a Fredholm operator of index* 0 *in* $H^1(\tilde{\Omega})$.

Proof. We shall prove that $\tilde{T}(\nu)$ is the sum of a completely continuous operator and of an operator whose real part is positive. Clearly, it is enough to prove this result for operator $U(\nu)$ defined by

$$(U(\nu)u\,|\,v)_{H^1(\tilde{\Omega})}=\langle \mathcal{T}_\nu u,\,\bar{v}\rangle_{H^{-1/2}(\Sigma),H^{1/2}(\Sigma)}.$$

Formula (11) shows that $\operatorname{Re}((H_n^{(1)})'(\sqrt{\nu}\,r)/H_n^{(1)}(\sqrt{\nu}\,r))$ is negative beyond some rank N; since

$$(U(\nu)u\,|\,u)_{H^1(\tilde{\Omega})}=-\sqrt{\nu}\left(\sum_{n=-N+1}^{N-1}|\tau_n(u)|^2\frac{(H_n^{(1)})'(\sqrt{\nu}\,r)}{H_n^{(1)}(\sqrt{\nu}\,r)}+\sum_{|n|\geqq N}|\tau_n(u)|^2\frac{(H_n^{(1)})'(\sqrt{\nu}\,r)}{H_n^{(1)}(\sqrt{\nu}\,r)}\right),$$

the statement of the proposition follows. □

From now on we shall denote $(I+\tilde{T}(\nu))^{-1}$ by $\tilde{\mathcal{R}}(\nu)$.

PROPOSITION 6. *For all $f\in L^2(\Gamma)$, function $\tilde{\mathcal{R}}(\nu)\tilde{M}(f)$ extends meromorphically to $\mathbb{C}\backslash\mathbb{R}^-$; moreover each pole of $\tilde{\mathcal{R}}(\nu)\tilde{M}(f)$ is a pole of $\tilde{R}(\nu)$.*

Proof. (i) From Propositions 1 and 3, the following identity holds for any $\nu\notin V(\Sigma,\lambda)$ with $\operatorname{Im}(\nu)>0$:

$$\tilde{R}(\nu)\tilde{F}(f,\nu)=\tilde{\mathcal{R}}(\nu)\tilde{M}(f),\qquad \forall f\in L^2(\Gamma).$$

This identity actually defines the analytical continuation of $\tilde{\mathcal{R}}(\nu)\tilde{M}(f)$ to $\mathbb{C}\backslash\mathbb{R}^-$, for all $f\in L^2(\Gamma)$.

(ii) The last assertion of the Proposition follows then from Theorem 1. □

THEOREM 2. (i) *$\tilde{\mathcal{R}}(\nu)$ extends meromorphically to $\mathbb{C}\backslash\mathbb{R}^-$.*

(ii) *For* $\operatorname{Im}(\nu_*)<0$, *the two following statements are equivalent: ν_* is a pole of $\tilde{\mathcal{R}}(\nu)$, and $\exists f\in L^2(\Gamma)$ such that ν_* is a pole of $\tilde{\mathcal{R}}(\nu)\tilde{M}(f)$.*

(iii) *For* $\operatorname{Im}(\nu_*)<0$ *and* $\operatorname{Im}(\lambda)>0$, *$\tilde{\mathcal{R}}(\nu)$ has the same poles as $\tilde{R}(\nu)$.*

Proof. (i) If ν_* is a singular value of $\tilde{\mathcal{R}}(\nu)$, then $\exists f\in L^2(\Omega)$ such that ν_* is a pole of $\tilde{\mathcal{R}}(\nu)M(f)$. As in the proof of Theorem 1, we shall assume that $\forall f\in L^2(\Gamma)$, $\tilde{M}(f)\in\operatorname{Im}(I+\tilde{T}(\nu_*))=(\operatorname{Ker}(I+\tilde{T}^*(\nu_*)))^\perp$. From formula (10), we get

$$(12)\quad (\tilde{T}^*(\nu_*)\tilde{u}\,|\,v)_{H^1(\tilde{\Omega})}=-(\mu_*+1)\int_{\tilde{\Omega}}\tilde{u}\bar{v}\,d\omega-\sqrt{\mu_*}\sum_{n\in\mathbb{Z}}\tau_n(\tilde{u}_{|\Sigma})\frac{(H_n^{(2)})'(\sqrt{\mu_*}\,r)}{H_n^{(2)}(\sqrt{\mu_*}\,r)}\overline{\tau_n(v_{|\Sigma})},$$

with $\mu_*=\overline{\nu_*}$. Moreover, if $\tilde{u}\in\operatorname{Ker}(I+\tilde{T}^*(\nu_*))$, then

$$\forall f\in L^2(\Gamma),\qquad (\tilde{M}(f)\,|\,\tilde{u})_{H^1(\tilde{\Omega})}=0,\quad \text{i.e.,}$$

$$\int_\Gamma f\tilde{u}\,ds=0.$$

It follows that $\tilde{u}_{|\Gamma}=0$. From (12), we deduce that $\tilde{u}$ is a solution of the following problem:

$$(13)\qquad \begin{aligned}&\Delta\tilde{u}+\mu_*\tilde{u}=0\quad\text{in }\tilde{\Omega},\\&\tilde{u}_{|\Gamma}=0,\\&\frac{\partial\tilde{u}}{\partial n}=-\mathcal{T}^2_{\mu_*}(\tilde{u}_{|\Sigma})\quad\text{on }\Sigma,\end{aligned}$$

where

$$\mathcal{T}^2_{\mu_*}(\chi)=-\sqrt{\mu_*}\sum_{n\in\mathbb{Z}}\tau_n(\chi)\frac{(H_n^{(2)})'(\sqrt{\mu_*}\,r)}{H_n^{(2)}(\sqrt{\mu_*}\,r)}\Phi_n.$$

Since $\operatorname{Im}(\mu_*) > 0$, the functions $H_n^{(2)}(\sqrt{\mu_*}\, r)$ decrease exponentially at infinity; therefore, $\mathcal{T}^2_{\mu_*}$ is nothing the Calderon operator associated with the problem

$$\text{Find } \check{u} \in H^1(\check{\Omega}), \quad \text{such that}$$

$$\Delta \check{u} + \mu_* \check{u} = 0 \quad \text{in } \check{\Omega},$$

$$\check{u} = \chi \quad \text{on } \Sigma,$$

and thus $\tilde{u}$ is the restriction to $\tilde{\Omega}$ of the solution of the following coercive problem:

$$\text{Find } u \in H^1(\Omega), \quad \text{such that}$$

$$\Delta u + \mu_* u = 0 \quad \text{in } \Omega,$$

$$u_{|\Gamma} = 0.$$

As a consequence, u, and thus $\tilde{u}$, vanish. We actually proved that $\operatorname{Ker}(I + \tilde{T}^*(\nu_*)) = \{0\}$, which is inconsistent with the hypothesis.

(ii) From Theorem 1 and Proposition 6, we infer that the poles of $\tilde{\mathcal{R}}(\nu)\tilde{M}(f)$ are poles of $\tilde{R}(\nu)$; as a consequence, $\tilde{\mathcal{R}}(\nu)$ has only isolated singularities which are poles of $\tilde{\mathrm{R}}(\nu)$. Moreover, if ν_* is a singular value of $\tilde{\mathcal{R}}(\nu)$, then -1 is an eigenvalue of $\tilde{T}(\nu)$, of finite multiplicity by Proposition 5. It follows (Kato [3, p. 574]) that $\tilde{\mathcal{R}}(\nu)$ is meromorphic in the vicinity of ν_*, and consequently on $\mathbb{C}\backslash\mathbb{R}^-$.

(iii) Finally it is clear that, if ν_* is a pole of $\tilde{\mathcal{R}}(\nu)\tilde{M}(f)$ for some $f \in L^2(\Omega)$, it is thus a pole of $\tilde{\mathcal{R}}(\nu)$. □

Remark 3. What we just proved is that the scattering frequencies are also the solutions of the following nonlinear eigenvalue problem:

$$\text{(E)} \qquad \int_{\tilde{\Omega}} \nabla \tilde{u}\, \nabla \bar{\tilde{v}}\, d\omega - \sum_{n \in \mathbb{Z}} \tau_n(\tilde{u}_{|\Sigma}) \frac{(H_n^{(1)})'(\sqrt{\nu}\, r)}{H_n^{(1)}(\sqrt{\nu}\, r)} \overline{\tau_n(\tilde{v}_{|\Sigma})} = \nu \int_{\tilde{\Omega}} \tilde{u} \bar{\tilde{v}}\, d\omega.$$

This is an alternative formulation to Problem (E_λ), a condition for this last formulation to hold is that Σ be a circle.

4.3. The set of the zeros of the Hankel functions. Up to now, $\tilde{T}(\nu)$ has only been defined outside the set Z of the zeros of the Hankel functions; it is thus necessary to know whether these zeros are scattering frequencies or not. Let V_n be the operator defined by

$$(V_n u \mid v)_{H^1(\tilde{\Omega})} = \tau_n(u)\overline{\tau_n(v)}.$$

PROPOSITION 7. *Operator* V_n *is of rank* 1.

Proof. By Riesz' theorem, there exists a unique e'_n in $H^1(\tilde{\Omega})$ such that

$$(e'_n \mid v)_{H^1(\tilde{\Omega})} = \int_\Sigma \Phi_n \bar{v}\, ds.$$

It follows that

$$(V_n u \mid v)_{H^1(\tilde{\Omega})} = \left(\int_\Sigma u \overline{\Phi_n}\, ds\right)\left(\int_\Sigma \Phi_n \bar{v}\, ds\right) = \left(\left(\int_\Sigma u \overline{\Phi_n}\, ds\right) e'_n \,\middle|\, v\right)_{H^1(\tilde{\Omega})}, \quad \text{i.e.,}$$

$$V_n u = \left(\int_\Sigma u \overline{\Phi_n}\, ds\right) e'_n. \qquad □$$

Remark 4. The function e'_n exhibited in the proof of the preceding proposition is actually the solution of the following coercive problem:

$$(14)\qquad \begin{aligned} &-\Delta e'_n + e'_n = 0 \quad \text{in } \tilde{\Omega},\\ &\frac{\partial e'_n}{\partial \nu} = \Phi_n \quad \text{on } \Sigma,\\ &\frac{\partial e'_n}{\partial \nu} = 0 \quad \text{on } \Gamma. \end{aligned}$$

From now on we shall scale e'_n in $H^1(\tilde{\Omega})$ and set $e_n = e'_n/\|e'_n\|_{H^1(\tilde{\Omega})}$.

Let E_n be the orthogonal supplementary to $\{e_n\}$ in $H^1(\tilde{\Omega})$:

$$H^1(\tilde{\Omega}) = \{e_n\} \overset{\perp}{\oplus} E_n,$$

and accordingly the matrix decomposition of operator $(I + \tilde{T}(\nu))$:

$$\begin{pmatrix} A_n(\nu) & B_n(\nu) \\ C_n(\nu) & D_n(\nu) \end{pmatrix}.$$

By ν^* we denote a zero of $H_n^{(1)}(\sqrt{\nu}\, r)$; from formula (10) we infer that $A_n(\nu)$ expands as

$$A_n(\nu) = A'_n(\nu) + \frac{V_n}{\varepsilon_n(\nu)}, \quad \text{where } \varepsilon_n(\nu) = \frac{H_n^{(1)}(\sqrt{\nu}\, r)}{(H_n^{(1)})'(\sqrt{\nu}\, r)}$$

tends to zero when $\nu \to \nu^*$, and $A'_n(\nu)$ depends continuously on ν in the vicinity of ν^*. Similarly, operators $B_n(\nu)$, $C_n(\nu)$, and $D_n(\nu)$ depend continuously on ν.

THEOREM 3. *A zero ν^* of $H_n^{(1)}(\sqrt{\nu}\, r)$ is a pole of $\tilde{\mathcal{R}}(\nu)$ if and only if operator $D_n(\nu)$ is not invertible at $\nu = \nu^*$.*

Proof. Since Z and the set $\tilde{E}$ of the poles of $\tilde{\mathcal{R}}(\nu)$ are discrete sets in $\mathbb{C}\backslash\mathbb{R}^-$, there exists a neighborhood $\mathcal{V}$ of ν^* which does not contain any other point of $Z \cup \tilde{E}$.

(i) Assume that $D_n(\nu^*)$ is invertible; in $\mathcal{V}\backslash\{\nu^*\}$, $(I + \tilde{T}(\nu))$ is also invertible and the matrix decomposition of its inverse has the following form:

$$\begin{pmatrix} \mathcal{A}_n(\nu) & \mathcal{B}_n(\nu) \\ \mathcal{C}_n(\nu) & \mathcal{D}_n(\nu) \end{pmatrix}, \quad \text{with}$$

$$\begin{aligned} &\mathcal{A}_n(\nu) = (A_n(\nu) - B_n(\nu) D_n(\nu)^{-1} C_n(\nu))^{-1},\\ &\mathcal{B}_n(\nu) = -\mathcal{A}_n(\nu) B_n(\nu) D_n(\nu)^{-1},\\ &\mathcal{C}_n(\nu) = -\mathcal{A}_n(\nu) D_n(\nu)^{-1} C_n(\nu), \quad \text{and}\\ &\mathcal{D}_n(\nu) = -A_n(\nu)(C_n(\nu) B_n(\nu) - A_n(\nu) D_n(\nu))^{-1}, \end{aligned}$$

according to the fact that $\mathcal{A}_n(\nu)$, which is a multiplication operator, interchanges with any other. When ν tends to ν^*, then $\mathcal{A}_n(\nu)$, $\mathcal{B}_n(\nu)$, and $\mathcal{C}_n(\nu)$ vanish; moreover,

$$\mathcal{D}_n(\nu) = -(\varepsilon_n(\nu) A'_n(\nu) + V_n)(\varepsilon_n(\nu) C_n(\nu) B_n(\nu) - (\varepsilon_n(\nu) A'_n(\nu) + V_n) D_n(\nu))^{-1}$$

tends to $D_n(\nu^*)^{-1}$ when $\nu \to \nu^*$. It follows that $\tilde{\mathcal{R}}(\nu)$ can be continuously extended at point ν^* as

$$\tilde{\mathcal{R}}(\nu) = \begin{pmatrix} 0 & 0 \\ 0 & D_n(\nu^*)^{-1} \end{pmatrix}.$$

(ii) Conversely, assume that $\nu^* \notin \tilde{E}$, then there exists $\mathscr{B}_n(\nu)$, $\mathscr{C}_n(\nu)$, and $\mathscr{D}_n(\nu)$ such that

$$\left(A_n'(\nu)+\frac{V_n}{\varepsilon_n(\nu)}\right)\mathscr{B}_n(\nu)+B_n(\nu)\mathscr{D}_n(\nu)=0. \tag{15}$$

$$C_n(\nu)\mathscr{B}_n(\nu)+D_n(\nu)\mathscr{D}_n(\nu)=I_{|E_n}. \tag{16}$$

When $\nu \to \nu^*$, $\mathscr{B}_n(\nu)$ vanishes by formula (15), and consequently (16) implies:

$$D_n(\nu^*)\mathscr{D}_n(\nu^*)=I_{|E_n},$$

showing that $D_n(\nu^*)$ is invertible. □

Remark 5. For practical purposes it is worthwhile to give an explicit expression for operator

$$\begin{pmatrix} 0 & 0 \\ 0 & D_n(\nu) \end{pmatrix},$$

which we shall now simply denote by $D_n(\nu)$. Assume u and v belong to $H^1(\tilde{\Omega})$, then

$$\begin{aligned}(D_n(\nu)u\,|\,v)_{H^1(\tilde{\Omega})} &= ((I+\tilde{T}(\nu))u\,|\,v)_{H^1(\tilde{\Omega})}-((I+\tilde{T}(\nu))u\,|\,e_n)_{H^1(\tilde{\Omega})}(e_n\,|\,v)_{H^1(\tilde{\Omega})} \\ &= \int_{\tilde{\Omega}} \nabla u\,\nabla\bar{v}\,d\omega-\nu\int_{\tilde{\Omega}} u\bar{v}\,d\omega-\sum_{k\in\mathbb{Z}}\frac{\tau_k(u)\overline{\tau_k(v)}}{\varepsilon_k(\nu)} \\ &\quad -\left(\int_{\tilde{\Omega}}\nabla u\,\nabla\overline{e_n}\,d\omega-\nu\int_{\tilde{\Omega}} u\overline{e_n}\,d\omega-\sum_{k\in\mathbb{Z}}\frac{\tau_k(u)\overline{\tau_k(e_n)}}{\varepsilon_k(\nu)}\right)\overline{\tau_n(v)}.\end{aligned}$$

As $\tau_k(e_n)=\delta_{kn}$, we obtain

$$\begin{aligned}(D_n(\nu)u\,|\,v)_{H^1(\tilde{\Omega})} &= \int_{\tilde{\Omega}} \nabla u\,\nabla\bar{v}\,d\omega-\nu\int_{\tilde{\Omega}} u\bar{v}\,d\omega-\sum_{k\neq n}\frac{\tau_k(u)\overline{\tau_k(v)}}{\varepsilon_k(\nu)} \\ &\quad -\frac{\tau_n(u)\overline{\tau_n(v)}}{\|e_n'\|}+(\nu+1)\left(\int_{\tilde{\Omega}} u\overline{e_n}\,d\omega\right)\overline{\tau_n(v)}.\end{aligned} \tag{17}$$ □

5. Approximation of the scattering frequencies and of the associated resonant states. In the preceding sections, we described two different methods for reducing the determination of the scattering frequencies to a nonlinear eigenvalue problem, for a compact operator. We introduce at first, the internal approximation of this problem, then we perform an asymptotic expansion of the resolvent in the vicinity of a scattering frequency; the approximate location of the resonant frequencies and of the associated maxima of the response of the system follow.

5.1. Discretisation of the nonlinear eigenvalue problem. By V_h we denote a finite-dimensional subspace of $H^1(\Omega)$, following, for example, from a finite element discretization. If $(\cdot|\cdot)_h$ denotes the scalar product in V_h, we define $\mathbb{J}_h$ and $\mathbb{S}_h(\nu)$ by the following formulas:

$$(\mathbb{J}_h\varphi_h\,|\,\psi_h)_h=\int_\Omega \nabla\varphi_h\,\nabla\overline{\psi_h}\,d\omega+\int_\Omega \varphi_h\overline{\psi_h}\,d\omega, \tag{18}$$

$$\begin{aligned}(\mathbb{S}_h(\nu)\varphi_h\,|\,\psi_h)_h &= -(1+\nu)\int_\Omega \varphi_h\overline{\psi_h}\,d\omega+\lambda\int_\Sigma \varphi_h\overline{\psi_h}\,ds \\ &\quad -\int_\Sigma \overline{\psi_h}(y)\int_\Gamma \varphi_h(x)\frac{\partial\gamma_\nu}{\partial n_x}(x,y)\,ds_x\,ds_y \quad \forall\varphi_h,\psi_h\in V_h.\end{aligned} \tag{19}$$

The approximate determination of the scattering frequencies by the method of coupling between variational formulation and integral representation consists in solving the set of equations:

$$\mu_h(\nu_{h*}) = -1, \tag{20}$$

where the $\mu_h(\nu)$ are the solutions of the following generalized eigenvalue problem:

$$\mathbb{S}_h(\nu)e_h(\nu) = \mu_h(\nu)\mathbb{J}_h e_h(\nu).$$

Remark 6. Using the localized finite element method consists in replacing $\mathbb{S}_h(\nu)$ defined at formula (19) by

$$(\mathbb{S}_h(\nu)\varphi_h \mid \psi_h)_h = -(1+\nu)\int_\Omega \varphi_h\overline{\psi_h}\,d\omega - \sum_{n\in\mathbb{Z}} \tau_n(\varphi_{h|\Sigma})\frac{(H_n^{(1)}(\sqrt{\nu}\mathrm{r}))'}{H_n^{(1)}(\sqrt{\nu}\mathrm{r})}\overline{\tau_n(\psi_{h|\Sigma})}. \tag{21}$$

The solution of (20) by Newton's method requires the derivative $\mu_h'(\nu)$ of $\mu_h(\nu)$. Assume, for example, that $\mu_h(\nu)$ is a simple eigenvalue of $\mathbb{J}_h^{-1}\mathbb{S}_h(\nu)$, then

$$\mu_h'(\nu) = (\mathbb{S}_h'(\nu)e_h(\nu) \mid g_h(\nu))_h,$$

where $g_h(\nu)$ is the associated eigenvector of the adjoint operator, i.e., satisfying:

$$\mathbb{S}_h^*(\nu)g_h(\nu) = \overline{\mu_h}(\nu)\mathbb{J}_h g_h(\nu), \quad \text{and}$$

$$(\mathbb{J}_h e_h(\nu) \mid g_h(\nu))_h = 1.$$

In the sequel, we shall give expressions for the evaluation of $\mu_h'(\nu)$ in more complicated situations.

The next step will consist of expanding the solution of the diffraction problem in the vicinity of a scattering frequency. Let

$$(F_h(f,\nu) \mid \psi_h)_h = \int_\Gamma f\overline{\psi_h}\,ds - \int_\Sigma \overline{\psi_h}(y)\int_\Gamma f(x)\frac{\partial\gamma_\nu}{\partial n_x}(x,y)\,ds_x\,ds_y; \tag{22}$$

the use of the coupling method between variational formulation and integral representation leads then to expand the solution $\varphi_h(\nu)$ of problem

$$(\mathbb{J}_h + \mathbb{S}_h(\nu))\varphi_h(\nu) = F_h(f,\nu). \tag{23}$$

Remark 7. When using the localized finite element method, we replace $F_h(f,\nu)$ by $F_h(f)$, defined by

$$(F_h(f) \mid \psi_h)_h = \int_\Gamma f\overline{\psi_h}\,ds. \tag{24}$$

In the sequel, we shall only consider the *approximate problem*, consequently we shall omit index "h"; ν_* will denote the approximate scattering frequency under consideration. We shall first construct the Laurent expansion of the approximate resolvent $\mathbb{R}(\nu) = [\mathbb{J} + \mathbb{S}(\nu)]^{-1}$ in the vicinity of ν_*: this is an application of the perturbation theory of Kato [3]. An explicit calculation of the expansion of the solution of problem (23) will then be produced, which will provide an approximation of the resonant frequency.

5.2. Canonical form of the resolvent. Let us begin with some notations. The following expansion holds in the vicinity of ν_*:

$$\mathbb{S}(\nu) = \mathbb{S} + \sum_{n=1}^{\infty}(\nu-\nu_*)^n\mathbb{S}^{(n)};$$

or, equivalently,

$$\mathbb{U}(\nu)=\mathbb{U}+\sum_{n=1}^{\infty}(\nu-\nu_*)^n\mathbb{U}^{(n)},\quad\text{with}$$
(25)
$$\mathbb{U}(\nu)=\mathbb{J}^{-1}\mathbb{S}(\nu),\quad \mathbb{U}=\mathbb{J}^{-1}\mathbb{S}\quad\text{and}\quad \mathbb{U}^{(n)}=\mathbb{J}^{-1}\mathbb{S}^{(n)}\quad \forall n>0.$$

Since $\mathbb{U}(\nu)$ is a holomorphic function of ν, the number of distinct eigenvalues of $\mathbb{U}(\nu)$ is a constant independent of ν, apart from isolated values of ν which are called exceptional points (ν_* may be such an exceptional point). Let $\mu_k(\nu)$, $k=1, K$ (respectively, μ_i, $i=1, I$ where $I\leqq K$) denote the eigenvalues of $\mathbb{U}(\nu)$ for $\nu\neq\nu_*$ (respectively, of $\mathbb{U}$), and $\mathbb{P}_k(\nu)$ (respectively, $\mathbb{P}_i$) the associated eigenprojections. The spectrum of $\mathbb{U}(\nu)$ for $\nu\neq\nu_*$ can be divided into I "μ_i-groups" $\{\mu_k(\nu), k\in\mathcal{J}_i\}$ such that:

$$\mu_k(\nu)\rightarrow\mu_i\quad\text{as } \nu\rightarrow\nu_*,\quad k\in\mathcal{J}_i,\quad i=1, I.$$

A scattering frequency ν_* makes -1 an eigenvalue of $\mathbb{U}$; we can choose, for example, $\mu_1=-1$ and $\mathcal{J}_1=\{1,\cdots,s\}$. $\mathbb{P}=\mathbb{P}_1$ denotes the eigenprojection for the eigenvalue -1. The "(-1)-group" is then defined by

$$\mu_k(\nu)\rightarrow -1\quad\text{as } \nu\rightarrow\nu_*,\quad k=1, s.$$

From now on we shall assume that $\mathbb{U}(\nu)$ is diagonalizable in the vicinity of ν_*. The canonical form of the resolvent is thus written:

$$\mathbb{R}(\nu)=\mathbb{R}_S(\nu)+\mathbb{R}_R(\nu),\quad \nu\neq\nu_*,\quad\text{where}$$
(26)
$$\mathbb{R}_S(\nu)=\sum_{k=1}^{s}\frac{\mathbb{P}_k(\nu)\mathbb{J}^{-1}}{1+\mu_k(\nu)}\quad\text{and}\quad \mathbb{R}_R(\nu)=\sum_{k=s+1}^{K}\frac{\mathbb{P}_k(\nu)\mathbb{J}^{-1}}{1+\mu_k(\nu)}.$$

$\mathbb{R}_S(\nu)$ is the singular part of the resolvent and $\mathbb{R}_{\text{R}}(\nu)$ the reduced resolvent, which is holomorphic near $\nu=\nu_*$. As a matter of fact

(27) $$\mathbb{R}_R(\nu)=\mathbb{Q}\mathbb{J}^{-1}+0(\nu-\nu_*),\quad\text{where } \mathbb{Q}=\sum_{i=2}^{I}\frac{\mathbb{P}_i}{1+\mu_i}.$$

Notice that $\mathbb{Q}$ is the operator which associates to a given G the only solution X of

$$(\mathbb{I}+\mathbb{U})X=(\mathbb{I}-\mathbb{P})G\quad\text{and}\quad \mathbb{P}X=0.$$

5.3. Expansion of the eigenvalues and eigenprojections of the (−1)-group. The question is now to compute the expansion of the singular part of the resolvent. Let us first assume that the (-1)-group has only one element $\mu_1(\nu)$ (i.e., $s=1$). In this simple case, the eigenvalue $\mu_1(\nu)$ and the eigenprojection $\mathbb{P}_1(\nu)$ are holomorphic near $\nu=\nu_*$:

(28) $$\begin{cases}\mu_1(\nu)=-1+(\nu-\nu_*)\mu^{(1)}+(\nu-\nu_*)^2\mu^{(2)}+0(\nu-\nu_*)^3,\\ \mathbb{P}_1(\nu)=\mathbb{P}+(\nu-\nu_*)\mathbb{P}^{(1)}+0(\nu-\nu_*)^2,\end{cases}$$

with

$$\mu^{(1)}=\frac{1}{M}\operatorname{tr}[\mathbb{U}^{(1)}\mathbb{P}],\qquad \mu^{(2)}=\frac{1}{M}\operatorname{tr}[\mathbb{U}^{(2)}\mathbb{P}-\mathbb{U}^{(1)}\mathbb{Q}\mathbb{U}^{(1)}\mathbb{P}],$$
(29)
$$\mathbb{P}^{(1)}=-\mathbb{P}\mathbb{U}^{(1)}\mathbb{Q}-\mathbb{Q}\mathbb{U}^{(1)}\mathbb{P},$$

where M is the multiplicity of the eigenvalue -1 of $\mathbb{U}$ (which agrees here with the multiplicity of the eigenvalue $\mu_1(\nu)$ of $\mathbb{U}(\nu)$). If $\mu^{(1)}$ is different from zero, the expansion of the singular part of the resolvent writes

(30) $$\mathbb{R}_S(\nu)=\frac{1}{(\nu-\nu_*)}\left[\frac{\mathbb{P}}{\mu^{(1)}}\right]\mathbb{J}^{-1}+\left[\frac{\mathbb{P}^{(1)}}{\mu^{(1)}}-\frac{\mu^{(2)}\mathbb{P}}{\mu^{(1)2}}\right]\mathbb{J}^{-1}+0(\nu-\nu_*).$$

If $\mu^{(1)}=0$, we have to find the order l of the first nonvanishing term in the expansion (28) of $\mu_1(\nu)$. An expansion similar to (30), beginning, however, at order $(\nu-\nu_*)^{-l}$ holds.

Let us now study the case where the (-1)-group has several elements (i.e., $s>1$). By $\mathbb{P}'(\nu)$ we denote the total projection associated with the (-1)-group:

$$\mathbb{P}'(\nu)=\sum_{k=1}^{s}\mathbb{P}_k(\nu), \qquad \nu\neq\nu_*.$$

As we assumed that $\mathbb{U}$ is diagonalizable, -1 is a semisimple eigenvalue of $\mathbb{U}$, and we can assert that the operator

$$\dot{\mathbb{U}}(\nu)=\frac{1}{\nu-\nu_*}(\mathbb{I}+\mathbb{U}(\nu))\mathbb{P}'(\nu)$$

is holomorphic near $\nu=\nu_*$:

$$\dot{\mathbb{U}}(\nu)=\dot{\mathbb{U}}+\sum_{n=1}^{\infty}(\nu-\nu_*)^n\dot{\mathbb{U}}^{(n)}, \quad \text{with}$$

(31) $\dot{\mathbb{U}}=\mathbb{P}\mathbb{U}^{(1)}\mathbb{P}$, and

(32) $\dot{\mathbb{U}}^{(1)}=\mathbb{P}\mathbb{U}^{(2)}\mathbb{P}-\mathbb{P}\mathbb{U}^{(1)}\mathbb{P}\mathbb{U}^{(1)}\mathbb{Q}-\mathbb{P}\mathbb{U}^{(1)}\mathbb{Q}\mathbb{U}^{(1)}\mathbb{P}-\mathbb{Q}\mathbb{U}^{(1)}\mathbb{P}\mathbb{U}^{(1)}\mathbb{P}.$

Let $\dot\mu_k(\nu)$, $k=1,s$ be the eigenvalues of $\dot{\mathbb{U}}(\nu)$ for $\nu\neq\nu_*$, and $\dot{\mathbb{P}}_k(\nu)$ the associated eigenprojections. We obviously have:

(33) $\mu_k(\nu)=-1+(\nu-\nu_*)\dot\mu_k(\nu)$ and $\dot{\mathbb{P}}_k(\nu)=\mathbb{P}_k(\nu)$, $\nu\neq\nu_*$, $k=1,s$.

These eigenvalues can be divided into several "$\dot\mu_j$-groups" where the $\dot\mu_j$ are the eigenvalues of $\dot{\mathbb{U}}$, their number being less or equal to s. For the sake of simplicity, we shall assume that ν_* is not an exceptional point for $\dot{\mathbb{U}}(\nu)$, i.e., that $\dot{\mathbb{U}}$ has exactly s eigenvalues. Consequently, the $\dot\mu_k(\nu)$ and $\dot{\mathbb{P}}_k(\nu)$ are holomorphic near $\nu=\nu_*$ and can be expanded as (28):

$$(34)\qquad \begin{cases}\dot\mu_k(\nu)=\dot\mu_k+(\nu-\nu_*)\dot\mu_k^{(1)}+0(\nu-\nu_*)^2,\\ \dot{\mathbb{P}}_k(\nu)=\dot{\mathbb{P}}_k+(\nu-\nu_*)\dot{\mathbb{P}}_k^{(1)}+0(\nu-\nu_*)^2,\end{cases}$$

with

$$(35)\qquad \begin{cases}\dot\mu_k^{(1)}=\dfrac{1}{M_k}\operatorname{tr}[\dot{\mathbb{U}}^{(1)}\dot{\mathbb{P}}_k],\\ \dot{\mathbb{P}}_k^{(1)}=-\dot{\mathbb{P}}_k\dot{\mathbb{U}}^{(1)}\dot{\mathbb{Q}}_k-\dot{\mathbb{Q}}_k\dot{\mathbb{U}}^{(1)}\dot{\mathbb{P}}_k,\end{cases}$$

where $\dot{\mathbb{P}}_k$ is the eigenprojection for the eigenvalue $\dot\mu_k$ of $\dot{\mathbb{U}}$, M_k its multiplicity (which is equal to the multiplicity of the eigenvalue $\mu_k(\nu)$ of $\mathbb{U}(\nu)$), and the $\dot{\mathbb{Q}}_k$ are given by:

$$\dot{\mathbb{Q}}_k=\sum_{\substack{j=1\\ j\neq k}}^{s}\frac{\dot{\mathbb{P}}_j}{\dot\mu_j-\dot\mu_k}, \qquad k=1,s.$$

If all the eigenvalues $\dot\mu_k$ are different from 0, $\mathbb{R}_S(\nu)$ thus writes:

$$(36)\qquad R_S(\nu)=\frac{1}{\nu-\nu_*}\left\{\sum_{k=1}^{s}\frac{\dot{\mathbb{P}}_k\mathbb{J}^{-1}}{\dot\mu_k}\right\}+\left\{\sum_{k=1}^{s}\left[\frac{\dot{\mathbb{P}}_k^{(1)}}{\dot\mu_k}-\frac{\dot\mu_k^{(1)}\dot{\mathbb{P}}_k}{\dot\mu_k^2}\right]\mathbb{J}^{-1}\right\}+0(\nu-\nu_*).$$

Otherwise a similar expansion beginning at order $-l=-\mathrm{Max}_{k=1,s}\,l(k)$, where $l(k)$ denotes the order of the first nonvanishing term in the expansion (34) of $\dot\mu_k(\nu)$, can be obtained.

We have thus studied the case where ν_* is an exceptional point for the eigenvalue -1 of $\mathbb{U}$ but not for the reduced operator $\dot{\mathbb{U}}(\nu)$. If ν_* is an exceptional point for at least one eigenvalue μ_j of $\dot{\mathbb{U}}$, this reduction process can go further on provided that $\dot{\mathbb{U}}$ is diagonalizable, and so on, as many times (but in finite number) as necessary.

5.4. Expansion of the solution in the vicinity of ν_*. We now come back to problem (23) for obtaining the expansion of the solution $\varphi(\nu)$ in the vicinity of ν_*. The expansion of the right-hand side is straightforward:

$$F(f,\nu)=\sum_{n=0}^{\infty}(\nu-\nu_*)^n F^{(n)}.$$

Since $\varphi(\nu)$ simply writes $\mathbb{R}(\nu)F(f,\nu)$, its expansion follows from the one of the resolvent $\mathbb{R}(\nu)$:

$$\mathbb{R}(\nu)=\sum_{n=-l}^{\infty}(\nu-\nu_*)\mathbb{R}^{(n)}.$$

We have

$$\varphi(\nu)=\sum_{n=-l}^{\infty}(\nu-\nu_*)^n\varphi^{(n)},\quad \text{with} \tag{37}$$

$$\varphi^{(-l)}=\mathbb{R}^{(-l)}F^{(0)}\quad\text{and}\quad \varphi^{(1-l)}=\mathbb{R}^{(1-l)}F^{(0)}+\mathbb{R}^{(-l)}F^{(1)}. \tag{38}$$

We shall give here the explicit calculation of the first two terms of this expansion; we shall restrict ourselves to the simple case where ν_* is not an exceptional point and $\mu^{(1)}$ (in (28)) differs from zero (formula (30)). Further cases can be treated exactly in the same way by using the proper expression of the expansion of $\mathbb{R}(\nu)$.

Let $e_m(=e_{1m})$, $m=1, M$ (respectively, e_{ij}, $j=1, M(i)$, $i=2, I$) be a basis of the eigenspace associated with the eigenvalue -1 (respectively, μ_i, $i=2, I$) of $\mathbb{U}$:

$$\mathbb{S}e_m=-\mathbb{J}e_m,\quad m=1,M\quad\text{and}\quad \mathbb{S}e_{ij}=\mu_i\mathbb{J}e_{ij},\quad j=1,M(i),\quad i=2,I,$$

and g_m (respectively, g_{ij}) the associated eigenvectors of the adjoint chosen such that:

$$(\mathbb{J}e_{ij}\,|\,g_{i'j'})=\delta_{ii'}\delta_{jj'},\quad j=1,M(i),\quad j'=1,M(i'),\quad i,i'=1,I.$$

The eigenprojections $\mathbb{P}$ and $\mathbb{P}_i$, $i=2, I$ can then be expressed in this basis:

$$\mathbb{P}X=\sum_{m=1}^{M}(\mathbb{J}X\,|\,g_m)e_m\quad\text{and}\quad \mathbb{P}_iX=\sum_{j=1}^{M(i)}(\mathbb{J}X\,|\,g_{ij})e_{ij},\quad i=2,I,$$

and $\mathbb{Q}$ (defined by (27)) writes:

$$\mathbb{Q}X=\sum_{i=2}^{I}\frac{1}{1+\mu_i}\sum_{j=1}^{M(i)}(\mathbb{J}X\,|\,g_{ij})e_{ij},$$

whence we infer the expression of the coefficients of the expansion of $\mu_1(\nu)$ (given by formula (29)):

$$\mu^{(1)}=\frac{1}{M}\sum_{m=1}^{M}(\mathbb{S}^{(1)}e_m\,|\,g_m);$$

$$\mu^{(2)}=\frac{1}{M}\sum_{m=1}^{M}\left[(\mathbb{S}^{(2)}e_m\,|\,g_m)-\sum_{i=2}^{I}\frac{1}{1+\mu_i}\sum_{j=1}^{M(i)}(\mathbb{S}^{(1)}e_m\,|\,g_{ij})(\mathbb{S}^{(1)}e_{ij}\,|\,g_m)\right].$$

From formulas (27) and (30) follow $\mathbb{R}^{(-1)}$ and $\mathbb{R}^{(0)}$, and finally the first two terms of the expansion of the solution:

$$\varphi^{(-1)}=\frac{1}{\mu^{(1)}}\sum_{m=1}^{M}(F^{(0)}|g_m)e_m \quad \text{and} \tag{39}$$

$$\begin{aligned}\varphi^{(0)}=&\frac{1}{\mu^{(1)}}\sum_{m=1}^{M}\left[\!\left[(F^{(1)}|g_m)-\frac{\mu^{(2)}}{\mu^{(1)}}(\mathrm{F}^{(0)}|g_m)-\sum_{i=2}^{I}\frac{1}{1+\mu_i}\sum_{j=1}^{M(i)}(F^{(0)}|g_{ij})(\mathbb{S}^{(1)}e_{ij}|g_m)\right]\!\right]e_m\\&+\sum_{i=2}^{I}\frac{1}{1+\mu_i}\sum_{j=1}^{M(i)}\left[\!\left[(F^{(0)}|g_{ij})-\frac{1}{\mu^{(1)}}\sum_{m=1}^{M}(F^{(0)}|g_m)(\mathbb{S}^{(1)}e_m|g_{ij})\right]\!\right]e_{ij}\end{aligned}\tag{40}$$

There is a simpler way for obtaining, in some respects, more general formulas, provided the existence of expansion (37) is assumed from the beginning; it will be worked out in the appendix.

5.5. Approximation of the resonant frequencies. The expansion (37) of the solution $\varphi(\nu)$ we obtained in the preceding paragraph makes it possible to compute an approximation of the (real) value ν_0 of the frequency where the maximum of $\|\varphi(\nu)\|$ is reached.

As a first approximation, we have

$$\varphi(\nu)\sim\frac{1}{(\nu-\nu_*)^l}(\varphi^{(-l)}+(\nu-\nu_*)\varphi^{(1-l)});$$

thus,

$$\|\varphi(\nu)\|^2\sim\frac{1}{(\alpha^2+\beta^2)^l}(\|X\|^2+2\alpha\ \mathrm{Re}\,(X\,|\,Y)+2\beta\ \mathrm{Im}\,(X\,|\,Y))=A(\alpha),$$

where $\nu-\nu_*=\alpha+i\beta$, $\varphi^{(-l)}=X$ and $\varphi^{(1-l)}=Y$.

The maximum of $A(\alpha)$ is reached when α is a solution of

$$\alpha^2(1-2l)\ \mathrm{Re}\,(X\,|\,Y)-\alpha(\|x\|^2+2\beta\ \mathrm{Im}\,(X\,|\,Y))+\beta^2\ \mathrm{Re}\,(X\,|\,Y)=0,$$

i.e., as a first approximation for

$$\alpha=\frac{\beta^2\ \mathrm{Re}\,(X\,|\,Y)}{l(\|X\|^2+2\beta\ \mathrm{Im}\,(X\,|\,Y))};$$

it follows that

$$\nu_0\sim\mathrm{Re}\,(\nu_*)+\frac{\mathrm{Re}\,(\varphi^{(-l)}\,|\,\varphi^{(1-l)})(\mathrm{Im}\,(\nu_*))^2}{l(\|\varphi^{(-l)}\|^2-2\ \mathrm{Im}\,(\nu_*)\ \mathrm{Im}\,(\varphi^{(-l)}|\varphi^{(1-l)}))}.\tag{41}$$

6. Some numerical results. In this section, we shall deal with a generalization of the previous problem: the coupling between an elastic structure and an unbounded acoustic fluid; numerical results will be presented, which show the efficiency of the method.

6.1. A simple coupled problem. The problem under consideration is the coupling between an infinite vibrating string and an elastic beam; the connection is realized through a density of springs (see Fig. 5). When the system is at rest, the beam is assumed rectilinear and parallel to the string; only simple bending motions will be

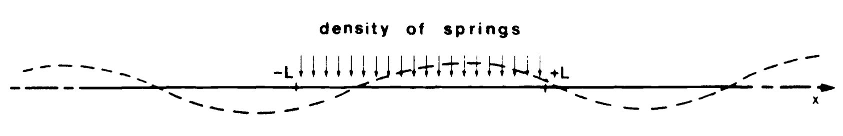


FIG. 5

considered. By $U(x, t)$ and $Y(x, t)$ we denote, respectively, the vertical displacement of a point on the string and the displacement of a point on the beam.

The system will be subject to a monochromatic incident wave of pulsation ω:

$$U_I(x, t) = \operatorname{Re}(u_I(x)\, e^{-i\omega t}), \quad \text{with } u_I(x) = e^{i(\omega/c)x};$$

it is a time-harmonic solution of

$$\frac{1}{c^2}\frac{\partial^2 U_I}{\partial t^2} - \frac{\partial^2 U_I}{\partial x^2} = 0 \quad \text{on } \mathbb{R}.$$

The following linearized equations hold then for the perturbation displacements of the string and of the beam:

$$(\mathcal{P}_t)\quad \begin{aligned}
&\frac{1}{c^2}\frac{\partial^2 U}{\partial t^2} - \frac{\partial^2 U}{\partial x^2} + K(x)(U - Y) = -K(x)U_I \quad \text{on } \mathbb{R},\\
&\rho(x)\frac{\partial^2 Y}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2 Y}{\partial x^2}\right) - K(x)(U - Y) = K(x)U_I \quad \text{on }]-L, +L[,\\
&\frac{\partial^2 Y}{\partial x^2}(\pm L) = \frac{\partial}{\partial x}\left(E(x)I(x)\frac{\partial^2 Y}{\partial x^2}\right)(\pm L) = 0, \quad \text{where}
\end{aligned}$$

c is the velocity of the waves in the string (assumed constant), $\rho(x)$ the mass of the beam per unit length, $E(x)$ its Young modulus, and $I(x)$ the geometrical inertia of one section. By $K(x)$ we denote the stiffness per unit length of the springs; it vanishes outside the segment $]-L, +L[$. The first two equations in $(\mathcal{P}_t)$ model, respectively, the propagation of waves in the string, and the bending motion of the beam; notice the additional coupling terms, taking into account the action of the springs. The last equations are the free-conditions at the ends of the beam. The study of the behaviour of the longtime solution leads to the following problem:

$$(\mathcal{P}_{\omega^2})\quad \begin{aligned}
&-\frac{\omega^2}{c^2}u - \frac{d^2 u}{dx^2} + K(x)(u - y) = f_u \quad \text{on } \mathbb{R},\\
&-\omega^2\rho(x)y + \frac{d^2}{dx^2}\left(E(x)I(x)\frac{d^2 y}{dx^2}\right) - K(x)(u - y) = f_y \quad \text{on }]-L, +L[,\\
&\frac{d^2 y}{dx^2}(\pm L) = \frac{d}{dx}\left(E(x)I(x)\frac{d^2 y}{dx^2}\right)(\pm L) = 0,\\
&\left|\frac{du}{d|x|} - i\frac{\omega}{c}u\right| \xrightarrow[|x|\to\infty]{} 0 \quad \text{radiation condition,}
\end{aligned}$$

where $f_u = -f_y = -K(x)u_I$.

6.2. Analytical continuation of the resolvent. We consider now the extension of problem $(\mathcal{P}_{\omega^2})$ to complex ν values of ω^2; for the time being ν is assumed to be of strictly positive imaginary part:

Find (u, y) in $\mathcal{H} = H^1(\mathbb{R}) \times H^2(]-L, +L[)$ such that

$$(\mathcal{P}_\nu)\quad \begin{aligned}
&-\frac{\nu}{c^2}u - \frac{d^2 u}{dx^2} + K(x)(u - y) = f_u \quad \text{on } \mathbb{R},\\
&-\nu\rho(x)y + \frac{d^2}{dx^2}\left(E(x)I(x)\frac{d^2 y}{dx^2}\right) - K(x)(u - y) = f_y \quad \text{on }]-L, +L[,\\
&\frac{d^2 y}{dx^2}(\pm L) = \frac{d}{dx}\left(E(x)I(x)\frac{d^2 y}{dx^2}\right)(\pm L) = 0.
\end{aligned}$$

Under variational form, problem $(\mathcal{P}_\nu)$ writes as

$$\text{find } (u,y) \text{ in } \mathcal{H}, \quad \text{such that } \forall (v,z) \in \mathcal{H},$$
$$a_\nu((u,y),(v,z)) = \int_{\mathbb{R}} f_u \bar{v}\,dx + \int_{-L}^{+L} f_y \bar{z}\,dx,$$

where

$$a_\nu((u,y),(v,z)) = \int_{\mathbb{R}} \frac{du}{dx}\frac{d\bar{v}}{dx}\,dx + \int_{-L}^{+L} E(x)I(x)\frac{d^2y}{dx^2}\frac{d^2\bar{z}}{dx^2}\,dx$$
$$-\nu\left[\frac{1}{c^2}\int_{\mathbb{R}} u\bar{v}\,dx + \int_{-L}^{+L}\rho(x)y\bar{z}\,dx\right] + \int_{-L}^{+L} K(x)(u-y)(\bar{v}-\bar{z})\,dx.$$

PROPOSITION 8. *The bilinear form a_ν is continuous and coercive on $\mathcal{H}$.*

Proof. From Young's inequality we infer that

$$|a_\nu((u,y),(u,y))|^2$$
$$\geqq \left(1-\frac{1}{\eta}\right)\left[\int_{\mathbb{R}}\left|\frac{du}{dx}\right|^2 dx + \int_{-L}^{+L} E(x)I(x)\left|\frac{d^2y}{dx^2}\right|^2 dx + \int_{-L}^{+L} K(x)|u-y|^2\,dx\right]^2$$
$$+[(1-\eta)(\mathrm{Re}\,(\nu))^2 + (\mathrm{Im}\,(\nu))^2]\left[\frac{1}{c^2}\int_{\mathbb{R}}|u|^2\,dx + \int_{-L}^{+L}\rho(x)|y|^2\,dx\right]^2,$$

for each $\eta \geqq 0$. Now let $\eta = 1+\frac{1}{2}(\mathrm{Im}\,(\nu)/\mathrm{Re}\,(\nu))^2$; from the previous inequality we deduce that

$$|a_\nu((u,y),(u,y))| \geqq M(\nu)\left[\frac{1}{c^2}\int_{\mathbb{R}}|u|^2\,dx + \int_{\mathbb{R}}\left|\frac{du}{dx}\right|^2 dx\right.$$
$$\left. + \int_{-L}^{+L}\rho(x)|y|^2\,dx + \int_{-L}^{+L}E(x)I(x)\left|\frac{d^2y}{dx^2}\right|^2 dx\right],$$

and by Lion's lemma the coerciveness of a_ν. □

By A_ν we denote the operator on $\mathcal{H}$ associated with the bilinear form a_ν:

$$(A_\nu(u,y),(v,z))_{\mathcal{H}} = a_\nu((u,y),(v,z)),$$

the resolvent thus being $R(\nu) = (A_\nu)^{-1}$.

The explicit continuation of $R(\nu)$ to $\mathbb{C}\backslash\mathbb{R}^-$ is performed along the same lines as in § 4, which is quite easy for one-dimensional problems. Let $d \geqq L$ and $\mathrm{Im}\,(\nu) > 0$; problem $(\tilde{\mathcal{P}}_\nu)$ is then well posed. Its solution is the restriction to $]-d, +d[$ of the solution (u,y) of problem $(\mathcal{P}_\nu)$:

$$\text{Find } (\tilde{u},\tilde{y}) \text{ in } \tilde{\mathcal{H}} = H^1(]-d,+d[) \times H^2(]-L,+L[) \quad \text{such that,}$$

$$(\tilde{\mathcal{P}}_\nu)\quad \begin{aligned} &-\frac{\nu}{c^2}\tilde{u} - \frac{d^2\tilde{u}}{dx^2} + K(x)(\tilde{u}-\tilde{y}) = f_u \quad \text{on } \mathbb{R}, \\ &-\nu\rho(x)\tilde{y} + \frac{d^2}{dx^2}\left(E(x)I(x)\frac{d^2\tilde{y}}{dx^2}\right) - K(x)(\tilde{u}-\tilde{y}) = f_y \quad \text{on }]-L,+L[, \\ &\frac{d^2\tilde{y}}{dx^2}(\pm L) = \frac{d}{dx}\left(E(x)I(x)\frac{d^2\tilde{y}}{dx^2}\right)(\pm L) = 0. \end{aligned}$$

A variational form of this problem in $\tilde{\mathcal{H}}$ is as follows:

$$(I+\tilde{S}(\nu))(\tilde{u},\tilde{y})=\tilde{F},\quad \text{where}$$

$$\begin{aligned}(\tilde{S}(\nu))(\tilde{u},\tilde{y})\,|\,(\tilde{v},\tilde{z}))_{\tilde{\mathcal{H}}}=&-(1+\nu)\left[\frac{1}{c^2}\int_{-d}^{+d}\tilde{u}\bar{\tilde{v}}\,dx+\int_{-L}^{+L}\tilde{y}\bar{\tilde{z}}\,dx\right]\\&+\int_{-L}^{+L}K(x)(\tilde{u}-\tilde{y})(\bar{\tilde{v}}-\bar{\tilde{z}})\,dx\\&+\frac{i\sqrt{\nu}}{c}(\tilde{u}(+d)\bar{\tilde{v}}(+d)+\tilde{u}(-d)\bar{\tilde{v}}(-d)),\end{aligned}$$

and

$$(\tilde{F},(\tilde{v},\tilde{z}))_{\tilde{\mathcal{H}}}=\int_{-d}^{+d}f_u\bar{\tilde{v}}\,dx+\int_{-L}^{+L}f_y\bar{\tilde{z}}\,dx.$$

Notice that $\tilde{\mathcal{H}}$ has been equipped with the following scalar product:

$$\begin{aligned}((\tilde{u},\tilde{y}),(\tilde{v},\tilde{z}))_{\tilde{\mathcal{H}}}=&\frac{1}{c^2}\int_{-d}^{+d}\tilde{u}\bar{\tilde{v}}\,dx+\int_{-L}^{+L}\rho(x)\tilde{y}\bar{\tilde{z}}\,dx\\&+\int_{-d}^{+d}\frac{d\tilde{u}}{dx}\frac{d\bar{\tilde{v}}}{dx}\,dx+\int_{-L}^{+L}E(x)I(x)\frac{d^2\tilde{y}}{dx^2}\frac{d^2\bar{\tilde{z}}}{dx^2}\,dx.\end{aligned}$$

We easily check that $\tilde{S}(\nu)$ defines a holomorphic family of compact operators; as $(I+\tilde{S}(\nu))$ is invertible for $\mathrm{Im}\,(\nu)>0$,

$$\tilde{R}(\nu)=(I+\tilde{S}(\nu))^{-1}$$

extends meromorphically to $\mathbb{C}\backslash\mathbb{R}^-$ (Steinberg [11]). The poles of $\tilde{R}(\nu)$ located in the half-plane $\mathrm{Im}\,(\nu)\leqq 0$ are exactly those of the analytical extension of $R(\nu)$; i.e., the scattering frequencies of the coupled problem or, equivalently, the solutions of the following nonlinear eigenvalue problem:

$$\begin{aligned}&\int_{-d}^{+d}\frac{d\tilde{u}}{dx}\frac{d\bar{\tilde{v}}}{dx}\,dx+\int_{-L}^{+L}E(x)I(x)\frac{d^2\tilde{y}}{dx^2}\frac{d^2\bar{\tilde{z}}}{dx^2}\,dx+\int_{-L}^{+L}K(x)(\tilde{u}-\tilde{y})(\bar{\tilde{v}}-\bar{\tilde{z}})\,dx\\&+\frac{i\sqrt{\nu}}{c}(\tilde{u}(+d)\bar{\tilde{v}}(+d)+\tilde{u}(-d)\bar{\tilde{v}}(-d))=\nu\left[\frac{1}{c^2}\int_{-d}^{+d}\tilde{u}\bar{\tilde{v}}\,dx+\int_{-L}^{+L}\rho(x)\tilde{y}\bar{\tilde{z}}\,dx\right].\end{aligned}\tag{42}$$

6.3. Numerical implementation. The discretization of the problem by the finite element method is carried out as follows. Let us consider a subdivision $S_N=\{x_k, k=-N, N\}$ of $[-L,+L]$ (where $x_k<x_{k+1}$ and $x_{\pm N}=\pm L$) and $S_{N+P}=\{x_k, k=-N-P, N+P\}$ a subdivision of $[-d,+d]$, which contains S_N. We denote by

$$\mathcal{H}_h\subset C^0([-d,+d])\times C^1([-L,+L])$$

the finite-dimensional subspace of $\tilde{\mathcal{H}}$ defined by the pairs (u_h, y_h) such that u_h (respectively, y_h) is a polynomial of order 1 (respectively, 3) on each interval $[x_k, x_{k+1}]$ of S_{N+P} (respectively, S_N). For the sake of simplicity, the mechanical data $E(x)$, $I(x)$, $\rho(x)$, and $K(x)$ are assumed constant. Let $(\cdot,\cdot)_h$ be the scalar product in $\mathcal{H}_h$; we define the matrices $\mathbb{M}_h$, $\mathbb{K}_h$, and $\mathbb{C}_h$ (respectively, the "mass," "stiffness," and "coupling" matrices of the system) by:

$$(\mathbb{M}_h(u_h, y_h), (v_h, z_h))_h = \frac{1}{c^2}\int_{-d}^{+d} u_h\overline{v_h}\, dx + \int_{-L}^{+L} \rho y_h\overline{z_h}\, dx,$$

$$(\mathbb{K}_h(u_h, y_h), (v_h, z_h))_h = \int_{-d}^{+d} \frac{du_h}{dx}\frac{d\overline{v_h}}{dx}\, dx + \int_{-L}^{+L} EI\frac{d^2y_h}{dx^2}\frac{d^2\overline{z_h}}{dx^2}\, dx$$
$$+ \int_{-L}^{+L} K(u_h - y_h)\overline{(v_h - z_h)}\, dx,$$

$$(\mathbb{C}_h(u_h, y_h), (v_h, z_h))_h = \frac{i}{c}\,(u_h(+d)\overline{v_h(+d)} + u_h(-d)\overline{v_h(-d)}).$$

These matrices are computed classically by assembling the associated elementary matrices of each element. One can easily check that $\mathbb{M}_h$ and $\mathbb{K}_h$ (but not $\mathbb{C}_h$) are hermitian matrices, $\mathbb{M}_h$ is positive, and $\mathbb{K}_h$ is nonnegative. Notice that the same matrices could be obtained by a standard difference scheme.

Let us first consider the initial problem $(\mathcal{P}_{\omega^2})$, i.e., when the system is subject to a monochromatic incident wave of frequency ω:

$$u_I(x) = e^{i(\omega/c)x}.$$

The approximate response of the system is thus the solution of the following linear problem:

$$(43)\qquad \mathbb{A}_h(\omega^2)(u_h, y_h) = F_h, \quad \text{where } \mathbb{A}_h(\omega^2) = [-\omega^2\mathbb{M}_h + \mathbb{K}_h + \omega\mathbb{C}_h],$$

and the second member $F_h \in \mathcal{H}_h$ is defined by:

$$(F_h, (v_h, z_h))_h = \int_{-L}^{+L} Ku_I\overline{(z_h - v_h)}\, dx.$$

This system is solved by the elementary method of Gaussian elimination. Once the solution has been computed, we can then calculate the total energy of the beam, which is given by:

$$(44)\qquad E_b(\omega) = \tfrac{1}{2}([\omega^2\mathbb{M}_h + \mathbb{K}_h](0, y_h), (0, y_h))_h.$$

The determination of this quantity for each ω over a given frequency range leads us to the "response curve" of the beam whose maxima show the resonant states associated with the beam.

On the other hand, let us consider the matrix nonlinear eigenvalue problem deduced from (42):

$$(45)\qquad [\mathbb{K}_h + \sqrt{\nu}\,\mathbb{C}_h](u_h, y_h) = \nu\mathbb{M}_h(u_h, y_h),$$

whose solutions are the approximate scattering elements. By $\lambda_k(\nu)$ we denote any of the eigenvalues of the problem:

$$(46)\qquad [\mathbb{K}_h + \sqrt{\nu}\,\mathbb{C}_h](u_h, y_h) = \lambda_k(\nu)\mathbb{M}_h(u_h, y_h),$$

which actually are branches of analytic functions of ν with only algebraic singularities (Kato [3]). An approximate scattering frequency is thus a solution of the fixed-point equation $\lambda_k(\nu) = \nu$. The application of the iterative Newton method consists in computing, for a given initial value ν_0 and a given k, the sequence:

$$\nu_{j+1} = \nu_j - (\lambda_k(\nu_j) - \nu_j)\bigg/\left(\frac{d\lambda_k}{d\nu}(\nu_j) - 1\right),$$

where the iterations will be terminated if $|\nu_{j+1}-\nu_j|$ becomes less than the expected error ε. However, from a numerical point of view, the algorithm cannot be implemented in this form, for we cannot a priori know at each iteration which eigenvalue of problem (46) (with $\nu=\nu_j$) corresponds to the chosen branch $\lambda_k(\nu)$, even if all the eigenvalues of (46) are computed. This leads us to modify the previous relation as follows:

$$\nu_{j+1}=\nu_j-(\lambda(\nu_j)-\nu_j)\Big/\left(\frac{d\lambda}{d\nu}(\nu_j)-1\right),$$

where $\lambda(\nu_j)$ is now the closest eigenvalue to the given initial value ν_0. In this case, the sequence (ν_j) will converge (if it actually does) to the closest scattering frequency to ν_0. Let us notice again that nothing can ensure that, for two successive iterations, $\lambda(\nu_j)$ and $\lambda(\nu_{j+1})$ are points of the same branch: in particular, the algorithm will probably fail if ν_0 is chosen in the vicinity of a branch point.

The eigenvalue $\lambda(\nu_j)$ and the corresponding eigenvector are computed by the inverse iteration method [15] (assuming that $\lambda(\nu_j)$ is simple); it consists in the determination of a sequence (u_h^r, y_h^r) defined by:

$$[\mathbb{K}_h+\sqrt{\nu_j}\,\mathbb{C}_h-\nu_0\mathbb{M}_h](u_h^{r+1}, y_h^{r+1})=m^r\mathbb{M}_h(u_h^r, y_h^r),$$

where m^r is such that $\|(u_h^{r+1}, y_h^{r+1})\|_h=1$ and (u_h^0, y_h^0) is an arbitrary unit vector. Of course, the termination criterion (see [15]) must be chosen so that the precision on $\lambda(\nu_j)$ is better than the expected precision ε of the Newton algorithm.

Since $\lambda(\nu_j)$ has been assumed simple (and thus ν_j cannot be a branch point), its derivative $d\lambda/d\nu$ is obtained quite easily. Let (u_h, y_h) be an eigenvector associated with the eigenvalue $\lambda(\nu_j)$, and (v_h, z_h) a left eigenvector (i.e., an eigenvector of the adjoint problem) chosen so that $(\mathbb{M}_h(u_h, y_h), (v_h, z_h))_h=1$; we have:

$$\frac{d\lambda}{d\nu}(\nu_j)=\left(\frac{1}{2\sqrt{\nu_j}}\mathbb{C}_h(u_h, y_h), (v_h, z_h)\right)_h.$$

In our case, the left eigenvectors are simply given by $\alpha\,\overline{(u_h, y_h)}$ (where α is a nonzero complex number), for $[\mathbb{K}_h+\sqrt{\nu_j}\,\mathbb{C}_h]$ is a symmetric (but not hermitian) matrix.

For a scattering frequency ν_* which is close enough to the positive real axis, the expansion of the solution of problem (43) in the vicinity of ν_* can be performed as described in § 5 with:

$$\mathbb{S}_h(\nu)=\mathbb{K}_h+\sqrt{\nu}\,\mathbb{C}_h+(1-\nu)\mathbb{M}_h \quad\text{and}\quad \mathbb{J}_h=\mathbb{M}_h.$$

This expansion is:

$$(47)\qquad (u_h, y_h)=(\omega^2-\nu_*)^{-1}X^{(-1)}+X^{(0)}+0(\omega^2-\nu_*),$$

where $X^{(-1)}$ and $X^{(0)}$ are given by formulas (39) and (40) (with $M=1$). The computation of the former is obvious; the latter requires the calculation of all the eigenvalues and corresponding right and left eigenvectors of problem (46) for $\nu=\nu_*$: they are obtained by the classical "QR" method [15]. The expansion of the total energy of the beam (44) follows directly from (47): it thus provides an approximation of the response curve in the vicinity of the real part of a scattering frequency.

6.4. Numerical results. We report in this last paragraph the outcome of one typical numerical application. It is related to the following data:

$$c=1,\quad K=1.4,\quad \rho=0.5,\quad EI=0.1,\quad L=1.$$

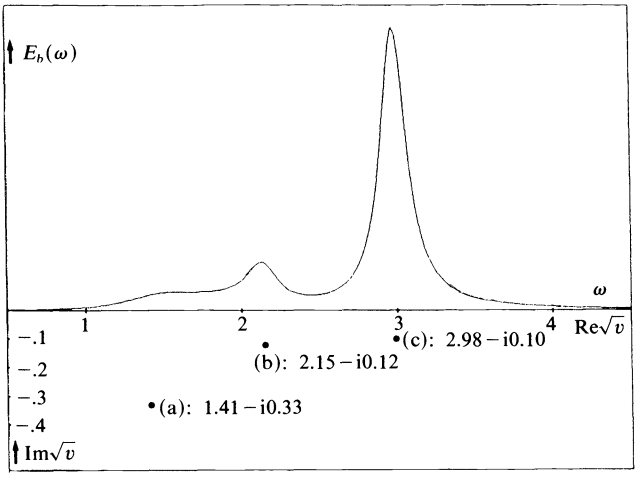


FIG. 6. *Response curve and location of scattering frequencies.*

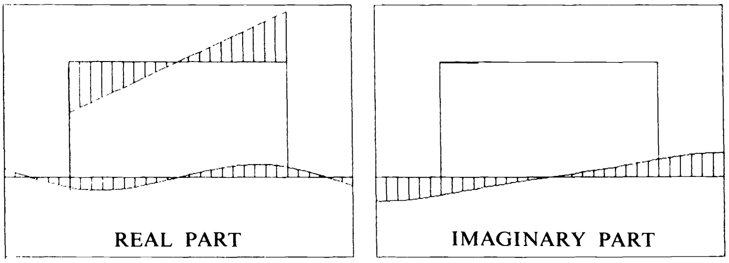


FIG. 7a. *Scattering mode—Case* (a).

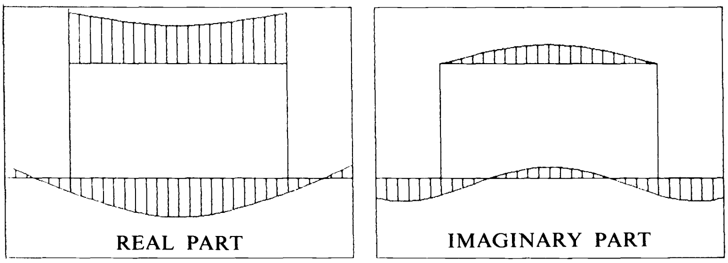


FIG. 7b. *Scattering mode—Case* (b).

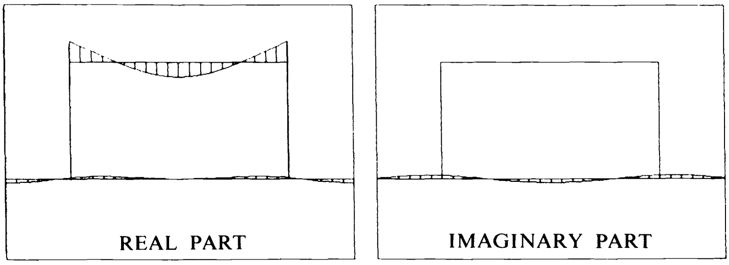


FIG. 7c. *Scattering mode—Case* (c).

The discretisation is shown in Fig. 7: 20 elements for the beam and $20+2\times 6$ elements for the string (75 degrees of freedom). For the computation of the scattering frequencies, the expected precision is 10^{-2} for the Newton algorithm and 10^{-4} for the inverse iteration method. The considered frequency range is $[0.5, 4.5]$.

The numerical experiments have shown that the choice of the initial value ν_0 is of the highest importance for the convergence of the Newton iterations. First, if ν_0 is of small modulus (say $|\nu_0|<1.$), the sequence (ν_j) generally does not converge: this is probably due to the fact that 0 is a branch point of the matrix $[\mathbb{K}_h+\sqrt{\nu}\,\mathbb{C}_h]$ in (46). When ν_0 is chosen in the vicinity of the positive real axis (say $\text{Im}\,(\nu_0)>-0.5$) and is not of small modulus, two situations occur. On one hand, if ν_0 is close enough to a scattering frequency (which is of course a priori unknown), the sequence (ν_j) converges to it (generally in less than 10 iterations); on the other hand, if ν_0 is unfortunately chosen about the middle of the gap between two scattering frequencies, the sequence may oscillate between them: actually, the computed eigenvalues $\lambda(\nu_j)$ result from two different but neighbouring branches $\lambda_k(\nu)$ of (46). At last, if the imaginary part of ν_0 increases, no convergence will be obtained.

The upper part of Fig. 6 is the response curve of the beam in the frequency range $[0.5, 4.5]$, i.e., the total energy (44) of the beam obtained by solving problem (43). The lower part of the same figure shows the location in the complex plane of the computed scattering frequencies (solutions of problem (46)). Notice that their real parts coincide approximately with the resonant frequencies, i.e., the locations of the maxima of the response.

Figure 7 shows, for each scattering frequency, the real and imaginary parts of the associated eigenvector: the upper and lower parts of each figure represent, respectively, the vertical displacement of the points on the beam and on the vibrating string. These "scattering modes" show what kind of motion is "excited":

(a) is a "roll" motion of the beam with very little bending strain;

(b) is a "heave" motion coupled with a bending mode of the beam;

(c) is a pure bending motion.

Figures 8 and 9 show the estimation of the response in the vicinity of the scattering frequencies (a), (b), and (c) using, respectively, the first term (order -1) and the two first terms (order -1 and 0) of the expansion (47) (the dotted curve reproduces the

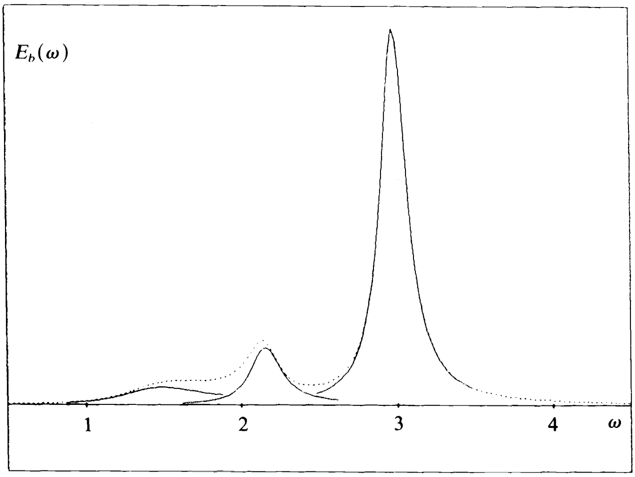


FIG. 8. *Approximation of the response at order* -1.

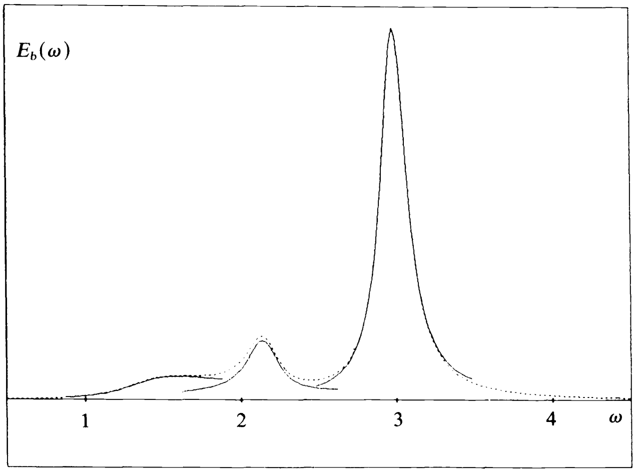


FIG. 9. *Approximation of the response at order* 0.

response curve of Fig. 6). In case (c), the first term provides a good approximation of the response and no significant improvement is given by including the second term (order zero) of the expansion. But in cases (a) and (b), this second term becomes necessary; in particular, it provides an approximation of the difference between the real part of the scattering frequency and the resonant frequency (see formula (41)): notice that although the imaginary parts of the scattering frequencies (b) and (c) have the same order of magnitude, this difference is negligible only in case (c).

7. Conclusion. The numerical results we obtained show the efficiency of our method: the knowledge of the scattering frequencies and of the associated scattering modes not only allows us to locate the resonant frequencies without computing the whole response curve, but also provides an approximation of the solution of the steady-state problem in the vicinity of these frequencies. However, as regards the numerical computation of the scattering frequencies, there still remains the question of which is worth studying: we actually need a priori bounds to provide good initial guesses for the iterative algorithm which solves the nonlinear eigenvalue problem.

In a forthcoming paper, we will show how the method can be extended to a more complicated fluid-structure interaction problem: the "sea-keeping" problem, i.e., the study of the motions of a floating body on the sea.

Appendix. A direct way of expanding the solution. In this section, we shall assume that the solution $\varphi(\nu)$ of (23) is subject to an expansion similar to (37): in the vicinity of ν_*, for some integer $\rho \geqq 0$, we have

$$\varphi(\nu) = \sum_{r \geqq -\rho} (\nu - \nu_*)^r \varphi^{(r)}. \tag{48}$$

The expansion of the right-hand side member $F(f, \nu)$ and of operator $\mathbb{S}(\nu)$ are straightforward:

$$F(f, \nu) = \sum_{n=0}^{\infty} (\nu - \nu_*)^n F^{(n)} \quad \text{and} \quad \mathbb{S}(\nu) = \sum_{n=0}^{\infty} (\nu - \nu_*)^n \mathbb{S}^{(n)}.$$

By replacing the different terms by their expansion in (23), we obtain the following relations between the coefficients:

$$\sum_{r=-p}^{q-1} \mathbb{S}^{(q-r)}\varphi^{(r)}+(\mathbb{J}+\mathbb{S}^{(0)})\varphi^{(q)}=F^{(q)}, \quad q \geqq -p. \tag{49}$$

In the sequel we shall give explicit expressions for the solution of system (49); the difficulty originates from the fact that the diagonal element $(\mathbb{J}+\mathbb{S}^{(0)})$ is singular.

Some definitions. We will use the same notations as defined in § 5. Moreover, for $m=1, M$, we set

$$h_m^{(0)}=e_m \quad \text{and, recursively}$$

$$H_m^{(\sigma)}=\sum_{s=0}^{\sigma} \mathbb{S}^{(\sigma+1-s)}h_m^{(s)} \quad \text{with} \tag{50}$$

$$h_m^{(\sigma+1)}=-\mathbb{Q}\mathbb{J}^{-1}H_m^{(\sigma)} \quad \sigma \geqq 0,$$

and

$$Z^{(\sigma)}=\sum_{s=-p}^{\sigma-1} \mathbb{S}^{(\sigma-s)}z^{(s)} \quad \sigma \geqq -p \quad \text{with}$$

$$z^{(s)}=\mathbb{Q}\mathbb{J}^{-1}(F^{(s)}-Z^{(s)}). \tag{51}$$

Notice that $Z^{(\sigma)}=0$ for $\sigma \leqq 0$ and $z^{(\sigma)}=0$ for $\sigma<0$, since $F^{(n)}=0$ for $n<0$; it follows that

$$Z^{(\sigma)}=\sum_{s=0}^{\sigma-1} \mathbb{S}^{(\sigma-s)}z^{(s)}.$$

Finally, by $\mathcal{H}^{(\sigma)}$ we denote the following matrix:

$$\mathcal{H}_{m'm}^{(\sigma)}=(H_m^{(\sigma)}|g_{m'}), \quad m, m'=1, M, \quad \sigma \geqq 0,$$

and by $\mathcal{S}^{(q)}$ the following vector:

$$\mathcal{S}_{m'}^{(q)}=(F^{(q+1)}-Z^{(q+1)}|g_{m'}).$$

Notice that $\mathcal{S}^{(q)}=0$ when $q<-1$.

PROPOSITION 9. *Provided that expansion* (48) *holds in the vicinity of* ν_*, *we have*

$$\text{(i)} \qquad \varphi^{(q)}=\sum_{m=1}^{M}\sum_{\sigma=-p}^{q} \varphi_m^{(\sigma)}h_m^{(q-\sigma)}+z^{(q)};$$

moreover, if $\psi^{(\sigma)}$ *denotes the vector of the* $\varphi_m^{(\sigma)}$, $m=1, M$,

$$\text{(ii)} \qquad \sum_{\sigma=-p}^{q} \mathcal{H}^{(q-\sigma)}\psi^{(\sigma)}=\mathcal{S}^{(q)}, \qquad q \geqq -p.$$

Proof. The proof is recursive; the first step consists in proving formula (i) for $q=-p$. By formula (49) we have

$$(\mathbb{J}+\mathbb{S}^{(0)})\varphi^{(-p)}=F^{(-p)}; \quad \text{consequently}$$

$$\mathbb{P}\mathbb{J}^{-1}F^{(-p)}=0 \quad \text{and thus}$$

$$\varphi^{(-p)}=\sum_{m=1}^{M} \varphi_m^{(-p)}e_m+\mathbb{Q}\mathbb{J}^{-1}F^{(-p)}.$$

(ii) We assume formula (i) is satisfied at ranks $q=-\rho$, Q, and we prove it at rank $Q+1$. By (49)

$$\sum_{r=-\rho}^{Q} \mathbb{S}^{(Q+1-r)}\left[\sum_{m=1}^{M} \sum_{\sigma=-\rho}^{r} \varphi_m^{(\sigma)} h_m^{(r-\sigma)} + z^{(r)}\right] + (\mathbb{J}+\mathbb{S}^{(0)})\varphi^{(Q+1)} = F^{(Q+1)},$$

which is nothing but

$$(\mathbb{J}+\mathbb{S}^{(0)})\varphi^{(Q+1)} + \sum_{m=1}^{M} \sum_{\sigma=-\rho}^{Q} \varphi_m^{(\sigma)} \sum_{r=\sigma}^{Q} \mathbb{S}^{(Q+1-r)} h_m^{(r-\sigma)} + \sum_{r=-\rho}^{Q} \mathbb{S}^{(Q+1-r)} z^{(r)} = F^{(Q+1)};$$

or, equivalently,

$$(\mathbb{J}+\mathbb{S}^{(0)})\varphi^{(Q+1)} + \sum_{m=1}^{M} \sum_{\sigma=-\rho}^{Q} \varphi_m^{(\sigma)} H_m^{(Q-\sigma)} = F^{(Q+1)} - Z^{(Q+1)}.$$

It follows that

$$\text{(52)} \qquad \begin{aligned} &\mathbb{P}\mathbb{J}^{-1}(F^{(Q+1)} - Z^{(Q+1)}) - \sum_{m=1}^{M} \sum_{\sigma=-\rho}^{Q} \varphi_m^{(\sigma)} \mathbb{P}\mathbb{J}^{-1} H_m^{(Q-\sigma)} = 0, \\ &\varphi^{(Q+1)} = \sum_{m=1}^{M} \varphi_m^{(Q+1)} e_m + \sum_{m=1}^{M} \varphi_m^{(\sigma)} h_m^{(Q+1-\sigma)} + z^{(Q+1)}. \end{aligned}$$

(iii) From formula (52), we infer that

$$\sum_{m=1}^{M} \sum_{\sigma=-\rho}^{Q} \varphi_m^{(\sigma)} (H_m^{(Q-\sigma)} \,|\, g_{m'}) = (F^{(Q+1)} - Z^{(Q+1)} \,|\, g_{m'}) \quad \forall m'=1, M, \forall Q \geqq -\rho, \quad \text{i.e.,}$$

$$\sum_{\sigma=-\rho}^{Q} \mathcal{H}^{(Q-\sigma)} \psi^{(\sigma)} = \mathcal{S}^{(Q)} \quad \forall Q \geqq -\rho. \qquad \square$$

This proposition allows the determination of the $\varphi_m^{(\sigma)}$. Three situations occur:

(i) If $\mathcal{H}^{(0)}=0$, then

$$\sum_{\sigma=-\rho}^{q-1} \mathcal{H}^{(q-\sigma)} \psi^{(\sigma)} = \mathcal{S}^{(q)}, \qquad q \geqq -\rho,$$

which is similar to formula (ii) of Proposition 9, with $\mathcal{H}^{(0)}$ replaced by $\mathcal{H}^{(1)}$. We are thus brought back to the initial situation.

(ii) If $\mathcal{H}^{(0)}$ is invertible, then

$$\text{(53)} \qquad \psi^{(\sigma)} = (\mathcal{H}^{(0)})^{-1}\left[\mathcal{S}^{(\sigma)} - \sum_{r=-\rho}^{\sigma-1} \mathcal{H}^{(\sigma-r)} \psi^{(r)}\right], \quad \text{for } \sigma \geqq -\rho;$$

notice that $\psi^{(\sigma)}$ vanishes for $\sigma < -1$.

(iii) If $\mathcal{H}^{(0)}$ is not invertible, and if we assume that its eigenvalue zero is semisimple, then the solution of (ii) of Proposition 9 is a similar problem to solving (49), where $\mathcal{H}^{(0)}$ replaces $(\mathbb{J}+\mathbb{S}^{(0)})$ and $\mathcal{H}^{(q)}$ replaces $\mathbb{S}^{(q)}$, $q>0$.

In the simple case where $M=1$, this discussion becomes straightforward: by l we denote the lowest integer such that $\mathcal{H}^{(l)} \neq 0$, and we obtain

$$\psi^{(q-l)} = \frac{1}{\mathcal{H}^{(l)}}\left[\mathcal{S}^{(q)} - \sum_{\sigma=-\rho}^{q+1-l} \mathcal{H}^{(q-\sigma)} \psi^{(\sigma)}\right];$$

it follows that $\psi^{(\sigma)}$ vanishes for $\sigma < -1 - l$.

Remark 8. The following quantities are needed in the preceding formulas:

$$h_m^{(\sigma+1)} = -\mathbb{Q}\mathbb{J}^{-1}H_m^{(\sigma)} \quad \text{and} \quad z^{(\sigma)} = \mathbb{Q}\mathbb{J}^{-1}(F^{(\sigma)} - Z^{(\sigma)});$$

we are thus led to the solution of singular systems:

$$(54) \qquad \begin{aligned} &(\mathbb{J} + \mathbb{S}^{(0)})h_m^{(\sigma+1)} = -\left[H_m^{(\sigma)} - \sum_{m'=1}^{M} (H_m^{(\sigma)} | g_{m'})\mathbb{J}e_{m'}\right] \\ &(\mathbb{J}h_m^{(\sigma+1)} | g_{m'}) = 0, \qquad m' = 1, M, \end{aligned}$$

and

$$(55) \qquad \begin{aligned} &(\mathbb{J} + \mathbb{S}^{(0)})z^{(\sigma)} = F^{(\sigma)} - Z^{(\sigma)} - \sum_{m'=1}^{M} (F^{(\sigma)} - Z^{(\sigma)} | g_{m'})\mathbb{J}e_{m'} \\ &(\mathbb{J}z^{(\sigma)} | g_{m'}) = 0, \qquad m' = 1, M. \end{aligned}$$

A special case. We now give some details about the case where $\mathscr{H}^{(0)}$ has only one eigenvalue, different from zero; moreover, assume that $\mathscr{H}^{(0)}$ is diagonalizable, then it is diagonal and eventually a multiple of the identity

$$\mathscr{H}^{(0)} = \frac{1}{M}\sum_{m=1}^{M} (H_m^{(0)} | g_m)\mathbb{I} = (H_{m'}^{(0)} | g_{m'})\mathbb{I} \quad \forall m'.$$

Furthermore, if $\mathbb{U}^{(0)}$ is assumed diagonalizable, an explicit expression of the operator $\mathbb{Q}$ is available:

$$\mathbb{Q}G = \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (\mathbb{J}G | g_{ij})e_{ij}.$$

As $\psi^{(\sigma)} = 0$ for $\sigma < -1$, it is enough to choose $\rho = 1$; as a consequence:

$$H_m^{(0)} = \mathbb{S}^{(1)}e_m,$$

$$h_m^{(1)} = -\sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (\mathbb{S}^{(1)}e_m | g_{ij})e_{ij},$$

$$H_m^{(1)} = \mathbb{S}^{(2)}e_m - \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (\mathbb{S}^{(1)}e_m | g_{ij})\mathbb{S}^{(1)}e_{ij},$$

and

$$z^{(0)} = \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (F^{(0)} | g_{ij})e_{ij},$$

$$Z^{(1)} = \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (F^{(0)} | g_{ij})\mathbb{S}^{(1)}e_{ij}.$$

It follows that

$$\mathscr{S}_m^{(-1)} = (F^{(0)} | g_m),$$

$$\mathscr{S}_m^{(0)} = (F^{(1)} | g_m) - \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (F^{(0)} | g_{ij})(\mathbb{S}^{(1)}e_{ij} | g_m),$$

and thus,

$$\varphi_m^{(-1)} = \frac{M(F^{(0)}|g_m)}{\sum_{m'=1}^{M} (\mathbb{S}^{(1)}e_{m'}|g_{m'})},$$

$$\begin{aligned}\varphi_m^{(0)} = \frac{M}{\sum_{m'=1}^{M} (\mathbb{S}^{(1)}e_{m'}|g_{m'})} \Bigg[& (F^{(1)}|g_m) - \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (F^{(0)}|g_{ij})(\mathbb{S}^{(1)}e_{ij}|g_m) \\ & - \frac{M}{\sum_{m'=1}^{M} (\mathbb{S}^{(1)}e_{m'}|g_{m'})} \sum_{m'=1}^{M} (F^{(0)}|g_{m'}) \\ & \cdot \Bigg[\!\!\Bigg[(\mathbb{S}^{(2)}e_{m'}|g_m) - \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (\mathbb{S}^{(1)}e_{m'}|g_{ij})(\mathbb{S}^{(1)}e_{ij}|g_m) \Bigg]\!\!\Bigg]\Bigg].\end{aligned}$$

Finally, we obtain

$$\varphi^{(-1)} = M \frac{\sum_{m=1}^{M} (F^{(0)}|g_m)e_m}{\sum_{m=1}^{M} (\mathbb{S}^{(1)}e_m|g_m)}, \quad \text{and}$$

$$\varphi^{(0)} = -\sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} \sum_{m=1}^{M} \varphi_m^{(-1)}(\mathbb{S}^{(1)}e_m|g_{ij})e_{ij} + \sum_{m=1}^{M} \varphi_m^{(0)}e_m + \sum_{i=2}^{I} \frac{1}{1+\mu_i} \sum_{j=1}^{M(i)} (F^{(0)}|g_{ij})e_{ij}.$$

Remark 9. It is worth noticing that these formulas agree with (39) and (40) if $\mathcal{H}^{(1)}$ is assumed diagonal.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] D. M. Eidus, *On the principle of limiting absorption*, Math. USSR-Sb., 57 (1962), pp. 13-44. (In Russian.) Amer. Math. Soc. (2), 47 (1965), pp. 157-191. (In English.)

[3] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1984.

[4] P. D. Lax and R. S. Phillips, *Scattering Theory*, Academic Press, New York, 1967.

[5] M. Lenoir, *Méthodes de couplage en Hydrodynamique Navale et application à la résistance de vagues bidimensionnelle, thèse*, Université Pierre et Marie Curie, Paris, 1982.

[6] M. Lenoir and A. Tounsi, *The localized finite element method and its application to the 2D sea-keeping problem*, SIAM J. Numer. Anal., 25 (1988), pp. 729-752.

[7] B. Nicolas-Vullierme, R. Ohayon, and G. Galbe, *Computation of the far-field radiation of submerged structures using asymptotic expansions starting from an incompressible fluid approximation*, Internat. Conf. Numer. Meth. for Transient and Coupled Problems, Venice, July 9th-13, 1984, pp. 238-248.

[8] R. Ohayon and E. Sanchez-Palencia, *On the vibration problem for an elastic body surrounded by a slightly compressible fluid*, RAIRO Numer. Anal., 17 (1983), pp. 311-326.

[9] N. Shenk and D. Thoe, *Outgoing solutions of $(-\Delta + q - k^2)u = f$ in exterior domains*, J. Math. Anal. Appl., 31 (1970), pp. 81-116.

[10] ———, *Resonant states and poles of the scattering matrix for perturbations of $-\Delta$*, J. Math. Anal. Appl., 37 (1972), pp. 467-491.

[11] S. Steinberg, *Meromorphic families of compact operators*, Arch. Rational Mech. Anal., 31 (1968), pp. 372-380.

[12] M. Vullierme-Ledard, *The limiting amplitude applied to the motion of floating bodies*, RAIRO Math. Model. Anal. Numer., 21 (1987), pp. 125-170.

[13] M. Wei, *Numerical computation of scattering frequencies*, Ph.D. thesis, Brown University, Providence, RI, 1986.

[14] C. H. Wilcox, *Scattering theory for the d'Alembert equation in exterior domains*, Lecture Notes in Math., 422, Springer-Verlag, New York, 1976.

[15] J. H. Wilkinson and C. Reinsch, *Linear Algebra*, Springer-Verlag, New York, 1971.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 609–634, May 1992

GLOBAL SOLUTIONS TO THE COMPRESSIBLE NAVIER–STOKES EQUATIONS FOR A REACTING MIXTURE*

GUI-QIANG CHEN†

Abstract. Existence theorems are established for global generalized solutions to the compressible Navier–Stokes equations for a reacting mixture with discontinuous Arrhenius functions, which describe dynamic combustion. Equivalence of the Navier–Stokes equations in the Euler coordinates and the Lagrange coordinates for the generalized solutions is verified. The asymptotic behavior of the generalized solutions with different boundary conditions is identified and proved.

Key words. global generalized solutions, asymptotic behavior, equivalence, combustion, discontinuous Arrhenius functions, a priori estimates, Navier–Stokes equations

AMS(MOS) subject classifications. 35B40, 35D05, 76V05, 35B45

1. Introduction. We are concerned with the existence and asymptotic behavior of global solutions to the compressible Navier–Stokes equations for a reacting mixture with discontinuous reacting rate functions, which describe dynamic combustion. These equations in the Euler coordinates are of the following form (cf. [1]–[3]):

$$\tag{1.1} \begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + p)_x = (\mu u_x)_x, \\ (\rho E)_t + (u(\rho E + p))_x = (\mu u u_x)_x + (\nu T_x)_x + (q d \rho Z_x)_x, \\ (\rho Z)_t + (\rho u Z)_x = -K\phi(T)\rho Z + (d\rho Z_x)_x, \end{cases}$$

where ρ, u, E, and Z are the density, the fluid velocity, the total specific energy, and the mass fraction of the reactant, respectively; the constants μ, ν, d, and K are the coefficients of bulk viscosity, heat conduction, species diffusion, and rate of reactant, respectively; and T is the temperature.

The total specific energy has the form

$$\tag{1.2} E = e + \frac{u^2}{2} + qZ,$$

where e is the specific internal energy and q is the difference in the heats of formation of the reactant and the product.

For an ideal gas mixture with the same γ-gas laws, the internal energy e and the temperature T are often given via thermodynamics through the following equations of state:

$$\tag{1.3} \begin{cases} e = pv/(\gamma - 1), \\ T = pv/a. \end{cases}$$

*Received by the editors November 5, 1990; accepted for publication (in revised form) September 3, 1991.

† Department of Mathematics, The University of Chicago, Chicago, Illinois 60637-1404, and Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439-4801. The research was supported in part by Department of Energy grant under contract W-31-109-Eng-38, by Office of Naval Research grant N00014-91-J-1384, and by National Science Foundation grant DMS-850403.

Here, $v = \frac{1}{\rho}$ (the specific volume) and $a = RM > 0$, where R is the Boltzmann's gas constant and M is the molecular weight.

The rate function $\phi(T)$ is typically determined by the Arrhenius law:

$$\phi(T) = \begin{cases} T^\alpha e^{-A/T}, & T > T_i > 0, \\ 0, & T < T_i, \end{cases} \tag{1.4}$$

which is generally a uniformly bounded and discontinuous function, where T_i is the ignition temperature and A is the activation energy.

We begin with a bounded region with impermeably insulated boundaries. We assume that the distribution of ρ, u, T, and Z is known at the initial instant $t = 0$. By scaling, we can form the following initial boundary value conditions:

$$(\rho, u, T, Z)|_{t=0} = (\rho_0(x), u_0(x), T_0(x), Z_0(x)), \quad 0 \le x \le 1, \tag{1.5}$$

$$\begin{cases} u(t,i) = 0, \\ T_x(t,i) = 0, \qquad i = 0,1, \quad t \ge 0, \\ Z_x(t,i) = 0, \end{cases} \tag{1.6}$$

$$0 < m_0 \le \rho_0(x), T_0(x) \le M_0 < \infty, \quad 0 \le Z_0(x) \le 1, \tag{1.7}$$

with the compatibility conditions:

$$\begin{cases} u_0(i) = T_{0x}(i) = Z_{0x}(i) = 0, \\ (a\rho_0 T_0 - (\mu u_{0x})_x)|_{x=i} = 0, \end{cases} \quad i = 0,1. \tag{1.8}$$

In addition to the impermeably insulated boundaries, we consider the source boundaries. In particular, we are concerned with the following inhomogeneous initial boundary value conditions:

$$(\rho, u, T, Z)|_{t=0} = (\rho_0(x), u_0(x), T_0(x), Z_0(x)), \quad 0 \le x \le 1, \tag{1.9}$$

$$\begin{cases} u(t,i) = 0, \\ T(t,i) = q_i(t), \qquad i = 0,1, \quad t \ge 0, \\ Z(t,i) = r_i(t), \end{cases} \tag{1.10}$$

$$0 < m_0 \le \rho_0(x), T_0(x) \le M_0 < \infty, \quad 0 \le Z_0(x) \le 1, \tag{1.11}$$

with

$$\begin{cases} u_0(i) = 0, \\ T_0(i) = q_i(0), \\ Z_0(i) = r_i(0), \qquad i = 0,1, \\ 0 < m_0 \le q_i(t) \le M_0 < \infty, \\ 0 \le r_i(t) \le 1. \end{cases} \tag{1.12}$$

Assume that the region filled with the reacting gas is the whole space, and that the distribution of ρ, u, T, and Z is known at the initial instant. For this case we can form the following initial conditions:

$$(\rho,u,T,Z)|_{t=0}=(\rho_0(x),u_0(x),T_0(x),Z_0(x)),\quad -\infty<x<\infty, \tag{1.13}$$

$$\begin{cases} \lim\limits_{|x|\to\infty}(\rho_0(x),u_0(x),T_0(x),Z_0(x))=(1,0,1,0), \\ 0<m_0\le\rho_0(x),T_0(x)\le M_0<\infty,\quad |u_0(x)|\le M_0,\quad 0\le Z_0(x)\le 1. \end{cases} \tag{1.14}$$

As a result of the ignition phenomena in combustion problems, the rate function $\phi(T)$ is generally a discontinuous function in the solution space, and we must consider nonclassical solutions. We denote $\Omega=(0,1)$ or $(-\infty,\infty)$. A function (ρ,u,T,Z) is called a generalized solution to the system (1.1) in $(0,L)\times\Omega$, with the initial boundary value conditions (1.5)–(1.8), or (1.9)–(1.12), or the initial conditions (1.13)–(1.14), if

$$\begin{cases} \rho-1\in L^\infty(0,L;H^1(\Omega)),\qquad \rho_t\in L^2(0,L;L^2(\Omega)), \\ (u,T-1,Z)\in L^\infty(0,L;H^1(\Omega))\cap L^2(0,L;H^2(\Omega)), \\ (u_t,T_t,Z_t)\in L^2(0,L;L^2(\Omega)), \end{cases}$$

and if the function satisfies the system (1.1) almost everywhere in $(0,L)\times\Omega$ and the prescribed initial boundary value in the trace sense of the function belonging to the above classes (see Adams [4]).

It is well known that there are two coordinate systems to describe the flow of fluid in fluid dynamics: the Euler coordinates and the Lagrange coordinates [5]. The Euler coordinates represent a coordinate system imposed on the physical space, while the Lagrange coordinates represent a coordinate system imposed on the flow field. A different choice of independent space coordinates will lead to a different system of partial differential equations. In §2 we introduce the Lagrange coordinates and translate the preceding initial boundary value problems and the Cauchy problem in the Euler coordinates into corresponding problems in the Lagrange coordinates. A calculation using the product rule and chain rule shows that these corresponding problems are equivalent for classical solutions (when such solutions exist). While the discontinuity of the reacting rate function precludes the use of the product rule and chain rule in the classical sense, we verify the equivalence of these corresponding problems in the two coordinates for the corresponding generalized solutions.

In §3 we first establish the existence of global classical solutions of typical initial boundary value problems for the compressible Navier–Stokes equations with a smooth reacting rate function in the Lagrange coordinates. The local solution is based on the Banach theorem as a result of contractivity of the operator, which is equivalent to the solution constructed by linearization of the problems on a small time interval. The global existence theorem is established by extending the local solution with respect to time, based on a priori global estimates. These estimates are obtained by using the maximal principle and the techniques of energy estimates. To deal with the discontinuous rate function ϕ, we concentrate our attention to the dependence of the a priori estimate constants for ϕ. The main finding of our analysis is the crucial independence of the a priori estimates in H^1 on the bounds of the derivative functions of ϕ. By smoothing ϕ and taking the limit, we finally obtain a global generalized solution for the initial boundary value problems with the discontinuous Arrhenius function (1.4).

Different smoothing approaches generate different forms of the discontinuous Arrhenius function near $T = T_i$. We also transform the general initial boundary value problems (cf. (1.9)–(1.10)) to typical ones and similarly establish a global existence theorem.

Our results show that there are global solutions for the initial boundary value problems in spite of the fact that the Arrhenius function is discontinuous. In fact, our analysis indicates that the H^1 norm of the solutions is independent of the bounds of the derivative functions of the Arrhenius function. This finding should be useful in the related problems in the combustion theory.

Our further objective is to study properties of the solutions with the discontinuous Arrhenius functions (1.4). In §4 we focus on the asymptotic behavior of the solutions as $t \to \infty$. We compare the asymptotic behavior of the solutions with two different boundary conditions: the impermeably insulated boundaries and the thermal source boundaries. We conclude that the asymptotic behavior of the solutions with the discontinuous Arrhenius functions (1.4) for the first case is determined not only by the initial data, but also by the Arrhenius functions and the amount of heat released by the given chemical reaction. However, for the second case, the asymptotic behavior of the solutions with the Arrhenius functions (1.4) is completely determined by the boundary data and the initial density, and is independent of the initial velocity, temperature, and mass fraction as well as the scale of the Arrhenius functions and the amount of heat released by the given chemical reaction. We remark that the discontinuous Arrhenius functions (1.4) can have different forms near the ignition temperature $T = T_i$. These results show that the asymptotic behavior of the solutions is independent of those different forms near $T = T_i$ for ϕ. From the difference between the two cases, we can also find a difference in the asymptotic behavior between a reacting mixture and a nonreacting gas [6].

Finally, in §5 we establish an existence theorem of global generalized solutions for the Cauchy problem of the compressible Navier–Stokes equations for the reacting mixture.

In connection with earlier work, we recall that Gardner [7] and Wagner [8] studied the existence and behavior of steady plane wave solutions to the compressible Navier–Stokes equations for a reacting gas and confirmed some phenomena observed in numerical calculations and predicted by the ZND theory, which had been developed independently by Zeldovich, von Neumann, and Döring (see [1] for details). Several theoretical and computational properties regarding the structure and stability of reacting shock waves of the system (1.1) are documented and analyzed in [9] and the references cited therein. For recent developments and strategies in the mathematical theory of combustion, we refer the reader to [2], [3], [10] and the references cited therein.

The existence of global solutions to the one-dimensional nonsteady equations of a viscous compressible gas was first studied in [11]–[12] for simple models. Kazhikhov and Shelukhin [6], [13] established the unique solvability and decay of the solutions for the initial boundary value problems of the viscous heat-conducting perfect gas with large initial data. For a systematic study of existence and decay of the solutions to the system of equations of a viscous compressible gas for small initial data, we refer the reader to [14]–[17] and the references cited therein. We also refer the reader to [18]–[19] for global well-posedness of the Cauchy problem for the nonreacting compressible flow with discontinuous initial data.

2. Equivalence of the Navier–Stokes equations in the Euler and Lagrange coordinates for generalized solutions. The relation between the Euler coordinates (t, x) and the Lagrange coordinates (t, y) is given by

$$P : (t, x) \to (t, y), \qquad y = \int_{x(t)}^{x} \rho(t, s) ds, \tag{2.1}$$

where $x(t)$ is a well-defined particle path satisfying $x'(t) = u(x(t), t)$. Using this transformation, we obtain the compressible Navier–Stokes equations for the reacting gas in the Lagrange coordinates from (1.1):

$$\begin{cases} v_t - u_y = 0, \\ u_t + \widetilde{p}_y = \left(\frac{\mu u_y}{v}\right)_y, \\ \widetilde{E}_t + (\widetilde{p}u)_y = \left(\frac{\mu u u_y}{v} + \frac{\nu T_y}{v} + \frac{q d Z_y}{v^2}\right)_y, \\ Z_t + K\phi(T)Z = \left(\frac{dZ_y}{v^2}\right)_y, \end{cases} \tag{2.2}$$

where $\widetilde{p}(v, T) = p\left(\frac{1}{\rho}, T\right)$, and $\widetilde{E}(u, v, T) = E\left(u, \frac{1}{\rho}, T\right)$.

The initial boundary value conditions (1.5)–(1.8) can be translated into similar conditions:

$$(v, u, T, Z)|_{t=0} = (v_0(y), u_0(y), T_0(y), Z_0(y)), \qquad 0 \le y \le 1, \tag{2.3}$$

$$\begin{cases} u(t, i) = 0, \\ T_y(t, i) = 0, \qquad i = 0, 1, \quad t \ge 0, \\ Z_y(t, i) = 0, \end{cases} \tag{2.4}$$

$$0 < m_0 \le v_0(y), T_0(y) \le M_0 < \infty, \qquad 0 \le Z_0(y) \le 1, \tag{2.5}$$

with the compatibility conditions:

$$\begin{cases} u_0(i) = 0, \\ T_{0y}(i) = 0, \\ Z_{0y}(i) = 0, \qquad i = 0, 1, \\ \left(\frac{aT_0}{v_0} - \left(\frac{\mu u_{0y}}{v_0}\right)_y\right)|_{y=i} = 0. \end{cases} \tag{2.6}$$

Here we take $x(t) = 0$ in (2.1) and, for simplicity, assume that $\int_0^1 \rho_0(x)dx = 1$.

Similarly, the boundary conditions (1.9) can be transformed into the following form:

$$\begin{cases} u(t, i) = 0, \\ T(t, i) = q_i(t), \qquad i = 0, 1, \quad t \ge 0, \\ Z(t, i) = r_i(t). \end{cases} \tag{2.7}$$

More general boundary conditions can be expressed by

$$(2.8)\qquad \begin{cases} u(t,i) = p_i(t), \\ T(t,i) = q_i(t), \qquad i = 0,1, \quad t \geq 0, \\ Z(t,i) = r_i(t), \end{cases}$$

with

$$(2.9)\qquad \begin{cases} u_0(i) = p_i(0), \quad T_0(i) = q_i(0), \quad Z_0(i) = r_i(0), \\ |p_i(t)| \leq M_0 < \infty, \quad 0 \leq r_i(t) \leq 1, \\ 0 < m_0 \leq q_i(t) \leq M_0 < \infty, \\ l(t) = \int_0^1 v_0(y)dy + \int_0^t (p_1(\tau) - p_0(\tau))d\tau \geq \delta > 0. \end{cases}$$

The initial conditions (1.13)–(1.14) can be written into the same form:

$$(2.10)\qquad (v,u,T,Z)|_{t=0} = (v_0(y), u_0(y), T_0(y), Z_0(y)), \qquad -\infty < y < \infty,$$

with

$$(2.11)\qquad \begin{cases} \lim_{|y|\to\infty} (v_0(y), u_0(y), T_0(y), Z_0(y)) = (1,0,1,0), \\ |u_0(y)| \leq M_0 < \infty, \qquad 0 \leq Z_0(y) \leq 1, \\ 0 < m_0 \leq v_0(y), \ T_0(y) \leq M_0 < \infty. \end{cases}$$

We have the following theorem.

THEOREM 1. *The coordinate transformation P given by* (2.1) *induces a one-to-one correspondence between the generalized solutions of* (1.1) *satisfying* $0 < \alpha \leq \rho(t,x) \leq M < \infty$, *almost everywhere for some α and M, and the generalized solutions of* (2.2) *satisfying* $0 < \beta \leq v(t,y) \leq N < \infty$, *almost everywhere for some β and N. In addition, there is a one-to-one correspondence between the corresponding prescribed boundary conditions and initial conditions for the corresponding generalized solutions.*

Proof. We first notice that the transformation $y = \int_{x(t)}^{x} \rho(t,s)ds$ is actually a classical formula for the solution to the gradient system:

$$\frac{\partial y}{\partial x} = \rho(t,x), \qquad \frac{\partial y}{\partial t} = -(\rho u)(t,x),$$

which is consistent because of $\rho_t = -(\rho u)_x$, and P is a bi-C^1 homeomorphism from $(0,L) \times \Omega$ onto itself because $0 < \alpha \leq \rho(t,x) \leq M < \infty$ and $(\rho, u) \in H^1((0,L) \times \Omega)$.

Similar to Wagner [20], we make the transformation P by using the change of variables for integrals with the Jacobian ρ for the generalized solutions (ρ, u, T, Z) to the system (1.1). We write the system (1.1) as

$$U_t + F(U)_x = (DU_x)_x + H(U),$$

with initial data $U(0,x) = U_0(x)$. Then

$$(2.12)\qquad \iint\limits_{(0,L)\times\Omega} (\phi_t U + \phi_x(F - DU_x) + \phi H)dxdt + \int\limits_{\Omega} \phi U_0|_{t=0}dx = 0$$

for all C^1 test functions ϕ with supp $\phi \subset\subset [0,L) \times \Omega$.

Using the transformation P for integrals, we obtain from (2.12)

$$\begin{aligned}
0 = \iint\limits_{(0,L)\times\Omega} &\{(\phi_t \circ P^{-1})(U \circ P^{-1}) + (\phi_x \circ P^{-1})(F \circ P^{-1} - (D \circ P^{-1})(U_x \circ P^{-1})) \\
&+ (\phi \circ P^{-1})(H \circ P^{-1})\} \frac{1}{(\rho \circ P^{-1})} dy\, dt \\
&+ \int_\Omega (\phi \circ P^{-1})(U_0 \circ P^{-1}) \frac{1}{(\rho_0 \circ P^{-1})}|_{t=0} dy \\
= \iint\limits_{(0,L)\times\Omega} &\{(\widetilde{\phi}_t - (\widetilde{\rho}\widetilde{u})\widetilde{\phi}_y)\widetilde{U} + (\widetilde{\phi}_y\widetilde{\rho})(\widetilde{F} - \widetilde{D}(\widetilde{U}_y\widetilde{\rho})) + \widetilde{\phi}\widetilde{H}\} \frac{1}{\widetilde{\rho}} dy\, dt + \int_\Omega \widetilde{\phi}\widetilde{U}_0 \frac{1}{\widetilde{\rho}_0} dy.
\end{aligned}$$

Note that P is a bi-C^1 homeomorphism of $[0,L) \times \Omega$ onto itself and that the induced map $\phi \mapsto \phi \circ P^{-1} = \widetilde{\phi}$ is a bijection on the set of test functions on $[0,L) \times \Omega$. We have

$$\begin{cases} \left(\frac{\widetilde{U}}{\widetilde{\rho}}\right)_t + (\widetilde{F} - \widetilde{u}\widetilde{U})_y = (\widetilde{\rho}\widetilde{D}\widetilde{U}_y)_y + \frac{\widetilde{H}}{\widetilde{\rho}}, \\ \frac{\widetilde{U}}{\widetilde{\rho}}|_{t=0} = \frac{\widetilde{U}_0}{\widetilde{\rho}_0}. \end{cases}$$

Thus the system (1.1) is transformed into the following form:

$$\begin{aligned}
1_t + 0_y &= 0, \\
u_t + \widetilde{p}(v,T)_y &= \left(\frac{\mu u_y}{v}\right)_y, \\
\widetilde{E}_t + (\widetilde{p}u)_y &= \left(\frac{\mu u u_y}{v} + \frac{\nu T_y}{v} + \frac{q d Z_y}{v^2}\right)_y, \\
Z_t + K\phi(T)Z &= \left(\frac{dZ_y}{v^2}\right)_y.
\end{aligned}$$

Moreover, from the conservation of volume, $1_t + 0_x = 0$, we have

$$\left(\frac{1}{\rho}\right)_t - u_y = 0,$$

that is,

$$v_t - u_y = 0.$$

Conversely, we can deduce the system (1.1) from the system (2.2) by checking the process step by step.

Furthermore, we can similarly prove the one-to-one correspondence between the corresponding prescribed boundary conditions for the generalized solutions, in the trace sense of the solution functions belonging to the corresponding classes. This completes the proof of Theorem 1.

In the following sections we restrict our attention to the existence and asymptotic behavior of the global generalized solutions to the compressible Navier–Stokes equations in the Lagrange coordinates. The results can be completely translated into the corresponding results for the compressible Navier–Stokes equations in the Euler coordinates. For instance, the existence theorems for the problems (2.2)–(2.3) and

(2.8)–(2.9) can be translated into existence theorems for the problems (1.1) and (1.5) with free boundary conditions of the following form.

Given $p_i(t) \in H^1(0,L)$, $i=0,1$, there exist unique C^1 functions $x_0(t)$ and $x_1(t)$ such that $x_i'(t) = p_i(t)$ and

$$\int_{x_0(t)}^{x_1(t)} \frac{1}{v(t,s)}ds = 1.$$

The boundary conditions are

$$\begin{cases} u(t,x_i(t)) = p_i(t), & \\ T(t,x_i(t)) = q_i(t), & i=0,1, \quad t \geq 0. \\ Z(t,x_i(t)) = r_i(t). & \end{cases}$$

In particular, if $p_i(t)=0$, $i=0,1$, and $\int_0^1 v_0(y)dy = 1$, then the existence theorems of the system (2.2) with the boundary conditions (2.3)–(2.6) and (2.8)–(2.9) are simply the existence theorems of the system (1.1) with the boundary conditions (1.5)–(1.8) and (1.9)–(1.12).

Now we introduce some notations for the subsequent development.

$$\begin{aligned} Q_L &= (0,L)\times\Omega, \\ B^{\sigma/2,\sigma}(Q_L) &= \left\{ u \in C(Q_L) : \sup_{Q_L}|u| + \sup_{\substack{t\neq s\\ x\neq y}} \frac{|u(t,x)-u(s,y)|}{|t-s|^{\sigma/2}+|x-y|^{\sigma}} < \infty \right\} \\ B^{1+\sigma}(Q_L) &= \{u \in B^{\sigma/2,\sigma}(Q_L) : u_t, u_y \in B^{\sigma/2,\sigma}(Q_L)\}, \\ B^{2+\sigma}(Q_L) &= \{u \in B^{s/2,\sigma}(Q_L) : u_t, u_y, u_{yy} \in B^{\sigma/2,\sigma}(Q_L)\}, \\ \|\cdot\|_{k+\sigma} &= \|\cdot\|_{B^{k+\sigma}}, \qquad k=1,2, \\ \|\cdot\| &= \|\cdot\|_{L^2}. \end{aligned}$$

3. Existence of global solutions of the initial boundary value problems. Scaling the variables, we can write the system (2.2) into the following form:

$$v_t - u_y = 0, \tag{3.1}$$

$$u_t + \left(\frac{aT}{v}\right)_y = \left(\frac{\mu u_y}{v}\right)_y, \tag{3.2}$$

$$\left(T + \frac{u^2}{2}\right)_t + \left(\frac{auT}{v}\right)_y = \left(\frac{\mu u u_y}{v} + \frac{\nu T_y}{v}\right)_y + qK\phi(T)Z, \tag{3.3}$$

$$Z_t + K\phi(T)Z = \left(\frac{d}{v^2}Z_y\right)_y. \tag{3.4}$$

For concreteness, we are concerned with the typical initial boundary value problems:

$$(v,u,T,Z)|_{t=0} = (v_0(y), u_0(y), T_0(y), Z_0(y)), \qquad 0 \leq y \leq 1, \tag{3.5}$$

$$\begin{cases} u(t,i) &= 0, & & \\ T_y(t,i) &= 0, & \text{or} \quad T(t,i)=1, & i=0,1, \quad t\geq 0, \\ Z_y(t,i) &= 0, & \text{or} \quad Z(t,i)=0, & \end{cases} \tag{3.6}$$

$$\begin{cases} 0 < m_0 \leq v_0(y), T_0(y) \leq M_0 < \infty, \quad 0 \leq Z_0(y) \leq 1, \\ \int_0^1 v_0(y)dy = 1, \end{cases} \tag{3.7}$$

with the compatibility conditions:

$$\begin{cases} u_0(i) = 0, & \\ T_{0y}(i) = 0, \quad \text{or} \quad T_0(i) = 1, \\ Z_{0y}(i) = 0, \quad \text{or} \quad Z_0(i) = 0, \\ \left(\frac{aT_0}{v_0} - \left(\frac{\mu u_{0y}}{v_0} \right)_y \right) |_{y=i} = 0. \end{cases} \tag{3.8}$$

At first, we assume that the reacting rate function $\phi(T)$ is a C^1 function and that

$$\begin{cases} ||\phi||_{C^1} & \leq \frac{1}{\epsilon} < \infty, \qquad 0 \leq \phi(T) \leq M < \infty, \\ \phi(T) & = 0 \quad \text{as } T \leq T_i, \end{cases} \tag{3.9}$$

where the constants ϵ and T_i are positive, and the constant M is independent of ϵ. Then we have the following theorem.

THEOREM 2 (Local solutions). *Suppose that there exists a constant $M_0 > 0$ such that*

$$\begin{aligned} &M_0^{-1} \leq \min(v_0, T_0), \quad \max(v_0, T_0) \leq M_0 < \infty, \quad 0 \leq Z_0(y) \leq 1, \\ &||v_0||_{1+\sigma}, \quad ||u_0, T_0, Z_0||_{2+\sigma} \leq M_0 < \infty, \end{aligned}$$

and that the conditions (3.8) *and* (3.9) *hold. Then, for any $M > M_0$, there exist constants $L_0 = L_0(M)$, $N = N(M, \epsilon) \geq M_0$ such that there exists a unique solution (v, u, T, Z) for every initial boundary value problem in* (3.1)–(3.9) *on Q_{L_0} that satisfies*

$$\begin{cases} v \in B^{1+\sigma}(Q_{L_0}), \qquad (u, T, Z) \in B^{2+\sigma}(Q_{L_0}), \\ M^{-1} \leq v \leq M, \qquad N^{-1} \leq T \leq M, \\ ||v||_{1+\sigma, L_0}, \ ||u, T, Z||_{2+\sigma, L_0} \leq N. \end{cases}$$

The proof of Theorem 2 is based on the Banach theorem and the contractivity of the operator:

$$\begin{gathered} B^{1+\sigma}(Q_{L_0}) \times (B^{2+\sigma}(Q_{L_0}))^3 \mapsto B^{1+\sigma}(Q_{L_0}) \times (B^{2+\sigma}(Q_{L_0}))^3, \\ (v, u, T, Z) \mapsto (\overline{v}, \overline{u}, \overline{T}, \overline{Z}), \end{gathered}$$

defined by the following linearization of the problems:

$$\begin{cases} \overline{v}(t, y) = v_0(y) + \displaystyle\int_0^t u_y(\tau, y)d\tau, \\ \overline{u}_t(t, y) = \dfrac{\mu}{v}\overline{u}_{yy} - \dfrac{\mu}{v^2}v_y\overline{u}_y - \dfrac{a}{v}\overline{T}_y + \dfrac{aT}{v^2}v_y, \\ \overline{T}_t(t, y) = \dfrac{\nu}{v}\overline{T}_{yy} - \dfrac{\nu}{v^2}v_y\overline{T}_y - \dfrac{au_y}{v}\overline{T} + \dfrac{\mu}{v}u_y^2 + qK\phi(T)Z, \\ \overline{Z}_t(t, y) = \dfrac{d}{v^2}\overline{Z}_{yy} - \dfrac{2d}{v^3}v_y\overline{Z}_y - K\phi(T)\overline{Z}, \end{cases}$$

with the initial boundary value conditions:

$$\begin{cases} (\overline{v},\overline{u},\overline{T},\overline{Z})|_{t=0} = (v_0(y),u_0(y),T_0(y),Z_0(y)), \\ \overline{u}|_{y=i} = 0, \\ \overline{T}_y|_{y=i} = 0 \quad \text{or} \quad \overline{T}|_{y=i} = 1, \quad i = 0,1, \quad t \geq 0, \\ \overline{Z}_y|_{y=i} = 0 \quad \text{or} \quad \overline{Z}|_{y=i} = 0, \end{cases}$$

on a small time interval $[0, L_0]$. We omit the details of the proof.

Based on the local solvability and the a priori estimates below, we can extend the local solutions in Theorem 2 and obtain the following existence theorem of global solutions.

THEOREM 3 (Classical solutions). *Let the initial data satisfy the conditions* (3.7)–(3.8), *and*

$$v_0 \in C^{1+\alpha}(0,1), \qquad (u_0,T_0,Z_0) \in C^{2+\alpha}(0,1),$$
$$||u_0,T_0,Z_0||_{H^1} \leq M_0,$$

and let the reacting rate function $\phi(T)$ satisfy the condition (3.9). *Then there exists a unique classical solution (v,u,T,Z) for every initial boundary value problem in* (3.1)–(3.9) *such that, for any $L>0$, there exist constants $M(M_0)$ (independent of ϵ and L), $N(M_0,\epsilon,L)>0$ and the following estimates hold on Q_L:*

$$(3.10) \quad \begin{cases} M^{-1} \leq v \leq M, \quad 0 < N^{-1} \leq T \leq M, \quad |u| \leq M, \quad 0 \leq Z(t,y) \leq 1, \\ ||v_t,v_y||^2(t) + \int_0^t ||v_t,v_y||^2(\tau)d\tau \leq M, \\ ||u_y,T_y,Z_y||^2(t) + \int_0^t ||u_{yy},T_{yy},Z_{yy},u_y,T_y,Z_y,u_t,T_t,Z_t||^2(\tau)d\tau \leq M, \\ ||v||_{1+\sigma,L}, \ ||u,T,Z||_{2+\sigma,L} \leq N. \end{cases}$$

Now we make these a priori estimates. The main insight from our analysis is the crucial independence of the a priori estimates on the parameter ϵ. For simplicity, we use M to denote all positive constants that depend only on α_0, M_0, and $||\phi||_\infty$, and are independent of ϵ, L, $||v_0||_{1+\sigma}$, $||u_0,T_0,Z_0||_{2+\sigma}$, and use N to denote the universal constant that specially depends on ϵ and L.

For concreteness, we first deal with the impermeably insulated boundaries.

We have from (3.2) and (3.3)

$$(3.11) \qquad T_t + \frac{aT}{v}u_y = \left(\frac{\nu T_y}{v}\right)_y + \frac{\mu u_y^2}{v} + qK\phi(T)Z.$$

We can immediately conclude the following lemma from (3.1)–(3.4) and (2.3)–(2.6).

LEMMA 1.

$$(3.12) \qquad \int_0^1 v(t,y)dy = \int_0^1 v_0(y)dy = 1,$$

$$(3.13) \qquad \int_0^1 Z(t,y)dy + \int_0^t \int_0^1 K\phi(T)Z\,dy\,d\tau = \int_0^1 Z_0(y)dy,$$

$$\int_0^1 \left(T + \frac{u^2}{2} + qZ\right) dy = \int_0^1 \left(T_0 + \frac{u_0^2}{2} + qZ_0\right) dy, \tag{3.14}$$

$$\begin{aligned} U(t) &+ \int_0^t (V(\tau) + W(\tau))d\tau \\ &= \int_0^1 \left[a(v_0 - 1 - \ln v_0) + \frac{u_0^2}{2} + (T_0 - 1 - \ln T_0)\right] dy \equiv E_0 < \infty, \end{aligned} \tag{3.15}$$

where

$$\begin{cases} U(t) = \displaystyle\int_0^1 \left[a(v - 1 - \ln v) + \frac{u^2}{2} + (T - 1 - \ln T)\right] dy, \\ V(t) = \displaystyle\int_0^1 \left(\frac{\mu u_y^2}{vT} + \frac{\nu T_y^2}{vT^2}\right) dy, \\ W(t) = -q \displaystyle\int_0^1 \frac{T-1}{T} K\phi(T) Z dy. \end{cases}$$

LEMMA 2. $0 \le Z(t,y) \le 1$.

Proof. Set $Y = Ze^{-\beta t}$, $\beta > 0$ constant. Then $Y(t,y)$ satisfies

$$\begin{cases} Y_t + (\beta + K\phi(T))Y = \left(\dfrac{d}{v^2} Y_y\right)_y, \\ Y_y|_{y=0,1} = 0, \\ Y|_{t=0} = Z_0(y). \end{cases}$$

We claim that

$$Y(t,y) \ge 0.$$

On the contrary, there exists $(t_0, y_0) \in [0, L] \times [0,1]$ such that

$$Y(t_0, y_0) = \min_{Q_L} Y(t,y) < 0.$$

Then

$$Y_y(t_0,y_0) = 0, \quad Y_t(t_0,y_0) \le 0, \quad Y_{yy}(t_0,y_0) \ge 0.$$

We have

$$\left(\frac{d}{v^2} Y_y\right)_y (t_0, y_0) \ge 0.$$

However,

$$(Y_t + (\beta + K\phi(T))Y)(t_0, y_0) < 0.$$

This is a contradiction. It follows that

$$Z(t,y) = Y(t,y)e^{\beta t} \ge 0.$$

On the other hand, if we multiply both sides of (3.4) by pZ^{p-1}, $p \ge 2$, then we have

$$(Z^p)_t = \left(\frac{pdZ^{p-1}Z_y}{v^2}\right)_y - p\left(K\phi(T)Z^p + (p-1)Z^{p-2}\frac{dZ_y^2}{v^2}\right) \le \left(\frac{dpZ^{p-1}Z_y}{v^2}\right)_y.$$

Integrating this inequality from zero to 1, we obtain

$$\frac{d}{dt}||Z||_{L^p}(t) \leq 0.$$

This means that

$$||Z||_{L^p}(t) \leq ||Z_0||_{L^p}.$$

Let $p \to +\infty$. We obtain

$$Z(t,y) \leq ||Z_0||_{L^\infty} \leq 1.$$

LEMMA 3. $M^{-1} \leq v(t,y) \leq M$.

Proof. We know from Kazhikov [6], and Kazhikov and Shelukhin [13] that $v(t,y)$ and $T(t,y)$ satisfy the following equalities and estimates using (3.1)–(3.2) and the boundary conditions $(u(t,x),T(t,y))|_{x=0,1} = (0,1)$:

(1) For any $t \geq 0$, there exists $y_1(t) \in [0,1]$ such that

(3.16)
$$\begin{aligned} v(t,y) =&D(t,y)\mathrm{exp}\left(-\frac{1}{\mu}\int_0^t\int_0^1(u^2+aT)d\xi d\tau\right) \\ &\cdot\left\{1+\frac{a}{\mu}\int_0^t\frac{T(\tau,y)}{D(\tau,y)}\mathrm{exp}\left(\frac{1}{\mu}\int_0^\theta au\int_0^1(u^2+aT)d\xi ds\right)d\tau\right\},\end{aligned}$$

where

$$\begin{aligned} D(t,y) =&v_0(y)\mathrm{exp}\left\{\frac{1}{\mu}\left(\int_{y_1(t)}^y u(t,\xi)d\ xi - \int_0^y u_0(\xi)d\xi\right.\right. \\ &\left.\left.+\int_0^1 v_0(\eta)\left(\int_0^\eta u_0(\xi)d\xi\right)d\eta\right)\right\}. \end{aligned} \tag{3.17}$$

(2) For any $t \geq 0$, there exists $y_2(t) \in [0,1]$ such that

$$v(t,y) = \frac{1+\int_0^t \frac{a}{\mu}T(\tau,y)P(\tau)Q(\tau,y)d\tau}{P(t)Q(t,y)}. \tag{3.18}$$

Here

$$\begin{cases} P(t) = v_0(y_2(t))\mathrm{exp}\left\{\displaystyle\int_0^t \frac{a}{\mu}\frac{T}{v}(\tau,y_2(t))d\tau\right\}, \\ Q(t,y) = \dfrac{1}{v_0(y)v(t,y_2(t))}\mathrm{exp}\left\{\dfrac{1}{\mu}\displaystyle\int_{y_2(t)}^y(u_0(\xi)-u(t,\xi))d\xi\right\}, \end{cases} \tag{3.19}$$

and

$$\alpha_0 \leq v(t,y_2(t)),T(t,y_2(t)) \leq \beta_0, \tag{3.20}$$

where the positive constants α_0 and β_0 are roots of

$$y-1-\mathrm{ln}y = E_1 \equiv \frac{1}{\min\{1,a\}}\left(E_0+q\int_0^1 Z_0(y)dy\right).$$

(3) $T(t,y)$ satisfies

$$\alpha_0 \le \int_0^1 T(t,y)dy \le \beta_0. \tag{3.21}$$

The estimate (3.21) is a direct corollary of the Jensen inequality and the convexity of the function $T - 1 - \ln T$. In fact, we have

$$\int_0^1 T(t,y)dy - 1 - \ln \int_0^1 T(t,y)dy \le \int_0^1 (T(t,y) - 1 - \ln T(t,y))dy \le E_1 < +\infty,$$

by using (3.13) and (3.15).

Set

$$\begin{cases} m_v(t) = \min\limits_{y\in[0,1]} v(t,y), & M_v(t) = \max\limits_{y\in[0,1]} v(t,y), \\ m_T(t) = \min\limits_{y\in[0,1]} T(t,y), & M_T(t) = \max\limits_{y\in[0,1]} T(t,y). \end{cases}$$

Using (3.15), we conclude that there exists a constant $M > 0$, which is independent of t, such that

$$M^{-1} \le D(t,y), Q(t,y) \le M < \infty. \tag{3.22}$$

Furthermore, we observe that

$$0 < a_1 \le \int_0^1 (u(t,y)^2 + aT(t,y))dy \le a_2 < +\infty, \tag{3.23}$$

by using (3.13), (3.15), and (3.21), where

$$\begin{cases} a_1 = \alpha_0 a, \\ a_2 = \beta_0 a + 2(E_0 + q\int_0^1 Z_0(y)dy). \end{cases}$$

Therefore, we have from (3.16) and (3.23) that

$$v(t,y) \le M\exp(-a_1 t)\left\{1 + \frac{a}{\mu}\int_0^t M_T(\tau)\exp(a_1\tau)d\tau\right\}.$$

Thus

$$M_v(t) \le M\exp(-a_1 t)\left\{1 + \frac{a}{\mu}\int_0^t M_T(\tau)\exp(a_1\tau)d\tau\right\}. \tag{3.24}$$

Moreover, for $y_2(t) \in [0,1]$ satisfying (3.20), we have

$$\begin{aligned} T(t,y)^{1/2} &= T(t,y_2(t))^{1/2} + \int_{y_2(t)}^{y} \frac{T_y}{2T^{1/2}}dy \\ &\le T(t,y_2(t))^{1/2} + \left(\frac{1}{4\nu}V(t)M_v(t)\int_0^1 Tdy\right)^{1/2} \\ &\le T(t,y_2(t))^{1/2} + \left(\frac{\beta_0}{4\nu}V(t)M_v(t)\right)^{1/2}. \end{aligned}$$

Similarly, we have

$$T(t,y)^{1/2} \geq T(t,y_2(t))^{1/2} - \left(\frac{\beta_0}{2\nu}V(t)M_v(t)\right)^{1/2}.$$

Therefore,

$$\left\{\begin{array}{l} M_T(t) \leq M(1+M_v(t)V(t)), \\ m_T(t) \geq M^{-1}(1-M_v(t)V(t)). \end{array}\right. \tag{3.25}$$

Using (3.24) and (3.25), we have

$$M_v(t) \leq M\left\{1+\int_0^t \exp(-a_1(t-\tau))V(\tau)M_v(\tau)d\tau\right\},$$

and

$$M_v(t) \leq M\exp\left(M\int_0^\infty V(\tau)d\tau\right)\{1+\exp(-a_1 t)\} \leq M. \tag{3.26}$$

Furthermore, we have

$$m_v(t) \geq M^{-1}\exp(-a_2 t)\left\{1+\frac{a}{\mu}\int_0^t m_T(\tau)\exp(a_2\tau)d\tau\right\}. \tag{3.27}$$

Substitute (3.25) into (3.27). We obtain

$$m_T(t) \geq M^{-1}\left\{1-M\int_0^t \exp(-a_2(t-\tau))V(\tau)d\tau\right\}.$$

Notice that

$$\lim_{t\to+\infty}\int_0^t \exp(-a_2(t-\tau))V(\tau)d\tau = 0.$$

Therefore, there exists an $L_0 > 0$ such that

$$m_v(t) \geq \frac{M^{-1}}{2} \quad \text{when } t > L_0. \tag{3.28}$$

On the other hand, we have from (3.12), (3.18), and (3.22) that

$$\begin{aligned} P(t) &= \int_0^1 P(t)v(t,y)dy \\ &= \int_0^1 \frac{1+\int_0^t \frac{a}{\mu}T(\tau,y)P(\tau)Q(\tau,y)d\tau}{Q(t,y)}dy \\ &\leq M\left(1+\int_0^t P(\tau)d\tau\int_0^1 T(\tau,y)dy\right) \\ &\leq M\left(1+\int_0^t P(\tau)d\tau\right). \end{aligned}$$

This means that

$$P(t) \leq Me^{Mt},$$

that is,

$$m_v(t) \geq P(t)^{-1}Q(t,y)^{-1} \geq M^{-1}e^{-ML_0} > 0 \quad \text{when} \quad 0 \leq t \leq L_0. \tag{3.29}$$

Combining (3.28) with (3.29), we have

$$m_v(t) \geq M^{-1} > 0.$$

This completes the proof of Lemma 3.

LEMMA 4. $T(t,y) \geq N^{-1} > 0$.

Proof. Set $\theta = T^{-1}$. Multiplying (3.11) by T^{-2}, we obtain

$$\theta_t = \left(\nu \frac{\theta_y}{v}\right)_y + \frac{a^2}{4\mu}\frac{1}{v} - \frac{qK\phi(T)}{T^2}Z - \left\{\frac{2\nu T}{v}\theta_y^2 + \frac{\mu\theta^2}{v}\left(u_y - \frac{aT}{\mu}\right)^2\right\}. \tag{3.30}$$

Multiplying (3.30) by $2l\theta^{2l-1}, l = 1,2,3,\cdots$, we have

$$\begin{aligned}(\theta^{2l})_t &\leq \frac{a^2}{4\mu}\frac{2l\theta^{2l-1}}{v} - 2l(2l-1)\theta^{2l-2}\frac{\theta_y^2}{v} + \left(\frac{\nu(\theta^{2l})_y}{v}\right)_y \\ &\leq \frac{a^2}{4\mu}\frac{2l\theta^{2l-1}}{v} + \left(\frac{\nu(\theta^{2l})_y}{v}\right)_y.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}(\|\theta\|_{L^{2l}}(t))^{2l-1}\frac{d}{dt}\|\theta\|_{L^{2l}}(t) &\leq \frac{a^2}{4\mu}\int_0^1 \frac{\theta^{2l-1}}{v}dy \\ &\leq (\|\theta\|_{L^{2l}}(t))^{2l-1}\frac{a^2}{4\mu}\left\|\frac{1}{v}\right\| L^{2l}(t).\end{aligned}$$

Using the Hölder's inequality, we have

$$\begin{aligned}\|\theta\|_{L^{2l}}(t) &\leq \|\theta\|_{L^{2l}}(0) + \frac{a^2}{4\mu}\int_0^t \|\frac{1}{v}\|_{L^{2l}}(\tau)d\tau \\ &\leq \|\theta\|_{L^{2l}}(0) + \frac{a^2}{4\mu}\int_0^t \frac{1}{m_v(\tau)}d\tau \\ &\leq \|\theta\|_{L^{2l}}(0) + Mt.\end{aligned}$$

Set $l \to +\infty$. We obtain

$$m_T(t)^{-1} \leq M(1+t),$$

that is,

$$m_T(t) \geq M^{-1}(1+t)^{-1} \geq N^{-1} > 0.$$

LEMMA 5.

$$\|v_y, u, T-1, Z\|^2(t) + \int_0^t \|v_y, u_y, T_y, Z_y\|^2(\tau)d\tau + \int_0^t\int_0^1 K\phi(T)Z^2 dy\, d\tau \leq M.$$

Proof. Set $w = T + \frac{u^2}{2} - 1$. Then

$$w_t = \left(\frac{\nu T_y}{v}\right)_y + \left(\frac{\mu u u_y}{v}\right)_y - \left(\frac{auT}{v}\right)_y + qK\phi(T)Z.$$

We multiply both sides of the above equality and (3.2) by w and u^3, respectively. We then integrate them with respect to y from zero to one and use the interpolation formula. We obtain

$$
\begin{aligned}
&\int_0^1 (w^2+u^4)dy + \int_0^t\int_0^1 (T_y^2+u^2u_y^2)dy\,d\tau \\
\le& M\left\{\int_0^1 (w_0(y)^2+u_0(y)^4)dy \right.\\
&\left. + \int_0^t\int_0^1 (u^2T^2+K^2\phi(T)^2Z^2+K\phi(T)Z)dy\,d\tau\right\} \\
\le& M\left\{1+\int_0^t\int_0^1 u^2(w^2+u^4)dy\,d\tau \right.\\
&\left. + \int_0^t\int_0^1 u^2(u^2+1)dy\,d\tau + \int_0^t\int_0^1 K\phi(T)Z dy\,d\tau\right\} \\
\le& M\left(1+\int_0^t \max_y u(\tau,y)^2\int_0^1 (w^2+u^4)dy\,d\tau \right.\\
&\left. + \int_0^t \max_y u(\tau,y)^2\int_0^1 (u^2+1)dy\,d\tau\right) \\
\le& M\left(1+\int_0^t \max_y u(\tau,y)^2 d\tau \right.\\
&\left. + \int_0^t \max_y u(\tau,y)^2\int_0^1 (w^2+u^4)dy\,d\tau\right)
\end{aligned}
$$

by using the estimate

$$\left(T+\frac{u^2}{2}-\int_0^1 T(t,y)dy\right)^2 \le \int_0^1 (T_y^2+u^2u_y^2)dy.$$

Notice that

$$u^2(t,y) \le \left(\int_0^y u_y^2 dy\right)^2 \le \int_0^1 \frac{u_y^2}{vT}dy\int_0^1 vTdy \le MV(t),$$

and $V(t)$ is uniformly integrable. We have

$$||T-1||^2(t) + \int_0^t ||T_y||(\tau)d\tau \le M$$

by using the Gronwall's inequality.

Furthermore, we have

$$u_t + a\frac{T_y}{v} = \frac{aT}{v}\frac{v_y}{v} + \mu\left(\frac{v_y}{v}\right)_t$$

from (3.1) and (3.2). We multiply both sides of the above formula and (3.2) by v_y/v and u, respectively, and then integrate them. We obtain

$$\mu\int_0^1\left(\frac{v_y}{v}\right)^2 dy + \int_0^t\int_0^1\left(\frac{a}{v}T\left(\frac{v_y}{v}\right)^2+\mu\frac{u_y^2}{v}\right)dy\,d\tau \le M.$$

Notice that there exists $0 < \beta < 1$ such that

$$\frac{\int_0^1 \frac{u^2}{2} dy}{\int_0^1 (T + \frac{u^2}{2}) dy} < \beta$$

from (3.21). We have

$$\begin{aligned}
\int_0^1 v_y^2 dy &= \left(\int_0^1 \left(T + \frac{u^2}{2}\right) dy\right)^{-1} \int_0^1 \left(\int_0^1 T + \frac{u^2}{2} dy\right) v_y^2 dy \\
&= \left(\int_0^1 \left(T + \frac{u^2}{2}\right) dy\right)^{-1} \left(\int_0^1 \left(\int_0^1 T dy - T\right) v_y^2 dy + \int_0^1 T v_y^2 dy \right. \\
&\qquad \left. + \int_0^1 \frac{u^2}{2} dy \int_0^1 v_y^2 dy\right) \\
&\le \frac{1+\beta}{2} \int_0^1 v_y^2 dy + M \left(\int_0^1 \left(\int_0^1 T dy - T\right)^2 v_y^2 dy + \int_0^1 T v_y^2 dy\right).
\end{aligned}$$

Therefore,

$$\int_0^1 v_y^2 dy \le M \left(\int_0^1 T_y^2 dy \int_0^1 v_y^2 dy + \int_0^1 T v_y^2 dy\right).$$

Notice that

$$\int_0^t \int_0^1 (T_y^2 + T v_y^2) dy\, dt \le M.$$

We immediately conclude that

$$\int_0^t \int_0^1 v_y^2 dy\, d\tau \le M.$$

Finally, we multiply (3.4) by Z and obtain

$$\left(\frac{Z^2}{2}\right)_t + \frac{d}{v^2} Z_y^2 + K\phi(T) Z^2 = \left(\frac{dZZ_y}{v^2}\right)_y.$$

Therefore, we have

$$\int_0^1 Z^2 dy + \int_0^t \int_0^1 (Z_y^2 + K\phi(T) Z^2) dy\, d\tau \le M.$$

This completes the proof of Lemma 5.

LEMMA 6.

$$||v_t, u_y, T_y, Z_y||^2(t) + \int_0^t ||v_{yt}, u_{yy}, T_{yy}, Z_{yy}, u_t, T_t, Z_t||^2(\tau) d\tau \le M.$$

Proof. We multiply both sides of (3.2) by u_{yy} and use the interpolation inequalities:

$$\begin{cases} \max\limits_y u_y^2 \le \delta \int_0^1 u_{yy}^2 dx + C_\delta \int_0^1 u_y^2 dy, \\ \max\limits_y T^2 \le 2 \int_0^1 T_y^2 dy + 2 \left(\int_0^1 T dy\right)^2. \end{cases}$$

We deduce that

$$||u_y||^2(t) + \int_0^t ||u_{yy}||^2(\tau)d\tau \le M$$

by using Lemma 5.

Next, we multiply both sides of (3.11) by T_{yy} and integrate it. We have

$$\left(\int_0^1 \frac{T_y^2}{2}dy\right)_t + \nu\int_0^1 \frac{T_{yy}^2}{v}dy = \int_0^1 \left(\frac{aT}{v}u_y - \frac{\mu u_y^2}{v} - qK\phi(T)Z\right)T_{yy}dy,$$

and then

$$\left(\int_0^1 \frac{T_y^2}{2}dy\right)_t + \int_0^1 T_{yy}^2dy \le M\int_0^1 (T^2u_y^2 + u_y^2 + q^2K^2\phi(T)^2Z^2)dy.$$

Notice that

$$\int_0^1 u_y^4dy \le \left(\delta\int_0^1 u_{yy}^2dy + C_\delta\int_0^1 u_y^2dy\right)\int_0^1 u_y^2dy,$$

$$\int_0^1 T^2u_y^2dy \le 2\left(\int_0^1 T_y^2dy + \left(\int_0^1 Tdy\right)^2\right)\int_0^1 v_y^2dy,$$

$$\int_0^t\int_0^1 K^2\phi(T)^2Z^2dy\,d\tau \le M\int_0^t\int_0^1 K\phi(T)Z^2dy\,d\tau \le M.$$

We obtain

$$||T_y||^2(t) + \int_0^t ||T_{yy}||^2(\tau)d\tau \le M.$$

Finally, multiplying (3.4) by Z_{yy}, we have

$$\int_0^1 \frac{Z_y^2}{2}dy + \int_0^t\int_0^1 \frac{d}{v^2}Z_{yy}^2dy\,d\tau = \int_0^t\int_0^1 \left(K\phi(T)ZZ_{yy} + \frac{2dv_y}{v^3}Z_yZ_{yy}\right)dy\,d\tau$$
$$\le M\int_0^t\int_0^1 (K\phi(T)Z^2 + Z_y^2)dy\,d\tau \le M$$

by using $\int_0^1 v_y^2dy \le M$ and the interpolation inequality

$$\max_y Z_y^2 \le \delta\int_0^1 Z_{yy}^2dx + C_\delta\int_0^1 Z_y^2dy.$$

This completes the proof of Lemma 6 by using (3.1).

LEMMA 7. $|u| \le M,\ 0 < N^{-1} \le T \le M < \infty$.

Proof. Notice that

$$u^2 \le \int_0^y (u^2)_ydy \le 2\left(\int_0^1 u^2dy\right)^{1/2}\left(\int_0^1 u_y^2dy\right)^{1/2} \le M.$$

We define an auxiliary function f by

$$f(T) = \int_0^T \sqrt{T - 1 - \ln T}dT.$$

Then

$$f(T) \to +\infty \quad \text{as } T \to +\infty.$$

However,

$$\begin{aligned}|f(T)| &\le \int_0^y \sqrt{T-1-\ln T}|T_y|dy\\ &\le \left(\int_0^1 (T-1-\ln T)dy\right)^{1/2}\left(\int_0^1 T_y^2 dy\right)^{1/2} \le M.\end{aligned}$$

Therefore,

$$0 < N^{-1} \le T \le M < \infty.$$

This proves Lemma 7.

With these a priori estimates and $\phi(T) \in C^1$, we obtain the following estimates by using the standard method (cf. [11]–[12]) and the Schauder estimates (cf. [21]).

LEMMA 8. $||v||_{1+\sigma,L}$, $||u,T,Z||_{2+\sigma,L} \le N$.

This proves Theorem 3 for the impermeably insulated boundary conditions.

For the Dirichlet-type boundary conditions, we must use other techniques to arrive at the result.

Notice that the equalities (3.12) and (3.15) still hold by using (3.1) and (3.2), and the boundary conditions $(u(t,y),T(t,y))|_{y=0,1} = (0,1)$. Now we consider (3.4) with the initial boundary value conditions:

$$\left\{\begin{array}{l} Z|_{y=0,1} = 0,\\ Z|_{t=0} = Z_0(x) \ge 0.\end{array}\right. \tag{3.31}$$

Then we have the following lemma.

LEMMA 9. *The solution* $Z(t,y)$ *of* (3.4) *and* (3.31) *satisfies*

$$\left\{\begin{array}{l} 0 \le Z(t,y) \le 1,\\ \int_0^1 Z(t,y)dy + K\int_0^t\int_0^1 \phi(T)Z dy\,d\tau \le \int_0^1 Z_0(y)dy.\end{array}\right. \tag{3.32}$$

Proof. Similarly to the proof of Lemma 2, we first conclude that

$$Z(t,y) \ge 0.$$

The estimate

$$Z(t,y) \le 1$$

is a direct corollary of the maximal principle.

We multiply (3.4) by $\beta Z^{\beta-1}, \beta \in (1,2)$. Then we have

$$(Z^\beta)_t + \beta K\phi(T)Z^\beta = \beta Z^{\beta-1}\left(\frac{d}{v^2}Z_y\right)_y.$$

Thus

$$\int_0^1 Z(t,y)^\beta dy + \beta\int_0^t\int_0^1 K\phi(T)Z^\beta dy\,d\tau \le \int_0^1 Z_0^\beta(y)dy.$$

Notice that, for any fixed point $(\tau, y) \in (0, t] \times (0, 1)$,

$$\begin{cases} Z(\tau, y)^{\beta} \longrightarrow Z(\tau, y), \\ Z_0(y)^{\beta} \longrightarrow Z_0(y), \end{cases}$$

when $\beta \longrightarrow 1$ and

$$|Z^{\beta}(\tau, y)| \leq 1.$$

Using the dominated convergence theorem, we obtain

$$\int_0^1 Z(t, y)dy + K \int_0^t \int_0^1 \phi(T) Z dy d\tau \leq \int_0^1 Z_0(y)dy.$$

This completes the proof of Lemma 9.

LEMMA 10.

$$T \geq N^{-1} > 0.$$

Proof. Set $X = Te^{-\beta t}, \beta = \max_{Q_L} |\frac{au_y}{v}|$. Then $X(t, y)$ satisfies

$$\begin{cases} X_t + (\beta + \frac{au_y}{v})X = \left(\frac{\nu X_y}{v}\right)_y + \frac{\mu u_y^2}{v} e^{-\beta t} + qK\phi(Xe^{\beta t})Ze^{-\beta t}, \\ X|_{y=0,1} = e^{-\beta t}, \\ X|_{t=0} = T_0(x). \end{cases}$$

We conclude that

$$X(t, y) \geq \min(e^{-\beta L}, \min_{y \in [0,1]} T_0(y)) > 0$$

by using the maximal principle, which means that

$$T(t, y) \geq N^{-1} > 0.$$

Using Lemma 9, we can obtain the uniform bound of $v(t, y)$, which is independent of t and ϵ.

LEMMA 11. $M^{-1} \leq v(t, y) \leq M$.

In fact, using Lemma 9 and (3.15), we can obtain the estimate (3.23). Following the proof of Lemma 3, we then can complete the proof of Lemma 11.

The remaining arguments are the same as those discussed for the insulated problem. For the mixed problems, we can arrive at the result in the same fashion. This completes the proof of the theorem.

We know that the reacting rate function $\phi(T)$ is typically of the form (1.4) because of the ignition phenomenon in the combustion process. We modify $\phi(T)$ as follows:

$$\begin{cases} \phi_\epsilon(T) = \begin{cases} T^{\alpha} e^{-A/T}, & T \geq T_i + \epsilon, \\ 0, & T \leq T_i - \epsilon, \end{cases} \\ ||\phi_\epsilon||_{C^1} \leq \frac{1}{\epsilon} < \infty. \end{cases}$$

Then we follow the scheme given above and obtain global solutions $(v^\epsilon, u^\epsilon, T^\epsilon, Z^\epsilon)$ such that

$$v^\epsilon \in B^{1+\sigma}(Q_L), \qquad (u^\epsilon, T^\epsilon, Z^\epsilon) \in B^{2+\sigma}(Q_L),$$

and

$$(3.33)\quad \begin{cases} M^{-1} \le v^\epsilon \le M, \quad N^{-1} \le T^\epsilon \le M, \quad |u^\epsilon| \le M, \quad 0 \le Z^\epsilon(t,y) \le 1, \\ ||v^\epsilon_t, v^\epsilon_y||^2(t) + \int_0^t ||v^\epsilon_t, v^\epsilon_y||^2(\tau)d\tau \le M, \\ ||u^\epsilon_y, T^\epsilon_y, Z^\epsilon_y||^2(t) + \int_0^t ||v^\epsilon_{yt}, u^\epsilon_{yy}, T^\epsilon_{yy}, Z^\epsilon_{yy}, u^\epsilon_y, T^\epsilon_y, Z^\epsilon_y, Z^\epsilon_y, u^\epsilon_t, T^\epsilon_t, Z^\epsilon_t||^2(\tau)d\tau \le M, \end{cases}$$

where M is the constant independent of ϵ.

Using this fact, we can obtain the following theorem.

THEOREM 4 (Generalized solution). *Let the initial data satisfy* (3.7)–(3.8), *and*

$$(v_0(y), u_0(y), T_0(y), Z_0(y)) \in H^1(0,1).$$

Then there exists a generalized solution (v, u, T, Z) *for every initial boundary value problem in* (3.1)–(3.8) *with the discontinuous Arrhenius function* (1.4) *satisfying the same estimates as* (3.33).

In fact, we can immediately assert from (3.2) that there exists a subsequence converging in the strong topology of L^2 such that

$$(v^\epsilon, u^\epsilon, T^\epsilon, Z^\epsilon) \to (v, u, T, Z) \in H^1,$$

where (v, u, T, Z) is a generalized solution for the corresponding problem.

Remark. These existence arguments can be generalized to general initial boundary value problems. As an example, we consider the problem (3.1), (2.3), and (2.8)–(2.9).

Construct function $(u_1(t,y), T_1(t,y), Z_1(t,y))$ as follows:

$$u_1(t,y) = \frac{k(t)}{l(t)} \int_0^y v(t,\xi)d\xi + p_0(t),$$

where $k(t) = p_1(t) - p_0(t)$.

$T_1(t,y)$ and $Z_1(t,y)$ are solutions of the following linear problems:

$$\begin{cases} \dfrac{\partial T_1}{\partial t} = \dfrac{\partial}{\partial y}\left(\dfrac{\nu}{v}\dfrac{\partial T_1}{\partial y}\right), \quad 0 < y < 1, \quad 0 < t < L, \\ T_1|_{y=i} = q_i(t), \quad i = 0, 1, \quad t \ge 0, \\ T_1|_{t=0} = T_0(y), \end{cases}$$

and

$$\begin{cases} \dfrac{\partial Z_1}{\partial t} = \dfrac{\partial}{\partial y}\left(\dfrac{d}{v^2}\dfrac{\partial Z_1}{\partial y}\right), \quad 0 < y < 1, \quad 0 < t < L, \\ Z_1|_{y=i} = r_i(t), \quad i = 0, 1, \quad t \ge 0, \\ Z_1|_{t=0} = Z_0(y), \end{cases}$$

respectively.

Using the functions and mapping

$$(u, T, Z) \to (w, X, Y)$$

by

$$\begin{cases} w(t,y) = u(t,y) - u_1(t,y), \\ X(t,y) = \dfrac{T(t,y)}{T_1(t,y)}, \\ Y(t,y) = Z(t,y) - Z_1(t,y), \end{cases}$$

we transform (3.1)–(3.4), (2.3), and (2.8)–(2.9) into the following problem:

$$\begin{cases} v_t = w_y + \dfrac{k}{l}v, \\ w_t = \left(\dfrac{\mu}{v}w_y\right)_y - \left(\dfrac{aT_1\theta}{v}\right)_y - u_{1t}, \\ X_t = \left(\dfrac{\nu X_y}{v}\right)_y + \dfrac{2\nu}{v}\dfrac{T_{1y}}{T_1}X_y + \dfrac{\mu u_y^2}{vT_1} - a\dfrac{X}{v}u_y + K\dfrac{\phi(T_1X)}{T_1}Z, \\ Y_t = \left(\dfrac{d}{v^2}Y_y\right)_y - K\phi(T_1X)(Y+Z_1), \end{cases} \tag{3.34}$$

$$\begin{cases} w|_{y=0,1} = 0, \\ X|_{y=0,1} = 1, \qquad t \geq 0, \\ Y|_{y=0,1} = 0, \end{cases} \tag{3.35}$$

and

$$(v,w,X,Y)|_{t=0} = \left(v_0(y), u_0(y) - \frac{k(0)}{l(0)}\int_0^y v_0(\xi)d\xi - p_0(0), 1, 0\right). \tag{3.36}$$

In the same manner we can solve the problem (3.34)–(3.36) and obtain similar estimates of the solutions to (3.21) for $0 \leq t \leq L < +\infty$.

4. Asymptotic behavior of the solutions for the initial boundary value problems. For concreteness, in this section we restrict our attention to the two typical boundary conditions:

$$(\mathrm{I})\begin{cases} u(t,i) = 0, \\ T_y(t,i) = 0, \qquad i = 0,1, \quad t \geq 0, \\ Z_y(t,i) = 0, \end{cases}$$

and

$$(\mathrm{II})\begin{cases} u(t,i) = 0, \\ T(t,i) = 1, \qquad i = 0,1, \quad t \geq 0, \\ Z(t,i) = 0. \end{cases}$$

We compare the asymptotic behavior of the solutions to the system (3.1)–(3.4), satisfying the boundary conditions (I) and (II) with the initial condition (3.5) and discontinuous Arrhenius functions (1.4). We have the following theorem.

THEOREM 5. *Let the initial data satisfy the conditions of Theorem 4, and let the reacting rate function be of the form* (1.4). *Then*

(A) *Any generalized solution* (v,u,T,Z) *to the system* (3.1)–(3.4) *with the boundary conditions* (I) *satisfies*

$$||v - \int_0^1 v_0(y)dy, u, T - T_\infty, Z - Z_\infty||_{H^1(0,1)}(t) \to 0 \quad \text{as } t \to \infty, \tag{4.1}$$

where the constants T_∞ *and* Z_∞ *satisfy*

$$\begin{cases} T_\infty + qZ_\infty = \displaystyle\int_0^1 \left(T_0(y) + qZ_0(y) + \frac{u_0^2(y)}{2}\right) dy, \\ \phi(T_\infty)Z_\infty = 0. \end{cases}$$

(B) *Any generalized solution* (v,u,T,Z) *to the system* (3.1)–(3.4) *with the boundary conditions* (II) *satisfies*

$$||v - \int_0^1 v_0(y)dy, u, T - 1, Z||_{H^1(0,1)}(t) \to 0 \quad \text{as } t \to \infty. \tag{4.2}$$

Proof. The generalized solutions of the two problems satisfy the estimate

$$\int_0^\infty ||v_y, u_y, T_y, Z_y, v_t, u_t, T_t, Z_t, v_{yt}, u_{yy}, T_{yy}, Z_{yy}||^2(t)dt \le M < \infty$$

from Theorem 4. Therefore, we have

$$\int_0^\infty \left(\left|\frac{d}{dt}||v_y||^2(t)\right| + \left|\frac{d}{dt}||u_y||^2(t)\right| + \left|\frac{d}{dt}||T_y||^2(t)\right| \right. \\ \left. + \left|\frac{d}{dt}||Z_y||^2(t)\right| \right) dt \le M$$

and

$$\int_0^\infty (||v_y||^2(t) + ||u_y||^2(t) + ||T_y||^2(t) + ||Z_y||^2(t))\, dt \le M.$$

This means that

$$||v_y, u_y, T_y, Z_y||(t) \to 0 \quad \text{as } t \to \infty. \tag{4.3}$$

For (A), we have

(4.4)
$$\begin{cases} \left(v - \displaystyle\int_0^1 v_0(y)dy\right)^2 = \left(v - \displaystyle\int_0^1 v(t,y)dy\right)^2 \le \displaystyle\int_0^1 v_y^2 dy = ||v_y||^2(t) \to 0, \\ u^2 = \left(\displaystyle\int_0^y u_y dy\right)^2 \le \displaystyle\int_0^1 u_y^2 dy = ||u_y||^2(t) \to 0, \qquad t \to \infty, \\ (T + qZ - (T_\infty + qZ_\infty))^2 \le \left(T + qZ + \dfrac{u^2}{2} - \displaystyle\int_0^1 (T + qZ + \frac{u^2}{2})dy\right)^2 + Mu^2 \\ \quad \le ||T_y||^2(t) + q^2||Z_y||^2(t) + M||u_y||^2(t) \to 0 \end{cases}$$

by using (3.12) and (3.14).

Furthermore, since

$$\int_0^\infty \int_0^1 \phi(T) Z \, dy \, dt \leq M < \infty,$$

we deduce that

$$\phi(T_\infty) Z_\infty = 0.$$

This gives us (4.1).

For (B), we have

$$\left\{ \begin{array}{l} \left(v - \int_0^1 v_0(y) dy\right)^2 = \left(v - \int_0^1 v(t,y) dy\right)^2 \leq ||v_y||^2(t) \rightarrow 0, \\ u^2 = \left(\int_0^y u_y dy\right)^2 \leq ||u_y||^2(t) \rightarrow 0, \\ (T-1)^2 = \left(\int_0^y T_y dy\right)^2 \leq \int_0^1 T_y^2 dy = ||T_y||^2(t) \rightarrow 0, \\ Z^2 = \left(\int_0^y Z_y dy\right)^2 \leq \int_0^1 Z_y^2 dy = ||Z_y||^2(t) \rightarrow 0 \end{array} \right. \qquad t \rightarrow \infty, \tag{4.5}$$

by using the boundary conditions (II) and (3.12). This completes the proof of Theorem 5.

The boundary conditions (I) correspond to the impermeably insulated boundaries. Theorem 5(A) shows that the asymptotic state of the solutions of the problem is determined by the initial data and the Arrhenius functions. In particular,

(i) If the initial data satisfy

$$\int_0^1 \left(T_0(y) + qZ_0(y) + \frac{u_0(y)}{2}\right) dy > T_i + q, \tag{4.6}$$

then $Z_\infty = 0$, that is,

$$||Z||_{H^1(0,1)}(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty;$$

(ii) If the initial data satisfy

$$\int_0^1 \left(T_0(y) + qZ_0(y) + \frac{u_0(y)}{2}\right) dy < T_i, \tag{4.7}$$

then

$$T_\infty \leq T_i.$$

Thus, in the impermeably insulated container, five factors determine whether the reacting process completes: the ignition temperature, the amount of heat released by the given chemical reaction, the initial temperature, the velocity, and the mass fraction of the reacting mixture after a sufficiently long time. In particular, if the initial temperature is appropriately high, and/or the initial velocity appropriately fast, and/or the initial mass fraction of the reacting mixture appropriately large such that (4.6) holds, then the reacting process will certainly complete, and all unburnt gas will be burnt out after a sufficiently long time. If the initial temperature is appropriately low, and the initial velocity appropriately slow, and the initial mass fraction of the reacting mixture appropriately small such that (4.7) holds, then the reacting process will certainly stop after a sufficiently long time.

In contrast, the boundary conditions (II) correspond to the thermal source boundaries. Theorem 5(B) shows that the asymptotic state of the temperature T and the mass fraction Z of the problem are completely determined by the boundary data instead of the initial data, the scale of the reacting rate functions, and the amount of heat released. This reacting process will certainly complete, and all unburnt gas will be burnt out after a sufficiently long time.

These arguments may be generalized to more general boundary conditions.

5. Existence of generalized solutions of the Cauchy problem. We now establish an existence theorem of global generalized solutions of the Cauchy problem.

THEOREM 6. *Let the initial data satisfy the condition* (2.11), *and let*

$$(v_0(y)-1, u_0(y), T_0(y)-1, Z_0(y)) \in H^1(-\infty,\infty).$$

Then there exists a generalized solution (v,u,T,Z) *of the problem* (3.1)–(3.4) *and* (2.10) *satisfying*

$$v(t,y), T(t,y) > 0, \qquad 0 \le Z(t,y) \le 1.$$

The proof of Theorem 6 is based on the same arguments as in §3 and the localization lemma of Kazhikhov [6].

LOCALIZATION LEMMA. *For each interval* $I_n = [n, n+1)$, *there is a point* $y_n(t) \in I_n$ *such that*

$$\begin{cases} \alpha_0 \le v(t,y_n(t)) \le \beta_0, \\ \alpha_0 \le T(t,y_n(t)) \le \beta_0, \\ \alpha_0 \le \int_{I_n} v(t,y)dy, \int_{I_n} T(t,y)dy \le \beta_0, \end{cases}$$

where the positive constants α_0 *and* β_0 *are two roots of the equation*

$$y-1-\ln y = \frac{1}{\min\{a,1\}}\left\{\int_0^\infty \left(\frac{u_0^2}{2} + a(v_0-1-\ln v_0) + (T_0-1-\ln T_0)\right)dy + q\int_{-\infty}^\infty Z_0(y)dy\right\}.$$

We omit the details of the proof of this theorem.

REFERENCES

[1] F. A. WILLIAMS, *Combustion Theory*, Addison–Wesley, Reading, MA, 1965.

[2] A. MAJDA, *High Mach number combustion*, Lecture Notes in Appl. Math., 24 (1986), pp. 109–184.

[3] J. GLIMM, *The continuous structure of discontinuities*, Lecture Notes in Phys., 344 (1989), pp. 177–186.

[4] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[5] R. COURANT AND K. O. FRIEDRICHS, *Supersonic Flow and Shock Waves*, Interscience, New York, 1948.

[6] A. V. KAZHIKHOV, *On the theory of initial-boundary-value problems for the equations of one-dimensional nonstationary motion of a viscous heat-conductive gas*, in Russian, Dinamika Sploshn. Sredy, 50 (1981), pp. 37–62.

[7] R. A. GARDNER, *On the detonation of a combustible gas*, Trans. Amer. Math. Soc., 277 (1983), pp. 431–468.

[8] D. H. WAGNER, *The existence and behavior of viscous structure for plane detonation waves*, SIAM J. Math. Anal., 20 (1989), pp. 1035–1054.

[9] P. COLLELA, A. MAJDA, AND V. ROYTBURD, *Theoretical and numerical structure for reacting shock waves*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1059–1080.

[10] G. S. S. LUDFORD, *Low Mach number combustion*, Lecture Notes in Appl. Math., 24 (1986), pp. 3–74.

[11] YA. KANEL', *On a model system of equations of one-dimensional gas motion*, Differential Equations, 4 (1968), pp. 374–380.

[12] N. ITAYA, *On the Cauchy problem for the system of fundamental equations describing the movement of compressible fluid*, Kodai Math. Sem. Rep., 23 (1971), pp. 60–120.

[13] A. V. KAZHIKHOV AND V. V. SHELUKHIN, *Unique global solution with respect to time of initial-boundary-value problems for one-dimensional equations of a viscous gas*, J. Appl. Math. Mech., 41 (1977), pp. 273–282.

[14] M. OKADA AND S. KAWASHIMA, *On the equations of one-dimensional motion of compressible viscous fluids*, J. Math. Kyoto Univ., 23 (1983), pp. 55–71.

[15] S. KAWASHIMA AND T. NISHIDA, *The initial-value problems for the equations of viscous compressible and perfect compressible gas*, Nonlinear Funct. Anal., (1981), pp. 34–59.

[16] A. MATSUMURA AND T. NISHIDA, *The initial-value problem for the equations of motion of viscous and heat-conductive gas*, J. Math. Kyoto Univ., 20 (1980), pp. 67–104.

[17] ———, *Initial boundary value problems for the equations of motion of compressible viscous and heat-conductive fluids*, Comm. Math. Phys., 89 (1983), pp. 445–464.

[18] D. HOFF, *Global well-posedness of the Cauchy problem for the Navier–Stokes equations of nonisentropic flow with discontinuous initial data*, J. Differential Equations, 95 (1992), pp. 33–74.

[19] D. HOFF, *Discontinuous solutions of the Navier–Stokes equations for compressible flow*, Arch. Rational Mech. Anal., 114 (1991), pp. 15–46.

[20] D. H. WAGNER, *Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions*, J. Differential Equations, 68 (1987), pp. 118–136.

[21] S. KRUZKOV, *A priori estimates for the derivative of a solution to a parabolic equation and some of its applications*, in Russian, Dokl. Akad. Nauk. SSSR, 170 (1966), pp. 501–504.

[22] G.-Q. CHEN AND Y.-G. LU, *The existence and asymptotic behavior of solutions to systems of gas dynamics with a class of sources*, Acta Math. Sci., 8 (1988), pp. 85–94.

[23] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer–Verlag, New York, 1983.

[24] C. M. DAFERMOS, *Global smooth solutions to the initial-boundary value problem for the equations of one-dimensional nonlinear thermoviscoelasticity*, SIAM J. Math. Anal., 13 (1982), pp. 397–408.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 635–648, May 1992

SOLUTION OF THE CAUCHY PROBLEM FOR A CONSERVATION LAW WITH A DISCONTINUOUS FLUX FUNCTION*

TORE GIMSE AND NILS HENRIK RISEBRO

Abstract. The Cauchy problem is solved for a conservation law arising in oil reservoir simulation where the flux function may depend discontinuously on the space variable. To do this front tracking is used as a method of analysis.

Key words. conservation laws, discontinuous coefficients, two-phase flow

AMS(MOS) subject classifications. 35A05, 35L65, 35R05, 76T05

Introduction. In this paper we study the Cauchy problem for two phase flow through a one-dimensional porous medium. Darcy's law together with the equations of mass balance gives

$$s_t + \Big\{f_0(s)\big(v - g(x)k(s)\big)\Big\}_x = 0, \tag{0.1}$$

where $s = s(x,t)$ denotes the saturation of one of the phases, f_0 is the fractional flow function, v is the total Darcy velocity, and $k(s)$ is the relative permeability of the phase not denoted by s. The gravitational term $g(x)$ includes the density differences between the phases as well as the absolute permeability of the rock and the angle of dip of the reservoir. This term is, therefore, not necessarily a continuous function of x. Equation (0.1) is an example of a conservation law

$$\begin{aligned} u_t + f(u,x)_x &= 0, \\ u(x,0) &= u_0(x), \end{aligned} \tag{0.2}$$

where u may be either a vector or a scalar variable. Such conservation laws do not in general possess continuous solutions, and by a solution of (0.2) we mean a solution in the distributional sense, such that for each $\phi \in C_0^1$

$$\int_{-\infty}^{\infty}\int_0^{\infty} \big(u\phi_t + f(u,x)\phi_x\big)\,dt\,dx + \int_{-\infty}^{\infty} \phi(x,0)u_0\,dx = 0. \tag{0.3}$$

The solution u is then called a *weak solution* of (0.2). Krushkov proved the existence of a weak solution to (0.2) for a scalar u under the assumption that $\partial f/\partial x$ was bounded [10]. This assumption does not hold for (0.1) since the geology often varies discontinuously in a porous medium. The analysis of Isaacson and Temple [9] shows that equations of the type (0.2) under reasonable restrictions on f, in general, possess a unique weak solution of the Riemann problem provided the initial states are close.

Here we are interested in the initial value problem for (0.1), and we prove the following theorem.

THEOREM. *If $g(x)$ has bounded total variation,* (0.1) *possesses a weak solution $s(x,t)$ for arbitrary initial data $s_0(x)$ of bounded total variation.*

This theorem is proved through construction of a sequence of approximate solutions. These solutions are constructed by the method of front tracking introduced by

*Received by the editors October 22, 1990; accepted for publication (in revised form) June 14, 1991.

†Department of Mathematics, University of Oslo, Oslo, Norway.

Dafermos [1] and developed by Holden et al. [4]. This front tracking method is based on the solution of the Riemann problem for (0.1) which was studied by the authors in [3], and here we give a brief review of its solution. The solution of the Riemann problem for (0.1) is similar to the solution of the Riemann problem for the oil-polymer system studied by Isaacson [6]. This similarity is sufficient for us to use some of the ideas developed by Temple [12] for the oil-polymer system, most notably, the construction of a mapping Ψ from (s, g) to (z, g) such that the total variation of the approximate solutions remains bounded in (z, g). We will define a functional $F = F(u_\delta)$ where u_δ is our approximate solution generated by the front tracking scheme. Then F is shown to be nonincreasing in time, and this enables us to show that the sequence of approximations is well defined in the sense that each approximate solution can be defined at any time. Furthermore, we show that each approximate solution is constant on a finite number of polygons in $x - t$ space. Via a standard compactness argument, we can now show that a subsequence of the approximate solutions converges. The approximate solutions are constructed in such a way that they are weak solutions of equations which are close to (0.1). This makes it straightforward to prove that the limit is a weak solution.

In §1 of this paper we give some of the "physics" of the problem which leads to (0.1). In §2 we review the solution of the Riemann problem. In §3 we present the front tracking scheme and introduce the mapping Ψ and the functional F. We then show that F is nonincreasing and that this implies that the approximations are well defined. In §4 we prove that a subsequence of the approximate solutions converges towards a weak solution. Finally we make a remark on the applicability of this method of analysis to the oil-polymer system.

1. Physical motivation. We want to study two-phase flow in porous media, assuming for each phase Darcy's law:

$$v = -\lambda(\nabla P - \rho G),$$

where v is the Darcy speed, λ is the mobility, P the phase pressure, ρ is the density, and G a gravitational term. Combining Darcy's law with the source-free equation of mass conservation for each phase

$$\rho_t + \nabla(v\rho) = 0,$$

we find (for a more detailed treatment of these equations see [11]):

$$\alpha(\phi\rho_w s_w)_t + (\alpha\rho_w F_w)_x = 0 \tag{1.1}$$

which is the one-dimensional saturation equation, ignoring capillary effects (diffusion). Here α is the one-dimensional cross-sectional area, ϕ is the rock porosity, ρ_w is the density of water, and s_w is the saturation of the water at position x at time t. Lower indices x and t indicate derivatives with respect to space and time, respectively. F_w is the flow function of water:

$$F_w = f_w\big(v - K\lambda_o(\rho_w - \rho_o)g\big), \tag{1.2}$$

where f_w is the fractional flow function of water, $f_w = \lambda_w/(\lambda_w + \lambda_o)$, λ_w and λ_o being the phase mobilities of water and oil, v is the total Darcy velocity, K is the absolute rock permeability, and g is the component of gravity along the reservoir. Even if α, ϕ, and ρ are constants, so that (1.1) simplifies to read:

$$(s_w)_t + \frac{1}{\phi}(F_w)_x = 0, \tag{1.3}$$

F_w may be a function of *position* as well as saturation, $F_w = F_w(s_w, x)$. Heterogeneities like a varying reservoir angle (and thereby changing g), or changes in the rock permeability, K, along the reservoir, may both affect the flow function. This positional dependence of F_w may be smooth, when the parameters vary continuously along the reservoir, or discontinuous. The latter is probably very important and perhaps more common, since the rock is usually layered to some extent throughout the reservoir. Between such layers, introducing abrupt changes in rock permeability, $K = K(x)$ should be modeled discontinuously.

In general, the phase mobility curves $\lambda = \lambda(s_w)$ are assumed to be convex functions, typically shaped as indicated in Fig. 1.1. This gives an s-shaped, increasing fractional flow function $f_w = f_w(s_w)$, as shown in Fig. 1.2. In general, with increasing gravity or permeability, f decreases, so that two different flow functions typically look like Fig. 1.3. We will be interested in the Cauchy problem for (1.3).

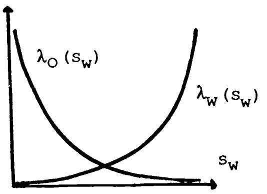


FIG. 1.1

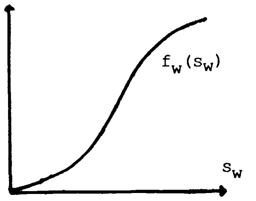


FIG. 1.2

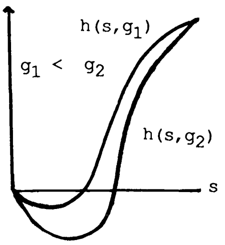


FIG. 1.3

2. The Riemann problem. We let s denote the saturation variable, introduce a variable $g = g(x)$ representing the geology, and let $u = (s, g)$, so that (1.3) may be written as a so-called triangular system ([2], [5]):

$$\begin{aligned} u_t + f(u)_x &= 0, \\ u(x,0) &= u_0(x). \end{aligned} \tag{2.1}$$

Here $f(u) = \big(h(s,g), 0\big)$ with

$$h(s,g) = f_0(s)\big(1 - gk(s)\big),$$

$f_0 = f_0(s)$ being a Lipschitz continuous, increasing function with one point of inflection (s-shaped) with $f_0(0) = 0$ and $f_0(1) = 1$ as in Fig. 1.2. The relative permeability $k(s)$ is usually assumed to be a decreasing, convex function of the saturation such that $k(0) = 1$ and $k(1) = 0$, cf. Fig. 1.1. Note that this implies that $h(1,g) = 1$ for all g. Also, each $h(\cdot, g)$ is Lipschitz continuous and has (possibly) one minimum and two points of inflection within the interval of definition, and finally $\partial h/\partial g \leq 0$, cf. Fig. 1.3. The Riemann problem for (1.3) or (2.1), which is the initial-value problem with initial constant states, denoted by $u_L = (s_l, g_l)$ and $u_R = (s_r, g_r)$, separated by a single geological discontinuity, has been studied by the authors in [3], where existence and uniqueness results are proved. Here we will need to know the solution of Riemann problems, so before proceeding with a more general treatment of (2.1), we will briefly summarize the main results of [3].

The one-dimensional Riemann problem for (2.1) may be written in the form

$$s_t + f_l(s)_x = 0 \quad \text{for } x < 0, \qquad s_t + f_r(s)_x = 0 \quad \text{for } x > 0, \tag{2.2}$$

with initial data

$$s(x,0) = \begin{cases} s_l & \text{if } x < 0, \\ s_r & \text{if } x > 0. \end{cases}$$

In the above notation, $f_l(s) = f_0(s)(1 - g_l k(s))$ and $f_r(s) = f_0(s)(1 - g_r k(s))$. Thus, to the left of the origin the flow function is f_l and to the right f_r. The saturation variable s is in the range $0 \leq s \leq 1$. We define s_- and s_+ to be the limits of the solution $s(x,t)$ as x approaches zero from below and above, respectively. Note that if the solution to (2.2) is unique, these values are independent of t. The Hugoniot relation at $x = 0$ gives

$$f_l(s_-) = f_r(s_+). \tag{2.3}$$

The procedure for determining possible values for s_- and s_+ is explained in full detail in [3], where it is proved that two such points always exist, and by introducing an additional entropy condition for the shock at $x = 0$, s_- and s_+ are uniquely determined. This entropy condition says that the jump $|s_- - s_+|$ at $x = 0$ should be the smallest possible jump here satisfying (2.3). This minimal jump condition is proved to be equivalent to the viscous profile entropy condition for an enlarged system of equations, in some extent equivalent to (2.2). The reader is referred to [3] for further details. We will now turn our attention to the two different waves involved in the solution of (2.1). First, s-waves are defined to be waves of constant g. Thus, in (s, g) phase space, these are found along horizontal lines. The other kind of waves are g-waves, which according to (2.3) have constant flow value, $f = const.$ Hence, it is useful to draw the level curves of f in the (s, g) diagram. This is done in Fig. 2.1. The bold curve labeled T in the figure is the transition curve, where $f_s = 0$. Solving the Riemann problem now consists of finding a sequence of s- and g-waves that go from u_L to u_R. The solution to the Riemann problem may now be found in the (s, g) phase space by the procedure indicated in Figs. 2.2 and 2.3 (use Fig. 2.2 for u_L to the left of T, and Fig. 2.3 for u_L to the right of T). Follow arrows that continuously connect u_L to u_R. Then find the solution by graphing the corresponding waves in the (x, t)-plane in the direction of the arrows. Note that any Riemann problem gives a solution consisting of at most three waves, an s-wave, a g-wave, and another s-wave. We write this composite solution wave as $[u_L u_R] = sgs'$. We close this section by displaying an example of how a Riemann problem like (2.2) is solved by the method indicated above. Given a Riemann problem as indicated in Fig. 2.4, in the (s, f) plane the solution looks like Fig. 2.5: Starting with a shock moving backwards along f_l from s_l (s-wave), crossing from s_- over to s_+ at f_r (minimal jump, g-wave), and finally continuing from s_+ to s_r with a rarefaction along f_r (another s-wave). In Fig. 2.6 we have indicated this solution in the (s, g) phase space, and finally in Fig. 2.7 the solution in the (x, t) plane is shown.

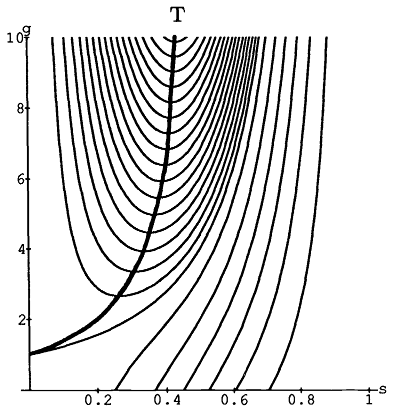


FIG. 2.1

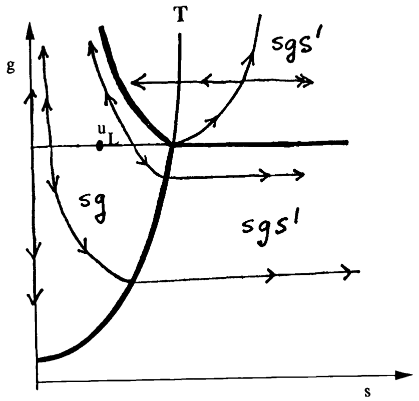


FIG. 2.2

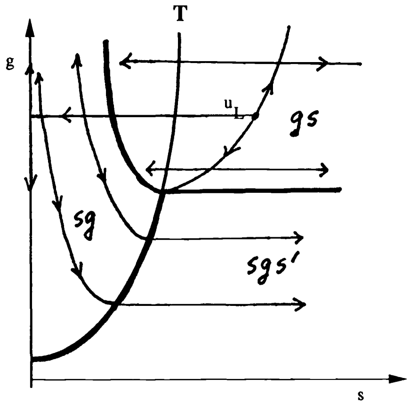


FIG. 2.3

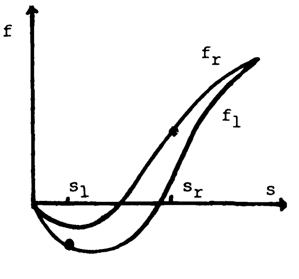


FIG. 2.4

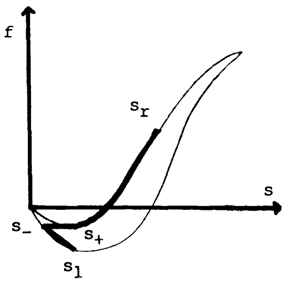


FIG. 2.5

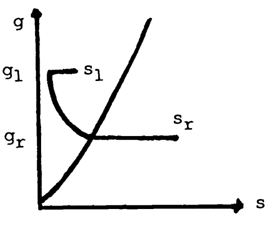


FIG. 2.6

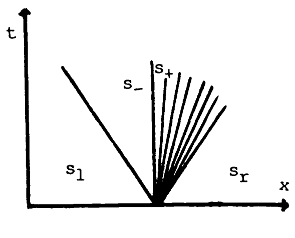


FIG. 2.7

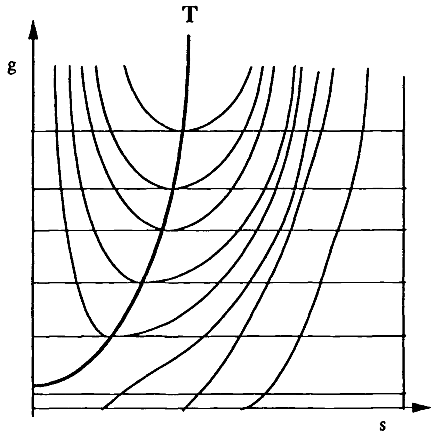


FIG. 3.1.

3. The front tracking scheme.

In this section we present the scheme we will use to generate a sequence of approximate solutions to (2.1). This scheme is a generalization of Dafermos' [1] scheme for the scalar conservation law. The basic idea of this scheme is to generate a sequence of exact solutions to approximate equations obtained by taking a piecewise linear approximation of the flux function. Via a standard compactness argument, we then show that this sequence possesses a convergent subsequence and furthermore that this converges towards a weak solution of (2.1).

In order to define our approximation we first have to define the approximate flux functions. Roughly speaking, these will be defined for a fixed g to be piecewise linear continuous in s. Assume that $g_0(x)$ is a function taking values in the interval $[0,G]$. We make a partition of this interval by choosing equally spaced points $\{\tilde{g}_i\}_1^{\tilde{n}}$ such that $\frac{1}{\tilde{n}} = \delta$ for some fixed δ. Let

$$s_T(g) = \min_s f(s,g) \tag{3.1}$$

and define $\tilde{f}_i = f\left(s_T(\tilde{g}_i), \tilde{g}_i\right)$. We choose $\{\hat{f}_i\}_1^{\hat{m}}$ to be an equally spaced partition of the interval $[f(s_T(G),G),1]$ such that $\hat{m} = \lceil\frac{1}{\delta}\rceil+1$, where $\lceil x\rceil$ = the largest integer smaller than or equal to x. Let

$$g_T(c) = \min_{f(s,g)=c} g \tag{3.2}$$

and define $\hat{g}_i = g_T(\hat{f}_i)$. Now we define

$$\{g_j\}_1^N = \{\tilde{g}_j\} \cup \{\hat{g}_j\} \qquad \{f_j\}_1^M = \{\tilde{f}_j\} \cup \{\hat{f}_j\}. \tag{3.3}$$

This defines a "mesh" in the (s,g) plane (cf. Fig. 3.1), and we see that the solution to a Riemann problem defined by two points in the mesh will have intermediate states that are in the mesh. We wish to simplify the solution to the Riemann problem further still by requiring that the s-waves consist only of states that are part of the mesh. This we do by approximating $f(s,g_i)$ by a piecewise linear function for each g_i. More precisely let

$$s_0(g) = 0 \quad \text{and} \quad s_{i+1}(g) = \min\left\{s > s_i(g) \mid f(s,g) \in \{f_i\}_1^M\right\} \tag{3.4}$$

for $i = 1, \cdots, n(g)$; note that $s_{n(g)}(g) = 1$. Let $s_{i,j}$ denote $s_i(g_j)$ and $f_{i,j}$ denote $f(s_{i,j}, g_j)$. We have that the intersections in the mesh have coordinates $(s_{i,j}, g_j)$, cf. Fig. 3.1. Finally we can define the approximate flux functions

$$(3.5)\qquad \begin{aligned} h_\delta(s,g_j) &= f_{i,j} + s\frac{f_{i+1,j}-f_{i,j}}{s_{i+1,j}-s_{i,j}},\\ f_\delta(s,g_j) &= \big(h_\delta(s,g_j),0\big)\end{aligned}$$

for $s \in [s_{i,j}, s_{i+1,j}]$ and for $j = 1,\cdots,N$.

The solution of the Riemann problem defined by

$$(3.6)\qquad \begin{aligned} &s_t{+}h_\delta(s,g_j)_x = 0,\\ &s_0(x) = \begin{cases} s_L & \text{if } x<0,\\ s_R & \text{if } x\ge 0\end{cases}\end{aligned}$$

consists of a number of constant states separated by discontinuities moving apart. Furthermore these constant states are a subset of $\{s_{i,j}\}_{i=1}^{n(g_j)}$. For a complete discussion of the Riemann problem for piecewise linear flux functions, see, e.g., [4].

In the following we let u denote the pair (s,g) and $u_{i,j}$ denote $(s_{i,j},g_j)$. Assume $u_0(x)$ to be a function taking values in the rectangle $[0,1]\times[0,G]$. We can construct an approximation to u_0, which we call $u_{0,\delta}(x)$, such that for each x; $u_{0,\delta}(x)\in\{u_{i,j}\}$ and

$$(3.7)\qquad \lim_{\delta\to 0}||u_0-u_{0,\delta}||_{L_1} = 0.$$

Condition (3.7) can be achieved since $g_j - g_{j-1}\le\delta$ and $s_{i,j}-s_{i-1,j} = O(\sqrt{\delta})$, where the right-hand side of the last equation depends on $\partial^2 f/\partial s^2$ on the T-curve.

We will now generate a weak solution $u_\delta(x,t)$ to the initial value problem

$$(3.8)\qquad u_{\delta t} + f_\delta(u_\delta)_x = 0,\qquad u_\delta(x,0) = u_{0,\delta}(x).$$

The initial function $u_{0,\delta}$ defines a series of finitely many Riemann problems, and by construction the solution to these problems are constant states (which are included in the set $\{u_{i,j}\}$) separated by discontinuities. We can track these discontinuities and thereby propagate the solution forward in time, until two of them collide. At this point we have a situation similar to what we had initially, namely a sequence of Riemann problems. Therefore we can solve these and propagate the solution until the next collision. Note that by construction u_δ is a weak solution of (3.8). We call this process front tracking, and it is clear that it can be repeated an arbitrary number of times. We do, however, need to justify that we can propagate the solution in this manner up to any given time by a finite number of operations. But in order to do this we first define a certain functional F which is nonincreasing for each collision of fronts.

We may think of a wave of u_δ either as a discontinuity in the (x,t) plane or as a directed path in (s,g) space. If the wave is an s-wave, this path is just the straight line from the state to the left of the discontinuity to the state to the right of the discontinuity. If the wave is a g-wave the path is the curve $f =$ const. from the left state to the right state. Thus u_δ can be thought of as a finite sequence of connected waves in the (s,g) plane, representing the discontinuities in u_δ as we move from left to right in the (x,t) plane. We will call any finite sequence of connected s- or g-waves in the (s,g) plane an I curve, where by connected we mean that the left state of a wave in the sequence is the right state of its predecessor, and we say that an I curve connects u_L to u_R if the left state of the first wave is u_L and the right state of the final wave is u_R. We will use the techniques developed for the oil-polymer system by Temple [12] and construct a certain $1-1$ mapping Ψ from (s,g) to (z,g), and a functional $F(I)$ such that $F(u_\delta)$ dominates the total variation of $\Psi\circ u_\delta$. We then prove that $F(u_\delta)$ is nonincreasing for each collision, and that this implies, first, that the approximation procedure can be continued to any time by finitely many operations, and, second, that a subsequence of the approximate solutions converges in L_1.

The mapping Ψ is similar to the mapping used by Temple in [12], and it involves the intersections of the T-curve with the level sets of f. Since f does not take all values on T, we must extend both T and the level sets outside $[0,1]\times[0,G]$. Assume for the moment that this is done in such a manner that for each point (s,g) in $[0,1]\times[0,G]$ we can find a unique point (s',g') on T such that $f(s',g')=f(s,g)$. Then z is defined as follows:

$$\begin{gathered}|z|=|g-g'|\\ \operatorname{sign} z=\begin{cases}-1 & \text{if } (s,g) \text{ is to the left of } T \text{ or above } T,\\ 1 & \text{if } (s,g) \text{ is to the right of } T \text{ or below } T.\end{cases}\end{gathered}\tag{3.9}$$

We have two cases of how to define the point (s',g') when f does not take the value $f(s,g)$ on T, depending on whether T intersects the s-axis or the g-axis. Assume first that T intersects the s-axis. Then we can extend T and the level curves of f in a smooth manner such that they intersect T at their minimum, cf. Fig. 3.2. If T intersects the g axis we make a smooth decreasing extension $g'(s)$ of T defined for negative s. If f does not take the value $f(s,g)$ on T, then f will take this value on the line $g=0$ at some point $\tilde{s}$. We then define $(s',g')=\bigl(-\tilde{s},g'(-\tilde{s})\bigr)$. Since the line $s=0$ is a level set for f, this mapping will be continuous and smooth, cf. Fig. 3.3. As in [12] we have that Ψ is $1-1$ and regular everywhere except on T. In the following we let $w=\Psi(u)$.

Now we can define the functional F. We define the strength of an s-wave to be

$$|s|=|\Delta z|\tag{3.10}$$

and the strength of a g-wave

$$|g|=\begin{cases}2|\Delta g| & \begin{cases}\text{if } g \text{ is to the right of } T \text{ and } g_L<g_R,\\ \text{or } g \text{ is to the left of } T \text{ and } g_L>g_R,\end{cases}\\ 4|\Delta g| & \begin{cases}\text{if } g \text{ is to the right of } T \text{ and } g_L>g_R,\\ \text{or } g \text{ is to the left of } T \text{ and } g_L<g_R.\end{cases}\end{cases}\tag{3.11}$$

We can write u_δ as $b_1,\cdots,b_n=I$, where b_i is either an s-wave or a g-wave, and we define

$$F(I)=\sum_i |b_i|.\tag{3.12}$$

Here follows the main lemma regarding F.

LEMMA 3.1. *Let J be any I curve connecting u_L to u_R, and let $[u_Lu_R]$ be the I curve that solves the Riemann problem defined by u_L and u_R. Then $F\bigl([u_Lu_R]\bigr)\le F(J)$.*

The proof of this lemma is analogous to the proof of the corresponding lemma (Lemma 5.1) in [12], and since it involves the study of a number of cases, it is presented in an appendix.

Now let F_0 denote $F\bigl(u_\delta(\cdot,t)\bigr)$, where t is taken to be so small that no collision has yet occured. The main theorem of this section then follows immediately from Lemma 3.1.

THEOREM 3.1. *Assume that F_0 is finite, and let $t_1\le t_2$, then*

$$F\bigl(u_\delta(\cdot,t_1)\bigr)\ge F\bigl(u_\delta(\cdot,t_2)\bigr).\tag{3.13}$$

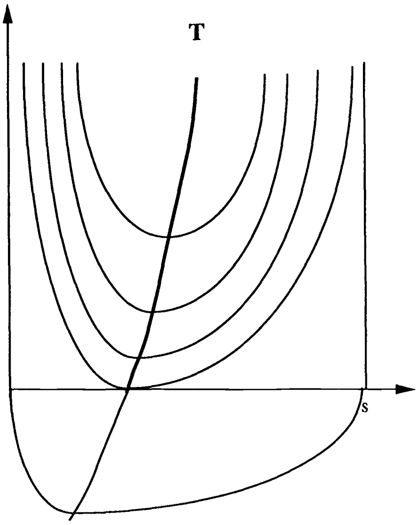


FIG. 3.2

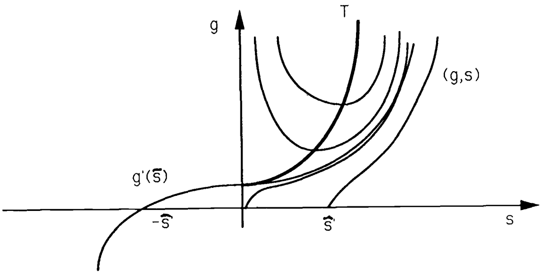


FIG. 3.3

Proof. It is clear that F only changes value when we have a collision of discontinuities in u_δ. At any one time we can have only finitely many collisions, and the change in F is a sum of the changes in F at each collision point. Consider, therefore, two discontinuities that collide, the one on the left separating states u_L and u_M, the one on the right separating states u_M and u_R. The theorem now follows, since by Lemma 3.1 $F\big([u_L u_M][u_M u_R]\big) \geq F\big([u_L u_R]\big)$. □

It is clear that F at each collision of u_δ either remains constant or changes by at least Δ, where Δ is the minimum distance between the states of which u_δ may consist, i.e.,

$$\Delta = \min_{(i,j)\neq(k,l)} |w_{i,j} - w_{k,l}|. \tag{3.14}$$

We wish to investigate those collisions which are possible if F remains constant for all time. Let s^+ (s^-) denote those s-waves over which s is increasing (decreasing), and let s_R (s_L) denote an s-wave with left and right state to the right (left) of T.

LEMMA 3.2. *Assume that $F(u_\delta)$ is constant and that u_δ contains the wave sequence gs_R^+ ($s_L^- g$). Then no s-wave will collide from the right (left) with g.*

Proof. Let the s_R^+ wave separate states $s_l < s_r$. Since F is constant it can only collide with s-waves that separate states $s_r < s'$. The result of this collision is a single s_R^+ wave separating states $s_l < s'$. If the s-wave collides with a g-wave the result of the collision must be gs_R^+ since F is constant. Since all s_R waves have positive speeds, the lemma follows. An analogous argument takes care of the case $s_L^- g$. □

LEMMA 3.3. *Assume that $F(u_\delta)$ is constant, then only three types of collisions can occur*:

(1) *An s-wave separating states $s_l \lessgtr s_m$, colliding with another s-wave separating states $s_m \lessgtr s_r$, giving as a result a single s-wave separating $s_l \lessgtr s_r$.*

(2) *An s-wave colliding with a g-wave from the right (left), giving sg (gs) as result ("an s-wave passing through a g-wave").*

(3) *An s-wave colliding with a g-wave, giving $s_L^- g s_R^+$ as a result.*

The proof of this lemma consists of checking a number of cases of Riemann solutions in the Figs. 2.2 and 2.3. It is straightforward and is, therefore, omitted. Combining the last two lemmas, we see that if F is constant; our approximation u_δ is well defined. However, Theorem 3.1 implies that after some finite number of collisions, F will change by an amount less than Δ for all subsequent collisions, i.e., F will remain constant for all collisions thereafter. Therefore, the approximation u_δ is well defined and u_δ is constant on a finite number of regions in $\mathbb{R} \times \mathbb{R}^+$. These regions are separated by a finite number of straight lines.

4. Convergence. Let $\mathrm{Var}_{ab}\, u$ denote the total variation of u with respect to the variables a and b. By construction, $F(I) \geq \mathrm{Var}_{zg}\, J$ for any I-curve J. Hence, by Theorem 3.1, we may find a uniform bound on $\mathrm{Var}_{zg}\, u_\delta(\cdot, t)$, provided F_0 is bounded. We show this by applying Temple's argument [12]: Let $S(\epsilon)$ be a strip of width ϵ around the T-curve. If u_L or u_R are outside the strip , the Riemann problem solution $[u_L u_R]$ has a finite number of waves, each globally bounded, and the waves intersect transversally in the (z, g) plane. Thus, $\mathrm{Var}_{zg}[u_L u_R] = O(1)|w_R - w_L|$ if $w_R = (z_R, g_R)$ and $w_L = (z_L, g_L)$ are the images of u_R and u_L under Ψ. Secondly, if u_L and u_R are in $S(\epsilon)$, $\mathrm{Var}_{zg} < 5|w_L - w_R|$ by construction, for ϵ sufficiently small. Therefore, for any Riemann problem, $\mathrm{Var}_{zg}\, I = O(1)|w_L - w_R|$. Let $\{J_i\}$ denote the solutions of the initial Riemann problems. This gives

$$\begin{aligned} F_0 &\leq 4 \sum_i \mathrm{Var}_{zg}\, J_i \\ &\leq O(1) \sum |w_L - w_R| \\ &\leq O(1)\, \mathrm{Var}_{zg}\, u_{0,\delta}, \end{aligned} \tag{4.1}$$

proving that $\mathrm{Var}_{zg}\, u_\delta(\cdot, t)$ is bounded for each fixed t.

Having proved boundedness in space for each time t, we want to prove Lipschitz continuity in t.

LEMMA 4.1.

$$\int_{-\infty}^{\infty} |w_\delta(x, t_2) - w_\delta(x, t_1)|\, dx \leq O(1)|t_2 - t_1|\, \mathrm{Var}_{zg}\, u_{0,\delta}.$$

Proof. Let M be the maximum speed at which a wave may propagate. M is given by the maximum slope of any $f_\delta(\cdot, g_j)$. Thus, if $t_1 < t_2$, $|w_\delta(x, t_2) - w_\delta(x, t_1)|$ is bounded by the spatial variation of $w_\delta(y, t_1)$, where $x - M|t_2 - t_1| < y < x + M|t_2 - t_1|$. However, as pointed out above, $w_\delta(\cdot, t)$ is of bounded variation, so that we may write:

$$\int_{-\infty}^{\infty} |w_\delta(x, t_2) - w_\delta(x, t_1)|\, dx = O(1) \int_{-\infty}^{\infty} \int_{x - M|t_2 - t_1|}^{x + M|t_2 - t_1|} \left| \frac{dw_\delta}{dy} \right| dx\, dy.$$

Here, $|dw_\delta/dy|\, dx\, dy$ is a measure of mass $\mathrm{Var}_{zg}\, w_\delta(x, t)$, and by changing the order of integration we have:

$$\begin{aligned}\int_{-\infty}^{\infty} |w_\delta(x,t_2) - w_\delta(x,t_1)|\,dx &= O(1)M|t_2 - t_1]\operatorname{Var}_{zg} w_\delta(x,t)\\ &\leq O(1)M|t_2 - t_1]\operatorname{Var}_{zg} u_{0,\delta}(x),\end{aligned}$$

the last inequality holds since

$$O(1)\operatorname{Var}_{zg} u_\delta(\cdot,t) = F\big(u_\delta(\cdot,t)\big) \leq F_0 \leq O(1)\operatorname{Var}_{zg} u_{0,\delta}. \qquad \square$$

It remains to prove the convergence of the sequence $\{u_\delta\}$.

THEOREM 4.1. *Let $w_0 = \Psi u_0$ be any initial data such that $\operatorname{Var}_{zg} w_0 < \infty$. Then for any sequence $\{\delta\}$, such that $\delta \to 0$, there exists a subsequence δ_j and a function u, such that for any finite time T, $u_{\delta_j}(\cdot,t)$ converges uniformly to $u(\cdot,t)$ in $L^1_{\text{loc}}(x)$ for any $t \leq T$.*

Proof. We have demonstrated that $w_\delta(\cdot,t)$ has uniformly bounded variation for each t, and so it follows from Helly's theorem, that a subsequence converges in $L^1_{\text{loc}}(x)$. By a diagonalization argument, such convergence is achieved on a countable dense set of t-values, t_j, $0 \leq t_j \leq T$. Let w_{δ_j} be this subsequence. By a further diagonalization argument, we may find a subsequence of $\{w_{\delta_j}\}$; w_μ, which converges uniformly in $L_1[-M,M]$ at a fixed t_j. Thus, for this sequence:

$$\begin{aligned}\int_{-M}^{M} |w_{\mu_1}(x,t) - w_{\mu_2}(x,t)|\,dx \leq &\int_{-M}^{M} |w_{\mu_1}(x,t) - w_{\mu_1}(x,t_j)|\,dx\\ &+ \int_{-M}^{M} |w_{\mu_1}(x,t_j) - w_{\mu_2}(x,t_j)|\,dx\\ &+ \int_{-M}^{M} |w_{\mu_2}(x,t_j) - w_{\mu_2}(x,t)|\,dx.\end{aligned}$$

Here the first and third term approach zero by Lemma 4.1, and the second term is small by the boundedness of $w_\mu(\cdot,t)$. Therefore, w_μ converges uniformly in $L_1[-M,M]$. This argument may now be applied a countable number of times, concluding the existence of a sequence, which we, for convenience, also label $\{w_\delta\}$, such that for any $t > 0$, $w_\delta(\cdot,t) \to w(\cdot,t)$ uniformly. The uniform continuity of Ψ^{-1} gives the theorem for $u_\delta = \Psi^{-1}w_\delta$ and $u = \Psi^{-1}w$. $\square$

Finally, we want to show that the limit obtained above is indeed a weak solution to the problem (2.1). For $T < \infty$ we define:

$$W(u) = \int_{-\infty}^{\infty}\int_0^T \big(\phi_t u + \phi_x f(u)\big)\,dx\,dt + \int_{-\infty}^{\infty} \phi u_o\,dx$$

for $\phi = \phi(x,t)$ an arbitrary function in C^1_0. For $u = \lim_{\delta\to 0} u_\delta$ we want to show that $W(u) = 0$. Since u_δ is a weak solution of (3.8):

$$\int_{-\infty}^{\infty}\int_0^T \big(\phi_t u_\delta + \phi_x f_\delta(u_\delta)\big)\,dx\,dt + \int_{-\infty}^{\infty} \phi u_{0,\delta}\,dx = 0,$$

which gives:

$$W(u) = \int_{-\infty}^{\infty}\int_0^T \Big(\phi_t(u - u_\delta) + \phi_x\big(f(u) - f_\delta(u_\delta)\big)\Big)\,dx\,dt + \int_{-\infty}^{\infty} \phi(u_0 - u_{0,\delta})\,dx$$

for all δ. Thus,

$$|W(u)| \leq \|\phi_t\|_\infty \|u - u_\delta\|_1 + \|\phi_x\|_\infty \|f(u) - f_\delta(u_\delta)\|_1 + \|\phi\|_\infty \|u_0 - u_{0,\delta}\|_1$$

$$\leq \|\phi_t\|_\infty \|u - u_\delta\|_1 \tag{4.2a}$$

$$+ \|\phi_x\|_\infty \|f(u) - f_\delta(u)\|_1 \tag{4.2b}$$

$$+ \|\phi_x\|_\infty \|f_\delta(u) - f_\delta(u_\delta)\|_1 \tag{4.2c}$$

$$+ \|\phi\|_\infty \|u_0 - u_{0,\delta}\|_1. \tag{4.2d}$$

Here, by Lipschitz continuity of f and f_δ, the terms (4.2b) and (4.2c) above are small. Furthermore, (4.2a) is small by the construction of u as the L_1 limit of u_δ, and (4.2d) is small by the construction of $u_{0,\delta}$. Hence, for any given $\epsilon > 0$, we may choose δ so that $|W(u)| < \epsilon$, concluding that $W(u) = 0$, and the limit u is a weak solution of (2.1). □

Remark. The system of conservation laws modeling polymer flow in porous media

$$\begin{aligned} s_t + \big(g(s,b)s\big)_x &= 0, \\ b_t + \big(g(s,b)b\big)_x &= 0 \end{aligned} \tag{4.3}$$

studied by Temple [12], and Temple and Isaacson [7], [8], has a structure of the solution to the Riemann problem that is remarkably similar to the Riemann solution used in this paper (Compare Figs. 8 and 9 in [12] with Figs. 2.2–2.3.). It is this similarity that enabled us to use essentially the same techniques as [12] to show that the functional F was nonincreasing and to obtain the estimates on $\mathrm{Var}_{zg}\, w_\delta$. This in turn guaranteed that our approximation u_δ was well defined and that the sequence $\{u_\delta\}$ possessed a subsequence which converged towards a weak solution of (2.1). We could have defined an analogue of u_δ, (s_δ, b_δ) as an approximation to the solution of (4.3). Since the whole subsequent argument hinges on the fact that F is nonincreasing, it applies equally well to (s_δ, b_δ) as to u_δ. Therefore, the front tracking method presented here gives an alternative proof of the existence of a solution to the Cauchy problem for (4.3).

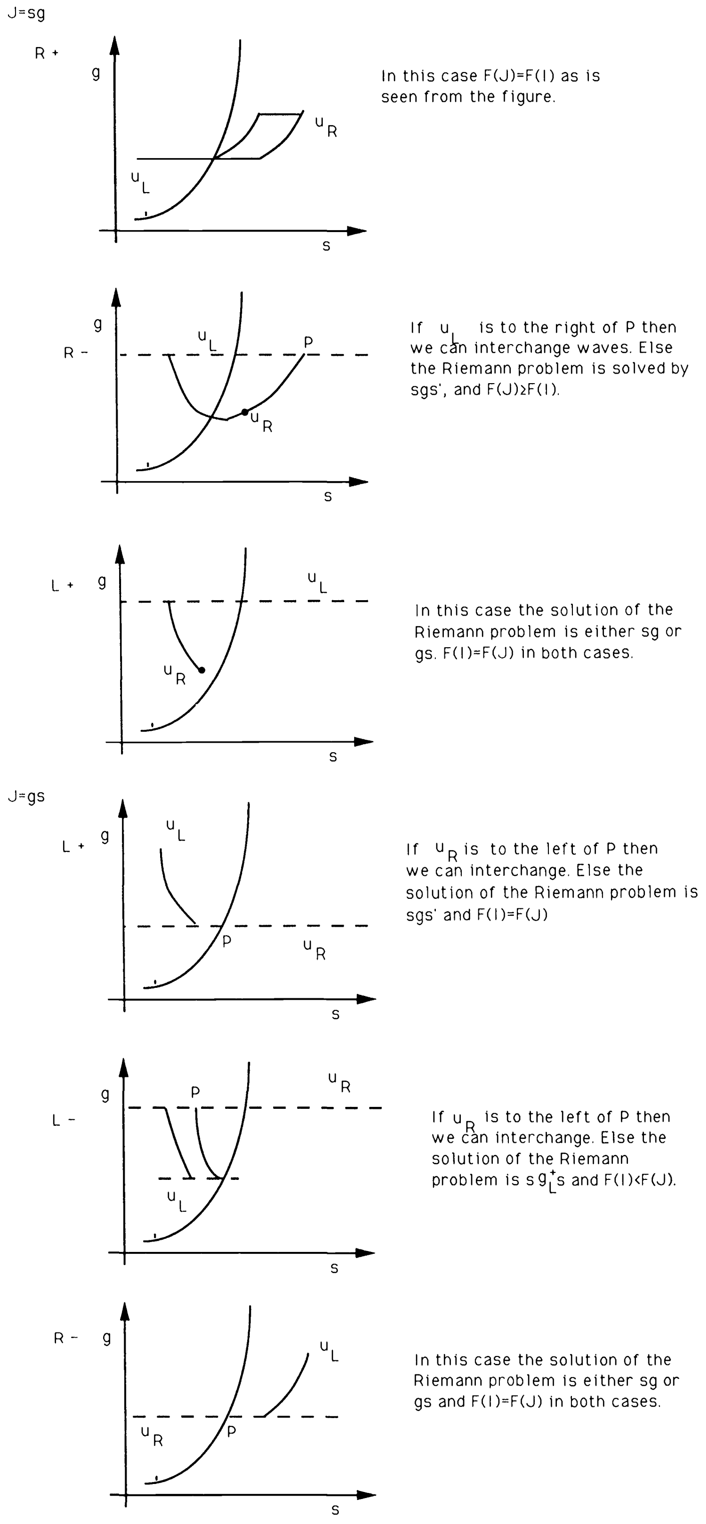
5. Appendix. Here we present the proof of Lemma 3.1.

LEMMA 3.1. *Let J be any I curve connecting u_L to u_R, and let $[u_L u_R]$ be the I curve that solves the Riemann problem defined by u_L and u_R. Then $F\big([u_L u_R]\big) \leq F(J)$.*

Since the structure of the solution of the Riemann problem is similar to the structure of the solution of the Riemann problem for the polymer system studied by Temple [12], Lemma 3.1 is proved by essentially the same arguments as the corresponding lemma in [12]. We first prove three lemmas; Lemma 3.1 will then follow from these.

Let g^+ (g^-) denote a g wave over which s is increasing (decreasing), and let g_R (g_L) denote a g-wave to the right (left) of T. We can now define the "addition" of waves; the addition of s_1 and s_2 is the s-wave that goes from the left state of s_1 to the right state of s_2. If g_1 and g_2 are both g_L or g_R waves then the addition of g_1 and g_2 is the combined g-wave. If two g-waves are of different type and $g(u_L) < g(u_R)$ $(g(u_L) > g(u_R))$ then their addition is the unique wave gs (sg) that goes from u_L to u_R.

Assume now that $J = sg$ $(J = gs)$ connects u_L to u_R. If u_L and u_R are on the same side of T and a "parallelogram" of s and g waves can be drawn with u_L and u_R as diagonally opposed corners, the interchange of J; $\bar{J}$ is defined to be the unique I curve $\bar{g}\bar{s}$ $(\bar{s}\bar{g})$ that connects u_L to u_R. If u_L and u_R are on different sides of T we can only define $\bar{J}$ if $J = sg_L^-$ or $J = g_R^+ s$. The interchange of sg_L^- is the unique I curve



In this case F(J)=F(I) as is seen from the figure.

If u_L is to the right of P then we can interchange waves. Else the Riemann problem is solved by sgs', and F(J)≥F(I).

In this case the solution of the Riemann problem is either sg or gs. F(I)=F(J) in both cases.

If u_R is to the left of P then we can interchange. Else the solution of the Riemann problem is sgs' and F(I)=F(J)

If u_R is to the left of P then we can interchange. Else the solution of the Riemann problem is sg_L^+s and F(I)<F(J).

In this case the solution of the Riemann problem is either sg or gs and F(I)=F(J) in both cases.

FIG. A1

$\bar{g}_R^+\bar{s}$ that connects the same endpoints. Other I-curves do not have an interchange. As in [12] it is easy to show that if J is the addition of b_1 and b_2 then $F(b_1b_2) \geq F(J)$, and if $\bar{J}$ is the interchange of J then $F(J) = F(\bar{J})$. Furthermore, if J connects u_L to u_R and J has an interchange $\bar{J}$, then $[u_Lu_R] = J$ or $[u_Lu_R] = \bar{J}$.

LEMMA A1. *If J connects u_L to u_R and $J = gs$ or $J = sg$ then $F(J) \geq F\big([u_Lu_R]\big)$.*

Proof. If J has an interchange then the lemma holds. Assuming that J does not have an interchange, we have eight cases to check: $J = sg$ or gs, $g = g_R$ or g_L, $g = g^+$ or g^-. But if $J = g_R^+s$ or sg_L^- then J has an interchange. This leaves six cases which are checked in Fig. A1. □

LEMMA A2. *If J connects u_L to u_R and $J = sgs'$ then $F(J) \geq F\big([u_Lu_R]\big)$.*

Proof. If gs or sg' can be interchanged, we can interchange and add waves so that Lemma A1 applies. Assume, therefore, that neither can be interchanged. This implies that s and s' both cross T and that $g = g_L^+$ or g_R^-. This leaves two cases to check as an exercise for the reader. □

LEMMA A3. *If J connects u_L to u_R and $J = gsg'$ then $F(J) \geq F\big([u_Lu_R]\big)$.*

Proof. We can assume that s crosses T and that $g \neq g_R^+$ and $g' \neq g_L^-$. Also if the variable g is increasing over g and decreasing over g' (or vice versa) it is easy to show that $F(J) \geq F(\tilde{s}\tilde{g})$ or $F(\tilde{g}\tilde{s})$, where $\tilde{s}\tilde{g}$ or $\tilde{g}\tilde{s}$ connects u_L to u_R. In this case, we can now use Lemma A1. Now s can cross from left to right or right to left, i.e., $gsg' = g_L^+sg_R^-$ or $g_R^-sg_L^+$. In both of these cases J contains a "strong" g-wave, whereas $[u_Lu_R] = sg^+s'$, i.e., the Riemann solution has a "weak" g-wave. The presence of this strong wave makes the lemma hold. □

Proof of Lemma 3.1. Once Lemmas A2 and A3 are established, the proof of this lemma carries over literally from the proof of Lemma 5.1 in [12] if "c-waves" are substituted with "g-waves." □

REFERENCES

[1] C.M. DAFERMOS, *Polygonal approximation of solution to the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33–41.

[2] T. GIMSE, *A numerical method for a system of equations modelling one-dimensional three-phase flow in a porous medium*, Notes Numer. Fluid Mech., 24 (1989), pp. 159–168.

[3] T. GIMSE AND N.H. RISEBRO, *Riemann problems with a discontinuous flux function*, Proc. Third Internat. Conf. on Hyperbolic Problems, Studentlitteratur, Lund, pp. 488–502.

[4] H. HOLDEN, L. HOLDEN, AND R. HØEGH-KROHN, *A numerical method for first order nonlinear scalar conservation laws in one dimension*, Comput. Math. Appl., 15 (1988), pp. 595–602.

[5] L. HOLDEN AND R. HØEGH-KROHN, *A class of N nonlinear hyperbolic conservation laws*, J. Differential Equations, 84 (1990), pp. 73–99.

[6] E. ISAACSON, *Global solution of a Riemann problem for a non-strictly hyperbolic system of conservation laws arising in enhanced oil recovery*, J. Comput. Phys., to appear.

[7] E. ISAACSON AND B. TEMPLE, *Analysis of a singular system of conservation laws*, J. Differential Equations, 65 (1986), pp. 250–268.

[8] ———, *Structure of asymptotic states in a singular system of conservation laws*, Adv. in Appl. Math., 11 (1990), pp. 205–219.

[9] ———, *Nonlinear resonance in inhomogeneous systems of conservation laws*, Contemp. Math., 108 (1990), pp. 63–77.

[10] N. KRUSHKOV, *Quasi-linear equations of the first order*, Mat. Sb., 2 (1970), pp. 217–243.

[11] D.W. PEACEMAN, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, Amsterdam, 1977.

[12] B. TEMPLE, *Global solution of the Cauchy problem for a class of* 2 × 2 *non-strictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 649-661, May 1992

LINEARIZED STABILITY AND IRREDUCIBILITY FOR A FUNCTIONAL DIFFERENTIAL EQUATION*

MARY E. PARROTT†

Abstract. A principle of linearized stability is given for the abstract functional differential equation $\dot{u}(t)=Bu(t)+Ku_t$, $t \geqq 0$, $u_0=f$, where B generates a strongly continuous semigroup of bounded linear operators on a Banach space X, and $K: E=C([-r_0, 0], X) \rightarrow X$ is a nonlinear, continuously Fréchet-differentiable operator. The strong positivity property of irreducibility is also investigated for the semigroup associated with solutions of the linearized equation. The theory is applied to the stability analysis of an equation from population dynamics.

Key words. linearized stability, functional differential equation, positive semigroup, irreducibility

AMS(MOS) subject classifications. 34K20, 34K30

1. Introduction. In a recent paper [11] a result of Desch and Schappacher [2] was used to develop a principle of linearized stability for the abstract functional differential equation

$$(\mathrm{FDE}_1) \qquad \begin{aligned} \dot{u}(t) &= Bu(t)+\phi u_t, \qquad t \geqq 0, \\ u_0 &= f, \end{aligned}$$

where B generates a strongly continuous semigroup of bounded linear operators on a Banach space X, ϕ is a nonlinear Lipschitz continuous operator from $E=C([-r_0, 0], X)$ to X, and the functional $u_t \in E$ is defined by $u_t(s)=u(t+s)$ for $s \in [-r_0, 0]$. As a consequence, recent results from positive semigroup theory can be used to study the stability of (FDE_1), as indicated in [11].

In the present work we are motivated by a widely used equation (see, for example, [7], [1], [5], [14]) which describes the growth of a spatially distributed population with delay in the birth process:

$$(\mathrm{E}) \qquad \begin{aligned} \frac{\partial u}{\partial t}(x, t) &= d \frac{\partial^2 u}{\partial x^2}(x, t)+a u(x, t) \\ &\quad \cdot\left[1-b u(x, t)-\int_{-1}^{0} u(x, t+r(s))\, d\eta(s)\right], \\ &\qquad x \in[0, \pi], \qquad t \geqq 0. \end{aligned}$$

In [7] positive semigroup theory is used to study the stability of the solution semigroup corresponding to the (abstract) linearization of (E) at a stationary solution. However, no conclusions regarding the stability of (E) could be drawn from this analysis. The connection between the stability of equilibria of the nonlinear functional differential equation (E) and the stability of the zero equilibrium of the corresponding linearized equation will be established in the present paper as a consequence of our abstract results. Because of the nature of the nonlinearity which appears in (E), this equation cannot be modeled by an abstract equation of the form (FDE_1). Our goal here is to

* Received by the editors April 24, 1989; accepted for publication (in revised form) August 14, 1991.

† Department of Mathematics, University of South Florida, Tampa, Florida 33620. This research was supported in part by National Science Foundation grant DMS-8706798.

develop a principle of linearized stability for abstract functional differential equations of the form

$$(\mathrm{FDE}_2)\qquad \begin{aligned} \dot{u}(t) &= Bu(t)+Ku_t, \qquad t\geqq 0,\\ u_0 &= f, \end{aligned}$$

where B generates a strongly continuous semigroup of bounded linear operators on a Banach space X and $K: E=C([-r_0,0],X)\to X$ is a nonlinear, continuously Fréchet-differentiable operator. (These assumptions are made more precise below.) This principle will be applied to (E) and, using the results of [7], we will be able to conclude an asymptotic stability result for an equilibrium of (E).

A second purpose of the present paper is to further examine the role of positive semigroup theory in the stability study of the linearizations of (FDE_1) and (FDE_2). More specifically, we will consider the consequences of the strong positivity property of irreducibility for the solution semigroup corresponding to the linear functional differential equation

$$\begin{aligned} \dot{u}(t) &= Bu(t)+\psi u_t, \qquad t\geqq 0,\\ u_0 &= f, \end{aligned}$$

where B generates a strongly continuous semigroup of bounded linear operators on X, and ψ is a bounded linear operator from E to X. The applicability of this theory to the study of the linearization of (E) is also given.

2. Stability result for (FDE_2). For $x\in X$, a Banach space, $\|x\|$ denotes the norm of x. For $f\in E=C([-r_0,0],X)$, where r_0 is a fixed positive constant, the norm of f, $\|f\|_E$, is defined by $\|f\|_E=\sup_{\theta\in[-r_0,0]}\|f(\theta)\|$. We use $|\cdot|$ to denote the norm of a bounded linear operator.

We consider the equation (FDE_2), where we assume the following hypotheses hold:

(H1) B generates a strongly continuous linear semigroup $T(t)$, $t\geqq 0$, on X.

(H2) $K: E=C([-r_0,0],X)\to X$ is a nonlinear, continuously Fréchet-differentiable operator at each $\hat{f}\in E$, that is,

$$K(f)=K(\hat{f})+K'(\hat{f})(f-\hat{f})+\circ(f-\hat{f})$$

for all $f\in E$, where $K'(\hat{f})$ is a bounded linear operator from E to X, $\circ$ is a continuous function from E to X, and b is a continuous increasing function from $[0,\infty)$ to $[0,\infty)$ such that $b(0)=0$ and $\|\circ(f)\|\leqq b(r)\|f\|_E$ for all $f\in E$ such that $\|f\|_E\leqq r$. In addition

$$|K'(f_1)-K'(f_2)|\leqq d(r)\|f_1-f_2\|_E$$

for all $f_1,f_2\in E$ such that $\|f_1\|_E$, $\|f_2\|_E\leqq r$, where d is a continuous increasing function from $[0,\infty)$ to $[0,\infty)$.

In the three lemmas that follow we show the existence of a unique local solution of (FDE_2). The method of proof is similar to that of Webb [16, Chap. 4], adapted for delay equations, and is therefore omitted. Basic existence and stability results for abstract semilinear functional differential equations (under various hypotheses) have also been obtained by other authors, including Fitzgibbon [3], Lightbourne [8], Rankin [12], Travis and Webb [15], Webb [17], and Martin and Smith [9].

LEMMA 2.1. *Assume that* (H1) *and* (H2) *are satisfied. Then for each $f\in E$ there exists a maximal interval of existence $[-r_0,T_f)$ and a unique continuous function*

$t \to u(t;f)$ from $[-r_0, T_f)$ to X such that

$$(1)\quad \begin{aligned} u(t;f) &= T(t)f(0) + \int_0^t T(t-s)Ku_s\,ds, \qquad t \in [0, T_f), \\ u(t;f) &= f(t), \qquad t \in [-r_0, 0], \end{aligned}$$

and either $T_f = \infty$ or $\lim_{t \to T_f^-} \sup \|u(t;f)\| = \infty$.

LEMMA 2.2. *Assume that* (H1) *and* (H2) *are satisfied. Then the function $u(t;f)$ of Lemma* 2.1 *is a continuous function of f in the sense that if $f \in E$ and $0 \leqq t < T_f$, there exist positive constants C and ε such that if $\hat{f} \in E$ and $\|f - \hat{f}\|_E < \varepsilon$, then $t < T_{\hat{f}}$ and $\|u(s;f) - u(s;\hat{f})\| \leqq C\|f - \hat{f}\|_E$ for all $-r_0 \leqq s \leqq t$.*

LEMMA 2.3. *Assume that* (H1) *and* (H2) *are satisfied. For $f \in E$ let $u(t;f)$ be the function given by Lemma* 2.1. *If $f(0) \in D(B), f' \in E$, and $f'^{-}(0) = Bf(0) + Kf$, then $u(t;f) \in D(B)$ for $0 \leqq t < T_f$, the function $t \to u(t;f)$ is continuously differentiable and satisfies $d/dt\, u(t;f) = Bu(t;f) + Ku_t$ on $[0, T_f)$, and $u_0 = f$.*

After the following definition we state a principle of linearized stability for (FDE_2).

DEFINITION 2.1. The *growth bound* ω of a linear semigroup $U(t)$, $t \geqq 0$, is defined by

$$\omega(U(t)) = \inf\{w \in R\text{: There exists } M \text{ such that } |U(t)| \leqq M e^{wt} \text{ for all } t \geqq 0\}.$$

THEOREM 2.1. *Let B and K satisfy* (H1) *and* (H2). *For each $f \in E$ let $u(t;f)$ be the solution of the integral equation* (1) *on the maximal interval of existence $[-r_0, T_f)$. Let $\hat{x} \in X$ satisfy $B\hat{x} + K\hat{f} = 0$, where $\hat{f} \in E$ is defined by $\hat{f}(s) = \hat{x}$ for all $s \in [-r_0, 0]$. Let $\hat{T}(t)$, $t \geqq 0$, be the strongly continuous semigroup of bounded linear operators in E with generator $\hat{A}$ defined by $\hat{A}f = f'$,*

$$D(\hat{A}) = \{f \in C^1([-r_0, 0], X) : f(0) \in D(B), f'(0) = Bf(0) + K'(\hat{f})f\}.$$

If $\omega(\hat{T}(t)) < 0$ then $\hat{f}$ is a locally exponentially stable equilibrium in the following sense: There exist $\varepsilon > 0$, $\tilde{M} \geqq 1$, and $\gamma < 0$ such that if $f \in E$ and $\|f - \hat{f}\|_E \leqq \varepsilon$, then $T_f = \infty$ and

$$\|u(t;f) - \hat{x}\| \leqq \tilde{M} e^{\gamma t} \|f - \hat{f}\|_E \quad \textit{for all } t \geqq 0.$$

Proof. If $\hat{x} \in X$ satisfies $B\hat{x} + K\hat{f} = 0$, then $u(t;\hat{f}) = \hat{x}$ for all $t \geqq 0$ by Lemmas 2.1 and 2.3. The fact that $\hat{A}$ generates a strongly continuous linear semigroup $\hat{T}(t)$, $t \geqq 0$, in E is proved in [10, B−IV, Thm. 3.1]. Also, $\hat{T}(t)$ is a translation semigroup, that is

$$(2)\quad \hat{T}(t)f(s) = \begin{cases} f(t+s) & \text{if } t+s \leqq 0, \\ \hat{T}(t+s)f(0) & \text{if } t+s \geqq 0, \end{cases}$$

and $\hat{T}(t)f(0)$ $(=(\hat{T}(t)f)(0))$ has the representation

$$\hat{T}(t)f(0) = T(t)f(0) + \int_0^t T(t-s)K'(\hat{f})(\hat{T}(s)f)\,ds, \qquad f \in E, \quad t \geqq 0.$$

In analogy to Webb [16, (4.100), p. 199] we claim the following fact:

(3) Let $\tilde{f} \in E$ and $t_1 > 0$. Let $h : [-r_0, t_1] \to X$ be continuous. Let $b : [-r_0, t_1] \to X$ satisfy

$$\begin{aligned} b(t) &= T(t)\tilde{f}(0) + \int_0^t T(t-s)(K'(\hat{f})b_s + h(s))\,ds, \qquad 0 \leqq t \leqq t_1 \\ b(t) &= \tilde{f}(0), \qquad -r_0 \leqq t \leqq 0. \end{aligned}$$

Then

$$b(t)=\hat{T}(t)\tilde{f}(0)+\int_0^t \hat{T}(t-s)h_s(0)\,ds, \qquad 0\leqq t\leqq t_1.$$

To prove (3), let $g:[0,t_1]\to E$ be defined by

$$g(t)(\theta)=\begin{cases}\left[\hat{T}(t)\tilde{f}+\displaystyle\int_0^{t+\theta}\hat{T}(t-s)h_s\,ds\right](\theta) & \text{if } t+\theta\geqq 0,\\ \tilde{f}(0) & \text{if } t+\theta\leqq 0,\end{cases}$$

and let $g_1:[0,t_1]\to E$ be defined by

$$g_1(t)(\theta)=\left[\hat{T}(t)\tilde{f}+\int_0^t \hat{T}(t-s)h_s\,ds\right](\theta), \qquad -r_0\leqq\theta\leqq 0.$$

If $\tilde{f}\in D(\hat{A})$ and h is continuously differentiable on $[-r_0,t_1]$, then by [6, p. 486],

$$\frac{d}{dt}g_1(t)=\hat{A}g_1(t)+h_t.$$

It can readily be verified that for $\hat{f}$ and h as above, and for every $\theta\in[-r_0,0]$ and $t\in[0,t_1]$ such that $t+\theta\geqq 0$, the function

$$\hat{T}(t)\tilde{f}+\int_0^{t+\theta}\hat{T}(t-s)h_s\,ds\in D(\hat{A}) \quad\text{and}\quad \hat{A}g(t)(0)=\hat{A}g_1(t)(0).$$

Thus,

$$\begin{aligned}\frac{d}{dt}g_1(t)(0)&=\hat{A}g(t)(0)+h_t(0)\\ &=Bg(t)(0)+K'(\hat{f})g(t)+h(t).\end{aligned}$$

Let

$$b_1(t)=\begin{cases}T(t)\tilde{f}(0)+\displaystyle\int_0^t T(t-s)(K'(\hat{f})g(s)+h(s))\,ds, & 0\leqq t\leqq t_1\\ \tilde{f}(0), \quad -r_0\leqq t\leqq 0.\end{cases}$$

Since

$$\begin{aligned}&g_1(t)(0)=\hat{T}(t)\tilde{f}(0)+\int_0^t\hat{T}(t-s)h_s(0)\,ds,\\ &g_1(0)(0)=\tilde{f}(0),\\ &\frac{d}{dt}g_1(t)(0)=Bg_1(t)(0)+K'(\hat{f})g(t)+h(t), \qquad 0\leqq t\leqq t_1,\end{aligned}$$

this implies, again by [6, p. 486],

$$g_1(t)(0)=g(t)(0)=b_1(t), \qquad 0\leqq t\leqq t_1.$$

For $\theta\in[-r_0,0]$, $t\in[0,t_1]$ such that $t+\theta\geqq 0$, $g(t)(\theta)=g(t+\theta)(0)=b_1(t+\theta)$ by the translation property (2).

For, θ, t such that $t+\theta\leqq 0$, $g(t)(\theta)=\tilde{f}(0)=b_1(t+\theta)$. Therefore, $g(t)=b_{1_t}$ and $b_1(t)=b(t)$, $0\leqq t\leqq t_1$. The fact (3) is thus proved for the case where $\tilde{f}\in D(\hat{A})$ and h is continuously differentiable. If $\tilde{f}\notin D(\hat{A})$ and h is not continuously differentiable, the argument of [16, p. 199] can be modified and applied to yield the conclusion again that $g(t)=b_t$, $0\leqq t\leqq t_1$.

Let $\bar\omega \in (\omega(\hat T(t)), 0)$. Then we can choose $M \geqq 1$ such that $|\hat T(t)| \leqq M e^{\bar\omega t}$ for $t \geqq 0$. Let $c_{\bar\omega} = e^{-2\bar\omega r_0}$, and let $r > 0$ such that $\|\circ(g)\| \leqq (-\bar\omega/2Mc_{\bar\omega})\|g\|_E$ for all $g \in E$ such that $\|g\|_E < r$. Let $\varepsilon = r/Mc_{\bar\omega}(=r/M e^{2\bar\omega r_0})$. Let $f \in E$ such that $\|f - \hat f\|_E < \varepsilon$, and let $t_1 \leqq \infty$ be the largest extended real number such that $\|u(t;f) - \hat x\| \leqq r$ for $0 \leqq t < t_1$. By (1) and (H2) we have for $0 \leqq t < t_1$,

$$\begin{aligned} u(t;f) &= T(t)f(0) + \int_0^t T(t-s)Ku_s\, ds \\ &= T(t)f(0) + \int_0^t T(t-s) \\ &\quad \cdot [K\hat f + K'(\hat f)(u_s - \hat f) + \circ(u_s - \hat f)]\, ds. \end{aligned} \tag{4}$$

We also have

$$\hat x = u(t;\hat f) = T(t)\hat x + \int_0^t T(t-s)K\hat f\, ds. \tag{5}$$

From (4) and (5) we have

$$u(t;f) - \hat x = T(t)(f(0) - \hat x) + \int_0^t T(t-s)[K'(\hat f)(u_s - \hat f) + \circ(u_s - \hat f)]\, ds.$$

Applying the fact (3) we obtain

$$u(t,f) - \hat x = \hat T(t)(f - \hat f)(0) + \int_0^t \hat T(t-s)h_s(0)\, ds,$$

where

$$h(s) := \begin{cases} \circ(u_s - \hat f), & s \in [0,t] \quad (t < t_1), \\ \circ(u_0 - \hat f), & s \in [-r_0, 0]. \end{cases}$$

Thus,

$$\|u(t;f) - \hat x\| \leqq \|\hat T(t)(f - \hat f)\|_E + \int_0^t \|\hat T(t-s)h_s\|_E\, ds. \tag{6}$$

For $s \in [0,t]$, $t < t_1$, $\|\hat T(t-s)h_s\|_E \leqq |\hat T(t-s)|\,\|h_s\|_E$.

Since

$$h(s+\theta) = \begin{cases} \circ(u_{s+\theta} - \hat f), & s+\theta \in [0,t], \\ \circ(u_0 - \hat f), & s+\theta \in [-r_0, 0], \end{cases}$$

$$\|h_s\|_E = \sup_{\theta \in [-r_0, 0]} \|h(s+\theta)\| \leqq \left(\frac{-\bar\omega}{2Mc_{\bar\omega}}\right)\|u_{\bar s} - \hat f\|_E$$

for some $\bar s \in [0, s]$. Let $\bar s = s + s_1$, $s_1 \in [-r_0, 0]$. If $s \geqq r_0$, then $s + \theta \in [-r_0, s]$ for $\theta \in [-2r_0, 0]$, and

$$\begin{aligned} \|u_{\bar s} - \hat f\|_E &= \sup_{\eta \in [-r_0, 0]} \|u(s + s_1 + \eta) - \hat f(\eta)\| \\ &\leqq \sup_{\theta \in [-2r_0, 0]} \|u(s+\theta) - \hat x\|. \end{aligned}$$

If $s < r_0$,

$$\begin{aligned} \|u_{\bar s} - \hat f\|_E &= \sup_{\eta \in [-r_0, 0]} \|u(s + s_1 + \eta) - \hat f(\eta)\| \\ &\leqq \sup_{\theta \in [-r_0 - s, 0]} \|u(s+\theta) - \hat x\|. \end{aligned}$$

For $s\in[0,t]$, $t<t_1$, we define a continuous function $\tilde{u}_s:[-2r_0,0]\to X$ by

$$\tilde{u}_s(\theta)=\tilde{u}(s+\theta;f)=\begin{cases}u(s+\theta;f) & \text{if } s+\theta\in[-r_0,s],\\ f(-r_0) & \text{if } s+\theta\in[-2r_0,-r_0].\end{cases}$$

Then for all $s\in[0,t]$, $t<t_1$,

$$\|u_{\bar{s}}-\hat{f}\|_E\leqq \sup_{\theta\in[-2r_0,0]}\|\tilde{u}(s+\theta)-\hat{x}\|,$$

and hence

$$\|h_s\|_E\leqq\left(\frac{-\bar{\omega}}{2Mc_{\bar{\omega}}}\right)\sup_{\theta\in[-2r_0,0]}\|\tilde{u}(s+\theta)-\hat{x}\|. \tag{7}$$

From (6) and (7) we obtain for $0\leqq t<t_1$,

$$\|u(t;f)-\hat{x}\|\leqq Me^{\bar{\omega}t}\|f-\hat{f}\|_E+\left(\frac{-\bar{\omega}}{2c_{\bar{\omega}}}\right)$$
$$\cdot\int_0^t e^{\bar{\omega}(t-s)}\sup_{\theta\in[-2r_0,0]}\|\tilde{u}(s+\theta)-\hat{x}\|\,ds.$$

If $t\in[0,t_1)$, $\theta\in[-2r_0,0]$ such that $0\leqq t+\theta\leqq t<t_1$, then replacing t by $t+\theta$ in the above we have

$$\begin{aligned}\|u&(t+\theta;f)-\hat{x}\|\\ &=\|\tilde{u}(t+\theta;f)-\hat{x}\|\\ &\leqq Me^{\bar{\omega}(t+\theta)}\|f-\hat{f}\|_E+\left(\frac{-\bar{\omega}}{2c_{\bar{\omega}}}\right)\int_0^{t+\theta}e^{\bar{\omega}(t+\theta-s)}\sup_{\theta\in[-2r_0,0]}\|\tilde{u}(s+\theta)-\hat{x}\|\,ds\\ &\leqq Me^{-2\bar{\omega}r_0}e^{\bar{\omega}t}\|f-\hat{f}\|_E+\left(\frac{-\bar{\omega}}{2c_{\bar{\omega}}}\right)\int_0^t e^{\bar{\omega}(t-s)}e^{-2\bar{\omega}r_0}\sup_{\theta\in[-2r_0,0]}\|\tilde{u}(s+\theta)-\hat{x}\|\,ds\\ &\leqq Me^{-2\bar{\omega}r_0}e^{\bar{\omega}t}\|f-\hat{f}\|_E+\left(\frac{-\bar{\omega}}{2}\right)e^{\bar{\omega}t}\int_0^t e^{-\bar{\omega}s}\sup_{\theta\in[-2r_0,0]}\|\tilde{u}(s+\theta;f)-\hat{x}\|\,ds.\end{aligned} \tag{8}$$

If $t\in[0,t_1)$, $\theta\in[-2r_0,0]$ such that $-r_0\leqq t+\theta<0$, then

$$\begin{aligned}\|\tilde{u}(t+\theta;f)-\hat{x}\|&=\|u(t+\theta;f)-\hat{x}\|\\ &=\|f(t+\theta)-\hat{f}(t+\theta)\|\\ &\leqq\|f-\hat{f}\|_E\\ &\leqq Me^{\bar{\omega}(t-2r_0)}\|f-\hat{f}\|_E.\end{aligned} \tag{9}$$

If $t\in[0,t_1)$, $\theta\in[-2r_0,0]$ such that $t+\theta\in[-2r_0,-r_0]$,

$$\begin{aligned}\|\tilde{u}(t+\theta;f)-\hat{x}\|&=\|f(-r_0)-\hat{x}\|\\ &=\|f(-r_0)-\hat{f}(-r_0)\|\\ &\leqq\|f-\hat{f}\|_E\\ &\leqq Me^{\bar{\omega}(t-2r_0)}\|f-\hat{f}\|_E.\end{aligned} \tag{10}$$

From (8)-(10) we can conclude that

$$\begin{aligned}e^{-\bar{\omega}t}&\sup_{\theta\in[-2r_0,0]}\|\tilde{u}(t+\theta;f)-\hat{x}\|\\ &\leqq Me^{-2\bar{\omega}r_0}\|f-\hat{f}\|_E+\left(\frac{-\bar{\omega}}{2}\right)\int_0^t e^{-\bar{\omega}s}\sup_{\theta\in[-2r_0,0]}\|\tilde{u}(s+\theta;f)-\hat{x}\|\,ds.\end{aligned}$$

Applying Gronwall's inequality we have

$$e^{-\bar{\omega}t} \sup_{\theta\in[-2r_0,0]} \|\tilde{u}(t+\theta;f)-\hat{x}\| \leqq M e^{-2\bar{\omega}r_0} \|f-\hat{f}\|_E e^{-(\bar{\omega}/2)t}$$

$$< \varepsilon M e^{-2\bar{\omega}r_0} e^{-(\bar{\omega}/2)t}.$$

Thus, for all $t\in[0, t_1)$

$$\|u(t;f)-\hat{x}\| < \varepsilon M e^{-2\bar{\omega}r_0} e^{(\bar{\omega}/2)t} \leqq r.$$

By Lemma 2.1, $t_1=\infty$ and we have shown that

$$\|u(t;f)-\hat{x}\| \leqq \tilde{M} e^{\gamma t} \|f-\hat{f}\|_E,$$

where $\tilde{M}=M e^{-2\bar{\omega}r_0}$ and $\gamma=\bar{\omega}/2$.

3. Positivity and the linearizations of (FDE$_1$) or (FDE$_2$). We consider the linearized abstract functional differential equation

$$\text{(FDE)}_\text{L} \qquad \begin{aligned} \dot{u}(t) &= Bu(t)+\psi u_t, \qquad t\geqq 0,\\ u_0 &= f, \end{aligned}$$

where we assume that B satisfies (H1) and ψ satisfies

(H2)′ $\psi\in\mathscr{L}(E, X)$ (that is, $\psi: E\to X$ is a bounded linear operator).

Let X be a Banach lattice. (The reader can refer to [10] or [13] for basic facts about Banach lattices and positive semigroups.) Then $E=C([-r_0,0],X)$ is also a Banach lattice with the natural pointwise order and the supremum norm. Let X_+ be the positive cone of X (that is, $X_+=\{x\in X: x\geqslant 0\}$, and let E_+ be the positive cone of E.

DEFINITION 3.1. An operator B with domain and range contained in a Banach lattice X is a *positive operator* on X if $Bx\in X_+$ whenever $x\in X_+$. A semigroup $T(t)$, $t\geqq 0$, on X is a *positive semigroup* if $T(t)x\in X_+$ whenever $x\in X_+$ and $t\in R_+$.

DEFINITION 3.2. Let A be the generator of a strongly continuous linear semigroup, $T(t)$, $t\geqq 0$, on a Banach space X. The *spectral bound* of A, $s(A)$, is defined by

$$s(A)=\sup\{\text{Re }\lambda: \lambda\in\sigma(A)\}.$$

(Here $\sigma(A)$ denotes the spectrum of A.)

If X is finite-dimensional, $s(A)=\omega(T(t))$, the growth bound of the semigroup (defined by Definition 2.1), and hence we obtain the classical Lyapunov stability theorem. It is well known that in general $s(A)\leqq\omega(T(t))<+\infty$, but strict inequality may occur. However, for positive linear semigroups, we can often conclude the uniform exponential stability of the zero equilibrium of the semigroup whenever $s(A)<0$. We state below some of the known important consequences of positivity for the stability study of (FDE)$_\text{L}$.

DEFINITION 3.3. A strongly continuous linear semigroup $T(t)$, $t\geqq 0$, on a Banach space X, with generator A, is called

(i) *exponentially stable* if there exists $\gamma>0$ such that $\lim_{t\to\infty} e^{\gamma t}\|T(t)f\|=0$ for every $f\in D(A)$.

(ii) *uniformly exponentially stable* if there exists $\gamma>0$ such that $\lim_{t\to\infty} e^{\gamma t}|T(t)|=0$.

We note that if $\omega(T(t))<0$, the semigroup $T(t)$ is uniformly exponentially stable.

PROPOSITION 3.1 [10, C-IV, Thms. 1.1 and 1.3]. *Let A be the generator of a positive linear semigroup $T(t)$, $t\geqq 0$, on a Banach lattice X. Then the following properties are equivalent*:

(a) *The semigroup $T(t)$ is exponentially stable.*

(b) *The spectral bound $s(A)$ is less than zero.*

If, in addition, X is a space $C(K)$, K compact, $C_0(Y)$, Y locally compact, or $L^1(Y, \mu)$ or $L^2(Y, \mu)$ for some measure space (Y, μ), then the above properties are equivalent to

(c) *The semigroup $T(t)$ is uniformly exponentially stable.*

PROPOSITION 3.2 [10, C-III, Cor. 1.4]. *Let $T(t)$, $t \geqq 0$, be a positive linear semigroup defined on a Banach lattice X, and let A be its generator. Then $s(A) \in \sigma(A)$ unless $s(A) = -\infty$.*

DEFINITION 3.4. For $\lambda \in \mathbb{C}$, $x \in X$, $g \in E = C([-r_0, 0], X)$, define the following operators:

(i) $\varepsilon_\lambda \otimes x \in E$ by $(\varepsilon_\lambda \otimes x)(s) = e^{\lambda s} \cdot x$, $s \in [-r_0, 0]$;

(ii) $H_\lambda \in \mathscr{L}(E)$ by $H_\lambda g(t) = \int_t^0 e^{\lambda(t-s)} g(s)\, ds$, $t \in [-r_0, 0]$;

(iii) $\psi_\lambda \in \mathscr{L}(X)$ by $\psi_\lambda(x) = \psi(\varepsilon_\lambda \otimes x)$.

PROPOSITION 3.3 [10, B-IV, Prop. 35 and Thm. 3.7]. *Let X be a Banach lattice. Assume, in addition to (H1) and (H2)$'$, that B generates a positive semigroup and ψ is a positive operator. Let $\hat{T}(t)$, $t \geqq 0$, be the strongly continuous semigroup of bounded linear operators in $E = C([-r_0, 0], X)$ with generator $\hat{A}$ defined by*

$$\hat{A}f = f',$$

$$D(\hat{A}) = \{f \in C^1([-r_0, 0], X) : f(0) \in D(B), f'(0) = Bf(0) + \psi f\}.$$

Then the semigroup $\hat{T}(t)$ is positive. For the generator $\hat{A}$ of $\hat{T}(t)$ and $\lambda \in R$ the following statements hold:

(a) *If $s(B + \psi_\lambda) < \lambda$, then $s(A) < \lambda$;*

(b) *If $s(B + \psi_\lambda) = \lambda$, then $s(A) = \lambda$;*

(c) *Suppose that B has compact resolvent and there exists $\mu \in R$ with $\sigma(B + \psi_\mu) \neq \varnothing$. Then*

$$s(B + \psi_\lambda) \leqq \lambda \text{ if and only if } s(\hat{A}) \leqq \lambda.$$

(*In particular, $s(B + \psi_0) < 0$ if and only if $s(\hat{A}) < 0$.*)

DEFINITION 3.5. A continuous map $r : [-1, 0] \to R_-$ satisfying $\min_{-1 \leqq s \leqq 0} r(s) = -r_0$ is called a *delay function* on $[-r_0, 0]$. If ψ is a bounded linear operator from $C([-1, 0], X)$ into X, the *delayed operator* $\psi_r \in \mathscr{L}(E, X)$ is defined by $\psi_r f = \psi(f \circ r)$, $f \in E$.

PROPOSITION 3.4 [7, Thm. 4.3]. *Let X be a Banach lattice. Assume that B satisfies (H1) and generates a positive semigroup on X, and ψ is a positive bounded linear operator from $C([-1, 0], X)$ to X. Let $\hat{T}_{B,\psi_r}(t)$, $t \geqq 0$, be the strongly continuous semigroup of bounded linear operators in E with generator $\hat{A}_{B,\psi_r}$ defined by $\hat{A}_{B,\psi_r} f = f'$,*

$$D(\hat{A}_{B,\psi_r}) = \{f \in C^1([-r_0, 0], X) : f(0) \in D(B), f'(0) = Bf(0) + \psi_r f\}.$$

If $s(B + \psi_0) < 0$ then $s(\hat{A}_{B,\psi_r}) < 0$ for every delay function r.

DEFINITION 3.6. If $\psi_1, \psi_2 \in \mathscr{L}(E, X)$ then ψ_1 is said to *dominate* ψ_2 if $|\psi_2 f| \leqq \psi_1 |f|$ for all $f \in E$. (Here $|f| = \sup(f, -f)$.) If $T_1(t)$, $T_2(t)$, $t \geqq 0$, are semigroups on E, $T_1(t)$ is said to *dominate* $T_2(t)$ if $|T_2(t) f| \leqq T_1(t)|f|$ for every $f \in E$, $t \geqq 0$.

PROPOSITION 3.5 [7, Cor. 4.4]. *Let X be a Banach lattice. Let B be the generator of a linear semigroup $T(t)$, $t \geqq 0$, on X, and let $\psi \in \mathscr{L}(C([-1, 0], X), X)$. Assume that there exists a semigroup $\tilde{T}(t)$, $t \geqq 0$, on X with generator $\tilde{B}$ which dominates $T(t)$, and an operator $\tilde{\psi}$ which dominates ψ. Then the linear semigroup $\hat{T}_{B,\psi_r}(t)$ (described in Proposition 3.4) is uniformly exponentially stable for all delay functions r if the spectral bound $s(\tilde{B} + \tilde{\psi}_0) < 0$ and one of the following conditions is satisfied:*

(a) *$X = C(K)$, K compact;*

(b) *$\tilde{T}(t)$ is norm continuous for $t > 0$ (that is, the function $t \to \tilde{T}(t)$ from $(0, \infty)$ into $\mathscr{L}(X)$ is norm continuous).*

We now examine the consequences of the strong positivity property of irreducibility for the stability study of $(\mathrm{FDE})_{\mathrm{L}}$.

DEFINITION 3.7. An operator B on a Banach lattice X is *strictly positive* if $x \in X_+$, $x \neq 0$ implies $Bx \in X_+$, $Bx \neq 0$. A linear semigroup $T(t)$, $t \geqq 0$, on X is *irreducible* if given $x \in X_+$, $x \neq 0$, $x^* \in X_+^*$, $x^* \neq 0$, there exists $t_0 > 0$ such that $\langle T(t_0)x, x^*\rangle > 0$. (For $y \in X$, $y^* \in X^*$, $\langle y, y^*\rangle = y^*(y)$.) A bounded operator T on X is *irreducible* if the semigroup $\{T^n\colon n \in N\}$ is irreducible.

It is well known that the eigenvalues of irreducible square matrices have special properties (see, for example, [13]). Likewise, positive irreducible semigroups possess special properties which are useful in the stability study of certain equations (see, for example, [10]).

We first give sufficient conditions which insure that the semigroup $T(t)$ corresponding to $(\mathrm{FDE})_{\mathrm{L}}$ (in the sense described in Proposition 3.3) is irreducible. A corresponding result for the functional equation $u(t) = \Phi u_t$, $t \geqq 0$, $u_0 = f$, was obtained in a recent paper by Grabosch [4, Prop. 3.7]. In that paper $E = L^1((-\infty, 0], F, e^{\eta s}\, ds)$ and F is a Banach lattice. The method of proof of the proposition below was motivated by the proof of [4, Prop. 3.7].

PROPOSITION 3.6. *Let X be a Banach lattice. Assume that B and ψ satisfy* (H1) *and* (H2)′. *In addition, assume that B generates a positive irreducible semigroup and ψ is strictly positive. Then the semigroup $\hat{T}(t)$, $t \geqq 0$, of Proposition* 3.3 *is irreducible.*

For the proof we will use the following two lemmas.

LEMMA 3.1 [10, C-III, Def. 3.1]. *Let E be a Banach lattice and $\hat{T}(t)$, $t \geqslant 0$, be a strongly continuous linear semigroup on E with generator $\hat{A}$. The following assertions are equivalent*:

(i) *$\hat{T}(t)$ is irreducible.*

(ii) *For some* (*every*) *$\lambda > s(\hat{A})$ the resolvent $R(\lambda, \hat{A})$ is irreducible.*

(iii) *For some* (*every*) *$\lambda > s(\hat{A})$, $R(\lambda, \hat{A})f$ is a quasi-interior point of E_+ whenever $f > 0$* (*that is, for all $\phi \in E^*$, $\phi > 0$, $\langle R(\lambda, \hat{A})f, \phi\rangle > 0$*).

LEMMA 3.2. *Let $\hat{x}$ be a quasi-interior point of X_+. Then $\varepsilon_\lambda \otimes \hat{x}$* (*as defined by Definition* 3.4) *is a quasi-interior point of E_+.*

Proof of Lemma 3.2. Let $\hat{x}$ be a quasi-interior point of X_+. Then the ideal $I_{\hat{x}}$ which is generated by $\hat{x}$ is dense in X (see [13, II, Def. 6.1]). Since ε_λ is a quasi-interior point of $C([-r_0, 0])_+$ [10, C-I, p. 238] the ideal I_{ε_λ} is dense in $C([-r_0, 0])$. Hence $I_{\varepsilon_\lambda} \otimes I_{\hat{x}}$ is dense in $C([-r_0, 0])\, \tilde{\otimes}_{\varepsilon} X$, which is isomorphic to $E = C([-r_0, 0], X)$ [13, p. 237]. (We are using here the tensor product notation from [13].) But $I_{\varepsilon_\lambda \otimes \hat{x}} \supseteq I_{\varepsilon_\lambda} \otimes I_{\hat{x}}$ implies that $I_{\varepsilon_\lambda \otimes \hat{x}}$ is dense in E, and hence that $\varepsilon_\lambda \otimes \hat{x}$ is a quasi-interior point of E_+.

Proof of Proposition 3.6. From [10, B-IV, Prop. 3.4] we have $\lambda \in \rho(B + \psi_\lambda)$ if and only if $\lambda \in \rho(\hat{A})$, and

(11) $R(\lambda, \hat{A})g = \varepsilon_\lambda \otimes [R(\lambda, B + \psi_\lambda)(g(0) + \psi H_\lambda g)] + H_\lambda g$, $g \in E$, where $\psi_\lambda \in \mathscr{L}(X)$ and $H_\lambda \in \mathscr{L}(E)$ are defined in Definition 3.4.

We will first show that

(12) $R(\lambda, B + \psi_\lambda)(g(0) + \psi H_\lambda g)$ is a quasi-interior point of X_+ (for sufficiently large λ) whenever $g \in E$, $g > 0$.

We note that if $g > 0$, then $g(0) \geqq 0$. Also, $H_\lambda g > 0$ if $g > 0$. Thus, due to the strict positivity of ψ, we have $\psi H_\lambda g > 0$ if $g > 0$. Therefore, if $g > 0$ then $g(0) + \psi H_\lambda g > 0$. So, if we show that $R(\lambda, B + \psi_\lambda)f$ is a quasi-interior point of X_+ whenever $f \in X$, $f > 0$, then this will establish (12). From [10, (1.12), p. 44] there exists $\lambda_0 \in R$ such that

$$R(\lambda, B + \psi_\lambda) = R(\lambda, B)(I - \psi_\lambda R(\lambda, B))^{-1} \quad \text{for } \lambda \geqq \lambda_0. \tag{13}$$

Since we are assuming that B generates an irreducible semigroup on X, this implies by Lemma 3.1 that $R(\lambda, B)\bar{f}$ is a quasi-interior point of X_+ whenever $\bar{f}\in X$, $\bar{f}>0$. Thus, if we show that $\bar{f}:=(I-\psi_\lambda R(\lambda, B))^{-1}f>0>0$ when $f>0$, then, by (13), $R(\lambda, B+\psi_\lambda)f$ is a quasi-interior point of X_+ when $f>0$. Since $1\in\rho(\psi_\lambda R(\lambda, B))$ for $\lambda\geqq\lambda_0$ and we know that $R(\lambda, B)$ is irreducible (hence strictly positive) and ψ_λ is strictly positive (since ψ is strictly positive), we have for $f>0$

$$(I-\psi_\lambda R(\lambda, B))^{-1}f=R(1, \psi_\lambda R(\lambda, B))f=\sum_{n=1}^{\infty}(\psi_\lambda R(\lambda, B))^n f>0.$$

Thus, $R(\lambda, B+\psi_\lambda)f$ is a quasi-interior point of X_+ when $f>0$. By Lemma 3.2, $\varepsilon_\lambda\otimes[R(\lambda, B+\psi_\lambda)(g(0)+\psi H_\lambda g)]$ is a quasi-interior point of E_+. Since $H_\lambda g>0$ for $g\in E$, $g>0$, $\varepsilon_\lambda\otimes[R(\lambda, B+\psi_\lambda)(g(0)+\psi H_\lambda g)]+H_\lambda g$ is a quasi-interior point of E_+ as well. Thus by (11) and Lemma 3.1, $R(\lambda, \hat{A})$ is irreducible and $\hat{T}(t)$ is irreducible.

Using Proposition 3.6 along with known facts from positive semigroup theory we obtain our main result for this section.

THEOREM 3.1. *Let X be a Banach lattice. Assume that B and ψ satisfy* (H1) *and* (H2)′. *In addition assume that B generates a positive irreducible semigroup $T(t)$, $t\geqq 0$, which is compact for each $t>0$, and ψ is strictly positive. Let $\hat{T}(t)$, $t\geqq 0$, be the strongly continuous semigroup in E with generator $\hat{A}$, as defined in Proposition* 3.3. *Then the following assertions hold*:

(i) *There exists a unique real number $\lambda_0=s(\hat{A})=\omega(\hat{T}(t))$ and a rank* 1 *projection P such that*

$$\|e^{-\lambda_0 t}\,\hat{T}(t)-P\|\leqq M e^{-\delta t}$$

for suitable constants $\delta>0$, $M\geqq 1$, and all $t\geqq 0$. The projection P has the form $P=\phi\otimes h$, where h is a quasi-interior point of E_+, and ϕ is a strictly positive linear form on E.

(ii) *λ_0 is the unique solution of the equation $\lambda=s(B+\psi_\lambda)$.*

Remark. Assertion (i) of Theorem 3.1 says that under the given conditions, solutions of $(\mathrm{FDE})_\mathrm{L}$ have *asynchronous exponential growth.* The constant λ_0 is called the *Malthusian parameter* and P is called the *exponential steady state.*

Proof of Theorem 3.1. The hypotheses that B generates a positive irreducible semigroup and ψ is strictly positive, along with (H1) and (H2)′, imply by Proposition 3.6 that $\hat{T}(t)$ is irreducible. The assumption that $T(t)$ is compact for each $t>0$ guarantees that $\hat{T}(t)$ is eventually compact (that is, compact for $t>r_0$) [15, Prop. 2.4]. By [10, C-III, Thm. 3.7(c)], $\sigma(\hat{A})\neq\varnothing$. Assertion (i) follows from [10, C-IV, Thm. 2.1 and Remarks 2.2(d), (e)]. Noting that $\sigma(B)\neq\varnothing$ (which again follows from [10, C-III, Thm. 3.7(c)], we have $-\infty<s(B)\leqq s(B+\psi_\lambda)$ for all $\lambda\in R$, which implies that $\sigma(B+\psi_\lambda)\neq\varnothing$. Thus assertion (ii) follows from [10, B-IV, Prop. 3.6].

If, in Theorem 3.1, ψ is not strictly positive but is dominated by a strictly positive bounded linear operator, then we have the following result.

COROLLARY 3.1. *Let X be a Banach lattice. Assume that B and ψ satisfy* (H1) *and* (H2)′. *Assume that B generates a positive irreducible semigroup $T(t)$, $t\geqq 0$, which is compact for each $t>0$. Assume there is a strictly positive linear operator $\tilde{\psi}\in\mathscr{L}(E, X)$ such that $|\psi f|\leqq\tilde{\psi}|f|$ for all $f\in E$. Let $\hat{T}(t)$, $t\geqq 0$, be the strongly continuous semigroup in E with generator $\hat{A}$, as defined in Proposition* 3.3. *Then there exists a unique real number $\lambda_0=s(\hat{A}_{\tilde{\psi}})=\omega(\hat{T}_{\tilde{\psi}}(t))$ and a positive constant M such that for $t\geqq 0$ and each $f\in E$,*

$$\|\hat{T}(t)f\|\leqq M e^{\lambda_0 t}\|f\|+o(e^{\lambda_0 t}).$$

(*Here $\hat{T}_{\tilde{\psi}}(t)$, $t\geqq 0$, is the linear semigroup, with generator $\hat{A}_{\tilde{\psi}}$, defined in a fashion analogous to $\hat{T}(t)$ and $\hat{A}$ except with $\tilde{\psi}$ replacing ψ* (*cf. Proposition* 3.3).)

Proof. By [7, Prop. 3.2] we know that $|\hat{T}(t)f| \leqq \hat{T}_{\tilde{\psi}}(t)|f|$. Since B and $\tilde{\psi}$ satisfy the hypotheses of Theorem 3.1, assertion (i) of that theorem holds for $\hat{T}_{\tilde{\psi}}(t)$. This assertion can also be formulated as follows: There exists $\phi \in E_+^*$, $\phi \neq 0$, and a quasi-interior element h of E_+ such that for $t \geqq 0$ and each $f \in E$,

$$\hat{T}_{\tilde{\psi}}(t)f = \phi(f)\, e^{\lambda_0 t}\, h + \circ(e^{\lambda_0 t}).$$

Thus,

$$\begin{aligned} \lim_{t\to\infty} \|e^{-\lambda_0 t}\, \hat{T}(t)f\| &\leqq \lim_{t\to\infty} \|e^{-\lambda_0 t}\, \hat{T}_{\tilde{\psi}}(t)|f|\| \\ &\leqq \|\phi(|f|)h\| \leqq M\|f\|, \end{aligned}$$

or, equivalently,

$$\|\hat{T}(t)f\| \leqq M e^{\lambda_0 t}\|f\| + \circ(e^{\lambda_0 t}).$$

4. Example. We now apply some of our results to the motivating population equation

$$\text{(E)}' \qquad \begin{aligned} \frac{\partial u}{\partial t}(x,t) &= d\Delta u(x,t) + au(x,t) \\ &\cdot\left[1 - bu(x,t) - \int_{-1}^{0} u(x, t + r(s))\, d\eta(s)\right], \end{aligned}$$

where $x \in [0, \pi]$, $t \geqq 0$, $\Delta = \partial^2/\partial x^2$ with Dirichlet boundary conditions, a, b, d are positive constants, η is a strictly positive measure on $[-1, 0]$ such that $b + \|\eta\| = 1$, and r is a delay function on $[-r_0, 0]$. If $a > d$ then it is known that there exists a stationary solution $h \in C^2[0, \pi]$ of (E) which is strictly positive on $(0, \pi)$ (see [14]). Thus $dh'' + ah(1-h) = 0$.

Let $X = \{f \in C[0, \pi]\colon f(0) = f(\pi) = 0\}$ and

$$B_0 = \frac{d^2}{dx^2} \quad \text{with maximal domain in } X.$$

As an abstract equation in X, (E)$'$ can be written as

$$\begin{aligned} \dot{u}(t) &= dB_0\, u(t) + au(t)\left[1 - bu(t) - \int_{-1}^{0} (u_t \circ r)(s)\, d\eta(s)\right] \\ &= dB_0\, u(t) + au(t) - abu_t^2(0) - au_t(0)\int_{-1}^{0} (u_t \circ r)(s)\, d\eta(s). \end{aligned}$$

Thus we have

$$\text{(14)} \qquad \dot{u}(t) = Bu(t) + Ku_t, \qquad t \geqq 0,$$

where $Bu = dB_0 u + au$, and for $g \in E = C([-r_0, 0], X)$,

$$\text{(15)} \qquad Kg = -abg^2(0) - ag(0)\int_{-1}^{0} (g \circ r)(s)\, d\eta(s).$$

It is well known that B_0 generates a contraction semigroup on X and B generates a strongly continuous semigroup on X. It is easy to check that K is continuously Fréchet-differentiable in the sense of (H2). Thus if we let $u_0 = f$, then (E)$'$ has the abstract form of (FDE$_2$). From Lemmas 2.1–2.3 we have the existence of a unique local solution of (14) for each $f \in E$. From Theorem 2.1 we know that the linearization of (14) at the stationary solution has the form

$$\dot{u}(t) = Bu(t) + \psi u_t, \qquad t \geqq 0, \tag{16}$$

where $\psi = K'(\hat{f})$, and $\hat{f}$ is the constant function in E defined by $\hat{f}(s) = h$ for all $s \in [-r_0, 0]$. We see from (15) that

$$\begin{aligned} \psi u_t &= K'(\hat{f})u_t \\ &= -2abhu(t) - au(t)\int_{-1}^{0} h\, d\eta(s) - ah\int_{-1}^{0} (u_t \circ r)(s)\, d\eta(s) \\ &= -abhu(t) - ahu(t)(b + \|\eta\|) - ah\int_{-1}^{0} (u_t \circ r)(s)\, d\eta(s) \\ &= (-ah - abh)u(t) - ah\int_{-1}^{0} (u_t \circ r)(s)\, d\eta(s). \end{aligned} \tag{17}$$

In [7], Kerscher and Nagel show that if $a > d$ and $b > \|\eta\|$ then the linear solution semigroup corresponding to (16), or the abstract equation in E, $v'(t) = \hat{A}v(t)$, is uniformly exponentially stable independent of the delay function r. From Theorem 2.1 we can conclude the following.

PROPOSITION 4.1. *If in* (E)$'$, $a > d$ *and* $b > \|\eta\|$, *then the equilibrium* $\hat{f} \in E$, *where* $\hat{f}(s) = h$, $s \in [-r_0, 0]$, *is locally exponentially stable in the sense of Theorem* 2.1, *independent of the delay function r.*

Proof. For completeness, and to illustrate the use of positivity in obtaining stability information for linear semigroups, we sketch the proof from [7, Prop., p. 48] that, under the given hypotheses, the linear solution semigroup corresponding to (16) is uniformly exponentially stable independent of the delay function.

It is well known that the operator B_0 generates a positive irreducible semigroup on X which is compact for $t > 0$ (see [10, B-III, § 3]). Let M_g denote the bounded multiplication operator with the function $g \in X$. Then the same properties hold for the semigroup generated by $B_0 + M_g$ (see [10, A-II, Thm. 1.30 and B-III, Prop. 3.3]). Since the function h is strictly positive and $dh'' + a(1-h)h = 0$, this implies that zero is an eigenvalue of $B_1 := dB_0 + M_{a(1-h)}$, which admits a strictly positive eigenfunction. It follows that the spectral bound of B_1 is zero (see [10, B-III, § 3]). Equation (16), with ψ as given in (17), is equivalent to the equation

$$\dot{u}(t) = \tilde{B}u(t) + \phi_r u_t, \qquad t \geqq 0, \tag{18}$$

where $\tilde{B} := dB_0 + M_{a(1-h-bh)}$, and $\phi_r f := -ah\int_{-1}^{0} (f \circ r)(s)\, d\eta(s)$ for $f \in E$. Let $B_2 := \tilde{B} + |\phi|_0$. Since $\phi_0(x) = \phi(1 \otimes x)$ for $x \in X$, where $\phi f := -ah\int_{-1}^{0} f(s)\, d\eta(s)$, $\phi_0(x) = -M_{ah}\|\eta\|x$. This implies $|\phi|_0 = M_{ah}\|\eta\|$. Thus,

$$\begin{aligned} B_2 &= dB_0 + M_{a(1-h-bh)} + M_{ah}\|\eta\| \\ &= dB_0 + M_{a(1-h)} - M_{ah(b-\|\eta\|)} \leqq B_1, \end{aligned}$$

by the assumption that $b > \|\eta\|$. Hence $s(B_2) \leqq s(B_1) = 0$. Assume that $s(B_2) = 0$. Then there exists a strictly positive fixed function $g \in X$ for $S_2(t)$, $t \geqq 0$, the semigroup generated by B_2; i.e., $S_2(t)g = g$ for $t \geqq 0$ (see [10, B-IV, § 2]). Thus, $S_1(t)g \geqq S_2(t)g = g$, $t \geqq 0$, for the semigroup $S_1(t)$, $t \geqq 0$, generated by B_1. But $S_1(t)$ possesses a strictly positive invariant linear form (see [10, B-II, § 3 and B-III, Prop. 1.5 and Thm. 1.6]). Hence $S_1(t)g = g$ and $B_1 g = 0 = B_2 g$, which is impossible since $M_{ah(b-\|\eta\|)} \neq 0$. We therefore conclude that $s(B_2)<0$. By Proposition 3.5, the solution semigroup corresponding to (18), and hence (16), is uniformly exponentially stable independent of the delay function r. The conclusion of the proposition follows from Theorem 2.1.

We can observe from the proof of Proposition 4.1 that the operator B in (16) generates a positive irreducible semigroup that is compact for each $t>0$. Also ψ, as given by (17), is dominated by the strictly positive operator $\tilde{\psi}$ defined by $\tilde{\psi}f=a(1+b)hf(0)+ah\int_{-1}^{0} f(r(s))\,d\eta(s)$ for $f\in E$. Therefore we can apply Corollary 3.1 to obtain the estimate

$$\|\hat{T}(t)f\| \leqq M e^{\lambda_0 t}\|f\| + \circ(e^{\lambda_0 t}) \tag{19}$$

for the linear semigroup $\hat{T}(t)$ corresponding to (16). From the analysis of [7] given in the proof of Proposition 4.1, the conclusion that

$$s(B_2)=s(\tilde{B}+|\phi|_0)<0$$

allows us to conclude that

$$\lambda_0=s(\hat{A}_{\tilde{\psi}})=\omega(\hat{T}_{\tilde{\psi}}(t))<0.$$

Thus the conclusion of uniform exponential stability of $\hat{T}(t)$ is also obtained here by using the estimate (19).

REFERENCES

[1] J. CUSHING, *Integrodifferential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomath. 20, Springer-Verlag, Berlin, 1977.

[2] W. DESCH AND W. SCHAPPACHER, *Linearized stability for nonlinear semigroups*, in Differential Equations in Banach Spaces, Bologna 1985, Lecture Notes in Math. 1223, Springer-Verlag, Berlin, New York, 1986, pp. 61-73.

[3] W. FITZGIBBON, *Semilinear functional differential equations in Banach space*, J. Differential Equations, 29 (1978), pp. 1-14.

[4] A. GRABOSCH, *Translation semigroups and their linearizations on spaces of integrable functions*, Trans. Amer. Math. Soc., 311 (1989), pp. 357-390.

[5] D. GREEN AND H. STECH, *Diffusion and hereditary effects in a class of population models*, in Differential Equations and Applications, Academic Press, 1981.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Grundlehren Math. Wiss. 132, Springer-Verlag, New York, 1966.

[7] W. KERSCHER AND R. NAGEL, *Positivity and stability for Cauchy problems with delay*, Semesterbericht Funktionalanalysis, Universität Tübingen, Tübingen, Sommersemester 1986, pp. 35-55.

[8] J. LIGHTBOURNE, *Nonlinear retarded perturbation of a linear evolution system*, in Integral Equations and Functional Differential Equations, T. Herdman, S. Rankin, and H. Stech, eds., Marcel Dekker, New York, 1980, pp. 201-212.

[9] B. MARTIN AND H. SMITH, *Abstract functional differential equations and reaction-diffusion systems*, preprint.

[10] R. NAGEL (ED.), *One-Parameter Semigroups of Positive Operators*, Lecture Notes in Math. 1184, Springer-Verlag, Berlin, New York, 1986.

[11] M. PARROTT, *Positivity and a principal of linearized stability for delay-differential equations*, J. Differential and Integral Equations, 2 (1989), pp. 170-182.

[12] S. RANKIN, *Existence and asymptotic behavior of a functional differential equation in Banach space*, J. Math. Anal. Appl., 88 (1982), pp. 531-542.

[13] H. SCHAEFER, *Banach Lattices and Positive Operators*, Grundlehren Math. Wiss. 215, Springer-Verlag, Berlin, Heidelberg, New York, 1974.

[14] H. STECH, *The effect of time lags on the stability of the equilibrium state of population growth equations*, J. Math. Biol., 5 (1978), pp. 115-120.

[15] C. TRAVIS AND G. WEBB, *Existence and stability for partial functional differential equations*, Trans. Amer. Math. Soc., 200 (1974), pp. 395-418.

[16] G. WEBB, *Theory of Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York, 1985.

[17] ———, *Asymptotic stability for abstract nonlinear functional differential equations*, Proc. Amer. Math. Soc., 54 (1974), pp. 225-230.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 662–688, May 1992

006

ANALYTICAL AND NUMERICAL RESULTS FOR THE AGE-STRUCTURED S-I-S EPIDEMIC MODEL WITH MIXED INTER-INTRACOHORT TRANSMISSION*

M. IANNELLI[†], F. A. MILNER[‡], AND A. PUGLIESE[†]

Abstract. A model which describes the dynamics of an $S \to I \to S$ epidemic in an age-structured population at the steady state is considered. The model consists of a nonlinear and nonlocal system of equations of hyperbolic type and has already been partly analyzed by other authors. Here, a special form for the force of infection is considered. Explicitly computable threshold conditions are given, and some regularity results for the solutions are proven. An implicit finite difference method of characteristics to approximate the solutions is used. Optimal error estimates are derived and results from numerical simulations are presented. The discrete dynamical system arising from the numerical algorithm, is also analyzed, showing that it shares many properties of the continuous model.

Key words. age structure, epidemic models, numerical method, discrete dynamical system

AMS(MOS) subject classifications. 35L60, 47H20, 65M25, 92D30

1. Introduction.

The importance of age structure in epidemic models has been recently stressed by many authors who have considered models for many different situations. Recently, Busenberg et al. [4], [5] have provided a complete analysis of a fairly general SIS model with age structure, showing existence of a threshold for endemic states. Our aim, in this paper, is to further develop this model by considering a special form for the force of infection, and to provide a numerical algorithm to approximate the solutions.

Since the model concerns diseases which do not impart immunity, it does not have many applications. Nonetheless, our results have a theoretical interest, and can be viewed as a preliminary step towards the study and simulations of more complex models such as the SIR models, which are used in the description of most childhood diseases. A significant advantage in the SIS case studied here is that the asymptotic behaviour of the model is completely known theoretically.

Let $p(a,t)$ be the age distribution of a population that is contaminated in part by a disease which does not impart immunity or affect the death rate. Within this population we distinguish the subpopulation of infected individuals and that of susceptibles. Let $i(a,t)$ and $s(a,t)$ denote, respectively, the age distributions of infected and susceptible individuals. The fact that the disease does not impart immunity means that

$$p(a,t) = i(a,t) + s(a,t). \tag{1.1}$$

We shall assume that infected and susceptible individuals interact with each other freely and uniformly. Thus, we shall assume that i and s are solutions of the following

*Received by the editors January 19, 1990; accepted for publication (in revised form) August 5, 1991.

[†] Dipartimento di Matematica, Università di Trento, 38050 Trento, Italy.

[‡] Dipartimento di Matematica, IIa Università di Roma, 00133 Roma, Italy, and Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

coupled system of equations:

$$(1.2)\qquad \begin{cases} \dfrac{\partial i}{\partial t}+\dfrac{\partial i}{\partial a}=-\mu i+\lambda s-\gamma i, & a>0,\quad t>0,\\[2mm] i(0,t)=B_i(t)=q\displaystyle\int_0^\infty \beta(a)i(a,t)da, & t\ge 0,\\[2mm] i(a,0)=i^0(a), \quad a\ge 0, \end{cases}$$

$$(1.3)\qquad \begin{cases} \dfrac{\partial s}{\partial t}+\dfrac{\partial s}{\partial a}=-\mu s-\lambda s+\gamma i, & a>0,\quad t>0,\\[2mm] s(0,t)=B_s(t)=\displaystyle\int_0^\infty \beta(a)\big[s(a,t)+(1-q)i(a,t)\big]da, & t\ge 0,\\[2mm] s(a,0)=s^0(a), \quad a\ge 0, \end{cases}$$

where $\mu=\mu(a)$ is the age-specific death-rate, $\lambda=\lambda(a,i)$ is the age-specific force of infection, $\beta=\beta(a)$ is the age-specific birth-rate, and $\gamma=\gamma(a)$ is the age-specific recovery rate. The constant q is the probability that the disease be transmitted vertically. When there is no vertical transmission $q=0$ and thus $B_i\equiv 0$, that is, all newborns are susceptible. Adding (1.2) and (1.3) we arrive at the well-known McKendrick–von Foerster equation for p:

$$(1.4)\qquad \begin{cases} \dfrac{\partial p}{\partial t}+\dfrac{\partial p}{\partial a}=-\mu p, & a>0,\quad t>0,\\[2mm] p(0,t)=B(t)=\displaystyle\int_0^\infty \beta(a)p(a,t)da, & t\ge 0,\\[2mm] p(a,0)=p^0(a), \quad a\ge 0. \end{cases}$$

On the demographic functions we make the hypothesis that there exists a maximum age $a_\dagger$ for the population so that we can restrict our attention to the age interval $[0,a_\dagger]$. We also assume

(1.5) $\beta(a)$ is a nonnegative continuous function on $[0,a_\dagger]$ and $\mu(a)$ is a nonnegative continuous function on $[0,a_\dagger$.

Under assumption (1.5), a steady state solution of (1.4) exists if and only if the *net reproduction rate* is equal to unity:

$$R=\int_0^{a_\dagger}\beta(a)e^{-\int_0^a\mu(\alpha)d\alpha}da=1.$$

In this case the steady state solutions are

$$(1.6)\qquad p(a,t)=p^\infty(a)=b^0e^{-\int_0^a\mu(\alpha)d\alpha},\qquad a\in[0,a_\dagger],$$

where b^0 is an arbitrary constant, representing the number of newborns.

Throughout the paper we shall assume that $R=1$ and that the population has reached its steady state, i.e., $p^0(a)=p^\infty(a)$. It should be noted that this restriction is a severe one in general, and one which is not satisfied by most animal species in our world. A further analysis for time dependent populations will be carried out elsewhere. However, for diseases that have a fairly rapid spread, it is not inadequate to assume the population at a steady state.

Note that in the case we consider here, since (1.6) is the known explicit solution of (1.4), we see from (1.1) that the unknown s can be eliminated in (1.2) to yield a single equation for the infective subpopulation. In this case (1.3) in unnecessary since, once we have solved (1.2) for i, we find s directly from (1.1) and (1.6).

As for the form of the force of infection, Busenberg et al. [3] considered

$$\lambda = \lambda\big(a; i(\cdot, t)\big) = \begin{cases} \kappa(a) i(a,t) & \text{(intracohort)}, \\ \kappa(a) I(t) & \text{(intercohort)}, \end{cases} \tag{1.7}$$

where $I(t) = \int_0^{a_\dagger} i(a,t)da$. They found that in either case there exists a threshold parameter T (the reproductive number of the epidemic) such that for $T \leq 1$ all nonnegative solutions of (1.2) tend to zero as t goes to infinity; for $T > 1$ there exists a unique positive stationary solution of (1.2) (an endemic state for the disease) which is locally asymptotically stable.

Recently, Busenberg et al. [4], [5] have analyzed the more general case

$$\lambda\big(a; i(\cdot,t)\big) = \kappa_0(a) i(a,t) + \int_0^{a_\dagger} \kappa(a,a') i(a',t)\, da'.$$

They proved that, under mild assumptions on $\kappa(a,a')$, the threshold phenomenon always holds. More precisely, they rewrote the equation (1.2) using as variable the fraction of infected individuals in the population

$$u(a,t) = \frac{i(a,t)}{p^\infty(a)}, \quad t \geq 0, \quad a \in [0, a_\dagger]. \tag{1.8}$$

It follows from (1.1), (1.2) (1.6), and (1.8) that u is a solution of the following initial-boundary value problem:

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} + \gamma u = \lambda(1-u), \quad t > 0, \quad a \in [0, a_\dagger], \\ &u(0,t) = \frac{q}{b^0} \int_0^{a_\dagger} \beta(a) p^\infty(a) u(a,t)\, da, \qquad t \geq 0, \\ &u(a,0) = u^0(a) = i^0(a)/p^\infty(a), \qquad a \in [0, a_\dagger]. \end{aligned} \tag{1.9}$$

Note that $0 \leq u^0(a) \leq 1$, $a \geq 0$, and (1.9) ensure that $0 \leq u(a,t) \leq 1$ for $a, t > 0$, as the model requires. The advantage of the formulation for the fraction of infected individuals (1.8) over (1.2)–(1.3) is that the death rate μ does not appear explicitly and, when there is no vertical transmission ($q = 0$), neither does the birth rate β appear in (1.9). On the other hand, (1.6) says that the death rate is given in terms of the population age-density function $p^\infty(a)$ by the relation $\mu(a) = \frac{d}{da}\log\big(b^0/p^\infty(a)\big)$, and thus we see that, both with or without vertical transmission, (1.9) really does involve μ.

In [5] (1.9) was formulated as an abstract semilinear equation in the space $E = L^1(0, a_\dagger)$. Let A and F be defined by

(1.10)
$$\begin{cases} D(A) = \left\{ f \in E \, : \, f \text{ is abs. continuous}, \, f(0) = \dfrac{q}{b^0} \displaystyle\int_0^{a_\dagger} \beta(a) p^\infty(a) f(a) \, da \right\}, \\ Af = -f' \end{cases}$$

and by

$$[F(f)](a) = \lambda(a, f(\cdot))(1 - f(a)) - \gamma(a) f(a). \tag{1.11}$$

If we define

$$C = \{ f \in E : 0 \le f(a) \le 1 \text{ a.e.} \} \tag{1.12}$$

then (1.9) can be written as the following Cauchy problem in the closed convex set C:

$$\begin{cases} \dfrac{d}{dt} u(t) = Au(t) + F(u(t)), \\ u(0) = u_0, \end{cases} \tag{1.13}$$

where $u(t) \equiv u(\cdot, t)$ and $u_0 \equiv u^0(\cdot)$.

Concerning this abstract problem, if we assume the following general conditions,

(1.14) $A \, : \, D_A \subset E \to E$ is the infinitesimal generator of a strongly continuous semigroup e^{tA} such that $e^{tA} C \subset C$.

(1.15) $F : C \to E$ is a Lipschitz continuous function, and there exists $\alpha \in (0,1)$ such that $(I + \alpha F) C \subset C$.

Then, for any $u_0 \in C$, problem (1.13) has a unique mild solution (see [12]), i.e., a solution $u \in C([0,T]; C)$ of the integral equation:

$$u(t) = e^{tA} u_0 + \int_0^t e^{(t-s)A} F(u(s)) \, ds. \tag{1.16}$$

In [4], [5] Busenberg et al. prove that A, F, and C, as defined in (1.10)–(1.12), do satisfy conditions (1.14)–(1.15) (see [4], [5] for the precise assumptions on the functions $\kappa_0(a)$ and $\kappa(a, a')$). Therefore, letting $S(t)u_0$ be the mild solution of (1.13), they prove the following asymptotic result.

THEOREM 1.1. *Let*

$$G = (I - \alpha A)^{-1} (I + \alpha F),$$

where $\alpha \in (0,1)$ is a constant chosen so that $(I + \alpha F)$ is positive. Let ρ be the spectral radius of $DG(0)$, the Gateaux derivative of G at zero with respect to C. Then, if $\rho \le 1$, for each $u_0 \in C$, $S(t)u_0 \to 0$ as $t \to \infty$. If $\rho > 1$, there exists a unique $u_\infty \in C$, $u_\infty \not\equiv 0$ such that $Gu_\infty = u_\infty$; moreover, for each nontrivial u_0, $S(t)u_0 \to u_\infty$ as $t \to \infty$.

Remark 1.2. The theorem does not provide an explicit calculation of ρ. In the next section we give, for a particular choice of $\kappa(a, a')$, a computable threshold condition equivalent to $\rho \le 1$.

2. The inter-intracohort case. Here, we shall restrict ourselves to a combination of the inter- and intracohort cases (1.7). Specifically, we assume

$$\lambda\big(a, i(\cdot,t)\big) = c^1(a)i(a,t) + c^2(a)I(t), \tag{2.1}$$

where

$$I(t) = \int_0^{a_\dagger} i(a,t)\,da$$

and

(2.2) $c^1(a)$ and $c^2(a)$ are nonnegative, continuous functions on $[0, a_\dagger]$.

Busenberg et al. [3] have established an explicit threshold condition for both the intracohort case and the intercohort case without vertical transmission. Here we compute the threshold for the inter-intracohort case with vertical transmission. The result is somewhat analogous to the threshold established in [2] and [7] for epidemics in a heterogeneous but not age-structured population.

THEOREM 2.1. *Let $c^2 \not\equiv 0$. The solution $u \equiv 0$ of* (1.9) *is globally stable if and only if*

$$T_1 = \frac{q}{b^0}\int_0^{a_\dagger} \beta(a)p^\infty(a)\exp\left\{\int_0^a [c^1(\sigma)p^\infty(\sigma) - \gamma(\sigma)]\,d\sigma\right\} da < 1 \tag{2.3}$$

and

$$\begin{aligned} T_2 = {} & \frac{q\int_0^{a_\dagger}\beta(a)p^\infty(a)\int_0^a \exp\{\int_\sigma^a [c^1(\tau)p^\infty(\tau) - \gamma(\tau)]\,d\tau\}c^2(\sigma)\,d\sigma\,da}{b^0 - q\int_0^{a_\dagger}\beta(a)p^\infty(a)\exp\{\int_0^a[c^1(\sigma)p^\infty(\sigma) - \gamma(\sigma)]\,d\sigma\}\,da} \\ & \times \int_0^{a_\dagger} p^\infty(a)\exp\left\{\int_0^a [c^1(\sigma)p^\infty(\sigma) - \gamma(\sigma)]\,d\sigma\right\} da \\ & + \int_0^{a_\dagger} p^\infty(a)\int_0^a \exp\left\{\int_\sigma^a [c^1(\tau)p^\infty(\tau) - \gamma(\tau)]\,d\tau\right\} c^2(\sigma)\,d\sigma\,da \le 1. \end{aligned} \tag{2.4}$$

If the previous condition does not hold, then there exists a unique positive stationary solution of (1.9), *which attracts all nontrivial initial data.*

The case $c^2 \equiv 0$ was studied in [3]. In that case $T_2 = 0$, and the condition for the global stability of $u \equiv 0$ becomes $T_1 \le 1$. When $c^1 \equiv 0$, (2.3) is automatically satisfied as long as $\gamma \not\equiv 0$ or $q < 1$; if $c^1 \equiv 0$ and $q = 0$, (2.4) reduces to the condition given in [3].

Note that the denominator of the first addendum in T_2 is $b^0(1 - T_1)$. Therefore, when $T_1 \ge 1$, T_2 is negative or undefined. In this case we consider T_2 to be undefined, since the condition $T_1 \ge 1$ is enough for establishing the existence of a stationary positive solution.

Proof. Because of Theorem 1.1, it is enough to study ρ, the spectral radius of $DG(0)$. Since $DG(0)$ is linear, completely continuous, and leaves invariant the cone of nonnegative functions, by Krein–Rutman's theorem, there exists an eigenvector $v \ge 0$ with eigenvalue ρ. By definition, v solves the problem

$$\left\{\begin{aligned} & \alpha\rho\frac{d}{da}v(a) = [\alpha\,(c^1(a)p^\infty(a) - \gamma(a)) + (1-\rho)]\,v(a) + \alpha c^2(a)\int_0^{a_\dagger} p^\infty(\tau)v(\tau)\,d\tau, \\ & v(0) = \frac{q}{b^0}\int_0^{a_\dagger} \beta(\tau)p^\infty(\tau)v(\tau)\,d\tau. \end{aligned}\right. \tag{2.5}$$

Let $\eta = \int_0^{a_\dagger} p^\infty(\tau)v(\tau)\,d\tau$. Then (2.5) can be solved explicitly. If we define

$$P(x,y,z) = \exp\left\{\int_y^x \frac{c^1(\tau)p^\infty(\tau)-\gamma(\tau)}{z}\,d\tau + \frac{1}{\alpha}\left(\frac{1}{z}-1\right)(x-y)\right\} \tag{2.6}$$

for $x \geq y \geq 0$ and $z > 0$, then, by integrating the first equation of (2.5), we obtain

$$v(a) = v(0)P(a,0,\rho) + \eta\int_0^a P(a,\sigma,\rho)\frac{c^2(\sigma)}{\rho}\,d\sigma. \tag{2.7}$$

The second equation of (2.5) then implies

$$v(0) = \frac{q}{b^0}v(0)Z(\rho) + \frac{q}{b^0}\eta W(\rho), \tag{2.8}$$

where we set, for $z > 0$,

$$W(z) = \int_0^{a_\dagger} \beta(\tau)p^\infty(\tau)\int_0^\tau P(\tau,\sigma,z)\frac{c^2(\sigma)}{z}\,d\sigma\,d\tau,$$
$$Z(z) = \int_0^{a_\dagger} \beta(\tau)p^\infty(\tau)P(\tau,0,z)\,d\tau.$$

Concerning these functions, we note that, if α is small enough (as it was to ensure the positivity of $I+\alpha F$), we can have $\partial/\partial z P(a,\sigma,z) < 0$ for all $a, z > 0$, $0 \leq \sigma < a$; therefore,

$$\begin{cases} W(z) \quad \text{and} \quad Z(z) \quad \text{are strictly decreasing in } z, \\ \lim_{z\to 0+} Z(z) = +\infty. \end{cases} \tag{2.9}$$

In the following we assume that α is chosen so that (2.9) holds. Since v must be nonnegative, (2.8) implies that ρ must be such that

$$\frac{q}{b^0}Z(\rho) < 1. \tag{2.10}$$

If $q > 0$, let z^* be the solution of $qZ(z^*) = b^0$, which is unique because of (2.9). If $q = 0$ we let $z^* = 0$. Because of (2.10), we have that

$$\rho > z^*.$$

Using (2.8), we finally obtain

$$v(a) = \eta\left[\frac{qW(\rho)}{b^0 - qZ(\rho)}P(a,0,\rho) + \int_0^a P(a,\sigma,\rho)\frac{c^2(\sigma)}{\rho}\,d\sigma\right], \tag{2.11}$$

and, by multiplying both sides of (2.11) by $p^\infty(a)$, integrating from 0 to $a_\dagger$, and using the definition of η, we get:

$$H_\alpha(\rho) = 1, \tag{2.12}$$

where we have defined, for $z > z^*$,

$$H_\alpha(z) = \frac{qW(z)}{b^0 - qZ(z)} \int_0^{a_\dagger} p^\infty(a)P(a,0,z)\,da + \int_0^{a_\dagger} p^\infty(a) \int_0^{a} P(a,\sigma,z)\frac{c^2(\sigma)}{z}\,d\sigma.$$

Therefore, ρ can be found as a solution in $(z^*, +\infty)$ of the scalar equation (2.12). As for the subscript α, we remind that $P(a,\sigma,\rho)$, and therefore ρ, depend on α. It is clear that $H_\alpha(z)$ is a decreasing function of z on $(z^*, +\infty)$, that $\lim_{z\to(z^*)^+} H_\alpha(z) = +\infty$ and $\lim_{z\to+\infty} H_\alpha(z) = 0$. Therefore, we obtain from (2.9), (2.10) and (2.12),

$$\rho \leq 1 \quad \text{if and only if} \quad \frac{q}{b^0}Z(1) < 1 \quad \text{and} \quad H_\alpha(1) \leq 1.$$

This is just the thesis. □

Remark 2.2. Note that, although ρ in general will depend on α, $Z(1)$ and $H_\alpha(1)$ do not depend on it; that is, the threshold condition is, as expected, independent of α.

3. Regularity of solutions. Here we prove some regularity results that we will use in the analysis of the numerical algorithm. A standard assumption that guarantees regularity of the solutions of (1.16) is $F \in C^1(E,E)$. Unfortunately, this assumption does not hold in this case. The restriction of F to $C \cap L^\infty$ is indeed in $C^1(L^\infty, L^\infty)$, but A is not the generator of a C^0-semigroup in L^∞. Our proof of the regularity rests upon another regularity theorem concerning equation (1.16).

THEOREM 3.1. *Let* (1.14) *and* (1.15) *hold. Assume that there exists a space* Y *densely embedded in* E *such that:*

(3.1) $C \subset Y$, *C is closed in* Y.

(3.2) $(I - \epsilon A)^{-1} \in L(Y)$ $\forall\, \epsilon > 0$ *small enough.*

(3.3) $F(C) \subset Y$ *and* F *can be extended to a mapping* $F : Y \to Y$ *such that* $F \in C^1(Y,Y)$.

(3.4) $\forall\, x \in Y$ $DF[x] : Y \to Y$ *can be extended to a linear bounded mapping* $DF[x] \in L(E)$.

(3.5) *The mapping* $x \to DF[x] : C \subset E \to L(E)$ *is strongly continuous in* E. *There exists* $M > 0$ *such that*

$$||DF[x]||_{L(E)} \leq M \quad \forall x \in C.$$

Then, if $u_0 \in D_A \cap C$ *and* $Au_0 \in Y$ *we have:*

$$u(\cdot) \in C^1(0,T;E) \cap C(0,T;D_A),$$
$$u'(t) = Au(t) + F(u(t))$$

Let us first point out the following.

LEMMA 3.2. *Let* (1.14) *hold. Let*

$$A_n = A\left(I - \frac{1}{n}A\right)^{-1} = -nI + n\left(I - \frac{1}{n}A\right)^{-1} \tag{3.6}$$

be the Yosida approximants of A. *Then, for* n *large enough,*

$$e^{tA_n}C \subset C\,.$$

Proof. We first note that, if A is the generator of a C^0-semigroup, the condition

$$e^{tA}C \subset C \quad \text{for all } t > 0$$

is equivalent to the following:

$$(I - \alpha A)^{-1}C \subset C \quad \text{for } \alpha > 0 \text{ small enough.}$$

Take in fact $x \in C$, and recall that

$$(I - \alpha A)^{-1}x = \frac{1}{\alpha}\int_0^\infty e^{-\frac{1}{\alpha}t}e^{tA}x\,dt\,.$$

Since $\frac{1}{\alpha}\int_0^\infty e^{-\frac{1}{\alpha}t}\,dt = 1$, if $e^{tA}x \in C$ for all $t > 0$, $(I - \alpha A)^{-1}x \in C$ also.

Conversely, we have

$$e^{tA}x = \lim_{n\to\infty}(I - \frac{t}{n}A)^{-n}x \quad \forall\, x \in E.$$

Thus if $(I - \alpha A)^{-1}C \subset C$, then also $e^{tA}C \subset C$.

Using this observation, we consider $(I - \alpha A_n)^{-1}x$ for $x \in C$. We have the identity:

$$\begin{aligned}(I - \alpha A_n)^{-1}x &= \frac{(n-A)}{n}\left(I - \frac{\alpha n + 1}{n}A\right)^{-1} x\\ &= \frac{1}{\alpha n + 1}x + \left(1 - \frac{1}{\alpha n + 1}\right)\left(I - \frac{\alpha n + 1}{n}A\right)^{-1} x.\end{aligned}$$

Thus $(I - \alpha A_n)^{-1}x$, being the convex combination of two elements of C (for n large enough), belongs to C. □

Proof of Theorem 3.1. Consider the problem

$$u_n(t) = e^{tA_n}u_0 + \int_0^t e^{(t-s)A_n}F(u_n(s))\,ds. \tag{3.7}$$

Since $A_n \in L(E)$, (3.7) has a solution $u_n(t) \in C^1(0,T;E)$; moreover, $u_n(t) \in C \quad \forall\, t \in [0,T]$, and

$$u_n \to u \quad \text{in } C(0,T;E).$$

Now, by (3.2), A_n can be restricted to a bounded linear operator in $L(Y)$, so that, by (3.1)–(3.3), problem (3.7) can be viewed as a problem in Y. Since $F \in C^1(Y,Y)$, we have

$$u_n(\cdot) \in C^1(0,T;Y).$$

Thus, setting $v_n(t) = u_n'(t)$, (3.7) yields

$$v_n(t) = e^{tA_n}(A_n u_0 + F(u_0)) + \int_0^t e^{(t-s)A_n}DF[u_n(s)]v_n(s)\,ds.$$

Now consider the limit problem in E (see (3.4)):

$$v(t) = e^{tA}(Au_0 + F(u_0)) + \int_0^t e^{(t-s)A}DF[u(s)]v(s)\,ds\,; \tag{3.8}$$

(3.8) has a unique solution $v \in C(0,T;E)$ that can be found, thanks to (3.5), as the limit in $C(0,T;E)$ of the iterates w_n defined by:

$$(3.9) \qquad \begin{cases} w^0(t) = e^{At}(Au_0 + F(u_0)), \\ w^{n+1}(t) = e^{At}(Au_0 + F(u_0)) + \displaystyle\int_0^t e^{A(t-s)}F'(u(s))w^n(s)\,ds. \end{cases}$$

It is easy to show, using (3.4)–(3.5), that

$$v_n \rightarrow v \quad \text{in } C(0,T;E),$$

proving that $u \in C^1(0,T;E)$ and $u'(t) = v(t)$.

Finally, differentiating (3.7) we get

$$u_n'(t) = A_n u_n(t) + F(u_n(t))$$

so that

$$A_n u_n(t) \rightarrow u'(t) - F(u(t)) \quad \text{in } C(0,T;E).$$

Using (3.6), since

$$\left(I - \frac{1}{n}A\right)^{-1} u_n(t) \rightarrow u(t) \quad \text{in } C(0,T;E)$$

and A is a closed operator, this implies

$$u \in C(0,T;D_A); \qquad u'(t) = Au(t) + F(u(t)). \qquad \square$$

We shall make the following assumption on the parameters:

$$(3.10) \qquad \begin{aligned} &\beta(a), \gamma(a), c^0(a), c^1(a), \text{ and } c^2(a) \text{ are Lipschitz functions on } [0,a_\dagger]; \\ &\mu(a)\exp\left\{-\int_0^a \mu(\sigma)\,d\sigma\right\} \text{ is bounded on } [0,a_\dagger). \end{aligned}$$

The first regularity result is the following.

PROPOSITION 3.3. *Let* (1.5), (2.2), *and* (3.10) *hold. Let* u_0 *be a Lipschitz function on* $[0,a_\dagger]$ *such that*

$$u_0(0) = \frac{q}{b^0}\int_0^{a_\dagger} \beta(a)p^\infty(a)u_0(a)\,da.$$

Then, for any $T > 0$,

$$(3.11) \qquad u(t) \in C^1([0,T];L^1(0,a_\dagger)) \cap C([0,T];D(A)),$$

$$(3.12) \qquad u'(t) \in L^\infty(0,a_\dagger), \quad a.e.\ t \in [0,T],$$

and there exists $K > 0$ *such that*

$$(3.13) \qquad \|u'(t)\|_\infty \le K, \quad a.e.\ t \in [0,T].$$

Proof. Equation (3.11) follows from Theorem 3.1. In fact, our assumptions imply that (3.1)–(3.5) are fulfilled with $Y = L^\infty(0, a_\dagger)$. In particular, (3.2) follows easily from the following formula for the operator $(I - \epsilon A)^{-1}$ (see (3.11) in [5]):

$$((I - \epsilon A)^{-1} f)(a) = q \frac{\int_0^{a_\dagger} \beta(a) p^\infty(a) \int_0^a e^{-\frac{1}{\epsilon}(a-s)} f(s)\, ds\, da}{\epsilon \left(b^0 - q \int_0^{a_\dagger} \beta(a) p^\infty(a) e^{-\frac{1}{\epsilon} a}\, da \right)} + \frac{1}{\epsilon} \int_0^a e^{-\frac{1}{\epsilon}(a-s)} f(s)\, ds.$$

To obtain (3.5), start from the expression

$$(DF[f]g)(a) = \lambda(a, g(\cdot))(1 - f(a)) - \lambda(a, f(\cdot)) g(a) - \gamma(a) g(a).$$

We want to prove that, if $f_n \in C$, $f_n \to f$ in $L^1(0, a_\dagger)$, then, for each $g \in L^1(0, a_\dagger)$, we have

$$DF[f_n]g \to DF[f]g \quad \text{in } L^1(0, a_\dagger). \tag{3.14}$$

To obtain (3.14) we compute

$$\begin{aligned} &(DF[f_n]g - DF[f]g)(a) \\ &\quad = 2c^1(a) p^\infty(a) g(a) (f(a) - f_n(a)) \\ &\quad + \int_0^{a_\dagger} c^2(\sigma) p^\infty(\sigma) g(\sigma)\, d\sigma (f(a) - f_n(a)) \\ &\quad + \int_0^{a_\dagger} c^2(\sigma) p^\infty(\sigma) (f(\sigma) - f_n(\sigma))\, d\sigma\, g(a). \end{aligned} \tag{3.15}$$

The L^1 norm of the last two terms is bounded by

$$b^0 ||c^2||_\infty ||g||_1 ||f - f_n||_1$$

As for the other term in the right-hand side of (3.15), suppose that it does not converge to zero in L^1. Then there would exist a constant $c > 0$ and a subsequence $\{f_{n_k}\}$ such that

$$\int_0^{a_\dagger} c^1(a) p^\infty(a) |g(a)| |f(a) - f_{n_k}(a)|\, da \to c$$

as k goes to ∞. Since $f_{n_k} - f$ converges to zero in L^1, there exists a subsequence, still denoted by the same name, such that $f_{n_k} - f$ converges to zero almost everywhere Then we have

$$c^1 p^\infty |g| |f - f_{n_k}| \to 0 \quad \text{a.e.}$$

and

$$c^1 p^\infty |g| |f - f_{n_k}| \le 2b^0 ||c^1||_\infty |g|$$

since f and f_n are in C. Lebesgue's theorem then implies that $c^1 p^\infty |g| |f - f_{n_k}|$ converges to 0 in L^1.

Next, $v(t) = u'(t)$ satisfies the integral equation (3.8) and can be obtained as the limit in $C(0, T; L^1(0, a_\dagger))$ of the iterates w^n defined in (3.9). We want to prove that these iterates satisfy the relations

$$\begin{cases} w^n(t) \in L^\infty(0, a_\dagger), & t \in [0, T], \\ ||w^n(t)||_\infty \le M e^{\omega t}, & \text{a.e. } t \in [0, T], \end{cases} \tag{3.16}$$

where M and ω are suitable positive constants depending on A, F, u_0, and G.

Once (3.16) is proven, (3.12) and (3.13) follow because a closed ball of $L^\infty(0, a_\dagger)$ is closed in $L^1(0, a_\dagger)$. In order to prove (3.16) we need two lemmas.

LEMMA 3.4. *If* $u_0 \in L^\infty(0,a_\dagger)$, *then* $e^{At}u_0 \in L^\infty(0,a_\dagger)$ *and*

$$\|e^{At}u_0\|_\infty \le e^{\bar\beta t}\|u_0\|_\infty,$$

where $\bar\beta = \max_{a\in[0,a_\dagger]}\{\beta(a)\}$.

Proof. The explicit representation of e^{At} is

$$(e^{At}u_0)(a) = \begin{cases} u_0(a-t), & a>t, \\ B(t-a)[u_0], & a<t, \end{cases}$$

where $B(t)[u_0]$ is the solution of the integral equation

$$B(t) = J(t) + \int_0^t K(t-s)B(s)\,ds,$$

with

$$J(t) = \frac{q}{b^0}\int_t^{a_\dagger} \beta(a)p^\infty(a)u_0(a-t)\,da,$$

and

$$K(t) = \frac{q}{b^0}\beta(t)p^\infty(t).$$

Since $|J(t)| \le q\|u_0\|_\infty$ and $|K(t)| \le \bar\beta$, Gronwall's lemma yields

$$|B(t)[u_0]| \le qe^{\bar\beta t}\|u_0\|_\infty.$$

Finally, as $\sup_{a>t}\{|u_0(a-t)|\} \le \|u_0\|_\infty$, the lemma follows. □

LEMMA 3.5. *Let S be any measure space and let $f \in C([a,b];L^1(S))$ be such that*

$$\begin{cases} f(t) \in L^\infty(S), & t\in[a,b], \\ \|f(t)\|_\infty \le g(t), & a.e.\ t\in[a,b], \end{cases}$$

where $g(t)$ is a continuous real function in $[a,b]$. Then,

$$\int_a^b f(t)\,dt \in L^\infty(S), \quad and \quad \left\|\int_a^b f(t)\,dt\right\|_\infty \le \int_a^b g(t)\,dt.$$

Proof. Since $f(t) \in L^1([a,b];L^1(S))$, there exists $\tilde f \in L^1([a,b]\times S)$ such that $f(t) \equiv \tilde f(t,\cdot)$; moreover, $\int_a^b f(t)\,dt \equiv \int_a^b \tilde f(t,\cdot)\,dt$, and the thesis follows easily. □

Now we are ready to prove (3.16). Observing that, by our assumptions, $Au_0 + F(u_0)$ lies in $L^\infty(0,a_\dagger)$, we set

$$M = e^{\bar\beta T}\|Au_0 + F(u_0)\|_\infty, \qquad \omega = e^{\bar\beta T}\sup_{t\in[0,T]}\|F'(u(t))\|_{L(Y)},$$

where C is the convex set defined before. Now (3.16) is clearly true for $n=0$; assuming that it holds for n, Lemmas 3.4 and 3.5 yield

$$\|w^{n+1}(t)\|_\infty \le M + \int_0^t \omega Me^{\omega s}\,ds = Me^{\omega t},$$

and (3.16) is proven. □

As far as the solution of (1.9) is concerned, since $u(t) \equiv u(\cdot,t)$, $u'(t) \equiv u_t(\cdot,t)$, $Au(t) \equiv -u_a(\cdot,t)$, the previous result states that

$$u_t \in L^\infty([0,a_\dagger]\times[0,T]).$$

Consequently, since $u_a = -u_t + \gamma u + \lambda(1-u)$, we also have

$$u_a \in L^\infty\big([0,a_\dagger]\times[0,T]\big).$$

This leads to the following result.

PROPOSITION 3.6. *Under the same assumptions as in Proposition* 3.3, $u(a,t)$ *has bounded derivatives through second order in the characteristic direction* $\tau = \frac{1}{\sqrt{2}}(1,1)$.

Proof. For a and t fixed, set

$$g(s) = u(a+s, t+s).$$

This function has a distributional derivative satisfying

$$g'(s) = A(s)g^2(s) + B(s)g(s) + C(s),$$

where

$$\begin{aligned}A(s) &= -c^1(a+s),\\ B(s) &= c^1(a+s) - c^0(a+s) - c^2(a+s)I(t+s) - \gamma(a+s),\\ C(s) &= c^0(a+s) + c^2(a+s)I(t+s).\end{aligned}$$

Since A, B and C are differentiable, then $g(s)$ is two times differentiable; moreover, since $A(0)$, $B(0)$, $C(0)$, $A'(0)$, $B'(0)$, and $C'(0)$ belong to $L^\infty([0,a_\dagger] \times [0,T])$, so do $g'(0) = \partial u/\partial\tau$ and $g''(0) = \partial^2 u/\partial\tau^2$, and the proof is complete. □

4. A numerical algorithm. We shall now describe an algorithm for the approximation of the solution of (1.9), based on a first-order implicit finite difference method along the characteristics. Higher-order methods that require more regularity of the solution (hence, more compatibility conditions on the data: see [11]) need a more sophisticated analysis, which we intend to carry out in the future. The scheme analyzed here has a special interest because it preserves, for any time step of the discretization, many properties of the continuous system, as discussed in §5.

Let $\Delta t > 0$ be the age-time discretization parameter. We shall find an approximation U_j^n of $u(j\Delta t, n\Delta t)$, $n \geq 0$, $0 \leq j \leq A_\dagger = \left[\frac{a_\dagger}{\Delta t} + 0.5\right]$ by the finite difference method of characteristics as follows:

$$\begin{aligned}&\frac{U_j^n - U_{j-1}^{n-1}}{\Delta t} + \gamma_j U_j^n = \Lambda_{j-1}^{n-1}(1-U_j^n), \qquad n \geq 1,\ 1 \leq j \leq A_\dagger,\\ &U_0^n = \frac{q}{B_0}\sum_{j=1}^{A_\dagger} \beta_j p_j^\infty U_j^{n-1}\Delta t, \qquad n > 0,\\ &U_j^0 = u_j^0, \qquad 0 \leq j \leq A_\dagger,\end{aligned} \tag{4.1}$$

where we have used the notation $f_j = f(j\Delta t)$ for any function $f = f(a)$, and where

$$\Lambda_j^n = c_j^0 + c_j^1 p_j^\infty U_j^n + c_j^2 \sum_{k=1}^{A_\dagger} p_k^\infty U_k^n \Delta t \tag{4.2}$$

is the discrete transmission rate, and

$$B_0 = \sum_{k=1}^{A_\dagger} \beta_k p_k^\infty \Delta t \tag{4.3}$$

is the discrete analogue of the newborn count

$$b^0 = p^\infty(0) = \int_0^{a_\dagger} \beta(a) p^\infty(a) da.$$

This algorithm is an adaptation of the one used in [8].

We can prove that, without any restrictions on Δt, if the initial datum is between zero and one, the numerical solution (just as the real solution) stays between zero and one. Since u represents a ratio, this is a necessary constraint for the model to make sense.

PROPOSITION 4.1. *If $0 \le u_0(a) \le 1$, then, there exists $K > 0$ such that, for all $\Delta t > 0$ and all $n \ge 0$, $0 \le j \le A_\dagger$, we have*

$$0 \le U_j^n \le 1, \tag{4.4}$$

and

$$0 \le \Lambda_j^n \le K. \tag{4.5}$$

Proof. Note that (4.1) defines U_j^n explicitly, for $j \ge 1$, as

$$U_j^n = \left(\Lambda_{j-1}^{n-1} \Delta t + U_{j-1}^{n-1}\right) / \left[1 + \Delta t(\gamma_j + \Lambda_{j-1}^{n-1})\right]. \tag{4.6}$$

Hence, $0 \le U_{j-1}^{n-1} \le 1$ implies that $0 \le U_j^n \le 1$ for $j \ge 1$ because $\gamma \ge 0$. Moreover, $0 \le U_j^{n-1} \le 1$ for all $j \ge 1$, by (4.3), implies that $0 \le U_0^n \le 1$. This proves (4.4) inductively. Finally, (4.5) follows immediately, with

$$K = \max_{0 \le k \le 2;\, j \ge 0} \{c_j^k\} \left[2 + \int_0^\infty p^\infty(a)\, da\right],$$

since $p^\infty(a) = b^0 \exp\{-\int_0^a \mu(\sigma)\, d\sigma\}$ is a nonincreasing function. □

Using the regularity results obtained in §3, we can prove that the discrete function U defined by (4.1)–(4.3) converges uniformly to the solution u of (1.8) at a first-order rate.

THEOREM 4.2. *Let* (1.5), (2.2), *and* (3.1) *hold. Then, for each $T > 0$, there exists a constant K, independent of Δt, such that, for $u_j^n = u(j\Delta t, n\Delta t)$, $0 \le j \le A_\dagger$, $0 \le n \le N = \left[\frac{T}{\Delta t}\right]$,*

$$|u_j^n - U_j^n| \le K\Delta t.$$

Proof. Note that Taylor's theorem implies that

$$\begin{aligned} \frac{u_j^n - u_{j-1}^{n-1}}{\Delta t} &= \sqrt{2}\left(\frac{\partial u}{\partial \tau}\right)_j^n + O\left(\Delta t \left\|\frac{\partial^2 u}{\partial \tau^2}\right\|_{L^\infty\left([0,a_\dagger]\times[0,T]\right)}\right) \\ &= \frac{\partial u}{\partial t}(j\Delta t, n\Delta t) + \frac{\partial u}{\partial a}(j\Delta t, n\Delta t) + O(\Delta t). \end{aligned} \tag{4.7}$$

If we set

$$\zeta_j^n = u_j^n - U_j^n, \qquad 0 \le j \le A_\dagger,\ 0 \le n, \tag{4.8}$$

we have, from (4.6)–(4.8), the error equations

$$\frac{\zeta_j^n - \zeta_{j-1}^{n-1}}{\Delta t} + \gamma_j \zeta_j^n = -\Lambda_{j-1}^{n-1}\zeta_j^n + \left(\lambda_j^n - \Lambda_{j-1}^{n-1}\right)(1 - u_j^n) + O(\Delta t), \qquad 1 \le j \le A_\dagger, 1 \le n, \tag{4.9}$$

$$\begin{aligned} \zeta_0^n &= \frac{q}{b_0}\left[\int_0^{a_\dagger} \beta(a)p^\infty(a)u(a,n\Delta t)\,da - \sum_{j=1}^{A_\dagger} \beta_j p_j^\infty U_j^{n-1}\Delta t\right] \\ &\quad + \frac{q(B_0 - b_0)}{b_0 B_0}\sum_{j=1}^{A_\dagger} \beta_j p_j^\infty U_j^{n-1}\Delta t, \qquad n > 0. \end{aligned} \tag{4.10}$$

It is also clear that, if f is a Lipschitz function,

$$\left|\int_0^{a_\dagger} f(a)\,da - \sum_{j=1}^{A_\dagger} f_j\,\Delta t\right| \le M\Delta t, \tag{4.11}$$

where M is the Lipschitz constant of f. Using (4.11) in the integral in (4.10) and in the definition of b_0, we obtain

$$\begin{aligned} \zeta_0^n &= \frac{q}{b_0}\sum_{j=1}^{A_\dagger} \beta_j p_j^\infty \zeta_j^{n-1}\Delta t + \sum_{j=1}^{A_\dagger} \beta_j p_j^\infty \left[u_j^n - u_j^{n-1}\right]\Delta t + O(\Delta t) \\ &= \frac{q}{b_0}\sum_{j=1}^{A_\dagger} \beta_j p_j^\infty \zeta_j^{n-1}\Delta t + O(\Delta t) \\ &\le C\left(\|\zeta^{n-1}\|_{l^1} + \Delta t\right), \end{aligned} \tag{4.12}$$

where $\|\zeta^n\|_{l^1} = \sum_{j=0}^{A_\dagger} |\zeta_j^n|\Delta t$, and the constant C depends on the Lipschitz constants of β and p^∞ and on the bounds for $|u_t|$ and $|u_a|$. Using (4.11) again, we see that

$$|\lambda_j^n - \Lambda_{j-1}^{n-1}| \le \tilde{C}\left(\|\zeta^{n-1}\|_{l^1} + |\zeta_{j-1}^{n-1}| + \Delta t\right), \tag{4.13}$$

where $\tilde{C}$ depends on the Lipschitz constants of c^0, c^1, c^2, and p^∞ and on the bounds for $|u_t|$ and $|u_a|$.

Using (4.9) and (4.13), we have that, modifying perhaps the constant $\tilde{C}$,

$$\begin{aligned} |\zeta_j^n| &\le |\zeta_j^n|\left(1 + (\gamma_j + \Lambda_{j-1}^{n-1})\Delta t\right) \\ &\le (1 + \tilde{C}\Delta t)|\zeta_{j-1}^{n-1}| + \tilde{C}\Delta t\|\zeta^{n-1}\|_{l^1} + \tilde{C}(\Delta t)^2, \quad 1 \le j \le A_\dagger, \quad 1 \le n. \end{aligned} \tag{4.14}$$

Multiplying (4.14) by Δt and summing on j, and adding to the resulting relation (4.12) multiplied by Δt, we obtain

$$\|\zeta^n\|_{l^1} \le (1 + K\Delta t)\|\zeta^{n-1}\|_{l^1} + K(\Delta t)^2, \tag{4.15}$$

for some constant $K > 0$. Gronwall's lemma applied to (4.15), together with $\zeta_j^0 = 0$, then imply that

$$\|\zeta^n\|_{l^1} \leq (e^{nK\Delta t} - 1)\,\Delta t,$$

and, therefore, for any $T > 0$, there exists $\tilde{K}$ such that

$$\|\zeta^n\|_{l^1} \leq \tilde{K}\Delta t \tag{4.16}$$

for all $n \leq T/\Delta t$.

Using (4.16) in (4.12), we directly obtain

$$|\zeta_0^n| \leq \tilde{K}\Delta t. \tag{4.17}$$

Substituting (4.16) in (4.14) we have

$$|\zeta_j^n| \leq (1 + \tilde{K}\Delta t)|\zeta_{j-1}^{n-1}| + \tilde{K}(\Delta t)^2, \qquad 1 \leq j \leq A_\dagger\ , 1 \leq n. \tag{4.18}$$

Applying Gronwall's lemma to (4.18), we have

$$|\zeta_j^n| \leq \begin{cases} e^{jc\Delta t}|\zeta_0^{n-j}| + (e^{jc\Delta t} - 1)\,\Delta t, & n \geq j, \\ e^{nc\Delta t}|\zeta_{j-n}^0| + (e^{nc\Delta t} - 1)\,\Delta t, & n \leq j. \end{cases} \tag{4.19}$$

Using (4.17) and $\zeta_j^0 = 0$ in (4.19), we have the thesis. □

5. The discrete dynamical system. Here we want to point out that the algorithm illustrated in the previous section itself defines a (discrete) dynamical system that inherits the behaviour of the originating continuous flow.

To this purpose, and in view of (4.1) we consider the mapping

$$F : \mathbf{R}^{A_\dagger + 1} \to \mathbf{R}^{A_\dagger + 1},$$

defined (denoting $x \in \mathbf{R}^{A_\dagger+1}$ as $x \equiv (x_0, x_1, \cdots, x_{A_\dagger})$) as follows

$$\begin{cases} F_0(x) = \dfrac{q}{B_0} \displaystyle\sum_{i=1}^{A_\dagger} \beta_i p_i^\infty x_i \Delta t \\ F_i(x) = \dfrac{x_{i-1} + \Lambda_{i-1}(x)\Delta t}{1 + (\gamma_i + \Lambda_{i-1}(x))\Delta t} \qquad i = 1, \ldots A_\dagger, \end{cases} \tag{5.1}$$

where B_0 is defined in (4.3) and

$$\Lambda_i(x) = c_i^1 p_i^\infty x_i + c_i^2 \sum_{k=1}^{A_\dagger} p_k^\infty x_k \Delta t. \tag{5.2}$$

We shall assume

(H) $q > 0$, and there exists $i = 0 \ldots A_\dagger - 1$ such that $c_i^1 + c_i^2 > 0$.

Also recall that β_i cannot vanish for all index $i > 0$ because the net reproduction rate is equal to 1.

We restrict F to the compact set

$$C = \left\{ x \in \mathbf{R}^{A_\dagger+1} : 0 \leq x_i \leq 1, \quad i = 0, 1, \dots A_\dagger \right\}.$$

In fact, C is left invariant by F; is essentially proven in Proposition 4.1.

We now enumerate several properties of F to be used later in the section.

First note that F is continuous in C and hence also bounded. Moreover, $F(0) = 0$. Next note that F is monotone nondecreasing, with respect to the usual componentwise partial ordering of $\mathbf{R}^{A_\dagger+1}$. Indeed, for $0 \leq j \leq A_\dagger$ and $0 < i \leq A_\dagger$ we have:

$$\frac{\partial}{\partial x_j} F_0(x) = (1 - \delta_{0,j}) \frac{q}{B_0} \beta_j p_j^\infty \Delta t,$$

$$\frac{\partial}{\partial x_j} F_i(x)$$

$$= \frac{\delta_{i-1,j}}{1 + (\gamma_i + \Lambda_{i-1}(x))\Delta t} + \frac{(1 + \Delta t \gamma_i - x_{i-1}) \left[c_{i-1}^1 \delta_{i-1,j} + c_{i-1}^2 (1 - \delta_{0,j}) p_j^\infty \Delta t \right] \Delta t}{\left[1 + (\gamma_i + \Lambda_{i-1}(x))\Delta t \right]^2},$$

both of which are nonnegative for $x \in C$.

We finally consider the Jacobian matrix $F'(0) \equiv (\alpha_{ij})$:

$$\alpha_{0,j} = (1 - \delta_{0,j}) \frac{q}{B_0} \beta_j p_j^\infty \Delta t, \tag{5.3}$$

$$\alpha_{i,j} = \frac{(1 + c_{i-1}^1 p_{i-1}^\infty \Delta t) \delta_{i-1,j} + p_j^\infty (1 - \delta_{0,j}) c_{i-1}^2 (\Delta t)^2}{(1 + \gamma_i \Delta t)}, \tag{5.4}$$

$(1 \leq i \leq A_\dagger,\ 0 \leq j \leq A_\dagger)$ and discuss its irreducibility.

For this purpose, we will write $P_i \longrightarrow P_j$ if $\alpha_{i,j} > 0$; then, a path from i to j is a sequence of indices $(i_0 = i, i_1, \cdots, i_n = j)$ such that $P_{i_k} \longrightarrow P_{i_{k+1}}$ for all $k = 0 \cdots n-1$. We remind the reader that the irreducibility of $F'(0)$ can be stated as the existence of a path from i to j for all couples (i,j). Now note that for $i = 1 \cdots A_\dagger$ we have $\alpha_{i,i-1} > 0$. Consequently, the connectivity graph of the matrix α contains at least the path

$$P_{A_\dagger} \longrightarrow \cdots \longrightarrow P_1 \longrightarrow P_0. \tag{5.5}$$

It follows that, in order to have $F'(0)$ irreducible, it is necessary and sufficient that there exists a path from P_0 to $P_{A_\dagger}$. It is then necessary to have $q > 0$ and $\beta_i > 0$ for some index i, as assumed in (H). To have a sufficient condition, consider

$$m = \max\{i = 1 \dots A_\dagger : \beta_i > 0\}$$

(the largest reproductive age); it is then enough to have:

(5.6) There exists $h < m$ such that $c_h^2 > 0$.

In fact, since $\beta_m > 0$, we have $\alpha_{0,m} > 0$ and, in view of (5.4), $\alpha_{h+1,A_\dagger} > 0$; then (5.6) means that we have the path:

$$P_0 \longrightarrow P_m \longrightarrow \cdots P_{h+1} \longrightarrow P_{A_\dagger}. \tag{5.7}$$

Equations (5.5) and (5.7) yield the irreducibility of $F'(0)$.

Condition (5.6) means that some age class below the maximum fertility age can be infected through intercohort transmission. Note that this corresponds exactly to the restriction imposed in [5] for the continuous case.

Furthermore, note that neither the case without vertical transmission nor the case of pure intracohort transmission result in an irreducible $F'(0)$, as F is defined in (5.1). However, in the case without vertical transmission, we may define

$$F \equiv (F_1, \cdots, F_{A_\dagger}) : \mathbf{R}^{A_\dagger} \to \mathbf{R}^{A_\dagger}.$$

In the pure intracohort case, we may define

$$F \equiv (F_0, \cdots, F_m) : \mathbf{R}^{m+1} \to \mathbf{R}^{m+1},$$

where m is, as above, the maximum reproductive age. In fact, it is clear that ages beyond m do not contribute to disease transmission to ages below m.

All the following considerations can be easily adjusted to these cases. For the sake of simplicity we will instead assume (H) and (5.6) for the rest of the section.

Now we define on C the discrete dynamical system:

$$\left\{\begin{array}{l} U^{n+1} = F(U^n) \\ U^0 \in C \end{array}\right. \tag{5.8}$$

for which the following holds.

THEOREM 5.1. *Let F be defined in* (5.1), *and let* (H) *and* (5.6) *hold. Then:*

(i) *If $\rho(F'(0)) \leq 1$, F has no nontrivial fixed points in C and $U^n \overset{n\to\infty}{\longrightarrow} 0$, for all $U^0 \in C$;*

(ii) *If $\rho(F'(0)) > 1$, F has one nontrivial fixed point $U^\infty \in C$; U^∞ is strictly positive, and we have $U^n \overset{n\to\infty}{\longrightarrow} U^\infty$ for all $U^0 \in C$, $U^0 \not\equiv 0$.*

The proof of this theorem is, for the main part, contained in Theorem 2.1 of [9]; however, we cannot rely completely upon that theorem because our mapping F does not satisfy a strict sublinearity condition needed there to prove uniqueness of the positive fixed point. Thus, to prove Theorem 5.1, we preliminarily prove uniqueness in an independent way.

PROPOSITION 5.2. *Under the assumptions of Theorem* 5.1, *F has at most one positive fixed point.*

Proof. Let $\bar{x} \neq \bar{y}$ be two positive fixed points of F. Without loss of generality we can assume $\bar{y} \not\geq \bar{x}$, so that, since $\bar{x}$ and $\bar{y}$ are strictly positive, it is possible to find the maximal $\xi \in (0,1)$ such that $\bar{y} \geq \xi\bar{x}$. Let

$$k = \max\{i = 0 \ldots A_\dagger : \bar{y}_i = \xi\bar{x}_i\}. \tag{5.9}$$

We will consider three cases, showing that each gives rise to a contradiction:

(a) $k > 0$ and $c^1_{k-1} + c^2_{k-1} > 0$.

(b) $k > 0$ and there exists h, $0 < h < k$ such that $c^1_{h-1} + c^2_{h-1} > 0$ and $c^1_i + c^2_i = 0$ for all $i = h \ldots k-1$.

(c) $k \geq 0$ and $c^1_{i-1} + c^2_{i-1} = 0$ for all i, $1 \leq i \leq k$.

Assume case (a). Then we have $\Lambda_{k-1}(\bar{x}) > 0$. The inequality $F_k(\bar{y}) \geq F_k(\xi\bar{x})$ gives:

$$\bar{y}_k \geq \frac{\xi\bar{x}_{k-1} + \xi\Lambda_{k-1}(\bar{x})\Delta t}{1 + (\gamma_k + \xi\Lambda_{k-1}(\bar{x}))\Delta t},$$

that is,

$$[1+(\gamma_k+\Lambda_{k-1}(\bar{x}))\Delta t]\,\bar{y}_k \geq \xi(\bar{x}_{k-1}+\Lambda_{k-1}(\bar{x})\Delta t)+(1-\xi)\bar{y}_k\Lambda_{k-1}(\bar{x})\Delta t.$$

Hence,

$$\bar{y}_k \geq \xi F_k(\bar{x})+\xi\bar{x}_k\frac{(1-\xi)\Lambda_{k-1}(\bar{x})\Delta t}{1+(\gamma_k+\xi\Lambda_{k-1}(\bar{x}))\Delta t},$$

and so,

$$\bar{y}_k \geq \xi\bar{x}_k\left[1+\frac{\Lambda_{k-1}(\bar{x})\Delta t}{1+(\gamma_k+\xi\Lambda_{k-1}(\bar{x}))\Delta t}\right] > \xi\bar{x}_k$$

in contradiction with (5.9).

In case (b), we note that it implies, for $h \leq i < k$, $\Lambda_i(x)=0\ \forall\ x \in C$; consequently, if $\bar{y}_{i+1}=\xi\bar{x}_{i+1}$, from

$$\text{(5.10)}\qquad \begin{cases} \bar{y}_{i+1}=F_{i+1}(\bar{y})=\dfrac{\bar{y}_i}{1+\gamma_{i+1}\Delta t} \\ \xi\bar{x}_{i+1}=\xi F_{i+1}(\bar{y})=\xi\dfrac{\bar{x}_i}{1+\gamma_{i+1}\Delta t} \end{cases}$$

we obtain

$$\text{(5.11)}\qquad \bar{y}_i=\xi\bar{x}_i$$

Applying this iteratively for $i=k-1\ldots h$, since $\bar{y}_k=\xi\bar{x}_k$, one gets:

$$\bar{y}_h=\xi\bar{x}_h \quad\text{and}\quad c^1_{h-1}+c^2_{h-1}>0.$$

We can therefore apply the argument of case (a).

In case (c), we have $\Lambda_i(x)=0$ for all $i=0\cdots k-1$; therefore, we can apply (5.10) iteratively for $i=k-1$ to $i=0$ to obtain

$$\text{(5.12)}\qquad \bar{y}_0=\xi\bar{x}_0.$$

Moreover, comparing (5.6) with (c), we see that $k<m$, that is,

$$\text{(5.13)}\qquad \bar{y}_m>\xi\bar{x}_m.$$

Then, we have

$$\text{(5.14)}\qquad \bar{y}_0=F_0(\bar{y})=\frac{q}{B_0}\sum_{i=1}^{A_\dagger}\beta_i p_i^\infty\bar{y}_i\Delta t>\frac{q}{B_0}\sum_{i=1}^{A_\dagger}\beta_i p_i^\infty\xi\bar{x}_i\Delta t=\xi F_0(\bar{x})=\xi\bar{x}_0,$$

contradicting (5.12). The strict inequality in (5.14) comes from (5.13), since $\beta_m>0$. □

Once uniqueness is proved, we use the following reduced version of Theorem 2.1 of [9].

THEOREM 5.3. *Let $\Phi(x)$ be a continuous, monotone, nondecreasing function that maps $[0,1]^n$ into itself. Assume that $\Phi(0) = 0$ and $\Phi'(0)$ exists and is irreducible. Suppose also that Φ is sublinear, i.e., if $x \in [0,1]^n$, then*

$$\Phi(\xi x) \geq \xi \Phi(x) \quad \forall\, \xi \in (0,1). \tag{5.15}$$

Then any nonzero fixed point of Φ is strictly positive. Moreover, if $\rho(\Phi'(0)) > 1$, Φ has a positive fixed point. Vice versa, if Φ has a positive fixed point, $\rho(\Phi'(0)) \geq 1$; if $\bar{x}$ is a positive fixed point and $\rho(\Phi'(0)) = 1$, then

$$\Phi'(0)\bar{x} = \bar{x}\,. \tag{5.16}$$

Proof. The proof of this theorem is essentially the same as that of Theorem 2.1 in [9]: here condition (5.15) replaces strict sublinearity.

The proof of existence does not require sublinearity. In order to take care of the fact that our domain is $[0,1]^n$ instead of $\mathbf{R}^n_+$, we have to note that, if $x \in [0,1]^n$ is such that $x_i = 1$, then $\Phi_i(x) \leq x_i$.

Finally, if Φ has a positive fixed point $\bar{x}$, we can apply the argument of Hethcote and Thieme to $\bar{x}$ and obtain

$$\Phi'(0)\bar{x} \geq \bar{x}. \tag{5.17}$$

Since $\bar{x}$ is positive, (5.17) implies, by Perron–Frobenius theory, that $\rho(\Phi'(0)) \geq 1$. If $\rho(\Phi'(0)) = 1$ (5.17) implies that $\bar{x}$ is an eigenvector of $\Phi'(0)$, i.e., (5.16). □

To apply this theorem to the proof of Theorem 5.1, we need to check that F satisfies (5.15), the other properties having already been established. From (5.1) we have

$$F_0(\xi \bar{x}) = \xi \frac{q}{B_0} \sum_{i=1}^{A_\dagger} \beta_i p_i^\infty \bar{x}_i \Delta t = \xi F_0(\bar{x}),$$

and, for $1 \leq i \leq A_\dagger$,

$$F_i(\xi \bar{x}) = \frac{\xi \bar{x}_{i-1} + \xi \Lambda_{i-1}(\bar{x})\Delta t}{1 + (\gamma_i + \xi \Lambda_{i-1}(\bar{x}))\Delta t} \geq \xi \frac{\bar{x}_{i-1} + \Lambda_{i-1}(\bar{x})\Delta t}{1 + (\gamma_i + \Lambda_{i-1}(\bar{x}))\Delta t} = \xi F_i(\bar{x}). \tag{5.18}$$

To exclude the case that there exists a positive fixed point $\bar{x}$ of (5.1) if $\rho(F'(0)) = 1$, we note that in this case (5.16) would hold, i.e., $F(\bar{x}) = F'(0)\bar{x}$. This means (see (5.3)–(5.4))

$$\frac{\bar{x}_{i-1} + \Lambda_{i-1}(\bar{x})\Delta t}{1 + (\gamma_i + \Lambda_{i-1}(\bar{x}))\Delta t} = \frac{\bar{x}_{i-1} + \Lambda_{i-1}(\bar{x})\Delta t}{1 + \gamma_i \Delta t} \tag{5.19}$$

for all $i = 1 \cdots A_\dagger$. Since $\bar{x}_i > 0$, (5.19) implies $\Lambda_{i-1}(\bar{x}) = 0$ for all $i = 1 \cdots A_\dagger$. Assumption (H) then yields $\bar{x} = 0$, in contradiction with its being positive.

Finally, as for the asymptotic behaviour of U^n, we note that it follows from Krasnoselskii's theory (Theorem 6.6 in [10]; see also [9]) that any $U^0 \in C$, $U^0 \neq 0$, will converge to the positive fixed point (when it exists), as long as there exists $N \geq 0$ such that $F^N(U^0)$ is strictly positive. This condition comes immediately from (5.1) and (5.6). □

Proceeding exactly as in §2, we can compute the threshold condition. More precisely, we have the following.

PROPOSITION 5.4. *The equilibrium $U \equiv 0$ is globally stable if and only if*

$$T_1(\Delta t) = \frac{q}{B_0} \sum_{i=1}^{A_\dagger} \beta_i p_i^\infty P_{i,0} \Delta t < 1, \tag{5.20}$$

and

$$T_2(\Delta t) = \sum_{i=1}^{A_\dagger} \left(\frac{qA\beta_i}{B_0(1 - T_1)} + 1 \right) p_i^\infty \left(\sum_{j=0}^{i-1} \frac{c_j^2 \Delta t}{1 + \gamma_{j+1}\Delta t} P_{i,j+1} \right) \Delta t \leq 1, \tag{5.21}$$

where

$$P_{i,j} = \prod_{k=j}^{i-1} \frac{1 + c_k^1 p_k^\infty \Delta t}{1 + \gamma_{k+1}\Delta t}$$

and

$$A = \sum_{i=1}^{A_\dagger} p_i^\infty P_{i,0} \Delta t.$$

Comparing Proposition 5.4 with Theorem 2.1, we realize that the threshold condition for the discrete dynamical system approaches that of the original continuous one as the time step tends to zero.

COROLLARY 5.5. *If Δt is small enough, there exists $C > 0$ such that*

$$|T_1(\Delta t) - T_1|, \qquad |T_2(\Delta t) - T_2| \leq C\Delta t. \tag{5.22}$$

Proof. It follows from the expressions (5.20)–(5.21) and (2.3)–(2.4), noting that, if $f \in L^\infty(0,a)$, $f \geq 0$, we have

$$0 \leq \exp\left\{ \sum_{k=j}^{i-1} f_k \Delta t \right\} - \prod_{k=j}^{i-1} (1 + f_k \Delta t) \leq \exp\left\{ \sum_{k=j}^{i-1} f_k \Delta t \right\} \left(e^{\|f\|_\infty^2 (\Delta t)^2} - 1 \right). \tag{5.23}$$

Equation (5.22) then follows from (5.23) and the approximation of integrals with Riemann sums. □

We finally prove that the endemic equilibrium U^∞ of the discrete dynamical system approaches the endemic equilibrium u^∞ of the continuous system as the time step Δt goes to zero.

For $h = \Delta t$, let

$$u_h^\infty(a) = \left(\left[\frac{a+h}{h} \right] - \frac{a}{h} \right) U_{\left[\frac{a}{h}\right]}^\infty + \left(\frac{a}{h} - \left[\frac{a}{h} \right] \right) U_{\left[\frac{a+h}{h}\right]}^\infty$$

i.e., u_h^∞ is the linear spline through U^∞.

PROPOSITION 5.5. *If $T_1 \geq 1$, or $T_2 > 1$, u_h^∞ converges to u^∞ uniformly in $[0, a_\dagger]$.*

Proof. We first note that u_h^∞ are everywhere differentiable from the right and from the left with

(5.24)

$$D^+ u_h^\infty(a) = \left[c^1 \left(\left[\frac{a}{h}\right] h \right) p^\infty \left(\left[\frac{a}{h}\right] h \right) u_h^\infty \left(\left[\frac{a}{h}\right] h \right) + c^2 \left(\left[\frac{a}{h}\right] h \right) \sum_{k=1}^{A_\dagger} p^\infty(kh) u_h^\infty(kh) h \right]$$
$$\times \left(1 - u_h^\infty \left(\left[\frac{a+h}{h} \right] h \right)\right) - \gamma \left(\left[\frac{a+h}{h} \right] h \right) u_h^\infty \left(\left[\frac{a+h}{h} \right] h \right)$$

and $D^-u_h^\infty(a)$ equal to $D^+u_h^\infty(a)$, except when $a/h \in \mathbf{N}$, in which case we substitute a/h in (5.24) with $(a-h)/h$, and $(a+h)/h$ with a/h.

From (5.24) it also comes out that u_h are Lipschitzian with common Lipschitz constant

$$L = \max\left\{ b^0\|c^1\|_\infty + \|c^2\|_\infty \int_0^{a_\dagger} p^\infty(a)\,da\,,\ \|\gamma\|_\infty \right\}.$$

Take a sequence $\{h_n\}$, $h_n \overset{n\to\infty}{\longrightarrow} 0$. Since $\{u_{h_n}^\infty\}$ are equi-Lipschitz (and so equicontinuous) and uniformly bounded, by Ascoli–Arzelà theorem, there exists a subsequence, still denoted by the same name, uniformly converging to $\bar{u}$.

Then, from (5.24), we see that both $D^+u_{h_n}^\infty$ and $D^-u_{h_n}^\infty$ converge uniformly (see Assumption 3.8) to

$$\left[c^1(a)p^\infty(a)\bar{u}(a) + c^2(a)\int_0^{a_\dagger} p^\infty(\sigma)\bar{u}(\sigma)\,d\sigma\right](1-\bar{u}(a)) - \gamma(a)\bar{u}(a), \tag{5.25}$$

and therefore $\bar{u}$ is differentiable, and its derivative is equal to (5.25).

Considering also the conditions on $u_h^\infty(0)$, we obtain that $G\bar{u} = \bar{u}$, with G as in Theorem 1.1. By Theorem 1.1 there exists a unique positive function u^∞ such that $Gu^\infty = u^\infty$. The only other possibility would be to have $\bar{u} = 0$, but this will be excluded below.

We have therefore proved that, from any sequence $\{h_n\}$, $h_n \overset{n\to\infty}{\longrightarrow} 0$, we can extract a subsequence uniformly converging to u^∞, proving, therefore, the convergence as h goes to zero.

We have still to exclude the case that there exists a sequence $\{h_n\}$, $h_n \overset{n\to\infty}{\longrightarrow} 0$, and $u_{h_n}^\infty \overset{n\to\infty}{\longrightarrow} 0$.

We find a lower bound, when Δt is sufficiently small, for U^∞ yielded by Theorem 5.1, using the fact, arising from the Proof of Theorem 2.1 of [9], that, if $\bar{v}$ is such that $F(\bar{v}) \geq \bar{v}$, then $U^\infty \geq \bar{v}$.

If $T_1 > 1$, there exists $z < 1$ and $h_0 > 0$ such that, for $\Delta t < h_0$,

$$T_{1,z}(\Delta t) = \frac{q}{B_0}\sum_{i=1}^{A_\dagger} \beta_i p_i^\infty P_{i,0} z^i \Delta t > 1,$$

with $P_{i,0}$ as in Proposition 5.4.

Now take $\bar{v}_0 = \epsilon$, $\bar{v}_i = \epsilon P_{i,0} z^i$, with ϵ such that $z\Lambda_{i-1}(\bar{v})\Delta t < 1 - z$ for all $i = 1 \ldots A_\dagger$.

Then, we have

$$F_0(\bar{v}) - \bar{v}_0 = \epsilon\left[\frac{q}{B_0}\sum_{i=1}^{A_\dagger} \beta_i p_i^\infty P_{i,0} z^i \Delta t - 1\right] > 0,$$

$$\begin{aligned}F_i(\bar{v}) - \bar{v}_i =& \frac{\bar{v}_{i-1}(1 + c_{i-1}^1 p_{i-1}^\infty \Delta t) + c_{i-1}^2 \Delta t \sum_{j=1}^{A_\dagger} p_j^\infty \bar{v}_j \Delta t}{1 + \gamma_i \Delta t + \Lambda_{i-1}(\bar{v})\Delta t} - z\frac{1 + c_{i-1}^1 p_{i-1}^\infty \Delta t}{1 + \gamma_i \Delta t}\bar{v}_{i-1} \\ =& \frac{\bar{v}_{i-1}(1 + c_{i-1}^1 p_{i-1}^\infty \Delta t)\left[(1+\gamma_i\Delta t)(1-z) - z\Lambda_{i-1}(\bar{v})\Delta t\right]}{(1 + \gamma_i\Delta t + \Lambda_{i-1}(\bar{v})\Delta t)(1+\gamma_i\Delta t)} \\ &+ \frac{c_{i-1}^2 \Delta t \sum_{j=1}^{A_\dagger} p_j^\infty \bar{v}_j \Delta t}{1 + \gamma_i\Delta t + \Lambda_{i-1}(\bar{v})\Delta t}.\end{aligned}$$

Therefore, we have $F(\bar{v}) \geq \bar{v}$ and so $U_0^\infty \geq \epsilon$, and $u_h^\infty(0) \geq \epsilon$ for all $h < h_0$.

Now suppose $T_1 = 1$, or $T_1 < 1$ and $T_2 > 1$. Then there exists $z < 1$ and $h_0 > 0$, such that for $\Delta t < h_0$,

$$T_{1,z}(\Delta t) < 1/z,$$

$$T_{2,z}(\Delta t) = \frac{A}{1/z - T_{1,z}} B + \sum_{i=1}^{A_\dagger} p_i^\infty D_i \Delta t > 1,$$

where

$$D_i = \sum_{j=0}^{i-1} \frac{c_j^2 \Delta t}{1 + \gamma_{j+1}\Delta t} P_{i,j+1} z^{i-j},$$

$$A = \sum_{i=1}^{A_\dagger} p_i^\infty P_{i,0} z^i \Delta t,$$

$$B = \frac{q}{B^0} \sum_{i=1}^{A_\dagger} \beta_i p_i^\infty D_i \Delta t.$$

Now take

$$\bar{v}_0 = \frac{\epsilon B}{1/z - T_{1,z}},$$

$$\bar{v}_i = P_{i,0} z^i \bar{v}_0 + \epsilon D_i, \qquad i \geq 1,$$

with ϵ as above.

Note that

$$\eta = \sum_{i=1}^{A_\dagger} p_i^\infty \bar{v}_i \Delta t = \frac{\epsilon B}{1/z - T_{1,z}} \sum_{i=1}^{A_\dagger} p_i^\infty P_{i,0} z^i \Delta t + \epsilon \sum_{i=1}^{A_\dagger} p_i^\infty D_i \Delta t \tag{5.26}$$
$$= \epsilon T_{2,z}(\Delta t).$$

We have

$$\begin{aligned}
F_0(\bar{v}) - \bar{v}_0 &= \frac{q}{B_0} \sum_{i=1}^{A_\dagger} \beta_i p_i^\infty P_{i,0} z^i \Delta t \bar{v}_0 + \epsilon \frac{q}{B_0} \sum_{i=1}^{A_\dagger} \beta_i p_i^\infty D_i \Delta t - \bar{v}_0 \\
&= \epsilon B \frac{1-z}{1 = zT_1} > 0 \\
F_i(\bar{v}) - \bar{v}_i &= \frac{\bar{v}_{i-1}(1 + c_{i-1}^1 p_{i-1}^\infty \Delta t) + c_{i-1}^2 \Delta t \sum_{j=1}^{A_\dagger} p_j^\infty \bar{v}_j \Delta t}{1 + \gamma_i \Delta t + \Lambda_{i-1}(\bar{v})\Delta t} \\
&\quad - z \frac{1 + c_{i-1}^1 p_{i-1}^\infty \Delta t}{1 + \gamma_i \Delta t} \bar{v}_{i-1} \\
&\quad - z \frac{c_{i-1}^2 \Delta t}{1 + \gamma_i \Delta t} \epsilon \\
&= \frac{\bar{v}_{i-1}(1 + c_{i-1}^1 p_{i-1}^\infty \Delta t)\left[(1 + \gamma_i \Delta t)(1-z) - z\Lambda_{i-1}(\bar{v})\Delta t\right]}{(1 + \gamma_i \Delta t + \Lambda_{i-1}(\bar{v})\Delta t)(1 + \gamma_i \Delta t)} \\
&\quad + \frac{\epsilon c_{i-1}^2 \Delta t \left((1 + \gamma_i \Delta t)(T_{2,z} - z) - z\Lambda_{i-1}(\bar{v})\Delta t\right)}{(1 + \gamma_i \Delta t + \Lambda_{i-1}(\bar{v})\Delta t)(1 + \gamma_i \Delta t)} \\
&> (1 - z - z\Lambda_{i-1}(\bar{v})\Delta t) \frac{\bar{v}_{i-1}(1 + c_{i-1}^1 p_{i-1}^\infty \Delta t) + \epsilon c_{i-1}^2 \Delta t}{(1 + \gamma_i \Delta t + \Lambda_{i-1}(\bar{v})\Delta t)(1 + \gamma_i \Delta t)} > 0.
\end{aligned}$$

Therefore we have $U^\infty \geq \bar{v}$, in particular, using (5.26) and $T_{2,z} > 1$,

$$\sum_{i=1}^{A_\dagger} p_i^\infty U_i^\infty \Delta t \geq \eta > \epsilon \tag{5.27}$$

for all $\Delta t < h_0$. Since the left-hand side of (5.27) are Riemann sums for u_h^∞, any limit function $\bar{u}$ must satisfy $\int_0^{a_\dagger} p^\infty(a)\bar{u}(a)\,da \geq \epsilon$. □

Note that, in contrast to the other results of this paper, through Proposition 5.5, we do not state anything about the order of convergence of u_h^∞ to u^∞.

6. Numerical experiments. The form of the force of infection λ we chose in (2.1) is $\lambda(a; i(\cdot, t)) = \chi_{[1,+\infty)}(a)\ \left[k_0 + k_1 T(a) i(a,t) + k_2 l(a) I(t)/P\right]$, where k_j, $0 \leq j \leq 2$ are nonnegative constants, and $l(a)$ and $T(a)$ are continuous piecewise linear functions given by

$$T(a) = \begin{cases} 0, & a \leq 1, \\ a/5, & 1 < a \leq 5, \\ 1, & 5 < a \leq 10, \\ 1 - (a-10)/5, & 10 < a \leq 15, \\ 0, & 15 < a, \end{cases}$$

and

$$l(a) = \begin{cases} 0, & a \leq 0, \\ a/15, & 0 < a \leq 15, \\ 1, & 15 < a. \end{cases}$$

The choice of these shapes is only indicative; they are a complication of the "catalytic logistic curve" introduced by Collins [6] and used more recently by Anderson and May [1].

We used the algorithm described above for the analysis of the asymptotic behavior of the subpopulation of infected individuals, and to see whether the convergence to the steady state is monotone or not. For the tests we took a steady state age distribution defined by (1.6) where $\mu(a)$ was the actual mortality rate of the population of the United States in 1980. For the sake of simplicity, the fertility function $\beta(a)$ was chosen as a sinusoidal centered at the age of 30 years, and with support between 15 and 45 years. Tests were run with and without vertical transmission. The cure rate γ was taken to be uniform in age, and such that the mean infective period is of 8 months.

We first studied the dependence of the threshold quantities on the coefficients of transmission k_1 and k_2. It is clear that T_1 is independent of k_2, while T_2 depends linearly on it. The dependence of T_1 and T_2 on k_1 is shown in Fig. 1. It appears that T_1 grows exponentially with k_1 (notice the logarithmic scale in the figure), while (when $k_2 > 0$) T_2 starts almost linearly, and then approaches an exponential. Notice also that, even for rather small values of k_2, the first quantity to pass the threshold is T_2. Finally note that, when $k_2 > 0$, a change from $q = 0$ to $q > 0$ increases T_2 by a very small amount.

We then performed several simulations of various cases (intracohort, intercohort, and mixed), with and without vertical transmission, below and above the threshold. The algorithm generally behaved in agreement with theoretical expectations: the total

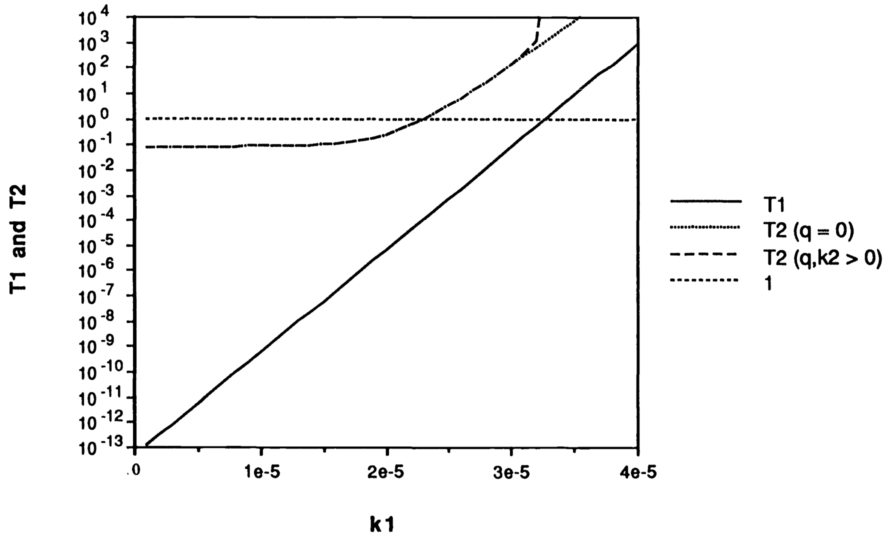


FIG. 1. *On the* y-*axis the threshold parameters* T_1 *and* T_2*; on the* x-*axis* k_1*. In all graphs* $\gamma = 1.5$*. The solid line represents* T_1*, computed with* $q = 0.111111$*;* T_2 *is computed with* $k_2 = 0.1$*, and either* $q = 0$*, or* $q = 0.111111$ *(these are the two dashed lines). Finally the threshold value* 1 *is shown in the figure.*

number of infectives converged to a positive equilibrium above the threshold, and to zero below the threshold; and the same was true for all age classes. Some plots of the total number of infectives vs. time are shown in Fig. 2; the solid lines corresponds to cases below the threshold, the dashed lines—above the threshold.

Note that in the intracohort case, it is not clear whether the lower simulation approaches zero or a positive equilibrium, despite a value of T_1 noticeably lower than one. In this case we have $T_1(\Delta t) < T_1$; therefore, theoretically, the discrete algorithm should also converge to zero. The lack of convergence to zero must thus be due to error accumulation.

In general, it seems that the intercohort case is much more well-behaved than the intracohort case; in the former case, changing the time step from 1/16 to 1/128 did not change the results noticeably, while in the intracohort case this caused differences of the order of 20% in certain cases at certain times. In this respect, the mixed case resembles the intercohort; a small amount of intercohort transmission is enough to stabilize the simulations. Probably, for the purely intracohort case, it should be worthwhile using a higher order method for integration along the characteristic lines; in order to do this a method where age discretization is kept distinct from time discretization should be used, such as the method introduced by de Roos [13].

From our simulations it appears also that, in the intercohort case, the convergence to the equilibrium is monotone (the same was true for all age classes, reaching at most something like a critically damped oscillation), while in the intracohort it can be oscillatory. The mixed case is intermediate in this respect.

Finally, in Fig. 3 we show the age distributions of the infectives at various times; we used rather high infectivities, in order to make these distributions visible. Clearly they are centered at the ages where the possibility of getting infected is highest: between the ages of 5 and 15 for the intracohort case; at later ages for the intercohort case. Notice that the addition of a small amount of intercohort transmission is enough

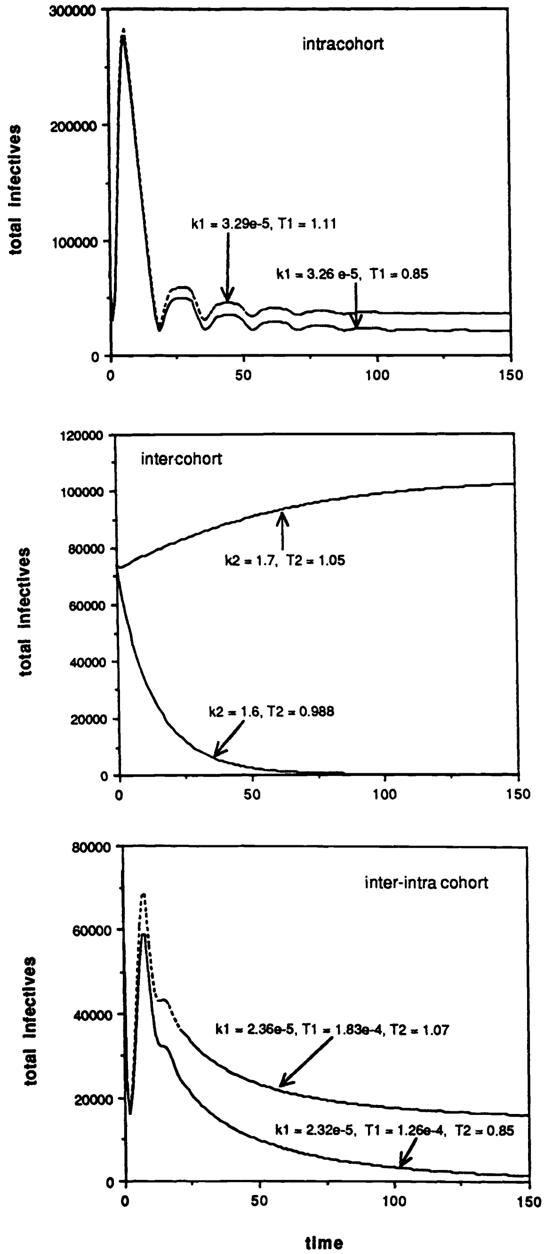


FIG. 2. *On the* y-*axis the total number of infectives; on the* x-*axis time. The three graphs correspond (from high to low) to the intracohort case* $(k_2 = 0)$, *intercohort* $(k_1 = 0)$, *inter-intracohort* $(k_1, k_2 > 0)$. *The two lines in each graph correspond to different values of a specified parameter. Other parameter values are* $\gamma = 1.5$, $q = 0.111111$.

for a very quick convergence to the equilibrium, but the age distribution remains similar to the intracohort case.

Acknowledgments. We thank an anonymous referee for several comments and suggestions that greatly improved the paper.

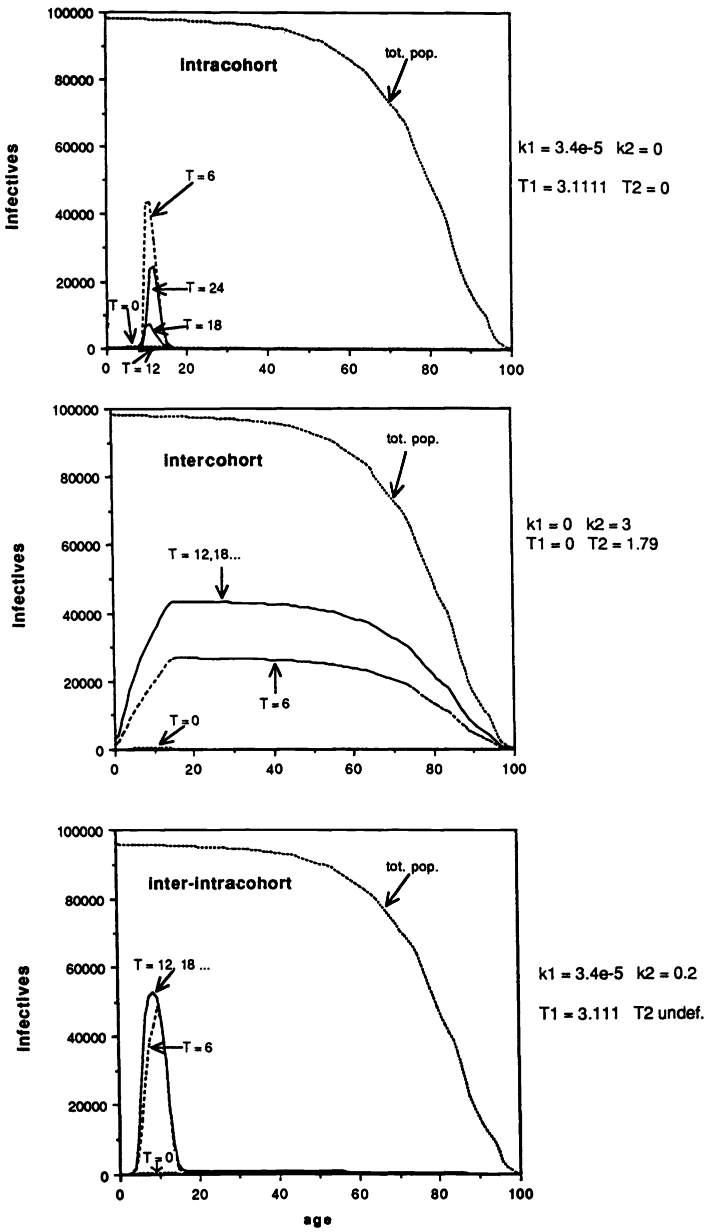


FIG. 3. *On the* x*-axis age, on the* y*-axis number of infectives. The three graphs are as in Fig. 2; the lines in each graph represent infective number at various times, starting with a given age distribution at* $T = 0$*. The stationary population distribution is also shown in the graphs. Some parameter values are specified; others are as above.*

REFERENCES

[1] R. M. ANDERSON AND R. M. MAY, *Directly transmitted infectious diseases: control by vaccination*, Science, 215 (1982), pp. 1053–1060.

[2] V. A. Andreasen and F. B. Christiansen, *Persistence of an infectious disease in a subdivided population*, Math. Biosci., 96 (1989), pp. 239–253.

[3] S. Busenberg, K. Cooke, and M. Iannelli, *Endemic thresholds and stability in a class of age-structured epidemics*, SIAM J. Appl. Math., 48 (1988), pp. 1379–1395.

[4] S. Busenberg, M. Iannelli, and H. Thieme, *Global behavior of an age-structured epidemic model*, SIAM J. Math. Anal., 22 (1991), pp. 1065–1080.

[5] ———, *Dynamics of an age-structured epidemic model*, in Proceedings on Dynamical Systems, Lecture Notes in Math. (Nankai Subseries), Springer-Verlag, Berlin, New York.

[6] S. D. Collins, *Age incidence of the common communicable diseases of children*, Publ. Health Reps., 44 (1929), pp. 763–826.

[7] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. Math. Biol., 28 (1990), pp. 365–382.

[8] J. Douglas, Jr. and F. A. Milner, *Numerical methods for a model of population dynamics*, Calcolo, XXIV (1987), pp. 247–254.

[9] H. Hethcote and H. R. Thieme, *Stability of the endemic equilibrium in epidemic models with subpopulations*, Math. Biosci., 75 (1985), pp. 205–227.

[10] M. A. Krasnoselskii, *Positive solutions of operator equations*, Noordhoff, Groningen, the Netherlands, 1964.

[11] F. A. Milner and G. Rabbiolo, *Rapidly converging numerical methods for models of population dynamics*, J. Math. Biol., to appear.

[12] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[13] A. M. de Roos, *Numerical methods for structured population models: the escalator boxcar train*, Numer. Methods for Partial Differential Equations, 4 (1988), pp. 173-195.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 689-701, May 1992

PERIODIC SOLUTIONS OF SINGLE-SPECIES MODELS WITH PERIODIC DELAY*

H. I. FREEDMAN† AND JIANHONG WU‡

Abstract. A single-species population growth model is considered, where the growth rate response to changes in its density has a periodic delay. It is shown that if the self-inhibition rate is sufficiently large compared to the reproduction rate, then the model equation has a globally asymptotically stable positive periodic solution.

Key words. single-species, population growth, oscillations, periodic solutions, delay equations, global stability, fixed point theorems

AMS(MOS) subject classifications. 92A17, 34K20

1. Introduction.

The main focus of this paper is on a model of single-species population growth which incorporates a periodic time delay in the birth process. In particular, we show the existence of a stable periodic solution of a retarded functional differential equation to be given later which has the feature of periodicity right in the time delay. To the best of our knowledge, this is the first time equations with such delays have been considered in the literature.

This paper is motivated by the laboratory work of the group led by U. Halbach (see [4], [21]-[24], [40], [44] and the references therein) on rotifers. They noticed that in laboratory populations, periodic phenomena due to time delays in gestation occurred, and that the length of delay was a function of the controlled temperature. These periodic variations in population numbers also occurred when the temperature itself was varied periodically (thereby inducing a periodic delay) on a daily basis. This led us to a conjecture that periodic solutions should exist for single species delay models with periodic delay.

Previous work has shown that periodic oscillations could occur in autonomous delay differential equations [5], [6], [10], [13], [15], [16], [20], [25], [29]-[31], [35], [37], [40], [43], [45], as well as delay equations for population growth in fluctuating environments [2], [9], [11], [12], [14], [19], [28], [32]-[34], [39], [45]. However, periodic oscillations are not automatic in single-species models with delay as shown in [3], [7], [8], [18].

In the case where the delay in growth rate is a constant, the mechanism causing oscillation is for the delay to be so significant in terms of the time length of the delay or the magnitude of the delayed effects that the positive equilibrium point (carrying capacity) loses its stability. For details, we refer to [13] and the references therein.

The technique used in the analysis of our model is to first show that due to the periodicity of the growth rate and of the delay, there exists a positive periodic carrying capacity which is not a solution, but yields a globally stable periodic oscillation in the species density. In contrast with the aforementioned research for the constant delay case, we find that the periodicity in various growth rates and in the delay can cause stable oscillation of the species density about the carrying capacity even when the delay is small.

* Received by the editors May 1, 1989; accepted for publication (in revised form) September 11, 1991.

† Applied Mathematics Institute, Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant NSERC 4823.

‡ Department of Mathematics, York University, North York, Ontario, Canada M3J 1P3.

The idea here is to then treat the periodic oscillation as being generated from the periodic carrying capacity by proving the existence of an attracting region containing the carrying capacity.

The organization of this paper is as follows. Model equations and our major results are described in § 2. We will state our results for the linear growth rate case in detail and briefly indicate the possible extension to the nonlinear growth rate case. The proofs of the theorems are contained in § 3. Under the assumption of the existence of a periodic carrying capacity, we construct a Lyapunov function about the carrying capacity and employ the Lyapunov-Razumikhin technique to obtain an attracting region. Section 4 contains a brief discussion of our results and some related open problems.

2. Model equations and main results.

2.1. Linear growth rates. We first consider the following single-species model involving a discrete periodic delay

$$\dot{x}(t)=x(t)[a(t)-b(t)x(t)+c(t)x(t-\tau(t))], \tag{2.1}$$

where the net birth rate $a(t)$, the self-inhibition rate $b(t)$, the reproduction rate $c(t)$, and the delay $\tau(t)$ are continuously differentiable, ω-periodic functions, and $a(t)>0$, $b(t)>0$, $c(t)\geqq 0$, $\tau(t)\geqq 0$ for $t\in R=(-\infty,+\infty)$. This model represents the case that when the population size is small, growth is proportional to the size, and when the population size is not so small, the positive feedback is $a(t)+c(t)x(t-\tau(t))$ while the negative feedback is $b(t)x(t)$. Such circumstances can arise when the resources are plentiful and the reproduction at time t is by individuals of at least age $\tau(t)$ units of time.

The above model, with constant coefficient and delay, and its variants, has been utilized by many authors as a model of single species growth (see [18] and the references therein). The delay in the term $c(t)x(t)x(t-\tau(t))$ is a delay due to gestation. Thinking of small animals such as rotifers (as in the work of Halbach and co-workers mentioned in the introduction), there is a small delay in the time between final feeding before reproduction and reproduction. Hence, the reproduction rate has a component which is proportional to those animals present a short time earlier and those animals currently present (random mating).

Let $\tau^*=\max_{t\in[0,\omega]}\tau(t)$. It is a well-known fact that for any given $\varphi\in C([-\tau^*,0]; R)$, there exist $\alpha\in(0,\infty)$ and a unique solution $x(t)=x(t;\varphi)$ of (2.1) on $[-\tau^*,\alpha)$; that is, $x(t)$ is continuous on $[-\tau^*,\alpha)$, continuously differentiable, and satisfies (2.1) on $(0,\alpha)$ and $x(\theta)=\varphi(\theta)$ on $[-\tau^*,0]$. Moreover, if $\varphi(t)\geqq 0$ on $[-\tau^*,0]$, then $x(t)$ remains nonnegative for all $t\in[0,\alpha)$, and if x is noncontinuable past α and $\alpha<+\infty$, then $|x(t)|\to\infty$ as $t\to\alpha^-$.

The following theorem sets forth the principal result of this paper.

THEOREM 2.1. *Suppose that the equation*

$$a(t)-b(t)K(t)+c(t)K(t-\tau(t))=0$$

has a positive, ω-periodic, continuously differentiable solution $K(t)$. Then the model equation (2.1) *has a positive ω-periodic solution $Q(t)$. Moreover, if $b(t)>c(t)Q(t-\tau(t))/Q(t)$ for all $t\in[0,\omega]$, then $Q(t)$ is globally asymptotically stable with respect to positive solutions of* (2.1).

Remark 2.1. $K(t)$ represents the carrying capacity of the environment. If all of the growth rates a, b, and c are constant in time, then $K=a/b-c$. In the case where τ is also a constant, it is shown in [18] that the condition $b>c$ guarantees the global asymptotic stability of the carrying capacity. Our result here indicates that such a global asymptotic stability holds even when τ is not a constant.

Remark 2.2. In the case where $a(t)/(b(t)-c(t))$ is not a constant, the carrying capacity $K(t)$ must be an ω-periodic function, and the periodic solution $Q(t)$ obtained in our results is nonconstant.

Remark 2.3. In the case where $b(t)>c(t)$ for $t\in[0,\omega]$, by iterating the equation $K(t)=a(t)/b(t)+c(t)/b(t)K(t-\tau(t))$, we can get an explicit expression for $K(t)$

$$K(t)=\frac{a(t)}{b(t)}+\sum_{j=0}^{\infty}\prod_{i=0}^{j}\frac{c\circ m^i(t)}{b\circ m^i(t)}\frac{a\circ m^{j+1}(t)}{b\circ m^{j+1}(t)}, \tag{2.2}$$

where $m^0(t)=t$, $m^1(t)=t-\tau(t)$, $m^i(t)=m\circ m^{i-1}(t)$ for $t\in R$ and $i\geqq 1$. It is easy to see from the formula (2.2) that if $a(t)/b(t)-c(t)$ is not a constant, then the major role of the periodicity of the delay is to cause a periodic fluctuation of the corresponding carrying capacity about the carrying capacity which occurs when τ is a constant.

Remark 2.4. In applications, it is useful to have an estimate for the location of the periodic solution $Q(t)$. The proof of this theorem in the next section will provide a rough estimate of the constants ε and $M>0$ such that

$$\varepsilon\leqq\frac{Q(t)}{K(t)}\leqq M \quad \text{for } t\in[0,\omega].$$

This inequality also indicates that we can regard the periodic oscillation as being generated from the carrying capacity, in contrast to Cushing's result [14], where the periodic oscillation bifurcates from the trivial solution.

There is some experimental evidence [9] which indicates that continuously distributed delays are more realistic and more accurate than those with instantaneous time delays. Inspired by this evidence, we consider the following Volterra integro-differential equation

$$\dot{x}(t)=x(t)\left[a(t)-b(t)x(t)+\int_{-\infty}^{t}p(t,s)x(s)\,ds\right], \tag{2.3}$$

where $p(t,s)$ is a nonnegative continuous function satisfying $p(t+\omega,s+\omega)=p(t,s)$ for $-\infty<s\leqq t<+\infty$, and there exists a constant $\gamma>0$ such that

$$\int_{-\infty}^{0}p(t,t+\theta)\,e^{-\gamma\theta}\,d\theta<\infty \quad \text{for } t\in[0,\omega]. \tag{2.4}$$

The above assumptions are motivated and satisfied by the following special delay kernel

$$k(t,s)=\frac{1}{\tau^2(t)}\cdot(t-s)\cdot\exp\left[-\frac{1}{\tau(t)}(t-s)\right], \tag{2.5}$$

which attains its maximum at $s=t-\tau(t)$ for any fixed t. Therefore, (2.3) represents a continuously distributed delay analog of the difference-differential equation (2.1) with periodic discrete delay.

Let

$$C_\gamma=\left\{\varphi\in C((-\infty,0];R);\ \lim_{\theta\to-\infty}e^{\gamma\theta}\varphi(\theta)\text{ exists}\right\}$$

with

$$|\varphi|_{C_\gamma}=\sup_{-\infty<\theta\leqq 0}e^{\gamma\theta}|\varphi(\theta)|, \qquad \varphi\in C_\gamma,$$

and define $F: R\times C_\gamma\to R$ by

$$F(t,\varphi)=\varphi(0)\left[a(t)-b(t)\varphi(0)+\int_{-\infty}^{0}p(t,t+\theta)\varphi(\theta)\,d\theta\right], \qquad (t,\varphi)\in R\times C_\gamma.$$

Then $(C_\gamma, |\cdot|_{C_\gamma})$ is a Banach space which satisfies all of the fundamental axioms described in [26], and F is a continuous functional which is Lipschitz in $\varphi \in C_\gamma$. We notice that (2.3) can be reformulated as $\dot{x}(t) = F(t, x_t)$. Therefore, by Theorems 2.1-2.5 of [26], for each $\varphi \in C_\gamma$ there exists $\alpha := \alpha(\varphi) > 0$ and a unique solution $x(t; \varphi)$ of (2.3) defined on $(-\infty, \alpha)$ with $x_0 = \varphi$, and the mapping $(t, \varphi) \in (0, \alpha(\varphi)) \times C_\gamma \subseteq R \times C_\gamma \to x_t(\varphi) \in C_\gamma$ is continuous. Moreover, if $x(t; \varphi)$ is noncontinuable past $\alpha(\varphi)$ and $\alpha(\varphi) < \infty$, then $\lim_{t\to\alpha^-} |x(t; \varphi)| = \infty$.

The following result represents an analog of Theorem 2.1 in the case of distributed delay.

THEOREM 2.2. *Assume that there exists a continuously differentiable positive ω-periodic function $K(t)$ satisfying*

$$a(t) - b(t)K(t) + \int_{-\infty}^{t} p(t, s)K(s)\,ds = 0, \qquad t \in R;$$

then the model equation (2.3) *has a positive ω-periodic solution $Q(t)$. Moreover, if*

$$b(t) > \int_{-\infty}^{t} p(t, s)\frac{Q(s)}{Q(t)}\,ds, \qquad t \in R,$$

then $Q(t)$ is globally asymptotically stable with respect to positive solutions of (2.3) *in the state space C_γ.*

2.2. Nonlinear growth rates. In this part, we indicate a possible extension of our previous results to nonlinear growth rates. We consider the following model

$$\dot{x}(t) = x(t)[-D(t, x(t)) + B(t, x_t)], \tag{2.6}$$

where the death rate $D(t, x)$ is continuous in $(t, x) \in R^2$, ω-periodic in t, increasing and continuously differentiable in x; the birth rate $B(t, \varphi)$ is continuous in $(t, \varphi) \in R \times C([-\tau, 0]; R)$ (τ is a constant), continuously differentiable in $\varphi \in C([-\tau, 0]; R)$, and is ω-periodic in t in the following sense.

(H1) For any continuous ω-periodic function $x: R \to R$, $B(t, x_t)$ is ω-periodic as a function of t.

This model represents the case where there is a delay in the per capita birth rate, whereas the death rate is instantaneous [3], [5]. We assume that all positive feedbacks are included in the birth processes and any negative feedback is included in the death rate. Our crucial assumption is the following.

(H2) There exists a positive ω-periodic continuously differentiable function $K(t)$ such that $D(t, K(t)) = B(t, K_t)$ for $t \in R$.

With this assumption, Theorem 2.1 can be modified so as to apply to our nonlinear case.

THEOREM 2.3. *Suppose that*

(i) (H1)-(H2) *are satisfied.*

(ii) *For all $t \in [0, \omega]$, we have*

$$\inf_{x\in R^+} D_x(t, x) - \sup_{\varphi\in C} \|B_\varphi(t, \varphi)\| \cdot \frac{\max_{\theta\in[0,\omega]} K(\theta)}{K(t)} > 0, \tag{H3}$$

where $\|B_\varphi(t, \varphi)\|$ denotes the operator norm of the bounded linear operator $B_\varphi(t, \varphi)$: $C \to C$.

(iii) *There exists a constant $\delta > 0$ such that for every $\delta_0 \in (0, \delta)$, and for any $\varphi \in C$ with $\varphi(s) \geqq \varphi(0) = \delta_0$, we have $B(t, \varphi) - D(t, \varphi(0)) \geqq 0$.*

Then the model equation (2.6) *has a positive ω-periodic solution $Q(t)$. Moreover, if*

$$\inf_{x\in R^+} D_x(t,x)-\sup_{\varphi\in C}\|B_\varphi(t,\varphi)\|\frac{\max_{\theta\in[0,\omega]}Q(\theta)}{Q(t)}>0$$

for all $t\in[0,\omega]$, then $Q(t)$ is globally asymptotically stable with respect to positive solutions of (2.6).

3. Proofs of theorems. In this section, we give detailed proofs for Theorems 2.1 and 2.2 and briefly indicate how to modify these proofs to the nonlinear case.

Let $C=C([-\tau^*,0];R)$ denote the Banach space of all continuous functions with the sup-norm

$$\|\varphi\|=\sup_{\theta\in[-\tau^*,0]}|\varphi(\theta)|\quad\text{for } \varphi\in C.$$

C^+ denotes a subset of C consisting of all nonnegative functions, $x(t;\varphi)$, $t\geqq-\tau^*$, $\varphi\in C^+$, denotes the unique solution of equation (2.1) satisfying $x(t;\varphi)=\varphi(t)$ on $[-\tau^*,0]$, and $x_t(\varphi)\in C$ is defined as $x_t(\varphi)(s)=x(t+s;\varphi)$ for all $s\in[-\tau^*,0]$.

LEMMA 3.1. *There exists a constant $\delta>0$ such that for every $\delta_0\in(0,\delta)$, the set*

$$B^C_{\delta_0}=\{\varphi\in C^+\colon \varphi(\theta)\geqq\delta_0\quad \textit{for } \theta\in[-\tau^*,0]\}$$

is invariant, that is, $\varphi\in B^C_{\delta_0}$ implies $x_t(\varphi)\in B^C_{\delta_0}$ for all $t\geqq0$.

Proof. We select a constant $\delta>0$ such that

$$\inf_{t\in[0,\omega]}\{a(t)-b(t)\delta\}>0.$$

Let $\delta_0\in(0,\delta)$ and $\varphi\in B^C_{\delta_0}$ be given. We consider the solution $x(t)=x(t;\varphi)$ of (2.1). If at an instant $t\geqq0$ we have $x^2(s)\geqq x^2(t)=\delta_0^2$ for $s\in[t-\tau^*,t]$, then $[x^2(t)]'\leqq0$. However, from (2.1) we have

$$\begin{aligned}[x^2(t)]'&=2x^2(t)[a(t)-b(t)x(t)+c(t)x(t-\tau(t))]\\&\geqq2x^2(t)[a(t)-b(t)\delta_0]\\&>0.\end{aligned}$$

This contradiction indicates that $\min\{\min_{\theta\in[-\tau^*,0]}x^2(t+\theta),\delta_0^2\}$ is nondecreasing, and therefore

$$\min\left\{\min_{\theta\in[-\tau^*,0]}x^2(t+\theta),\delta_0^2\right\}\geqq\min\left\{\min_{\theta\in[-\tau^*,0]}\varphi^2(\theta),\delta_0^2\right\}=\delta_0^2$$

for all $t\geqq0$. This completes the proof.

LEMMA 3.2. *For any $\rho>1$, we have*

$$\rho x-\ln(\rho x)\geqq\beta[x-\ln x]\quad \textit{for all } x\geqq1$$

where $\beta=\rho-\ln\rho$.

Proof. Let $G(x)=\rho\ln x-\ln(\rho x)+(1/x)\ln\rho$. Then $G(1)=0$, $G(\infty)=\infty$, and

$$G'(x)=\frac{1}{x^2}[(\rho-1)x-\ln\rho]$$

from which we know that $G'(x)>0$ for $x>\ln\rho/(\rho-1)$ and $G'(x)<0$ for $x<\ln\rho/(\rho-1)$. Therefore there exists a unique $x^*>1$ such that $G(x^*)=\rho-1$, $G(x)>1$ if $x>x^*$ and $G(x)<\rho-1$ for $x<x^*$.

Consider now $f(x)=(\rho x-\ln(\rho x))/(x-\ln x)$. Then

$$f'(x)=\frac{\rho-1-\rho\ln x+\ln(\rho x)-(\ln\rho)/x}{(x-\ln x)^2}$$
$$=\frac{\rho-1-G(x)}{(x-\ln x)^2}$$

which implies that $f'(x)>0$ if $x<x^*$, and $f'(x)<0$ if $x>x^*$. Therefore

$$f(x)\geqq\min\{f(1),f(\infty)\}=\min\{\rho-\ln\rho,\rho\}=\rho-\ln\rho$$

for all $x\geqq 1$. This completes the proof.

The following result describes a dissipative property of the equation, where the existence of an attracting region is essential for our main results.

LEMMA 3.3. *Assume that*

$$\frac{c(t)}{K(t)}K(t-\tau(t))<b(t)\quad \textit{on } [0,\omega].$$

Then

(i) *For any $\xi\geqq\delta$, there exists a constant $d:=d(\xi)>0$ such that for any $\varphi\in C$ with $\delta\leqq\varphi(\theta)\leqq\xi$ on $[-\tau^*,0]$, we have $\delta\leqq x(t;\varphi)\leqq d(\xi)$ for all $t\geqq 0$;*

(ii) *There exists a constant $M\geqq\delta$ such that for any $\beta\geqq\delta$ there is a constant $T=T(\beta)>0$ such that for any $\varphi\in C$ with $\delta\leqq\varphi(\theta)\leqq\beta$ on $[-\tau^*,0]$ we have $\delta\leqq x(t;\varphi)\leqq M$ for all $t\geqq T(\beta)$.*

Proof. According to the assumptions, we can find a constant $\rho>1$ such that

$$\min_{t\in[0,\omega]}\left\{b(t)-\rho\cdot\frac{c(t)}{K(t)}\cdot K(t-\tau(t))\right\}=\delta_1>0.$$

For such $\gamma>1$, define

$$M^*=\frac{2}{\delta_1}\max_{0\leqq t\leqq\omega}\left\{(\rho-1)c(t)\max_{\theta\in[0,\omega]}K(\theta)+\frac{|\dot K(t)|}{K(t)}\right\}+\max_{\theta\in[0,\omega]}K(\theta).$$

Define a continuous map $V: R\times(0,\infty)\to R$ by

$$V(t,x)=\frac{x}{K(t)}-\ln\frac{x}{K(t)}\quad\text{for } (t,x)\in R\times(0,\infty).$$

Suppose $x(t)=x(t;\varphi)$ is a solution of (2.1) with $\min_{\theta\in[-\tau^*,0]}\varphi(\theta)\geqq\delta$. By Lemma 3.1, $x(t)\geqq\delta$ for all $t\geqq 0$, and therefore $V(t,x(t))$ is well defined and differentiable for $t\geqq 0$. Moreover, we have

$$\frac{d}{dt}V(t,x(t))=\left[1-\frac{K(t)}{x(t)}\right]\left\{\frac{x(t)}{K(t)}[a(t)-b(t)x(t)+c(t)x(t-\tau(t))]-\frac{\dot K(t)}{K^2(t)}x(t)\right\}$$
$$=\frac{x(t)-K(t)}{K(t)}\left\{a(t)-b(t)x(t)+c(t)x(t-\tau(t))-\frac{\dot K(t)}{K(t)}\right\}$$
$$=-\frac{x(t)-K(t)}{K(t)}\left\{b(t)[x(t)-K(t)]\right.$$
$$\left.-c(t)[x(t-\tau(t))-K(t-\tau(t))]-\frac{\dot K(t)}{K(t)}\right\}.$$

Suppose at some $t \geqq 0$, we have

$$V(t+s, x(t+s)) \leqq (\rho - \ln \rho) V(t, x(t)) \quad \text{for } s \in [-\tau^*, 0]$$

and $x(t) \geqq M^*$. Then by Lemma 3.2, we have

$$\frac{x(t+s)}{K(t+s)} - \ln \frac{x(t+s)}{K(t+s)} \leqq \frac{\rho x(t)}{K(t)} - \ln \left(\frac{\rho x(t)}{K(t)} \right)$$

for all $s \in [-\tau^*, 0]$. From the choice of M^*, it follows that

$$\frac{x(t)}{K(t)} \geqq \frac{M^*}{K(t)} \geqq 1,$$

and therefore by the increasing property of the function $u - \ln u$ for $u \geqq 1$, we get

$$\frac{x(t+s)}{K(t+s)} \leqq \frac{\rho x(t)}{K(t)} \quad \text{for } s \in [-\tau^*, 0].$$

Hence

$$x(t+s) - K(t+s) \leqq \frac{K(t+s)}{K(t)} \rho [x(t) - K(t)] + (\rho - 1) K(t+s)$$

for all $s \in [-\tau^*, 0]$. This implies that

$$\begin{aligned}
-K(t) \frac{d}{dt} V(t, x(t)) &= b(t)[x(t) - K(t)]^2 - c(t)[x(t) - K(t)][x(t - \tau(t)) - K(t - \tau(t))] \\
&\quad - \frac{\dot{K}(t)}{K(t)} [x(t) - K(t)] \\
&\geqq b(t)[x(t) - K(t)]^2 - c(t) \cdot \frac{K(t - \tau(t))}{K(t)} [x(t) - K(t)]^2 \rho \\
&\quad -(\rho - 1) c(t) K(t - \tau(t)) |x(t) - K(t)| - \frac{\dot{K}(t)}{K(t)} [x(t) - K(t)] \\
&\geqq \left[b(t) - \rho \frac{c(t)}{K(t)} \cdot K(t - \tau(t)) \right] [x(t) - K(t)]^2 \\
&\quad - \left[(\rho - 1) c(t) \max_{\theta \in [0, \omega]} K(\theta) + \frac{|\dot{K}(t)|}{K(t)} \right] |x(t) - K(t)| \\
&\geqq \delta_1 [x(t) - K(t)]^2 \\
&\quad - \left[(\rho - 1) c(t) \max_{\theta \in [0, \omega]} K(\theta) + \frac{|\dot{K}(t)|}{K(t)} \right] |x(t) - K(t)| \\
&\geqq \frac{1}{2} \delta_1 |x(t) - K(t)|^2.
\end{aligned}$$

That is,

$$\frac{d}{dt} V(t, x(t)) \leqq - \frac{\delta_1}{2 \max_{\theta \in [0, \omega]} K(\theta)} |x(t) - K(t)|^2$$

whenever $V(t+s, x(t+s)) \leqq (\gamma - \ln \gamma) V(t, x(t))$ for $s \in [-\tau^*, 0]$ and $x(t) \geqq M^*$. Therefore, employing a variation of the standard argument of the uniform boundedness and uniform ultimate boundedness theorem of Lyapunov-Razumikhim type [25], we can prove the conclusion with any given constant $M > M^*$.

The following result from [27] is our major tool used in guaranteeing the existence of a ω-periodic solution.

LEMMA 3.4 (Horn's fixed-point theorem). *Let $S_0 \subset S_1 \subset S_2$ be convex subsets of the Banach space X, with S_0 and S_2 compact and S_1 open relative to S_2. Let $P: S_2 \to X$ be a continuous mapping such that, for some integer $m > 0$,*

(a) $P^j(S_1) \subseteq S_2$, $1 \leqq j \leqq m-1$,

and

(b) $P^j(S_1) \subseteq S_0$, $m \leqq j \leqq 2m-1$.

Then P has a fixed point in S_0.

Now we are in a position to prove our major results.

Proof of Theorem 2.1. Let $M \geqq \delta$ be given according to (ii) of Lemma 3.3. By (i) of Lemma 3.3, we can find a constant $M_1 > M+1$ such that $\delta \leqq \varphi(\theta) \leqq M+1$ on $[-\tau^*, 0]$ implies $\delta \leqq x(t; \varphi) \leqq M_1$ for all $t \geqq 0$. By (ii) of Lemma 3.3, we can find a constant $T_1 > 0$ such that $\delta \leqq \varphi(\theta) \leqq M_1 + 1$ on $[-\tau^*, 0]$ implies $\delta \leqq x(t; \varphi) \leqq M$ for all $t \geqq T_1$. Similarly, we can find constants M_2 and $M_3 > \delta$ such that

$$\delta \leqq \varphi(\theta) \leqq M_1 + 1 \quad \text{on } [-\tau^*, 0] \quad \text{implies } \delta \leqq x(t; \varphi) \leqq M_2$$

for all $t \geqq 0$, *and*

$$\delta \leqq \varphi(\theta) \leqq M_2 \quad \text{on } [-\tau^*, 0] \quad \text{implies } \delta \leqq x(t; \varphi) \leqq M_3$$

for all $t \geqq 0$.

Define

$$L = M_3 \sup_{t \in [0, \omega]} \{a(t) + b(t) M_3 + c(t) M_3\}$$

and

$$S_0 = \{\varphi \in C;\ \delta \leqq \varphi(\theta) \leqq M+1, |\varphi(\theta) - \varphi(\eta)| \leqq L|\theta - \eta| \quad \text{for } \theta, \eta \in [-\tau^*, 0]\},$$

$$S_1 = \{\varphi \in C;\ \delta \leqq \varphi(\theta) < M_1+1, |\varphi(\theta) - \varphi(\eta)| \leqq L|\theta - \eta| \quad \text{for } \theta, \eta \in [-\tau^*, 0]\},$$

$$S_2 = \{\varphi \in C;\ \delta \leqq \varphi(\theta) \leqq M_2, |\varphi(\theta) - \varphi(\eta)| \leqq L|\theta - \eta| \quad \text{for } \theta, \eta \in [-\tau^*, 0]\}.$$

As well, define a Poincaré map $P: S_2 \to C$ by

$$P(\varphi) = x_\omega(\varphi) \quad \text{for } \varphi \in S_2.$$

Then by the uniqueness and continuous dependence of solutions and the periodicity of a, b, c and τ, we have $P^n(\varphi) = x_{n\omega}(\varphi)$ for all integers $n \geqq 0$, and furthermore P is a continuous map. Evidently, $S_0 \subset S_1 \subset S_2$ are convex subsets of the Banach space C, with S_0 and S_2 compact (Arzola-Ascoli's theorem) and S_1 open relative to S_2. Choose an integer $m > 0$ such that $m\omega > T_1$. Then

$$P^j(S_1) \subseteq S_2 \quad \text{for all } j \geqq 1$$

and

$$P^j(S_1) \subseteq S_0 \quad \text{for all } j \geqq m.$$

Now by Horn's asymptotic fixed point theorem, P has a fixed point in S_0. That is, there is a ω-periodic solution $Q(t)$ of (2.1) with $Q(t) \geqq \delta$ for $t \in [0, \omega]$.

To prove the global asymptotic stability of $Q(t)$ with respect to positive solutions of (2.1), we note that

$$\left[\ln \frac{x(t)}{Q(t)}\right]' = -b(t)[x(t) - Q(t)] + c(t)[x(t - \tau(t)) - Q(t - \tau(t))].$$

Let $u(t)=x(t)-Q(t)$. We get

$$\left[\ln\left(1+\frac{u(t)}{Q(t)}\right)\right]'=-b(t)u(t)+c(t)u(t-\tau(t)).$$

The change of variable

$$\ln\left(1+\frac{u(t)}{Q(t)}\right)=y(t)$$

or equivalently,

$$u(t)=[e^{y(t)}-1]Q(t)$$

leads to the equation

$$\dot{y}(t)=-b(t)Q(t)[e^{y(t)}-1]+c(t)Q(t-\tau(t))[e^{y(t-\tau(t))}-1]. \tag{3.5}$$

Consider the function $W(t)=[e^{y(t)}-1]^2$ and choose a constant $\rho^*>1$ such that

$$\rho^*c(t)Q(t-\tau(t))<b(t)Q(t) \quad \text{for } t\in[0,\omega].$$

If at an instant $t\geqq 0$, we have

$$W(y(s))\leqq \rho^* W(y(t)) \quad \text{for all } s\in[t-\tau(t),t],$$

then

$$\begin{aligned}\frac{d}{dt}W(y(t))&=-2\,e^{y(t)}\{b(t)Q(t)[e^{y(t)}-1]^2-c(t)Q(t-\tau(t))[e^{y(t-\tau(t))}-1][e^{y(t)}-1]\}\\ &\leqq -2\,e^{y(t)}\{b(t)Q(t)[e^{y(t)}-1]^2-\rho^*c(t)Q(t-\tau(t))[e^{y(t)}-1]^2\}\\ &\leqq -2\varepsilon\, e^{y(t)}[e^{y(t)}-1]^2,\end{aligned}$$

where $\varepsilon=\inf_{t\in[0,\omega]}\{b(t)Q(t)-\rho^*c(t)Q(t-\tau(t))\}>0$. Therefore, by the uniform asymptotic stability theorem of Lyapunov-Razumikhin type, we are assured that the zero solution of (3.5) is globally uniformly asymptotically stable, that is, the ω-periodic solution $Q(t)$ of (2.1) is uniformly globally asymptotically stable with respect to positive solutions of (2.1). The proof is completed.

Proof of Theorem 2.2. First of all, using an argument similar to that for Lemma 3.1, we can show that if $\delta>0$ is sufficiently small, so that $a(t)-b(t)\delta>0$ for $t\in[0,\omega]$, then for any $\varphi\in C_\gamma$ with $\varphi(\theta)\geqq\delta$ for $\theta\leqq 0$, we have $x(t;\varphi)\geqq\delta$ for $t\geqq 0$.

Let $BC_\gamma=\{\varphi\in C_\gamma;\ \sup_{\theta\leqq 0}|\varphi(\theta)|<\infty\}$. We next prove that for any $\xi\geqq\delta$ there exists $d(\xi)>0$ such that if $\varphi\in BC_\gamma$ is given, so that $\delta\leqq\varphi(\theta)\leqq\xi e^{-\gamma\theta}$ for $\theta\leqq 0$, then $\delta\leqq x(t;\varphi)\leqq d(\varphi)$ for $t\geqq 0$. In fact, for the function $V: R\times(0,\infty)\to R$ defined by $V(t,x)=(x/K(t))-\ln(x/K(t))$, and for $x(t):=x(t;\varphi)$, we have

$$\begin{aligned}\frac{d}{dt}V(t,x(t))&=-\frac{x(t)-K(t)}{K(t)}\\ &\quad\cdot\left\{b(t)[x(t)-K(t)]-\int_{-\infty}^{t}p(t,s)[x(s)-K(s)]\,ds-\frac{\dot{K}(t)}{K(t)}\right\}.\end{aligned}$$

Since $u-\ln u$ is an increasing and unbounded function for $u\geqq 1$, we can find a constant $N_1\geqq\max_{\delta\leqq\varphi(0)\leqq\xi}V(0,\varphi(0))$ such that if $\max\{N_1, V(s,x(s))\}\leqq V(t,x(t))$ for $s\leqq t$, then $\max\{N_2, x(s)/K(s)\} \leqq x(t)/K(t)$ for $s \leqq t$, where

$$N_2 = \max_{t\in[0,\omega]} \frac{|\dot{K}(t)|/K(t)}{b(t) - \int_{-\infty}^{t} p(t,s)[K(s)/K(t)]\,ds}.$$

Therefore, if $\max\{N_1, V(s, x(s))\} \leqq V(t, x(t))$ for $s \leqq t$, then $x(s) - K(s) \leqq (K(s)/K(t))[x(t) - K(t)]$ for $s \leqq t$, and

$$\begin{aligned} &\frac{d}{dt} V(t, K(t)) \\ &\quad \leqq -\frac{1}{K(t)}\left\{\left[b(t) - \int_{-\infty}^{t} p(t,s)\frac{K(s)}{K(t)}\,ds\right][x(t)-K(t)]^2 - \frac{|\dot{K}(t)|}{K(t)}|x(t)-K(t)|\right\} \\ &\quad \leqq 0. \end{aligned}$$

Therefore, $V(t, x(t)) \leqq N_1$ for $t \geqq 0$, which implies the existence of $d(\xi)$.

We then show that there exists a constant $M \geqq \delta$ such that for any $\xi \geqq \delta$ there is a constant $T(\xi) > 0$ such that if $\varphi \in BC_\gamma$ and $\delta \leqq \varphi(\theta) \leqq \xi e^{-\gamma\theta}$ for $\theta \leqq 0$, then $\delta \leqq x(t; \varphi) \leqq M$ for all $t \geqq T(\xi)$. In fact, from the condition (2.4), for any $\varphi \geqq \delta$, we can choose $q(\xi) > 0$ such that

$$\int_{-\infty}^{-q(\xi)} p(t, t+\theta)\, e^{-\gamma\theta}\, d\theta \leqq \frac{1}{\xi + d(\varphi) + |K_0|_\gamma + \max_{0\leqq s\leqq\omega} K(s)}, \quad t \in [0, \omega].$$

Therefore,

$$\begin{aligned} &\int_{-\infty}^{-q(\xi)} p(t, t+\theta)|x(t+\theta) - K(t+\theta)|\, d\theta \\ &\quad \leqq \int_{-\infty}^{-q(\xi)} p(t, t+\theta)\, e^{-\gamma\theta}\, d\theta \left[\sup_{s\leqq -t} |x(t+s) - K(t+s)|\, e^{\gamma(t+s)}\, e^{-\gamma t}\right. \\ &\qquad\qquad\qquad\qquad \left. + \sup_{-t\leqq s\leqq 0} |x(t+s) - K(t+s)|\right] \\ &\quad \leqq \int_{-\infty}^{-q(\xi)} p(t, t+\theta)\, e^{-\gamma\theta}\, d\theta \left[\xi + |K_0|_{C_\gamma} + d(\xi) + \max_{0\leqq s\leqq\omega} K(s)\right] \\ &\quad \leqq 1. \end{aligned}$$

We now find a constant $\rho > 1$ such that

$$\min_{t\in[0,\omega]} \left\{b(t) - \rho \int_{-\infty}^{t} p(t,s)\frac{K(s)}{K(t)}\,ds\right\} = \delta_1 > 0$$

and define

$$M^* = \frac{2}{\delta_1}\left[1 + (\rho - 1) \sup_{0\leqq t\leqq\omega} \left\{\int_{-\infty}^{t} p(t,s)K(s)\,ds + \frac{|\dot{K}(t)|}{K(t)}\right\}\right].$$

Then using an argument similar to that for Lemma 3.3, we can see that if at some $t \geqq 0$, $V(s, x(s)) \leqq (\rho - \ln\rho)V(t, x(t))$ for $s \in [t - q(\xi), t]$ and $x(t) \geqq M^*$, then

$$x(s) - K(s) \leqq \frac{K(s)}{K(t)}\rho[x(t) - K(t)] + (\rho - 1)K(s)$$

for $s \in [t-q(\xi), t]$, and hence

$$\begin{aligned} -K(t)\frac{d}{dt}V(t, x(t)) &\geqq b(t)[x(t)-K(t)]^2 - \int_{-\infty}^{t-q(\xi)} p(t, s)[x(s)-K(s)]\,ds|x(t)-K(t)| \\ &\quad -\int_{t-q(\xi)}^{t} p(t, s)\frac{K(s)}{K(t)}\rho[x(t)-K(t)]^2\,ds \\ &\quad -\int_{t-q(\xi)}^{t} p(t, s)(\rho-1)K(s)|x(t)-K(t)|\,ds \\ &\quad -\frac{|\dot{K}(t)|}{K(t)}|x(t)-K(t)| \\ &\geqq \frac{\delta_1}{2}|x(t)-K(t)|^2. \end{aligned}$$

Therefore, employing a variation of the standard argument of the uniform ultimate boundedness theorem of Lyapunov-Razumikhin type [25], we can prove the existence of M.

The rest of the proof is similar to the proofs for Theorem 2.1 and Theorem 3.1 of [1], and therefore is omitted.

Proof of Theorem 2.3. We construct a Lyapunov function $V(t, x) = (x/K(t)) - \ln(x/K(t))$ for $(t, x) \in R \times (0, \infty)$, and select a constant $\rho > 0$ such that

$$\inf_{x\in R^+} D_x(t, x) - \rho \sup_{\varphi\in C} \|B_\varphi(t, \varphi)\| \frac{\max_{\theta\in[0,\omega]} K(\theta)}{K(t)} \geqq \delta_1 > 0,$$

where $\delta_1 > 0$ is a constant. It is easy to obtain

$$\begin{aligned} \frac{d}{dt}V(t, x(t)) &= -\frac{x(t)-K(t)}{K(t)}[D(t, x(t)) - B(t, x_t)] - \frac{x(t)-K(t)}{K^2(t)}\dot{K}(t) \\ &= -\frac{x(t)-K(t)}{K(t)}[D(t, x(t)) - D(t, K(t))] \\ &\quad + \frac{x(t)-K(t)}{K(t)}[B(t, x_t) - B(t, K_t)] - \frac{x(t)-K(t)}{K^2(t)}\dot{K}(t) \\ &\leqq -\frac{[x(t)-K(t)]^2}{K(t)} \inf_{x\in R^+} D_x(t, x) \\ &\quad + \sup_{\varphi\in C} \|B_\varphi(t, \varphi)\| \frac{|x(t)-K(t)|}{K(t)} \|x_t - K_t\| - \frac{x(t)-K(t)}{K^2(t)}\dot{K}(t). \end{aligned}$$

Therefore, if

$$V(t+s, x(t+s)) \leqq (\rho - \ln \rho) V(t, x(t)) \quad \text{for } s \in [-\tau^*, 0]$$

and

$$\begin{aligned} x(t) \geqq M^* = \frac{2}{\delta_1} \sup_{t\in[0,\omega]} &\left\{(\rho-1) \sup_{\varphi\in C} \|B_\varphi(t, \varphi)\| \max_{\varphi\in[0,\omega]} K(\theta) + \frac{|\dot{K}(t)|}{K(t)}\right\} \\ &+ \max_{\theta\in[0,\omega]} K(\theta). \end{aligned}$$

Then by Lemma 3.2, we can obtain

$$\frac{dV(t, x(t))}{dt} \leqq -\frac{\delta_1}{2\max_{\theta\in[0,\omega]} K(\theta)}|x(t)-K(t)|^2.$$

Hence results in Lemma 3.3 are valid. The rest of the proof is exactly the same as that for Theorem 2.1 and therefore is omitted.

Discussion. In this paper we have considered several single-species models with time delays where both the coefficients and the delays are periodic functions. These models are based on laboratory evidence in observing the population growth of rotifers.

The model given by (2.1) is of retarded type, whereas the model described by (2.3) incorporates a distributed periodic delay. However, both of these are of Lotka-Volterra type. It would be of interest to consider equations of single-species which are more general. Unfortunately, we are not able to do so at this time, since some of the technical steps in our method of proofs of the existence of positive periodic solutions require the Lotka-Volterra format.

It would also be of interest to consider higher-dimensional systems with periodic delays, representing predator-prey or competitive systems. Again, this is likely to be considerably more difficult, and we leave this for future work.

Acknowledgment. The authors thank the referees for their valuable comments and suggestions which led to a substantial improvement of our original manuscript.

REFERENCES

[1] O. A. Arino, T. A. Burton, and J. R. Haddock, *Periodic solutions to functional differential equations*, Proc. Royal Soc. Edinburgh Sect. A, 101 (1985), pp. 253-271.

[2] M. Arrigoni and A. Stein, *Logistiches Wachstum in fluktuierender Umwelt*, J. Math. Biol., 21 (1985), pp. 237-241.

[3] J. R. Bedington and R. M. May, *Time delays are not necessarily destabilizing*, Math. Biosci., 27 (1975), pp. 109-117.

[4] K. J. Beutler, C. Wissel, and U. Halbach, *Correlation and spectral analyses of the dynamics of a controlled rotifer population*, Quant. Population Dynamics, 13 (1981), pp. 61-82.

[5] S. P. Blythe, R. M. Nisbet, and W. S. C. Gurney, *Instability and complex dynamic behaviour in population models with long time delays*, Theoret. Population Biol., 22 (1982), pp. 147-176.

[6] S. P. Blythe, R. M. Nisbet, W. S. C. Gurney, and N. MacDonald, *Stability switches in distributed delay models*, J. Math. Anal. Appl., 109 (1985), pp. 388-396.

[7] T. A. Burton, *Uniform stabilities for Volterra equations*, J. Differential Equations, 36 (1980), pp. 40-53.

[8] ———, *Boundedness in functional differential equations*, Funkcial. Ekvac., 25 (1982), pp. 51-77.

[9] J. Caperon, *Time lag in population growth response of Isochrysis Galbana to a variable nitrate environment*, Ecology, 50 (1969), pp. 188-192.

[10] D. S. Cohen, E. Coutrias, and J. C. Neu, *Stable oscillations in single species growth models with hereditary effects*, Math. Biosci., 44 (1979), pp. 255-268.

[11] B. D. Coleman, *Nonautonomous logistic equations as models of the adjustment of populations to environmental change*, Math. Biosci., 45 (1979), pp. 159-173.

[12] B. D. Coleman, Y. H. Ysieh, and G. P. Knowles, *On the optimal choice of τ for a population in a periodic environment*, Math. Biosci., 46 (1979), pp. 71-85.

[13] J. M. Cushing, *Integrodifferential equations and delay models in population dynamics*, in Lecture Notes in Biomath., 20, Springer-Verlag, New York, 1977.

[14] ———, *Stable positive periodic solutions of the time-dependent logistic equation under possible hereditary influences*, J. Math. Anal. Appl., 60 (1977), pp. 747-754.

[15] G. M. Dunkel, *Some mathematical models for population growth with lags*, Ph.D. thesis, University of Maryland, College Park, MD, 1968.

[16] ———, *Single species model for population growth depending on past history*, in Seminar on Differential Equations and Dynamical Systems, Lecture Notes in Math., Vol. 60, Springer-Verlag, New York, 1968, pp. 92-99.

[17] H. I. Freedman, *Deterministic Mathematical Models in Population Ecology*, HIFR Consulting Ltd., Edmonton, Canada, 1987.

[18] H. I. Freedman and K. Gopalsamy, *Global stability in time-delayed single-species dynamics*, Bull. Math. Biol., 48 (1986), pp. 485-492.

[19] H. I. Freedman, V. S. H. Rao, and J. W.-H. So, *Asymptotic behaviour of a time-dependent single-species model*, Analysis, 9 (1989), pp. 217-223.

[20] K. GRIMMER, *Existence of periodic solutions of functional differential equations*, J. Math. Anal. Appl., 72 (1979), pp. 666-667.

[21] U. HALBACH, *Einfluss der Temperatur auf die Populationsdynamik des Planktischen Rödertieres Brachionus Calyciflorus Pallas*, Oecologia (Berl.), 4 (1970), pp. 176-207.

[22] ———, *Life table data and population dynamics of the rotifer Brachionus Calyciflorus Pallas as influenced by periodically oscillating temperature*, in Effects of Temperature on Ectothermic Organisms, W. Wieser, ed., Springer-Verlag, Heidelberg, 1973, pp. 217-228.

[23] ———, *Introductory remarks: strategies in population research exemplified by rotifer population dynamics*, Fortschr. Zool., 25 (1979), pp. 1-27.

[24] U. HALBACH, M. SIEBERT, C. WISSEL, M. KLAUS, K. BEUTLER, AND M. DELION, *Population dynamics of rotifers as bioassay tool for toxic effects of organic pollutants*, Verh. Intermat. Verein Limnol., 21 (1981), pp. 1147-1152.

[25] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Heidelberg, 1977.

[26] J. K. HALE AND J. KATO, *Phase space for retarded equations with infinite delay*, Funkcial. Ekvac., 21 (1978), pp. 11-41.

[27] W. A. HORN, *Some fixed point theorems for compact maps and flows in Banach spaces*, Trans. Amer. Math. Soc., 149 (1970), pp. 391-404.

[28] D. A. JILSON, *Inspect populations respond to fluctuating environments*, Nature, 288 (1980), pp. 699-700.

[29] G. S. JONES, *On the nonlinear differential-difference equation* $f'(x) = -\alpha f(x-1)[1+f(x)]$, J. Math. Anal. Appl., 4 (1962), pp. 440-469.

[30] ———, *The existence of periodic solutions of* $f'(x) = -\alpha f(x-1)[1+f(x)]$, J. Math. Anal. Appl., 5 (1962), pp. 435-450.

[31] S. KAKUTANI AND L. MARKUS, *On the non-linear difference-differential equations* $y'(t) = [A - By(t-\tau)]y(t)$, in Contributions to the Theory of Nonlinear Oscillations, Princeton University Press, Princeton, 1958, pp. 1-18.

[32] J. KAPLAN AND J. A. YORKE, *On the stability of a periodic solution of a differential delay equation*, SIAM J. Math. Anal., 6 (1975), pp. 268-282.

[33] N. D. KAZARINOFF AND P. VAN DEN DRIESSCHE, *Control of oscillations in hematopoiesis*, Science, 203 (1979), pp. 1348-1349.

[34] W. L. KILMER AND T. H. PROBERT, *Oscillatory depletion models for renewable ecosystems*, Math. Biosci, 36 (1977), pp. 25-29.

[35] E. SOUTHWOOD, *Time delays, density-dependence and single-species oscillations*, J. Anim. Ecol., 43 (1974), pp. 747-770.

[36] A. J. NICHOLSON, *An outline of the dynamics of animal populations*, Austral. J. Zool., 2 (1954), pp. 9-65.

[37] R. D. NUSSBAUM, *Periodic solutions of some nonlinear autonomous functional differential equations* II, J. Differential Equations, 14 (1973), pp. 360-394.

[38] V. A. PLISS, *Nonlocal Problems of the Theory of Oscillations*, Academic Press, New York, 1966.

[39] G. SEIFERT, *On a delay-differential equation for single species population variations*, Nonlinear Anal., Theory Math. Appl., 11 (1987), pp. 1051-1059.

[40] A. SEITZ AND V. HALBACH, *How is the population density regulated?*, Die Naturwissenschaften, 60 (1973), pp. 1-2.

[41] J. M. SMITH, *Models in Ecology*, Cambridge University Press, Cambridge, 1974.

[42] O. J. STAFFANS, *The null space and the range of a convolution operator in a fading memory space*, Trans. Amer. Math. Soc., 281 (1984), pp. 361-388.

[43] H. O. WALTHER, *Existence of a non-constant periodic solution of a non-linear autonomous functional differential equation representing the growth of a single species population*, J. Math. Biol., 1 (1975), pp. 227-240.

[44] C. WISSEL, K. BEUTLER, AND U. HALBACH, *Correlation functions for the evaluation of repeated time series with fluctuations*, ISEM J., 3 (1981), pp. 11-29.

[45] E. M. WRIGHT, *A non-linear difference-differential equation*, J. Reine Angew. Math., 194 (1955), pp. 66-87.

[46] B. G. ZHANG AND K. GOPALSAMY, *Global attractivity and oscillation in a periodic delay-logistic equation*, preprint.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 702–721, May 1992

HOMOCLINIC SOLUTIONS FOR AUTONOMOUS DYNAMICAL SYSTEMS IN ARBITRARY DIMENSION*

JOSEPH GRUENDLER†

Abstract. An autonomous differential equation in $\mathcal{R}^n$ with two parameters, μ_1, μ_2, is considered and it is assumed that when both parameter values are zero the equation has a hyperbolic equilibrium and a homoclinic solution. Curves are sought through the origin in the μ_1-μ_2 plane along which the homoclinic solution persists for nonzero parameter values. By using the method of Lyapunov–Schmidt, a function, H, is obtained between two finite-dimensional spaces where the zeros of H represent homoclinic solutions for nonzero parameter values. The implicit function theorem is applied to H in various cases. For $n = 2$ a single curve is obtained as in the work of Melnikov. When $n > 2$ and the stable and unstable manifolds of the hyperbolic equilibrium have an intersection of dimension one, a result of Palmer is achieved which, again, yields a single curve. When this dimension of intersection is greater than one, original results give multiple curves. The various cases of the theory are illustrated by seven examples: one in $\mathcal{R}^2$, one in $\mathcal{R}^3$, four in $\mathcal{R}^4$, and one in $\mathcal{R}^6$.

Key words. differential equations, dynamical systems, homoclinic solutions, bifurcations

AMS(MOS) subject classification. 34C35

1. Introduction.

The problem of determining parameter values for which an autonomous dynamical system possesses a homoclinic or heteroclinic solution arises in a variety of applications. Some of these are predator-prey systems [3], diffusion problems [7], traveling waves in neurons [8], [12], [13], shock waves in gas dynamics [9], [15], climate models [10], [24], chemical stirred tank reactors [16], [17], and chemical kinetics [22].

Various techniques have been used for detecting homoclinic solutions. In [4], Carr uses bifurcation from a center equilibrium. In [18], Kopell and Howard follow the bifurcation of one equilibrium to two, which can produce a heteroclinic solution. In [21], Mock uses topological degree.

In this work we consider an autonomous dynamical system involving two parameters, μ_1, μ_2, and assume that when both parameter values are zero the system has a known homoclinic solution. We then locate curves through the origin in the μ_1-μ_2 plane along which the homoclinic solution persists for small nonzero values of the parameters.

The seminal work along these lines is by Melnikov [20]. He assumes that the dynamical system is in $\mathcal{R}^2$ and is analytic, that the unperturbed system is autonomous, and that the perturbation is periodic. A generalization of [20] to $\mathcal{R}^n$ can be found in [11].

In the present work we consider a $\mathcal{C}^3$, autonomous system in $\mathcal{R}^n$ and utilize the method of Lyapunov–Schmidt. Previous work of a similar nature is by Chow and Hale [5] and Chow, Hale, and Mallet–Paret [6]. Our work is a generalization, in the autonomous case, of [5, §11.3] from $\mathcal{R}^2$ to $\mathcal{R}^n$.

*Received by the editors November 8, 1989; accepted for publication (in revised form) September 18, 1991.

†Department of Mathematics, North Carolina A&T State University, Greensboro, North Carolina 27401.

We assume the unperturbed system has a hyperbolic equilibrium and a homoclinic solution. The assumption of a homoclinic solution is equivalent to the assumption that the stable and unstable manifolds for the equilibrium intersect. Now, in the plane the intersection of the stable and unstable manifolds is equal to the homoclinic orbit itself (or, possibly, two homoclinic orbits).

In higher dimensions, on the other hand, the invariant manifolds need not meet along a single orbit. We give examples below where the invariant manifolds in $\mathcal{R}^4$ meet along an entire family of homoclinic orbits. Another way of describing this situation is in terms of the dimension of intersection of the stable and unstable manifolds. In $\mathcal{R}^2$ or $\mathcal{R}^3$ this dimension of intersection is always one. In general, the dimension of intersection in $\mathcal{R}^n$ can be as much as $n/2$.

We consider a dynamical system in $\mathcal{R}^n$ with the dimension of intersection of the stable and unstable manifolds arbitrary and use the method of Lyapunov–Schmidt to obtain a function, H, between two finite-dimensional spaces where the zeros of H represent homoclinic solutions. The independent variables for H include the parameters from the dynamical system along with some additional variables. These additional variables represent excess tangent directions for the intersection of the stable and unstable manifolds; that is, directions tangent to both manifolds other than the tangent direction provided by the homoclinic orbit itself.

When the invariant manifolds meet in dimension one the extra variables drop out. We can then apply the implicit function theorem to H and obtain, when $n = 2$, an autonomous version of Melnikov's result [20] or, when $n > 2$, a result of Palmer [23]. These results are illustrated in the first three examples below.

When the invarient manifolds meet in dimension greater than one the function H has a singularity at the origin which can result in multiple curves through the origin along which H is zero. The latter three examples below have, respectively, two, three, and five such curves.

Theory. We shall consider dynamical systems of the form

$$\dot{x}(t) = f(x(t), \mu), \tag{1}$$

where $x \in \mathcal{R}^n, \mu \in \mathcal{R}^2$. We could just as easily consider μ restricted to some connected open set containing the origin.

We make the following assumptions about (1):

(i) f is $\mathcal{C}^3$ in all its arguments.

(ii) $x = 0$ is a hyperbolic equilibrium. That is, $f(0, \mu) = 0$ for all μ and the eigenvalues of $D_1 f(0,0)$ lie off the imaginary axis.

(iii) The system has a homoclinic solution when $\mu = 0$. That is, there exists a differentiable function $t \rightarrow \gamma(t)$ such that $\dot{\gamma}(t) = f(\gamma(t), 0)$ and $\lim_{t\rightarrow+\infty} \gamma(t) = \lim_{t\rightarrow-\infty} \gamma(t) = 0$.

We shall adopt the standard notation of W^s, W^u for the stable and unstable manifolds, respectively, of the origin and $d_s = \dim(W^s)$, $d_u = \dim(W^u)$. Since $x = 0$ is a hyperbolic equilibrium, γ must approach the origin along W^s as $t \rightarrow +\infty$ and along W^u as $t \rightarrow -\infty$. Thus, γ lies on $W^s \cap W^u$. Letting $P = \gamma(0)$ we shall denote $d_b = \dim(T_P W^s \cap T_P W^u)$.

Our approach in locating homoclinic solutions for nonzero values of μ will be to define two Banach spaces, $\mathcal{Z}^1$ and $\mathcal{Z}^0$, consisting of homoclinic paths in $\mathcal{R}^n$ and try to use the usual function $F : \mathcal{Z}^1 \times \mathcal{R}^2 \rightarrow \mathcal{Z}^0$ defined as $F(z, \mu) = \dot{\gamma} + \dot{z} - f(\gamma + z, \mu)$. The idea will be to solve the equation $F(z, \mu) = 0$ for z as a function of μ. Actually, modifications of this idea will be necessary, but we introduce the basic concept here to motivate our preliminary results.

Let $\{\lambda_1, \cdots, \lambda_n\}$ denote the eigenvalues of $D_1 f(0,0)$. Since $x = 0$ is a hyperbolic equilibrium we can choose $M > 0$ such that $|\Re(\lambda_i)| \geq 4M$ for all i. For sufficiently small $|\mu|$ any homoclinic solution which exists must decay at least as fast as e^{-Mt} as $t \to +\infty$ and e^{Mt} as $t \to -\infty$. Accordingly, we define

$$\mathcal{Z}^0 = \left\{ z \in \mathcal{C}^0(\mathcal{R}, \mathcal{R}^n) : \sup_t |z(t)| e^{M|t|} < \infty \right\},$$

$$\mathcal{Z}^1 = \left\{ z \in \mathcal{C}^1(\mathcal{R}, \mathcal{R}^n) : \sup_t |z(t)| e^{M|t|} < \infty, \sup_t |\dot{z}(t)| e^{M|t|} < \infty \right\}.$$

We take as a norm on each space the maximum of the sups in the respective definition. With these norms each $\mathcal{Z}^r$ is a Banach space.

Once the function F above has been properly defined we will need to look at the derivative, $D_1 F(0,0)$. To determine the Kernel of this map we must solve the variational equation $\dot{u} = D_1 f(\gamma, 0) u$. Note that $\dot{\gamma}$ is one solution in $\mathcal{Z}^1$ to this equation. There may be others.

THEOREM 1. *Let $\dot{x} = f(x, \mu)$ be as in (1). Let $\{\lambda_1, \cdots, \lambda_n\}$ denote the eigenvalues of $D_1 f(0,0)$. Then there exists a fundamental solution, U, for the variational equation $\dot{u}(t) = D_1 f(\gamma(t), 0) u(t)$; nonnegative integers k_i; real, continuous, bounded functions v_i^+, v_i^-; and a permutation, σ, on n symbols such that, if u_j denotes the jth column of U,*

$$\lim_{t \to +\infty} \left(u_j(t) t^{-k_j} \exp(-\Re(\lambda_j) t) - v_j^+(t) \right) = 0,$$

$$\lim_{t \to -\infty} \left(u_j(t) t^{-k_{\sigma(j)}} \exp(-\Re(\lambda_{\sigma(j)}) t) - v_j^-(t) \right) = 0.$$

Proof. Let $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ denote the eigenvalues of $D_1 f(0,0)$ repeated according to algebraic multiplicity and numbered so that $\Re(\lambda_j) \geq \Re(\lambda_{j+1})$ and so that $k_j \geq k_{j+1}$ when $\Re(\lambda_j) = \Re(\lambda_{j+1})$. By standard asymptotic theory we have the existence of two real fundamental solutions, $\{\alpha_1, \cdots, \alpha_n\}$ and $\{\beta_1, \cdots, \beta_n\}$, to the variational equation such that, for $j = 1, \cdots, n$, we have

$$\lim_{t \to +\infty} (\alpha_j(t) t^{-k_j} \exp(-\Re(\lambda_j) t) - w_j(t)) = 0,$$

$$\lim_{t \to -\infty} (\beta_j(t) t^{-k_{n-j+1}} \exp(-\Re(\lambda_{n-j+1}) t) - w_{n-j+1}(t)) = 0$$

for some set of nonnegative integers k_j and nonzero, continuous, periodic functions w_j.

We denote by $A(t)$ and $B(t)$ the matrices with, respectively, $\alpha_j(t)$ and $\beta_j(t)$ as column j. Since these are both fundamental solutions we can write $A = BP$ for some constant matrix P. We now operate on P by means of elementary column operations. The objective is to obtain $AQ = B\tilde{P}$ with Q lower-triangular and $\tilde{P}$ such that the first nonzero entry in each column is one with each column-leading one in a different row.

Suppose we have reached the point where the transformed P has the following property: there exist distinct integers $j_1, j_2, \cdots, j_{r-1}$ such that

$$\begin{aligned} &p_{ij_k} = 0 \quad \text{if } i < k, \\ &p_{kj_k} = 1, \\ &p_{ik} = 0 \quad \text{for } 1 \leq i < r-1, \quad k \notin \{j_1, j_2, \cdots, j_{r-1}\}. \end{aligned}$$

In row r pick the maximum $j_r \notin \{j_1, \cdots j_{r-1}\}$ such that $p_{rj_r} \neq 0$. Such a j_r must exist as P is nonsingular. Now divide column j_r by p_{rj_r} so now $p_{rj_r} = 1$. Next, use column operations to get $p_{rk} = 0$ for $k \notin \{j_1, j_2, \cdots, j_r\}$. Notice we need operate only on columns to the left of column j_r.

Continuing this process through $r = n$ yields a non-singular, lower triangular constant matrix Q such that $A(t)Q = B(t)\tilde{P}$ where $\tilde{P}$ has the property that given j, $1 \leq j \leq n$, there exists $\sigma(j)$ defined by $j_{\sigma(i)} = i$ such that $\sigma(i) \neq \sigma(j)$ for $i \neq j$, $\tilde{p}_{ij} = 0$ for $i < \sigma(j)$, and $\tilde{p}_{\sigma(j),j} = 1$.

Let u_j denote column j of the matrix $U = A(t)Q = B(t)\tilde{P}$. Then

$$u_j(t) = \sum_{r=j}^{n} q_{rj}\alpha_r(t) = \sum_{r=\sigma(j)}^{n} \tilde{p}_{rj}\beta_r(t),$$

so that

$$\lim_{t \to +\infty} \left(u_j(t)t^{-k_j} \exp(-Re(\lambda_j)t) - v_j^+(t)\right) = 0$$

where $v_j^+ = \sum q_{rj}w_r$, the sum taken over

$$\{r : j \leq r \leq n, \Re(\lambda_r) = \Re(\lambda_j), k_r = k_j\}.$$

Also,

$$\lim_{t \to -\infty} \left(u_j(t)t^{-k_{\sigma(j)}} \exp(-\Re(\lambda_{\sigma(j})t) - v_j^-(t)\right) = 0$$

where $v_j^- = \sum \tilde{p}_{rj}w_r$, the sum taken over

$$\{r : \sigma(j) \leq r \leq n, \Re(\lambda_r) = \Re(\lambda_j), k_r = k_j\}. \qquad \square$$

We shall refer to U as a dichotomous variational solution along γ. We classify the solutions into four types and introduce notation useful for solution of the nonhomogeneous variational equation.

DEFINITION 2. Let $\dot{x} = f(x, \mu)$ be as in (1). Let U denote a dichotomous variational solution along γ with u_j the jth column of U.

(i) We shall say that u_j connects λ_j to $\lambda_{\sigma(j)}$.

(ii) We write the index set, $I = \{1, 2, \cdots, n\}$, as the disjoint union of four sets according to the behavior of the u_j's at $+\infty$ and $-\infty$:

$$\begin{aligned}
j \in I_{ab}, \quad a = s \quad &\text{iff} \lim_{t \to +\infty} u_j(t) = 0,\\
j \in I_{ab}, \quad a = u \quad &\text{iff} \lim_{t \to +\infty} u_j(t) = \infty,\\
j \in I_{ab}, \quad b = s \quad &\text{iff} \lim_{t \to -\infty} u_j(t) = 0,\\
j \in I_{ab}, \quad b = u \quad &\text{iff} \lim_{t \to -\infty} u_j(t) = \infty.
\end{aligned}$$

(iii) We let n_{ab} denote the order of I_{ab} and use "u_j is of type ab" to mean $j \in I_{ab}$.

(iv) For each $i = 1, \cdots, n$ we define $u_i^\perp(t)$ by $\langle u_i^\perp(t), u_j(t)\rangle = \delta_{ij}$.

It may occur that $\dot{\gamma} = u_k$ for some k. The actual numbering of the u_j's is unimportant. The essential thing is that each u_j connects a distinct pair of eigenvalues.

The vectors $u_j^\perp$ can be computed from the formula $U^{\perp t} = U^{-1}$, where $U^\perp$ denotes the matrix with $u_j^\perp$ as column j. Differentiating $UU^{\perp t} = I$ we obtain $\dot{U}U^{\perp t} + U\dot{U}^{\perp t} = 0$ so that $\dot{U}^\perp = -(U^{-1}\dot{U}U^{\perp t})^t = -D_1 f(\gamma, 0)^t U^\perp$. Thus, $U^\perp$ is the adjoint of U.

We also have

$$\det(U(t)) = \det(U(0)) \exp\left(\int_0^t (\nabla \cdot f)(\gamma(s), 0)\, ds\right),$$

and, for $w \in \mathcal{C}^r(\mathcal{R}, \mathcal{R}^n)$, the formula

$$\langle u_i^\perp(t), w(t)\rangle = \det(u_1(t), \cdots, u_{i-1}(t), w(t), u_{i+1}(t), \cdots, u_n(t)) \det(U(t))^{-1}. \tag{2}$$

We shall construct a solution to the nonhomogeneous variational equation $\dot{x} = D_1 f(\gamma, 0)x + w$ in the form

$$x = \sum_{j=1}^n u_j(t) \int \langle u_j^\perp(s), w(s)\rangle\, ds = U(t) \int U^{-1}(s) w(s)\, ds. \tag{3}$$

For breaking up this formula over the various I_{ab} we define, for each $a = s, u$, $b = s, u$, an $n \times n$ matrix, P_{ab}, defined as $(P_{ab})_{ii} = 1$ if $i \in I_{ab}$ and $(P_{ab})_{ij} = 0$, otherwise. We have then

$$\sum_{j \in I_{ab}} u_j(t) \int \langle u_j^\perp(s), w(s)\rangle\, ds = U(t) \int P_{ab} U^{-1}(s) w(s)\, ds. \tag{4}$$

LEMMA 3. *There exists a constant K_0 such that for every $z \in \mathcal{Z}^0$ we have:*

(i) $$\left|\langle u_i^\perp(t), z(t)\rangle\right| \leq K_0 \|z\| e^{-Mt}$$
for $t \geq 0$ when $i \in I_{us} \cup I_{uu}$.

(ii) $$\left|\langle u_i^\perp(t), z(t)\rangle\right| \leq K_0 \|z\| e^{Mt}$$
for $t \leq 0$ when $i \in I_{su} \cup I_{uu}$.

(iii) $$\left|\int_0^t U(t)(P_{ss} + P_{su}) U^{-1}(s) z(s)\, ds\right| \leq K_0 \|z\| e^{-Mt}$$
for $t \geq 0$.

(iv) $$\left|\int_t^\infty U(t)(P_{us} + P_{uu}) U^{-1}(s) z(s)\, ds\right| \leq K_0 \|z\| e^{-Mt}$$
for $t \geq 0$.

(v) $$\left|\int_0^t U(t)(P_{ss} + P_{us}) U^{-1}(s) z(s)\, ds\right| \leq K_0 \|z\| e^{Mt}$$
for $t \leq 0$.

(vi) $$\left|\int_{-\infty}^t U(t)(P_{su} + P_{uu}) U^{-1}(s) z(s)\, ds\right| \leq K_0 \|z\| e^{Mt}$$
for $t \leq 0$.

Proof. Choose $K_0 > 0$ such that

$$|u_i(t)| \leq K_0^{1/n} \exp\left(\left(\Re(\lambda_i) + M/(n-1)\right) t\right)$$

for all $t \geq 0$ and

$$\det\left(U^{-1}(t)\right) = \det(U(0)) \exp\left(-\int_0^t (\nabla \cdot f)(\gamma(s), 0)\, ds\right) \leq K_0^{1/n} e^{(-\lambda+M)t},$$

where $\lambda = \Re(\lambda_1) + \cdots + \Re(\lambda_n)$. Since $i \in I_{us} \cup I_{uu}$, $\Re(\lambda_i) \geq 4M$, so, using (2),

$$\left\langle u_i^{\perp}(t), z(t) \right\rangle \leq K_0 \|z\| \exp((-\Re(\lambda_i))t) \leq K_0 \|z\| e^{-Mt}.$$

This proves (i). Part (ii) follows similarly. Note that (i) and (ii) imply the convergence of the improper integrals in (iv) and (vi) when these are viewed via (4).

We now turn to (iii). Assume the eigenvalues are numbered so that $I_{ss} \cup I_{su} = \{1, 2, \cdots, d_s\}$. Then $P_{ss} + P_{su} = \left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$ where I is the $d_s \times d_s$ identity matrix. Let $A = D_1 f(0,0)$ and let Q be a real, $n \times n$ matrix such that the matrix $B = Q^{-1}AQ$ has the form $B = \left(\begin{smallmatrix} B_1 & 0 \\ 0 & B_2 \end{smallmatrix}\right)$ with B_1 a $d_s \times d_s$ matrix whose eigenvalues all have negative real parts. We can choose $K_B > 0$ such that

$$\left|\begin{pmatrix} e^{tB_1} & 0 \\ 0 & 0 \end{pmatrix}\right| \leq |e^{tB_1}| \leq K_B e^{-2Mt}.$$

Since

$$Q^{-1}\dot{U} = \left(Q^{-1} D_1 f(\gamma, 0) Q\right) Q^{-1} U$$

and

$$\lim_{t \to \infty} Q^{-1} D_1 f(\gamma, 0) Q = B$$

we can choose $K_1 > 0$ such that $|Q^{-1}U - e^{tB}| \leq K_1 e^{-2Mt}$, and using the adjoint equation we obtain $|U^{-1}Q - e^{-tB}| \leq K_2 e^{-2Mt}$, both for all $t \geq 0$.

Now write

$$\begin{aligned}
&U(t)(P_{ss} + P_{su})U^{-1}(s)z(s) \\
&\quad = Q\left[\left(Q^{-1}U(t) - e^{tB}\right)(P_{ss} + P_{su})\left(U^{-1}(s)Q - e^{-sB}\right)\right. \\
&\qquad + \left(Q^{-1}U(t) - e^{tB}\right)(P_{ss} + P_{su})e^{-sB} + e^{tB}(P_{ss} + P_{su})e^{-sB} \\
&\qquad \left. + e^{tB}(P_{ss} + P_{su})\left(U^{-1}(s)Q - e^{-sB}\right)\right] Q^{-1}z(s) \\
&\quad = Q\left[\left(Q^{-1}U(t) - e^{tB}\right)(P_{ss} + P_{su})\left(U^{-1}(s)Q - e^{-sB}\right)\right. \\
&\qquad + \left(Q^{-1}U(t) - e^{tB}\right)\begin{pmatrix} e^{-sB_1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} e^{(t-s)B_1} & 0 \\ 0 & 0 \end{pmatrix} \\
&\qquad \left. + \begin{pmatrix} e^{tB_1} & 0 \\ 0 & 0 \end{pmatrix}\left(U^{-1}(s)Q - e^{-sB}\right)\right] Q^{-1}z(s).
\end{aligned}$$

From this we obtain

$$|U(t)(P_{ss} + P_{su})U^{-1}(s)z(s)| \leq K_3 \|z\| e^{-2Mt} e^{-3Ms} + K_B \|z\| e^{-2Mt} e^{Ms},$$

where $K_3 = K_1 \left|P_{ss} + P_{su}\right| K_2 + K_1 K_B + K_B K_2$.

Integration of the preceding estimate yields (iii) with $K_0 = \frac{K_3}{3M} + \frac{K_B}{M}$. This proves (iii). Parts (iv)–(vi) follow similarly. □

One consequence of the preceding result is that the integrals $\int_{-\infty}^{\infty} \left\langle u_i^\perp, z \right\rangle dt$ are convergent for $z \in \mathcal{Z}^0, i \in I_{uu}$.

Furthermore, since

$$\begin{aligned}\left\langle u_i^\perp, \dot{z} - D_1 f(\gamma, 0) z \right\rangle &= \left\langle u_i^\perp, \dot{z} \right\rangle - \left\langle D_1 f(\gamma, 0)^t u_i^\perp, z \right\rangle \\ &= \left\langle u_i^\perp, \dot{z} \right\rangle + \left\langle \dot{u}_i^\perp, z \right\rangle \\ &= \frac{d}{dt} \left\langle u_i^\perp, z \right\rangle,\end{aligned}$$

it follows that

$$\int_{-\infty}^{\infty} \left\langle u_i^\perp, \dot{z} - D_1 f(\gamma, 0) z \right\rangle dt = 0 \quad \text{for } z \in \mathcal{Z}^1,\ i \in I_{uu}. \tag{5}$$

THEOREM 4. *Let $\dot{x} = f(x, \mu)$ be as in* (1) *and consider the nonhomogeneous variational equation*

$$\dot{h}(t) = D_1 f(\gamma(t), 0) h(t) + w(t) \tag{6}$$

with $w \in \mathcal{Z}^0$. Then (6) *has a solution in $\mathcal{Z}^1$ if and only if*

$$\int_{-\infty}^{\infty} \left\langle u_i^\perp, w \right\rangle dt = 0$$

for all $i \in I_{uu}$. In this case (6) *has a unique solution in $\mathcal{Z}^1$ satisfying $\left\langle u_i^\perp(0), h(0) \right\rangle = 0$ for all $i \in I_{ss}$.*

Proof. If (6) has a solution $h \in \mathcal{Z}^1$ then $\int_{-\infty}^{\infty} \left\langle u_i^\perp, w \right\rangle dt = 0$ for all $i \in I_{uu}$ by (5). Now suppose $\int_{-\infty}^{\infty} \left\langle u_i^\perp, w \right\rangle dt = 0$ for all $i \in I_{uu}$. Then using variation of parameters we get a particular solution, h_p, to (6) given by

$$\begin{aligned} & h_p(t) \\ &= U(t) \left[\int_{-\infty}^{0} P_{su} U^{-1}(s) w(s)\, ds + \int_0^t (P_{ss} + P_{su}) U^{-1}(s) w(s)\, ds \right. \\ &\qquad \left. - \int_t^{\infty} (P_{us} + P_{uu}) U^{-1}(s) w(s)\, ds \right] \\ &= U(t) \left[- \int_0^{\infty} P_{us} U^{-1}(s) w(s)\, ds + \int_0^t U(t)(P_{ss} + P_{us}) U^{-1}(s) w(s)\, ds \right. \\ &\qquad \left. + \int_{-\infty}^{t} U(t)(P_{su} + P_{uu}) U^{-1}(s) w(s)\, ds \right], \end{aligned}$$

where the first and second forms are intended for $t \geq 0$ and $t \leq 0$, respectively.

Now $h_p \in \mathcal{Z}^1$ by Lemma 3. The general solution in $\mathcal{Z}^1$ to (6) is $h = h_p + \sum_{i \in I_{ss}} c_i u_i$ for constants c_i. But $\left\langle u_i^\perp(0), h(0) \right\rangle = c_i$ for $i \in I_{ss}$ so h_p is the unique solution in $\mathcal{Z}^1$ to (6), satisfying $\left\langle u_i^\perp(0), h(0) \right\rangle = 0$ for all $i \in I_{ss}$. □

The preceding result has two consequences. First, it characterizes the image of $D_1F(0,0)$ as $\{w \in \mathcal{Z}^0 : \int_{-\infty}^{\infty} \langle u_i^\perp, z\rangle\, dt = 0,\ i \in I_{uu}\}$. Further, we see that we can make $D_1F(0,0)$ injective by restricting the domain of F to the subspace

$$\tilde{\mathcal{Z}}^1 = \{z \in \mathcal{Z}^1 : \langle u_i^\perp(0), z(0)\rangle = 0,\ i \in I_{ss}\}.$$

As the Kernel of a continuous linear map $\tilde{\mathcal{Z}}^1$ is closed and hence a Banach space.

We can proceed to our main result once we settle the question of the domain for F. On the one hand, F cannot be defined on all $\mathcal{Z}^1$ since this would distinguish solutions in $\mathcal{Z}^1$ which are time-translates of each other. There is a way to handle this problem. Let U be a dichotomous variational solution along γ and choose $k \in I_{ss}$ so that $\langle u_k^\perp(0), \dot\gamma(0)\rangle \neq 0$. We then require z to satisfy $\langle u_k^\perp(0), z(0)\rangle = 0$ so that $\langle u_k^\perp(0), \gamma(0) + z(0)\rangle$ is independent of z. This construction has a geometric meaning. Since $\langle u_i^\perp(0), \dot\gamma(0)\rangle$ is the coordinate of $\dot\gamma(0)$ along $u_i(0)$, the vectors $u_i(0)$, $i \neq k$ span an affine linear subspace, Π, in $\mathcal{R}^n$ through $P = \gamma(0)$ and transverse to $\dot\gamma(0)$. We are requiring that $\gamma(0) + z(0)$ remain in Π. In each example below we have $u_k = \dot\gamma$ for some k so then $\langle u_k^\perp(0), \dot\gamma(0)\rangle = \langle u_k^\perp(0), u_k(0)\rangle = 1$. This situation can always be assumed for $d_b = 1$ but not necessarily when $d_b > 1$.

On the other hand, for $D_1F(0,0)$ to be injective we need to have $z \in \tilde{\mathcal{Z}}^1$. Now, when $d_b = 1$ these two requirements agree but, for $d_b > 1$, requiring z to be in $\tilde{\mathcal{Z}}^1$ is unnecessarily restrictive. We reconcile this by taking $z \in \tilde{\mathcal{Z}}^1$ but then adding extra variables to take into account the tangent directions $u_j(0)$, $j \in I_{ss}$, $j \neq k$.

THEOREM 5. *Let $\dot x = f(x,\mu)$ be as in (1). Let U be a dichotomous variational solution along γ numbered so that $I_{uu} = \{1, \cdots, d_b\}$ along with a bijection $\alpha : \{1, \cdots, d_b\} \to I_{ss}$ and such that $\left\langle u_{\alpha(d_b)}^\perp(0), \dot\gamma(0)\right\rangle \neq 0$. Then there exists a connected open set $V \subset \mathcal{R}^{d_b-1} \times \mathcal{R}^2$ with $(0,0) \in V$ and a differentiable function $H : V \to \mathcal{R}^{d_b}$ denoted $(\beta,\mu) \to H(\beta,\mu)$ with the following properties:*

(i) *If $H(\beta^*,\mu^*) = 0$ then $\dot x = f(x,\mu^*)$ has a homoclinic solution*

(ii) $H(0,0) = 0.$

(iii) $$\frac{\partial H_i}{\partial \mu_j}(0,0) = -\int_{-\infty}^{\infty} \left\langle u_i^\perp, \frac{\partial f}{\partial \mu_j}(\gamma,0)\right\rangle dt.$$

(iv) $$\frac{\partial H_i}{\partial \beta_j}(0,0) = 0.$$

(v) $$\frac{\partial^2 H_i}{\partial \beta_j \partial \beta_k}(0,0) = -\int_{-\infty}^{\infty} \langle u_i^\perp, D_{11}f(\gamma,0)u_{\alpha(j)}u_{\alpha(k)}\rangle\, dt.$$

Proof. We define a differentiable function $F : \tilde{\mathcal{Z}}^1 \times \mathcal{R}^{d_b-1} \times \mathcal{R}^2 \to \mathcal{Z}^0$ as follows:

$$F(z,\beta,\mu)(t) = \dot\gamma(t) + \dot z(t) + \sum_{i=1}^{d_b-1} \beta_i \dot u_{\alpha(i)}(t) - f\left(\gamma(t) + z(t) + \sum_{i=1}^{d_b-1} \beta_i u_{\alpha(i)}(t), \mu\right).$$

A sufficient condition for a homoclinic solution is $F(z,\beta,\mu) = 0$. We use the method of Lyapunov–Schmidt to eliminate z from this equation.

Let $\phi : \mathcal{R} \to \mathcal{R}$ be a continuous function satisfying $\sup_t |\phi(t)|e^{M|t|} < \infty$ and $\int_{-\infty}^{\infty} \phi(t)dt = 1$. Then define a projection $P : \mathcal{Z}^0 \to \mathcal{Z}^0$ by

$$P(z) = \phi \sum_{i \in I_{uu}} \left(\int_{-\infty}^{\infty} \langle u_i^{\perp}, z \rangle \, dt \right) u_i.$$

It is easy to check that for any $z \in \mathcal{Z}^0$ we have

$$\int_{-\infty}^{\infty} \langle u_i^{\perp}, (I-P)z \rangle \, dt = 0 \quad \text{for } i \in I_{uu}. \tag{7}$$

So now F resolves into two parts:

$$\begin{aligned} (I-P)F &: \tilde{\mathcal{Z}}^1 \times \mathcal{R}^{d_b - 1} \times \mathcal{R}^2 \to \mathrm{Im}(I-P), \\ PF &: \tilde{\mathcal{Z}}^1 \times \mathcal{R}^{d_b - 1} \times \mathcal{R}^2 \to \mathrm{Im}(P). \end{aligned}$$

The idea is to solve the equation $(I-P)F = 0$ for z and substitute the result into $PF = 0$. Using (5) we find that the equation

$$D_1[(I-P)F](0)h = w$$

is

$$\dot{h} = D_1 f(\gamma, 0)h + w.$$

Combining (7) and Theorem 4 we see that this last equation has a unique solution in $\tilde{\mathcal{Z}}^1$ since $w \in \mathrm{Im}(I-P)$. We can now apply the implicit function theorem to obtain a connected open set $V \subset \mathcal{R}^{d_b - 1} \times \mathcal{R}^2$ and a differentiable function $\psi : V \to \tilde{\mathcal{Z}}^1$ such that $\psi(0,0) = 0$ and

$$((I-P)F)(\psi(\beta, \mu), \beta, \mu) = 0 \tag{8}$$

for all $(\beta, \mu) \in V$.

The conditions for a homoclinic solution now become

$$(PF)(\psi(\beta, \mu), \beta, \mu) = 0.$$

These conditions can be stated in the form $H(\beta, \mu) = 0$ by defining the differentiable function $H : V \to \mathcal{R}^{d_b}$ as follows:

$$\begin{aligned} H_i(\beta, \mu) = \int_{-\infty}^{\infty} \Bigg\langle u_i^{\perp}, \dot{\gamma} + \dot{\psi}(\beta, \mu) + \sum_{j=1}^{d_b - 1} \beta_j \dot{u}_{\alpha(j)} \\ - f\left(\gamma + \psi(\beta, \mu) + \sum_{j=1}^{d_b - 1} \beta_j u_{\alpha(j)}, \mu \right) \Bigg\rangle \, dt. \end{aligned}$$

Result (ii) follows from the fact that $\psi(0,0) = 0$. Differentiation of H combined with (5) yields (iii) and (iv).

To show (v) we first differentiate (8) with respect to β and use (5) to obtain

$$\frac{\partial\dot{\psi}}{\partial\beta_j}(0,0) = D_1 f(\gamma,0)\frac{\partial\psi}{\partial\beta_j}(0,0),$$

and this result combined with Theorem 4 yields

$$\frac{\partial\psi}{\partial\beta_j}(0,0) = 0.$$

We now obtain (v) by differentiation of H twice with respect to β, combined with this last result and (5). □

Let us look at some special cases of the preceding theorem. If $x \in \mathcal{R}^2$ we must have $d_b = 1$, $I_{uu} = \{1\}$, $I_{ss} = \{2\}$, and $u_2 = \dot{\gamma}$. Denoting $\gamma = (\gamma_1, \gamma_2)$ we can take

$$u_1^{\perp}(t) = (-\dot{\gamma}_2(t), \dot{\gamma}_1(t)) \exp\left(-\int_0^t (\nabla \cdot f)(\gamma(s),0)\,ds\right),$$

and knowledge of u_1 is not required. There is no β and the condition for a homoclinic solution is the scalar equation $H(\mu_1, \mu_2) = 0$. With the appropriate hypothesis the implicit function theorem can be used to solve this last equation for μ_1 in terms of μ_2. This yields the following result which is an autonomous version of Melnikov's theorem [20]. See also [5] and [6]. This next result is illustrated in Example 1 below.

COROLLARY 6. *Let $\dot{x} = f(x,\mu)$ be as in* (1) *with $x \in \mathcal{R}^2$. Let*

$$a_i = -\int_{-\infty}^{\infty} \det\left(\dot{\gamma}(t), \frac{\partial f}{\partial\mu_i}(\gamma(t),0)\right) \exp\left(-\int_0^t (\nabla \cdot f)(\gamma(s),0)\,ds\right) dt$$

for $i = 1,2$ and suppose $a_1 \neq 0$. Then there exists an open interval W containing zero and a differentiable function $\phi : W \to \mathcal{R}$ such that $\phi(0) = 0$, $\phi'(0) = -a_2/a_1$ and such that (1) *has a homoclinic solution for $\mu_1 = \phi(\mu_2)$, $\mu_2 \in W$.*

The preceding result generalizes with little difficulty to higher n as long as we have $d_b = 1$. The difference is that for $n > 2$ it is necessary to utilize a dichotomous variational solution. This leads to the following result obtained by Palmer [23]. This next result is illustrated in Examples 2 and 3 below.

COROLLARY 7. *Let $\dot{x} = f(x,\mu)$ be as in* (1) *with $d_b = 1$. Let U be a dichotomous variational solution along γ with $I_{uu} = \{1\}$. Let*

$$a_i = \frac{\partial H}{\partial\mu_i}(0,0) = -\int_{-\infty}^{\infty} \left\langle u_1^{\perp}, \frac{\partial f}{\partial\mu_i}(\gamma,0)\right\rangle dt$$

for $i = 1,2$ and suppose $a_1 \neq 0$. Then there exists an open interval W containing zero and a differentiable function $\phi : W \to \mathcal{R}$ such that $\phi(0) = 0$, $\phi'(0) = -a_2/a_1$ and such that (1) *has a homoclinic solution for $\mu_1 = \phi(\mu_2)$, $\mu_2 \in W$.*

We now wish to consider $d_b > 1$. A special case occurs for $d_b = 2$ for then we have $\beta \in \mathcal{R}$, $\mu \in \mathcal{R}^2$, and the conditions for a homoclinic solution are $H_i(\beta, \mu_1, \mu_2) = 0$; $i = 1,2$. With the appropriate assumption these equations can be solved for μ in terms of β to yield the following result.

COROLLARY 8. *Let $\dot{x} = f(x,\mu)$ be as in* (1) *with $d_b = 2$. Let U be a dichotomous variational solution along γ with $I_{uu} = \{1,2\}$, $I_{ss} = \{3,4\}$, $\left\langle u_4^\perp(0), \dot{\gamma}(0)\right\rangle \neq 0$. Define a 2×2 matrix $A = [a_{ij}]$ by*

$$a_{ij} = \frac{\partial H_i}{\partial \mu_j}(0,0) = -\int_{-\infty}^{\infty} \left\langle u_i^\perp, \frac{\partial f}{\partial \mu_j}(\gamma, 0)\right\rangle dt, \qquad i = 1,2, \quad j = 1,2$$

and a 2-vector $b = [b_i]$ by

$$b_i = \frac{\partial^2 H_i}{\partial \beta^2}(0,0) = -\int_{-\infty}^{\infty} \langle u_i^\perp, D_{11} f(\gamma, 0) u_3 u_3\rangle \, dt, \qquad i = 1,2.$$

If $\det(A) \neq 0$ *then there exists an open interval W containing zero and a differentiable function $\phi : W \to \mathcal{R}^2$ such that $\phi(0) = 0$, $\phi'(0) = 0$, $\phi''(0) = -A^{-1}b$ and such that* (1) *has a homoclinic solution for $\mu = \phi(\beta)$, $\beta \in W$.*

When this result is interpreted in terms of the μ's we obtain a curve in the μ_1–μ_2 plane passing through the origin with slope

$$m = \lim_{\beta \to 0} \frac{\mu_2(\beta)}{\mu_1(\beta)} = \frac{-a_{21}b_1 + a_{11}b_2}{a_{22}b_1 - a_{12}b_2}.$$

This is illustrated in Example 4 below.

The case of $d_b \geq 3$ is a bit harder. Our approach is, first, to set to zero the terms of lowest degree in $H(\beta, \mu)$.

DEFINITION 9. Let $\dot{x} = f(x,\mu)$ be as in (1). Let U be a dichotomous variational solution along γ numbered in such a way that $I_{uu} = \{1, \cdots, d_b\}$ together with a bijection $\alpha : \{1, \cdots, d_b\} \to I_{ss}$ such that $\left\langle u_{\alpha(d_b)}^\perp(0), \dot{\gamma}(0)\right\rangle \neq 0$. Let $V \subset \mathcal{R}^{d_b - 1} \times \mathcal{R}^2$ and $H : V \to \mathcal{R}^{d_b}$ be as in Theorem 5.

(i) We define

$$a_{ij} = \frac{\partial H_i}{\partial \mu_j}(0,0) = -\int_{-\infty}^{\infty} \left\langle u_i^\perp, \frac{\partial f}{\partial \mu_j}(\gamma, 0)\right\rangle dt$$

$$\text{for } i = 1, \cdots, d_b, \quad j = 1,2.$$

$$b_{ijk} = \frac{\partial^2 H_i}{\partial \beta_j \partial \beta_k}(0,0) = -\int_{-\infty}^{\infty} \langle u_i^\perp, D_{11} f(\gamma, 0) u_{\alpha(j)} u_{\alpha(k)}\rangle \, dt$$

$$\text{for } i = 1, \cdots, d_b, \quad j,k = 1, \cdots, d_b - 1.$$

(ii) We define a second degree map $M : \mathcal{R}^{d_b - 1} \times \mathcal{R}^2 \to \mathcal{R}^{d_b}$ by

$$M_i(\beta, \mu) = \sum_{j=1}^{2} a_{ij}\mu_j + \frac{1}{2} \sum_{j=1}^{d_b - 1} \sum_{k=1}^{d_b - 1} b_{ijk}\beta_j\beta_k.$$

(iii) We shall say that $(\beta^0, \mu^0) \in \mathcal{R}^{d_b - 1} \times \mathcal{R}^2$ is a characteristic vector for (1) if $M(\beta^0, \mu^0) = 0$.

If (β^0, μ^0) is a characteristic vector for (1) then $(s\beta^0, s^2\mu^0)$ is a characteristic vector for (1) for all scalars s. We can specify a choice for s by requiring $\|\beta^0\|^2 + \|\mu^0\| = 1$. We now use the implicit function theorem to find a solution to the exact equation $H(\beta, \mu) = 0$.

THEOREM 10. *Let $\dot{x} = f(x, \mu)$ be as in* (1). *Let a_{ij}, b_{ijk}, and $M(\beta, \mu)$ be as in Definition* 9. *Let (β^0, μ^0) be a characteristic vector for* (1) *and define a $d_b \times (d_b + 1)$ matrix $C = [c_{ij}]$ by*

$$\begin{aligned} c_{ij} &= \sum_{k=1}^{d_b-1} b_{ijk}\beta_k^0, && i = 1, \cdots, d_b, \quad j = 1, \cdots, d_b - 1, \\ c_{ij} &= a_{i,j-d_b+1}, && i = 1, \cdots, d_b, \quad j = d_b, \ d_b + 1. \end{aligned}$$

If C has maximal rank there exists an open interval W containing zero and a differentiable function $\phi : W \to \mathcal{R}^2$ with $\phi(0) = 0$ such that $\dot{x} = f(x, \mu)$ has a homoclinic solution for $\mu = s^2(\mu^0 + \phi(s))$, $s \in W$.

Proof. Let $V \subset \mathcal{R}^{d_b-1} \times \mathcal{R}^2$ and $H : V \to \mathcal{R}^{d_b}$ be as in Theorem 5. Let $V_1 \subset \mathcal{R}^{d_b-1} \times \mathcal{R}^2 \times \mathcal{R}$ be a sufficiently small open neighborhood of $(0,0)$ and define a differentiable function $F : V_1 \to \mathcal{R}^{d_b}$ by

$$F(\psi, \phi, s) = \begin{cases} \dfrac{1}{s^2} H(s\beta^0 + s\psi, s^2\mu^0 + s^2\phi) & \text{for } s \neq 0, \\ M(\beta^0 + \psi, \mu^0 + \phi) & \text{for } s = 0. \end{cases}$$

Observe that $F(0) = 0$ and $\tilde{D}F(0) = C$, where $\tilde{D}$ denotes differentiation with respect to (ψ, ϕ). By hypothesis, C has maximal rank which means that it has an invertible $d_b \times d_b$ submatrix. One of the columns of this submatrix must correspond to one of the components ϕ_1, ϕ_2 of ϕ. Assume the component is ϕ_1. Then from the implicit function theorem we can solve the equation $F(\psi_1, \cdots, \psi_{d_b-1}, \phi_1, 0, s) = 0$ for ψ and ϕ_1 in terms of s. The result now follows from Theorem 5. □

Note that Theorem 10 can yield multiple curves in the μ_1–μ_2 plane along which we have a homoclinic solution. For each (β^0, μ^0) we get a curve passing through the origin with slope $m = \mu_2^0 / \mu_1^0$. This is illustrated in Example 5.

Case of a manifold of homoclinic orbits. The preceding results utilize the second derivatives of H with respect to β; when these derivatives are zero we need an alternate approach. One situation worth considering occurs when all or part of $W^s \cap W^u$ is a manifold which we can term the bistable manifold, W^b. This case arises in certain integrable Hamiltonian systems [19].

Suppose $W^s \cap W^u$ has a branch which is a manifold, denoted W^b, of dimension d_b; let $P : \mathcal{R}^{d_b-1} \to W^b$ be differentiable; and let γ_θ denote a homoclinic solution for $\dot{x} = f(x, 0)$ satisfying $\gamma_\theta(0) = P(\theta)$. Assume P is constructed so that $(\theta, t) \to \gamma_\theta(t)$ establishes local coordinates on W^b. Since

$$\dot{\gamma}_\theta(t) = f(\gamma_\theta(t), 0),$$

differentiation with respect to θ yields

$$\frac{\partial \dot{\gamma}_\theta}{\partial \theta_i}(t) = D_1 f(\gamma_\theta(t), 0) \frac{\partial \gamma_\theta}{\partial \theta_i}(t).$$

Thus, the functions $\partial \gamma_\theta / \partial \theta_j$, along with $\dot{\gamma}_\theta$ provide a natural set of solutions of type ss to the variational equation along γ_θ.

Let U_θ denote a dichotomous variational solution along γ_θ with $u_{\theta j}$ the jth column of U_θ, and define $u_{\theta i}^\perp$ by $\langle u_{\theta i}^\perp(t), u_{\theta j}(t) \rangle = \delta_{ij}$.

Also, let $A(\theta) = [a_{ij}(\theta)]$ denote the $d_b \times 2$ matrix defined as

$$a_{ij}(\theta) = -\int_{-\infty}^{\infty} \left\langle u_{\theta i}^\perp, \frac{\partial f}{\partial \mu_j}(\gamma_\theta, 0) \right\rangle dt.$$

THEOREM 11. *Let $\dot{x} = f(x,\mu)$ be as in* (1). *Suppose $W^s \cap W^u$ has a branch, W^b, which is a manifold and let $A(\theta)$ be as above. Suppose there exist $\theta^0 \in \mathcal{R}^{d_b-1}$ and a nonzero $\mu^0 \in \mathcal{R}^2$ such that $A(\theta^0)\mu^0 = 0$. Define a $d_b \times (d_b+1)$ matrix $C = [c_{ij}]$ by*

$$c_{ij} = \begin{cases} \sum_{k=1}^2 \dfrac{\partial a_{ik}}{\partial \theta_j}(\theta^0)\mu_k^0 & \textit{for } i = 1, \cdots, d_b, \quad j = 1, \cdots, d_b - 1 \\ a_{i,j-d_b+1}(\theta^0) & \textit{for } i = 1, \cdots, d_b, \quad j = d_b,\ d_b+1. \end{cases}$$

If C has maximal rank there exists an open interval W containing zero and a differentiable function $\phi : W \to \mathcal{R}^2$ with $\phi(0) = 0$ such that (1) *has a homoclinic solution for $\mu = s(\mu^0 + \phi(s))$, $s \in W$.*

Proof. We begin by modifying the proof of Theorem 5. We define a map F : $\tilde{\mathcal{Z}}^1 \times \mathcal{R}^{d_b-1} \times \mathcal{R}^2 \to \mathcal{Z}^0$ by

$$F(z,\theta,\mu)(t) = \dot{\gamma}_\theta(t) + \dot{z}(t) - f(\gamma_\theta(t) + z(t), \mu).$$

In other words, we replace the β coordinates in Theorem 5 with the local manifold coordinates θ. Proceeding as before we obtain a connected open set $V \subset \mathcal{R}^{d_b-1} \times \mathcal{R}^2$ with $(0,0) \in V$ and a differentiable function $H : V \to \mathcal{R}^{d_b}$ denoted $(\theta,\mu) \to H(\theta,\mu)$ with the following properties:

(i) If $H(\theta^*,\mu^*) = 0$, then $\dot{x} = f(x,\mu^*)$ has a homoclinic solution;

(ii) $H(\theta,0) = 0$;

(iii) $$\frac{\partial H_i}{\partial \mu_j}(\theta,0) = -\int_{-\infty}^{\infty} \left\langle u_{\theta i}^{\perp}, \frac{\partial f}{\partial \mu_j}(\gamma_\theta, 0) \right\rangle dt = a_{ij}(\theta).$$

The next step is to modify the proof of Theorem 10. For a sufficiently small neighborhood $V_1 \subset \mathcal{R}^{d_b-1} \times \mathcal{R}^2 \times \mathcal{R}$ we define a differentiable map $G : V_1 \to \mathcal{R}^{d_b}$ by

$$G(\psi,\phi,s) = \begin{cases} \dfrac{1}{s} H(\theta^0 + \psi, s\mu^0 + s\phi) & \text{for } s \neq 0, \\ A(\theta^0 + \psi)(\mu^0 + \phi) & \text{for } s = 0. \end{cases}$$

We have $G(0) = A(\theta^0)\mu^0 = 0$ and $\tilde{D}G(0) = C$, where $\tilde{D}$ denotes differentiation with respect to (ψ,ϕ). As in the proof of Theorem 10 we can, after renumbering if necessary, use the implicit function theorem to solve the equation $G(\psi_1, \cdots, \psi_{d_b-1}, \phi_1, 0, s) = 0$ for ψ and ϕ_1 in terms of s. □

This result is very similar to Theorem 10 and generalizes Theorem 1 of [19]. In terms of parameter space we have, for each (θ^0, μ^0), a curve in the μ_1–μ_2 plane passing through the origin with slope $m = \mu_2^0/\mu_1^0$. Note that here there is no restriction on the number of resulting curves. Theorem 11 is illustrated in Examples 6 and 7 below with, respectively, three and five curves.

Examples. We now proceed to illustrate the above theory with a number of examples. We begin with a well-known example in $\mathcal{R}^2$ and consider various generalizations to higher dimensions. In all the following calculations we adopt the notation $r(t) = \operatorname{sech} t$. Note that we have $\ddot{r} = r - 2r^3$. The following numerical values and defined functions will be needed.

$$\int_{-\infty}^{\infty} r^2\,dt = 2, \qquad \int_{-\infty}^{\infty} r^3\,dt = \frac{\pi}{2},$$

$$\int_{-\infty}^{\infty} \dot{r}^2\,dt = \frac{2}{3}, \qquad \int_{-\infty}^{\infty} r\dot{r}^2\,dt = \frac{\pi}{8}, \qquad \int_{-\infty}^{\infty} r^2\dot{r}^2\,dt = \frac{4}{15},$$

$$
\begin{aligned}
I_1(\theta) &= -\int_{-\infty}^{\infty} \dot r(t) r(t-\theta)\,dt = 2\left(\frac{\theta\cosh\theta - \sinh\theta}{\sinh^2\theta}\right),\\
I_2(\theta) &= \int_{-\infty}^{\infty} \dot r(t)\dot r(t-\theta)\,dt = 2\left(\frac{2\sinh\theta\cosh\theta - 2\theta - \theta\sinh^2\theta}{\sinh^3\theta}\right),\\
I_3(\theta) &= -\int_{-\infty}^{\infty} r(t)^2\dot r(t) r(t-\theta)\,dt\\
&= 2\left(\frac{2\sinh^3\theta - 3\sinh\theta\cosh^2\theta + 3\theta\cosh\theta}{\sinh^4\theta}\right).
\end{aligned}
$$

Example 1. Consider the equation

$$\ddot x = x - 2x^3 + \mu_1\dot x(1-x^2) - \mu_2\dot x \tag{9}$$

which we regard as a first-order system in phase space $(x,\dot x)$. This equation has been studied by Holmes and Rand [14]. Similar equations appear in [1], [2, p. 280], and [4]. In meteorology this equation is sometimes referred to as the Saltzman oscillator [24].

Equation (9) has a hyperbolic fixed point at $(0,0)$ for all sufficiently small $|\mu|$. When $\mu = 0$ we get the familiar Duffing's equation with negative stiffness, which has two homoclinic solutions. We will consider the one given by $x = r$.

In the notation of Corollary 6 we compute $a_1 = -\frac{2}{5}$ and $a_2 = \frac{2}{3}$. Thus, the corollary applies so there exists a function ϕ with $\phi(0) = 0$ and $\phi'(0) = \frac{5}{3}$ such that (9) has a homoclinic solution when $\mu_1 = \phi(\mu_2)$.

Example 2. We now wish to generalize to $\mathcal{R}^3$ the equation of the preceding example. We do this by using the μ_2 term to couple the equation to an additional first-order equation which maintains the hyperbolic equilibrium. Thus consider

$$
\begin{aligned}
\ddot x &= x - 2x^3 + \mu_1\dot x(1-x^2) + \mu_2(\dot y - \dot x),\\
\dot y &= -2y + \mu_2(\dot x - \dot y).
\end{aligned}
$$

Letting $x_1 = x,\ x_2 = \dot x,\ x_3 = y$ we get

$$
\begin{aligned}
\dot x_1 &= x_2,\\
\dot x_2 &= x_1 - 2x_1^3 + \mu_1 x_2(1-x_1^2) - \left(\frac{2\mu_2}{1+\mu_2}\right)x_3 - \left(\frac{\mu_2}{1+\mu_2}\right)x_2,\\
\dot x_3 &= \left(\frac{\mu_2}{1+\mu_2}\right)x_2 - \left(\frac{2}{1+\mu_2}\right)x_3.
\end{aligned}
$$

The eigenvalues of $D_1 f(0,0)$ are $\lambda_1 = -2$, $\lambda_2 = -1$, $\lambda_3 = 1$. When $\mu = 0$ this system has a homoclinic solution given by $\gamma = (r, \dot r, 0)$. For a dichotomous variational solution along γ we first take $u_2 = \dot\gamma = (\dot r, \ddot r, 0)$. Using variation of parameter we obtain a second solution of the form $u_1 = (P\dot r, (P\dot r)^{\cdot}, 0)$, where P is a differentiable function which satisfies $(P\dot r)^{\cdot}\dot r - P\dot r\ddot r = \dot P\dot r^2 = 1$, an arbitrary constant.

For a third solution we have $u_3 = (0, 0, e^{-2t})$.

Notice that u_1 connects λ_3 to λ_2 (type uu), u_2 connects λ_2 to λ_3 (type ss), and u_3 connects λ_1 to λ_1 (type su). Thus, we have a dichotomous variational solution, necessarily, with $n_{ss} = d_b = 1$.

Using $u_1^{\perp} = (-\ddot{r}, \dot{r}, 0)$ we compute $a_1 = -\frac{2}{5}$ and $a_2 = \frac{2}{3}$ so Corollary 7 applies. There exists a function ϕ such that $\phi(0) = 0$, $\phi'(0) = \frac{5}{3}$ and such that the dynamical system has a homoclinic solution for $\mu_1 = \phi(\mu_2)$.

Example 3. The preceding example can easily be generalized to higher dimension. For example, consider the equations

$$\begin{aligned}\ddot{x} &= x - 2x^3 + \mu_1\dot{x}(1 - x^2) + \mu_2(\dot{y} - \dot{x}),\\ \ddot{y} &= y + \mu_2(\dot{x} - \dot{y}).\end{aligned}$$

The phase space for this system is $\mathcal{R}^4$ with coordinates $(x, \dot{x}, y, \dot{y})$. The eigenvalues of $D_1 f(0,0)$ are $\lambda_1 = -1$, $\lambda_2 = -1$, $\lambda_3 = 1$, $\lambda_4 = 1$ and, when $\mu = 0$, a homoclinic solution is given by $\gamma = (r, \dot{r}, 0, 0)$. A dichotomous variational solution is given by

$u_1 = (P\dot{r}, (P\dot{r})^{\cdot}, 0, 0),$	connects λ_3 to λ_2,	type uu,
$u_2 = (\dot{r}, \ddot{r}, 0, 0),$	connects λ_2 to λ_3,	type ss,
$u_3 = (0, 0, e^{-t}, -e^{-t}),$	connects λ_1 to λ_1,	type su,
$u_4 = (0, 0, e^{t}, e^{t}),$	connects λ_2 to λ_2,	type us,

where P is as in the preceding example. Things are numbered so that $I_{uu} = \{1\}$, $I_{ss} = \{2\}$, and $u_2 = \dot{\gamma}$. Since $n_{ss} = 1$, $d_b = 1$, and Corollary 7 still applies. A straightforward calculation with $u_1^{\perp} = (-\ddot{r}, \dot{r}, 0, 0)$ yields the same values for each a_i as in the previous examples.

Example 4. We now turn to some examples with $d_b = 2$. The most natural way to achieve this is to replace the second equation in Example 3 with a Duffing's equation. The problem with this is that then $W^s \cap W^u$ is a manifold which is a higher degree of degeneracy. Such a system is considered in Example 7 below. For now we consider

$$\begin{aligned}\ddot{x} &= x - \tfrac{4}{3}x^3 - \tfrac{2}{3}y^3 + \mu_1(x + a\dot{x}),\\ \ddot{y} &= y - \tfrac{4}{3}y^3 - \tfrac{2}{3}x^3 + \mu_2(y + b\dot{y}),\end{aligned}$$

where a, b are constants.

A homoclinic solution, when $\mu = 0$, is given by $x = y = r$. We work in the phase space $(x, \dot{x}, y, \dot{y})$ and find the eigenvalues of $D_1 f(0,0)$ to be $\lambda_1 = -1$, $\lambda_2 = -1$, $\lambda_3 = 1$, $\lambda_4 = 1$. Letting $u = (v, \dot{v}, w, \dot{w})$ denote a typical solution to the variational equation we get

$$\begin{aligned}\ddot{v} &= v - 4r^2 v - 2r^2 w,\\ \ddot{w} &= w - 4r^2 w - 2r^2 v.\end{aligned}$$

One variational solution is given by $v = w = \dot{r}$, and variation of parameter leads to a second of the form $v = w = P\dot{r}$, where P is as in Example 2. A third solution is given by $v = -w = r$ and, once again turning to variation of parameter, a solution $v = -w = Qr$, where Q satisfies $(Qr)^{\cdot} r - Qr\dot{r} = \dot{Q}r^2 = 1$, an arbitrary constant.

We now have a dichotomous variational solution:

$u_1 = (Qr, (Qr)^{\cdot}, -Qr, -(Qr)^{\cdot}),$	connects λ_3 to λ_2,	type uu,
$u_2 = (P\dot{r}, (P\dot{r})^{\cdot}, P\dot{r}, (P\dot{r})^{\cdot}),$	connects λ_4 to λ_1,	type uu,
$u_3 = (r, \dot{r}, -r, -\dot{r}),$	connects λ_2 to λ_3,	type ss,
$u_4 = (\dot{r}, \ddot{r}, \dot{r}, \ddot{r}),$	connects λ_1 to λ_4,	type ss.

The u_i's have been numbered so that $I_{uu} = \{1, 2\}$, $I_{ss} = \{3, 4\}$, and $u_4 = \dot{\gamma}$. We see that $n_{ss} = 2$ which implies that $d_b = 2$. In the notation of Corollary 8 we compute

$$\begin{aligned} u_1^{\perp} &= \tfrac{1}{2}(-\dot{r}, r, \dot{r}, -r), \\ u_2^{\perp} &= \tfrac{1}{2}(-\ddot{r}, \dot{r}, -\ddot{r}, \dot{r}) \end{aligned}$$

and $a_{11} = -1$, $a_{12} = 1$, $a_{21} = -a/3$, $a_{22} = -b/3$, $b_1 = 2\pi$, $b_2 = 0$.

Since $\det(A) = (a + b)/3$ and

$$\langle u_4^{\perp}(0), \dot{\gamma}(0)\rangle = \langle u_4^{\perp}(0), u_4(0)\rangle = 1,$$

Corollary 8 applies as long as $a + b \neq 0$. In this case, there exists an interval, W, containing zero and a function $\phi : W \to \mathcal{R}^2$ with $\phi(0) = 0$, $\phi'(0) = 0$, $\phi''(0) = \frac{2\pi}{a+b}(b, -a)$ such that the dynamical system has a homoclinic solution when $\mu = \phi(\beta)$. We have a curve in the μ_1–μ_2 plane passing through the origin with slope $m = \phi_2''(0)/\phi_1''(0) = -a/b$.

Example 5. To obtain a value of $d_b = 3$ requires a minimum of $n = 6$ and, hence, the possibility of extensive calculations. We present a model problem which illustrates the principle. Consider the system:

$$\begin{aligned} \ddot{x} &= x - 2xz^2 + \dot{x}^2 + \mu_1 z, \\ \ddot{y} &= y - 2yz^2 + \dot{x}\dot{y}, \\ \ddot{z} &= z - 2z^3 + y\dot{y} + \mu_2 \dot{z}. \end{aligned}$$

We work in the phase space $(x, \dot{x}, y, \dot{y}, z, \dot{z})$ and easily compute the eigenvalues of $D_1 f(0, 0)$ to be $\lambda_1 = \lambda_2 = \lambda_3 = -1$, $\lambda_4 = \lambda_5 = \lambda_6 = 1$. A homoclinic solution when $\mu = 0$ is given by $x = y = 0$, $z = r$, i.e., $\gamma = (0, 0, 0, 0, r, \dot{r})$. The variational equation along γ uncouples so that it is easy to compute the following dichotomous variational solution:

$$\begin{array}{lll} u_1 = (Qr, (Qr)^{\cdot}, 0, 0, 0, 0), & \text{connects } \lambda_4 \text{ to } \lambda_1, & \text{type uu,} \\ u_2 = (0, 0, Qr, (Qr)^{\cdot}, 0, 0), & \text{connects } \lambda_5 \text{ to } \lambda_1, & \text{type uu,} \\ u_3 = (0, 0, 0, 0, P\dot{r}, (P\dot{r})^{\cdot}), & \text{connects } \lambda_6 \text{ to } \lambda_3, & \text{type uu,} \\ u_4 = (r, \dot{r}, 0, 0, 0, 0), & \text{connects } \lambda_1 \text{ to } \lambda_4, & \text{type ss,} \\ u_5 = (0, 0, r, \dot{r}, 0, 0), & \text{connects } \lambda_2 \text{ to } \lambda_5, & \text{type ss,} \\ u_6 = (0, 0, 0, 0, \dot{r}, \ddot{r}), & \text{connects } \lambda_3 \text{ to } \lambda_6, & \text{type ss.} \end{array}$$

The functions P and Q are as in earlier examples. Since u_4, u_5, and u_6 are each of type ss we have $d_b = n_{ss} = 3$. Things are numbered so that $I_{uu} = \{1, 2, 3\}$, $I_{ss} = \{4, 5, 6\}$, and $u_6 = \dot{\gamma}$. We take $\alpha(1) = 4$, $\alpha(2) = 5$, $\alpha(3) = 6$ so then $\langle u_{\alpha(3)}^{\perp}(0), \dot{\gamma}(0)\rangle = \langle u_6^{\perp}(0), u_6(0)\rangle = 1$.

Using

$$\begin{aligned} u_1^{\perp} &= (-\dot{r}, r, 0, 0, 0, 0), \\ u_2^{\perp} &= (0, 0, -\dot{r}, r, 0, 0), \\ u_3^{\perp} &= (0, 0, 0, 0, -\ddot{r}, \dot{r}), \end{aligned}$$

the equations $\sum a_{ij}\mu_j + \frac{1}{2}\sum\sum b_{ijk}\beta_j\beta_k = 0$ in Definition 9 become:

$$\begin{aligned} -2\mu_1 - \frac{\pi}{8}{\beta_1}^2 &= 0, \\ -\frac{\pi}{8}\beta_1\beta_2 &= 0, \\ -\tfrac{2}{3}\mu_2 - \frac{\pi}{8}{\beta_2}^2 &= 0. \end{aligned}$$

We see that these equations yield two characteristic vectors which, in the notation of Theorem 10, are summarized below.

$$\text{(i)} \qquad \beta^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mu^0 = \begin{pmatrix} -\frac{\pi}{16} \\ 0 \end{pmatrix}, \qquad C = \begin{pmatrix} -\frac{\pi}{4} & 0 & -2 & 0 \\ 0 & -\frac{\pi}{8} & 0 & 0 \\ 0 & 0 & 0 & -\frac{2}{3} \end{pmatrix},$$

$$\text{(ii)} \qquad \beta^0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \mu^0 = \begin{pmatrix} 0 \\ -\frac{3\pi}{16} \end{pmatrix}, \qquad C = \begin{pmatrix} 0 & 0 & -2 & 0 \\ -\frac{\pi}{8} & 0 & 0 & 0 \\ 0 & -\frac{\pi}{4} & 0 & -\frac{2}{3} \end{pmatrix}.$$

In each case $\text{rank}(C) = 3$ so there are two curves through the origin in the μ_1-μ_2 plane along which the perturbed system has a homoclinic solution. The curves have slopes 0 and ∞; that is, each curve is tangent to one of the axes.

Example 6. Example 4 used reflective symmetry to obtain $d_b = 2$. We now use rotational symmetry. Consider:

$$\begin{aligned} \ddot{x} &= x + 2x(x^2 + y^2) + \mu_1(x + \dot{x}), \\ \ddot{y} &= y + 2y(x^2 + y^2) + \mu_2(y - \dot{y}). \end{aligned}$$

A similar example is given in [19]. When $\mu = 0$ the eigenvalues of the origin are $\lambda_1 = -1$, $\lambda_2 = -1$, $\lambda_3 = 1$, $\lambda_4 = 1$ and the system has a homoclinic solution given by $x(t) = r(t)\cos\theta$, $y(t) = r(t)\sin\theta$ for all θ. We have a 2-manifold of homoclinic solutions so, necessarily, $d_b = 2$.

The necessary calculations are simplified by introduction of the rotated coordinates $\tilde{x} = x\cos\theta + y\sin\theta$, $\tilde{y} = -x\sin\theta + y\cos\theta$. In the phase space $(\tilde{x}, \dot{\tilde{x}}, \tilde{y}, \dot{\tilde{y}})$ we have $\gamma_\theta = (r, \dot{r}, 0, 0)$ and $\partial\gamma/\partial\theta = (0, 0, r, \dot{r})$. A dichotomous variational solution is given by

$$\begin{aligned} &u_1 = (0, 0, Qr, (Qr)^{\cdot}), && \text{connects } \lambda_3 \text{ to } \lambda_2, && \text{type uu}, \\ &u_2 = (P\dot{r}, (P\dot{r})^{\cdot}, 0, 0), && \text{connects } \lambda_4 \text{ to } \lambda_1, && \text{type uu}, \\ &u_3 = (0, 0, r, \dot{r}), && \text{connects } \lambda_2 \text{ to } \lambda_3, && \text{type ss}, \\ &u_4 = (\dot{r}, \ddot{r}, 0, 0), && \text{connects } \lambda_1 \text{ to } \lambda_4, && \text{type ss}. \end{aligned}$$

In the notation preceding Theorem 11 we use

$$\begin{aligned} u_1^{\perp} &= (0, 0, -\dot{r}, r), \\ u_2^{\perp} &= (-\ddot{r}, \dot{r}, 0, 0), \end{aligned}$$

and compute $a_{11} = \sin 2\theta$, $a_{12} = -\sin 2\theta$, $a_{21} = -\frac{2}{3}\cos^2\theta$, $a_{22} = \frac{2}{3}\sin^2\theta$. We find that the equation $A(\theta)\mu^0 = 0$ has a nontrivial solution for μ^0 when $\det(A(\theta)) = -\frac{2}{3}\sin 2\theta \cos 2\theta = 0$. We get three distinct results for $-\pi \leq \theta \leq \pi$, summarized as follows:

$$\text{(i)} \qquad \theta^0 = \pm\tfrac{\pi}{2}; \qquad \mu^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \qquad C = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 0 & \frac{2}{3} \end{pmatrix},$$

$$\text{(ii)} \qquad \theta^0 = 0, \pi; \qquad \mu^0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \qquad C = \begin{pmatrix} -2 & 0 & 0 \\ 0 & -\frac{2}{3} & 0 \end{pmatrix},$$

$$\text{(iii)} \qquad \theta^0 = \pm\tfrac{\pi}{4}, \mp\tfrac{3\pi}{4}; \qquad \mu^0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}; \qquad C = \begin{pmatrix} 0 & \pm 1 & \mp 1 \\ \pm\frac{4}{3} & -\frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

In each case $\operatorname{rank}(C) = 2$ so Theorem 11 can be applied three times. The result is three curves in the μ_1–μ_2 plane along which the system has a homoclinic solution. There is a curve tangent to each axis and one with a slope of one.

Example 7. When a system in $\mathcal{R}^4$ is such that the unperturbed system uncouples into a pair of equations, each with a homoclinic solution, there is always a 2-manifold of homoclinic solutions. For example, consider

$$\begin{aligned} \ddot{x} &= x - 2x^3 + \mu_1 y + \mu_2(3\dot{y} - \dot{x}), \\ \ddot{y} &= y - 2y^3 + \mu_1 x + \mu_2(\dot{x} + 2\dot{y}). \end{aligned}$$

The phase space for this system is $(x, \dot{x}, y, \dot{y})$. When $\mu = 0$ the eigenvalues are $\lambda_1 = -1$, $\lambda_2 = -1$, $\lambda_3 = 1$, $\lambda_4 = 1$, and a 2-manifold of homoclinic solutions is given by $x(t) = r(t)$, $y(t) = r(t - \theta)$. Using the function P from Example 2 we get a dichotomous variational solution:

$$\begin{aligned} u_{\theta 1}(t) &= ((P\dot{r})(t), (P\dot{r})^{\cdot}(t), 0, 0), && \lambda_3 \text{ to } \lambda_2, && \text{type uu}, \\ u_{\theta 2}(t) &= (0, 0, (P\dot{r})(t - \theta), (P\dot{r})^{\cdot}(t - \theta)), && \lambda_4 \text{ to } \lambda_1, && \text{type uu}, \\ u_{\theta 3}(t) &= (\dot{r}(t), \ddot{r}(t), 0, 0), && \lambda_2 \text{ to } \lambda_3, && \text{type ss}, \\ u_{\theta 4}(t) &= (0, 0, \dot{r}(t - \theta), \ddot{r}(t - \theta)), && \lambda_1 \text{ to } \lambda_4, && \text{type ss}. \end{aligned}$$

From this we get

$$\begin{aligned} u_{\theta 1}^{\perp}(t) &= (-\ddot{r}(t), \dot{r}(t), 0, 0), \\ u_{\theta 2}^{\perp}(t) &= (0, 0, -\ddot{r}(t - \theta), \dot{r}(t - \theta)). \end{aligned}$$

For the purposes of Theorem 11 we compute

$$\begin{aligned} A(\theta) &= \begin{pmatrix} I_1(\theta) & \frac{2}{3} - 3I_2(\theta) \\ -I_1(\theta) & -\frac{4}{3} - I_2(\theta) \end{pmatrix}, \\ \frac{dA}{d\theta}(\theta) &= \begin{pmatrix} I_1(\theta) & -3(I_1(\theta) - 6I_3(\theta)) \\ -I_2(\theta) & -I_1(\theta) + 6I_3(\theta) \end{pmatrix}. \end{aligned}$$

We have introduced the definitions preceding Example 1 and have used the relationships $dI_1/d\theta = I_2$, $dI_2/d\theta = I_1 - 6I_3$. Application of Theorem 11 requires the solution of the equation $\det(A(\theta)) = -\frac{2}{3}I_1(\theta)(6I_2(\theta) + 1) = 0$. It is easy to check that

$I_1(\theta) = 0$ has the solution $\theta = 0$ and a little computer work shows that this is the only solution. The equation $6I_2(\theta) + 1 = 0$ can be shown by computer calculation to have four solutions. The resulting five solutions to $A(\theta^0)\mu^0 = 0$ are:

$$\text{(i)} \qquad \theta^0 = 0, \qquad \mu^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$\text{(ii)} \qquad \theta^0 = \pm 2.199071, \qquad \mu^0 = \begin{pmatrix} .9005686 \\ \mp .4347139 \end{pmatrix},$$

$$\text{(iii)} \qquad \theta^0 = \pm 3.741983, \qquad \mu^0 = \begin{pmatrix} .9759556 \\ \mp .2179694 \end{pmatrix}.$$

In each case a computer calculation determines that $\operatorname{rank}(C) = 2$ so Theorem 12 can be applied five times. In the μ_1–μ_2 plane we have five curves through the origin along which the system has a homoclinic solution. One curve is tangent to the μ_1–axis, the others have slopes ±.4827105 and ±.2233395.

Acknowledgments. This work has benefitted greatly from the generous assistance of Stephen Schecter who read countless earlier drafts and pointed out errors and means of improvement. In addition, the SIAM referee provided two sets of useful comments including an additional reference.

REFERENCES

[1] A. A. Andronov, E. A. Leontovich, I. I. Gordon, and A. G. Maier, *Theory of Bifurcations of Dynamical Systems on A Plane*, John Wiley, New York, 1973.

[2] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[3] F. Brauer and A. C. Soudak, *Stability regions and transition phenomena for harvested predator-prey systems*, J. Math. Biol., 8 (1979), pp. 319–337.

[4] J. Carr, *Applications of Center Manifold Theory*, Springer-Verlag, New York, 1981.

[5] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[6] S. N. Chow, J. K. Hale, and J. Mallet-Paret, *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37 (1980), pp. 351–373.

[7] D. S. Cohen and E. A. Stanley, *Gaseous diffusion in glassy polymers*, SIAM J. Appl. Math., 43 (1983), pp. 949–970.

[8] G. B. Ermentrout and J. D. Cowan, *Temporal oscillations in neuronal nets*, Math. Biol., 7 (1979), pp. 265–280.

[9] A. Gaalswyk, *Limit behavior and the existence of combustion shock layers*, Stud. in Appl. Math., 67 (1982), pp. 141–168.

[10] M. Ghil and J. Tavantzis, *Global Hopf-bifurcation in a simple climate model*, SIAM J. Appl. Math., 43 (1983), pp. 1019–1041.

[11] J. R. Gruendler, *The existence of homoclinic orbits and the method of Melnikov for systems in $\mathcal{R}^n$*, SIAM J. Math. Anal., 16 (1985), pp. 907–931.

[12] S. P. Hastings, *On the existence of homoclinic and periodic orbits for the Fitzhugh–Nagumo equations*, Quart. J. Math. Oxford, 27 (1976), pp. 123–134.

[13] ———, *Single and multiple pulse waves for the Fitzhugh–Nagumo equations*, SIAM J. Appl. Math., 42 (1982), pp. 247–260.

[14] P. Holmes and D. Rand, *Phase portraits and bifurcations of the non-linear oscillator $\ddot{x}+(a+\gamma x^2)\cdot\dot{x} + \beta x + \delta x^3 = 0$*, Internat. J. Non-Linear Mech., 15 (1980), pp. 449–458.

[15] P. Holmes and D. S. Stewart, *The existence of one dimensional steady detonation waves in a simple model problem*, Stud. in Appl. Math., 66 (1982), pp. 121–143.

[16] J. P. KEENER, *Chaotic behavior in slowly varying systems of nonlinear differential equations*, Stud. Appl. Math., 67 (1982), pp. 25–44.

[17] ———, *Infinite periodic bifurcation and global bifurcation branches*, SIAM J. Appl. Math, 41 (1981), pp. 127–144.

[18] N. KOPELL AND L. N. HOWARD, *Bifurcations and trajectories joining critical points*, Adv. in Math., 18 (1975), pp. 306–358.

[19] L. M. LERMAN AND IA. L. UMANSKII, *On the existence of separatrix loops in four-dimensional systems*, J. Appl. Math. Mech., 47 (1983), pp. 335–340.

[20] V. K. MELNIKOV, *On the stability of the center for time-periodic perturbations*, Trans. Moscow Math. Soc., 12 (1963), pp. 1–56.

[21] M. S. MOCK, *A topological degree for orbits connecting critical points of autonomous systems*, J. Differential Equations, 38 (1980), pp. 176–191.

[22] G. NICOLIS, *Bifucations and symmetry breaking in far-from-equilibrium systems: toward a dynamics of complexity*, Adv. in Chem. Phys., 55 (1984), pp. 177–199.

[23] K. J. PALMER, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225–256.

[24] B. SALTZMAN, *Structural stocastic stability of a simple auto-oscillatory climatic feedback system*, Atmos. Sci., 38 (1981), pp. 494–503.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 722-736, May 1992

009

DIFFERENTIAL OPERATORS AND THE LAGUERRE TYPE POLYNOMIALS*

W. N. EVERITT†, A. M. KRALL‡, L. L. LITTLEJOHN§, AND V. P. ONYANGO-OTIENO¶

Abstract. In 1940, all fourth-order differential equations which have a sequence of orthogonal polynomial eigenfunctions were classified by H. L. Krall, up to a linear change of variable. One of these equations was subsequently named the Laguerre type equation and various properties of the orthogonal polynomial solutions and the right-definite boundary value problem were studied by A. M. Krall in 1981. In this paper, the Laguerre type expression is further studied in the right-definite setting and the appropriate left-definite problem associated with the fourth-order Laguerre type differential expression is discussed in detail.

Key words. orthogonal polynomials, differential equations, right-definite boundary value problems, left-definite boundary value problems, self-adjoint operators

AMS(MOS) subject classifications. 33A65, 34B20, 41A10

1. Introduction. The existence of the Laguerre type differential equation and the corresponding orthogonal polynomial solutions was first observed by H. L. Krall in 1940; see [15] and [16] for the methods used by Krall to discover this and other differential equations having orthogonal polynomial solutions. Of course, these polynomials may be constructed by using the standard definition given in Szegö [23, §§ 2.1 and 2.2] with respect to the nondecreasing function $\hat{\sigma}$ where, for some given positive number A,

$$\hat{\sigma}(x)=\begin{cases}-1/A & \text{if } x\in(-\infty,0],\\ 1-e^{-x} & \text{if } x\in(0,\infty).\end{cases}\tag{1.1}$$

Let σ denote the regular, nonnegative measure generated by $\hat{\sigma}$ on the Borel sets of the real line $\mathbb{R}$. Further, let $L^2_\sigma[0,\infty)$ denote the Hilbert function space derived from this measure σ, i.e., if $\mathbb{C}$ represents the complex field,

$$L^2_\sigma[0,\infty):=\left\{f\colon[0,\infty)\to\mathbb{C}\,|\,f \text{ is Borel measurable on } [0,\infty) \text{ and } \int_{[0,\infty)}|f(x)|^2\,d\sigma(x)<\infty\right\}\tag{1.2}$$

with inner product given by

$$(f,g)_\sigma:=\frac{f(0)\bar{g}(0)}{A}+\int_0^\infty f(x)\bar{g}(x)\,e^{-x}\,dx.\tag{1.3}$$

Note that in (1.3), the integral is Lebesgue in view of the fact that $\hat{\sigma}$ is locally absolutely continuous with respect to Lebesgue measure on the open interval $(0,\infty)$; i.e., the functions f and g belong to the weighted Hilbert space $L^2(0,\infty;e^{-x})$ defined by

$$L^2(0,\infty;e^{-x})=\left\{f\colon(0,\infty)\to\mathbb{C}\,|\,f \text{ is Lebesgue measurable and } \int_0^\infty|f(x)|^2\,e^{-x}\,dx<\infty\right\}.$$

Since the monomials $x^n\in L^2_\sigma[0,\infty)$ $(n=0,1,2,\cdots)$, the Laguerre type polynomials, denoted in this paper by $\{R_n(x;A)|x\in[0,\infty);\,n=0,1,2,\cdots\}$, may be defined

* Received by the editors August 20, 1990; accepted for publication (in revised form) September 3, 1991.

† Department of Mathematics, University of Birmingham, Birmingham, B15 2TT, England.
‡ Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802.
§ Department of Mathematics, Utah State University, Logan, Utah, 84322-3900.
¶ Department of Physical Sciences, Egerton University, P.O. Box 536, Njoro, Kenya.

through the Gram–Schmidt orthogonalization process; see [23, § 2.1]. For convenience, when the parameter $A>0$ is fixed, we shall write $\{R_n(x)\}$ or $\{R_n\}$ for this system of orthogonal polynomials. Alternatively, the Laguerre type polynomials may be found readily from the differential equation that they satisfy. Indeed, the Laguerre type polynomials are solutions of:

$$\begin{aligned} x^2y^{(4)}+(-2x^2+4x)y^{(3)}+(x^2-(2A+6)x)y'' \\ +((2A+2)x-2A)y'+ky=(\lambda_n+k)y, \end{aligned} \tag{1.4}$$

where $x\in(0,\infty)$, k is a fixed, nonnegative constant and

$$\lambda_n=n(n+2A+1), \qquad n=0,1,2,\cdots. \tag{1.5}$$

We note that (1.4) may be put into formally symmetric form when multiplied by e^{-x}:

$$M_k[y]=(\lambda_n+k)\,e^{-x}y,$$

where, for $x\in(0,\infty)$,

$$M_k[y](x):=(x^2\,e^{-x}y''(x))''-(((2A+2)x+2)\,e^{-x}y'(x))'+k\,e^{-x}y(x). \tag{1.6}$$

By using elementary power series techniques in (1.4), we can easily find an explicit formula for the Laguerre type polynomials. Indeed, we see that

$$R_n(x):=\sum_{j=0}^{n}\frac{(-1)^j}{(j+1)!}\binom{n}{j}((A+n+1)j+A)x^j, \qquad n=0,1,2,\cdots, \tag{1.7}$$

normalized so that $R_n(0)=A$ for all $n=0,1,2,\cdots$. These polynomials satisfy the orthogonality relation

$$\begin{aligned} (R_n,R_m)_\sigma=\int_{[0,\infty)}R_n(x)R_m(x)\,d\sigma(x)=(A+n+1)(A+n)\delta_{nm}, \\ n,m=0,1,2,\cdots, \end{aligned} \tag{1.8}$$

where δ_{nm} is the Kronecker delta function.

For general information concerning orthogonal polynomial solutions to differential equations of the form

$$\sum_{k=1}^{r}a_k(x)y^{(k)}(x)=\lambda y(x), \tag{1.9}$$

the reader is referred to the review articles [13] and [14]. Observe that, in (1.4) and (1.6), the spectral parameter λ and the degree of the polynomial solution do not appear in the coefficients on the left-hand side of the equations, a feature which is necessary for the study of the spectral properties of the differential expression (1.6). Furthermore, we note that it is a remarkable property that the Laguerre type polynomials satisfy a linear, fourth-order differential equation with this property. Certainly, it is not uncommon for orthogonal polynomials to satisfy differential equations (see, for example, [14]), but it is rare indeed that orthogonal polynomials satisfy a differential equation of the form (1.9), where the coefficients $a_k(x)$ are functions of x only. For example, only the classical orthogonal polynomials of Jacobi, Laguerre, and Hermite and the Bessel polynomials satisfy second-order differential equations of this type. In the fourth-order case, Krall [16] showed that the only differential equations of the form (1.9) with orthogonal polynomial solutions, apart from the formal squares of the known second-order equations, are the Legendre type, Laguerre type, and Jacobi type equations.

In the paper by A. M. Krall [12], the properties of the Laguerre type polynomials $\{R_n(x)\}$ are developed in detail. In particular, information is given on the three-term recurrence relation, the Rodrigues type formula, and the generating function for the polynomials. The paper [12] is also concerned with studying some of the properties of a certain self-adjoint operator generated by $M_k[\cdot]$ in the Hilbert space $L^2(0,\infty;\, e^{-x})\otimes\mathbb{C}$, which is isometrically isomorphic to $L^2_\sigma[0,\infty)$, having the Laguerre type polynomials as eigenfunctions.

The purpose of this present paper is to extend the ideas and methods in the earlier papers of Everitt, Krall, and Littlejohn [6], [7], and [12], to a study of the right-definite and left-definite spectral problems associated with the differential equation

$$M_k[y](x)=\lambda\, e^{-x}y(x) \qquad (x\in(0,\infty)).$$

The right-definite problem is studied in the Hilbert space $L^2_\sigma[0,\infty)$ and the left-definite problem in the Sobolev space $H^2_\sigma[0,\infty)$, where

$$\begin{aligned}H^2_\sigma[0,\infty)&:=\{f\colon [0,\infty)\to\mathbb{C}\,|\, f\in AC_{\text{loc}}[0,\infty);\, f'\in AC_{\text{loc}}(0,\infty);\\ &\qquad f,\,(x+1)^{1/2}f',\, xf''\in L^2(0,\infty,\, e^{-x})\}\end{aligned} \tag{1.10}$$

and the inner product is defined, for $f,\ g\in H^2_\sigma[0,\infty)$, by

$$(f,g)_H:=\int_0^\infty \{x^2\, e^{-x}f''(x)\bar{g}''(x)+((2A+2)x+2)\, e^{-x}f'(x)\bar{g}'(x)\}\,dx+k(f,g)_\sigma. \tag{1.11}$$

We are able to show that there is a self-adjoint representation of both the right-definite and left-definite problems for which, respectively, the Laguerre type polynomials $\{R_n(x)\}$ are determined as a complete set of eigenfunctions in the spaces $L^2_\sigma[0,\infty)$ and $H^2_\sigma[0,\infty)$.

Earlier work on the spectral representation of some of the classical orthogonal polynomials is given in Titchmarsh [24, Chap. IV], together with results which connect the Titchmarsh-Weyl m-coefficient theory with self-adjoint differential operators as given by Chaudhuri and Everitt [2]. The ideas and methods of Titchmarsh were extended to cover the right-definite and left-definite problems for the classical Laguerre orthogonal polynomials by Onyango-Otieno [19] and [20].

The contents of the paper are as follows. In § 2, we develop the properties of the maximal domain Δ_k in $L^2(0,\infty;\, e^{-x})$ associated with the differential expression $M_k[\cdot]$. In §§ 3 and 4, we establish the self-adjointness of the right-definite operator $T_k[\cdot]$ in the right-definite space $L^2_\sigma[0,\infty)$ and determine explicitly the spectrum $\sigma(T_k)$ of $T_k[\cdot]$. By taking advantage of the spectral properties of $T_k[\cdot]$ in $L^2_\sigma[0,\infty)$ we show, in § 5, that all self-adjoint operators generated by $M_k[\cdot]$ in $L^2(0,\infty;\, e^{-x})$ have discrete spectrums that are bounded below. Section 6 contains a proof of the density of the Laguerre type polynomials in the left-definite space $H^2_\sigma[0,\infty)$. Finally, in § 7, we present the left-definite theory for the Laguerre type orthogonal polynomials, including the definition and various properties of the self-adjoint operator $S_k[\cdot]$ in the Sobolev space $H^2_\sigma[0,\infty)$.

2. The differential expression M_k. In this section, we shall study properties of the differential expression $M_k[\cdot]$ defined in (1.6); i.e., for $x\in(0,\infty)$,

$$M_k[f](x):=(x^2\, e^{-x}f''(x))''-(((2A+2)x+2)\, e^{-x}f'(x))'+k\, e^{-x}f(x). \tag{2.1}$$

Let Δ_k denote the *maximal domain* of $M_k[\cdot]$ in $L^2(0,\infty;\, e^{-x})$; i.e.,

$$\Delta_k:=\{f\colon (0,\infty)\to\mathbb{C}\,|\, f^{(r)}\in AC_{\text{loc}}(0,\infty),\ r=0,1,2,3;\, f,\, e^xM_k[f]\in L^2(0,\infty;\, e^{-x})\}. \tag{2.2}$$

Standard techniques show that Δ_k is dense in $L^2(0, \infty; e^{-x})$; see [18, Chap. 5]. The *maximal operator* $T_{\max}[\cdot]$ generated by $M_k[\cdot]$ in $L^2(0, \infty; e^{-x})$ is defined to be:

$$T_{\max}[f](x) = e^x M_k[f](x) \qquad (x \in (0, \infty)),$$
$$\mathscr{D}(T_{\max}) = \Delta_k.$$

For f, $g \in \Delta_k$, and $[\alpha, \beta] \subset (0, \infty)$, we have *Green's formula*:

$$\int_\alpha^\beta \{e^x M_k[f](x)\bar{g}(x) - e^x \overline{M_k[g]}(x) f(x)\}\, e^{-x}\, dx = [f, g](x)|_\alpha^\beta, \tag{2.3}$$

where, for $x \in (0, \infty)$,

$$\begin{aligned} [f, g](x) := {} & \{(x^2\, e^{-x} f''(x))' - ((2A+2)x+2)\, e^{-x} f'(x)\}\bar{g}(x) - x^2\, e^{-x} f''(x)\bar{g}'(x) \\ & -\{(x^2\, e^{-x}\bar{g}''(x))' - ((2A+2)x+2)\, e^{-x}\bar{g}'(x)\} f(x) \\ & + x^2\, e^{-x} f'(x)\bar{g}''(x), \end{aligned} \tag{2.4}$$

and *Dirichlet's formula*:

$$\begin{aligned} & \int_\alpha^\beta \{x^2\, e^{-x} f''(x)\bar{g}''(x) + ((2A+2)x+2)\, e^{-x} f'(x)\bar{g}'(x) + k\, e^{-x} f(x)\bar{g}(x)\}\, dx \\ & \quad = \{-(x^2\, e^{-x} f''(x))'\bar{g}(x) + ((2A+2)x+2)\, e^{-x} f'(x)\bar{g}(x))\}|_\alpha^\beta \\ & \qquad + x^2\, e^{-x} f''(x)\bar{g}'(x)|_\alpha^\beta + \int_\alpha^\beta M_k[f](x)\bar{g}(x)\, dx. \end{aligned} \tag{2.5}$$

By Green's formula and the definition of Δ_k, note that the limits $[f, g](\infty) := \lim_{x\to\infty} [f, g](x)$ and $[f, g](0) := \lim_{x\to 0^+} [f, g](x)$ exist and are finite, for all f, $g \in \Delta_k$.

The *minimal operator* $T_{\min}[\cdot]$, generated by $M_k[\cdot]$ in $L^2(0, \infty; e^{-x})$, may be defined to be

$$T_{\min}[f](x) = e^x M_k[f](x) \qquad (x \in (0, \infty)),$$
$$\mathscr{D}(T_{\min}) = \{f \in \mathscr{D}(T_{\max}) | [f, g](\infty) = [f, g](0) \text{ for all } g \in \mathscr{D}(T_{\max})\}.$$

There is a strong similarity between the behavior of $M_k[\cdot]$ near the singular endpoint $x = 0$ and that of the Legendre type expression (see [6] and [7]) near the singular endpoints $x = \pm 1$. In fact, $M_k[\cdot]$ is limit-3 at the point $x = 0$ with the same form of Frobenius solutions as that of the Legendre type expression at $x = \pm 1$; the reader is referred to [11, pp. 396-403] for details on the Frobenius method. We shall therefore simply quote, without proof, appropriate results corresponding to Theorem 2.1 in [6], and Theorem 1.1 and Corollary 2.1 of [7], which continue to hold for f, $g \in \Delta_k$ at $x = 0$. We state these results as follows.

THEOREM 2.1. *Let f, $g \in \Delta_k$ and $a \in (0, \infty)$. Then*

(i) $\lim_{x\to 0^+} \{(x^2\, e^{-x} f''(x))' - ((2A+2)x+2)\, e^{-x} f'(x)\}$ *exists and is finite*;

(ii) $f'' \in L^2(0, a]$;

(iii) $f, f' \in AC[0, a]$; *i.e.*, $f, f' \in AC_{\mathrm{loc}}[0, \infty)$;

(iv) $\lim_{x\to 0^+} (x^2\, e^{-x} f''(x))' = 0$;

(v) $x\, e^{-x} f'' \in L^2(0, a]$;

(vi) $\lim_{x\to 0^+} x^2\, e^{-x} f''(x)\bar{g}'(x) = 0$;

(vii) $\lim_{x\to 0^+} [f, 1](x) = -2f'(0)$; $\lim_{x\to 0^+} [f, x](x) = 2f(0)$; $\lim_{x\to 0^+} [f, x^2](x) = 0$;

(viii) $\lim_{x\to 0^+} [f, g](x) = 2(f(0)\bar{g}'(0) - f'(0)\bar{g}(0))$.

Unfortunately, the methods used to establish the analogue of Theorem 2.1 for the Legendre type expression at $x = \pm 1$ do not seem to carry over in analyzing the Laguerre type expression at $x=\infty$. Recently, however, Race [21] has developed some sufficient conditions, based on ideas from [4], in order for a fourth-order differential expression to be strong limit-2 and Dirichlet at $x=\infty$. The Laguerre type differential expression satisfies Race's criteria, as he shows in [21]; we are grateful to David Race for allowing us to publish from his manuscript before publication. We can state the following theorem.

THEOREM 2.2. *Let $f, g \in \Delta_k$. Then*

(i) *The differential expression $M_k[\cdot]$ is strong limit-2 at $x=\infty$; i.e.,*

$$\lim_{x\to\infty}\{x^2 e^{-x}f''(x)\bar{g}'(x)+((2A+2)x+2)e^{-x}f'(x)\bar{g}(x)-(x^2 e^{-x}f''(x))'\bar{g}(x)\}=0.$$

(ii) *$M_k[\cdot]$ is Dirichlet at $x=\infty$; i.e., xf'', $x^{1/2}f' \in L^2(1,\infty; e^{-x})$.*

As a consequence of Theorems 2.1 and 2.2, we obtain the following simplified limiting forms of, respectively, Green's formula and Dirichlet's formula: for $f, g \in \Delta_k$,

$$\int_0^\infty \{e^x M_k[f](x)\bar{g}(x)-e^x\overline{M_k[g]}(x)f(x)\}\,e^{-x}\,dx = 2\bar{g}(0)f'(0)-2f(0)\bar{g}'(0), \tag{2.6}$$

and

$$\begin{aligned}\int_0^\infty \{x^2 e^{-x}f''(x)\bar{g}''(x)+((2A+2)x+2)e^{-x}f'(x)\bar{g}'(x)+k\,e^{-x}f(x)\bar{g}(x)\}\,dx \\ = -2f'(0)\bar{g}(0)+\int_0^\infty e^x M_k[f](x)\bar{g}(x)\,e^{-x}\,dx.\end{aligned} \tag{2.7}$$

3. A certain self-adjoint operator in $L^2(0,\infty; e^{-x})$. In this section, we find a certain self-adjoint extension of the minimal operator $T_{\min}[\cdot]$ generated by $M_k[\cdot]$ in $L^2(0,\infty; e^{-x})$. As the reader will see, this particular self-adjoint operator plays a key role in establishing the self-adjointness of the right-definite operator defined and discussed below in § 4. A characterization of all self-adjoint extensions of $T_{\min}[\cdot]$ in $L^2(0,\infty; e^{-x})$ will be given in § 5. The reader is encouraged to consult [18, Chap. 5] for details on the general theory of self-adjoint extensions of formally symmetric differential expressions.

Since $M_k[\cdot]$ is limit-2 at $x=\infty$ in $L^2(0,\infty; e^{-x})$, we note that the domain of the minimal operator $T_{\min}[\cdot]$ generated by $M_k[\cdot]$ in $L^2(0,\infty; e^{-x})$ is given by (see [18, § 18.3]):

$$\mathscr{D}(T_{\min})=\{f\in\mathscr{D}(T_{\max})\,|\,[f,g](0)=0 \text{ for all } g\in\mathscr{D}(T_{\max})\}. \tag{3.1}$$

In fact, from (3.1) and (viii) of Theorem 2.1, we see that

$$f\in\mathscr{D}(T_{\min}) \text{ if and only if } f(0)=f'(0)=0. \tag{3.2}$$

From [18, § 18.1], we see that all self-adjoint extensions of $T_{\min}[\cdot]$ can be determined by imposing on Δ_k one boundary condition at $x=0$ of the form

$$[f,g](0)=0, \qquad f\in\Delta_k,$$

where g is any function in $\Delta_k\backslash\mathscr{D}(T_{\min})$ satisfying $[g,g](0)=0$. In particular, we can take as a boundary condition $[f,1](0)=0$, or equivalently $f'(0)=0$, by using Theorem 2.1 (vii). From the theory of Naimark, we then have that the following operator in $L^2(0,\infty; e^{-x})$ is self-adjoint:

$$\begin{aligned}A_k[f](x)&=e^x M_k[f](x) \qquad (x\in(0,\infty)),\\ \mathscr{D}(A_k)&=\{f\in\Delta_k\,|\,f'(0)=0\}.\end{aligned} \tag{3.3}$$

Since $\mathcal{D}(A_k) \subset \Delta_k$ it follows, from (2.7) and the fact that the functions $x^2 e^{-x}$ and $((2A+2)x+2)\,e^{-x}$ are positive on $(0,\infty)$, that

$$(3.4)\qquad \begin{aligned}(A_k[f],f) &= \int_0^\infty M_k[f](x)\bar{f}(x)\,dx\\ &= \int_0^\infty \{x^2 e^{-x}|f''(x)|^2 + ((2A+2)x+2)\,e^{-x}|f'(x)|^2 + k\,e^{-x}|f(x)|^2\}\,dx\\ &\geqq k(f,f), \qquad f \in \mathcal{D}(A_k),\end{aligned}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(0,\infty;\, e^{-x})$. Thus $A_k[\cdot]$ is bounded below by kI, where I is the identity operator in $L^2(0,\infty;\, e^{-x})$. Hence, if $k>0$, we see that $0 \in \rho(A_k)$, the resolvent set of $A_k[\cdot]$. Consequently, when $k>0$, the resolvent operator $R_0(A_k) = A_k^{-1}$ exists and is a bounded operator from $L^2(0,\infty;\, e^{-x})$ onto $\mathcal{D}(A_k)$. We shall assume, for the remainder of this paper, that $k>0$.

4. The right-definite self-adjoint operator in $L^2_\sigma[0,\infty)$. In this section we shall discuss the appropriate self-adjoint operator $T_k[\cdot]$ in $L^2_\sigma[0,\infty)$ generated by $M_k[\cdot]$ that has the Laguerre type polynomials as eigenfunctions and whose spectrum is given by $\{\lambda_n + k \mid n = 0,1,2,\cdots\}$.

Let T_k: $\mathcal{D}(T_k) \to L^2_\sigma[0,\infty)$ be the operator defined by

$$(4.1)\qquad \begin{aligned}T_k[f](x) &:= \begin{cases} -2Af'(0) + kf(0) & \text{if } x = 0,\\ e^x M_k[f](x) & \text{if } x \in (0,\infty),\end{cases}\\ \mathcal{D}(T_k) &= \Delta_k.\end{aligned}$$

From Theorem 2.1 (iii), we see that $\Delta_k \subset L^2_\sigma[0,\infty)$ so that $T_k[\cdot]$ does indeed map $\mathcal{D}(T_k)$ into $L^2_\sigma[0,\infty)$. Observe, from (1.4), that if $f \in C^4[0,\infty)$, then $T_k[f](x) = e^x M_k[f](x)$ for all $x \in [0,\infty)$. It is precisely this property that prompts the definition of $T_k[\cdot]$ in (4.1). Furthermore, it is easy to check that $R_n \in \Delta_k$ and $T_k[R_n](x) = (\lambda_n + k)R_n(x)$, $x \in [0,\infty)$ and $n = 0,1,2,\cdots$, where R_n is the nth Laguerre type polynomial defined in (1.7). That is to say, the Laguerre type polynomials are eigenfunctions of $T_k[\cdot]$. We first prove the following theorem.

THEOREM 4.1.

(i) *$T_k[\cdot]$ in $L^2_\sigma[0,\infty)$ is a symmetric operator.*

(ii) *$T_k[\cdot]$ is bounded below by kI, where I denotes the identity operator in $L^2_\sigma[0,\infty)$.*

Proof. Let f, $g \in \Delta_k$. Then

$$(4.2)\qquad \begin{aligned}(T_k[f],g)_\sigma &= \int_0^\infty e^x M_k[f](x)\bar{g}(x)\,e^{-x}\,dx + \frac{T_k[f](0)\bar{g}(0)}{A}\\ &= 2\bar{g}(0)f'(0) - 2\bar{g}'(0)f(0) + \left(\frac{-2Af'(0)+kf(0)}{A}\right)\bar{g}(0)\\ &\quad + \int_0^\infty e^x\overline{M_k[g]}(x)f(x)\,e^{-x}\,dx,\\ &= -2\bar{g}'(0)f(0) + \frac{kf(0)\bar{g}(0)}{A} + \int_0^\infty e^x\overline{M_k[g]}(x)f(x)\,e^{-x}\,dx\\ &= \left(\frac{-2A\bar{g}'(0)+k\bar{g}(0)}{A}\right)f(0) + \int_0^\infty e^x\overline{M_k[g]}(x)f(x)\,e^{-x}\,dx\\ &= (f, T_k[g])_\sigma,\end{aligned}$$

by (2.6) and the definition of $T_k[\cdot]$. Hence $T_k[\cdot]$ in $L^2_\sigma[0,\infty)$ is hermitian. Since $C_0^\infty[0,\infty)\subset\Delta_k$ and $C_0^\infty[0,\infty)$ is dense in $L^2_\sigma[0,\infty)$, it follows that $T_k[\cdot]$ is symmetric in $L^2_\sigma[0,\infty)$.

Moreover, it follows from Dirichlet's formula (2.7) that

$$\begin{aligned}\int_0^\infty & e^x\overline{M_k[g]}(x)f(x)\,e^{-x}\,dx\\ &=2\bar g'(0)f(0)\\ &\quad+\int_0^\infty\{x^2\,e^{-x}f''(x)\bar g''(x)+((2A+2)x+2)\,e^{-x}f'(x)\bar g'(x)+k\,e^{-x}f(x)\bar g(x)\}\,dx.\end{aligned}$$

Combining this last equality with (4.2) yields for all $f,\ g\in\Delta_k$:

$$\begin{aligned}&(T_k[f],g)_\sigma\\ &=\frac{kf(0)\bar g(0)}{A}+\int_0^\infty\{x^2\,e^{-x}f''(x)\bar g''(x)+((2A+2)x+2)\,e^{-x}f'(x)\bar g'(x)\\ &\qquad\qquad+k\,e^{-x}f(x)\bar g(x)\}\,dx.\\ &=\int_0^\infty\{x^2\,e^{-x}f''(x)\bar g''(x)+((2A+2)x+2)\,e^{-x}f'(x)\bar g'(x)\}\,dx+k(f,g)_\sigma.\end{aligned}\tag{4.3}$$

In particular, as in (3.4), we see that

$$(T_k[f],f)_\sigma\geqq k(f,f)_\sigma.$$

Hence $T_k[\cdot]$ is bounded below by kI in $L^2_\sigma[0,\infty)$. This completes the proof. □

Remarks.

(1) We shall discuss, in detail, the left-definite problem associated with $M_k[\cdot]$ in § 7. However, we point out now that the left-definite inner product $(\cdot,\cdot)_H$ defined in (1.11) coincides with the right-hand side of (4.3) above. Indeed, it is precisely this identity in (4.3) which prompts the definition of $(f,g)_H$ in (1.11).

(2) Since $T_k[R_n]=(\lambda_n+k)R_n$, $n=0,1,2,\cdots$, we see from (1.5), (1.8), (1.11), and (4.3) that

$$\begin{aligned}&(R_n,R_m)_H\\ &=\int_0^\infty\{x^2\,e^{-x}R_n''(x)R_m''(x)+((2A+2)x+2)\,e^{-x}R_n'(x)R_m'(x)\}\,dx+k(R_n,R_m)_\sigma\\ &=n(n+2A+1)(A+n+1)(A+n)\delta_{nm};\end{aligned}\tag{4.4}$$

i.e., the Laguerre type polynomials $\{R_n(x)\}$ form an orthogonal set in the weighted Sobolev space H defined by (1.10) and (1.11). We shall show that they, in fact, form a complete orthogonal set in $H^2_\sigma[0,\infty)$ in § 6.

To show that $T_k[\cdot]$ is self-adjoint in $L^2_\sigma[0,\infty)$, we shall first establish the self-adjointness of two operators T_k' and T_k'' which form a decomposition of T_k; i.e., $T_k=T_k'+T_k''$. This requires us to consider the solutions of the equation

$$M_k[y](x)=0\qquad(x\in(0,\infty))\tag{4.5}$$

which are in $L^2(0,\infty;e^{-x})$. First, we note that the deficiency index of the minimal operator $T_{\min}[\cdot]$ generated by $M_k[\cdot]$ in $L^2(0,\infty;e^{-x})$ is (1, 1); i.e., for any $\lambda\in\mathbb{C}$ with nonzero imaginary part, the equation

$$M_k[y](x)=\lambda\,e^{-x}y(x)\qquad(x\in(0,\infty))$$

has only one linearly independent solution in $L^2(0,\infty; e^{-x})$. This follows from the limit classification in $L^2(0,\infty; e^{-x})$ of the singular endpoints $x=0$ and $x=\infty$ discussed in § 3; see [18, § 17.5]. Secondly, from (2.7) and (3.2), we see that

$$(T_{\min}[f], f) \geqq k(f, f), \qquad (f \in \mathcal{D}(T_{\min})), \tag{4.6}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(0,\infty; e^{-x})$. That is to say, $T_{\min}[\cdot]$ is bounded below by kI, where I is the identity operator on $L^2(0,\infty; e^{-x})$. Using the Cauchy-Schwarz inequality together with (4.6), we have for any $c \in \mathbb{R}$, with $c<k$:

$$(k-c)\|f\|^2 \leqq (T_{\min}[f]-cf, f) \leqq \|T_{\min}[f]-cf\|\,\|f\|,$$

where $\|\cdot\| = (\cdot,\cdot)^{1/2}$.
Hence,

$$\|T_{\min}[f]-cf\| \geqq (k-c)\|f\| \qquad (f \in \mathcal{D}(T_{\min})).$$

For $k>0$, it follows then that $c=0$ is in the domain of regularity of the minimal operator $T_{\min}[\cdot]$ (see [18, § 14.9]). Furthermore, from [18, § 14.10, Cor. 2], we can conclude that there is only one linearly independent solution $\theta(x)$ of (4.5) which belongs to $L^2(0,\infty; e^{-x})$. Clearly, $\theta \in \Delta_k$. We claim that $\theta'(0) \neq 0$. For if $\theta'(0)=0$, then $\theta \in \mathcal{D}(A_k)$, where $A_k[\cdot]$ is the operator defined in (3.3). However, this means that θ is an eigenfunction of $A_k[\cdot]$ corresponding to the eigenvalue $\lambda=0$. This contradicts the fact that $0 \in \rho(A_k)$, the resolvent set of $A_k[\cdot]$. Consequently, $\theta'(0)\neq 0$. Without loss of generality, we may assume that $\theta'(0)=1$.

Define the operator T'_k in $L^2_\sigma[0,\infty)$ by:

$$\begin{aligned} T'_k[f](x) &= \begin{cases} -2Af'(0) & \text{if } x=0, \\ e^x M_k[f](x) & \text{if } x \in (0,\infty), \end{cases} \\ \mathcal{D}(T'_k) &= \Delta_k. \end{aligned} \tag{4.7}$$

Similar to the proof of Theorem 4.1, we find that $T'_k[\cdot]$ is a symmetric operator in $L^2_\sigma[0,\infty)$. It remains to show that $T'_k[\cdot]$ is self-adjoint in $L^2_\sigma[0,\infty)$. We do this by following the analysis given in [6, § 4] which requires the following theorem (see [1, § 46]).

THEOREM 4.2. *Let A be a symmetric operator in a Hilbert space H. If the range of A is all of H, then A is self-adjoint in H.*

THEOREM 4.3. *The operator $T'_k[\cdot]$ in $L^2_\sigma[0,\infty)$ is self-adjoint.*

Proof. Let $f \in L^2_\sigma[0,\infty)$ and define g: $[0,\infty)\to\mathbb{C}$ by

$$g(x) = \frac{-f(0)}{2A}\theta(x) + [R_0(A_k)f](x), \qquad x \in [0,\infty),$$

where $R_0(A_k) = A_k^{-1}$ is the resolvent operator of A_k at the regular point $\lambda=0$. We claim that $g \in \mathcal{D}(T'_k)$ and $T_k[g](x)=f(x)$, $x\in[0,\infty)$. Since $\theta \in \Delta_k = \mathcal{D}(T'_k)$ and $R_0(A_k)f \in \mathcal{D}(A_k) \subset \mathcal{D}(T'_k)$, we see that $g \in \mathcal{D}(T'_k)$. Also, from the definition of θ and the fact that $R_0(A_k)f \in \mathcal{D}(A_k)$, we see that $g'(0) = -f(0)/2A$. Hence $T'_k[g](0)=f(0)$. Since θ is a solution of the homogeneous equation (4.5) and $R_0(A_k)$ is the inverse of $e^xM_k[\cdot]$ on $(0,\infty)$, it follows that, for all $x \in (0,\infty)$,

$$T'_k[g](x) = e^xM_k[(-f(0)/2A)\theta(x) + [R_0(A_k)f](x)] = f(x);$$

i.e., $T'_k[g](x)=f(x)$, $x \in [0,\infty)$. Hence by Theorem 4.2, $T'_k[\cdot]$ is self-adjoint in $L^2_\sigma[0,\infty)$. □

Consider now the operator T_k'' in $L_\sigma^2[0,\infty)$ defined by

$$T_k''[f](x)=\begin{cases}kf(0) & \text{if } x=0,\\ 0 & \text{if } x\in(0,\infty),\end{cases}$$

$$\mathcal{D}(T_k'')=L_\sigma^2[0,\infty).$$

It is easy to check that $T_k''[\cdot]$ is symmetric in $L_\sigma^2[0,\infty)$ and since its domain is all of $L_\sigma^2[0,\infty)$, we have, in fact, that $T_k''[\cdot]$ is self-adjoint in $L_\sigma^2[0,\infty)$. By following similar arguments in [6, § 5], we have that $T_k=T_k'+T_k''$ is self-adjoint in $L_\sigma^2[0,\infty)$.

As Theorem 4.5 below states, the spectrum $\sigma(T)$ of $T_k[\cdot]$ consists of only the point spectrum $\{\lambda_n+k\,|\,n=0,1,2,\cdots\}$. To see this, it suffices to know that the Laguerre type polynomials are complete in $L_\sigma^2[0,\infty)$ or, equivalently, that the set of polynomials $P[0,\infty)$, where

$$P[0,\infty):=\left\{\sum_{k=0}^{n} a_kx^k\,|\,a_k\in\mathbb{C},\, n=0,1,2,\cdots,x\in[0,\infty)\right\}, \tag{4.8}$$

is dense in $L_\sigma^2[0,\infty)$. For, if $P[0,\infty)$ is dense in $L_\sigma^2[0,\infty)$, then a simple argument shows that the point spectrum of $T_k[\cdot]$ is given by:

$$\sigma_p(T_k)=\{\lambda_n+k\,|\,n=0,1,2,\cdots\}.$$

Indeed, suppose $\tilde\lambda\in\sigma_p(T_k)$ and $\tilde\lambda\neq\lambda_n+k$ for any $n=0,1,2,\cdots$. Let $\tilde f$ be an eigenfunction of $T_k[\cdot]$ associated with the eigenvalue $\tilde\lambda$; in particular, note that $\tilde f\not\equiv 0$ in $L_\sigma^2[0,\infty)$. Since the eigenfunctions of a self-adjoint operator are necessarily orthogonal, we have that $(\tilde f, R_n)_\sigma=0$, $n=0,1,2,\cdots$. However, from the completeness of $\{R_n(x)\}$ in $L_\sigma^2[0,\infty)$, this forces $\tilde f\equiv 0$ in $L_\sigma^2[0,\infty)$ which is a contradiction. By appealing to a well-known result (see [22, p. 361]), we conclude that the spectrum $\sigma(T_k)$ of $T_k[\cdot]$ is the closure in $\mathbb{C}$ of the point spectrum. Since $\lim_{n\to\infty}(\lambda_n+k)=\infty$, we have

$$\sigma(T_k)=\{\lambda_n+k\,|\,n=0,1,2,\cdots\}.$$

There are many routes to take in showing that $P[0,\infty)$ is dense in $L_\sigma^2[0,\infty)$. For example, this fact follows from theorems of M. Riesz and H. Hamburger, which we state below in Theorem 4.4 (see [10, Thm. 4.2, Prob. 19, Thm. 5.2, and the remark on p. 87]). To introduce this theorem, we use the same notation that Freud [10] uses.

Suppose $\hat\mu:\mathbb{R}\to\mathbb{R}$ is a distribution; i.e., $\hat\mu$ is a monotonic, nondecreasing function with infinitely many points of increase in $\mathbb{R}$. Let μ be the regular, positive, Borel measure generated by $\hat\mu$ on the Borel subsets of $\mathbb{R}$. The nth moment of $d\mu$ is defined to be

$$\mu_n:=\int_{-\infty}^{\infty} x^n\,d\mu(x),\qquad n=0,1,2,\cdots;$$

we shall assume that these numbers exist and are finite. Of course, this implies that $P[0,\infty)\subset L_\mu^1[0,\infty)\cap L_\mu^2[0,\infty)$. In this case, we say that the moment sequence $\{\mu_n\}$ is generated or determined by $\hat\mu(x)$. We say that the moment sequence is *determinate* if, whenever it is also generated by a distribution $\hat\nu(x)$, we have $\hat\mu(x)=\hat\nu(x)+c$ at all common points of continuity, where c is a fixed constant. If $\{\mu_n\}$ is determinate, we write $d\mu\in\mathscr{C}$ and say that $\hat\mu$ is *substantially unique*. We state the following theorem.

THEOREM 4.4. *With the notation as above.*

(a) (*M. Riesz*, 1923): $P[0,\infty)$ *is dense in* $L_\mu^2[0,\infty)$ *if and only if* $d\mu/(1+x^2)\in\mathscr{C}$.

(b) (*Hamburger*, 1919): *If there exists a* $\beta > 0$ *such that*

$$\int_{-\infty}^{\infty} e^{\beta|x|}\, d\mu(x) < \infty,$$

then $d\mu \in \mathscr{C}$.

In the case of the Laguerre type polynomials, it is easy to check that $d\sigma(x)/(1+x^2)$ satisfies Hamburger's condition (b) for any $0<\beta<1$. Hence, by Riesz' condition (a), we have that $P[0,\infty)$ is dense in $L^2_\sigma[0,\infty)$.

We summarize the main results of this section.

THEOREM 4.5. *The operator* $T_k[\cdot]$, *defined in* (4.1), *is self-adjoint in* $L^2_\sigma[0,\infty)$. *The Laguerre type polynomials* $\{R_n(x)\}$ *are eigenfunctions of* $T_k[\cdot]$ *and they form a complete orthogonal set in* $L^2_\sigma[0,\infty)$. *Furthermore, the spectrum of* $T_k[\cdot]$ *is given by*

$$\sigma(T_k) = \{n(n+2A+1)+k \mid n = 0, 1, 2, \cdots\}.$$

Note. Since $T_k[R_n](0) = (\lambda_n + k)R_n(0)$, we see, from the definition of $T_k[\cdot]$, that

$$-2AR_n'(0) = \lambda_n R_n(0). \tag{4.9}$$

This is the λ-dependent boundary condition discussed in [12] and [13]. Notice that it is satisfied by the eigenfunctions of $T_k[\cdot]$ and not, in general, by each element of $\mathscr{D}(T_k)$. Hence, (4.9) is seen as a property of the Laguerre type polynomials and not as an essential element in the definition of $T_k[\cdot]$; see also [6; § 5].

5. The spectrum of self-adjoint extensions of the minimal operator generated by M_k in $L^2(0,\infty; e^{-x})$. The main result in this section is concerned with the form and the spectrum of all self-adjoint operators generated by the minimal operator $T_{\min}[\cdot]$ of $M_k[\cdot]$ in $L^2(0,\infty;\, e^{-x})$. We prove the following theorem.

THEOREM 5.1. *Let* $T[\cdot]$ *be any self-adjoint extension of the minimal operator* $T_{\min}[\cdot]$ *generated by* $M_k[\cdot]$ *in* $L^2(0,\infty;\, e^{-x})$.

(i) *Then there exists a nonzero vector* $(\alpha,\beta)\in\mathbb{R}^2$ *such that* $T[\cdot]$ *is given by*

$$\begin{gathered} T[f](x) = e^x M_k[f](x) \qquad (x\in(0,\infty)), \\ \mathscr{D}(T) = \{f\in\mathscr{D}(T_{\max}) \mid \alpha f(0)+\beta f'(0) = 0\}. \end{gathered} \tag{5.1}$$

Conversely, each operator $T[\cdot]$ *in* $L^2(0,\infty;\, e^{-x})$ *of the form* (5.1), *where* $(\alpha,\beta)\in\mathbb{R}^2\backslash\{(0,0\}$, *is self-adjoint.*

(ii) *The spectrum of T is discrete and bounded below in* $L^2(0,\infty;\, e^{-x})$.

Proof.

(i) Define $\{h_1, h_2\}\subset\Delta_k = \mathscr{D}(T_{\max})$ such that

$$h_1(x) = \begin{cases} -\frac{1}{2} & \text{for } x \text{ near } 0, \\ 0 & \text{for } x \text{ sufficiently large,} \end{cases}$$

$$h_2(x) = \begin{cases} x/2 & \text{for } x \text{ near } 0, \\ 0 & \text{for } x \text{ sufficiently large.} \end{cases}$$

Note that $[h_1,h_1](0) = [h_2,h_2](0) = 0$, $[h_1,h_2](0) = -\frac{1}{2}$, and $[h_2,h_1](0) = \frac{1}{2}$. From (3.2), it is clear that $\{h_1,h_2\}$ is linearly independent modulo $\mathscr{D}(T_{\min})$. In fact, since the deficiency index of $T_{\min}$ in $L^2(0,\infty;\, e^{-x})$ is (1, 1), the set $\{h_1,h_2\}$ forms a basis for the quotient space $\mathscr{D}(T_{\max})/\mathscr{D}(T_{\min})$. Consequently, from [18, § 18.1], every self-adjoint extension T of the minimal operator $T_{\min}[\cdot]$ generated by $M_k[\cdot]$ in $L^2(0,\infty;e^{-x})$ has the form

$$T[f](x)=e^x M_k[f](x), \qquad x\in(0,\infty),$$
$$\mathscr{D}(T)=\{f\in\mathscr{D}(T_{\max})\,|\,[f,w](x)|_0^\infty=0\},$$

where $w(x)=ah_1(x)+bh_2(x)$ satisfies

$$(5.2) \qquad [w,w](x)|_0^\infty=0,$$

and $(a,b)\in\mathbb{C}^2$ is a nonzero vector. However, since $M_k[\cdot]$ is strong limit-2 at $x=\infty$ (actually, limit-2 at $x=\infty$ is sufficient), we see that $[f,w](\infty)=\lim_{x\to\infty}[f,w](x)=0$ for all $f\in\mathscr{D}(T_{\max})$. Hence the boundary condition in the definition above for $\mathscr{D}(T)$ reduces to requiring that

$$(5.3) \qquad [f,w](0)=0,$$

while the symmetry condition (5.2) reduces to

$$(5.4) \qquad [w,w](0)=0.$$

Written out in full, (5.4) yields the requirement

$$\begin{aligned}0=[w,w](0)&=[ah_1+bh_2,\,ah_1+bh_2](0)\\&=\bar{a}b[h_2,h_1](0)+a\bar{b}[h_1,h_2](0)\\&=\tfrac{1}{2}(\bar{a}b-a\bar{b}).\end{aligned}$$

This last requirement implies, without loss of generality, that both a and b can be taken as real numbers. For the boundary condition (5.3) is homogeneous in w, and so if $a\neq 0$ we can take $a=1$, and it then follows that $b=\bar{b}$ and so b is real; similarly if $b\neq 0$ then take $b=1$, and we obtain $a=\bar{a}$, and so a is real. We now define $\alpha=b$ and $\beta=a$. Returning to the boundary condition (5.3) and using (vii) of Theorem 2.1 above, we see that

$$[f,w](0)=\bar{a}[f,h_1](0)+\bar{b}[f,h_2](0)=\alpha f(0)+\beta f'(0),$$

and we obtain the given form of the boundary condition at the endpoint 0, as required in the statement above of Theorem 5.1. This completes the proof of (i).

(ii) We shall prove that the operator $T[\cdot]$ in $L^2(0,\infty;e^{-x})$ has a discrete spectrum by relating $T[\cdot]$ to the operator $T_k[\cdot]$ in the space $L^2_\sigma[0,\infty)$. This argument is identical to the one given in [7, § 4]. Recall, from (4.6), that the minimal operator $T_{\min}[\cdot]$ is bounded below by kI in $L^2(0,\infty;e^{-x})$. Now, if a semibounded operator has equal, finite deficiency indices, then every self-adjoint extension of this operator is also semibounded (see [18, § 14]). Hence every self-adjoint extension $T[\cdot]$ of $T_{\min}[\cdot]$ is bounded below in $L^2(0,\infty;e^{-x})$.

To show that the spectrum of $T[\cdot]$ in $L^2(0,\infty;e^{-x})$ is discrete, we need only show that the essential spectrum $\sigma_e(T)=\emptyset$, the empty set. Since $T[\cdot]$ is a finite-dimensional extension of $T_{\min}[\cdot]$, we have that $\sigma_e(T)=\sigma_e(T_{\min})$ (see [18, § 14.9, Thm. 9]. Suppose then, for some real number λ, we have $\lambda\in\sigma_e(T)=\sigma_e(T_{\min})$. Then there must be some bounded, noncompact sequence $\{f_n\,|\,n=1,2,\cdots\}$ in $L^2(0,\infty;e^{-x})$ such that

$$(5.5) \qquad f_n\in\mathscr{D}(T_{\min}),\ \|f_n\|=1 \text{ and } \lim_{n\to\infty}\|T_{\min}[f_n]-\lambda f_n\|=0$$

(see [25; Thm. 7.24]); here $\|\cdot\|$ refers to the norm of an element in the space $L^2(0,\infty;e^{-x})$. Since $T_{\min}[\cdot]$ is the closure of the operator restricted to those elements of the maximal domain Δ_k having compact support in the open interval $(0, \infty)$ (see, [18, § 17]), there is no loss of generality in supposing that each f_n above has compact support in $(0, \infty)$.

Embed each f_n in $L^2_\sigma[0, \infty)$ by setting $f_n(0)=0$ for all $n=1, 2, \cdots$; by doing this we have a sequence $\{f_n\} \subset \mathscr{D}(T_k)$. Moreover, we see that

$$\|f_n\| = \|f_n\|_\sigma \qquad n=1, 2, 3 \cdots, \tag{5.6}$$

$$\|T_k[f_n] - \lambda f_n\|_\sigma = \|T_{\min}[f_n] - \lambda f_n\|. \tag{5.7}$$

It follows from (5.5), (5.6), and (5.7) that $\{f_n\}$ is bounded and noncompact in $L^2_\sigma[0, \infty)$ and

$$\lim_{n\to\infty} \|T_k[f_n] - \lambda f_n\| = 0.$$

Thus we must have $\lambda \in \sigma_e(T_k)$. However, as we showed in Theorem 4.5, the spectrum of $T_k[\cdot]$ is discrete and thus $\sigma_e(T_k) = \emptyset$.

Hence $\sigma_e(T) = \emptyset$ and therefore the spectrum of every self-adjoint extension of the minimal operator $T_{\min}[\cdot]$ generated by $M_k[\cdot]$ in $L^2(0, \infty; e^{-x})$ is discrete. □

6. The density of polynomials in the Sobolev space $H^2_\sigma[0, \infty)$. In this section, we prove that $P[0, \infty)$, the space of polynomials defined in (4.8), is dense in the Hilbert space $H^2_\sigma[0, \infty)$, defined in (1.10). We refer the reader to [7, § 5] and [9] where different proofs are given for the completeness of polynomials in the Legendre type left-definite spaces. The proof that $H^2_\sigma[0, \infty)$ is complete in the topology generated from the inner product $(\cdot, \cdot)_H$, defined in (1.11), follows from standard results in the theory of Lebesgue integration. The reader is referred to [17] for the proof of the completeness of the left-definite Sobolev spaces associated with the Legendre type expressions.

Notice that $P[0, \infty)$ and Δ_k are all linear manifolds of the Hilbert space $H^2_\sigma[0, \infty)$. In fact, from Theorem 2.1, these spaces satisfy the following inclusion

$$P[0, \infty) \subset \Delta_k \subset H^2_\sigma[0, \infty). \tag{6.1}$$

Before establishing the density of $P[0, \infty)$ in $H^2_\sigma[0, \infty)$, we note that (4.3) simplifies to

$$(T_k[f], g)_\sigma = (f, g)_H \qquad (f, g \in \Delta_k).$$

From Theorems 2.1 and 2.2, it is not too difficult to generalize this identity and establish the following (see also [21]):

$$(T_k[f], g)_\sigma = (f, g)_H, \qquad (f \in \Delta_k, g \in H^2_\sigma[0, \infty)), \tag{6.2}$$

$$(f, T_k[g])_\sigma = (f, g)_H, \qquad (f \in H^2_\sigma[0, \infty), g \in \Delta_k). \tag{6.3}$$

From (4.4), we see that the Laguerre type polynomials $\{R_n(x)\}$ are orthogonal in $H^2_\sigma[0, \infty)$. Consequently, to prove the density of $P[0, \infty)$ in $H^2_\sigma[0, \infty)$ it suffices to show, equivalently, that $\{R_n(x)\}$ forms a complete orthogonal set in $H^2_\sigma[0, \infty)$.

Let $g \in H^2_\sigma[0, \infty)$ and suppose that

$$(R_n, g)_H = 0 \qquad (n=0, 1, \cdots). \tag{6.4}$$

We shall show that $g=0$ in the space $H^2_\sigma[0, \infty)$. From (6.2), we see that (6.4) yields

$$(\lambda_n + k)(R_n, g)_\sigma = 0 \qquad (n=0, 1, \cdots), \tag{6.5}$$

where we have used the fact that

$$T_k[R_n](x) = (\lambda_n + k)R_n(x) \qquad (x \in [0, \infty), n=0, 1, \cdots).$$

Since $\lambda_n + k > 0$ $(n=0, 1, \cdots)$, we see that (6.5) simplifies to

$$(R_n, g)_\sigma = 0 \qquad (n=0, 1, \cdots). \tag{6.6}$$

However, from Theorem 4.5, the Laguerre type polynomials $\{R_n(x)\}$ form a complete orthogonal set in $L^2_\sigma[0,\infty)$. Thus it follows, from (6.6), that $g=0$ in the space $L^2_\sigma[0,\infty)$; that is to say, $g(0)=0$ and $g(x)=0$ for almost all $x\in(0,\infty)$. But since $g\in AC_{\text{loc}}[0,\infty)$, we must have $g(x)\equiv 0$ on $[0,\infty)$ and hence $g=0$ in the space $H^2_\sigma[0,\infty)$. This completes the proof of the completeness of $\{R_n(x)\}$ in $H^2_\sigma[0,\infty)$.

We summarize by the following.

THEOREM 6.1. *The Laguerre type polynomials, defined in* (1.7), *form a complete set of orthogonal polynomials in the weighted Sobolev space* $H^2_\sigma[0,\infty)$. *Equivalently, the set* $P[0,\infty)$ *of polynomials, defined in* (4.8), *is dense in* $H^2_\sigma[0,\infty)$.

7. The Laguerre type left-definite theory. In this section, we shall discuss the left-definite theory associated with the Laguerre type differential expression $M_k[\cdot]$, defined in (1.6). The reader is referred to the contributions [5], [6], [7], [8], [9], [17], [19], and [20] for further studies of left-definite theory with applications in the theory of orthogonal polynomials.

Recall, from § 4, that the self-adjoint operator $T_k[\cdot]$ is bounded below by kI, where I is the identity operator on $L^2_\sigma[0,\infty)$. Hence, if $k>0$, then $0\in\rho(T_k)$, the resolvent set of $T_k[\cdot]$. Consequently, in this case, we see that $R_0(T_k)=T_k^{-1}$ is a bounded operator from $L^2_\sigma[0,\infty)$ onto Δ_k. Furthermore, note the following inclusion between the spaces Δ_k, $H^2_\sigma[0,\infty)$ and $L^2_\sigma[0,\infty)$:

$$\Delta_k\subset H^2_\sigma[0,\infty)\subset L^2_\sigma[0,\infty). \tag{7.1}$$

Define the operator B_k: $H^2_\sigma[0,\infty)\to H^2_\sigma[0,\infty)$ by

$$B_k[f](x)=R_0(T_k)[f](x)\qquad (x\in(0,\infty)),$$
$$f\in\mathscr{D}(B_k)=H^2_\sigma[0,\infty).$$

Note, from (7.1), that $B_k[\cdot]$ does indeed map $H^2_\sigma[0,\infty)$ into $H^2_\sigma[0,\infty)$. Furthermore from (6.2) and the fact that $B_k[f]\in\Delta_k$ for all $f\in H^2_\sigma[0,\infty)$, we see that:

$$(B_k[f],g)_H=(T_k(B_k[f]),g)_\sigma=(f,g)_\sigma,\qquad f,g\in H^2_\sigma[0,\infty).$$

Similarly, from (6.3), it follows that

$$(f,B_k[g])_H=(f,g)_\sigma,\qquad f,g\in H^2_\sigma[0,\infty).$$

Consequently, $B_k[\cdot]$ is a symmetric operator in $H^2_\sigma[0,\infty)$ and, since $\mathscr{D}(B_k)=H^2_\sigma[0,\infty)$, it follows that $B_k[\cdot]$ is self-adjoint in $H^2_\sigma[0,\infty)$. Moreover, if $B_k[f]=0$ for some $f\in H^2_\sigma[0,\infty)$, then $f=T_k(B_k[f])=0$ in $H^2_\sigma[0,\infty)$; i.e., $B_k[\cdot]$ is an injective map and hence $S_k:=B_k^{-1}$ exists. Furthermore, from [1, *I*, § 41], we see that $S_k[\cdot]$ is a self-adjoint operator in $H^2_\sigma[0,\infty)$.

Now, from the equality

$$R_n=B_k(T_k[R_n])=(\lambda_n+k)B_k[R_n],$$

we see that

$$S_k[R_n]=(\lambda_n+k)R_n,\qquad n=0,1,2,\cdots.$$

That is to say, the nth Laguerre type polynomial $R_n(x)$ is an eigenfunction of $S_k[\cdot]$. Since $\lim_{n\to\infty}(\lambda_n+k)=\infty$, we see that $S_k[\cdot]$ is an unbounded self-adjoint operator in $H^2_\sigma[0,\infty)$. Of course, the general theory of self-adjoint operators in a Hilbert space says that these eigenfunctions are orthogonal in $H^2_\sigma[0,\infty)$; we remind the reader that the explicit orthogonality relation of the Laguerre type polynomials in $H^2_\sigma[0,\infty)$ is given in (4.4). From § 6, we know that the Laguerre type polynomials are complete in

$H^2_\sigma[0, \infty)$. Following the argument mutatis mutandis that is given in § 4, we deduce that the spectrum of $S_k[\cdot]$ is given by

$$\sigma(S_k) = \{\lambda_n + k \mid n = 0, 1, 2, \cdots\}.$$

Is there an identification of $S_k[\cdot]$ as a differential operator in terms of the differential expression $M_k[\cdot]$? The answer is yes; recently, it has been shown that $S_k[f](x) = e^x M_k[f](x)$ for all $f \in \mathcal{D}(S_k) \subset \Delta_k$ and $x \in (0, \infty)$. The details of this result will be reported in a future paper by Everitt and Littlejohn.

We summarize the results of this section.

THEOREM 7.1. *The operator $S_k[\cdot]$ defined above is a self-adjoint differential operator in the weighted Sobolev space $H^2_\sigma[0, \infty)$ generated by the differential expression $M_k[\cdot]$. The Laguerre type polynomials $\{R_n(x)\}$ form a complete set of eigenfunctions of $S_k[\cdot]$ in $H^2_\sigma[0, \infty)$. The spectrum of $S_k[\cdot]$ is given by*:

$$\sigma(S_k) = \{\lambda_n + k \mid n = 0, 1, 2, \cdots\}.$$

Acknowledgments. V. P. Onyango-Otieno expresses his gratitude to Professors Everitt, Krall, and Littlejohn for introducing him to the problems discussed in this paper. He is also grateful to Professor Abdus Salam of the International Centre for Theoretical Physics, the International Atomic Energy Agency, and the United Nations Educational, Scientific and Cultural Organization for kind hospitality and financial support during his stay in Trieste, Italy.

REFERENCES

[1] N. I. AKHIEZER AND I. M. GLAZMAN, *Theory of Linear Operators in Hilbert Space*: I *and* II, Pitman, London, 1981.

[2] J. CHAUDHURI AND W. N. EVERITT, *The spectrum of a fourth-order differential operator*, Proc. Roy. Soc. Edinburgh Sect. A, 68 (1969), pp. 185-210.

[3] R. S. CHISHOLM AND W. N. EVERITT, *On bounded integral operators in the space of integrable square functions*, Proc. Roy. Soc. Edinburgh Sect. A, 69 (1971), pp. 199-204.

[4] W. N. EVERITT, *Some positive definite differential operators*, J. London Math. Soc., 43 (1968), pp. 465-473.

[5] ———, *Legendre polynomials and singular differential operators*, Lecture Notes in Math., 827, Springer-Verlag, New York, 1980, pp. 83-106.

[6] W. N. EVERITT AND L. L. LITTLEJOHN, *Differential operators and the Legendre type polynomials*, Differential and Integral Equations, 1 (1988), pp. 97-116.

[7] W. N. EVERITT, A. M. KRALL, AND L. L. LITTLEJOHN, *On some properties of the Legendre type differential expression*, Quaestiones Math., 13 (1990), pp. 83-106.

[8] W. N. EVERITT, L. L. LITTLEJOHN, AND S. C. WILLIAMS, *Orthogonal polynomials in weighted Sobolev spaces*, Lecture Notes in Pure and Appl. Math., Jaime Vinuesa, ed., 117, Dekker, New York, 1989, pp. 53-72.

[9] ———, *The left-definite Legendre type boundary value problem*, Constr. Approx., 7 (1991), pp. 485-500.

[10] G. FREUD, *Orthogonal Polynomials*, Pergamon, New York, 1971.

[11] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1956.

[12] A. M. KRALL, *Orthogonal polynomials satisfying fourth order differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 87 (1981), pp. 271-288.

[13] ———, A review of orthogonal polynomials satisfying boundary value problems, Lecture Notes in Math., 1329, Springer-Verlag, New York, 1986, pp. 73-97.

[14] A. M. KRALL AND L. L. LITTLEJOHN, *Differential equations and Bochner-Krall orthogonal polynomials*, Utah State Tech. Rep., Utah State University, Logan, Utah, 1988.

[15] H. L. KRALL, *Certain differential equations for Tchebycheff polynomials*, Duke Math. J., 4 (1938), pp. 705-718.

[16] ———, *On orthogonal polynomials satisfying a certain fourth order differential equation*, The Pennsylvania State College Studies, 6, The Pennsylvania State College, State College, PA, 1940.

[17] S. M. LOVELAND, *Spectral analysis of the Legendre equations*, Ph.D. thesis, Utah State University, Logan, UT 1990.

[18] M. A. NAIMARK, *Linear differential operators* II, Ungar, New York, 1968.

[19] V. P. ONYANGO-OTIENO, *The application of ordinary differential operators to the study of classical orthogonal polynomials*, Ph.D. thesis, University of Dundee, Dundee, Scotland, 1980.

[20] ———, *Laguerre polynomials and singular differential operators*, Indian J. Pure Appl. Math., 18 (1987), pp. 515–535.

[21] D. RACE, *Some strong limit-2 and Dirichlet criteria for fourth-order differential expressions*, Math. Proc. Cambridge Philos. Soc., 108 (1990), pp. 409–416.

[22] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Ungar, New York, 1978.

[23] G. SZEGÖ, *Orthogonal polynomials*, Fourth ed., Amer. Math. Soc. Colloq. Publ., Providence, RI, 1978.

[24] E. C. TITCHMARSH, *Eigenfunction expansions associated with second-order differential equations* I, Clarendon Press, Oxford, 1962.

[25] J. WEIDMANN, *Linear operators in Hilbert spaces*, Springer-Verlag, Heidelberg, 1980.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 737-757, May 1992

010

ON ORTHOGONAL POLYNOMIALS OF SOBOLEV TYPE: ALGEBRAIC PROPERTIES AND ZEROS*

M. ALFARO†, F. MARCELLÁN‡, M L. REZOLA†, AND A. RONVEAUX§

Abstract. In this paper the inner product $\langle f, g\rangle = \int_I fg\,d\mu + Mf(c)g(c) + Nf'(c)g'(c)$ is considered, where μ is a positive measure on the interval I, $c \in \mathbf{R}$ and $M, N \geqq 0$. General algebraic properties of the orthogonal polynomials associated with $\langle \cdot, \cdot \rangle$ as well as the zeros and their location are studied. In particular, the case of a symmetric measure μ is analyzed. Finally, a second-order linear differential equation and two applications are given.

Key words. orthogonal polynomials, inner product, kernels, zeros, differential equations

AMS(MOS) subject classification. 33C45

1. Introduction. Problems concerning the approximation of $C^{(k)}$ functions by polynomials, using the method of least squares, had been considered by Lewis [16], Gröbner [8], and Lesky [15]. In these papers, orthogonal polynomials associated to inner products involving derivatives appear in a natural way.

On the other hand, the study of the families of orthogonal polynomials related to inner products defined by

$$\langle f, g\rangle = \int_I fg\,d\mu + \lambda \int_I f'g'\,d\mu$$

and the properties of their zeros was begun by Althammer [1], Cohen [6], and Schäfke [22] in the case of Lebesgue measure with $I = (-1, 1)$, by Brenner [4] in the case $d\mu = e^{-x}\,dx$ with $I = (0, +\infty)$, and by Schäfke and Wolf [23] for the classical weights in the corresponding intervals I.

More recently, a group of Dutch mathematicians have considered similar problems for inner products

$$\langle f, g\rangle = \int_I fg\,d\mu + \sum_{k=0}^{n} \lambda_k f^{(k)}(0)g^{(k)}(0)$$

when $I = (0, +\infty)$ and μ is the Laguerre measure [11] or a q-discrete measure [12], as well as when μ is the Gegenbauer measure and $\langle \cdot, \cdot \rangle$ is given by

$$\langle f, g\rangle = \int_{-1}^{1} fg\,d\mu + M[f(-1)g(-1) + f(1)g(1)] + N[f'(-1)g'(-1) + f'(1)g'(1)]$$

with $I = (-1, 1)$, (see [2], [3]).

* Received by the editors February 27, 1991; accepted for publication (in revised form) September 18, 1991.

† Departamento de Matématicas, Universidad de Zaragoza, 50009 Zaragoza, Spain. The research of these authors was partially supported by Proyecto Ayuda a la Investigación 1990 (University of Zaragoza), 227-36.

‡ Departamento de Matemática Aplicada, Universidad Politecnica de Madrid, E. T. S. Ingenieros Industriales, C. Jose Gutierrez Abascal, 2, 28006 Madrid, Spain. This research was partially supported by Comisión Interministerial de Ciencia y Tecnología. PB 89-0181-CO2-01.

§ Laboratoire de Physique Mathematique, Facultés Universitaires N. D. de la Paix, Rue de Bruxelles, 61, 5000 Namur, Belgium. The research was partially supported by North Atlantic Treaty Organization Scientific Affairs Division grant 0564/88.

Besides, Marcellán and Ronveaux [17] have studied the most general situation when the inner product is

$$\langle f, g\rangle = \int_I fg\, d\mu + \lambda f^{(r)}(c) g^{(r)}(c),$$

where $\lambda \in \mathbf{R}^+$ and $c \in \mathbf{R}$.

Finally, results relative to zeros have been the object of a very recent work by Meijer [20] and asymptotic properties have been obtained by Marcellán and van Assche [18].

The aim of this paper is to present the most general possible treatment of the families of orthogonal polynomials associated to an inner product of type

$$\langle f, g\rangle = \int_I fg\, d\mu + Mf(c)g(c) + Nf'(c)g'(c)$$

with $c \in \mathbf{R}$ and $M, N \geqq 0$.

In § 2, we study the algebraic properties of these orthogonal polynomials. An explicit representation in terms of the orthogonal polynomials associated to μ is given, as well as a five term recurrence relation, which is based on the self-adjoint character of a certain multiplication operator in the space of the polynomials. Moreover, a relation between the corresponding kernels and an analog of the Christoffel-Darboux formula is presented.

In § 3, we obtain results related to the distribution of the zeros, showing the dependence of this distribution from the position of the point c with respect to the support of the measure μ. Estimations about the position of the greatest zero are given.

In § 4, we consider a particularly simple situation corresponding to symmetric measures. In this case we can improve the results related to the zeros.

In § 5, we expose an application for semiclassical measures, deriving a second-order linear differential equation satisfied by the new orthogonal polynomials. Finally, two particularly interesting cases are considered: one of them deals with the case of Poisson's distribution, as an example of a discrete measure, and the other one corresponds to the case of Gegenbauer measure with $c = 0$. In the latter, the mass is placed in an interior point of the support, unlike the usual location of masses in the ends of the support. This simplifies the calculations very much.

2. Algebraic properties.

2.1. Representation formulas. Let μ be a positive Borel measure on an interval (finite or infinite) $I \subset \mathbf{R}$ with infinite support such that all the moments $\int_I x^n\, d\mu$ exist. We define the following real inner product in the linear space of real polynomials $\mathcal{P}$:

$$\langle f, g\rangle = \int_I fg\, d\mu + Mf(c)g(c) + Nf'(c)g'(c),$$

where $c \in \mathbf{R}$ and $M, N \geqq 0$. This inner product cannot be associated to any positive measure on I in the standard sense [7], whenever $N > 0$.

Let $(P_n(x)) = (P_n)$ and $(Q_n(x)) = (Q_n)$ be the sequences of monic orthogonal polynomials (SMOP) with respect to μ and with the inner product $\langle \cdot, \cdot \rangle$, respectively. If we consider the representation of Q_n in terms of P_j,

$$Q_n(x) = P_n(x) + \sum_{j=0}^{n-1} \alpha_{nj} P_j(x)$$

from the orthogonality of Q_n with respect to P_j, $j=0, 1, \cdots, n-1$, it follows that

$$\alpha_{nj}=\frac{\int_I Q_nP_j\,d\mu}{\int_I P_j^2\,d\mu}=-\frac{MQ_n(c)P_j(c)+NQ'_n(c)P'_j(c)}{\|P_j\|^2} \qquad 0\leqq j\leqq n-1.$$

Then

$$Q_n(x)=P_n(x)-MQ_n(c)K_{n-1}(x, c)-NQ'_n(c)K^{(0,1)}_{n-1}(x, c), \tag{2.1}$$

where $(K_n(x, y))$ is the sequence of kernels associated to (P_n), and $K_n^{(r,s)}(x, y)$ denotes the generalized kernel

$$K_n^{(r,s)}(x, y)=\sum_{j=0}^{n}\frac{P_j^{(r)}(x)P_j^{(s)}(y)}{\|P_j\|^2}.$$

If we derive in (2.1) with respect to x and evaluating at $x=c$, the values $Q_n(c)$ and $Q'_n(c)$ can be expressed as the solutions of the system,

$$\begin{aligned} P_n(c)&=Q_n(c)[1+MK_{n-1}(c, c)]+Q'_n(c)NK^{(0,1)}_{n-1}(c, c),\\ P'_n(c)&=Q_n(c)MK^{(0,1)}_{n-1}(c, c)+Q'_n(c)[1+NK^{(1,1)}_{n-1}(c, c)], \end{aligned} \tag{2.2}$$

whose determinant:

$$\begin{aligned} D=1&+MK_{n-1}(c, c)+NK^{(1,1)}_{n-1}(c, c)\\ &+MN[K_{n-1}(c, c)K^{(1,1)}_{n-1}(c, c)-K^{(0,1)}_{n-1}(c, c)^2] \end{aligned}$$

is positive from the Cauchy-Schwartz inequality. Therefore,

$$\begin{aligned} Q_n(c)&=\frac{P_n(c)[1+NK^{(1,1)}_{n-1}(c, c)]-P'_n(c)NK^{(0,1)}_{n-1}(c, c)}{D},\\ Q'_n(c)&=\frac{-P_n(c)MK^{(0,1)}_{n-1}(c, c)+P'_n(c)[1+MK_{n-1}(c, c)]}{D}. \end{aligned} \tag{2.3}$$

Then (2.1) becomes

$$\begin{aligned} Q_n(x)=P_n(x)&-M\frac{P_n(c)[1+NK^{(1,1)}_{n-1}(c, c)]-P'_n(c)NK^{(0,1)}_{n-1}(c, c)}{D}K_{n-1}(x, c)\\ &-N\frac{-P_n(c)MK^{(0,1)}_{n-1}(c, c)+P'_n(c)[1+MK_{n-1}(c, c)]}{D}K^{(0,1)}_{n-1}(x, c). \end{aligned} \tag{2.4}$$

We need establish some auxiliary results.

LEMMA 2.1. *Let $(P_n^c(x))$ and $(P_n^{c,c}(x))$ be the* SMOP *with respect to the measures $(x-c)^2\,d\mu$ and $(x-c)^4\,d\mu$, respectively. Then*:

$$(x-c)P^c_{n-1}(x)=P_n(x)-\frac{P_n(c)}{K_{n-1}(c, c)}K_{n-1}(x, c), \tag{2.5}$$

$$P^c_{n-1}(c)=P'_n(c)-\frac{P_n(c)}{K_{n-1}(c, c)}K^{(0,1)}_{n-1}(c, c), \tag{2.6}$$

$$(x-c)P^{c,c}_{n-2}(x)=P^c_{n-1}(x)-\frac{P^c_{n-1}(c)}{K^c_{n-2}(c, c)}K^c_{n-2}(x, c), \tag{2.7}$$

$$(x-c)(y-c)K^c_{n-1}(x, y)=K_n(x, y)-\frac{K_n(x, c)K_n(c, y)}{K_n(c, c)}, \tag{2.8}$$

$$(2.9) \qquad (x-c)K^c_{n-1}(x,c)=K^{(0,1)}_n(x,c)-\frac{K^{(0,1)}_n(c,c)}{K_n(c,c)}K_n(x,c),$$

$$(2.9') \qquad K^c_{n-1}(c,c)=K^{(1,1)}_n(c,c)-\frac{[K^{(0,1)}_n(c,c)]^2}{K_n(c,c)},$$

where $(K^c_n(x,y))$ *denotes the sequence of kernels associated to* (P^c_n).

Proof. Let us consider the representation of $(x-c)P^c_{n-1}(x)$ in terms of $P_j(x)$:

$$(x-c)P^c_{n-1}(x)=P_n(x)+\sum_{j=0}^{n-1}\alpha_{n-1,j}P_j(x).$$

By using the orthogonality of the sequence (P^c_{n-1}) with respect to the measure $(x-c)^2\,d\mu$ we get:

$$\alpha_{n-1,0}=\frac{P_0(c)}{\|P_0\|^2}\int_I (x-c)P^c_{n-1}(x)\,d\mu$$

and

$$\begin{aligned}\alpha_{n-1,j}&=\frac{1}{\|P_j\|^2}\int_I (x-c)P^c_{n-1}(x)P_j(x)\,d\mu\\ &=\frac{1}{\|P_j\|^2}\left[\int_I P^c_{n-1}(x)\frac{P_j(x)-P_j(c)}{x-c}(x-c)^2\,d\mu+P_j(c)\int_I (x-c)P^c_{n-1}(x)\,d\mu\right]\\ &=\frac{P_j(c)}{\|P_j\|^2}\int_I (x-c)P^c_{n-1}(x)\,d\mu\end{aligned}$$

if $j=1,\cdots,n-1$.

Then

$$(x-c)P^c_{n-1}(x)=P_n(x)+K_{n-1}(x,c)\int_I (t-c)P^c_{n-1}(t)\,d\mu(t).$$

Evaluating at $x=c$, it follows the value of the last integral and, therefore, (2.5).

In order to prove (2.7) it suffices to consider the representation of $(x-c)P^{c,c}_{n-2}(x)$ in terms of $P^c_j(x)$ and to repeat the above argument.

If we derive (2.5) with respect to x and evaluating at $x=c$, we deduce (2.6).

Formula (2.8) can be obtained from the representation

$$(x-c)(y-c)K^c_{n-1}(x,y)=\sum_{j=0}^{n}\beta_{n-1,j}(y)P_j(x).$$

By using the reproducing property of the kernels and the orthogonality of P_n we have:

$$\beta_{n-1,0}(y)=\frac{(y-c)}{\|P_0\|^2}P_0(c)\int_I (x-c)K^c_{n-1}(x,y)\,d\mu(x)$$

and

$$\begin{aligned}\beta_{n-1,j}(y)&=\frac{(y-c)}{\|P_j\|^2}\int_I (x-c)K^c_{n-1}(x,y)P_j(x)\,d\mu(x)\\ &=\frac{(y-c)}{\|P_j\|^2}\left[\int_I K^c_{n-1}(x,y)\frac{P_j(x)-P_j(c)}{x-c}(x-c)^2\,d\mu(x)\right.\\ &\qquad\left.+P_j(c)\int_I (x-c)K^c_{n-1}(x,y)\,d\mu(x)\right]\end{aligned}$$

$$=\frac{1}{\|P_j\|^2}\left[P_j(y)-P_j(c)+(y-c)P_j(c)\int_I (x-c)K^c_{n-1}(x,y)\,d\mu(x)\right]$$

for every $j=1,\cdots,n$.

Then,

$$(x-c)(y-c)K^c_{n-1}(x,y)=K_n(x,y)-K_n(x,c)$$
$$+(y-c)K_n(x,c)\int_I (t-c)K^c_{n-1}(x,t)\,d\mu(t).$$

Now, formula (2.8) can be derived directly from the last one.

By derivation in (2.8) with respect to y and evaluating at $y=c$ we get (2.9). In a similar way, we deduce (2.9′) from (2.9). □

The above lemma allows us to represent the kernels $K_{n-1}(x,c)$ and $K^{(0,1)}_{n-1}(x,c)$ in terms of the polynomials $P_n(x)$, $P^c_{n-1}(x)$, and $P^{c,c}_{n-2}(x)$. By substitution of these values in (2.4) we obtain the following.

PROPOSITION 2.2. *Let c be such that the condition $P_n(c)P^c_{n-1}(c)\neq 0$ is satisfied for every $n\in N$. Then, the formula*

$$\begin{aligned}Q_n(x)&=(1-\alpha_n)P_n(x)+(\alpha_n-\beta_n)(x-c)P^c_{n-1}(x)\\&\quad+\beta_n(x-c)^2P^{c,c}_{n-2}(x),\end{aligned}\tag{2.10}$$

where

$$\alpha_n=1-\frac{Q_n(c)}{P_n(c)}=1-\frac{[1+NK^{(1,1)}_{n-1}(c,c)]P_n(c)-NK^{(0,1)}_{n-1}(c,c)P'_n(c)}{DP_n(c)},$$

$$\beta_n=\frac{NQ'_n(c)K^c_{n-2}(c,c)}{P^c_{n-1}(c)}$$

holds.

Remarks. (1) Since all the zeros of the polynomials $P_n(x)$ and $P^c_{n-1}(x)$ are in the interior of the interval I, we conclude that if c is not an interior point of I, then the formula (2.10) is true.

(2) The polynomials P^c_n have been identified by Kautsky and Golub (see [10]). By using methods of linear algebra they prove the fact that, if J is the Jacobi matrix associated with (P_n), a single step of the QR algorithm with the (Wilkinson) shift c corresponds to finding the Jacobi matrix associated with (P^c_n). A proof of this result by an analytic technique can be found in [5].

Let us consider the Christoffel-Darboux formula for the kernel $K_{n-1}(x,y)$ ([7, Thm. 4.5, p. 23] or [24, Thm. 3.2.2, p. 43]). For the first consequence below we evaluate at $y=c$, and for the second we derive with respect to y and evaluate at $y=c$.

$$(x-c)K_{n-1}(x,c)=\frac{1}{\|P_{n-1}\|^2}[P_n(x)P_{n-1}(c)-P_{n-1}(x)P_n(c)],$$

$$\begin{aligned}(x-c)^2K^{(0,1)}_{n-1}(x,c)&=\frac{1}{\|P_{n-1}\|^2}[P_n(x)\{P_{n-1}(c)+(x-c)P'_{n-1}(c)\}\\&\quad-P_{n-1}(x)\{P_n(c)+(x-c)P'_n(c)\}].\end{aligned}\tag{2.11}$$

Multiplying the formula (2.4) by $(x-c)^2$ and substituting (2.11) we obtain

$$(x-c)^2Q_n(x)=q_2(x,n)P_n(x)+q_1(x,n)P_{n-1}(x),\tag{2.12}$$

where

$$q_2(x, n) = (x-c)^2 - \frac{MQ_n(c)}{\|P_{n-1}\|^2} \sum_{k=0}^{1} P_{n-1}^{(k-1)}(c)(x-c)^k$$

$$-\frac{NQ_n'(c)}{\|P_{n-1}\|^2} \sum_{k=0}^{1} P_{n-1}^{(k)}(c)(x-c)^k,$$

$$q_1(x, n) = \frac{MQ_n(c)}{\|P_{n-1}\|^2} \sum_{k=0}^{1} P_n^{(k-1)}(c)(x-c)^k$$

$$+\frac{NQ_n'(c)}{\|P_{n-1}\|^2} \sum_{k=0}^{1} P_n^{(k)}(c)(x-c)^k.$$

(We denote $P_n^{(-1)}(c) = 0$.)

From (2.12), it follows that the sequence (Q_n) is strictly quasi orthogonal of order 2 with respect to the measure $(x-c)^2\, d\mu$ [19] and, therefore,

$$(x-c)^2 Q_n(x) = P_{n+2}(x) + \sum_{j=n-2}^{n+1} a_{nj} P_j(x), \tag{2.13}$$

where the numbers a_{nj} can be expressed in terms of the coefficients of the polynomials $q_2(x, n)$, $q_1(x, n)$, and the coefficients of the three term recurrence relation satisfied by the SMOP (P_n).

Now, we can obtain a recurrence relation for the orthogonal polynomials Q_n.

PROPOSITION 2.3. *The polynomials Q_n satisfy a five term recurrence relation*:

$$(x-c)^2 Q_n(x) = Q_{n+2}(x) + \sum_{j=n-2}^{n+1} \gamma_{nj} Q_j(x) \qquad n \geqq 0, \tag{2.14}$$

where $\gamma_{n,n-2} > 0$ $(n \geqq 2)$ and the convention $Q_{-1} = Q_{-2} = 0$.

Proof. Let

$$(x-c)^2 Q_n(x) = \sum_{j=0}^{n+2} \gamma_{nj} Q_j(x)$$

be the expansion of the polynomial $(x-c)^2 Q_n(x)$ with respect to the sequence (Q_n).

Obviously, $\gamma_{n,n+2} = 1$ because Q_n is monic. On the other hand, if $0 \leqq j < n-2$, $\gamma_{n,j} = 0$ from the orthogonality of the sequence (Q_n).

The remaining coefficients $\gamma_{n,j}$ can be found as follows: from the definition of the inner product, if $n-2 \leqq j \leqq n+1$

$$\gamma_{n,j} = \frac{\langle (x-c)^2 Q_n(x), Q_j(x) \rangle}{\langle Q_j, Q_j \rangle}$$

$$= \frac{1}{\langle Q_j, Q_j \rangle} \int_I (x-c)^2 Q_n(x) Q_j(x)\, d\mu(x).$$

But from (2.13),

$$(x-c)^2 Q_j(x) = \sum_{h=j-2}^{j+2} a_{jh} P_h(x) \tag{2.15}$$

with $a_{j,j+2} = 1$, and from (2.1)

$$Q_n(x) = \sum_{h=0}^{n} \beta_{nh} P_h(x),$$

where $\beta_{nn}=1$, and if $h<n$

$$\beta_{nh}=\frac{-1}{\|P_h\|^2}[MQ_n(c)P_h(c)+NQ'_n(c)P'_h(c)],$$

then,

$$\begin{aligned}\int_I (x-c)^2 Q_n(x)Q_j(x)\,d\mu(x)\\ &=a_{jn}\|P_n\|^2+\sum_{h=j-2}^{n-1}\beta_{nh}a_{jh}\|P_h\|^2\\ &=a_{jn}\|P_n\|^2-\sum_{h=j-2}^{n-1}a_{jh}[MQ_n(c)P_h(c)+NQ'_n(c)P'_h(c)].\end{aligned}$$

Also, from (2.15)

$$\sum_{h=j-2}^{j+2}a_{jh}P_h(c)=\sum_{h=j-2}^{j+2}a_{jh}P'_h(c)=0,$$

and hence

$$(2.16)\qquad \gamma_{nj}=\langle Q_j,Q_j\rangle^{-1}\left[a_{jn}\|P_n\|^2+MQ_n(c)\sum_{h=n}^{j+2}a_{jh}P_h(c)+NQ'_n(c)\sum_{h=n}^{j+2}a_{jh}P'_h(c)\right]$$

holds. Finally, from the definition of the inner product $\langle\ ,\ \rangle$

$$\langle Q_n,Q_n\rangle=\|P_n\|^2+MQ_n(c)P_n(c)+NQ'_n(c)P'_n(c).$$

So, if $j=n-2$ we get

$$\gamma_{n,n-2}=\frac{\langle Q_n,Q_n\rangle}{\langle Q_{n-2},Q_{n-2}\rangle}>0. \qquad \square$$

Remark. In the above proposition we have pointed out that

$$\langle Q_n,Q_n\rangle=\|P_n\|^2+MQ_n(c)P_n(c)+NQ'_n(c)P'_n(c).$$

An explicit expression of $\langle Q_n,Q_n\rangle$ in terms of M, N, and the polynomials P_n can be derived by using (2.3). We find, by straightforward calculations,

$$(2.17)\qquad \langle Q_n,Q_n\rangle=\|P_n\|^2\frac{\gamma_{n-1}\lambda_n+\gamma_n\lambda_{n-1}-2MNK_n^{(0,1)}(c,c)K_{n-1}^{(0,1)}(c,c)-D}{D},$$

where $\gamma_n=1+MK_n(c,c)$ and $\lambda_n=1+NK_n^{(1,1)}(c,c)$.

2.2. Kernels. We are going to derive a formula relating the kernel associated to the new polynomials Q_n with the kernels $K_n(x,c)$ and $K_{n-1}^{(0,1)}(x,c)$.

Let

$$L_n(x,y)=\sum_{h=0}^{n}\frac{Q_h(x)Q_h(y)}{\langle Q_h,Q_h\rangle}.$$

If we consider its expansion in terms of the polynomials $P_j(x)$,

$$L_n(x,y)=\sum_{j=0}^{n}\alpha_{nj}(y)P_j(x),$$

we have:

$$\begin{aligned}\alpha_{nj}(y) &= \int_I L_n(x, y)\frac{P_j(x)}{\|P_j\|^2}\,d\mu \\ &= \left\langle L_n(x, y), \frac{P_j(x)}{\|P_j\|^2}\right\rangle - ML_n(c, y)\frac{P_j(c)}{\|P_j\|^2} - NL_n^{(1,0)}(c, y)\frac{P_j'(c)}{\|P_j\|^2} \\ &= \frac{1}{\|P_j\|^2}[P_j(y) - ML_n(c, y)P_j(c) - NL_n^{(1,0)}(c, y)P_j'(c)].\end{aligned}$$

This proves that the kernels $L_n(x, y)$ and $K_n(x, y)$ satisfy the following formula:

$$\begin{aligned}L_n(x, y) = K_n(x, y) - ML_n(c, y)K_n(x, c) \\ - NL_n^{(1,0)}(c, y)K_n^{(0,1)}(x, c).\end{aligned} \tag{2.18}$$

Explicit expressions for $L_n(c, y)$ and $L_n^{(1,0)}(c, y)$ can be obtained as solutions of the system

$$\begin{aligned}K_n(c, y) &= L_n(c, y)[1 + MK_n(c, c)] + L_n^{(1,0)}(c, y)NK_n^{(1,0)}(c, c), \\ K_n^{(1,0)}(c, y) &= L_n(c, y)MK_n^{(1,0)}(c, c) \\ &\quad + L_n^{(1,0)}(c, y)[1 + NK_n^{(1,1)}(c, c)].\end{aligned} \tag{2.19}$$

Now, we obtain an analog of the Christoffel-Darboux formula for the new polynomials.

PROPOSITION 2.4. *The relation*

$$\begin{aligned}(x + y &- 2c)(x - y)L_n(x, y) \\ &= \frac{1}{\langle Q_n, Q_n\rangle}[Q_{n+2}(x)Q_n(y) - Q_{n+2}(y)Q_n(x)] \\ &+ \frac{\gamma_{n,n+1}}{\langle Q_n, Q_n\rangle}[Q_{n+1}(x)Q_n(y) - Q_{n+1}(y)Q_n(x)] \\ &+ \frac{1}{\langle Q_{n-1}, Q_{n-1}\rangle}[Q_{n+1}(x)Q_{n-1}(y) - Q_{n+1}(y)Q_{n-1}(x)]\end{aligned} \tag{2.20}$$

and its confluent form

$$\begin{aligned}2(x - c)L_n(x, x) &= \frac{1}{\langle Q_n, Q_n\rangle}[Q_{n+2}'(x)Q_n(x) - Q_{n+2}(x)Q_n'(x)] \\ &+ \frac{\gamma_{n,n+1}}{\langle Q_n, Q_n\rangle}[Q_{n+1}'(x)Q_n(x) - Q_{n+1}(x)Q_n'(x)] \\ &+ \frac{1}{\langle Q_{n-1}, Q_{n-1}\rangle}[Q_{n+1}'(x)Q_{n-1}(x) - Q_{n+1}(x)Q_{n-1}'(x)]\end{aligned} \tag{2.21}$$

hold.

Proof. Multiplying in the relation (2.14) by $Q_n(y)$ and multiplying the same relation evaluated at $x = y$ by $Q_n(x)$, we obtain after subtraction:

$$(x + y - 2c)(x - y)Q_n(x)Q_n(y) = \sum_{0\neq k=-2}^{2} \gamma_{n,n+k}[Q_{n+k}(x)Q_n(y) - Q_{n+k}(y)Q_n(x)], \tag{2.22}$$

where $\gamma_{n,n+2} = 1$.

On the other hand, the inner product $\langle\, ,\rangle$ is such that

$$\langle (x-c)^2 Q_n(x),\, Q_m(x)\rangle = \langle Q_n(x),\, (x-c)^2 Q_m(x)\rangle$$

for all $n, m \in N$. Hence, as

$$(x-c)^2 Q_{n-i}(x) = \sum_{k=i-2}^{i+2} \gamma_{n-i,n-k} Q_{n-k}(x)$$

we get

$$\gamma_{n-i,n-k}\langle Q_{n-k}, Q_{n-k}\rangle = \gamma_{n-k,n-i}\langle Q_{n-i}, Q_{n-i}\rangle, \qquad k-2 \leqq i \leqq k+2. \tag{2.23}$$

From (2.22) and (2.23), by straightforward calculations, we get (2.20).

The result in (2.21) follows immediately from (2.20). □

3. Zeros of Q_n. In this section, we always consider $N>0$. It is well known that the zeros of P_n are real, simple, and belong to $\mathring{I}$ ($\mathring{I}$ denotes the interior of the true interval of orthogonality I). But this result may be false for polynomials Q_n. In fact, the general result we can prove is the following.

PROPOSITION 3.1. *If $n \geqq 3$, the polynomial Q_n has at least $n-2$ different zeros with odd multiplicity in $\mathring{I}$.*

Proof. Let $\xi_{n1}, \cdots, \xi_{nk}$ denote all the distinct zeros of Q_n of odd multiplicity which are in $\mathring{I}$. Define $p(x) = (x-\xi_{n1})\cdots(x-\xi_{nk})$. The polynomial $(x-c)^2 p(x) Q_n(x)$ does not change sign in the interval I; hence,

$$\langle (x-c)^2 p(x) Q_n(x), 1\rangle = \int_I (x-c)^2 p(x) Q_n(x)\, d\mu(x) \neq 0.$$

Since (Q_n) is a quasi-orthogonal sequence of order 2 with respect to $(x-c)^2\, d\mu$, it follows that $\deg p(x) \geqq n-2$. □

PROPOSITION 3.2. *The zeros of the polynomial Q_n are real, simple, and at least $n-1$ of them belong to $\mathring{I}$, whenever either $c = \inf I$ or $c = \sup I$.*

Proof. Suppose $c = \sup I$. Let $\xi_{n1}, \cdots, \xi_{nk}$ denote all the zeros of Q_n in $\mathring{I}$. From Proposition 3.1, it follows that $k \geqq n-2$. Set $p(x) = (x-\xi_{n1})\cdots(x-\xi_{nk})$; then, the polynomials $p(x)Q_n(x)$ and $(x-c)p(x)Q_n(x)$ have constant but opposite signs in $\mathring{I}$.

If $Q_n'(c) = 0$, we have

$$\langle (x-c)p(x),\, Q_n(x)\rangle = \int_I (x-c)p(x)Q_n(x)\, d\mu(x) \neq 0,$$

and hence, $\deg p(x) \geqq n-1$.

Let $Q_n'(c) \neq 0$. If we suppose $k = n-2$, the following formulas hold:

$$0 = \langle (x-c)p(x),\, Q_n(x)\rangle = \int_I (x-c)p(x)Q_n(x)\, d\mu + Np(c)(Q_n'(c)$$

$$0 = \langle (p(x),\, Q_n(x)\rangle = \int_I p(x)Q_n(x)\, d\mu + Mp(c)Q_n(c) + Np'(c)Q_n'(c).$$

Hence, $p(c)Q_n'(c)$ and $p'(c)Q_n'(c)$ have opposite signs, which is a contradiction. Thus $k \geqq n-1$. As a consequence, all the zeros of $Q_n(x)$ are real and simple.

If $c = \inf I$, the proof is similar. □

Remark. We want to note that if we consider the inner product

$$\langle f, g\rangle = \int_I f(x)g(x)\, d\mu(x) + Mf(c)g(c) + Nf^{(r)}(c)g^{(r)}(c)$$

with $r \in N$, by using the above arguments, we can deduce that the polynomial Q_n associated to the new inner product has at least $n-(r+1)$ different zeros in $\mathring{I}$. Whenever either $c = \sup I$ or $c = \inf I$, then Q_n has at least $n-1$ zeros in $\mathring{I}$ and, therefore, all the zeros are real and simple (see [11]).

Note that if $c = \sup I$ and all the roots of $Q_n(x)$ are located in the interior of I, then both conditions

$$Q_n(c) > 0 \quad \text{and} \quad Q_n'(c) > 0 \tag{3.1}$$

hold. In the similar way, if $c = \inf I$ and all the roots of $Q_n(x)$ belong to $\mathring{I}$, then

$$\operatorname{sgn} Q_n(c) = (-1)^n \quad \text{and} \quad \operatorname{sgn} Q_n'(c) = (-1)^{n-1} \tag{3.2}$$

hold.

This remark allows us to easily deduce sufficient conditions to assure a zero of $Q_n(x)$ is not in $\mathring{I}$, and besides we can give some results about its location.

From now on, if $c = \sup I$ or $c = \inf I$, we shall denote the zeros of $Q_n(x)$ being ordered by increasing size: $\xi_{n1} < \cdots < \xi_{nn}$.

PROPOSITION 3.3. *The following statements hold*:

(a) *Let $c = \sup I$. If the property* (3.1) *is not true then the greatest zero of $Q_n(x)$ satisfies*

$$c \leqq \xi_{nn} < c + \frac{c - \xi_{n1}}{n-1} \quad \textit{and} \quad |\xi_{nn} - c| < |\xi_{n,n-1} - c|.$$

Moreover, if $M \neq 0$, $\xi_{nn} - c < \frac{1}{2}\sqrt{N/M}$.

(b) *Let $c = \inf I$. If the property* (3.2) *is not true then the lowest zero of $Q_n(x)$ satisfies*

$$c - \frac{\xi_{nn} - c}{n-1} < \xi_{n1} \leqq c \quad \textit{and} \quad |\xi_{n1} - c| < |\xi_{n2} - c|.$$

Moreover, if $M \neq 0$, $c - \xi_{n1} < \frac{1}{2}\sqrt{N/M}$.

Proof. It suffices to prove (a). It is easy to deduce that if (3.1) is not true, we have $c \leqq \xi_{nn}$.

Assume $c < \xi_{nn}$, then $Q_n(c) = (c - \xi_{n1}) \cdots (c - \xi_{nn}) < 0$. In this situation,

$$K_{n-1}^{(0,1)}(c, c) = \sum_{h=0}^{n-1} \frac{P_h(c) P_h'(c)}{\|P_h\|^2} > 0,$$

from (2.2) it follows $Q_n'(c) > 0$. Since

$$\frac{Q_n'(c)}{Q_n(c)} = \sum_{j=1}^{n-1} \frac{1}{c - \xi_{nj}} - \frac{1}{\xi_{nn} - c},$$

we get

$$\frac{1}{\xi_{nn} - c} > \sum_{j=1}^{n-1} \frac{1}{c - \xi_{nj}} > \frac{n-1}{c - \xi_{n1}}.$$

Hence,

$$\xi_{nn} < c + \frac{c - \xi_{n1}}{n-1} \quad \text{and} \quad |\xi_{nn} - c| < |\xi_{n,n-1} - c|.$$

Now, let us set $Q_n(x) = (\xi_{nn} - x)\varphi(x)$. Then,

$$\langle Q_n, \varphi \rangle = \int_I Q_n \varphi \, d\mu + M Q_n(c) \varphi(c) + N Q_n'(c) \varphi'(c) = 0.$$

As $Q_n(x)\varphi(x)>0$ in I, in the above formula the integral is positive and so

$$MQ_n(c)\varphi(c)+NQ_n'(c)\varphi'(c)=(\xi_{nn}-c)[M\varphi(c)^2+N\varphi'(c)^2]\\ -N\varphi(c)\varphi'(c)<0.$$

Whenever $M>0$, taking into account that $\varphi(c)<0$ and $\varphi'(c)<0$ and by using the Cauchy-Schwarz inequality, we obtain

$$\xi_{nn}-c<\frac{1}{2}\sqrt{\frac{N}{M}}. \qquad \square$$

Remark. The same results for $c=0$ and the Laguerre weight have been obtained in [13], and for some generalizations of the Laguerre weight see [20].

4. Analysis of the symmetric case. If I is a symmetric interval and the measure μ is symmetric on I (i.e., $\mu(A)=\mu(-A)$ for every $A\subset I$ measurable), it is well known [7, Thm. 4.3] that the SMOP (P_n) associated to μ satisfies $P_n(-x)=(-1)^nP_n(x)$ for all $n\in N$. As examples of this situation we have Hermite polynomials and Gegenbauer polynomials. We want to emphasize that the condition $P_n(-x)=(-1)^nP_n(x)$ for all $n\in N$ is equivalent to $K_n^{(0,1)}(0,0)=0$ for all $n\in N$.

Let us consider the condition

$$K_n^{(0,1)}(c,c)=0 \quad \text{for every } n\in N \tag{4.1}$$

is satisfied. Let us remark that

(i) $P_n(c)P_n'(c)=0$ for every $n\in N$;

(ii) $P_n(c)P_{n-1}(c)=0$ and $P_n'(c)P_{n-1}'(c)=0$ for every $n\in N$

are separately equivalent to (4.1). From (i) or (ii), it follows that c must belong to $\mathring{I}$. Furthermore there exists at most one c, which is determined by $P_1(c)=0$. Then, in general, we have

$$P_{2n-1}(c)=0 \quad \text{and} \quad P_{2n}'(c)=0 \quad \text{for every } n\in N.$$

We point out that no number c satisfies (4.1) for Jacobi polynomials with $a\neq\beta$ or for Laguerre polynomials.

Now, it is not difficult to prove the polynomials P_n are symmetric with respect to the point c is equivalent to the condition (4.1). Since translation of the centre of symmetry is trivial, in the sequel, we assume (with absolutely no loss of generality) that $c=0$.

Since the determinant D is

$$D=[1+MK_{n-1}(0,0)][1+NK_{n-1}^{(1,1)}(0,0)],$$

we achieve

$$\begin{aligned} Q_n(0)&=\frac{P_n(0)}{1+MK_{n-1}(0,0)},\\ Q_n'(0)&=\frac{P_n'(0)}{1+NK_{n-1}^{(1,1)}(0,0)}. \end{aligned} \tag{4.2}$$

Then (2.4) becomes

$$\begin{aligned} Q_{2n}(x)&=P_{2n}(x)-\frac{MP_{2n}(0)}{1+MK_{2n-1}(0,0)}K_{2n-1}(x,0),\\ Q_{2n+1}(x)&=P_{2n+1}(x)-\frac{NP_{2n+1}'(0)}{1+NK_{2n}^{(1,1)}(0,0)}K_{2n}^{(0,1)}(x,0). \end{aligned} \tag{4.3}$$

Some properties about the quantities $Q_n(0)$ and $Q'_n(0)$ can be derived directly from (4.1) and (4.2). For instance:

(a) $Q_{2n}(0) \neq 0$ and $Q_{2n-1}(0) = 0$ for every $n \in N$;

(b) $Q'_{2n}(0) = 0$ and $Q'_{2n-1}(0) \neq 0$ for every $n \in N$;

(c) sign $Q_n(0) =$ sign $P_n(0)$ and sign $Q'_n(0) =$ sign $P'_n(0)$ for every $n \in N$.

In order to obtain Proposition 2.2 we might impose $P_n(0)P'_n(0) \neq 0$ for every $n \in N$. This restriction is not necessary now. Indeed, from (2.5)

$$P_{2n}(0)K_{2n-1}(x, 0) = K_{2n-1}(0, 0)[P_{2n}(x) - xP^c_{2n-1}(x)],$$

and from (2.9), (2.6), (2.7), (2.5), and (2.9′)

$$P'_{2n+1}(0)K^{(0,1)}_{2n}(x, 0) = K^{(1,1)}_{2n}(0, 0)[P_{2n+1}(x) - x^2P^{c,c}_{2n-1}(x)].$$

By substituting these values in (4.3) we obtain the following.

PROPOSITION 4.1. *The decomposition*:

$$Q_n(x) = (1 - \alpha_n)P_n(x) + (\alpha_n - \beta_n)xP^c_{n-1}(x) + \beta_n x^2 P^{c,c}_{n-2}(x) \tag{4.4}$$

where

$$\alpha_{2n} = \frac{MK_{2n-1}(0, 0)}{1 + MK_{2n-1}(0, 0)} \qquad \beta_{2n} = 0$$

$$\alpha_{2n+1} = \frac{NK^{(1,1)}_{2n}(0, 0)}{1 + NK^{(1,1)}_{2n}(0, 0)} \qquad \beta_{2n+1} = \alpha_{2n+1}$$

holds.

Remark. It is interesting to point out that α_n and β_n are nonnegative and bounded by 1. Consequently, all the coefficients in (4.4) are nonnegative and bounded.

By substituting the values of $Q_n(0)$ and $Q'_n(0)$ (see (4.2)) in (2.12) and simplifying, we obtain:

$$x^2Q_n(x) = [x^2 - a_n]P_n(x) + b_n xP_{n-1}(x), \tag{4.5}$$

where

$$a_n = \frac{1}{\|P_{n-1}\|^2} \frac{NP'_n(0)P_{n-1}(0)}{1 + NK^{(1,1)}_{n-1}(0, 0)},$$

$$b_n = \frac{1}{\|P_{n-1}\|^2}\left[\frac{MP_n(0)^2}{1 + MK_{n-1}(0, 0)} + \frac{NP'_n(0)^2}{1 + NK^{(1,1)}_{n-1}(0, 0)}\right].$$

Note that from the above formula it follows that the polynomials Q_n are also symmetric.

To deduce the recurrence relation we shall employ the expansion of $x^2Q_n(x)$ in terms of the polynomials P_n. Using the three term recurrence formula verified by the SMOP (P_n),

$$xP_n(x) = P_{n+1}(x) + B_{n+1}P_{n-1}(x)$$

and

$$x^2Q_n(x) = P_{n+2}(x) + \sum_{j=n-2}^{n+1} a_{n,j}P_j(x),$$

we find:

$$\begin{aligned} a_{n,n+1} &= 0, \\ a_{n,n} &= B_{n+2} + B_{n+1} - a_n + b_n, \\ a_{n,n-1} &= 0, \\ a_{n,n-2} &= B_n(B_{n+1} + b_n). \end{aligned} \tag{4.6}$$

Substituting these values (2.16) we obtain the coefficients of the five term recurrence relation verified by the SMOP (Q_n). To do this, it suffices to note that if $|m-n|$ is odd,

$$P_m(0)P_n(0) = P'_m(0)P'_n(0) = 0$$

holds. Thus, by using the notations $\gamma_n = 1 + MK_n(0,0)$ and $\lambda_n = 1 + NK_n^{(1,1)}(0,0)$ we can give the following.

PROPOSITION 4.2. *The* SMOP (Q_n) *satisfies the formula*

$$x^2 Q_n(x) = Q_{n+2}(x) + \sum_{j=n-2}^{n+1} \gamma_{n,j} Q_j(x),$$

where

$$\begin{aligned} \gamma_{n,n+1} &= 0, \\ \gamma_{n,n} &= a_{n,n} + \frac{1}{\langle Q_n, Q_n\rangle}\left[\frac{MP_n(0)P_{n+2}(0)}{1+MK_{n-1}(0,0)} + \frac{NP'_n(0)P'_{n+2}(0)}{1+NK_{n-1}^{(1,1)}(0,0)}\right], \\ \gamma_{n,n-1} &= 0, \\ \gamma_{n,n-2} &= \frac{\langle Q_n, Q_n\rangle}{\langle Q_{n-2}, Q_{n-2}\rangle}, \\ \langle Q_n, Q_n\rangle &= \|P_n\|^2\left[\frac{\gamma_n}{\gamma_{n-1}} + \frac{\lambda_n}{\lambda_{n-1}} - 1\right]. \end{aligned} \tag{4.7}$$

Note that, in the symmetric case, the recurrence formula satisfied by the polynomials Q_n is

$$x^2 Q_n(x) = Q_{n+2}(x) + \gamma_{n,n}Q_n(x) + \gamma_{n,n-2}Q_{n-2}(x). \tag{4.8}$$

Moreover, the explicit expression concerning the kernels $L_n(0,y)$ and $L_n^{(1,0)}(0,y)$ is very simple. Then (2.18) becomes

$$\begin{aligned} L_n(x,y) = K_n(x,y) &- \frac{MK_n(0,y)}{1+MK_n(0,0)}K_n(x,0) \\ &- \frac{NK_n^{(1,0)}(0,y)}{1+NK_n^{(1,1)}(0,0)}K_n^{(0,1)}(x,0). \end{aligned} \tag{4.9}$$

PROPOSITION 4.3. *The kernel* $L_n(x,y)$ *associated to the the polynomials* Q_n *can be expressed, in terms of the kernels associated to the polynomials* P_n, P_n^c, *and* $P_n^{c,c}$:

$$L_n(x,y) = r_n K_n(x,y) + s_n xy K_{n-1}^c(x,y) + t_n x^2y^2 K_{n-2}^{c,c}(x,y), \tag{4.10}$$

where

$$\begin{aligned} r_n &= \frac{1}{1+MK_n(0,0)}, \\ s_n &= \frac{MK_n(0,0)}{1+MK_n(0,0)} - \frac{NK_n^{(1,1)}(0,0)}{1+NK_n^{(1,1)}(0,0)}, \\ t_n &= \frac{NK_n^{(1,1)}(0,0)}{1+NK_n^{(1,1)}(0,0)}. \end{aligned}$$

Proof. Using the formulas (2.9), the analog of (2.8) for $K_{n-2}^{c,c}(x,y)$, and (2.9') we obtain

$$K_n^{(0,1)}(x,0)K_n^{(1,0)}(0,y) = K_n^{(1,1)}(0,0)[xyK_{n-1}^c(x,y) - x^2y^2K_{n-2}^{c,c}(x,y)].$$

Then the decomposition (4.10) holds and the explicit expression of coefficients r_n, s_n, t_n is obtained. □

Remarks.

(a) The coefficients in (4.10) are bounded and besides r_n, s_n are nonnegative.

(b) If $N=0$ there is always a decomposition as (4.10). But, if $N\neq 0$ there is such a decomposition if and only if $K_n^{(0,1)}(0,0)=0$ for all $n\in N$.

Next, we shall work in the symmetric case to obtain some strong results about zeros.

PROPOSITION 4.4. *All the zeros of Q_n are real, simple and belong to $\mathring{I}$.*

Proof. By Proposition 3.1, Q_n has at least $n-2$ different zeros in $\mathring{I}$, and all of them have odd multiplicity. As, $Q_n(-x)=(-1)^n Q_n(x)$ for every $x\in I$ and $Q'_{2n-1}(0)\neq 0$ for every $n\in N$, all the zeros of Q_n are simple. Suppose ξ is a complex zero of Q_n; then $\bar{\xi}$ is also a zero of Q_n and hence $-\xi=\bar{\xi}$. Thus $\xi=ir$ with $r\in\mathbf{R}$. Let us denote ξ_{nj}, $j=1,\cdots,n-2$, the remaining zeros of Q_n. Setting $p(x)=(x-\xi_{n1})\cdots(x-\xi_{n,n-2})$ we can write $Q_n(x)=p(x)(x^2+r^2)$. Then

$$\langle p, Q_n\rangle=\int_I p^2(x)(x^2+r^2)\,d\mu(x)+Mr^2p(0)^2+Nr^2p'(0)^2>0,$$

which is a contradiction; hence all the zeros are real.

Finally, we are going to show that ξ and $-\xi$ belong to $\mathring{I}$. Indeed, as $Q_n(x)=p(x)(x^2-\xi^2)$, it follows that $\langle p, Q_n\rangle=0$. But if we suppose $\xi\notin\mathring{I}$, then

$$\langle p, Q_n\rangle=\int_I p^2(x)(x^2-\xi^2)\,d\mu(x)-Mp(0)^2\xi^2-N[p'(0)]^2\xi^2<0.$$

Therefore, $\xi\in\mathring{I}$ holds. □

It is possible as well to deduce a separation property of the zeros. In order to prove it we will use the following.

LEMMA 4.5. *Between two consecutive zeros of $P_n(x)$ there is exactly one zero of $P_{n-1}^c(x)$.* (see [20, Lemma 6.1] or [9, Prop. 1.4.9]).

Since P_n and Q_n have symmetric zeros it suffices to consider the positive zeros. Let M, N be positive, real numbers.

PROPOSITION 4.6. *The positive zeros of P_n and Q_n mutually separate each other and the greatest positive zero of Q_n is less than the greatest positive zero of P_n. Moreover, the positive zeros of Q_{2n} alternate with the positive zeros of P_{2n-1}^c and the positive zeros of Q_{2n+1} alternate with the positive zeros of $P_{2n-1}^{c,c}$.*

Proof. Let us consider $n=2m$. As in (4.4) $\beta_{2m}=0$, we may write

$$Q_{2m}(x)=(1-\alpha_{2m})P_{2m}(x)+\alpha_{2m}xP_{2m-1}^c(x). \tag{4.11}$$

We denote $(x_{2m-1,j})_1^{m-1}$, $(x_{2m,j})_1^m$, $(\xi_{2m,j})_1^m$ the systems of the positive zeros of polynomials P_{2m-1}^c, P_{2m}, and Q_{2m}, respectively, each system arranged by increasing order.

From (4.11) and Lemma 4.5 it follows that whenever $x\geqq x_{2m,m}$, $Q_{2m}(x)>0$, and so $\xi_{2m,m}<x_{2m,m}$. On the other hand, as by Lemma 4.5 $P_{2m}(x_{2m-1,m-1}^c)\leqq 0$, we have $Q_{2m}(x_{2m-1,m-1}^c)\leqq 0$ and so $x_{2m-1,m-1}^c\leqq\xi_{2m,m}$. Since the roots of P_{2m} and P_{2m-1}^c are real and simple using, once more, Lemma 4.5 we have that the sign of $P_{2m-1}^c(x)$ changes in every $x_{2m,j}$ $(j=1,\cdots,m)$ and by (4.11) the sign of $Q_{2m}(x)$ changes in $x_{2m,j}$. Therefore, in each interval $(x_{2m,j-1}, x_{2m,j})$ there exists only one root of Q_{2m}.

In a similar way, the sign of $P_{2m}(x)$ changes in the roots of $P_{2m-1}^c(x)$, and, consequently, the sign of $Q_{2m}(x)$. Hence the positive roots of Q_{2m} and P_{2m-1}^c are interlaced.

If we suppose $n=2m+1$, then $\beta_{2m+1}=\alpha_{2m+1}$ and $P_{2m+1}(x)=xP^c_{2m}(x)$. Thus

$$Q_{2m+1}(x)=(1-\alpha_{2m+1})xP^c_{2m}(x)+\alpha_{2m+1}x^2P^{c,c}_{2m-1}(x).$$

Using the above argument and taking into account that the positive zeros of P_{2m+1} coincide with the positive zeros of P^c_{2m}, the result follows. □

Remark. Note that if $M=0$, $Q_{2m}(x)=P_{2m}(x)$, and if $N=0$, $Q_{2m+1}(x)=P_{2m+1}(x)$.

5. Differential properties.

5.1. Differential equation. Let us consider the case of (P_n) being a sequence of semiclassical orthogonal polynomials (see [19]). This means that the linear functional $\mathscr{L}$ defined by

$$\int_I P\,d\mu=\langle\mathscr{L},P\rangle,\qquad P\in\mathscr{P} \tag{5.1}$$

is characterized by polynomials ϕ and ψ such that a functional equation for $\mathscr{L}$

$$D(\phi\mathscr{L})+\psi\mathscr{L}=0 \tag{5.2}$$

holds with

$$\begin{gathered}\langle\psi\mathscr{L},P\rangle=\langle\mathscr{L},\psi P\rangle,\\ \langle D(\phi\mathscr{L}),P\rangle=-\langle\phi\mathscr{L},P'\rangle\end{gathered} \tag{5.3}$$

for every $P\in\mathscr{P}$.

It is easy to construct a second-order linear differential equation for the SMOP (Q_n) using the representation (2.12), where the polynomials q_2 and q_1 are known explicitly in terms of P_n.

Let us use the structure relation (see [19]) for semiclassical orthogonal polynomials P_n of class s ($s=\max\{(\deg\psi)-1,(\deg\phi)-2\}$).

$$\phi P'_{n+1}=\sum_{k=n-s}^{n+t}\theta_{nk}P_k, \tag{5.4}$$

where $t=\deg\phi$ and θ_{nk} are constants. This relation can be writen

$$\phi P'_{n+1}=C_nP_n+D_nP_{n-1}, \tag{5.5}$$

where the polynomials $C_n=C(x,n)$ and $D_n=D(x,n)$ are computed from the three term recurrence relation for the SMOP (P_n).

The usual 3 step procedure (see [21]) now give the relations

$$(x-c)^2Q_n=q_2P_n+q_1P_{n-1}, \tag{5.6}$$

$$\begin{aligned}\phi[(x-c)^2Q_n]'&=\phi(q_2'P_n+q_1'P_{n-1})+q_2(C_nP_n+D_nP_{n-1})\\ &\quad+q_1(C_{n-1}P_{n-1}+D_{n-1}P_{n-2})\\ &=q_{2,1}P_n+q_{1,1}P_{n-1},\end{aligned} \tag{5.7}$$

$$\phi[\phi[(x-c)^2Q_n]']'=q_{2,2}P_n+q_{1,2}P_{n-1}. \tag{5.8}$$

In the computation of the polynomials $q_{i,j}$ $(i,j=1,2)$, we need again the recurrence relation of the P_n in order to eliminate P_{n-2} in terms of P_n and P_{n-1}.

The following determinant gives the expected differential equation for the sequence (Q_n):

$$\begin{vmatrix}(x-c)^2Q_n & q_2 & q_1\\ \phi[(x-c)^2Q_n]' & q_{2,1} & q_{1,1}\\ \phi\{\phi[(x-c)^2Q_n]'\}' & q_{2,2} & q_{1,2}\end{vmatrix}=0. \tag{5.9}$$

This differential equation becomes particularly simple in the symmetric case with $c=0$. The Hermite case was already treated in [17], so we study here the Gegenbauer case. Bavinck and Meijer also analyze this situation (Gegenbauer case), but with two mass points located at the endpoints of the interval (see [2]).

5.2. Applications. As a first example, we consider the inner product of Sobolev type

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)(1-x^2)^{\lambda-1/2}\,dx + Mf(0)g(0) + Nf'(0)g'(0) \tag{5.10}$$

with $\lambda > -\frac{1}{2}$. In this case, the point c $(c=0)$ is in the support of the measure, and the symmetric character is preserved.

It is well known that the monic Gegenbauer polynomials verify a three term recurrence relation.

$$xP_{n+1}^{(\lambda)}(x) = P_{n+2}^{(\lambda)}(x) + \frac{(n-1)(n+2\lambda)}{4(n+\lambda)(n+\lambda+1)} P_n^{(\lambda)}(x) \qquad n \geqq 0,$$

$$P_0^{(\lambda)}(x) = 1 \qquad P_1^{(\lambda)}(x) = x$$

and

$$P_{2n}^{(\lambda)}(0) = \frac{(-1)^n}{2^{2n}} \frac{(2n)!}{n!} \frac{\Gamma(n+\lambda)}{\Gamma(2n+\lambda)},$$

$$P_{2n}^{(\lambda)\prime}(0) = P_{2n}^{(\lambda)\prime\prime\prime}(0) = 0,$$

$$P_{2n+1}^{(\lambda)}(0) = P_{2n+1}^{(\lambda)\prime\prime}(0) = 0,$$

$$P_{2n+1}^{(\lambda)\prime}(0) = \frac{(2n+1)(n+\lambda)}{2n+\lambda} P_{2n}^{(\lambda)}(0),$$

$$P_{2n}^{(\lambda)\prime\prime}(0) = -4n(n+\lambda)P_{2n}^{(\lambda)}(0),$$

$$P_{2n+1}^{(\lambda)\prime\prime\prime}(0) = -\frac{4n(2n+1)(n+\lambda)(n+\lambda+1)}{2n+\lambda} P_{2n}^{(\lambda)}(0),$$

$$\|P_n^{(\lambda)}\|^2 = 2^{1-2(\lambda+n)}\pi \frac{n!\Gamma(n+2\lambda)}{(n+\lambda)[\Gamma(n+\lambda)]^2}.$$

Moreover, they satisfy a structure relation

$$(x^2-1)P_{n+1}^{(\lambda)\prime}(x) = (n+1)xP_{n+1}^{(\lambda)}(x) - \frac{(n+1)(n+2\lambda)}{2(n+\lambda)} P_n^{(\lambda)}(x) \tag{5.11}$$

(see [24, formula 4.7.27, p. 83]). Thus (4.3) becomes

$$Q_{2n}(x) = P_{2n}^{(\lambda)}(x) + M_n \frac{P_{2n-1}^{(\lambda)}(x)}{x}, \tag{5.12}$$

$$Q_{2n+1}(x) = P_{2n+1}^{(\lambda)}(x) - N_n \frac{P_{2n+1}^{(\lambda)}(x) - \dfrac{P_{2n+1}^{(\lambda)\prime}(0)}{P_{2n}^{(\lambda)}(0)} xP_{2n}^{(\lambda)}(x)}{x^2}, \tag{5.13}$$

where

$$M_n = M \frac{[P_{2n}^{(\lambda)}(0)]^2}{\|P_{2n-1}^{(\lambda)}\|^2[1+MK_{2n-1}(0,0)]},$$

$$N_n = N \frac{P_{2n+1}^{(\lambda)\prime}(0)P_{2n}^{(\lambda)}(0)}{\|P_{2n}^{(\lambda)}\|^2[1+NK_{2n}^{(1,1)}(0,0)]},$$

but,

$$P_{2n}^{(\lambda)}(x) = S_n(x^2); \qquad P_{2n+1}^{(\lambda)}(x) = xS_n^*(x^2)$$

and

$$K_{2n}^{(0,1)}(x, 0) = -\frac{P_{2n}(0)}{\|P_{2n}^{(\lambda)}\|^2}\,\frac{n(2n-1+2\lambda)}{2n+\lambda}\,xS_{n-1}^{**}(x^2)$$

(see [7, Chap. 1, § 8]). Then

$$Q_{2n}(x) = S_n(x^2) + M_nS_{n-1}^*(x^2),$$

$$Q_{2n+1}(x) = x\left[S_n^*(x^2) + N_n\frac{n(2n-1+2\lambda)}{2n+\lambda}S_{n-1}^{**}(x^2)\right].$$

The following proposition can easily be derived from the above comments and from (4.7) and (4.8).

PROPOSITION 5.1. *For the* SMOP (Q_n) *corresponding to the inner product defined by* (5.10), $Q_n(-x) = (-1)^nQ_n(x)$. *If*

$$Q_{2n}(x) = U_n(x^2) \quad \textit{and} \quad Q_{2n+1}(x) = xV_n(x^2),$$

then

(5.14) $$U_n(x) = S_n(x) + M_nS_{n-1}^*(x),$$

(5.15) $$V_n(x) = S_n^*(x) + N_n\frac{n(2n-1+2\lambda)}{2n+\lambda}S_{n-1}^{**}(x),$$

and U_n, V_n *satisfy a three term recurrence relation in the standard sense.*

Remark. In general, for a symmetric SMOP associated to a Sobolev type inner product, we can define two SMOP in the standard sense. They satisfy a decomposition in terms of (5.14) and (5.15).

PROPOSITION 5.2. *The* SMOP (Q_n) *verifies a second-order linear differential equation*

$$A(x;\, n)Q_n''(x) + B(x;\, n)Q_n'(x) + C(x;\, n)Q_n(x) = 0,$$

where A, B, C *are polynomials of degree independent of* n. *More precisely,* $\deg B(x;\, n) \leqq \deg A(x;\, n) - 1$; $\deg C(x;\, n) \leqq \deg A(x;\, n) - 2$; $\deg A(x;\, 2n) = 6$ *and* $\deg A(x;\, 2n+1) = 8$.

Proof. From (5.11) and (5.12)

$$\begin{aligned} xQ_{2n}(x) &= xP_{2n}^{(\lambda)}(x) + M_n\frac{2n-1+\lambda}{(2n-1+2\lambda)n} \\ &\quad\cdot [2nxP_{2n}^{(\lambda)}(x) - (x^2-1)P_{2n}^{(\lambda)\prime}(x)] \\ &= \left(1 + 2M_n\frac{2n-1+\lambda}{2n-1+2\lambda}\right)xP_{2n}^{(\lambda)}(x) \\ &\quad - M_n\frac{2n-1+\lambda}{n(2n-1+2\lambda)}(x^2-1)P_{2n}^{(\lambda)\prime}(x). \end{aligned}$$

On the other hand, from (5.11) and (5.13)

$$\begin{aligned} x^2Q_{2n+1}(x) &= (x^2 - N_n)P_{2n+1}^{(\lambda)}(x) - N_n\left[\frac{-P_{2n+1}^{(\lambda)\prime}(0)}{P_{2n}^{(\lambda)}(0)}xP_{2n}^{(\lambda)}(x)\right] \\ &= (x^2 - N_n)P_{2n+1}^{(\lambda)}(x) - N_n \\ &\quad\cdot [x(x^2-1)P_{2n+1}^{(\lambda)\prime}(x) - (2n+1)x^2P_{2n+1}^{(\lambda)}(x)] \\ &= ([1 + (2n+1)N_n]x^2 - N_n)P_{2n+1}^{(\lambda)}(x) - N_nx(x^2-1)P_{2n+1}^{(\lambda)\prime}(x). \end{aligned}$$

Then

$$Q_n(x) = \tilde{M}_n(x)P_n^{(\lambda)}(x) + \tilde{N}_n(x)P_n^{(\lambda)\prime}(x), \tag{5.16}$$

where

$$\tilde{M}_{2n}(x) = 1 + 2M_n \frac{2n-1+\lambda}{2n-1+2\lambda},$$

$$\tilde{M}_{2n+1}(x) = 1 + (2n+1)N_n - \frac{N_n}{x^2},$$

$$\tilde{N}_{2n}(x) = M_n \frac{2n-1+\lambda}{n(2n-1+2\lambda)} \frac{1-x^2}{x},$$

$$\tilde{N}_{2n+1}(x) = N_n \frac{1-x^2}{x}.$$

Using derivatives in (5.16),

$$\begin{aligned} Q_n'(x) = {} & \tilde{M}_n'(x)P_n^{(\lambda)}(x) + [\tilde{M}_n(x) + \tilde{N}_n'(x)]P_n^{(\lambda)\prime}(x) \\ & + \tilde{N}_n(x)P_n^{(\lambda)\prime\prime}(x). \end{aligned} \tag{5.17}$$

But, from the second-order linear differential equation satisfied by Gegenbauer polynomials (see [24, formula 4.7.5, p. 80]),

$$(x^2-1)P_n^{(\lambda)\prime\prime}(x) + (2\lambda+1)xP_n^{(\lambda)\prime}(x) - n(n+2\lambda)P_n^{(\lambda)}(x) = 0$$

formula (5.17) becomes

$$Q_n'(x) = \hat{M}_n(x)P_n^{(\lambda)}(x) + \hat{N}_n(x)P_n^{(\lambda)\prime}(x), \tag{5.18}$$

where

$$\hat{M}_n(x) = \tilde{M}_n'(x) + \frac{\tilde{N}_n(x)}{x^2-1} n(n+2\lambda),$$

$$\hat{N}_n(x) = \tilde{M}_n(x) + \tilde{N}_n'(x) - (2\lambda+1)x\frac{\tilde{N}_n(x)}{x^2-1}.$$

From (5.16) and (5.18)

$$P_n^{(\lambda)}(x) = \frac{\begin{vmatrix} Q_n(x) & \tilde{N}_n(x) \\ Q_n'(x) & \hat{N}_n(x) \end{vmatrix}}{\Delta_n}, \tag{5.19}$$

$$P_n^{(\lambda)\prime}(x) = \frac{\begin{vmatrix} \tilde{M}_n(x) & Q_n(x) \\ \hat{M}_n(x) & Q_n'(x) \end{vmatrix}}{\Delta_n}, \tag{5.20}$$

where $\Delta_n = \tilde{M}_n(x)\hat{N}_n(x) - \hat{M}_n(x)\tilde{N}_n(x)$ is a rational function.

From derivation in (5.19) and taking into account (5.20), the result follows. □

Remark. The above result should be compared with Proposition 6.1 in [14].

We consider, as a second example, an inner product of Sobolev type when μ is a discrete positive measure. More precisely, μ is a step function whose jumps are

$$d\mu(x) = \frac{e^{-a}a^x}{x!} \quad \text{at } x = 0, 1, 2, \cdots \quad \text{and} \quad a \in \boldsymbol{R}^+.$$

This corresponds to Poisson distribution in Probability Theory. The corresponding sequence $(C_n^{(a)})$ of monic orthogonal polynomials is called Charlier polynomials in the literature (see [7, p. 170]).

They can be expressed in terms of Laguerre polynomials as $C_n^{(a)}(x) = n!L_n^{(x-n)}(a)$ and satisfy a three term recurrence relation

$$C_{n+1}^{(a)}(x) = (x-n-a)C_n^{(a)}(x) - anC_{n-1}^{(a)}(x) \qquad n \geqq 0,$$
$$C_{-1}^{(a)}(x) = 0 \qquad C_0^{(a)}(x) = 1.$$

Moreover, Charlier polynomials can be characterized as the only SMOP belonging to Δ-Appell class, i.e.,

$$\Delta C_n^{(a)}(x) = nC_{n-1}^{(a)}(x) \qquad n \geqq 1,$$

where

$$\Delta p(x) = p(x+1) - p(x).$$

In this case, (2.12) becomes

$$x^2 Q_n(x) = q_2(x;\, n)C_n^{(a)}(x) + q_1(x;\, n)C_{n-1}^{(a)}(x), \tag{5.21}$$

where

$$q_2(x;\, n) = x^2 - a_n x - b_n,$$
$$q_1(x;\, n) = c_n x + d_n$$

and

$$a_n = \frac{MQ_n(0)C_{n-1}^{(a)}(0) + NQ_n'(0)C_{n-1}^{(a)\prime}(0)}{\|C_{n-1}^{(a)}\|^2},$$
$$b_n = \frac{NQ_n'(0)C_{n-1}^{(a)}(0)}{\|C_{n-1}^{(a)}\|^2},$$
$$c_n = \frac{MQ_n(0)C_n^{(a)}(0) + NQ_n'(0)C_n^{(a)\prime}(0)}{\|C_{n-1}^{(a)}\|^2},$$
$$d_n = \frac{NQ_n'(0)C_n^{(a)}(0)}{\|C_{n-1}^{(a)}\|^2} = -ab_n.$$

If in (5.21) we apply the Δ-operator and the recurrence relation for $C_n^{(a)}$, we get

$$\begin{aligned}
&(x+1)^2\Delta Q_n(x) + (2x+1)Q_n(x)\\
&\quad = q_2(x+1;\, n)nC_{n-1}^{(a)}(x) + \Delta q_2(x;\, n)C_n^{(a)}(x)\\
&\qquad + q_1(x+1;\, n)(n-1)C_{n-2}^{(a)}(x) + \Delta q_1(x;\, n)C_{n-1}^{(a)}(x)\\
&\quad = \left[\Delta q_2(x;\, n) - \frac{1}{a}q_1(x+1;\, n)\right]C_n^{(a)}(x)\\
&\qquad + \left[nq_2(x+1;\, n) + \Delta q_1(x;\, n) + \frac{1}{a}(x+1-n-a)q_1(x+1;\, n)\right]C_{n-1}^{(a)}(x).
\end{aligned}$$

Thus,

$$(x+1)^2\Delta Q_n(x) + (2x+1)Q_n(x) = A(x;\, n)C_n^{(a)}(x) + B(x;\, n)C_{n-1}^{(a)}(x) \tag{5.22}$$

with

$$A(x;\, n)=\left(2-\frac{c_n}{a}\right)x+1-a_n-\frac{1}{a}(c_n+d_n),$$

$$B(x;\, n)=\left(n+\frac{c_n}{a}\right)(x+1)^2$$
$$+\left[\frac{1}{a}\{d_n-(n+a)c_n\}-na_n\right](x+1)+c_n-d_n.$$

Then, from (5.21) and (5.22), Cramer's rule gives

$$C_n^{(a)}(x)=\frac{E_n(x)}{S_n(x)}Q_n(x)+\frac{F_n(x)}{S_n(x)}\Delta Q_n(x),$$

$$C_{n-1}^{(a)}(x)=\frac{G_n(x)}{S_n(x)}Q_n(x)+\frac{H_n(x)}{S_n(x)}\Delta Q_n(x),$$

where

$$E_n(x)=x^2B(x;\, n)-(2x+1)q_1(x;\, n),$$
$$F_n(x)=-(x+1)^2q_1(x;\, n),$$
$$S_n(x)=q_2(x;\, n)B(x;\, n)-q_1(x;\, n)A(x;\, n),$$
$$G_n(x)=(2x+1)q_2(x;\, n)-x^2A(x;\, n),$$
$$H_n(x)=(x+1)^2q_2(x;\, n).$$

Finally, using $\Delta C_n^{(a)}(x)=nC_{n-1}^{(a)}(x)$

$$\frac{E_n(x+1)}{S_n(x+1)}\Delta Q_n(x)+\left(\Delta\frac{E_n(x)}{S_n(x)}\right)Q_n(x)$$
$$+\frac{F_n(x+1)}{S_n(x+1)}\Delta^2Q_n(x)+\left(\Delta\frac{F_n(x)}{S_n(x)}\right)\Delta Q_n(x)$$
$$=n\left(\frac{G_n(x)}{S_n(x)}Q_n(x)+\frac{H_n(x)}{S_n(x)}\Delta Q_n(x)\right).$$

Therefore,

$$F_n(x+1)S_n(x)\Delta^2Q_n(x)$$
$$+([E_n(x+1)+F_n(x+1)]S_n(x)-[nH_n(x)+F_n(x)]S_n(x+1))\Delta Q_n(x)$$
$$+(E_n(x+1)S_n(x)-[E_n(x)+nG_n(x)]S_n(x+1))Q_n(x)=0.$$

In conclusion, we present the following.

PROPOSITION 5.3. *The* SMOP (Q_n) *satisfies a second-order linear difference equation*

$$U_n(x;\, n)\Delta^2Q_n(x)+V_n(x;\, n)\Delta Q_n(x)+W_n(x;\, n)Q_n=0,$$

where U, V and W are polynomials with degree independent of n. More precisely, deg $U=7$, deg $V\leqq 8$ *and* deg $W\leqq 7$.

Acknowledgments. The authors thank Professor Walter Van Assche (Katholieke Universiteit Leuven) for his useful suggestions and comments. They also wish to thank the referee who provided them with reference [10].

REFERENCES

[1] P. Althammer, *Eine Erweiterung des Orthogonalitätsbegriffes bei Polynomen und deren Anwendung auf die beste Approximation*, J. Reine Angew. Math., 211 (1962), pp. 192-204.

[2] H. Bavinck and H. G. Meijer, Orthogonal polynomials with respect to a symmetric inner product involving derivatives, Appl. Anal., 33 (1989), pp. 103-117.

[3] ———, *On orthogonal polynomials with respect to an inner product involving derivatives: zeros and recurrence relations*, Indag. Math. (N.S.), 1 (1990), pp. 7-14.

[4] J. Brenner, *Über eine Erweiterung des Orthogonalitätsbegriffes bei Polynomen*, in Proc. Conf. Constructive Theory of Functions, Budapest, 1969, G. Alexits and S. B. Stechkin, eds., Akadémiai Kiadó, Budapest, 1972, pp. 77-83.

[5] M. D. Buhmann and A. Iserles, *On orthogonal polynomials transformed by the QR algorithm*, Num. Analysis Reports 7, DAMTP, Cambridge University, 1991.

[6] E. A. Cohen, *Zero distribution and behavior of orthogonal polynomials in the Sobolev space $W^{1,2}[-1,1]$*, SIAM J. Math. Anal., 6 (1975), pp. 105-116.

[7] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[8] W. Gröbner, *Orthogonale Polynomsysteme, die gleichzeitig mif $f(x)$ auch deren Ableitung $f'(x)$ approximieren*, in I.S.N.M., Vol. 7, Birkhäuser-Verlag, Basel, Switzerland, 1967, pp. 24-32.

[9] R. Guadalupe, *Operador deslizamiento en matrices hermitianas definidas positivas. Aplicaciones a ceros de polinomios ortogonales y distribuciones en* **R**, Ph.D. thesis, Universidad de Santander, Santander, Spain, 1984.

[10] J. Kautsky and G. H. Golub, *On the calculation of Jacobi matrix*, Linear Algebra Appl., 52/53 (1983), pp. 439-455.

[11] R. Koekoek, *Generalizations of Laguerre polynomials*, J. Math. Anal. Appl., 153 (1990), pp. 576-590.

[12] ———, *Generalizations of the classical Laguerre polynomials and some q-analogues*, Ph.D. thesis, Technical University of Delft, Delft, the Netherlands, 1990.

[13] R. Koekoek and H. G. Meijer, *A generalization of Laguerre polynomials*, SIAM J. Math. Anal., submitted.

[14] T. H. Koornwinder, *Orthogonal polynomials with weight function $(1-x)^\alpha(1+x)^\beta+M\delta(x+1)+N\delta(x-1)$*, Canad. Math. Bull., 27 (1984), pp. 205-214.

[15] P. Lesky, *Zur Konstruktion von Orthogonalpolynomen*, in Proc. Conf. Constructive Theory of Functions, Budapest, 1969, G. Alexits and S. B. Stechkin, eds., Akadémiai Kiadó, Budapest, 1972, pp. 289-298.

[16] D. C. Lewis, *Polynomial least square approximations*, Amer. J. Math., 69 (1947), pp. 273-278.

[17] F. Marcellán and A. Ronveaux, *On a class of polynomials orthogonal with respect to a Sobolev inner product*, Indag. Math. (N.S.), 1 (1990), pp. 451-464.

[18] F. Marcellán and W. Van Assche, *Relative asymptotic for orthogonal polynomials with a Sobolev inner product*, J. of Approx. Theory, to appear.

[19] P. Maroni, *Prolégomènes à l'étude des polynômes orthogonaux semiclassiques*, Ann. Mat. Pura Appl., 149 (1987), pp. 165-184.

[20] H. G. Meijer, *Laguerre polynomials generalized to a certain discrete Sobolev inner product space*, 1990, preprint.

[21] A. Ronveaux and F. Marcellán, *Differential equation for classical-type orthogonal polynomials*, Canad. Math. Bull., 32 (1989), pp. 404-411.

[22] F. W. Schäfke, *Zu den Orthogonalpolynomen von Althammer*, J. Reine Angew. Math., 252 (1972), pp. 195-199.

[23] F. W. Schäfke and G. Wolf, *Einfache verallgemeinerte klassische Orthogonalpolynome*, J. Reine Angew. Math., 262/263 (1973), pp. 339-355.

[24] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., 23, American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 758-765, May 1992

011

MULTIDIMENSIONAL q-BETA INTEGRALS*

RONALD J. EVANS†

Abstract. A multidimensional extension of a q-beta integral of Andrews and Askey is evaluated. As an application, a short new proof of an important q-Selberg integral formula is given.

Key words. q-integral, Selberg integral, beta integrals

AMS(MOS) subject classification. 33A15

1. Introduction. This paper has been motivated by Anderson's wonderfully innovative proof [2] of Selberg's multidimensional beta integral formula [17]. In § 2 (see Theorem 1), we present a new n-dimensional q-beta integral formula which reduces to that of Andrews and Askey [4, eqn. (2.2)] when $n = 1$ and that of Anderson [2, "claim"] when $q = 1$. Our proof is self-contained and in particular makes no appeal to the results of the aforementioned papers. In § 3, we apply Theorem 1 to give a surprisingly short, self-contained proof of the q-Selberg integral formula (1.8). Finally, we indicate in § 4 the modifications that can be made in § 3 to give a short proof of Kadell's extension of the q-Selberg integral formula containing the extra parameter m of Aomoto [5]; see Theorem 2. It is hoped that this method will lead to a short proof of a q-extension of the Selberg-Jack integral formula [15].

For some of the many applications and extensions of Selberg's integral, see the papers of Askey [6]-[8] and Kadell [14]-[16]. For character sum analogues of Selberg's integral, see the papers of Anderson [1], Evans [10] and van Wamelen [18].

Let

$$0<q<1, \tag{1.1}$$

and define, for complex x, α,

$$(\alpha)_\infty := \prod_{r=0}^{\infty} (1-\alpha q^r), \qquad (\alpha)_x := (\alpha)_\infty/(\alpha q^x)_\infty. \tag{1.2}$$

Define the q-gamma function

$$\Gamma_q(x) := (q)_{x-1}(1-q)^{1-x}, \qquad x \in \mathbb{C}. \tag{1.3}$$

As $q \to 1$, $\Gamma_q(x) \to \Gamma(x)$ [11, eqn. (1.10.3)]. For $\alpha, \beta \in \mathbb{C}$ and a (say) continuous function $f: \mathbb{C} \to \mathbb{C}$, define the q-integral

$$\int_\alpha^\beta f(x)\, d_qx := \int_0^\beta f(x)\, d_qx - \int_0^\alpha f(x)\, d_qx, \tag{1.4}$$

where

$$\int_0^\beta f(x)\, d_qx := (1-q) \sum_{m=0}^{\infty} f(\beta q^m)\beta q^m. \tag{1.5}$$

As $q \to 1$, $\int_\alpha^\beta f(x)\, d_qx \to \int_\alpha^\beta f(x)\, dx$ [11, p. 19]. For example, for $m > 0$,

$$\int_\alpha^\beta x^{m-1}\, d_qx = \frac{(\beta^m - \alpha^m)(1-q)}{(1-q^m)} \to \frac{\beta^m - \alpha^m}{m} \tag{1.6}$$

* Received by the editors December 17, 1990; accepted for publication (in revised form) June 21, 1991.

† Department of Mathematics 0112, University of California, San Diego, La Jolla, California 92093-0112.

as $q \to 1$. The following q-integral extension of Euler's beta function integral is essentially a version of the q-binomial theorem [11, pp. 18-19]:

$$\int_0^1 t^{a-1}(tq)_{b-1}\, d_qt = \Gamma_q(a)\Gamma_q(b)/\Gamma_q(a+b), \qquad \operatorname{Re}(a), \operatorname{Re}(b) > 0. \tag{1.7}$$

This is the case $n=1$ of the following n-dimensional q-Selberg integral formula [13, eqn. (4.18)]:

$$\begin{aligned} S_n(a, b, c) &:= \frac{1}{n!}\int_0^1 \cdots \int_0^1 \prod_{i=1}^n t_i^{a-1}(t_iq)_{b-1} \prod_{1 \leqq i<j \leqq n} \prod_{k=1-c}^{c-1} (t_i - q^k t_j)\, d_qt_1 \cdots d_qt_n \\ &= q^{ac\binom{n}{2}+2c^2\binom{n}{3}} \prod_{j=0}^{n-1} \frac{\Gamma_q(a+jc)\Gamma_q(b+jc)\Gamma_q(c+jc)}{\Gamma_q(a+b+(n-1+j)c)\Gamma_q(c)}, \end{aligned} \tag{1.8}$$

where n, c are positive integers and $\operatorname{Re}(a)$, $\operatorname{Re}(b) > 0$. This reduces to Selberg's integral formula [17] when $q \to 1$. Note that the integrand in (1.8) is symmetric in the variables t_i. It is not difficult to show that the nonsymmetric version of (1.8) originally conjectured by Askey [6, Conj. 1] is equivalent to (1.8); see Kadell [13, p. 953]. Proofs of (1.8) have been given independently by Habsieger [12] and Kadell [14].

We observe here for later use that the value of the integral in (1.8) is unchanged if the upper limits of integration are replaced by q^{-u}, when u and b are integers such that $0 \leqq u \leqq b-1$. This is because $(tq)_{b-1}$ vanishes for $t = q^{-1}, q^{-2}, \cdots, q^{-u}$. It follows that the integral in (1.8) changes only by a factor of a power of q when the variables t_i are replaced by t_iq^{-u}.

2. Extension of the Andrews–Askey q-integral.

THEOREM 1. *Let u_i, s_i be integers such that*

$$0 \leqq u_i \leqq s_i - 1, \qquad i = 0, 1, \cdots, n, \tag{2.1}$$

and let z_i, w_i be complex variables with

$$w_i = z_iq^{-u_i}, \qquad i = 0, 1, \cdots, n. \tag{2.2}$$

Then

$$\begin{aligned} L &:= \int_{t_n=w_{n-1}}^{w_n} \cdots \int_{t_2=w_1}^{w_2} \int_{t_1=w_0}^{w_1} \prod_{i=0}^n \prod_{j=1}^n \prod_{k=1}^{s_i-1} (z_i - q^k t_j) \\ &\quad \cdot \prod_{1 \leqq i<j \leqq n} (t_j - t_i)\, d_qt_1\, d_qt_2 \cdots d_qt_n \\ &= (-1)^\sigma q^\tau \frac{\Gamma_q(s_0)\Gamma_q(s_1)\cdots\Gamma_q(s_n)}{\Gamma_q(s_0+s_1+\cdots+s_n)} \prod_{0 \leqq i<j \leqq n} \prod_{k=1-s_j}^{s_i-1} (z_i - q^k z_j), \end{aligned} \tag{2.3}$$

where

$$\sigma = \sum_{i=1}^n is_i, \qquad \tau = \sum_{i=1}^n i\binom{s_i}{2}. \tag{2.4}$$

Remark 1. Suppose that all z_i are nonzero and all u_i are zero. Then the integral formula in Theorem 1 can be written in the form

$$\begin{aligned} &\int_{t_n=z_{n-1}}^{z_n} \cdots \int_{t_2=z_1}^{z_2} \int_{t_1=z_0}^{z_1} \prod_{i=0}^n \prod_{j=1}^n \left(\frac{qt_j}{z_i}\right)_{s_i-1} \\ &\quad \cdot \prod_{1 \leqq i<j \leqq n} (t_j - t_i)\, d_qt_1\, d_qt_2 \cdots d_qt_n \\ &= \frac{\Gamma_q(s_0)\Gamma_q(s_1)\cdots\Gamma_q(s_n)}{\Gamma_q(s_0+s_1+\cdots+s_n)} \prod_{0 \leqq i<j \leqq n} z_j\left(\frac{z_i}{z_j}\right)_{s_j}\left(\frac{qz_j}{z_i}\right)_{s_i-1}. \end{aligned} \tag{2.5}$$

Since (2.5) is valid for all positive integers s_i by Theorem 1, it follows by analytic continuation (cf. [3, p. 115]) that it holds for all complex s_i with

$$\operatorname{Re}(s_i) > \max_{0 \leqq j \leqq n} \frac{\log |z_j/z_i|}{|\log q|}, \qquad i = 0, 1, 2, \cdots, n.$$

If $n = 1$, (2.5) reduces to the Andrews-Askey q-integral [4, (2.2)].

Remark 2. From (2.5) and [9, Thm. 2.2], it may be deduced that the constant term of the Laurent polynomial

$$P(z_1, \cdots, z_n) := \int_0^1 \cdots \int_0^1 \prod_{1 \leqq i, j \leqq n} (qt_j z_j/z_i)_{s_i - 1} \cdot \prod_{1 \leqq i < j \leqq n} (t_j - t_i z_i/z_j) \, d_q t_1 \cdots d_q t_n$$

equals

$$\prod_{i=1}^{n} (1-q)/(1-q^{s_i + s_{i+1} + \cdots + s_n}). \tag{2.6}$$

It would be interesting to find a proof independent of [9].

Proof of Theorem 1. Assume that each z_i is an integral power of q and that the sequence $w_0, w_1, w_2, \cdots, w_n$ is monotone. It suffices to prove (2.3) under these assumptions, since both sides of (2.3) are polynomials in $z_0, \cdots, z_n$.

Consider any one of the rightmost factors in (2.3), say

$$z_\alpha - q^\gamma z_\beta, \tag{2.7}$$

with

$$0 \leqq \alpha < \beta \leqq n, \qquad 1 - s_\beta \leqq \gamma \leqq s_\alpha - 1. \tag{2.8}$$

We will show that $z_\alpha - q^\gamma z_\beta$ is also a factor of L by showing that L vanishes under the assumption

$$z_\alpha = q^\gamma z_\beta. \tag{2.9}$$

The q-integral L is a series by definition, and it suffices to show that each summand in this series vanishes. This will be accomplished if we can show

$$\prod_{k=1}^{s_\alpha - 1} (z_\alpha - q^k t) \prod_{m=1}^{s_\beta - 1} (z_\beta - q^m t) = 0 \quad \text{for all } t \in S, \tag{2.10}$$

where S is the set of integral powers of q between w_α and w_β including $\max(w_\alpha, w_\beta)$ but not $\min(w_\alpha, w_\beta)$. Define

$$A = \{z_\alpha q^{-k} : 1 \leqq k \leqq s_\alpha - 1\}, \qquad B = \{z_\beta q^{-m} : 1 \leqq m \leqq s_\beta - 1\}. \tag{2.11}$$

Since $z_\alpha = q^\gamma z_\beta$ by (2.9), there is no integral power of q lying strictly between the sets A and B on the real axis. It is thus seen that $A \cup B \supset S$, and (2.10) follows. We have now proved that L is divisible by each of the linear factors in (2.7), and hence by the polynomial

$$\prod_{0 \leqq i < j \leqq n} \prod_{k = 1 - s_j}^{s_i - 1} (z_i - q^k z_j). \tag{2.12}$$

By definition of L, if we view L as a polynomial in z_0 with leading term $C_n z_0^\nu$ (with C_n independent of z_0), then

$$\nu = n(s_0 - 1) + (s_1 + \cdots + s_n). \tag{2.13}$$

Viewing (2.12) as a polynomial in z_0, we see that it also has degree ν. Thus it remains to prove that

$$C_n=(-1)^\sigma q^\tau \frac{\Gamma_q(s_0)\cdots\Gamma_q(s_n)}{\Gamma_q(s_0+\cdots+s_n)} \prod_{1\leqq i<j\leqq n} \prod_{k=1-s_j}^{s_i-1} (z_i-q^k z_j). \tag{2.14}$$

First consider the case $n=1$. Then C_1 is the coefficient of $z_0^{s_0+s_1-1}$ in

$$\int_{t=w_0}^{w_1} \prod_{k=1}^{s_0-1} (z_0-q^k t) \prod_{m=1}^{s_1-1} (z_1-q^m t)\, d_q t, \tag{2.15}$$

so C_1 is the coefficient of $z_0^{s_0+s_1-1}$ in

$$-\prod_{m=1}^{s_1-1} (-q^m) \int_{w_1}^{w_0} t^{s_1-1} z_0^{s_0-1} (qt/z_0)_{s_0-1}\, d_q t. \tag{2.16}$$

Replace t by $z_0 t$ to see that C_1 is the constant term in the expansion in z_0 of

$$(-1)^{s_1} q^{\binom{s_1}{2}} \int_{w_1/z_0}^{q^{-u_0}} t^{s_1-1} (qt)_{s_0-1}\, d_q t. \tag{2.17}$$

The constant term in (2.17) is unchanged if the lower limit of q-integration is replaced by 0. It is further unchanged if the upper limit of q-integration is replaced by 1, since

$$(qt)_{s_0-1}=0 \quad \text{for} \quad t=q^{-i} \qquad (i=1,2,\cdots,s_0-1). \tag{2.18}$$

It now follows from (1.7) that (2.14) holds for $n=1$, so the proof of Theorem 1 is complete in the case $n=1$.

Suppose now that $n>1$ and that Theorem 1 holds with $(n-1)$ in place of n. Directly from (2.3), we see that C_n is the coefficient of $z_0^{(s_0+\cdots+s_n)-1}$ in

$$\begin{aligned}&\int_{t_n=w_{n-1}}^{w_n} \cdots \int_{t_2=w_1}^{w_2} \prod_{i=1}^{n} \prod_{j=2}^{n} \prod_{k=1}^{s_i-1} (z_i-q^k t_j) \cdot \prod_{2\leqq i<j\leqq n} (t_j-t_i)\\ &\cdot (-1)^{s_1+\cdots+s_n} q^{\binom{s_1}{2}+\cdots+\binom{s_n}{2}} \int_{t=w_1}^{w_0} t^{(s_1+\cdots+s_n)-1}\\ &\cdot \prod_{k=1}^{s_0-1} (z_0-q^k t)\, d_q t\, d_q t_2 \cdots d_q t_n.\end{aligned} \tag{2.19}$$

The inner integral on t in (2.19) may be replaced by

$$z_0^{(s_0+\cdots+s_n)-1} \int_{w_1/z_0}^{q^{-u_0}} t^{(s_1+\cdots+s_n)-1} (qt)_{s_0-1}\, d_q t, \tag{2.20}$$

and just as with (2.17), the desired coefficient is unchanged if we further replace the lower and upper limits of q-integration in (2.20) by zero and 1, respectively. Thus by (1.7), C_n is the constant term of the polynomial in z_0 obtained from (2.19) by replacing the inner integral on t by

$$\frac{\Gamma_q(s_0)\Gamma_q(s_1+\cdots+s_n)}{\Gamma_q(s_0+s_1+\cdots+s_n)}. \tag{2.21}$$

By induction on n, the proof of Theorem 1 is complete.

3. Proof of the q-Selberg integral formula. In this section we apply Theorem 1 to give a short proof of the q-Selberg integral formula (1.8). The result is true for $n=1$ by (1.7), so let $n>1$. We may assume that a and b are positive integers, as the result can be extended by analytic continuation to hold whenever Re (a), Re $(b)>0$.

Given polynomials

$$E(t)=\prod_{i=1}^{n}(t-e_i), \qquad H(t)=\prod_{i=1}^{n-1}(t-h_i) \tag{3.1}$$

with

$$0\leqq e_1\leqq h_1\leqq e_2\leqq h_2\leqq\cdots\leqq h_{n-1}\leqq e_n\leqq 1, \tag{3.2}$$

use for brevity the symbolic notation

$$\begin{aligned}\int_{E\in D_n}\{\ \}\,d_qE := \int_{e_n=0}^{1}\cdots\int_{e_2=0}^{e_3}\int_{e_1=0}^{e_2}\{\ \}\\ \cdot\prod_{1\leqq i<j\leqq n}(e_i-e_j)\,d_qe_1\,d_qe_2\cdots d_qe_n\end{aligned} \tag{3.3}$$

and

$$\begin{aligned}\int_{H\in D_{n-1}(E)}\{\ \}\,d_qH := \int_{h_{n-1}=e_{n-1}}^{e_n}\cdots\int_{h_2=e_2}^{e_3}\int_{h_1=e_1}^{e_2}\{\ \}\\ \cdot\prod_{1\leqq i<j\leqq n-1}(h_i-h_j)\,d_qh_1\,d_qh_2\cdots d_qh_{n-1}.\end{aligned} \tag{3.4}$$

Note that

$$\int_{E\in D_n}\int_{H\in D_{n-1}(E)}=\int_{H\in D_{n-1}}\int_{E\in D_n(V)}, \tag{3.5}$$

where

$$V(t)=\prod_{i=0}^{n}(t-v_i)\quad\text{with } v_0=0,\quad v_n=1,\quad v_i=qh_i\quad(1\leqq i\leqq n-1). \tag{3.6}$$

Define

$$\begin{aligned}I_n(a,b,c) := \int_{E\in D_n}\int_{H\in D_{n-1}(E)}\prod_{i=1}^{n}e_i^{a-1}(qe_i)_{b-1}\\ \cdot\prod_{i=1}^{n}\prod_{j=1}^{n-1}\prod_{k=1}^{c-1}(q^{c-1}e_i-q^kh_j)\,d_qH\,d_qE.\end{aligned} \tag{3.7}$$

If we replace n by $n-1$ in Theorem 1 and then further take $t_i=h_i$, $s_i=c$, $u_i=c-1$, $w_i=e_{i+1}$, $z_i=q^{c-1}e_{i+1}$, then Theorem 1 yields

$$\begin{aligned}\int_{H\in D_{n-1}(E)}\prod_{i=1}^{n}\prod_{j=1}^{n-1}\prod_{k=1}^{c-1}(q^{c-1}e_i-q^kh_j)\,d_qH\\ =(-1)^{\binom{n-1}{2}+c\binom{n}{2}}q^{\binom{n}{2}\binom{c}{2}}\frac{\Gamma_q(c)^n}{\Gamma_q(cn)}\\ \cdot\prod_{1\leqq i<j\leqq n}\prod_{k=1-c}^{c-1}(q^{c-1}e_i-q^{k+c-1}e_j).\end{aligned} \tag{3.8}$$

Thus, by definition of $S_n(a,b,c)$ and $I_n(a,b,c)$,

$$I_n(a,b,c)=(-1)^{\binom{n-1}{2}+c\binom{n}{2}}q^{\binom{n}{2}\binom{c}{2}+\binom{n}{2}\binom{2c-1}{2}}\frac{\Gamma_q(c)^n}{\Gamma_q(cn)}S_n(a,b,c). \tag{3.9}$$

By (3.5) and (3.6), interchange of integration in (3.7) yields

$$\begin{aligned} I_n(a,b,c) = \int_{H\in D_{n-1}} \int_{E\in D_n(V)} & (-1)^{n(a-1)} q^{-n\binom{a}{2}} \prod_{j=1}^{n}\prod_{k=1}^{a-1} (0-q^k e_j) \\ & \cdot \prod_{j=1}^{n}\prod_{k=1}^{b-1} (1-q^k e_j) \\ & \cdot q^{2\binom{n}{2}\binom{c-1}{2}} \prod_{i=1}^{n-1}\prod_{j=1}^{n}\prod_{k=1}^{c-1} (v_i - q^k e_j)\, d_qE\, d_qH. \end{aligned} \tag{3.10}$$

Apply Theorem 1 with $t_i = e_i$, $s_0 = a$, $s_n = b$, $s_i = c$ $(1 \leqq i \leqq n-1)$, $u_i = 0$, $w_i = v_i$, and $z_i = v_i$ to see that the inner integral on E equals

$$\begin{aligned} & (-1)^{\binom{n-1}{2}+c\binom{n}{2}} q^{\binom{n}{2}\binom{c}{2}+2\binom{n}{2}\binom{c-1}{2}} \\ & \cdot \frac{\Gamma_q(a)\Gamma_q(b)\Gamma_q(c)^{n-1}}{\Gamma_q(a+b+(n-1)c)} \prod_{j=1}^{n-1} v_j^{a+c-1} \prod_{j=1}^{n-1}\prod_{k=1-c}^{b-1} (1-q^k v_j) \\ & \cdot \prod_{1\leqq i<j\leqq n-1}\prod_{k=1-c}^{c-1} (v_i - q^k v_j). \end{aligned} \tag{3.11}$$

Before integrating (3.11) on H, make the change of variables $h_i \to q^{c-1}h_i$ (so $v_i \to q^c v_i$). As a result,

$$\begin{aligned} I_n(a,b,c) = & (-1)^{\binom{n-1}{2}+c\binom{n}{2}} \\ & \cdot q^{\binom{n}{2}\binom{c}{2}+2\binom{n}{2}\binom{c-1}{2}+(c-1)\binom{n}{2}+\binom{n-1}{2}\binom{2c}{2}+c(a+c-1)(n-1)} \\ & \cdot \frac{\Gamma_q(a)\Gamma_q(b)\Gamma_q(c)^{n-1}}{\Gamma_q(a+b+(n-1)c)} S_{n-1}(a+c, b+c, c). \end{aligned} \tag{3.12}$$

Comparison of (3.9) and (3.12) yields

$$\begin{aligned} S_n(a,b,c) = q^{ac(n-1)+c^2\binom{n-1}{2}} & \frac{\Gamma_q(a)\Gamma_q(b)\Gamma_q(cn)}{\Gamma_q(a+b+(n-1)c)\Gamma_q(c)} \\ & \cdot S_{n-1}(a+c, b+c, c) \end{aligned} \tag{3.13}$$

and the result follows by induction on n. □

4. Extension of the q-Selberg integral. Let $S_{n,m}(a,b,c)$ denote the extension of the q-Selberg integral $S_n(a,b,c)$ obtained by inserting the factor $t_1t_2\cdots t_m$ in the integrand in (1.8), where $0 \leqq m \leqq n$. In Theorem 2 below, we evaluate $S_{n,m}(a,b,c)$. It is not difficult to show that Theorem 2 is equivalent to the case $l=0$ of [14, Thm. 2]; see [14, eqns. (4.17), (4.19)].

THEOREM 2. *For positive integers n, c and* $\operatorname{Re}(a)$, $\operatorname{Re}(b) > 0$,

$$S_{n,m}(a,b,c) = \frac{S_n(a,b,c)T_{n,m}(a,b,c)}{\binom{n}{m}}, \tag{4.1}$$

where

$$T_{n,m}(a,b,c) := q^{c\binom{m}{2}} \prod_{i=n-m}^{n-1} \frac{(1-q^{a+ci})(1-q^{c+ci})}{(1-q^{a+b+c(n-1+i)})(1-q^{cn-ci})}. \tag{4.2}$$

Proof. We proceed as in the proof in § 3, with the following modifications. Let u be an indeterminate and let $S_n(a, b, c, u)$ be the extension of the q-Selberg integral $S_n(a, b, c)$ obtained by inserting the factor $\prod_{i=1}^n (u-t_i)$ in the integrand of (1.8). We must show that

$$\frac{S_n(a, b, c, u)}{S_n(a, b, c)} = \sum_{m=0}^{n} (-1)^m T_{n,m}(a, b, c) u^{n-m}. \tag{4.3}$$

Let $I_n(a, b, c, u)$ be the extension of $I_n(a, b, c)$ obtained by inserting the factor $q^{c(n-1)}H(u/q)$ in the integrand in (3.7). By Lagrange interpolation,

$$q^{c(n-1)}H\left(\frac{u}{q}\right) = \sum_{r=1}^{n} q^{c(n-1)}H\left(\frac{e_r}{q}\right) \prod_{i\neq r} \frac{u-e_i}{e_r-e_i}, \tag{4.4}$$

for distinct e_i. Thus, from (3.7),

$$\begin{aligned} I_n(a, b, c, u) = &\int_{E\in D_n} \sum_{r=1}^{n} \prod_{i\neq r} \frac{u-e_i}{e_r-e_i} \prod_{i=1}^{n} e_i^{a-1}(qe_i)_{b-1} \\ &\cdot \int_{H\in D_{n-1}(E)} \prod_{i=1}^{n} \prod_{j=1}^{n-1} \prod_{k=1}^{c-1+\delta(i,r)} (q^{c-1}e_i - q^k h_j)\, d_qH\, d_qE, \end{aligned} \tag{4.5}$$

where $\delta(i, r) = 1$ if $i = r$ and $\delta(i, r) = 0$ if $i \neq r$. If for each fixed r we replace n by $n-1$ in Theorem 1, and then further take $t_i = h_i$, $s_i = c + \delta(i, r)$, $u_i = c-1$, $w_i = e_{i+1}$, and $z_i = q^{c-1}e_{i+1}$, then Theorem 1 shows that the inner integral on H in (4.5) equals

$$\text{RHS (3.8) } q^{(n-1)(2c-1)} \frac{(1-q^c)}{(1-q^{cn})} \prod_{i\neq r} (q^{-c}e_r - e_i), \tag{4.6}$$

where RHS (3.8) denotes the right-hand side of (3.8). Thus

$$\begin{aligned} I_n(a, b, c, u) = &\, q^{(n-1)(2c-1)} \frac{(1-q^c)}{(1-q^{cn})} \int_{E\in D_n} \text{RHS (3.8)} \\ &\cdot \prod_{i=1}^{n} e_i^{a-1}(qe_i)_{b-1} \sum_{r=1}^{n} \prod_{i\neq r} \frac{u-e_i}{e_r-e_i} (q^{-c}e_r - e_i)\, d_qE. \end{aligned} \tag{4.7}$$

Given a polynomial $F(u)$, let $F^*(u)$ denote its q^{-c}-derivative [11, p. 22], namely

$$F^*(u) = \frac{F(u) - F(q^{-c}u)}{u - q^{-c}u}. \tag{4.8}$$

Since

$$E^*(e_r) = \prod_{i\neq r} (q^{-c}e_r - e_i), \tag{4.9}$$

the inner sum on r in (4.7) equals $E^*(u)$. Thus

$$I_n(a, b, c, u) = \text{RHS (3.9) } q^{(n-1)(2c-1)} \frac{(1-q^c)}{(1-q^{cn})} \frac{S_n^*(a, b, c, u)}{S_n(a, b, c)}. \tag{4.10}$$

After interchanging the order of integration, we obtain

$$I_n(a, b, c, u) = \text{RHS (3.12) } q^{(n-1)(2c-1)} \frac{S_{n-1}(a+c, b+c, c, uq^{-c})}{S_{n-1}(a+c, b+c, c)}. \tag{4.11}$$

Comparing (4.10) and (4.11), we arrive at the "differential equation"

$$\frac{S_n^*(a, b, c, u)}{S_n(a, b, c)} = \frac{1-q^{cn}}{1-q^c} \frac{S_{n-1}(a+c, b+c, c, uq^{-c})}{S_{n-1}(a+c, b+c, c)}. \tag{4.12}$$

By induction on n, (4.3) furnishes a solution to (4.12). Moreover, (4.3) is valid for $u=0$, by (1.8) with $a+1$ in place of a. Hence (4.3) is proved.

Acknowledgments. The author is grateful to Professors G. Anderson, G. Andrews, R. Askey, D. Bressoud, and K. Kadell for helpful correspondence.

REFERENCES

[1] G. ANDERSON, *The evaluation of Selberg sums*, C.R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 469-472.

[2] ———, *A short proof of Selberg's generalized beta formula*, Forum Math., 3 (1991), pp. 415-417.

[3] G. ANDREWS, *q-Series: their development and application in analysis, number theory, combinatorics, physics, and computer algebra*, Regional Conference Series in Math., 66, American Mathematical Society, Providence, RI, 1986.

[4] G. ANDREWS AND R. ASKEY, *Another q-extension of the beta function*, Proc. Amer. Math. Soc., 81 (1981), pp. 97-100.

[5] K. AOMOTO, *Jacobi polynomials associated with Selberg integrals*, SIAM J. Math. Anal., 18 (1987), pp. 545-549.

[6] R. ASKEY, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938-951.

[7] ———, *Computer algebra and definite integrals*, in *Computer Algebra*, D. Chudnovsky and R. Jenks, eds., pp. 121-128, Dekker, New York, 1989.

[8] ———, *Integration and computers*, in *Computers in Mathematics*, D. Chudnovsky and R. Jenks, eds., pp. 35-82, Dekker, New York, 1990.

[9] D. BRESSOUD AND I. GOULDEN, *Constant term identities extending the q-Dyson theorem*, Trans. Amer. Math. Soc., 291 (1985), pp. 203-228.

[10] R. EVANS, *The evaluation of Selberg character sums*, Enseign. Math., (2), to appear.

[11] G. GASPAR AND M. RAHMAN, *Basic hypergeometric series*, Encyclopedia of Mathematics and Its Applications, Vol. 35, Cambridge University Press, NY, 1990.

[12] L. HABSIEGER, *Une q-intégrale de Selberg et Askey*, SIAM J. Math. Anal., 19 (1988), pp. 1475-1489.

[13] K. KADELL, *A proof of some q-analogues of Selberg's integral for $k=1$*, SIAM J. Math. Anal., 19 (1988), pp. 944-968.

[14] ———, *A proof of Askey's conjectured q-analogue of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal., 19 (1988), pp. 969-986.

[15] ———, *The Selberg-Jack symmetric functions*, Adv. Math., to appear.

[16] ———, *A proof of the q-Macdonald-Morris conjecture for BC_n*, Trans. Amer. Math. Soc., to appear.

[17] A. SELBERG, *Bemerkninger om et multipelt integral*, Norsk Mat. Tidsskrift, 26 (1944), pp. 71-78 (Collected Papers, I, No. 14).

[18] P. VAN WAMELEN, *Proof of Evans-Root conjectures for Selberg character sums*, to appear.

SIAM J. MATH. ANAL.
Vol. 23, No. 3, pp. 766-784, May 1992

ON REFINEMENT EQUATIONS DETERMINED BY PÓLYA FREQUENCY SEQUENCES*

T. N. T. GOODMAN† AND CHARLES A. MICCHELLI‡

Abstract. The refinement equation

$$\phi(x)=\sum_{i\in\mathbb{Z}} a_i\phi(2x-i), \qquad x\in\mathbb{R}$$

for a given sequence $\mathbf{a}=\{a_j\colon i\in\mathbb{Z}\}$ has found important application in the study of both *Stationary Subdivision Schemes* for the generation of curves and surfaces as well as the construction of *orthonormal wavelets* by means of *multiresolution analysis.* The main goal here is to study properties of the solution of this equation when the sequence $\mathbf{a}$ is a *Pólya frequency sequence.* In the case that $\operatorname{supp}\mathbf{a}:=\{k\colon a_k\neq 0, k\in\mathbb{Z}\}$ is finite the refinement equation is also considered when

$$a(z)=\sum_{j=0}^{n} a_j z^j$$

is a Hurwitz polynomial (has all zeros in the left-half plane).

Key words. subdivision, wavelets, refinement equation, total positivity, Pólya frequency sequences

AMS(MOS) subject classifications. primary 41A15, 39B20; secondary 15A48

1. Introduction.

The refinement equation

$$\varphi(x)=\sum_{i\in\mathbb{Z}} a_i\varphi(2x-i), \qquad x\in\mathbb{R} \tag{1.1}$$

for a given sequence $\mathbf{a}=\{a_i\colon i\in\mathbb{Z}\}$ has found important application in the study of both *Stationary Subdivision Schemes* for the generation of curves and surfaces [3] as well as the construction of *orthonormal wavelets* by means of *multiresolution analysis* [5], [11].

Our main goal here is to study properties of the solution (1.1) when the sequence $\mathbf{a}$ is a *Pólya frequency sequence.* Recall that this requires *all* the minors of the bi-infinite matrix A

$$A_{ij}=a_{j-i} \qquad i,j\in\mathbb{Z} \tag{1.2}$$

to be nonnegative, that is,

$$A\begin{pmatrix} i_1,\cdots,i_p \\ j_1,\cdots,j_p\end{pmatrix}:=\det_{k,l=1,\cdots,p} A_{i_kj_l}\geqq 0 \tag{1.3}$$

for all integers $i_1<\cdots<i_p, j_1<\cdots<j_p$.

Pólya frequency sequences have been studied by numerous authors [1], [7], [8], and a *complete* characterization of such sequences is available in terms of the symbol

$$a(z)=\sum_{i\in\mathbb{Z}} a_i z^i, \qquad z\in\mathbb{C}. \tag{1.4}$$

* Received by the editors September 17, 1990; accepted for publication (in revised form) June 25, 1991.

† Department of Mathematics and Computer Science, The University, Dundee DD1 4HN, Scotland, United Kingdom.

‡ Department of Mathematical Sciences, IBM T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.

We will describe this later as it will be important to our analysis. Recall that the symbol of every Pólya frequency sequence (other than the trivial sequence $\{\gamma^n\colon n \in \mathbb{Z}\}$ for some $\gamma > 0$) converges in some annulus [9, p. 418]. We assume throughout that $\mathbf{a}$ is not a trivial Pólya frequency sequence.

A basic example of a solution to (1.1) comes from the theory of *spline functions.* We let χ denote the characteristic function of the interval $[0, 1]$ and define

$$M_n = \chi * \cdots * \chi, \quad n \text{ factors}, \quad n \geqq 2, \tag{1.5}$$

where $*$ signifies convolution. It is known [14] that M_n is a polynomial of degree $\leqq n-1$ on each interval $(j, j+1)$, $j \in \mathbb{Z}$, has $n-2$ continuous derivatives on $\mathbb{R}$, is zero outside of $(0, n)$, and is positive otherwise. Clearly, the Fourier transform of M_n is

$$\hat{M}_n(t) := \int_{\mathbb{R}} e^{itx} M_n(x)\,dx = \left(\frac{e^{it}-1}{it}\right)^n, \qquad t \in \mathbb{R}, \tag{1.6}$$

and so $\phi := M_n$ satisfies (1.1) with

$$m(z) = \sum_{j \in \mathbb{Z}} m_j z^j = 2^{-n+1}(1+z)^n, \qquad z \in \mathbb{C}, \tag{1.7}$$

that is,

$$M_n(x) = \sum_{j \in \mathbb{Z}} m_j M_n(2x - j), \qquad x \in \mathbb{R}. \tag{1.8}$$

This equation is central in the development of the *line average algorithm* for the fast computation of curves, cf. [4].

The function M_n also has a remarkable *variation diminishing* property. Specifically, if $S^-(f) =$ the number of sign changes of f on $\mathbb{R}$ and similarly for a sequence $\mathbf{c} = \{c_j\colon j \in \mathbb{Z}\}$, $S^-(\mathbf{c}) =$ the number of (strict) sign changes in $\mathbf{c}$ we have

$$S^-\left(\sum_{j \in \mathbb{Z}} c_j M_n(\bullet - j)\right) \leqq S^-(\mathbf{c}). \tag{1.9}$$

It is known that variation diminishing is a consequence of determinental inequalities cf. [9]. In fact, given a function φ such that

$$\Phi\begin{pmatrix} x_1, \cdots, x_p \\ i_1, \cdots, i_p \end{pmatrix} := \det_{l,j=1,\cdots,p} \varphi(x_l - i_j) \tag{1.10}$$

is nonnegative for all $x_1 < \cdots < x_p$ and integers $i_1 < \cdots < i_p$, φ is variation diminishing in the sense that for any sequence $\mathbf{c}$ of finite support

$$S^-\left(\sum_{j \in \mathbb{Z}} c_j \varphi(\bullet - j)\right) \leqq S^-(\mathbf{c}). \tag{1.11}$$

We will call any function, such that the determinants in (1.10) are nonnegative, a *ripplet.* Ripplets arise in various contexts, for instance, the B-spline with integer knots determined by a constant coefficient differential operator, whose characteristic polynomial only has real zeros is a ripplet, cf. [13]. Also, the B-spline for geometrically continuous splines studied in [6] is a ripplet.

Functions ϕ satisfying the stronger requirement

$$\Phi\begin{pmatrix} x_1, \cdots, x_p \\ y_1, \cdots, y_p \end{pmatrix} \geqq 0 \tag{1.12}$$

for all $x_1 < \cdots < x_p$ and $y_1 < \cdots < y_p$ are called *Pólya frequency functions* and have been studied extensively, cf. [9].

One useful consequence of (1.11) is the following, which has applications in computer-aided design. Given a ripplet φ with $\sum_{j\in\mathbb{Z}} \varphi(x-j)=1$, $x\in\mathbb{R}$, and a sequence $\{V_j: j\in\mathbb{Z}\}$ in $\mathbb{R}^s$, $s\geqq 2$, we may define a curve in $\mathbb{R}^s$ by

$$R(t)=\sum_{j=\mathbb{Z}} V_j\varphi(t-j), \qquad t\in\mathbb{R}. \tag{1.13}$$

The variation diminishing property (1.11) then implies that the curve R cuts any given straight line no more than the polygonal arc P, with consecutive vertices $\{V_j: j\in\mathbb{Z}\}$. This ensures that in a sense the shape of the curve R mimics that of the polygonal arc P and so the points $\{V_j: j\in\mathbb{Z}\}$ can be used to predict or control the shape of the curve R. (For further details see Goodman, "Shape preserving representatives," in *Mathematical Methods in Computer Aided Geometric Design.*)

Among other things, we will show here that any Pólya frequency sequence whose symbol can be factored as

$$a(z)=(1+z)q(z),\ q(1)=1, \qquad z\in\mathbb{C}, \tag{1.14}$$

where $q(z)\neq z^k$ for any $k\in\mathbb{Z}$ determines a ripplet solution to (1.1).

Returning to the B-spline M_n, it is known, cf. [9], that the determinants

$$M_n\begin{pmatrix} x_1, \cdots, x_p \\ i_1, \cdots, i_p \end{pmatrix} = \det_{i,l=1,\cdots,p} M_n(x_i-i_l) \tag{1.15}$$

are *strictly positive* if and only if

$$i_l<x_l<i_l+n, \qquad l=1,\cdots,p. \tag{1.16}$$

It was recently conjectured in [12] that the same result holds for the ripplet satisfying (1.1) when **a** is a finite Pólya frequency sequence $\cdots,0,a_0,\cdots,a_n,0,\cdots$. We will prove this is in fact the case even when $a(z)$ is left-half plane stable as well as delineate when the determinants appearing in (1.10) are strictly positive for the solution of (1.1) corresponding to any Pólya frequency sequence.

2. Existence of refinable ripplets. In this section we prove the existence of a unique ripplet solution to the refinement equation when **a** is a Pólya frequency sequence. Specifically, we have the following.

THEOREM 2.1. *Let $a=\{a_j: j\in\mathbb{Z}\}$ be a sequence whose symbol satisfies*:

$$a(z)=(1+z)q(z),\ q(1)=1, \tag{2.1}$$

where $q(z)=\sum_{j\in\mathbb{Z}} q_j z^i$ satisfies $q_j\geqq 0$, $j\in\mathbb{Z}$, and

$$\rho:=\max\left(\sum_{i\in\mathbb{Z}} q_{2i}, \sum_{i\in\mathbb{Z}} q_{2i+1}\right)<1. \tag{2.2}$$

Then there exists a continuous function ϕ such that

$$\varphi(x)=\sum_{j\in\mathbb{Z}} a_j\varphi(2x-j), \qquad x\in\mathbb{R}, \tag{2.3}$$

and

$$\sum_{j\in\mathbb{Z}} \varphi(x-j)=1, \qquad x\in\mathbb{R}. \tag{2.4}$$

Moreover, if a is a Pólya frequency sequence satisfying (2.1) *with $q(z)\neq z^l$ for any $l\in\mathbb{Z}$, then there is a continuous ripplet φ satisfying* (2.3) *and* (2.4), *and $\varphi(x)=0$ if $x\notin I^0$ where $I:=$ the smallest closed interval containing* $\operatorname{supp}\mathbf{a}:=\{k: k\in\mathbb{Z}, a_k>0\}$.

Remark 2.1. Our condition (2.1) means that $a(1)=2$ and $a(-1)=0$. It is important to realize that the factor $q(z)$ is also the symbol of a Pólya frequency sequence. This follows from the factorization theorem for the symbol of a Pólya frequency sequence which we now describe.

We shall write $\mathbb{Z}_+=\{j\in\mathbb{Z}: j\geqq 0\}$ and $\mathbb{R}_+=\{t\in\mathbb{R}: t\geqq 0\}$ in what follows. Moreover, for any sequence $\boldsymbol{\alpha}=\{\alpha_j: j\in\mathbb{Z}_+\}\subseteq\mathbb{R}_+$, which is summable $\sum_{j\in\mathbb{Z}_+}\alpha_j<\infty$, i.e., $\boldsymbol{\alpha}\in l^1(\mathbb{Z}_+)$ we set

$$f(z;\boldsymbol{\alpha})=\prod_{j=1}^{\infty}(1+\alpha_j z),\qquad z\in\mathbb{C}. \tag{2.5}$$

For later use we let

$$\boldsymbol{\alpha}_+=|\{j: \alpha_j>0, j\in\mathbb{Z}_+\}| \tag{2.6}$$

so that $\boldsymbol{\alpha}_+=0$ means $f(\bullet;\boldsymbol{\alpha})=1$ and $\boldsymbol{\alpha}_+<\infty$ means $f(\bullet;\boldsymbol{\alpha})$ is a polynomial with negative zeros.

The fundamental fact about Pólya frequency sequences is the following result, cf. [9, Thm. 9.5, p. 427].

THEOREM A. *A necessary and sufficient condition for* $\mathbf{a}=\{a_j: j\in\mathbb{Z}\}$ *to be a Pólya frequency sequence is that*

$$a(z)=rz^k\exp(sz+tz^{-1})\frac{f(z;\boldsymbol{\alpha})f(z^{-1};\boldsymbol{\delta})}{f(z;-\boldsymbol{\beta})f(z^{-1};-\boldsymbol{\gamma})},\qquad z\in\mathbb{C} \tag{2.7}$$

for some nonnegative $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta}\in l_1(\mathbb{Z}_+)$ *as above and scalars* $r>0$, $s, t\geqq 0$ *where* k *is some integer.*

Therefore, if $a(-1)=0$, $1+z$ must be a factor and the remaining function still has the form (2.7).

Now that we know that $q(z)$ is also the symbol of a Pólya frequency sequence q we are assured by our hypothesis $q(1)=1$ that (2.2) is satisfied. This will be important in the proof of Theorem 2.1.

Let us now proceed to the proof which uses ideas from [3], where the result is proved for the case that $\mathbf{a}$ is of the finite support.

Proof of Theorem 2.1. We introduce the operator

$$(F_{\mathbf{a}}h)(x):=\sum_{k\in\mathbb{Z}}a_k h(2x-k),\qquad x\in\mathbb{R}. \tag{2.8}$$

The linear map F takes $C(\mathbb{R})$ into itself and has norm two; specifically, we have

$$\|F_{\mathbf{a}}h\|_\infty\leqq 2\|h\|_\infty,\qquad h\in C(\mathbb{R}), \tag{2.9}$$

where $\|h\|_\infty=\sup\{|h(x)|: x\in\mathbb{R}\}$.

A basic formula from [3] states that for any $\lambda\in l_1(\mathbb{Z})$, $h\in C(\mathbb{R})$

$$\sum_{k\in\mathbb{Z}}\lambda_k(F_{\mathbf{a}}h)(x-k)=\sum_{k\in\mathbb{Z}}(S_{\mathbf{a}}\lambda)_k h(2x-k),\qquad x\in\mathbb{R}, \tag{2.10}$$

where

$$(S_{\mathbf{a}}\lambda)_k=\sum_{j\in\mathbb{Z}}a_{k-2j}\lambda_j,\qquad k\in\mathbb{Z}. \tag{2.11}$$

Therefore, inductively, for any $j\in\mathbb{Z}_+$

$$\sum_{k\in\mathbb{Z}}\lambda_k(F_{\mathbf{a}}^j h)(x-k)=\sum_{k\in\mathbb{Z}}(S_{\mathbf{a}}^j\lambda)_k h(2^j x-k),\qquad x\in\mathbb{R}. \tag{2.12}$$

We choose any integer shift of the B-spline $N_n = M_n(\bullet - r)$, $r \in \mathbb{Z}$ (of degree $n \geqq 2$), and consider the sequence of functions

$$g_j = F_{\mathbf{a}}^j N_n, \qquad j \in \mathbb{Z}_+. \tag{2.13}$$

We will explain later how we choose r and n.

According to (2.12), g_j has a B-spline expansion given by

$$g_j(x) = \sum_{k \in \mathbb{Z}} (S_{\mathbf{a}}^j \delta)_k N_n(2^j x - k), \qquad x \in \mathbb{R}, \tag{2.14}$$

where

$$\delta_k := \begin{cases} 1, & k = 0 \\ 0, & k \in \mathbb{Z} \backslash \{0\}. \end{cases}$$

Let us show that g_m converges to a solution to the refinement equation (1.1). Following arguments in [3] we can show there exists a positive constant $k > 0$ such that

$$\|g_{j+1} - g_j\|_\infty \leqq k \|\Delta S_{\mathbf{a}}^j \delta\|_\infty. \tag{2.15}$$

The next step is to realize that the equation (2.1) implies that

$$\Delta S_{\mathbf{a}} = S_{\mathbf{q}} \Delta, \tag{2.16}$$

so that

$$\|\Delta S_{\mathbf{a}} \lambda\|_\infty \leqq \|S_{\mathbf{q}}\| \, \|\Delta \lambda\|_\infty = \rho \|\Delta \lambda\|_\infty, \tag{2.17}$$

where ρ is defined by (2.2); consequently we obtain

$$\|g_{j+1} - g_j\|_\infty \leqq k \rho^j, \qquad j \in \mathbb{Z}_+. \tag{2.18}$$

Since $\rho < 1$, there is a $\varphi \in C(\mathbb{R})$ such that

$$\lim_{j \to \infty} g_j(x) = \varphi(x) \tag{2.19}$$

uniformly on $\mathbb{R}$. This is the function we seek. In fact, since $g_{j+1} = F_{\mathbf{a}} g_j$ we see that φ satisfies the refinement equation (1.1).

Let us now show that the sum of integer translates of φ is one. For this purpose we return to (2.10) and note that when h is of compact support, the right-hand side of (2.10) is finite for all $x \in \mathbb{R}$ even when $\lambda \in l_\infty(\mathbb{Z})$. As for the left-hand side, we note that the mapping $F_{\mathbf{a}}$ is *nonexpansive* relative to the norm

$$|h| := \max \left\{ \sum_{l \in \mathbb{Z}} |h(x - l)| : x \in [0, 1] \right\},$$

that is,

$$|F_a h| \leqq |h|,$$

since $\sum_{l \in \mathbb{Z}} a_{k-2l} = 1$, $k \in \mathbb{Z}$. Obviously then, the left-hand side of (2.10) is also finite for all x when $\lambda \in l_\infty(\mathbb{Z})$. Thus (2.10), as well as (2.12), remain valid for $\lambda \in l_\infty(\mathbb{Z})$ and h of compact support. Thus, choosing for λ, h in (2.12) l and N_n, respectively, where $e_k = 1$, $k \in \mathbb{Z}$, we get

$$\sum_{k \in \mathbb{Z}} g_j(x - k) = \sum_{k \in \mathbb{Z}} (S_a^j e)_k N_n(2^j x - k) = \sum_{k \in \mathbb{Z}} N_n(2^j x - k) = 1, \qquad x \in \mathbb{R}.$$

This confirms the claim (2.4).

Henceforward, we assume a is a Pólya frequency sequence. It remains to verify that φ is a ripplet. As we already pointed out $g_0 = N_n$ is a ripplet. Therefore, we can see inductively, in the following way, that each g_j, $j \in \mathbb{Z}$ is a ripplet. We write

$$g_{j+1}(x-i) = \sum_{k\in\mathbb{Z}} a_k g_j(2x-2i-k) = \sum_{k\in\mathbb{Z}} a_{k-2i} g_j(2x-k), \qquad x \in \mathbb{R}. \tag{2.20}$$

Therefore, by the Cauchy-Binet formula (cf. [9, p.1])

$$\begin{aligned} G_{j+1}\begin{pmatrix} x_1, \cdots, x_s \\ i_1, \cdots, i_s \end{pmatrix} &:= \det_{l,r=1,\cdots,s} g_{j+1}(x_l - i_r) \\ &= \sum_{k_1<\cdots<k_s} A\begin{pmatrix} 2i_1, \cdots, 2i_s \\ k_1, \cdots, k_s \end{pmatrix} G_j\begin{pmatrix} 2x_1, \cdots, 2x_s \\ k_1, \cdots, k_s \end{pmatrix}. \end{aligned} \tag{2.21}$$

Since $\mathbf{a} = \{a_j : j \in \mathbb{Z}\}$ is a Pólya frequency sequence, it follows inductively that each g_j is a ripplet and hence so is φ.

For the last claim we recall that for a Pólya frequency sequence the set $\operatorname{supp} \mathbf{a} := \{k : k \in \mathbb{Z}, a_k > 0\}$ consists of a set of *consecutive integers*, cf. [9, p. 418]. Thus $\operatorname{supp} \mathbf{a} = I \cap \mathbb{Z}$ and $I = [k_-, k_+]$ where k_-, k_+ are integers, either of which may be infinite. Moreover, because of our hypothesis that $q(z) \neq z^l$ for any l we conclude that $k_+ - k_- \geqq 2$. Referring back to the iterative relation (2.20) we see that if $g_j(x) = 0$, for $x \notin I^0$, then also $g_{j+1}(x) = 0$ for $x \notin I^0$. Now, we may choose the degree of the B-spline N_n and the shift $r \in \mathbb{Z}$ so that for $j = 0$, $g_0(x) = N_n(x) = M_n(x-r)$ is zero for $x \notin I^0$, that is, $r = k_-$ and $n = k_+ - k_-$. Consequently, we conclude that $\varphi(x) = 0$, for $x \notin I^0$ as well. This completes the proof of Theorem 2.1.

3. Strict positivity of minors. In this section we develop criteria for the determinants (1.10) of the ripplet φ to be *strictly* positive. The proof is long and is distinguished by several cases. A basic ingredient is a theorem of Karlin [9, p. 428] which describes, in terms of the parameters of the factorization (2.6), when the minors of the matrix A defined by (1.2), (1.3) are strictly positive. Because we will refer to this fact several times during the course of our analysis, we begin by describing it below. For simplicity, we assume the integer k appearing in (2.6) is zero. This can always be arranged by an integer shift of $\mathbf{a}$ and φ.

THEOREM B. *Let $\mathbf{a} = \{a_j : j \in \mathbb{Z}\}$ be a Pólya frequency sequence whose symbol has the factorization* (2.6) *with $k = 0$. Then*

(a) *If $s > 0$ and $t > 0$ then the determinant* (1.3) *is always positive*;

(b) *If $s > 0$ and $t = 0$ then the determinant* (1.3) *is positive if and only if*

$$i_k < j_{k+\gamma_+} + \delta_+, \qquad k = 1, \cdots, p, \tag{3.1}$$

where equality is allowed for any k if $\gamma_+ = 0$;

(c) *If $s = 0$ and $t > 0$ then the determinant* (1.3) *is positive if and only if*

$$j_{k-\beta_+} - \alpha_+ < i_k, \qquad k = 1, \cdots, p, \tag{3.2}$$

where equality is allowed for any k, if $\beta_+ = 0$;

(d) *If $s = t = 0$ then the determinant* (1.3) *is positive if and only if*

$$j_{k-\beta_+} - \alpha_+ < i_k < j_{k+\gamma_+} + \delta_+, \qquad k = 1, \cdots, p, \tag{3.3}$$

if $\beta_+ = 0$ equality is allowed on the left-hand side of (3.3); *if $\gamma_+ = 0$ equality is allowed on the right-hand side of* (3.3).

In the equalities (3.1)-(3.3) we interpret i_k, j_k for $k < 1$ as $-\infty$, and as ∞ if $k > p$.

We use this theorem to examine when the powers of the submatrix D of A defined by

$$D_{ij} := a_{j-2i}, \qquad i, j \in \mathbb{Z} \tag{3.4}$$

are positive. It will become clear why we need this later. At this point we might note that D already appeared in the proof of Theorem 2.1, see (2.25).

Obviously,

$$D\begin{pmatrix} i_1, \cdots, i_p \\ j_1, \cdots, j_p \end{pmatrix} = A\begin{pmatrix} 2i_1, \cdots, 2i_p \\ j_1, \cdots, j_p \end{pmatrix}, \tag{3.5}$$

and so we can easily decide from Theorem B which minors of D are positive. For any $j \in \mathbb{Z}_+$ the symbol for the matrix A^j is $(a(z))^j$, and so Theorem B easily tells us which minors of A^j are positive. However, generally the minors of D^j are not minors of A^j.

Specifically, if we define the sequences $\mathbf{a}^r = \{a_i^r \colon i \in \mathbb{Z}\}$ by

$$a(z)a(z^2) \cdots a(z^{2^{r-1}}) = \sum_{i \in \mathbb{Z}} a_i^r z^i, \tag{3.6}$$

then it is straightforward to verify by induction on r that

$$D_{ij}^r = a_{j-2^r i}^r, \qquad r = 1, 2, \cdots.$$

Therefore, minors of D^r correspond to minors of the Toeplitz matrix determined by the sequence $\mathbf{a}^r$. It is apparent that the generating function of $\mathbf{a}^r$, $r \geqq 2$ does not have an expansion in the form (2.6) and so is not a Pólya frequency sequence. Nevertheless we have the following, employing the notation of Theorem A.

THEOREM 3.1. *Let* $\mathbf{a}$ *be a Pólya frequency sequence, and if* $s = t = 0$ *and* $\boldsymbol{\beta}_+ + \boldsymbol{\gamma}_+ > 0$, *then either* $p \leqq \boldsymbol{\beta}_+ + \boldsymbol{\gamma}_+$ *or* $p \leqq \boldsymbol{\alpha}_+ + \boldsymbol{\delta}_+ + \max(\boldsymbol{\beta}_+ - 1, 0) + \max(\boldsymbol{\gamma}_+ - 1, 0)$. *Then given* $r \in \mathbb{Z}_+ \backslash \{0\}$

$$D^r\begin{pmatrix} i_1, \cdots, i_p \\ j_1, \cdots, j_p \end{pmatrix} \geqq 0 \tag{3.7}$$

for $i_1 < \cdots < i_p$, $j_1 < \cdots < j_p$. *Strict inequality holds in* (3.7) *if and only if*

(a) $s > 0$ *and* $t > 0$;

or

(b) $s > 0$, $t = 0$ *and*

$$i_l \leqq 2^{-r} j_{l + r\boldsymbol{\gamma}_+} + (1 - 2^{-r})(\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \qquad l = 1, \cdots, p; \tag{3.8}$$

or

(c) $s = 0$, $t > 0$ *and*

$$2^{-r} j_{l - r\boldsymbol{\beta}_+} + (1 - 2^{-r})(\min(\boldsymbol{\beta}_+, 1) - \boldsymbol{\alpha}_+) \leqq i_l, \qquad l = 1, \cdots, p; \tag{3.9}$$

or

(d) $s = t = 0$ *and*

$$\begin{aligned} &2^{-r} j_{l - r\boldsymbol{\beta}_+} + (1 - 2^{-r})(\min(\boldsymbol{\beta}_+, 1) - \boldsymbol{\alpha}_+) \\ &\quad \leqq i_l \leqq 2^{-r} j_{l + r\boldsymbol{\gamma}_+} + (1 - 2^{-r})(\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \qquad l = 1, \cdots, p. \end{aligned} \tag{3.10}$$

To prove Theorem 3.1 we shall need the following.

LEMMA 3.1. *Take* $\beta, \gamma, \mu, \nu \in \mathbb{Z}_+$, $r, p \in \mathbb{Z}_+ \backslash \{0\}$, $m, n \in \mathbb{Z}$ *and* $a_i, b_i \in \mathbb{Z} \cup \{\pm\infty\}$ *with* $a_{i+1} \geqq a_1 + 1$, $b_{i+1} \geqq b_i + 1$, *for* $i \in \mathbb{Z}$. *Suppose that* $m + n \geqq 0$ *and if* $\beta + \gamma + \mu + \nu > 0$, *then* $m + n \geqq p$. *If for* $l = 1, \cdots, p$,

$$a_{l-\gamma} - n \leqq 2^{-r-1}(b_{l+\nu} - n), \tag{3.11}$$

$$2^{-r-1}(b_{l-\mu} + m) \leqq a_{l+\beta} + m, \tag{3.12}$$

then there exist $k_1 < \cdots < k_p$ *in* $\mathbb{Z}$ *such that for* $l = 1, \cdots, p$,

$$a_{l-\gamma} - n \leqq 2^{-1}(k_l - n) \leqq 2^{-r-1}(b_{l+\nu} - n), \tag{3.13}$$

$$2^{-r-1}(b_{l-\mu} + m) \leqq 2^{-1}(k_l + m) \leqq a_{l+\beta} + m. \tag{3.14}$$

Proof. This is by induction on p. We take p as in Lemma 3.1 and suppose that the result is true for all smaller values of p (a vacuous assumption if $p=1$). If $a_{1-\gamma}=b_{1-\mu}=-\infty$, then the inductive hypothesis ensures the existence of $k_2,\cdots,k_p$, satisfying (3.13) and (3.14), and k_1 can be chosen with $k_1<k_2$ to satisfy the upper bounds in (3.13), (3.14). So we may assume $a_{1-\gamma}>-\infty$ or $b_{1-\mu}>-\infty$.

Now for $l=1$, define k_l to be the smallest integer satisfying

$$a_{l-\gamma}-n\leqq 2^{-1}(k_l-n), \tag{3.15}$$

$$2^{-r-1}(b_{l-\mu}+m)\leqq 2^{-1}(k_l+m), \tag{3.16}$$

and for $l=2,\cdots,p$, define k_l to be the minimum integer satisfying (3.15), (3.16), and $k_l>k_{l-1}$. By our inductive hypothesis, (3.13), (3.14) are satisfied for $l=1,\cdots,p-1$, and so it remains only to prove

$$k_p\leqq 2^{-r}(b_{p+\nu}-n)+n, \tag{3.17}$$

$$k_p\leqq 2a_{p+\beta}+m. \tag{3.18}$$

Now suppose that for some i, $1\leqq i\leqq p-1$, $k_{i+1}\geqq k_i+2$. Then k_{i+1} is the minimum integer satisfying (3.15), (3.16) with $l=i+1$ and so, by our inductive hypothesis, k_l satisfies (3.13), (3.14) for $l=i+1,\cdots,p$, which gives (3.17) and (3.18). Thus, we may assume

$$k_{l+1}=k_l+1,\qquad l=1,\cdots,p-1. \tag{3.19}$$

Suppose that (3.18) does not hold, i.e.,

$$k_p\geqq 2a_{p+\beta}+m+1. \tag{3.20}$$

If $p\geqq 2$, then

$$k_p=k_{p-1}+1\leqq 2a_{p-1+\beta}+m+1\leqq 2a_{p+\beta}+m-1,$$

which contradicts (3.20). Suppose $p=1$, then (3.20) gives

$$k_1-1-n\geqq 2a_{1+\beta}+m-n\geqq 2a_{1-\gamma}-2n$$

since $m+n\geqq 0$. Also (3.20) gives

$$k_1-1+m\geqq 2a_{1+\beta}+2m\geqq 2^{-r}(b_{1-\mu}+m)$$

by (3.12). This contradicts the definition of k_1. Thus we have proved (3.18) and it remains to prove (3.17). Suppose that (3.17) is not true, which by (3.19) gives

$$k_1>2^{-r}(b_{p+\nu}-n)+n+1-p. \tag{3.21}$$

Then by (3.11),

$$k_1>2a_{p-\gamma}-n+1-p\geqq 2a_{1-\gamma}+p-1-n,$$

and so

$$k_1\geqq 2a_{1-\gamma}-n+1,$$

which gives

$$2^{-1}(k_1-1-n)\geqq a_{1-\gamma}-n.$$

By definition, k_1 is the smallest integer satisfying (3.16) and so

$$k_1<2^{-r}(b_{1-\mu}+m)+1-m. \tag{3.22}$$

Then

$$\begin{aligned} k_1 &\leqq 2^{-r}(b_{1-\mu}+m-1)+1-m \\ &\leqq 2^{-r}(b_{p+\nu}-p+m)+1-m \\ &= 2^{-r}(b_{p+\nu}-n)+n+1-p+(p-m-n)(1-2^{-r}). \end{aligned}$$

If $p \leqq m+n$, then this contradicts (3.21). So we may assume $\beta=\gamma=\mu=\nu=0$ and $m+n \leqq p-1$. Note that by (3.19) and (3.15),

$$\begin{aligned} k_1+p-1=k_p &\geqq 2a_p-n \\ &\geqq 2a_1+2p-2-n \\ &\geqq 2^{-r}(b_1+m)+2p-2-n-2m \\ &> k_1+2p-3-n-m \end{aligned}$$

by (3.12) and (3.22). Thus, $m+n \geqq p-1$ and it remains to consider the case $m+n=p-1$. Let

$$b_1+m-1=2^{r+1}a+t$$

for some $s, t \in \mathbb{Z}$, such that $0 \leqq t < 2^{r+1}$.

By (3.12),

$$a_1+m \geqq 2^{-r-1}(b_1+m)=s+2^{-r-1}(t+1)>s$$

and so

$$a_1+m \geqq s+1.$$

Also by (3.11),

$$2^{-r-1}(b_p-n) \geqq a_p-n \geqq a_1+p-1-n=a_1+m \geqq s+1. \tag{3.23}$$

Now by (3.22),

$$k_1 < 2s+2^{-r}(t+1)+1-m \leqq 2s+3-m$$

and so

$$k_1 \leqq 2s+2-m \leqq 2^{-r}(b_p-n)-m$$

by (3.23), which contradicts (3.21) and completes the proof of the Lemma.

Proof of Theorem 3.1. The proof of (a)-(d) is by induction on r. The case $r=1$ follows from Theorem B and (3.5). The induction step is based on the formula

$$D^{r+1}\begin{pmatrix} i_1, \cdots, i_p \\ j_1, \cdots, j_p \end{pmatrix} = \sum_{k_1<\cdots<k_p} D\begin{pmatrix} i_1, \cdots, i_p \\ k_1, \cdots, k_p \end{pmatrix} D^r\begin{pmatrix} k_1, \cdots, k_p \\ j_1, \cdots, j_p \end{pmatrix}, \tag{3.24}$$

which is a consequence of the Cauchy-Binet formula. This formula immediately establishes the nonnegativity of the minors of D^r as stated by (3.7). The proof of (a)-(d) is more involved.

To advance the induction hypothesis we must show that there is a choice of integers $k_1^0 < \cdots < k_p^0$, such that the corresponding summand in the right-hand side of (3.24) is positive. This is an easy matter for case (a) as there are no constraints on $i_1 < \cdots < i_p$ and $j_1 < \cdots < j_p$. Let's look now at (b). There are some cases we may easily handle. For instance, when $\delta_+ = \infty$ or $\gamma_+ = \infty$ there are again no conditions on $i_1 < \cdots < i_p$ and $j_1 < \cdots < j_p$, and so the induction follows immediately from (3.8). Thus, we consider only the case when $\boldsymbol{\delta}_+ < \infty$ and $\boldsymbol{\gamma}_+ < \infty$. We suppose that (3.8) is valid for r replaced by $r+1$. Our goal is to find $k_1^0 < \cdots < k_p^0$, such that

$$i_l \leqq 2^{-1} k^0_{l+\boldsymbol{\gamma}_+} + 2^{-1}(\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \qquad l = 1, \cdots, p \tag{3.25}$$

and

$$k_l^0 \leqq 2^{-r} j_{l+r\boldsymbol{\gamma}_+} + (1 - 2^{-r})(\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \qquad l = 1, \cdots, p. \tag{3.26}$$

The inequalities (3.25) and (3.26) insure the summand on the right-hand side of (3.11), corresponding to $k_i = k_i^0$, $i = 1, \cdots, p$, is nonzero by the induction hypothesis. As for (3.25) and (3.26) we choose

$$k_l^0 = 2i_{l-\boldsymbol{\gamma}_+} - (\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \qquad l > \boldsymbol{\gamma}_+ \tag{3.27}$$

and *any* $k_1^0 < \cdots < k^0_{\boldsymbol{\gamma}_+}$ so that

$$k^0_{\boldsymbol{\gamma}_+} \leqq 2i_{1-\boldsymbol{\gamma}_+} - (\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)). \tag{3.28}$$

For this choice, note that (3.25) is automatically satisfied. As for (3.26) we observe that the induction hypothesis implies that

$$\begin{aligned} &2i_{l-\boldsymbol{\gamma}_+} - (\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)) \le 2^{-r} j_{l+r\boldsymbol{\gamma}_+} + (1 - 2^{-r})(\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad l = 1, \cdots, p, \end{aligned} \tag{3.29}$$

and hence (3.26) follows for $l > \boldsymbol{\gamma}_+$. For $l \leqq \boldsymbol{\gamma}_+$ we use (3.29) for $l = 1$ to obtain

$$\begin{aligned} k_l^{sz} \leqq k^{sz}_{\boldsymbol{\gamma}_+} &\leqq 2i_{1-\boldsymbol{\gamma}_+} - (\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)) \\ &\leqq 2^{-r} j_{l+r\boldsymbol{\gamma}_+} + (1 - 2^{-r})(\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)) \end{aligned}$$

because obviously $j_{1+r\boldsymbol{\gamma}_+} \leqq j_{l+r\boldsymbol{\gamma}_+}$. Thus, our choice of $k_1^0, \cdots, k_p^0$ satisfies (3.25) and (3.26). Therefore, the induction has been advanced and sufficiency of (b) has been established. For the necessity of (3.8) we note that according to (3.24) and the induction hypotheses the determinant on the left-hand side of (3.24) is positive if and only if there exists $k_1^0 < \cdots < k_p^0$ satisfying the inequalities of (3.25) and (3.26). These clearly imply (3.8) for r replaced by $r+1$. The proof of (c) follows analogously and we omit the details.

The final case (d) introduces further difficulties which need some explanation. First we observe that all the quantities $\boldsymbol{\alpha}_+$, $\boldsymbol{\beta}_+$, $\boldsymbol{\gamma}_+$, and $\boldsymbol{\delta}_+$ can be assumed to be finite, otherwise (d) reduces to a previously considered case. Now, by induction, we assume

$$\begin{aligned} &2^{-r-1} j_{l-(r+1)\boldsymbol{\beta}_+} + (1 - 2^{-r-1})(\min(\boldsymbol{\beta}_+, 1) - \boldsymbol{\alpha}_+) \\ &\quad \leqq i_l \leqq 2^{-r-1} j_{l+(r+1)\boldsymbol{\gamma}_+} + (1 - 2^{-r-1}) \\ &\qquad\qquad \cdot (\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \qquad l = 1, \cdots, p, \end{aligned} \tag{3.30}$$

and seek integers $k_1^0 < \cdots < k_p^0$, such that

$$\begin{aligned} &2^{-1} k^0_{l-\boldsymbol{\beta}_+} + 2^{-1}(\min(\boldsymbol{\beta}_+, 1) - \boldsymbol{\alpha}_+) \\ &\quad \leqq i_l \leqq 2^{-1} k^0_{l+\boldsymbol{\gamma}_+} + 2^{-1}(\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1), \qquad l = 1, \cdots, p, \end{aligned} \tag{3.31}$$

and

$$\begin{aligned} &2^{-r} j_{l-r\boldsymbol{\beta}_+} + (1 - 2^{-r})(\min(\boldsymbol{\beta}_+, 1) - \boldsymbol{\alpha}_+) \\ &\quad \leqq k_l^0 \leqq 2^{-r} j_{l+r\boldsymbol{\gamma}_+} + (1 - 2^{-r}) \\ &\qquad\qquad \cdot (\boldsymbol{\delta}_+ - \min(\boldsymbol{\gamma}_+, 1)), \qquad l = 1, \cdots, p. \end{aligned} \tag{3.32}$$

It will be convenient to write in the remaining part of the proof

$$m=\alpha_+-\min(\beta_+,1), \qquad n=\delta_+-\min(\gamma_+,1)$$

and to rewrite (3.30) as

$$i_{l-\gamma_+}-n\leqq 2^{-r-1}(j_{l+r\gamma_+}-n), \quad 2^{-r-1}(j_{l-r\beta_+}+m)\leqq i_{l+\beta_+}+m, \quad l=1,\cdots,p, \tag{3.33}$$

and rewrite (3.31), (3.32) as

$$i_{l-\gamma_+}-n\leqq 2^{-1}(k_l^0-n)\leqq 2^{-r-1}(j_{l+r\gamma_+}-n), \qquad l=1,\cdots,p, \tag{3.34}$$

$$2^{-r-1}(j_{l-r\beta_+}+m)\leqq 2^{-1}(k_l^0+m)\leqq i_{l+\beta_+}+m, \qquad l=1,\cdots,p. \tag{3.35}$$

If $p\geqq\beta_++\gamma_++1$, then by assumption $p\leqq m+n+\beta_++\gamma_+$, if $\beta_++\gamma_+>0$, and clearly $m+n\geqq 0$ if $\beta_+=\gamma_+=0$. So we can apply Lemma 3.1 to define $k^0_{\beta_++1}<\cdots<k^0_{p-\gamma_+}$, satisfying (3.34), (3.35) for $l=\beta_++1,\cdots,p-\gamma_+$.

If $\gamma_+<\beta_+$, we define

$$k_l^0=2i_{l-\gamma_+}-n, \qquad l=\gamma_++1,\cdots,\beta_+.$$

For $l=1,\cdots,\min(\beta_+,\gamma_+)$, we define k_l^0 as any strictly increasing sequence, satisfying the upper bounds in (3.34), (3.35) and $k_l^0<k_{l+1}^0$ for $l=\min(\beta_+,\gamma_+)$. Now for $l=\gamma_++1,\cdots,\beta_+$,

$$2^{-1}(k_l^0-n)=i_{l-\gamma_+}-n\leqq 2^{-r-1}(j_{l+r\gamma_+}-n)$$

by (3.33), and

$$\begin{aligned}2^{-1}(k_l^0+m)&=i_{l-\gamma_+}+2^{-1}(m-n)\\&\leqq i_{l+\beta_+}-\beta_+-\gamma_++m-2^{-1}(m+n)\\&\leqq i_{l+\beta_+}+m,\end{aligned}$$

since $m+n+2\beta_++2\gamma_+\geqq 0$. Thus $k_1^0,\cdots,k^0_{\beta_+}$ satisfy (3.34) and (3.35). Similarly we define k_l^0 to satisfy (3.34), (3.35) for $l>\max(p-\gamma_+,\beta_+)$. For $l=\gamma_++1,\cdots,\beta_+$, (3.34) gives

$$k^0_{l+1}\geqq 2i_{l+1-\gamma_+}-n\geqq 2i_{l-\gamma_+}+2-n=k_l^0+2.$$

Thus $k_1^0<\cdots<k^0_{\beta_++1}$, and similarly we have $k^0_{\beta_++1}<\cdots<k_p$. This advances the induction and establishes the sufficiency of (3.10). The necessity of (3.10) follows easily as in case (b), and so the proof of Theorem 3.1 is complete.

Remark 3.1. Note that, in general, part (d) of Theorem 3.1 is not true without a restriction on p. For example, take $\alpha_+=\delta_+=\beta_+=0$, $\gamma_+=1$, $p=2$, $(i_1,i_2)=(0,1)$, $(j_1,j_2)=(-1,0)$. It is easily checked that for $r=1$, (3.30) is satisfied, but (3.32) implies $k_1^0=-\frac{1}{2}$, which is impossible. Thus

$$D^2\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}=0,$$

but for $r=2$ (3.10) is satisfied.

Before we formulate the main result of this section, we need some further facts about the function φ.

LEMMA 3.2. *Let φ be as in Theorem* 2.1. *Then*

$$\lim_{x\to\pm\infty}\varphi(x)=0 \tag{3.36}$$

and

$$\varphi(x)>0, \qquad x\in I^0, \tag{3.37}$$

where I is the smallest closed interval containing $\operatorname{supp}\mathbf{a}=\{k\colon k\in\mathbb{Z},\ a_k>0\}$.

Proof. We first prove (3.36). According to (2.3),

$$\lim_{j\to\pm\infty}\varphi(j)=0, \tag{3.38}$$

and since φ is a ripplet we also have

$$0\leqq\varphi(x)\leqq 1, \qquad x\in\mathbb{R}.$$

Suppose, also, that k is an integer such that $\varphi(k)\neq 0$. Then for $x>j+k$, $j\in\mathbb{Z}_+$ we have

$$0\leqq\Phi\begin{pmatrix} j+k & x\\ 0 & j\end{pmatrix}=\begin{vmatrix}\varphi(j+k-0) & \varphi(j+k-j)\\ \varphi(x-0) & \varphi(x-j)\end{vmatrix},$$

and hence

$$\varphi(x)\leqq\frac{\varphi(j+k)}{\varphi(k)}\varphi(x-j)\leqq\frac{\varphi(j+k)}{\varphi(k)}.$$

Thus from (3.38) we conclude that $\lim_{x\to\infty}\varphi(x)=0$. Similarly, by considering the determinant

$$\Phi\begin{pmatrix} x & j\\ j-k & 0\end{pmatrix}$$

we conclude that $\lim_{x\to-\infty}\varphi(x)=0$.

As for (3.37) we will recall that the integers for which $a_k>0$ are consecutive. Therefore, $I^0=(k_-, k_+)$ and $a_k>0$ if and only if $k=k_-,\cdots,k_+$ for some integers k_-, k_+, either of which may be infinite.

The proof of (3.37) is conveniently described by several cases. When I is a finite interval, the result appears in [12]. Here we modify the argument. First we show φ is positive on some closed interval of length one. For instance, when I is finite

$$\sum_{k_-}^{k^+}\varphi(x-j)=\sum_{j\in\mathbb{Z}}\varphi(x-j), \qquad x\in[k_-+k_+-1,\ k_-+k_++1],$$

and so the refinement equations (1.1) and (2.3) give

$$\varphi\left(\frac{x}{2}\right)\geqq\min\{a_j\colon k_-\leqq j\leqq k_+\}>0,$$

for $x\in[k_-+k_+-1,\ k_-+k_++1]$. When I is doubly infinite, we consider the sequence of functions

$$\Psi_N(x):=\sum_{-N}^{N}\varphi(x-j).$$

Clearly, Ψ_N is a nondecreasing sequence of nonnegative functions which converge pointwise to one, in view of (2.3). Hence, by Dini's theorem the convergence is uniform on any compact interval. Consequently, there is an integer p such that $\Psi_p(x)>0$ for $x\in[-1,1]$. Again the refinement equation gives

$$\varphi\left(\frac{x}{2}\right)\geqq\sum_{-p}^{p}a_j\varphi(x-j)\geqq\min\{a_j\colon |j|\leqq p\}\Psi_p(x)>0.$$

Thus, in both cases we have established the existence of a closed interval J_0 of length one on which φ is positive. Now inductively we suppose φ is positive on a closed interval J_r of length at least one. Choose any y in the interval

$$J_{r+1} := \bigcup_{j \in \operatorname{supp} \mathbf{a}} (j + J_r)/2 = \tfrac{1}{2} J_r + \tfrac{1}{2}(k_-, k_+)$$

so that $y = (x+k)/2$ for some $x \in J_r$ and $k \in \operatorname{supp} \mathbf{a}$. Then according to the refinement equation

$$\varphi(y) = \sum_{j \in \mathbb{Z}} a_j \varphi(2y - j) \geqq a_k \varphi(2y - k) = a_k \varphi(x) > 0,$$

that is, φ is positive on J_{r+1}. When $I^0 = (-\infty, \infty)$ we see that $J_1 = (-\infty, \infty)$, while in the case that $I^0 = (k_-, k_+)$ is finite, we observe that $\lim_{r \to \infty} J_r = (k_-, k_+)$, and so again in both cases we conclude φ is positive on I.

For the cases when $k_- > -\infty$, $k_+ = \infty$ or $k_- = -\infty$, $k_+ < -\infty$ we argue differently. In the first instance (the argument in the other case is the same and we omit it) we pick any $x_0 > k_-$ such that $\varphi(x_0) > 0$. Then, as above, $\varphi(y_l) > 0$, $l = k_-, k_- + 1, k_- + 2, \cdots$, where $y_l := \frac{1}{2}(x_0 + l)$ because by the refinement equation

$$\varphi(y_l) \geqq a_l \varphi(2y_l - l) = a_l \varphi(x_0) > 0.$$

Similarly, we see that φ is positive on the sequence $x_l := 2^{-l} x_0 + (1 - 2^{-l}) k_-$, $l = 0, 1, 2, \cdots$, since $x_{l+1} = \frac{1}{2}(x_l + k_-)$. Now, choose any $x \in (k_-, \infty)$. There exists an $i \in \mathbb{Z}_+$ such that

$$x_i < x < y_i.$$

Pick any k, j such that $k < 0 < j$, then

$$0 \leqq \begin{vmatrix} \varphi(x-0) & \varphi(x-j) \\ \varphi(y_i - 0) & \varphi(y_i - j) \end{vmatrix} \tag{3.39}$$

and

$$0 \leqq \begin{vmatrix} \varphi(x_i - k) & \varphi(x_i - 0) \\ \varphi(x-k) & \varphi(x-0) \end{vmatrix}. \tag{3.40}$$

We want to show $\varphi(x) > 0$. Suppose to the contrary $\varphi(x) = 0$, then from (3.39) we get $\varphi(x - j) = 0$, $j = 0, 1, 2, \cdots$, while (3.40) gives us $\varphi(x - k) = 0$, $k = -1, -2, \cdots$. But this contradicts (2.3), which says that

$$\sum_{j \in \mathbb{Z}} \varphi(x - j) = 1.$$

This establishes that $\varphi(x) > 0$ and proves the lemma.

We are now ready to state and prove the main result in this section.

THEOREM 3.2. *Let* $\mathbf{a} = \{a_j : j \in \mathbb{Z}\}$ *be a Pólya sequence whose symbol satisfies the hypothesis of Theorem* 2.1 *and has the factorization* (2.6). *Suppose that if* $s = t = 0$ *and* $\boldsymbol{\beta}_+ + \boldsymbol{\gamma}_+ > 0$, then either $p \leqq \boldsymbol{\beta}_+ + \boldsymbol{\gamma}_+$ *or* $p \leqq \boldsymbol{\alpha}_+ + \boldsymbol{\delta}_+ + \max(\boldsymbol{\beta}_+ - 1, 0) + \max(\boldsymbol{\gamma}_+ - 1, 0)$. *Then the determinant* (1.10) *is positive if and only if* $x_l - i_l \in I^0$, $l = 1, \cdots, p$. *Equivalently, the determinant* (1.10) *is positive if and only if the diagonal elements of the matrix are all positive.*

Proof. We begin the proof of this theorem by relating the parameters of the factorization (2.6) to the smallest closed interval I containing $\{k \colon k \in \mathbb{Z}, a_k > 0\}$. As before we let

$$\operatorname{supp} \mathbf{a} = \{k \colon k \in \mathbb{Z}, k_- \leqq k \leqq k_+\},$$

where k_-, k_+ are integers which may be $-\infty$, ∞, respectively. It follows directly from (2.6) that

$$k_- = \begin{cases} -\boldsymbol{\delta}_+ & \text{if } t = \boldsymbol{\gamma}_+ = 0 \\ -\infty & \text{otherwise} \end{cases} \tag{3.41}$$

and

$$k_+ = \begin{cases} \boldsymbol{\alpha}_+, & s = \boldsymbol{\beta}_+ = 0 \\ \infty & \text{otherwise.} \end{cases} \tag{3.42}$$

Next we use the refinement equation (1.1) to get

$$\begin{aligned} \varphi(x-i) &= \sum_{j\in\mathbb{Z}} a_j\varphi(2x-2i-j) = \sum_{j\in\mathbb{Z}} D_{ij}\varphi(2x-j) \\ &= \cdots = \sum_{j\in\mathbb{Z}} D^r_{ij}\varphi(2^rx-j), \qquad i\in\mathbb{Z}. \end{aligned}$$

Therefore, we obtain by the Cauchy-Binet formula

$$\Phi\begin{pmatrix} x_1, \cdots, x_p \\ i_1, \cdots, i_p \end{pmatrix} = \sum_{j_1<\cdots<j_p} D^r\begin{pmatrix} i_1, \cdots, i_p \\ j_1, \cdots, j_p \end{pmatrix} \Phi\begin{pmatrix} 2^rx_1, \cdots, 2^rx_p \\ j_1, \cdots, j_p \end{pmatrix}. \tag{3.43}$$

First, we consider the sufficiency of the conditions $x_l - i_l \in I^0$, $l = 1, \cdots, p$. The idea of the proof is to choose r sufficiently large as to approximate each x_l by a dyadic rational $2^{-r}j_{l,r}$ in such a way that the summand on the right-hand side of (3.43) corresponding to $j_{1,r} < \cdots < j_{p,r}$ is nonzero.

It is important to first note that if $2^{-r}j_{l,r}$ approximates x_l as $r\to\infty$ then the off diagonal terms of the determinant

$$\Phi\begin{pmatrix} 2^rx_1, \cdots, 2^rx_p \\ j_{1,r}, \cdots, j_{p,r} \end{pmatrix} \tag{3.44}$$

tend to zero because of (3.36) of Lemma 3.2. To insure that the diagonal terms appearing in the determinant (3.44) are positive, we pick any integer ρ such that $[\rho, \rho+1] \subseteq I$ and choose $j_{l,r} := -\rho + [2^rx_l]$. Then we see that

$$\lim_{r\to\infty} 2^{-r}j_{l,r} = x_l, \qquad l = 1, \cdots, p \tag{3.45}$$

as well as $2^rx_l - j_{l,r} \in [\rho, \rho+1] \subseteq I$, and so the diagonal terms of the determinant (3.44) are positive by (3.37) of Lemma 3.2.

To confirm that the determinant

$$D^r\begin{pmatrix} i_1, \cdots, i_p \\ j_{1,r}, \cdots, j_{p,r} \end{pmatrix} \tag{3.46}$$

is positive, we will use Theorem 3.1. It is best to verify the inequalities in four distinct cases.

Case A. The first case we consider is $k_- = -\infty$ and $k_+ = \infty$. According to (3.41) and (3.42) this occurs if one of the following four situations hold:

(i)$_1$ $\boldsymbol{\delta}_+ = \infty$, $t = 0$, $\boldsymbol{\gamma}_+ = 0$, $s > 0$ or $\boldsymbol{\beta}_+ > 0$.

(ii)$_1$ $\boldsymbol{\delta}_+ = \infty$, $t = 0$, $\boldsymbol{\gamma}_+ = 0$, $s = 0$, $\boldsymbol{\beta}_+ = 0$, and $\boldsymbol{\alpha}_+ = \infty$.

(iii)$_1$ $t > 0$ or $\boldsymbol{\gamma}_+ > 0$ and $s = 0$, $\boldsymbol{\beta}_+ = 0$, and $\boldsymbol{\alpha}_+ = \infty$.

(iv)$_1$ $t > 0$ or $\boldsymbol{\gamma}_+ > 0$ and $s > 0$ or $\boldsymbol{\beta}_+ > 0$.

In each of these cases we claim that for *r sufficiently large* the determinant (3.46) is positive. In the case (i)$_1$ we see if $s > 0$ then (3.8) implies (3.46) is positive for all r

while $s=0$ (3.10) requires only a lower bound for i_l which for r large ceases to constrain i_l as well. The other cases follow similarly.

Case B. Now, for $k_- > -\infty$ and $k_+ = \infty$ we must have one of the following:

(i)$_2$ $\boldsymbol{\delta}_+ < \infty$, $t=0$, $\boldsymbol{\gamma}_+ = 0$, $\boldsymbol{\alpha}_+ = \infty$, $s=0$, and $\boldsymbol{\beta}_+ = 0$.

(ii)$_2$ $\boldsymbol{\delta}_+ < \infty$, $t=0$, $\boldsymbol{\gamma}_+ = 0$, and $s>0$ or $\boldsymbol{\beta}_+ > 0$.

When (i)$_2$ holds, (3.10) reduces to

$$i_l \leqq 2^{-r} j_{l,r} + (1-2^{-r})\boldsymbol{\delta}_+. \tag{3.47}$$

However, by hypothesis, $0 < x_l - i_l + \boldsymbol{\delta}_+$ and so (3.47) is valid for r sufficiently large. Case (ii)$_2$ follows similarly.

Case C. When $k_- = -\infty$ and $k_+ < \infty$, which occurs if one of the following holds:

(i)$_3$ $\boldsymbol{\delta}_+ = \infty$, $t=0$, $\boldsymbol{\gamma}_+ = 0$, $\boldsymbol{\alpha}_+ < \infty$, $s=0$, and $\boldsymbol{\beta}_+ = 0$.

(ii)$_3$ $t>0$ or $\boldsymbol{\gamma}_+ > 0$ and $\boldsymbol{\alpha}_+ < \infty$, $s=0$, and $\boldsymbol{\beta}_+ = 0$.

The proof is the same.

Case D. The final case $k_- < \infty$ and $k_+ < \infty$ can occur if and only if

(i)$_4$ $\boldsymbol{\delta}_+ < \infty$, $\boldsymbol{\alpha}_+ < \infty$, $t=0$, $s=0$, $\boldsymbol{\gamma}_+ = 0$ and $\boldsymbol{\beta}_+ = 0$.

In this case inequalities (d) of Theorem 3.1 become

$$2^{-r} j_{l,r} - (1-2^{-r})\boldsymbol{\alpha}_+ \leqq i_l \leqq 2^{-r} j_{l,r} + (1-2^{-r})\boldsymbol{\delta}_+, \qquad l = 1, \cdots, r,$$

which is valid for r sufficiently large, since $-\boldsymbol{\delta}_+ < x_l - i_l < \boldsymbol{\alpha}_+$, $l = 1, 2, \cdots, p$ by hypothesis. This establishes the sufficiency of the condition that $x_l - i_l \in I^0$, $l = 1, \cdots, p$ for the positivity of lower determinants (3.43).

For the necessity, we again consider the four cases above. In Case A there is nothing to prove. In Case B if $x_\mu - i_\mu \leqq k_- = -\boldsymbol{\delta}_+$ for some μ, $1 \leqq \mu \leqq p$. Hence $\varphi(x_i - i_j) = 0$ for $i = 1, \cdots, \mu$, $j = \mu, \cdots, p$ and so the first rows of the determinant are linearly dependent. Case C is similar and we omit the details. For the last, Case D, we know $\varphi(x) = 0$ if either $x \leqq k_-$ or $x \geqq k_+$. Thus the proof here uses both the argument used in Case B and in Case C. Thus we have established the theorem.

The next result proves the conjecture made in [12].

COROLLARY 3.1. *Given* $\mathbf{a} = \{a_j : 0 \leqq j \leqq n\}$ *such that the polynomial*

$$a(z) = \sum_{j=0}^{n} a_j z^j$$

only has negative zeros, vanishes at $z = -1$, *and has the value* 2 *at* $z = 1$, *then there exists a unique solution of the refinement equation*

$$\varphi\left(\frac{x}{2}\right) = \sum_{j=0}^{n} a_j \varphi(x-j), \qquad x \in \mathbb{R}$$

such that

$$\sum_{j \in \mathbb{Z}} \varphi(x-j) = 1.$$

Moreover, the determinants

$$\Phi\begin{pmatrix} x_1, \cdots, x_p \\ i_1, \cdots, i_p \end{pmatrix} := \det_{l,j=1,\cdots,p} \varphi(x_l - i_j) \tag{3.48}$$

are nonnegative for all $x_1 < \cdots < x_p$ *and integers* $i_1 < \cdots < i_p$ *with strict positivity holding if and only if*

$$i_l < x_l < i_l + n, \qquad l = 1, \cdots, p. \tag{3.49}$$

4. Symbols which are left-half plane stable. The previous result actually holds under a much weaker hypothesis. Recall that a polynomial is called left-half plane stable if all its zeros are in the (open) left-half plane (Hurwitz polynomial).

THEOREM 4.1. *Suppose that the polynomial*

$$a(z)=\sum_{j=0}^{n} a_j z^j,$$

with real coefficients a_j, if left-half plane stable, vanishes at $z=-1$, and has the value two at $z=1$. Then all the conclusions of Corollary 3.1 *are valid.*

Proof. The proof of this result follows the pattern of Corollary 3.1 with some important differences. The existence of the function φ follows from Theorem 2.1 on noting that $a_j>0, j=0, 1, \cdots, n$ since $a(z)$ can be factored as a product of linear and quadratic factors which have positive coefficients. Hence, when we write $a(z)=(1+z)q(z)$, the polynomial $q(z)$ is also left-half plane stable and so has positive coefficients. Thus (2.2) is satisfied and Theorem 2.1 can be applied.

The main fact needed to complete the proof is a result of Kemperman [10]. We write $a(z)$ in the form

$$a(z)=d_0z^n+d_1z^{n-1}+\cdots+d_n, \qquad d_j=a_{n-j}$$

to conform with the notation of [10]. Asner [2] and Kemperman [10] proved that the Hurwitz matrix

$$H=(H_{ij}\colon i,j\in\mathbb{Z}), \qquad H_{ij}:=d_{2j-i}$$

is totally positive. In addition, Kemperman proved the following.

THEOREM C. *Let $a(z)$ be left-half plane stable. Then*

$$H\begin{pmatrix} i_1, \cdots, i_p \\ j_1, \cdots, j_p \end{pmatrix} \geqq 0$$

for any integers $i_1<\cdots<i_p$ and $j_1<\cdots<j_p$, and equality holds if and only if all the diagonal elements $d_{2j_r-i_r}$, $r=1, \cdots, p$ are positive, equivalently, that

$$0\leqq 2j_l-i_l\leqq n, \qquad l=1, \cdots, p.$$

In our notation,

$$D_{ij}=a_{j-2i}=d_{n+2i-j}=d_{2(n+i)-(n+j)}=H_{n+j,\,n+i}.$$

Thus, Kemperman's result is equivalent to saying that

$$D\begin{pmatrix} i_1, \cdots, i_p \\ j_1, \cdots, j_p \end{pmatrix} = H\begin{pmatrix} n+j_1, \cdots, n+j_p \\ n+i_1, \cdots, n+i_p \end{pmatrix} \geqq 0$$

with equality if and only if

$$0\leqq j_l-2i_l\geqq n, \qquad l=1, \cdots, p.$$

From the total positivity of D it follows, by the arguments in Theorem 2.1, that

$$\Phi\begin{pmatrix} x_1, \cdots, x_p \\ i_1, \cdots, i_p \end{pmatrix} \geqq 0 \tag{4.1}$$

for $x_1<\cdots<x_p$, $i_1<\cdots<i_p$. The methods used in the proof of Theorem 3.1, especially the case $s=t=0$, $\boldsymbol{\delta}_+=\boldsymbol{\beta}_+=\boldsymbol{\gamma}_+=0$, and $\boldsymbol{\alpha}_+=n$ implies

$$D^r\begin{pmatrix} i_1, \cdots, i_p \\ j_1, \cdots, j_p \end{pmatrix} > 0$$

if and only if $0 \leqq 2^{-r}j_l - i_l \leqq (1-2^{-r})n$, $l=1, \cdots, p$, whenever $a(z)$ is left-half plane stable. Therefore formula (3.43) can be used just as before to conclude that the determinants (3.48) are positive if and only if (3.49) holds.

As a final result we derive a fact which has applications for the study of the *planar curve*:

$$S(x) = \sum_{j \in \mathbb{Z}} \mathbf{c}_j \varphi(x-j), \qquad x \in \mathbb{R}, \tag{4.2}$$

where each $\mathbf{c}_j$, $j \in \mathbb{Z}$ is a vector in $\mathbb{R}^2$.

THEOREM 4.2. *Suppose*

$$a(z) = \sum_{j=0}^{n} a_j z^j$$

is left-half plane stable. Then the function φ of Theorem 4.1 *is $C^k(\mathbb{R})$, $0 \leqq k \leqq n-1$, if and only if $a(z)$ can be factored as*

$$a(z) = (1+z)^{k+1} q(z), \qquad q(1) = 2^{-k}. \tag{4.3}$$

Moreover, in this case

$$S^-\left(\sum_{j \in \mathbb{Z}} c_j \varphi^{(l)}(\bullet - j) \right) \leqq S^-(\Delta^l c), \qquad 0 \leqq l \leqq k, \tag{4.4}$$

where $c = \{c_j; j \in \mathbb{Z}\}$ is a sequence in $\mathbb{R}$ and ∇^l is the lth order forward difference operator defined inductively as

$$(\Delta^l c)_j = (\Delta^{l-1} c)_{j+1} - (\Delta^{l-1} c)_j, \qquad j \in \mathbb{Z}. \tag{4.5}$$

Proof. The necessity of the factorization (4.3) is a consequence of two results from [3, Cor. 6.3, 8.2] concerning a sequence a of finite support, which we state here for the convenience of the reader.

PROPOSITION A. *There exists a polynomial p of degree $\leqq k$ such that*

$$\sum_{j \in \mathbb{Z}} a_{i-2j} j^k = p(i), \qquad i \in \mathbb{Z} \tag{4.6}$$

if and only if $a^{(l)}(-1) = 0$, $l = 0, 1, \cdots, k$.

PROPOSITION B. *Let φ satisfy the refinement equation*

$$\varphi(x) = \sum_{j \in \mathbb{Z}} a_j \varphi(2x-j).$$

If $\varphi \in C^k(\mathbb{R})$ and the functions $\{\varphi(\bullet - j) : j \in \mathbb{Z}\}$ are linearly independent on $\mathbb{R}$ then (4.6) *holds for some polynomial p of degree $\leqq k$.*

To use Propositions A and B we note that Theorem 4.1 implies that the functions $\{\varphi(\bullet - j) : j \in \mathbb{Z}\}$ are linearly independent, and hence it follows that $a(z)$ can be factored as (4.3).

The converse also follows from a result from [3, Cor. 8.1]. However, let us point out that we already have enough information available in Theorem 4.1 to prove it directly.

Pick any l, $1 \leqq l \leqq k$. Then we may express $a(z)$ as

$$a(z) = 2^{-l}(1+z)^l b(z), \tag{4.7}$$

where $b(z)$ is left-half plane stable, $b(-1)=0$ and $b(1)=2$. Hence, by Theorem 4.1 there is a $\Psi \in C(\mathbb{R})$, of support in $(0, n-l)$ such that

$$\Psi(x)=\sum_{j=0}^{n-l} b_j\Psi(2x-j), \tag{4.8}$$

$$\sum_{j\in\mathbb{Z}} \Psi(x-j)=1, \qquad x\in\mathbb{R}, \tag{4.9}$$

and

$$S^-\left(\sum_{j\in\mathbb{Z}} d_j\Psi(\bullet-j)\right)\leqq S^-(d), \tag{4.10}$$

where $d=\{d_j: j\in\mathbb{Z}\}$. From (4.9) we conclude, by integrating both sides of (4.9) from zero to one, that

$$\int_{\mathbb{R}} \Psi(x)\,dx=1. \tag{4.11}$$

Using the refinement equation (4.8) for Ψ and the refinement equations (1.7), (1.8) (when $n=l$) for M_l, and the factorization (4.7), we conclude that the convolution $M_l * \Psi$ satisfies the refinement equation (1.1). This can be seen by substituting into the integral

$$(M_l * \Psi)(x)=\int_{\mathbb{R}} M_l(x-t)\Psi(t)\,dt,$$

the refinement equations for M_l, Ψ, and then by simplifying. By (4.11) we also have

$$\sum_{j\in\mathbb{Z}} (M_l * \Psi)(x-j)=1, \qquad x\in\mathbb{R},$$

and we conclude (by uniqueness) that $\varphi=M_l*\Psi$ is the function of Theorem 4.1.

By construction the B-spline satisfies the recurrence relation

$$M_l(x)=\int_0^1 M_{l-1}(x-t)\,dt$$

or equivalently $M_l'(x)=M_{l-1}(x)-M_{l-1}(x-1)$. Consequently,

$$\varphi'=M_{l-1}*(\Psi(\bullet)-\Psi(\bullet-1))$$

and therefore inductively we obtain

$$\varphi^{(l)}(x)=(\nabla^l\Psi(x-\bullet))_0,$$

where ∇ is the difference operator, $(\nabla c)_j := c_j - c_{j+1}$.

Hence it follows that

$$\sum_{j\in\mathbb{Z}} c_j\varphi^{(l)}(x-j)=\sum_{j\in\mathbb{Z}} (\Delta^l c)_j\Psi(x-j), \qquad x\in\mathbb{R},$$

and so from (4.10) we get

$$S^-\left(\sum_{j\in\mathbb{Z}} c_j\varphi^{(l)}(\bullet-j)\right)\leqq S^-(\Delta^l c),$$

which proves the theorem.

Acknowledgement. We wish to thank Allan Pinkus for bringing the paper by J. H. B. Kemperman to our attention.

REFERENCES

[1] M. Aissen, A. Edrei, I. J. Schoenberg, and A. Whitney, *On the generating functions of totally positive sequences*, Proc. Nat. Acad. Sci. U.S.A., 37, pp. 303-307.

[2] B. A. Asner, Jr., *On the total nonnegativity of Hurwitz matrix*, SIAM J. Appl. Math., 18 (1970), pp. 407-414.

[3] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, IBM Res. Report, 1989, in Mem. Amer. Math. Soc., to appear.

[4] W. Dahmen and C. A. Micchelli, *Subdivision algorithms for the generation of box spline surfaces*, Comput. Aided Geom. Design, 1 (1984), pp. 115-129.

[5] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909-996.

[6] A. Edelman, N. Dyn, and C. A. Micchelli, *On locally supported basis functions for the representation of geometrically continuous curves*, Analysis, 7 (1987), pp. 313-341.

[7] A. Edrei, *On the generating function of a doubly infinite totally positive sequence*, Trans. Amer. Math. Soc., 74 (1952), pp. 367-383.

[8] ———, *On the generating function of totally positive matrices* II, J. Analyse Math., 2 (1953), pp. 86-94.

[9] S. Karlin, *Total Positivity*, Stanford University Press, Stanford, CA, 1968.

[10] J. H. B. Kemperman, *A Hurwitz matrix is totally positive*, SIAM J. Math. Anal., 13 (1982), pp. 331-341.

[11] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69-87.

[12] C. A. Micchelli and A. Pinkus, *Descartes systems from corner cutting*, Constr. Approx., 7 (1991), pp. 161-194.

[13] C. A. Micchelli, *Cardinal L-splines*, in Studies in Spline Functions and Approximation Theory, S. Karlin, C. A. Micchelli, A. Pinkus, and I. J. Schoenberg, eds., Academic Press, New York, 1976, pp. 163-189.

[14] I. J. Schoenberg, *Cardinal Interpolation*, CBMS-NSF Regional Conf. Ser. in Appl. Math., 12, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1973.

WEIGHTED FOURIER TRANSFORM INEQUALITIES FOR RADIALLY DECREASING FUNCTIONS*

C. CARTON-LEBRUN† AND H. P. HEINIG‡

Abstract. Weighted norm inequalities are established for the Fourier transform of certain radially decreasing functions. For large classes of weights and indices, the involved conditions are proved to be necessary. Similar characterizations are also given for the Hankel- and K-transform of functions satisfying monotonicity conditions.

Key words. radially decreasing functions, Fourier transform, B_p-weights, Hankel transform, K-transformation

AMS(MOS) subject classifications. 42A38, 42B10

1. Introduction. Let $\hat{f}$ be the Fourier transform of f defined by

$$\hat{f}(x)=\int_{\mathbb{R}^n} e^{-2\pi i x\cdot y} f(y)\,dy, \qquad x\in\mathbb{R}^n,$$

provided the integral converges. Weighted L^p-estimates of the form

$$\|\hat{f}\|_{q,u}\leqq C\|f\|_{p,v}, \qquad 0<p,\, q<\infty, \tag{1.1}$$

where u and v are positive weight functions, have been studied in recent years with a view to characterize the weights for which (1.1) is satisfied. While much progress has been made in this direction, the complete solution is still elusive. (For a discussion of these and related questions we refer to [2], [3] and the recent work of Strömberg and Wheeden [13].)

The object of this note is to consider the Fourier transform of certain radially decreasing functions and prove weighted norm inequalities of the form (1.1). For $n\geqq 1$, $u=v$, and $q=p\geqq 1$, a complete characterization of the weight is given for which the inequality holds. Similar characterizations are given in the case of the Hankel transform which may be viewed as a generalization of the n-dimensional Fourier transform result (cf. Remark 4.2). We also consider generalizations of the Laplace transform (the K-transform) of functions satisfying certain monotonicity conditions, and characterize the weights u and v for which similar (L_v^p, L_u^q) norm estimates hold, $1<p\leqq q<\infty$. This result extends work given in [8].

Our result contrasts with previous studies, where typically monotonicity conditions on the weights are imposed to provide the characterizations (cf. [2], [5], [9]).

The plan of the paper is as follows. The next section contains the one weight characterizations involving the Fourier transform, while § 3 considers the two weighted Fourier inequalities in the index ranges $1<p\leqq q<\infty, 0<q<p<\infty$ with $p>1$ and $0<p<1<q$. If $1<p\leqq q<\infty$, then under an auxiliary condition on the range weight (which is satisfied by power weights), the results are also necessary. The final section contains the characterizations involving the Hankel- and K-transformations.

* Received by the editors October 17, 1990; accepted for publication (in revised form) September 16, 1991. This research was supported by FNRS of Belgium and Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics, University of Mons, Mons, Belgium.

‡ Department of Mathematics and Statistics, McMaster University, Hamilton, Canada.

The notation and conventions used here are as follows: If $x, y \in \mathbb{R}^n$, then $x \cdot y = \sum_{i=1}^n x_i y_i$ and $|x| = (\sum_{i=1}^n x_i^2)^{1/2}$. If $x \neq 0$, we write $x = |x|x'$ where $x' \in S_{n-1}$, the unit sphere in $\mathbb{R}^n$. We also use $|S_{n-1}| = \int_{S_{n-1}} d\sigma$, with the convention $|S_0| = 2$. For $0 < p < \infty$, the conjugate index p' is defined by $1/p + 1/p' = 1$ with $p' = \infty$ if $p = 1$, and similarly for other indices. Of course $p' < 0$ if $0 < p < 1$. χ_E denotes the characteristic function of the set E and inequalities are interpreted in the sense that if the right side is finite, so is the left side and the inequality holds. A, B, C denote constants (sometimes with subscripts) which may be different at different places. Finally, we adhere to the convention that positive means nonnegative and decreasing means nonincreasing.

2. One weight characterizations. We begin with the weight characterization for the Fourier transform of positive even functions decreasing to zero on $(0, \infty)$.

THEOREM 2.1. *Let v be a positive weight function and $1 \leqq p < \infty$. Then*

$$\int_0^\infty v(1/x)x^{p-2}|\hat{f}(x)|^p\,dx \leqq C \int_0^\infty v(x)f(x)^p\,dx \tag{2.1}$$

holds for all positive even functions f decreasing to zero on $(0, \infty)$, if and only if $v \in B_p$, i.e., for each $r > 0$,

$$\int_r^\infty x^{-p}v(x)\,dx \leqq Ar^{-p}\int_0^r v(x)\,dx. \tag{2.2}$$

Proof. Since f is even, positive, and decreasing on $(0, \infty)$,

$$\begin{aligned}\hat{f}(x) &= 2\int_0^\infty \cos(2\pi xt)f(t)\,dt \\ &= 2\left[\int_0^{1/x} \cos(2\pi xt)f(t)\,dt + f(1/x)\int_{1/x}^{\xi} \cos(2\pi xt)\,dt\right].\end{aligned}$$

Here we applied the second mean value theorem. But since

$$x^{-1}f(1/x) \leqq \int_0^{1/x} f(t)\,dt,$$

it follows that

$$|\hat{f}(x)| \leqq (2 + 1/\pi)\int_0^{1/x} f(t)\,dt,$$

and therefore

$$\int_0^\infty v(x)|x^{-1}\hat{f}(1/x)|^p\,dx \leqq C\int_0^\infty v(x)\left[x^{-1}\int_0^x f(t)\,dt\right]^p dx.$$

A result of Ariño and Muckenhoupt [1, Thm. 1.7] shows that the integral on the right is dominated by the right side of (2.1) provided $v \in B_p$. This proves the sufficiency.

To prove necessity, let $f(x) = \chi_{(0,r)}(|x|)$, $r > 0$, fixed in (2.1). Then

$$\begin{aligned}C\int_0^r v(x)\,dx &\geqq \int_0^\infty v(1/x)x^{p-2}\left|2\int_0^r \cos(2\pi xt)\,dt\right|^p dx \\ &= 2^p\int_0^\infty v(x)x^{-p}\left|\int_0^r \cos(2\pi t/x)\,dt\right|^p dx \\ &\geqq 2^p\int_{2\pi r}^\infty v(x)x^{-p}\left|\int_0^r \cos(2\pi t/x)\,dt\right|^p dx.\end{aligned}$$

But since $(2\pi t/x)\leqq 1$, it follows that $\cos(2\pi t/x)\geqq \cos 1$ and hence

$$C\int_0^r v(x)\,dx \geqq (2\cos 1)^p \int_{2\pi r}^{\infty} v(x)\cdot (r/x)^p\,dx.$$

This implies (2.2).

The condition $v\in B_p$ (i.e., (2.2)) of Theorem 2.1 is equivalent to the L_v^p-boundedness of the averaging operator of positive decreasing functions [1, Thm. 1.7]. It should be noted that this L^p-boundedness property was shown by Boyd [4] (see also [10, Chap. 2, Thm. 6.6]) to be equivalent with the weight condition

$$\sup_{s>0}\frac{V(rs)}{V(s)} = o(r^p) \quad \text{as } r\to\infty, \tag{2.3}$$

where $V(s)=\int_0^s v(x)\,dx$. Therefore, we obtain the following.

COROLLARY 2.2. *Inequality* (2.1) *holds for all positive even functions decreasing to zero on* $(0,\infty)$ *if and only if* (2.3) *holds.*

Observe that with $v(x)=x^{p-2}, p>1$, Theorem 2.1 and Corollary 2.2 reduce to Theorem 82 of [14].

We now give a weighted n-dimensional characterization for the Fourier transform of certain radially decreasing functions on $\mathbb{R}^n$, $n>1$.

THEOREM 2.3. *Let* $1\leqq p<\infty$ *and* v *be a positive weight function on* $\mathbb{R}^n$. *Then*

$$\int_{\mathbb{R}^n} v(x'/|x|)|x|^{n(p-2)}|\hat{f}(x)|^p\,dx \leqq C\int_{\mathbb{R}^n} v(x)|f(x)|^p\,dx \tag{2.4}$$

holds for all positive radial functions $f(x)=f_0(|x|)$, $x\in\mathbb{R}^n$, *with* $g(t)=t^{n-1}f_0(t)$ *decreasing to zero on* $(0,\infty)$, *if and only if* $v\in B_p^*(n)$, *i.e., for each* $r>0$,

$$\int_{|x|\geqq r} |x|^{-np}v(x)\,dx \leqq Br^{-p}\int_{|x|\leqq r}|x|^{-p(n-1)}v(x)\,dx. \tag{2.5}$$

Proof (*Sufficiency*). Clearly

$$\begin{aligned}
\hat{f}(x) &= \int_{\mathbb{R}^n} f_0(|t|)\,e^{-2\pi i x\cdot t}\,dt\\
&= \int_0^\infty \rho^{n-1}\left(\int_{S_{n-1}} f_0(\rho)\,e^{-2\pi i\rho|x|(x'\cdot\omega)}\,d\sigma\right)d\rho\\
&= 2|S_{n-2}|\int_0^\infty g(\rho)\int_0^1 \cos(2\pi\rho|x|t)(1-t^2)^{(n-3)/2}\,dt\,d\rho\\
&= 2|S_{n-2}|\left(\int_0^{1/|x|}+\int_{1/|x|}^{\infty}\right)g(\rho)\int_0^1 \cos(2\pi\rho|x|t)(1-t^2)^{(n-3)/2}\,dt\,d\rho\\
&= 2|S_{n-2}|(A_1(|x|)+A_2(|x|)), \quad \text{respectively.}
\end{aligned}$$

To estimate $A_1(s)$, $s>0$, let $\psi_n(t)=\int_0^t(1-u^2)^{(n-3)/2}\,du$. Then

$$\begin{aligned}
&\int_0^1 \cos(2\pi\rho st)(1-t^2)^{(n-3)/2}\,dt\\
&\qquad = \cos(2\pi\rho s)\psi_n(1)+2\pi\rho s\int_0^1 \sin(2\pi\rho st)\psi_n(t)\,dt.
\end{aligned}$$

Therefore

$$|A_1(s)| = \left|\psi_n(1)\int_0^{1/s} g(\rho)\cos(2\pi\rho s)\,d\rho + 2\pi s\int_0^{1/s}\rho g(\rho)\int_0^1 \sin(2\pi\rho st)\psi_n(t)\,dt\,d\rho\right|$$

$$\leqq C\int_0^{1/s} g(\rho)\,d\rho.$$

On the other hand,

$$A_2(s) = \int_{1/s}^{\infty} g(\rho)\varphi_n(s\rho)\,d\rho,$$

where

$$\varphi_n(s\rho) = \int_0^1 \cos(2\pi\rho st)\cdot(1-t^2)^{(n-3)/2}\,dt.$$

By the second mean value theorem,

$$A_2(s) = g(1/s)\int_{1/s}^{\xi}\varphi_n(s\rho)\,d\rho$$

for some $\xi > 1/s$. Therefore,

$$|A_2(s)| \leqq Cs^{-1}g(1/s) \leqq C\int_0^{1/s} g(t)\,dt \tag{2.6}$$

provided we show that

$$\left|\int_{1/s}^{\xi}\varphi_n(s\rho)\,d\rho\right| \leqq C/s \tag{2.7}$$

for all $s>0$ and C independent of ξ and s.

To prove (2.7), note that, for $n \geqq 2$,

$$\int_{1/s}^{\xi}\varphi_n(s\rho)\,d\rho = \int_0^1 (1-t^2)^{(n-3)/2}\left(\int_{1/s}^{\xi}\cos(2\pi\rho st)\,d\rho\right)dt$$

$$= (1/2\pi s)\int_0^1 (1-t^2)^{(n-3)/2}[(\sin 2\pi\xi st)/t - (\sin 2\pi t)/t]\,dt$$

$$= (1/2\pi s)[T_1 - T_2], \quad \text{respectively.}$$

Clearly, $|T_2| \leqq C$.

Now, for $n \geqq 3$, a change of variable and an application of the second mean value theorem show that for any $\xi s > 1$, there exists an $\eta \in (0, 2\pi\xi s)$ such that

$$T_1 = \int_0^{2\pi\xi s}[1-(u/(2\pi\xi s))^2]^{(n-3)/2}\cdot[(\sin u)/u]\,du = \int_0^{\eta}[(\sin u)/u]\,du.$$

But since $\int_0^{\to\infty}(\sin u/u) = \pi/2$, it follows that $T_1 \leqq C$.

If $n=2$, then

$$|T_1| \leqq \left|\int_0^{1/2}\frac{\sin(2\pi\xi st)}{t(1-t^2)^{1/2}}\,dt\right| + \left|\int_{1/2}^1 \frac{\sin(2\pi\xi st)}{t(1-t^2)^{1/2}}\,dt\right|$$

Here, where the second term is clearly bounded and for the first integral, we apply the second mean value theorem to the increasing function $(1-t^2)^{-1/2}$ on $(0, 1/2)$. Thus there exists some $\lambda \in (0, 1/2)$ such that

$$\left|\int_0^{1/2} \frac{\sin(2\pi st\xi)}{t(1-t^2)^{1/2}}\,dt\right| = \left|(2/\sqrt{3})\int_\lambda^{1/2} \frac{\sin(2\pi st\xi)}{t}\,dt\right|$$

$$= \left|(2/\sqrt{3})\int_{2\pi\xi s\lambda}^{\pi\xi s} \frac{\sin u}{u}\,du\right| \leqq (4/\sqrt{3})\int_0^{\to\infty} \frac{\sin u}{u}\,du = C.$$

Therefore, (2.7) holds for all $n \geqq 2$.

From the expression of $\hat{f}(x)$ in terms of $A_i(|x|)$, $i=1,2$, we deduce, for γ to be determined later,

$$J = \int_{\mathbb{R}^n} |\hat{f}(x)|^p v(x'/|x|)|x|^\gamma\,dx$$

$$\leqq (2|S_{n-2}|)^p \int_0^\infty s^{n-1+\gamma}|A_1(s)+A_2(s)|^p V(1/s)\,ds,$$

where $V(1/s) = \int_{S_{n-1}} v(x'/s)\,d\sigma$, $s>0$.

By the estimates of $A_i(s)$, $i=1,2$, this yields

$$J \leqq C\int_0^\infty s^{n-1+\gamma}\left(\int_0^{1/s} g(\rho)\,d\rho\right)^p V(1/s)\,ds = C\int_0^\infty \left(s^{-1}\int_0^s g(\rho)\,d\rho\right)^p v_1(s)\,ds,$$

where $v_1(s) = s^{p-n-\gamma-1}V(s)$.

Applying [1, Thm. 1.7] to g and v_1, we obtain

$$J \leqq C_1 \int_0^\infty g(\rho)^p v_1(\rho)\,d\rho \tag{2.8}$$

provided that for each $r>0$,

$$\int_r^\infty s^{-p}v_1(s)\,ds \leqq C_2 r^{-p}\int_0^r v_1(s)\,ds. \tag{2.9}$$

A simple calculation shows that the integral in (2.8) is equal to $\int_{\mathbb{R}^n} v(x)|f(x)|^p\,dx$ provided $\gamma = n(p-2)$. With this γ, (2.9) is clearly equivalent to (2.5). The sufficiency part of the theorem is thus proved.

Necessity. Let $f(x) = |x|^{1-n}\chi_{(0,r)}(|x|)$ in (2.4); then we obtain

$$C\int_0^r \rho^{n-1+(1-n)p}\left(\int_{S_{n-1}} v(x'\rho)\,d\sigma\right)d\rho$$

$$= C\int_0^r \rho^{(n-1)(1-p)}V(\rho)\,d\rho = \int_{|x|\leqq r} |x|^{-p(n-1)}v(x)\,dx$$

$$\geqq \int_0^{1/(2\pi r)} \rho^{(n-1)+n(p-2)}V(1/\rho)|\hat{f}(\rho\theta)|^p\,d\rho$$

for all $\theta \in S_{n-1}$, where

$$\hat{f}(\rho\theta) = \int_0^r \left(\int_{S_{n-1}} e^{-2\pi i\rho t(\omega\cdot\theta)}\,d\sigma(\omega)\right)dt.$$

But, $0<t<r$ and $0<\rho<1/(2\pi r)$ imply $0<2\pi t\rho<1$ so that

$$\left|\int_{S_{n-1}} e^{-2\pi i\rho t(\omega\cdot\theta)}\,d\sigma(\omega)\right| = 2|S_{n-2}|\left|\int_0^1 \cos(2\pi\rho ts)(1-s^2)^{(n-3)/2}\,ds\right|$$

$$\geqq 2|S_{n-2}|(\cos 1)\int_0^1 (1-s^2)^{(n-3)/2}\,ds \equiv C>0.$$

Hence for $r>0$

$$\int_{|x|\leqq r} |x|^{-p(n-1)}v(x)\,dx \geqq C^p r^p \int_0^{1/(2\pi r)} \rho^{np-n-1}V(1/\rho)\,d\rho$$

$$= C^p r^p \int_{|x|\geqq 2\pi r} |x|^{-np}v(x)\,dx$$

which implies (2.5). This completes the proof of the theorem. □

3. The two weighted case. In this section we discuss two weighted Fourier inequalities in the index ranges

$$1<p\leqq q<\infty,\quad 0<q<p<\infty,\quad p>1 \quad\text{and}\quad 0<p<1<q.$$

PROPOSITION 3.1. *Suppose f is a positive even function decreasing to zero on* $(0,\infty)$. *Let u and v the weight functions and $V(x)=\int_0^x v(t)\,dt$, $x>0$.*

(i) *If* $1<p\leqq q<\infty$ *and*

$$A_0 := \sup_{r>0}\left(\int_0^r u(x)\,dx\right)^{1/q}\left(\int_0^r v(x)\,dx\right)^{-1/p}<\infty, \tag{3.1}$$

$$A_1 := \sup_{r>0}\left(\int_r^\infty x^{-q}u(x)\,dx\right)^{1/q}\left(\int_0^r x^{p'}V(x)^{-p'}v(x)\,dx\right)^{1/p'}<\infty, \tag{3.2}$$

then

$$\left\{\int_0^\infty u(1/x)x^{q-2}|\hat f(x)|^q\,dx\right\}^{1/q} \leqq C\left\{\int_0^\infty v(x)f(x)^p\,dx\right\}^{1/p}. \tag{3.3}$$

(ii) *If* $0<q<p<\infty, p>1$ *and* $1/r=1/q-1/p$, *then*

$$\int_0^\infty\left[\left(\int_0^t u(x)\,dx\right)^{1/p}\left(\int_0^t v(x)\,dx\right)^{-1/p}\right]^r u(t)\,dt<\infty$$

and

$$\int_0^\infty\left[\left(\int_t^\infty x^{-q}u(x)\,dx\right)^{1/q}\left(\int_0^t x^{p'}V(x)^{-p'}v(x)\,dx\right)^{1/q'}\right]^r\cdot t^{p'}V(t)^{-p'}v(t)\,dt<\infty$$

imply (3.3).

(iii) *If* $0<p\leqq q<\infty, 0<p<1$ *then* $A_0<\infty$ *and*

$$C_0 := \sup_{r>0} r\left(\int_r^\infty x^{-q}u(x)\,dx\right)^{1/q}\left(\int_0^r v(x)\,dx\right)^{-1/p}<\infty$$

imply (3.3).

Proof. The proof of Theorem 2.1 shows that

$$|\hat{f}(1/x)\cdot x^{-1}| \leqq C\left(x^{-1}\int_0^x f(t)\,dt\right), \qquad x>0,$$

so that

$$\int_0^\infty u(x)|\hat{f}(1/x)\cdot x^{-1}|^q\,dx \leqq C\left\{\int_0^\infty u(x)\left(x^{-1}\int_0^x f(t)\,dt\right)^q dx\right\}.$$

The result then follows if there are corresponding results for the two weighted averaging operator of decreasing functions. If $1<p\leqq q<\infty$ and $1<q<p<\infty, p>1$: this was given in [11, Thm. 2], and the case $0<p\leqq q<\infty$, $0<p<1$ was given in [12, Thm. 3(b)].

Our next result shows that for power weights in the range space of the operator, Proposition 3.1(i) and (iii) is sharp. In fact, with A_0, A_1 and C_0 defined as in Proposition 3.1, we have the following.

PROPOSITION 3.2. *Suppose* (3.3) *is satisfied for all positive even f decreasing to zero on* $(0,\infty)$.

(i) *If* $1<p\leqq q<\infty$, *then* $A_1<\infty$. *If, in addition,* $x^{q-2}u(1/x)\in B_q$ *i.e.*,

$$\int_r^\infty x^{-q}u(x)\,dx \geqq Cr^{-q}\int_0^r u(x)\,dx \tag{3.4}$$

then $A_0<\infty$.

(ii) *If* $0<p\leqq q<\infty$, $0<p<1$, *then* $C_0<\infty$. *If, in addition,* (3.4) *holds, then* $A_0<\infty$.

Proof. Take $f(x)=\chi_{(0,r)}(|x|)$ in (3.3), then for $0<p,\ q<\infty$,

$$\begin{aligned}
C\left(\int_0^r v(x)\,dx\right)^{1/p} &\geqq \left(\int_0^{1/(2\pi r)} u(1/x)x^{q-2}\left|\int_0^r \cos(2\pi xy)\,dy\right|^q dx\right)^{1/q}\\
&\geqq (\cos 1)r\left(\int_0^{1/(2\pi r)} u(1/x)x^{q-2}\,dx\right)^{1/q}\\
&= (\cos 1)r\left(\int_{2\pi r}^\infty u(x)x^{-q}\,dx\right)^{1/q}
\end{aligned}$$

which implies $C_0<\infty$. If (3.4) also holds, this yields at once $A_0<\infty$.

It remains to show that $A_1<\infty$ if $1<p\leqq q<\infty$. Arguing as in [12, Thm. 2], let

$$f_r(|s|)=\left(\int_{2\pi|s|}^r y^{p'}V(y)^{-p'-1}v(y)\,dy\right)^{1/p}\chi_{(0,r)}(2\pi|s|),$$

$r>0$, in (3.3). Then an interchange of order of integration shows that

$$\begin{aligned}
&C\left[\int_0^r y^{p'}V(y)^{-p'}v(y)\,dy\right]^{1/p}\\
&\quad\geqq C\left[\int_0^r y^{p'}V(y)^{-p'-1}v(y)\left(\int_0^{y/2\pi} v(x)\,dx\right)dy\right]^{1/p}\\
&\quad= C\left[\int_0^\infty v(x)f_r(|x|)^p\,dx\right]^{1/p}\\
&\quad\geqq \left\{\int_0^\infty u(x)\left|(2/x)\int_0^\infty \cos(2\pi y/x)f_r(y)\,dy\right|^q dx\right\}^{1/q}\\
&\quad\geqq 2\cos 1\left(\int_r^\infty x^{-q}u(x)\,dx\right)^{1/q}\int_0^{r/2\pi}\left[\int_{2\pi y}^{2\pi r} t^{p'}V(t)^{-p'-1}v(t)\,dt\right]^{1/p} dy.
\end{aligned}$$

Here we used the fact that $(2\pi y)/x \leqq 1$ implies $\cos(2\pi y/x) \geqq \cos 1$. On making the change of variable $2\pi y$ to y, we see that the integral on the right dominates:

$$\int_0^r y^{p'/p}\left[\int_y^r V(t)^{-p'-1}v(t)\,dt\right]^{1/p} dy$$

$$=(1/p)\int_0^r y^{p'/p}\int_y^r\left(\int_s^r V^{-p'-1}v\,d\alpha\right)^{-1/p'} V(s)^{-p'-1}v(s)\,ds\,dy$$

$$=(1/p)\int_0^r V(s)^{-p'-1}v(s)\left(\int_0^s y^{p'/p}\,dy\right)\left[\int_s^r V^{-p'-1}v\,d\alpha\right]^{-1/p'} ds$$

$$\geqq (pp')^{-1}\int_0^r V(s)^{-p'-1}v(s)\left[\int_s^\infty V^{-p'-1}v\,d\alpha\right]^{-1/p'} s^{p'}\,ds$$

$$=(1/p)(p')^{-1/p}\int_0^r V(s)^{-p'}v(s)s^{p'}\,ds.$$

The last equality was obtained under the assumption that $V(\infty)=\infty$. Thus on substituting we have shown that

$$C\left[\int_0^r y^{p'}V(y)^{-p'}v(y)\,dy\right]^{1/p}$$

$$\geqq (2\cos 1)p^{-1}(p')^{-1/p}\left(\int_{2\pi r}^\infty x^{-q}u(x)\,dx\right)^{1/q}\left(\int_0^r V(s)^{-p'}v(s)s^{p'}\,ds\right)$$

and this implies $A_1<\infty$.

Now if $V(\infty)<\infty$, replace v by $v_\varepsilon = v+\varepsilon$, $\varepsilon>0$. Then $V_\varepsilon(\infty)=\int_0^\infty v_\varepsilon\,dt=\infty$ and the above argument shows that (3.2) holds with A_1 independent of ε. Let $\varepsilon\to 0$; then the result follows from Fatou's lemma. □

Since $u(x)=|x|^\alpha$, $-1<\alpha<q-1$ satisfies (3.4) we single out the following.

COROLLARY 3.3. *The inequality* (3.3) *holds for all positive even functions, decreasing to zero on* $(0,\infty)$ *if and only if*

(i) *For* $0<p<1$, $0<p<q<\infty$,

$$r^{(\alpha+1)/q} \leqq C\left(\int_0^r v(x)\,dx\right)^{1/p}, \quad r>0, \quad -1<\alpha<q-1 \tag{3.5}$$

holds.

(ii) *For* $1<p\leqq q<\infty$, (3.5) *and*

$$\left(\int_0^r x^{p'}V(x)^{-p'}v(x)\,dx\right)^{1/p'} \leqq Cr^{1/q'-\alpha/q}, \qquad r>0,$$

hold.

4. The Hankel- and *K*-transformations. In this section we characterize weights for which the Hankel- and K-transformation of certain monotone functions is bounded on weighted Lebesgue spaces.

The Hankel transformation is defined by

$$(H_\lambda f)(x)=\int_0^\infty (xt)^{1/2}J_\lambda(xt)f(t)\,dt, \qquad \lambda>-1/2,$$

where J_λ is the Bessel function [6, p. 81(8)],

$$J_\lambda(z)=C_\lambda z^\lambda \int_0^1 (1-t^2)^{\lambda-1/2}\cos(zt)\,dt,$$

and $C_\lambda = 2^{1-\lambda}\pi^{1/2}/\Gamma(\lambda+1/2)$. Our first result in this section is the following.

THEOREM 4.1. *Let $\lambda>-1/2$, $1\leqq p<\infty$, and w be a positive weight function on $(0,\infty)$. Then*

$$\int_0^\infty w(1/s)s^{-2}|(H_\lambda h)(2\pi s)|^p\,ds \leqq C\int_0^\infty w(s)|sh(s)|^p\,ds \tag{4.1}$$

holds for all positive h with $s^{\lambda+1/2}h(s)$ decreasing to zero on $(0,\infty)$ if and only if $w_\lambda = s^{-p(\lambda-1/2)}w\in B_p^*$, i.e., for each $r>0$

$$\int_r^\infty w(s)s^{-p(\lambda+1/2)}\,ds \leqq Cr^{-p}\int_0^r w(s)\cdot s^{-p(\lambda-1/2)}\,ds. \tag{4.2}$$

Proof. Suppose $g(s)=s^{\lambda+1/2}h(s)$ is positive and decreasing to zero on $(0,\infty)$, then

$$\begin{aligned}(H_\lambda h)(2\pi s) &= C_\lambda(2\pi s)^{\lambda+1/2}\left[\int_0^{1/s}+\int_{1/s}^\infty\right]g(\rho)\int_0^1(1-t^2)^{\lambda-1/2}\cos(2\pi st\rho)\,dt\,d\rho\\ &= C'_\lambda s^{\lambda+1/2}(A_1(s)+A_2(s)).\end{aligned}$$

For $\lambda=(n-2)/2$ with $n\in\mathbb{N}$, $n\geqq 2$, the above functions A_i $(i=1,2)$ are of the same type as the A_i functions in the proof of Theorem 2.3. From the latter proof, it results that

$$A_i(s)\leqq C\int_0^{1/s} g(\rho)\,d\rho,\qquad i=1,2. \tag{4.3}$$

Noting now that all the arguments used to prove these estimates also apply if $n=2\lambda+2>1$ is not an integer, we can conclude that (4.3) holds for any $\lambda>-1/2$.

Therefore, by [1, Thm. 1.7],

$$\begin{aligned}\int_0^\infty w(s)|(H_\lambda h)(2\pi/s)|^p\,ds &\leqq C\int_0^\infty w_\lambda(s)\left[(1/s)\int_0^s t^{\lambda+1/2}h(t)\,dt\right]^p ds\\ &\leqq C'\int_0^\infty w(s)|sh(s)|^p\,ds\end{aligned}$$

provided (4.2) is satisfied.

To prove necessity, set $h(s)=s^{-(\lambda+1/2)}\chi_{(0,r)}(s)$, $r>0$, in (4.1). Then

$$\begin{aligned}&C\int_0^r w(s)s^{p(-\lambda+1/2)}\,ds\\ &\quad\geqq\int_0^\infty w(s)|(H_\lambda h)(2\pi/s)|^p\,ds\\ &\quad= C\int_0^\infty w(s)s^{-\lambda-1/2}\left|\int_0^r\int_0^1(1-y^2)^{\lambda-1/2}\cos(2\pi ty/s)\,dy\,dt\right|^p ds\\ &\quad\geqq C\int_{2\pi r}^\infty w(s)s^{-p(\lambda+1/2)}r^p(\cos 1)^p\left|\int_0^1(1-y^2)^{\lambda-1/2}\,dy\right|^p ds\\ &\quad= Cr^p\int_{2\pi r}^\infty w(s)s^{-p(\lambda+1/2)}\,ds,\end{aligned}$$

and this implies (4.2). □

Remark 4.2. Theorem 2.3 can be viewed as a corollary of Theorem 4.1. Indeed, if v is a positive weight on $\mathbb{R}^n$, $n>1$, and if w is defined by

$$w(s)=s^{n-1-(n+1)p/2}\int_{S_{n-1}} v(\theta s)\,d\sigma, \qquad s>0,$$

then v satisfies (2.5) if and only if w satisfies (4.2) with $\lambda=(n-2)/2$ (i.e., $v\in B_p^*(n)$ if and only if $w_\lambda=s^{-p(\lambda+1/2)}w\in B_p$, $\lambda=(n-2)/2$). Note also that (2.4) holds for $f(x)=f_0(|x|)$, $x\in\mathbb{R}^n$, if and only if (4.1) holds with w as above, for $h(t)=t^{(n-1)/2}f_0(t)$, $\lambda=(n-2)/2$.

Next, we consider the K-transform defined by

$$(K_\lambda f)(x)=\int_0^\infty (xy)^{1/2}k_\lambda(xy)f(y)\,dy, \qquad x>0, \quad \lambda\geqq -1/2,$$

where k_λ is the modified Bessel function of the third kind [6, Chap. X]. If $\lambda=\pm 1/2$, the K-transform reduces to the Laplace transform: $(K_{\pm 1/2}f)(x)=\sqrt{\pi/2}\int_0^\infty e^{-xt}f(t)\,dt$ and if $\lambda>-1/2$ the kernel has the representations

$$k_\lambda(x)=C_\lambda x^\lambda\int_1^\infty e^{-xt}(t^2-1)^{\lambda-1/2}\,dt=C_\lambda x^\lambda\varphi_\lambda(x),$$

$$x>0, \quad C_\lambda=2^{-\lambda}\Gamma(1/2)\cdot\Gamma(\lambda+1/2)^{-1},$$

[7, p. 958(3)], and

$$k_\lambda(x)=C_\lambda^{-1}x^{-\lambda}\int_0^\infty (1+t^2)^{-\lambda-1/2}\cos(xt)\cdot dt=C_\lambda^{-1}x^{-\lambda}\psi_\lambda(x), \qquad x>0$$

[7, p. 959(5)].

Hence we can define the K-transform for $\lambda>-1/2$ by

$$(K_\lambda f)(x)=C_\lambda x^{\lambda+1/2}\int_0^\infty y^{\lambda+1/2}f(y)\cdot\varphi_\lambda(xy)\,dy \tag{4.4}$$

and

$$(K_\lambda f)(x)=C_\lambda^{-1}x^{-\lambda+1/2}\int_0^\infty y^{-\lambda+1/2}f(y)\cdot\psi_\lambda(xy)\,dy. \tag{4.5}$$

We shall need both these representations in the next result.

THEOREM 4.3. *Let $\lambda\geqq -1/2$, $\lambda\neq 0$, $1<p\leqq q<\infty$, and u, v positive weight functions on $(0,\infty)$. Then*

$$\left(\int_0^\infty u(1/x)x^{q-2}|(K_\lambda f)(x)|^q\,dx\right)^{1/q}\leqq C\left(\int_0^\infty v(x)f(x)^p\,dx\right)^{1/p} \tag{4.6}$$

holds for all positive f, with $f(x)\cdot x^{-|\lambda|+1/2}$ decreasing to zero on $(0,\infty)$, if and only if

$$A_0^\lambda:=\sup_{r>0}\left(\int_0^r x^{q(|\lambda|-1/2)}u(x)\,dx\right)^{1/q}\left(\int_0^r v_\lambda(x)\,dx\right)^{-1/p}<\infty$$

and

$$A_1^\lambda:=\sup_{r>0}\left(\int_r^\infty x^{q(|\lambda|-3/2)}u(x)\,dx\right)^{1/q}\left(\int_0^r x^{p'}v_\lambda(x)V_\lambda(x)^{-p'}\,dx\right)^{1/p'}<\infty,$$

where $v_\lambda(x)=x^{p(|\lambda|-1/2)}v(x)$ and $V_\lambda(x)=\int_0^x v_\lambda$.

Proof. For $\lambda = -1/2$: this result is Theorem 1.12 of [8]. We assume therefore that $\lambda > -1/2$.

(i) *Sufficiency.* If $-1/2 < \lambda < 0$, then by (4.4),

$$(K_\lambda f)(x) = C_\lambda x^{\lambda+1/2}\left(\int_0^{1/x} + \int_{1/x}^{\infty}\right) y^{\lambda+1/2} f(y)\varphi_\lambda(xy)\, dy$$
$$= C_\lambda x^{\lambda+1/2}[A_1(x) + A_2(x)], \text{ respectively.}$$

Since $\varphi_\lambda(s) \leqq C$, then with $g(y) = y^{-|\lambda|+1/2} f(y)$, we obtain

$$A_1(x) \leqq C \int_0^{1/x} g(y)\, dy.$$

To estimate A_2, the second mean value theorem shows that for some $\xi > 1/x$

$$\begin{aligned} A_2(x) &= g(1/x) \int_{1/x}^{\xi} \varphi_\lambda(xy)\, dy \\ &= x^{-1} g(1/x) \int_1^{x\xi} \int_1^{\infty} e^{-st}(t^2-1)^{\lambda-1/2}\, dt\, ds \\ &= x^{-1} g(1/x) \int_1^{\infty} (t^2-1)^{\lambda-1/2}(e^{-t} - e^{-x\xi t}) t^{-1}\, dt \\ &\leqq C x^{-1} g(1/x) \\ &\leqq C \int_0^{1/x} g(y)\, dy. \end{aligned}$$

Therefore, if $-1/2 < \lambda < 0$,

$$(K_\lambda f)(x) \leqq C x^{-|\lambda|+1/2} \int_0^{1/x} g(y)\, dy. \tag{4.7}$$

If $0 < \lambda < \infty$, then, by (4.5),

$$(K_\lambda f)(x) = C_\lambda^{-1} x^{-\lambda+1/2}\left(\int_0^{1/x} + \int_{1/x}^{\infty}\right) g(y)\psi_\lambda(xy)\, dy$$
$$= C_\lambda^{-1} x^{-\lambda+1/2}[A_3(x) + A_4(x)], \text{ respectively.}$$

Since $\psi_\lambda(s) \leqq C$, $A_3(x) \leqq C \int_0^{1/x} g(y)\, dy$, again the second mean value theorem shows that for some $\xi > 1/x$,

$$A_4(x) = g(1/x) \int_{1/x}^{\xi} \psi_\lambda(xy)\, dy = x^{-1} g(1/x) \int_1^{x\xi} \psi_\lambda(s)\, ds.$$

But

$$\left|\int_1^{x\xi} \psi_\lambda(s)\, ds\right| \leqq \left|\int_0^{\infty} (1+t^2)^{-\lambda-1/2}[(\sin x\xi t)/t]\, dt\right| + \left|\int_0^{\infty} (1+t^2)^{-\lambda-1/2}[(\sin t)/t]\, dt\right|.$$

The integral on the right is clearly convergent, while the first integral is seen to be convergent if we apply the second mean value theorem to $(1+t^2)^{-\lambda-1/2}$. Therefore

$$A_4(x) \leqq C x^{-1} g(1/x) \leqq C \int_0^{1/x} g(y)\, dy,$$

and this implies that (4.7) holds also for $0<\lambda<\infty$. Now with $u_\lambda(x)=x^{q(|\lambda|-1/2)}u(x)$, (4.7) implies that

$$\left(\int_0^\infty u(x)|x^{-1}(K_\lambda f)(1/x)|^q\,dx\right)^{1/q} \leqq C\left(\int_0^\infty u_\lambda(x)\left[x^{-1}\int_0^x g(y)\,dy\right]^q dx\right)^{1/q}$$

and by [11, Thm. 2] the last integral expression is dominated by

$$\left(\int_0^\infty v_\lambda(x)g(x)^p\,dx\right)^{1/p}=\left(\int_0^\infty v(x)f(x)^p\,dx\right)^{1/p}$$

provided $A_0^\lambda<\infty$ and $A_1^\lambda<\infty$. This proves the first part of the theorem.

(ii) *Necessity.* For $-1/2<\lambda<\infty$, $\lambda\neq 0$, substitute $f(x)=x^{|\lambda|-1/2}\chi_{(0,r)}(x)$, $r>0$, in (4.6). Then from the representation (4.4) we obtain

$$\begin{aligned}
C\left(\int_0^r v_\lambda(x)\,dx\right)^{1/p} &\geqq C_\lambda\left\{\int_0^\infty u(x)\cdot x^{-q(\lambda+3/2)}\left[\int_0^r y^{\lambda+|\lambda|}\varphi_\lambda(y/x)\,dy\right]^q dx\right\}^{1/q}\\
&\geqq C_\lambda\left\{\int_0^r u(x)\cdot x^{-q(\lambda+3/2)}\left[\int_0^x y^{\lambda+|\lambda|}\varphi_\lambda(1)\,dy\right]^q dx\right\}^{1/q}\\
&=C_\lambda'\left\{\int_0^r u(x)x^{q(|\lambda|-1/2)}\,dx\right\}^{1/q}.
\end{aligned}$$

Hence $A_0^\lambda<\infty$. To prove that $A_1^\lambda<\infty$, let

$$g_r^\lambda(x)=\left(\int_x^r y^{p'}V_\lambda(y)^{-p'-1}v_\lambda(y)\,dy\right)^{1/p}$$

and substitute $f(x):=f_r^\lambda(x)=x^{|\lambda|-1/2}g_r^\lambda(x)\chi_{(0,r)}(x)$, $r>0$, in (4.6). Then for $-1/2<\lambda<\infty$, $\lambda\neq 0$, the right side of (4.6) is

$$C\left\{\int_0^r v_\lambda(x)\left(\int_x^r y^{p'}V_\lambda(y)^{-p'-1}v_\lambda(y)\,dy\right)dx\right\}^{1/p}=C\left\{\int_0^r v_\lambda(y)y^{p'}V_\lambda(y)^{-p'}\,dy\right\}^{1/p}.$$

If $-1/2<\lambda<0$, we use the representation (4.4) of the K-transform. The left side of (4.6) then has the form

$$\begin{aligned}
&C_\lambda\left\{\int_0^\infty u(x)\cdot x^{-q(\lambda+3/2)}\left|\int_0^r g_r^\lambda(y)\cdot\varphi_\lambda(y/x)\,dy\right|^q dx\right\}^{1/q}\\
&\qquad\geqq C_\lambda'\left\{\int_r^\infty u(x)\cdot x^{-q(\lambda+3/2)}\left|\int_0^r g_r^\lambda(y)\,dy\right|^q dx\right\}^{1/q}.
\end{aligned}$$

As seen in the proof of Proposition 3.2, the inner integral dominates

$$\int_0^r V_\lambda(s)^{-p'}v_\lambda(s)s^{p'}\,ds$$

whenever $V_\lambda(\infty)=\infty$ and this on substituting implies $A_1^\lambda<\infty$.

If $0<\lambda<\infty$, we substitute (4.5) in the left side of (4.6) so that it has the form

$$C_\lambda^{-1}\left\{\int_0^\infty u(x)\cdot x^{q(\lambda-3/2)}\left|\int_0^\infty g_r^\lambda(y)\psi_\lambda(y/x)\,dy\right|^q dx\right\}^{1/q}. \tag{4.8}$$

But since ψ_λ is strictly positive on $[0,\infty)$ for $0<\lambda<\infty$, there exists δ, $0<\delta<\pi/2$, such that

$$\psi_\lambda(y/x)=\int_0^\infty(1+t^2)^{-\lambda-1/2}\cos(yt/x)\,dt\geqq\int_0^\delta(1+t^2)^{-\lambda-1/2}\cos(yt/x)\,dt \tag{4.9}$$

for all $(y/x)\in[0,1]$ with δ independent of (y/x). To see this, note first that for each $\xi^*\in[0,1]$ there exists $\delta_{\xi^*}, 0<\delta_{\xi^*}<\pi/2$ such that $\int_{\delta_{\xi^*}}^{\infty}(1+t^2)^{-\lambda-1/2}\cos(\xi^* t)\,dt>0$.

Then, for $\xi^*\in[0,1]$ fixed, there exists a neighborhood V_{ξ^*} of ξ^* such that

$$\int_{\delta_{\xi^*}}^{\infty}(1+t^2)^{-\lambda-1/2}(\cos \xi t)\,dt>0$$

for all $\xi\in V_{\xi^*}$. From the covering family $\{V_{\xi^*}: \xi^*\in[0,1]\}$ of $[0,1]$, we can then extract a finite covering subfamily and define δ as the minimum of the corresponding $\delta_{\xi_j^*}, j=1,2,\cdots,N$, N finite. This δ satisfies $0<\delta<\pi/2$, which implies

$$\int_{\delta}^{\delta_{\xi_j^*}}(1+t^2)^{-\lambda-1/2}(\cos \xi t)\,dt\geqq 0$$

for each j and every $\xi\in V_{\xi_j^*}$. This shows that (4.9) holds for all $(y/x)\in[0,1]$ and, as a consequence, we can minorize (4.8) as follows:

$$\begin{aligned}&\left\{\int_r^{\infty}u(x)\cdot x^{q(\lambda-3/2)}\left[\int_0^{\infty}(1+t^2)^{-\lambda-1/2}\int_0^r\cos(yt/x)\cdot g_r^{\lambda}(y)\,dy\,dt\right]^q dx\right\}^{1/q}\\ &\quad\geqq(\cos\delta)\left(\int_0^{\delta}(1+t^2)^{-\lambda-1/2}\,dt\right)\left\{\int_r^{\infty}u(x)\cdot x^{q(\lambda-3/2)}\left[\int_0^r g_r^{\lambda}(y)\,dy\right]^q dx\right\}^{1/q}\\ &\quad\geqq C'\left\{\int_r^{\infty}u(x)\cdot x^{q(\lambda-3/2)}\,dx\right\}^{1/q}\left\{\int_0^r V_{\lambda}(s)^{-p'}v_{\lambda}(s)s^{p'}\,ds\right\},\end{aligned}$$

where the last inequality is again obtained as in the proof of Proposition 3.2, provided $V_\lambda(\infty)=\infty$. Thus again under this condition, $A_1^\lambda<\infty$.

Finally, if $V_\lambda(\infty)<\infty$, replace v_λ by $v_\lambda^\varepsilon=v_\lambda+\varepsilon$, $\varepsilon>0$. Then $V_\lambda^\varepsilon(\infty)=\int_0^\infty v_\lambda^\varepsilon\,dt=\infty$ and the above arguments show that the result holds in this case with A_1^λ independent of ε. As before, Fatou's lemma then implies the result. □

Remark 4.4. Since (4.7) holds for all $\lambda\in(-1/2,\infty)$, $\lambda\neq 0$, the sufficiency part of the theorem holds also in the index ranges $0<q<p<\infty$, $p>1$, and $0<p\leqq q<\infty$, $0<p<1$, under suitably modified weight conditions, by applying [11, Thm. 2] and [12, Thm. 3], respectively (cf. Proposition 3.1). We leave the details.

REFERENCES

[1] M. A. Ariño and B. Muckenhoupt, *Maximal functions on classical Lorentz spaces and Hardy's inequality with weights for non-increasing functions*, Trans. Amer. Math. Soc., 320 (1990), pp. 727-735.

[2] J. J. Benedetto, H. P. Heinig, and R. Johnson, *Weighted Hardy spaces and the Laplace transform* II, Math. Nachr. 132 (1987), pp. 29-55.

[3] J. J. Benedetto and H. P. Heinig, *Fourier transform inequalities with measure weights*, Adv. in Math., to appear.

[4] D. W. Boyd, *The Hilbert transform on rearrangement invariant spaces*, Canad. J. Math., 19 (1967), pp. 599-616.

[5] S. A. Emara and H. P. Heinig, *Weighted norm inequalities for the Hankel- and K-transformations*, Proc. Royal Soc. Edinburgh Sect. A, 103 (1986), pp. 325-333.

[6] A. Erdélyi et al., *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.

[7] I. G. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products*, Academic Press, New York, 1980.

[8] H. P. Heinig, *Weighted inequalities in Fourier analysis*, in Nonlinear Analysis, Functional Spaces and Applications, Vol. 4, Proc. Conf. Roudnice nad Labem, 1990; Teubner-Texte Math., 119 (1990), pp. 42-85.

[9] P. Heywood and P. G. Rooney, *A weighted norm inequality for the Hankel transformation*, Proc. Royal Soc. Edinburgh Sect. A, 99 (1984), pp. 45-50.

[10] S. G. Krein, Ju. I. Petunin, and E. M. Semenov, *Linear Operators*, Transl. Math. Monographs 54, Providence, RI, 1982.

[11] E. Sawyer, *Boundedness of classical operators on classical Lorentz spaces*, Studia Math., 96 (1990), pp. 145-158.

[12] V. D. Stepanov, *The weighted Hardy's inequality for nonincreasing functions*, Trans. Amer. Math. Soc., to appear.

[13] J.-O. Strömberg and R. L. Wheeden, *Weighted norm estimates for the Fourier transform with a pair of weights*, Trans. Amer. Math. Soc., 318 (1990), pp. 355-372.

[14] E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Second ed., Oxford University Press, 1959.

SIAM J. MATH. ANAL
Vol. 23, No. 3, pp. 799-819, May 1992

014

SOME TRACE THEOREMS IN ANISOTROPIC SOBOLEV SPACES*

PATRICK JOLY†

Abstract. Anisotropic Sobolev spaces are functional spaces of Sobolev's type in which different space directions have different roles. In the case of dimension 2, some new trace theorems in such spaces for very general open sets are proved. A sense is also given of the corresponding Green's formula via a generalized concept of Cauchy principal value.

Key words. Sobolev spaces, trace theorems, approximations, Green's formulas

AMS(MOS) subject classifications. 26B99, 46E35

Introduction. In the usual Sobolev spaces such as $H^m(\Omega)$ or $W^{m,p}(\Omega)$, all derivatives in all space directions play the same role. This explains why only the regularity properties of the boundary (Γ) of the domain Ω play a role for trace theorems and extension properties (see the classical references [1], [8], [10]). In some domains of physics, we encounter partial differential equations leading us to work in functional spaces in which only partial derivatives with respect to some space directions have a privileged role. This is in particular the case for paraxial approximations of the wave equation [2], [3], [5] or of the Stokes equations [9]. The simplest example is the parabolic approximation of the wave equation in two dimensions, which is written

$$\frac{1}{c}\frac{\partial^2 u}{\partial t^2}-\frac{\partial^2 u}{\partial x_2\,\partial t}-\frac{1}{2}\frac{\partial}{\partial x_1}\left(c\frac{\partial u}{\partial x_1}\right)=0. \tag{0.1}$$

The energy associated to (0.1) is

$$E(t)=\frac{1}{2}\int_\Omega\left|\frac{\partial u}{\partial t}\right|^2\frac{dx}{c}+\frac{1}{4}\int_\Omega c\left|\frac{\partial u}{\partial x_1}\right|^2 dx, \tag{0.2}$$

which leads us naturally to work in the anisotropic Sobolev space

$$H^{1,0}(\Omega)=\left\{v\in L^2(\Omega)\Big/\frac{\partial v}{\partial x_1}\in L^2(\Omega)\right\},\qquad \Omega\subset\mathbb{R}^2. \tag{0.3}$$

Trace theorems in anisotropic Sobolev spaces of type (0.3) are well known in the case where Ω is a product set [8]. The case where Ω is an arbitrary open set appears to be more delicate.

In this article, we intend to treat completely the case of the space $H^{1,0}(\Omega)$ in two dimensions (see also [7]). The outline of the paper is as follows. In § 1 we state our two main results (Theorems 1 and 2). The first is the trace theorem for $H^{1,0}(\Omega)$; the second one concerns the corresponding Green's formula. Section 2 is devoted to the detailed proof of these two theorems. In § 3 we give the extension of the previous results to the spaces $H^{s,0}(\Omega)$ and $W_p^{1,0}(\Omega)$. The interest of this paper lies rather in the results themselves, which may appear surprising compared to more classical results, than in the techniques, which remain very elementary (we use mainly some variations around Poincaré's inequality in dimension 1). In particular, the orientation of the boundary (Γ) with respect to the direction x_2 and the relative position of Ω with respect

* Received by the editors November 20, 1989; accepted for publication (in revised form) August 5, 1991.

† Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceaú-Rocquencourt, BP 105, 78153 Le Chesnay Cédex, France.

to (Γ) have an influence on the result, and not the regularity of (Γ). So, even if Γ is smooth, the space $H^{1,0}(\Omega)$ does not in general satisfy the extension property.

A complete study of density, compactness, and embedding theorems is presented in the book by Besov, Il'jin, and Nicolskii [4]. An application of our results to the study of the regularity of the solutions of paraxial approximations of the wave equation can be found in [6].

1. The two main theorems in the space $H^{1,0}(\Omega)$.

1.1. Notation and definition. In what follows, Ω will denote an open set of $\mathbb{R}^2$ whose boundary $(\Gamma)=(\partial\Omega)$ is supposed to satisfy the following assumptions (see Fig. 1):

(H1) Ω is locally the epigraph of a Lipschitz function (i.e., (Γ) is locally the graph of a Lipschitz function, and Ω is locally "above" Γ);

(H2) Γ is the reunion of a finite number of:

- parts G_l in the form

$$G_l=\{(x_1,x_2):x_1=f_l(x_2),\ a_1<x_2<b_1,\ f_l \text{ is Lipschitz continuous}\},$$

- their summits $A_l=(a_l,f_l(a_l))$, $B_l=(b_l,f_l(b_l))$,
- horizontal parts H_k:

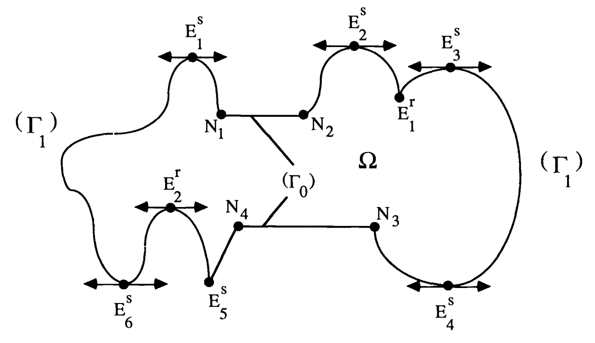
$$H_k=\{(x_1,x_2):x_2=X_k,\ c_k<x_1<d_k\}.$$

We shall note:

$$\Gamma_1=\bigcup_l \bar{G}_l,\quad \Gamma_0=\bigcup_k H_k,\quad \Gamma=\Gamma_0\cup\Gamma_1.$$

We can define the unit normal vector to Γ, outgoing with respect to Ω, almost everywhere on Γ, and we shall denote by $n(M)=(n_1(M),n_2(M))$, M belonging to Γ, the corresponding vector field. Γ_0 and Γ_1 differ by the value of the first component n_1 since

$$\text{a.e. } M\in\Gamma_0,\quad n_1(M)=0,$$
$$\text{a.e. } M\in\Gamma_1,\quad n_1(M)\neq 0.$$



- ● (N_1, N_2, N_3, N_4) : nonstrict extremal point of Γ_1
- ● $(E_1^s, E_2^s, E_3^s, E_4^s, E_5^s, E_6^s)$: strict x_2-extremal outgoing points
- ● (E_2^r, E_1^r) : strict x_2-extremal incoming points

FIG. 1. *Illustration of Hypotheses* (H1) *and* (H2).

As we shall see later, the traces will be defined only on Γ_1. We shall use in the sequel the notion of x_2-extremal point.

DEFINITION 1. A point M of Γ is a (strict) x_2-extremal point if and only if the coordinate x_2 realizes a (strict) local extremum on Γ. Such a point is said to be outgoing with respect to Ω (otherwise it is said to be incoming) if x_2 realizes at this point a local extremum in Ω, i.e., if there exists a ball $B(M, \varepsilon)$ such that

$$x_2(M) > x_2(M') \quad \forall M' \in B(M, \varepsilon),$$

or

$$x_2(M) < x_2(M') \quad \forall M' \in B(M, \varepsilon).$$

As an example, consider the case where (Γ) locally coincides with the graph of a Lipschitzian map, $x_2 = f(x_1)$, f admitting an extremum at $x_1 = 0$.

Note that the notion of x_2-extremal point makes use of the orientation of (Γ) with respect to the direction Ox_2, while the notion of an outgoing (or incoming) x_2-extremal point involves the relative position of Ω with respect to (Γ). Because of (H2), the number of strict x_2-extremal points is finite and each of them is one of the summits (A_1, B_1). We shall denote by $\{M_j, 1 \leqq j \leqq N\}$ the set of these points.
Finally, we are led to introduce a weight function $l(M)$ on Γ_1 as follows.

We shall say that a point M of Γ_1 belongs to Γ_1^* if it exists a point M^* (which is unique if it exists) of Γ such that the open segment MM^* is horizontal and included in Ω (see Fig. 3). Note that M^* may not exist if Ω is unbounded, as illustrated in Fig. 2, but that $\Gamma_1^* = \Gamma_1$ as soon as Ω is bounded. Then we introduce

$$\lambda(M) = \text{length of } MM^* = |x_1(M^*) - x_1(M)| \quad \text{for } M \text{ in } \Gamma_1^*$$

and then define the weight function $l(M)$ on Γ_1 by

$$\begin{aligned} l(M) &= \inf(\lambda(M), 1) && \text{if } M \text{ belongs to } \Gamma_1^*, \\ l(M) &= 1 && \text{if not.} \end{aligned}$$

DEFINITION 2. A function $\phi(M)$ defined on Γ_1 is said to be (Γ, x_1)-even if and only if

$$\phi(M) = \phi(M^*) \quad \text{if } M \text{ belongs to } \Gamma_1^*,$$

and (Γ, x_1)-odd if and only if

$$\begin{aligned} \phi(M) + \phi(M^*) - 0 && \text{if } M \text{ belongs to } \Gamma_1^*, \\ \phi(M) = 0 && \text{if not.} \end{aligned}$$

Of course, any function ϕ defined on Γ_1 can be uniquely decomposed as $\phi = \phi_e + \phi_o$ where ϕ_e is (Γ, x_1)-even and ϕ_0 is (Γ, x_1)-odd.

Denoting by σ a curvilinear abcissa along Γ and by $d\sigma$ the corresponding superficial measure, we introduce the following two Hilbert spaces:

$$(1.1) \qquad \begin{aligned} L^2_{\text{even}}(\Gamma_1; l|n_1|) &= \left\{\phi: \int_{\Gamma_1} |\phi|^2 l |n_1| \, d\sigma), \phi \text{ is } (\Gamma, x_1)\text{-even}\right\}, \\ L^2_{\text{odd}}(\Gamma_1; l^{-1}|n_1|) &= \left\{\phi: \int_{\Gamma_1} |\phi|^2 l^{-1} |n_1| \, d\sigma), \phi \text{ is } (\Gamma, x_1)\text{-odd}\right\} \end{aligned}$$

and the space of traces $T(\Omega, \Gamma_1)$:

$$(1.2) \qquad T(\Gamma_1, \Omega) = L^2_{\text{even}}(\Gamma_1; l|n_1|) \oplus L^2_{\text{odd}}(\Gamma_1; l^{-1}|n_1|)$$

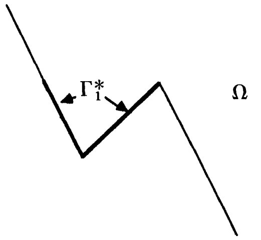


FIG. 2. *A case where* $\Gamma_1^* \neq \Gamma_1$.

equipped with the norm

$$(1.3)\qquad |\phi_e+\phi_o|^2_{T(\Omega,\Gamma_1)}=\int_{\Gamma_1}|\phi_e|^2 l|n_1|\,d\sigma+\int_{\Gamma_1}|\phi_o|^2 l^{-1}|n_1|\,d\sigma.$$

Remarks. (1) The weight function $l(M)$ is strictly positive almost everywhere on Γ_1 and vanishes only at the strict outgoing x_2-extremal points of Γ. In some sense, the singularity of the function $l(M)^{-1}$ measures the sharpness of the open set Ω in the neighborhood of such an extremal point.

(2) If Γ_1 has no strict outgoing x_2-extremal point, the space simply coincides with the space $L^2(\Gamma_1, |n_1|)$. More generally, if we exclude the neighborhoods of the points of Γ_1 whose tangent is horizontal (where the function $|n_1|$ degenerates) and the neighborhoods of the strict outgoing x_2-extremal points of Γ_1 (where $l(M)$ degenerates) the space $T(\Gamma_1, \Omega)$ simply coincides with the space $L^2(\Gamma_1)$.

(3) The space $T(\Gamma_1, \Omega)$ is intrinsically linked not to the curve Γ_1 but to the pair (Γ_1, Ω), since the function $l(M)$ depends on the outgoing x_2-extremal points of Γ_1 which themselves depend on the relative position of Γ_1 with respect to Ω.

1.2. The two main results.

THEOREM 1. *The trace mapping* $\gamma_0: D(\bar{\Omega}) \to L^2(\Gamma_1)$ *defined by* $(\gamma_0 u)(M) = u(M)$ *extends in a unique way to a linear and continuous map, which we still denote by* γ_0, *from* $H^{1,0}(\Omega)$ *onto* $T(\Gamma_1, \Omega)$. *Moreover, the application* γ_0 *is surjective from* $H^{1,0}(\Omega)$ *onto* $T(\Gamma_1, \Omega)$.

As a direct consequence of Theorem 1, we have the following corollary.

COROLLARY 1. *The space* $H^{1,0}(\Omega)$ *has the extension property if and only if* Ω *has no strict* x_2*-extremal point.*

Indeed if $\mathcal{O} = \mathbb{R}^2\text{-}\Omega$, it is clear that any strict x_2-extremal of Γ_1 that is outgoing with respect to Ω is incoming with respect to $\mathcal{O}$, and conversely. So the space $T(\Gamma_1, \Omega)$ and $T(\Gamma_1, \mathcal{O})$ coincide if and only if Γ_1 has no strict x_2-extremal point. As an example, note that the boundary of the unit square has no strict x_2-extremal point while the unit circle has two extremal points.

It is interesting to make the following comments:

- The trace of a function of $H^{1,0}(\Omega)$ is defined only on the part (Γ_1) of the boundary (Γ).

- The trace can have singularities which can be stronger than L^2-singularities the x_2-extremal points of (Γ_1). Moreover, such a singularity can be stronger at the neighborhood of a strict outgoing x^2-extremal point. However, this singularity only affects the (Γ, x_1)-odd part of the function, while the (Γ, x_1)-even part must be more regular because of the presence of the weight function l^{-1}. In some sense, this condition traduces a sort of continuity of the trace at the extremal point, due to the effect of the regularity of the original function in the x_1 direction.

Our second result will express how the classical Green's formula

$$(1.4)\qquad \int_\Omega \left(u\frac{\partial v}{\partial x_1}+\frac{\partial u}{\partial x_1}v\right)dx=\int_\Gamma uvn_1\,d\sigma\quad \forall(u,v)\in D(\bar{\Omega})^2$$

can be extended to functions in $H^{1,0}(\Omega)^2$. For this we are led to define the notion of principal value in the direction x_2 of a curvilinear integral along (Γ_1). Let us denote by $\{M_j, j\in J_o\}$ the strict outgoing x_2-extremal points of Γ_1 and by $\{M_j, j\in J_i\}$ the strict incoming x_2-extremal points of Γ_1 (which will not play any role in what follows).

For $\delta>0$ small enough and j in J_0 we denote by $\Gamma_1^j(\delta)$ the connected component of the set

$$\{M\in\Gamma_1/|x_2(M)-x_2(M_j)|<\delta\},$$

which contains M_j (see Fig. 3), and we set

$$\Gamma_1(\delta)=\bigcup_{j\in J_o}\Gamma_1^j(\delta).$$

The essential property of $\Gamma_1^j(\delta)$ is the fact that it is "x_2-symmetry with respect to M_j" in the sense that it is made of two arcs of curve whose common extremity is the point M_j and which are of equal size in the x_2-direction.

THEOREM 2. (i) *For any (ϕ,ψ) in $T(\Gamma_1,\Omega)^2$, the limit*

$$(1.5)\qquad \lim_{\delta\downarrow 0}\int_{\Gamma_1\setminus\Gamma_1(\delta)}\phi\psi n_1\,d\sigma\overset{\text{def}}{=}\text{v.p. }x_2\int_{\Gamma_1}\phi\psi n_1\,d\sigma$$

exists and the bilinear form $(\phi,\varphi)\to\text{v.p. }x_2\int_{\Gamma_1}\phi\varphi n_1\,d\sigma$ is continuous on $T(\Gamma_1,\Omega)^2$.

(ii) *We have in $H^{1,0}(\Omega)$ the following Green's formula*:

$$(1.6)\qquad \forall(u,v)\in H^{1,0}(\Omega)^2,\quad \int_\Omega\left(u\frac{\partial v}{\partial x_1}+\frac{\partial u}{\partial x_1}v\right)dx=\text{v.p. }x_2\int_{\Gamma_1}\gamma_0u\,\gamma_0v\,n_1\,d\sigma.$$

2. Proofs of the two main results.

2.1. Proof of Theorem 2. Rather than give a direct proof, we have chosen to break down our presentation into several lemmas. This has the advantage of being expository and of illustrating the importance of the notions given in § 1.1. As a preliminary result we can give a first simple result, which can be found, for instance, in [13].

PROPOSITION 1. *The map γ_{n_1}: $D(\bar{\Omega})\to L^2(\Gamma)$ defined by $\gamma_{n_1}u(M)=n_1(M)u(M)$ can be extended in a unique way to a linear and continuous mapping from $H^{1,0}(\Omega)$ in $H^{-1/2}(\Gamma)$ and we have the Green's formula*

$$(2.1)\qquad \forall(u,v)\in H^{1,0}(\Omega)\times H^1(\Omega),\quad \int_\Omega\left(\frac{\partial u}{\partial x_1}v+u\frac{\partial v}{\partial x_1}\right)dx=\langle\gamma_{n_1}u,\gamma_0u\rangle_\Gamma,$$

where $\langle\cdot,\cdot\rangle_\Gamma$ denotes the duality $H^{-1/2}(\Gamma)-H^{1/2}(\Gamma)$.

Proof. Just note that $u\in H^{1,0}(\Omega)$ implies that $(u,0)\in H(\text{div};\Omega)$ and apply the standard trace theorem in $H(\text{div};\Omega)$ (see [12]). □

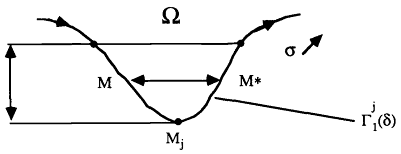


FIG. 3. *The set* $\Gamma_1^j(\delta)$.

Such a result is not completely satisfactory. Indeed, in (2.1), (u, v) have a symmetric role on the right-hand side and not on the left-hand side. Moreover, this result concerns the product $n_1 \times u$ and not u itself. In particular, as $n_1 = 0$ on Γ_0, it is clear that nothing can be said about the trace of u along Γ_0. This is not surprising since a function in $H^{1,0}(\Omega)$ can be discontinuous through a line $x_1 = \text{cste}$.

Our purpose now is to show precisely that the trace of u can be defined as a function almost everywhere (for the Lebesgue measure on Γ) defined on Γ_1 and, in order to get an optimal result, to characterize the image of this trace operator. For the sake of clarity, we shall distinguish several steps. Moreover, for the proofs, we shall use the curvilinear abcissa σ along Γ and the corresponding parametric representation

$$\sigma \to M(\sigma) = (X_1(\sigma), X_2(\sigma)),$$

and suppose that the curve Γ is oriented in such a way that, if $n(\sigma)$ denotes $n(M(\sigma))$,

$$n_1(\sigma) = -\frac{dX_2}{d\sigma}(\sigma), \qquad n_2(\sigma) = \frac{dX_1}{d\sigma}(\sigma).$$

2.1.1. The case where $n_1(\sigma)$ has a constant sign along (Γ_1).

LEMMA 1. *If $n_1(\sigma)$ has a constant sign along (Γ_1) (see Fig. 4), the trace mapping $\gamma_0 : D(\bar{\Omega}) \to L^2(\Gamma_1)$ can be uniquely extended to a linear continuous mapping from $H^{1,0}(\Omega)$ onto $L^2(\Gamma_1; |n_1|\, d\sigma)$ and we have*

$$\text{(i)} \qquad \forall (u, v) \in H^{1,0}(\Omega)^2, \qquad \int_\Omega \left(u \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_1} v \right) dx = \int_{\Gamma_1} \gamma_0 u \cdot \gamma_0 v n_1 \, d\sigma,$$

$$\text{(ii)} \qquad \forall u \in H^{1,0}(\Omega), \qquad \|\gamma_0 u\|_{n_1, \Gamma_1} \leqq \|u\|_{H^{1,0}(\Omega)}.$$

Proof of Lemma 1. When we take $v = u$ in the usual Green's formula, we get

$$\forall u \in D(\bar{\Omega}), \qquad \int_{\Gamma_1} |u|^2 n_1 \, d\sigma = 2 \int_\Omega u \frac{\partial u}{\partial x_1} dx.$$

Then it suffices to remark that the absolute value of the left-hand side is,if $n_1$has a constant sign, the square of the norm of $\gamma_0 u$ in $L^2(\Gamma_1; |n_1|\, d\sigma)$; we also use the Schwar z's inequality to estimate the right-hand side to obtain

$$\text{(2.2)} \qquad \forall u \in D(\bar{\Omega}), \qquad \|\gamma_0 u\|^2_{n_1, \Gamma_1} \leqq \|u\|^2_{H^{1,0}(\Omega)}.$$

Formula (ii) follows then by density and continuity. It is then easy to deduce the unique extension result and the Green's formula (i) for the same reasons.

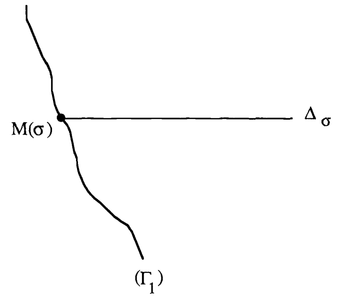


FIG. 4. *The case where n_1 has a constant sign.*

It remains to prove the surjectivity result. For this we note that, thanks to the assumption about $n_1(\sigma)$, Ω can be seen as the reunion of half lines. For instance, if $n_1(\sigma) \leqq 0$,

$$\bar{\Omega} = \bigcup_{M(\sigma)\in\Gamma_1} \bar{\Delta}_\sigma, \qquad \Delta_\sigma = \{(x_1, X_2(\sigma)), x_1 \geqq X_1(\sigma)\}.$$

Let us consider ϕ in $L^2(\Gamma_1; |n_1|\, d\sigma)$. As $|n_1(\sigma)| = |\, dX_2/d\sigma(\sigma)|$, we can write

$$\|\phi\|^2_{|n_1|,\Gamma_1} = \int_{\Gamma_1} |\phi(\sigma)|^2 \left| \frac{dX_2}{d\sigma}(\sigma) \right| d\sigma. \tag{2.3}$$

If $I_2 = \{x_2 = X_2(\sigma)/M(\sigma) \in \Gamma_1\}$, as $n_1(\sigma) \geqq 0$, the map $\sigma \to X_2(\sigma)$ is a bijection between I_2 and Σ_1, the set of the values of σ when $M(\sigma)$ describes Γ_1. Let $\Phi_\eta(x)$ be the function defined in Fig. 5.

We construct the following extension of $\phi(\sigma)$:

$$u(x_1, x_2) = \phi(\sigma)\Phi_\eta(x_1 - X_1(\sigma)) \quad \text{if } x_2 = X_2(\sigma). \tag{2.4}$$

We evaluate the integral $\int_\Omega |u|^2\, dx$ with the help of the change of variable $(x_1, \sigma) \to (x_1, x_2 = X_2(\sigma))$, whose Jacobian is equal to $|n_1(\sigma)|$. By Fubini's theorem, we have

$$\begin{aligned} \int_\Omega |u(x_1, x_2)|^2\, dx &= \int_{\Sigma_1} \left(\int_{X_1(\sigma)}^{+\infty} |\Phi_\eta(x_1 - X_1(\sigma))|^2\, dx_1 \right) |\phi(\sigma)|^2 |n_1(\sigma)|\, d\sigma \\ &= \|\Phi_\eta\|^2_{L^2} \|\phi\|^2_{|n_1|,\Gamma_1}. \end{aligned} \tag{2.5}$$

By the same approach, we obtain for $\partial u/\partial x_1$

$$\int_\Omega \left| \frac{\partial u}{\partial x_1}(x_1, x_2) \right|^2 dx = \left\| \frac{d\Phi_\eta}{dx} \right\|^2_{L^2} \|\phi\|^2_{|n_1|,\Gamma}. \tag{2.6}$$

Equations (2.5) and (2.6) show that the map $\phi \to u$ is a continuous linear extension operator from $L^2(\Gamma_1; |n_1|\, d\sigma)$ in $H^{1,0}(\Omega)^2$. □

Remark 1. Playing with the parameter η, we can localize the extension of ϕ in an arbitrary neighborhood of (Γ_1).

Figure 6 illustrates the optimality of Lemma 1.

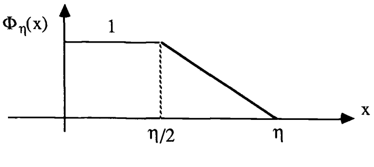


FIG. 5. *The cut-off function* Φ_η.

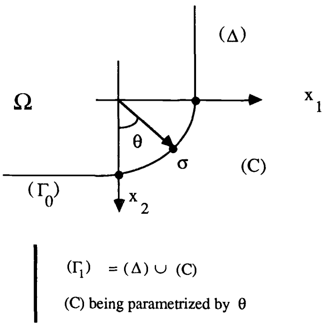


FIG. 6. *A first example.*

If (C) is parametrized by θ we easily check that $n_1(\theta)=\sin\theta$ and that

$$L^2(C;|n_1|\,d\sigma)\equiv\left\{g(\theta):\left[0,\frac{\pi}{2}\right]\to\mathbb{R}\Big/\int_0^{\pi/2}|g(\theta)|^2\sin\theta\,d\theta<+\infty\right\}.$$

Consider the function, for $\alpha>0$,

$$u_\alpha(x_1,x_2)=\Phi(x_1,x_2)\frac{1}{x_2^\alpha},$$

where Φ is a smooth function with compact support, identically equal to 1 in a neighborhood of (C). The trace of u_α can be identified to the function

$$g_\alpha(\theta)=(\sin\theta)^{-\alpha},$$

and we verify immediately that

$$u_\alpha\in H^{1,0}(\Omega)\Leftrightarrow\alpha<\tfrac{1}{2},$$
$$g_\alpha\in L^2(C,|n_1|\,d\sigma)\Leftrightarrow\alpha<\tfrac{1}{2}.$$

2.1.2. An intermediate result: Application to the case where Γ_1 has no strict x_2-extremal point. Let us establish a general result that will be useful in the sequel. For any $1\leqq j\leqq N$, we denote by $B_j(\rho)$ the ball of center M_j and radius ρ (see Fig. 7) and we set

$$\tilde{\Gamma}_{1,\rho}=\Gamma_1\cap\left[\mathbb{R}^2\Big\backslash\bigcup_{j=1}B_j(\rho)\right]. \tag{2.7}$$

We can state the following lemma.

LEMMA 2. *For any $\rho>0$ small enough, the trace operator γ_0 extends uniquely in a linear, continuous, and surjective map from $H^{1,0}(\Omega)$ in $L^2(\tilde{\Gamma}_{1,\rho};|n_1|\,d\sigma)$.*

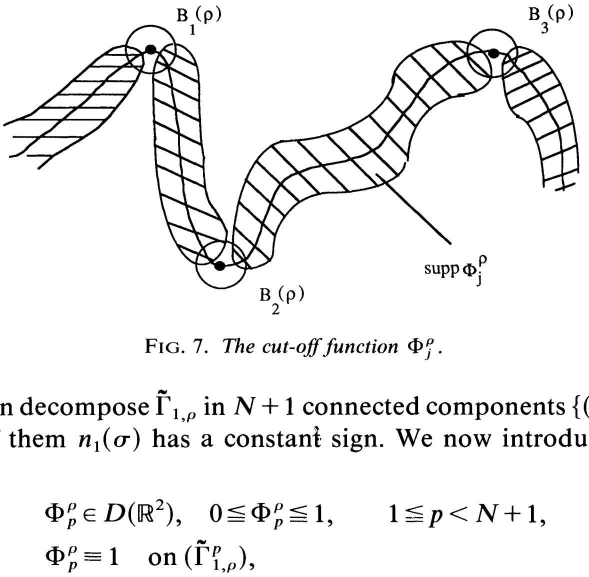


FIG. 7. *The cut-off function Φ_j^ρ.*

Proof. We can decompose $\tilde{\Gamma}_{1,\rho}$ in $N+1$ connected components $\{(\tilde{\Gamma}_{1,\rho}^p),1\leqq p\leqq N+1\}$; along each of them $n_1(\sigma)$ has a constant sign. We now introduce smooth cut-off functions:

$$\begin{aligned}&\Phi_p^\rho\in D(\mathbb{R}^2),\quad 0\leqq\Phi_p^\rho\leqq1,\qquad 1\leqq p<N+1,\\ &\Phi_p^\rho\equiv1\quad\text{on}\,(\tilde{\Gamma}_{1,\rho}^p),\\ &\operatorname{supp}\Phi_p^\rho\cap\operatorname{supp}\Phi_m^\rho=\phi\quad\text{for } p\neq m.\end{aligned}\tag{2.8}$$

On $\tilde{\Gamma}_{1,\rho}$, $\sum_{j=1}^{N+1}\Phi_\rho^p=1$, and we can write

$$\forall u\in D(\bar{\Omega}),\quad \|u\|^2_{|n_1|,\tilde{\Gamma}_{1,\rho}}=\sum_{p=1}^{N+1}\|\Phi_p^\rho u\|_{|n_1|,\tilde{\Gamma}_{1,\rho}^p} \tag{2.9}$$

We apply the Green's formula to each of the functions $\Phi^\rho_p u$:

$$\int_{\tilde\Gamma^p_{1,\rho}} |\Phi^\rho_p u|^2 |n_1|\, d\sigma = 2\left|\int_\Omega \Phi^\rho_p u \frac{\partial}{\partial x_1}(\Phi^\rho_p u)\,dx\right|$$

$$= 2\left|\int_\Omega \left(|\Phi^\rho_p|^2 u\frac{\partial u}{\partial x_1} + \Phi^\rho_p \frac{\partial \Phi^\rho_p}{\partial x_1}|u|^2\right)dx\right|.$$

After summing on p, we get

$$\int_{\tilde\Gamma_{1,\rho}} |u|^2 n_1\,d\sigma \leqq 2\sum_{p=1}^{N+1}\int_\Omega |\Phi^\rho_p|^2\left|u\frac{\partial u}{\partial x_1}\right|dx + 2\sum_{p=1}^{N+1}\int_\Omega \left|\Phi^\rho_p\frac{\partial\Phi^\rho_p}{\partial x_1}\right||u|^2\,dx.$$

So, if we introduce the constant

$$C(\rho) = \max_{1\leqq p\leqq N+1}\left(\max_{\mathbb{R}^2}\left|\frac{\partial\Phi^\rho_p}{\partial x_1}\right|\right), \tag{2.10}$$

it is easy to derive, using Young's and Cauchy-Schwarz's inequalities, the following estimate:

$$\left(\int_{\tilde\Gamma_{1,\rho}} |u|^2|n_1|\,d\sigma\right)^{1/2} \leqq (1+C(\rho))^{1/2}\|u\|_{H^{1,0}}, \tag{2.11}$$

from which the first part of the lemma follows immediately. To prove that γ_0 is surjective it is sufficient to extend locally in Ω each of the restrictions to $(\tilde\Gamma^p_{1,\rho})$ of a function ϕ in the space $L^2(\tilde\Gamma^p_{1,\rho}; |n_1|\,d\sigma)$. This is possible thanks to Remark 1. □

As a direct consequence of Lemma 2, we have the following corollary.

COROLLARY 2. *When Γ_1 has no strict x_2-extremal point, the trace mapping γ_0 is linear, continuous, and surjective from $H^{1,0}(\Omega)$ on to $L^2(\Gamma_1; |n_1|\,d\sigma)$.*

Of course, the constant $C(\rho)$ appearing in the estimate (2.11) blows up when $\rho\to 0$ as soon as $N\neq 0$; indeed, we have to keep the supports of the cut-off functions disjoint. This remark suggests that the result of Corollary 2 cannot be extended to the general case. This is confirmed by the counterexample in Fig. 8.

Consider:

$$u_\alpha(x_1,x_2) = \frac{1}{x_2^\alpha};$$

we see immediately that

$$u_\alpha \in H^{1,0}(\Omega) \Leftrightarrow \alpha < 1,$$

$$\gamma_0 u_\alpha \in L^2(\Gamma_1; |n_1|\,d\sigma) \Leftrightarrow \alpha < \tfrac{1}{2}.$$

For $\frac{1}{2}\leqq\alpha<1$, u_α belongs to $H^{1,0}(\Omega)$, whose trace on Γ_1 is not in $L^2(\Gamma_1; |n_1|\,d\sigma)$.

2.1.3. Study of the traces in a neighborhood of strict x_2-extremal points. Clearly it remains to treat the neighborhoods of the strict x_2-extremal points of (Γ). We shall

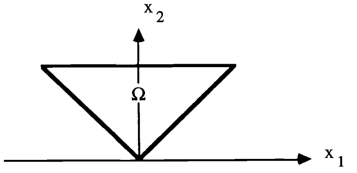


FIG. 8. *A second example.*

need some preliminary notation. Let $M_j = M(\sigma_j)$ be a strict x_2-extremal point of (Γ). As such a point is isolated, the function $\sigma \to X_2(\sigma)$ has a unique relative extremum (strict) at $\sigma = \sigma_j$. So there exists, for any $\varepsilon > 0$ small enough, a pair $\{\sigma_j^-(\varepsilon), \sigma_j^+(\varepsilon)\}$ such that, if for instance $X_2(\sigma_j)$ is minimum of $X_2(\sigma)$,

$$X_2(\sigma)\colon\ [\sigma_j^-(\varepsilon), \sigma_j] \to [X_2(\sigma_j), X_2(\sigma_j)+\varepsilon] \quad \text{is decreasing and surjective,}$$

$$X_2(\sigma)\colon\ [\sigma_j, \sigma_j^+(\varepsilon)] \to [X_2(\sigma_j), X_s(\sigma_j)+\varepsilon] \quad \text{is increasing and surjective.}$$

see Fig. 9.

Therefore, for any x_2 in $[X_2(\sigma_j), X_2(\sigma_j)+\varepsilon]$, there exists a unique pair $(\sigma_-, \sigma_+) \in [\sigma_j^-(\varepsilon), \sigma_j] \times [\sigma_j, \sigma_j^+(\varepsilon)]$, such that $X_2(\sigma^-) = X_2(\sigma^+) = x_2$. In this way we define a bijection $\sigma \to \sigma^+$ between the intervals $[\sigma_j^-(\varepsilon), \sigma_j]$ and $[\sigma_j, \sigma_j^+(\varepsilon)]$, and we can associate to σ_j the open set

$$\Omega(\sigma_j, \varepsilon) = \{]M(\sigma^-), M(\sigma^+)[/x_2 = X_2(\sigma^-) \in]X_2'(\sigma_j), X_2(\sigma_j)+\varepsilon[\}. \tag{2.12}$$

According to the notation of § 1.1, if $M(\sigma^-) = M$, $M(\sigma^+) = M^*$, and conversely. It is clear that:

$$M_j \text{ is outgoing} \Rightarrow \Omega(\sigma_j, \varepsilon) \subset \Omega,$$

$$M_j \text{ is incoming} \Rightarrow \Omega(\sigma_j, \varepsilon) \not\subset \Omega.$$

Let us note that, up to a local change the orientation of (Γ), we can assume that $X_1(\sigma^+) > X_1(\sigma^-)$ as soon as σ^- and σ^+ are related by the bijection shown in Fig. 10. By definition, $\Gamma_1^j(\varepsilon)$ is the part of (Γ) described by $M(\sigma)$ when σ varies in the interval $[\sigma_j^-(\varepsilon), \sigma_j^+(\varepsilon)]$. $\Gamma_1^j(\varepsilon)$ coincides with the "Γ_1 part" of the boundary of $\Omega(\sigma_j, \varepsilon)$.

First of all, let us note that the incoming strict x_2-extremal points do not play any role. This is due to the fact that we can move inside Ω a "horizontal" segment (i.e., parallel to the line $x_2 = 0$) of fixed length η, one of the extremities of the segment describing $\Gamma_1^j(\varepsilon)$ as illustrated in Fig. 11.

In this way, we have defined two disjoint open sets, namely, $\Omega_j^-(\varepsilon, \eta)$ and $\Omega_j^+(\varepsilon, \eta)$, included in Ω, to which we can apply the techniques of Lemmas 1 and 2. Let us define the cut-off functions ϕ_η^+ and ϕ_η^-, respectively, on $\Omega_j^-(\varepsilon, \eta)$ and $\Omega_j^+(\varepsilon, \eta)$ by

$$\phi_\eta^\pm(X_1(\sigma) \pm \xi_1, X_2(\sigma)) = \Phi_\eta(\xi_1), \qquad \xi_1 \in [0, \eta], \tag{2.13}$$

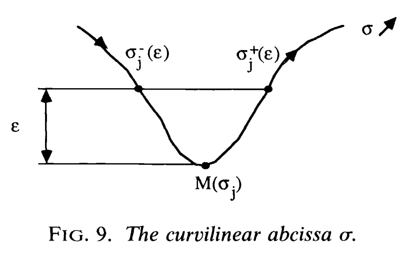


FIG. 9. *The curvilinear abcissa σ.*

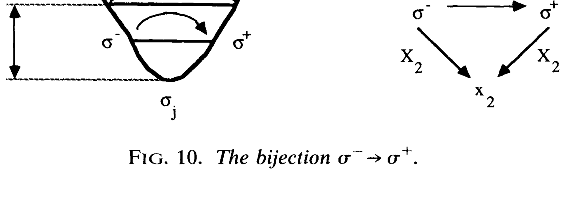


FIG. 10. *The bijection $\sigma^- \to \sigma^+$.*

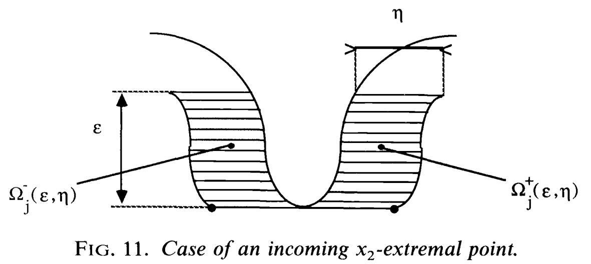

η
ε
$\Omega_j^-(\varepsilon,\eta)$
$\Omega_j^+(\varepsilon,\eta)$

FIG. 11. *Case of an incoming x_2-extremal point.*

where Φ_η has been defined in the proof of Lemma 1. To show that the trace of a function $H^{1,0}(\Omega)$ belongs to $L^2(\Gamma_1^j(\varepsilon), |n_1|\, d\sigma)$, we apply the Green's formula (1.12) to the functions $\phi_\eta^- u$ and $\phi_\eta^+ u$ when u is smooth. To get the surjectivity result, we construct an extension of $\phi(\sigma)$ in $L^2(\Gamma_1^j(\varepsilon), |n_1|\, d\sigma)$ by constructing independently two extensions, respectively in $\Omega_j^-(\varepsilon, \eta)$ and $\Omega_j^+(\varepsilon, \eta)$, of the restrictions of $\phi(\sigma)$ to $\Gamma_1^j(\varepsilon) \cap \partial\Omega_j^-(\varepsilon, \eta)$ and $\Gamma_1^j(\varepsilon) \cap \partial\Omega_j^+(\varepsilon, \eta)$. Therefore the results stated in Lemma 2 and Corollary 2 remain valid if we replace the strict x_2-extremal points of (Γ) by the only outgoing strict x_2-extremal points of (Γ), as described in the following corollary.

COROLLARY 3. *When Γ_1 has no strict outgoing x_2-extremal point, the trace mapping γ_0 is linear, continuous, and surjective from $H^{1,0}(\Omega)$ on to $L^2(\Gamma_1; |n_1|\, d\sigma)$.*

It remains to examine what happens in a neighborhood of strict outgoing x_2-extremal points, which means to make precise the trace theorem for $H^{1,0}(\Omega(\sigma_j, \varepsilon))$. For this, we shall use the function $l(\sigma)$ defined in § 1.1. Restricting ourselves to $\Gamma_1^j(\varepsilon)$, we have

$$(2.14) \qquad l(\sigma^-) = l(\sigma^+) = |X_1(\sigma^+) - X_1(\sigma^-)| \quad \text{for } X_2(\sigma^-) = X_2(\sigma^+) = x_2.$$

We have set here $l(\sigma) = l(M(\sigma))$ and assumed that ε is small enough to ensure that

$$|X_1(\sigma^+) - X_1(\sigma^-)| < 1 \quad \forall \sigma^- \in [\sigma_j^-(\varepsilon), \sigma_j].$$

The example in Fig. 8 suggests that the subspace of $H^{1,0}(\Omega)$ of functions independent of x_1

$$(2.15) \qquad V_1 = \left\{ v \in H^{1,0}(\Omega(\sigma_j, \varepsilon)) \Big/ \frac{\partial v}{\partial x_1} = 0 \right\}$$

has a particular role. Indeed, let us use the following decomposition of $H^{1,0}(\Omega(\sigma_j, \varepsilon))$:

$$H^{1,0}(\Omega(\sigma_j, \varepsilon)) = V_1 + V_1^\perp,$$

where $V_1^\perp$ is nothing but the space of functions in $H^{1,0}$ whose mean value along each segment $[M(\sigma), M(\sigma^+)]$ is equal to zero

$$v \in V_1^\perp \Leftrightarrow \text{a.e. } x_2 = X_2(\sigma^-) = X_2(\sigma^+) \Rightarrow \int_{X_1(\sigma^-)}^{X_1(\sigma^+)} u(x_1, x_2)\, dx_1 = 0.$$

For technical reasons, we shall use another decomposition (nonorthogonal) of $H^{1,0}(\Omega(\sigma_j, \varepsilon))$:

$$(2.16) \qquad H^{1,0}(\Omega(\sigma_j, \varepsilon)) = V_1 + V_2,$$

where V_2 is defined by

$$(2.17) \qquad v \in V_2 \Leftrightarrow \text{p.p. } x_2 = X_2(\sigma^-) = X_2(\sigma^+) \Rightarrow v(M(\sigma^-)) + v(M(\sigma^+)) = 0.$$

Note that the trace of a function in V_1 is necessarily (Γ, x_1)-even while the trace of a function in V_2 is by construction (Γ, x_1)-odd.

LEMMA 3. *The trace mapping* $\gamma_0: D(\overline{\Omega(\sigma_j, \varepsilon)}) \to L^2(\Gamma_1^j(\varepsilon))$ *can be uniquely extended to a linear, continuous, and surjective map from* $H^{1,0}(\Omega(\sigma_j, \varepsilon))$ *onto* $T(\Gamma_1^j(\varepsilon), \Omega(\sigma_j, \varepsilon))$. *More precisely,* γ_0 *is linear, continuous, and surjective from* V_1 *onto* $L^2_{\text{even}}(\Gamma_1^j(\varepsilon), l|n_1|\, d\sigma)$ *and from* V_2 *onto* $L^2_{\text{odd}}(\Gamma_1^j(\varepsilon), l^{-1}|n_1|\, d\sigma)$.

Proof.

(i) *Traces of functions in* V_1. Let v be a function in V_1; its trace on $\Gamma_1^j(\varepsilon)$ is given by

$$\phi(\sigma^-) = \phi(\sigma^+) = v(x_1, x_2) \quad \text{for } x_2 = X_2(\sigma^-) = X_2(\sigma^+). \tag{2.18}$$

The norm of v in $H^{1,0}(\Omega(\sigma_j, \varepsilon))$ coincides with its L^2-norm:

$$\|v\|^2_{H^{1,0}} = \int_{X_2(\sigma_j)}^{X_2(\sigma_j)+\varepsilon} \left(\int_{X_1(\sigma^-)}^{X_1(\sigma^+)} |v(x_1, x_2)|^2\, dx_1 \right) dx_2. \tag{2.19}$$

On the other hand, we have

$$\int_{X_1(\sigma^-)}^{X_1(\sigma^+)} |v(x_1, x_2)|^2\, dx_1 = l(\sigma^-)|\phi(\sigma^-)|^2 = l(\sigma^+)|\phi(\sigma^+)|^2. \tag{2.20}$$

With the help of the two following changes of variables:

$$x_2 = X_2(\sigma^-), \qquad \sigma^- \in [\sigma_j^-(\varepsilon), \sigma_j],$$
$$x_2 = X_2(\sigma^+), \qquad \sigma^+ \in [\sigma_j, \sigma_j^+(\varepsilon)],$$

we obtain the two equalities

$$\begin{aligned} \|v\|^2_{H^{1,0}} &= \int_{\sigma_j^-(\varepsilon)}^{\sigma_j} |\phi(\sigma^-)|^2 l(\sigma^-)|n_1(\sigma^-)|\, d\sigma^-, \\ \|v\|^2_{H^{1,0}} &= \int_{\sigma_j}^{\sigma_j^+(\varepsilon)} |\phi(\sigma^+)|^2 l(\sigma^+)|n_1(\sigma^+)|\, d\sigma^+, \end{aligned} \tag{2.21}$$

which we can add to finally obtain

$$\|\phi\|^2_{l|n_1|, \Gamma_1^j(\varepsilon)} = 2\|v\|^2_{H^{1,0}}, \tag{2.22}$$

from which both the continuity and the bijectivity of γ_0 from V_1 onto $L^2_{\text{even}}(\Gamma_1^j(\varepsilon), l|n_1|\, d\sigma)$ follow immediately.

(ii) *Traces of functions in* V_2. Let v be in V_2. For almost every x_2 in $]X_2(\sigma_j), X_2(\sigma_j)+\varepsilon[$, the function $x_1 \to v(x_1, x_2)$ belongs to $H^1(]X_1(\sigma^-), X_1(\sigma^+)[)$ and satisfies, if $x_2 = X_2(\sigma^-) = X_2(\sigma^+)$,

$$v(X_1(\sigma^-), x_2) + v(X_1(\sigma^+), x_2) = 0.$$

Therefore, as $H^1 \to C^0$ in dimension 1, we know that

$$\text{p.p. } x_2 = X_2(\sigma^-) = X_2(\sigma^+),$$
$$\exists x_0 = X_0(\sigma^-) = X_0(\sigma^+) \in]X_1(\sigma^-), X_1(\sigma^+)[/ v(x_1, x_0) = 0.$$

Thus, if we set $\phi = \gamma_0 v$, we can write

$$\begin{aligned} \phi(\sigma^-) &= -\int_{X_1(\sigma^-)}^{X_0(\sigma^-)} \frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^-))\, dx_1, \\ \phi(\sigma^+) &= -\int_{X_0(\sigma^+)}^{X_1(\sigma^+)} \frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^+))\, dx_1. \end{aligned} \tag{2.23}$$

By the Cauchy-Schwarz inequality, using the fact that $|X_0(\sigma^-)-X_1(\sigma^-)|$ and $|X_0(\sigma^+)-X_1(\sigma^+)|$ are bounded by $l(\sigma^-)=l(\sigma^+)$, we deduce the inequalities

$$\begin{aligned}
&|\phi(\sigma^-)|^2 \leqq l(\sigma^-)\int_{X_1(\sigma^-)}^{X_1(\sigma^+)}\left|\frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^-))\right|^2 dx_1,\\
&|\phi(\sigma^+)|^2 \leqq l(\sigma^+)\int_{X_1(\sigma^-)}^{X_1(\sigma^+)}\left|\frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^+))\right|^2 dx_1.
\end{aligned}\tag{2.24}$$

On the other hand, using successively two change of variables,

$$\begin{aligned}
&(x_1, \sigma^-)\rightarrow(x_1, x_2)=(x_1, X_2(\sigma^-)), \qquad \sigma^-\in[\sigma_j^-(\varepsilon), \sigma_j],\\
&(x_1, \sigma^+)\rightarrow(x_1, x_2)=(x_1, X_2(\sigma^+)), \qquad \sigma^+\in[\sigma_j, \sigma_j^+(\varepsilon)],
\end{aligned}$$

with respective Jacobian $|n_1(\sigma^-)|$ and $|n_1(\sigma^+)|$, we obtain two expressions for the L^2-norm of $\partial v/\partial x_1$:

$$\begin{aligned}
\int_{\Omega(\sigma_j,\varepsilon)}\left|\frac{\partial v}{\partial x_1}\right|^2 dx &= \int_{\sigma_j^-(\varepsilon)}^{\sigma_j}\left(\int_{X_1(\sigma^-)}^{X_1(\sigma^+)}\left|\frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^-))\right|^2 dx_1\right)|n_1(\sigma^-)|\,d\sigma^-\\
&= \int_{\sigma_j}^{\sigma_j^+(\varepsilon)}\left(\int_{X_1(\sigma^-)}^{X_1(\sigma^+)}\left|\frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^+))\right|^2 dx_1\right)|n_1(\sigma^+)|\,d\sigma^+.
\end{aligned}\tag{2.25}$$

Consequently, from (2.24) and (2.25), we deduce

$$\begin{aligned}
&\int_{\sigma_j^-(\varepsilon)}^{\sigma_j}|\phi(\sigma^-)|^2\frac{|n_1(\sigma^-)|}{|l(\sigma^-)|}\,d\sigma^- \leqq \int_{\Omega(\sigma_j,\varepsilon)}\left|\frac{\partial v}{\partial x_1}\right|^2 dx,\\
&\int_{\sigma_j}^{\sigma_j^+(\varepsilon)}|\phi(\sigma^+)|^2\frac{|n_1(\sigma^+)|}{|l(\sigma^+)|}\,d\sigma^+ \leqq \int_{\Omega(\sigma_j,\varepsilon)}\left|\frac{\partial v}{\partial x_1}\right|^2 dx.
\end{aligned}\tag{2.26}$$

Adding these two inequalities, we obtain

$$\|\phi\|^2_{l^{-1}|n_1|,\Gamma_1^j(\varepsilon)} \leqq 2\|v\|^2_{H^{1,0}},\tag{2.27}$$

which proves the continuity of γ_0 from V_2 on to the space $L^2_{\text{odd}}(\Gamma_1^j(\varepsilon), l^{-1}|n_1|\,d\sigma)$ to consider the function v defined by (see Fig. 12)

$$\begin{aligned}
&\partial^2 v/\partial x_1^2=0 \quad \text{in } \Omega(\sigma_j, \varepsilon),\\
&v(X_1(\sigma), X_2(\sigma))=\phi(\sigma) \qquad \sigma\in[\sigma_j^-(\varepsilon), \sigma_j^+(\varepsilon)].
\end{aligned}\tag{2.28}$$

Then a simple calculation gives

$$\begin{aligned}
&\int_{\Omega(\sigma_j,\varepsilon)}|v|^2\,dx=\frac{1}{3}\int_{\Gamma_1^j(\varepsilon)}|\phi(\sigma)|^2|n_1(\sigma)|l(\sigma)\,d\sigma,\\
&\int_{\Omega(\sigma_j,\varepsilon)}\left|\frac{\partial v}{\partial x}\right|^2 dx=\int_{\Gamma_1^j(\varepsilon)}|\phi(\sigma)|^2\frac{|n_1(\sigma)|}{l(\sigma)}\,d\sigma,
\end{aligned}\tag{2.29}$$

which completes the proof of the lemma.

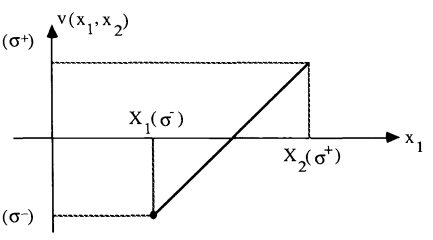


FIG. 12. *The solution of* (2.28).

Now to prove Theorem 1, the main point is to prove the estimate

$$\|\gamma_0 u\|_{T(\Gamma_1,\Omega)} \leqq C\|u\|_{H^{1,0}(\Omega)}$$

for any smooth function u. For this we proceed by localization with the help of a partition of unity. In this way we are able to write u as a finite sum of functions u_k, each of them having compact support, to which we can apply either Corollary 2 or Lemma 3. The difficulties are purely technical and we shall omit the details.

2.2. The proof of Theorem 2. To show that the limit of $\int_{\Gamma\backslash\Gamma_1(\delta)} \phi\psi n_1\, d\sigma$ exists when δ tends to zero, we first write, for $\delta < \varepsilon$,

$$\int_{\Gamma\backslash\Gamma_1(\delta)} \phi\psi n_1\, d\sigma = \int_{\Gamma\backslash\Gamma_1(\varepsilon)} \phi\psi n_1\, d\sigma + \sum_{j\in J_0} \int_{\Gamma_1^j(\varepsilon)\backslash\Gamma_1^j(\delta)} \phi\psi n_1\, d\sigma. \tag{2.30}$$

The only difficulty consists in proving that the limit

$$\lim_{\delta\searrow 0} \int_{\Gamma_1^j(\varepsilon)\backslash\Gamma_1^j(\delta)} \phi\psi n_1\, d\sigma \tag{2.31}$$

exists when j belongs to J_0. Indeed, the integral

$$\int_{\Gamma_1^j(\varepsilon)} \phi\psi n_1\, d\sigma$$

is not defined in the Lebesgue sense because of the possible singularities of ϕ and ψ near $\sigma = \sigma_j$. To overcome this difficulty we write:

$$\phi = \phi_0 + \phi_e, \qquad \psi = \psi_0 + \psi_e,$$

where $((\phi_0, \phi_e), (\psi_0, \psi_e)) \in \{L^2_{\text{odd}}(\Gamma_1^j(\varepsilon), l^{-1}|n_1|\, d\sigma) \times L^2_{\text{even}}(\Gamma_1^j(\varepsilon), l|n_1|\, d\sigma)\}^2$. We have, setting $\Gamma_1^j(\varepsilon, \delta) = \Gamma_1^j(\varepsilon)\backslash\Gamma_1^j(\delta)$,

$$\begin{aligned}\int_{\Gamma_1^j(\varepsilon,\delta)} \phi\psi n_1\, d\sigma &= \int_{\Gamma_1^j(\varepsilon,\delta)} (\phi_0(\sigma)\psi_0(\sigma) + \phi_e(\sigma)\psi_e(\sigma))\, d\sigma \\ &\quad + \int_{\Gamma_1^j(\varepsilon,\delta)} (\phi_0(\sigma)\psi_e(\sigma) + \phi_e(\sigma)\psi_0(\sigma))\, d\sigma.\end{aligned} \tag{2.32}$$

To estimate the second term of the right-hand side of (2.32), we use Cauchy–Schwarz's inequality:

$$\begin{aligned}&\int_{\Gamma_1^j(\varepsilon,\delta)} |\phi_o(\sigma)\psi_e(\sigma) + \phi_e(\sigma)\psi_o(\sigma)|\,|n_1(\sigma)|\, d\sigma \\ &\quad \leqq \|\phi_0\|_{\Gamma_1^j(\varepsilon),l^{-1}|n_1|}\|\psi_e\|_{\Gamma_1^j(\varepsilon),l|n_1|} + \|\phi_e\|_{\Gamma_1^j(\varepsilon),l|n_1|}\|\psi_0\|_{\Gamma_1^j(\varepsilon),l^{-1}|n_1|}.\end{aligned} \tag{2.33}$$

We can thus apply the dominated convergence theorem of Lebesgue and Cauchy–Schwarz's inequality to deduce

$$\begin{aligned}&\lim_{\delta\searrow 0} \int_{\Gamma_1^j(\varepsilon,\delta)} (\phi_o(\sigma)\psi_e(\sigma) + \phi_e(\sigma)\psi_o(\sigma))\, d\sigma \\ &\quad = \int_{\Gamma_1^j(\varepsilon)} (\phi_o(\sigma)\psi_e(\sigma) + \phi_e(\sigma)\psi_o(\sigma))\, d\sigma\end{aligned} \tag{2.34}$$

$$\begin{aligned}&\int_{\Gamma_1^j(\varepsilon)} (\phi_o(\sigma)\psi_e(\sigma) + \phi_e(\sigma)\psi_o(\sigma))\, d\sigma \\ &\quad \leqq \|\phi\|_{T(\Gamma_1^j(\varepsilon),\Omega(\sigma_j,\varepsilon))}\|\psi\|_{T(\Gamma_1^j(\varepsilon),\Omega(\sigma_j,\varepsilon))}\end{aligned} \tag{2.35}$$

For the first term of the right-hand side of (2.32), we remark that the function $\theta(\sigma) = \phi_o(\sigma)\psi_o(\sigma) + \phi_e(\sigma)\psi_e(\sigma)$ is (Γ, x_1)-even. Now suppose that for instance $X_2(\sigma_j)$

is a local minimum of $X_2(\sigma)$; with a suitable orientation of (Γ) we can assume that (see Fig. 13)

$$n_1(\sigma)=\frac{dX_2}{d\sigma}(\sigma)<0 \quad \text{if } \sigma\in[\sigma_j^-(\varepsilon), \sigma_j],$$

$$n_1(\sigma)=\frac{dX_2}{d\sigma}(\sigma)>0 \quad \text{if } \sigma\in[\sigma_j, \sigma_j^+(\varepsilon)].$$

We then write

$$\int_{\Gamma_1^j}\theta(\sigma)n_1(\sigma)\,d\sigma=\int_{\sigma_j^-(\varepsilon)}^{\sigma_j^-(\delta)}\theta(\sigma)n_1(\sigma)\,d\sigma+\int_{\sigma_j^+(\delta)}^{\sigma_j^+(\varepsilon)}\theta(\sigma)n_1(\sigma)\,d\sigma.$$

With the following changes of variable:

$$x_2=X_2(\sigma^-) \quad \text{if } \sigma^-\in[\sigma_j^-(\varepsilon), \sigma_j^-(\delta)],$$

$$x_2=X_2(\sigma^+) \quad \text{if } \sigma^+\in[\sigma_j^+(\delta), \sigma_j^+(\varepsilon)],$$

we obtain, taking into account the fact that the sign of $n_1(\sigma)$ changes from $[\sigma_j^-(\varepsilon), \sigma_j^-(\delta)]$ to $[\sigma_j^+(\delta), \sigma_j^+(\varepsilon)]$,

$$\int_{\Gamma_1^j(\varepsilon,\delta)}\theta(\sigma)n_1(\sigma)\,d\sigma=\int_{X_2(\sigma_j)+\delta}^{X_2(\sigma_j)+\varepsilon}[\theta(\sigma^+)-\theta(\sigma^-)]\,dx_2.$$

But, as θ is (Γ, x_1)-even, $\theta(\sigma^+)=\theta(\sigma^-)$, so that we have:

$$\forall\delta<\varepsilon, \quad \int_{\Gamma_1^j(\varepsilon,\delta)}(\phi_0(\sigma)\psi_o(\sigma)+\phi_e(\sigma)\psi_e(\sigma))\,d\sigma=0. \tag{2.36}$$

Joining this result to (2.34) shows that limit (2.31) does exist and is given by

$$\text{v.p. } x_2\int_{\Gamma_1^j(\varepsilon)}\phi(\sigma)\psi(\sigma)n_1(\sigma)\,d\sigma=\int_{\Gamma_1^j(\varepsilon)}(\phi_o(\sigma)\psi_e(\sigma)+\phi_e(\sigma)\psi_o(\sigma))\,d\sigma.$$

Moreover, the inequality (2.35) shows that the bilinear form

$$(\phi, \psi)\to\text{v.p. } x_2\int_{\Gamma_1^j(\varepsilon)}\phi(\sigma)\psi(\sigma)n_1(\sigma)\,d\sigma$$

is continuous on $T(\Gamma_1^j(\varepsilon), \Omega(\sigma_j, \varepsilon))$.

The general result is then a consequence of the equality (2.30). To obtain formula (1.14), it suffices to take the limit in Green's formula (1.12) written with a sequence

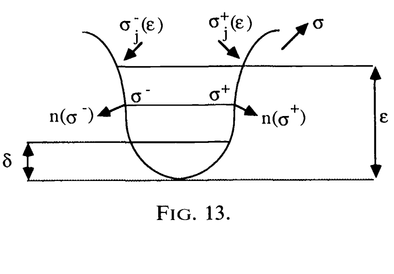


FIG. 13.

$(u_\varepsilon, v_\varepsilon)$ in $D(\bar{\Omega})^2$ converging to (u, v) in $H^{1,0}(\Omega)$ when $\varepsilon \to 0$, and to use the continuity of the bilinear forms

$$(u, v) \to \int_\Omega \left(\frac{\partial u}{\partial x_1} v + u \frac{\partial v}{\partial x_1}\right) dx,$$

$$(\phi, \psi) \to \text{v.p. } x_2 \int_{\Gamma_1} \phi(\sigma)\psi(\sigma) n_1(\sigma)\, d\sigma$$

in $H^{1,0}(\Omega)$ and $T(\Gamma_1, \Omega)$, respectively. □

3. Generalization to the spaces $W_p^{1,0}(\Omega)$ and $H^{S,0}(\Omega)$.

3.1. The trace theorem in $W_p^{1,0}(\Omega)$. For any $1 < p < +\infty$, we define the space $W_p^{1,0}(\Omega)$ by

$$W_p^{1,0}(\Omega) = \left\{ u \in L^p(\Omega) \Big/ \frac{\partial u}{\partial x_1} \in L^p(\Omega) \right\}, \tag{3.1}$$

which is a Banach space equipped with the norm

$$\|u\|_{W_p^{1,0}} = \left(\|u\|_{L^p}^p + \left\| \frac{\partial u}{\partial x_1} \right\|_{L^p}^p \right)^{1/p}. \tag{3.2}$$

With the notation and definitions of § 1, we define

$$T_p(\Gamma_1, \Omega) = L^p_{\text{odd}}(\Gamma_1; l^{1-p}|n_1|) \oplus L^p_{\text{even}}(\Gamma_1; l|n_1|) \tag{3.3}$$

where the weighted L^p-spaces $L^p_{\text{odd}}(\Gamma_1^j(\varepsilon), m)$ and $L^p_{\text{even}}(\Gamma_1^j(\varepsilon), m)$ are defined as $L^2_{\text{odd}}(\Gamma_1^j(\varepsilon), m)$ and $L^2_{\text{even}}(\Gamma_1^j(\varepsilon), m)$ (see § 1.1) by simply replacing 2 by p.

Of course, $T_p(\Gamma_1, \Omega)$ is a Banach space for the norm ($\phi = \phi_o + \phi_e$):

$$\|\phi\|_{T_p(\Gamma_1,\Omega)} = \|\phi_o\|_{L^p(\Gamma_1, l^{1-p}|n_1|)} + \|\phi_e\|_{L^p(\Gamma_1, l|n_1|)} \tag{3.4}$$

and, for ϕ in $L^p(C, m)$, $C \subset \Gamma$:

$$\|\phi\|_{L^p(C,m)} = \left(\int_C |\phi(\sigma)|^p m(\sigma)\, d\sigma \right)^{1/p}. \tag{3.5}$$

We can now state our two results in Theorems 3 and 4.

THEOREM 3. *The trace mapping $\gamma_0: D(\bar{\Omega}) \to L^p(\Gamma_1)$ can be uniquely extended to a linear, continuous and surjective map, still denoted by γ_0, from $W_p^{1,0}(\Omega)$ onto $T_p(\Gamma_1, \Omega)$.*

Proof of Theorem 3. The proof is very similar to that of Theorem 2 (§ 2). We simply point out the differences with the L^2-case.

Step 1. *Case where $n_1(\sigma)$ has a constant sign.* Take $v = |u|^{p-2}u$ in Green's formula (1.12) and use Holder's inequality instead of the Cauchy-Schwarz inequality to get the extension and continuity results. The surjectivity result can be obtained exactly as in the two-dimensional case.

Step 2. *An intermediate result.* There is no significant difference with the L^2-case.

Step 3. *The general case.* As in the L^2-case, only the strict outgoing x_2-extremal points have a role. We must study the space $W_p^{1,0}(\Omega(\sigma_j, \varepsilon))$, which we break down as follows:

$$W_p^{1,0}(\Omega(\sigma_j, \varepsilon)) = V_{1,p} + V_{2,p},$$

where $V_{1,p}$ and $V_{2,p}$ are defined as V_1 and V_2 (see § 2.1) by simply replacing 2 by p. The specific result concerning $W_p^{1,0}(\Omega)$ can be stated as in the following lemma.

LEMMA 4. *The trace mapping γ_0 extends in a unique way to a linear continuous and surjective map from $W_p^{1,0}(\Omega(\sigma_j, \varepsilon))$ onto $T_p(\Gamma_1^j(\varepsilon), \Omega(\sigma_j, \varepsilon))$. γ_0 is bijective and continuous from $V_{1,p}$ onto $L^p_{\text{even}}(\Gamma_1^j(\varepsilon), l|n_1|)$ and from $V_{2,p}$ onto $L^p_{\text{odd}}(\Gamma_1^j\varepsilon), l^{1-p}|n_1|)$.*

Proof. We only point out the differences with the proof of Lemma 3 (§ 2.1).

(i) *Traces of functions in $V_{1,p}$.* The L^p and $W_p^{1,0}$ norms of such functions coincide. With the notation of the proof of Lemma 3, we have

$$\int_\Omega |v(x_1, x_2)|^p \, dx = \int_{\sigma_j-\varepsilon}^{\sigma_j} |\phi(\sigma)|^p l(\sigma)|n_1(\sigma)| \, d\sigma$$

$$= \int_{\sigma_j}^{\sigma_j+\varepsilon} |\phi(\sigma)|^p l(\sigma)|n_1(\sigma)| \, d\sigma,$$

from which we deduce the identity

$$\|\gamma_0 v\|_{L^p(\Gamma_1^j(\varepsilon), l|n_1|)} = 2^{-1/p} \|v\|_{W_p^{1,0}};$$

from this the first part of the lemma is a direct consequence.

(ii) *Traces of functions in $V_{2,p}$.* Keeping the notation of the proof of Lemma 3, we write

$$\phi(\sigma^-) = -\int_{X_1(\sigma^-)}^{X_0(\sigma^-)} \frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^-)) \, dx_1, \qquad \sigma^- \in [\sigma_j^-(\varepsilon), \sigma_j],$$

$$\phi(\sigma^+) = \int_{X_0(\sigma^+)}^{X_1(\sigma^+)} \frac{\partial v}{\partial x_1}(x_1, X_2(\sigma^+)) \, dx_1, \qquad \sigma^+ \in [\sigma_j, \sigma_j^+(\varepsilon)].$$

Using Holder's inequality we get, if $1/p + 1/q = 1$,

$$|\phi(\sigma^-)| \leqq l(\sigma^-)^{1/q} \left(\int_{X_1(\sigma^-)}^{X_1(\sigma^+)} \left| \frac{\partial v}{\partial x_1}(x_1, X_2) \right|^p d\sigma \right)^{1/p},$$

$$|\phi(\sigma^+)| \leqq l(\sigma^+)^{1/q} \left(\int_{X_1(\sigma^-)}^{X_1(\sigma^+)} \left| \frac{\partial v}{\partial x_1}(x_1, X_2) \right|^p d\sigma \right)^{1/p}.$$

Following the proof of Lemma 3, we easily get the two inequalities

$$\int_{\sigma_j^-(\varepsilon)}^{\sigma_j} |\phi(\sigma^-)|^p l(\sigma^-)^{-p/q} \, d\sigma^- \leqq \int_{\Omega(\sigma_j, \varepsilon)} \left| \frac{\partial v}{\partial x_1} \right|^p dx,$$

$$\int_{\sigma_j}^{\sigma_j^+(\varepsilon)} |\phi(\sigma^+)|^p l(\sigma^+)^{-p/q} \, d\sigma^+ \leqq \int_{\Omega(\sigma_j, \varepsilon)} \left| \frac{\partial v}{\partial x_1} \right|^p dx.$$

As $-p/q = 1 - p$, we finally get

$$\int_{\Gamma_1^j(\varepsilon)} |\phi(\sigma)|^p l(\sigma)^{1-p} \, d\sigma \leqq 2\|v\|^p_{W_0^{1,p}},$$

which proves the continuity of γ_0 from $V_{2,p}$ in $L^p_{\text{odd}}(\Gamma_1^j(\varepsilon), l^{1-p}|n_1|)$. To get the surjectivity result, we take exactly the same extension (2.28) as in the L^2-case.

Explicit calculations give

$$\int_{\Omega(\sigma_j, \varepsilon)} |v|^p \, dx = \frac{2^{-p}}{p+1} \int_{\Gamma_1^j(\varepsilon)} |\phi(\sigma)|^p l(\sigma)|n_1(\sigma)| \, d\sigma,$$

$$\int_{\Omega(\sigma_j, \varepsilon)} \left| \frac{\partial v}{\partial x_1} \right|^p dx = 2^{-p} \int_{\Gamma_1^j(\varepsilon)} |\phi(\sigma)|^p l(\sigma)|n_1(\sigma)| \, d\sigma,$$

which proves that the map $\phi \to v$ defined by (2.28) is a continuous extension operator from $L^2_{\text{odd}}(\Gamma_1^j(\varepsilon), l^{1-p}|n_1|)$ in $V_{2,p}$. This completes the proof of Lemma 4. □

From steps 1, 2, and 3, it is then easy to complete the proof of Theorem 3. □

THEOREM 4. *Let (p, q) in $[1, \infty)^2$ such that $1/p + 1/q = 1$.*

(i) *For any (ϕ, ψ) in $T_p(\Gamma_1, \Omega) \times T_q(\Gamma_1, \Omega)$ the limit*

$$(3.6) \qquad \lim_{\delta \searrow 0} \int_{\Gamma_1 \setminus \Gamma_1(\delta)} \phi\psi n_1 \, d\sigma \overset{\text{def}}{=} \text{v.p. } x_2 \int_{\Gamma_1} \phi\psi n_1 \, d\sigma$$

exists and the bilinear form $(\phi, \psi) \to \text{v.p. } x_2 \int_{\Gamma_1} \phi\psi n_1 \, d\sigma$ is continuous.

(ii) *We have the following Green's formula:*

$$(3.7) \qquad \begin{gathered} \forall (u, v) \in W_p^{1,0}(\Omega) \times W_q^{1,0}(\Omega) \\ \int_\Omega \left(u \frac{\partial v}{\partial x_1} + v \frac{\partial u}{\partial x_1} \right) dx = \text{v.p. } x_2 \int_{\Gamma_1} \gamma_0 u \gamma_0 v n_1 \, d\sigma. \end{gathered}$$

Proof of Theorem 4. This proof is very similar to that of Theorem 2. The only point is to check that the spaces $T_p(\Gamma_1, \Omega)$ and $T_q(\Gamma_1, \Omega)$ are each other's dual. To obtain this property, take for instance ϕ in $L^p(\Gamma_1^j(\varepsilon), l|n_1|)$ and ψ in $L^q(\Gamma_1^j(\varepsilon), l^{1-q}|n_1|)$ such that ϕ and ψ are zero in a neighborhood of σ_j. We can write

$$\int_{\Gamma_1^j} \phi(\sigma)\psi(\sigma)|n_1(\sigma)| \, d\sigma = \int_{\Gamma_1^j} \phi(\sigma) l(\sigma)^{1/p} |n_1(\sigma)|^{1/p} \psi(\sigma) l(\sigma)^{-1/p} |n_1(\sigma)|^{1/q} \, d\sigma.$$

By Holder's inequality, we have

$$\left|\left|\int_{\Gamma_1^j(\varepsilon)} \phi(\sigma)\psi(\sigma)n_1(\sigma) \, d\sigma\right| \leqq \left(\int_{\Gamma_1^j(\varepsilon)} |\phi(\sigma)|^p l(\sigma) |n_1(\sigma)| \, d\sigma \right)^{1/p} \cdot \left(\int_{\Gamma_1^j(\varepsilon)} |\psi(\sigma)|^q l(\sigma)^{-q/p} |n_1(\sigma)| \, d\sigma \right)^{1/q}.\right.$$

Remarking that $-q/p = 1 - q$, we get, by density,

$$(3.8) \qquad \begin{gathered} \forall (\phi, \psi) \in L^p(\Gamma_1^j(\varepsilon), l|n_1| \, d\sigma) \times L^q(\Gamma_1^j(\varepsilon), l^{1-q}|n_1| \, d\sigma), \\ \left|\left|\int_{\Gamma_1^j(\varepsilon)} \phi(\sigma)\psi(\sigma)n_1(\sigma) \, d\sigma\right| \leqq \|\phi\|_{L^p(\Gamma_1^j(\varepsilon), l|n_1|)} \|\psi\|_{L^q(\Gamma_1^j(\varepsilon), l^{1-q}|n_1|)}.\right. \end{gathered}$$

Reciprocally, it is also easy to prove that

$$(3.9) \qquad \begin{gathered} \forall (\phi, \psi) \in L^p(\Gamma_1^j(\varepsilon), l^{1-p}|n_1| \, d\sigma) \times L^q(\Gamma_1^j(\varepsilon), l|n_1| \, d\sigma), \\ \left.\int_{\Gamma_1^j(\varepsilon)} \phi(\sigma)\psi(\sigma)n_1(\sigma) \, d\sigma\right| \leqq \|\phi\|_{L^p(\Gamma_1^j(\varepsilon), l^{1-p}|n_1|)} \|\psi\|_{L^q(\Gamma_1^j(\varepsilon), l^q|n_1|)}. \end{gathered}$$

With the help of (3.10) and (3.11) it is easy to conclude as in the proof of Theorem 2. □

3.2. The trace theorem in $H^{s,0}(\Omega)$. We first define the spaces $H^{m,0}(\Omega)$, when m is an integer:

$$(3.10) \qquad H^{m,0}(\Omega) = \left\{ v \in L^2(\Omega) / \forall k \leqq m, \frac{\partial^k v}{\partial x_1^k} \in L^2(\Omega) \right\},$$

$$(3.11) \qquad \|u\|^2_{H^{m,0}} = \sum_{k=1}^{m} \left\| \frac{\partial^k v}{\partial x_1^k} \right\|^2_{L^2},$$

and the space $H^{s,0}(\Omega)$ for $s>0$, with the help of the interpolation theory (see [8]) in Hilbert spaces:

$$\theta \in]0,1[\; H^{m+\theta,0}(\Omega)=[H^{m,0}(\Omega), H^{m+1,0}(\Omega)]_\theta. \tag{3.12}$$

We define by this way a family of Hilbert spaces. If we set

$$I_2=\{x_2\in \mathbb{R}/\exists x_1 \in \mathbb{R} \text{ s.t. } (x_1,x_2)\in\Omega\},$$
$$\Omega(x_2)=\{x_1\in\mathbb{R}/(x_1,x_2)\in\Omega\} \quad \text{when } x_2\in\Omega,$$

it is not difficult to see that the space $H^{s,0}(\Omega)$ can be characterized by:

$$H^{s,0}(\Omega)=\{v\in L^2(\Omega)/v(\cdot,x_2)\in L^2(I_2;H^s(\Omega(x_2)))\} \tag{3.13}$$

and that the norm of $H^{s,0}(\Omega)$ (given by the interpolation theory) is equivalent to the norm

$$\|v\|^2_{H^{s,0}}=\int_{I_2}\|v(\cdot,x_2)\|^2_{H^s(\Omega(x_2))}\,dx_2. \tag{3.14}$$

We now define the trace spaces $T^s(\Gamma_1,\Omega)$ for $s>\frac{1}{2}$ by

$$\begin{aligned} T^s(\Gamma_1,\Omega)&=L^2_{\text{even}}(\Gamma_1, l|n_1|)\oplus L^2_{\text{odd}}(\Gamma_1, l^{1-2s}|n_1|) && \text{if } \tfrac12<s\leqq 1,\\ T^s(\Gamma_1,\Omega)&=T^1(\Gamma_1,\Omega) && \text{if } s\geqq 1,\end{aligned} \tag{3.15}$$

which we equip with its natural Hilbert-space norm.

Our precise result is given in the following theorem.

THEOREM 5. *The trace mapping γ_0 extends uniquely to a linear, continuous mapping from $H^{s,0}(\Omega)$ onto $T^s(\Gamma_1,\Omega)$ for any real $s>\frac{1}{2}$.*

Proof of Theorem 5. The case $s\geqq 1$ is trivial and so we shall treat only the case $\frac12<s<1$.

As for Theorem 3, we shall only indicate the technical differences from the proof of Theorem 1.

Step 1. *Case where $n_1(s)$ has a constant sign.* We use the notation of the proof of Theorem 1. We use the trace theorem in $H^s(\mathbb{R}^+)$,

$$|u(0)|\leqq C(s)\|u\|_{H^s(\mathbb{R}^+)},$$

to write that, if $x_2=X_2(s)\in I_2$,

$$|u(X_1(\sigma),X_2(\sigma))|^2\leqq C(s)^2\|u(\cdot,x_2)\|^2_{H^s(X_1(s),+\infty)}.$$

Integrating over x_2 in I_2 and using the change of variable $x_2=X_2(s)$, we obtain

$$\int_{\Gamma_1}|\gamma_0u(\sigma)|^2|n_1(\sigma)|\,d\sigma\leqq C(s)\|u\|^2_{H^{s,0}},$$

which gives the unique extension and continuity results. The surjectivity result is obtained exactly as in the L^2 case, thanks to interpolation arguments.

Step 2. *An intermediate result.* To obtain the equivalent of Lemma 2, we apply the previous result to each of the functions $\Phi^\rho_p u$ (see the proof of Lemma 2) for which we have

$$\int_{\Gamma_1}|\Phi^\rho_p u(\sigma)|^2|n_1(\sigma)|\,d\sigma\leqq C(s)^2\|\Phi^\rho_p u\|^2_{H^{s,0}}.$$

We have (with the same notation as in the proof of Lemma 2)

$$\|\Phi_p^\rho u\|_{L^2} \leqq C(s)\|u\|_{L^2},$$
$$\|\Phi_p^\rho u\|_{H^{1,0}} \leqq C(s)(1+C(\rho))^{1/2}\|u\|_{H^{1,0}}.$$

By interpolation we easily get

$$\|\Phi_p^\rho u\|_{H^{s,0}} \leqq C(s)(1+C(\rho))^{s/2}\|u\|_{H^{s,0}}.$$

It is then easy to see that the result of Lemma 2 can be generalized to $H^{s,0}(\Omega)$ by simply replacing $(1+C(\rho))^{s/2}$ by $C(s)(1+C(\rho))^{s/2}$.

Step 3. *General case.* The main difference with the L^2 case comes from the study of the space $H^{s,0}(\Omega(\sigma_j, \varepsilon))$. The essential of the proof is contained in the following technical lemma.

LEMMA 5. *Let* $s > \frac{1}{2}$ *and let* u *in* $H^s(0, L)$ *be such that* $u(0)+u(L)=0$; *then we have the estimate*

$$|u(0)| \leqq C(s)L^{s-1/2}\|u\|_{H^s(0,L)}.$$

Proof. Let us use a particular representation of $H^s(0, L)$. Let us introduce the self-adjoint operator A in $L^2(0, L)$ defined by:

$$D(A) = \left\{u \in H^2(0, L) \Big/ \frac{du}{dx}(0) = \frac{du}{dx}(L) = 0\right\}, \tag{3.16}$$
$$Au = -\frac{d^2u}{dx^2}.$$

It is well known (see [11], for instance) that the form-domain of A is given by

$$D(A^{1/2}) = H^1(0, L) \tag{3.17}$$

and that the spectrum of A is

$$\sigma(A) = \left\{\frac{n^2\pi^2}{L^2}, n \in \mathbb{N}\right\}.$$

The corresponding basis of eigenfunctions of A is given by

$$w_n(x) = \sqrt{\frac{2}{L}}\cos\left(\frac{n\pi x}{L}\right).$$

Of course we have, for $u = \sum_{n=0}^{+\infty} u_n w_n$ in $D(A)$,

$$\|u\|^2_{L^2(0,L)} = \sum_{n=0}^{+\infty} |u_n|^2,$$

$$\|u\|^2_{H^1(0,L)} = \sum_{n=0}^{+\infty}\left(1+\frac{n^2\pi^2}{L^2}\right)^2 |u_n|^2.$$

From the interpolation theory, we know that:

$$D(A^{s/2}) = \left\{u = \sum_{n=0}^{+\infty} u_n w_n \Big/ \sum n^{2s}|u_n|^2 < +\infty\right\}.$$

But as $D(A^{1/2}) = H^1(0, L)$ by interpolation, $D(A^{s/2}) = H^s(0, L)$ and we can write

$$\|u\|^2_{H^s(0,L)} = \sum_{n=0}^{+\infty}\left(1+\frac{n^2\pi^2}{L^2}\right)^s |u_n|^2. \tag{3.18}$$

Now, let $u(x)$ be a smooth function such that $u(0)+u(L)=0$. There exists x_0 such that $u(x_0)=0$, so that we can write

$$u(0)=u(0)-u(x_0)=\sqrt{\frac{2}{L}}\sum_{n=1}^{+\infty}\left(1-\frac{\cos n\pi x_0}{L}\right)u_n.$$

Therefore,

$$|u(0)|\leqq 2\sqrt{\frac{2}{L}}\sum_{n=1}^{+\infty}|u_n|=2\sqrt{\frac{2}{L}}\sum_{n=1}^{+\infty}\frac{L^s}{\pi^s n^s}\,\frac{\pi^s n^s u_n}{L^s}.$$

By Cauchy–Schwartz's inequality, it follows that

$$|u(0)|^2\leqq 8L^{2s-1}\left(\sum_{n=1}^{+\infty}\frac{1}{n^{2s}}\right)\left(\sum_{n=1}^{+\infty}\frac{n^{2s}\pi^{2s}u_n^2}{L^{2s}}\right),$$

which gives the lemma with $C(s)=2\sqrt{2}(\sum_{n=1}^{+\infty}1/n^{2s})^{1/2}$. □

Now we break down $H^{s,0}(\Omega(\sigma_j,\varepsilon))$ as

$$H^{s,0}(\Omega(\sigma_j,\varepsilon))=V_1^s+V_2^s,$$

$$V_1^s=\left\{u\in H^s(\Omega(\sigma_j,\varepsilon))\Big/\frac{\partial u}{\partial x_1}=0\right\},$$

$$V_2^s=\{u\in H^s(\Omega(\sigma_j,\varepsilon))/\text{p.p. } x_2=X_2(\sigma^-)=X_s(\sigma^+),$$
$$u(M(\sigma^-))+u(M(\sigma^+))=0\}.$$

As the L^2 norm and the $H^{s,0}$ norm of a function in V_1^s coincide, it is easy to see that γ_0 is an isomorphism between V_1^s and $L^2_{\text{even}}(\Gamma_1^j(\varepsilon), l|n_1|)$.

With the help of Lemma 5, it is easy to prove that γ_0 is continuous from V_2^s in $L^2_{\text{odd}}(\Gamma_1^j(\varepsilon), l^{1-2s}|n_1|)$. Finally, the surjectivity result is obtained exactly with the extension operator defined by (2.28).

The complete result is proved by localization with the help of a partition of unity. □

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. Bamberger, B. Engquist, L. Halpern, and P. Joly, *Parabolic wave equation approximation in heterogeneous media*, SIAM J. Appl. Math., 48 (1988), pp. 99–128.

[3] ———, *Higher order wave equation approximation in heterogeneous media*, SIAM J. Appl. Math., 48 (1988), pp. 129–154.

[4] O. Besov, V. Il'Jin, and S. Nikol'ski, *Integral representations of functions and imbedding theorems*, Vol. 1 Scripta Ser. Math., Winston, WA, 1978.

[5] P. Joly, *Etude mathématique de l'approximation parabolique de l'équation des ondes en milieu stratifié*, Rapport INRIA 229, Institut National de Recherche en Informatique et en Automatique, 1984.

[6] ———, *Analyse numérique et mathématique de problèmes liés à la propagation d'ondres acoustiques, élastiques et électromagnétiques*, Thése d'Etat, Université Paris-Dauphine, 1987.

[7] ———, *Un théorème de traces dans un espace de Sobolev anisotrope*, Rapport INRIA 772, 1987.

[8] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications* Vols. I and II—Dunod, Paris, 1968.

[9] F. Nataf, *Approximation paraxiale pour les fluides incompressibles*, Thèse de Doctorat de l'Ecole Polytechnique, Paris, 1989.

[10] J. Necas, *Méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[11] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vols. II and IV, Academic Press, New York, 1981.

[12] J. E. Roberts and J. -M. Thomas, *Mixed and hybrid finite element methods*, Rapport INRIA 737, 1987.

[13] R. Temam, *Sur la convergence de la méthode des pas fractionnaires*, Thèse d'Etat, Université de Paris, 1971.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 821-835, July 1992

001

GLOBAL EXISTENCE OF WEAK SOLUTIONS FOR INTERFACE EQUATIONS COUPLED WITH DIFFUSION EQUATIONS*

YOSHIKAZU GIGA†, SHUN'ICHI GOTO‡, AND HITOSHI ISHII§

Abstract. A weak formulation for an interface dynamics coupled with a diffusion equation is introduced. A global-in-time weak solution is constructed for an arbitrary initial data under a periodic boundary condition. The result applies to the interface equation obtained as a certain singular limit of some reaction-diffusion systems including the activator-inhibitor model.

Key words. interface equation with diffusion equation, global existence, viscosity solution

AMS(MOS) subject classifications. 35K55, 35K57, 35K65

1. Introduction. In this paper we are concerned with interface equations coupled with diffusion equations. A typical example is formally obtained as a certain singular limit of a class of reaction diffusion systems [XYC]. Our main objective is to construct a global-in-time weak solution for the initial value problem of these interface equations.

Let $\Omega_\pm(t)$ be two disjoint open sets in $\mathbb{R}^n$ depending on time t. The complement of the union of $\Omega_+(t)$ and $\Omega_-(t)$ is called the interface and denoted by $\Gamma(t)$. To write down the equation we assume that the interface $\Gamma(t)$ is a smooth hypersurface so that $\Gamma(t)$ is the boundary of $\Omega_\pm(t)$. Let $V=V(t,x)$ denote the speed of $\Gamma(t)$ at $x\in\Gamma(t)$ in the normal direction $\vec{n}$ from $\Omega_+(t)$ to $\Omega_-(t)$. Let $\kappa(=\operatorname{div}\vec{n})$ denote ($n-1$ times) the mean curvature of $\Gamma(t)$ at $x\in\Gamma(t)$. We consider an interface equation for $\Gamma(t)$:

$$V=W(v)-c\kappa \quad \text{on } \Gamma(t) \tag{1.1}$$

coupled with a diffusion equation for $v=v(t,x)$:

$$v_t=D\Delta v+g_\pm(v), \qquad x\in\Omega_\pm(t), \quad t>0, \tag{1.2}$$

where $c\geqq 0$ and $D>0$ are constants. Here $g_\pm$ and W are given bounded continuous functions on $\mathbb{R}$. We also impose a condition that $v(t)=v(t,\cdot)$ is continuous in $\mathbb{R}^n$ with its first derivatives, i.e.,

$$v(t)=v(t,\cdot)\in C^1(\mathbb{R}^n) \quad \text{for } t>0. \tag{1.3}$$

Our goal is to construct a global solution of the initial value problem for interface equations coupled with diffusion equations—a typical example of which is (1.1)-(1.3). It is intrinsically difficult to construct a global solution $(\Omega_\pm(t), v(t))_{t\geqq 0}$ since $\Gamma(t)$ may have singularities in finite time. If v is a constant so that $g_\pm(v)=0$, (1.1)-(1.3) becomes

$$V=C-c\kappa, \tag{1.4}$$

where C is a constant. If $C=0$ and $c>0$, (1.4) becomes

$$V=-c\kappa, \tag{1.5}$$

* Received by the editors April 22, 1991; accepted for publication July 19, 1991.

† Department of Mathematics, Hokkaido University, Sapporo 060, Japan. The work of this author was partially supported by the Grant-in-Aid for Scientific Research 01740076, 01540092, Japan Ministry of Education, Science, and Culture.

‡ Department of Applied Science, Faculty of Engineering, Kyushu University 36, Fukuoka 812, Japan.

§ Department of Mathematics, Chuo University, Bunkyo-ku, Tokyo 112, Japan. The work of this author was partially supported by the Grant-in-Aid for Scientific Research 02640150, Japan Ministry of Education, Science, and Culture.

which is called the mean curvature flow equation. Even for (1.5) Grayson [Gr] gives an example of a barbell in $\mathbb{R}^3$ with a long, thin handle that actually pinches off in finite time. To track the whole evolution of interface we interpret $\Gamma(t)$ as a level set of viscosity solution of some second order evolution equations as in [CGG]. In fact Y.-G. Chen and the first two authors [CGG] constructed a *unique* global weak solution with arbitrary initial data for a class of interface equations, including (1.4) and (1.1), where v only depends on time (see [GG1] for interface equations that the theory in [CGG] applies to). At nearly the same time, Evans and Spruck [ES1] constructed the same solution but only for (1.5). Another formulation closely related to [CGG] and [ES1] is given in [S]. We refer to [ES2] and [GG2] for further development of the theory and references. We note that the idea of using level sets of viscosity solutions for $V=C$ is also found in an unpublished paper of Barles [B].

Although the interface equation admits a global weak solution, $\Gamma(t)$ may develop an interior (Remark 2.5). We introduce a generalized formulation of (1.2)-(1.3). For technical reasons we impose a periodic boundary condition. In this paper, we construct a global weak solution of the initial value problem for (1.1)-(1.3) with arbitrary initial data $(\Gamma(0), v(0,x))$, $v(0,x)\in C^2$ under the periodic boundary condition (Theorem 4.6). (We may assume $D=1$ without loss of generality.) For this purpose, for a given v, we construct a unique global weak solution for (1.1). The basic idea is the same in [CGG], but we are forced to use results in [GGIS] since v may depend on x as well as t. We solve (1.2)-(1.3) with the above v and $\Omega_\pm(t)$ determined by (1.1) with the initial condition. If we write the solution by w, we have a mapping $v\mapsto w$. Since our weak formulation forces us to interpret the mapping $v\mapsto w$ as a multivalued mapping, we use Kakutani (-Ky Fan's) fixed point theory (see [AF]) to get a global generalized solution as a fixed point of the mapping $v\mapsto w$. Our results apply to the system (1.2), (1.3) with (1.1) replaced by more general interface equations, including anisotropic motion (cf. [Gu1], [Gu2], [C]). We do not know the uniqueness of our solutions.

Let us mention some results on (1.1)-(1.3) that are related to ours. In [XYC], X.-Y. Chen constructed a unique local smooth solution for a smooth initial data $(\Gamma(0), v(0,x))$ in $\mathbb{R}^n$ when $c>0$. When $n=1$, the curvature term in (1.1) disappears. Hilhorst, Nishiura, and Mimura [HNM] constructed a global unique solution for (1.1)-(1.3) when the interface is a point and $n=1$ under the Neumann boundary condition. Their interface is C^1 in time. After this work was completed, we learned of the recent paper of X. Chen [XC2], which extends the local existence results [XYC] to the case $c=0$. Our result seems to be a first global result even for (1.1)-(1.3) with $c>0$ or $c=0$ when $n>1$.

Interface equations and reaction-diffusion equations are closely related (see [F]). Typical examples of the system (1.1)-(1.3) are formally provided as a singular limit of reaction-diffusion equations (see [OMK] and [XYC]). We will explain it more explicitly by following [XYC]. We consider a reaction-diffusion system describing the activator-inhibitor model:

$$u_t=\varepsilon\Delta u+\frac{1}{\varepsilon}f(u,v), \qquad x\in\mathbb{R}^n, \quad t>0, \tag{1.6}$$

$$v_t=D\Delta v+g(u,v), \qquad x\in\mathbb{R}^n, \quad t>0, \tag{1.7}$$

with

$$f(u,v)=f_0(u)-v, \qquad f_0(u)=u(1-u)(u-a),$$
$$g(u,v)=u-\gamma v,$$

where $\gamma > 0$, $0 < a < 1$, and ε is a small positive parameter. The zero set of f consists of three branches

$$\begin{aligned} u &= h_-(v) \quad \text{for } u < a_-, \\ u &= h_+(v) \quad \text{for } a_+ < u, \\ u &= h_0(v) \quad \text{for } a_- < u < a_+, \end{aligned}$$

where $a_- < a_+$ and $f_0'(a_-) = f_0'(a_+) = 0$. When $\varepsilon \to 0$, it is expected that u tends to $h_\pm(v)$ in some region $\Omega_\pm(t)$ in $\mathbb{R}^n$ since $h_\pm(v)$ is a stable zero of $u_t = f(u, v)$. From (1.6), it is also expected that the interface $\Gamma(t)$ moves by (1.1) with $c = \varepsilon$. Here $W(b)$ for b, $f_0(a_-) < b < f_0(a_+)$ is the speed of the travelling wave of

$$u_t = \Delta u + f(u, b)$$

and is given by

$$W(b) = \frac{1}{\sqrt{2}}(h_+(b) + h_-(b) - h_0(b))$$

(see Aronson and Weinberger [AW]). The equation (1.7) now becomes (1.2) as $\varepsilon \to 0$ by taking $g_\pm(v) = g(h_\pm(v), v)$. For more details we refer to [OMK] and [XYC] and references therein. We note that anisotropic interface equations are also derived by a singular limit of some reaction-diffusion equation [C].

Extensive literature exists on the behavior u^ε as $\varepsilon \downarrow 0$ in (1.6) and its relation to the solutions of interface equation when v is given and the space dimension $n = 1$ (see, e.g., [FH], [BK], [CP]). Recently, some results were extended to the case $n > 1$, where the curvature effect comes in. If v is a constant and $W(v) = 0$, (1.6) is called the Allen-Cahn equation, whose relation to (1.5) with $c > 0$ is rigorously analyzed by Bronsard and Kohn [BK] and DeMottoni and Schatzman [DS]. X. Chen [XC1] extended results of [DS] and simplified the argument. After this work was completed, we learned that X. Chen [XC2] derived (1.1)-(1.3) with $c = 0$ rigorously as a singular limit of (1.6)-(1.7). There is also an argument to interpret the case $c = \varepsilon > 0$ in [XC2]. His method is an extension of his work [XC1]. All results in [BK], [DS], [XC1], [XC2] assume that the solution of the interface equation is smooth enough to get the behavior of u^ε as $\varepsilon \downarrow 0$. Very recently we learned that Evans, Soner, and Songanidis [ESS] obtained the behavior of u^ε even after singularities appear on the interface for the Allen-Cahn equation.

In § 2 we solve a general interface equation including (1.1) for a given function v globally in time under a periodic boundary condition. In § 3 we give a generalized formulation of (1.2)-(1.3). In § 4 we state our main existence results and prove them by a fixed point argument. In the Appendix, we state a stability property of the viscosity solutions used in § 4.

After this work was completed, X.-Y. Chen kindly informed us that he had found another proof for global existence for (4.1), (4.2), (4.3′) with $c > 0$ without using a fixed point argument for multivalued mappings.

2. Interface equations. We consider interface equations under periodic boundary conditions. The periodic boundary condition is important because it is often used in numerical experiments. For $\alpha_i > 0$ $(1 \leqq i \leqq n)$ let R be a rectangle in $\mathbb{R}^n$ of the form

$$R = \{(x_1, \cdots, x_n) \in \mathbb{R}^n; 0 \leqq x_i \leqq \alpha_i, 1 \leqq i \leqq n\}.$$

We identify faces $x_i = 0$ and $x_i = \alpha_i (1 \leqq i \leqq n)$ of R to obtain an n-dimensional flat torus $\mathbb{T}$. Motion of interfaces in R under periodic boundary conditions is interpreted as the motion in $\mathbb{T}$. We consider $\mathbb{T}$ rather than $\mathbb{R}^n$ for later technical convenience because $\mathbb{T}$ is compact and has no boundary.

Let $\Omega_\pm(t)$ be an open set in $\mathbb{T}$ depending on time $t \geqq 0$ such that $\Omega_+(t) \cap \Omega_-(t) = \phi$. Let $\Gamma(t)$ denote the complement of $\Omega_+(t) \cup \Omega_-(t)$ in $\mathbb{T}$. Physically speaking, $\Gamma(t)$ is called an interface bounding two phases $\Omega_\pm(t)$ of material, e.g., solid and liquid region. Suppose that $\Gamma(t)$ is a smooth hypersurface, and let $\vec{n}$ denote the unit normal vector field pointing from $\Omega_+(t)$ to $\Omega_-(t)$. Let $V = V(t, x)$ denote the speed of $\Gamma(t)$ at $x \in \Gamma(t)$ in the direction $\vec{n}$. It is convenient to extend $\vec{n}$ to a vector field (still denoted by $\vec{n}$) on a tubular neighborhood of $\Gamma(t)$ such that $\vec{n}$ is constant in the normal direction of $\Gamma(t)$. The equation for $\Gamma(t)$ that we consider here is of the form

$$\begin{aligned} V &= \xi(t, x, \vec{n}, \nabla \vec{n}) \\ &:= \eta(\vec{n}, \nabla \vec{n}) + \omega(t, x, \vec{n}) \quad \text{on } \Gamma(t), \end{aligned} \tag{2.1}$$

where η and ω are given functions and ∇ stands for the spatial gradient in $\mathbb{T}$. A typical example is

$$V = -c \operatorname{div} \vec{n} + \omega(t, x), \tag{2.2}$$

where c is a nonnegative constant and ω is independent of $\vec{n}$. Equation (2.2) is called the mean curvature flow equation if $c > 0$ and $\omega \equiv 0$. The reason we consider general (2.1) is to include anisotropic motion as in [Gu1], [Gu2].

Next, we introduce a weak formulation for (2.1) following [CGG], [GG1]. For η we set

$$\begin{aligned} F_\eta(p, X) &:= -|p|\eta(-\bar{p}, -Q_{\bar{p}}(X)), \qquad \bar{p} = p/|p|, \\ Q_{\bar{p}}(X) &= R_{\bar{p}} X R_{\bar{p}} \quad \text{with } R_{\bar{p}} = I - \bar{p} \otimes \bar{p}, \end{aligned} \tag{2.3}$$

where $p \in \mathbb{R}^n \backslash \{0\}$ and $X \in \mathbb{S}_n$, the space of $n \times n$ real symmetric matrices. We also set

$$F_\xi(t, x, p, X) := F_\eta(p, X) - \omega(t, x, -\bar{p})|p|. \tag{2.4}$$

For example, a calculation shows

$$F_\eta(p, X) = -c \operatorname{trace}((I - \bar{p} \otimes \bar{p})X) \tag{2.5}$$

if

$$\eta(\vec{n}, \nabla \vec{n}) = -c \operatorname{div} \vec{n} \tag{2.6}$$

as in (2.2). The following definition of weak solutions for (2.1) is a variant of that in [CGG], [GG1]. For the definition of (viscosity) sub- and supersolutions and viscosity solutions; see, e.g., [GGIS].

Definition 2.1. Let $\{\Omega_\pm(t)\}_{0 \leqq t < T}$ be a one parameter family of open sets in $\mathbb{T}$ such that $\Omega_+(t) \cap \Omega_-(t) = \phi$. Suppose that there is a viscosity solution $u \in C([0, T) \times \mathbb{T})$ of

$$u_t + F_\xi(t, x, \nabla u, \nabla^2 u) = 0 \quad \text{in } (0, T) \times \mathbb{T} \tag{2.7}$$

such that

$$\Omega_\pm(t) = \{x \in \mathbb{R}^n;\ u(t, x) \gtrless 0\} \quad \text{for } 0 \leqq t < T. \tag{2.8}$$

We say $\{\Omega_\pm(t)\}_{0 \leqq t < T}$ is a *weak solution* of (2.1) in $(0, T)$ with initial data $\Omega_\pm(0)$. Here F_ξ is defined by (2.3)-(2.4).

Roughly speaking, if (2.1) is parabolic (not necessarily strictly parabolic), and η grows linearly in $\nabla\vec{n}$, then we can claim the unique global existence of weak solutions for (2.1) with given initial data $\Omega_\pm(0)$, provided that η and ω are continuous. If ω is independent of x and $\mathbb{T}$ is replaced by $\mathbb{R}^n$, the unique global existence is now well known if one of $\Omega_\pm(0)$ is bounded (cf. [CGG], [GG1]). We now list our assumptions on η and ω:

$$\begin{gathered}\eta \text{ is a real valued continuous function on the vector bundle}\\ E=\{(\bar{p}, Q_{\bar{p}}(X)); \bar{p}\in S^{n-1}, X\in \mathbb{S}_n\}\\ \text{over a unit sphere } S^{n-1}.\end{gathered} \tag{2.9}$$

$$\begin{gathered}\eta(-\bar{p}, -Q_{\bar{p}}(X)) \geqq \eta(-\bar{p}, -Q_{\bar{p}}(Y)) \quad \text{for } X\geqq Y,\ \bar{p}\in S^{n-1},\\ \text{where } \mathbb{S}_n \text{ is equipped with usual ordering.}\end{gathered} \tag{2.10}$$

$$\begin{gathered}\liminf_{\rho\downarrow 0} \rho \inf_{|\bar{p}|=1} \eta\left(-\bar{p}, \frac{I-\bar{p}\otimes\bar{p}}{\rho}\right) > -\infty,\\ \limsup_{\rho\downarrow 0} \rho \sup_{|\bar{p}|=1} \eta\left(-\bar{p}, \frac{-I+\bar{p}\otimes\bar{p}}{\rho}\right) < \infty.\end{gathered} \tag{2.11}$$

(2.12) ω is continuous from $[0, T)\times\mathbb{T}\times S^{n-1}$ to $\mathbb{R}$ with a bound on $|\nabla\omega|$.

All assumptions on η are found in [GG1]; (2.10) means that $-\eta$ is degenerate elliptic and (2.11) restricts the growth of η in $\nabla\vec{n}$. The only assumption for ω is (2.12).

THEOREM 2.2. *Assume* (2.9)-(2.12) *for η and ω. Let $\Omega_\pm(0)$ be mutually disjoint open sets in $\mathbb{T}$. Then there is a unique global weak solution $\{\Omega_\pm(t)\}_{0\leqq t<T}$ of* (2.1) *in* $(0, T)$ *with initial data $\Omega_\pm(0)$.* (*The case $T=\infty$ is included.*)

The basic idea of the proof is the same as [CGG, Thms. 6.8, 7.1]; see also [GG1] for the relation between assumptions on η and F_η. The major technical difference is that the comparison theorem in [CGG] does not apply to (2.7) because F_ξ depends on x. Instead we apply [GGIS, Thm. 4.1] to get a comparison principle for (2.7). For the reader's convenience, we state a version of the comparison principle which follows from [GGIS, Thm. 4.1] and give a brief proof of Theorem 2.2.

PROPOSITION 2.3. *Assume* (2.9)-(2.12). *Let u and v be, respectively,* (*viscosity*) *sub- and supersolutions of* (2.7). *Assume that u and v are, respectively, upper and lower semicontinuous functions on* $[0, T)\times\mathbb{T}$. *If*

$$u(0,x)\leqq v(0,x) \quad \textit{on } \mathbb{T},$$

then $u(t,x)\leqq v(t,x)$ *on* $[0, T)\times\mathbb{T}$.

Proof. To apply [GGIS, Thm. 4.1] we extend u, v, and ω periodically in space variables outside R and regard (2.7) as

$$u_t + F_\xi(t, x, \nabla u, \nabla^2 u) = 0 \quad \text{in } (0, T']\times\mathbb{R}^n,$$

where T' is an arbitrary positive number less than T.

We first check assumptions of equation in [GGIS]. By (2.9)-(2.10) we know F_η satisfies all assumptions on F in [GGIS, Thm. 4.1]. Except for the boundedness of $F_\eta(p, X)$ on a bounded set in $(\mathbb{R}^n\backslash\{0\})\times\mathbb{S}_n$, the proof is found in [GG1]. This boundedness of F_η can be proved similarly to the proof of [GG1, Lemma 3.5].

By (2.12) we see ω is continuous in $[0, T']\times\mathbb{R}^n\times S^{n-1}$ with a bound on $|\nabla\omega|$ so F_ξ satisfies the uniform continuity assumption in x of [GGIS, (F8)]; there is a modulus σ (i.e., σ: $[0,\infty)\to[0,\infty)$ is continuous, nondecreasing and $\sigma(0)=0$) such that

$$|F_\xi(t,x,p,X) - F_\xi(t,y,p,X)| \leqq \sigma(|x-y|(|p|+1))$$

for x, $y \in \mathbb{R}^n$, $t \in [0, T']$, $p \in \mathbb{R}^n \backslash \{0\}$, $X \in \mathbb{S}_n$. All other assumptions on F in [GGIS, Thm. 4.1] are fulfilled since ω satisfies (2.12) and F_η satisfies all assumptions on F.

Since u and v are extended periodically and R is bounded, it is not difficult to see that u and v satisfy all the assumptions of [GGIS, Thm. 4.1].

We now apply [GGIS, Thm. 4.1] to conclude $u \leqq v$ on $[0, T'] \times \mathbb{R}^n$. Since $T' < T$ is arbitrary, the proof is complete. □

Proof of Theorem 2.2. (*Uniqueness.*) Suppose that u, $v \in C([0, T) \times \mathbb{T})$ solves (2.7) such that

$$\Omega_\pm(0) = \{x \in \mathbb{T}, u(0, x) \gtrless 0\} = \{x \in \mathbb{T}; v(0, x) \gtrless 0\}.$$

By [CGG, Lemma 7.2] there is a continuous nondecreasing function $\theta : \mathbb{R} \to \mathbb{R}$ with $\theta(0) = 0$ such that

$$u(0, x) \leqq \theta(v(0, x)).$$

Since F_ξ is geometric, i.e.,

$$(2.13) \qquad \begin{gathered} F_\xi(t, x, \lambda p, \lambda X + \sigma p \otimes p) = \lambda F_\xi(t, x, p, X) \\ \text{for } \lambda > 0,\ \sigma \in \mathbb{R},\ t \in (0, T),\ x \in \mathbb{T},\ p \in \mathbb{R}^n \backslash \{0\},\ X \in \mathbb{S}_n, \end{gathered}$$

by [CGG, Thm. 5.2] we see $\theta(v(t, x))$ also solves (2.7). From Proposition 2.3, it follows $u \leqq \theta(v)$ on $[0, T) \times \mathbb{T}$. Thus, we observe that $u > 0$ implies $v > 0$ and $v < 0$ implies $u < 0$. A parallel argument yields the converse implication so $\Omega_\pm(t)$ is determined by $\Omega_\pm(0)$ and is independent of the choice of u. This proves the uniqueness of weak solutions.

(*Existence.*) For given $\Omega_\pm(0)$, we take $u_0(x) \in C(\mathbb{T})$ such that

$$\Omega_\pm(0) = \{x \in \mathbb{T}; u_0(x) \gtrless 0\}.$$

Since (2.11) is assumed, we may apply [CGG, Prop. 6.4] to (2.7) in $[0, T'] \times \mathbb{R}^n$ with periodic initial data and find sub- and supersolutions v_-, v_+ of (2.7) on $[0, T'] \times \mathbb{R}^n$ such that

$$\begin{gathered} v_\pm(0, x) = u_0(x) \quad \text{on } \mathbb{R}^n \\ v_-(t, x) \leqq u_0(x) \leqq v_+(t, x) \quad \text{on } [0, T'] \times \mathbb{R}^n, \end{gathered}$$

where $T' < T$. The dependence of x in F_ξ is allowed in [CGG, Prop. 6.4]. A trivial modification of the argument enables us to take v_-, v_+ as functions on $[0, T'] \times \mathbb{T}$.

Existence of $v_\pm$ yields a viscosity solution $u \in C([0, T'] \times \mathbb{T})$ of (2.7) with $u(0, x) = u_0(x)$ by Perron's method and Proposition 2.3. Since $T' < T$ is arbitrary and the solution is unique, we now obtain a weak solution $\{\Omega_\pm(t)\}_{0 \leqq t < T}$ for initial data $\Omega_\pm(0)$.

Note that the scaling property (2.13) is also used to construct $v_\pm$. □

Remark 2.4. The family $\{\Omega_+(\mathrm{t})\}$ is determined by $\Omega_+(0)$ and is independent of $\Omega_-(0)$. Indeed, if u solves (2.7) with (2.8), then $\theta(u)$ solves (2.7) for continuous nondecreasing $\theta : \mathbb{R} \to \mathbb{R}$ since F_ξ is geometric. Take $\theta(\sigma) = \sigma_+ = \max(\sigma, 0)$ to observe that $u_+ = \theta(u)$ solves (2.7). By (2.8) u_+ gives a weak solution $\{\Omega'_\pm(t)\}_{0 \leqq t < T}$ with initial data $(\Omega_+(0), \phi)$. By the definition of u_+, we see $\Omega'_+(t) = \Omega_+(t)$ and $\Omega'_-(t) = \phi$. We thus observe that $\Omega_+(t)$ is determined by $\Omega_+(0)$.

Remark 2.5. The interface $\Gamma(t)$ is defined by the complement of $\Omega_+(t) \cup \Omega_-(t)$ in $\mathbb{T}$. There is a chance that $\Gamma(t)$ develops an interior even if $\Gamma(0)$ is a smooth hypersurface in $\mathbb{T}$. For example, consider the equation $V = -1$ and

$$R = \{(x_1, x_2) \in \mathbb{R}^2; 0 \leqq x_1 \leqq 2, 0 \leqq x_2 \leqq 2\}.$$

Suppose that

$$\Omega_+(0)=\{x\in\mathbb{T};\, x_1\neq 1\}, \Omega_-(0)=\phi$$

so that $\Gamma(0)=\{x_1=1\}$. Then $\Omega_+(t)=\{x\in\mathbb{T};\, 0\leqq x_1\leqq 2, |x_1\leqq 1|>t\}$ and $\Omega_-(t)=\phi$. Indeed, equation (2.7) for this example is

$$u_t+|\nabla u|=0. \tag{2.14}$$

By the definition of viscosity solutions, we can check that

$$u(t,x)=\begin{cases}0 & \text{for } |x_1-1|\leqq t,\\ x_1-1-t & \text{for } x_1-1>t\ (x\in\mathbb{R}^2),\\ 1-x_1-t & \text{for } x_1-1<-t\end{cases} \tag{2.15}$$

is a viscosity solution of (2.14) on $(0,\infty)\times\mathbb{T}$. We now observe that $\Omega_\pm(t)$ is given by (2.8) with u of (2.15).

For the mean curvature flow equation

$$V=-\operatorname{div}\vec{n},$$

we do not know whether or not $\Gamma(t)$ develops an interior if $\Gamma(0)$ is a smooth hypersurface. As pointed out in [ES1], we know $\Gamma(t)$ may develop an interior if $\Gamma(0)$ has a singularity.

3. Diffusion equations across interfaces. This section gives a generalized formulation of

$$v_t=\Delta v+g_\pm(v) \quad \text{in } Q_T^\pm=\bigcup_{0<t<T}\{t\}\times\Omega_\pm(t), \tag{3.1}$$

$$v(t):=v(t,\cdot)\in C^1(\mathbb{T}) \quad \text{for } 0<t<T, \tag{3.2}$$

where

$$Q_T^\pm=\{(t,x)\in Q_T;\, u(t,x)\gtrless 0\}, \qquad Q_T=(0,T)\times\mathbb{T} \tag{3.3}$$

with some function $u\in C(\bar{Q}_T)$. The interpretation of the equation on the interface is crucial.

We introduce a multivalued function Φ associated with continuous function $g_\pm(\sigma)$. For $(s,\sigma)\in\mathbb{R}^2$, we define a closed interval $\Phi(s,\sigma)$ such that

$$\Phi(s,\sigma)=\begin{cases}\{g_+(\sigma)\} & \text{if } s>0,\\ [\underline{g}(\sigma),\bar{g}(\sigma)] & \text{if } s=0,\\ \{g_-(\sigma)\} & \text{if } s<0,\end{cases} \tag{3.4}$$

where $\underline{g}(\sigma)=\min(g_+(\sigma),g_-(\sigma))$, $\bar{g}=\max(g_+,g_-)$. This correspondence defines a mapping $\Phi:\mathbb{R}^2\to 2^{\mathbb{R}}$. For $u,v\in C(\bar{Q}_T)$ we define a subset $G(u,v)$ such that

$$G(u,v)=\{q\in L^\infty(Q_T);\, q(z)\in\Phi(u(z),v(z)) \text{ a.e. } z\in Q_T\}, \tag{3.5}$$

where $z=(t,x)$. This correspondence defines a mapping $G: C(\bar{Q}_T)\times C(\bar{Q}_T)\to 2^{L^\infty(Q_T)}$.

DEFINITION 3.1. Suppose that $u\in C(\bar{Q}_T)$ is given and the $Q_T^\pm$ is defined by (3.3). Suppose also that $g_\pm:\mathbb{R}\to\mathbb{R}$ is continuous. We say $v\in C(\bar{Q}_T)$ is a *generalized solution* of (3.1)-(3.2) if

$$v_t-\Delta v\in G(u,v) \quad \text{in } Q_T$$

i.e., there is $q\in G(u,v)$ such that

$$v_t-\Delta v=q \quad \text{in } Q_T$$

in the distribution sense. Since $G(u,v)$ depends on u only through its signature, this definition depends only on $Q_T^\pm$ and is independent of the choice of u.

PROPOSITION 3.2. *For $u, v \in C(\bar{Q}_T)$ the set $G(u, v)$ is a nonempty, bounded convex subset of $L^\infty(Q_T)$.*

Proof. Since $\Phi(s, \sigma)$ is convex in $\mathbb{R}$,

$$\lambda q_1(z)+(1-\lambda)q_2(z) \in \Phi(u(z), v(z)) \quad \text{for a.e. } z$$

if $q_1, q_2 \in G(u, v)$ and $0<\lambda<1$. This implies that $\lambda q_1+(1-\lambda)q_2 \in G(u, v)$ so $G(u, v)$ is convex in $L^\infty(Q_T)$.

The Borel measurable function

$$\psi(s, \sigma)=\chi_{(-\infty,0)}(s)g_-(\sigma)+\chi_{[0,\infty)}(s)g_+(\sigma)$$

on $\mathbb{R}^2$ satisfies $\psi(s, \sigma) \in \Phi(s, \sigma)$ for all $s, \sigma \in \mathbb{R}$, and therefore $\psi(u, v) \in G(u, v)$.

Since $g_\pm$ is locally bounded, we see $G(u, v)$ is bounded in $L^\infty(Q_T)$. □

LEMMA 3.3. *Suppose that $u_m \to u$ in $C(\bar{Q}_T)$ and that $v_m \to v$ in $C(\bar{Q}_T)$. Suppose that $q_m \in G(u_m, v_m)$. Then there is a subsequence $\{m_j\}$ and $q \in G(u, v)$ such that $q_{m_j} \rightharpoonup q$ *-weakly in $L^\infty(Q_T)$.*

Proof. Since $g_\pm$ is continuous, $\bigcup_{m=1}^\infty G(u_m, v_m)$ is bounded in $L^\infty(Q_T)$. In particular, $\{q_m\}$ is bounded in $L^\infty(Q_T)$. By the Banach-Alaoglu theorem, $\{q_m\}$ has a *-weak convergent subsequence (still denoted $\{q_m\}$), i.e.,

$$q_m \rightharpoonup q \text{ *-weakly in } L^\infty(Q_T).$$

In particular $q_m \rightharpoonup q$ weakly in $L^2(Q_T)$ since Q_T is bounded. Applying Mazur's theorem (see, e.g., [Y]), we see that there is $\lambda_m^m, \cdots, \lambda_m^{lm} \geqq 0$ with

$$\sum_{j=m}^{l_m} \lambda_m^j=1$$

such that

$$\tilde{q}_m := \sum_{j=m}^{l_m} \lambda_m^j q_j \to q \text{ strongly in } L^2(Q_T) \quad \text{as } m \to \infty.$$

Taking a subsequence if necessary we may conclude

$$\tilde{q}_m(z) \to q(z)(m \to \infty) \quad \text{for a.e. } z. \tag{3.6}$$

We fix $z \in Q_T$ such that (3.6) and

$$q_m(z) \in \Phi(u_m(z), v_m(z)) \quad \text{for all } m \geqq 1. \tag{3.7}$$

Suppose that $u(z)=0$. By (3.4) and (3.7)

$$q_m(z) \in [\underline{g}(v_m(z)), \bar{g}(v_m(z))] \tag{3.8}$$

since $\{g_\pm(v_m(z))\}$ lies in the interval in (3.8). Since $\underline{g}$ and $\bar{g}$ are continuous and $v_m(z) \to v(z)$, for each $\varepsilon>0$ there is m_0 such that if $m \geqq m_0$, then

$$[\underline{g}(v_m(z)), \bar{g}(v_m(z))] \subset (a-\varepsilon, b+\varepsilon)$$

with

$$[a, b] := [\underline{g}(v(z)), \bar{g}(v(z))].$$

By (3.8), we now observe that

$$\tilde{q}_m(z) \in (a-\varepsilon, b+\varepsilon).$$

From (3.6) it follows that

$$q(z) \in (a-\varepsilon, b+\varepsilon).$$

Since $\varepsilon>0$ is arbitrary, this implies

$$q(z) \in [\underline{g}(v(z)), \bar{g}(v(z))]. \tag{3.9}$$

Suppose that $u(z)>0$. For sufficiently large m, say $m\geqq m_0$, we may assume $u_m(z)>0$. It follows that

$$\Phi(u_m(z), v_m(z))=\{g_+(v_m(z))\} \quad \text{for } m\geqq m_0.$$

By (3.7) we have

$$q_m(z)=g_+(v_m(z)) \quad \text{for } m\geqq m_0. \tag{3.10}$$

Since $\tilde{q}_m(z)\to q(z)$ by (3.6) and g_+ is continuous, (3.10) yields

$$q(z)=g_+(v(z)). \tag{3.11}$$

The proof for $u(z)<0$ parallels that for $u(z)>0$. By (3.9) and (3.11) we can conclude that

$$q(z)\in\Phi(u(z), (v(z)) \quad \text{a.e. } z\in Q_T,$$

which completes the proof □

COROLLARY 3.4. *The set $G(u, v)$ is weak $*$ compact in $L^\infty(Q_T)$.*

Proof. By Lemma 3.3 we see $G(u, v)$ is weak $*$ sequentially closed. Since $G(u, v)$ is bounded by Proposition 3.2 and since the predual $L^1(Q_T)$ is separable, we can drop the word "sequentially." The boundedness of $G(u, v)$ now implies that $G(u, v)$ is weak $*$ compact. □

Remark 3.5. The condition (3.2) is implicit in Definition 3.1. We will see that all generalized solutions v have the regularity property (3.2).

4. Main results. We consider a system (3.1)-(3.2), coupled with an interface equation:

$$v_t=\Delta v+g_\pm(v) \quad \text{in } Q_T^\pm=\bigcup_{0<t<T}\{t\}\times\Omega_\pm(t), \tag{4.1}$$

$$v(t)=v(t,\cdot)\in C^1(\mathbb{T}) \quad \text{for } 0<t<T, \tag{4.2}$$

$$V=\eta(\vec{n},\nabla\vec{n})+W(v)\alpha(\vec{n}) \quad \text{on } \Gamma(t)=\mathbb{T}\backslash(\Omega_+(t)\cup\Omega_-(t)), \tag{4.3}$$

with given initial data

$$v(0,x)=v_0(x) \quad \text{in } \mathbb{T}, \tag{4.4}$$

$$\Omega_\pm(t)|_{t=0}=\Omega_\pm(0). \tag{4.5}$$

Here we assume that

(4.6a) $g_\pm:\mathbb{R}\to\mathbb{R}$ is continuous and bounded,

(4.6b) η satisfies (2.9)-(2.11),

(4.6c) $W:\mathbb{R}\to\mathbb{R}$ is locally Lipschitz continuous,

(4.6d) $\alpha:S^{n-1}\to\mathbb{R}$ is continuous.

We say $(\Omega_\pm(t), v(t))$ is a *weak solution* of (4.1)-(4.5) if $\{\Omega_\pm(t)\}_{0\leqq t<T}$ is a weak solution of (4.3), (4.5) with $v\in C(\bar{Q}_T)$ and v is a generalized solution of (4.1)-(4.2) with (4.4); see Definitions 2.1 and 3.1. We now state one of our main results.

THEOREM 4.1. *Let $T>0$. Assume that $g_\pm$, η, W, α satisfy* (4.6a-d). *Suppose that $\Omega_+(0)$ and $\Omega_-(0)$ are mutually disjoint open sets in $\mathbb{T}$ and that $v_0(x)\in C^2(\mathbb{T})$. Then there exists a (global) weak solution $(\Omega_\pm(t), v(t))$ of* (4.1)-(4.5) *such that $v\in C^{1,0}(\bar{Q}_T)=\{v\in C(\bar{Q}_T);\ \nabla v\in C(\bar{Q}_T)\}$.*

Remark 4.2. We note that (4.3) includes

$$V=-c\operatorname{div}\vec{n}+W(v),\qquad c\geqq 0 \tag{4.3'}$$

as a special example. If (4.3′) replaces (4.3) in (4.1)-(4.5), then it is known that there is a unique smooth local solution. This is proved by X.-Y. Chen [XYC] for $c>0$ and by X. Chen [XC2] for $c=0$ where $\mathbb{R}^n$ replaces $\mathbb{T}$. Our result is the first global existence result ever for this special system if the space dimension $n\geqq 2$. For $n=1$ see [HNM].

We shall construct a solution using Kakutani's fixed point theory for a multivalued mapping. We take a Banach space

$$X:=C^{1,0}(\bar{Q}_T).$$

For $v\in X$ we solve (4.3), (4.5) by applying Theorem 2.2. Since v can be extended as an element of $C^{1,0}$ for $t>T$, we have a unique weak solution $\{\Omega_\pm(t)\}_{0\leqq t\leqq T}$ for (4.3), (4.5) with given $\Omega_\pm(0)$. If we set

$$\tilde{Q}_T^\pm=\bigcup_{0\leqq t\leqq T}\{t\}\times\Omega_\pm(t),$$

then we have a mapping

$$\mathcal{T}:X\to\mathcal{O},\ v\mapsto(\tilde{Q}_T^+,\tilde{Q}_T^-),$$

where $\mathcal{O}$ denotes the set of disjoint pair of open sets in $[0,T]\times\mathbb{T}$.

For $q\in L^\infty(Q_T)$, let $w=E(q)$ be the unique solution of

$$\begin{aligned} &w_t-\Delta w=q\quad\text{in } Q_T\\ &w(0,x)=v_0(x)\in C^2(\mathbb{T}).\end{aligned}\tag{4.7}$$

By the parabolic theory [LUS], E defines a continuous affine map from $L^\infty(Q_T)$ to $\bigcap_{p>1}W_p^{2,1}(Q_T)$, which is continuously embedded in X by the Sobolev inequality. Thus

$$E:L^\infty(Q_T)\to X$$

is a continuous affine operator. For u, $v\in C(\bar{Q}_T)$, we define a subset of X by

$$\mathcal{P}(u,v)=\{E(q);\ q\in G(u,v)\}.$$

This correspondence defines a mapping

$$\mathcal{P}:C(\bar{Q}_T)\times C(\bar{Q}_T)\to 2^X.$$

For given $(\tilde{Q}_T^+,\tilde{Q}_T^-)\in\mathcal{O}$, we take $u\in C(\bar{Q}_T)$ such that

$$\tilde{Q}_T^\pm=\{(t,x)\in\bar{Q}_T;\ u\gtrless 0\}.$$

Since G depends on u through its signature, we may regard the mapping $\mathcal{P}$ as

$$\mathcal{P}:\mathcal{O}\times C(\bar{Q}_T)\to 2^X.$$

For given v_0 and $\Omega_\pm(0)$, we define

$$\mathcal{S}:X\to 2^X$$

by $\mathcal{S}(v) = \mathcal{P}(\mathcal{T}(v), v)$. If $\mathcal{S}$ has a fixed point $\bar{v} \in X$, i.e.,

$$\bar{v} \in \mathcal{S}(\bar{v}),$$

we observe that $(\Omega_\pm(t), \bar{v}(t))$ is a weak solution of (4.1)-(4.5), where

$$\mathcal{T}(\bar{v}) = \bigcup_{0 \leqq t \leqq T} \{t\} \times \Omega_\pm(t).$$

We shall prove that $\mathcal{S}$ has a fixed point.

PROPOSITION 4.3. *The set $\mathcal{P}(u, v)$ is nonempty, compact and convex in $C(\bar{Q}_T)$ (and in X).*

Proof. Since E defined by (4.7) is affine and $G(u, v)$ is nonempty and convex by Proposition 3.2 we see that $\mathcal{P}(u, v)$ is convex.

We next observe that E is continuous from a bounded set of $L^\infty(\bar{Q}_T)$ (equipped with weak $*$ topology) to X. Indeed, if $q_m \rightharpoonup q$ $*$ weakly in $L^\infty(Q_T)$ then $\{E(q_m)\}$ has a weakly convergent subsequence in $W_p^{2,1}(Q_T)$ for $p > 1$. Since the inclusion

$$(4.8) \qquad W_p^{2,1}(Q_T) \to X = C^{1,0}(\bar{Q}_T) \text{ is compact if } p > n+1$$

(see, e.g., [LUS]), $E(q_m) \to w$ strongly in X by taking a subsequence. Since $w_m = E(q_m)$ satisfies

$$(\partial_t - \Delta) w_m = q_m \quad \text{in } Q_T, \qquad w_m(0, x) = v_0(x),$$

w solves

$$(\partial_t - \Delta) w = q \quad \text{in } Q_T$$

in the distribution sense with $w(0, x) = v_0(x)$. This implies $w = E(q)$. Since the limit w is independent of the choice of subsequences, we observe that

$$E(q_m) \to E(q) \quad \text{in } X.$$

This sequential continuity implies the continuity on a bounded set of $L^\infty(Q_T)$.

Since $G(u, v)$ is weak $*$ compact in $L^\infty(Q_T)$, the continuous image of $G(u, v)$ is compact. The above continuity of E implies that $\mathcal{P}(u, v)$ is compact in $C^0(\bar{Q}_T)$ as well as in X. □

Since $g_\pm$ is bounded by (4.6a), we see that

$$\bigcup_{u, v \in C(\bar{Q}_T)} G(u, v)$$

is bounded in $L^\infty(Q_T)$. Therefore, by the parabolic theory for (4.7) [LUS],

$$\mathcal{S}(v) \subset K = \{w \in W_p^{2,1}(Q_T);\ \|w\|_{W_p^{2,1}} \leqq M\}, \qquad p > 1$$

if M is taken sufficiently large. We fix $p > n+1$ so that K is compact in X by (4.8). The mapping $\mathcal{S}$ is now interpreted as

$$\mathcal{S}: X \to 2^K.$$

The graph of $\mathcal{S}$ is defined by

$$\operatorname{gr} \mathcal{S} = \{(v, w);\ w \in \mathcal{S}(v)\} \subset X \times K.$$

Since K is compact, $\operatorname{gr} \mathcal{S}$ is closed if and only if $\mathcal{S}$ is upper semicontinuous. For the definition of upper semicontinuity see [AF].

PROPOSITION 4.4. *The set $\operatorname{gr} \mathcal{S}$ is closed in $X \times K$.*

Proof. Suppose that v_m, $v \in X$, $w_m \in \mathcal{S}(v_m)$, $w \in X$ such that $v_m \to v$ in X and $w_m \to w$ in X. Our goal is to prove $w \in \mathcal{S}(v)$. By the definition of weak solutions for (4.3), there is a viscosity solution $u_m \in C(\bar{Q}_T)$ of

$$u_t + F_m(t, x, \nabla u, \nabla^2 u) = 0 \quad \text{in } Q_T$$

with

$$F_m(t, x, p, X) = F_\eta(p, X) - W(v_m(t, x))\alpha(-p/|p|)|p|$$

such that

$$\mathcal{T}(v_m) = (\{u_m(t, x) > 0\}, \{u_m(t, x) < 0\}).$$

We can arrange $u_m(0, x) = u_0(x)$ independent of m such that

$$\Omega_\pm(0) = \{x \in \mathbb{T};\ u_0(x) \gtrless 0\}.$$

Since $v_m \to v$ in X, by the stability of viscosity solutions, there is $u \in C(\bar{Q}_T)$ such that $u_m \to u$ in $C(\bar{Q}_T)$ and u solves (in the viscosity sense)

$$u_t + F(t, x, \nabla u, \nabla^2 u) = 0 \quad \text{in } Q_T$$

with $F(t, x, p, X) = F_\eta(p, X) - W(v(t, x))\alpha(-p/|p|)|p|$ (see the Theorem in the Appendix), where $u(0, x) = u_0(x)$. This implies

$$\mathcal{T}(v) = (\{u > 0\}, \{u < 0\}). \tag{4.9}$$

By the definition of $\mathcal{P}$ there is $q_m \in G(u_m, v_m)$ such that

$$\begin{aligned} &(\partial_t - \Delta)w_m = q_m \quad \text{in } Q_T, \\ &w_m(0, x) = v_0(x) \quad \text{on } \mathbb{T}. \end{aligned} \tag{4.10}$$

Applying Lemma 3.3, we may conclude that

$$q_m \rightharpoonup q \ *\text{-weakly in } L^\infty(Q_T)$$

with some $q \in G(u, v)$ by taking a subsequence if necessary. Since $w_m \to w$ in X, (4.10) implies that

$$(\partial_t - \Delta)w = q \quad \text{in } Q_T$$

in the distribution sense, and

$$w(0, x) = v_0(x).$$

This yields $w \in \mathcal{S}(v)$ by (4.9) so that the proof is now complete. □

Proof of Theorem 4.1. Since K is compact and convex, by Propositions 4.3 and 4.4 we can apply the following fixed point theorem to conclude that there is $\bar{v} \in \mathcal{S}(\bar{v}) \cap K$. By the definition of $\mathcal{S}$, we see $\bar{v}$ together with $\mathcal{T}(\bar{v})$ is a desired weak solution of (4.1)-(4.5).

KAKUTANI'S FIXED POINT THEOREM [AF, THM. 3.2.3]. *Let K be a convex compact subset of a Banach space X and $\mathcal{S}: X \to 2^K$. If $\mathcal{S}$ is upper semicontinuous and $\mathcal{S}(v)$ is a nonempty convex closed set in K for $v \in X$, then $\mathcal{S}$ has a fixed point $\bar{v} \in K \cap \mathcal{S}(\bar{v})$.*

Remark 4.5. The assumption $v_0(x) \in C^2(\mathbb{T})$ in Theorem 4.1 is weakened as $v_0(x) \in W^{2-2/p}(\mathbb{T})$, $p > n+1$ because the regularity condition on v_0 is only used to solve (4.7) in $W_p^{2,1}(\bar{Q}_T)$.

We conclude this paper by stating an existence result of a global solution on the time interval $(0, \infty)$.

THEOREM 4.6. *Assume the same hypotheses of Theorem* 4.1 *for* $g_\pm$, η, W, α, $\Omega_\pm(0)$. *Suppose that* $v_0 \in W^{2-2/p}(\mathbb{T})$ *for* $p > n+1$. *Then there exists* $\{(\Omega_\pm(t), v(t))\}_{t \geqq 0}$ *which is a weak solution of* (4.1)-(4.5) *for arbitrary* $T > 0$.

Proof. For fixed $T > 0$ by Remark 4.5 there is v_T such that $v_T \in \mathcal{S}(v_T)$. This implies

$$(\partial_t - \Delta) v_T = q_T \quad \text{in } Q_T,$$

$$v_T(0, x) = v_0(x),$$

with

$$q_T \in \bigcup_{u, v \in C(\bar{Q}_T)} G(u, v).$$

Since $g_\pm$ is bounded by (4.6a) we observe that

$$|q_T|_{L^\infty(Q_T)} \leqq M = \sup_\sigma |g_\pm(\sigma)|.$$

By the parabolic regularity theory [LUS], $\{v_T\}_{T \geqq 1}$ is bounded in $W_p^{2,1}(\bar{Q}_{t_0})(p > n+1)$ for each $t_0 > 0$. By (4.8) and a diagonal argument, there is a subsequence $\{v_{T'}\}$ and $v \in C([0, \infty) \times \mathbb{T})$ such that

$$v_{T'} \to v \text{ in } X_{t_0} = C^{1,0}(\bar{Q}_{t_0}). \tag{4.11}$$

Since $v_{T'} \in \mathcal{S}(v_{T'}) \subset X_{t_0}$ and $\operatorname{gr} \mathcal{S}$ is closed, (4.11) implies $v \in \mathcal{S}(v) \subset X_{t_0}$ where $\mathcal{S}$ depends on t_0. Since t_0 is arbitrary, this yields a desired global solution on $[0, \infty)$. □

Appendix. We shall state stability properties of viscosity solutions used in the proof of Proposition 4.4 for the reader's convenience. We use the following notation. For $h_m : L \to \mathbb{R}$, $L \subset Z$ we define

$$\lim_* h_m : \bar{L} \to \mathbb{R} \cup \{-\infty\},$$

$$\lim^* h_m : \bar{L} \to \mathbb{R} \cup \{+\infty\},$$

by

$$\left(\lim_* h_m\right)(z) = \lim_{\substack{m \to \infty \\ \varepsilon \downarrow 0}} \inf \{h_j(y), d(z, y) < \varepsilon, j \geqq m, y \in L\}$$

and

$$\lim^* h_m = -\lim_* (-h_m),$$

where Z is a metric space with metric d. If h is independent of m, we write $h_* = \lim_* h_m$, $h^* = \lim^* h_m$. We shall suppress the word "viscosity."

LEMMA. *Suppose that* $F_m : Q_T \times \mathbb{R}^n \times \mathbb{S}_n \to \mathbb{R}$ *is lower semicontinuous and that* $F = \lim_* F_m$. *Suppose that* u_m *is a subsolution of*

$$u_t + F_m(t, x, \nabla u, \nabla^2 u) = 0 \quad \textit{in } Q_T.$$

Then $u = \lim^* u_m$ *is a subsolution of*

$$u_t + F(t, x, \nabla u, \nabla^2 u) = 0 \quad \textit{in } Q_T,$$

provided that u *does not take* $+\infty$ *in* $\bar{Q}_T$.

Similar results are proved by Barles and Perthame [BP] for first-order differential equations and formulated in Ishii [I] in the general case. Since the proof is easily modified for our setting, we omit the proof. The following is a simple application of the lemma, the comparison Proposition 2.3, and construction of sub- and supersolutions.

THEOREM. *Suppose that* η, W, α *satisfy* (4.6b-d). *Suppose that* $v_m \to v$ *in* X. *We set*

$$F_m = F_\eta(p, X) - W(v_m(t, x))\alpha(-p/|p|)|p|,$$

$$F = F_\eta(p, X) - W(v(t, x))\alpha(-p/|p|)|p|,$$

where F_η *is defined by* (2.3). *Suppose that* $u_m \in C(\bar{Q}_T)$ *is a solution of*

$$u_t + F_m(t, x, \nabla u, \nabla^2 u) = 0 \quad \textit{in } Q_T \tag{1}$$

with $u_m(0, x) = u_0(x) \in C(\mathbb{T})$. *Then* $u_m \to u$ *in* $C(\bar{Q}_T)$ *for some* $u \in C(\bar{Q}_T)$ *and* u *is a solution of*

$$u_t + F(t, x, \nabla u, \nabla^2 u) = 0 \quad \textit{in } Q_T \tag{2}$$

with $u(0, x) = u_0(x)$.

Proof. Since $v_m \to v$ in $C(\bar{Q}_T)$ there are bounded sub- and supersolutions $w_\pm$ of (1) such that

$$\begin{gathered} w_\pm(0, x) = u_0(x), \\ w_-(t, x) \leqq u_0(x) \leqq w_+(t, x) \quad \text{in } Q_T \\ Q_T \end{gathered} \tag{3}$$

and that $w_\pm$ is independent of m; see [CGG, Prop. 6.4] and the proof of Theorem 2.2. By Proposition 2.3 we see that

$$w_-^* \leqq u_m \leqq w_{+*} \quad \text{in } Q_T. \tag{4}$$

Since u_m is a subsolution of

$$u_t + (F_m)_*(t, x, \nabla u, \nabla^2 u) = 0 \quad \text{in } Q_T$$

by definition, applying the lemma yields that $\bar{u} = \lim^* u_m$ is a subsolution of

$$u_t + F_*(t, x, \nabla u, \nabla^2 u) = 0 \quad \text{in } Q_T.$$

(This is the definition that $\bar{u}$ is a subsolution of (2).) Similarly $\underline{u} = \lim_* u_m$ is a supersolution of

$$u_t + F^*(t, x, \nabla u, \nabla^2 u) = 0 \quad \text{in } Q_T.$$

By (3) and (4) we observe that

$$\bar{u}(0, x) = \underline{u}(0, x) = u_0(x).$$

Applying Proposition 2.3 implies $\bar{u} = \underline{u}$ and $u = \bar{u}$ is a solution of (2). The property $\bar{u} = \underline{u}$ implies that $u_m \to u$ in $C(\bar{Q}_T)$. The proof is now complete. □

Acknowledgment. The work of the first author was completed while he visited the Institute for Mathematics and Its Applications. Its hospitality and support are gratefully acknowledged.

REFERENCES

[AW] D. G. ARONSON AND H. F. WEINBERGER, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33-76.

[AF] J. P. AUBIN AND H. FRANKOWSKA, *Set-valued Analysis*, Birkhäuser, Boston, Basel, Berlin, 1990.

[B] G. BARLES, *Remarks on a flame propagation model*, Rapport INRIA, 464, 1985.

[BP] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modèl. Math. Anal. Numèr., 21 (1987), pp. 557-579.

[BK] L. BRONSARD AND R. V. KOHN, *Motion by mean curvature as the singular of Ginzburg-Landau dynamics*, J. Differential Equations, 90 (1991), pp. 211-237.

[C] G. CAGINALP, *The role of microscopic anisotropy in the macroscopic behavior of a phase boundary*, Ann. Physics, 172 (1986), pp. 136-155.

[CP] J. CARR AND R. L. PEGO, *Metastable patterns in solutions of* $u_t = \varepsilon^2 u_{xx} - f(u)$, Comm. Pure Appl. Math., 42 (1989), pp. 523-576.

[XC1] X. CHEN, *Generation and propagation of the interface for reaction-diffusion equations*, IMA preprint 637, Institute for Mathematics and Its Applications, Minneapolis, MN, 1990.

[XC2] ———, *General and propagation of interface in reaction-diffusion systems*, IMA preprint 708, Institute for Mathematics and Its Applications, Minneapolis, MN, 1990.

[XYC] X.-Y. CHEN, *Dynamics of interfaces in reaction diffusion systems*, Hiroshima Math. J., 21 (1991), pp. 47-83.

[CGG] Y.-G. CHEN, Y. GIGA, AND S. GOTO, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geometry, 33 (1991), pp. 749-786; Announcement: Proc. Japan Acad. Ser. A, 65 (1989), pp. 207-210.

[DS] P. DEMOTTONI AND M. SCHATZMAN, *Geometrical evolution of developed interfaces*, 1990, preprint; Announcement: *Evolution géometric d'interfaces*, C.R. Acad. Sci. Paris, 309 (1989), pp. 453-458.

[ESS] L. C. EVANS, H. M. SONER, AND P. E. SONGANIDIS, *Phase transitions and generalized motion by mean curvature*, Comm. Pure Appl. Math., to appear.

[ES1] L. C. EVANS AND J. SPRUCK, *Motion of level sets by mean curvature* I, J. Differential Geometry, 33 (1991), pp. 635-681.

[ES2] ———, *Motion of level sets by mean curvature* II, Trans. Amer. Math. Soc., to appear.

[F] P. C. FIFE, *Dynamics of Interfacial Layers and Diffusive Interfaces*, CBMS-NSF Regional Conf. Series Appl. Math. 53, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982.

[FH] P. C. FIFE AND L. HSIAO, *The generation and propagation of internal layers*, Nonlinear Anal., TMA 12 (1988), pp. 19-41.

[GG1] Y. GIGA AND S. GOTO, *Motion of hypersurfaces and geometric equations*, J. Math. Soc. Japan, 44 (1992), pp. 99-111.

[GG2] ———, *Geometric evolution of phase-boundaries*, IMA preprint 738, Institute for Mathematics and Its Applications, Minneapolis, MN, 1990.

[GGIS] Y. GIGA, S. GOTO, H. ISHII, AND M.-H. SATO, *Comparison principle and convexity preserving properties for singular degenerate parabolic equations on unbounded domains*, Indiana Univ. Math. J., 40 (1991), pp. 443-470.

[Gr] M. GRAYSON, *A short note on the evolution of a surface by its mean curvature*, Duke Math. J., 58 (1989), pp. 555-558.

[Gu1] M. GURTIN, *Towards a nonequilibrium thermodynamics of two-phase materials*, Arch. Rational Mech. Anal., 100 (1988), pp. 275-312.

[Gu2] ———, *Multiphase thermomechanics with interfacial structure*, 1. *Heat conduction and the capillary balance law*, Arch. Rational Mech. Anal., 104 (1988), pp. 195-221.

[HNM] D. HILHORST, Y. NISHIURA, AND M. MIMURA, *A free boundary problem arising from some reaction-diffusion system*, Proc. Roy. Soc. Edinburgh, Sect. A, 118 (1991), pp. 335-378.

[I] H. ISHII, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, Ann. Scuola Norm. Sup. Pisa, Cl. Sci. (4), 16 (1989), pp. 105-135.

[LUS] O. LADYZHENSKAYA, V. SOLONNIKOV, AND N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monographs, Vol. 23, American Mathematical Society, Providence, RI, 1968.

[OMK] T. OHTA, M. MIMURA, AND R. KOBAYASHI, *Higher dimensional localized patterns in excitable media*, Phys. D, 34 (1989), pp. 115-144.

[S] H. M. SONER, *Motion of a set by the curvature of its boundary*, J. Differential Equations, to appear.

[Y] K. YOSIDA, *Functional Analysis*, Fourth ed., Springer-Verlag, Berlin, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 836–851, July 1992

ON A NONLINEAR ELLIPTIC-PARABOLIC PARTIAL DIFFERENTIAL EQUATION SYSTEM IN A TWO-DIMENSIONAL GROUNDWATER FLOW PROBLEM*

PH. CLÉMENT†, C. J. VAN DUIJN†, AND SHUANHU LI†

Abstract. In this paper a nonlinear elliptic-parabolic system which arises in a two-dimensional groundwater flow problem is studied. Abstract results on evolution equations are employed to obtain existence and uniqueness results. Regularity and stability properties of the solution are also considered.

Key words. elliptic-parabolic system, analytic semigroups, semilinear and quasilinear evolution equations

AMS(MOS) subject classifications. 35K50, 47D05

1. Introduction.

Let $\Omega \in \mathbf{R}^2$ be a bounded domain with smooth boundary. In this paper we study the following nonlinear elliptic-parabolic system:

$$(E)\begin{cases} -\Delta v = \partial_1 u & \text{in } \Omega \times (0,\infty), \\ v = 0 & \text{on } \partial\Omega \times (0,\infty), \end{cases}$$

and

$$(P)\begin{cases} \partial_t u + \operatorname{div}\vec{F} = 0 & \text{in } \Omega \times (0,\infty), \\ \vec{F}\cdot\vec{\nu} = 0 & \text{on } \partial\Omega \times (0,\infty), \\ u(\cdot,0) = u_0(\cdot) & \text{in } \Omega. \end{cases}$$

Here we have

$$\vec{F} = \vec{q}\,u - D\cdot \operatorname{grad} u,$$
$$\vec{q} = \operatorname{curl} v,$$
$$D = (D_{ij}),$$

where $D_{ij}(q_1,q_2)$ are uniformly Lipschitz continuous functions on $\mathbf{R}^2$.

This system arises in the description of the movement of a fluid of variable density (u) through a porous medium under the influence of gravity and hydrodynamic dispersion. In §2 we set up the model and we discuss the physical background.

In a slightly different form, Problem (E), (P) was studied by Su [16] using classical partial differential equation (PDE) methods. In this paper we present an approach in the spirit of abstract evolution equations in Banach spaces. This turns out to be quite efficient because of the particular form of the problem.

We consider two cases of the model separately. In the first (approximate) case we take $D_{ij} = \delta_{ij}$ (δ_{ij} is the Kronecker symbol). Then the system can be considered as a semilinear evolution equation. Clearly, there are many results on abstract semilinear evolution equations, and these results can be well applied to partial differential equations of parabolic type; see, e.g., Friedman [7], Henry [9], Pazy [12], or von Wahl [19]. Here we choose one theorem from von Wahl [20], which fits precisely to the abstract formulation of Problem (E), (P) with constant (D). By this theorem we obtain the global existence of the solution in $L^p(\Omega)$. This is done in §3. There we also study the

*Received by the editors November 28, 1990; accepted for publication (in revised form) November 8, 1991.

† Department of Mathematics and Informatics, Delft University of Technology, the Netherlands.

regularity and asymptotic properties of the solution. We show that the solution is in fact a classical solution of $(E), (P)$, and u converges to the mean value in sup-norm as $t \to \infty$. A first draft of §3 was made by de Roo [13].

In §4, we study the full problem, i.e., D is nonconstant and velocity dependent. Then the abstract formulation leads to a quasilinear evolution equation. The abstract results on such equations are not as complete as the results on semilinear equations. Moreover the application to partial differential equations is much harder. In this paper we use the framework of quasilinear evolution equations due to Amann [2], see also Sobolevskii [15]. As a result, we obtain local existence of weak solutions in $W^{1,p}(\Omega)$. As for this moment, we are not able to obtain global existence. Because the coefficients D_{ij} are not differentiable at the origin, see (2.13), we can not expect to have classical solutions.

2. The physical background. Let $\Omega = (-L, L) \times (0, H)$, with $L, H > 0$, denote a rectangular region in the x_1, x_2 plane which is occupied by a homogeneous and isotropic porous medium. This medium is characterized by a permeability $\kappa \in (0, \infty)$ and a porosity $\phi \in (0, 1)$. It is saturated by an incompressible fluid. The fluid is characterized by a constant viscosity $\mu \in (0, \infty)$ and a variable density ρ (or a specific weight $\gamma = \rho g$, where g is the accelaration of gravity). Here the coordinate system is chosen such that the gravity is pointing in the negative x_2-direction. A typical example of this situation arises in the flow of fresh and salt groundwater in a two-dimensional vertical aquifer. In this application it is natural to assume that γ satisfies

$$0 < \gamma_f \leq \gamma(x_1, x_2, t) \leq \gamma_s \quad \forall (x_1, x_2, t) \in \Omega \times (0, \infty). \tag{2.1}$$

Here γ_f and γ_s are constants, denoting the specific weight of the fresh and the salt groundwater, respectively.

The basic equations for flow in a porous medium are the continuity equation

$$\operatorname{div} \vec{q} = 0 \quad \text{in } \Omega \times (0, \infty) \tag{2.2}$$

and the momentum balance equation (Darcy's law), see, e.g., Bear [5],

$$\frac{\mu}{\kappa} \vec{q} + \operatorname{grad} p + \gamma \vec{e}_2 \ = 0 \quad \text{in } \Omega \times (0, \infty). \tag{2.3}$$

Here we denote by the vector $\vec{q}$ the specific discharge of the fluid and by the scalar p the fluid pressure. Finally, $\vec{e}_2$ denotes the unit vector in the positive x_2-direction (i.e., pointing upwards).

In this paper we are interested in describing the distribution of the specific weight γ in the domain Ω under the action of gravity and hydrodynamic dispersion, without any other influence from outside. Therefore, we impose on the boundary $\partial\Omega$ the no-flow condition

$$\vec{q} \cdot \vec{\nu} = 0 \quad \text{on } \partial\Omega \times (0, \infty), \tag{2.4}$$

where $\vec{\nu}$ is the outward normal unit vector on $\partial\Omega$.

For a given specific weight distribution γ, (2.2)–(2.4) determine the discharge field $\vec{q}$. To obtain a single equation for this relation we can use either the pressure or, because of (2.2), the stream function. Here we use a formulation in terms of the stream function. It satisfies

$$\vec{q} = (q_1, q_2) = \operatorname{curl} \psi := (-\partial_2 \psi, \ \partial_1 \psi), \tag{2.5}$$

where ∂_i denotes the partial derivative with respect to the variable x_i for $i = 1, 2$. Note that the operator curl in (2.5) acts on a scalar function. Therefore this definition differs from the usual one. It is introduced here only for convenience.

Substituting (2.5) into Darcy's law (2.3) and taking the curl in the usual sense (i.e., curl $\vec{q} = \partial_2 q_1 - \partial_1 q_2$) gives

$$-\Delta\psi = \frac{\kappa}{\mu}\partial_1\gamma \quad \text{in } \Omega \times (0,\infty). \tag{2.6}$$

Combining (2.4) and (2.5) implies that ψ is constant on the boundary $\partial\Omega$. Without loss of generality, we take the boundary condition

$$\psi = 0 \quad \text{on } \partial\Omega \times (0,\infty). \tag{2.7}$$

The boundary value problem (2.6), (2.7) gives the stream function and thus the specific discharge, in terms of the specific weight γ. Conversely, the mass balance equation for the fluid gives the density ρ (and thus the specific weight) in terms of the fluid field $\vec{q}$. According to Bear [5], we have

$$\phi\partial_t\rho + \operatorname{div}\vec{F} = 0 \quad \text{in } \Omega \times (0,\infty), \tag{2.8}$$

where the flux $\vec{F}$ is given by

$$\vec{F} = \vec{q}\rho - D \cdot \operatorname{grad}\rho. \tag{2.9}$$

In (2.9), $D = (D_{ij})_{2\times 2}$ is the hydrodynamic dispersion matrix with $D_{ij} : \mathbf{R}^2 \to \mathbf{R}$ given by

$$D_{ij}(q_1,q_2) = \begin{cases} (\alpha_T \mid \vec{q} \mid + \tau\ \phi D_{\text{mol}})\delta_{ij} + (\alpha_L - \alpha_T)\dfrac{q_i q_j}{\mid \vec{q} \mid} & \text{if } (q_1,q_2) \neq 0, \\ \tau\phi D_{\text{mol}}\delta_{ij} & \text{if } (q_1,q_2) = 0. \end{cases} \tag{2.10}$$

Here $\alpha_L, \alpha_T, D_{\text{mol}}$ and τ are positive constants: α_L is the longitudinal and α_T is the the transversal dispersion length ($\alpha_T < \alpha_L$), D_{mol} is the molecular diffusion coefficient and the constant τ describes the tortuosity of the porous medium. Further, $|\cdot|$ denotes the Euclidean norm on $\mathbf{R}^2$ and δ_{ij} the Kronecker symbol.

In order to determine ρ (or γ) from (2.8) we have to specify boundary and initial conditions. We consider the no-flux condition

$$\vec{F} \cdot \vec{\nu} = 0 \quad \text{on } \partial\Omega \times (0,\infty), \tag{2.11}$$

and initially

$$\rho(\cdot,0) = \rho_0(\cdot) \quad \text{on } \Omega. \tag{2.12}$$

Next we rescale the equations into a dimensionless form.
Setting

$$\begin{aligned} x_1 &:= x_1/H, \\ x_2 &:= x_2/H, \\ t &:= t\frac{\kappa}{\mu}(\gamma_s - \gamma_f)/(H\phi), \\ u &:= (\gamma - \gamma_f)/(\gamma_s - \gamma_f), \\ v &:= \psi/(\frac{\kappa}{\mu}(\gamma_s - \gamma_f)H), \\ \Omega &:= (-L/H, L/H) \times (0,1), \end{aligned}$$

we find for u, v the elliptic-parabolic system

$$(E)\begin{cases} -\Delta v = \partial_1 u & \text{in } \Omega \times (0,\infty), \\ v = 0 & \text{on } \partial\Omega \times (0,\infty), \end{cases}$$

$$(P)\begin{cases} \partial_t u + \operatorname{div}\vec{F} = 0 & \text{in } \Omega \times (0,\infty), \\ \vec{F}\cdot\vec{\nu} = 0 & \text{on } \partial\Omega \times (0,\infty), \\ u(\cdot,0) = u_0(\cdot) & \text{on } \Omega. \end{cases}$$

Here we have

$$\begin{aligned} &\vec{F} = \vec{q}\,u - D\cdot\operatorname{grad} u, \\ &\vec{q} = \operatorname{curl} v, \\ &D = (D_{ij}) \end{aligned}$$

with

$$(2.13)\qquad D_{ij}(q_1,q_2) = \begin{cases} (a\mid\vec{q}\mid+m)\delta_{ij} \ + (b-a)\dfrac{q_iq_j}{\mid\vec{q}\mid} & \text{if } (q_1,q_2)\neq 0, \\ m\delta_{ij} & \text{if } (q_1,q_2) = 0, \end{cases}$$

where $a = \alpha_T/H$, $b = \alpha_L/H$ and $m = \phi D_{\mathrm{mol}T}/[\frac{\kappa}{\mu}(\gamma_s-\gamma_f)H]$.

The dispersion matrix D satisfies the following.

PROPOSITION 2.1. *Let $D = (D_{ij})$ be given by (2.13). Then*

(i) *D is uniformly positive definite on $\mathbf{R}^2$, i.e., there exists $\mu > 0$ such that*

$$\sum_{i,j=1}^{2} D_{ij}(q_1,q_2)\xi^i\xi^j \geq \mu|\xi|^2 \quad \forall\xi = (\xi^1,\xi^2), (q_1,q_2)\in\mathbf{R}^2;$$

(ii) *D_{ij} is uniformly Lipschitz continuous.*

Proof. The proof of (i) is immediate. To prove (ii) we have to show that the functions $f_{ij} : \mathbf{R}^2 \to \mathbf{R}$, defined by

$$f_{ij}(x) = \begin{cases} \dfrac{x_ix_j}{|x|} & \text{if } x\neq(0,0), \\ 0 & \text{if } x = (0,0), \end{cases}$$

are uniformly Lipschitz continuous. A straightforward computation shows that there exists a constant $L > 0$ such that

$$|\nabla f_{ij}(x)| \leq L \quad \forall x\in\mathbf{R}^2\backslash\{0\}$$

and

$$|f_{ij}(x) - f_{ij}(0)| \leq |x-0| \quad \forall x\in\mathbf{R}^2.$$

This implies the Lipschitz continuity for f_{ij} and thus for D_{ij}. □

The purpose of this paper is to study the elliptic-parabolic system $(E), (P)$. We do this in two steps. First, in §3 we consider the case, where

$$a = b = 0 \quad \text{and} \quad m = 1.$$

This situation describes the mixing of fresh and salt groundwater with dominant molecular diffusion. It implies $D_{ij} = \delta_{ij}$ which means that the problem is of semilinear type. In §4, we consider the full problem, where

$$0 < a < b < \infty \quad \text{and} \quad m > 0.$$

In this case the mixing is due to mechanical dispersion and molecular diffusion. It implies that D is velocity dependent which means that the problem is of quasilinear type.

3. The semilinear case.

3.1. The abstract setting. In this section we consider the case where the dispersion matrix D is independent of the velocity $\vec{q}$. This can be achieved by setting $a = b = 0$ in (2.13). For simplicity, we also set $m = 1$, which implies that $D_{ij} = \delta_{ij}$. Noting that $\vec{q} \cdot \vec{\nu} = 0$ on $\partial\Omega$, we arrive at the problem

$$(E)\begin{cases} -\Delta v = \partial_1 u & \text{in } \Omega \times (0,\infty), \\ v = 0 & \text{on } \partial\Omega \times (0,\infty), \end{cases}$$

$$(P')\begin{cases} \partial_t u - \Delta u + \operatorname{grad} u \cdot \operatorname{curl} v = 0 & \text{in } \Omega \times (0,\infty), \\ \dfrac{\partial u}{\partial \vec{\nu}} = 0 & \text{on } \partial\Omega \times (0,\infty), \\ u(\cdot, 0) = u_0(\cdot) & \text{in } \Omega. \end{cases}$$

Throughout this section we suppose that Ω is a bounded domain in $\mathbf{R}^2$ with smooth boundary $\partial\Omega$.

In order to formulate problem (E), (P') into an abstract form, we need to introduce some operators and Banach spaces.

Throughout this paper all vector spaces are over $\mathbf{R}$. If we use complex quantities (for example, in connection with spectral theory), it is always understood that we work with the natural complexifications (of spaces and operators). Thus by $\rho(A)$, the resolvent set of a linear operator with domain $D(A)$ and range $R(A)$, we mean always the resolvent set of its complexifications.

Let $p \in (2, \infty)$. By inverting (E) we obtain the operator (see the appendix)

$$E_p : D(E_p) = W^{1,p}(\Omega) \to W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega),$$

given by

$$E_p v = (-\Delta)^{-1} \partial_1 v.$$

Then we define

$$M_p(u) = (\partial_1 E_p u)\partial_2 u - (\partial_1 u)\partial_2 E_p u - u$$

for $u \in W^{1,p}(\Omega)$. Furthermore, we define operator A_p by

$$D(A_p) = \left\{ u \in W^{2,p}(\Omega) : \frac{\partial u}{\partial \vec{\nu}} = 0 \right\},$$

$$A_p : D(A_p) \to L^p(\Omega)$$

with

$$A_p u = -\Delta u + u.$$

Observe that in the definition of A_p, due to the imbedding $W^{2,p}(\Omega) \hookrightarrow C^1(\overline{\Omega})$, the boundary condition $\partial u/\partial\vec{\nu} = 0$ is satisfied in the classical sense. By using the operators introduced above, Problem (E), (P') can be formulated as

$$(CP)\begin{cases} u' + A_p u + M_p(u) = 0 & \text{for } t \in (0,\infty), \\ u(0) = u_0. \end{cases}$$

Here u' denotes the derivative of u with respect to t.

It is known that $-A_p$ generates an analytic semigroup on $L^p(\Omega)$. We shall show that M_p is a locally Lipschitz perturbation (in an appropriate sense) of A_p. Then we can apply abstract results for proving existence of solutions of (CP).

We recall the following results.

Let $\Sigma_\omega := \{\lambda \in \mathbf{C} : \mathrm{Re}\lambda \geq \omega\}$ for $\omega \in \mathbf{R}$. Furthermore, let X be a Banach space with norm $\|\cdot\|$, and let A be a given linear operator satisfying

(A1) A is densely defined and closed;
(A2) $\Sigma_0 \subset \rho(-A)$, where $\rho(-A)$ is the resolvent set of $-A$;
(A3) There exists a constant $M > 0$, such that

$$\|(\lambda + A)^{-1}\| \leq \frac{M}{1 + |\lambda|} \quad \forall \lambda \in \Sigma_0.$$

The fractional powers A^α of A are well defined for $0 < \alpha \leq 1$, and A^α is a closed linear operator whose domain $D(A^\alpha) \supset D(A)$. In this section we denote by X_α the Banach space obtained by endowing $D(A^\alpha)$ with the graph norm of A^α. Since $0 \in \rho(A)$, A^α is invertible and the norm of X_α is equivalent to $\|u\|_\alpha := \|A^\alpha u\|$ for $u \in D(A^\alpha)$. Also, for $0 < \beta < \alpha \leq 1$, $X_\alpha \hookrightarrow X_\beta$ with continuous imbedding.

Concerning the solvability of semilinear evolution equations of the form

$$u' + Au + M(u) = 0 \tag{3.1}$$

with initial value $u(0) = \varphi$, under the assumptions (A1)–(A3), we recall the following result (see von Wahl [20]).

THEOREM 3.1. *Let $0 \leq \beta < \alpha < 1$, and let $M : X_\alpha \to X$ satisfy $M(0) = 0$ and*

$$\|M(u) - M(v)\| \leq g(\|u\|_\beta + \|v\|_\beta)[\|u - v\|_\alpha + \|u - v\|_\beta(\|u\|_\alpha + \|v\|_\alpha + 1)]$$

for some continuous function g: $\mathbf{R}^+ \to \mathbf{R}^+$ and for all u, $v \in X_\alpha$. For $\varphi \in X_\beta$, there exists a $T = T(\varphi) \in (0,\infty]$ such that

(i) *there is one and only one mapping u: $[0,T) \to X$ fulfilling*

$$u \in C([0,T), X_\beta) \cap C((0,T), X_\alpha),$$

and

$$\sup_{0<t\leq T'} \|t^{\alpha-\beta} A^\alpha u(t)\| < \infty$$

for all $0 < T' < T$;

(ii)

$$u(t) = e^{-tA}\varphi - \int_0^t e^{-(t-s)A} M(u(s))ds$$

for $t \in (0, T)$;

(iii)

$$u(0) = \varphi;$$

(iv) *if* $T < \infty$, *then*

$$\lim_{t \uparrow T} \|u(t)\|_\beta = \infty.$$

Moreover, on $(0, T)$, u *fulfills* (3.1) *in the sense that* $u \in C^1((0, T), X)$, $u(t) \in D(A)$ *for* $t \in (0, T)$ *and* $Au(\cdot) \in C((0, T), X)$.

About the solution obtained in Theorem 3.1 we also have the following (see Henry [9]).

PROPOSITION 3.2. *Under the assumptions of Theorem* 3.1, *the solution* u *satisfies*

$$u'(t) \in X_\gamma$$

for $t \in (0, T)$ *and for any* $\gamma \in (0, 1)$.

3.2. The existence results. It follows from Agmon [1] that A_p satisfies (A1)–(A3). Moreover, we have the imbedding properties (see Henry [9]):

PROPOSITION 3.3. (i) $D(A_p^\alpha) \hookrightarrow W^{1,p}(\Omega)$ *for* $\alpha \in (\frac{1}{2}, 1)$;

(ii) $D(A_p^\alpha) \hookrightarrow W^{1,\infty}(\Omega)$ *for* $\alpha \in (\frac{1}{2} + \frac{1}{p}, 1)$.

We use Theorem 3.1 to obtain the existence for (CP). In this application we take $X = L^p(\Omega)$ with norm $\|\cdot\|_p$, X_α ($\alpha \in (0, 1)$) the Banach space induced by the operator A_p and $\beta = 0$ with $\|\cdot\|_\beta = \|\cdot\|_p$.

PROPOSITION 3.4. *Let* $\alpha \in (\frac{1}{2} + \frac{1}{p}, 1)$. *Then there exists a constant* $C \geq 1$ *such that*

$$\|M_p(u) - M_p(v)\|_p \leq C[\|u - v\|_\alpha \|u\|_p + \|u - v\|_p(\|v\|_\alpha + 1)]$$

for all $u, v \in D(A_p^\alpha)$.

Proof. By the definition of M_p we have

$$\|M_p(u) - M_p(v)\|_p \leq \|u - v\|_p + \|\text{grad}\,(u - v) \cdot \text{curl}\, E_p u\|_p + \|\text{grad}\, v \cdot \text{curl}\, E_p(u - v)\|_p. \tag{3.2}$$

From the Appendix and Proposition 3.3, we obtain:

$$\|\text{curl}\, E_p u\|_p \leq C\|u\|_p \tag{3.3}$$

and

$$\|E_p u\|_{1,\infty} \leq C\|u\|_\alpha \tag{3.4}$$

for all $u \in D(A_p^\alpha)$ and for some constant $C \geq 1$. Combining (3.2), (3.3), and (3.4), the desired estimate follows. □

Combining Theorem 3.1 and Proposition 3.4, we obtain that, for every $u_0 \in L^p(\Omega)$, there exists a solution u of (CP) on some interval $[0,T)$.

According to Theorem 3.1 (iv), the global existence of the solution follows if we can show that

$$\overline{\lim_{t\uparrow T}}\|u(t)\|_p < \infty.$$

By the imbedding $W^{2,p}(\Omega) \hookrightarrow C^1(\overline{\Omega})$, we can define $u(x,t) = u(t)(x)$ pointwise on $\Omega \times (0,T)$. Further, we have the following.

PROPOSITION 3.5. *Let $u_0 \in L^p(\Omega)$ and u be the corresponding solution of (CP) on $[0,T)$ in the sense of Theorem 3.1. Let $J \in C^2(\mathbf{R},\mathbf{R}^+)$ be a convex function; then we have*

$$\int_\Omega J(u(x,t))dx \leq \int_\Omega J(u(x,s))dx$$

for any $0 < s \leq t < T$.

Proof. Note that $J(u)$ is well defined due to the imbedding $W^{2,p}(\Omega) \hookrightarrow C^1(\overline{\Omega})$.

Multiplying the differential equation in (P') by $J'(u)$ and integrating the result over Ω gives

$$\frac{d}{dt}\int_\Omega J(u)dx = \int_\Omega J'(u)\Delta u\, dx + \int_\Omega J'(u)\text{grad}\, u \cdot \text{curl}\, v\, dx$$

for $0 < t < T$. Using Green's formula we know that

$$\int_\Omega J'(u)\Delta u\, dx = -\int_\Omega J''(u)[(\partial_1 u)^2 + (\partial_2 u)^2]dx \leq 0$$

and

$$\int_\Omega J'(u)\text{grad}\, u \cdot \text{curl}\, v\, dx = \int_{\partial\Omega} J(u)\frac{\partial v}{\partial \vec{\tau}}ds = 0,$$

where $\vec{\tau}$ is the tangential unit vector along $\partial\Omega$. Therefore,

$$\frac{d}{dt}\int_\Omega J(u)dx \leq 0,$$

which implies the required inequality. □

COROLLARY 3.6. *Let $u_0 \in L^p(\Omega)$ with $p \in (2,\infty]$ and u be the solution of (CP) on $[0,T)$ in the sense of Theorem 3.1. For any $q \in [2,p]$ we have*

$$\|u(t)\|_q \leq \|u_0\|_q \tag{3.5}$$

for $t \in [0,T)$.

Proof. This estimate follows directly from Proposition 3.5 by taking $J(s) = |s|^q$ and from the fact that $u \in C([0,T),X)$ for $p < \infty$. We obtain the estimate (3.5) for $p = q = \infty$ by using a limit argument. □

Using Theorem 3.1, Proposition 3.4, and Corollary 3.6, we obtain the following existence result for (CP).

THEOREM 3.7. *Let $\alpha \in (\frac{1}{2}+\frac{1}{p},1)$ and $u_0 \in L^p(\Omega)$. Then the initial value problem (CP) has a unique global solution $u(\cdot)$, i.e.,*

$$u \in C([0,\infty),X) \cap C((0,\infty),X_\alpha),$$

$$\sup_{0<t\leq 1} \|t^{\alpha} A_p^{\alpha} u(t)\| < \infty,$$

$$u(t) = e^{-tA_p} u_0 - \int_0^t e^{-(t-s)A_p} M(u(s)) ds$$

for $t \in (0, \infty)$, *and*

$$u(0) = u_0.$$

Moreover, u fulfills the equation $u' + A_p u + M_p(u) = 0$ *on* $(0, \infty)$ *in the sense that* $u \in C^1((0, \infty), X), u(t) \in D(A_p)$ *for* $t \in (0, \infty)$ *and* $A_p u \in C((0, \infty), X)$.

3.3. Regularity and asymptotic properties. In the preceding section we obtained the solution of the abstract problem (CP). Here we consider the original system (E), (P'). Let u be the solution of (CP). Then we have

$$u(t) \in W^{2,p}(\Omega), v(t) = E_p u(t) \in W^{2,p}(\Omega) \quad \forall t \in (0, \infty).$$

By the imbedding $W^{2,p}(\Omega) \hookrightarrow C^1(\overline{\Omega})$, we can define $u(x, t) = u(t)(x)$ and $v(x, t) = E_p u(t)(x)$ for $(x, t) \in \overline{\Omega} \times (0, \infty)$. The pair (u, v) satisfies the following

THEOREM 3.8. *Let* $\theta \in (0, 1 - \frac{2}{p})$, $\partial\Omega \in C^{2+\theta}$ *and suppose* $u_0 \in L^p(\Omega)$. *Let u,v be defined as above. Then* (u, v) *is the unique classical solution of the system* (E), (P'), *which satisfies*

(i) $u(\cdot, t) \in C^{2+\theta}(\overline{\Omega})$, $\partial_t u(\cdot, t) \in C^{\theta}(\overline{\Omega})$, $\forall t \in (0, \infty)$;

(ii) $u(x, \cdot) \in C^{1+\frac{\theta}{2}}(0, \infty)$ $\forall x \in \overline{\Omega}$;

(iii) $v(\cdot, t) \in C^{2+\theta}(\overline{\Omega})$, $\forall t \in (0, \infty)$.

Proof. (i) By the imbedding $W^{2,p}(\Omega) \hookrightarrow C^{1+\theta}(\overline{\Omega})$, we have

$$u(\cdot, t), v(\cdot, t) \in C^{1+\theta}(\overline{\Omega}) \quad \forall t \in (0, \infty).$$

Using Propositions 3.2 and 3.3, we also have

$$\partial_t u(\cdot, t) \in C^{\theta}(\overline{\Omega}) \quad \forall t \in (0, \infty).$$

Let $t_0 \in (0, \infty)$ be fixed. The regularity for u and v implies that

$$F(\cdot) = -\text{grad}\, u(\cdot, t_0) \cdot \text{curl}\, v(\cdot, t_0) + u(\cdot, t_0) - \partial_t u(\cdot, t_0)$$

satisfies

$$F(\cdot) \in C^{\theta}(\overline{\Omega}).$$

Next, consider the problem

$$\begin{cases} -\Delta w + w = F & \text{in } \Omega, \\ \dfrac{\partial w}{\partial \vec{\nu}} = 0 & \text{on } \partial\Omega. \end{cases}$$

By Gilbarg and Trudinger [8] this problem has a unique solution $w \in C^{2+\theta}(\overline{\Omega})$. A standard argument gives $w(\cdot) = u(\cdot, t_0)$, hence $u(\cdot, t_0) \in C^{2+\theta}(\overline{\Omega})$.

(ii) This is a direct result of (i) and Ladyzenskaja et al. [10, Thm. 5.3].

(iii) The regularity for v is a direct result of the Dirichlet problem (E). □

Remark. If the boundary $\partial\Omega$ is smooth, then the solution is smooth in $\overline{\Omega}\times(0,\infty)$. This follows from Theorem 3.8 together with a bootstrapping argument.

Let (u, v) be the solution of (E), (P'), a straightforward computation shows

$$\overline{u} := \frac{1}{|\Omega|}\int_\Omega u(x,t)dx = \frac{1}{|\Omega|}\int_\Omega u_0(x)dx$$

for all $t \in (0,\infty)$. Here $|\Omega|$ denotes the measure of Ω.

LEMMA 3.9. *We have*

$$\lim_{t\longrightarrow\infty} \|u(\cdot,t)-\overline{u}\|_2 = 0.$$

Proof. Taking $J(s) = s^2$ in the proof of Propositon 3.5, we obtain

$$\frac{d}{dt}\|u(\cdot,t)-\overline{u}\|_2^2 \leq -(\|\partial_1 u\|_2^2 + \|\partial_2 u\|_2^2).$$

We estimate the right-hand side by Poincaré's inequality. This gives

$$\|u(\cdot,t)-\overline{u}\|_2^2 \leq K(\|\partial_1 u\|_2^2 + \|\partial_2 u\|_2^2)$$

for some constant $K > 0$. Therefore,

$$\frac{d}{dt}\|u(\cdot,t)-\overline{u}\|_2^2 \leq -\frac{1}{K}\|u(\cdot,t)-\overline{u}\|_2^2,$$

which can be integrated to yield

$$\|u(\cdot,t)-\overline{u}\|_2^2 \leq e^{-t/K}\|u_0(\cdot)-\overline{u}\|_2^2, \tag{3.6}$$

for all $t \geq 0$. □

We now consider the asymptotic behavior of the solution in the sup-norm.

THEOREM 3.10. *Let $u_0 \in L^p(\Omega)$ for any $p \in (2,\infty]$. Then*

$$\lim_{t\longrightarrow\infty} \|u(\cdot,t)-\overline{u}\|_\infty = 0.$$

Proof. We put

$$\omega = \{U \in C(\overline{\Omega}) : \exists\{t_m\}, \text{ s.t. } \lim_{m\longrightarrow\infty} t_m = \infty \text{ and } \lim_{m\longrightarrow\infty} \|u(\cdot,t_m) - U(\cdot)\|_\infty = 0\}$$

and

$$F = \{u(\cdot,t) : t \in (0,\infty)\}.$$

From Corollary 3.6 and Theorem 3.8, it follows that F is a uniformly bounded and equicontinuous subset in $C(\overline{\Omega})$. Therefore, ω is nonempty. Next we show that ω contains only one single point. Let $U \in \omega$. Then there exists a sequence $\{t_m\}$ with

$$\lim_{m\longrightarrow\infty} t_m = \infty$$

and

$$\lim_{m\longrightarrow\infty} \|u(\cdot,t_m) - U(\cdot)\|_\infty = 0.$$

This implies

$$u(x, t_m) \to U(x)$$

as $m \to \infty$, uniformly in $x \in \overline{\Omega}$.

On the other hand, we obtain from Lemma 3.9 that

$$u(x, t_m) \to \overline{u}$$

as $m \to \infty$, for almost all $x \in \Omega$. Thus

$$U(x) = \overline{u} \quad \forall x \in \overline{\Omega},$$

which completes the proof. □

4. The quasilinear case.

4.1. The abstract setting. In this section we study Problem (E), (P). As in §3, we treat this system as an abstract evolution equation in a suitably chosen Banach space. In this part we collect some results on quasilinear evolution equations.

Let $\overline{E} = (E_0, E_1)$ be a pair of Banach spaces with E_1 continuously and densely imbedded in E_0. We denote by $\mathcal{H}(\overline{E})$ the set of all $A \in \mathcal{L}(E_1, E_0)$ such that $-A$, considered as a linear operator on E_0, is the infinitesimal generator of a strongly continuous analytic semigroup on E_0. For $\theta \in (0,1)$, let E_θ be the complex interpolation space $[\overline{E}]_\theta$, and $\|\cdot\|_\theta$ be the norm on E_θ. (The notation here is different from the previous section.)

Let $T > 0$ be fixed. We assume (Q) $\beta \in (0,1), V \subset E_\beta$ is open and $A \in C^{1-}(V, \mathcal{H}(\bar{E}))$, i.e., A is locally Lipschitz continuous.

Under these assumptions, we consider the following quasilinear Cauchy problem

$$(QCP)_{(u_0)} : \begin{cases} \dot{u}(t) + A(u(t))u(t) = 0, \ 0 < t \le T, \\ u(0) = u_0, \end{cases}$$

where $u_0 \in V$.

Let $\tau \in (0, T]$; u is called a solution of $(QCP)_{(u_0)}$ on $[0, \tau]$ if the following conditions are satisfied:

(i) $u \in C([0,\tau], V) \cap C((0,\tau], E_1) \cap C^1((0,\tau], E_0)$,

(ii) $\dot{u}(t) + A(u(t))u(t) = 0, \forall t \in (0, \tau]$,

(iii) $u(0) = u_0$.

A solution u is *maximal* if there does not exist a solution of $(QCP)_{(u_0)}$ which is a proper extension of u. In this case the interval of existence is called the *maximal interval of existence.*

The following fundamental theorem can be found in Amann [2] (see also Sobolevskii [15]).

THEOREM 4.1. *Suppose that* $0 < \beta < \alpha < 1$, *and* $u_0 \in V_\alpha := E_\alpha \cap V$. *Furthermore, suppose that the assumption* (Q) *holds. Then there exists* $\tau > 0$, *such that* $(QCP)_{u_0}$ *has a unique solution* $u(\cdot)$ *on* $[0, \tau]$, *satisfying* $u \in C([0, \tau], V_\alpha)$. *Moreover, the maximal interval of existence is open in* $[0, T]$. □

4.2. Local existence. Again we put the system into an abstract form.

Let $\Omega \subset \mathbf{R}^2$ be a bounded domain with smooth boundary $\partial\Omega$. For $p \in (1, \infty)$ and $r \in (-\infty, \infty)$, we denote by $H_p^r(\Omega)$ the so-called Lebesgue spaces (see Triebel [17] or Bergh and Löfström [6]). In this section the norm on $H_p^r(\Omega)$ is denoted by $\|\cdot\|_{r,p}$.

It should be observed that $H_p^m(\Omega) = W^{m,p}(\Omega)$ for integer m. Moreover, we have the interpolation property

$$[H_{p_0}^{s_0}(\Omega), H_{p_1}^{s_1}(\Omega)]_\theta = H_p^s(\Omega) \tag{4.1}$$

for $s_0, s_1 \in \mathbf{R}$, $p_0, p_1 \in (1,\infty)$ with $\frac{1}{p} = (1-\theta)/p_0 + \theta/p_1$ and $s = (1-\theta)s_0 + \theta s_1$.

Let $a_{jk} := D_{jk} \circ Q$ and $a_j = -Q_j$ for $j, k = 1, 2$ (see Appendix). Then problem $(E), (P)$ can be formulated as

$$(QCP) \begin{cases} \partial_t u - \partial_j(a_{jk}(u)\partial_k u + a_j(u)u) = 0 & \text{in } \Omega \times (0,T], \\ \nu^j a_{jk}(u)\partial_k u + a_j(u)\nu^j u = 0 & \text{on } \partial\Omega \times (0,T], \\ u(\cdot, 0) = u_0 & \text{in } \Omega. \end{cases}$$

Here $T > 0$ and $\vec{\nu} = (\nu^1, \nu^2)$. Note that in this section the summation convention is used and the indices run from 1 to 2.

We use Theorem 4.1 to obtain the existence result for Problem (QCP). In this application we take

$$E_0 = (H_{p'}^1(\Omega))'$$

and

$$E_1 = H_p^1(\Omega),$$

where $p \in (1,\infty)$ and $\frac{1}{p} + \frac{1}{p'} = 1$. It should be observed that

$$E_\theta = [E_0, E_1]_\theta \hookrightarrow L^p(\Omega) \tag{4.2}$$

for $\theta \in [\frac{1}{2}, 1]$; see Amann [4, Thm. 3.3].

Let $\mathcal{M}(\Omega) \subset C(\overline{\Omega})^4 \times C(\overline{\Omega})^2$ be the subset whose elements $m(\cdot) = (b_{jk}(\cdot), b_j(\cdot))$ are chosen such that $(b_{jk}(\cdot))_{2\times 2}$ is uniformly positive definite on $\overline{\Omega}$. Assume we set

$$\langle f, g \rangle = \int_\Omega f(x)g(x)dx$$

for $f \in L^p(\Omega)$, $g \in L^{p'}(\Omega)$. With this notation we define

$$a(m)(v,u) = \langle \partial_j v, b_{jk}\partial_k u + b_j u \rangle$$

for $v \in H_{p'}^1(\Omega)$, $u \in H_p^1(\Omega)$, and $m \in \mathcal{M}(\Omega)$.

Furthermore, given $m \in \mathcal{M}(\Omega)$, we define the operator

$$A(m) : E_1 \to E_0$$

such that

$$\langle A(m)u, v \rangle = a(m)(v,u) \quad \forall v \in H_{p'}^1(\Omega).$$

Then we have the following generation theorem; see Amann [3] or Lunardi and Vespri [11].

THEOREM 4.2.

$$[m \to A(m)] \in C^{1-}(\mathcal{M}(\Omega), \mathcal{H}(\overline{E})).$$

For $p \in (2, \infty)$ and $r > \frac{2}{p}$, we have

$$H_p^r(\Omega) \hookrightarrow C(\overline{\Omega}), \tag{4.3}$$

Therefore, the coefficients $a_{jk}(u)$, $a_j(u)$ are defined pointwise on $\overline{\Omega}$ for each $u \in H_p^r(\Omega)$. Consequently,

$$m(u)(\cdot) := (a_{jk}(u)(\cdot), a_j(u)(\cdot))$$

is well defined on $\overline{\Omega}$. For m we also have the following.

LEMMA 4.3. *Let $p \in (2, \infty)$ and $1 \geq r > \frac{2}{p}$. Then $[u \to m(u)] : H_p^r(\Omega) \to \mathcal{M}(\Omega)$ is uniformly Lipschitz continuous.*

Proof. From the appendix we have

$$Q_i \in \mathcal{L}(H_p^r(\Omega)). \tag{4.4}$$

We combine this with imbedding (4.3) and Proposition 2.1 to obtain

$$m(u) \in \mathcal{M}(\Omega).$$

On the other hand, by Proposition 2.1, (4.3), and (4.4), there exists a constant $C > 0$ such that

$$\|a_{jk}(u) - a_{jk}(v)\|_{C(\overline{\Omega})} \leq C\|u - v\|_{r,p}$$

and

$$\|a_j(u) - a_j(v)\|_{C(\overline{\Omega})} \leq C\|u - v\|_{r,p}$$

for any $u, v \in H_p^r(\Omega)$ and for $j, k = 1, 2$. This completes the proof. □

Let us put $A(u) := A(m(u)(\cdot))$ We are now in a position to prove the main existence result.

THEOREM 4.4. *Let $p \in (2, \infty)$ and $\frac{1}{2} + \frac{1}{p} < \beta < \alpha < 1$. For every $u_0 \in E_\alpha$, there exists a $\tau > 0$ such that*

$$\begin{cases} \dot{u}(t) + A(u(t))u(t) = 0, & 0 < t \leq \tau, \\ u(0) = u_0, \end{cases}$$

has a unique solution $u(\cdot)$ on $[0, \tau]$, i.e.,

(i) $u \in C([0, \tau], E_\alpha) \cap C((0, \tau], E_1) \cap C^1((0, \tau], E_0)$,
(ii) $\dot{u}(t) + A(u(t))u(t) = 0, \forall t \in (0, \tau]$,
(iii) $u(0) = u_0$.

Proof. For $\beta = \frac{1}{2} + \frac{r}{2} \in (0, 1)$ we have

$$E_\beta \hookrightarrow [E_{\frac{1}{2}}, E_1]_r,$$

by the reiteration theorem (see Triebel [17] or Bergh and Löfström [6]). Using (4.2), we have

$$E_\beta \hookrightarrow [L^p(\Omega), H_p^1(\Omega)]_r = H_p^r(\Omega)$$

with $r \in (0, 1)$. Finally, if $1 > \beta > \frac{1}{2} + \frac{1}{p}$, then $1 > r > \frac{2}{p}$ and $H_p^r(\Omega) \hookrightarrow C(\overline{\Omega})$. From Lemma 4.3 we know $[u \to m(u)]$ is uniformly Lipschitz continuous from E_β to $\mathcal{M}(\Omega)$.

On the other hand, it follows from Theorem 4.2 that

$$[m \to A(m)] \in C^{1-}(\mathcal{M}(\Omega), \mathcal{H}(\overline{E})).$$

Hence

$$[u \to A(u)] \in C^{1-}(E_\beta, \mathcal{H}(\overline{E})).$$

The conclusion then follows directly from Theorem 4.1. □

4.3. Some properties of the weak solution. Up to now we have obtained a local solution for Problem (QCP) in $H_p^1(\Omega)$-sense. We now come back to the original system.

Let $\tau > 0$, $u_0 \in E_\alpha$ for some $\alpha \in (\frac{1}{2} + \frac{1}{p}, 1)$ and we suppose $u \in C^1((0,\tau], E_0) \cap C((0,\tau], E_1)$ is the weak solution mentioned in Theorem 4.4. By the appendix we know that $v(t) = K \circ \partial_1 u(t) \in H_p^2(\Omega)$. Using the imbedding $H_p^1(\Omega) \hookrightarrow C(\overline{\Omega})$, we can define

$$u(x,t) := u(t)(x)$$

and

$$v(x,t) := v(t)(x)$$

pointwise on $\overline{\Omega} \times (0,\tau]$. Obviously, we have

$$\partial_t u(x,t) = \dot{u}(t) \in L^p(\Omega).$$

From Theorem 4.4 we know that problem (P) is satisfied in the following sense:

$$\frac{d}{dt}\int_\Omega u(x,t) f(x) dx - \int_\Omega \vec{F}(x,t) \text{grad } f(x) dx = 0 \tag{4.5}$$

for all $f \in H_{p'}^1(\Omega)$ and $t \in (0,\tau]$. Moreover, $u(x,0) = u_0$.

As in the semilinear case we can prove the following.

THEOREM 4.5. *Let (u,v) be the weak solution of $(E), (P)$ as constructed above. Then*

$$\|u(\cdot,t)\|_p \leq \|u_0\|_p$$

for all $t \in (0,\tau]$.

Proof. Using the facts

$$u(\cdot,t) \in C(\overline{\Omega})$$

and

$$L^p(\Omega) \hookrightarrow L^{p'}(\Omega),$$

we obtain immediately that $f := p|u|^{p-1}\text{sgn } u \in H_{p'}^1(\Omega)$.

Substitution into (4.5) gives

$$\frac{d}{dt}\|u(\cdot,t)\|_p^p = \int_\Omega (u \text{ curl } v - D \cdot \text{grad } u) \cdot p(p-1)|u|^{p-2} \text{grad } u \, dx \tag{4.6}$$

Since the matrix D is positive definite,

$$-\int_\Omega (D \cdot \operatorname{grad} u) \cdot p(p-1)|u|^{p-2}\operatorname{grad} u \, dx \leq 0.$$

On the other hand, by using the Green's formula we have

$$\int_\Omega (u \operatorname{curl} v) \cdot p(p-1)|u|^{p-2}\operatorname{grad} u \, dx = 0.$$

Therefore, the conclusion follows directly from (4.6). □

Final remark. In this paper we assumed Ω to be a bounded domain of $\mathbf{R}^2$ with smooth boundary. On the other hand, the domain in the motivating problem is a rectangle. For such a domain, the same existence results will hold. This is a consequence of the fact that the generation theorems for the operators A_p in §3 and A in §4.2, as well as the proposition in Appendix also hold for such a domain (Vespri [18]).

Appendix. Here we state some results on the Laplace operator with Dirichlet boundary condition, which are related to problem (E).

Let γ denote the trace operator. It is known that the operator $-\Delta$ with Dirichlet boundary condition zero is invertible in $L^p(\Omega)$. We denote this inverse operator by

$$K := (-\Delta \mid \gamma)^{-1}.$$

Further, we introduce operator

$$Q = (Q_1, Q_2) = \operatorname{curl} K \, \partial_1.$$

Let $H^r_p(\Omega)$ be the Lebesgue spaces, with indices $-\infty < r < +\infty$, $1 < p < \infty$.

The operator Q satisfies the following.

PROPOSITION. *Let $r \in [0,1]$ and $1 < p < \infty$. Then*

$$Q_i \in \mathcal{L}(H^r_p(\Omega))$$

for $i = 1, 2$.

Proof. Let $f \in L^p(\Omega)$. We define

$$Fv = \int_\Omega f \partial_1 v \, dx$$

for $v \in W^{1,p'}_0(\Omega)$. Clearly, $F \in (W^{1,p'}_0(\Omega))'$. By the representation theorem in Simader [14, p. 91], we know that there exists $u \in W^{1,p}_0(\Omega)$ such that

$$Fv = \int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v \, dx$$

for $v \in W^{1,p'}_0(\Omega)$. Moreover, there exists a constant C independent of u and f such that

$$\|u\|_{1,p} \leq C\|f\|_p.$$

Therefore,

$$Q_i \in \mathcal{L}(L^p(\Omega)).$$

On the other hand, it is well known that

$$Q_i \in \mathcal{L}(H_p^1(\Omega)).$$

By the interpolation property,

$$[H_p^{s_0}(\Omega), H_p^{s_1}(\Omega)]_\theta = H_p^s(\Omega)$$

for $\theta \in [0,1]$ and s_0, s_1, $s \in \mathbf{R}$ with $s = (1-\theta)s_0 + \theta s_1$; the conclusion follows.

Acknowledgment. The authors are very grateful to Professor H. Amann for his help during his visit in Delft.

REFERENCES

[1] S. Agmon, *On the eigenfunctions and on the eigenvalues of general elliptic boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 119–147.

[2] H. Amann, *Quasilinear evolution equation and parabolic systems*, Trans. Amer. Math. Soc. 293 (1986), pp. 191–227.

[3] ———, *Dynamic theory of quasilinear parabolic equations* III. *Global existence*, Math. Z., 202 (1989), pp. 219–250. Erratum, Math. Z., 205 (1990), p. 331.

[4] ———, *On abstract parabolic fundamental solutions*, J. Math. Soc. Japan, 39 (1987), pp. 93–116.

[5] J. Bear, *Dynamics of Fluids in Porous Media*, American Elsevier, New York, 1972.

[6] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Springer-Verlag, Berlin, 1976.

[7] A. Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, Chicago, San Francisco, 1969.

[8] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1983.

[9] D. Henry, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1968.

[10] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[11] A. Lunardi and V. Vespri, *Hölder regularity in variational parabolic non-homogeneous equations*, J. Differential Equations, 94 (1991), pp. 1–40.

[12] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[13] R. de Roo, *Globale oplossingen van een abstract Cauchy probleem*, Afstudeerverslag, Delft University of Technology, the Netherlands, 1987.

[14] C. G. Simader, *On Dirichlet's boundary value problem*, Lecture Notes in Math. 268, Springer-Verlag, New York, Berlin, 1972.

[15] P. E. Sobolevskii, *Equations of parabolic type in a Banach space*, Amer. Math. Soc. Transl., 49 (1966), pp. 1–62.

[16] N. Su, *The mathematical problems on the fluid-solute-heat flow through porous media*, Ph.D. thesis, Tsinghua University, Beijing, 1987.

[17] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

[18] V. Vespri, personal communication.

[19] W. von Wahl, *The Equations of Navier–Stokes and Abstract Parabolic Equations*, Aspects of Math., Vieweg, Braunschweig, 1985.

[20] ———, *On the Cahn–Hilliard Equation* $u' + \Delta^2 u - \Delta f(u) = 0$, Delft Progr. Rep., 10 (1985), pp. 291–310.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 852-869, July 1992

FINAL TIME BLOWUP PROFILES FOR SEMILINEAR PARABOLIC EQUATIONS VIA CENTER MANIFOLD THEORY*

J. BEBERNES† AND S. BRICHER‡

Abstract. This paper considers the semilinear parabolic equation $u_t = \Delta u + f(u)$ in $\mathbb{R}^n \times (0, \infty)$, where $f(u) = e^u$ or $f(u) = u^p$, $p > 1$. For any initial data that is a positive, radially decreasing lower solution, and that causes the corresponding solution $u(x, t)$ to blow up at $(0, T) \in \mathbb{R}^n \times (0, \infty)$, the authors prove by using techniques from center manifold theory that the final time blowup profiles satisfy

$$u(x, T) = -2 \ln |x| + \ln |\ln |x|| + \ln 8 + o(1) \quad \text{for } f(u) = e^u,$$

$$u(x, T) = \left(\frac{8\beta^2 p |\ln |x||}{|x|^2}\right)^\beta (1 + o(1)) \quad \text{for } f(u) = u^p$$

as $|x| \to 0$.

Key words. semilinear parabolic equations, blowup, asymptotic behavior of solutions, center manifold theory

AMS(MOS) subject classifications. 35B40, 35K55, 35K57, 34C30

1. Statement of problem and results. Consider the initial-value problem

$$u_t - \Delta u = f(u), \qquad (x, t) \in \mathbb{R}^n \times (0, \infty), \tag{1.1}$$

$$u(x, 0) = \phi(x), \qquad x \in \mathbb{R}^n, \tag{1.2}$$

where

$$f(u) = e^u \quad \text{or} \quad f(u) = u^p, \quad p > 1 \tag{1.3}$$

and ϕ satisfies

$$\begin{cases} \phi \in C^2(\mathbb{R}^n; [0, \infty)) \text{ is radially symmetric,} \\ \phi = \phi(r), \qquad r = |x|, \\ \phi'(0) = 0, \phi''(0) < 0, \phi'(r) < 0 \quad \text{for } r \in (0, \infty), \\ \Delta\phi + f(\phi) \geqq 0 \text{ with } \phi \text{ not a steady-state solution.} \end{cases} \tag{1.4}$$

The following facts are well known for (1.1)-(1.2) [2], [9].

(i) There exists a unique nonnegative solution.

(ii) The solution $u(x, t)$ is radially symmetric; thus (1.1)-(1.2) is equivalent to

$$u_t = u_{rr} + \frac{n-1}{r} u_r + f(u), \qquad (0, \infty) \times \{t : t > 0\}, \tag{1.1'}$$

$$\begin{aligned} u(r, 0) &= \phi(r), \qquad r \in (0, \infty), \\ u_r(0, t) &= 0, \qquad t > 0. \end{aligned} \tag{1.2'}$$

(iii) $u(\cdot, t)$ is radially decreasing, $u_r(r, t) < 0$, and $u_t(r, t) > 0$.

We assume $\phi(x) \geqq 0$ is such that $u(x, t)$ blows up in finite time $T < \infty$; that is, $\sup_{\mathbb{R}^n} |u(x, t)| \to \infty$ as $t \to T^-$ and $\sup_{\mathbb{R}^n \times [0,\tau]} |u(x, t)| < \infty$ for all $0 < \tau < T$.

* Received by the editors August 27, 1990; accepted for publication (in revised form) November 7, 1991. This research was supported by U.S. Army Research Office contract DAAL 03-88-K0111.

† Program in Applied Mathematics, University of Colorado, Boulder, Colorado 80309-0526.

‡ Department of Mathematics, University of Colorado, Boulder, Colorado 80309-0426.

(iv) Blowup occurs only at $x=0$. Moreover; if $f(u)=e^u$, then for any $\alpha\in(0,1)$,

$$u(x,t)\leqq -\frac{1}{\alpha}\ln\left(\frac{\alpha\varepsilon|x|^2}{2}\right),$$

and if $f(u)=u^p$, then for any $\gamma\in(1,p)$,

$$u(x,t)\leqq\left(\frac{(\gamma-1)\varepsilon|x|^2}{2}\right)^{1/1-\gamma}$$

for $(x,t)\in B_R\times(0,T]$, where B_R is some ball centered at the origin and $\varepsilon>0$ is sufficiently small.

(v)

$$(1.5)\qquad \begin{aligned} &(a)\quad -\ln(T-t)\leqq u(0,t),\ t\in[0,T) &&\text{for } f(u)=e^u,\\ &(b)\quad \beta^\beta(T-t)^{-\beta}\leqq u(0,t),\ t\in[0,T) &&\text{for } f(u)=u^p \end{aligned}$$

where $\beta=1/(p-1)$.

(vi)

$$(1.6)\qquad \begin{aligned} &(a)\quad u(x,t)\leqq -\ln[\delta(T-t)],\ t\in[0,T),\ \delta>0 &&\text{for } f(u)=e^u,\\ &(b)\quad u(x,t)\leqq\left(\frac{\beta}{\delta}\right)^\beta(T-t)^{-\beta},\ t\in[0,T) &&\text{for } f(u)=u^p. \end{aligned}$$

(vii) There exists $\bar{t}<T$ such that

$$(1.7)\qquad \begin{aligned} &(a)\quad |\nabla u(x,t)|\leqq[2e^{u(0,t)}]^{1/2} &&\text{for } f(u)=e^u,\\ &(b)\quad |\nabla u(x,t)|\leqq\left[\frac{2}{p+1}(u(0,t))^{p+1}\right]^{1/2} &&\text{for } f(u)=u^p, \end{aligned}$$

where $(x,t)\in\mathbb{R}^n\times[\bar{t},T)$.

The purpose of this paper is to give a precise description of the asymptotic behavior of solutions $u(x,t)$ of (1.1)-(1.2) as the blowup time T is approached. There has been a considerable effort to resolve this problem in recent years (see [2], [9], [12]). Until very recently, the best rigorous result in this direction for the problem under consideration is the following theorem.

THEOREM 1.1. *If $u(x,t)$ is a solution of* (1.1)-(1.2) *which blows up at* $(0,T)$, *then*

$$(1.8)\qquad \begin{aligned} &(a)\quad \lim_{t\to T}[\ln(T-t)+u(x,t)]=0 &&\text{for } f(u)=e^u,\\ &(b)\quad \lim_{t\to T}u(x,t)(T-t)^\beta=\beta^\beta &&\text{for } f(u)=u^p \end{aligned}$$

uniformly for $|x|\leqq C(T-t)^{1/2}$, $C\geqq0$, *as* $t\to T^-$.

Since $u(x,t)$ blows up only at $x=0$, $u_t(x,t)\geqq0$, and $\sup_{[0,T]}u(x,t)<\infty$ for each $x\neq0$, $u(x,t)\to u_F(x)$ as $t\to T^-$ for all $x\in\mathbb{R}^n-\{0\}$. This final time solution profile $u_F(x)$ should be describable in a neighborhood of the blowup point $x=0$. This observation led Kassoy and Poland [17], [18] and Kapila [19] to formally attempt to describe $u_F(x)$ using singular perturbation techniques for $f(u)=e^u$. Dold [6] for $f(u)=e^u$ and Galaktionov [11] for $f(u)=u^p$ extended these singular perturbation ideas to predict the behavior of $u(x,t)$ near the blowup time, but their analyses are nonrigorous and formal.

Filippas and Kohn [8] are the first to have observed that a center manifold approach to such problems can be used to precisely describe, to higher-order terms, the asymptotic behavior near blowup. Bressan [4] considers (1.1) with $f(u)=e^u$ in a convex domain $\Omega\subset\mathbb{R}^n$. Given any $b\in\Omega$, he proves the existence of solutions that blow up in finite time exactly at b, and whose final profile satisfies

$$u_F(x)=u(x,T)=-2\ln|x-b|+\ln|\ln|x-b||+\ln 8+O([\ln|x-b|]^{-1/3})$$

and proves that this asymptotic behavior is stable with respect to small perturbations of initial conditions. Herrero and Velázquez [15], [16] consider the one-dimensional problem (1.1)-(1.2). Without using center manifold theory, they prove some results suggested by perturbation techniques that will be described in more detail in conjunction with our results.

In this paper, we prove the following three theorems.

THEOREM 1.2. *Let $u(x,t)$ be the solution of* (1.1)-(1.2); *then*

$$\begin{aligned}&\textit{for } f(u)=e^u,\\ &\text{(a)}\quad u(x,t)\sim\ln\left(\frac{1}{T-t}\right)+\frac{1}{4\ln(T-t)}\left(\frac{|x|^2}{T-t}-2n\right)+o\left(\frac{1}{\ln(T-t)}\right)\\ (1.9)\qquad &\textit{for } f(u)=u^p,\\ &\text{(b)}\quad (T-t)^\beta u(x,t)\sim\beta^\beta+\frac{\beta^\beta}{4p\ln(T-t)}\left(\frac{|x|^2}{T-t}-2n\right)+o\left(\frac{1}{\ln(T-t)}\right)\end{aligned}$$

uniformly on $|x|\leqq C(T-t)^{1/2}$, $C\geqq 0$, as $t\to T^-$.

For $n=1$ and $f(u)=u^p$, (1.9(b)) was first obtained by Filippas and Kohn [8], substantiating the conjecture of Galaktionov [11]. Herrero and Velázquez [15] proved both (1.9(a)) and (1.9(b)) for $n=1$. Our proof given in § 2 is influenced by that of Filippas and Kohn [8], where they utilize ideas related to a center manifold theory for infinite-dimensional dynamical systems.

THEOREM 1.3. *Let $u(x,t)$ be the solution of* (1.1)-(1.2); *then*

$$\begin{aligned}&\textit{for } f(u)=e^u,\\ &\text{(a)}\quad \lim_{t\to T}\left[u(\eta((T-t)|\ln(T-t)|)^{1/2},t)+\ln(T-t)\right]=-\ln\left(1+\frac{|\eta|^2}{4}\right)\\ (1.10)\qquad &\textit{for } f(u)=u^p,\\ &\text{(b)}\quad \lim_{t\to T}(T-t)^\beta u(\eta((T-t)|\ln(T-t)|)^{1/2},t)=\beta^\beta\left[1+\frac{|\eta|^2}{4p\beta}\right]^{-\beta}\end{aligned}$$

uniformly on compacts in η.

This result was conjectured and formally verified by Dold [6], [7] in the "ignition-kernel" variable $\eta=x/((T-t)|\ln(T-t)|)^{1/2}$ for $f(u)=e^u$ and by Galaktionov [11] for $f(u)=u^p$. Bressan [4] proved (1.10(a)) for the Cauchy-Dirichlet problem and for some initial conditions $\phi(x)$ whose corresponding solution blows up at the origin. He also showed that the same holds for all initial conditions sufficiently close to ϕ. Bressan's proof includes higher-order terms, and so for certain initial conditions, his result is an improvement of Theorem 1.3 for the case $f(u)=e^u$. This allows him to obtain a more precise estimate of the error term in the final time profile. For $n=1$, Herrero and Velázquez [15] proved (1.10(a)) and (1.10(b)). Theorem 1.3 extends their result to the multidimensional case for any ϕ satisfying (1.4), but is proven by using their techniques. For this reason, we only outline the ideas of the proof in § 3.

THEOREM 1.4. *Let $u(x, t)$ be any solution of* (1.1)-(1.2) *with initial data $\phi(x)$ satisfying* (1.4); *then*

$$\text{(1.11)}\quad \begin{aligned}&\textit{for } f(u)=e^u,\\ &\text{(a)}\quad u_F(x)=u(x,T)=-2\ln|x|+\ln|\ln|x||+\ln 8+o(1)\\ &\textit{for } f(u)=u^p,\\ &\text{(b)}\quad u_F(x)=u(x,T)=\left(\frac{8\beta^2p|\ln|x||}{|x|^2}\right)^{\beta}(1+o(1))\end{aligned}$$

as $|x|\to 0$.

By fact (iv) (see [2, p. 66]), we have the upper bounds

$$\text{(1.12)}\quad \begin{aligned}&\text{for } f(u)=e^u,\\ &\text{(a)}\quad u(x,t)\leqq -2\ln|x|+\ln|\ln|x||+C\\ &\text{for } f(u)=u^p,\\ &\text{(b)}\quad u(x,t)\leqq\left(\frac{k|\ln|x||}{|x|^2}\right)^{\beta}\end{aligned}$$

for all $(x, t)\in B_R\times(0, T]$, where k and C are constants. Theorem 1.4 gives a precise description of $u_F(x)$ in a neighborhood of the singularity $x=0$. Bressan [4] has the first rigorous result of this type provided $f(u)=e^u$ and ϕ is sufficiently spiked. Herrero and Velázquez [16] proved (1.11(b)) when $\phi\in C_b(\mathbb{R}; [0,\infty))$ and the corresponding solution $u(x, t)$ of (1.1)-(1.2) satisfies (1.10(b)). If we assume ϕ satisfies (1.4) and extend their proof to higher dimensions, we obtain (1.11(b)) provided $n\leqq 2$ or $n\geqq 3$ and $p\leqq(n+2)/(n-2)$.

Our proof uses ideas from [16], but avoids the spiked condition on ϕ for $f(u)=e^u$ and the requirement $p\leqq(n+2)/(n-2)$ when $n\geqq 3$ for $f(u)=u^p$ by making use of the upper bounds (1.12). Furthermore, it shows that for any ϕ satisfying (1.4), for which the solution $u(x, t)$ of (1.1)-(1.13) blows up in finite time, $u_F(x)$ satisfies the estimate (1.11).

In § 2, we prove Theorem 1.2 using center manifold theory. We then extend to the ignition variable grouping to obtain Theorem 1.3 in § 3. Finally, we prove Theorem 1.4 in § 4.

Throughout, we will use the following function space notation. For $1\leqq p<\infty$, let $L^p_\rho\equiv\{f\in L^p_{\text{loc}}(\mathbb{R}^n):\int_{\mathbb{R}^n}|f|^p e^{-|x|^2/4}\,dx<\infty\}$, where $\rho=e^{-|x|^2/4}$, $L^p_{\rho 0}=\{f\in L^p_\rho: f$ radially symmetric on $\mathbb{R}^n\}$, $H^m_\rho\equiv\{f\in L^2_{\text{loc}}(\mathbb{R}^n): f^{(j)}\in L^2_{\text{loc}}(\mathbb{R}^n)$, and $\int_{\mathbb{R}^n}|f^{(j)}|\,e^{-|x|^2/4}\,dx<\infty$, $j\in[0, m]\}$ is the weighted Sobolev space, and $H^m_{\rho 0}\equiv\{f\in H^m_\rho: f$ radially symmetric on $\mathbb{R}^n\}$. We will use the standard asymptotic notation $o(\cdot)$, $O(\cdot)$, $\ll$, $\sim$ when convenient.

2. Center-unstable manifold analysis of ignition models. Suppose that $u(x, t)$ is a solution of (1.1)-(1.2), which blows up at $(0, T)\in\mathbb{R}^n\times(0,\infty)$. To analyze the asymptotic behavior of $u(x, t)$, make the "hot-spot" change of variables:

$$\text{(2.1)}\qquad s=-\ln(T-t),\qquad y=\frac{x}{(T-t)^{1/2}},$$

and let

$$\text{(2.2)}\quad \begin{aligned}&\text{(a)}\quad w(y,s)=u(x,t)+\ln(T-t)\quad \text{for } f(u)=e^u,\\ &\text{(b)}\quad w(y,s)=(T-t)^{\beta}u(x,t)\qquad \text{for } f(u)=u^p.\end{aligned}$$

Then $w(y, s)$ satisfies

$$w_s = \Delta w - \tfrac{1}{2} y \cdot \nabla \omega + F(w), \qquad \mathbb{R}^n \times (s_0, \infty),\ s_0 \equiv -\ln (T) \tag{2.3}$$

or, equivalently,

$$w_s = \frac{1}{\rho} \nabla \cdot (\rho \nabla w) + F(w), \qquad \rho = e^{-|y|^2/4}, \tag{2.4}$$

where

$$\begin{aligned} &\text{(a)} \quad F(w) = e^w - 1 && \text{for } f(u) = e^u, \\ &\text{(b)} \quad F(w) = w^p - \frac{1}{p-1} w && \text{for } f(u) = u^p. \end{aligned} \tag{2.5}$$

Then $w(y, s)$ satisfies

$$\nabla w(0, s) = 0 \quad \text{for } s \geqq s_0, \tag{2.6}$$

and

$$\begin{aligned} &\text{(a)} \quad w(y, s_0) = \phi(T^{1/2} y) + \ln (T) && \text{for } f(u) = e^u, \\ &\text{(b)} \quad w(y, s_0) = T^\beta \phi(T^{1/2} y) && \text{for } f(u) = u^p. \end{aligned} \tag{2.7}$$

As a consequence of facts (v)-(vii),

$$|\nabla w| \leqq c, \tag{2.8}$$

and

$$\begin{aligned} &\text{(a)} \quad w(y, s) \leqq c,\ |w(y, s)| \leqq c(1 + |y|) && \text{for } f(u) = e^u, \\ &\text{(b)} \quad 0 < w(y, s) \leqq c && \text{for } f(u) = u^p. \end{aligned} \tag{2.9}$$

The following theorem can be found in [2] and [20].

THEOREM 2.1. *The solution $w(y, s)$ converges to $S(y)$ uniformly on compacts $|y| \leqq C$ as $s \to \infty$, where*

$$S(y) = 0 \quad \text{for } f(u) = e^u \quad \text{or} \quad S(y) = \beta^\beta \quad \text{for } f(u) = u^p. \tag{2.10}$$

This theorem gives as an immediate consequence Theorem 1.1. It gives a good description of the temporal evolution as the blowup time T is approached, but the spatial variable is stretched too much to give any information concerning the spatial variable. To get more information about u near blowup, we need to analyze more precisely how u approaches $S(y)$.

To do this, a natural first step would be to linearize about the steady-state solution $S(y)$. We set

$$v(y, s) = w(y, s) - S(y). \tag{2.11}$$

By Theorem 2.1, $\lim_{s \to \infty} v(y, s) = 0$ uniformly on compacts in $\mathbb{R}^n$. From (2.4) and (2.5), $v(y, s)$ satisfies

$$v_s = \frac{1}{\rho} \nabla \cdot (\rho \nabla v) + g(v), \tag{2.12}$$

where

$$(2.13)\qquad \begin{aligned}&\text{(a)}\quad g(v)=v+\frac{v^2}{2} &&\text{for } f(u)=e^u,\\ &\text{(b)}\quad g(v)=v+\frac{p}{2\beta^\beta}v^2 &&\text{for } f(u)=u^p,\end{aligned}$$

neglecting order three and higher terms.

Abstractly, we can write (2.12) as

$$(2.14)\qquad v'=Av+Nv,$$

where $Av=1/\rho\nabla\cdot(\rho\nabla v)+v$ and $Nv=g(v)-v$. We observe that A generates a semigroup, and the spectrum of A, $\sigma(A)$, is the point spectrum $\sigma_p(A)$.

We begin by determining the eigenvalues and eigenvectors of A. From $Av=\lambda v$, we have

$$(2.15)\qquad \Delta v-\tfrac{1}{2}y\cdot\nabla v+(1-\lambda)v=0.$$

It can be shown (see, for example, [8]) that the eigenvalues are

$$(2.16)\qquad \lambda_m=-\frac{m}{2}+1,\qquad m=0,1,2,\cdots$$

with the associated eigenfunctions being in dimension $n=1$,

$$(2.17)\qquad h_m(y)=H_m\left(\frac{y}{2}\right),$$

where H_m is the mth Hermite polynomial. The first three eigenfunctions are

$$(2.18)\qquad h_0(y)=1,\quad h_1(y)=y,\quad h_2(y)=y^2-2.$$

In higher dimensions the eigenfunctions are formed by taking products of the polynomials $\{h_m\}_{m=0}^{\infty}$. It is easy to see that the products $h_{m_1}(y_1)\cdots h_{m_n}(y_n)$ form an orthogonal basis for L^2_ρ.

Let $Z=L^2_{\rho_0}$ be the Hilbert space of radially symmetric functions in L^2_ρ with inner product $\langle f,g\rangle\equiv\int_{\mathbb{R}^n}fg\rho$, where $\rho=e^{-|y|^2/4}$.

For Z, we can ignore all odd eigenvalues and corresponding eigenfunctions due to the radial symmetry. Let

$$(2.19)\qquad \begin{aligned}Z_c&\equiv\operatorname{sp}(|y|^2-2n),\\ Z_u&\equiv\operatorname{sp}(1),\\ Z_s&\equiv\overline{Z-(Z_c\oplus Z_u)}.\end{aligned}$$

Because the linearization of (2.14) has neutral as well as stable and unstable modes, a natural tool to employ is the center manifold theory for infinite-dimensional systems [5], [14], [22]. In Z, the Hilbert space of radially symmetric functions in L^2_ρ with inner product $\langle\cdot,\cdot\rangle$, consider the given abstract problem

$$(2.14)\qquad v'=Av+Nv,$$

where $Nv=g(v)-v$ and $g(v)$ is given in (2.13). The operator A is the generator of a strongly continuous semigroup $S(s)$ and has the following spectral properties:

(1) $Z = Z_c \oplus Z_s \oplus Z_u$, where Z_c, Z_u are finite-dimensional, Z_s is closed, and all are defined in (2.19). Associated with this splitting of Z, there exist projections $\pi_s : Z \to Z_s$, $\pi_u : Z \to Z_u$, $\pi_c : Z \to Z_c$ with ker $\pi_s = Z_c \oplus Z_u = Z_{cu}$, ker $\pi_c = Z_s \oplus Z_u = Z_{su}$, and ker $\pi_u = Z_c \oplus Z_s = Z_{cs}$.

(2) Z_c and Z_u are A-invariant. If $A^0 = A|_{Z_c}$ and $A^+ = A|_{Z_u}$, then Re $\sigma_p(A^0) = 0$ and Re $\sigma_p(A^+) > 0$.

(3) If $U(s) = S(s)|_{Z_s}$, then Z_s is $U(s)$-invariant and for some $a, b > 0$,

$$\|U(s)\| \leqq a\, e^{-bs}, \qquad s \geqq 0. \tag{2.20}$$

For $v_c \in Z_c$, $v_s \in Z_s$, $v_u \in Z_u$, let $f(v_c, v_s, v_u) = \pi_c N(v_c + v_s + v_u)$, $g(v_c, v_s, v_u) = \pi_s N(v_c + v_s + v_u)$, $h(v_c, v_s, v_u) = \pi_u N(v_c + v_s + v_u)$, and $A^- = \pi_s A$. Then (2.14) can be written as

$$\begin{aligned} v_c' &= A^0 v_c + f(v_c, v_s, v_u), \\ v_s' &= A^- v_s + g(v_c, v_s, v_u), \\ v_u' &= A^+ v_u + h(v_c, v_s, v_u). \end{aligned} \tag{2.21}$$

If the nonlinear term N maps Z into Z with $N(0) = 0$ and $N'(0) = 0$, where N' is the Fréchet derivative of N so that Nv can be considered as a small perturbation, then the following theorem holds (see [5], [14], [22]).

THEOREM 2.2. (1) *System* (2.21) *has the following invariant manifolds, both of which are tangent at the origin*:

$$M_c = \{(v_c, v_s, v_u) : \|v_c\| < \delta,\ v_s = W^s(v_c),\ v_u = W^u(v_c)\},$$

the center manifold, and

$$M_{cu} = \{(v_c, v_s, v_u) : v_s = W^+(v_c, v_u),\ \|v_c\| + \|v_u\| < \delta\},$$

the center-unstable manifold.

(2) *There exists $\gamma > 0$ and $H_{cu} \in C(Z; M_{cu})$ such that for each $v_0^{cu} \in M_{cu}$ and $\varepsilon > 0$, there exists $\delta > 0$ such that $v_0 \in Z$ with $\|v_0 - v_0^{cu}\| < \delta$, and $\|v(s; v_0)\| < \varepsilon$ for large s implies $\|v(s; v_0) - v(s; H_{cu}(v_0))\| \leqq \varepsilon\, e^{-\gamma s}$, s large.*

Unfortunately, our nonlinear term does not have the required properties in any of the obvious function spaces $L^2_{\rho_0}$ or $H^m_{\rho_0}$ as observed by Filippas and Kohn in [8]. Because of this, we are unable to apply Theorem 2.2 directly. By using the properties of the known trajectory $v(y, s)$, we can avoid this difficulty and still obtain the same conclusions for $v(y, s)$ as would be given by the existence of a center-unstable manifold.

For $Z = Z_s \oplus Z_c \oplus Z_u$, we will use the following notation:

$$\begin{gathered} \pi_* = \text{projection onto } Z_* \quad \text{for } * = s, c, u, \\ P_u v = \langle v, 1 \rangle, \qquad P_c v = \langle v, |y|^2 - 2n \rangle. \end{gathered} \tag{2.22}$$

Let

$$v = v(y, s) = a(s) + b(s)(|y|^2 - 2n) + \theta, \tag{2.23}$$

where $a(s), b(s) \in \mathbb{R}$, $\theta \in Z_s$.

We claim that if v solves (2.14), then it does not decay exponentially fast. To see this, assume v decays exponentially fast, i.e., $\|v\| \leqq M e^{-\alpha s}$ for some $M, \alpha > 0$. Herrero and Velázquez [15] and Liu [21] for dimensions $n \geqq 2$ have proven that either $v \equiv 0$ or there exists an integer $m \geqq 3$, which is even by symmetry, such that

$$v = C\, e^{(1-(m/2))s} h_{m,n}(|y|) + o(e^{(1-(m/2))s}), \tag{2.24}$$

where $h_{m,n}(|y|) = \sum_{i=1}^n h_m(y_i)$ and C is some constant. If $v \equiv 0$, then for $f(u) = e^u$,

$u(x, t) = -\ln(T-t)$ and for $f(u) = u^p$, $u(x, t) = \beta^\beta(T-t)^\beta$. But this contradicts $u_r < 0$. For the other possibility, define $\bar{r} = |y|$; then $v(\bar{r}, s) = C e^{(1-(m/2))s} h_{m,n}(\bar{r}) + o(e^{(1-(m/2))s})$. Since $h_{m,n}$ has $m/2$ maxima, taking R large enough, our solution should have at least one maximum on $0 < \bar{r} < R$ for large enough s, and this is impossible because $v_{\bar{r}}(\bar{r}, s) < 0$.

The following two theorems are essentially due to Filippas and Kohn [8] and Herrero and Velázquez [15].

THEOREM 2.3. *Let $v(y, s)$ be a solution of* (2.14). *Then given any $\varepsilon > 0$, there exists $\hat{s}$ such that*

$$\begin{aligned} \dot{a} &= a + \|1\|^{-2} P_u N(\pi_u v + \pi_c v) + \varepsilon O(2), \\ \dot{b} &= \| |y|^2 - 2n \|^{-2} P_c N(\pi_u v + \pi_c v) + \varepsilon O(2) \end{aligned} \tag{2.25}$$

for $s \geqq \hat{s}$, where $O(2)$ denotes quadratic terms in a and b.

Recall that $w(y, s)$ satisfies (2.3) on $\mathbb{R}^n \times (s_0, \infty)$, is smooth (see [10]), $|w(y, s)| \leqq c(1 + |y|)$, and $|\nabla \omega| \leqq c$ on $\mathbb{R}^n \times (s_0, \infty)$. By considering the parabolic equations satisfied by the first-order spatial derivatives of $w(y, s)$, we can conclude all spatial derivatives of the same order satisfy a uniform bound in (y, s) (see [12]). Since $v = w - S$, the same is true for v. This implies v is in the Sobolev space $H^m_{\rho_0}$ for any $m \geqq 0$, where we use the convention of $H^0_{\rho_0} = L^2_{\rho_0}$. By the Lebesgue dominated convergence theorem, $v \to 0$ as $s \to \infty$ in $H^m_{\rho_0}$.

THEOREM 2.4. *Let $v(y, s)$ be a solution of* (2.14). *Then given $\varepsilon > 0$, there exists $\hat{s}$ such that*

$$\|\pi_s v\|_{H^m_{\rho_0}} \leqq \varepsilon(\|\pi_u v\|_{H^m_{\rho_0}} + \|\pi_c v\|_{H^m_{\rho_0}}) \tag{2.26}$$

for $s \geqq \hat{s}$ and any $m \geqq 0$.

The proofs of the theorems are easier than those found in [8] and [15] because we only require the stable mode to be dominated by the center and unstable components as opposed to the stable and unstable components being dominated by the center mode, which corresponds to a center manifold.

Recalling the definition of $N(\pi_u v + \pi_c v)$, we have that

$$\begin{aligned} &\text{(a)} \quad N(\pi_u v + \pi_c v) = \tfrac{1}{2}(\pi_u v + \pi_c v)^2 \qquad \text{for } f(u) = e^u, \\ &\text{(b)} \quad N(\pi_u v + \pi_c v) = \frac{p}{2\beta^\beta}(\pi_u v + \pi_c v)^2 \quad \text{for } f(u) = u^p. \end{aligned} \tag{2.27}$$

Equation (2.25) corresponds to the reduced equation for (2.14), which would determine the flow on the center-unstable manifold if it existed. Discarding $\varepsilon O(2)$ terms, using the definitions of P_u, P_c, and (2.27), we have that (2.25) reduces to the following theorem.

THEOREM 2.5. *Let $v(y, s)$ be a solution of* (2.14). *Then*

$$\left.\begin{aligned} \dot{a} &= a + \tfrac{1}{2}(a^2 + 8nb^2) \\ \dot{b} &= ab + 4b^2, \end{aligned}\right\} \quad \textit{for } f(u) = e^u, \tag{2.28a}$$

$$\left.\begin{aligned} \dot{a} &= a + \frac{p}{2\beta^\beta}(a^2 + 8nb^2) \\ \dot{b} &= \frac{p}{\beta^\beta}(ab + 4b^2) \end{aligned}\right\} \quad \textit{for } f(u) = u^p, \tag{2.28b}$$

Using Theorem 2.5, we can now obtain more information as to how $v(y, s)$ tends to zero as $s \to \infty$. By Theorem 2.1 and (2.11), we know that $\lim_{s\to\infty} v(y, s) = 0$ uniformly on compacts. This can be immediately extended to $\lim_{s\to 0} v(y, s) = 0$ in Z by observing that $v(y, s) \to 0$ pointwise as $s \to \infty$, $|v(y, s)| \leqq c(1+|y|)$ on $\mathbb{R}^n \times (s_0, \infty)$, and then applying the Lebesgue dominated convergence theorem. By the Pythagorean theorem, $\|v\|^2 \geqq \|\pi_u v\|^2 = \|a(s)\|^2 = a^2(s)\|1\|^2$, which in turn implies $a(s) \to 0$ as $s \to \infty$. Similarly, $b(s) \to 0$ as $s \to \infty$.

We now can prove the following theorem.

THEOREM 2.6. *On compacts in y,*

$$(2.29)\qquad \begin{aligned} &\text{(a)} \qquad v(y, s) \sim -\frac{1}{4s}(|y|^2 - 2n) + o\left(\frac{1}{s}\right) \qquad \textit{for } f(u) = e^u, \\ &\text{(b)} \qquad v(y, s) \sim -\frac{\beta^\beta}{4ps}(|y|^2 - 2n) + o\left(\frac{1}{s}\right) \qquad \textit{for } f(u) = u^p, \end{aligned}$$

uniformly as $s \to \infty$.

Proof. We only prove (2.29(b)) and observe that the proof of (2.29(a)) is similar. We begin by solving (2.28(b)). By a simple phase-plane analysis, we observe that $a(s) \leqq 0$ and $b(s) \leqq 0$ for all $s \geqq s_0$ since a and b decay to zero. If $a(\hat{s}) = 0$ for some $\hat{s}$, then $a(s) > 0$ for $s > \hat{s}$ unless $b(s) \equiv 0$ for $s \geqq \hat{s}$ by (2.28(b_1)). If $b(s)$ is identically zero for $s \geqq \hat{s}$, then from (2.28(b_1)) we have $\dot{a} = a + (p/2\beta^\beta)a^2$. By uniqueness we conclude $a(s) \equiv 0$ for $s \geqq \hat{s}$ since $a(s)$ decays to zero. This implies that $v(y, s) \equiv 0$ for $s \geqq \hat{s}$ by Theorem 2.4, and hence v would decay exponentially fast, which cannot happen. Thus, $a(s) < 0$ for $s \geqq s_0$. If b has a zero, then since $b \equiv 0$ solves (2.28(b_2)), we conclude by uniqueness that b must vanish identically. Thus $b(s) < 0$ for $s \geqq s_0$.

We now show that $a(s)$ does not decay exponentially fast.

If $a(s) \to 0$ exponentially fast, then $b(s) \to 0$ exponentially fast. To prove this, we assume there exists $\alpha > 0$ such that $\lim_{s\to\infty} e^{\alpha s}|a(s)| = 0$. We can express $a(s)$ as

$$a(s) = e^s\left[a(s_0)\, e^{-s_0} + \frac{p}{2\beta^\beta}\int_{s_0}^{s} e^{-\xi}(a^2(\xi) + 8nb^2(\xi))\, d\xi\right].$$

Since $a(s) \to 0$ and in fact is exponentially fast by assumption,

$$\lim_{s\to\infty}\left[a(s_0)\, e^{-s_0} + \frac{p}{2\beta^\beta}\int_{s_0}^{s} e^{-\xi}(a^2(\xi) + 8nb^2(\xi))\, d\xi\right] = 0,$$

and

$$0 = \lim_{s\to\infty} e^{\alpha s}a(s) = -\frac{4np}{\beta^\beta(1+\alpha)}\lim_{s\to\infty} e^{\alpha s}b^2(s)$$

by L'Hôpital's rule. This implies $b(s) = o(e^{-(\alpha/2)s})$. Equation (2.26) then implies $\|v\| \to 0$ exponentially fast, which is not the case. Therefore, $a(s)$ cannot decay exponentially.

Define $m(s) = b^2(s)/a(s)$. Then $m(s) < 0$ for all $s \geqq s_0$ and satisfies

$$(2.30)\qquad \dot{m} = -\frac{4np}{\beta^\beta}\, m\left(\frac{\beta^\beta}{4np} - \frac{3}{8n}a(s) - \frac{2}{n}b(s) + m\right).$$

Choose $\hat{s} > 16n^2p/\beta^\beta$ such that $|(3/8n)a(s) + (2/n)b(s)| < \beta^\beta/16n^2p$ for all $s \geqq \hat{s}$. For

$s \geqq \hat{s}$ define

$$G(s) = \sup_{\xi \geqq s} \left[\frac{3}{8n} a(\xi) + \frac{2}{n} b(\xi) \right] - \frac{\beta^\beta}{4np} + \frac{1}{s}, \tag{2.31}$$
$$F(s) = \inf_{\xi \geqq s} \left[\frac{3}{8n} a(\xi) + \frac{2}{n} b(\xi) \right] - \frac{\beta^\beta}{4np} - \frac{1}{s};$$

then $F(s) < (3/8n)a(s) + (2/n)b(s) - \beta^\beta/4np < G(s) < 0$ with F increasing, G decreasing, and both continuous for $s \geqq \hat{s}$.

We now observe that $m(s)$ has a limit at infinity. Suppose $m(\hat{s}) \geqq G(\hat{s})$, then $m(s)$ is increasing and bounded above so it has a finite limit. If $m(\hat{s}) \leqq F(\hat{s})$, then $m(s)$ is decreasing and so it has a limit, which could possibly be $-\infty$. Suppose $F(\hat{s}) < m(\hat{s}) < G(\hat{s})$. Then, either $m(s)$ remains in the funnel $(F(s), G(s))$ for all $s \geqq \hat{s}$, in which case $\lim_{s\to\infty} m(s) = -(\beta^\beta/4np)$, or $m(s)$ intersects either $F(s)$ or $G(s)$. If $m(s)$ intersects $F(s)$ or $G(s)$ at some $s^* > \hat{s}$, then $m(s)$ is either decreasing or increasing, respectively, for $s \geqq s^*$ and so it will have a limit at infinity.

Suppose $\lim_{s\to\infty} m(s) = -\infty$. Then by (2.28($\mathrm{b}_1$)), $\lim_{s\to\infty} (\dot{a}(s)/a(s)) = -\infty$. This implies that $a(s)$ decays exponentially fast, which cannot happen. Hence, $\lim_{s\to\infty} m(s)$ exists and is finite.

If $\lim_{s\to\infty} m(s) = 0$, then by (2.28($\mathrm{b}_1$)), there exists $\hat{s}$ such that $(\dot{a}(s)/a(s)) \geqq \frac{1}{2}$ for $s \geqq \hat{s}$. This implies $a(s) \leqq a(\hat{s})\, e^{1/2(s-\hat{s})}$, which is a contradiction. Therefore, $\lim_{s\to\infty} m(s) < 0$ and $b(s) = O(|a(s)|^{1/2})$ with $b \neq o(|a|^{1/2})$.

Since $\lim_{s\to\infty} (a(s)/b(s)) = 0$, there exists $\hat{s}$ such that $b(s) < a(s) < 0$ for $s \geqq \hat{s}$. By (2.28(b_2)), we have that

$$\frac{4p}{\beta^\beta} b^2(s) \leqq \dot{b}(s) \leqq \frac{5p}{\beta^\beta} b^2(s) \tag{2.32}$$

for $s \geqq \hat{s}$, which implies $b(s) = O(s^{-1})$, and so $a(s) = O(s^{-2})$. Up to order $O(s^{-3})$, (2.28(b_2)) becomes $\dot{b} = (4p/\beta^\beta)b^2$, which implies $b(s) = -(\beta^\beta/4ps) + o(s^{-1})$.

By Theorem 2.4,

$$\left\| v(y,s) + \frac{\beta^\beta}{4ps}(|y|^2 - 2n) \right\|_{H^m_{\rho_0}} \leqq \left\| \pi_u v + \pi_c v + \frac{\beta^\beta}{4ps}(|y|^2 - 2n) \right\|_{H^m_{\rho_0}} + \|\pi_s v\|_{H^m_{\rho_0}}$$
$$\leqq o(s^{-1}) + \varepsilon(\|\pi_u v\|_{H^m_{\rho_0}} + \|\pi_c v\|_{H^m_{\rho_0}}) \tag{2.33}$$
$$\leqq o(s^{-1}).$$

Therefore,

$$v(y,s) \sim -\frac{\beta^\beta}{4ps}(|y|^2 - 2n) + o(s^{-1}) \quad \text{in } H^m_{\rho_0}. \tag{2.34}$$

This can be extended to uniform convergence on compacts in y. Given the spatial dimension n, choose $m > n/2$. Then the Sobolev imbedding theorem (see [1]) implies that (2.34) holds in $C^0_{b\rho_0} = \{u \in C^0\colon u \text{ is symmetric}, \sup[\mathrm{e}^{(-|y|^2/4)}|u(y)|] < \infty\}$. On any compact set $|y| \leqq c$:

$$\sup_{|y|\leqq c} e^{-(c^2/4)} \left| v(y,s) + \frac{\beta^\beta}{4ps}(|y|^2 - 2n) \right| \leqq \sup_{|y|\leqq c} e^{-(|y|^2/4)} \left| v(y,s) + \frac{\beta^\beta}{4ps}(|y|^2 - 2n) \right|$$
$$\leqq \sup_{\mathbb{R}^n} e^{-(|y|^2/4)} \left| v(y,s) + \frac{\beta^\beta}{4ps}(|y|^2 - 2n) \right|$$
$$= o(s^{-1}),$$

which proves (2.29(b)). □

In the original variables, we can restate Theorem 2.6 as the following corollary, which is Theorem 1.2.

COROLLARY 2.7. *Let* $u(x, t)$ *be the solution of* (1.1)-(1.2), *then*

$$
\begin{aligned}
&\textit{for } f(u)=e^u,\\
&\text{(a)}\quad u(x,t)\sim \ln\left(\frac{1}{T-t}\right)+\frac{1}{4\ln(T-t)}\left(\frac{|x|^2}{T-t}-2n\right)+o\left(\frac{1}{\ln(T-t)}\right),\\
\text{(2.35)}\quad&\textit{for } f(u)=u^p,\\
&\text{(b)}\quad u(x,t)\sim\frac{1}{(T-t)^\beta}\left[\beta^\beta+\frac{\beta^\beta}{4p\ln(T-t)}\left(\frac{|x|^2}{T-t}-2n\right)+o\left(\frac{1}{\ln(T-t)}\right)\right]
\end{aligned}
$$

uniformly on $|x|\leqq C(T-t)^{1/2}$, $C\geqq 0$ *as* $t\to T^-$.

3. Extension to the ignition variable domain. We now indicate how to extend (2.29) from the hot-spot variable grouping to the ignition variable grouping first suggested by Dold [6] and independently by Galaktionov [11]. By this, we mean that we can get a spatial description of how the blowup singularity evolves not only in parabolic domains $|x|\leqq C(T-t)^{1/2}$ with vertex at $(0, T)\in\mathbb{R}^n\times(0,\infty)$, but in the larger domains $|x|\leqq C((T-t)|\ln(T-t)|)^{1/2}$, $C>0$ arbitrary.

Our procedure for doing this in the multidimensional case is due to Herrero and Velázquez ([15], § 6). We summarize their method for the case $f(u)=e^u$, along with the extension to higher dimensions. The case $f(u)=u^p$ is similar.

Let $u(x, t)$ be a solution of (1.1)-(1.2) for ϕ satisfying (1.4) and $f(u)=e^u$, which blows up at $(0, T)$. Make the change of variables:

$$
\text{(3.1)}\qquad s=-\ln(T-t),\quad y=\frac{x}{(T-t)^{1/2}},\quad \eta=\frac{x}{((T-t)|\ln(T-t)|)^{1/2}},
$$

and as before let

$$
\text{(3.2)}\qquad v(y,s)=u(x,t)+\ln(T-t).
$$

In § 2, we proved that

$$
\text{(3.3)}\qquad v(y,s)=-\frac{1}{4s}(|y|^2-2n)+o\left(\frac{1}{s}\right)
$$

in $C^0_{b\rho_0}$ or uniformly on compacts $|y|\leqq C$ as $s\to\infty$.

We now show that (3.3) actually holds on the ignition variable domain suggested by Dold.

Without loss of generality, we may assume $T=1$.

LEMMA 3.1. *For a solution* $u(x, t)$ *of* (1.1)-(1.2) *that blows up at* $T=1$,

$$
\text{(3.4)}\qquad \ln(1-t)+u(\eta((1-t)|\ln(1-t)|)^{1/2},t)\geqq-\ln\left(1+\frac{|\eta|^2}{4}\right)+o(1)
$$

uniformly for $|\eta|\leqq R$, $R>0$, *as* $t\to 1^-$.

The proof of Lemma 3.1 for dimension $n=1$ given in Herrero and Velázquez [15, Lemma 6.1] relies on expansions in terms of the Hermite polynomials. In the following proof, we are able to avoid the difficulties in extending their method to higher dimensions by making use of the semigroup associated with the heat equation in $\mathbb{R}^n$.

Proof. Let $v(y, s)$ be given by (3.2) with $T=1$; then v satisfies

$$v_s=\Delta v-\frac{y}{2}\cdot\nabla v+e^v-1$$

$$=Av+e^v-v-1.$$

From the facts of § 1, we can immediately infer that v is bounded above, $v(y, s)\to 0$ as $s\to\infty$ uniformly on compacts, and $|\nabla v(\cdot, s)|$ is bounded. Therefore, $|v(y, s)|\leqq C(1+|y|)$, and hence $\|v(\cdot, s)\|\to 0$ as $s\to\infty$.

Let α be a parameter, $0<\alpha<1$, and consider

$$\psi_\alpha(x, t)=\ln(1-\alpha)+u(x\sqrt{(1-\alpha)}, \alpha+t(1-\alpha)). \tag{3.5}$$

For any α fixed, ψ_α solves (1.1). Moreover,

$$\psi_\alpha(x, 0)=\ln(1-\alpha)+u(x\sqrt{1-\alpha}, \alpha)=\frac{1}{4\ln(1-\alpha)}(|x|^2-2n)+o\left(\frac{1}{\ln(1-\alpha)}\right)$$

as $\alpha\to 1^-$ uniformly for $|x|\leqq R$, $R>0$, by (3.3).

Now consider the function

$$F_\alpha(x, t)=-\ln(e^{-S(t)\psi_\alpha(x,0)}-t), \tag{3.6}$$

where $S(t)$ is the linear semigroup corresponding to the heat equation in the strip $\mathbb{R}^n\times[0, 1)$. It is obvious that $F_\alpha(x, 0)=\psi_\alpha(x, 0)$, and we can easily verify that

$$(F_\alpha)_t-\Delta F_\alpha-e^{F_\alpha}\leqq 0.$$

We immediately conclude $\psi_\alpha(x, t)\geqq F_\alpha(x, t)$ for fixed α whenever $x\in\mathbb{R}^n$, $0<t<1$.

Set $s_\alpha=-\ln(1-\alpha)$; then

$$\left\|S(t)\psi_\alpha(x, 0)-S(t)\left(-\frac{1}{4s_\alpha}(|x|^2-2n)\right)\right\|_{C^0_{b\rho_0}}\leqq\|S(t)\|\left\|\psi_\alpha(x, 0)+\frac{1}{4s_\alpha}(|x|^2-2n)\right\|_{C^0_{b\rho_0}}$$

$$=o\left(\frac{1}{s_\alpha}\right)$$

as $\alpha\to 1^-$. Thus

$$S(t)\psi_\alpha(x, 0)=\frac{1}{4\ln(1-\alpha)}(1-t)\left(\frac{|x|^2}{1-t}-2n\right)+o\left(\frac{1}{\ln(1-\alpha)}\right) \tag{3.7}$$

as $\alpha\to 1^-$ uniformly for $|x|\leqq R$, $R>0$.

We can write $\psi_\alpha(x, t)=\ln(1-\alpha)+u(x, \hat{t})$, where $r=x\sqrt{(1-\alpha)}$, $\hat{t}=\alpha+t(1-\alpha)$. Note that $1-\hat{t}=(1-\alpha)(1-t)$. To get a lower bound on $u(r, \hat{t})$ along sets where $r=\eta(1-\hat{t})^{1/2}|\ln(1-\hat{t})|^{1/2}$ in terms of x and t, i.e., $x=\eta(1-t)^{1/2}|\ln(1-t)(1-\alpha)|^{1/2}$, we first select $t=t(\alpha)$ by

$$1=(1-t)|\ln(1-\alpha)(1-t)| \tag{3.8}$$

so that $(1-t)\approx|\ln(1-\alpha)|^{-1}$ as $\alpha\to 1^-$. Then by (3.7) and (3.8)

$$S(t)\psi_\alpha(x, 0)=-(1-t)\frac{|\eta|^2}{4}+o(1-t)\quad\text{as } t\to 1^-,$$

uniformly on sets $|\eta|\leqq R$, $R>0$.

Using (3.6), $\psi_\alpha(x,t) \geqq F_\alpha(x,t) = -\ln(1-t) - \ln(1+(|\eta|^2/4)) + o(1)$, and thus

$$\psi_\alpha(x,t) = \ln(1-\alpha) + u(\eta(1-\hat{t})^{1/2}|\ln(1-\hat{t})|^{1/2}, \hat{t})$$
$$\geqq -\ln(1-t) - \ln\left(1+\frac{|\eta|^2}{4}\right) + o(1).$$

Therefore,

$$\ln(1-\hat{t}) + u(\eta(1-\hat{t})^{1/2}|\ln(1-\hat{t})|^{1/2}, \hat{t}) \geqq -\ln\left(1+\frac{|\eta|^2}{4}\right) + o(1)$$

as $\hat{t} \to 1^-$ uniformly on sets $|\eta| \leqq R$, $R > 0$. This proves (3.4). □

We now turn to the task of showing we have equality in (3.4).

LEMMA 3.2. *For any $R>0$ there exists $C>0$ such that*

$$|\nabla_y v(\eta\sqrt{s}, s)| \leqq \frac{C}{\sqrt{s}} \tag{3.9}$$

uniformly for $|\eta| \leqq R$ and large enough $s>0$.

The proof given in [15, Lemma 6.2] extends immediately to higher dimensions and so we do not include a proof of this lemma.

Now set

$$J = e^{-v},$$

which leads to the equation

$$J_s = \Delta J - \frac{y}{2}\cdot\nabla J + J - \frac{|\nabla J|^2}{J} - 1.$$

Since $v \to 0$ as $s \to \infty$, $J \to 1$ as $s \to \infty$.

LEMMA 3.3. *There exists $C>0$ such that*

$$\|J(\cdot, s) - 1\| \leqq \frac{C}{s} \quad \text{as } s \to \infty. \tag{3.10}$$

The proof given in [15, Lemma 6.3] uses the fact that the solution $u(x,t)$ of (1.1) is a supercaloric function. We avoid using this observation in the following proof, which allows for an immediate extension to higher dimensions.

Proof. By definition we have that

$$\|J(\cdot,s)-1\|^2 = \int_{\mathbb{R}^n} |e^{-v}-1|^2 \, e^{-(|y|^2/4)}\, dy.$$

Since $|e^{-v}-1|^2 \leqq e^{-2v} + 2e^{-v} + 1$ and $|v| \leqq C(1+|y|)$ we have

$$|e^{-v}-1|^2 \leqq k\, e^{k(1+|y|)} \in L^1_p \quad \text{for all } s \text{ and some } k>0.$$

Also, $|e^{-v}-1|^2 = O(1/s^2)$ as $s \to \infty$ pointwise in y by (3.3). The Lebesgue dominated convergence theorem gives the desired result. □

Let us write

$$G = J - 1;$$

then G solves

$$G_s = \Delta G - \frac{y}{2}\cdot\nabla G + G - \frac{|\nabla G|^2}{1+G}.$$

We set

$$L(y, s)=\frac{|\nabla G|^2}{1+G};$$

then for $s \geqq s_0$, $G(y, s)$ can be written in the form

$$\begin{aligned} G(y, s)&=\frac{e^{s-s_0}}{(4\pi(1-e^{-(s-s_0)}))^{n/2}}\int_{\mathbb{R}^n}\exp\left(-\frac{|y\,e^{-(s-s_0)/2}-\lambda|^2}{4(1-e^{-(s-s_0)})}\right)G(\lambda, s_0)\,d\lambda \\ &\quad -\int_{s_0}^{s}\int_{\mathbb{R}^n}\frac{e^{(s-\beta)}}{(4\pi(1-e^{-(s-\beta)}))^{n/2}}\exp\left(-\frac{|y\,e^{-(s-\beta)/2}-\lambda|^2}{4(1-e^{-(s-\beta)})}\right)L(\lambda, \beta)\,d\lambda\,d\beta \qquad (3.11)\\ &\equiv I_1(y, s)+I_2(y, s). \end{aligned}$$

We now indicate the estimates on I_1 and I_2 as $s\to\infty$.

LEMMA 3.4. *Let $I_1(y, s)$ be as in* (3.11); *then*

$$\lim_{s\to\infty} I_1(\eta\sqrt{s}, s)=\frac{|\eta|^2}{4} \tag{3.12}$$

uniformly on set $|\eta|\leqq C$ with $C>0$.

LEMMA 3.5. *Let $I_2(y, s)$ be as in* (3.11), *then*

$$\lim_{s\to\infty} I_2(\eta\sqrt{2}, s)=0 \tag{3.13}$$

uniformly on sets $|\eta|\leqq C$ with $C>0$.

The proofs given in [15, Lemmas 6.4, 6.5] can easily be extended to cover the multidimensional case and so we do not include the proofs of the above two lemmas.

THEOREM 3.6.

$$\lim_{t\to 1^-}[\ln(1-t)+u(\eta\sqrt{(1-t)|\ln(1-t)|}, t)]=-\ln\left(1+\frac{|\eta|^2}{4}\right) \tag{3.14}$$

uniformly on compact sets $|\eta|\leqq K$, $K>0$.

Proof. By Lemmas 3.4 and 3.5, we have that

$$\lim_{s\to\infty} G(\eta\sqrt{s}, s)=\frac{|\eta|^2}{4} \tag{3.15}$$

uniformly on compact sets $|\eta|\leqq K$. Therefore,

$$\lim_{s\to\infty} v(\eta\sqrt{s}, s)=-\ln\left(1+\frac{|\eta|^2}{4}\right)$$

uniformly on compacts. Since $v(x/\sqrt{1-t}, s)=\ln(1-t)+u(x, t)$, if

$$x=\eta\sqrt{(1-t)|\ln(1-t)|}=\eta\sqrt{1-t}\,\sqrt{s},$$

we then have the desired result. □

This is precisely Theorem 1.3 for the case $f(u)=e^u$ with the blowup time normalized to one. For the case $f(u)=u^p$, the proof proceeds in an analogous manner.

4. Final time solution profiles. We now prove Theorem 1.4, which gives us a precise description of the final time solution profile $u_F(x)$ in a neighborhood of the singularity $x=0$. Our proof follows that of Herrero and Velázquez [16].

Let $u(x, t)$ be a solution of (1.1)-(1.3) that blows up at $(0, T)\in\mathbb{R}^n\times(0,\infty)$, where ϕ satisfies (1.14). Let $\eta\neq 0$ and consider the auxiliary functions

$$\text{(4.1)}\quad \begin{aligned} &\text{(a)}\quad \psi_\alpha(x,t)=\ln(T-\alpha)+u(\lambda(\alpha)+x\sqrt{T-\alpha},\ \alpha+(T-\alpha)t) &&\text{for } f(u)=e^u,\\ &\text{(b)}\quad \psi_\alpha(x,t)=(T-\alpha)^\beta u(\lambda(\alpha)+x\sqrt{T-\alpha},\ \alpha+(T-\alpha)t) &&\text{for } f(u)=u^p,\end{aligned}$$

where

$$\text{(4.2)}\qquad \lambda(\alpha)=\sqrt{T-\alpha}\,|\ln(T-\alpha)|^{1/2}\eta,\qquad 0<\alpha<T,$$

and

$$\text{(4.3)}\qquad |x|\leqq\frac{|\eta|}{2}|\ln(T-\alpha)|^{1/2},\qquad 0<t<1.$$

We first prove the following lemma, which gives a uniform bound on the auxiliary functions (4.1) as $\alpha\to T^-$.

Lemma 4.1. *Let $m\in\mathbb{N}$ be fixed and α sufficiently close to T. Then there exists a constant M_m such that*

$$\text{(4.4)}\qquad |\psi_\alpha(x,t)|\leqq M_m$$

for $|x|\leqq m/2$, $t\in[0,1]$ uniformly as $\alpha\to T^-$.

Proof. We point out that (4.3) implies $|\lambda(\alpha)+x\sqrt{T-\alpha}|\geqq|\lambda(\alpha)|/2$. For $f(u)=u^p$, since $u_r\leqq 0$ and $u_t\geqq 0$, (4.1(b)) gives

$$\psi_\alpha(x,t)\leqq(T-\alpha)^\beta u\left(\frac{\lambda(\alpha)}{2},\ \alpha+(T-\alpha)t\right)\leqq(T-\alpha)^\beta u\left(\frac{\lambda(\alpha)}{2},\ T\right).$$

Using the upper bound (1.12(b)) and (4.2) we have that

$$\begin{aligned}\psi_\alpha(x,t)&\leqq(T-\alpha)^\beta\left(\frac{K|\ln(|\lambda(\alpha)|/2)|}{(|\lambda(\alpha)|/2)^2}\right)^\beta\\ &=\left(\frac{4K}{|\eta^2|}\left|\frac{1}{2}+\frac{1}{\ln(T-\alpha)}\left(\frac{1}{2}\ln|\ln(T-\alpha)|+\ln|\eta|-\ln 2\right)\right|\right)^\beta\leqq C\end{aligned}$$

for all $|x|\leqq|\eta|/2|\ln(T-\alpha)|^{1/2}$, $t\in[0,1]$ uniformly as $\alpha\to T^-$. Since $\psi_\alpha(x,t)\geqq 0$, the result follows.

For $f(u)=e^u$, the proof of (4.4) is more difficult since $\psi_\alpha(x,t)$ could approach $-\infty$. Since $u_r\leqq 0$ and $u_t\geqq 0$, (4.1(a)) gives

$$\psi_\alpha(x,t)\leqq\ln(T-\alpha)+u\left(\frac{\lambda(\alpha)}{2},\ T\right).$$

Using the upper bound (1.12(a)) and (4.2) we have that

$$\begin{aligned}\psi_\alpha(r,t)&\leqq\ln(T-\alpha)-2\ln\left|\frac{\lambda(\alpha)}{2}\right|+\ln\left|\ln\left|\frac{\lambda(\alpha)}{2}\right|\right|+C\\ &=\ln\left(\frac{4(t-\alpha)|\ln|\lambda(\alpha)/2||}{|\lambda(\alpha)|^2}\right)+C\\ &=\ln\left[\frac{4}{|\eta|^2}\left|\frac{1}{2}+\frac{1}{\ln(T-\alpha)}\left(\frac{1}{2}\ln|\ln(T-\alpha)|+\ln|\eta|-\ln 2\right)\right|\right]+C,\end{aligned}$$

which implies

$$\psi_\alpha(x, t) \leqq K \tag{4.5}$$

for all $|x| \leqq |\eta|/2|\ln(T-\alpha)|^{1/2}$, $t \in [0, 1]$ uniformly as $\alpha \to T^-$.

By Friedman and McLeod [9],

$$|\nabla \psi_\alpha|^2 = (T-\alpha)|\nabla u|^2 \leqq K \tag{4.6}$$

for all $|x| \leqq |\eta|/2|\ln(T-\alpha)|^{1/2}$, $t \in [0, 1]$ uniformly as $\alpha \to T^-$.

Since $u_t \geqq 0$, we can use Lemma 3.1 to obtain a lower bound for $\psi_\alpha(0, t)$.

$$\psi_\alpha(0, t) \geqq \ln(T-\alpha) + u(\lambda(\alpha), \alpha) \geqq -\ln\left(1 + \frac{|\eta|^2}{4}\right) + o(1)$$

as $\alpha \to T^-$. Therefore, for α sufficiently close to T, (4.5) gives

$$|\psi_\alpha(0, t)| \leqq K \tag{4.7}$$

for all $t \in [0, 1]$, uniformly as $\alpha \to T^-$.

Now let $m \in \mathbb{N}$ be fixed and α sufficiently close to T. For $|x| \leqq m/2$ and $t \in [0, 1]$ define $\theta_\alpha(\nu) = \psi_\alpha((1-\nu)x, t)$, $0 \leqq \nu \leqq 1$. Then

$$\psi_\alpha(x, t) - \psi_\alpha(0, t) = \theta_\alpha(1) - \theta_\alpha(0) = \int_0^1 \frac{d}{d\nu}\,\theta_\alpha(\nu)\,d\nu = -\int_0^1 \nabla\psi_\alpha((1-\nu)x, t) \cdot x\,d\nu.$$

By (4.6) and (4.7),

$$|\psi_\alpha(x, t)| \leqq K(1 + |x|) \leqq M_m$$

for some constant M_m uniformly as $\alpha \to T^-$. □

Remark. The proof for $f(u) = u^p$ given by Herrero and Velázquez [16], when extended to higher dimensions, establishes (4.4) provided $n \leqq 2$ or $n \geqq 3$ and $1 < p \leqq (n+2)/(n-2)$. This restriction is coming from an application of the results of Giga and Kohn [12] to the nonradially symmetric function $\psi_\alpha(x, t)$.

We can easily check

$$(\psi_\alpha)_t = \Delta\psi_\alpha + f(\psi_\alpha), \qquad x \in \mathbb{R}^n, \quad 0 < t < 1, \tag{4.8}$$

and, by (1.10),

$$\text{(4.9)}\quad \begin{aligned} &\text{(a)} \quad \psi_\alpha(x, 0) = -\ln\left(1 + \frac{1}{4}\left|\eta + \frac{x}{\sqrt{|\ln(T-\alpha)|}}\right|^2\right) + o(1) \qquad \text{for } f(u) = e^u, \\ &\text{(b)} \quad \psi_\alpha(x, 0) = \beta^\beta\left[1 + \frac{1}{4p\beta}\left|\eta + \frac{x}{\sqrt{|\ln(T-\alpha)|}}\right|^2\right]^{-\beta} + o(1) \quad \text{for } f(u) = u^p \end{aligned}$$

as $\alpha \to T^-$ uniformly in x for $\eta \neq 0$ fixed.

By Schauder's interior estimates all partial derivatives of ψ_α remain bounded in the set $Q_m = \{(x, t) : |x| \leqq m/3, \frac{1}{2} \leqq t \leqq 1\}$ uniformly as $\alpha \to T^-$. It follows that there exists a subsequence, also denoted by $\psi_\alpha(x, t)$, and a function $\bar{\psi}_m(x, t)$ such that

$$\psi_\alpha(x, t) \to \bar{\psi}_m(x, t) \quad \text{as } \alpha \to T^- \quad \text{uniformly on } Q_m,$$

$$(\bar{\psi}_m)_t = \Delta\bar{\psi}_m + f(\bar{\psi}_m) \quad \text{on int}\,(Q_m).$$

Repeating the construction for all m and taking a diagonal subsequence, we can conclude there exists a subsequence, again labeled $\psi_\alpha(x, t)$, and a function $\bar{\psi}(x, t)$ such that

$$\psi_\alpha(x, t) \to \bar{\psi}(x, t) \tag{4.10}$$

uniformly as $\alpha \to T^-$ on compacts of $\mathbb{R}^n \times [\frac{1}{2}, 1]$,

$$\bar{\psi}_t = \Delta\bar{\psi} + f(\bar{\psi}) \quad \text{in } \mathbb{R}^n \times (\tfrac{1}{2}, 1), \tag{4.11}$$

$$\begin{aligned} &\text{(a)} \quad \bar{\psi}(x, 0) = -\ln(1 + \tfrac{1}{4}|\eta|^2) &&\text{for } f(u) = e^u, \\ &\text{(b)} \quad \bar{\psi}(x, 0) = \beta^\beta\left[1 + \frac{1}{4p\beta}|\eta|^2\right]^{-\beta} &&\text{for } f(u) = u^p. \end{aligned} \tag{4.12}$$

It is easy to verify that $\bar{\psi}$ is given by

$$\begin{aligned} &\text{(a)} \quad \bar{\psi}(x, t) = -\ln\left((1-t) + \frac{1}{4}|\eta|^2\right) &&\text{for } f(u) = e^u \\ &\text{(b)} \quad \bar{\psi}(x, t) = \beta^\beta\left[(1-t) + \frac{1}{4p\beta}|\eta|^2\right]^{-\beta} &&\text{for } f(u) = u^p. \end{aligned} \tag{4.13}$$

From (4.10) and (4.13) we deduce that

$$\begin{aligned} &\text{(a)} \quad \psi_\alpha(0, 1) = -\ln\left(\frac{1}{4}|\eta|^2\right) + o(1) &&\text{for } f(u) = e^u, \\ &\text{(b)} \quad \psi_\alpha(0, 1) = (4p\beta^2)^\beta |\eta|^{-2\beta} + o(1) &&\text{for } f(u) = u^p, \end{aligned} \tag{4.14}$$

or recalling (4.1),

$$\begin{aligned} &\text{for } f(u) = e^u, \\ &\text{(a)} \quad \ln(T-\alpha) + u(\eta\sqrt{T-\alpha}\,|\ln(T-\alpha)|^{1/2}, T) = -\ln(\tfrac{1}{4}|\eta|^2) + o(1), \\ &\text{for } f(u) = u^p, \\ &\text{(b)} \quad (T-\alpha)^\beta u(\eta\sqrt{T-\alpha}\,|\ln(T-\alpha)|^{1/2}, T) = (4p\beta^2)^\beta|\eta|^{-2\beta} + o(1) \end{aligned} \tag{4.15}$$

as $\alpha \to T^-$.

Now set

$$y = \eta\sqrt{T-\alpha}\,|\ln(T-\alpha)|^{1/2}$$

so that

$$|\ln|y|| = \tfrac{1}{2}|\ln(T-\alpha)| + O(\ln|\ln(T-\alpha)|) \quad \text{as } \alpha \to T^-,$$

and

$$y = \eta\,\sqrt{2}\,\sqrt{T-\alpha}\,\sqrt{|\ln|y||} + O(\sqrt{T-\alpha}\,\ln|\ln|y||) \quad \text{as } \alpha \to T^-.$$

Therefore,

$$(T-\alpha) \approx \frac{|y|^2}{2|\eta|^2|\ln|y||} \quad \text{as } \alpha \to T^-.$$

Substituting this into (4.15) gives

$$u(y, T) \approx -2\ln|y| + \ln|\ln|y|| + \ln 8 \quad \text{for } f(u) = e^u,$$

$$\left(\frac{|y|^2}{2|\ln|y||}\right)^\beta u(y, T) \approx (4p\beta^2)^\beta \quad \text{for } f(u) = u^p$$

as $|y| \to 0$. Theorem 1.4 now follows immediately.

Acknowledgment. The authors express their appreciation to the referee for the many helpful comments, and, in particular, for pointing out a cleaner proof of Theorem 2.6.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Appl. Math. Sci., 83, Springer-Verlag, Berlin, 1989.

[3] M. Berger and R. Kohn, *A rescaling algorithm for the numerical calculation of blowing up solutions*, Comm. Pure Appl. Math., 41 (1988), pp. 841-863.

[4] A. Bressan, *Stable blowup patterns*, preprint.

[5] J. Carr, *Applications of Centre Manifold Theory*, Appl. Math. Sci., 35, Springer-Verlag, Berlin, 1981.

[6] J. W. Dold, *Analysis of the early stage of thermal runaway*, Quart. J. Mech. Appl. Math., 38 (1985), pp. 361-387.

[7] ———, *Analysis of thermal runaway in the ignition process*, SIAM J. Appl. Math., 49 (1989), pp. 459-480.

[8] S. Filippas and R. V. Kohn, *Refined asymptotics for the blowup of $u_t - \Delta u = u^p$*, preprint.

[9] A. Friedman and B. McLeod, *Blowup of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425-447.

[10] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[11] V. Galaktionov and S. A. Posashkov, *The equation $u_t = u_{xx} + u^\beta$. Localization and asymptotic behavior of unbounded solutions*, Differentsial'nye Uravneniya, 22 (1986), pp. 1165-1173.

[12] Y. Giga and R. Kohn, *Asymptotically self-similar blowup of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297-319.

[13] ———, *Characterizing blowup using similarity variables*, Indiana Univ. Math. J., 36 (1987), pp. 1-40.

[14] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1981.

[15] M. Herrero and J. J. L. Velázquez, *Blow-up behavior of one-dimensional semilinear parabolic equations*, preprint.

[16] ———, *Blow-up profiles in one-dimensional semilinear problems*, preprint.

[17] D. Kassoy and J. Poland, *The thermal explosion confined by a constant temperature boundary*: I. *The induction period solution*, SIAM J. Appl. Math., 39 (1980), pp. 412-430.

[18] ———, *The thermal explosion confined by a constant temperature boundary*: II. *The extremely rapid transient*, SIAM J. Appl. Math., 41 (1981), pp. 231-246.

[19] A. K. Kapila, *Reactive-diffusive system with Arrhenius kinetics: Dynamics of ignition*, SIAM J. Appl. Math., 39 (1980), pp. 21-36.

[20] W. Liu, *The blow-up rate of solutions of semilinear heat equations*, J. Differential Equations, 77 (1989), pp. 104-122.

[21] ———, *Blow-up behavior for semilinear heat equations: multi-dimensional case*, IMA Preprint series #711, 1900.

[22] A. Vanderbauwhede, *Centre manifolds, normal forms and elementary bifurcations*, Dynam. Report., 2 (1989), pp. 89-169.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 870–888, July 1992

004

EXISTENCE OF TRAVELLING WAVE SOLUTIONS FOR A BISTABLE EVOLUTIONARY ECOLOGY MODEL*

JACK D. DOCKERY† AND ROGER LUI‡

Abstract. The existence of travelling wave solutions for a density-dependent selection migration model in population genetics is proven. A single locus and two alleles are assumed. It is also assumed that the fitnesses of the heterozygotes in the population are below those of the homozygotes. The method of proof is by constructing an isolating neighborhood and computing a connection index.

Key words. population genetics, travelling waves, wave speed, connection index, isolated invariant set, homotopy

AMS(MOS) subject classifications. 35K57, 92A10

1. Introduction. During the past two decades, a considerable amount of mathematics has been done on the following nonlinear diffusion equation,

$$u_t = u_{xx} + h(u) \tag{1}$$

where $h \in C^1[0,1], h(0) = 0$ and $h(1) = 0$ [1], [10]. This equation is popular because it has numerous applications, one of which is to describe the dynamics of a certain gene frequency in a population subject to selection pressure and random migration [11]. In such an application, many simplifying assumptions had to be made in order that the situation can be modeled by (1). Some of these assumptions are more serious than others, but the most restrictive is probably the assumption that the population density remains constant throughout space and time. We would like to develop and analyze a model that does not have this requirement, and to compare our results to those of (1). We shall develop such a model in this section. The rest of the paper is devoted to proving the existence of travelling wave solutions for an important case of the model. A complete discussion of selection-migration models and the mathematics of (1) may be found in [9].

Consider a population of diploid individuals living in a one-dimensional homogeneous habitat which we assume to be the entire real line. Suppose a particular pair of chromosomes carries at one of its loci a particular gene that occurs in two forms, called alleles, which we denote by A and a. Then the population may be divided into three classes or genotypes: AA, aa, and Aa. Individuals with the first two genotypes are called homozygotes while individuals with the last genotype are called heterozygotes.

Let $\rho_1(x,t), \rho_2(x,t), \rho_3(x,t)$ be the densities of genotypes AA, Aa, and aa at point x and time t, respectively. We assume that the population mates randomly without regard to genotype, produces offspring at the rate r and that the population diffuses with a constant rate 1. Let τ_1, τ_2, τ_3 denote the death-rates of the individuals with genotypes AA, Aa, and aa, respectively, and let $n(x,t)$ denote the total population

*Received by the editors June 4, 1990; accepted (in revised form) December 2, 1991.

†Department of Mathematical Sciences, Montana State University, Bozeman, Montana, 59715. This research was supported in part by the Institute for Mathematics and Its Applications, University of Minnesota.

‡Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts 01609. This research was supported in part by National Science Foundation grant DMS-8920597.

density. Then, under the above assumptions, ρ_1, ρ_2, ρ_3 satisfy the following system of partial differential equations,

$$
(2) \qquad \begin{aligned}
\rho_{1,t} &= \rho_{1,xx} - \tau_1\rho_1 + \frac{r}{n}\left(\rho_1 + \frac{1}{2}\rho_2\right)^2, \\
\rho_{2,t} &= \rho_{2,xx} - \tau_2\rho_2 + \frac{2r}{n}\left(\rho_1 + \frac{1}{2}\rho_2\right)\left(\rho_3 + \frac{1}{2}\rho_2\right), \\
\rho_{3,t} &= \rho_{3,xx} - \tau_3\rho_3 + \frac{r}{n}\left(\rho_3 + \frac{1}{2}\rho_2\right)^2.
\end{aligned}
$$

These equations hold without any assumptions on the dependence of the birth and death rates on x, t and ρ_i.

In population genetics, the frequency of an allele is more interesting than the densities of the genotypes. Let $p(x,t) = (\rho_1 + \frac{1}{2}\rho_2)/n$ be the frequency of allele A in the population. Then a straightforward but tedious calculation yields the following equation for p,

$$
(3) \quad p_t = p_{xx} + 2\frac{p_x n_x}{n} + f(p,n)p(1-p) + \frac{1}{4}\{(\tau_2 - \tau_3)p - (\tau_2 - \tau_1)(1-p)\}\sigma.
$$

In the above equation $f(p,n) = p(\tau_2 - \tau_1) + (1-p)(\tau_3 - \tau_2)$, $\sigma = (\rho_2^2 - 4\rho_1\rho_3)/n^2$, and we have assumed that r and τ_i depend on p and n only. We can also obtain an equation for n by simply adding the equations in (2). Doing so, we obtain,

$$
(4) \qquad n_t = n_{xx} + g(p,n)n + (\tau_1 - 2\tau_2 + \tau_3)\frac{\sigma n}{4},
$$

where $g(p,n) = r - p^2\tau_1 - 2p(1-p)\tau_2 - (1-p)^2\tau_3$. The above method of deriving (3) and (4) from (2) is contained in the appendix of [1].

Equations (3) and (4) are insufficient to determine p and n; we need another equation for σ. The quantity σ measures the deviation of the population from Hardy–Weinberg equilibrium. For discrete-time models, the Hardy–Weinberg principle says that with random mating, and in the absence of factors which affect the gene frequencies, the genotype frequencies will arrive at and remain in the proportion $p^2 : 2p(1-p) : (1-p)^2$ after one generation. Such a proportion is called the Hardy–Weinberg equilibrium. Note that $\sigma = 0$ in this case. In a continuous-time model, Hardy–Weinberg equilibrium is attained only asymptotically [6]. In this paper we shall make the assumption that $\sigma = 0$. Hence we obtain the following reaction-diffusion system,

$$
(5) \qquad \begin{aligned}
p_t &= p_{xx} + 2\frac{p_x n_x}{n} + f(p,n)p(1-p), \\
n_t &= n_{xx} + g(p,n)n.
\end{aligned}
$$

It is worthwhile to see how (1) can be derived from (2) using a scaling argument. Let $v = \rho_2/n$. From (2), an equation for v can be derived which we shall not display here. Let $\epsilon = |\tau_1 - \tau_2| + |\tau_2 - \tau_3|$ and assume that it is sufficiently small. (This is called weak selection in population genetic theory.) Then by rescaling time by ϵ and space by $\sqrt{\epsilon}$, some of the terms in the equation for v will contain an ϵ in front. As a first approximation we set $\epsilon = 0$. Then the terms without ϵ in the equation for v imply that $\sigma = 0$. In other words, Hardy–Weinberg equilibrium is achieved. From (4), under the same scaling of space and time and setting $\epsilon = 0$, we obtain $g(p,n) = 0$. Suppose the birth and death rates are functions of n only. Then since $\tau_1 = \tau_2 = \tau_3 \equiv \tau$, we have $n = K$ which is the root of the equation $r(n) = \tau(n)$. For $\epsilon > 0$, we substitute $\sigma = 0$ and $n = K$ into (3) and obtain (1) where $h(p) = p(1-p)\{p(\tau_2(K) - \tau_1(K)) + (1-p)(\tau_3(K) - \tau_2(K))\}$. The above scaling argument is taken from §2.3 of [9].

By relabeling the two alleles A and a, it can always be assumed that $\tau_3 \geq \tau_1$ so that there are three cases to consider in (1), depending on whether τ_2 lies between, above or below τ_1 and τ_3. These are called the heterozygote intermediate, superior and inferior cases, respectively. In the last two cases, h has an intermediate zero between zero and 1.

The mathematical theory of (1) is very rich and well understood. One of the most intriguing properties is the existence of travelling waves. A travelling wave solution of (1) with speed θ is a nonconstant function $\tilde{u}(z)$ such that $\tilde{u}(x + \theta t)$ satisfies (1) for all x and $t > 0$. For example, in the heterozygote inferior case, if $\int_0^1 h > 0$, then there exists $\theta^* > 0$ such that a monotone travelling wave solution connecting zero to 1 exists if and only if $\theta = \theta^*$.

In this paper, we shall consider the heterozygote inferior case of (5); that is, $\tau_2(n)$ lies above $\tau_i(n)$ for $i = 1, 3$. We prove the existence of travelling wave solutions for (5) under additional assumptions on f and g. We shall discuss these assumptions in the next section. The proof of our existence theorem is based on the connection index from the Conley index theory. For the sake of completeness, we have provided a brief description of this index in §3. The computation of the connection index as well as the proof of our theorem are given in §4. To compute the connection index, we continue the original problem to a problem where the computation is much easier. In the last section, we provide a specific example and show that we can easily follow the above-mentioned continuation method numerically. In a forthcoming paper we shall prove that the travelling wave shown to exist here is stable in the case of weak selection.

2. Hypotheses and result. There are two types of hypotheses for our theorem, those that are motivated by our model ((A1) and (A2) below) and those that are necessary to complete our mathematical argument ((A3) and (A4) below). We begin by listing the hypotheses for f and g.

(A1) f and g are C^1 in p and n with $f_p > 0, f_n > 0, g_n < 0$ for $0 \leq p \leq 1$ and $n \geq 0$. Also, the relation

$$g_p(p,n) = f(p,n) \tag{6}$$

holds.

(A2) The nullclines $f = 0$ and $g = 0$ intersect at a point (p^*, n^*) in the region

$$Q \equiv \{(p,n) | 0 < p < 1 \text{ and } n > 0\} \quad \text{with } p^* < \tfrac{1}{2}.$$

(A3) Let the curve $g = 0$ intersect the line $p = 0$ at K_3 and the line $p = 1$ at K_1. We assume that $0 < K_3 < K_1$ and

$$\int_0^{p^*} f(p, K_3)p(1-p)dp < 0.$$

(A4) There exists $\alpha > 0$ such that $f(p, n^*) \geq \alpha(p - p^*)$ for $0 \leq p \leq 1$ and

$$g(0, n^*) < \frac{\min(1, \alpha)}{2}\left(\frac{1}{2} - p^*\right)^2. \tag{7}$$

From (A1) and (A2), one can determine the form of the nullclines $f(p,n) = g(p,n) = 0$. They are shown in Fig. 1.

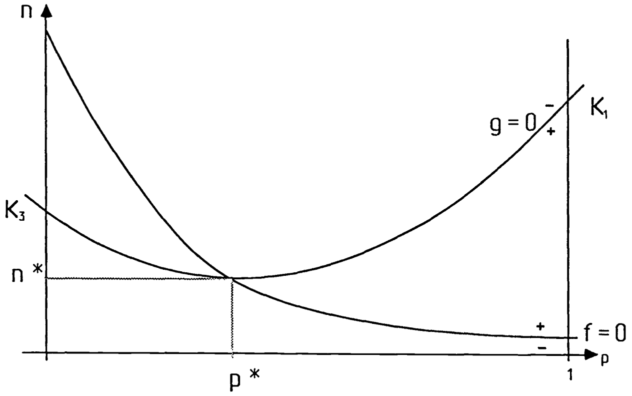


FIG. 1. *The form of the nullclines* $f = 0$ *and* $g = 0$.

We now explain how hypotheses (A1) and (A2) can be satisfied by our model.

Suppose the functions r and $\tau_i, i = 1,2,3$ depend only on n (density-dependent selection) and that they are continuously differentiable on the interval $[0, \infty)$. It is more convenient to write f and g in terms of the fitness functions η_i where $\eta_i(n) = r(n) - \tau_i(n)$. Doing so, we obtain,

$$\begin{aligned} f(p,n) &= p(\eta_1 - \eta_2) + (1-p)(\eta_2 - \eta_3), \\ g(p,n) &= p^2\eta_1 + 2p(1-p)\eta_2 + (1-p)^2\eta_3. \end{aligned} \tag{8}$$

Condition (6) is therefore satisfied. The function g represents the fitness of the entire population.

We are interested in the heterozygote inferior case of (5). A weaker condition than heterozygote inferiority is $\eta_1 + \eta_3 > 2\eta_2$ for $n \geq 0$. From (8), this is equivalent to the condition $f_p > 0$ for $n \geq 0$ in (A1).

In ecological models, it is frequently assumed that resources are scarce so that the growth rate of the population decreases with increase in population size. Thus, it is reasonable to assume that η_i is a decreasing function of n, positive near zero and negative for large n. This implies that $g_n < 0$ and that for each p, $g(p,n) = 0$ has a (unique) positive root.

Finally, from (8), the condition $f_n > 0$ for $0 \leq p \leq 1$ and $n \geq 0$ in (A1) is equivalent to $\eta_1' > \eta_2' > \eta_3'$. Such an assumption is important because it allows us to use the comparison principle on the first equation of (5). The comparison principle is not valid for (5). For an example where these inequalities and $\eta_1 + \eta_3 > 2\eta_2$ are satisfied let $\eta_i(n) = r_i(1 - (n/K_i)), i = 1,2,3$ where r_i, K_i are positive constants chosen so that $2r_2 < r_1+r_3$, $r_1/K_1+r_3/K_3 < 2(r_2/K_2)$ and $r_1/K_1 < r_2/K_2 < r_3/K_3$.

Since $f_p > 0$ and $g_n < 0$, the implicit function theorem implies that there exist functions $\tilde{n}$ and $\hat{n}$ such that $f(p, \tilde{n}(p)) = 0$ and $g(p, \hat{n}(p)) = 0$ for p in the unit interval. Since $f_n > 0$, $\tilde{n}$ is decreasing in p. We assume that the graphs of $\tilde{n}$ and $\hat{n}$ intersect at some point (p^*, n^*) where $0 < p^* < \frac{1}{2}$ and $n^* > 0$. From (6), it is easy to see that (p^*, n^*) is unique and $\hat{n}$ achieves a minimum at p^*. From (8), the function f can be written as $f(p,n) = C_1(n)(p - C_2(n))$, where $C_1(n) = \eta_1 + \eta_3 - 2\eta_2$ and $C_2(n) = (\eta_3 - \eta_2)/(\eta_1 + \eta_3 - 2\eta_2)$. Thus $p^* < \frac{1}{2}$ if and only if $\eta_3(n^*) < \eta_1(n^*)$.

Hypotheses (A1) and (A2) are not enough to prove our theorem. Two technical assumptions, (A3) and (A4), have to be added. Assumption (A3) is used only in the proof of Lemma 4.4 while (A4) is used only in the proof of Lemma 4.9. Recall the definition of $C_2(n)$ from the above paragraph. From the form of f given above, we see that the condition $f(p,n^*) \geq \alpha(p - p^*)$ is an equality and is always satisfied. Also, (7) is satisfied for sufficiently small p^* since $g(0,n^*) = \alpha(p^*)^2$. To see this, solve p^* in terms of $\eta_i(n^*)$ by writing $g(p^*,n^*) = 0$ as a quadratic equation in p^*. From the above paragraph, $p^* = C_2(n^*)$. Setting these two quantities equal, we obtain $\eta_2^2(n^*) = \eta_1(n^*)\eta_3(n^*)$ which is equivalent to $g(0,n^*) = \alpha(p^*)^2$. Finally, substituting $\eta_2(n^*) = -\sqrt{\eta_1(n^*)\eta_3(n^*)}$ into $p^* = C_2(n^*)$, we obtain

$$p^* = \frac{\sqrt{\eta_3(n^*)}}{\sqrt{\eta_1(n^*)} + \sqrt{\eta_3(n^*)}}$$

so that p^* is small if and only if $\eta_3(n^*)/\eta_1(n^*)$ is small.

It is obvious that the constant solutions of (5) in $\mathrm{cl}(Q)$ are $(0,0)$, $(1,0)$, (p^*,n^*), $(0,K_3)$ and $(1,K_1)$. If we only consider solutions that are spatially homogeneous, then (5) becomes a system of ordinary differential equations. From assumption (A1), it is easily checked that the first three solutions are unstable and the last two are stable. This type of system where there are exactly two stable equilibria is better known as a bistable system.

By a travelling wave solution of (5) with speed θ, we mean a nonconstant, bounded solution $(\tilde{p}, \tilde{n})$ such that $(\tilde{p}, \tilde{n})(x + \theta t)$ satisfies (5) for all x and $t > 0$. Equivalently, $(\tilde{p}, \tilde{n})$ satisfies the system of ordinary differential equations,

$$\begin{aligned} p'' - \theta p' + 2\frac{p'n'}{n} + f(p,n)p(1-p) &= 0, \\ n'' - \theta n' + g(p,n)n &= 0, \end{aligned} \tag{9}$$

on $\mathbf{R}$ where $' = d/dz$. As in the case of a single equation, we look for a travelling wave solution of (5) which connects the two stable equilibria $(0,K_3)$ and $(1,K_1)$; i.e., $(\tilde{p}, \tilde{n})$ satisfies the boundary conditions:

$$\lim_{z\to-\infty}(p(z),n(z))=(0,K_3),$$
$$\lim_{z\to\infty}(p(z),n(z))=(1,K_1). \tag{10}$$

Under the hypotheses (A1)–(A4) we can prove the following theorem.

THEOREM 2.1. *There exists a positive wave speed θ such that* (9) *has a solution* $(\tilde{p},\tilde{n})$ *which satisfies* (10). *Furthermore,* $\tilde{p}'>0$ *while* $\tilde{n}$ *has at most one local minimum on* $\mathbf{R}$.

The proof of Theorem 2.1 is based on the connection index theory.

3. The connection index. In this section we shall provide a cursory description of the connection index so that readers who are unfamiliar with such concepts can understand the proof of our theorem quickly. Many technical details are therefore omitted, but they can all be found in the papers [4], [5], [12]. The connection index is actually based on the Conley index [3] which we now describe.

3.1. The Conley index. Consider a flow defined by an autonomous system of differential equations on $\mathbf{R}^n$. Suppose $N\subset\mathbf{R}^n$ is compact. Let $I(N)$ denote the set of all points $x\in\mathbf{R}^n$ whose entire orbit (solution curve) through x is contained in N. If $S=I(N)$ is interior to N, then S is an isolated invariant set and N an isolating neighborhood. It is clear that a compact set N is an isolating neighborhood for S if every orbit which hits the boundary of N eventually leaves N in either forward or backward time, and if no orbit in S gets arbitrarily close to the boundary of N.

DEFINITION 1. Let S be an isolated invariant set with isolating neighborhood N. An *index pair* for S is a pair of compact sets (N_1,N_0) with $N_0\subset N_1\subset N$ such that:

(i) $\mathrm{cl}(N_1\backslash N_0)$ is an isolating neighborhood for S.

(ii) N_i is positively invariant relative to N for i=0,1, i.e., given $x\in N_i$ and $x\cdot[0,t]\subset N$, then $x\cdot[0,t]\subset N_i$.

(iii) N_0 is an exit set for N_1, i.e., if $x\in N_1$, $x\cdot[0,\infty)\not\subset N_1$, then there is a $T\geq 0$ such that $x\cdot[0,T]\subset N_1$ and $x\cdot T\in N_0$.

Given an index pair, the Conley (homotopy) index of S is defined to be the homotopy type of the pointed space N_1/N_0 obtained by collapsing N_0 to a point. This homotopy index is well defined and depends only on the invariant set S [3]. We shall denote the Conley index of S by $h(S)$.

The easiest example is when $S=\emptyset$. Then $(\emptyset,\emptyset)$ is an index pair. On collapsing the empty set to a point, a pointed one-point space is obtained. The homotopy type of this space, hence $h(\emptyset)$, is denoted by $\overline{0}$. If S is a hyperbolic rest point for the flow with a k-dimensional unstable manifold, then $h(S)=\Sigma^k$, the homotopy type of a pointed k-sphere.

An important property of the index is the sum formula. The sum of two pointed spaces (A,a) and (B,b) is defined as $A\cup B/\{a,b\}$, the pointed space obtained by taking the union of A and B and identifying the distinguished points a and b. This sum is denoted by $(A,a)\vee(B,b)$. If S_1 and S_2 are two isolated invariant sets with $S_1\cap S_2=\emptyset$, then $h(S_1\cup S_2)=h(S_1)\vee h(S_2)$. Furthermore, $\overline{0}\vee h(S)=h(S)$ for any isolated invariant set S.

The product of two pointed spaces can also be defined. If S_1 and S_2 are isolated invariant sets for the two flows $x^{'}=f(x)$ and $y^{'}=g(y)$, respectively, then $S_1\times S_2$

is an isolated invariant set for the product flow. The index of $S_1 \times S_2$ is given by $h(S_1 \times S_2) = h(S_1) \wedge h(S_2)$ where $\wedge$ is the smash product. For pointed spheres, we have $\Sigma^m \wedge \Sigma^n = \Sigma^{m+n}$ for all m and $n \geq 0$. Furthermore, $\overline{0} \wedge h(S) = \overline{0}$.

Finally, the Conley index also has the continuation property. Suppose S is an isolated invariant set with isolating neighborhood N and we continuously deform the flow so that N remains an isolating neighborhood throughout. Then the index of S before and after the deformation are the same.

3.2. The connection index. Suppose that a one-parameter family of flows on $\mathbf{R}^n$ is given by

$$x' = f(x,\theta) \tag{11}$$

where f depends continuously on $\theta \in [\theta_1, \theta_2]$. By appending the equation

$$\theta' = 0, \tag{12}$$

we obtain a flow, denoted by Φ, on $X = \mathbf{R}^n \times [\theta_1, \theta_2]$. Let S, S', and S'' be isolated invariant sets for the flow Φ and let $S(\theta)$ denote the θ slice of S.

DEFINITION 2. The triple (S, S', S'') is called a connection triple if:

(i) $S' \cup S'' \subset S$,

(ii) $S' \cap S'' = \emptyset$,

(iii) $S(\theta) = S'(\theta) \cup S''(\theta)$ for $\theta = \theta_1$ and θ_2.

A homotopy invariant index, called the connection index, can be defined for the connection triples [5]. It has many properties similar to the Conley index. We denote the connection index by $\overline{h}(S, S', S'')$ and postpone its definition together with an example to the end of this section.

Our proof of Theorem 2.1 relies on the following result in [5].

THEOREM 3.1. *Let (S, S', S'') be a connection triple for the flow Φ and suppose that $S = S' \cup S''$. Then $\overline{h}(S, S', S'') = (\Sigma^1 \wedge h(S')) \vee h(S'')$, where $h(S')$ and $h(S'')$ are the Conley indices of S' and S'' for the flow Φ, respectively.*

This theorem clearly implies that if one can prove that $\overline{h} \neq (\Sigma^1 \wedge h(S')) \vee h(S'')$ for some connection triple (S, S', S''), then there exists $\theta \in (\theta_1, \theta_2)$ such that $S(\theta) \not\subset S'(\theta) \cup S''(\theta)$. Our connection triple is constructed so that $S'(\theta)$ and $S''(\theta)$ are the rest points. Therefore there must be another orbit in $N(\theta)$ besides $S'(\theta)$ and $S''(\theta)$. From our construction of the isolating neighborhood, this orbit in $N(\theta)$ is the desired travelling wave with wave speed θ.

To show that $\overline{h} \neq (\Sigma^1 \wedge h(S')) \vee h(S'')$, we compute $\overline{h}$, $h(S')$ and $h(S'')$ via a continuation argument. The idea is similar to that of the Conley index described at the end of the previous section. We parameterize the flow Φ on X by $\lambda \in [0,1]$ and call it $\Phi(\lambda)$ where $\Phi(1) = \Phi$. Let $Y = X \times [0,1]$ with the obvious flow defined on Y and let S_0 and S_1 be isolated invariant sets for the flows $\Phi(0)$ and $\Phi(1)$, respectively. Then S_0 and S_1 are said to be related by continuation if there is an isolating neighborhood N for the flow on Y such that $S_0 = I(N(0))$ and $S_1 = I(N(1))$. ($N(\lambda)$ is the λ slice of N and is an isolating neighborhood for $\Phi(\lambda)$.) According to §3.1, $h(S_0) = h(S_1)$. Presumably, $h(S_0)$ is easier to compute than $h(S_1)$. Similarly, we can define a continuation of connection triples. Suppose S', S'', S''' are isolated invariant sets for the flow on Y such that for each λ, $(S'(\lambda), S''(\lambda), S'''(\lambda))$ is a connection triple. Then the connection triples at $\lambda = 0$ and at $\lambda = 1$ are related by continuation and have the same connection index.

We now give the definition of a connection triple (S, S', S''). The following is taken from [5].

Extend the flow (11) so that it is defined for $\theta \in [\theta_1 - \epsilon, \theta_2 + \epsilon]$ for some $\epsilon > 0$. Let U' and U'' be open neighborhoods in $\mathbf{R}^n \times [\theta_1 - \epsilon, \theta_2 + \epsilon]$ of $S'(\theta_1) \cup S'(\theta_2)$ and $S''(\theta_1) \cup S''(\theta_2)$, respectively, and choose them so that they have disjoint closures. Let ϕ be a continuous real-valued function on $\mathbf{R}^n$ which is positive on U' and negative on U'' and zero everywhere else. Append to the above given family of equations the equation $\theta' = \mu\phi(x)[\theta - (\theta_1 + \theta_2)/2]$, where μ is a small positive parameter. Let N be a compact neighborhood in $\mathbf{R}^n \times (\theta_1 - \epsilon, \theta_2 + \epsilon)$ such that $N(\theta)$ is an isolating neighborhood of $S(\theta)$ for each θ. Then there is a $\mu_0 > 0$ such that if $\mu \in (0, \mu_0)$, then N is an isolating neighborhood for the appended equation. Let h_μ be the Conley index of $I(N)$ for $\mu \in (0, \mu_0)$. Then h_μ is independent of μ and in fact depends only on the triple (S, S', S''). We define $\overline{h}(S, S', S'') = h_\mu$.

By way of example, consider the following system of equations,

$$
\begin{aligned}
u' &= v, \\
v' &= \theta v - u(1-u)(u - u^*)
\end{aligned}
$$

where $\theta > 0$ and $u^* \in \frac{1}{2}$. This is the bistable equation which has an increasing travelling wave solution connecting $(u, v) = (0,0)$ to $(1,0)$ for a positive wave speed $\theta = \theta^*$. The points $(0,0)$ and $(1,0)$ are saddles for all values of θ while $(u^*, 0)$ changes from an unstable spiral to an unstable node as θ increases, say from θ_1 near zero to $\theta_2 > \theta^*$. The phase plane diagram for the two values of θ are shown in Fig. 2 and 3 below. An isolating neighborhood is also shown where the exit set is marked. Let N denote this set cross $[\theta_1, \theta_2]$. Then N is an isolating neighborhood for the above flow and $\theta' = 0$. Let $S = I(N)$, $S' = (0,0) \times [\theta_1, \theta_2]$, and $S'' = (1,0) \times [\theta_1, \theta_2]$. Then (S, S', S'') is a connection triple according to Definition 2. To calculate its index, one can follow the recipe described in the above paragraph. An easier, though somewhat incorrect, method is the following (see §4B of [12]).

Let N_0 be a subset of N such that $(N(\theta), N_0(\theta))$ forms an index pair for $S(\theta)$. N_0 is just the exit set of N. Let $\hat{N}_0$ be N_0 together with the closure of all orbit segments in $N(\theta_1)$ and $N(\theta_2)$ which tend to $S'(\theta_1)$ and $S'(\theta_2)$ in negative time, respectively. Then $\overline{h}(S, S', S'')$ is the homotopy type of the pointed space $N/\hat{N}_0$. The reason why this is somewhat incorrect is because $S'(\theta_i) \not\subset N \backslash \hat{N}_0$, $i = 1, 2$, so that $(N, \hat{N}_0)$ is technically not an index pair according to the definition in the previous section. Therefore, we have to modify the flow near $S'(\theta_i)$ and $S''(\theta_i)$,$i = 1, 2$. It turns out that the homotopy type of N/N_0 for the modified flow is the same as the homotopy type of $N/\hat{N}_0$ as defined above.

Now the exit set at θ_i, $i = 1, 2$, consists of three disjoint parts and two of these parts are connected to the unstable manifold at $S'(\theta_i)$. (See Figs. 2 and 3.) Therefore, $\hat{N}_0$ is contractable to a point on the boundary of N, which is homotopic to the surface of a ball. Thus $N/\hat{N}_0$ is homotopy equivalent to the one-point space, or $\overline{0}$. The connection index is $\overline{0}$ for this example.

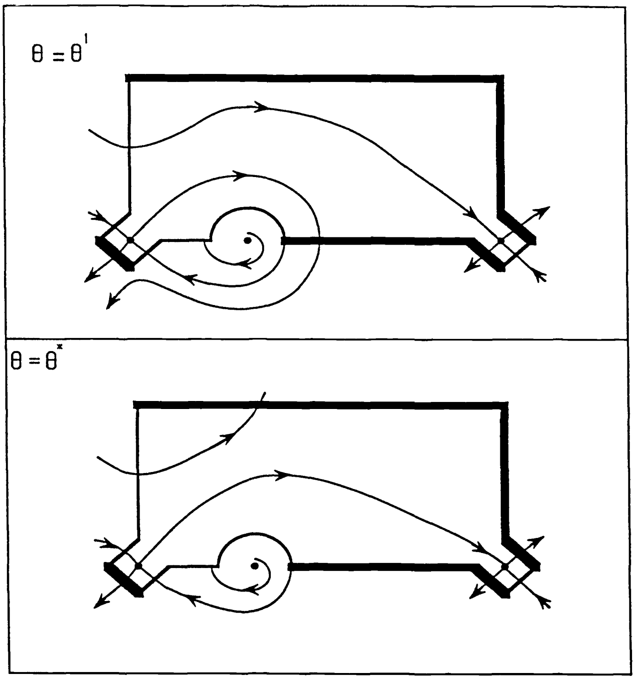


FIG. 2. *The phase planes for the bistable equation at $\theta = \theta_1$ and θ^*.*

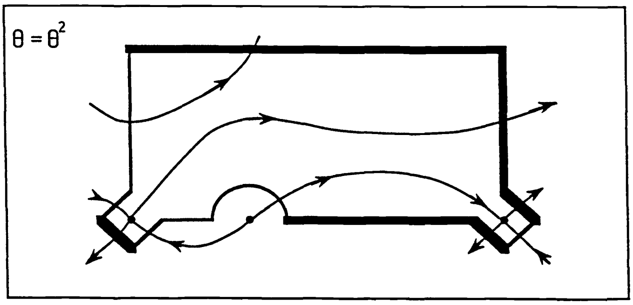


FIG. 3. *The phase plane for the bistable equation at $\theta = \theta_2$.*

4. Proof of Theorem 2.1. In this section we shall use the connection index to prove the existence of travelling waves. It is difficult to compute this index for the original model, so we continue the model to a system for which the connection index is easy to compute and then apply the continuation theorem in the previous section.

4.1. The homotopy. Let $\lambda \in [0,1]$ and consider the system,

$$\begin{aligned} p'' - \theta p' + 2\lambda \frac{p'n'}{n} + f^\lambda(p,n)p(1-p) = 0, \\ n'' - \theta n' + g^\lambda(p,n)n = 0, \end{aligned} \tag{13}$$

where $f^\lambda(p,n) = \lambda f(p,n) + (1-\lambda)(p-p^*)$ and $g^\lambda(p,n) = \lambda g(p,n) + (1-\lambda)(n^* - n)$. When $\lambda = 1$, we recover our original model and when $\lambda = 0$, (13) decouples into the

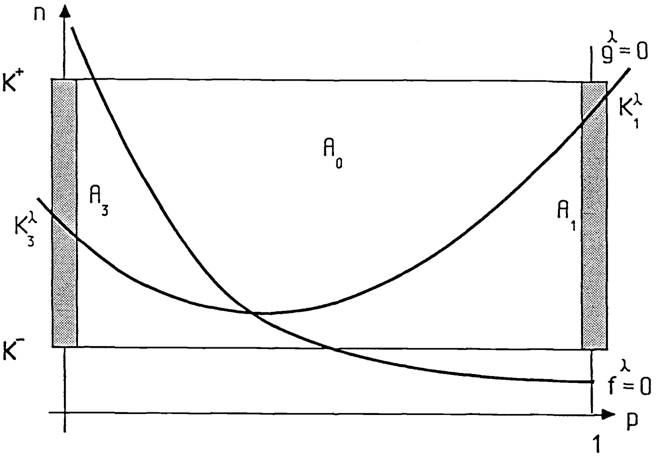


FIG. 4. *Projection of the isolating neighborhood onto the p-n plane.*

bistable equation,

$$p'' - \theta p' + p(1-p)(p-p^*) = 0 \tag{14}$$

and the Fisher equation,

$$n'' - \theta n' + n(n^* - n) = 0. \tag{15}$$

It is easy to check that except for (6), conditions (A1) to (A3) of §2 hold for f^λ and g^λ with K_i replaced by K_i^λ, $i = 1, 3$. We define K_i^λ by $g^\lambda(0, K_3^\lambda) = 0$ and $g^\lambda(1, K_1^\lambda) = 0$. The nullclines $f^\lambda = 0$ and $g^\lambda = 0$ are similar in form to the nullclines $f = 0$ and $g = 0$, respectively. In fact, $f^\lambda = 0$ lies between $f = 0$ and $p = p^*$ while $g^\lambda = 0$ lies between $g = 0$ and $n = n^*$. They intersect only at (p^*, n^*).

It is convenient to write (13) as a first order system:

$$\begin{aligned} p' &= v_1, \\ v_1' &= \theta v_1 - 2\lambda \frac{v_1 v_2}{n} - f^\lambda(p,n)p(1-p), \\ n' &= v_2, \\ v_2' &= \theta v_2 - g^\lambda(p,n)n. \end{aligned} \tag{16}$$

For each θ and λ, (16) defines a flow on $\mathbf{R}^4$. The rest points are $Y_1^\lambda = (1, 0, K_1^\lambda, 0)$, $Y^* = (p^*, 0, n^*, 0)$, $Y_3^\lambda = (0, 0, K_3^\lambda, 0)$, $(1,0,0,0)$ and $(0,0,0,0)$. A travelling wave solution of (5) corresponds to a solution of (16) which connects Y_3^1 to Y_1^1.

4.2. The isolating neighborhood. We first find a set N in $\mathbf{R}^4$, independent of θ and λ, such that it is an isolating neighborhood for the flow (16). The only rest points in N are Y_1^λ and Y_3^λ, and p is increasing along any nonconstant orbit in N.

Let K^+ and K^- be such that $0 < K^- < K_1^\lambda, K_3^\lambda < K^+$ for all $\lambda \in [0,1]$, $f = 0$ intersects $p = 0$ above K^+ and intersects $p = 1$ below K^-. Let $A_0 = \{(p,n) \mid 0 \leq p \leq 1 \text{ and } K^- \leq n \leq K^+\}$ (see Fig. 4).

If $Y = (p, v_1, n, v_2)$ is a solution of (16) with $\theta \geq \theta_0 > 0$ and $-1 \leq p(z) \leq 2$, $K^- \leq n(z) \leq K^+$ for all $z \in \mathbf{R}$, then $|v_i| \leq L$ for $i = 1, 2$. To see this, choose C such that $|g^\lambda(p,n)n| \le C$ for all $\lambda \in [0,1]$ and above values of p, n and define $L = C/\theta_0$. Suppose $v_2(z_0) > L$. Then from (16), $v_2'(z_0) > 0$ which implies that $v_2(z) > L$ for all $z \ge z_0$. Hence n is unbounded which is a contradiction. Thus, $|v_2| \le L$. A similar argument can be used to show that $|n^{2\lambda}p'|$, and hence $|p'|$, is bounded if we observe that p satisfies the equation,

$$(n^{2\lambda}p')' - \theta(n^{2\lambda}p') + f^\lambda(p,n)p(1-p)n^{2\lambda} = 0.$$

Let $N_0 = \{(p,v_1,n,v_2)|(p,n) \in A_0, 0 \le v_1 \le L \text{ and } |v_2| \le L\}$. N_0 is not an isolating neighborhood since Y_1^λ, Y_3^λ and Y^* belong to the boundary of N_0. We need to add to N_0 neighborhoods of Y_1^λ and Y_3^λ and remove a neighborhood of Y^* to obtain an isolating neighborhood.

To add a neighborhood of Y_3^λ, let $A_3 = \{(p,n) \mid |p| \le \delta, K^- \le n \le K^+\}$ where δ is independent of λ. By assumption (A1) we can choose $\delta > 0$ small enough so that $g^\lambda = 0$ intersects the boundary of A_3 only in the $|p| = \delta$ faces for all $\lambda \in [0,1]$ and A_3 lies below the curve $f = 0$ (see Fig. 4). Let $N_3 = \{(p,v_1,n,v_2) \mid (p,n) \in A_3 \text{ and } |v_i| \le L \text{ for } i = 1,2\}$. Recall that $I(N)$ is the set of all orbits of (16) that lie in N for all z.

LEMMA 4.1. $I(N_0 \cup N_3) = I(N_0)$.

Proof. We need to show that $p \ge 0$ and $v_1 \ge 0$ along any orbit in $I(N_0 \cup N_3)$ so that the orbit actually lies in $I(N_0)$.

Let $Y = (p,v_1,n,v_2)$ be an orbit in $I(N_0 \cup N_3)$. Suppose p has a negative minimum at z_0 where $v_1(z_0) = 0$ and $v_1'(z_0) \ge 0$. Since $f^\lambda(p,n) < 0$ if $p < 0$ and $n \in [K^-, K^+]$, (16) implies that $v_1'(z_0) < 0$ which is a contradiction. Therefore, $p \ge 0$.

Now suppose $v_1(z_0) < 0$. If $v_1(z) < 0$ for all $z < z_0$, then Y must tend to a rest point in N_3 as $z \to -\infty$. There is only one rest point in N_3, namely Y_3^λ. Since $p \ge 0$, we must have $p(-\infty) > p(0) \ge 0$ which contradicts the fact that $p = 0$ at Y_3^λ. Therefore, $v_1(z_1) = 0$ for some $z_1 < z_0$ and $v_1(z) < 0$ on $(z_1, z_0]$. Hence $v_1'(z_1) \le 0$. This assumption also implies that $(p,n)(z) \in A_3$ for all $z \in [z_1, z_0]$ and hence $f^\lambda(p,n) < 0$ at $z = z_1$. From (16), $v_1'(z_1) \ge 0$ and hence $v_1'(z_1) = 0$ and $p(z_1) = 0$. But then Y must lie in the invariant manifold $p \equiv 0$, $v_1 \equiv 0$ because of uniqueness which then contradicts the assumption that $v_1(z_0) < 0$. Therefore, $v_1 \ge 0$. This completes the proof of the lemma. □

We add a neighborhood of Y_1^λ in a similar manner. Let $A_1 = \{(p,n) \mid |p-1| < \delta_1, K^- \le n \le K^+\}$. We choose δ_1 independent of λ and sufficiently small so that $g^\lambda = 0$ intersects the boundary of A_1 only in the $p = 1 - \delta_1$ and $p = 1 + \delta_1$ faces (see Fig. 4). Let $N_1 = \{(p,v_1,n,v_2) \mid (p,n) \in A_1 \text{ and } |v_i| \le L \textit{ for } i = 1,2\}$. The proof of the following lemma is similar to that of Lemma 4.1 and is omitted.

LEMMA 4.2. $I(N_0 \cup N_1 \cup N_3) = I(N_0)$.

Let $\hat{N} = N_0 \cup N_1 \cup N_3$. $\hat{N}$ is still not an isolating neighborhood since Y^* is on its boundary. We need to remove a neighborhood of Y^*. To do this we first show that $v_1 > 0$ along any nonconstant orbit in $I(\hat{N})$.

LEMMA 4.3. *For any orbit $Y \in I(\hat{N})$, $n \ge n^*$ for all $z \in \mathbf{R}$.*

Proof. Suppose n has a local minimum less than n^* at say $z = 0$. Then $(p,n)(0)$ must lie in the region where $g^\lambda > 0$. From (16), $v_2'(0) < 0$ which is a contradiction. Therefore $n \ge n^*$ for all z. □

LEMMA 4.4. *Let $\theta > 0$. Then $v_1 > 0$ along any orbit in $I(\hat{N})$ except when the orbit is Y^*, Y_3^λ, or Y_1^λ.*

Proof. Suppose Y is an orbit in $I(\hat{N}) = I(N_0)$ with $v_1(z) = 0$ at say $z = 0$. Then $v_1'(0) = 0$ for otherwise Y would leave N_0 in either forward or backward time. From (16), $f^\lambda(p,n)p(1-p) = 0$ at $z = 0$. If $p(0) = 0$ or 1, then Y lies in the invariant manifold $p \equiv 0, v_1 \equiv 0$ or $p \equiv 1, v_1 \equiv 0$. In each case, a simple phase plane analysis reveals that the only orbit that lies between K^+ and K^- for all z is the rest point Y_3^λ or Y_1^λ, respectively. Thus we need only consider the case $f^\lambda(p(0), n(0)) = 0$. From Lemma 4.3, the point $(p(0), n(0))$ must lie on or above the curve $g^\lambda = 0$ which implies that $p(0) \leq p^*$. If it lies on $g^\lambda = 0$ and $v_2(0) = 0$, then $Y \equiv Y^*$. Otherwise, Y must tend to Y_3^λ as $z \to -\infty$. To see this, differentiate the equation for p to obtain $p''' + [f(p,n)]'p(1-p) = 0$ at $z = 0$. Since $p'''(0) \geq 0$, we have $n'(0) \leq 0$. If $n'(0) = 0$, then (16) implies that n has a local minimum at $z = 0$. Otherwise, $n'(0) < 0$. In both cases, since n cannot have a local maximum at a point above $g^\lambda = 0$ or a local minimum below it, we conclude that $(p,n)(-\infty) = (0, K_3^\lambda)$ and that $n' \leq 0$ for $z < 0$.

Multiply the first equation in (13) by $p' = v_1$ and integrate to obtain:

$$(17) \qquad \theta \int_{-\infty}^{0} (p')^2 dz = \int_{0}^{p(0)} f^\lambda(p, n(z^{-1}(p)))p(1-p)dp + 2\lambda \int_{-\infty}^{0} \frac{n'(p')^2}{n} dz.$$

The right side of (17) is less than $\int_0^{p^*} f^\lambda(p, K_3^\lambda)p(1-p)dp < 0$ because $n' < 0$ for $z < 0$, $f_n > 0$ and assumption (A3). This contradicts the assumption that $\theta > 0$. Therefore $v_1 > 0$ along any nonconstant orbit in $I(\hat{N})$. □

It now follows that the only orbit of (16) in $I(\hat{N})$ that hits the boundary of $\hat{N}$ is the constant solution Y^*. To see this, we may assume that the orbit Y lies entirely in N_0 because of Lemma 4.2. By our choice of L, $|v_i| < L$ along the orbit. The n-component of the orbit cannot hit the $n = K^\pm$ faces of $\partial\hat{N}$. (For example, if n has a local maximum at $z = 0$ and $n(0) = K^+$, then the last equation of (16) is contradicted.) If Y hits the boundary $p = 0$ or $p = 1$ of N_0, then $v_1 = 0$ also since Y lies in N_0. According to Lemma 4.4, $Y \equiv Y_3^\lambda$ or Y_1^λ which are in the interior of $\hat{N}$ because of the added neighborhood. Finally, if Y hits the boundary $v_1 = 0$ of N_0 and $p \neq 0$ or 1, then Lemma 4.4 implies that $Y \equiv Y^*$. Hence our assertion at the beginning of the paragraph is proved. We record it as a lemma.

LEMMA 4.5. *If $\theta \geq \theta_0 > 0$, then the only orbit of (16) in $I(\hat{N})$ which hits the boundary of $\hat{N}$ is $Y \equiv Y^*$.*

Finally, we must remove a neighborhood of Y^*. To do this, we use a result of [5] concerning the excision of a portion of an invariant set. Let S be a compact invariant set of a flow and $S_r \subset S$ an isolated set relative to S; that is, there is a compact relative neighborhood N_r of S_r in S such that $S_r = I(N_r)$. Let $A^+ = A^+(S, S_r)$ be the points on solutions in $S \backslash S_r$ that tend to S_r in forward time. Similarly, let $A^- = A^-(S, S_r)$ be those points which tend to S_r in backward time. A proof of the following lemma can be found in §4D of [5].

LEMMA 4.6. *Suppose $\tilde{N}$ is compact and $S = I(\tilde{N})$. Let $S_r \subset S$ be isolated relative to S and suppose that at least one of the sets A^+ and A^- is empty. Then for all sufficiently small neighborhoods W of S_r in $\tilde{N}$,*

$$(18) \qquad I(\tilde{N} \backslash W) \cap \partial(\tilde{N} \backslash W) = I(\tilde{N}) \cap \partial\tilde{N} \backslash (S_r \cup A^+ \cup A^-).$$

We will take $\tilde{N}$ to be $\hat{N}$ and S_r the rest point Y^*. In this case, $A^+(I(\hat{N}), Y^*)$ is the component of the stable manifold at Y^* that is contained in $I(\hat{N})$.

LEMMA 4.7. *For all $\lambda \in [0,1]$ and $\theta > 0$, $A^+(I(\hat{N}), Y^*) = \emptyset$.*

Proof. First note that if Y is a nonconstant orbit in $I(\hat{N})$, then by Lemma 4.2, $p \in [0,1]$ and by Lemma 4.4, $v_1 > 0$ for all z. Thus the only orbit in A^+ must be a heteroclinic connection from Y_3^λ to Y^*. For $\theta > 0$, there can be no such connection. The proof of this fact uses the same argument as in the second half of the proof of Lemma 4.4 and is omitted. Thus $A^+ = \emptyset$. □

The set $A^-(I(\hat{N}), Y^*)$ is nonempty since it is possible to find a connecting orbit, as for example when $\lambda = 0$, from Y^* to K_1^λ for sufficiently large θ. The unstable manifold at Y^* has dimension three and connecting orbits may be found using a shooting argument.

By Lemmas 4.6 and 4.7, for all sufficiently small neighborhoods W of Y^*, we have (18) holding. Since the λ-interval is compact, we can choose a neighborhood W so that (18) holds for all $\lambda \in [0,1]$. Finally, let $N = \hat{N} \backslash W$.

PROPOSITION 4.8. *N is an isolating neighborhood for the flow* (16) *for each $\lambda \in [0,1]$ and $\theta \geq \theta_0 > 0$.*

Proof. We must show that $I(N)$ is in the interior of N. From (18) and Lemma 4.7, we have

$$I(N) \cap \partial N = I(\hat{N}) \cap \partial \hat{N} \backslash (Y^* \cup A^-).$$

Suppose P belongs to $I(\hat{N}) \cap \partial \hat{N}$. If an orbit in $I(\hat{N})$ hits P in finite time, then by Lemma 4.5, $P = Y^*$. If P is approached by a nonconstant orbit Y in $I(\hat{N})$, then by Lemma 4.4 and the fact that the n-component of Y cannot have a local maximum (minimum) above (below) $g^\lambda = 0$, Y must connect two of the three rest points Y^*, Y_3^λ and Y_1^λ. It cannot connect Y_3^λ to Y^* because of Lemma 4.7. It cannot connect Y_3^λ and Y_1^λ since neither points are on $\partial \hat{N}$. Hence it must connect Y^* to Y_1^λ in which case $P = Y^*$. Finally, we have to consider the case when there exist orbits Y_n in $I(\hat{N})$ such that $Y_n(z_n) \rightarrow P$ as $n \rightarrow \infty$. Let $\hat{Y}_n(z) = Y_n(z_n - z)$. Then $\hat{Y}_n$ belongs to $I(\hat{N})$ and $\hat{Y}_n(0) \rightarrow P$ as $n \rightarrow \infty$. By the Arzela–Ascoli theorem, a subsequence of $\hat{Y}_n$ converges to an orbit $\hat{Y}$ in $I(\hat{N})$ where $\hat{Y}(0) = P$. From above, $P = Y^*$. Therefore, $I(N) \cap \partial N = \emptyset$ which completes the proof of the proposition. □

4.3. The proof of Theorem 2.1. We first derive a priori bounds on the wave speed θ.

LEMMA 4.9. *There exist $0 < \theta_* < \theta^*$, independent of $\lambda \in [0,1]$, such that if $(\hat{p}, \hat{n})$ is a nonconstant solution of* (13) *for some θ with $\hat{p} \in [0,1]$, $\hat{p}' > 0$, $\hat{n} \in [n^*, K^+]$, $\limsup_{z \to -\infty} \hat{p}(z) < p_1^*$ and $\liminf_{z \to \infty} \hat{p}(z) > p^*$, then $\theta \in (\theta_*, \theta^*)$. Here, p_1^* is the root of the equation $f(p, K^+) = 0$ that lies between zero and* 1.

Proof. We first show how to obtain the upper bound θ^* assuming that the lower bound θ_* has been found. Recall from the beginning of §4.2 that $\theta \geq \theta_*$ implies that $|\hat{n}'|$ is bounded independently of λ. Since $\hat{n} \geq n^*$, we can choose L_1 such that $|\hat{n}'/\hat{n}| \leq L_1$ for all z.

Consider the equation $w_t = w_{xx} + f^\lambda(w, K^+)w(1-w)$. Since

$$\int_0^1 f^\lambda(w, K^+)w(1-w)dw > 0,$$

this equation has a monotone travelling wave solution W^λ with positive wave speed

$\theta^{*,\lambda}$ such that $W^{\lambda}(-\infty)=0$ and $W^{\lambda}(\infty)=1$. Furthermore,

$$\theta^{*,\lambda}\int_{\mathbf{R}}([W^{\lambda}]')^2dz=\int_0^1 f^{\lambda}(w,K^+)w(1-w)dw.$$

This relation implies that $\theta^{*,\lambda}$ depends continuously on λ since $[W^{\lambda}]'$ depends on λ uniformly on $\mathbf{R}$. The travelling wave solution W^{λ} also has strong stability properties. Fife and McLeod showed in [10] that if the initial data w_0 satisfies the conditions $\limsup_{x\to-\infty} w_0(x)<p_1^*$ and $\liminf_{x\to\infty} w_0(x)>p_1^*$, then $w(x-\theta^{*,\lambda}t,t)$ is essentially bounded between two translates of W^{λ}.

Let $u(x,t)=\hat{p}(x+\hat{\theta}t)$ and $v(x,t)=\hat{n}(x+\hat{\theta}t)$ where $\hat{\theta}=\theta-2\lambda L_1$. From (13), $\hat{p}''-\hat{\theta}\hat{p}'+f^{\lambda}(\hat{p},\hat{n})\hat{p}(1-\hat{p})\geq 0$ so that u satisfies the inequality $u_t\leq u_{xx}+f^{\lambda}(u,v)u(1-u)$. Since $f_n>0$ and $n\leq K^+$, we have $u_t\leq u_{xx}+f^{\lambda}(u,K^+)u(1-u)$. Let w satisfy $w_t=w_{xx}+f^{\lambda}(w,K^+)w(1-w)$ with initial data $w_0=\hat{p}$. From the maximum principle, $u(x,t)=\hat{p}(x+\hat{\theta}t)\leq w(x,t)$ for all x and $t>0$. From the stability properties of W^{λ} mentioned above and our hypotheses on $\hat{p}$, there exist positive constants C,μ, and α such that

$$w(x-\theta^{*,\lambda}t,t)\leq W^{\lambda}(x-\alpha)+Ce^{-\mu t}$$

for all x and $t\geq 0$. Letting $y=x+(\hat{\theta}-\theta^{*,\lambda})t$, we have

$$\hat{p}(y)\leq W^{\lambda}(y+(\theta^{*,\lambda}-\hat{\theta})t-\alpha)+Ce^{-\mu t}.$$

If $\hat{\theta}>\theta^{*,\lambda}$, then letting $t\to\infty$ in the above inequality we obtain $\hat{p}(y)\leq 0$ for all $y\in\mathbf{R}$, which contradicts our hypotheses. Thus $\hat{\theta}\leq\theta^{*,\lambda}$. If we choose θ^* to be larger than $\max_{\lambda\in[0,1]}(\theta^{*,\lambda})+2L_1$, we obtain an upper bound for θ for all $\lambda\in[0,1]$. We now turn to finding θ_*.

Let $u=\hat{n}'/n$ and let $-\hat{\theta}=\min_z(-\theta+2\lambda u(z))$. From the definition of f^{λ}, assumption (A4), and the hypotheses of our lemma, we have

$$\hat{p}''-\hat{\theta}\hat{p}'+\beta(\hat{p}-p^*)\hat{p}(1-\hat{p})\leq 0$$

where $\beta=1-\lambda+\lambda\alpha\geq\min(1,\alpha)$. This inequality implies that $\hat{\theta}=\theta-2\lambda\min_z u(z)\geq\theta_*^{\lambda}$ where θ_*^{λ} is the wave speed of the bistable equation $w_t=w_{xx}+\beta(w-p^*)w(1-w)$. This fact may be proved using the same method we used to prove the upper bound θ^*. It is also known [2] that for the above bistable equation, θ_*^{λ} is given by $\sqrt{2\beta}(\frac{1}{2}-p^*)$. We now proceed to find the minimum of u.

If $n'\geq 0$, then $\min_z u(z)\geq 0$ so that $\theta\geq\sqrt{2\beta}(\frac{1}{2}-p^*)\geq\sqrt{2\min(1,\alpha)}(\frac{1}{2}-p^*)$. Suppose u is not monotone. Then u cannot have a local maximum. For if u achieves a local maximum at, say, $z=0$, then $(\hat{p},\hat{n})(0)$ lies below $g^{\lambda}=0$. Since $\hat{n}$ cannot have a local minimum below $g^{\lambda}=0$, $\hat{n}$ is either increasing for $z<0$ or decreasing for $z>0$. But then this would imply that either $(\hat{p},\hat{n})(-\infty)=(0,K_3^{\lambda})$ or $(\hat{p},\hat{n})(\infty)=(1,K_1^{\lambda})$ which is impossible. Therefore, we assume that $\hat{n}$ has a unique minimum at $z=0$, is decreasing on $(-\infty,0)$, is increasing on $(0,\infty)$ and the minimum of u occurs at some point $z_0<0$.

From (13), u satisfies the equation $u'=-u^2+\theta u-g^{\lambda}$ where $g^{\lambda}(z)=g^{\lambda}(\hat{p},\hat{n})(z)$. Therefore, $u(z_0)=(\theta-\sqrt{\theta^2-4g^{\lambda}(z_0)})/2$. Substituting this into the inequality $\theta-2\lambda\min_z u(z)\geq\theta_*^{\lambda}$, we obtain

$$(1-\lambda)\theta+\lambda\sqrt{\theta^2-4g^{\lambda}(z_0)}\geq\sqrt{2\beta}\left(\tfrac{1}{2}-p^*\right).$$

To estimate $g^\lambda(z_0)$, we claim that if $(p_i, n_j), i, j = 1, 2$, are four corners of a rectangle in the p-n plane where $0 < p_1 < p_2 < 1$ and $0 < n_1 < n_2 < K_3^\lambda$, then $g^\lambda(p_1, n_1) + g^\lambda(p_2, n_2) \geq g^\lambda(p_1, n_2) + g^\lambda(p_2, n_1)$. Assuming this for the moment, let $(p_2, n_2) = (\hat{p}, \hat{n})(z_0)$ and $(p_1, n_1) = (0, n^*)$. Then $g^\lambda(z_0) \geq -\lambda g(0, n^*)$ which according to (7) is greater than $-(\epsilon^2 \min(1, \alpha))(\frac{1}{2} - p^*)^2/2$ for some $0 < \epsilon < 1$. Therefore, by completing the square and rearranging the last inequality of the last paragraph, we have $\theta \geq \sqrt{2}\min(1, \alpha)(1 - \epsilon)(\frac{1}{2} - p^*) \equiv \theta_*$.

To prove our claim, let $(\tilde{p}(z), \tilde{n}(z)), 0 \leq z \leq 1$ be a line segment joining the points (p_1, n_2) to (p_2, n_1) so that $\tilde{p}' \geq 0$ and $\tilde{n}' \leq 0$. Then, since $g_{pn}^\lambda = 2\lambda f_n \geq 0$, we have,

$$
\begin{aligned}
g^\lambda(p_2, n_1) &= g^\lambda(p_1, n_2) + \int_0^1 g^\lambda(\tilde{p}(z), \tilde{n}(z))_z dz \\
&= g^\lambda(p_1, n_2) + \int_0^1 (g_p^\lambda \tilde{p}' + g_n^\lambda \tilde{n}') dz \qquad (19) \\
&\geq g^\lambda(p_1, n_2) + \int_0^1 g^\lambda(\tilde{p}(z), n_1)_z dz + \int_0^1 g^\lambda(p_2, \tilde{n}(z))_z dz, \qquad (20)
\end{aligned}
$$

which is the same as our claim. The proof of the lemma is complete. □

We are now ready to apply the connection index theory to prove Theorem 2.1. Let $X = \mathbf{R}^4 \times [\theta_*, \theta^*]$ and $Y = X \times [0, 1]$ with the obvious flow Φ defined on Y. Let N be the set defined near the end of §4.2 with $\theta_0 = \theta_*$. Then Proposition 4.8 implies that $\hat{N} = N \times [\theta_*, \theta^*] \times [0, 1]$ is an isolating neighborhood for the flow Φ. Let $S' = Y_3^\lambda \times [\theta_*, \theta^*] \times [0, 1]$, $S'' = Y_1^\lambda \times [\theta_*, \theta^*] \times [0, 1]$ and $S = I(\hat{N})$. Then S, S', S'' are isolated invariant sets for the flow Φ on Y. We claim that for each λ, $(S(\lambda), S'(\lambda), S''(\lambda))$ is a connection triple. Conditions (i) and (ii) of Definition 2 are obvious and condition (iii) follows from Lemma 4.9 above. According to §3.2, the connection index $\overline{h}(\lambda) = \overline{h}(S(\lambda), S'(\lambda), S''(\lambda))$ is independent of λ. Hence, $\overline{h}(1) = \overline{h}(0)$. When $\lambda = 0$, (16) uncouples and $(n, v_2) = (n^*, 0)$ is a saddle point for the two-dimensional flow obtained by writing (15) as a first-order system. From the example given in §3.2, $\overline{h}(0) = \Sigma^1 \wedge \overline{0} = \overline{0}$. Therefore, $\overline{h}(1) = \overline{0}$.

To compute the Conley index $h(S'(1))$, we first observe that $S'(0)$ and $S'(1)$ are related by continuation. When $\lambda = 0$, $Y_3^0 = (0, 0, n^*, 0)$ and $Y_1^0 = (1, 0, n^*, 0)$. If we write (14) as a first-order system, then $(p, v_1) = (0, 0)$ and $(1, 0)$ are both saddle points. Each has a one-dimensional unstable manifold. Hence $h(S'(1)) = h(S'(0)) = \Sigma^1 \times \Sigma^1 = \Sigma^2$, according to §3.1. Similarly, $h(S''(1)) = \Sigma^2$. Since $(\Sigma^1 \wedge h(S')) \vee h(S'') = \Sigma^3 \vee \Sigma^2 \not\simeq \overline{0} = \overline{h}(\lambda)$, Theorem 3.1 implies that there exists θ such that $I(N_\theta)$ contains a nonconstant orbit. From Lemma 4.4, this orbit must be a travelling wave solution connecting Y_3^1 to Y_1^1 with speed $\theta \in (\theta_*, \theta^*)$. Furthermore, p is increasing. Using the fact that n cannot have a maximum above the curve $g = 0$ or a minimum below the curve $g = 0$, it is easy to see that n can have at most one minimum on $\mathbf{R}$. The proof of Theorem 2.1 is complete.

5. A numerical example.

In the previous section we showed how the homotopy invariance of the connection index allowed us to ascertain the existence of a travelling wave solution for the model (5). In this section we shall show that one can follow this heteroclinic connection for a specific example using numerical continuation techniques.

Recall from the introduction that

$$f(p, n) = p(\eta_1 - \eta_2) + (1 - p)(\eta_1 - \eta_3)$$

and

$$g(p,n) = p^2\eta_1 + 2p(1-p)\eta_2 + (1-p)^2\eta_3.$$

Following [13] we assume that the fitness functions have the form:

$$\eta_i(n) = r_i(1 - n/K_i) \quad \text{for } i = 1, 2, \text{ and } 3.$$

To obtain an example of the heterozygote inferior case we use the following values of parameters:

i	1	2	3
r_i	0.6	0.7	0.8
K_i	12000	7300	8000

For the numerical computations we have found it convenient to scale n by 10^3. It is easily checked that this corresponds to scaling the K_i's by 10^{-3} in the above table. Since g is linear in n and quadratic in p one can write down explicit formula for p^* and n^*. Furthermore, it is easy to check that assumptions (A1)–(A4) are satisfied for this model with the parameters given above.

We seek a heteroclinic orbit for (16) connecting Y_3^λ at $z = -\infty$ to Y_1^λ at $z = +\infty$, where $Y_3^\lambda = (0, 0, K_3^\lambda, 0)$, $Y_1^\lambda = (1, 0, K_1^\lambda, 0)$, and

$$K_i^\lambda = K_i \frac{(1-\lambda)n^* - \lambda r_i}{(1-\lambda)K_i - \lambda r_i} \quad \text{for } i = 1 \text{ and } 3.$$

It is easy to show that for each $\theta > 0$, (16) has a two-dimensional unstable manifold and a two-dimensional stable manifold at Y_i^λ for $i = 1$ and 3. One can also write down explicit expressions for the eigenvalues as well as a set of orthonormal eigenvectors. Let $\{\phi_3^\lambda, \psi_3^\lambda\}$ denote an orthonormal bases for the tangent space to the unstable manifold at Y_3^λ and $\{\phi_1^\lambda, \psi_1^\lambda\}$ an orthonormal bases for the tangent space of the stable manifold at Y_1^λ.

The numerical method used here is similar to the method given in [7]. The approximation is based on the following equations:

$$Y' = TF^\lambda(Y, \theta) \quad \text{for } 0 < z < 1, \tag{21}$$

with the boundary conditions:

$$Y(0) = Y_3^\lambda + \epsilon_3(m_{11}\phi_3^\lambda + m_{12}\psi_3^\lambda) \quad \text{with } m_{11}^2 + m_{12}^2 = 1, \tag{22}$$

and

$$Y(1) = Y_1^\lambda + \epsilon_1(m_{21}\phi_1^\lambda + m_{22}\psi_1^\lambda) \quad \text{with } m_{21}^2 + m_{22}^2 = 1. \tag{23}$$

Here, $F^\lambda(Y, \theta)$ is the vector field on the right-hand side of (16) and T is a large positive constant.

Equation (21) is just the differential equation (16) with z scaled by T. The boundary condition (22) is the requirement that the initial value $Y(0)$ lies on the sphere of radius ϵ_3 intersect the linear approximation to the unstable manifold at

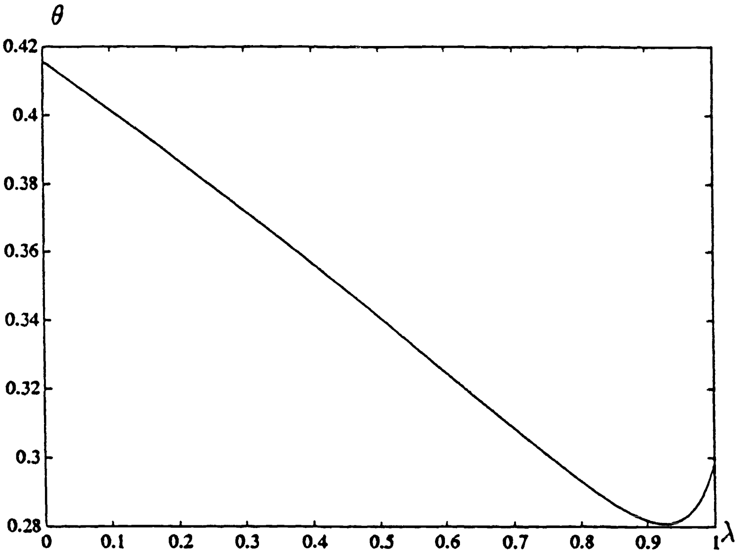


FIG. 5. *The wave speed θ versus the homotopy parameter λ.*

Y_3^λ. Likewise, (23) implies that $Y(1)$ lies on the sphere of radius ϵ_1 intersect the linear approximation to the stable manifold of Y_1^λ. The constant T is the time of travel between these two points for the original unscaled variable z. The m_{ij}'s are the projections onto the unstable and stable directions. For T large and ϵ_i's small, each solution of (20)–(22) represents an approximate heteroclinic connection.

In [7], the parameter T is fixed and the ϵ_i's are allowed to vary. Since the translate of a travelling wave is also a travelling wave, another constraint is needed to fix the phase of the solution. In [7], it was required that the L^2-norm of the difference between the derivatives of two successive approximations be at a minimum. This requirement takes the form of an integral condition. In the presence of sharp fronts, which occur for singularly perturbed equations, this condition is derived so as to economize the numerical calculations. For this model it is more economical to simply set $\epsilon_1 = \epsilon_3 \equiv \epsilon$, a fixed positive number, forego the integral constraint and allow T to be a free parameter. Along the solution branch we need to compute for each λ the solution vector Y, the wave speed θ, T and the m_{ij}'s so that the boundary conditions (21)–(22) hold.

To find a starting solution of (21) at $\lambda = 0$ we use $Y(z) = (u(z), u'(z), n^*, 0)$ where

$$u(z) = \frac{\exp(T\sqrt{2}/2(z-1/2))}{1+\exp(T\sqrt{2}/2(z-1/2))},$$

with $\theta = \sqrt{2}(1/2 - p^*)$. This is a solution of (16) at $\lambda = 0$ connecting Y_3^0 to Y_1^0. At $\lambda = 0$ we fix $T = 50$ and compute ϵ from the exact solution, thereafter holding ϵ fixed and allow T to vary. The numerical continuation is computed using the continuation program AUTO [8]. The results of this computation are shown below.

In Fig. 5 we have plotted the wave speed θ versus the homotopy parameter λ. In

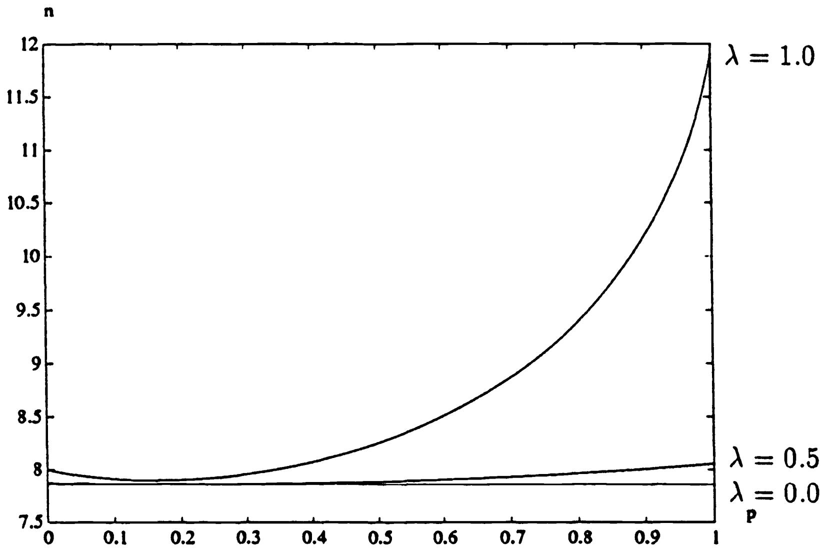


FIG. 6. *The n-component of the solution for* $\lambda = 0.0,\ 0.5,$ *and* 1.0.

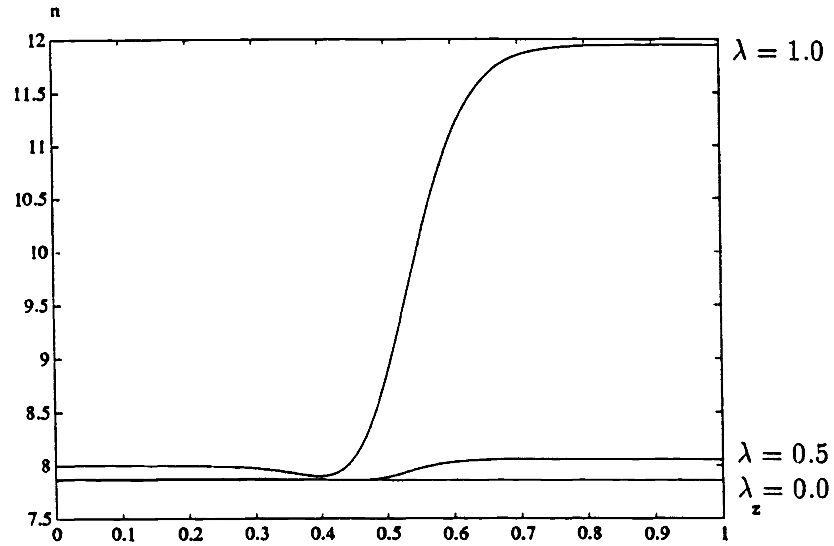


FIG. 7. *The projection onto the p-n plane of the solution for* $\lambda = 0.0,\ 0.5,$ *and* 1.0.

Fig. 6 we have plotted the n-component of the solution as a function of z for $\lambda = 0, \frac{1}{2}$, and 1. We see that n is clearly not monotone at $\lambda = 1$. In Fig. 7 we have plotted the projection of the solutions for $\lambda = 0, \frac{1}{2}$, and 1 onto the p-n phase plane.

Acknowledgment. Part of this work was performed while the first author was visiting the Institute for Applied Mathematics and Its Applications and the second author was visiting the University of Utah. Both authors are indebted to their host institutions for their support.

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion and nerve propagation*, in Partial Differential Equations and Related Topics, J. Goldstein, ed., Lecture Notes in Math. 446, Springer-Verlag, Berlin, New York, 1975, pp. 5–49.

[2] R. Casten, H. Cohen, and P. Lagerstrom, *Perturbation analysis of an approximation to the Hodgkin–Huxley theory*, Quart. Appl. Math., 32 (1975), pp. 365–402.

[3] C. Conley, *Isolated invariant sets and the Morse index*, AMS Reg. Conf. Ser. Math., Vol. 38, American Mathematical Society, Providence, RI, 1978.

[4] C. Conley and J. Smoller, *Isolated invariant sets of parameterized systems of differential equations*, in The Structure of Dynamical Systems, N. G. Markley, J. C. Martin, and W. Perrizo, eds., Lecture Notes in Math. 668, Springer-Verlag, Berlin, 1978.

[5] C. Conley and R. Gardner, *An application of the generalized Morse index to travelling wave solution of a competitive reaction–diffusion model*, Indiana Univ. Math. J., 33 (1984), pp. 319–343.

[6] J. Crow and M. Kimura, *An Introduction to Population Genetics Theory*, Burgess, Minneapolis, MN, 1970.

[7] E. J. Doedel and M. J. Friedman, *Numerical computation of heteroclinic orbits*, J. Comput. and Appl. Math., 26 (1989), pp. 159–170.

[8] E. J. Doedel and J. P. Kerevez, AUTO: *Software for continuation and bifurcation problems in ordinary differential equation*, Applied Mathematics Report, California Institute of Technology, Pasadena, CA, 1986.

[9] P. C. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer-Verlag, Berlin, New York, 1979.

[10] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[11] R. A. Fisher, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 355–369.

[12] R. Gardner, *Existence of travelling wave solutions of predator-prey systems via the connection index*, SIAM J. Appl. Math., 44 (1984), pp. 56–79.

[13] J. Roughgarden, *Density-dependent natural selection*, Ecology, 52 (1971), pp. 453–468.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 889-899, July 1992

ON THE ENERGY DECAY OF A LINEAR THERMOELASTIC BAR AND PLATE*

JONG UHN KIM†

Abstract. It is shown that the energy of a thermoelastic bar and plate decays exponentially fast. The energy method, combined with a multiplier technique and compactness property, is used.

Key words. energy decay, thermoelastic bar, thermoelastic plate

AMS(MOS) subject classifications. 35M05, 35B40

Introduction. In this paper we shall prove exponential decay of the energy of a one-dimensional linear thermoelastic bar and a linear thermoelastic plate. Since the pioneering work of Dafermos [1] on linear thermoelasticity, significant progress has been made on the mathematical aspect of thermoelasticity; see [4], [5], [8], [9], and [10], among others. Most studies focused on the existence, regularity, and asymptotic behavior of solutions to the equations of nonlinear thermoelasticity. Surprisingly, nothing has been known on exponential decay of the energy for the one-dimensional linear equations with the Dirichlet boundary condition. In the above cited works and references therein, some results on decay rate can be found. However, these are not in the form from which we can infer exponential decay. When the displacement and the temperature satisfy the Dirichlet and Neumann boundary conditions, respectively, or the other combination, Hansen [3] proved exponential decay by using nonharmonic Fourier series. But his argument does not seem to extend to the case where both the displacement and the temperature satisfy the Dirichlet boundary condition.

The purpose of the present work is to resolve this open question for the one-dimensional equations. The main results are Theorems 1.6 and 1.7 below. It is known that the energy in higher-dimensional thermoelasticity does not decay to zero under certain circumstances. There can exist nontrivial time-periodic solutions. A precise statement can be found in [1].

We shall also prove exponential decay of the energy of the linear thermoelastic plate in any space dimensions. Lagnese [6] discussed stabilization of various plate models, and showed that the energy of a linear thermoelastic plate decays exponentially fast with a certain dissipative boundary condition. We shall show that exponential decay can be achieved with the homogeneous Dirichlet boundary condition. The main results are Theorems 2.6 and 2.7.

Our main tool is the energy method, combined with a multiplier technique and compactness property. The general strategy of proof is the same for both a linear thermoelastic bar and plate. It is quite different from that of the earlier works, which also employed the energy method. Our approach is somewhat indirect in that the key estimates (1.32) and (2.32) below are established by the argument of contradiction, which fully exploits compactness property and inherent boundary regularity associated with the wave and plate equations.

1. Linear thermoelastic bar.

In this section, we let $\Omega = (0, 1)$ and consider the following initial-boundary value problem.

$$u_{tt} - au_{xx} + b\theta_x = 0 \quad \text{in } \Omega \times (0, \infty), \tag{1.1}$$

* Received by the editors November 26, 1990; accepted for publication (in revised form) November 13, 1991. This research, performed at Mathematical Sciences Research Institute, was supported in part by National Science Foundation grant DMS-8505550.

† Department of Mathematics, Virginia Polytechnic Institute, Blacksburg, Virginia 24061.

$$\theta_t - \theta_{xx} + bu_{tx} = 0 \quad \text{in } \Omega \times (0, \infty), \tag{1.2}$$

$$u = 0, \quad \theta = 0 \quad \text{at } x = 0 \text{ and } 1, \tag{1.3}$$

$$u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x), \quad \theta(x, 0) = \theta_0(x) \quad \text{in } \Omega, \tag{1.4}$$

where $a > 0$ and $b \neq 0$ are constants, and u and θ denote the displacement and the temperature, respectively. The derivation of (1.1) and (1.2) can be found in [2].

LEMMA 1.1. *For $u_0 \in H_0^1(\Omega)$, $u_1 \in L^2(\Omega)$, and $\theta_0 \in L^2(\Omega)$, there is a unique solution (u, θ) such that*

$$u \in C([0, \infty); H_0^1(\Omega)) \cap C^1([0, \infty); L^2(\Omega)), \tag{1.5}$$

$$\theta \in C([0, \infty); L^2(\Omega)) \cap L^2(0, \infty; H_0^1(\Omega)). \tag{1.6}$$

This is a known fact. A typical proof is to construct a sequence of smooth solutions whose initial data approximate (u_0, u_1, θ_0). Then, we apply the a priori estimates derived from the identity

$$\frac{d}{dt} \int_\Omega (u_t^2 + u_x^2 + \theta^2)\, dx + 2 \int_\Omega \theta_x^2\, dx = 0 \tag{1.7}$$

to the difference between any two smooth solutions of this sequence and conclude that the sequence is strongly convergent in the function spaces in (1.5) and (1.6). We shall omit the details.

LEMMA 1.2. *For $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$, $u_1 \in H_0^1(\Omega)$, and $\theta_0 \in H_0^1(\Omega) \cap H^2(\Omega)$, there is a unique solution such that*

$$u \in C([0, \infty); H_0^1(\Omega) \cap H^2(\Omega)) \cap C^1([0, \infty); H_0^1(\Omega)), \tag{1.8}$$

$$\theta \in C([0, \infty); H_0^1(\Omega) \cap H^2(\Omega)) \cap C^1([0, \infty); L^2(\Omega)), \tag{1.9}$$

$$\theta, \theta_t \in L^2(0, \infty; H_0^1(\Omega)). \tag{1.10}$$

Proof. This can be derived from Lemma 1.1 through an equivalent problem. Suppose that $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$, $u_1 \in H_0^1(\Omega)$, and $\theta_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ are given. Let (v, ϕ) be a solution of (1.1)-(1.3) and

$$v(x, 0) = v_0(x), \quad v_t(x, 0) = v_1(x), \quad \phi(x, 0) = \phi_0(x), \tag{1.11}$$

where v_0, v_1, and ϕ_0 are given by

$$v_0 = u_1, \tag{1.12}$$

$$v_1 = au_{0xx} - b\theta_{0x}, \tag{1.13}$$

$$\phi_0 = \theta_{0xx} - bu_{1x}. \tag{1.14}$$

The existence and uniqueness of (v, ϕ) follow from Lemma 1.1. Then, we set

$$u(x, t) = u_0(x) + \int_0^t v(x, s)\, ds, \tag{1.15}$$

$$\theta(x, t) = \theta_0(x) + \int_0^t \phi(x, s)\, ds. \tag{1.16}$$

It follows from the regularity conditions (1.5) and (1.6) applied to (v, ϕ) that

$$u \in C^1([0, \infty); H_0^1(\Omega)) \cap C^2([0, \infty); L^2(\Omega)), \tag{1.17}$$

$$\theta \in C^1([0, \infty); L^2(\Omega)) \cap C([0, \infty); H_0^1(\Omega)), \tag{1.18}$$

$$\theta_t \in L^2(0, \infty; H_0^1(\Omega). \tag{1.19}$$

By virtue of (1.12)-(1.14), (u, θ) defined by (1.15) and (1.16) also satisfies (1.1) and (1.2). By means of (1.17) and (1.18), we can infer from (1.1) and (1.2) that

$$u \in C([0,\infty); H_0^1(\Omega) \cap H^2(\Omega)), \tag{1.20}$$

$$\theta \in C([0,\infty); H_0^1(\Omega) \cap H^2(\Omega)). \tag{1.21}$$

The property that $\theta \in L^2(0,\infty; H_0^1(\Omega))$ follows directly from (1.7). Now the proof is complete.

For later use, we also need to consider the following initial-boundary value problem.

$$v_{tt} - av_{xx} + b\theta_{tx} = 0 \quad \text{in } \Omega \times (0,\infty), \tag{1.22}$$

$$\theta_t - \theta_{xx} + bv_x = 0 \quad \text{in } \Omega \times (0,\infty), \tag{1.23}$$

$$v = 0, \quad \theta = 0 \quad \text{at } x = 0 \text{ and } 1 \tag{1.24}$$

$$v(x,0) = v_0(x), \quad v_t(x,0) = v_1(x), \quad \theta(x,0) = \theta_0(x) \quad \textit{in } \Omega. \tag{1.25}$$

LEMMA 1.3. *For* $v_0 \in H_0^1(\Omega)$, $v_1 \in L^2(\Omega)$, *and* $\theta_0 \in H_0^1(\Omega) \cap H^2(\Omega)$, *there is a unique solution* (v, θ) *of* (1.22)-(1.25) *such that*

$$v \in C([0,\infty); H_0^1(\Omega)) \cap C^1([0,\infty); L^2(\Omega)), \tag{1.26}$$

$$\theta \in C([0,\infty): H_0^1(\Omega) \cap H^2(\Omega)) \cap C^1([0,\infty); L^2(\Omega)), \tag{1.27}$$

$$\theta, \theta_t \in L^2(0,\infty; H_0^1(\Omega)). \tag{1.28}$$

Proof. For given $v_0 \in H_0^1(\Omega)$, $v_1 \in L^2(\Omega)$, and $\theta_0 \in H_0^1(\Omega) \cap H^2(\Omega)$, we can determine $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ uniquely from

$$au_{0xx} = v_1 + b\theta_{0x}. \tag{1.29}$$

Then, let (u, θ) be a solution in Lemma 1.2 with

$$u(x,0) = u_0, \quad u_t(x,0) = v_0, \quad \theta(x,0) = \theta_0, \tag{1.30}$$

and set $v = u_t$. Obviously, (v, θ) is a solution of (1.22)-(1.25) satisfying (1.26)-(1.28). Uniqueness follows from the identity

$$\frac{d}{dt}\int_\Omega (v_t^2 + av_x^2 + \theta_t^2)\,dx + 2\int_\Omega \theta_{tx}^2\,dx = 0. \tag{1.31}$$

LEMMA 1.4. *Let* (v, θ) *be a solution of* (1.22)-(1.25) *with* $v_0 \in H_0^1(\Omega)$, $v_1 \in L^2(\Omega)$, *and* $\theta_0 \in H_0^1(\Omega) \cap H^2(\Omega)$, *and choose any* $T > 0$. *Then, it holds that*

$$\int_0^T \int_\Omega \theta_{tx}^2\,dx\,dt \geqq M(\|v_0\|_{H_0^1(\Omega)}^2 + \|v_1\|_{L^2(\Omega)}^2 + \|\theta_0\|_{H^2(\Omega)}^2) \tag{1.32}$$

for some positive constant M independent of v_0, v_1, *and* θ_0.

Proof. Assume that (1.32) is false. Then, there are sequences $\{v_0^m\}_{m=1}^\infty \subset H_0^1(\Omega)$, $\{v_1^m\}_{m=1}^\infty \subset L^2(\Omega)$, and $\{\theta_0^m\}_{m=1}^\infty \subset H_0^1(\Omega) \cap H^2(\Omega)$ such that

$$\|v_0^m\|_{H_0^1(\Omega)}^2 + \|v_1^m\|_{L^2(\Omega)}^2 + \|\theta_0^m\|_{H^2(\Omega)}^2 = 1 \quad \text{for each } m, \tag{1.33}$$

$$\int_0^T \int_\Omega (\theta_{tx}^m)^2\,dx\,dt \to 0 \quad \text{as } m \to \infty, \tag{1.34}$$

$$v_0^m \to v_0^\infty \text{ weakly} \quad \text{in } H_0^1(\Omega) \quad \text{as } m \to \infty, \tag{1.35}$$

$$v_1^m \to v_1^\infty \text{ weakly} \quad \text{in } L^2(\Omega) \quad \text{as } m \to \infty, \tag{1.36}$$

$$\theta_0^m \to \theta_0^\infty \text{ weakly} \quad \text{in } H_0^1(\Omega) \cap H^2(\Omega) \quad \text{as } m \to \infty \tag{1.37}$$

for some $v_0^\infty \in H_0^1(\Omega)$, $v_1^\infty \in L^2(\Omega)$, and $\theta_0^\infty \in H_0^1(\Omega) \cap H^2(\Omega)$. Here (v^m, θ^m) denotes a solution of (1.22), (1.23), (1.24), and

$$v^m(x,0) = v_0^m, \quad v_t^m(x,0) = v_1^m, \quad \theta^m(x,0) = \theta_0^m. \tag{1.38}$$

Similarly, $(v^\infty, \theta^\infty)$ is a solution with

$$v^\infty(x,0) = v_0^\infty, \quad v_t^\infty(x,0) = v_1^\infty, \quad \theta^\infty(x,0) = \theta_0^\infty. \tag{1.39}$$

It follows from a priori estimates that can be deduced from (1.31), (1.35)-(1.37) that

$$v^m \to v^\infty \text{ weak * } \quad \text{in } L^\infty(0, T; H_0^1(\Omega)), \tag{1.40}$$

$$v_t^m \to v_t^\infty \text{ weak * } \quad \text{in } L^\infty(0, T; L^2(\Omega)), \tag{1.41}$$

$$\theta^m \to \theta^\infty \text{ weak * } \quad \text{in } L^\infty(0, T; H_0^1(\Omega) \cap H^2(\Omega)), \tag{1.42}$$

$$\theta_t^m \to \theta_t^\infty \text{ weak * } \quad \text{in } L^\infty(0, T; L^2(\Omega)), \tag{1.43}$$

$$\theta_t^m \to \theta_t^\infty \text{ weakly } \quad \text{in } L^2(0, T; H_0^1(\Omega)). \tag{1.44}$$

Since (1.34) implies that

$$\theta_{tx}^\infty = 0 \quad \text{in } \Omega \times (0, T) \tag{1.45}$$

and $(v^\infty, \theta^\infty)$ satisfies

$$v_{tt}^\infty - a v_{xx}^\infty + b\theta_{tx}^\infty = 0 \quad \text{in } \Omega \times (0, T), \tag{1.46}$$

$$\theta_t^\infty - \theta_{xx}^\infty + b v_x^\infty = 0 \quad \text{in } \Omega \times (0, T), \tag{1.47}$$

we find that v_x^∞ is independent of t. Since $v^\infty \in C([0, T]; H_0^1(\Omega))$, v^∞ is also independent of t. Consequently, (1.46) yields

$$v^\infty = 0 \quad \text{in } \Omega \times (0, T), \tag{1.48}$$

which, together with (1.47), implies

$$\theta^\infty = 0 \quad \text{in } \Omega \times (0, T). \tag{1.49}$$

Next, let us rewrite (1.22) and(1.23) in (v^m, θ^m):

$$v_{tt}^m - a v_{xx}^m + b\theta_{tx}^m = 0 \quad \text{in } \Omega \times (0, T), \tag{1.50}$$

$$\theta_t^m - \theta_{xx}^m + b v_x^m = 0 \quad \text{in } \Omega \times (0, T). \tag{1.51}$$

Combining (1.40)-(1.43) and (1.50), we find that for any $\varepsilon > 0$,

$$v^m \to 0 \text{ strongly } \quad \text{in } C([0, T]; H^{1-\varepsilon}(\Omega)), \tag{1.52}$$

$$v_t^m \to 0 \text{ strongly } \quad \text{in } C([0, T]; H^{-\varepsilon}(\Omega)), \tag{1.53}$$

$$\theta^m \to 0 \text{ strongly } \quad \text{in } C([0, T]; H^{2-\varepsilon}(\Omega)). \tag{1.54}$$

Since it holds that

$$\|v^m\|_{C([0, T]; H_0^1(\Omega))} + \|v_t^m\|_{C([0, T]; L^2(\Omega))} + \|\theta_{tx}^m\|_{L^2(0, T; L^2(\Omega))} \leqq M \tag{1.55}$$

for all m, where M is a constant, we obtain

$$\|v_x^m(0, t)\|_{L^2(0, T)} + \|v_x^m(1, t)\|_{L^2(0, T)} \leqq M, \tag{1.56}$$

where M is a constant independent of m. This is a well-known fact that can be proved by a multiplier technique; see [7]. Next, we multiply (1.50) by θ_x^m and integrate over $\Omega\times(0, T)$ to obtain

$$\begin{aligned}&\int_\Omega v_t^m(x, T)\theta_x^m(x, T)\,dx-\int_\Omega v_t^m(x,0)\theta_x^m(x,0)\,dx-\int_0^T\int_\Omega v_t^m\theta_{xt}^m\,dx\,dt\\ &-a\int_0^T v_x^m(1,t)\theta_x^m(1,t)\,dt+a\int_0^T v_x^m(0,t)\theta_x^m(0,t)\,dt\\ &+a\int_0^T\int_\Omega v_x^m\theta_{xx}^m\,dx\,dt+b\int_0^T\int_\Omega \theta_{xt}^m\theta_x^m\,dx\,dt=0\quad\text{for each } m.\end{aligned}\tag{1.57}$$

By virtue of (1.34), (1.53), (1.54), and (1.56), we derive from (1.57) that

$$\int_0^T\int_\Omega v_x^m\theta_{xx}^m\,dx\,dt\to 0\quad\text{as } m\to\infty.\tag{1.58}$$

Multiplying (1.51) by θ_{xx}^m and integrating over $\Omega\times(0, T)$, we get

$$\int_0^T\int_\Omega \theta_t^m\theta_{xx}^m\,dx\,dt-\int_0^T\int_\Omega(\theta_{xx}^m)^2\,dx\,dt+b\int_0^T\int_\Omega v_x^m\theta_{xx}^m\,dx\,dt=0\quad\text{for each } m.\tag{1.59}$$

On account of (1.34), (1.42), and (1.58), we have

$$\int_0^T\int_\Omega(\theta_{xx}^m)^2\,dx\,dt\to 0\quad\text{as } m\to\infty,\tag{1.60}$$

which, combined with (1.34) and (1.51), yields

$$\int_0^T\int_\Omega(v_x^m)^2\,dx\,dt\to 0\quad\text{as } m\to\infty.\tag{1.61}$$

We then multiply (1.50) by v^m, integrate over $\Omega\times(0, T)$ and use (1.34), (1.52), (1.53), and (1.61) to find that

$$\int_0^T\int_\Omega(v_t^m)^2\,dx\,dt\to 0\quad\text{as } m\to\infty.\tag{1.62}$$

Finally, we set

$$E^m(t)=\frac{1}{2}\int_\Omega\{(v_t^m)^2+a(v_x^m)^2+(\theta_t^m)^2\}\,dx.\tag{1.63}$$

Then, it follows from (1.50) and (1.51) that

$$E^m(s)-E^m(0)=-\int_0^s\int_\Omega(\theta_{xt}^m)^2\,dx\,dt,\tag{1.64}$$

from which it follows that

$$TE^m(0)-T\int_0^T\int_\Omega(\theta_{xt}^m)^2\,dx\,dt\leqq\int_0^T E^m(t)\,dt.\tag{1.65}$$

By virtue of (1.34), (1.61), and (1.62), we conclude that

$$E^m(0)\to 0\quad\text{as } m\to\infty.\tag{1.66}$$

This contradicts (1.33) because

$$\theta_t^m(x,0)=\theta_{0xx}^m-bv_{0x}^m. \tag{1.67}$$

Now the proof of (1.32) is complete.

LEMMA 1.5. *Let* (v,θ) *be a solution of* (1.22)-(1.25) *with* $v_0\in H_0^1(\Omega)$, $v_1\in L^2(\Omega)$, *and* $\theta_0\in H_0^1(\Omega)\cap H^2(\Omega)$. *Then, it holds that*

$$\begin{aligned}&\|v(t)\|_{H_0^1(\Omega)}+\|v_t(t)\|_{L^2(\Omega)}+\|\theta(t)\|_{H^2(\Omega)}\\&\quad\leqq M\exp(-\alpha t)(\|v_0\|_{H_0^1(\Omega)}+\|v_1\|_{L^2(\Omega)}+\|\theta_0\|_{H^2(\Omega)})\end{aligned} \tag{1.68}$$

for some positive constants M *and* α *independent of* v_0, v_1, *and* θ_0.

Proof. Let us define

$$E_1(t)=\int_\Omega(v_t^2+av_x^2+\theta_t^2)\,dx, \tag{1.69}$$

$$E_2(t)=\int_\Omega(v_t^2+av_x^2+\theta_{xx}^2)\,dx. \tag{1.70}$$

Then, by (1.23), there are positive constants β_1 and β_2 independent of t and (v,θ) such that

$$\beta_1E_1(t)\leqq E_2(t)\leqq\beta_2E_1(t). \tag{1.71}$$

Fix any $T>0$. Then, (1.32) and (1.71) give

$$E_1(T)-E_1(0)=-2\int_0^T\int_\Omega\theta_{xt}^2\,dx\,dt\leqq -cE_1(0) \tag{1.72}$$

for some positive constant c independent of (v,θ). Without loss of generality, we may assume $c<1$, so that

$$E_1(T)\leqq\varepsilon E_1(0)\quad\text{for } 0<\varepsilon<1. \tag{1.73}$$

This implies (1.68) by the semigroup property of solution.

THEOREM 1.6. *Let* (u,θ) *be a solution of* (1.1)-(1.4) *with* $u_0\in H_0^1(\Omega)\cap H^2(\Omega)$, $u_1\in H_0^1(\Omega)$, *and* $\theta_0\in H_0^1(\Omega)\cap H^2(\Omega)$. *Then, we have*

$$\begin{aligned}&\|u(t)\|_{H^2(\Omega)}+\|u_t(t)\|_{H_0^1(\Omega)}+\|\theta(t)\|_{H^2(\Omega)}\\&\quad\leqq M\exp(-\alpha t)(\|u_0\|_{H^2(\Omega)}+\|u_1\|_{H_0^1(\Omega)}+\|\theta_0\|_{H^2(\Omega)})\end{aligned} \tag{1.74}$$

for all $t\geqq 0$, *with some positive constants* M *and* α *independent of* u_0, u_1, *and* θ_0.

Proof. It is enough to set $v=u_t$ and use Lemma 1.5.

THEOREM 1.7. *Let* (u,θ) *be a solution of* (1.1)-(1.4) *with* $u_0\in H_0^1(\Omega)$, $u_1\in L^2(\Omega)$, *and* $\theta_0\in L^2(\Omega)$. *Then, it holds that*

$$\begin{aligned}&\|u(t)\|_{H_0^1(\Omega)}+\|u_t(t)\|_{L^2(\Omega)}+\|\theta(t)\|_{L^2(\Omega)}\\&\quad\leqq M\exp(-\alpha t)(\|u_0\|_{H_0^1(\Omega)}+\|u_1\|_{L^2(\Omega)}+\|\theta_0\|_{L^2(\Omega)})\end{aligned} \tag{1.75}$$

for all $t\geqq 0$, *with some positive constants* M *and* α *independent of* u_0, u_1, *and* θ_0.

Proof. We determine $U_0\in H_0^1(\Omega)\cap H^2(\Omega)$, $U_1\in H_0^1(\Omega)$, and $\Theta_0\in H_0^1(\Omega)\cap H^2(\Omega)$ by

$$U_1=u_0, \tag{1.76}$$

$$\Theta_{0xx}=\theta_0+bU_{1x}, \tag{1.77}$$

$$aU_{0xx}=u_1+b\Theta_{0x}. \tag{1.78}$$

Then, we denote by(U, Θ) a solution of (1.1), (1.2), (1.3), and

$$U(x,0)=U_0, \quad U_t(x,0)=U_1, \quad \Theta(x,0)=\Theta_0. \tag{1.79}$$

Then, apply Theorem 1.6 to (U, Θ). Since $u = U_t$ and $\theta = \Theta_t$, the assertion (1.75) follows.

2. Linear thermoelastic plate. Let Ω be a bounded open subset of R^n with smooth boundary $\partial\Omega$. We consider the following initial-boundary value problem.

$$u_{tt}+\Delta^2 u+\alpha\Delta\theta=0 \quad \text{in } \Omega\times(0,\infty), \tag{2.1}$$

$$\theta_t-\beta\Delta\theta+\gamma\theta-\alpha\Delta u_t=0 \quad \text{in } \Omega\times(0,\infty), \tag{2.2}$$

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

$$u(x,0)=u_0(x), \quad u_t(x,0)=u_1(x), \quad \theta(x,0)=\theta_0(x) \quad \text{in } \Omega, \tag{2.4}$$

where $\partial/\partial\nu$ denotes the outward normal derivative on $\partial\Omega$. Here $\alpha\neq 0$, $\beta>0$, and $\gamma\geqq 0$ are constants, and u and θ denote vertical deflection of the plate and the temperature, respectively. The derivation of (2.1) and (2.2) can be found in [6]. Our purpose is to establish exponential decay of the energy, and the argument is the same as in the previous section except some minor technical details.

LEMMA 2.1. *For* $u_0\in H_0^2(\Omega)$, $u_1\in L^2(\Omega)$, *and* $\theta_0\in L^2(\Omega)$, *there is a unique solution* (u, θ) of (2.1)-(2.4) *such that*

$$u\in C([0,\infty); H_0^2(\Omega))\cap C^1([0,\infty); L^2(\Omega)), \tag{2.5}$$

$$\theta\in C([0,\infty); L^2(\Omega))\cap L^2(0,\infty; H_0^1(\Omega)). \tag{2.6}$$

This can be proven by constructing a sequence of smooth solutions whose initial data approximate (u_0, u_1, θ_0). Then, this sequence can be shown to be strongly convergent in the function spaces in (2.5) and (2.6) by means of the identity

$$\frac{d}{dt}\int_\Omega (u_t^2+(\Delta u)^2+\theta^2)\,dx+2\int_\Omega (\beta|\nabla\theta|^2+\gamma\theta^2)\,dx=0. \tag{2.7}$$

Since this is a well-known procedure, we omit the details.

LEMMA 2.2. *For* $u_0\in H_0^2(\Omega)\cap H^4(\Omega)$, $u_1\in H_0^2(\Omega)$, *and* $\theta_0\in H_0^1(\Omega)\cap H^2(\Omega)$, *there is a unique solution* (u, θ) *of* (2.1)-(2.4) *such that*

$$u\in C([0,\infty); H_0^2(\Omega)\cap H^4(\Omega))\cap C^1([0,\infty); H_0^2(\Omega)), \tag{2.8}$$

$$\theta\in C([0,\infty); H_0^1(\Omega)\cap H^2(\Omega))\cap C^1([0,\infty); L^2(\Omega)), \tag{2.9}$$

$$\theta, \theta_t\in L^2(0,\infty; H_0^1(\Omega)). \tag{2.10}$$

Proof. This is a well-known fact. But for our purposes later, we shall derive this from Lemma 2.1 as in the proof of Lemma 1.2. Let us set

$$v_0=u_1, \tag{2.11}$$

$$v_1=-\Delta^2 u_0-\alpha\Delta\theta_0, \tag{2.12}$$

$$\varphi_0=\beta\Delta\theta_0-\gamma\theta_0+\alpha\Delta u_1, \tag{2.13}$$

and let (v, φ) be a solution of (2.1)-(2.3), and

$$v(x,0)=v_0(x), \quad v_t(x,0)=v_1(x), \quad \varphi(x,0)=\varphi_0(x) \quad \text{in } \Omega, \tag{2.14}$$

according to Lemma 2.1. Then, we set

$$u(x, t) = u_0(x) + \int_0^t v(x, s)\, ds, \tag{2.15}$$

$$\theta(x, t) = \theta_0(x) + \int_0^t \varphi(x, s)\, ds. \tag{2.16}$$

By means of the regularity conditions (2.5) and (2.6) applied to (v, φ), we find that

$$u \in C^1([0, \infty); H_0^2(\Omega)) \cap C^2([0, \infty); L^2(\Omega)), \tag{2.17}$$

$$\theta \in C^1([0, \infty); L^2(\Omega)) \cap C([0, \infty); H_0^1(\Omega)), \tag{2.18}$$

$$\theta_t \in L^2(0, \infty; H_0^1(\Omega)). \tag{2.19}$$

By virtue of (2.11)-(2.13), it is evident that (u, θ) is also a solution of (2.1)-(2.4). Then it follows from (2.2), (2.17)-(2.19) that

$$\theta \in C([0, \infty); H_0^1(\Omega) \cap H^2(\Omega)), \tag{2.20}$$

which, combined with (2.1) and (2.17), yields

$$u \in C([0, \infty); H_0^2(\Omega) \cap H^4(\Omega)). \tag{2.21}$$

The condition that $\theta \in L^2(0, \infty; H_0^1(\Omega))$ follows directly from (2.7) and the proof is complete.

We next consider the following initial-boundary value problem:

$$v_{tt} + \Delta^2 v + \alpha \Delta \theta_t = 0 \quad \text{in } \Omega \times (0, \infty), \tag{2.22}$$

$$\theta_t - \beta \Delta \theta + \gamma \theta - \alpha \Delta v = 0 \quad \text{in } \Omega \times (0, \infty), \tag{2.23}$$

$$v = \frac{\partial v}{\partial \nu} = 0, \theta = 0 \quad \text{on } \partial\Omega \times (0, \infty), \tag{2.24}$$

$$v(x, 0) = v_0(x), \quad v_t(x, 0) = v_1(x), \quad \theta(x, 0) = \theta_0(x) \quad \text{in } \Omega. \tag{2.25}$$

LEMMA 2.3. *For* $v_0 \in H_0^2(\Omega)$, $v_1 \in L^2(\Omega)$, *and* $\theta_0 \in H_0^1(\Omega) \cap H^2(\Omega)$, *there is a unique solution of* (2.22)-(2.25) *such that*

$$v \in C([0, \infty); H_0^2(\Omega)) \cap C^1([0, \infty); L^2(\Omega)), \tag{2.26}$$

$$\theta \in C([0, \infty); H_0^1(\Omega) \cap H^2(\Omega)) \cap C^1([0, \infty); L^2(\Omega)), \tag{2.27}$$

$$\theta, \theta_t \in L^2(0, \infty; H_0^1(\Omega)). \tag{2.28}$$

Proof. Let us set

$$u_1 = v_0, \tag{2.29}$$

and determine $u_0 \in H_0^2(\Omega) \cap H^4(\Omega)$ by

$$\Delta^2 u_0 = -v_1 - \alpha \Delta \theta_0. \tag{2.30}$$

We then denote by (u, θ) a solution of (2.1)-(2.4) with u_0 and u_1 determined by (2.29) and (2.30). By setting $v = u_t$, (v, θ) is a solution of (2.22)-(2.25) satisfying (2.26)-(2.28).

Uniqueness follows from the identity

$$\frac{d}{dt}\int_\Omega (v_t^2+(\Delta v)^2+\theta_t^2)\,dx+2\int_\Omega (\beta|\nabla\theta_t|^2+\gamma\theta_t^2)\,dx=0. \tag{2.31}$$

LEMMA 2.4. *Let* (v, θ) *be a solution of* (2.22)-(2.25) *with* $v_0\in H_0^2(\Omega)$, $v_1\in L^2(\Omega)$, *and* $\theta_0\in H_0^1(\Omega)\cap H^2(\Omega)$. *Choose any* $T>0$. *Then, it holds that*

$$\int_0^T\int_\Omega |\nabla\theta_t|^2\,dx\,dt \geqq M(\|v_0\|^2_{H_0^2(\Omega)}+\|v_1\|^2_{L^2(\Omega)}+\|\theta_0\|^2_{H^2(\Omega)}) \tag{2.32}$$

for some positive constant M *independent of* v_0, v_1, *and* θ_0.

Proof. The method of proof is the same as that for Lemma 1.4. Assume that (2.32) is false. Then, there are sequences $\{v_0^m\}_{m=1}^\infty\subset H_0^2(\Omega)$, $\{v_1^m\}_{m=1}^\infty\subset L^2(\Omega)$, and $\{\theta_0^m\}_{m=1}^\infty\subset H_0^1(\Omega)\cap H^2(\Omega)$ such that

$$\|v_0^m\|^2_{H_0^2(\Omega)}+\|v_1^m\|^2_{L^2(\Omega)}+\|\theta_0^m\|^2_{H^2(\Omega)}=1 \quad \text{for all } m, \tag{2.33}$$

$$\int_0^T\int_\Omega |\nabla\theta_t^m|^2\,dx\,dt\to 0 \quad \text{as } m\to\infty, \tag{2.34}$$

$$v_0^m\to v_0^\infty \text{ weakly} \quad \text{in } H_0^2(\Omega), \tag{2.35}$$

$$v_1^m\to v_1^\infty \text{ weakly} \quad \text{in } L^2(\Omega), \tag{2.36}$$

$$\theta_0^m\to\theta_0^\infty \text{ weakly} \quad \text{in } H_0^1(\Omega)\cap H^2(\Omega) \tag{2.37}$$

for some $v_0^\infty\in H_0^2(\Omega)$, $v_1^\infty\in L^2(\Omega)$, and $\theta_0^\infty\in H_0^1(\Omega)\cap H^2(\Omega)$. Here (v^m, θ^m) denotes a solution of (2.22)-(2.24), and

$$v^m(x,0)=v_0^m(x),\quad v_t^m(x,0)=v_1^m(x),\quad \theta^m(x,0)=\theta_0^m(x) \quad \text{in } \Omega. \tag{2.38}$$

Similarly, (v^∞, u^∞) stands for a solution satisfying

$$v^\infty(x,0)=v_0^\infty(x),\quad v_t^\infty(x,0)=v_1^\infty(x),\quad \theta^\infty(x,0)=\theta_0^\infty(x) \quad \text{in } \Omega. \tag{2.39}$$

By virtue of a priori estimates that can be derived from (2.31), (2.35)-(2.37), we find that

$$v^m\to v^\infty \text{ weak* } \quad \text{in } L^\infty(0, T; H_0^2(\Omega)), \tag{2.40}$$

$$v_t^m\to v_t^\infty \text{ weak* } \quad \text{in } L^\infty(0, T; L^2(\Omega)), \tag{2.41}$$

$$\theta_t^m\to\theta_t^\infty \text{ weak* in } L^\infty(0, T; L^2(\Omega)), \tag{2.42}$$

$$\theta_t^m\to\theta_t^\infty \text{ weakly} \quad \text{in } L^2(0, T; H_0^1(\Omega)). \tag{2.43}$$

In order to obtain further estimates, we rewrite (2.22) and (2.23) in (v^m, θ^m):

$$v_{tt}^m+\Delta^2 v^m+\alpha\Delta\theta_t^m=0 \quad \text{in } \Omega\times(0, T), \tag{2.44}$$

$$\theta_t^m-\beta\Delta\theta^m+\gamma\theta^m-\alpha\Delta v^m=0 \quad \text{in } \Omega\times(0, T). \tag{2.45}$$

It follows from these equations that

$$v_{tt}^m\to v_{tt}^\infty \text{ weak* } \quad \text{in } L^\infty(0, T; H^{-2}(\Omega)), \tag{2.46}$$

$$\theta^m\to\theta^\infty \text{ weak* } \quad \text{in } L^\infty(0, T; H_0^1(\Omega)\cap H^2(\Omega)). \tag{2.47}$$

Next, let $h \in (h_1, \cdots, h_n)$ be a vector field such that $h \in [C^2(\bar{\Omega})]^n$ and $h(x)$ coincide with the outward unit normal vector on $\partial\Omega$. Then, multiplying (2.44) by $h \cdot \nabla v^m$ and integrating over $\Omega \times (0, T)$, we can obtain the identity

$$\begin{aligned}\frac{1}{2}\int_0^T \int_{\partial\Omega} (\Delta v^m)^2\,dx\,dt = &\int_\Omega v^m(x, T)h \cdot \nabla v^m(x, T)\,dx - \int_\Omega v^m(x, 0)h \cdot \nabla v^m(x, 0)\,dx \\ &+ \frac{1}{2}\int_0^T \int_\Omega (\nabla \cdot h)\{(v_t^m)^2 - (\Delta v^m)^2\}\,dx\,dt \\ &+ 2\int_0^T \int_\Omega \sum_{j,k=1}^n \frac{\partial h_k}{\partial x_j}\frac{\partial^2 v^m}{\partial x_k\,\partial x_j}\Delta v^m\,dx\,dt \\ &+ \int_0^T \int_\Omega (\Delta v^m)(\Delta h) \cdot \nabla v^m\,dx\,dt \\ &- \int_0^T \int_\Omega \alpha \nabla \theta_t^m \cdot \nabla(h \cdot \nabla v^m)\,dx\,dt. \end{aligned} \tag{2.48}$$

The proof of this can be found in [7, p. 244]. On account of (2.40)-(2.43), it is easy to see that

$$\int_0^T \int_{\partial\Omega} (\Delta v^m)^2\,dx\,dt \leqq M \tag{2.49}$$

for a constant M independent of m. In the meantime, (2.34) implies that

$$\nabla \theta_t^\infty = 0 \quad \text{in } \Omega \times (0, T). \tag{2.50}$$

By the same argument as in the previous section, we can conclude that

$$v^\infty = 0 \quad \text{and} \quad \theta^\infty = 0 \quad \text{in } \Omega \times (0, T). \tag{2.51}$$

Consequently, we combine (2.40)-(2.43), (2.46), and (2.47) to find that for any $\varepsilon > 0$,

$$v^m \to 0 \text{ strongly} \quad \text{in } C([0, T]; H^{2-\varepsilon}(\Omega)), \tag{2.52}$$

$$v_t^m \to 0 \text{ strongly} \quad \text{in } C([0, T]; H^{-\varepsilon}(\Omega)), \tag{2.53}$$

$$\theta^m \to 0 \text{ strongly} \quad \text{in } C([0, T]; H^{2-\varepsilon}(\Omega)). \tag{2.54}$$

We then multiply (2.44) by θ^m, integrate over $\Omega \times (0, T)$ and use (2.49), (2.52)-(2.54) to derive that

$$\int_0^T \int_\Omega (\Delta v^m)\Delta\theta^m\,dx\,dt \to 0 \quad \text{as } m \to \infty. \tag{2.55}$$

Next, multiplying (2.45) by $\Delta\theta^m$ and integrating over $\Omega \times (0, T)$, we use (2.55) to find that

$$\int_0^T \int_\Omega (\Delta\theta^m)^2\,dx\,dt \to 0 \quad \text{as } m \to \infty. \tag{2.56}$$

By the same argument as in the proof of Lemma 1.4, we can derive from (2.34), (2.44), (2.45), (2.56) that

$$\int_0^T \int_\Omega \{(v_t^m)^2 + (\Delta v^m)^2\}\,dx\,dt \to 0 \quad \text{as } m \to \infty. \tag{2.57}$$

By means of an inequality analogous to (1.65), we arrive at

$$\|v_0^m\|^2_{H_0^2(\Omega)}+\|v_1^m\|^2_{L^2(\Omega)}+\|\theta_0^m\|^2_{H^2(\Omega)}\to 0 \quad \text{as } m\to\infty, \tag{2.58}$$

which contradicts (2.33). Now the proof is complete.

LEMMA 2.5. *Let (v,θ) be a solution of* (2.22)-(2.25) *with $v_0\in H_0^2(\Omega)$, $v_1\in L^2(\Omega)$, and $\theta_0\in H_0^1(\Omega)\cap H^2(\Omega)$. Then, it holds that*

$$\begin{aligned}&\|v(t)\|_{H_0^2(\Omega)}+\|v_t(t)\|_{L^2(\Omega)}+\|\theta(t)\|_{H^2(\Omega)}\\&\leqq M\exp(-\alpha t)(\|v_0\|_{H_0^2(\Omega)}+\|v_1\|_{L^2(\Omega)}+\|\theta_0\|_{H^2(\Omega)})\end{aligned} \tag{2.59}$$

for some positive constants M and α independent of v_0, v_1, and θ_0.

As in the proof of Lemma 1.5, (2.59) follows from (2.32). We shall omit the details of the proof. Finally, we present the main results of this section. These can be proved exactly in the same way as Theorem 1.6 and 1.7.

THEOREM 2.6. *Let (u,θ) be a solution of* (2.1)-(2.4) *with $u_0\in H_0^2(\Omega)\cap H^4(\Omega)$, $u_1\in H_0^2(\Omega)$, and $\theta_0\in H_0^1(\Omega)\cap H^2(\Omega)$. Then, it holds that*

$$\begin{aligned}&\|u(t)\|_{H^4(\Omega)}+\|u_t(t)\|_{H_0^2(\Omega)}+\|\theta(t)\|_{H^2(\Omega)}\\&\leqq M\exp(-\alpha t)(\|u_0\|_{H^4(\Omega)}+\|u_1\|_{H_0^2(\Omega)}+\|\theta_0\|_{H^2(\Omega)})\end{aligned} \tag{2.60}$$

for all $t\geqq 0$, with some positive constants M and α independent of u_0, u_1, and θ_0.

THEOREM 2.7. *Let (u,θ) be a solution of* (2.1)-(2.4) *with $u_0\in H_0^2(\Omega)$, $u_1\in L^2(\Omega)$, and $\theta_0\in L^2(\Omega)$. Then, it holds that*

$$\begin{aligned}&\|u(t)\|_{H_0^2(\Omega)}+\|u_t(t)\|_{L^2(\Omega)}+\|\theta(t)\|_{L^2(\Omega)}\\&\leqq M\exp(-\alpha t)(\|u_0\|_{H_0^2(\Omega)}+\|u_1\|_{L^2(\Omega)}+\|\theta_0\|_{L^2(\Omega)})\end{aligned} \tag{2.61}$$

for all $t\geqq 0$, with some positive constants M and α independent of u_0, u_1, and θ_0.

Acknowledgment. The author would like to acknowledge the warm hospitality of the Mathematical Sciences Research Institute in Berkeley, where this research was performed.

REFERENCES

[1] C. DAFERMOS, *On the existence and the asymptotic stability of solutions to the equations of linear thermoelasticity*, Arch. Rational Mech. Anal., 29 (1968), pp. 241-271.

[2] W. DAY, *Heat Conduction within Linear Thermoelasticity*, Springer Tracts Nat. Philos., 30 (1985).

[3] S. HANSEN, *Exponential energy decay in a linear thermoelastic rod*, J. Math. Anal. Appl., to appear.

[4] W. HRUSA AND M. TARABEK, *On smooth solutions of the Cauchy problem in one-dimensional nonlinear thermoelasticity*, Quart. Appl. Math., 47 (1989), pp. 631-644.

[5] S. JIANG, *Global existence of smooth solutions in one-dimensional nonlinear thermoelasticity*, Proc. Roy. Soc. Edinburgh, 115 (1990), pp. 257-274.

[6] J. LAGNESE, *Boundary stabilization of thin plates*, SIAM Stud. Appl. Math., 10 (1989).

[7] J. L. LIONS, *Contrôlabilitè exacte perturbations et stabilisation de systèmes distribués*, Tome 1, Masson, Paris, 1988.

[8] G. PONCE AND R. RACKE, *Global existence of small solutions to the initial value problem for nonlinear thermoelasticity*, J. Differential Equations, 87 (1990), pp. 70-83.

[9] R. RACKE AND Y. SHIBATA, *Global smooth solutions and asymptotic stability in one-dimensional thermoelasticity*, Arch. Rational Mech. Anal., 116 (1992), pp. 1-34.

[10] M. SLEMROD, *Global existence, uniqueness and asymptotic stability of classical smooth solutions in one-dimensional non-linear thermoelasticity*, Arch. Rational Mech. Anal., 76 (1981), pp. 97-133.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 900-922, July 1992

006

SOLUTIONS TO THE CUBIC SCHRÖDINGER EQUATION BY THE INVERSE SCATTERING METHOD*

AMY COHEN† AND THOMAS KAPPELER‡

Abstract. Weak solutions to the cubic Schrödinger equation are constructed by the inverse scattering method for a large class of initial data u_0 such that $(1+|x|^\alpha)u_0 \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and $u_0 \in H^\alpha(\mathbb{R})$ for an α with $\frac{1}{4} < \alpha < \frac{1}{2}$. These solutions are shown to evolve in $L^2(\mathbb{R}) \cap L^4(\mathbb{R})$. This construction is valid, in particular, if the initial data is the characteristic function on an interval of length not an odd multiple of $\pi/2$.

Key words. cubic Schrödinger equation, inverse scattering method

AMS(MOS) subject classification. 35Q20

1. Introduction and summary. We consider the initial value problem for the cubic Schrödinger equation

$$iu_t + \tfrac{1}{2}u_{xx} + |u|^2 u = 0 \quad \text{for } x \in \mathbb{R}, \tag{1.1}$$

$$u(x, 0) = u_0(x). \tag{1.2}$$

This initial value problem has been considered by many authors, for example [GV], [HNT1, 2], [K], [T1, 2], to cite only a few. Recently in [T2], Tsutsumi proved by functional analytic methods that for $u_0 \in L^2(\mathbb{R})$ there exists a solution $u(x, t)$ of (1.1), (1.2) evolving in $L^\infty([0, T], L^2(\mathbb{R}) \cap L^4(\mathbb{R}))$. It follows from results of Kato [K] that this solution is unique in this space.

On the other hand, Zakharov and Shabat [ZS] developed a representation theorem for smooth solutions of (1.1), (1.2) by the Marachenko-type inverse scattering method. Tanaka [Ta] used their method to construct solutions evolving in Schwartz class, provided that u_0 was of Schwartz class.

The purpose of this paper is to generalize Tanaka's results in order to construct, by the inverse scattering method, a large class of the solutions obtained by Tsutsumi [T2]. The inverse scattering representation is important because it reveals the soliton structure of the solutions and provides a method for the rigorous analysis of the long-time asymptotics of the solutions. We intend in a subsequent paper to analyze, in particular, the asymptotics of solutions evolving from box-shaped initial potentials.

In this paper a continuous map $t \to u(\cdot, t)$ from $[0, T]$ into L^2_{loc} is said to be a *weak solution* of (1.1), (1.2) if for all φ in $C^\infty([0, T] \times \mathbb{R})$, with compact support in $(0, T) \times \mathbb{R}$,

$$\int_0^\infty \int_{-\infty}^\infty u(x, t)\{i\varphi_t(x, t) + \tfrac{1}{2}\varphi_{xx}(x, t) + |u(x, t)|^2 \varphi(x, t)\}\, dx\, dt = 0, \tag{1.3}$$

and if

$$u(\cdot, t) \to u_0(\cdot) \quad \text{in } L^2_{\text{loc}}(\mathbb{R}) \quad \text{as } t \to 0^+.$$

In this paper we prove the following.

THEOREM. *Suppose there is an* α *with* $\frac{1}{4} < \alpha < \frac{1}{2}$ *such that*

$$(1+|x|^\alpha)u_0(x) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}),$$

* Received by the editors July 30, 1990; accepted for publication (in revised form) October 10, 1991.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The research of this author was partially supported by the National Science Foundation.

$u_0 \in$ the Sobolev space H^α, and u_0 meets the technical assumption (2.12). *Then the inverse scattering method produces a weak solution of* (1.1), (1.2) *evolving in* $L^2(\mathbb{R}) \cap L^4(\mathbb{R})$.

Remark. It follows from the uniqueness theorem of Kato [K] that the constructed solution is unique within the space $L^2(\mathbb{R}) \cap L^4(\mathbb{R})$, and coincides with the solution constructed by Tsutsumi [T2]. In particular it follows that

$$\int_{-\infty}^{\infty} |u(x,t)|^2 \, dx$$

is preserved in time.

The method of analysis applied by Zakharov and Shabat [ZS] to (1.1), (1.2) is analogous to the inverse scattering analysis introduced by Gardner, Greene, Kruskal, and Miura [GGKM] for the Korteweg-de Vries equation. Zakharov and Shabat associate (1.1) with the scattering problem

$$i\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\frac{d\vec{\psi}}{dx} - i\begin{bmatrix}0 & u\\ u^* & 0\end{bmatrix}\vec{\psi} = \zeta\vec{\psi}. \tag{1.4}$$

They showed that if $u(x,t)$ solves (1.1) and $u(\cdot,t)$ is of Schwartz class for each t, then the "scattering data" of (1.4) evolve according to certain linear first-order ordinary differential equations in t. They showed further that for $t>0$, $u(x,t)$ could be represented in terms of the scattering data at time t, and thus in terms of the initial scattering data which came from $u(x,t)$ at $t=0$. In particular,

$$u(x,t) = -B_{+1}(x,0,t),$$

where $B_{+1}(x,y,t)$ solves the Marchenko equation

$$\begin{aligned} &B_{+1}(x,y,t) + \int_0^\infty \Omega_+^*(x+y+z,t)\int_0^\infty \Omega_+(x+z+w,t)B_{+1}(x,w,t)\,dw\,dz \\ &\qquad + \Omega_+^*(x+y,t) = 0 \quad \text{for } y \geqq 0. \end{aligned} \tag{1.5}$$

The kernel $\Omega_+(s,t)$ is the sum of the inverse Fourier transform of the reflection coefficient at time t, and some linear combinations of products of the form $s^\nu e^{-2i\mathrm{Im}(\eta_j)s}$, where the η_j are the poles of the transmission coefficient.

Let us also point out that it would be possible to use the reformulation of inverse scattering due to Beals and Coifman [BC1], [BC2] in terms of the ∂-bar operator and the Riemann-Hilbert problem. While they consider the case of simple poles in the transmission coefficient, their formalism has been extended subsequently to higher order poles, cf. [SZ].

In § 2 we analyze the forward scattering theory of (1.4) for general $u = u_0 \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, introducing the necessary added hypothesis as (2.12). The assumption (2.12) is needed to guarantee that the transmission coefficient as defined in § 2 has a finite number of poles and that none of them are on the real axis. In § 3 we discuss the time evolution of "scattering data at $t>0$." The definitions are suggested by Zakharov and Shabat [ZS] and by Tanaka [Ta]. In § 4 we solve the Marchenko equation (1.5) and begin to analyze $u(x,t) := -B_{1+}(x,0,t)$. In § 5 we treat the inverse scattering problem on the full line, and show that, under the assumption of the theorem, u evolves in $L^2 \cap L^4$. In § 6 we show that this $u(x,t)$ is a weak solution of (1.1), (1.2). In § 7 we compute the scattering data associated with a u_0 in (1.4), where u_0 is the characteristic function of an interval and analyze the solution to (1.1), (1.2) in that case.

Notational conventions. $\|\cdot\|$ denotes the norm in $L^2(\mathbb{R}^+)$; $\|\cdot\|_{\mathrm{op}}$ the operator norm on $L^2(\mathbb{R}^+)$. Other norms are noted explicitly, e.g., $\|f\|_{L^2(\mathbb{R})}$ or $\|f\|_{L^\infty(\mathbb{R}^+)}$.

$L^p(+\infty)=\{f: f\in L^p([a,\infty))$ for all finite $a\}$,

$C^b(\mathbb{R})=\{f: f$ is a bounded continuous function on $\mathbb{R}\}$,

H^{2+} is the Hardy space; $\{f\in L^2(\mathbb{R})\colon \mathscr{F}[f]$ has support in $\mathbb{R}^+\}$.

The Fourier transform $\mathscr{F}$ and its inverse $\mathscr{F}^{-1}$ are taken in the form

$$\mathscr{F}[g](\xi)=\int_{-\infty}^{\infty} g(x)\,e^{-2i\xi x}\,dx \quad\text{and}\quad \mathscr{F}^{-1}[g](x)=\frac{1}{\pi}\int_{-\infty}^{\infty} g(\xi)\,e^{+2i\xi x}\,d\xi.$$

2. The forward scattering problem. The scattering problem associated with the cubic Schrödinger equation is

$$i\begin{bmatrix}1&0\\0&-1\end{bmatrix}\frac{d\vec{\Psi}}{dx}-i\begin{bmatrix}0&u\\u^*&0\end{bmatrix}\vec{\Psi}=\zeta\vec{\Psi}, \tag{2.1}$$

where $\vec{\Psi}$ has components Ψ_1, Ψ_2, and $\zeta\in\mathbb{C}$. In general, we assume only that

$$u\in L^1(\mathbb{R})\cap L^2(\mathbb{R}). \tag{2.2}$$

The purpose of this section is to define the Jost functions and the scattering data of (2.1) and to obtain the fundamental integral equation relating them.

The Jost functions for (2.1) are the solutions $\vec{\Psi}_+$ and $\vec{\Psi}_-$ of (2.1) for $\operatorname{Im}(\zeta)>0$ such that

$$\vec{\Psi}_+(x,\zeta)\sim\begin{bmatrix}0\\1\end{bmatrix}e^{i\zeta x}\quad\text{as } x\to+\infty,\quad\text{and } \vec{\Psi}_-(x,\zeta)\sim\begin{bmatrix}1\\0\end{bmatrix}e^{-i\zeta x}\quad\text{as } x\to-\infty.$$

The existence of the Jost functions for all ζ with $\operatorname{Im}(\zeta)\geqq 0$ and their key properties are established in the Lemmas 2.1-2.7 below.

The scattering coefficients, $a_+(\xi)$ and $b_+(\xi)$, are defined for real ξ by

$$\vec{\Psi}_-(x,\xi)=a_+(\xi)\vec{\Psi}_+^{\#}(x,\xi)+b_+(\xi)\vec{\Psi}_+(x,\xi), \tag{2.3}$$

where for any 2-vector $\vec{v}$, $v^{\#}$ denotes the transpose of $[v_2^*,-v_1^*]$. Similarly, the coefficients a_- and b_- are determined by the relation

$$\vec{\Psi}_+(x,\xi)=a_-(\xi)\vec{\Psi}_-^{\#}(x,\xi)+b_-(\xi)\vec{\Psi}_-(x,\xi).$$

The formal reflection and transmission coefficients, $r_+(\xi)$ and $t_+(\xi)$, are defined by

$$r_+(\xi)=b_+(\xi)/a_+(\xi)\quad\text{and}\quad t_+(\xi)=1/a_+(\xi).$$

In Theorem 2.11 we will prove that if $u\in L^1(\mathbb{R})\cap L^2(\mathbb{R})$ and u meets (2.12), then $r_+\in L^2(\mathbb{R})\cap C^b(\mathbb{R})$ and $r_+(\xi)\to 0$ as $\xi\to\pm\infty$. This theorem will also tell how weak regularity of u gives some decay in r_+ and how decay in u gives some weak regularity in r_+. Next, the analysis of the zeros of a_+ leads to the definition of the full set of scattering data associated to u. Finally, we derive the fundamental integral equations, or Marchenko equations, which relate the Jost functions and the scattering data.

Construction of the Jost functions. For $\operatorname{Im}(\zeta)\geqq 0$, $\vec{\Psi}_+$ must satisfy the system

$$\begin{aligned} i\partial_x\Psi_{+1}(x,\zeta)-iu(x)\Psi_{+2}(x,\zeta)&=\zeta\Psi_{+1}(x,\zeta),\\ -i\partial_x\Psi_{+2}(x,\zeta)-iu^*(x)\Psi_{+1}(x,\zeta)&=\zeta\Psi_{+2}(x,\zeta), \end{aligned} \tag{2.4}$$

with the boundary conditions

$$\Psi_{+1}(x, \zeta) \sim 0 \quad \text{as } x \to +\infty,$$
$$\Psi_{+2}(x, \zeta) \sim e^{i\zeta x} \quad \text{as } x \to +\infty.$$

Similarly, $\vec{\Psi}_-$ must satisfy the system

$$\begin{aligned} i\partial_x \Psi_{-1}(x, \zeta) - iu(x)\Psi_{-2}(x, \zeta) &= \zeta \Psi_{-1}(x, \zeta), \\ -i\partial_x \Psi_{-2}(x, \zeta) - iu^*(x)\Psi_{-1}(x, \zeta) &= \zeta \Psi_{-2}(x, \zeta), \end{aligned} \tag{2.5}$$

with the boundary conditions

$$\Psi_{-1}(x, \zeta) \sim e^{i\zeta x} \quad \text{as } x \to -\infty,$$
$$\Psi_{-2}(x, \zeta) \sim 0 \quad \text{as } x \to -\infty.$$

Now write

$$m_{+j}(x, \zeta) = e^{-i\zeta x}\Psi_{+j}(x, \zeta) \quad \text{for } j = 1, 2.$$

We get

$$\begin{aligned} m_{+1}(x, \zeta) = &-\int_{y=x}^{\infty} e^{2i\zeta(y-x)}u(y)\, dy \\ &-\int_{y=x}^{\infty} u(y)\, e^{2i\zeta(y-x)} \int_{z=y}^{\infty} m_{+1}(z, \zeta)u^*(z)\, dz\, dy, \end{aligned} \tag{2.6}$$

$$m_{+2}(x, \zeta) = 1 - \int_{y=x}^{\infty} u^*(y) \int_{z=y}^{\infty} m_{+2}(z, \zeta)\, e^{2i\zeta(z-y)}u(z)\, dz\, dy. \tag{2.7}$$

Similarly, putting

$$m_{-j}(x, \zeta) = e^{i\zeta x}\Psi_{-j}(x, \zeta) \quad \text{for } j = 1, 2,$$

we get

$$m_{-1}(x, \zeta) = 1 - \int_{y=-\infty}^{x} u(y) \int_{z=-\infty}^{y} u^*(z)m_{-1}(z, \zeta)\, e^{2i\zeta(y-z)}\, dz\, dy, \tag{2.8}$$

$$\begin{aligned} m_{-2}(x, \zeta) = &-\int_{y=-\infty}^{x} u^*(y)\, e^{2i\zeta(x-y)}\, dy \\ &-\int_{y=-\infty}^{x} u^*(y)\, e^{2i\zeta(x-y)} \int_{z=-\infty}^{y} u(z)m_{-2}(z, \zeta)\, dz\, dy. \end{aligned} \tag{2.9}$$

Because (2.8) and (2.9) are similar to (2.6) and (2.7), it suffices to study (2.6) and (2.7). Here we introduce the maps T_1 and T_2, defined by the formulas

$$T_1[g](x, \zeta) = \int_{y=x}^{\infty} u(y)\, e^{2i\zeta(y-x)} \int_{z=y}^{\infty} g(z, \zeta)u^*(z)\, dx\, dy,$$

$$T_2[g](x, \zeta) = \int_{y=x}^{\infty} u^*(y) \int_{z=y}^{\infty} g(z, \zeta)\, e^{2i\zeta(z-y)}u(z)\, dz\, dy,$$

in the appropriate spaces defined below. Let

$$\mathbb{H} := \{\zeta: \operatorname{Im}(\zeta) > 0\} \quad \text{and} \quad \bar{\mathbb{H}} := \{\zeta: \operatorname{Im}(\zeta) \geqq 0\}.$$

For each real x_0 let E_{x_0} denote the space of functions $f\colon [x_0, +\infty) \times \bar{\mathbb{H}} \to \mathbb{C}$ such that

(E.1) $x \geqq x_0, f(x, \cdot)$ is continuous on $\bar{\mathbb{H}}$ and analytic on $\mathbb{H}$;

(E.2) The map $x \mapsto f(x, \cdot)$ is bounded and continuous from $[x_0, +\infty)$ into the Hardy space H^{2+};

(E.3) The map $x \mapsto f(x, \cdot)$ is bounded and continuous from $[x_0, +\infty)$ into $C^b(\bar{\mathbb{H}})$;

(E.4) The map $\zeta \mapsto f(\cdot, \zeta)$ is bounded and continuous from $\bar{\mathbb{H}}$ into $C^b([x_0, +\infty))$;

(E.5) $f(x, \xi) \to 0$ uniformly in $x \geqq x_0$ as real $\xi \to \pm\infty$.

Now (2.6) and (2.7) take the form

$$(I + T_1)[m_{+1}](x, \zeta) = -\int_{y=x}^{\infty} e^{2i\zeta(y-x)} u(y)\, dy, \tag{2.10}$$

$$(I + T_2)[m_{+2}](x, \zeta) = 1. \tag{2.11}$$

Thus, formally,

$$m_{+1} = \sum_{n=0}^{\infty} (-T_1)^n [g_1], \quad \text{where } g_1(x, \zeta) := -\int_{y=x}^{\infty} e^{2i\zeta(y-x)} u(y)\, dy,$$

and

$$m_{+2} = \sum_{n=0}^{\infty} (-T_2)^n [g_2], \quad \text{where } g_2(x, \zeta) := 1.$$

The next lemmas show that m_{+1} and m_{+2} exist in E_{x_0}. Their proofs consist of extensive, but straightforward, analysis using the definitions of m_{+1} and m_{+2}.

LEMMA 2.1. *Keep x_0 fixed. Assume that $u \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Then*

(i) $T_2[g_2] \in E_{x_0}$.

(ii) *If u is also in the Sobolev space $H^\alpha(\mathbb{R})$ with $0 < \alpha < \frac{1}{2}$, then*

$$|\xi|^\alpha T_2[g_2](x, \xi) \in L^2(\mathbb{R})$$

for all $x \geqq x_0$, with L^2-norms uniformly bounded for $x \geqq x_0$.

(iii) *If $x^\alpha u(x) \in L^2(\mathbb{R})$ for $0 < \alpha < \frac{1}{2}$, then*

$$T_2[g_2](x, \cdot) \in H^\alpha(\mathbb{R})$$

for all $x \geqq x_0$, with H^α-norms uniformly bounded for $x \geqq x_0$.

LEMMA 2.2. *Recall that $g_1(x, \zeta) = \int_{y=x}^{\infty} e^{2i\zeta(y-x)} u(y)\, dy$. Assume that $u \in L^1 \cap L^2$.*

(i) *For each x_0, $g_1 \in E_{x_0}$.*

(ii) *If also $u \in H^\alpha$ with $0 < \alpha < \frac{1}{2}$, then $|\xi|^\alpha g_1(x, \xi) \in L^2(-\infty < \xi < \infty)$ uniformly in $x \geqq x_0$.*

(iii) *If $|s|^\alpha u(s) \in L^2$ with $0 < \alpha < \frac{1}{2}$, then $g_1(x, \xi) \in H^\alpha$ uniformly in $x \geqq x_0$.*

LEMMA 2.3. *Assume $u \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Pick x_0 in $\mathbb{R}$.*

(i) *If $f \in E_{x_0}$, then $T_2 f \in E_{x_0}$.*

(ii) *If also $|\xi|^\alpha f(x, \xi) \in L^2(-\infty < \xi < \infty)$ uniformly for $x \geqq x_0$, then $|\xi|^\alpha (T_2 f)(x, \xi) \in L^2(-\infty < \xi < \infty)$ uniformly for $x \geqq x_0$.*

(iii) *If also $f(x, \cdot) \in H^\alpha(\mathbb{R})$ uniformly for $x \geqq x_0$ and $x^\alpha u(x) \in L^1(\mathbb{R})$, and $0 < \alpha < 1$, then $T_2 f(x, \cdot) \in H^\alpha(\mathbb{R})$ uniformly for $x \geqq x_0$.*

LEMMA 2.4. *Fix x_0 in $\mathbb{R}$. Assume that $u \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.*

(i) *If $f \in E_{x_0}$, then $T_1(f) \in E_{x_0}$.*

(ii) *If also $\xi^\alpha f(x, \xi) \in L^2(\mathbb{R})$ uniformly in $x \geqq x_0$ and $0 < \alpha < \frac{1}{2}$, then $\xi^\alpha T_1(f)(x, \xi) \in L^2(\mathbb{R})$ uniformly in $x \geqq x_0$.*

(iii) *If also $f(x, \cdot) \in H^\alpha(\mathbb{R})$ uniformly in $x \geqq x_0$ and $x^\alpha u(x) \in L^1(\mathbb{R})$, then $T_1(f)(x, \cdot) \in H^\alpha(\mathbb{R})$ uniformly in $x \geqq x_0$.*

PROPOSITION 2.5. *Let $x_0 \in \mathbb{R}$. Assume that $u \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.*

(i) *There is a function $m_{+2}(x, \zeta)$ such that*

$$m_{+2}(x, \zeta) - 1 \in E_{x_0}, \quad \textit{and}$$

$$(I + T_2)[m_{+2} - 1] = -T_2[1],$$

whence m_{+2} solves (2.7) *and* (2.11).

(ii) *If, in addition, $u \in H^\alpha(\mathbb{R})$ with $0 < \alpha < \frac{1}{2}$, then for all $n \geqq 1$,*

$$\|\xi^\alpha T_2^n[1](x, \xi)\|_{L^2(\mathbb{R})} \leqq C\sigma^{2n-2}(x)/(2n-2)!,$$

where

$$C = \|u\|_{L^1(\mathbb{R})}\|u\|_{H^\alpha(\mathbb{R})} \quad \textit{and} \quad \sigma(x) = \int_x^\infty |u(s)|\, ds.$$

Thus

$$\|\xi^\alpha T_2[m_{+2}](x, \xi)\|_{L^2(\mathbb{R}, d\xi)} \leqq \sum_1^\infty \|\xi^\alpha T_2^n[1](x, \xi)\| \leqq C\, e^{\sigma(x)}.$$

(iii) *If $u \in H^\alpha$ and $(1 + |x|^\alpha)u(x) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ with $0 < \alpha < \frac{1}{2}$, then*

$$\|T_2^n[1](x, \cdot)\|_{H^\alpha} \leqq 2^n K \mu^{2n-1}(x)/(2n-1)!,$$

where

$$K = \|(1 + |y|^\alpha)u(y)\|_{L^2(\mathbb{R})} \quad \textit{and} \quad \mu(x) = \int_x^\infty (1 + |s|^\alpha)|u(s)|\, ds,$$

whence

$$\|T_2[m_{+2}](x, \xi)\|_{H^\alpha} \leqq \sum_{n \geqq 1} \|T_2^n[1](x, \xi)\|_{H^\alpha} \leqq 2^{3/2} K\, e^{\mu(x)}.$$

Remark 2.6. The analogous results hold for m_{+1} and T_1. These lemmas yield the following theorem.

THEOREM 2.7. *Assume $u \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Consider the functions $m_{+1}(x, \zeta)$ and $m_{+2}(x, \zeta)$ defined above. For any x_0 in $\mathbb{R}$*

(i) $T_1[m_{+1}](x, \zeta) \in E_{x_0}$; $\quad T_2[m_{+2}](x, \zeta) \in E_{x_0}$.

(ii) *If, in addition, $u \in H^\alpha(\mathbb{R})$ and $0 < \alpha < \frac{1}{2}$, then for each x with $x \geqq x_0$,*

$$\xi^\alpha T_1[m_{+1}](x, \xi) \in L^2(\mathbb{R}); \qquad \xi^\alpha T_2[m_{+2}](x, \xi) \in L^2(\mathbb{R})$$

with L^2-norms uniformly bounded for $x \geqq x_0$.

(iii) *If, in addition, $u \in H^\alpha$ and $x^\alpha u(x) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ with $0 < \alpha < \frac{1}{2}$, then*

$$T_1[m_{+1}](x, \xi) \in H^\alpha(\mathbb{R}); \qquad T_2[m_{+2}](x, \xi) \in H^\alpha(\mathbb{R})$$

with H^α-norms uniformly bounded in $x \geqq x_0$.

Similar arguments are used to construct the left-side Jost functions $\psi_{-1}(x, \zeta)$ and $\psi_{-2}(x, \zeta)$. By analogy with E_{x_0} we define spaces $E_{x_0}^-$ on left half-lines and operators

$$T_{-1}[f](x, \zeta) := \int_{-\infty}^x u(y) \int_{-\infty}^y u^*(z) f(z, \zeta)\, e^{2i\zeta(y-z)}\, dz\, dy,$$

$$T_{-2}[f](x, \zeta) := \int_{-\infty}^x u^*(y)\, e^{2i\zeta(x-y)} \int_{-\infty}^y u(z) f(z, \zeta)\, dz\, dy.$$

We then construct the solutions m_{-1} and m_{-2} of the equations

$$(I+T_{-1})m_{-1}(x,\zeta)=1,$$

$$(I+T_{-2})m_{-2}(x,\zeta)=-\int_{-\infty}^{x} u^*(y)\, e^{2i\zeta(x-y)}\, dy.$$

This finally gives the Jost function $\psi_-(x,\zeta)$ the form

$$\psi_-(x,\zeta)=\begin{bmatrix}\psi_{-1}(x,\zeta)\\ \psi_{-2}(x,\zeta)\end{bmatrix}=e^{i\zeta x}\begin{bmatrix}m_{-1}(x,\zeta)\\ m_{-2}(x,\zeta)\end{bmatrix}.$$

COROLLARY 2.8. *The analogue of Theorem* 2.7 *holds for m_{-1} and m_{-2}.*

The scattering data, definitions, and properties. There are three components to the scattering data associated to a potential u in (1.4). The first is the *reflection coefficient* r_+, which has already been defined. The second is the set of zeros of $a_+(\zeta)$ in $\mathbb{H}$, or, equivalently, the set of *poles of the transmission coefficient* $t_+(\zeta)=1/a_+(\zeta)$. The third is the set of *normalizing chains*, which we will define below.

Since $a_+(\zeta)$ is equal to the Wronskian $W[\Psi_-,\Psi_+]$, it is continuous in $\bar{\mathbb{H}}$ and analytic in $\mathbb{H}$. Following Tanaka in § 2 of [Ta], we see that $a_+(\zeta)\to 1$ as $|\zeta|\to\infty$ in $\bar{\mathbb{H}}$, and we make the technical assumption mentioned in the Introduction, namely,

$$\text{for all real } \xi,\ a_+(\xi)\neq 0. \tag{2.12}$$

Observe that condition (2.12) is satisfied for $u\equiv 0$, just as $a_+(\xi)\equiv 1$ in this case. Note that for ξ fixed in $\mathbb{R}$, $a_+(\xi)$ is a real analytic function of Re (u) and Im (u). Therefore, for a given ξ in $\mathbb{R}$, the condition $a_+(\xi)\neq 0$ is satisfied generically. It is not likely that the stronger condition (2.12) is generic.

It follows that the set of zeros of $a_+(\zeta)$ in $\mathbb{H}$ is finite; enumerate them as

$$\zeta_1,\zeta_2,\cdots,\zeta_J.$$

By convention, $J=0$ will mean that a_+ has no zeros in $\mathbb{H}$. Let $m(j)$ denote the multiplicity of the jth zero of a_+ in $\mathbb{H}$.

The normalization chains c_j^+ and c_j^- relate the Jost functions $\Psi_+(x,\zeta_j)$ and $\Psi_-(x,\zeta_j)$. To define these chains, we give a simpler version of Tanaka's Theorem 2.3 in [Ta] and prove it by a more elementary argument.

THEOREM 2.9. *Suppose that a_+ has a zero of multiplicity $m(j)$ at ζ_j in $\mathbb{H}$. Then there are sequences*

$$c_j^+=(c_{j,0}^+,c_{j,1}^+,\cdots,c_{j,m(j)-1}^+),$$

$$c_j^-=(c_{j,0}^-,c_{j,1}^-,\cdots,c_{j,m(j)-1}^-),$$

such that

$$c_{j,0}^+\neq 0 \quad \textit{and} \quad c_{j,0}^-\neq 0,$$

and, for each k with $0\leqq k\leqq m(j)-1$,

$$\frac{1}{k!}\left(\frac{d}{d\zeta}\right)^k[\psi_-(x,\zeta)]\bigg|_{\zeta_j}=\sum_{\nu=0}^{k}\frac{1}{\nu!}c_{j,k-\nu}^+\left(\frac{d}{d\zeta}\right)^\nu[\psi_+(x,\zeta)]\bigg|_{\zeta_j}, \tag{2.13}$$

$$\frac{1}{k!}\left(\frac{d}{d\zeta}\right)^k[\psi_+(x,\zeta)]\bigg|_{\zeta_j}=\sum_{\nu=0}^{k}\frac{1}{\nu!}c_{j,k-\nu}^-\left(\frac{d}{d\zeta}\right)^\nu[\psi_-(x,\zeta)]\bigg|_{\zeta_j}. \tag{2.14}$$

Furthermore, c_j^+ is related to $\bar{c}_j$ by the relations

$$(2.15)\qquad \sum_{\lambda=\sigma}^{\tau} c^-_{j,\tau-\lambda}c^+_{j,\lambda-\sigma} = \begin{cases} 1 & \text{if } \sigma=\tau, \\ 0 & \text{if } \sigma<\tau, \end{cases}$$

whenever $0 \leqq \sigma \leqq \tau \leqq m(j)-1$.

Proof. For real ξ we have the relation

$$a_+(\xi) = W[\psi_-, \psi_+].$$

Extend ψ_+, ψ_-, and thus a_+ also, analytically to $\mathbb{H}$, obtaining

$$a_+(\zeta) = W[\psi_-(x,\zeta), \psi_+(x,\zeta)].$$

Since ζ_j is a zero of order $m(j)$ for $a_+(\zeta)$, we learn that for $0 \leqq k \leqq m(j)-1$

$$0 = \left(\frac{d}{d\zeta}\right)^k [a_+(\zeta)]\Big|_{\zeta_j},$$

whence

$$0 = \sum_{\nu=0}^{k} \binom{k}{\nu} W[\psi_-^{(\nu)}, \psi_+^{(k-\nu)}]\Big|_{\zeta_j}.$$

At $k=0$ we get, in particular,

$$0 = W[\psi_-(x,\zeta_j), \psi_+(x,\zeta_j)],$$

so each of $\psi_-(x,\zeta_j)$ and $\psi_+(x,\zeta_j)$ is a nonzero multiple of the other. Thus we can define $c_{j,0}^{\pm}$ by the relations

$$\psi_-(x,\zeta_j) = c^+_{j,0}\psi_+(x,\zeta_j) \quad \text{with } c^+_{j,0} \neq 0,$$
$$\psi_+(x,\zeta_j) = c^-_{j,0}\psi_-(x,\zeta_j) \quad \text{with } c^-_{j,0} \neq 0.$$

The proof now continues by induction. Assume that we have proved (2.13) and (2.14) for all $k=0,1,\cdots,N$, where N is less than $m(j)-1$. Now

$$0 = \left(\frac{d}{d\zeta}\right)^{N+1} [a_+(\zeta)]\Big|_{\zeta_j} = \sum_{k=0}^{N+1} \binom{N+1}{k} W[\psi_-^{(k)}, \psi_+^{(N+1-k)}]\Big|_{\zeta_j}$$
$$= W[\psi_-^{(N+1)}, \psi_+^{(0)}]|_{\zeta_j} + \sum_{k=0}^{N} \frac{(N+1)!}{k!(N+1-k)!} W[\psi_-^{(k)}, \psi_+^{(N+1-k)}]\Big|_{\zeta_j}.$$

Use the induction hypothesis to get

$$0 = W[\psi_-^{(N+1)}, \psi_+^{(0)}]|_{\zeta_j} + \sum_{k=0}^{N} \frac{(N+1)!}{k!(N+1-k)!} W\left[\sum_{\mu=0}^{k} \frac{k!}{\mu!} c^+_{j,k-\mu}\psi_+^{(\mu)}, \psi_+^{(N+1-k)}\right]\Bigg|_{\zeta_j}.$$

Treat $\mu=0$ separately from $\mu \geqq 1$.

$$0 = W[\psi_-^{(N+1)}, \psi_+^{(0)}]|_{\zeta_j} + \sum_{k=0}^{N} \frac{(N+1)!}{k!(N+1-k)!} W[k!\,c^+_{j,k}\psi_+^{(0)}, \psi_+^{(N+1-k)}]\Big|_{\zeta_j}$$
$$+ \sum_{k=0}^{N} \frac{(N+1)!}{k!(N+1-k)!} W\left[\sum_{\mu=1}^{k} \frac{k!}{\mu!} c^+_{j,k-\mu}\psi_+^{(\mu)}, \psi_+^{(N+1-k)}\right]\Bigg|_{\zeta_j}.$$

In the middle term use the linearity and skew symmetry of W to get

$$0 = W\left[\psi_-^{(N+1)} - \sum_{k=0}^{N} \frac{(N+1)!\,k!}{k!(N+1-k)!} c^+_{j,k}\psi_+^{(N+1-k)}, \psi_+^{(0)}\right]\Bigg|_{\zeta_j}$$
$$+ \sum_{k=0}^{N} \frac{(N+1)!}{k!(N+1-k)!} W\left[\sum_{\mu=1}^{k} \frac{k!}{\mu!} c^+_{j,k-\mu}\psi_+^{(\mu)}, \psi_+^{(N+1-k)}\right]\Bigg|_{\zeta_j}.$$

Suppose we could show that the second line is zero. Then there is a nonzero $\gamma_{j,N+1}$ such that

$$\psi_+^{(0)}(x,\zeta_j)=\gamma_{j,N+1}\left\{\psi_-^{(N+1)}-\sum_{k=0}^{N}\frac{(N+1)!}{(N+1-k)!}c_{j,k}^+\psi_+^{(N+1-k)}\right\}\Bigg|_{\zeta_j}.$$

Let $c_{j,N+1}^+=[(N+1)!\,\gamma_{j,N+1}]^{-1}$. Setting $n=N+1-k$, we find that

$$\frac{1}{(N+1)!}\psi_-^{(N+1)}(x,\zeta_j)=\sum_{n=0}^{N+1}\frac{1}{n!}c_{j,N+1-n}^+\psi_+^{(n)}(x,\zeta_j).$$

It remains to show that

$$0=\mathscr{S}_N:=\sum_{k=0}^{N}\frac{(N+1)!}{k!(N+1-k)!}W\left[\sum_{\mu=1}^{k}\frac{k!}{\mu!}c_{j,k-\mu}^+\psi_+^{(\mu)},\ \psi_+^{(N+1-k)}\right]\Bigg|_{\zeta_j}.$$

Now, since the $k=0$ term is vacuous,

$$\mathscr{S}_N=\sum_{k=1}^{N}\sum_{\mu=1}^{k}\frac{(N+1)!}{(N+1-k)!\,\mu!}c_{j,k-\mu}^+W[\psi_+^{(\mu)},\psi_+^{(N+1-k)}]|_{\zeta_j}.$$

Set $l=k-\mu$ for each fixed k.

$$\begin{aligned}\mathscr{S}_N&=\sum_{k=1}^{N}\sum_{l=0}^{k-1}\frac{(N+1)!}{(N+1-k)!(k-l)!}c_{j,l}^+W[\psi_+^{(k-l)},\psi_+^{(N+1-k)}]\Bigg|_{\zeta_j}\\&=(N+1)!\sum_{l=0}^{N-1}c_{j,l}^+\sum_{k=l+1}^{N}\frac{1}{(N+1-k)!(k-l)!}W[\psi_+^{(k-l)},\psi_+^{(N+1-k)}]\Bigg|_{\zeta_j}.\end{aligned}$$

Let S_l denote the sum over k.

$$S_l=\sum_{k=l+1}^{N}\frac{1}{(N+1-k)!(k-l)!}W[\psi_+^{(k-l)},\psi_+^{(N+1-k)}]\Bigg|_{\zeta_j}.$$

Note that as k runs from $l+1$ to N, $\mu:=N+1-k$ runs from $N-l$ to 1, and $\nu:=k-l$ runs from 1 to $N-l$. Thus

$$S_l=\sum_{\mu=1}^{N-l}\frac{1}{\mu!(N+1-l-\mu)!}W[\psi_+^{(N+1-l-\mu)},\psi_+^{(\mu)})]\Bigg|_{\zeta_j}$$

and

$$S_l=\sum_{\nu=1}^{\nu-l}\frac{1}{(N+1-l-\nu)!\,\nu!}W[\psi_+^{(\nu)},\psi_+^{(N-1-l-\nu)}]\Bigg|_{\zeta_j}.$$

So $S_l=-S_l$, whence $S_l=0$, and $\mathscr{S}_N=0$.

Thus (2.13) is proved for $k=N+1$ and the induction is complete. (2.14) is proved similarly, and (2.15) follows by comparing (2.13) and (2.14). □

Definition. The sequences c_j^+ and c_j^- are the normalizing chains associated to the jth pole of $T_+(\zeta)$. Because of (2.15), c_j^- is determined by c_j^+.

Properties of the reflection coefficient r_+. Since $W[\psi_-(x,\xi),\psi_+(x,\xi)]$ is independent of x for real ξ, this independence of ξ persists into the upper half plane, and

$$\begin{aligned}a_+(\zeta)&=W[\psi_-(x,\zeta),\psi_+(x,\zeta)]|_{x=0}\\&=m_{-1}(0,\zeta)m_{+2}(0,\zeta)-m_{+1}(0,\zeta)m_{-2}(0,\zeta).\end{aligned}$$

PROPOSITION 2.10. *Assume that* $u \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$.

(i) $a_+(\xi) - 1 \in C_0(\mathbb{R})$, *i.e.*, $a_+(\xi) - 1$ *is continuous form* $\mathbb{R}$ *to* $\mathbb{C}$ *and has limit* 0 *as* $\xi \to \pm\infty$; $a_+(\zeta) - 1 \in H^{2+} \cap H^{\infty+}$.

(ii) *If, in addition,* $u \in H^\alpha(\mathbb{R})$ *for* $0 < \alpha < \frac{1}{2}$, *then* $\xi^\alpha(a_+(\xi) - 1) \in L^2(\mathbb{R})$.

(iii) *If, in addition,* $|x|^\alpha u(x) \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ *for* $0 < \alpha < \frac{1}{2}$, *then* $a_+(\xi) - 1 \in H^\alpha(\mathbb{R})$.

Proof. This can be read off from the properties of T_1, T_2, m_+, and m_-. □

Similarly, we get the expansion

$$b_+(\zeta) = m_{-1}(0, \zeta) m^*_{+1}(0, \zeta) - m_{-2}(0, \zeta) m^*_{+2}(0, \zeta),$$

and find the following result.

PROPOSITION 2.11. *Let* $u \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$. *Then*

(i) $b_+(\xi) \in L^2(\mathbb{R}) \cap C_0(\mathbb{R})$;

(ii) *If, in addition,* $u \in H^\alpha(\mathbb{R})$ *with* $0 < \alpha < \frac{1}{2}$, *then* $\xi^\alpha b_+(\xi) \in L^2(\mathbb{R})$;

(iii) *If, in addition,* $u \in H^\alpha$ *and* $x^\alpha u(x) \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ *with* $0 < \alpha < \frac{1}{2}$, *then* $b_+(\xi) \in H^\alpha(\mathbb{R})$.

COROLLARY 2.12. *Assume* $u \in L^2(\mathbb{R}) \in L^1(\mathbb{R})$ *and* u *satisfies* (2.12). *Then*

(i) $r_+(\xi) \in L^2(\mathbb{R}) \cap C_0(\mathbb{R})$.

(ii) *If, in addition,* $u \in H^\alpha(\mathbb{R})$ *with* $0 < \alpha < \frac{1}{2}$, *then* $\xi^\alpha r_+(\xi) \in L^2(\mathbb{R})$.

(iii) *If, in addition,* $u \in H^\alpha$ *and* $x^\alpha u(x) \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ *with* $0 < \alpha < \frac{1}{2}$, *then* $r_+(\xi) \in H^\alpha(\mathbb{R})$.

Proof. Parts (i) and (ii) follow immediately from Propositions 2.11 and 2.12 since by (2.12) $a_+(\xi)$ is bounded away from zero for ξ in $\mathbb{R}$.

By Theorem 10.2 of [LM, p. 58] we see that (iii) is equivalent to

$$r_+(\xi) \in L^2(\mathbb{R}) \quad \text{and} \quad \int_0^\infty t^{-(1+2\alpha)} \int_{-\infty}^\infty |r_+(\xi + t) - r_+(\xi)|^2 \, d\xi \, dt < \infty.$$

We already know that $r_+ \in L^2$. We also know that $a_+(\xi)$ is continuous and never zero on $\mathbb{R}$, and that $a_+(\xi) \to 1$ as $\xi \to \pm\infty$. Thus there is an M such that $|a_+(\xi)|^{-1} \leqq M$ for all real ξ. Since $b_+(\xi)$ is continuous on $\mathbb{R}$ and goes to zero as $\xi \to \pm\infty$, we may take M larger if necessary to get $|b_+(\xi)| \leqq M$ for all real ξ. Finally, because of Propositions 2.11 and 2.12, we know that

$$\int_0^\infty t^{-(1+2\alpha)} \int_{-\infty}^\infty |a_+(\xi + t) - a_+(\xi)|^2 \, d\xi \, dt < \infty$$

and

$$\int_0^\infty t^{-(1+2\alpha)} \int_{-\infty}^\infty |b_+(\xi + t) - b_+(\xi)|^2 \, d\xi \, dt < \infty.$$

Now

$$\begin{aligned} &\int_0^\infty t^{-(1+2\alpha)} \int_{-\infty}^\infty |r_+(\xi + t) - r_+(\xi)|^2 \, d\xi \, dt \\ &\leqq \int_0^\infty t^{-(1+2\alpha)} \int_{-\infty}^\infty \left| \frac{b_+(\xi + t)}{a_+(\xi + t)} - \frac{b_+(\xi)}{a_+(\xi)} \right|^2 d\xi \, dt \\ &\leqq \int_0^\infty t^{-(1+2\alpha)} \int_{-\infty}^\infty \{[1 + M^2] M^2 [|b_+(\xi + t) - b_+(\xi)|^2 + |a_+(\xi + t) - a_+(\xi)|^2] \, d\xi \, dt < \infty. \end{aligned}$$

□

3. The scattering data and Marchenko kernel for positive time. At $t = 0$ we have identified the components of the scattering data associated to (1.4), where the potential

u is the initial function u_0 in the cubic Schrödinger problem. For $t>0$ we define nominal scattering data, "nominal" because we do not assert that these objects really arise as the scattering data of any potential in $L^1(\mathbb{R})\cap L^2(\mathbb{R})$.

DEFINITIONS.

$$r(\xi,t):=r_\pm(\xi)\,e^{\pm 2i\xi^2 t}, \tag{3.1a}$$

$$\zeta_j(t):=\zeta_j \quad \text{for } 1\leqq j\leqq J, \tag{3.1b}$$

$$(c^\pm_{j,\mu}(t))_{\mu=0}^{m(j)-1}, \tag{3.1c}$$

are determined as in Tanaka's paper [Ta] by the ordinary differential equations

$$\frac{d}{dt}[c^\pm_{j,\mu}]=\pm 2i(\zeta_j^2 c^\pm_{j,\mu}+2\zeta_j c^\pm_{j,\mu-1}+c^\pm_{j,\mu-2})$$

with the conventions

$$c^\pm_{j,-1}(t)\equiv c^\pm_{j,-2}(t)\equiv 0$$

and the initial conditions

$$c^\pm_{j,\mu}(0)=c^\pm_{j,\mu}.$$

Tanaka remarks that

$$c^\pm_{j,\mu}(t)=\mathcal{P}^\pm_{j,\mu}(t)\,e^{2i\zeta_j^2 t},$$

where $\mathcal{P}^\pm_{j,\mu}(t)$ is a polynomial of degree μ in t. The ordinary differential equations (3.1c) insure that the relations (2.15) for the normalizing chains persist for $t>0$.

Now, following Tanaka [Ta], and Zakharov and Shabat [ZS], we define

$$\Omega_+(x,t):=F_+(x,t)+\sum_{j=1}^{J} f_j(x,t), \tag{3.2}$$

where

$$F_+(x,t)=\pi^{-1}\int_{-\infty}^{\infty} r_+(\xi,t)\,e^{2i\xi x}\,d\xi \tag{3.3a}$$

and

$$f_{+j}(x,t)=-2i\sum_{\mu=0}^{m(j)-1}\frac{1}{\mu!}c^+_{m(j)-1-\mu}(t)\left(\frac{d}{d\zeta}\right)^{\mu}\left[\frac{(\zeta-\zeta_j)^{m(j)}\,e^{2i\zeta x}}{a_+(\zeta)}\right]\Bigg|_{\zeta_j}. \tag{3.3b}$$

Introduce $Q^+_{j,\mu}(x)$ by the equation

$$Q^+_{j,\mu}(x)\,e^{2i\zeta_j x}=\left(\frac{d}{d\zeta}\right)^{\mu}\left[\frac{(\zeta-\zeta_j)^{m(j)}\,e^{2i\zeta x}}{a_+(\zeta)}\right]\Bigg|_{\zeta_j}.$$

Similarly, we define Ω_-, F_-, f_{-j}, and $Q^-_{j,\mu}$.

Observe that $Q^+_{j,\mu}(x)$ is a polynomial in x of degree μ, and

$$f_{+j}(x,t)=-2i\sum_{\mu=1}^{m(j)-1}\frac{1}{\mu!}\mathcal{P}^+_{j,m(j)-1-\mu}(t)\,Q^+_{j,\mu}(x)\,e^{2i\zeta_j^2 t+2i\zeta_j x}.$$

The ordinary differential equations for the normalizing chains were chosen to insure that each f_{+j} satisfied the partial diffeerential equation

$$i\omega_t-\tfrac{1}{2}\omega_{xx}=0. \tag{3.4}$$

The $f_{+j}(x,t)$ are smooth functions of x and t and decay exponentially as $x\to+\infty$ for any fixed t.

Since $r_+(\cdot, t)$ evolves in $L^2(\mathbb{R})$, so does $F_+(\cdot, t)$. Further, F_+ satisfies the initial value problem (3.4) with

$$\omega(x, 0) = F_+(x).$$

Note that (3.4) differs from the linear part of the cubic Schrödinger equation (1.1) in the sign in front of the second space derivative.

PROPOSITION 3.1. *Suppose there is an* α *with* $0 < \alpha < 1$ *such that*

$$F_+(x) \in H^\alpha(\mathbb{R}) \quad \textit{and} \quad (1+|x|^\alpha)F_+(x) \in L_2(\mathbb{R}).$$

Then for each T *with* $0 < T < \infty$,

$$\text{(a)} \qquad F_+(x, t) \in L^\infty([0, T], H^\alpha(\mathbb{R}))$$

and

$$\text{(b)} \qquad (1+|x|^\alpha)F_+(x, t) \in L^\infty([0, T], L^2(\mathbb{R})).$$

Proof. It suffices to prove that

$$\text{(i)} \qquad r_+(\xi, t) \in L^\infty([0, T], H^\alpha(\mathbb{R}))$$

and

$$\text{(ii)} \qquad (1+|x|^\alpha)r_+(\xi, t) \in L^\infty([0, T], L^2(\mathbb{R})).$$

But $r_+(\xi, t) = r_+(\xi) \exp(2i\xi^2 t)$, and thus (ii) is trivial. For (i) we must show that the H^α-norm of $r_+(\xi, t)$ is bounded uniformly in $0 \leqq t \leqq T$. Following [LM, Thm. 10.2, p. 52] the H^α-norm of $f(x)$ is equivalent to

$$\left[\|f\|^2_{L^2(\mathbb{R})} + \int_0^\infty ds\, s^{-(1+2\alpha)} \int_{\mathbb{R}} dx\, |f(x+s) - f(x)|^2\right]^{1/2}.$$

Clearly $\|r_+(\xi, t)\|_{L^2(\mathbb{R})}$ is independent of t and thus it remains to prove that

$$\int_0^\infty ds\, s^{-(1+2\alpha)} \int_{\mathbb{R}} d\xi\, |r_+(x+s, t) - r_+(\xi, t)|^2$$

is uniformly bounded for $0 \leqq t \leqq T$. Write

$$\int_0^\infty ds\, s^{-(1+2\alpha)} \int_{-\infty}^\infty d\xi\, |r_+(\xi+s) \exp(2i(\xi+s)^2 t) - r_+(\xi) \exp(2i\xi^2 t)|^2$$

$$\leqq 2 \int_0^\infty ds\, s^{-(1+2\alpha)} \int_{-\infty}^\infty |r_+(\xi+s) - r_+(\xi)|^2\, d\xi + 2 \int_0^\infty ds\, s^{-(1+2\alpha)}$$

$$\cdot \int_{-\infty}^\infty |r_+(\xi)|^2 \exp(i(\xi+s)^2 t - i\xi^2 t) - \exp(-i(\xi+s)^2 t + i\xi^2 t)|^2\, d\xi$$

$$\leqq 2C\|r_+(\cdot)\|^2_{H^\alpha} + 2 \int_0^\infty ds\, s^{-(1+2\alpha)} \int_{-\infty}^\infty |r_+(\xi)|^2 4 \sin^2((\xi+s)^2 t - \xi^2 t)\, d\xi$$

$$\leqq I_1 + I_2.$$

It remains to estimate I_2. Divide the outer integral at $s=1$.

$$I_3=2\int_1^\infty ds\, s^{-(1+2\alpha)}\int_{-\infty}^{\infty}|r_+(\xi)|^2 4\sin^2((s^2+2\xi s)t)\, d\xi$$

$$\leqq 8\|r_+\|_{L^2(\mathbb{R})}\int_1^\infty s^{-(1+2\alpha)}\, ds<\infty.$$

$$I_4=8\int_0^1 ds\, s^{-(1+2\alpha)}\int_{-\infty}^{\infty}|r_+(\xi)|^2\sin^2((s^2+2\xi s)t)\, d\xi$$

$$=8\int_{-\infty}^{\infty} d\xi\, |r_+(\xi)|^2\int_0^1 ds\, s^{-(1+2\alpha)}\sin^2((s^2+2\xi s)t)=I_5+I_6,$$

where I_5 is the piece of the outer integral where $|\xi|\leqq 1$, and I_6 is the rest.

$$I_5\leqq 8\int_{|\xi|\leqq 1} d\xi\, |r_+(\xi)|^2\int_0^1 ds\, s^{-(1+2\alpha)}(s^2+2\xi s)^2t^2$$

$$\leqq 8\int_{|\xi|\leqq 1} d\xi\, |r_+(\xi)|^2\int_0^1 ds\, s^{1-2\alpha}(s+2\xi)^2t^2<\infty$$

since $0<\alpha<1$. Now

$$I_6=8\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2\int_0^1 ds\, s^{-(1+2\alpha)}\sin^2((s^2+2\xi s)^2t^2)=I_7+I_8,$$

where we have divided the integral at $s=|\xi|^{-1}$.

$$I_7=8\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2\int_{1/|\xi|}^1 ds\, s^{-(1+2\alpha)}\sin^2((s^2+2\xi s)t)$$

$$\leqq 8\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2\int_{1/|\xi|}^1 ds\, s^{-(1+2\alpha)}$$

$$\leqq 8\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2(|\xi|^{2\alpha}-1)/2\alpha<\infty$$

since we know that $r_+(\xi)|\xi|^\alpha\in L^2(\mathbb{R})$.

$$I_8=8\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2\int_0^{1/|\xi|} ds\, s^{-(1+2\alpha)}\sin^2((s^2+2\xi s)t)$$

$$\leqq 8\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2\int_0^{1/|\xi|} ds\, s^{-(1+2\alpha)}s^2(s+2\xi)^2t^2.$$

Since $0\leqq s\leqq 1/|\xi|$ and $|\xi|\geqq 1$, we see that

$$I_8\leqq 8\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2\int_0^{1/|\xi|} ds\, s^{1-2\alpha}(3\xi)^2t^2$$

$$\leqq 36t^2(1-\alpha)^{-1}\int_{|\xi|\geqq 1} d\xi\, |r_+(\xi)|^2|\xi|^{2\alpha}<\infty.$$

This concludes the proof of Theorem 3.1. □

4. Solution of the Marchenko equation for $t>0$. Here we obtain a function that will be shown in § 6 to be the solution to the initial value problem (1.1), (1.2), and we establish some of its properties.

We consider further the kernel $\Omega_+(x,t)$ discussed in § 3. For each real x and each nonnegative t, let F_x^t, G_{xj}^t, and Ω_x^t be the operators defined by

$$F_x^t[g](y) := \int_0^\infty F_+(x+y+z,t)g(z)\,dz,$$

$$G_{xj}^t[g](y) := \int_0^\infty f_{+j}(x+y+z,t)g(z)\,dz,$$

$$\Omega_x^t[g](y) := F_x^t[g](y) + \sum_{j=1}^J G_{xj}^t[g](y).$$

It can be shown that for each x the map $t \to \Omega_x^t$ is continuous from $[0,\infty)$ into $\mathscr{L}(L^2(\mathbb{R}^+))$ with the operator norm topology. The map $x \to \Omega_x^t$ is continuous in the strong operator topology, but not in the uniform topology. Note that $(I+\Omega_x^{t*}\Omega_x^t)^{-1}$ exists in $L^2(\mathbb{R}^+)$ with operator norm no larger than one because $\Omega_x^{t*}\Omega_x^t$ is positive and self-adjoint.

It is easy to check that the Marchenko equation

$$0 = B_+^{\#}(x,y,t) + \Omega_+(x+y,t)\begin{bmatrix}0\\1\end{bmatrix} + \int_0^\infty \Omega_+(x+y+z,t)B_+(x,z,t)\,dz \tag{4.1}$$

is equivalent to the system of equations

$$B_{+2}(x,y,t) = -\int_0^\infty \Omega_+^*(x+y+z,t)B_{+1}(x,z,t)\,dz, \tag{4.2a}$$

$$\Omega_+^*(x+y,t) = (I+\Omega_x^{t*}\Omega_x^t)[B_{+1}(x,\cdot,t)](y). \tag{4.2b}$$

Thus we get solutions to (4.1) by setting

$$B_{+1}(x,\cdot,t) = (I+\Omega_x^{t*}\Omega_x^t)^{-1}[\Omega_+^*(x+[\cdot],t)] \tag{4.3a}$$

and

$$B_{+2}(x,\cdot,t) = -\Omega_x^{t*}[B_{+1}(x,\cdot,t)]. \tag{4.3b}$$

It is easy to check that the map $(x,t) \to B_+(x,\cdot,t)$ is continuous from $\mathbb{R}\times[0,\infty)$ into $L^2(\mathbb{R})$.

DEFINITION. The function that we shall eventually prove to be the solution of our initial value problem for the cubic Schrödinger equation is

$$u(x,t) := u_+(x,t) := -\Omega_+^*(x,t) + \Omega_x^{t*}\Omega_x^t[B_+(x,\cdot,t)](0). \tag{4.4}$$

Observe that if $g \in L^2(\mathbb{R})$, then $\Omega_x^t[g](y)$ depends continuously on x, y, and t. Thus $\Omega_x^t\Omega_x^t[B_+(x,\cdot,t)](y)$ is continuous in x, y, t and can be evaluated at $y=0$.

For the rest of this section we derive some elementary properties of $u(x,t)$.

PROPOSITION 4.1. *Suppose that $u_0 \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and that u_0 satisfies the generic condition* (2.12). *Then the map $t \to u(x,t)$ is continuous from* $[0,\infty)$ *into* $L^2_{\text{loc}}(\mathbb{R})$.

Proof. We need to prove that for $t \geqq 0$, $u(\cdot,t) \in L^2_{\text{loc}}(\mathbb{R})$. The continuity follows from the remarks above.

It is clear from the definition of $\Omega_+(x,t)$ that $\Omega_+^*(\cdot,t) \in L^2(+\infty)$. It will suffice to show that the other term in $u(x,t)$ is bounded on all half lines $[a,\infty)$ in the x-axis. Now

$$|\Omega_x^{t*}\Omega_x^t[B_{+1}(x,\cdot,t)](0)| \leqq \int_0^\infty |\Omega_+(x+w,t)|\,|\Omega_x^t[B_{+1}(x,\cdot,t)](w)|\,dw$$

$$\leqq \|\Omega_+(\cdot,t)\|_{L^2([x,\infty))}\|\Omega_x^t\|_{op}\|B_{+1}(x,\cdot,t)\|_{L^2(\mathbb{R}^+)}.$$

Since $\|(I+\Omega_x^{t*}\Omega_x^t)^{-1}\|_{op}\leqq 1$, we conclude from (4.3a) that

$$\|B_{+1}(x,\cdot,t)\|_{L^2(\mathbb{R}^+)}\leqq\|\Omega_+(\cdot,t)\|_{L^2([x,\infty))}.$$

To control the operator norm of

$$\Omega_x^t=F_x^t-2i\sum_1^J G_{jx}^t,$$

we check that

$$\|F_x^t\|_{op}\leqq\|r_+\|_{L^\infty(\mathbb{R})}$$

and that

$$\|G_{jx}\|_{op}\leqq P(x,t)\,e^{-cx}$$

for a polynomial P and some positive c. Thus, finally, $u(\cdot,t)\in L^2_{\text{loc}}(\mathbb{R})$. □

PROPOSITION 4.2. *Suppose that* $(1+|x|^\alpha)u_0\in L^2(\mathbb{R})\cap L^1(\mathbb{R})$ *and* $u_0\in H^\alpha(\mathbb{R})$ *for some* α *with* $\frac{1}{4}<\alpha$. *Then the map* $t\to u(\cdot,t)$ *is continuous from* $[0,\infty)$ *into* $L^2(+\infty)$.

Proof. Since we know that $u(\cdot\;,t)\in L^2_{\text{loc}}$ for each t, it suffices to show that $u(\cdot,t)\in L^2([1,\infty))$. From (4.4) we have already seen that

$$|u(x,t)|\leqq|\Omega_+(x,t)|+\{\|\Omega_+(\cdot,t)\|_{L^2([x,\infty))}\}^2\|\Omega_x^t\|_{op}$$

The first term we know is in $L^2([1,\infty))$. We also know that the operator norm is bounded on $[1,\infty)$. Now, using the properties of f_{+j}, we get

$$\begin{aligned}\|\Omega_+(\cdot,t)\|^2_{L^2([x,\infty))}&\leqq\left\{\|F_+(\cdot,t)\|_{L^2([x,\infty))}+2\sum_1^J\|f_{+j}(\cdot,t)\|_{L^2([x,\infty))}\right\}^2\\&\leqq\{1+4J\}\left\{\|F_+(\cdot,t)\|^2_{L^2([x,\infty))}+\sum_1^J\|f_{+j}(\cdot,t)\|^2_{L^2([x,\infty))}\right\}\\&\leqq\{1+4J\}\left\{\|F_+(\cdot,t)\|^2_{L^2([x,\infty))}+\sum_1^J K(t)\,e^{-cx}\right\}.\end{aligned}$$

Using Proposition 3.1, we find that $\|F_+(\cdot,t)\|^2_{L^2([x,\infty))}\leqq K^2x^{-2\alpha}$. □

5. Uniqueness results. The inverse scattering method provides two candidates for the solution to the initial value problem of this paper. One is

$$u_+(x,t):=-B_{+1}(x,0,t),\tag{5.1}$$

where B_+ solves the right-side Marchenko equation

$$B_+^{\#}(x,y,t)+\begin{bmatrix}1\\0\end{bmatrix}\Omega_+(x+y,t)+\int_0^\infty\Omega_+(x+y+z,t)B_+(x,z,t)\,dz=0.\tag{5.2}$$

The other is

$$u_-(x,t):=-B^*_{-2}(x,0,t),\tag{5.3}$$

where B_- solves the left-side Marchenko equation

$$B_-^{\#}(x,y,t)+\begin{bmatrix}0\\1\end{bmatrix}\Omega_-(x+y,t)+\int_{-\infty}^0\Omega_-(x+y+z,t)B_-(x,z,t)\,dz=0.\tag{5.4}$$

In the KdV problem for $t>0$, the left-side Marchenko equation is far less tractable than the right-side equation. In the cubic Schrödinger equation problem, the two Marchenko equations are both solvable for $t>0$, provided only that

$$u_0\in L^1(\mathbb{R})\cap L^2(\mathbb{R}); \tag{5.5}$$

$a_+(\zeta)$ has no zeros on the real axis.

THEOREM 5.1. *If* u_0 *satisfies* (5.5), *then*

$$u_+(x,t)=u_-(x,t)\quad \textit{for}\, x\in\mathbb{R},\ t>0. \tag{5.6}$$

Proof. We first show that it suffices to establish the following identities:

$$a_+^{-1}(\xi)\psi_-(x,\xi,t)=\psi_+^{\#}(x,\xi,t)+r_+(\xi,t)\psi_+(x,\xi,t)\quad \text{for } x>0 \tag{5.7}$$

and

$$a_-^{-1}(\xi)\psi_+(x,\xi,t)=\psi_-^{\#}(x,\xi,t)+r_-(\xi,t)\psi_-(x,\xi,t)\quad \text{for } x<0. \tag{5.8}$$

The proof of these identities is deferred.

By construction we know that

$$L_{u_+}[\psi_+]=\zeta\psi_+, \tag{5.9}$$

$$L_{u_-}[\psi_-]=\zeta\psi_-. \tag{5.10}$$

From (5.10) with $\zeta=\xi\in\mathbb{R}$ and $x>0$, we get

$$L_{u_-}[\psi_-]=\xi\psi_-.$$

From (5.7) and verification that $L_{u_+}[\psi_+^{\#}]=\xi\psi_+^{\#}$, we get

$$\begin{aligned}L_{u_+}[\psi_-]&=L_{u_+}[a_+\psi_+^{\#}+b_+\psi_+]\\&=a_+L_{u_+}[\psi_+^{\#}]+b_+L_{u_+}[\psi_+]\\&=\xi(a_+\psi_+^{\#}+b_+\psi_+)=\xi\psi_-.\end{aligned}$$

Thus,

$$0=L_{u_-}[\psi_-]-L_{u_+}[\psi_-]=-i\begin{bmatrix}0&(u_--u_+)\\(u_--u_+)&0\end{bmatrix}\psi_-.$$

Since ψ_- cannot vanish, $u_-=u_+$ for $x>0$. By working with (5.8), we can show similarly that $u_-=u_+$ for $x<0$.

Proof of (5.7). Keep $x>0$ and $t>0$ throughout. Recall that

$$a_+(\zeta,t)=a_+(\zeta)\quad\text{and}\quad r_+(\xi,t)=r_+(\xi)\,e^{2i\xi^2t}.$$

Now (5.7) is equivalent to

$$m_-(\xi,t)/a_+(\xi)=m_+^{\#}(x,\xi,t)+r_+(\xi,t)\,e^{2i\xi x}m_+(x,\xi,t). \tag{5.11}$$

Let g_0 denote the right side of (5.11):

$$g_0(x,\xi,t):=m_+^{\#}(x,\xi,t)+r_+(\xi,t)\,e^{2i\xi x}m_+(x,\xi,t)$$

and set

$$\mathcal{G}_0(x,y,t):=\pi^{-1}\int_{-\infty}^{\infty}\left\{g_0(x,\xi,t)-\begin{bmatrix}1\\0\end{bmatrix}\right\}e^{2i\xi y}\,d\xi.$$

Thus

$$\mathcal{G}_0(x, y, t) = B_+^{\#}(x, y, t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} F_+(x+y, t) + \int_{-\infty}^{\infty} B_+(x, z, t) F_+(x+y+z, t)\, dz.$$

From the definitions and the Marchenko equation, we get

$$\begin{aligned}\mathcal{G}_0(x, y, t) &= -\begin{bmatrix} 0 \\ 1 \end{bmatrix} G_+(x+y, t) - \int_{-\infty}^{\infty} B_+(x, z, t) G_+(x+y+z, t)\, dz \\ &= -\sum_{j=1}^{J} \left\{ \begin{bmatrix} 0 \\ f_{+j}(x+y, t) \end{bmatrix} + \int_{-\infty}^{\infty} B_+(x, z, t) f_{+j}(x+y+z, t)\, dz \right\},\end{aligned}$$

where

$$f_{+j}(x, t) := -2i \sum_{\mu=0}^{m(j)-1} \frac{1}{\mu!} c^+_{m(j)-1-\mu}(t) \left(\frac{d}{d\zeta}\right)^{\mu} \left[\frac{(\zeta-\zeta_j)^{m(j)}\, e^{2i\zeta x}}{a_+(\zeta)}\right]\Bigg|_{\zeta=\zeta_j}.$$

Now keep $y > 0$, as well as $x > 0$. Recall that $B_+(x, z, t) = 0$ when $z < 0$. Let $f_{+j}^{\sim}(x, t)$ be the result of cutting off f_{+j} with the characteristic function of $\mathbb{R}^+$. Careful analysis of $f_{+j}^{\sim}$ and contour integration in the manner of Tanaka [Ta] verifies that g_0 is meromorphic in the upper half plane, with poles exactly at the ζ_j with orders $m(j)$. Introduce the new functions

$$m_0(x, \zeta, t) := a_+(\zeta) g_0(x, \zeta, t),$$
$$\psi_0(x, \zeta, t) := e^{-i\zeta x} m_0(x, \zeta, t).$$

We will show that $m_0 = m_-$; hence $\psi_0 = \psi_-$.

Since the zeros of a_+ kill off the poles of g_0, the function m_0 is holomorphic in $\mathbb{H}$. Indeed, a detailed analysis shows that

$$m_0(x, \zeta, t) - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \text{the Hardy space } H^{2+}, \text{provided that } x > 0.$$

Setting

$$C(x, y, t) := \pi^{-1} \mathcal{F}\left[m_0(x, \cdot, t) - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right](-y) \quad \text{for } y > 0,$$

we verify that

$$m_0(x, \cdot, t) - \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \int_{-\infty}^{0} C(x, y, t)\, e^{-2i\xi y}\, dy.$$

To show that $m_0 = m_-$ it suffices to show that $C = B_-$. Since the left-hand Marchenko equation (5.4) has a unique solution, it suffices to show that C satisfies (5.4).

Careful computation using the properties of a_+, b_+, and g_0 leads to

$$-m_+(x, \xi, t)/a_+(\xi) = m_0^{\#}(x, \xi, t) + r_-(\xi, t)\, e^{-2i\xi x} m_0(x, \xi, t). \tag{5.12}$$

Taking a Fourier transform and evaluating the integrals by contour integration, we get

$$\begin{aligned}&-\sum_{j=1}^{J} 2i \operatorname{Res} \frac{m_+(x, \zeta)\, e^{-2i\zeta y \cdot}}{a_+(\zeta)} \quad \text{at } \zeta_j \\ &\quad = C^{\#}(x, y, t) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Omega_-(x, z, t) + \int_{-\infty}^{0} C(x, z, t) \Omega_-(x+y+z, t)\, dz \\ &\quad\quad - \sum_{j=1}^{J} \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} f_{-j}(x+y, t) + \int_{-\infty}^{0} C(x, z, t) f_{-j}(x+y+z, t)\, dz \right\}.\end{aligned}$$

We will show that the jth terms of the sums are equal in order to conclude that C solves (5.4). It turns out that it suffices to prove that

$$\frac{1}{\alpha!}\psi_+^{(\alpha)}(x, \zeta_j, t) = \sum_{\lambda=0}^{\alpha} c_{j,\alpha-\lambda}^{-}\frac{1}{\lambda!}\psi_0^{(\lambda)}(x, \zeta_j, t) \quad \text{for all } \alpha \quad \text{with } 0<\alpha<m(j)-1. \tag{5.13}$$

Here the superscripts (α) and (λ) denote orders of differentiation with respect to ζ. Because of the conditions

$$\sum_{\lambda=\sigma}^{\nu} c_{j,\nu-\lambda}^{-} c_{j,\lambda-\sigma}^{+} = \begin{cases} 1 & \text{if } \sigma=\nu, \\ 0 & \text{if } \sigma<\nu, \end{cases}$$

on the normalizing chains, (5.13) is equivalent to the identity

$$\frac{1}{\gamma!}\psi_0^{(\gamma)}(x, \zeta_j, t) = \sum_{\beta=0}^{\gamma} c_{j,\gamma-\beta}^{+}\frac{1}{\beta!}\psi_+^{(\beta)}(x, \zeta_j, t). \tag{5.14}$$

Finally, the verification of (5.14) is a straightforward argument based on (5.12).

The proof of (5.8) is similar.

PROPOSITION 5.2. *If* $u_0 \in H^\alpha(\mathbb{R})$ *and* $(1+|x|^\alpha)u_0 \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ *for an* α *with* $\frac{1}{4}<\alpha<\frac{1}{2}$, *then the map* $t \to u(\cdot, t)$ *is continuous from* $[0, \infty)$ *into* $L^2(\mathbb{R}) \cap L^4(\mathbb{R})$.

Proof. Here we use the representation (4.4) to show that the map $t \to u(\cdot, t)$ is continuous from $[0, \infty)$ into $L^2(\mathbb{R}^+) \cap L^4(\mathbb{R}^+)$. The proof for the other half-line is similar and uses Theorem 5.1.

It follows from Proposition 3.1 that the maps $t \to F_+(\cdot, t)$ and $t \to \Omega_+(\cdot, t)$ are continuous from $[0, \infty)$ into $H^\alpha(\mathbb{R}^+)$. By the Sobolev theorem, $H^\alpha(\mathbb{R}^+)$ is continuously included in $L^2(\mathbb{R}^+) \cap L^4(\mathbb{R}^+)$, and thus $t \to \Omega_+^*(\cdot, t)$ maps $[0, \infty)$ continuously into $L^2(\mathbb{R}^+) \cap L^4(\mathbb{R}^+)$.

We already know from the proof of Proposition 4.2 that the map $t \to \Omega_+^{t*}\Omega_x^t[B_+(x, \cdot, t)](0)$ is continuous from $[0, \infty)$ into $L^2(\mathbb{R}^+)$, and we know from the proof of Proposition 4.1 that this map is continuous from $[0, \infty)$ into $L^\infty(\mathbb{R}^+)$.

6. Verification that *u* solves the problem. In this section we show that the function $u(x, t) := -B_+(x, 0, t)$ constructed in § 4 solves the cubic Schrödinger equation (1.1) in the sense that

$$\iint u(x, t)\left\{i\varphi_t + \frac{1}{2}\varphi_{xx} + |u|^2\varphi\right\} dx\, dt = 0 \quad \text{for all } \varphi \in C_0^\infty(\mathbb{R} \times [0, \infty)). \tag{6.1}$$

We already know that the initial condition (1.2) is satisfied in the sense that

$$u(\cdot, t) \to u_0(\cdot) \quad \text{in } L^2_{\text{loc}}(\mathbb{R}) \quad \text{as } t \to 0^+. \tag{6.2}$$

Return to $r_+(\xi)$, the reflection coefficient of the initial profile $u_0(x)$. From Corollary 2.12, it follows that if $u_0 \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap H^\alpha(\mathbb{R})$ for $\frac{1}{4}<\alpha<\frac{1}{2}$, then $r_+ \in L^2(\mathbb{R}) \cap C_0(\mathbb{R})$ and $|\xi|^\alpha r_+(\xi) \in L^2(\mathbb{R})$. It follows further that $r_+ \in L^{3/2}(\mathbb{R})$. By cutting off and convolving with an appropriate approximate identity, we can obtain a sequence $\{r_n(\xi)\colon n \in \mathbb{N}\}$ in $C_0^\infty(\mathbb{R})$ such that

$$\|r_n\|_{L^\infty(\mathbb{R})} \leqq \|r_+\|_{L^\infty(\mathbb{R})},$$

and

$$r_n \to r_+ \quad \text{in } L^p(\mathbb{R}) \quad \text{as } n \to +\infty \quad \text{for } p=\tfrac{3}{2}, 2, \text{and } \infty.$$

Now set

$$F_n(x, t) := \pi^{-1} \int_{-\infty}^{\infty} r_n(k)\, e^{2ik^2t+2ikx}\, dk.$$

These F_n all satisfy

$$iF_t - \tfrac{1}{2}F_{xx} = 0; \tag{6.3}$$

they all evolve in $L^2(\mathbb{R})$; and at $t=0$,

$$F_n(\cdot, 0) = \mathscr{F}^{-1}[r_n] \to \mathscr{F}^{-1}[r_+] = F_+(\cdot, 0) \quad \text{in } L^2(\mathbb{R}) \quad \text{as } n \to +\infty.$$

But the initial value problem for (6.3) is well posed in $L^2(\mathbb{R})$, so we also have

$$F_n(\cdot, t) \to F_+(\cdot, t) \text{ in } L^2(\mathbb{R}) \quad \text{as } n \to \infty \quad \text{for each } t \geqq 0.$$

Finally note that

$$F_n(x, t) = \mathscr{F}^{-1}[r_n(k) \exp(4ik^2t)]$$

and that

$$\|r_n(k) \exp(4ik^2t)\|_{L^\infty(\mathbb{R})} \leqq \|r_+\|_{L^\infty(\mathbb{R})}.$$

Define kernels $\Omega_n(x, t)$ by

$$\Omega_n(x, t) = F_n(x, t) - 2i \sum_1^J f_{+j}(x, t),$$

where the $f_{+j}(x, t)$ are the functions defined in (3.3b).

We define operators on $L^2(\mathbb{R}^+)$ by

$$F^t_{n,x}[g](y) := \int_0^\infty F_n(x+y+z, t) g(z)\, dz,$$

$$\Omega^t_{n,x}[g](y) := F^t_{n,x} - 2i \sum_1^J G^t_{jx},$$

where the G^t_{jx} were defined in § 4.

Tanaka, in [Ta], has shown that smooth solutions to (1.1) are obtained by setting

$$u_n(x, t) = -B_n(x, 0, t),$$

where B_n solves

$$(I + \Omega^{t*}_{n,x}\Omega^t_{n,x})[B_n(x, \cdot, t)] = \Omega^*_n(x + \cdot, t). \tag{6.4}$$

The remainder of this section shows how these smooth solutions u_n converge to u as $n \to +\infty$.

LEMMA 6.1. *$B_n(x, \cdot, t) \to B_+(x, \cdot, t)$ in $L^2(\mathbb{R}^+)$ as $n \to +\infty$, uniformly for $x \geqq X$ and $0 \leqq t < T$ for any positive X and T.*

Proof. The proof is straightforward.

PROPOSITION 6.2. *As $n \to +\infty$, $u_n \to u$ weakly; i.e.,*

$$\iint u_n(x, t)\varphi(x, t)\, dx\, dt \to \iint u(x, t)\varphi(x, t)\, dx\, dt \quad \textit{for all } \varphi \in C_0^\infty(\mathbb{R} \times [0, \infty)).$$

Proof. This proof is also straightforward.

THEOREM 6.3. *If $u_0 \in L^1 \cap L^2 \cap H^\alpha$ for $\frac{1}{4} < \alpha$, then u is a weak solution of* (1.1), (1.2) *in $t \geqq 0$.*

Proof. Consider a ψ from $C_0^\infty(\mathbb{R}\times[0,\infty))$. Since the u_n are smooth classical solutions of (1.1) and evolve in Schwartz class $\mathcal{S}$, we get

$$0=\iint u\left\{iu_{n,t}+\frac{1}{2}u_{n,xx}+|u_n|^2u_n\right\}\varphi\,dx\,dt$$
$$=\iint u\left\{i\varphi_t+\frac{1}{2}\varphi_{xx}+|u_n|^2\varphi\right\}dz\,dt$$

for each n and each φ in $C_0^\infty(\mathbb{R}\times[0,\infty))$. In light of Proposition 6.2 we get

$$\lim_{n\to\infty}\iint u_n\left\{i\varphi_t+\frac{1}{2}\varphi_{xx}\right\}dx\,dt=\iint u\left\{i\varphi_t+\frac{1}{2}\varphi_{xx}\right\}dx\,dt.$$

It remains to show that

$$\lim_{n\to\infty}\iint u_n|u_n|^2\varphi\,dx\,dt=\iint u|u|^2\varphi\,dx\,dt. \tag{6.5}$$

Set

$$\mathcal{M}_n(x,t)=\Omega^{t*}_{n,x}\Omega^{t}_{n,x}[B_n(x,\cdot,t)](0)$$

and

$$\mathcal{M}(x,t)=\Omega^{t*}_{x}\Omega^{t}_{x}[B(x,\cdot,t)](0).$$

Thus,

$$u_n(x,t)=\mathcal{M}_n(x,t)+\Omega^*_n(x,t)$$

and

$$u(x,t)=\mathcal{M}(x,t)+\Omega^*_+(x,t).$$

To prove (6.5) it suffices to prove that

$$u_n|u_n|^2-u|u|^2\to 0\quad\text{in } L^1_{\text{loc}}(\mathbb{R}\times[0,\infty)). \tag{6.6}$$

Now

$$u_n|u_n|^2-u|u|^2=(u_n-u)|u_n|^2+u(|u_n|^2-|u|^2)$$
$$=(u_n-u)|u_n|^2+u(|u_n|-|u|)(|u_n|+|u|).$$

So,

$$|u_n|u_n|^2-u|u|^2|\leqq|u_n-u|^3+3|u_n-u|^2|u|+3|u_n-u|\,|u|^2.$$

Pick any compact interval $I=[a,b]$ in the x-axis.

$$\int_a^b|u_n|u_n|^2-u|u|^2|\,dx$$
$$\leqq\int_a^b|u_n-u|^3\,dx+3\left(\int_a^b|u_n-u|^3\,dx\right)^{2/3}\left(\int_a^b|u|^3\,dx\right)^{1/3}$$
$$+3\left(\int_a^b|u_n-u|^3\,dx\right)^{1/3}\left(\int_a^b|u|^3\,dx\right)^{2/3}.$$

It will suffice to show that $u_n-u\to 0$ in $L^3_{\text{loc}}(\mathbb{R})$ uniformly on any compact time interval $[0,T]$. Now

$$u_n-u=\Omega_n-\Omega+\mathcal{M}_n-\mathcal{M}=F_n-F+\mathcal{M}_n-\mathcal{M}.$$

We already know that $\mathcal{M}_n-\mathcal{M}\to 0$ in $L^\infty(\mathbb{R}\times[0,\infty))$. It remains to have $F_n-F\to 0$ in $L^3(\mathbb{R})$ uniformly in t. By the Hausdorff-Young inequality, we get

$$\begin{aligned}\|F_n(\cdot,t)-F_n(\cdot,t)\|_{L^3(\mathbb{R})}&=\|\mathcal{F}^{-1}[r_n(\cdot,t)-r_+(\cdot,t)]\|_{L^3(\mathbb{R})}\\&\leqq\|\{r_n(k)-r_+(k)\}\,e^{4ik^3t}\|_{L^{3/2}(\mathbb{R})}\\&=\|r_n(k)-r_+(k)\|_{L^{3/2}(\mathbb{R})}.\end{aligned}$$

We know that $r_n(k)\to r_+(k)$ in $L^{3/2}(\mathbb{R})$. Thus $F_n(\cdot,t)-F_n(\cdot,t)\to 0$ in $L^3(\mathbb{R})$ uniformly in t. □

7. Examples of initial data satisfying the general hypotheses of the existence and uniqueness theorem. In this section we consider the problem (1.1), (1.2) with initial data of the form

$$u_0(x)=\begin{cases}A & \text{when } 0<x<X,\\ 0 & \text{otherwise,}\end{cases}\tag{7.1}$$

where A is a nonzero constant and X is positive. In order to apply the general theorem we need to compute the Jost functions for u_0 as potential in (1.4), determine when the (2.12) is satisfied, and determine the regularity of u_0.

The scattering problem is

$$\begin{aligned}\psi_{1,x}-u\psi_2&=-i\zeta\psi_1,\\ \psi_{2,x}+u^*\psi_1&=i\zeta\psi_2.\end{aligned}$$

The boundary conditions determining the Jost functions $\vec{\psi}_+$ and $\vec{\psi}_-$ are

$$\begin{bmatrix}\psi_{+1}\\ \psi_{+2}\end{bmatrix}\sim\begin{bmatrix}0\\ e^{i\zeta x}\end{bmatrix}\quad\text{as } x\to+\infty;\qquad\begin{bmatrix}\psi_{-1}\\ \psi_{-2}\end{bmatrix}\sim\begin{bmatrix}e^{i\zeta x}\\ 0\end{bmatrix}\quad\text{as } x\to-\infty.$$

We now compute $\vec{\psi}_+$. It is easy to see that

$$\vec{\psi}_+(x,\zeta)=\begin{bmatrix}0\\ e^{i\zeta x}\end{bmatrix}\quad\text{throughout } X<x<+\infty.$$

In $0<x<X$, the scattering ordinary differential equations become

$$\begin{aligned}\psi_{1,x}-A\psi_2&=-i\zeta\psi_1,\\ \psi_{2,x}-A^*\psi_1&=i\zeta\psi_2,\end{aligned}$$

whence

$$\psi_2=\psi_{1,x}+i\zeta\psi_1/A$$

and

$$\psi_{1,xx}+\{|A|^2+\zeta^2\}\psi_1=0.$$

For the moment, restrict attention to $\zeta=\xi\in\mathbb{R}$. Let $\mathcal{A}=\sqrt{|A|^2+\xi^2}$. We will see that it is not necessary to resolve the sign ambiguity in the square root. Writing

$$\psi_{+1}=a_m\,e^{i\mathcal{A}x}+b_m\,e^{-i\mathcal{A}x}$$

and matching boundary conditions at $x=X$, we find that

$$\psi_{+1}(x,\xi)=A\,e^{i\xi X}(x-X)\left[\frac{\sin[\mathcal{A}(x-X)]}{\mathcal{A}(x-X)}\right]$$

and

$$\psi_{+2}(x,\xi)=e^{i\xi X}\cos[\mathcal{A}(x-X)]+\xi\,e^{i\xi X}\left[\frac{\sin[\mathcal{A}(x-X)]}{\mathcal{A}(x-X)}\right].$$

Note that both ψ_{+1} and ψ_{+2} are, in fact, functions of $\mathcal{A}^2$; therefore, both are well defined despite any sign ambiguity in $\mathcal{A}$.

Now in $x<0$, $u_0(x)\equiv 0$ and ψ_{+1}, ψ_{+2} must have the form

$$\psi_{+1}(x,\xi)=\beta\,e^{i\xi X},$$

$$\psi_{+2}(x,\xi)=\alpha\,e^{i\xi X}.$$

Matching boundary conditions at $x=0$, we find that

$$\alpha=e^{i\xi X}[\cos(\mathcal{A}X)-i\xi\sin(\mathcal{A}X)/\mathcal{A}],$$

$$\beta=-A\,e^{i\xi X}\sin(\mathcal{A}X)/\mathcal{A}.$$

Now the scattering coefficients $a_+(\xi)$ and $b_+(\xi)$ are defined by

$$\vec{\psi}_-=a_+\psi_-^{\#}+b_+\psi_+.$$

It is easy to see that

$$\psi_-(x,\xi)\equiv\begin{bmatrix}e^{-i\xi x}\\0\end{bmatrix}\quad\text{if } x<0.$$

Thus we can determine a_+ and b_+ from α and β:

$$a_+=\frac{\alpha}{|\alpha|^2+|\beta|^2};\qquad b_+=\frac{\beta}{|\alpha|^2+|\beta|^2}.$$

It also turns out to be easy to confirm that $|\alpha|^2+|\beta|^2=1$ for all real ξ. The following results are now immediate.

PROPOSITION 7.1.

(i) *The only possible real zero of $a_+(\xi)$ is $\xi=0$.*

(ii) *Unless AX is an odd multiple of $\pi/2$, $a_+(\xi)$ has no real zeros at all.*

Observe that

$$u_0(x)\in H^{\alpha}(\mathbb{R})\quad \textit{for all } \alpha\quad \textit{with } 0<\alpha<\tfrac{1}{2}$$

and

$$|x|^{\alpha}u_0\in L^2(\mathbb{R})\cap L^1(\mathbb{R})\quad \textit{for } 0\leqq\alpha.$$

Thus by the existence and uniqueness results of §§ 5 *and* 6, *we get the following.*

THEOREM 7.2. *Let u_0 be a function of the form*

$$u_0(x)=\begin{cases}A & \textit{when } 0<x<X,\\ 0 & \textit{otherwise},\end{cases}$$

where $X>0$ and $0\neq A\in\mathbb{R}$. If AX is not an odd multiple of $\pi/2$, then there is a solution of (1.1), (1.2) *in the sense of the theorem stated in the Introduction.*

REFERENCES

[BC1] R. Beals and R. Coifman, *Scattering and inverse scattering for first order systems*, Comm. Pure Appl. Math., 37 (1984), pp. 39-90.

[BC2] ———, *Inverse scattering and evolution equations*, Comm. Pure Appl. Math., 38 (1985), pp. 29-42.

[GGKM] O. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, *Method for solving the Korteweg-deVries equation*, Phys. Rev. Lett., 19 (1967), pp. 1095-1097.

[GV] J. Ginibre and G. Velo, *On a class of nonlinear Schrödinger equations* I: *the Cauchy problem*, J. Funct. Anal., 32 (1979), pp. 1-32.

[HNT1] N. Hayashi, K. Nakamitsu, and M. Tsutsumi, *On solutions of the initial value problem for the nonlinear Schrödinger equations in one space dimension*, Math. Z., 192 (1986), pp. 637-650.

[HNT2] ———, *On solutions of the initial value problem for the nonlinear Schrödinger equations*, J. Funct. Anal., 71 (1987), pp. 218-245.

[K] T. Kato, *On nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987) pp. 113-129.

[LM] J.-L. Lions and E. Magenes, *Problemes aux limites non-homogenes et applications*, Vol. I. Springer-Verlag, New York, 1972.

[SZ] D. Sattinger and V. Zurkowski, *Gague theory of Backlund transformations* II, Phys. D, 26 (1987), pp. 225-250.

[Ta] S. Tanaka, *Nonlinear Schrödinger equation and modified Korteweg-deVries equation*: *construction of solutions in terms of scattering data*, Publ. Res. Inst. Math. Sci., Kyoto Univ., 10 (1975), pp. 329-357.

[T1] Y. Tsutsumi, *Global strong solutions for nonlinear Schrödinger equations*, Nonlinear Anal. Theory, Meth. Appl., 11 (1987), pp. 1143-1154.

[T2] ———, *L^2 solutions for nonlinear Schrödinger equations and nonlinear groups*, Funkcial. Ekvac., 30 (1987), pp. 115-125.

[ZS] V. E. Zakharov and A. B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, J. Exp. Theor. Phys., 61 (1971), pp. 118-134. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 923–940, July 1992

GENERALIZED SOLUTIONS TO THE KORTEWEG–DE VRIES AND THE REGULARIZED LONG-WAVE EQUATIONS*

H. A. BIAGIONI[†] AND M. OBERGUGGENBERGER[‡]

Abstract. In this article generalized solutions to two model equations describing nonlinear dispersive waves are studied. The solutions are found in certain algebras of new generalized functions containing spaces of distributions. On the one hand, this allows the handling of initial data with strong singularities. On the other hand, suitable scaling allows one to introduce an infinitesimally small coefficient; thereby the authors produce generalized solutions in the sense of Colombeau to the inviscid Burgers equation.

Key words. nonlinear dispersive waves, algebras of generalized functions, singular initial data, small dispersion limit

AMS(MOS) subject classifications. 35Q20, 35D05, 46F10, 35L65

1. Introduction.

The purpose of this paper is to study generalized solutions to the Korteweg–de Vries equation

$$u_t + uu_x + u_{xxx} = 0 \tag{1}$$

as well as to the so-called regularized long-wave equation

$$u_t + u_x + uu_x - u_{xxt} = 0 \tag{2}$$

(proposed and investigated by Benjamin, Bona, and Mahony in [2]; see also [5], [6]) in the framework of generalized functions introduced by Colombeau [9], [10].

This article is the second part of a program to obtain generalized solutions to hyperbolic conservation laws by adding a viscous or dispersive term that is associated with zero; see [4].

A soliton described by the KdV equation leads to an example of a nonzero solution to (1) in the Colombeau algebra $\mathcal{G}(I\!R \times [0,\infty))$ whose restriction to $t = 0$ is zero in $\mathcal{G}(I\!R)$. This shows that we do not have uniqueness of solutions to the Cauchy problem with initial data in $\mathcal{G}(I\!R)$ for the KdV equation in the algebra $\mathcal{G}(I\!R \times [0,\infty))$. Accordingly, we define new algebras of generalized functions, denoted by $\mathcal{G}_{p,q}(\Omega)$, where $1 \leq p, q \leq \infty$ and Ω is a domain in $I\!R^n$ with the cone property, whose elements have representatives with bounds taken in terms of the L^p-norms. If $\Omega = I\!R^n$, this algebra contains the space $W^{-\infty,p}(I\!R^n)$ and it has $W^{\infty,q}(I\!R^n)$ as a subalgebra, if $q \leq p$. The algebra $\mathcal{G}_{s,g}(\overline{\Omega})$, defined in [4], is the particular case where $p = q = \infty$, if Ω has the strong local Lipschitz property.

We prove that there exists a solution to (1) in $\mathcal{G}_{\infty,2}(I\!R \times (0,\infty))$, given initial data $g \in \mathcal{G}_{2,2}(I\!R)$. Further, for every $T > 0$, there is at most one solution in $\mathcal{G}_{2,2}(I\!R \times (0,T))$ such that its partial x-derivative has a logarithmic dependence on the regularization parameter.

* Received by the editors March 19, 1990; accepted for publication (in revised form) November 20, 1991.

[†] Departamento de Matemática, UNICAMP, 13081 Campinas SP, Brasil. This work was supported in part by Conselho Nacional de Pesquisas.

[‡] Institut für Mathematik und Geometrie, Universität Innsbruck, 6020 Innsbruck, Austria. This work was supported in part by Conselho Nacional de Pesquisas and Fundação de Amparo a Pesquisa do Estado de São Paulo.

For equation (2) we prove existence of solutions in $\mathcal{G}_{p,q}(\mathbb{R}\times(0,T))$ for each $T>0$, provided the initial data are in $\mathcal{G}_{p,q}(\mathbb{R})$ and have moderate H^1-bounds. There is at most one solution $u \in \mathcal{G}_{p,q}(\mathbb{R}\times(0,T))$ that is of "logarithmic-type" in the following cases: $q>2$ and $\frac{1}{p}+\frac{1}{q}\geq 1$; or $1\leq q\leq 2$ and $p<\infty$; or $1\leq q\leq 2$ and $p=\infty$ (in this last case a further decay requirement has to be imposed).

We prove also existence and uniqueness of solutions to

$$u_t + uu_x + \nu u_{xxx} = 0 \tag{3}$$

in $\mathcal{G}_{\infty,2}(\mathbb{R}\times(0,\infty))$ with initial data in $\mathcal{G}_{2,2}(\mathbb{R})$, where ν is a generalized constant. As a consequence, if ν is associated with zero and the initial data have a representative with the L^2-norm bounded independently of ε, then we obtain a solution to the conservation law

$$u_t + uu_x \approx 0, \tag{4}$$

written with association, which arises in the study of shock waves in Colombeau's setting. In particular, the generalized solution to (3) with classical initial data and ν associated with zero also satisfies (4), in spite of the fact that, according to the results of Lax and Levermore [11] and Venakides [15] on zero dispersion limits, its associated distribution will generally not be a weak solution to the conservation law $v_t + (\frac{1}{2}v^2)_x = 0$. Our setting also allows us to model "infinitely narrow solitons" in the sense of Maslov et al. [12], [13], as well as some source type solutions with zero impact at positive times.

2. The space $\mathcal{G}_{p,q}(\Omega)$.

2.1. Notation. In our notation concerning spaces of functions and distributions we follow Adams [1]. Thus for $\Omega\subset\mathbb{R}^n$ open, $m\in\mathbb{Z}$, and $1\leq p\leq\infty$, $W^{m,p}(\Omega)$ is the usual Sobolev space, $W^{\infty,p}(\Omega)=\cap_m W^{m,p}(\Omega)$, $W^{-\infty,p}(\Omega)=\cup_m W^{-m,p}(\Omega)$, $H^m(\Omega)=W^{m,2}(\Omega)$. If E is a Banach space, $C^j(\Omega;E)$ denotes the space of j-times continuously differentiable functions in Ω with values in E, $C^j_B(\Omega;E)$ those functions whose derivatives are bounded, in addition. For $E=\mathbb{R}$ we simply use $C^j_B(\Omega)$.

All functions and distribution spaces are assumed to be real valued in this paper.

Let $1\leq p,q\leq\infty$ and Ω an open subset of $\mathbb{R}^n$. We set

$$\begin{aligned}
\mathcal{E}[\Omega] &= \{u:(0,\infty)\times\Omega\to\mathbb{R} \text{ such that } u(\varepsilon,x) \text{ is } C^\infty \\
&\qquad \text{in the variable } x\in\Omega, \text{ for each } \varepsilon>0\};\\
\mathcal{E}_p[\Omega] &= \{u\in\mathcal{E}[\Omega] \text{ such that } u(\varepsilon,\cdot)\in W^{\infty,p}(\Omega) \text{ for all } \varepsilon>0\};\\
\mathcal{E}_{M,p}[\Omega] &= \{u\in\mathcal{E}_p[\Omega] \text{ such that for all } \alpha\in\mathbb{N}^n \text{ there is}\\
&\qquad N\in\mathbb{N} \text{ such that}\\
&\qquad \|\partial^\alpha u(\varepsilon,\cdot)\|_p = O(\varepsilon^{-N}) \text{ as } \varepsilon\to 0\}; \qquad (5)\\
\mathcal{N}_{p,q}(\Omega) &= \{u\in\mathcal{E}_{M,p}[\Omega]\cap\mathcal{E}_q[\Omega] \text{ such that for all } \alpha\in\mathbb{N}^n\\
&\qquad \text{and } M\in\mathbb{N},\\
&\qquad \|\partial^\alpha u(\varepsilon,\cdot)\|_q = O(\varepsilon^{M}) \text{ as } \varepsilon\to 0\}, \qquad (6)
\end{aligned}$$

where $\|\cdot\|_p$ denotes the L^p-norm.

Remarks 2.2. (i) If Ω has the strong local Lipschitz property and $u\in\mathcal{E}_p[\Omega]$, then $u(\varepsilon,\cdot)\in C^\infty(\overline{\Omega})$ for every ε. In particular, the sets $\mathcal{E}_{M,\infty}[\Omega]$ and $\mathcal{N}_{\infty,\infty}(\Omega)$ are, respectively, the same as $\mathcal{E}_{M,s,g}[\overline{\Omega}]$ and $\mathcal{N}_{s,g}(\overline{\Omega})$ defined in [4].

(ii) If $\Omega = \mathbb{R}^n$, $p < \infty$, and $u \in \mathcal{E}_p[\Omega]$ then, for every $\varepsilon > 0, \lim_{|x|\to\infty} u(\varepsilon, x) = 0$.

PROPOSITION 2.3. *Let Ω have the cone property. Then*

(i) *if* $p_1 \leq p_2$, $\mathcal{E}_{M,p_1}[\Omega] \subset \mathcal{E}_{M,p_2}[\Omega]$;

(ii) $\mathcal{E}_{M,p}[\Omega]$ *is an algebra with partial derivatives;*

(iii) $\mathcal{N}_{p,q}(\Omega)$ *is an ideal in* $\mathcal{E}_{M,p}[\Omega]$ *which is invariant under partial derivatives.*

Proof. (i) We have $\mathcal{E}_{M,p_1}[\Omega] \subset \mathcal{E}_{M,\infty}[\Omega]$. Indeed, by Sobolev's imbedding theorem [1, Chap. V], we have $W^{j+m,p_1}(\Omega) \subset C^j_B(\Omega)$ for all m such that $mp_1 > n$. Thus, given $j \in \mathbb{N}$ we have, for $u \in \mathcal{E}_{M,p_1}[\Omega]$,

$$\max_{|\alpha|\leq j} ||\partial^\alpha u(\varepsilon, \cdot)||_\infty \leq c \max_{|\beta|\leq m+j} ||\partial^\beta u(\varepsilon, \cdot)||_{p_1}.$$

Thus $u \in \mathcal{E}_{M,\infty}[\Omega]$.

Given $p_2 \geq p_1$, since $L^{p_1}(\Omega) \cap L^\infty(\Omega) \subset L^{p_2}(\Omega)$, we have, if $u \in \mathcal{E}_{M,p_1}[\Omega]$ and $\alpha \in \mathbb{N}^n$, $|\alpha| \leq j$,

$$\begin{aligned} ||\partial^\alpha u(\varepsilon, \cdot)||^{p_2}_{p_2} &= \int_\Omega |\partial^\alpha u(\varepsilon, x)|^{p_2-p_1} |\partial^\alpha u(\varepsilon, x)|^{p_1} dx \\ &\leq ||\partial^\alpha u(\varepsilon, \cdot)||^{p_2-p_1}_\infty ||\partial^\alpha u(\varepsilon, \cdot)||^{p_1}_{p_1} \\ &\leq c^{p_2-p_1} \max_{|\beta|\leq m+j} ||\partial^\beta u(\varepsilon, \cdot)||^{p_2-p_1}_{p_1} ||\partial^\alpha u(\varepsilon, \cdot)||^{p_1}_{p_1} \\ &\leq c^{p_2-p_1} \max_{|\beta|\leq m+j} ||\partial^\beta u(\varepsilon, \cdot)||^{p_2}_{p_1}. \end{aligned}$$

Thus $u \in \mathcal{E}_{M,p_2}[\Omega]$. The proofs of (ii) and (iii) follow from the fact that $\mathcal{E}_{M,p}[\Omega] \subset \mathcal{E}_{M,\infty}[\Omega]$. □

DEFINITION 2.4. We define, for $1 \leq p, q \leq \infty$,

$$\mathcal{G}_{p,q}(\Omega) = \mathcal{E}_{M,p}[\Omega]/\mathcal{N}_{p,q}(\Omega).$$

If Ω has the cone property, then it is clear from Proposition 2.3 that $\mathcal{G}_{p,q}(\Omega)$ is an algebra with partial derivatives.

PROPOSITION 2.5. *For $1 \leq p_1 \leq p_2 \leq \infty$ and $1 \leq q \leq \infty$ we have that*

$$\mathcal{G}_{p_1,q}(\Omega) \subset \mathcal{G}_{p_2,q}(\Omega).$$

Proof. Consider the commutative diagram

$$\begin{tikzcd} \mathcal{E}_{M,p_1}[\Omega] \arrow[r] & \mathcal{E}_{M,p_2}[\Omega] \\ \mathcal{N}_{p_1,q}(\Omega) \arrow[u] \arrow[r] & \mathcal{N}_{p_2,q}(\Omega). \arrow[u] \end{tikzcd}$$

The result follows from the fact that $\mathcal{N}_{p_2,q}(\Omega) \cap \mathcal{E}_{M,p_1}[\Omega] = \mathcal{N}_{p_1,q}(\Omega)$. □

Remark 2.6. It is not true that $\mathcal{G}_{p_1,p_1}(\Omega) \subset \mathcal{G}_{p_2,p_2}(\Omega)$ if $p_1 < p_2$, since $\mathcal{N}_{p_2,p_2}(\Omega) \cap \mathcal{E}_{M,p_1}[\Omega] \neq \mathcal{N}_{p_1,p_1}(\Omega)$ in general. We have, though, a canonical map $\mathcal{G}_{p_1,p_1}(\Omega) \to \mathcal{G}_{p_2,p_2}(\Omega)$, which, however, is not injective.

For example, it is simple to construct an element $u \in \mathcal{E}_{M,2}[\mathbb{R}] \cap \mathcal{N}_{\infty,\infty}(\mathbb{R})$, $u \notin \mathcal{N}_{2,2}(\mathbb{R})$ as follows: let $\tilde{u} : (0,\infty) \times \mathbb{R} \to [0,1]$ be such that $\tilde{u}(\varepsilon, x) \equiv 1$ for $|x| \leq \exp(\frac{2}{\varepsilon})$, $\tilde{u}(\varepsilon, x) \equiv 0$ for $|x| \geq 1 + \exp(\frac{2}{\varepsilon})$, $\tilde{u}(\varepsilon, \cdot) \in C^\infty(\mathbb{R})$, and all derivatives of $\tilde{u}(\varepsilon, \cdot)$ are bounded independently of ε, i.e., $||\tilde{u}^{(n)}(\varepsilon, \cdot)||_\infty \leq c_n$ for all $\varepsilon > 0$. Let

$u(\varepsilon, x) = \exp(-\frac{1}{\varepsilon})\tilde{u}(\varepsilon, x)$. We have, for $p \geq 2$,

$$2\exp\left(\frac{2}{\varepsilon}\right)\exp\left(-\frac{p}{\varepsilon}\right) \leq \int_{-\infty}^{\infty} |u(\varepsilon,x)|^p dx = \exp\left(-\frac{p}{\varepsilon}\right)\int_{|x|\leq 1+\exp(2/\varepsilon)} |\tilde{u}(\varepsilon,x)|^p dx$$

$$\leq \exp\left(-\frac{p}{\varepsilon}\right)\left(2+2\exp\left(\frac{2}{\varepsilon}\right)\right) = 2\exp\left(-\frac{p}{\varepsilon}\right) + 2\exp\left(\frac{2-p}{\varepsilon}\right),$$

and, if $n \geq 1$,

$$\int_{-\infty}^{\infty} |u^{(n)}(\varepsilon,x)|^p dx = \exp\left(-\frac{p}{\varepsilon}\right)\int_{-\infty}^{\infty} |\tilde{u}^{(n)}(\varepsilon,x)|^p dx \leq 2\exp\left(-\frac{p}{\varepsilon}\right) c_n^p.$$

These inequalities show that $u \in \mathcal{E}_{M,2}[\mathbb{R}] \cap \mathcal{N}_{p,p}(\mathbb{R})$ for every $p > 2$, but $u \notin \mathcal{N}_{2,2}(\mathbb{R})$.

THEOREM 2.7. (i) *There is an imbedding of* $W^{-\infty,p}(\mathbb{R}^n)$ *into* $\mathcal{G}_{p,q}(\mathbb{R}^n)$.

(ii) *If* $q \leq p$, *this imbedding turns* $W^{\infty,q}(\mathbb{R}^n)$ *into a subalgebra of* $\mathcal{G}_{p,q}(\mathbb{R}^n)$.

Proof. (i) Fix $\rho \in \mathcal{S}(\mathbb{R}^n)$ such that $\int_{\mathbb{R}^n} x^i \rho(x)dx = 0$ for all $i \in \mathbb{N}^n$, $|i| > 0$, and $\int_{\mathbb{R}^n} \rho(x)dx = 1$. Define

$$\iota : W^{-\infty,p}(\mathbb{R}^n) \rightarrow \mathcal{E}_p[\mathbb{R}^n] \tag{7}$$

by $\iota(w)(\varepsilon,x) = (w * \rho_\varepsilon)(x)$. Let us prove first that $\iota(w) \in \mathcal{E}_{M,p}[\mathbb{R}^n]$ for all $w \in W^{-\infty,p}(\mathbb{R}^n)$. There is $m \in \mathbb{N}$ such that $w \in W^{-m,p}$, so there exist $w_\alpha \in L^p(\mathbb{R}^n)$, $|\alpha| \leq m$, such that $w = \sum_{|\alpha|\leq m} \partial^\alpha w_\alpha$. Given $\beta \in \mathbb{N}^n$, we have by Young's inequality that

$$||\partial^\beta(w * \rho_\varepsilon)||_p \leq \sum_{|\alpha|\leq m} ||w_\alpha||_p ||\partial^{\alpha+\beta}\rho_\varepsilon||_1 = \sum_{|\alpha|\leq m} ||w_\alpha||_p \frac{1}{\varepsilon^{|\alpha+\beta|}} ||\partial^{\alpha+\beta}\rho||_1.$$

On the other hand, if $\iota(w) \in \mathcal{N}_{p,q}(\mathbb{R}^n)$, then $w * \rho_\varepsilon \rightarrow 0$ in $\mathcal{D}'(\mathbb{R}^n)$ as $\varepsilon \rightarrow 0$, so $w = 0$. It follows that ι imbeds $W^{-\infty,p}(\mathbb{R}^n)$ into $\mathcal{G}_{p,q}(\mathbb{R}^n)$.

(ii) Let $f \in W^{\infty,q}(\mathbb{R}^n)$. We show that $f * \rho_\varepsilon - f \in \mathcal{N}_{p,q}(\mathbb{R}^n)$ if $q \leq p$. We have

$$||f * \rho_\varepsilon - f||_q = \left(\int_{\mathbb{R}^n} |f * \rho_\varepsilon - f|^q dx\right)^{1/q}$$

$$= \left(\int_{\mathbb{R}^n} \left|\int_{\mathbb{R}^n} [f(x-\varepsilon y) - f(x)]\rho(y)dy\right|^q dx\right)^{1/q}.$$

By Taylor's formula up to order m applied to f, and since $\int y^\alpha \rho(y)dy = 0$ for $|\alpha| \leq m$, this in turn equals

$$\left(\int_{\mathbb{R}^n} \left|\sum_{|\alpha|=m+1} \int_{\mathbb{R}^n} \frac{(-\varepsilon y)^\alpha}{m!} \int_0^1 (1-\sigma)^m \partial^\alpha f(x - \sigma\varepsilon y) d\sigma \rho(y) dy\right|^q dx\right)^{1/q}$$

$$\leq C(m,q) \max_{|\alpha|=m+1} \left(\int_{\mathbb{R}^n} \left|\int_{\mathbb{R}^n} \frac{(-\varepsilon y)^\alpha}{m!} \rho(y) \int_0^1 (1-\sigma)^m \partial^\alpha f(x-\sigma\varepsilon y) d\sigma dy\right|^q dx\right)^{1/q}.$$

By the generalized Minkowski inequality, the last expression is estimated by

$$C(m,q) \max_{|\alpha|=m+1} \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} \left|\frac{(\varepsilon y)^\alpha}{m!} \rho(y) \int_0^1 (1-\sigma)^m \partial^\alpha f(x-\sigma\varepsilon y) d\sigma\right|^q dx\right)^{1/q} dy$$

$$\leq \frac{\varepsilon^{m+1}}{m!} C(m,q) \max_{|\alpha|=m+1} \int_{\mathbb{R}^n} |y^\alpha \rho(y)| \left(\int_{\mathbb{R}^n} \int_0^1 |\partial^\alpha f(x-\sigma\varepsilon y)|^q d\sigma dx\right)^{1/q} dy$$

$$\leq c\varepsilon^{m+1} \int_{\mathbb{R}^n} |y|^{m+1} |\rho(y)| dy \max_{|\alpha|=m+1} ||\partial^\alpha f||_q.$$

Thus we have, for every $m \in I\!N$ and small ε that

$$\|f * \rho_\varepsilon - f\|_q \leq c\varepsilon^m.$$

The same holds for all partial derivatives of f. Hence $f * \rho_\varepsilon - f \in \mathcal{N}_{p,q}(I\!R^n)$. □

DEFINITION 2.8. Let $u \in \mathcal{G}_{p,q}(I\!R \times (0,T))$. We define the *restriction of u to $I\!R \times \{0\}$* as follows: Let $\hat{u}$ be a representative of u. By Remark 2.2(i), $\hat{u}(\varepsilon,\cdot,\cdot) \in C^\infty(I\!R \times [0,T])$ for each $\varepsilon > 0$. Since the restriction map $W^{m+1,p}(I\!R \times (0,T)) \to W^{m,p}(I\!R)$ is continuous, we have that $\hat{u}(\varepsilon,\cdot,0)$ belongs to $\mathcal{E}_{M,p}[I\!R]$. Also, $\hat{u}(\varepsilon,\cdot,0) \in \mathcal{N}_{p,q}(I\!R)$ if $\hat{u} \in \mathcal{N}_{p,q}(I\!R \times (0,T))$. Thus we may define the restriction of u to $I\!R \times \{0\}$ as the class of $\hat{u}(\varepsilon,\cdot,0)$ in $\mathcal{G}_{p,q}(I\!R)$.

DEFINITION 2.9. We say that $u \in \mathcal{G}_{p,q}(\Omega)$ *is associated with the distribution* $w \in \mathcal{D}'(\Omega)$ if there is a representative $\hat{u}$ of u such that $\hat{u}(\varepsilon,\cdot) \to w$ in $\mathcal{D}'(\Omega)$ as $\varepsilon \to 0$. Notation: $u \approx w$.

DEFINITION 2.10. We say that $u \in \mathcal{G}_{p,q}(\Omega)$ is of *r-$\sqrt[j]{\log}$-type*, $r \geq p, j \geq 1$, if it has a representative $\hat{u} \in \mathcal{E}_{M,p}[\Omega]$ such that

$$\|\hat{u}(\varepsilon,\cdot)\|_r = O(\sqrt[j]{|\log\varepsilon|}) \quad \text{as } \varepsilon \to 0. \tag{8}$$

3. Generalized solutions to the KdV equation.

THEOREM 3.1. *Let $g \in \mathcal{G}_{2,2}(I\!R)$. Then there is a solution u of* (1) *in $\mathcal{G}_{\infty,2}(I\!R \times (0,\infty))$ such that*

$$u|_{t=0} = g \tag{9}$$

and, for every $T > 0$, $u|_{I\!R\times(0,T)} \in \mathcal{G}_{2,2}(I\!R \times (0,T))$.

Proof. Let $\hat{g} \in \mathcal{E}_{M,2}[I\!R]$ be a representative of g. Since $\hat{g}(\varepsilon,\cdot) \in H^\infty(I\!R)$ for each $\varepsilon > 0$, according to the existence theory of Bona and Smith [7, Thm. 3, p. 578], there is a unique solution $\hat{u}_\varepsilon$ of (1), with initial data $\hat{g}(\varepsilon,\cdot)$, which belongs to $\bigcap_{0\leq j\leq[s/3]} C_B^j([0,\infty); H^{s-3j}(I\!R))$ for arbitrary s. Observe that for every $m \in I\!N$ there is $s \in I\!N$ large enough such that

$$\bigcap_{0\leq j\leq[s/3]} C^j([0,T]; H^{s-3j}(I\!R)) \subset H^m(I\!R \times (0,T))$$

and

$$\bigcap_{0\leq j\leq[s/3]} C_B^j([0,\infty); H^{s-3j}(I\!R)) \subset C_B^m(I\!R \times [0,\infty)).$$

It follows that the solution thus constructed belongs to $H^\infty(I\!R \times (0,T))$ for every $T > 0$ as well as to $W^{\infty,\infty}(I\!R \times (0,\infty))$; in particular, the map $\hat{u} : (\varepsilon,x,t) \mapsto \hat{u}_\varepsilon(x,t)$ belongs to $\mathcal{E}_2[I\!R \times (0,T)] \cap \mathcal{E}_\infty[I\!R \times (0,\infty)]$. It remains to show that $\hat{u} \in \mathcal{E}_{M,\infty}[I\!R \times (0,\infty)] \cap \mathcal{E}_{M,2}[I\!R \times (0,T)]$. For this, it suffices to prove that for all $k \in I\!N$ there are $c > 0$ and $\eta > 0$ such that

$$\sup_{t\geq 0} \|\partial_x^k \hat{u}(\varepsilon,\cdot,t)\|_2 \leq \frac{c}{\varepsilon^N}, \qquad 0 < \varepsilon < \eta. \tag{10}$$

In fact, if we have (10), since $\hat{u}(\varepsilon,\cdot)$ satisfies (1), we get an analogous estimate for $\partial_t \hat{u}$ and then, by successive differentiations in the equation we get, for all $\alpha \in I\!N^2$,

$$\sup_{t\geq 0} \|\partial^\alpha \hat{u}(\varepsilon,\cdot,t)\|_2 \leq \frac{c}{\varepsilon^N}, \qquad 0 < \varepsilon < \eta.$$

Accordingly, this implies an analogous estimate for $||\partial^\alpha \hat{u}(\varepsilon,\cdot)||_{L^\infty(I\!R\times(0,\infty))}$ and $||\partial^\alpha \hat{u}(\varepsilon,\cdot)||_{L^2(I\!R\times(0,T))}$. Then the class of $\hat{u}$ will be an element of $\mathcal{G}_{\infty,2}(I\!R\times(0,\infty))$, a solution to (1),(9), whose restriction to any strip belongs to $\mathcal{G}_{2,2}(I\!R\times(0,T))$.

But inequality (10) is an immediate consequence of the following lemma, and the proof is completed.

LEMMA 3.2. *For every $k \in I\!N$ there is a polynomial P_k such that*

$$(11)\qquad ||\partial_x^k \hat{u}(\varepsilon,\cdot,t)||_2 \le P_k(||\hat{g}(\varepsilon,\cdot)||_2, ||\hat{g}'(\varepsilon,\cdot)||_2, \cdots, ||\hat{g}^{(k)}(\varepsilon,\cdot)||_2).$$

Proof. In order to simplify the notation we drop the ε and the "hat" on the representatives of u and g. By Miura, Gardner, and Kruskal [14], the KdV equation has a sequence of conserved quantities

$$(12)\qquad I_k(u) = \int_{-\infty}^{\infty} [(\partial_x^k u)^2 - c_k u(\partial_x^{k-1}u)^2 + Q_k(u,\partial_x u,\cdots,\partial_x^{k-2}u)]dx,$$

$k = 0,1,\cdots$ (here $\partial_x^{-1}u = \partial_x^{-2}u \equiv 0$), and Q_k is seen as a sum of monomials $u^{a_o}(\partial_x u)^{a_1}\cdots(\partial_x^{k-2}u)^{a_{k-2}}$ such that

$$(13)\qquad \sum_{i=0}^{k-2}\left(1+\frac{i}{2}\right)a_i = k+2.$$

We will prove the assertion by induction over k. For $k=0$ it is true since $I_o(u) = \int_{-\infty}^{\infty} u^2(x,t)dx$, i.e., $d/dt\int_{-\infty}^{\infty} u^2(x,t)dx = 0$ and $||u(\cdot,t)||_2 = ||g||_2$. Assume that (11) holds for $j \le k-1$. Thus also

$$(14)\qquad ||\partial_x^j u(\cdot,t)||_\infty \le R_j(||g||_2,\cdots,||g^{(j+1)}||_2), \qquad 0 \le j \le k-2.$$

By (12) we have

$$\int_{-\infty}^{\infty}(\partial_x^k u)^2 dx = \int_{-\infty}^{\infty}[c_k u(\partial_x^{k-1}u)^2 - Q_k(u,\partial_x u,\cdots,\partial_x^{k-2}u)]dx + C,$$

where, since $(d/dt)I_k(u) = 0$, $I_k(u) = I_k(g) = C$ for all $t \ge 0$. We have, by (14) for $j=0$ and the induction hypothesis,

$$\begin{aligned}\int_{-\infty}^{\infty}|u(\partial_x^{k-1}u)^2|dx &\le ||u||_\infty ||\partial_x^{k-1}u||_2^2\\ &\le R_o(||g||_2,||g'||_2)\left(P_{k-1}(||g||_2,\cdots,||g^{(k-1)}||_2)\right)^2.\end{aligned}$$

By (13) we have that, in all monomials of Q_k, $\sum_{j=0}^{k-2} a_j \ge 2$. Thus in each monomial we may take two factors and apply Hölder's inequality to them, while the other ones are estimated by the L^∞-norm. Thus

$$\begin{aligned}\int |Q_k(u,\ldots,\partial_x^{k-2}u)|dx &\le \tilde{Q}_k(||u||_\infty,\cdots,||\partial_x^{k-2}u||_\infty,||u||_2,\cdots,||\partial_x^{k-2}u||_2)\\ &\le \tilde{\tilde{Q}}_k(||g||_2,\cdots,||g^{(k-1)}||_2)\end{aligned}$$

by induction and by (14). Thus we have proved (11). □

Remark 3.3. Theorem 3.1 establishes the existence of a representative $\hat{u} \in \mathcal{E}_{M,2}[\mathbb{R} \times (0,T)]$. Since the ideal does not enter in the proof, we might as well infer that there exists a solution in $\mathcal{G}_{2,q}(\mathbb{R} \times (0,T))$ for every $T > 0$ and $q \geq 1$, if the initial data belong to $\mathcal{G}_{2,q}(\mathbb{R})$.

THEOREM 3.3. *Let $g \in \mathcal{G}_{2,2}(\mathbb{R})$. Then, for every $T > 0$ there is at most one solution $u \in \mathcal{G}_{2,2}(\mathbb{R} \times (0,T))$ of* (1), (9) *such that $\partial_x u$ is of ∞-log-type* (*see* 2.10).

Proof. Let $u_1, u_2 \in \mathcal{G}_{2,2}(\mathbb{R} \times (0,T))$ be two solutions to (1), (9) with respective representatives $\hat{u}_1, \hat{u}_2 \in \mathcal{E}_{M,2}[\mathbb{R} \times (0,T)]$ such that $\partial_x \hat{u}_i$ satisfies (8) with $r = \infty$, $j = 1$, $i = 1,2$. There are $N \in \mathcal{N}_{2,2}(\mathbb{R} \times (0,T))$ and $n \in \mathcal{N}_{2,2}(\mathbb{R})$ such that, setting $w = \hat{u}_1 - \hat{u}_2$, $h = \frac{1}{2}(\hat{u}_1 + \hat{u}_2)$:

$$\begin{aligned} [w_t + (hw)_x + w_{xxx}](\varepsilon,x,t) &= N(\varepsilon,x,t), \\ w(\varepsilon,x,0) &= n(\varepsilon,x). \end{aligned} \tag{15}$$

By changing representatives, we may assume that $n(\varepsilon, x) \equiv 0$. For simplicity we drop the ε in our notation. Multiplying (15) by w and integrating with respect to x, we obtain

$$\int_{-\infty}^{\infty} w\, w_t\, dx + \int_{-\infty}^{\infty} (hw)_x w\, dx + \int_{-\infty}^{\infty} w_{xxx} w\, dx = \int_{-\infty}^{\infty} w\, N\, dx.$$

By Remark 2.2 (ii) we get

$$\frac{1}{2}\frac{d}{dt}\int_{-\infty}^{\infty} w^2 dx + \frac{1}{2}\int_{-\infty}^{\infty} h_x w^2 dx = \int_{-\infty}^{\infty} wN dx.$$

Integrating from zero to $t \leq T$ we have, since $w(x,0) \equiv 0$,

$$\begin{aligned} \int_{-\infty}^{\infty} w^2(x,t)dx &= 2\int_0^t \int_{-\infty}^{\infty} w(x,\tau)N(x,\tau)dx\, d\tau \\ &\quad - \int_0^t \int_{-\infty}^{\infty} h_x(x,\tau)w^2(x,\tau)dx\, d\tau \\ &\leq 2||w||_{L^2(\mathbb{R}\times(0,T))}||N||_{L^2(\mathbb{R}\times(0,T))} \\ &\quad + ||h_x||_\infty \int_0^t \int_{-\infty}^{\infty} w^2(x,\tau)dx\, d\tau. \end{aligned}$$

By Gronwall's inequality, we get

$$||w(\cdot,t)||_2^2 \leq 2||w||_{L^2(\mathbb{R}\times(0,T))}||N||_{L^2(\mathbb{R}\times(0,T))} \exp(T||h_x||_\infty).$$

Since $w \in \mathcal{E}_{M,2}[\mathbb{R} \times (0,T)]$, $N \in \mathcal{N}_{2,2}(\mathbb{R} \times (0,T))$, and h_x is of ∞-log-type, it follows that, for every $M > 0$,

$$\sup_{t\in[0,T]} ||w(\cdot,t)||_2 = O(\varepsilon^M) \ \text{ as } \ \varepsilon \to 0.$$

For the x-derivatives, we get, by differentiation of (15):

$$\partial_x^k w_t + \partial_x^{k+1}(hw) + \partial_x^{k+3} w = \partial_x^k N.$$

Multiplying the above equation by $\partial_x^k w$ and integrating with respect to x we get

$$\int_{-\infty}^{\infty} \partial_x^k w_t\, \partial_x^k w\, dx + \int_{-\infty}^{\infty} \sum_{j=0}^{k+1} \binom{k+1}{j} \partial_x^{k+1-j} h\, \partial_x^j w\, \partial_x^k w\, dx \tag{16}$$

$$+\int_{-\infty}^{\infty} \partial_x^{k+3} w \, \partial_x^k w \, dx = \int_{-\infty}^{\infty} \partial_x^k N \, \partial_x^k w \, dx.$$

The third term on the left-hand side is zero and the second one equals

$$\left(k+\frac{1}{2}\right)\int_{-\infty}^{\infty} \partial_x h (\partial_x^k w)^2 dx + \int_{-\infty}^{\infty} \sum_{j=0}^{k-1} \binom{k+1}{j} \partial_x^{k+1-j} h \, \partial_x^j w \, \partial_x^k w \, dx.$$

Then (16) becomes

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{-\infty}^{\infty} (\partial_x^k w)^2 dx = &-\left(k+\frac{1}{2}\right)\int_{-\infty}^{\infty} \partial_x h (\partial_x^k w)^2 dx + \int_{-\infty}^{\infty} \partial_x^k N \, \partial_x^k w \, dx \\ &-\sum_{j=0}^{k-1} \binom{k+1}{j} \int_{-\infty}^{\infty} \partial_x^{k+1-j} h \, \partial_x^j w \, \partial_x^k w \, dx.\end{aligned}$$

Then we have, by integrating from zero to t:

$$\begin{aligned}\int_{-\infty}^{\infty} (\partial_x^k w)^2 dx \le\ & 2\left(k+\frac{1}{2}\right) ||\partial_x h||_\infty \int_0^t \int_{-\infty}^{\infty} (\partial_x^k w(x,s))^2 dx\, ds \\ &+2\int_0^t \int_{-\infty}^{\infty} |\partial_x^k N(x,s) \, \partial_x^k w(x,s)| dx\, ds \\ &+2\sum_{j=0}^{k-1} \binom{k+1}{j} \int_0^t \int_{-\infty}^{\infty} |\partial_x^{k+1-j} h \, \partial_x^k w \, \partial_x^j w| dx\, ds.\end{aligned}$$

The last two terms on the right-hand side can be estimated by

$$2\left[||\partial_x^k N||_2 ||\partial_x^k w||_2 + c\sum_{j=0}^{k-1} ||\partial_x^{k+1-j} h||_\infty ||\partial_x^k w||_2 \sup_{0\le t\le T} ||\partial_x^j w(\cdot,t)||_2 \right].$$

Then, if we assume that $\sup_t ||\partial_x^j w(\cdot,t)||_2 = O(\varepsilon^M)$ for any given $M > 0$ and $0 \le j < k$, we get, since all derivatives of h and w satisfy (5),

$$\sup_{0\le t\le T} ||\partial_x^k w(\cdot,t)||_2 \le c\varepsilon^M \exp[(2k+1)||\partial_x h||_\infty T].$$

Since $\partial_x h$ is of ∞-log-type we infer that $\sup_t ||\partial_x^k w(\cdot,t)||_2 = O(\varepsilon^M)$.

For the mixed derivatives the result follows from the equation, as in the proof of Theorem 3.1. □

Remark 3.5. If $g \in \mathcal{G}_{2,2}(\mathbb{R})$ is, together with g' and g'', of 2-$\sqrt[4]{\log}$-type, then the solution $u \in \mathcal{G}_{2,2}(\mathbb{R} \times (0,T))$ to (1),(9) is such that $\partial_x u$ is of ∞-log-type. In fact, since

$$||\partial_x \hat{u}(\varepsilon,\cdot,\cdot)||_\infty \le \sup_{0\le t\le T} (||\partial_x \hat{u}(\varepsilon,\cdot,t)||_2 + ||\partial_x^2 \hat{u}(\varepsilon,\cdot,t)||_2),$$

we need a logarithmic estimate on both $||\partial_x \hat{u}(\varepsilon,\cdot,t)||_2$ and $||\partial_x^2 \hat{u}(\varepsilon,\cdot,t)||_2$. Using the conserved quantities (12) for $k = 1, 2$, we get that

$$||\partial_x u||_2^2 \le \tfrac{1}{3}(||g||_2 + ||\partial_x u||_2)||g||_2^2 + ||g'||_2^2 + \tfrac{1}{3}||g||_2^3 + \tfrac{1}{3}||g'||_2||g||_2^2$$

and

$$||\partial_x^2 u||_2^2 \le c(||u||_2 + ||\partial_x u||_2)||\partial_x u||_2^2 + c(||u||_2 + ||\partial_x u||_2)^2||u||_2^2$$
$$+||g''||_2^2 + c(||g||_2 + ||g'||_2^2)||g'||_2^2 + c(||g||_2 + ||g'||_2)^2||g||_2^2.$$

Solving the quadratic inequality for $||\partial_x u||_2$ shows that the 2-$\sqrt[4]{\log}$-type of g, g' gives a 2-$\sqrt[2]{\log}$-estimate for $\partial_x u$. This, together with the hypotheses on g'', yields the 2-log-type of $\partial_x^2 u$, and so $\partial_x u$ is of ∞-log-type.

Remark 3.6. As noted in the Introduction, the solutions to the KdV equation are not unique in the algebra $\mathcal{G}_s(\mathbb{R} \times [0,\infty))$, as defined, e.g., in Biagioni [3]. This algebra is constructed in the same way as $\mathcal{G}_{\infty,\infty}$ but the bounds are only required to hold locally. For example, the generalized function u with a representative given by

$$\hat{u}(\varepsilon, x, t) = \frac{3}{\varepsilon}\ \mathrm{sech}^2\left[\frac{1}{2\sqrt{\varepsilon}}\left(x + \frac{1}{\varepsilon} - \frac{1}{\varepsilon}t\right)\right] \tag{17}$$

is a nonzero solution to (1) which belongs to $\mathcal{G}_s(\mathbb{R}\times(0,\infty))$, but its restriction to $t = 0$ is zero in $\mathcal{G}_s(\mathbb{R})$. Indeed, (17) defines a soliton described by the KdV equation (1), where $\frac{1}{\varepsilon}$ is the amplitude and the speed of the wave. At time $t = 0$, its peak is located at $x_o = -\frac{1}{\varepsilon}$. Introducing the notation $\psi(x) = \tanh(x)$, $\varphi(x) = \mathrm{sech}^2(x) = \psi'(x)$, we check immediately that each derivative of φ is of the form $\sum a_{mn}\psi^m\varphi^n$, where $m \ge 0$ and $n \ge 1$. Thus the absolute value of any derivative of $\hat{u}(\varepsilon,\cdot,0)$ is bounded by $c\varepsilon^{-j}|\varphi[\frac{1}{2\sqrt{\varepsilon}}(x + \frac{1}{\varepsilon})]|$ for some $j \in \mathbb{N}$ and some constant c. Since

$$\left|\varphi\left(\frac{1}{2\sqrt{\varepsilon}}\left(x + \frac{1}{\varepsilon}\right)\right)\right| \le \frac{1}{\cosh[(1/2\sqrt{\varepsilon})(x + \frac{1}{\varepsilon})]} \le \frac{1}{\cosh(1/\sqrt{\varepsilon})} \le 2\exp\left(-\frac{1}{\sqrt{\varepsilon}}\right)$$

for $x \ge -\frac{1}{\varepsilon} + 2$, we have that all derivatives of $\hat{u}(\varepsilon,\cdot,0)$ are bounded from above by any positive power of ε, uniformly for x in compact sets, as $\varepsilon \to 0$. Thus $u(\cdot,0)$ is zero in $\mathcal{G}_s(\mathbb{R})$. On the other hand, $\hat{u}(\varepsilon,0,1) = \frac{3}{\varepsilon} \to \infty$ as $\varepsilon \to 0$; thus u is not equal to zero in $\mathcal{G}_s(\mathbb{R} \times [0,\infty))$. □

Our last result in this section relates the generalized solution to the classical solution, if the latter exists.

PROPOSITION 3.7. *If $g \in H^2(\mathbb{R})$, then the solution to* (1), (9) *in* $\mathcal{G}_{2,2}(\mathbb{R}\times(0,T))$ *given in Theorem* 3.1 *is associated with the classical solution* $v \in C([0,T]; H^2(\mathbb{R}))$ *given by* [7], *Corollary* 2 *of Theorem* 8.

Proof. Consider the imbedding ι given in (7); the L^2-norms of $\iota(g)(\varepsilon,\cdot)$, $\iota(g')(\varepsilon,\cdot)$ and $\iota(g'')(\varepsilon,\cdot)$ are bounded independently of ε. By Remark 3.5 and Theorem 3.4, there is a unique solution $u \in \mathcal{G}_{2,2}(\mathbb{R} \times 0, T))$ to (1), with initial data the class of $\iota(g)$ in $\mathcal{G}_{2,2}(\mathbb{R})$.

By Theorem 3 in [7], there is a unique classical solution $\hat{u}_\varepsilon$ to (1) with initial data $\iota(g)(\varepsilon,\cdot)$, which is in $C([0,T]; H^\infty(\mathbb{R}))$. Our generalized solution u has, by construction, $\hat{u}_\varepsilon$ as a representative. Since $\iota(g)(\varepsilon,\cdot) \to g$ in $H^2(\mathbb{R})$ as $\varepsilon \to 0$, it follows from the continuous dependence result in [8] for the case $s = 2$ that the net $(\hat{u}_\varepsilon)$ converges to v in $C([0,T]; H^2(\mathbb{R}))$, hence also in $\mathcal{D}'(\mathbb{R} \times (0,T))$. □

4. Generalized solutions to the regularized long-wave equation.

THEOREM 4.1. *Let $g \in \mathcal{G}_{p,q}(\mathbb{R})$ have a representative $\hat{g} \in \mathcal{E}_{M,p}[\mathbb{R}]$ satisfying*

$$||\hat{g}(\varepsilon,\cdot)||_{1,2} = \left(\int_{-\infty}^{\infty} [\hat{g}(\varepsilon,x)^2 + \hat{g}'(\varepsilon,x)^2]dx\right)^{1/2} = O(\varepsilon^{-N}) \tag{18}$$

as $\varepsilon \to 0$ for some $N \in \mathbb{N}$. Then, for each $T > 0$ there is a solution $u \in \mathcal{G}_{p,q}(\mathbb{R} \times (0,T))$ to (2), (9).

Remarks 4.2. (i) Equation (18) is always satisfied if $p \leq 2$, since then $\mathcal{E}_{M,p}[\mathbb{R}]$ is contained in $\mathcal{E}_{M,2}[\mathbb{R}]$.

(ii) The solutions constructed in Theorem 4.1 for different T are consistent, i.e., if u_T is the solution on the strip $\mathbb{R} \times (0,T)$ and $T_1 < T_2$, then $u_{T_2}|_{\mathbb{R}\times(0,T_1)} = u_{T_1}$. This will follow from the proof below.

Proof of 4.1. For ε small enough, $\hat{g}(\varepsilon,\cdot) \in C^\infty(\mathbb{R})$ satisfies the hypotheses of Theorem 1 in [2]. Thus there is $u_\varepsilon \in C^\infty_B(\mathbb{R} \times [0,T])$, which solves (2) and satisfies $u_\varepsilon(x,0) = \hat{g}(\varepsilon,x)$ for all $x \in \mathbb{R}$, and, for any $t \geq 0$,

$$\|u_\varepsilon(\cdot,t)\|_{1,2} = \|\hat{g}(\varepsilon,\cdot)\|_{1,2}. \tag{19}$$

This, together with (18), implies that, for every $r \in [2,\infty]$,

$$\|u_\varepsilon(\cdot,t)\|_r = O(\varepsilon^{-N}) \tag{20}$$

as $\varepsilon \to 0$. In particular, in the case $p \geq 2$,

$$\|u_\varepsilon(\cdot,t)\|_p = O(\varepsilon^{-N}). \tag{21}$$

We wish to show that (21) is true in the case $1 \leq p < 2$ also. A simple calculation [2, p. 58] shows that u_ε satisfies the integral equation

$$u_\varepsilon(x,t) = \hat{g}(\varepsilon,x) + \int_0^t \int_{-\infty}^\infty K(x-\xi)\left[u_\varepsilon(\xi,\tau) + \frac{1}{2}u_\varepsilon^2(\xi,\tau)\right] d\xi d\tau, \tag{22}$$

where

$$K(x) = \tfrac{1}{2}\,\mathrm{sgn}\, x\, \exp(-|x|).$$

Fixing $R > 0$ and integrating (22), we obtain

$$\begin{aligned}\left(\int_{-R}^R |u_\varepsilon(x,t)|^p dx\right)^{1/p} &\leq \left(\int_{-\infty}^\infty |\hat{g}(\varepsilon,x)|^p dx\right)^{1/p} \\ &\quad + \left(\int_{-R}^R \left|\int_0^t \int_{-\infty}^\infty K(\xi) u_\varepsilon(x-\xi,\tau) d\xi d\tau\right|^p dx\right)^{1/p} \\ &\quad + \frac{1}{2}\left(\int_{-R}^R \left|\int_0^t \int_{-\infty}^\infty K(\xi) u_\varepsilon^2(x-\xi,\tau) d\xi d\tau\right|^p dx\right)^{1/p}.\end{aligned}$$

By the generalized Minkowski inequality, this is estimated by

$$\begin{aligned}&\|\hat{g}(\varepsilon,\cdot)\|_p + \int_0^t \int_{-\infty}^\infty |K(\xi)| \left(\int_{-R-\xi}^{R-\xi} |u_\varepsilon(z,\tau)|^p dz\right)^{1/p} d\xi\, d\tau \\ &\quad + \frac{1}{2}\int_0^t \int_{-\infty}^\infty |K(\xi)| \left(\int_{-\infty}^\infty |u_\varepsilon(x-\xi,\tau)|^{2p} dx\right)^{1/p} d\xi\, d\tau \\ &\leq \|\hat{g}(\varepsilon,\cdot)\|_p + \int_0^t \int_{-\infty}^\infty |K(\xi)| d\xi \left(\int_{-R}^R |u_\varepsilon(z,\tau)|^p dz\right)^{1/p} d\tau \\ &\quad + \int_0^t \int_{-\infty}^\infty |K(\xi)| \left|\int_{-R-\xi}^{-R} |u_\varepsilon(z,\tau)|^p dz - \int_{R-\xi}^R |u_\varepsilon(z,\tau)|^p dz\right|^{1/p} d\xi\, d\tau \\ &\quad + \frac{1}{2}T \int_{-\infty}^\infty |K(\xi)| d\xi \cdot \frac{c}{\varepsilon^N}.\end{aligned}$$

(Here we have used (20), noting that $2p \geq 2$, as well as the inequality $|a+b|^{1/p} \leq |a|^{1/p} + |b|^{1/p}$.) Thus

$$\begin{aligned}\left(\int_{-R}^{R} |u_\varepsilon(x,t)|^p dx\right)^{1/p} &\leq ||\hat{g}(\varepsilon,\cdot)||_p + \int_0^t \left(\int_{-R}^{R} |u_\varepsilon(x,\tau)|^p dx\right)^{1/p} d\tau \\ &\quad + 2^{1/p} T ||u_\varepsilon||_\infty \int_{-\infty}^{\infty} |\xi|^{1/p} |K(\xi)| d\xi + \frac{1}{2} T \frac{c}{\varepsilon^N} \\ &\leq \frac{c'}{\varepsilon^N} + \int_0^t \left(\int_{-R}^{R} |u_\varepsilon(x,\tau)|^p dx\right)^{1/p} d\tau.\end{aligned}$$

By Gronwall's inequality, we obtain

$$\left(\int_{-R}^{R} |u_\varepsilon(x,t)|^p dx\right)^{1/p} \leq \frac{c'}{\varepsilon^N} \exp(T).$$

This holds uniformly in $R > 0$ and $0 \leq t \leq T$. It follows that $u_\varepsilon(\cdot,t) \in L^p(I\!R)$, and (21) holds also in the case $1 \leq p < 2$.

In order to prove (5) for the derivatives of u_ε, we differentiate (22). Thus

$$\partial_t u_\varepsilon(x,t) = [K * (u_\varepsilon + \tfrac{1}{2}u_\varepsilon^2)(\cdot,t)](x).$$

We have that $v_\varepsilon(\cdot,t) := (u_\varepsilon + \frac{1}{2}u_\varepsilon^2)(\cdot,t)$ belongs to $H^1(I\!R)$ and

$$||v_\varepsilon(\cdot,t)||_{1,2} \leq ||u_\varepsilon(\cdot,t)||_{1,2} + c||u_\varepsilon(\cdot,t)||_{1,2}^2. \tag{23}$$

Before going on, let us collect a few useful relations in a lemma.

LEMMA 4.2. *If* $u \in H^1(I\!R)$, *then*

(a) $(K * u)'(x) = u(x) + (E * u)(x)$, *where* $E(x) = -\frac{1}{2}\exp(-|x|)$;

(b) $(E * u)' = K * u$;

(c) $K * u \in H^1(I\!R)$ *and* $||K * u||_{1,2} \leq \sqrt{5}||u||_2$;

(d) $E * u \in H^1(I\!R)$ *and* $||E * u||_{1,2} \leq \sqrt{2}||u||_2$.

These assertions follow from the fact that E is a fundamental solution of $\partial_x^2 - 1$ and $\partial_x E = K$, as well as Young's inequality.

Proof. Going on with the proof of the theorem, we see from Lemma 4.3 (c) that $\partial_t u_\varepsilon(\cdot,t) \in H^1(I\!R)$ and

$$||\partial_t u_\varepsilon(\cdot,t)||_{1,2} \leq \sqrt{5}(||u_\varepsilon(\cdot,t)||_{1,2} + c||u_\varepsilon(\cdot,t)||_{1,2}^2).$$

Employing (20) with $r = p$ and $r = \infty$, we see that

$$||v_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N})$$

as $\varepsilon \to 0$, and so also

$$||\partial_t u_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N})$$

uniformly in $0 \leq t \leq T$. For the mth derivatives ($m = 2, 3, \cdots$) with respect to t, we have

$$\begin{aligned}\partial_t^m u_\varepsilon(x,t) &= [K * \partial_t^{m-1} v_\varepsilon(\cdot,t)](x) \\ &= K * \left[\partial_t^{m-1} u_\varepsilon(\cdot,t) + \sum_{k=0}^{m-2} \binom{m-2}{k} \partial_t^k u_\varepsilon \partial_t^{m-1-k} u_\varepsilon\right](x).\end{aligned} \tag{24}$$

Assuming that, for $k < m$, $\partial_t^k u_\varepsilon(\cdot,t) \in H^1(\mathbb{R})$ and $||\partial_t^k u_\varepsilon(\cdot,t)||_{1,2} \leq c\varepsilon^{-N}$, Lemma 4.3 and the fact that $H^1(\mathbb{R})$ is an algebra imply the same for $\partial_t^m u_\varepsilon(\cdot,t)$. So, if $p \geq 2$, $\partial_t^m u_\varepsilon(\cdot,t) \in L^p(\mathbb{R})$ and

$$||\partial_t^m u_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N}). \tag{25}$$

If $1 \leq p < 2$, assuming that, for $k \leq m-1$, $||\partial_t^k u_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N})$, we have from the inclusion $W^{1,p}(\mathbb{R}) \subset L^\infty(\mathbb{R})$ that $||\partial_t^\ell u_\varepsilon(\cdot,t)||_\infty = O(\varepsilon^{-N})$ for $\ell \leq m-2$. Then the L^p-norm of $\partial_t^{m-1} v_\varepsilon(\cdot,t)$ is less than or equal to

$$||\partial_t^{m-1} u_\varepsilon(\cdot,t)||_p + \sum_{k=0}^{m-2} \binom{m-2}{k} ||\partial_t^k u_\varepsilon||_\infty ||\partial_t^{m-1-k} u_\varepsilon||_p,$$

and thus we get (25) for $1 \leq p < 2$.

For the x-derivatives of u_ε we have, from (22) and Lemma 4.3 (a),

$$\partial_x u_\varepsilon(x,t) = \hat{g}'(\varepsilon,x) + \int_0^t v_\varepsilon(x,\tau)d\tau + \int_0^t (E * v_\varepsilon(\cdot,\tau))d\tau,$$

and for $m \geq 2$, again using (22) and Lemma 4.3 (b),

$$\begin{aligned} \partial_x^m u_\varepsilon(x,t) = \hat{g}^{(m)}(\varepsilon,x) + \int_0^t \partial_x^{m-1} v_\varepsilon(x,\tau)d\tau \\ + \partial_x^{m-2} u_\varepsilon(x,t) - \hat{g}^{(m-2)}(\varepsilon,x). \end{aligned} \tag{26}$$

Since $\hat{g} \in \mathcal{E}_{M,p}[\mathbb{R}]$ and $v_\varepsilon(\cdot,t) \in L^p(\mathbb{R})$, by induction as above we get $||\partial_x^m u_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N})$, uniformly in $0 \leq t \leq T$.

Let us prove (5) now for the mixed derivatives $\partial_t^n \partial_x^m u_\varepsilon$. For $n = 1$, from (26), we obtain

$$\begin{aligned} \partial_t \partial_x^m u_\varepsilon &= \partial_x^{m-1} v_\varepsilon + \partial_t \partial_x^{m-2} u_\varepsilon \\ &= \begin{cases} \partial_x^{m-1} v_\varepsilon + \partial_x^{m-3} v_\varepsilon + \ldots + \partial_x v_\varepsilon + K * v_\varepsilon & \text{if } m \text{ is even} \\ \partial_x^{m-1} v_\varepsilon + \partial_x^{m-3} v_\varepsilon + \ldots + \partial_x^2 v_\varepsilon + v_\varepsilon + E * v_\varepsilon & \text{if } m \text{ is odd.} \end{cases} \end{aligned} \tag{27}$$

Since $||\partial_x^k v_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N})$ for all k, we get the same for $||\partial_t \partial_x^m u_\varepsilon(\cdot,t)||_p$.

Noting that this implies a similar estimate for $||\partial_t \partial_x^m v_\varepsilon(\cdot,t)||_p$ as well, we can now differentiate (27) again with respect to t and obtain that

$$||\partial_t^2 \partial_x^m u_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N}),$$

and so on. This completes the proof of the estimate

$$||\partial^\alpha u_\varepsilon(\cdot,t)||_p = O(\varepsilon^{-N})$$

for every $\alpha \in \mathbb{N}^2$, uniformly in $0 \leq t \leq T$.

Since $||\partial^\alpha u_\varepsilon||_{L^p(\mathbb{R}\times(0,T))} \leq T^{1/p} \sup_{0\leq t\leq T} ||\partial^\alpha u_\varepsilon(\cdot,t)||_p$, we get $\hat{u}(\varepsilon,x,t) := u_\varepsilon(x,t)$ in $\mathcal{E}_{M,p}[\mathbb{R} \times (0,T)]$, and its class in $\mathcal{G}_{p,q}(\mathbb{R} \times (0,T))$ is a solution of (2), (9). □

THEOREM 4.3. (a) *Let* $1 \leq q \leq 2$, $g \in \mathcal{G}_{\infty,q}(\mathbb{R})$, *and* $T > 0$. *Then there exists at most one solution* $u \in \mathcal{G}_{\infty,q}(\mathbb{R} \times (0,T))$ *to* (2), (9) *which is of* ∞*-log-type and has a representative* $\hat{u}$ *such that*

$$\lim_{|x|\to\infty} \hat{u}(\varepsilon,x,t) = 0 \tag{28}$$

for every $\varepsilon > 0$ *and almost all* t, $0 \le t \le T$.

(b) *In particular, if* $1 \le q \le 2$, $1 \le p < \infty$, *and* $g \in \mathcal{G}_{p,q}(I\!R)$, *then there exists at most one solution* $u \in \mathcal{G}_{p,q}(I\!R \times (0,T))$ *to* (2), (9) *which is of* ∞*-log-type.*

(c) *If* $q > 2$, $\frac{1}{p} + \frac{1}{q} \ge 1$, *and* $g \in \mathcal{G}_{p,q}(I\!R)$, *then there exists at most one solution* $u \in \mathcal{G}_{p,q}(I\!R \times (0,T))$ *which is of* ∞*-log-type.*

Proof. We note first that (a) implies (b) since if there are two solutions to (2), (9), $u_1, u_2 \in \mathcal{G}_{p,q}(I\!R \times (0,T))$ which are of ∞-log-type, then all representatives $\hat{u}_1, \hat{u}_2 \in \mathcal{E}_{M,p}[I\!R \times (0,T)]$ satisfy (28). Thus, by (a), $\hat{u}_1 - \hat{u}_2 \in \mathcal{N}_{\infty,q}(I\!R \times (0,T))$, and so $\hat{u}_1 - \hat{u}_2 \in \mathcal{N}_{p,q}(I\!R \times (0,T))$ as well.

Also, in (c) we may assume that $\frac{1}{p} + \frac{1}{q} = 1$. Indeed, if $\frac{1}{p} + \frac{1}{q} \ge 1$ and $u \in \mathcal{G}_{p,q}$, then $u \in \mathcal{G}_{\tilde{p},q}$ for all $\tilde{p} \ge p$, in particular for $\tilde{p} = (1 - \frac{1}{q})^{-1}$.

Let $u_1, u_2 \in \mathcal{G}_{p,q}(I\!R \times (0,T))$ be two solutions of (2),(9) with respective representatives $\hat{u}_1, \hat{u}_2 \in \mathcal{E}_{M,p}[I\!R \times (0,T)]$ satisfying (28) and (8) with $r = \infty$ and $j = 1$. There are $n_1 \in \mathcal{N}_{p,q}(I\!R \times (0,T))$ and $n_2 \in \mathcal{N}_{p,q}(I\!R)$ such that, if $w = \hat{u}_1 - \hat{u}_2$,

$$\begin{aligned} \left\{w_t + w_x + \tfrac{1}{2}[(\hat{u}_1 + \hat{u}_2)w]_x - w_{xxt}\right\}(\varepsilon,x,t) &= n_1(\varepsilon,x,t), \\ w(\varepsilon,x,0) &= n_2(\varepsilon,x). \end{aligned} \tag{29}$$

By a change of representatives we can assume $n_2 \equiv 0$. Multiplying (29) by w and integrating from $-R$ to R, we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\int_{-R}^{R}(w^2 + w_x^2)dx = &\left[w_{xt}w - \frac{1}{2}w^2 - \frac{1}{2}(\hat{u}_1 + \hat{u}_2)w^2\right]_{x=-R}^{x=R} \\ &+\frac{1}{2}\int_{-R}^{R}(\hat{u}_1 + \hat{u}_2)ww_x dx + \int_{-R}^{R} wn_1 dx. \end{aligned}$$

Thus

$$\begin{aligned} \int_{-R}^{R}(w^2 + w_x^2)dx \le 2\int_0^T &\left[w_{xt}w - \frac{1}{2}w^2 - \frac{1}{2}(\hat{u}_1 + \hat{u}_2)w^2\right]_{x=-R}^{x=R} d\tau \\ &+\int_0^t\int_{-R}^{R}(\hat{u}_1 + \hat{u}_2)w\, w_x\, dx\, d\tau \\ &+2\int_0^t\int_{-R}^{R} w\, n_1\, dx\, d\tau. \end{aligned}$$

If $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\int_0^T\int_{-R}^{R}|wn_1|dx\, d\tau \le ||w||_p||n_1||_q;$$

if $p = \infty$ and $1 \le q \le 2$, since $\mathcal{N}_{p,q} \subset \mathcal{N}_{p,2}$ we have

$$\begin{aligned} \int_0^t\int_{-R}^{R}|wn_1|dx\, d\tau &\le \int_0^t\left(\int_{-R}^{R}w^2dx\right)^{1/2}\left(\int_{-\infty}^{\infty}n_1^2(x,\tau)dx\right)^{1/2}d\tau \\ &\le \int_0^t\left(1 + \int_{-R}^{R}w^2(x,\tau)dx\right)||n_1||_2 d\tau. \end{aligned}$$

We have used the estimate $\sqrt{a} \le 1 + a$ in the last inequality. We end up with

$$\int_{-R}^{R}(w^2+w_x^2)(x,t)dx \le 2\int_0^T \left[w_{xt}w - \frac{1}{2}w^2 - \frac{1}{2}(\hat{u}_1+\hat{u}_2)w^2\right]_{x=-R}^{x=R} d\tau$$
$$+\begin{cases} ||\hat{u}_1+\hat{u}_2||_\infty \int_0^t\int_{-R}^R (w^2+w_x^2)(x,\tau)d\tau + 2||w||_p||n_1||_q \\ \text{or} \\ (||\hat{u}_1+\hat{u}_2||_\infty + 2||n_1||_2)\int_0^t\int_{-R}^R(w^2+w_x^2)(x,\tau)d\tau + 2T||n_1||_2\,. \end{cases}$$

Here the first line applies to the case $\frac{1}{p}+\frac{1}{q}=1$, while the second applies to $p=\infty, 1\le q\le 2$. In any case, Gronwall's inequality gives that

$$\int_{-R}^{R}(w^2+w_x^2)(x,t)dx \le \left\{2\int_0^T \left[w_{xt} - \frac{1}{2}w^2 - \frac{1}{2}(\hat{u}_1+\hat{u}_2)w^2\right]_{x=-R}^{x=R} d\tau \right.$$
$$\left. + \begin{bmatrix} 2||w||_p||n_1||_q \\ \text{or} \\ 2T||n_1||_2 \end{bmatrix}\right\} \exp T \begin{bmatrix} ||\hat{u}_1+\hat{u}_2||_\infty \\ \text{or} \\ ||\hat{u}_1+\hat{u}_2||_\infty + 2||n_1||_2 \end{bmatrix}.$$

This estimate holds for $0\le t\le T$ and $R>0$. In particular, we may let $R\to\infty$ and obtain, since $w(\varepsilon,\cdot,t)\to 0$ as $|x|\to\infty$,

$$||w(\cdot,t)||_{1,2} \le \begin{bmatrix} 2||w||_p||n_1||_q \\ \text{or} \\ 2T||n_1||_2 \end{bmatrix} \exp(||\hat{u}_1+\hat{u}_2||_\infty + 2)T.$$

Introducing the dependence on ε again, we obtain from (8) that

$$||w(\varepsilon,\cdot,t)||_{1,2} = O(\varepsilon^M)$$

for every M, uniformly in $0\le t\le T$.

Thus it is clear that $||w(\varepsilon,\cdot,t)||_\infty = O(\varepsilon^M)$ as well. In order to go on with case (c), we want to show that actually $||w(\varepsilon,\cdot,t)||_q = O(\varepsilon^M)$ for every $M\in I\!N$. The following identity is derived similar to (22):

$$w(\varepsilon,x,t) = \int_0^t\int_{-\infty}^{\infty} K(\xi)\left[1+\frac{1}{2}(\hat{u}_1+\hat{u}_2)\right] w(\varepsilon,x-\xi,\tau)d\xi\, d\tau$$
$$-\int_0^t\int_{-\infty}^{\infty} E(x-\xi)n_1(\xi,\tau)d\xi\, d\tau. \tag{30}$$

Thus, setting $A = 1+\frac{1}{2}||\hat{u}_1(\varepsilon,\cdot)||_\infty + \frac{1}{2}||\hat{u}_2(\varepsilon,\cdot)||_\infty$,

$$\left(\int_{-R}^{R}|w(\varepsilon,x,t)|^q dx\right)^{1/q} \le A\left(\int_{-R}^{R}\left|\int_0^t\int_{-\infty}^{\infty}K(\xi)w(\varepsilon,x-\xi,\tau)d\xi d\tau\right|^q dx\right)^{1/q}$$
$$+\left(\int_{-R}^{R}\left|\int_0^t[E*n_1(\varepsilon,\cdot,\tau)](x)d\tau\right|^q dx\right)^{1/q}$$
$$\le A\int_0^t\int_{-\infty}^{\infty}\left(\int_{-R}^{R}|K(\xi)w(\varepsilon,x-\xi,\tau)|^q dx\right)^{1/q} d\xi\, d\tau$$
$$+\int_0^t ||E*n_1(\varepsilon,\cdot,\tau)||_q d\tau.$$

We have used the generalized Minkowski inequality in both terms. Now we will use Young's inequality and Hölder's inequality in the second term. The first term has already been calculated in the proof of Theorem 4.1 with u_ε instead of $w(\varepsilon,\cdot)$. Thus the last expression is estimated by

$$\begin{aligned}
&\leq A\Big(\int_0^t \Big(\int_{-R}^{R} |w(\varepsilon,x,\tau)|^q dx\Big)^{1/q} d\tau + 2^{1/q}T||w(\varepsilon,\cdot)||_\infty \int_{-\infty}^{\infty} |\xi|^{1/q}|K(\xi)|d\xi\Big) \\
&\quad + t^{1/p}||n_1(\varepsilon,\cdot)||_q \\
&\leq T^{1/p}||n_1(\varepsilon,\cdot)||_q + cT||w(\varepsilon,\cdot)||_\infty + A\int_0^t \Big(\int_{-R}^{R} |w(\varepsilon,x,\tau)|^q dx\Big)^{1/q} d\tau\ .
\end{aligned}$$

From Gronwall's inequality,

$$\Big(\int_{-R}^{R} |w(\varepsilon,x,t)|^q dx\Big)^{1/q} \leq \Big(T^{1/p}||n_1(\varepsilon,\cdot)||_q + cT||w(\varepsilon,\cdot)||_\infty\Big)\exp(AT).$$

This holds for every $R > 0$. Thus, since u_1, u_2 are of ∞-log-type, we get

$$||w(\varepsilon,\cdot,t)||_q = O(\varepsilon^M)$$

for every $M \in I\!N$.

In order to get estimates for the derivatives of w, we differentiate the integral equation (30). The t-derivatives of w satisfy:

$$\begin{aligned}
\partial_t^m w(\varepsilon,x,t) = &\int_{-\infty}^{\infty} K(x-\xi)\partial_t^{m-1}\Big[w + \frac{w}{2}(\hat{u}_1 + \hat{u}_2)\Big](\varepsilon,\xi,t)d\xi \\
&-\int_{-\infty}^{\infty} E(x-\xi)\partial_t^{m-1}n_1(\xi,t)d\xi.
\end{aligned}$$

The x-derivatives of w satisfy

$$\begin{aligned}
\partial_x w = &\int_0^t \int_{-\infty}^{\infty} E(x-\xi)\Big[w + \frac{w}{2}(\hat{u}_1 + \hat{u}_2)\Big] d\xi d\tau + \int_0^t \Big[w + \frac{w}{2}(\hat{u}_1 + \hat{u}_2)\Big] d\tau \\
&-\int_0^t \int_{-\infty}^{\infty} K(x-\xi)n_1 d\xi d\tau;
\end{aligned}$$

and, for $m \geq 2$ (using Lemma 4.3),

$$\partial_x^m w = \partial_x^{m-2}w + \int_0^t \partial_x^{m-1}\Big[w + \frac{w}{2}(\hat{u}_1 + \hat{u}_2)\Big] d\tau - \int_0^t \partial_x^{m-2} n_1 dx.$$

We now prove that $||\partial^\alpha w(\varepsilon,\cdot)||_q = O(\varepsilon^M)$ by induction, using the L^q-estimates on w, n_1, and the L^∞-estimates on $\hat{u}_1, \hat{u}_2$. □

Remark 4.5. If $g \in \mathcal{G}_{p,q}(I\!R)$ has a representative $\hat{g}$ such that

$$||\hat{g}(\varepsilon,\cdot)||_{1,2} = O(|\log\varepsilon|)$$

as $\varepsilon \to 0$, then the solution u to (2), (9), assured by Theorem 4.1, is of ∞-log-type; thus it is unique by Theorem 4.4. This follows since the representative $\hat{u}$ given in the proof of Theorem 4.1 satisfies

$$||\hat{u}(\varepsilon,\cdot)||_\infty \leq ||\hat{g}(\varepsilon,\cdot)||_{1,2}.$$

Remark 4.6. The question of uniqueness remains open in the case $q > 2$ and $\frac{1}{p} + \frac{1}{q} < 1$.

5. The equation $u_t + uu_x + \nu u_{xxx} = 0$.

PROPOSITION 5.1. *Let ν be a generalized constant with a representative $\hat{\nu}$ satisfying: there are $N \in I\!N$ and $\eta > 0$ such that*

$$\varepsilon^N \leq \hat{\nu}(\varepsilon) \leq \varepsilon^{-N}, \qquad 0 < \varepsilon < \eta. \tag{31}$$

Let $g \in \mathcal{G}_{2,2}(I\!R)$. Then there is a solution $u \in \mathcal{G}_{\infty,2}(I\!R \times (0,\infty))$ to the problem

$$\begin{cases} u_t + uu_x + \nu u_{xxx} = 0, \\ u|_{t=0} = g \end{cases} \tag{32}$$

such that $u|_{I\!R\times(0,T)} \in \mathcal{G}_{2,2}(I\!R \times (0,T))$ for every $T > 0$. Moreover, for every $T > 0$ there is at most one solution $u \in \mathcal{G}_{2,2}(I\!R \times (0,T))$ such that $\partial_x u$ is of ∞-log-type.

Proof. Let

$$\hat{h}(\varepsilon, x) = \frac{1}{\sqrt[3]{\hat{\nu}(\varepsilon)}}\hat{g}(\varepsilon, \sqrt[3]{\hat{\nu}(\varepsilon)}x),$$

where $\hat{g}$ is a representative of g. By Theorem 3.1 there is a solution $v \in \mathcal{G}_{\infty,2}(I\!R \times (0,\infty))$ to (1) with initial data $h \in \mathcal{G}_{2,2}(I\!R)$, the class of $\hat{h}$, such that $v|_{I\!R\times(0,T)} \in \mathcal{G}_{2,2}(I\!R \times (0,T))$. Set

$$\hat{u}(\varepsilon, x, t) = \sqrt[3]{\hat{\nu}(\varepsilon)}\hat{v}\left(\varepsilon, \frac{x}{\sqrt[3]{\hat{\nu}(\varepsilon)}}, t\right), \tag{33}$$

where $\hat{v} \in \mathcal{E}_{M,\infty}[I\!R \times (0,\infty)]$ is the representative of v which satisfies

$$\begin{aligned} (\hat{v}_t + \hat{v}\hat{v}_x + \hat{v}_{xxx})(\varepsilon, x, t) &= 0, \\ \hat{v}(\varepsilon, x, 0) &= \hat{h}(\varepsilon, x). \end{aligned}$$

Then

$$(\hat{u}_t + \hat{u}\hat{u}_x + \hat{\nu}\hat{u}_{xxx})(\varepsilon, x, t) = 0$$

and

$$\hat{u}(\varepsilon, x, 0) = \hat{g}(\varepsilon, x).$$

Let us prove that $\hat{u} \in \mathcal{E}_{M,\infty}[I\!R \times (0,\infty)] \cap \mathcal{E}_{M,2}[I\!R \times (0,T)]$. As we observed in the proof of Theorem 3.1, it suffices to prove (10). But that obviously holds since

$$||\partial_x^k \hat{u}(\varepsilon, \cdot, t)||_2 = (\sqrt[3]{\hat{\nu}(\varepsilon)})^{\frac{3}{2}-k}||\partial_x^k \hat{v}(\varepsilon, \cdot, t)||_2.$$

To prove uniqueness, it suffices to observe that if u solves (32) and is related to v via (33), then v is a solution of the KdV equation. Also, if $\partial_x u$ is of ∞-log-type, so is $\partial_x v$. By Theorem 3.4, v is unique; hence u is unique. □

COROLLARY 5.2. *Let ν and g be as in Proposition 5.1, and assume further that $\nu \approx 0$ and g has a representative $\hat{g}$ such that $||\hat{g}(\varepsilon, \cdot)||_2$ is bounded independently of ε. Then the solution $u \in \mathcal{G}_{\infty,2}(I\!R \times (0,\infty))$ to (32) satisfies*

$$\begin{cases} u_t + uu_x \approx 0, \\ u|_{t=0} = g. \end{cases} \tag{34}$$

Proof. Indeed, $\|\hat{u}(\varepsilon,\cdot,t)\|_2 = \|\hat{g}(\varepsilon,\cdot)\|_2$ is bounded independently of ε and t, thus $\nu u_{xxx} \approx 0$. □

Corollary 5.2 has two interesting aspects, the first one relating to results of Lax and Levermore [11] and Venakides [15] on the zero dispersion limit, the second one to the possibility of modeling what Maslov and his collaborators [12], [13] have called infinitely narrow solitons.

To discuss the first, assume that the initial data are classical, say, $g \in W^{\infty,2}(\mathbb{R})$, and let $\nu \approx 0$. By Corollary 5.2, the solution $u \in \mathcal{G}_{\infty,2}(\mathbb{R}\times(0,\infty))$ to (32) solves (34) as well. On the other hand, if g satisfies the hypotheses required in the articles of Lax and Levermore [11] or Venakides [15], the representative $\hat{u}(\varepsilon,\cdot,\cdot)$ will converge weakly to a function v which, however, will generally not solve the equation $v_t + vv_x = 0$ in the sense of distributions. Thus, in this situation our generalized solution u satisfies (4), while for its associated distribution v we will have as a rule that

$$v_t + (\tfrac{1}{2}v^2)_x \neq 0 \quad \text{in } \mathcal{D}'(\mathbb{R}\times(0,\infty)).$$

This is in contrast to the situation arising with zero viscosity limits, i.e., solutions of Burgers' equation $u_t + uu_x = \nu u_{xx}$ with $\nu \approx 0$. There the generalized solution u with classical initial data g both solves (34) and admits an associated distribution w which always satisfies $w_t + (\frac{1}{2}w^2)_x = 0$ in $\mathcal{D}'(\mathbb{R}\times(0,\infty))$; see [4]. It is clear from this discussion that solutions to (34) are not unique, although some limited uniqueness in the sense of association can be obtained in the class of zero viscosity limits [4].

Secondly, following Maslov and Omel'yanov [12], Maslov and Tsupin [13], we shall see that the notion of "infinitely narrow solitons" can be accommodated in our approach as well. To simplify the presentation, we shall take ν in (32) as the class of $\hat{\nu}(\varepsilon) = \varepsilon$. We note that for each $c \geq 0$, the function

$$\hat{u}(\varepsilon,x,t) = 3c\,\mathrm{sech}^2\left(\frac{\sqrt{c}}{2\sqrt{\varepsilon}}(x-ct)\right)$$

is a representative of a solution u in $\mathcal{G}_{\infty,2}(\mathbb{R}\times(0,\infty))$ to (32); actually, a soliton for each fixed $\varepsilon > 0$. As $\varepsilon \to 0$, the width of the soliton decreases to zero, while its amplitude remains fixed. In fact, in the sense of generalized pointvalues of elements of $\mathcal{G}_{\infty,2}(\mathbb{R}\times(0,\infty))$, we have that

$$\begin{aligned} u(x,t) &= 0 \quad \text{if } x \neq ct, \\ u(ct,t) &= 3c \end{aligned}$$

for all $x \in \mathbb{R}$ and $t \geq 0$, where the evaluation of the generalized function u at (x,t) is naturally defined as a generalized constant and the equality is understood in the corresponding sense [3, §1.5.1]. It is not difficult to check that $u \neq 0$ in $\mathcal{G}_{\infty,2}(\mathbb{R}\times(0,T))$ for every $T > 0$, while $u \approx 0$, and (4) is satisfied. In addition, the generalized function

$$\frac{1}{3\sqrt{c\varepsilon}}\hat{u}(\varepsilon,x,t)$$

is associated with the Dirac measure along the line $x = ct$. This is a special case of the asymptotic developments elaborated in the quoted articles [12], [13].

Finally, we point out that it is also possible to have solutions to (32) and (34), which are associated with the Dirac measure at $t = 0$, but associated with zero on

$t > 0$: the impact of the initial singularity is removed immediately after time $t = 0$. It is easily verified that the class u in $\mathcal{G}_{\infty,2}(I\!\!R \times (0,\infty))$ of

$$\hat{u}(\varepsilon, x, t) = \frac{\alpha^2}{48\hat{\nu}(\varepsilon)} \operatorname{sech}^2\left(\frac{\alpha}{24\hat{\nu}(\varepsilon)}(x - \frac{\alpha^2}{144\hat{\nu}(\varepsilon)}t)\right)$$

with $\nu \approx 0$ and $\alpha \in \mathbb{C}$ may serve as an example: u solves (32) and (34) with $u|_{t=0} = g \approx \alpha\delta$, and $u \approx 0$ on $I\!\!R \times (0,\infty)$, where $\delta \in \mathcal{D}'(I\!\!R)$ denotes the Dirac measure at the origin.

Acknowledgment. The authors are indebted to the referee for pointing out the connections with the results of Lax, Levermore, and Venakides in §5.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans. Roy. Soc. London, A 272 (1972), pp. 47–78.

[3] H. A. Biagioni, *A Nonlinear Theory of Generalized Functions*, Lecture Notes in Math. 1421, Springer-Verlag, Berlin, 1990.

[4] H. A. Biagioni and M. Oberguggenberger, *Generalized solutions to Burgers' equation*, J. Differential Equations, to appear.

[5] J. L. Bona, *On solitary waves and their role in the evolution of long waves*, in Applications of Nonlinear Analysis in the Physical Sciences, H. Amann, N. Bazley, and K. Kirchgässner, eds., Pitman, London, 1981, pp. 183–205.

[6] J. L. Bona, W. G. Pritchard, and L. R. Scott, *A comparison of solutions of two model equations for long waves*, Lectures in Appl. Math., 20 (1983), pp. 235–267.

[7] J. L. Bona and R. Smith, *The initial-value problem for the Korteweg–de Vries equation*, Philos. Trans. Roy. Soc. London, A 278 (1975), pp. 555–601.

[8] J. L. Bona and R. Scott, *Solutions of the Korteweg–de Vries equation in fractional order Sobolev spaces*, Duke Math. J., 43 (1976), pp. 87–99.

[9] J. F. Colombeau, *New Generalized Functions and Multiplication of Distributions*, North-Holland Math. Stud. 84, North-Holland, Amsterdam, 1984.

[10] ———, *Elementary Introduction to New Generalized Functions*, North-Holland Math. Stud. 113, Amsterdam, 1985.

[11] P. D. Lax and C. D. Levermore, *The small dispersion limit of the Korteweg–de Vries equation*, I–III, Comm. Pure Appl. Math., 36 (1983), pp. 253–290, 571–593, 809–830.

[12] V. P. Maslov and G. A. Omel'yanov, *Asymptotic soliton-form solutions of equations with small dispersion*, Russian Math. Surveys, 36 (1981), pp. 73–119.

[13] V. P. Maslov and V. A. Tsupin, *Necessary conditions for the existence of infinitely narrow solitons in gas dynamics*, Soviet Phys. Dokl., 24 (1979), pp. 354–356.

[14] R. M. Miura, C. S. Gardner, and M. D. Kruskal, *The Korteweg–de Vries equation and generalizations. II: Existence of conservation laws and constants of motion*, J. Math. Phys., 9 (1968) pp. 1204–1209.

[15] S. Venakides, *The zero dispersion limit of the Korteweg–de Vries equation for initial potentials with non-trivial reflection coefficient*, Comm. Pure Appl. Math., 38 (1985), pp. 125–155.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 941-949, July 1992

008

NONCONSERVATIVE PRODUCTS IN BOUNDED VARIATION FUNCTIONS*

JEAN FRANÇOIS COLOMBEAU† AND ARNAUD HEIBIG†

Abstract. There exist two definitions of products of a bounded variation function by a derivative of another bounded variation function. One of them follows from a concept of generalized functions in which arbitrary products of distributions make sense: one has only one product but its understanding involves a nonclassical concept contained in each generalized function. Another one has been recently introduced by Dal Maso, Le Floch, and Murat as a generalization of a definition of Volpert; one has a family of different products indexed by a "path ϕ"; each ϕ-product is well defined as a measure, and the scenario takes place in the familiar framework of functions of bounded variation and measures. In spite of their apparent great difference, these products are closely related: the purpose of this paper is to prove a clear correspondence between the two approaches: the path ϕ is the nonclassical ingredient inherent in the concept of generalized functions.

Key words. products of distributions, nonconservative hyperbolic equations, shock waves, generalized solutions

AMS(MOS) subject classifications. 35L65, 46F10

1. Introduction. Nonconservative shocks appear in some formulations of engineering problems; see [1]-[4], [8]-[10], [19], among other references. Since they may be viewed as problems of multiplications of distributions, a first attempt is to use the theory developed in the expository texts [3], [5]-[7], [18]. This attempt led to a numerical and physical theory of nonconservative shocks and generalized solutions of PDEs (see [3], [5], [11] and the references therein). This approach is based on a new concept of generalized functions whose construction from the classical concept of C^∞ functions is suggestive of the construction of the real number system from the rational numbers. Therefore, the price to pay is the acceptance of new objects of pure mathematics. Attempts to show how this concept of "new generalized functions" can be handled on an intuitive level in applications to physics and engineering are given in [2], [7]-[10].

Another definition of a nonconservative product was introduced in [20], [14], [15], [13], [16], [17]. The first version [14], [15] applies a definition introduced by Volpert [20]. The second version [13], [16], [17] is more general and allows various different results for the product $Y^n\delta$, according to the variant in use (note that the fact that products $Y^n\delta$ appearing in equations of physics give different results according to the context has been stressed in [3, pp. 107, 119] and [2], [5], [8], [9]). At first sight it does not involve any concept of generalized functions since it takes place in the familiar setting of bounded variation functions on the real line. However, a choice has to be made concerning the variant in use for defining any nonconservative product (this choice is called a "path" in [13], [16], [17]).

In this paper we prove that this second definition is deeply connected with the definition using generalized functions. The choice of the path in the second definition is the nonclassical ingredient contained in a new generalized function. This result is interesting since the two approaches are often presented and considered as quite different. The definition in the setting of boundary value (BV) functions and paths is

* Received by the editors July 23, 1990; accepted for publication (in revised form) October 10, 1991.

† Ecole Normale Supérieure de Lyon, 46, allée d'Italie, 69364 Lyon Cedex 07, France.

not compatible with the differential-algebraic structure of the set of generalized functions, useful for calculations: it reflects the association in $\mathcal{G}$, not the equality. This differential and integral calculus of generalized functions (and its novelty relative to the differential and integral calculus of distributions) is at the origin of most of its applications [references above]; in particular, it is at the basis of the modeling assumptions proposed in [2], [3], [5], [8] for the resolution of ambiguities. The result in this paper and its proof show that these assumptions amount to the choice of a path ϕ to define a nonconservative product according to [13], [16], [17]. Therefore, there is a clear correspondence between the two approaches, when the theory of generalized functions is restricted to the setting of bounded variation functions. In the opinion of the authors, the explicit use of generalized functions is clear and simple, and can be easily dealt with on an intuitive level. Also, the differential and integral calculus of generalized functions contains basic ingredients not yet considered in the path formulation that could perhaps be productively incorporated into the path formulation. Note that the situation looks somewhat similar when the approach of generalized functions is compared with the concept of Di Perna's measure valued solutions for systems of conservation laws [12].

2. Statement of the Theorem. The notation in the setting of BV functions is that of [13], [16], [17] (briefly recalled in this paper), and the notation in the setting of "generalized functions" is that of [8, Chap. 1] (recalled in the Appendix). The basic idea concerning the connection between the two approaches can be understood without a detailed knowledge of the definitions (§ 3).

Let $\phi:[0,1]\times\mathbb{R}^p\times\mathbb{R}^p\to\mathbb{R}^p$ be a fixed "family of paths" (a "path" for short) satisfying the assumptions (Hypothesis 1-2-3) in [13], [16], [17]: essentially the map ϕ has to be differentiable in the variable $s\in[0,1]$ and $\partial\phi/\partial s$ has to be Lipschitz. Let $u:\,]a,b[\,\to\mathbb{R}^p$ be a locally bounded Borel function. The nonconservative product $\mu=[g(u,x)(du/dx)]_\phi$ is defined in [13], [16], [17] as a real valued bounded Borel measure, for any BV function $u:\,]a,b[\,\to\mathbb{R}^p$ ($g(u,x)(du/dx)$ has to be read as $\sum_{1\leqq i\leqq p} g^i(u,x)(du^i/dx)$).

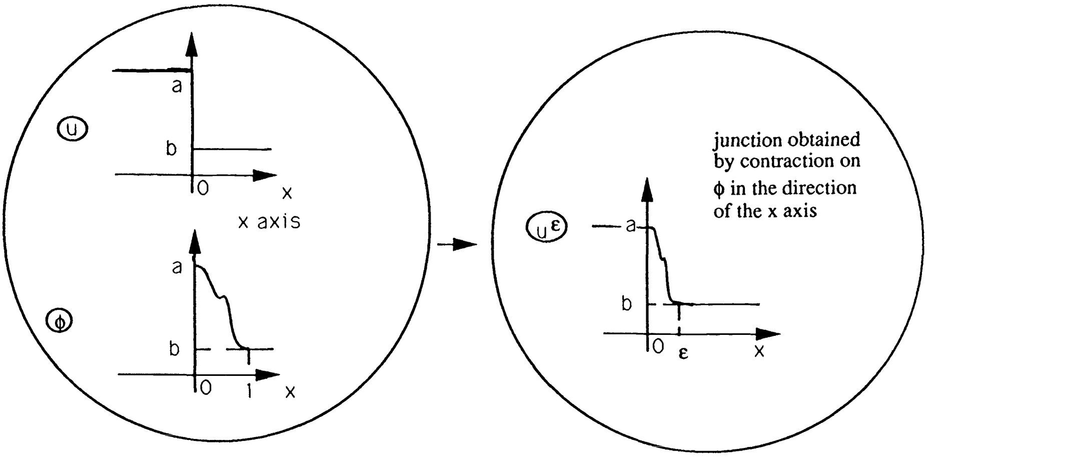
Let us look at a special case: the points of discontinuity of u. If x_0 is a point of discontinuity of u then we have by definition

$$\mu(\{x_0\})=\int_0^1 g(\phi(s;u(x_0^-),u(x_0^+)),x_0)\cdot\frac{\partial\phi}{\partial s}(s;u(x_0^-),u(x_0^+))\,ds. \tag{1}$$

The path ϕ is used inside this formula to smoothly join $u(x_0^-)$ and $u(x_0^+)$, as is made clear in Fig. 1.

In the theory of generalized functions of [3], [5]-[7], [18], the nonconservative product $g(U,x)(dU/dx)$ makes sense as a generalized function, if U is a generalized function from $]a,b[$ to $\mathbb{R}^p$ (vector valued generalized functions are of course defined as a set of components; for arbitrary $U\in\mathcal{G}(]a,b[;\mathbb{R}^p)$ some growth property on g is requested, but for the particular U considered here this restriction does not enter.) Here we shall consider only that case g is a C^∞ function; this is done only to simplify the exposition: the case of non-$C^\infty g$ can be handled as in [6, pp. 258-259].

The space $\mathcal{G}(]a,b[)$ of generalized functions from $]a,b[$ into $\mathbb{R}$ contains the space $\mathcal{D}'(]a,b[)$ of (real valued) distributions. Hence the measure $[g(u,x)(du/dx)]_\phi$ is a well-defined element of $\mathcal{G}(]a,b[)$. Any generalized function G is a class of representatives $\{R(\varepsilon,\cdot)\}_{0<\varepsilon<1}$ (see the Appendix), where the functions $R(\varepsilon,\cdot)$ are C^∞ functions. If $G\in\mathcal{G}(]a,b[)$ and if $\theta\in\mathcal{D}(]a,b[)$, the brackets $\langle G,\theta\rangle\in\mathbb{R}$ are defined as the limit—when it exists—of $\int R(\varepsilon,x)\theta(x)\,dx$ when $\varepsilon\to0$ (then it is independent on the choice



u
a
b
0
x
x axis
ϕ
a
b
0
1
x
u^ε
a
b
0
ε
x
junction obtained by contraction on ϕ in the direction of the x axis

FIG. 1. *Passage from* (u, ϕ) *to* u^ε *or "How a pair* (u, ϕ) *gives a generalized function* U *(defined as the class of the family* (u^ε)*)"; the figure is drawn in the case* $p = 1$.

of a representative of G). If $G \in \mathscr{D}'(]a, b[)$, then the above $\langle G, \theta\rangle$ coincides with the classical brackets of distribution theory.

Now we are able to state our comparison theorem; note that the case of $g(u, x)(dv/dx)$ can be reduced at once to the case $u = v$. (Since we consider vector valued objects.)

THEOREM. *Let* (u, ϕ) *be a given pair* (BV *function, path*). *Then there exists a generalized function* $U \in \mathscr{G}(]a, b[)$ *such that, for every* C^∞ *function* g, *the measure* $[g(u, x)(du/dx)]_\phi \in \mathscr{D}'(]a, b[)$ *and the generalized function* $g(U, x)(dU/dx) \in \mathscr{G}(]a, b[)$ *are equal in the following sense: for every test function* $\theta \in \mathscr{D}(]a, b[)$,

$$\left\langle \left[g(u, \cdot)\frac{du}{dx}\right]_\phi, \theta \right\rangle = \left\langle g(U, \cdot)\frac{dU}{dx}, \theta \right\rangle.$$

The above equality can be considered as an equality in the sense of distributions, although $g(U) \cdot (dU/dx)$ is not a distribution. The function u is—in a sense to be made precise later—the "macroscopic aspect" of the generalized function U, while ϕ is its "microscopic aspect" on its points of discontinuity. Note that the theorem gives only a correspondence between (u, ϕ) and U; this seems to be so because an analogue of the concept of (strong) equality in $\mathscr{G}$ does not exist in the other approach: this would require a concept more precise than the pairs (u, ϕ).

The proof is somewhat technical because the set of discontinuities of an arbitrary BV function can be rather complicated. In the case of step functions, if some details are dropped the proof is quite simple and gives a clear understanding of the connection between the two frameworks. The proof of this special case is given in the next section; the complete proof is in § 4. The x-dependence of g is always dropped to simplify the notation.

3. Sketch of proof in a particular case. Let $u = (u^1, \cdots, u^p)$ be a piecewise constant function of one real variable with a discontinuity at the origin; let $\phi = (\phi^1, \cdots, \phi^p)$ be a "path." For $1 \leqq i \leqq p$, $0 < \varepsilon \leqq 1$, and $0 \leqq \xi \leqq \varepsilon$, let

$$\Psi^{i,\varepsilon}(\xi) = \phi^i\left(\frac{\xi}{\varepsilon}, u(0^-), u(0^+)\right). \tag{2}$$

Define the functions $u^{i,\varepsilon}$ by

$$(2') \qquad u^{i,\varepsilon}(x)=\begin{cases} u^i(x) & \text{if } x<0 \text{ or } x>\varepsilon, \\ \Psi^{i,\varepsilon}(x) & \text{if } 0\leqq x\leqq \varepsilon. \end{cases}$$

From definition [13], [16], [17] of a path the functions $u^{i,\varepsilon}$ are continuous. For simplicity (i.e., to avoid a regularization, as in the general proof), let us assume that the functions $u^{i,\varepsilon}$ are C^∞ functions.

Let $g=(g^i)_{1\leqq i\leqq p}$ be a C^∞ function from $\mathbb{R}^p$ to $\mathbb{R}^p$. By definition [13], [16], [17], the measure $[g(u)(du/dx)]_\phi$ is equal to $\sum_{1\leqq i\leqq p}\mu^i(\{0\})\delta$, δ the Dirac measure, where

$$(1') \qquad \mu^i(0)=\int_0^1 g^i(\phi(s,a,b))\frac{\partial\phi^i}{\partial s}(s,a,b)\,ds \quad \text{if } a=u(0^-),\, b=u(0^+).$$

By definition in the context of generalized functions, $\langle g(U)(dU/dx),\theta\rangle$ is the limit when $\varepsilon\to 0$ of

$$I=\sum_{1\leqq i\leqq p} I_\varepsilon^i, \qquad I_\varepsilon^i=\int_{-\infty}^{+\infty} g^i(u^\varepsilon(x))(u^{i\varepsilon})'(x)\theta(x)\,dx$$

if $\theta\in\mathcal{D}(\mathbb{R})$.

$$\begin{aligned} I_\varepsilon^i &= \int_0^\varepsilon g^i((\Psi^{i,\varepsilon}(x))_{1\leqq i\leqq p})\frac{d}{dx}\psi^{i,\varepsilon}(x)\theta(x)\,dx \\ &= \int_0^\varepsilon g^i\left(\phi\left(\frac{x}{\varepsilon},a,b\right)\right)\frac{\partial\phi^i}{\partial s}(s,a,b)\theta(x)\,dx \\ &= \int_0^1 g^i(\phi(s,a,b))\frac{\partial\phi^i}{\partial s}(s,a,b)\theta(\varepsilon s)\,ds. \end{aligned}$$

Thus $I_\varepsilon^i\to\mu(\{0\})\theta(0)$ when $\varepsilon\to 0$: we obtain

$$\left\langle\left[g(u)\frac{du}{dx}\right]_\phi,\theta\right\rangle.$$

Note that the construction of U is not canonical because the replacement of $[0,\varepsilon]$ by, for instance, $[-\varepsilon/2,\varepsilon/2]$ would work as well (and would produce another U). One sees very clearly that the role of the path ϕ is merely to join smoothly—by its contraction to $[0,\varepsilon]-u(0^-)$ and $u(0^+)$; this concept of path, called "microscopic profile" of the shock in [3], [9], [10] is contained in the generalized function U, but not contained in the BV function u. Different ϕ's produce different U's; the same u and ϕ may produce different U's. Thus the pairs (u,ϕ) are not generalized functions, but rather classes of generalized functions having the same macroscopic aspect u and giving the same products $[g(u)(du/dx)]_\phi$.

4. Proof of the theorem. The construction of $U=\text{class }(u^\varepsilon)_{0<\varepsilon<1}$ in $\mathcal{G}(]a,b[,\mathbb{R}^p)$ will be done in several steps:

$$\begin{tikzcd} \begin{pmatrix} u \\ \phi \end{pmatrix} \arrow[r, "\text{Step 1}"] & \begin{pmatrix} v^\varepsilon \\ \phi \end{pmatrix} \arrow[r, "\text{Step 2}"] & \tilde{v}^\varepsilon \arrow[r, "\text{Step 3}"] & \tilde{v}^\varepsilon * \rho_{h(\varepsilon)} \arrow[r, "\text{Step 4}"] & (u^\varepsilon)_{0<\varepsilon<1} \\ \text{u: BV function} & \text{v^ε: piecewise constant} & \text{$\tilde{v}^\varepsilon$: Lipschitz continuous junctions by $\psi^{i,\varepsilon}$ as in Fig. 1} & \text{regularization $\rho\in\mathcal{D}$, $\int\rho=1$, $h(\varepsilon)\to 0$ if $\varepsilon\to 0$, h suitably chosen} & \text{gives U as its class} \end{tikzcd}$$

Step 1. Given the BV function u, there is a family $(v^\varepsilon)_{0<\varepsilon<1}$ of step functions, $v^\varepsilon\to u$ in sup norm when $\varepsilon\to 0$ and $\text{TV}(v^\varepsilon)<c$ (i.e., the total variation of v^ε is less

than a constant independent of ε). Theorem 2.2 in [13] asserts that, for every test function $\theta \in \mathscr{D}(]a, b[)$,

$$\int \left[g(v^\varepsilon)\frac{dv^\varepsilon}{dx}\right]_\phi (x)\theta(x)\,dx \to \int \left[g(u)\frac{du}{dx}\right]_\phi (x)\theta(x)\,dx \tag{3}$$

when $\varepsilon \to 0$. These v^ε do not depend continuously on ε: in the Appendix we do not state such a dependence in the definition of $\mathscr{G}$. Such a dependence can be useful for other purposes (see [5]); it could be incorporated here at the price of a more complicated construction.

Step 2. We construct the continuous function $\tilde{v}^\varepsilon$ from v^ε and ϕ as described above in Fig. 1. Indeed, v^ε is piecewise constant, with a discrete set of discontinuities x_i^ε, $1 \leqq i \leqq n$, on each compact interval. In order to avoid a possible overlapping of the junctions, the segments $[x_i^\varepsilon, x_i^\varepsilon + \varepsilon]$ as in Fig. 1 are replaced by $[x_i^\varepsilon, x_i^\varepsilon + f_i^\varepsilon]$ with $0 < f_i^\varepsilon \leqq \varepsilon$ small enough. In a neighborhood of x_i^ε, we have the following:

$$\begin{cases} \tilde{v}^\varepsilon(x) = v^\varepsilon(x) & \text{if } x \leqq x_i^\varepsilon \text{ or } x \geqq x_i^\varepsilon + f_i^\varepsilon, \\ \tilde{v}^\varepsilon(x) = \phi(\lambda, v^\varepsilon(x_i^{\varepsilon-}), v^\varepsilon(x_i^{\varepsilon+})) & \text{if } x = x_i^\varepsilon + \lambda f_i^\varepsilon, \quad 0 \leqq \lambda \leqq 1. \end{cases} \tag{4}$$

Now let μ_ε be the measure $\mu_\varepsilon = [g(v^\varepsilon)(dv^\varepsilon/dx)]_\phi$ defined in [18].

Then $\mu_\varepsilon = \sum_{i \in \mathbb{N}} \mu_\varepsilon(\{x_i\})\delta_{x_i}$, where

$$\mu_\varepsilon(\{x_i\}) = \int_0^1 g(\phi(s, v^\varepsilon(x_i^-), v^\varepsilon(x_i^+)))\frac{\partial \phi}{\partial s}(s, v^\varepsilon(x_i^{\varepsilon-}), v^\varepsilon(x_i^{\varepsilon+}))\,ds.$$

An obvious change of variables and (4) give (see § 3)

$$\mu_\varepsilon(\{x_i\}) = \int_{x_i^\varepsilon}^{x_i^\varepsilon + f_i^\varepsilon} g(\tilde{v}^\varepsilon(x))\frac{d\tilde{v}^\varepsilon}{dx}(x)\,dx.$$

Therefore, if $\theta \in \mathscr{D}(]a, b[)$,

$$\int \left[g(v^\varepsilon)\frac{dv^\varepsilon}{dx}\right]_\phi (x)\theta(x)\,dx = \sum_i \theta(x_i^\varepsilon)\int_{x_i^\varepsilon}^{x_i^\varepsilon + f_i^\varepsilon} g(\tilde{v}^\varepsilon(x))\frac{d\tilde{v}^\varepsilon}{dx}(x)\,dx. \tag{5}$$

On the other hand,

$$\int g(v^\varepsilon)\frac{d\tilde{v}^\varepsilon}{dx}\theta(x)\,dx = \sum_i \int_{x_i^\varepsilon}^{x_i^\varepsilon + f_i^\varepsilon} g(\tilde{v}^\varepsilon(x))\frac{d\tilde{v}^\varepsilon}{dx}(x)\theta(x)\,dx. \tag{6}$$

The difference D of the quantities in (5) and (6) is

$$D = \sum_i \int_{x_i^\varepsilon}^{x_i^\varepsilon + f_i^\varepsilon} g(\tilde{v}^\varepsilon(x))\frac{d\tilde{v}^\varepsilon}{dx}(x)(\theta(x) - \theta(x_i^\varepsilon))\,dx.$$

Further,

$$\int_{x_i^\varepsilon}^{x_i^\varepsilon + f_i^\varepsilon} \left|\frac{d\tilde{v}^\varepsilon}{dx}\right| d\lambda = \int_0^1 \left|\frac{\partial \phi}{\partial s}(s, v^\varepsilon(x_i^{\varepsilon-}), v^\varepsilon(x_i^{\varepsilon+}))\right| ds \leqq k|v^\varepsilon(x_i^{\varepsilon-}) - v^\varepsilon(x_i^{\varepsilon+})|,$$

k independent of ε ([13, Hypothesis 2]).

Thus

$$\begin{aligned} |D| &\leqq \sum_i \int_{x_i^\varepsilon}^{x_i^\varepsilon + f_i^\varepsilon} \left| g(\tilde{v}^\varepsilon(x))\frac{d\tilde{v}^\varepsilon(x)}{dx}(\theta(x) - \theta(x_i^\varepsilon))\right| dx \\ &\leqq \|g(\tilde{v}^\varepsilon)\|_\infty \|\theta'\|_\infty \sup_i (f_i^\varepsilon) k\, \mathrm{TV}(v^\varepsilon). \end{aligned}$$

Since $(f_i^\varepsilon) \leqq \varepsilon$ and $\mathrm{TV}(v^\varepsilon) \leqq c$ one obtains

$$\int \left[g(v^\varepsilon)\frac{dv^\varepsilon}{dx}\right]_\phi (x)\theta(x)\,dx - \int g(\tilde{v}^\varepsilon(x))\frac{d\tilde{v}^\varepsilon}{dx}(x)\theta(x)\,dx \to 0 \quad \text{as } \varepsilon \to 0. \tag{7}$$

Step 3. The objective of this step is to regularize the continuous functions $\tilde{v}^{\varepsilon}$ so as to obtain C^{∞} functions.

Let $\rho \in \mathscr{D}(\mathbb{R})$, $\int \rho(x)\,dx = 1$; we set $\rho_{\varepsilon}(x) = (1/\varepsilon)\rho(x/\varepsilon)$. Let $h:]0,1] \to]0,1]$ be a decreasing function such that $h(\varepsilon) \to 0$ when $\varepsilon \to 0$. We set $\tilde{\tilde{v}}^{\varepsilon} = \tilde{v}^{\varepsilon} * \rho_{h(\varepsilon)}$. We want to prove that for a suitably chosen function h, then, for any $\theta \in \mathscr{D}(]a,b[)$ and $g \in C^{\infty}(\mathbb{R}^p, \mathbb{R}^p)$,

$$(8) \qquad \int g(\tilde{\tilde{v}}^{\varepsilon,\lambda}(x)) \frac{d\tilde{\tilde{v}}^{\varepsilon,\lambda}}{dx}(x)\theta(x)\,dx - \int g(\tilde{v}^{\varepsilon}(x)) \frac{d\tilde{v}^{\varepsilon}}{dx}(x)\theta(x)\,dx \to 0 \quad \text{as } \varepsilon \to 0.$$

Let us fix ε and set $\tilde{\tilde{v}}^{\varepsilon,\lambda} = \tilde{v}^{\varepsilon} * \rho_{\lambda}$, $\lambda > 0$. Then it is easy to prove that for fixed ε, g, and θ

$$\int g(\tilde{\tilde{v}}^{\varepsilon,\lambda}(x)) \frac{d\tilde{\tilde{v}}^{\varepsilon,\lambda}}{dx}(x)\theta(x)\,dx - \int g(\tilde{v}^{\varepsilon}(x)) \frac{d\tilde{v}^{\varepsilon}}{dx}(x)\theta(x)\,dx \to 0 \quad \text{as } \lambda \to 0.$$

Equation (8) follows (nontrivially) since we are free to make a choice of the function h (i.e., $h(\varepsilon) \to 0$ as fast as needed when $\varepsilon \to 0$); the details are left to the reader (the above limit depends on bounds of g and θ that have a countable character: for this there exists a suitable h). A similar argument has been elaborated in [12].

The conjunction of (3), (7), and (8) gives

$$(9) \qquad \int g(\tilde{\tilde{v}}^{\varepsilon}(x)) \frac{d\tilde{\tilde{v}}^{\varepsilon}}{dx}(x)\theta(x)\,dx \to \int \left[g(u)\frac{du}{dx} \right]_{\phi} (x)\theta(x)\,dx \quad \text{as } \varepsilon \to 0$$

for any $\theta \in \mathscr{D}(]a,b[)$.

Step 4. The family $(\tilde{\tilde{v}}^{\varepsilon})_{0<\varepsilon<1}$ is a family of C^{∞} functions in the x variable. In the course of the construction of this family we have lost any control on its dependence in ε. Let $\mathscr{E}_M[\,]a,b[,\mathbb{R}^p]$ be the requested reservoir of representatives for defining $\mathscr{G}(]a,b[;\mathbb{R}^p)$. A family $\{u^{\varepsilon}\}_{0<\varepsilon<1}$ of C^{∞} functions on $]a,b[$ valued in $\mathbb{R}^p$ lies in this reservoir if and only if (owing to the simplification due to the dependence on ε only) for every compact interval K in $]a,b[$ and for every $n \in \mathbb{N}$ there are $c > 0$, $\eta > 0$ and $N \in \mathbb{N}$ such that

$$\sup_{x \in K} \left\| \frac{d^n}{dx^n} u^{\varepsilon}(x) \right\| \leqq \frac{c}{\varepsilon^N} \quad \text{if } 0 < \varepsilon < \eta,$$

where $\|\ \|$ is the norm in $\mathbb{R}^p$. The family $(\tilde{\tilde{v}}^{\varepsilon})_{0<\varepsilon<1}$ does not a priori satisfy these bounds. But there is a suitable function $k:]0,1[\to]0,1[$, $k(\varepsilon) \to 0$ if $\varepsilon \to 0$, such that $u^{\varepsilon} = \tilde{\tilde{v}}^{k(\varepsilon)}$ satisfies the above bounds (k is obtained by a diagonal process since one has a countable set of requirements: countability of an exhaustive sequence of compact sets in $]a,b[$, and countability of the set of successive derivatives). Then $(u^{\varepsilon})_{0<\varepsilon<1}$ lies in the requested reservoir of representatives. Its class defines an element U of $\mathscr{G}(]a,b[;\mathbb{R}^p)$. Of course (9) holds with u^{ε} in place of $\tilde{\tilde{v}}^{\varepsilon}$. □

5. Comments. In the terminology of the concept of generalized functions in use, the weak equality in the statement of the theorem is formulated by saying that the measure $[g(u,x)(du/dx)]_{\phi}$ and the generalized function $g(U,x)(dU/dx)$ are associated; see the Appendix.

The generalized function U can—as is quite apparent from Fig. 1—be (noncanonically) considered as the pair (u, ϕ), and then, of course, the product depends only on U: the path ϕ enters into the definition of the generalized function U.

Conversely, given a generalized function U, with—to fix the idea—$U(x) = a$ if $x < 0$ and $U(x) = b$ if $x > 0$, is it possible to find a family of paths ϕ in the sense of [13], [16], [17]? In various concrete circumstances this is the case, and examples from numerical analysis and physics are shown in [3], [8]-[10], where these paths were called "the microscopic profile of the shock." In general this is impossible: U is the class of a family $(u^\varepsilon)_{0<\varepsilon<1}$ of C^∞ functions. For each u^ε the junction between the left- and right-hand side values can be assumed (at least intuitively) to take place on an interval $[\alpha(\varepsilon), \beta(\varepsilon)]$, $\alpha(\varepsilon)\to 0$, $\beta(\varepsilon)\to 0$ as $\varepsilon\to 0$. By a linear change of variable we can transform this interval into the interval [0, 1]: this gives some function ϕ^ε defined on [0, 1], which, of course, stands for the path ϕ in Fig. 1. If, when $\varepsilon\to 0$, the function ϕ^ε tends to a function ϕ, then we can say that the reverse process works: we obtain a pair (u, ϕ) from which the process in Fig. 1 gives a generalized function $U(U\approx u)$, having ϕ as a "microscopic profile." But it may happen that the ϕ^ε's vary endlessly when $\varepsilon\to 0$, or are unbounded when $\varepsilon\to 0$; even in case of a weak convergence of the ϕ^ε's to some ϕ, ϕ need not satisfy the assumptions in [13], [16], [17]. Examples have been observed in numerical tests; see [3], Figs. 6, 7 in Appendix 2 of Chapter 3, [8], [10]. Further, in [13], [16], [17] the path ϕ depends only on the left- and right-hand side values of the BV function u. This excludes step functions with time independent step values and time dependent microscopic profiles that have been observed, for example, in [1]. Also, it may happen that in the solution of a Cauchy problem (with a BV initial condition) the solution has several shocks, two of them with same left- and right-hand side values but different "microscopic profiles."

Therefore, the concept of generalized functions can be considered as more general than the concept of pairs (u, ϕ)—at least when ϕ has the properties in [13], [16], [17]. However, in the theorem of [17] a remarkable property of conservation of the path ϕ at the limit of Glimm's random choice method has been proven; then the consideration of more general objects to describe the limit is avoided. Note also that the space $\mathscr{G}$ has primarily been presented as an algebra of functions: notions of topology are exposed in [3, § 1.7], but they are not significantly used. When we restrict it to the setting of functions of bounded variation, the classical topologies in these spaces can be used; see [1], [3], [4]. Then it appears that both approaches are closely related on their common domain.

Appendix. The purpose of this Appendix is to offer a definition of the concept of the "generalized functions" we use. If Ω is an open set in $\mathbb{R}^n$, we construct the space $\mathscr{G}(\Omega)$ of (real or complex valued) generalized functions on Ω from the space $\mathscr{C}^\infty(\Omega)$ of C^∞ functions on Ω in a way that looks like the construction of the real numbers from the rational numbers (each real number is an equivalence class of (Cauchy) sequences of rational numbers). Slightly different constructions lead to slightly different spaces $\mathscr{G}(\Omega)$, but at the level of the results in this paper, there is no significant difference between them. We set

$$\mathscr{A}_q=\left\{\varphi\in\mathscr{D}(\mathbb{R}^n)\text{ such that }\int_{\mathbb{R}^n}\varphi(\lambda)\,d\lambda=1,\ \int_{\mathbb{R}^n}\lambda^i\varphi(\lambda)\,d\lambda=0\text{ if }1\leqq|i|\leqq q\right\}.$$

If $0<\varepsilon<1$, we set as usual $\varphi_\varepsilon(\lambda)=(1/\varepsilon^n)\varphi(\lambda/\varepsilon)$, and it is clear that $\varphi\in\mathscr{A}_q$ if and only if $\varphi_\varepsilon\in\mathscr{A}_q$.

Let us define our "reservoir of representatives" (i.e., the analogue of the set of all Cauchy sequences of rational numbers). This reservoir, denoted by $\mathscr{E}_M[\Omega]$, can be intuitively introduced as follows: let there be given a distribution T on $\mathbb{R}^n$; T can be viewed as a class of regularizations $T*\varphi\in\mathscr{C}^\infty(\mathbb{R}^n)$, with $\varphi\in\mathscr{D}$, φ tending to the Dirac measure δ in the sense of distributions. For a canonical regularization it is necessary to consider all possible regularizing functions φ, and, therefore, all maps $(\varphi, x)\to R(\varphi, x)(R(\varphi, x)=(T*\varphi)(x)$ in the above particular case of a distribution). $\mathscr{E}_M[\Omega]$ is the set of all maps

$$R:\mathscr{A}_0x\Omega\to\mathbb{R},\qquad(\varphi, x)\to R(\varphi, x),$$

such that:

(1) For any φ the map $R : x \to (\varphi, x)$ is a C^∞ function of the variable $x \in \Omega$;

(2) if $D = (\partial^{|k|}/\partial_{x_1}^{k_1} \cdots \partial_{x_n}^{k_n})$ is any partial derivation operator (D = identity is allowed), if $\varphi \in \mathcal{A}_0$ and K is any compact subset of Ω, then there are constants $c > 0$, $N \in \mathbb{N}$, and $0 < \eta \leqq 1$ such that

$$\sup_{x \in K} |DR(\varphi_\varepsilon, x)| \leqq \frac{c}{\varepsilon^N}$$

if $0 < \varepsilon < \eta$. Note that if $R \in \mathscr{E}_M[\Omega]$, then $DR \in \mathscr{E}_M[\Omega]$ for any D. Also, if R, $Q \in \mathscr{E}_M[\Omega]$, then $R + Q \in \mathscr{E}_M[\Omega]$ and $R.\ Q \in \mathscr{E}_M[\Omega]$: $\mathscr{E}_M[\Omega]$ is a differential algebra (partial derivatives, addition, and multiplication).

Now let us define our "set of null elements" (i.e., the analogue of all zero sequences of rational numbers). This set is denoted by $\mathcal{N}[\Omega]$ and is an ideal of the algebra $\mathscr{E}_M[\Omega]$. $\mathcal{N}[\Omega]$ is the set of all maps $R \in \mathscr{E}_M[\Omega]$ such that for all D and K as above there is $N \in \mathbb{N}$ and an increasing map $\gamma : \mathbb{N} \to \mathbb{N}$, $\gamma(q) \to +\infty$ if $q \to +\infty$, such that $\forall q \geqq N, \forall \varphi \in \mathcal{A}_q\ \exists c > 0$ and $0 < \eta \leqq 1$ such that $\sup_{x \in K} |DR(\varphi_\varepsilon, x)| \leqq c\varepsilon^{\gamma(q)}$ if $0 < \varepsilon < \eta$.

A generalized function G on Ω is an element of the quotient space $\mathscr{G}(\Omega) = \mathscr{E}_M[\Omega]/\mathcal{N}[\Omega]$, which is a differential algebra (since the operations of partial differentiation, of addition and multiplication are coherent with this quotient).

$\mathscr{C}^\infty(\Omega)$ is contained in $\mathscr{G}(\Omega)$ in the following way: to $f \in \mathscr{C}^\infty(\Omega)$ associate the "representative" $R(\varphi, x) = f(x)$. $\mathscr{D}'(\Omega)$ is contained in $\mathscr{G}(\Omega)$ in the following way: to $T \in \mathscr{D}'(\Omega)$ associate the "representative" $R(\varphi, x) = \langle T_\lambda, \varphi(\lambda - x)\rangle = (T * \varphi)(x)$. We have the inclusions $\mathscr{C}^\infty(\Omega) \subset \mathscr{D}'(\Omega) \subset \mathscr{G}(\Omega)$, and $\mathscr{C}^\infty(\Omega)$ is a subalgebra.

The origin of this somewhat unexpected construction is described in [5]. Indeed this construction relies on the classical Taylor expansion of a C^∞ function.

We state $G \approx 0$ if it has a representative R (then this works for all representatives) such that for large enough $N \in \mathbb{N}$, for any $\theta \in \mathscr{D}(\Omega)$,

$$\int R(\varphi_\varepsilon, x)\theta(x)\, dx \to 0$$

if $\varepsilon \to 0$. We state $G_1 \approx G_2$ if and only if $G_1 - G_2 \approx 0$. We say that $G \in \mathscr{G}(\Omega)$ admits $T \in \mathscr{D}'(\Omega)$ as "macroscopic aspect" or "associated distribution" if and only if $G \approx T$. The importance of this concept is due to the fact that because the ideal $\mathcal{N}[\Omega]$ is very small, the concept of equality in $\mathscr{G}(\Omega)$ is a very strong relation; the concept of association plays the role of a weaker concept of equality, however incoherent with the multiplication.

Simplified formulation. Practice often shows that the formal replacement of $R(\varphi_\varepsilon, x)$ by $R(\varepsilon, x)$ (whatever φ may be) amounts only to a simplification in notation. Also, we can rigorously consider a subalgebra $\mathscr{G}_s(\Omega)$—in which the subscript s stands for "simplified"—for which the representatives depend only on ε and x. We have the following situation [3]:

$$\begin{tikzcd}
& \mathscr{D}'(\Omega) \arrow[dr, phantom, "\subset"] & \\
\mathscr{C}^\infty(\Omega) \arrow[ur, phantom, "\subset"] \arrow[dr, phantom, "\subset"] & & \mathscr{G}(\Omega). \\
& \mathscr{G}_S(\Omega) \arrow[ur, phantom, "\subset"] &
\end{tikzcd}$$

We pay the price of simplification by a loss of the canonical inclusion of $\mathscr{D}'$ into the space $\mathscr{G}_S$ of generalized functions. In this paper we have adopted $R(\varepsilon, x)$ as a simplified notation for $R(\varphi_\varepsilon, x)$. Furthermore, the U's constructed in the proof can be considered as elements of $\mathcal{G}_S(\Omega)$. This makes no difference for the contents of this paper. Note that the simplified concept is introduced directly in [9], [10].

Acknowledgments. The authors are very much indebted to A. Y. Le Roux, M. Oberguggenberger, and D. Serre for their help during the preparation of this paper. Significant improvements in the text are due to the referee.

Note added in proof. A very simplified formulation is presented in Egorov, "On the theory of generalized functions," *Russian Math. Surveys* 5, (1990), pp. 1-49.

REFERENCES

[1] M. Adamczewski, J. F. Colombeau, and A. Y. Le Roux, *Convergence of numerical schemes involving powers of the Dirac delta function*, J. Math. Anal. Appl., 145 (1990), pp. 172-185.

[2] Y. A. Barka, J. F. Colombeau, and B. Perrot, *A numerical modelling of the fluid/fluid acoustic dioptra*, J. d'Acoustique, 2 (1989), pp. 333-346.

[3] H. A. Biagioni, *Introduction to a nonlinear theory of generalized functions*, Notas Mat., UNICAMP, Campinas, SP Brazil, 1988; *A nonlinear theory of generalized functions*, Lecture Notes in Math. 1421, Springer-Verlag, 1990.

[4] J. J. Cauret, J. F. Colombeau, and A. Y. Le Roux, *Discontinuous generalized solutions of nonlinear nonconservative hyperbolic equations*. J. Math. Anal. Appl., 139 (1989), pp. 552-573.

[5] J. F. Colombeau, *Multiplication of Distributions*, Bull. Amer. Math. Soc., 23 (1990), pp. 251-268.

[6] ———, *Elementary Introduction to New Generalized Functions*, North-Holland, Amsterdam, the Netherlands, 1985.

[7] ———, *Introduction to "new generalized functions" and multiplication of distributions*, in Functional Analysis and Its Applications, H. Hogbé Nlend, ed., ICPAM Lecture Notes, World Scientific, Singapore, 1988, pp. 338-380.

[8] ———, *The elastoplastic shock problem as an example of the resolution of ambiguities in the multiplication of distributions*, J. Math. Phys., 30 (1989), pp. 2273-2279.

[9] J. F. Colombeau and A. Y. Le Roux, *Multiplication of distributions in elasticity and hydrodynamics*, J. Math. Phys., 29 (1988), pp. 315-319.

[10] J. F. Colombeau, A. Y. Le Roux, A. Noussair, and B. Perrot, *Microscopic profiles or shock waves and ambiguities in multiplications of distributions*, SIAM J. Numer. Anal. 26 (1989), pp. 871-883.

[11] J. F. Colombeau and M. Oberguggenberger, *On a hyperbolic system with a compatible quadratic term; generalized solutions, delta waves and multiplication of distributions*, Comm. Partial Differential Equations, 15 (1990), pp. 905-938.

[12] ———, *Approximate generalized solutions and measure valued solutions to conservation laws*, preprint.

[13] G. Dal Maso, Ph. le Floch, and F. Murat, *Definition and weak stability of a nonconservative product*, Ecole Polytechnique, 1989, preprint.

[14] Ph. le Floch, *Entropy weak solutions to nonlinear hyperbolic systems in nonconservation form*, Comm. Partial Differential Equations, 13 (1988), pp. 669-722.

[15] ———, *Entropy weak solutions to nonlinear hyperbolic systems in nonconservation form*, in Nonlinear Hyperbolic Equations, J. Ballman and R. Jeltsch, eds., Notes on Numerical Fluid Mechanics, Vol. 24, Vieweg, 1989, pp. 362-373.

[16] ———, *Shock waves for nonlinear hyperbolic systems in nonconservation form*, preprint.

[17] T. P. Liu and Ph. le Floch, *Convergence of the random choice method for systems in nonconservative form*, 1990, preprint.

[18] E. E. Rosinger, *Generalized solutions of nonlinear* PDEs, North-Holland, Amsterdam, the Netherlands, 1988.

[19] H. B. Stewart, B. Wendroff, *Two-phase flows: models and methods*, J. Comp. Phys., 56 (1984), pp. 363-409.

[20] A. I. Volpert, *The space BV and quasilinear equations*, Math. Sb., 73 (1967), pp. 225-267.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 950–958, July 1992

009

A UNIQUENESS RESULT CONCERNING THE IDENTIFICATION OF A COLLECTION OF CRACKS FROM FINITELY MANY ELECTROSTATIC BOUNDARY MEASUREMENTS*

KURT BRYAN† AND MICHAEL VOGELIUS‡

Abstract. The problem of identification of a collection of finitely many cracks inside a planar domain is considered. The data used for the identification consist of measurements of the electrostatic boundary potentials induced by prescribed current fluxes. It is shown that a collection of n or fewer cracks is uniquely identified by boundary measurements corresponding to $n+1$ specific current fluxes, consisting entirely of electrode pairs.

Key words. cracks, nondestructive testing, electrical impedance imaging

AMS(MOS) subject classifications. 35R30, 35J25

Introduction. In a recent paper, [1], Friedman and Vogelius proved that the presence of a single crack, and its shape and location inside a planar domain, may be determined from measurements of the steady state boundary voltage potentials corresponding to two specific boundary current fluxes. In the present paper we extend this result to any finite number of cracks. We show that voltage measurements corresponding to $n+1$ specific fluxes suffice to determine the location and shape of a collection of n (or fewer) cracks. In contrast to [1] the fluxes we use here all consist of electrode pairs—exactly the type of fluxes which were used for the computational algorithm developed in [2].

Let Ω be a simply connected domain in $\mathbb{R}^2$ with a smooth boundary. In order to describe our result in detail we need to define the notion of a collection of cracks. By a C^2-curve, σ, we understand a twice continuously differentiable map: $[0,1] \to \Omega$ with non-vanishing derivative. *A collection of cracks consists of a finite number of mutually disjoint, nonselfintersecting C^2-curves σ_k , $k = 1, \cdots, n$.* We use capital Greek letters to denote collections of cracks, e.g., $\Sigma = \{\sigma_k\}_{k=1}^n$; note that n may possibly be zero, so that Σ is empty. We shall also use the notation σ_k and Σ for the image of each of the individual curves and the union of all the images, respectively (i.e., $\Sigma = \cup_{k=1}^n \sigma_k$). Let $\gamma : \overline{\Omega} \to \mathbb{R}$ be a positive function (the known reference conductivity). Throughout this paper we assume that

$$\gamma \text{ is real-analytic on } \quad \overline{\Omega}.$$

In the following, when a function is called analytic, it shall always mean real-analytic. Quite frequently in the literature the term crack is used synonymously with an electrically insulating crack: if ϕ represents the boundary voltage, then the steady state

*Received by the editors April 8, 1991; accepted for publication August 19, 1991.

†ICASE, NASA Langley Research Center, Hampton, Virginia 23665. This work was partially supported by National Aeronautics and Space Administration contract NAS1-18605.

‡Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. This work was partially supported by National Science Foundation grant DMS-89-02532 and Air Force Office of Scientific Research contract 89-NM-605.

voltage potential satisfies

$$\begin{aligned} \nabla\cdot(\gamma\nabla v) &= 0 \quad \text{in } \Omega\setminus\Sigma,\\ \gamma\frac{\partial v}{\partial\nu} &= 0 \quad \text{on } \Sigma,\\ v &= \phi \quad \text{on } \partial\Omega. \end{aligned}$$

In this framework the inverse problem is to determine Σ from knowledge of several pairs $(\phi, \gamma\frac{\partial v}{\partial\nu}|_{\partial\Omega})$. We shall, instead of working with the potential v, opt to work with its "γ-harmonic" conjugate, u. This function is related to v by

$$(0.1) \qquad (\nabla u)^{\perp} = \gamma\nabla v,$$

where $\perp$ indicates counterclockwise rotation by $\pi/2$. Let T be a fixed point on $\partial\Omega$, in a neighborhood of which ϕ is smooth. Let τ_k be a smooth curve in $\Omega\setminus\Sigma$ connecting T to an interior point of the crack σ_k, and let s denote the unit tangent direction along τ_k, pointing from T towards σ_k. Define constants

$$c^{(k)} = -\int_{\tau_k}\gamma\frac{\partial v}{\partial\nu}ds \;+\; u(T),$$

where ν denotes the normal field $\nu = -s^{\perp}$. The "γ-harmonic" conjugate, u, solves

$$(0.2) \qquad \begin{aligned} \nabla\cdot(\gamma^{-1}\nabla u) &= 0 \quad \text{in } \Omega\setminus\Sigma,\\ u &= c^{(k)} \quad \text{on } \sigma_k \;\; k = 1,\cdots,n,\\ \gamma^{-1}\frac{\partial u}{\partial\nu} &= \frac{\partial\phi}{\partial s} = \psi \quad \text{on } \partial\Omega, \end{aligned}$$

where s denotes the counterclockwise tangent direction on $\partial\Omega$.

From (0.1) it follows immediately that knowledge of $u|_{\partial\Omega}$ leads to knowledge of $-\gamma\frac{\partial v}{\partial\nu}|_{\partial\Omega} = \frac{\partial}{\partial s}(u|_{\partial\Omega})$, and vice versa. Therefore, knowledge of pairs $(u|_{\partial\Omega}, \psi)$ is equivalent to knowledge of corresponding pairs $(\phi, \gamma\frac{\partial v}{\partial\nu}|_{\partial\Omega})$, where ϕ and ψ are related by $\psi = \frac{\partial\phi}{\partial s}$. Physically (0.2) corresponds to a collection of perfectly conducting cracks. One way to solve (0.2) is to minimize the energy

$$\frac{1}{2}\int_{\Omega}\gamma^{-1}|\nabla w|^2\,dx - \int_{\partial\Omega}\psi w\,ds$$

in the space $H^1(\Omega)\cap\{w = \text{ const on each } \sigma_k \in \Sigma\}$ (such minimization gives, modulo a single undetermined constant, exactly the values on σ_k, $k = 1,\cdots,n$ described above). This method works provided $\psi \in H^{-1/2}(\partial\Omega)$. The fluxes we shall apply here, however, correspond to single pairs of electrodes, i.e., we shall take ψ of the form $\psi = \delta_{P_0} - \delta_{P_1}$, where P_0 and P_1 are two distinct points on $\partial\Omega$. Such ψ are not in $H^{-1/2}(\partial\Omega)$—the solution, u, is, therefore, not in $H^1(\Omega)$, and it is not obtained as a minimizer of energy. Rather, u is a weak solution to (0.2); it is smooth everywhere except at P_0 and P_1 and at the endpoints of the cracks. At P_0 and P_1 the function u has singularities of the form $-\gamma(P_0)/\pi\ \log r$, $r = |x - P_0|$, and $\gamma(P_1)/\pi\ \log r$, $r = |x - P_1|$, respectively; at the endpoints of the cracks u has in general $r^{1/2}$-type singularities, cf. [1].

For our uniqueness result it is not necessary that we have solutions which attain exactly the constant values on the cracks described above—any set of constants will do. To construct the specific boundary currents, let $P_0,\cdots,P_M$ be $M+1$ different (fixed) points on $\partial\Omega$; we assume that these points are labeled in order of counterclockwise appearance, starting from P_0. In our first uniqueness result we utilize solutions to the boundary value problems

$$\begin{aligned} \nabla\cdot(\gamma^{-1}\nabla u_j) &= 0 \quad \text{in } \Omega\setminus\Sigma, \\ u_j &= \text{ constant on each } \sigma_k \in \Sigma, \\ \gamma^{-1}\frac{\partial u_j}{\partial\nu} &= \delta_{P_0} - \delta_{P_j} \quad \text{on } \partial\Omega. \end{aligned} \tag{0.3}$$

THEOREM 0.1. *Let $\Sigma = \{\sigma_k\}_{k=1}^n$ and $\tilde{\Sigma} = \{\tilde{\sigma}_k\}_{k=1}^m$ denote two collections of cracks contained in the domain Ω, with $\max(m,n)+1 \le M$. Let u_j, $j = 1,\cdots,M$ denote solutions to (0.3) and let $\tilde{u}_j$, $j = 1,\cdots,M$ denote solutions to (0.3) with Σ replaced by $\tilde{\Sigma}$. Then $u_j = \tilde{u}_j$ on $\partial\Omega\setminus\cup_{i=0}^M\{P_i\}$ for $j = 1,\cdots,M$ implies that $\Sigma = \tilde{\Sigma}$.*

Instead of prescribing fixed fluxes $\gamma^{-1}\frac{\partial u_j}{\partial\nu}|_{\partial\Omega} = \psi_j$ and measuring $u_j|_{\partial\Omega}$, we can prescibe fixed boundary voltages $w_j|_{\partial\Omega} = \phi_j$ equally well and measure $\gamma^{-1}\frac{\partial w_j}{\partial\nu}|_{\partial\Omega}$. For that purpose we utilize solutions to the following boundary value problems

$$\begin{aligned} \nabla\cdot(\gamma^{-1}\nabla w_j) &= 0 \quad \text{in } \Omega\setminus\Sigma, \\ w_j &= \text{const on each } \sigma_k \in \Sigma, \\ w_j &= 1_{\overset{\frown}{P_{j-1}P_j}} \quad \text{on } \partial\Omega, \end{aligned} \tag{0.4}$$

where $1_{\overset{\frown}{P_{j-1}P_j}}$ denotes the characteristic function of the counterclockwise curve from P_{j-1} to P_j. The function w_j is a weak solution to (0.4)—it is not in $H^1(\Omega)$, and, therefore, not obtained as a minimizer of energy. w_j has a singularity of the form $-\theta/\pi$ at P_{j-1}, $\theta = \arg(x-P_{j-1})$ and has a singularity of the form θ/π at P_j, $\theta = \arg(x-P_j)$. At the endpoints of the cracks w_j has, in general, $r^{1/2}$-type singularities.

THEOREM 0.2. *Let $\Sigma = \{\sigma_k\}_{k=1}^n$ and $\tilde{\Sigma} = \{\tilde{\sigma}_k\}_{k=1}^m$ denote two collections of cracks contained in the domain Ω, with $\max(m,n)+1 \le M$. Let w_j, $j = 1,\cdots,M$ denote solutions to (0.4) and let $\tilde{w}_j$, $j = 1,\cdots,M$ denote solutions to (0.4) with Σ replaced by $\tilde{\Sigma}$. Then $\gamma^{-1}\frac{\partial w_j}{\partial\nu} = \gamma^{-1}\frac{\partial\tilde{w}_j}{\partial\nu}$ on $\partial\Omega\setminus\cup_{i=0}^M\{P_i\}$ for $j = 1,\cdots,M$ implies that $\Sigma = \tilde{\Sigma}$.*

REMARK 0.1. As was the case with the first theorem, this second theorem also has an alternative formulation in terms of cracks that are insulating. In that case we would prescribe normal boundary fluxes $-\partial(1_{\overset{\frown}{P_{j-1}P_j}})/\partial s = \delta_{P_j} - \delta_{P_{j-1}}$ and measure the corresponding boundary voltages.

1. Preliminaries. The proof of our main results consists of a very detailed analysis of the structure of the level curves of solutions to the equation $\nabla\cdot(\gamma^{-1}\nabla u) = 0$. For that purpose we shall need two auxiliary lemmas.

LEMMA 1.1. *Let u satisfy $\nabla\cdot(\gamma^{-1}\nabla u) = 0$ in $\Omega\setminus\Sigma$ with $\gamma^{-1}\partial u/\partial\nu = \sum_j \beta_j\delta_{P_j}$ on $\partial\Omega$, and u constant on each σ_k. Let ρ be a nonempty analytic curve in Ω with $\rho\cap\Sigma = \emptyset$ along which u is constant. Then there exists an analytic curve ρ' with $\rho\subset\rho'$ such that*

$$u \text{ is constant on } \rho', \tag{1.1a}$$

$$\rho' \text{ has one endpoint on } \partial\Omega \text{ or } \sigma_k \text{ for some } k, \tag{1.1b}$$

(1.1c) *ρ' has the other endpoint on $\partial\Omega$ or σ_l for some l with $l \neq k$.*

The proof of this lemma is identical to the proof of Lemma 2.3 in [1]. We shall not repeat the the proof here. The second lemma we need concerns the existence of intersecting level curves. Some of the details of the proof of this result are not unlike those found in the proof of Lemma 2.3 in [1], but for the convenience of the reader we give a complete proof here.

LEMMA 1.2. *Let u satisfy $\nabla \cdot (\gamma^{-1}\nabla u) = 0$ in $\Omega \setminus \Sigma$ with $\gamma^{-1}\partial u/\partial\nu = \sum_j \beta_j \delta_{P_j}$ on $\partial\Omega$, and u constant on each σ_k. Let ρ be a nonempty analytic curve in Ω with $\rho \cap \Sigma = \emptyset$, along which u is constant, and assume that x^* is an interior point of ρ, where $\nabla u(x^*) = 0$. Then there exists an analytic curve ρ' which has x^* as an interior point such that*

$$\rho' \cap \rho = \{x^*\}, \tag{1.2a}$$

$$u \text{ is constant on } \rho'. \tag{1.2b}$$

Proof. Let $(r, \theta) \in [0, \epsilon] \times [0, 2\pi]$ denote polar coordinates at x^*. Since $\nabla u(x^*) = 0$ we know that $\frac{\partial}{\partial r}u(0, \theta) \equiv 0$, and by expanding in a Taylor series in r we get

$$u(x) = u(x^*) + r^N(a \sin N\theta + b \cos N\theta + rA(r, \theta))$$

for some a, b (not both zero) and some $N \geq 2$. Here we have used that u is nonconstant and satisfies $\nabla \cdot (\gamma^{-1}\nabla u) = 0$ near x^*, and we have used that γ^{-1} is analytic (the case of a constant u is trivial). We may, without loss of generality, assume that the tangent to ρ at x^* is $\{(r, 0),\ r > 0\} \cup \{(r, \pi),\ r > 0\}$. It follows that $b = 0$, i.e.,

$$u(x) = u(x^*) + ar^N(\sin N\theta + rA(r, \theta)).$$

Since u is analytic near x^*, it is well known that u as a function of (r, θ) is analytic on $[0, \epsilon] \times [0, 2\pi]$, the main point being that it is analytic at $r = 0$ and, therefore, also has an analytic extension to $[-\epsilon, \epsilon]$ for a sufficiently small ϵ (indeed the analytic extension for negative r is given by $\tilde{u}(r, \theta) = u(-r, \theta + \pi)$, $\theta + \pi$ taken modulo 2π). It follows that $A(r, \theta)$ also has an analytic extension near $r = 0$; we denote this extension by $\tilde{A}(r, \theta)$, $(r, \theta) \in [-\epsilon, \epsilon] \times [0, 2\pi]$. The function $F(r, \theta) = \sin N\theta + r\tilde{A}(r, \theta)$ satisfies

$$F(0, \pi/N) = 0 \quad \text{and} \quad \frac{\partial}{\partial\theta}F(0, \pi/N) = -N,$$

and therefore, by the implicit function theorem, it is possible to find a unique analytic function $\theta(r)$ such that $\theta(0) = \pi/N$ and $\{(r, \theta) :\ F(r, \theta) = 0\}$ coincides with $\{(r, \theta(r))\}$ in a neighborhood of $(0, \pi/N)$. The curve, ρ', given by

$$(r \cos\theta(r), r \sin\theta(r)) + x^*,$$

is an analytic curve through x^*, which satisfies $\rho' \cap \rho = x^*$ and which by its very definition is a level curve for u. □

2. Proof of Theorem 0.1. Let O be the open set enclosed by Σ and $\tilde{\Sigma}$, i.e., the set of points in $\Omega \setminus (\Sigma \cup \tilde{\Sigma})$ from which it is only possible to reach $\partial\Omega$ by crossing Σ or $\tilde{\Sigma}$. Since $\Omega \setminus (O \cup \Sigma \cup \tilde{\Sigma})$ has only one connected component, it follows from the assumptions about the boundary data (by unique continuation) that

$$u_j = \tilde{u}_j \quad \text{in } \Omega \setminus (O \cup \Sigma \cup \tilde{\Sigma}), \quad j = 1, \cdots, M. \tag{2.1}$$

If O is nonempty then ∂O consists of pieces of curves from Σ and $\tilde{\Sigma}$; on each of these pieces either u_j or $\tilde{u}_j$ is constant. Due to (2.1) it now follows that u_j is constant on each of the pieces that make up ∂O. Each function u_j therefore assumes finitely many values on ∂O (at most $m + n$). Since u_j is continuous in Ω (cf. [1]) we get that u_j is constant on each connected component of ∂O. Each connected component of O is simply connected, and it now follows, by the maximum principle, that u_j is constant in each connected component of O. This implies that u_j is constant in all of Ω—a contradiction. We thus conclude that O is empty, so that $u_j = \tilde{u}_j$ in $\Omega \setminus (\Sigma \cup \tilde{\Sigma})$; by continuity it follows that

$$u_j = \tilde{u}_j \quad \text{in } \Omega, \quad j = 1, \cdots, M. \tag{2.2}$$

Let us assume that Σ and $\tilde{\Sigma}$ are not identical. We may assume that there exists a curve ρ contained in $\tilde{\sigma}_k$ for some k with $\rho \cap \Sigma = \emptyset$. Based on (2.2) we conclude that the functions u_j are all constant on ρ. There must exist a point on ρ where $\nabla u_1 \neq 0$, since otherwise u_1 is constant in Ω (by unique continuation); the implicit function theorem asserts that ρ must be analytic near this point. By shortening, if necessary, we may assume that the entire curve ρ is analytic. Let ν be a unit normal vector field on the curve ρ and let $x_1, \cdots, x_{M-1}$ be distinct interior points on ρ. Let $\alpha_1, \cdots, \alpha_M$ denote numbers, *not all zero* , satisfying the underdetermined set of linear equations

$$\sum_{j=1}^{M} \frac{\partial u_j}{\partial \nu}(x_i)\alpha_j = 0, \qquad i = 1, \cdots, M-1.$$

Define the function

$$u(x) = \sum_{j=1}^{M} \alpha_j u_j(x) \tag{2.3}$$

for $x \in \Omega$. The curve ρ is also a level curve for u. Applying Lemma 1.1 we obtain an analytic curve ρ_0 containing ρ, which satisfies (1.1a)–(1.1c), i.e.,

$$u \text{ is constant on } \rho_0, \tag{2.4a}$$

$$\rho_0 \text{ has one endpoint on } \partial\Omega \text{ or } \sigma_k \text{ for some } k, \tag{2.4b}$$

$$\rho_0 \text{ has the other endpoint on } \partial\Omega \text{ or } \sigma_l \text{ for some } l \text{ and } l \neq k. \tag{2.4c}$$

For $x \in \rho_0$ we have $|\nabla u(x)| = |\partial u/\partial \nu(x)|$. From (2.3) it follows that $\partial u/\partial \nu(x_i) = 0$, so that $\nabla u(x_i) = 0$ for $i = 1, \cdots, M-1$. Using Lemma 1.2 we may now, for each of the critical points x_i, construct an analytic curve ρ_i such that

$$\rho_i \cap \rho_0 = \{x_i\}, \tag{2.5a}$$

(2.5b) $\qquad u \quad$ is constant on ρ_i.

Lemma 1.1 permits us to extend each of the curves ρ_i until it hits the boundary or one of the cracks in Σ; this way we obtain curves ρ_i which, in addition to (2.5a) and (2.5b), satisfy

(2.5c) $\qquad \rho_i$ has one endpoint on $\partial\Omega$ or σ_k for some k,

(2.5d) $\qquad \rho_i$ has the other endpoint on $\partial\Omega$ or σ_l for some l and $l \neq k$.

The fact that the extended curve ρ_i still only intersects ρ_0 at x_i is proven as follows. If ρ_i intersected ρ_0 at some other point x_i' then there would be some nonempty region O enclosed by ρ_0 and ρ_i with u constant on ∂O. By the maximum principle u would be constant on O, and hence it would be constant on all of Ω. This clearly contradicts the fact that $\gamma^{-1}\partial u/\partial\nu = \sum_{j=1}^{M}\alpha_j(\delta_{P_0}-\delta_{P_j})$ on $\partial\Omega$, where the P_j are distinct and at least one α_j is nonzero. Since all the curves ρ_i intersect ρ_0, the function u assumes the same constant value on $\cup_{i=0}^{M-1}\rho_i$. Note that no two of the $M-1$ curves $\rho_1, \cdots, \rho_{M-1}$ can intersect, for then we would have some nonempty region O enclosed by the M curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$, with u constant on ∂O—a contradiction. For a similar reason none of the curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$ can self-intersect. Between the M curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$ we have a total of $2M$ endpoints. Note that no two of these curves can terminate on the same crack σ_k, for then we would have some region O bounded by these curves and the crack σ_k, with u constant on ∂O—a contradiction. There are n cracks in Σ, so it follows that there must be at least $2M-n$ points on $\partial\Omega$ at which the curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$ terminate. Since the curves $\rho_1, \cdots, \rho_{M-1}$ do not intersect and each only intersect ρ_0 at one point, it is easy to see that any connected component of $\Omega \setminus \cup_{i=0}^{M-1}\rho_i$ has a part of its boundary in common with $\partial\Omega$, and that the number of connected components is exactly equal to the number of terminal points of the curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$ that lie on $\partial\Omega$ (at least $2M-n$). A situation corresponding to $n=2$ and $M=4$ is schematically shown in Fig. 1.

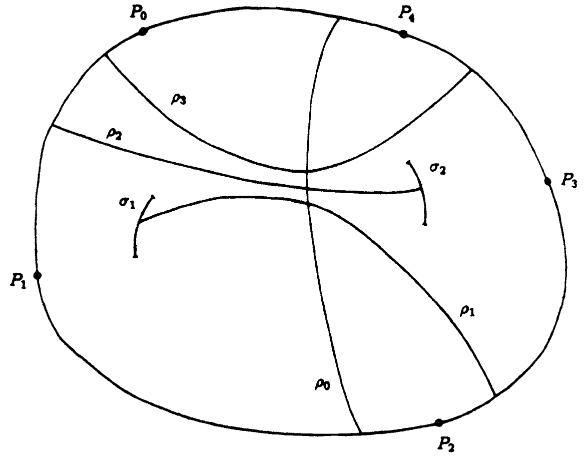


FIG. 1

The Neumann data for u has the form

$$\gamma^{-1}\frac{\partial u}{\partial\nu} = \sum_{j=0}^{M}\beta_j\delta_{P_j} \quad \text{on } \partial\Omega.$$

Let $M'+1 \leq M+1$ be the number of nonzero $\beta's$ in the above sum, i.e., $M'+1$ is the total number of sources and sinks (for u) on $\partial\Omega$. We note that none of the curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$ can terminate at a source or a sink, since $|u|$ approaches ∞ there.

If $M+(M-M') > n+1$ then it follows that $2M-n = M+(M-M')-n+M' > M'+1$, and, therefore, we conclude that at least one of the connected components of $\Omega \backslash \cup_{i=0}^{M-1} \rho_i$ is bounded by a level curve for u, and a portion of $\partial\Omega$ on which the normal derivative of u vanishes. This forces u to be a constant—a contradiction. Hence we see that if $M+(M-M') > n+1$, then the assumption that Σ and $\tilde{\Sigma}$ are different is incorrect.

Since $M \geq n+1$ and $M \geq M'$ we always have that $M+(M-M') \geq n+1$. The only case we have not analyzed yet is $M+(M-M') = n+1$, or, equivalently, $M' = M = n+1$. None of the curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$ can now terminate at any of the points P_j, $j = 0, \cdots, M$ (since $|u|$ approaches ∞ there). Furthermore, each of the connected components of $\Omega \backslash \cup_{i=0}^{M-1} \rho_i$ has at least one of the points P_j on its boundary. If not, the argument from the case $M+(M-M') > n+1$ leads to a contradiction. There cannot be one connected component, the boundary of which contains two or more of the points P_j because then, due to the identity $2M-n = M+1$, there would automatically have to be one connected component, the boundary of which contained none of the points P_j—a contradiction. In summary, each connected component of $\Omega \backslash \cup_{i=0}^{M-1} \rho_i$ has exactly one of the points P_j on its boundary. This means that there are exactly $M+1$ connected components and, therefore, exactly $M+1$ terminal points of the curves $\rho_0, \rho_1, \cdots, \rho_{M-1}$ on $\partial\Omega$. This leaves $2M - (M+1) = M-1 = n$ terminal points that fall on cracks—one on each crack of Σ. From the inequality $n+1 = M \geq \max(m,n)+1$ we conclude that $n \geq m$. If $n = 0$ it would follow that $m = 0$, so that both $\Sigma = \tilde{\Sigma} = \emptyset$—a contradiction. There is, therefore, at least one crack σ_{k_0} in the collection Σ. The crack σ_{k_0} is contained in the closure of one of the connected components of $\Omega \setminus \cup_{i=0}^{M-1} \rho_i$; we denote this connected component by O. Furthermore, we denote by ρ' that part of $\cup_{i=1}^{M-1} \rho_i$, which connects σ_{k_0} to ρ_0. ρ', must necessarily, due to the construction of the curves ρ_i, be an interior boundary of O. A situation corresponding to n=2 and M=3 is illustrated in Fig. 2.

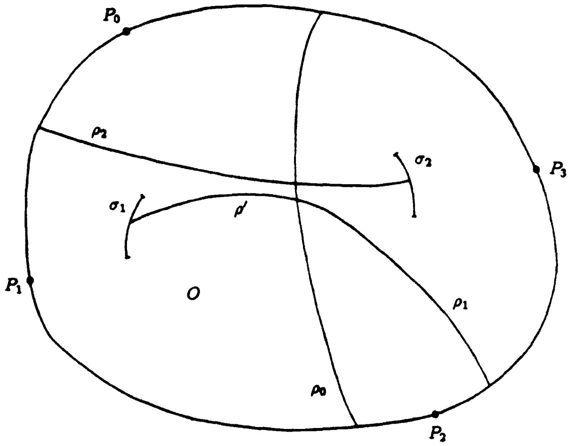


FIG. 2

Let P_{j_0} denote the point which lies on ∂O. We may, without loss of generality, assume that β_{j_0} is negative (so that P_{j_0} is a sink). Consider the maximum of u on

$\overline{O}$. This maximum must be achieved at a boundary point, and it cannot be near P_{j_0}, since u behaves like $-\beta_{j_0}\gamma(P_{j_0})/\pi \ \log r$ there. Since $\frac{\partial u}{\partial \nu}$ is zero on $\partial\Omega \setminus \cup_{j=0}^{M}\{P_j\}$ it now follows from the strong version of the maximum principle that the maximum of u on $\overline{O}$ is attained on that part of ∂O which comes from $\cup_{i=0}^{M-1}\rho_i$; in particular, the maximum is attained all along ρ'. Let x_0 be an interior point on ρ' and let $B \subset O \cup \rho'$ be a ball centered at x_0 which does not intersect Σ. The function u satisfies the elliptic equation $\nabla \cdot (\gamma^{-1}\nabla u) = 0$ in B (and is not constant), and therefore, by the maximum principle,

$$\inf_B u(x) < u(x_0) < \sup_B u(x). \tag{2.6}$$

On the other hand, $B \subset \overline{O}$ so

$$u(x_0) = \max_{\overline{O}} u(x) \geq \sup_B u(x),$$

and this immediately leads to a contradiction with (2.6). Hence we conclude that also in the case $M + (M - M') = n + 1$ we cannot have that Σ and $\tilde{\Sigma}$ are different, and this completes the proof of Theorem 0.1. □

3. Proof of Theorem 0.2. The proof of Theorem 0.2 goes entirely along the lines of the previous proof up to and including the construction of the function u and the curves ρ_i (using the equivalent of Lemmas 1.1 and 1.2 with Dirichlet boundary conditions of the form $\sum \beta_j 1_{\overset{\frown}{P_{j-1}P_j}}$). From there the proof proceeds as outlined below.

The points $P_0, \cdots, P_M$ divide the boundary $\partial\Omega$ into $M+1$ half-open curves

$$\overset{\frown}{P_0P_1} \cup \{P_0\}, \cdots, \overset{\frown}{P_{M-1}P_M} \cup \{P_{M-1}\}, \quad \text{and} \quad \overset{\frown}{P_MP_0} \cup \{P_M\}.$$

Here we have used the notation $\overset{\frown}{PQ}$ for the counterclockwise curve from P to Q, excluding the endpoints. We shall use the notation $[P,Q)$ for the counterclockwise curve from P to Q, including P: $[P,Q) = \overset{\frown}{PQ} \cup \{P\}$. The function u is constant on each of the curves $\overset{\frown}{P_0P_1}, \cdots, \overset{\frown}{P_{M-1}P_M}$, and $\overset{\frown}{P_MP_0}$ (on the last curve, u is actually zero). The curves $\rho_0, \cdots, \rho_{M-1}$ have at least $2M-n$ terminal points on the boundary of Ω, and we note that in this case the curves may very well terminate at one or more of the points $P_0, \cdots, P_M$. If $2M - n > M+1$ (i.e., $M > n+1$) it therefore follows that at least one of the curves $[P_0, P_1), \cdots, [P_{M-1}, P_M)$, and $[P_M, P_0)$ contains two terminal points of $\cup_{i=0}^{M-1}\rho_i$. Consequently, there is a connected component of $\Omega \setminus \cup_{i=0}^{M-1}\rho_i$ which as its boundary has a level curve of u—a contradiction.

We now consider the remaining case: $M = n+1$. In this case we conclude that any one of the curves $[P_0, P_1), \cdots, [P_{M-1}, P_M)$, and $[P_M, P_0)$ contains exactly one terminal point of $\cup_{i=0}^{M-1}\rho_i$. This leaves $2M - (M+1) = M - 1 = n$ terminal points of the curves $\rho_0, \cdots, \rho_{M-1}$ that fall on cracks—one on each crack of Σ. We also see that there are exactly $M+1$ connected components of $\Omega \setminus \cup_{i=0}^{M-1}\rho_i$. For each connected component, that part of the boundary which is shared with $\partial\Omega$ consists of a single curve between two adjacent terminal points of $\cup_{i=0}^{M-1}\rho_i$ (these points lie on adjacent curves $[P_{j-1}, P_j)$ and $[P_j, P_{j+1})$, indices counted modulo M+1). As in the proof of Theorem 0.1, we may argue that $n \geq 1$, so that Σ contains at least one crack σ_{k_0}. Let O denote the connected component of $\Omega \setminus \cup_{i=0}^{M-1}\rho_i$, whose closure contains σ_{k_0}, and let ρ' denote that part of $\cup_{i=0}^{M-1}\rho_i$ which connects σ_{k_0} to ρ_0. ρ' must necessarily, due to the construction of the curves ρ_i, be an interior boundary of O. The (interior) part of the boundary of O, which is shared with $\partial\Omega$, consists of a curve $\overset{\frown}{PQ}$, with $P \in [P_{j_0-1}, P_{j_0})$ and $Q \in [P_{j_0}, P_{j_0+1})$ for some j_0 (mod $M+1$). The rest of the boundary of O is a level curve for u. It is now very easy to see that u takes at most two values on ∂O. Consequently, either the minimum or the maximum of u on $\overline{O}$ is attained on ρ'. This leads to a contradiction, just as in the proof of Theorem 0.1 □

Acknowledgment. This work was completed while the second author visited the Mathematical Sciences Research Institute, Berkeley, California, which is supported by a grant from the National Science Foundation.

REFERENCES

[1] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements*, Indiana Univ. Math. J., 38 (1989), pp. 527–556.

[2] F. Santosa and M. Vogelius, *A computational algorithm to determine cracks from electrostatic boundary measurements*, Internat. J. Engrg. Sci., 29 (1991), pp. 917–937.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 959-964, July 1992

ORTHOGONAL POLYNOMIALS AND A DISCRETE BOUNDARY VALUE PROBLEM I*

RYSZARD SZWARC†

Abstract. Let $\{P_n\}_{n=0}^{\infty}$ be a system of orthogonal polynomials with respect to a measure μ on the real line. Sufficient conditions are given under which any product P_nP_m is a linear combination of P_k's with positive coefficients.

Key words. orthogonal polynomials, recurrence formula

AMS(MOS) subject classifications. 33A65, 39A70

Let us consider the following problem: we are given a probability measure μ on the real line $\mathbf{R}$ all of whose moments $\int x^{2n}d\mu(x)$ are finite. Let $\{P_n(x)\}$ be an orthonormal system in $L^2(\mathbf{R}, d\mu)$ obtained from the sequence $1, x, x^2, \cdots$ by the Gram-Schmidt procedure. We assume that the support of μ is an infinite set so that $1, x, x^2, \ldots$ are linearly independent. Clearly P_n is a polynomial of degree n which is orthogonal to all polynomials of degree less than n. It can be taken to have positive leading coefficients. The product P_nP_m is a polynomial of degree $n+m$ and it can be expressed uniquely as a linear combination of polynomials $P_0, P_1, \cdots, P_{n+m}$,

$$P_nP_m = \sum_{k=0}^{n+m} c(n, m, k)P_k$$

with real coefficients $c(n, m, k)$. Actually, if $k<|n-m|$ then $c(n, m, k)=0$. This is because

$$c(n, m, k) = \langle P_nP_m, P_k\rangle_{L^2(d\mu)} = \langle P_n, P_mP_k\rangle_{L^2(d\mu)} = \langle P_m, P_nP_k\rangle_{L^2(d\mu)}.$$

Hence if $k<|n-m|$ then either $k+m<n$ or $k+n<m$ and one of the above scalar products vanishes. Finally we get

$$P_nP_m = \sum_{k=|n-m|}^{n+m} c(n, m, k)P_k. \tag{1}$$

We ask when the coefficients $c(n, m, k)$ are nonnegative for all $n, m, k = 0, 1, 2, \cdots$. The positivity of coefficients $c(n, m, k)$ (called also the linearization coefficients) gives rise to a convolution structure on $l^1(N)$ and if some additional boundedness condition is satisfied then l^1 with this new operation resembles l^1 of the circle (see [2]).

Analogously to (1), we have

$$xP_n = \gamma_nP_{n+1} + \beta_nP_n + \alpha_nP_{n-1} \quad \text{for } n = 0, 1, 2, \cdots \tag{2}$$

(we apply the convention $\alpha_0 = \gamma_{-1} = 0$). The coefficients α_n and γ_n are strictly positive. If the measure μ is symmetric, i.e., $d\mu(x) = d\mu(-x)$, then $\beta_n = 0$. When P_n are normalized so that $\|P_n\|_{L^2(\mu)} = 1$ then we can check easily that $\alpha_{n+1} = \gamma_n$. Hence, if we put $\lambda_n = \gamma_n$ we get

$$xP_n = \lambda_nP_{n+1} + \beta_nP_n + \lambda_{n-1}P_{n-1} \quad \text{for } n = 0, 1, 2, \cdots. \tag{3}$$

* Received by the editors January 1, 1990; accepted for publication August 16, 1991.

† Mathematical Institute, University of Wrocław, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland.

Favard [4] proved that the converse is also true, i.e., any system of polynomials satisfying (3) is orthonormal with respect to a probability measure μ (not necessarily unique). In case of bounded sequences λ_n and β_n we can recover the measure μ in the following way. Consider a linear operator L on $l^2(N)$ given by

$$(4) \qquad La_n = \lambda_n a_{n+1} + \beta_n a_n + \lambda_{n-1} a_{n-1}, \qquad n = 0, 1, 2, \cdots.$$

Then L is a self-adjoint operator on $l^2(N)$. Let $dE(x)$ be the spectral resolution associated with L. Then the system $\{P_n\}$ is orthonormal with respect to the measure $d\mu(x) = d\langle E(x)\delta_0, \delta_0\rangle$.

The statement of the positivity of $c(n, m, k)$ does not require orthonormalization of the polynomials P_n. We can as well consider another normalization, i.e., let $\tilde{P}_n = \sigma_n P_n$ where σ_n is a sequence of positive numbers. The problem of positive coefficients in the product of $\tilde{P}'_n s$ is equivalent to that of $P'_n s$. Moreover, it is easy to check that the polynomials $\tilde{P}_n$ satisfy the recurrence relation of the form

$$(5) \qquad x\tilde{P}_n = \gamma_n \tilde{P}_{n+1} + \beta_n \tilde{P}_n + \alpha_n \tilde{P}_{n-1} \quad \text{for } n = 0, 1, 2, \cdots$$

and the unique relation connecting α_n, γ_n and the coefficients λ_n from (3) is $\alpha_{n+1}\gamma_n = \lambda_n^2$; the sequence of diagonal coefficients β_n remains unchanged. From this observation it follows that if polynomials $\tilde{P}_n$ satisfy (5) then after appropriate renormalization the polynomials $\bar{P}_n = c_n \tilde{P}_n$ satisfy

$$(6) \qquad x\bar{P}_n = \alpha_{n+1}\bar{P}_{n+1} + \beta_n \bar{P}_n + \gamma_{n-1}\bar{P}_{n-1}.$$

Consider the particular case of monic normalization, i.e., assume that the leading coefficient of any P_n is 1. Then the recurrence formula is

$$(7) \qquad xP_n = P_{n+1} + \beta_n P_n + \lambda_{n-1}^2 P_{n-1}.$$

In 1970 Askey proved the following theorem concerning the monic case.

THEOREM (Askey [1]). *Let P_n satisfy (6) and let the sequences λ_n and β_n be increasing ($\lambda_n \geqq 0$); then the linearization coefficients in the formula*

$$P_n P_m = \sum_{k=|n-m|}^{n+m} c(n, m, k) P_k$$

are nonnegative.

This theorem applies to the Hermite, Laguerre, and Jacobi polynomials with $\alpha + \beta \geqq 1$ (see [7]). However, it does not cover the symmetric Jacobi polynomials with $\alpha = \beta$ when $-\frac{1}{2} \leqq \alpha \leqq \frac{1}{2}$ (and, in particular, the Legendre polynomials when $\alpha = \beta = 0$). Recall that the problem of positive linearization for Jacobi polynomials was completely solved by Gaspar in [5] and [6]. In particular, $c(n, m, k)$ are positive for $\alpha \geqq \beta$ and $\alpha + \beta + 1 \geqq 0$.

The aim of this paper is to give a generalization of Askey's result so it would cover the symmetric Jacobi polynomials for $a \geqq -\frac{1}{2}$. One of the results is as follows.

THEOREM 1. *If polynomials P_n satisfy*

$$xP_n = \gamma_n P_{n+1} + \beta_n P_n + \alpha_n P_{n-1}$$

and

(i) *α_n, β_n, and $\alpha_n + \gamma_n$ are increasing sequences (γ_n, $\alpha_n \geqq 0$),*

(ii) *$\alpha_n \leqq \gamma_n$ for $n = 0, 1, 2, \cdots$,*

then $c(n, m, k) \geqq 0$ (see (1)).

It is remarkable that the assumptions on α_n's and γ_n's are separated from that on β_n.

Before giving a proof let us explain how Askey's theorem can be derived from Theorem 1. If polynomials $\tilde{P}_n$ satisfy the assumptions of Askey's theorem then after orthonormalization of $\tilde{P}_n$'s we get the system of polynomials P_n satisfying (3), i.e.,

$$xP_n=\lambda_nP_{n+1}+\beta_nP_n+\lambda_{n-1}P_{n-1} \quad \text{for } n=0,1,2,\cdots,$$

and if λ_n and β_n are increasing then putting $\alpha_n=\lambda_{n-1}$ and $\gamma_n=\lambda_n$ we can see easily that the assumptions of Theorem 1 are also satisfied.

Example. Consider the symmetric Jacobi polynomials $R_n^{(\alpha,\alpha)}$ normalized by $R_n^{\alpha,\alpha}(1)=1$. They satisfy the following recurrence formula:

$$xR_n^{(\alpha,\alpha)}=\frac{n+2\alpha+1}{2n+2\alpha+1}R_{n+1}^{(\alpha,\alpha)}+\frac{n}{2n+2\alpha+1}R_{n-1}^{(\alpha,\alpha)}.$$

In this case

$$\alpha_n=\frac{n}{2n+2\alpha+1},\quad \gamma_n=\frac{n+2\alpha+1}{2n+2\alpha+1},\quad \beta_n=0.$$

Observe that $\alpha_n+\gamma_n=1$ and α_n is increasing when $\alpha\geqq-\frac{1}{2}$. We have also $\alpha_n\leqq\gamma_n$ when $\alpha\geqq-\frac{1}{2}$.

Instead of showing Theorem 1 we will prove a more general result.

THEOREM 2. *Let polynomials P_n satisfy*

$$xP_n=\gamma_nP_{n+1}+\beta_nP_n+\alpha_nP_{n-1}$$

and let for some sequence of positive numbers σ_n polynomials $\bar{P}_n=\sigma_nP_n$ satisfy

$$x\bar{P}_n=\gamma'_n\bar{P}_{n+1}+\beta_n\bar{P}_n+\alpha'_n\bar{P}_{n-1}.$$

Assume also that

(i) $\beta_m\leqq\beta_n$ *for any* $m\leqq n$,
(ii) $\alpha_m\leqq\alpha'_n$ *for any* $m<n$,
(iii) $\alpha_m+\gamma_m\leqq\alpha'_n+\gamma'_n$ *for any* $m<n-1$,
(iv) $\alpha_m\leqq\gamma'_n$ *for any* $m\leqq n$.

Then the linearization coefficients $c(n,m,k)$ in the formula

$$P_nP_m=\sum_{k=|n-m|}^{n+m}c(n,m,k)P_k$$

are nonnegative.

Setting $\alpha'_n=\alpha_n$ and $\gamma'_n=\gamma_n$, we can easily see that Theorem 2 implies Theorem 1.

Proof. First observe that we have $\alpha_{n+1}\gamma_n=\alpha'_{n+1}\gamma'_n$. Moreover, by the remarks preceding (6) we may assume that P_n and $\bar{P}_n$ satisfy

$$xP_n=\alpha_{n+1}P_{n+1}+\beta_nP_n+\gamma_{n-1}P_{n-1},$$
$$x\bar{P}_n=\alpha'_{n+1}\bar{P}_{n+1}+\beta_n\bar{P}_n+\gamma'_{n-1}\bar{P}_{n-1}.$$

The rest of the proof will follow from the maximum principle for a discrete boundary value problem.

Let L and L' be linear operators acting on sequences $\{a_n\}_{n\in N}$ by the rule

$$La_n=\alpha_{n+1}a_{n+1}+\beta_na_n+\gamma_{n-1}a_{n-1},$$
(8)
$$L'a_n=\alpha'_{n+1}a_{n+1}+\beta'_na_n+\gamma'_{n-1}a_{n-1}.$$

Let L_m and L'_n denote the operators acting on complex functions $u(n, m)$, n, $m \in N$, as L and L' but with respect to the m- or n-variable treating the other variable as a parameter.

Let us consider the following problem: $N \times N \ni (n, m) \mapsto u(n, m) \in C$ and

$$(L'_n - L_m)u = 0, \qquad u(n, 0) \geqq 0. \tag{9}$$

THEOREM 3. *Assume that* $\alpha_n > 0$ *for* $n \geqq 1$ (*we follow the convention* $\alpha_0 = \alpha'_0 = 0$) *and*

(i) $\beta_m \leqq \beta'_n$ *for any* $m \leqq n$,

(ii) $\alpha_m \leqq \alpha'_n$ *for any* $m < n$,

(iii) $\alpha_m + \gamma_m \leqq \alpha'_n + \gamma'_n$ *for any* $m < n - 1$,

(iv) $\alpha_m \leqq \gamma'_n$ *for any* $m \leqq n$.

Then $u(n, m) \geqq 0$ *for* $m \leqq n$.

Proof. On the contrary, assume that u is negative at some points. Let $(n, m+1)$ be the lowest point in the domain $\{(s, t)\colon s \geqq t\}$ for which $u(n, m+1) < 0$. It means that $u(s, t)$ is nonnegative if $t \leqq m$. Consider the rectangular triangle with vertices $A(n, m)$, $B(n-m, 0)$ and $C(n+m, 0)$, as illustrated in Fig. 1.

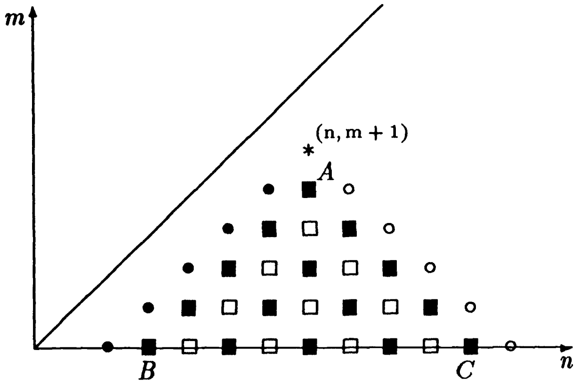


FIG. 1

All lattice points in ΔABC we divide into two subsets: Ω_1, consisting of the points (k, l) such that $k - l = n - m (\text{mod } 2)$, and the rest Ω_2. In the figure the points of Ω_1 are marked by ■ while the points of Ω_2 are marked by □. Let Ω_3 denote the lattice points connecting $(n-m-1, 0)$ and $(n, m+1)$ (except $(n, m+1)$) and Ω_4 denote those which connect $(n+m+1, 0)$ with $(n, m+1)$ (except $(n, m+1)$). The points of Ω_3 and Ω_4 are marked by ● and ○, respectively.

Assume that $(L'_n - L_m)u = 0$. Thus $\sum_{(x,y)\in\Omega_1} (L'_n - L_m)u(x, y) = 0$. If we calculate the terms $(L'_n - L_m)u(x, y) = 0$ and we sum them up we will obtain a sum of the values of the function $u(s, t)$ with some coefficients $c_{s,t}$ where (s, t) runs throughout the sets $\Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4 \cup \{(n, m+1)\}$. Namely,

$$\begin{aligned} 0 &= \sum_{(x,y)\in\Omega_1} (L'_n - L_m)u(x, y) \\ &= \sum_{i=1}^{4} \sum_{(s,t)\in\Omega_1} c_{s,t}u(s, t) + c_{n,m+1}u(n, m+1). \end{aligned}$$

It is not hard to compute the coefficients $c_{s,t}$, so we just list them below.

(i) $(s, t) \in \Omega_1$; $c_{s,t} = \beta'_s - \beta_t$.
(ii) $(s, t) \in \Omega_2$; $c_{s,t} = \alpha'_s + \gamma'_s - (\alpha_t + \gamma_t)$.
(iii) $(s, t) \in \Omega_3$; $c_{s,t} = \gamma'_s - \alpha_t$.
(iv) $(s, t) \in \Omega_4$; $c_{s,t} = \alpha'_s - \alpha_t$.
(v) $c_{n,m+1} = -\alpha_{m+1}$.

By the assumptions of the theorem all coefficients $c_{s,t}$ are nonnegative while $c_{n,m+1}$ is strictly negative. Since $u(s, t) \geqq 0$ for $(s, t) \in \Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4$ and $u(n, m+1) < 0$ then the sum we were dealing with cannot be zero. It gives a contradiction.

Let us return to the proof of Theorem 2. Let P_n and $\bar{P}_n$ satisfy (8) and $\bar{P}_n = \sigma_n P_n$ for a strictly positive sequence σ_n. If

$$P_n P_m = \sum_{k=|n-m|}^{n+m} c(n, m, k) P_k,$$

then

$$\bar{P}_n P_m = \sum_{k=|n-m|}^{n+m} \bar{c}(n, m, k) P_k,$$

where $\bar{c}(n, m, k) = c(n, m, k)\sigma_n$. Therefore in order to prove $c(n, m, k) \geqq 0$ it suffices to show that $\bar{c}(n, m, k) \geqq 0$ for $n > m$. Since

$$L'_n(\bar{P}_n P_m) = x\bar{P}_n P_m = L_m(\bar{P}_n P_m)$$

and the polynomials P_n are linearly independent then for any k the function $u(n, m) = \bar{c}(n, m, k)$ is a solution of (9). Obviously,

$$u(n, 0) = c(n, 0, k)\sigma_n = \begin{cases} \sigma_n & \text{if } n = k, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, $u(n, 0) \geqq 0$. Hence by Theorem 3 we get $u(n, m) = \bar{c}(n, m, k) \geqq 0$. This completes the proof of Theorem 2.

COROLLARY. *Let polynomials P_n satisfy $xP_n = \gamma_n P_{n+1} + \beta_n P_n + \alpha_n P_{n-1}$ and let*
(i) *β_n and α_n be increasing $(\alpha_n > 0$ for $n \geqq 1$, $\alpha_0 = 0)$;*
(ii) *$\alpha_m + \gamma_m \leqq \alpha_{n+1} + \gamma_{n-1}$ for $m < n-1$;*
(iii) *$\alpha_m \leqq \gamma_n$ for $m < n$.*

Then the linearization coefficients $c(n, m, k)$ in (1) *are nonnegative.*

Proof. By remarks preceding (6) after appropriate renormalization of P_n we obtain polynomials P'_n satisfying (6). Then we get the required result by applying Theorem 2.

Example. Consider Jacobi polynomials $P_n^{(\alpha,\beta)}$. They satisfy the recurrence formula

$$xP_n^{(\alpha,\beta)} = \frac{2(n+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)} P_{n+1}^{(\alpha,\beta)}$$
$$+ \frac{\beta^2 - \alpha^2}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)} P_n^{(\alpha,\beta)}$$
$$+ \frac{2(n+\alpha)(n+\beta)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)} P_{n-1}^{(\alpha,\beta)}.$$

Applying the corollary yields that for $\alpha \geqq \beta$ and $\alpha + \beta \geqq 0$ we get positive linearization coefficients. However, for $\alpha \geqq \beta$ and $\alpha + \beta < 0$ the sequence

$$\beta_n = \frac{\beta^2 - \alpha^2}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)}$$

is decreasing and we cannot apply any of the preceding results, although we know from [5] and [6] that the condition $\alpha+\beta+1\geqq 0$ is sufficient.

In part II of this paper we will discuss the problem of positive linearization under assumption β_n is decreasing when starting from $n=1$. This is more delicate because assumptions on α_n's and γ_n's cannot be separated from those on β_n's.

REFERENCES

[1] R. Askey, *Linearization of the product of orthogonal polynomials*, in Problems in Analysis, R. Gunning, ed., Princeton University Press, Princeton, NJ, 1970, pp. 223-228.

[2] ———, *Orthogonal Polynomials and Special Functions*, Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.

[3] R. Askey and G. Gasper, *Linearization of the product of Jacobi polynomials*. III, Canad. J. Math., 23 (1971), pp. 119-122.

[4] J. Favard, *Sur les polynômes de Tchebycheff*, C. R. Acad. Sci. Paris, 200 (1935), pp. 2052-2055.

[5] G. Gasper, *Linearization of the product of Jacobi polynomials*. I, Canad. J. Math., 22 (1970), pp. 171-175.

[6] ———, *Linearization of the product of Jacobi polynomials*. II, Canad. J. Math., 22 (1970), pp. 582-593.

[7] G. Szegö, *Orthogonal Polynomials*, Fourth ed., Amer. Math. Soc. Colloq. Publ. 23, American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 965-969, July 1992

011

ORTHOGONAL POLYNOMIALS AND A DISCRETE BOUNDARY VALUE PROBLEM II*

RYSZARD SZWARC†

Abstract. Let $\{P_n\}_{n=0}^{\infty}$ be a system of polynomials orthogonal with respect to a measure μ on the real line. Then P_n satisfy the three-term recurrence formula $xP_n = \gamma_n P_{n+1} + \beta_n P_n + \alpha_n P_{n-1}$. Conditions are given on the sequence α_n, β_n, and γ_n under which any product $P_n P_m$ is a linear combination of P_k with positive coefficients. The result is applied to the measures $d\mu(x) = (1-x^2)^{\alpha}|x|^{2\beta+1}\,dx$ and $d\mu(x) = |x|^{2\alpha+1}e^{-x^2}\,dx$, α, $\beta > -1$. As a corollary, a Gasper result is derived on the Jacobi polynomials $P_n^{(\alpha,\beta)}$ with $\alpha \geqq \beta$ and $\alpha + \beta + 1 \geqq 0$.

Key words. orthogonal polynomials, recurrence formula

AMS(MOS) subject classifications. primary 33A65, 39A70

The present paper is a continuation of our earlier work [9]. We were concerned in part I with the following question. Given a probability measure μ on the real line $\mathbb{R}$ such that all its moments are finite, let $\{P_n\}_{n=0}^{\infty}$ be a system of orthogonal polynomials obtained from the sequence of consecutive monomials $1, x, x^2, \cdots$ by the Gram-Schmidt procedure. We do not impose any special normalization upon P_n except that its leading coefficient be positive. The product $P_n P_m$ is a polynomial of degree $n+m$ and it can be expressed as

$$P_n P_m = \sum_{k=|n-m|}^{n+m} c(n, m, k) P_k \tag{1}$$

with some real coefficients $c(n, m, k)$. We are asking when $c(n, m, k)$ are nonnegative for any n, m, $k \in \mathbb{N}$. The coefficients $c(n, m, k)$ from (1) are called the *linearization coefficients* of $\{P_n\}$ and if they are nonnegative we simply say that the linearization coefficients are nonnegative.

It is well known that P_n that P_n obey a three-term recurrence formula of the form

$$xP_n = \gamma_n P_{n+1} + \beta_n P_n + \alpha_n P_{n-1}, \tag{2}$$

where α_n, γ_n are positive, except $\alpha_0 = 0$, and β_n are real. In [9, Thm. 1], we proved that if $\{\alpha_n\}$, $\{\beta_n\}$, $\{\alpha_n + \gamma_n\}$ are increasing sequences and $\gamma_n \geqq \alpha_n$, for $n = 0, 1, 2, \cdots$, then the linearization coefficients of $\{P_n\}$ are nonnegative.

Our aim now is to get rid in some way of the condition of the monotonicity of the sequence $\{\beta_n\}$. Roughly the idea consists in reducing the problem to the case $\beta_n = 0$. This can be done in the following way. Consider first polynomials P_n satisfying

$$xP_n = \gamma_n P_{n+1} + \alpha_n P_{n-1}, \qquad P_0 = 1. \tag{3}$$

Then, of course, P_{2n} are even functions while P_{2n+1} are odd ones. Equivalently, this means that the corresponding measure, which orthogonalizes $\{P_n\}$ (and which exists by the Favard theorem [5]) is symmetric with respect to zero. An easy calculation gives the following:

$$\begin{aligned} x^2 P_{2n}(x) = {} & \gamma_{2n+1}\gamma_{2n} P_{2n+2}(x) + (\alpha_{2n+1}\gamma_{2n} + \alpha_{2n}\gamma_{2n-1}) P_{2n}(x) \\ & + \alpha_{2n}\alpha_{2n-1} P_{2n-2}(x). \end{aligned} \tag{4}$$

* Received by the editors May 1, 1990; accepted for publication August 16, 1991.

† Mathematical Institute, University of Wroclaw, pl. Grunwaldzki 2/4, 50-384 Wroclaw, Poland.

Let us define the polynomials Q_n by

$$Q_n(y)=P_{2n}(\sqrt{y}). \tag{5}$$

Then by (4) the polynomials Q_n satisfy

$$yQ_n(y)=\gamma_{2n+1}\gamma_{2n}Q_{n+1}(x)+(\alpha_{2n+1}\gamma_{2n}+\alpha_{2n}\gamma_{2n-1})Q_n(x)+\alpha_{2n}\alpha_{2n-1}Q_{n-1}(x). \tag{6}$$

Observe that (6) is again a three-term recurrence formula. Moreover, if the polynomials P_n have nonnegative linearization coefficients, then by (5) the polynomials Q_n do as well.

We can go the other way around. Assume we are given a sequence of polynomials Q_n orthogonal with respect to a measure ν supported on $[0, +\infty)$. Instead of studying the Q_n we can examine the polynomials P_n satisfying (3) and (5) with regard to the question of nonnegative linearization coefficients. Those are easier to handle, because in (3) the coefficients β_n are missing, unlike in the recurrence formula for Q_n.

First we will sharpen Theorem 1 from [9] in case of symmetric measures.

THEOREM 1. *Let orthogonal polynomials P_n satisfy*

$$xP_n=\gamma_nP_{n+1}+\alpha_nP_{n-1}, \qquad n=0,1,2,\cdots, \tag{7}$$

where $\alpha_0=0$, α_n, $\gamma_n\geqq 0$. Assume that the sequences $\{\alpha_{2n}\}$, $\{\alpha_{2n+1}\}$, $\{\alpha_{2n}+\gamma_{2n}\}$, $\{\alpha_{2n+1}+\gamma_{2n+1}\}$ are increasing and $\alpha_n\leqq\gamma_n$ for $n=0,1,2,\cdots$. Then the linearization coefficients of P_n are nonnegative.

Proof. As in [9], Remark 1, we can renormalize P_n (i.e., multiply each P_n by a positive number σ_n) so as to satisfy

$$xP_n=\alpha_{n+1}P_{n+1}+\gamma_{n-1}P_{n-1}. \tag{8}$$

Of course, it does not affect the conclusion of the theorem, so we introduce no new symbols for the renormalized polynomials. Let μ be a symmetric probability measure that orthogonalizes the polynomials P_n. Then by (1)

$$c(n,m,k)\int P_k^2\,d\mu=\int P_nP_mP_k\,d\mu. \tag{9}$$

Hence the quantity $c(n,m,k)\int P_k^2\,d\mu$ is invariant under permutations of n, m, k. Since μ is symmetric, then $c(n,m,k)=0$ if n, m, k are all odd numbers. Thus if $c(n,m,k)\neq 0$ then one of n, m, k is an even number. By invariance, we can always assume that k is such. Collecting all of the above it suffices to show that in the formulas

$$\begin{aligned} P_{2n}P_{2m}&=\sum c(2n,2m,2k)P_{2k},\\ P_{2n+1}P_{2m+1}&=\sum c(2n+1,2m+1,2k)P_{2k}\end{aligned} \tag{10}$$

the coefficients $c(2n,2m,2k)$ and $c(2n+1,2m+1,2k)$ are nonnegative. It automatically implies that they are also nonnegative in the formula

$$P_{2n+1}P_{2m}=\sum c(2n+1,2m,2k+1)P_{2k+1}. \tag{11}$$

Let L be the linear operator acting on the sequences $\{a_n\}_{n=0}^{\infty}$ by

$$La_n=\alpha_{n+1}a_{n+1}+\gamma_{n-1}a_{n-1}. \tag{12}$$

Let L_n and L_m denote the linear operators acting on the matrices $\{u(n,m)\}_{n,m=0}^{\infty}$ as the operator L does but according to the n or m variable (cf. [9]). Fix $k\in\mathbb{N}$ and consider the matrix $u(n,m)=c(n,m,k)$. By (8) and (9) (cf. [9]) we have $(L_n-L_m)u=0$. Moreover, $u(n,0)=1$ for $n=2k$ and $u(n,0)=0$ otherwise. Hence the following maximum principle would complete the proof.

LEMMA 1. *Let the matrix $u(n, m)$, $n, m = 0, 1, 2, \cdots$ satisfy*

$$(13)\qquad \begin{gathered}(L_n - L_m)u = 0\\ u(2n, 0) \geqq 0, \quad u(2n+1, 0) = 0, \quad n = 0, 1, 2, \cdots.\end{gathered}$$

Then (under the assumptions of Theorem 1) $u(n, m) \geqq 0$ for $n \geqq m$.

For the proof of Lemma 1 we refer the reader to [9] (the proof of Theorem 3). It suffices to observe that (10) and (11) imply $u(n, m) = 0$ whenever $n + m$ is an odd number. Hence, scanning the proof of Theorem 3 from [9], we can observe that the coefficients $c_{s,t}$, which are computed there, have the property that $s + r$ is an even number.

Combining Theorem 1, (4), (5), and (6) immediately gives the following corollary.

COROLLARY 1. *Let the orthogonal polynomials $Q_n(y)$ satisfy the recurrence formula*

$$yQ_n = \tilde{\gamma}_n Q_{n+1} + \tilde{\beta}_n Q_n + \tilde{\alpha}_n Q_{n-1}.$$

Assume that there exist sequences α_n, γ_n of nonnegative numbers ($\alpha_0 = 0$) and a real constant β such that

$$(14)\qquad \tilde{\gamma}_n = \gamma_{2n+1}\gamma_{2n}, \quad \tilde{\alpha}_n = \alpha_{2n}\alpha_{2n-1}, \quad \tilde{\beta}_n = \alpha_{2n+1}\gamma_{2n} + \alpha_{2n}\gamma_{2n-1} + \beta,$$

and α_n, γ_n satisfy the assumptions of Theorem 1. Then the linearization coefficients of Q_n are nonnegative.

Before giving applications of Corollary 1 let us study the relation between orthogonal polynomials P_n and Q_n connected by (3) and (5). Let μ be a measure that orthogonalizes the polynomials P_n. Then

$$\begin{aligned}0 &= \int_{-\infty}^{+\infty} P_{2n}(x)P_{2m}(x)\,d\mu(x) = 2\int_0^{+\infty} P_{2n}(x)P_{2m}(x)\,d\mu(x)\\ &= 2\int_0^{+\infty} Q_n(y)Q_m(y)\,d\mu(\sqrt{y}).\end{aligned}$$

Hence Q_n are orthogonal with respect to the measure $d\nu(y) = 2d\mu(\sqrt{y})$, $y \geqq 0$. Note that the measure μ can be recovered back from ν by $d\mu(x) = \frac{1}{2}d\nu(x^2)$, $x \geqq 0$, and $d\mu(-x) = d\mu(x)$.

It is worthwhile to look at the polynomials R_n defined by

$$S_n(y) = \frac{1}{\sqrt{y}} P_{2n+1}(\sqrt{y}).$$

Then

$$\begin{aligned}0 &= \int_{-\infty}^{+\infty} P_{2n+1}(x)P_{2m+1}(x)\,d\mu(x)\\ &= 2\int_0^{+\infty} x^2 \frac{P_{2n+1}(x)}{x}\frac{P_{2m+1}(x)}{x}\,d\mu(x)\\ &= 2\int_0^{+\infty} S_n(y)S_m(y)y\,d\mu(\sqrt{y}).\end{aligned}$$

Hence the measure that orthogonalizes the S_n is $2y\,d\mu(\sqrt{y})$ or simply $y\,d\nu(y)$.

THEOREM 2. *Let $\{P_n\}_{n=0}^{\infty}$ be the system of polynomials orthogonal with respect to the measure $d\mu(x) = (1-x^2)^{\alpha}|x|^{2\beta+1}\,dx$, $x \in (-1, 1)$, $\alpha, \beta > -1$. If $\alpha \geqq \beta$ and $\alpha + \beta + 1 \geqq 0$, then the coefficients $c(n, m, k)$ in $P_nP_m = \sum c(n, m, k)P_k$ are nonnegative.*

Proof. It suffices to find a three-term recurrence formula for P_n so as to fulfill the assumptions of Theorem 1.

LEMMA 2. *The polynomials* $\{P_n\}_{n=0}^{\infty}$ *satisfying*

$$xP_{2n}=\frac{n+\alpha+\beta+1}{2n+\alpha+\beta+1}P_{2n+1}+\frac{n}{2n+\alpha+\beta+1}P_{2n-1}, \tag{15}$$

$$xP_{2n-1}=\frac{n+\alpha}{2n+\alpha+\beta}P_{2n}+\frac{n+\beta}{2n+\alpha+\beta}P_{2n-2} \tag{16}$$

for $n=0,1,2,\cdots$, $(P_0=1)$ *are orthogonal with respect to the measure* $d\mu(x)=(1-x^2)^{\alpha}|x|^{2\beta+1}\,dx$.

Proof of Lemma 2. Let $R_n^{(\alpha,\beta)}(y)$ denote the Jacobi polynomials normalized by $R_n^{(\alpha,\beta)}(1)=1$. Let

$$\tilde{Q}_n(y)=R_n^{(\alpha,\beta)}(2y-1). \tag{17}$$

Then $\tilde{Q}_n$ are orthogonal with respect to the measure $d\nu(y)=(1-y)^{\alpha}y^{\beta}\,dy$. By the recurrence formula for $R_n^{(\alpha,\beta)}$ (see [6, (4) p. 172] or [4, (3) and (11), p. 169]), $\tilde{Q}_n$ satisfy

$$y\tilde{Q}_n=\frac{(n+\alpha+\beta+1)(n+\alpha+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}\tilde{Q}_{n+1}$$
$$+\frac{1}{2}\left(1+\frac{\beta^2-\alpha^2}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)}\right)\tilde{Q}_n$$
$$+\frac{n(n+\beta)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta)}\tilde{Q}_{n-1}.$$

Let P_n be the polynomials satisfying (13). Then by (4) and (6) the polynomials $Q_n(y)=P_{2n}(\sqrt{y})$ satisfy the same recurrence formula as $\tilde{Q}_n$ do. Indeed, in both recurrence formulas the coefficients of Q_{n+1}, Q_{n-1} *and* $\tilde{Q}_{n+1}$, $\tilde{Q}_{n-1}$ coincide. Then the coefficients of Q_n, $\tilde{Q}_n$ must also coincide because in both formulas the sum of coefficients is equal to 1 (for $\tilde{Q}_n=R_n^{(\alpha,\beta)}(1)=1$ and $Q_n(1)=P_{2n}(1)=1$ by (14)). Hence we have just proved that $Q_n=\tilde{Q}_n$. This means Q_n are orthogonal with respect to the measure $d\nu(y)=(1-y)^{\alpha}y^{\beta}\,dy$. Thus by the reasoning of Corollary 1 the polynomials P_n are orthogonal with respect to the measure $d\mu(x)=\frac{1}{2}d\nu(x^2)=(1-x^2)^{\alpha}|x|^{2\beta+1}\,dx$, as was required.

Let us return to the proof of Theorem 2. From Lemma 1 we can easily see that if $\alpha\geqq\beta$ *and* $\alpha+\beta+1\geqq0$ then the assumptions of Theorem 1 are satisfied. This completes the proof.

COROLLARY 2 (Gasper [6]). *Let* $R_n^{(\alpha,\beta)}$ *be the Jacobi polynomials normalized so that* $R_n^{(\alpha,\beta)}(1)=1$. *If* $\alpha\geqq\beta$ *and* $\alpha+\beta+1\geqq0$ *then*

$$R_n^{(\alpha,\beta)}R_m^{(\alpha,\beta)}=\sum_{k=|n-m|}^{n+m}c(n,m,k)R_k^{(\alpha,\beta)}$$

with nonnegative coefficients $c(n,m,k)$.

Proof. Let P_n be the polynomials orthogonal with respect to the measure $d\mu(x)=(1-x^2)^{\alpha}|x|^{2\beta+1}\,dx$ and satisfying (15) and (16). Then by Theorem 2 we have $P_nP_m=\sum d(n,m,k)P_k$, where $d(n,m,k)\geqq0$. From the proof of Lemma 2 we know that $P_{2n}(\sqrt{y})=R_n^{(\alpha,\beta)}(2y-1)$. Hence we get $R_n^{(\alpha,\beta)}R_m^{(\alpha,\beta)}=\sum d(2n,2m,2k)R_k^{(\alpha,\beta)}$, where $d(2n,2m,2k)\geqq0$.

COROLLARY 3. *Let $\alpha \geqq \beta$ and $\alpha+\beta+1 \geqq 0$. Then*

$$(y+1)R_n^{(\alpha,\beta+1)}R_m^{(\alpha,\beta+1)} = \sum_{k=|n-m|}^{n+m} c(n,m,k)R_k^{(\alpha,\beta)},$$

$$R_n^{(\alpha,\beta)}R_m^{(\alpha,\beta+1)} = \sum_{k=|n-m|}^{n+m} d(n,m,k)R_k^{(\alpha,\beta+1)},$$

where $c(n,m,k)$ and $d(n,m,k)$ are nonnegative coefficients.

Proof. Let P_n be the orthogonal polynomials corresponding to the measure $d\mu(x) = (1-x^2)^\alpha |x|^{2\beta+1}\,dx$. Then, as we have seen in the proof of Lemma 2, $P_{2n}(\sqrt{y}) = R_n^{(\alpha,\beta)}(2y-1)$. Let the polynomials $S_n(y)$ be defined as $S_n(y) = (1/\sqrt{y})\,P_{2n+1}(\sqrt{y})$. By the considerations following Corollary 1 we know that $S_n(y)$ are orthogonal with respect to the measure $2y\,d\mu(\sqrt{y}) = (1-y)^\alpha y^{\beta+1}\,dy$ and $S_n(1)=1$. This yields $S_n(y) = R_n^{(\alpha,\beta)}(2y-1)$. Now both required formulas coincide with (10) and (11). The latter have nonnegative coefficients if $\alpha \geqq \beta$ and $\alpha+\beta+1 \geqq 0$.

Now we turn to the so called generalized Hermite polynomials.

THEOREM 3. *Let P_n be the polynomials orthogonal with respect to the measure $d\mu(x) = |x|^{2\alpha+1} e^{-x^2}\,dx$, $\alpha > -1$. Then the P_n have nonnegative linearization coefficients.*

Proof. First we show that P_n satisfy the following recurrence formulas.

$$xP_{2n} = (n+\alpha+1)P_{2n+1} + nP_{2n-1}, \tag{18}$$

$$xP_{2n-1} = P_{2n} + P_{2n-2}. \tag{19}$$

Indeed, let P_n satisfy (18) and (19). Then

$$x^2P_{2n} = (n+\alpha+1)P_{2n+2} + (2n+\alpha+1)P_{2n} + nP_{2n-2}.$$

Hence, putting $Q_n(y) = P_{2y}(\sqrt{y})$ gives

$$yQ_n = (n+\alpha+1)Q_{n+1} + (2n+\alpha+1)Q_n + nQ_{n-1}.$$

Therefore, the polynomials Q_n coincide with the Laguerre polynomials $(-1)^n L_n^{(\alpha)}$, so they are orthogonal with respect to the measure $d\nu(y) = y^\alpha e^{-y}\,dy$. This implies that P_n are orthogonal with respect to the measure $d\mu(x) = \frac{1}{2}d\nu(x^2) = |x|^{2\alpha+1} e^{-x^2}\,dx$. Combining (18), (19) and Theorem 2 yields the conclusion.

REFERENCES

[1] R. ASKEY, *Linearization of the product of orthogonal polynomials*, in Problems in Analysis, R. Gunning, ed., Princeton University Press, Princeton, NJ, 1970, pp. 223-228.

[2] ———, *Orthogonal polynomials and special functions*, Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.

[3] R. ASKEY AND G. GASPER, *Linearization of the product of Jacobi polynomials*, III, Canad. J. Math., 23 (1971), pp. 119-122.

[4] A. ERDÉLVI, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.

[5] J. FAVARD, *Sur les polynômes de Tchebycheff*, C.R. Acad. Sci. Paris, 200 (1935), pp. 2052-2055.

[6] G. GASPER, *Linearization of the product of Jacobi polynomials*, I, Canad. J. Math., 22 (1970), pp. 171-175.

[7] ———, *Linearization of the product of Jacobi polynomials*, II, Canad. J. Math., 22 (1970), pp. 582-593.

[8] G. SZEGÖ, *Orthogonal Polynomials*, Fourth ed., Amer. Math. Soc. Colloq. Publ. 23, American Mathematical Society, Providence, RI, 1975.

[9] R. SZWARC, *Orthogonal polynomials and a discrete boundary value problem*, I, SIAM J. Math. Anal., this issue (1992), pp. 959-964.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 970-983, July 1992

012

THE GROWTH OF POLYNOMIALS BOUNDED AT EQUALLY SPACED POINTS*

DON COPPERSMITH† AND T. J. RIVLIN†

Abstract. If the absolute value of a (real) polynomial of degree d is bounded by 1 at k equally spaced points of the real line, it is of interest to know how large its absolute value on the interval spanned by the points can be. This work provides a fairly definitive answer to this question.

Key words. polynomials, comparison of norms

AMS(MOS) subject classification. 26C05

Introduction. Let $t_0 < t_1 < \cdots < t_m$ be equally spaced points on the real line. Consider the set $\mathbf{P} = P(d, m)$ of real polynomials, $p(x)$, of degree at most d such that $|p(t_i)| \leq 1$, $i = 0, \cdots, m$. Put

$$\|p\| = \|p\|_{[t_0, t_m]} = \max \{|p(x)|: t_0 \leqq x \leqq t_m\},$$

and

$$B(d, m) = \sup_{p \in P(d,m)} \|p\|. \tag{1}$$

This paper is devoted to an investigation of the behavior of $B(d, m)$. Since $B(d, m)$ is infinite for $m < d$ we suppose, henceforth, that $m \geqq d$, and consequently the "sup" in (1) may be replaced by "max". Moreover, since $B(d, m)$ is invariant under linear transformations, $s_i = at_i + b$, $(a \neq 0)$, $i = 0, \cdots, m$, we shall choose convenient sets of equally spaced points in what follows. There is a considerable literature about the properties of $B(d, m)$. Schönhage [4] showed that $B(d, m)$ is bounded if $m > d^2$. Ehlich and Zeller [1], in a brief and elegant paper, improved on Schönhage's result by showing that for $m > d^2/\sqrt{6}$, $B(d, m)$ is bounded by $(1-\rho)^{-1}$, where $\rho = d^2(d^2-1)/(6m^2)$. (In this same paper, they also showed that when the absolute value of a polynomial of degree d is bounded by one at the zeros or extrema of the Chebyshev polynomial, $T_m(x)$, its norm on $[-1, 1]$ is bounded by $(\cos(\pi d/(2m)))^{-1}$, hence certainly bounded if $m \geqq cd$, $c > 1$, a result which will motivate our approach to the problem we shall be considering.) Subsequently Ehlich and Zeller [2] showed that $B(d, m)$ remains bounded if $m > (2/\pi^2)d^2$. As counterpoint to these results, Ehlich [3], by an explicit construction involving Zolotarev polynomials, showed that if $m = o(d^2)$ as $d \to \infty$, then $B(d, m) \to \infty$ as $d \to \infty$. We wish to close the gap between the boundedness results of Schönhage and Ehlich and Zeller, and Ehlich's unboundedness result by establishing a converse to Ehlich's result. Namely, for fixed $\delta > 0$ and d and m growing subject to $m \geqq \delta d^2$, $B(d, m)$ remains bounded, with the bound depending only on δ. In fact, we show that there are positive constants a, b such that

$$B(d, m) < a\, e^{b/\delta}, \tag{2}$$

and, conversely, there are positive constants γ, β with

$$B(d, m) > \gamma\, e^{\beta/\delta},$$

$\delta = m/d^2$.

* Received by the editors February 4, 1991; accepted for publication (in revised form) November 8, 1991.

† IBM Research Division, T. J. Watson Research Center, Mathematical Sciences Department, Yorktown Heights, New York 10598.

In § 1 we gather a small syllabus of elementary material to be used in establishing (2) so that the proof given in § 2 will not be excessively fragmented. Section 3 contains the construction of the polynomials that show that (2) cannot be improved upon in its dependence on δ.

1. Some preliminary information.

(i)

$$f(\varphi)=\frac{1-\cos \pi\varphi/d}{\varphi^2} \tag{3}$$

is monotonically decreasing for $0\leqq \varphi \leqq (d/2)$.

(ii) Suppose $0<j<d$, then

$$\prod_{\substack{i=0\\ i\neq j}}^{d}\left|\cos\frac{\pi i}{d}-\cos\frac{\pi j}{d}\right|=\frac{d}{2^{d-1}}. \tag{4}$$

Proof of (4). Put $\eta_i=\cos(\pi i/d)$, $i=0,\cdots,d$, and

$$\omega(x)=\prod_{i=0}^{d}(x-\eta_i)=\frac{x^2-1}{d2^{d-1}}T_d'(x),$$

where $T_d(x)$ is the Chebyshev polynomial of degree d. Then

$$\prod_{\substack{i=0\\ i\neq j}}^{d}(\eta_j-\eta_i)=\omega'(\eta_j)=\frac{(\eta_j^2-1)}{d2^{d-1}}T_d''(\eta_j)+\frac{2\eta_j}{d2^{d-1}}T_d'(\eta_j)$$

$$=(-1)^j\frac{d}{2^{d-1}},\qquad 0<j<d,$$

since $T_d'(\eta_j)=0$, $0<j<d$, $(1-x^2)T_d''(x)=xT_d'(x)-d^2T_d(x)$ and $T_d(\eta_j)=(-1)^j$. □

(iii)

$$\prod_{i=1}^{d}\sin^2\frac{\pi i}{2d}=\frac{d}{2^{2d-2}}. \tag{5}$$

Proof of (5).

$$\prod_{i=1}^{d-1}\sin^2\frac{\pi i}{2d}=\prod_{i=1}^{d-1}\left(1-\cos^2\frac{\pi i}{2d}\right).$$

But, if $U_k(x)$ is the Chebyshev polynomial of the second kind, then

$$2d=U_{2d-1}(1)=2^{2d-1}\prod_{i=1}^{2d-1}\left(1-\cos\frac{\pi i}{2d}\right)$$

$$=2^{2d-1}\prod_{i=1}^{d-1}\left(1-\cos^2\frac{\pi i}{2d}\right).$$ □

(iv) Suppose $m\leqq n$, then $B(d,n)\leqq B(d,m)$.

Proof.

$$B(d,n)=\max\{\|p\|_{[0,n]}: |p(i)|\leqq 1, i=0,1,\cdots,n\}$$

$$=|\bar{p}(x)|, 0\leqq x\leqq n.$$

Let $k, k+1, \cdots, k+m$ be chosen so that $0 \leqq k$, $k+m \leqq n$ and $k \leqq x \leqq k+m$. Then $|\bar{p}(i)| \leqq 1$, $i = k, \cdots, k+m$, and so

$$\begin{aligned} B(d, m) &= \max \{\|p\|_{[k,k+m]} : |p(i)| \leqq 1, i = k, \cdots, k+m\} \\ &\geqq \|\bar{p}\|_{[k,k+m]} = |\bar{p}(x)|. \end{aligned} \qquad \square$$

2. The upper bound. Let p be a polynomial of degree d. If $d = 0, 1$, the determination of $B(d, m)$ is trivial. If $d = 2$, Schönhage [4] gives the simple result

$$B(2, 2k+1) = 1 + \frac{1}{2k(k+1)},$$

$$B(2, 2k) = 1 + \frac{1}{2k(k+1)}.$$

With no loss of generality, we choose the $m+1$ points at which $|p|$ is bounded by 1 to be the integers $0, 1, \cdots, m$. Given d and m put $\varepsilon = m/d^2$. Suppose $d \leqq m \leqq 8d$ so that $\varepsilon \leqq 8/d$. Fix $p \in \mathbf{P}$ and $x \in [0, m]$. Select $d+1$ consecutive integers, $a_0 < a_1 < \cdots < a_d$, contained in $[0, m]$, such that $a_0 \leqq x \leqq a_d$. Then since $|p(a_i)| \leqq 1$, $i = 0, \cdots, d$, the Lagrange interpolation formula implies that $|p(x)| \leqq 2^d$. Thus

$$B(d, m) \leqq 2^d \leqq 2^{8/\varepsilon}, \qquad d \leqq m \leqq 8d. \tag{6}$$

So we shall assume that

$$m > 8d$$

in the remainder of this section. Furthermore, we assume that $\varepsilon < 1/10$ (hence $\varepsilon > 8/d$ and $d > 80$), m is divisible by 4, and d is even.

Fix $p \in \mathbf{P}$ and choose any $x \in [0, m]$. Put $M = m/4 = \varepsilon d^2/4$. Since $2M = m/2$, either the interval $[\lfloor x \rfloor, \lfloor x \rfloor + 2M]$ or the interval $[\lceil x \rceil - 2M, \lceil x \rceil]$ is contained in $[0, m]$, where $\lfloor x \rfloor$ is the greatest integer $\leqq x$, and $\lceil x \rceil$ is the least integer $\geqq x$. In either case, by a linear transformation of the argument of p, we arrive at the following setting: $x \in [M-1, M]$, $|p(i)| \leqq 1$, $i = -M, -M+1, \cdots, M-1, M$. With this setting, we obtain our bound on $B(d, m)$ by bounding $|p(x)|$, where x is chosen to be the point at which $|p(x)|$ attains its maximum.

Our plan is to choose $d+1$ integers in $[-M, M]$ in such a way that the required bound on $|p(x)|$ can be obtained from the Lagrange interpolation formula for $p(x)$ with respect to these $d+1$ integers. To this end, we next define $d+1$ *integers*, $x_j, j = 0, 1, \cdots, d$, in the interval $[-M, M]$. We would like these integers to be close to the extrema of the Chebyshev polynomial (relative to interval $[-M, M]$), $M \cos((j\pi)/d)$, $j = 0, \cdots, d$, guided by the observation made in the introduction about the efficiency of these nodes. As the Chebyshev points are bunched closely towards the endpoints of the interval, and we require distinct integers there, we will choose the x_j to be *consecutive* integers near $\pm M$. Formally, our choice is

$$x_j = \min \{M - j, \max \{-M + (d-j), \text{Round}\ (M \cos((j\pi)/d))\}\}, \qquad 0 \leqq j \leqq d,$$

where "Round" denotes rounding to the nearest integer. The following description is more useful. Define K to be the largest integer such that

$$M - K \leqq M \cos \frac{K\pi}{d}. \tag{7}$$

K is well approximated by $8/(\pi^2 \varepsilon)$. More precisely, we have the following.

LEMMA 1. *If K is the largest integer such that $M-K\leqq M\cos(K\pi/d)$, then*

$$\frac{8}{\varepsilon\pi^2}-1<K<\frac{8}{\varepsilon\pi^2}\cdot\frac{1}{1-(\pi^2/108)}<\frac{0.893}{\varepsilon}. \tag{8}$$

Moreover,

$$7<K<\frac{d}{9.7}. \tag{9}$$

Proof. Note that if $d\geqq K>d/3$, then $M\cos((K\pi)/d)<M/2$. Thus $M-K<M/2$, but $d\leqq(M/2)<K$, a contradiction which implies that $K\leqq d/3$.

(i) $M-K\leqq M\cos((K\pi)/d)$ implies

$$M-K\leqq M\left(1-\frac{\pi^2K^2}{2d^2}+\frac{\pi^4K^4}{24d^4}\right),$$

which, by the definition of M, and the inequality $K\leqq d/3$, yields

$$K\leqq\frac{8}{\varepsilon\pi^2}+\frac{\pi^2K^3}{12d^2}\leqq\frac{8}{\varepsilon\pi^2}+\frac{\pi^2d^2K}{108d^2}=\frac{8}{\varepsilon\pi^2}+\frac{\pi^2}{108}K, \tag{10}$$

from which the right-hand inequality in (8) follows.

Since $M-(K+1)>M\cos((K+1)\pi/d)>M-M(\pi^2(K+1)^2/2d^2)$, we obtain

$$\frac{M\pi^2}{2d^2}(K+1)>1,$$

and the left-hand inequality in (8) is verified.

(ii) The right-hand inequality in (9) is obtained by iteration of the first inequality in (10), recalling that $(8/\varepsilon)<d$ and beginning with

$$K^3<K\frac{d^2}{\pi^4(1-(\pi^2/108))^2}.$$

The left-hand inequality in (9) follows from (8), since $\varepsilon<1/10$. □

Thus

$$x_j=\begin{cases}M-j, & j\leqq K,\\ -M+(d-j), & (d-j)\leqq K,\\ \text{Round}\left(M\cos\frac{j\pi}{d}\right), & K<j<d-K.\end{cases} \tag{11}$$

The x_j are clearly integers. Furthermore, they are distinct since: (i) $M>2d$, (ii) $(M-K)-M\cos((K+1)\pi/d)>1$, by the definition of K, and (iii) if $K+1\leqq j\leqq d/2$, the mean-value theorem gives

$$M\cos\frac{j\pi}{d}-M\cos\frac{(j+1)\pi}{d}>1$$

(similarly for $(d/2)<j\leqq d$).

Since the x_j are integers in $[-M, M]$ we know that $|p(x_j)|\leqq 1, j=0, 1, \cdots, d$, and the Lagrange interpolation formula yields

$$\begin{aligned}|p(x)|&=\left|\sum_{j=0}^{d}p(x_j)\prod_{i\neq j}\frac{x-x_i}{x_j-x_i}\right|\\ &\leqq\sum_{j=0}^{d}|l_j(x)|,\end{aligned}$$

where

$$l_j(x)=\prod_{\substack{i\neq j\\ i=0}}^{d}\frac{x-x_i}{x_j-x_i}.$$

Our task is now reduced to bounding $|l_j(x)|$, $x\in[M-1, M]$, $j=0, 1, \cdots, d$. Since each factor $|(x-x_i)/(x_0-x_i)|\leqq 1$, we see that $|l_0(x)|\leqq 1$.

(i) *Suppose* $1\leqq j\leqq K$. We break the product

$$\prod_{\substack{i=0\\ i\pm j}}^{d}\left|\frac{x-x_i}{x_j-x_i}\right|$$

into several factors. First, note that if $i=0$,

$$\left|\frac{x-x_0}{x_j-x_0}\right|\leqq\frac{1}{j}.$$

If $1\leqq i\leqq j-1$, then

$$\prod_{i=1}^{j-1}\left|\frac{x-x_i}{x_j-x_i}\right|\leqq\prod_{i=1}^{j-1}\frac{i}{j-i}=1. \tag{12}$$

If $j+1\leqq i\leqq 2K$, then

$$\left|\frac{x-x_i}{x_j-x_i}\right|\leqq\frac{M-x_i}{x_j-x_i}=1+\frac{j}{(M-x_i)-j}\leqq 1+\frac{j}{i-j}=\frac{i}{i-j}.$$

Hence

$$\prod_{i=j+1}^{2K}\left|\frac{x-x_i}{x_j-x_i}\right|\leqq\binom{2K}{j}. \tag{13}$$

Suppose $2K<i\leqq d/2$. Then, in view of § 1 (i),

$$M\left(1-\cos\frac{\pi i}{d}\right)\geqq\frac{4Mi^2}{d^2}. \tag{14}$$

We next show that the slack in (14) more than compensates for the rounding in the x_i. Namely,

$$M-\text{Round}\left(M\cos\frac{\pi i}{d}\right)\geqq\frac{4Mi^2}{d^2}. \tag{15}$$

To establish (15) is suffices to prove

$$M\left(1-\cos\frac{\pi i}{d}-4\frac{i^2}{d^2}\right)\geqq\frac{1}{2} \tag{16}$$

because

$$\text{Round}\left(M\cos\frac{\pi i}{d}\right)-M\cos\frac{\pi i}{d}\leqq\frac{1}{2}.$$

Inequality (15) holds when $i=d/2$. Suppose $2K<i\leqq(d-2)/2$. Put

$$f(i)=\sqrt{2}\sin\frac{\pi i}{2d}-2\frac{i}{d};\qquad g(i)=\sqrt{2}\sin\frac{\pi i}{2d}+2\frac{i}{d}.$$

Then (16) may be written as

$$Mf(i)g(i)\geqq\tfrac{1}{2}. \tag{17}$$

Note that $f(t)$ is concave and nonnegative on $0 \leqq t \leqq d/2$, and $f(t)=0$ if and only if $t=0,\ d/2$. Thus

$$f(i) \geqq \min\left(f(2K+1), f\left(\frac{d-2}{2}\right)\right). \tag{18}$$

But

$$f(2K+1) > \left(\frac{\sqrt{2}\pi}{2}-2\right)\left(\frac{2K+1}{d}\right) - \frac{\sqrt{2}\pi^3}{48}\left(\frac{2K+1}{d}\right)^3 > \frac{2}{d}, \tag{19}$$

the first inequality following from the truncated power series expansion of $\sin(((2K+1)\pi)/2d)$ and the second from Lemma 1, and our assumption that $\varepsilon < 1/10$, and hence $d > 80$. Also,

$$\begin{aligned} f\left(\frac{d-2}{2}\right) &= \cos\frac{\pi}{2d} - \sin\frac{\pi}{2d} - 1 + \frac{2}{d} \\ &> \left(2-\frac{\pi}{2}\right)\frac{1}{d} - \frac{\pi^2}{8}\cdot\frac{1}{d^2} \\ &> \frac{0.413}{d}, \end{aligned}$$

so that, in view of (19), (18) yields

$$f(i) \geqq \frac{0.413}{d}. \tag{20}$$

In a similar manner, we obtain

$$g(i) > (2+\sqrt{2})\left(\frac{2K+1}{d}\right) > (2+\sqrt{2})\left(\frac{16}{\pi^2\varepsilon}-1\right)\frac{1}{d}.$$

Hence

$$\varepsilon g(i) > \frac{5.19}{d},$$

and, finally,

$$Mf(i)g(i) = \frac{\varepsilon d^2}{4} f(i)g(i) > \frac{1}{2},$$

thus establishing (16), and, therefore, (15).

Next we observe that

$$\frac{jd^2}{4M} \leqq \frac{K}{\varepsilon} < (2K)^2,$$

and so, mindful of (15), we obtain

$$\begin{aligned} \prod_{i=2K+1}^{d/2} \left|\frac{x-x_i}{x_j-x_i}\right| &\leqq \prod_{i=2K+1}^{d/2} \frac{M-x_i}{x_j-x_i} = \prod_{i=2K+1}^{d/2} \frac{M-x_i}{(M-x_i)-j} \\ &\leqq \prod_{i=2K+1}^{d/2} \frac{4Mi^2/d^2}{(4Mi^2/d^2)-j} = \prod_{i=2K+1}^{d/2} \frac{i^2}{i^2-(jd^2/4M)} \\ &< \prod_{i=2K+1}^{d/2} \frac{i^2}{(i+2K)(i-2K)} \leqq \binom{4K}{2K} < 2^{4K}. \end{aligned} \tag{21}$$

Finally, for $d/2 < i \leqq d$, we have

$$\prod_{d/2<i\leqq d}\left|\frac{x-x_i}{x_j-x_i}\right| \leqq \prod_{d/2<i\leqq d}\frac{M/2}{(M/2)-K} < \left(\frac{d}{d-K}\right)^d \tag{22}$$

$$\leqq \left(1+\frac{2K}{d}\right)^d < e^{2K},$$

where we have used $M \geqq 2d$ and $K < d/2$.

Upon putting the bounds for the various domains of i together, we find

$$|l_j(x)| \leqq \frac{1}{j} \times 1 \times \binom{2K}{j} \times 2^{4K} \times e^{2K}, \qquad 1 \leqq j \leqq K.$$

Hence

$$\sum_{j=1}^{K} |l_j(x)| \leqq 2^{4K} e^{2K} \sum_{j=1}^{K} \binom{2K}{j} < (2^6 e^2)^K. \tag{23}$$

(ii) *Suppose* $K < j \leqq d/2$. In this range of j the situation is a bit more delicate. The terms, $|l_j(x)|$, are bounded by an exponential in K times the corresponding terms when the "pure" Chebyshev points are employed. The latter are bounded by c/j^2 for some constant c, so the sum is bounded by an exponential in K. We turn to the details.

Let y_i denote the Chebyshev points:

$$y_i = M \cos\frac{\pi i}{d}, \qquad i = 0, \cdots, d.$$

The denominator of $|l_j(x)|$,

$$D_x = \prod_{\substack{i=0\\ i\neq j}}^{d} |x_i - x_j|,$$

will be related to

$$Dy = \prod_{\substack{i=0\\ i\neq j}}^{d} |y_i - y_j| = M^d \prod_{\substack{i=0\\ i\neq j}}^{d} \left|\cos\frac{\pi i}{d} - \cos\frac{\pi j}{d}\right| = 2d\left(\frac{M}{2}\right)^d, \tag{24}$$

the last equality following from (4).

The difference between D_x and D_y is due to two effects: namely, the rounding of y_i to produce x_i for $K < i < d-K$, and the replacement of y_i by consecutive integers for $0 \leqq i \leqq K$ and $d-K \leqq i \leqq d$.

If $0 \leqq i \leqq K$, then

$$\left|\frac{x_i - x_j}{y_i - y_j}\right| = \frac{M - x_j - i}{M\cos\frac{\pi i}{d} - M\cos\frac{\pi j}{d}} \geqq \frac{M - x_j - i}{M\cos\frac{\pi i}{d} - \left(x_j - \frac{1}{2}\right)}. \tag{25}$$

But by the definition of K there exists α, $K \leqq \alpha < K+1$, such that

$$M\cos\frac{\pi\alpha}{d} = M - \alpha,$$

and hence, in view of § 1 (i)

$$\frac{1-\cos\dfrac{\pi i}{d}}{i^2} \geqq \frac{1-\cos\dfrac{\pi \alpha}{d}}{\alpha^2} = \frac{1}{\alpha M}.$$

Thus, since $\alpha < K+1$,

$$M\cos\frac{\pi i}{d} \leqq M - \frac{i^2}{K+1},$$

and, in view of (25)

$$\left|\frac{x_i-x_j}{y_i-y_j}\right| > \frac{(M-x_j)-i}{(M-x_j)-\left(\dfrac{i^2}{K+1}-\dfrac{1}{2}\right)} > \frac{K+1-i}{K+\dfrac{3}{2}-\dfrac{i^2}{K+1}} \geqq \frac{1}{3}, \tag{26}$$

the last inequality holding since $(K+1-i)/(K+(3/2)-(i^2/(K+1)))$ takes its minimum when $i=K$. Inequality (26) yields

$$\prod_{i=0}^{K}\left|\frac{x_i-x_j}{y_i-y_j}\right| \geqq \left(\frac{1}{3}\right)^{K+1}. \tag{27}$$

Similarly, if $d-K \leqq i \leqq d$ we obtain

$$\prod_{i=d-K}^{d}\left|\frac{x_i-x_j}{y_i-y_j}\right| \geqq \left(\frac{1}{3}\right)^{K+1}. \tag{28}$$

If $K<i<d-K$ $(i \neq j)$ we need only worry about rounding. Now

$$\left|\frac{x_i-x_j}{y_i-y_j}\right| \geqq 1-\frac{1}{|y_i-y_j|}. \tag{29}$$

Moreover,

$$|y_i-y_j| \geqq \frac{4}{3}. \tag{30}$$

For,

$$|y_i-y_j| = |i-j|\frac{M\pi}{d}\sin\frac{\varphi\pi}{d} > \frac{M\pi}{d}\sin\frac{K\pi}{d}$$

by the mean-value theorem, $|i-j| \geqq 1$, and φ is between i and j. Thus

$$|y_i-y_j| > \frac{M\pi}{d}\left(\frac{\pi K}{d}-\frac{\pi^3K^3}{6d^3}\right) > \frac{\varepsilon\pi^2K}{4}\left(1-\frac{\pi^2}{486}\right)$$

since $K<d/9$. Inequality (30) now follows from the first inequality in (8).

We now get

$$\log\left|\frac{y_i-y_j}{x_i-x_j}\right| \leqq -\log\left(1-\frac{1}{|y_i-y_j|}\right) \leqq \frac{2}{|y_i-y_j|} = \frac{2}{2M\sin\dfrac{\pi(i+j)}{2d}\sin\dfrac{\pi(|i-j|)}{2d}},$$

where the first inequality follows from (29) and the second from (30) and the observation that (if we put $|y_i-y_j|^{-1}=x$) $f(x)=2x+\log(1-x)$ is positive for $0<x<3/4$, since $f(x)$ is concave on $[0,3/4]$, $f(0)=0$ and $f(3/4)>0$ (as $e^{3/4}>2$).

Since $j \leqq d/2$ we have

$$\frac{|\pi(i \pm j)|}{2d} \leqq \frac{3\pi}{4},$$

and for $|t| \leqq 3\pi/4$ the inequality

$$\left|\frac{\sin t}{t}\right| \geqq \frac{2\sqrt{2}}{3\pi}$$

holds. Thus

$$\log\left|\frac{y_i - y_j}{x_i - x_j}\right| \leqq \frac{2}{2M\left(\frac{2\sqrt{2}}{3\pi}\right)\frac{\pi(i+j)}{2d}\left(\frac{2\sqrt{2}}{3\pi}\right)\frac{\pi(|i-j|)}{2d}} = \frac{18}{\varepsilon(i+j)(|i-j|)}.$$

Then

$$\begin{aligned}\sum_{\substack{K<i<d-K\\ i\neq j}} \log\left|\frac{y_i - y_j}{x_i - x_j}\right| &\leqq \frac{18}{\varepsilon}\left(\sum_{i=K+1}^{j-1}\frac{1}{(i+j)(j-i)} + \sum_{i=j+1}^{d-K-1}\frac{1}{(i+j)(i-j)}\right)\\ &\leqq \frac{18}{\varepsilon}\left(\sum_{i=K+1}^{j-1}\frac{1}{j} + \sum_{i=j+1}^{\infty}\frac{1}{(i+j)(i-j)}\right)\\ &\leqq \frac{18}{\varepsilon}\left(1 + \frac{1}{2j}\sum_{i=j+1}^{\infty}\left(\frac{1}{i-j} - \frac{1}{i+j}\right)\right)\\ &= \frac{18}{\varepsilon}\left(1 + \frac{1}{2j}\sum_{h=1}^{2j}\frac{1}{h}\right)\\ &\leqq \frac{18}{\varepsilon}\left(1 + \frac{2j}{2j}\right) \leqq \frac{36}{\varepsilon},\end{aligned}$$

and so

$$\prod_{\substack{K<i<d-K\\ i\neq j}} \left|\frac{y_i - y_j}{x_i - x_j}\right| \leqq e^{36/\varepsilon}. \tag{31}$$

Thus, for $K < j \leqq d/2$ we obtain

$$\prod_{\substack{i=0\\ i\neq j}}^{d} |x_i - x_j| = D_x = D_y \times \frac{D_x}{D_y} \geqq 2d\left(\frac{M}{2}\right)^d 9^{-(K+1)}\, e^{-36/\varepsilon} \tag{32}$$

in view of (24), (27), (28), and (31).

We turn next to the numerator,

$$N_x = \prod_{\substack{i=0\\ i\neq j}}^{d} |x - x_i|.$$

N_x is bounded by

$$F_x = \prod_{\substack{i=1\\ i\neq j}}^{d} (M - x_i).$$

Next, we relate F_x to a product resembling the numerator in the pure Chebyshev case

$$F_y=\prod_{\substack{i=1\\i\neq j}}^{d}(M-y_i)=\frac{1}{M-y_j}\prod_{i=1}^{d}(M-y_i).$$

Since $j\leqq d/2$, we have

$$\sin\frac{\pi j}{2d}\geqq\frac{\sin\frac{\pi}{4}}{\frac{\pi}{4}}\frac{\pi j}{2d}=\frac{j\sqrt{2}}{d},\tag{33}$$

so that

$$F_y=\frac{1}{2M\left(\sin\frac{\pi j}{2d}\right)^2}\prod_{i=1}^{d}2M\sin^2\frac{\pi i}{2d}\leqq\frac{1}{2M\left(\frac{2j^2}{d^2}\right)}2(2d)\left(\frac{M}{2}\right)^d,\tag{34}$$

where the inequality follows from (33) and (5).

If $1\leqq i\leqq K$, then $M-x_i=i$ and $M-y_i=M(1-\cos((\pi i)/d))$. But

$$\frac{M\left(1-\cos\frac{\pi i}{d}\right)}{i^2}\geqq\frac{M\left(1-\cos\frac{\pi K}{d}\right)}{K^2}=\frac{\varepsilon d^2}{2K^2}\sin^2\frac{\pi K}{2d}$$

$$>\frac{\varepsilon d^2}{2K^2}\frac{K^2}{d^2}=\frac{\varepsilon}{2}>\frac{1}{4K}.$$

Therefore, $M-y_i>i^2/(4K)$ and

$$\prod_{i=1}^{K}\frac{M-x_i}{M-y_i}\leqq\prod_{i=1}^{K}\frac{4i}{i^2/K}=\frac{4^KK^K}{K!}<(4e)^K.\tag{35}$$

If $d-K\leqq i\leqq d$, it is easy to see that $x_i\geqq y_i$, and so

$$\prod_{i=d-K}^{d}\frac{M-x_i}{M-y_i}\leqq 1.\tag{36}$$

Lastly, suppose that $K<i<d-K$. Then

$$\frac{M-x_i}{M-y_i}=1+\frac{y_i-x_i}{M-y_i}\leqq 1+\frac{1}{M-y_i}=1+\frac{1}{2M\left(\sin\frac{\pi i}{2d}\right)^2},$$

so that

$$\log\frac{M-x_i}{M-y_i}\leqq\frac{1}{2M\left(\sin\frac{\pi i}{2d}\right)^2}\leqq\frac{1}{2M(i/d)^2}=\frac{2}{\varepsilon i^2}.$$

Thus

$$\sum_{\substack{i=K+1\\i\neq j}}^{d-K-1}\log\frac{M-x_i}{M-y_i}\leqq\sum_{i=K+1}^{d-K-1}\frac{2}{\varepsilon i^2}<\frac{2}{\varepsilon K}<\frac{\pi^2}{3}.\tag{37}$$

Hence, for $K < j \leqq d/2$ we obtain

$$\begin{aligned}\prod_{\substack{i=0\\ i\neq j}}^{d} |x-x_i| = N_x \leqq F_x = F_y \times \frac{F_x}{F_y} \\ \leqq \frac{2}{\varepsilon j^2}(2d)\left(\frac{M}{2}\right)^d (4e)^K e^{\pi^2/3},\end{aligned} \tag{38}$$

according to (34)-(37). Therefore, recalling (32), we get

$$|l_j(x)| \leqq \frac{18}{\varepsilon j^2} \exp\left(K + \frac{\pi^2}{3} + \frac{36}{\varepsilon}\right) 6^{2K}. \tag{39}$$

But

$$\sum_{j=K+1}^{d/2} \frac{1}{j^2} < \frac{1}{K},$$

so that

$$\sum_{j=K+1}^{d/2} |l_j(x)| \leqq \frac{18}{\varepsilon K} \exp\left(K + \frac{\pi^2}{3} + \frac{36}{\varepsilon}\right) 6^{2K} \leqq c_1\, e^{c_2/\varepsilon}, \tag{40}$$

where c_1, c_2 are positive constants independent of d, m, ε.

Finally, we note that if $d/2 \leqq j \leqq d$, then, by symmetry,

$$|l_j(x)| \leqq |l_{d-j}(x)|,$$

which gives

$$\sum_{j=0}^{d} |l_j(x)| \leqq 2\left(\sum_{j=0}^{K} |l_j(x)| + \sum_{j=K+1}^{d/2} |l_j(x)|\right) \leqq C_3\, e^{C_4/\varepsilon}, \tag{41}$$

where C_3, C_4 are positive constants independent of d, m, and ε. Inequality (41) was established under the assumptions that m was divisible by 4 and d was even. But if we drop these assumptions about d and m, then, since (41) holds for the greatest integer not exceeding m which is divisible by 4, and the least integer not less than d which is even, (41) is valid with appropriately chosen constants replacing C_3 and C_4, say c_3 and c_4.

We can now present our main result.

THEOREM. *If $\delta > 0$ and $n \geqq \delta\, d^2$ $(d > 0)$ there exist positive constants, a and b, independent of d, n, and δ, such that*

$$B(d, n) < a\, e^{b/\delta}. \tag{42}$$

Proof. If $m = \varepsilon d^2$, where $\varepsilon < 1/10$ and $d > 80$, then (41) gives

$$B(d, m) \leqq c_3\, e^{c_4/\varepsilon}. \tag{43}$$

Put $k = \lceil \delta d^2 \rceil$. If $\delta > 1/11$ and $d > 80$, we have

$$B(d, k) \leqq B\left(d, \left\lceil \frac{d^2}{11} \right\rceil\right) \leqq c_3\, e^{11 c_4} \tag{44}$$

in view of (43) and § 1 (iv). However, if $\delta \leqq 1/11$ and $d > 80$, we have

$$B(d, k) \leqq c_3\, e^{c_4/\delta} \leqq c_3\, e^{c_4/\delta}\, e^{11 c_4}.$$

Thus if $\delta > 0$,

$$B(d, k) \leqq c_5\, e^{c_4/\delta}$$

for $d > 80$ and c_5 a positive constant independent of d, k, and δ. Now when $1 \leqq d \leqq 80$, $B(d, d) \leqq 2^d \leqq 2^{80}$ implies that $B(d, k) \leqq 2^{80}$, since $k \geqq d$. Since $n \geqq k$, (42) now follows.

3. Lower bounds. We have exhibited an upper bound which is exponential in $d^2/m = 1/\varepsilon$. We next obtain a matching lower bound for $B(d, m)$ by constructing a polynomial of degree d whose absolute value is bounded by 1 on $m+1$ uniformly spaced points, and reaches heights exponential in d^2/m in the interior of the span of the points.

Assume that $\varepsilon < 1/10$ and $m > 8d$, so that $d > 80$. When we put $x = (m/2)(1 - \cos\theta)$,

$$Q(x) = C \prod_{j=0}^{d-1} \left(x - \frac{m}{2}\left(1 - \cos\frac{\pi(2j+1)}{2d}\right)\right) \tag{45}$$

and $\cos d\theta$ are polynomials of degree d in $\cos\theta$, having the same set of d zeros. We choose C so that

$$Q(x) = \cos d\theta, \qquad x = \frac{m}{2}(1 - \cos\theta). \tag{46}$$

Now put

$$P(x) = Q(x)\frac{K-1}{K}\prod_{j=0}^{K} \frac{x-j}{x - \frac{m}{2}\left(1 - \cos\frac{\pi(2j+1)}{2d}\right)}, \tag{47}$$

where $K = \lfloor \hat{K} \rfloor$ and $\hat{K}$ is the smallest solution of

$$\hat{K} = \frac{m}{2}\left(1 - \cos\frac{\pi(2\hat{K}+1)}{2d}\right)$$

satisfying $\hat{K} > 1$. It is easy to see that $\hat{K} < d/8$, and by methods similar to those used to establish (8) (i.e., choosing appropriate bounds for $\cos\theta$ from its power series expansion) we get the inequalities

$$1 - \frac{13\pi^2}{480} < \frac{\hat{K}}{\left(\frac{4}{\varepsilon\pi^2}\right)} < \frac{1}{1 - \frac{\pi^2}{768}},$$

so that

$$\frac{2}{\varepsilon\pi^2} < K < \frac{8}{\varepsilon\pi^2}.$$

Thus, for $1 \leqq j \leqq K$ we have

$$j \geqq \frac{m}{2}\left(1 - \cos\frac{\pi(2j+1)}{2d}\right),$$

and the factor $(K-1)/K$ in (47) has been chosen so that for $i > K$,

$$\frac{K-1}{K}(i-0) < i - \frac{m}{2}\left(1 - \cos\frac{\pi}{2d}\right).$$

Notice that $P(x)$ is of degree d and its zeros $(m/2)\,(1 - \cos(\pi(2j+1)/2d))$ are related linearly to the zeros of the Chebyshev polynomials, except near the left endpoint where they have been replaced by consecutive integers. (It is not necessary to alter those near the right endpoint.) Note the similarity to the upper bound construction.

Now $P(i)=0$, $i=0,1,\cdots,K$. For $i=K+1,\cdots,m$ we have

$$|P(i)|\leqq|Q(i)|\leqq 1.$$

since for each $j=1,\cdots,K$ the factor

$$\frac{i-j}{i-\dfrac{m}{2}\left(1-\cos\dfrac{\pi(2j+1)}{2d}\right)}$$

is bounded by 1, as is the term for $j=0$. Hence

$$|P(i)|\leqq 1,\qquad i=0,\cdots,m.$$

Let us evaluate Q and P at the point

$$\xi=\frac{m}{2}\left(1-\cos\frac{\pi}{d}\right),$$

corresponding to

$$\theta=\frac{\pi}{d}.$$

We have

$$\frac{\varepsilon\pi^2}{5}<\xi<\frac{\pi^2\varepsilon}{4}<\frac{1}{2}.$$

Then

$$Q(\xi)=\cos d\theta=\cos\pi=-1,$$

and

$$|P(\xi)|=|Q(\xi)|\frac{K-1}{K}\prod_{j=0}^{K}\frac{|\xi-j|}{\left|\xi-\dfrac{m}{2}\left(1-\cos\dfrac{\pi(2j+1)}{2d}\right)\right|}$$

$$=\frac{K-1}{K}\frac{\xi}{\xi-\dfrac{m}{2}\left(1-\cos\dfrac{\pi}{2d}\right)}\prod_{j=1}^{K}\frac{j-\xi}{\dfrac{m}{2}\left(1-\cos\dfrac{\pi(2j+1)}{2d}\right)-\xi}$$

$$\geqq\frac{K-1}{K}\prod_{j=1}^{K}\frac{j}{\dfrac{m}{2}\left(1-\cos\dfrac{\pi(2j+1)}{2d}\right)}\geqq\frac{K-1}{K}\prod_{j=1}^{K}\frac{j}{(2j+1)^2\dfrac{\pi^2\varepsilon}{16}}$$

$$=\frac{K-1}{K(2K+1)^2}\prod_{j=1}^{K}\frac{j}{(2j-1)^2\dfrac{\pi^2\varepsilon}{16}}$$

$$>\frac{1}{8K^2}\prod_{j=1}^{K}\frac{2Kj}{(2j-1)^2}>\frac{1}{8K^2}\frac{(2K)^K K!}{2^{2K}(K!)^2}=\frac{1}{8K^2}\frac{K^K}{K!}\frac{1}{2^K}$$

$$>\frac{a_1(e/2)^K}{K^{5/2}}>a_2\,e^{a_3/\varepsilon},$$

with a_2, a_3 positive and independent of d, m, and ε, and $0<a_3<1-\log 2$. Thus, if $\varepsilon<1/10$ and $\varepsilon d^2=m>8d$ we have

$$B(d, m)>a_2\, e^{a_3/\varepsilon}. \tag{48}$$

Furthermore: (i) If $m=\varepsilon d^2$, $d>80$, and $d \leqq m \leqq 8d$, then

$$B(d, m) \geqq B(d, 8d+1)>a_2\, e^{a_3/\varepsilon'}, \qquad \varepsilon'<1/10,$$

and since $\varepsilon'=\varepsilon(8d+1)/m \leqq \varepsilon(8+(1/d)) \leqq 9\varepsilon$ $(\varepsilon<1/10)$ we get

$$B(d, m)>a_2\, e^{a_3/9\varepsilon}. \tag{49}$$

(ii) If $m=\varepsilon d^2$, $d \leqq m \leqq 8d$, and $d \leqq 80$, then, since $m \leqq 8d$, there are a *finite* number of pairs d, m to consider. Let

$$a_4=\min_{d,m} \frac{B(d, m)}{e^{a_3/(m/d^2)}},$$

so that

$$B(d, m) \geqq a_4\, e^{a_3/\varepsilon}. \tag{50}$$

As a consequence of (48), (49), and (50), we have, for $\varepsilon<1/10$, $m=\varepsilon d^2$, and $d \geqq 1$,

$$B(d, m)>\alpha\, e^{\beta/\varepsilon}, \tag{51}$$

where α, β are positive and independent of ε and d.

Finally, if $\varepsilon \geqq 1/10$ put $\gamma=\min(\alpha, e^{-10\beta})$. Then we obtain, for $m=\varepsilon d^2$, $d \geqq 1$,

$$B(d, m) \geqq 1 \geqq e^{-10\beta}\, e^{\beta/\varepsilon} \geqq \gamma\, e^{\beta/\varepsilon}. \tag{52}$$

Inequalities (51) and (52) now give, for $m=\delta d^2$, $d \geqq 1$, $\delta>0$,

$$B(d, m) \geqq \gamma\, e^{\beta/\delta},$$

with γ, β positive and independent of δ and d, as we promised.

We may summarize our work as follows: $B(d, m)$ is bounded as $d \rightarrow \infty$ if and only if

$$\liminf_{d \rightarrow \infty} \frac{m}{d^2}=\delta>0. \tag{53}$$

Moreover, if (53) holds $B(d, m)<a\, e^{b/\delta}$, while for $m=\delta d^2$ we have $B(d, m)>\gamma\, e^{\beta/\delta}$, where a, b, γ, β are positive constants.

REFERENCES

[1] H. Ehlich and K. Zeller, *Schwankung von Polynomen zwischen Gitterpunkten*, Math. Z., 86 (1964), pp. 41–44.

[2] ———, *Numerische Abschätzung von Polynomen*, Z. Angew. Math. Mech., 45 (1965), pp. T20–T22.

[3] H. Ehlich, *Polynome zwischen Gitterpunkten*, Math. Z., 93 (1966), pp. 144–153.

[4] A. Schönhage, *Fehlerfortpflantzung bei Interpolation*, Numer. Math., 3 (1961), pp. 62–71.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 984-994, July 1992

COMMUTANT LIFTING AND SIMULTANEOUS H^∞ AND L^2 SUBOPTIMIZATION*

C. FOIAS† AND A. E. FRAZHO‡

Abstract. In this paper the commutant lifting theorem is used to obtain simultaneously a suboptimal solution to the two-sided Nehari optimization problem, with respect to the L^∞ norm and the L^2 norm. For the rational Nehari case a computational procedure in terms of a minimal realization is also given.

Key words. commutant lifting theorem, H^∞ optimization, L^2 optimization, Nehari problem, minimal realization

AMS(MOS) subject classifications. 47A20, 47A57, 93B36

1. Introduction. In studying subalgebras of C^*-algebras Kaftal, Larson, and Weiss [16] discovered that given any f in L^∞ and $\delta > 1$, there exists a function h in H^∞ satisfying

$$\|f+h\|_\infty \leqq \delta d_\infty \quad \text{and} \quad \|f+h\|_2 \leqq \frac{\delta d_2}{\sqrt{\delta^2-1}}, \tag{1.1}$$

where d_∞ is the distance from f to H^∞ in the L^∞ norm and d_2 is the distance from f to H^2 in the L^2 norm. In this paper, we will use the commutant lifting theorem to generalize their $H^\infty - L^2$ result to the two sided multidimensional Nehari setting. Both H^∞ and L^2 optimization have played an important role in control theory [5], [7], [11], [13], [18], [19], [20]. In fact, [7], [19], [20] develop some nice relationships between H^∞ and L^2 optimization problems arising in control theory. For this reason, we will also give explicit state space formulas to compute a solution h when f is rational, which may be useful in control theory.

To establish some notation, let $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ be the set of all strongly Lebesgue measurable functions, uniformly bounded almost everywhere on the unit circle, whose values are linear operators from $\mathscr{E}_1$ to $\mathscr{E}_2$. Throughout, we always assume that both $\mathscr{E}_1$ and $\mathscr{E}_2$ are finite-dimensional. If G is in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$, then the L^∞ norm of G is denoted by $\|G\|_\infty$, that is, $\|G\|_\infty$ is the essential supremum of $\|G(e^{it})\|$ for $0 \leqq t < 2\pi$. The L^2 norm of G is given by

$$\|G\|_2^2 = (G, G)_2 = \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(G(e^{it})^* G(e^{it}))\, dt = \sum_{-\infty}^{\infty} \operatorname{tr}(G_n^* G_n),$$

where tr denotes the trace and $G = \sum G_n e^{\text{int}}$ is the Fourier series expansion of G. Obviously, the set of all strongly Lebesgue measurable G satisfying $\|G\|_2 < \infty$ defines a Hilbert space. Recall that if M is any operator from $\mathscr{E}_1$ to $\mathscr{E}_2$, then

$$\|M\|_2^2 = \operatorname{tr}(M^*M) = \sum_1^n \|M\phi_i\|^2 = \operatorname{tr}(MM^*), \tag{1.2}$$

where $\{\phi_i\}_1^n$ is any orthonormal basis for $\mathscr{E}_1$. The trace of M^*M is independent of the choice of the orthonormal basis $\{\phi_i\}$. Finally, $H^\infty(\mathscr{E}_1, \mathscr{E}_2)$ is the subspace of $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ identified with the set of all uniformly bounded analytic functions in the open unit disc, whose values are operators from $\mathscr{E}_1$ to $\mathscr{E}_2$.

* Received by the editors July 1, 1991; accepted for publication (in revised form) November 8, 1991.

† Department of Mathematics, Indiana University, Bloomington, Indiana.

‡ School of Aeronautics and Astronautics, Purdue University, West Lafayette, Indiana 47907.

Throughout this paper, F is a specified function in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ and Θ_1 is a specified $*$-inner function in $H^\infty(\mathscr{E}_1, \mathscr{F}_1)$ and Θ_2 is a specified inner function in $H^\infty(\mathscr{F}_2, \mathscr{E}_2)$. Recall that a function Θ in $H^\infty(\mathscr{E}, \mathscr{F})$ is *inner*, respectively *$*$-inner*, if $\Theta(e^{it})$ is almost everywhere an isometry, respectively, a co-isometry. Now consider the following H^∞ and, respectively, L^2 optimization problems:

$$\begin{aligned} d_\infty &= d_\infty(F) = \inf\{\|F+\Theta_2 H\Theta_1\|_\infty : H \in H^\infty(\mathscr{F}_1, \mathscr{F}_2)\}, \\ d_2 &= d_2(F) = \inf\{\|F+\Theta_2 H\Theta_1\|_2 : H \in H^\infty(\mathscr{F}_1, \mathscr{F}_2)\}, \end{aligned} \tag{1.3}$$

where d_∞ is the distance from F to $\Theta_2 H^\infty(\mathscr{F}_1, \mathscr{F}_2)\Theta_1$ in the L^∞ norm, and d_2 is the corresponding distance in the L^2 norm. These optimization problems naturally arise in control theory [7], [11]. For the sake of the reader we shall present in § 2, the N. J. Young formula for d_∞ in [23]. In most problems the same H does not minimize both d_∞ and d_2. However, in this note we will use the commutant lifting theorem to establish the following relationship between d_∞ and d_2.

THEOREM 1.1. *Let $\delta > 1$. Then there exists a function H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying*

$$\|F+\Theta_2 H\Theta_1\|_\infty \leqq \delta d_\infty(F) \quad \text{and} \quad \|F+\Theta_2 H\Theta_1\|_2 \leqq \frac{\delta d_2(F)}{\sqrt{\delta^2-1}}. \tag{1.4}$$

In particular, if $\delta = \sqrt{2}$, then there exists a function H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying

$$\|F+\Theta_2 H\Theta_1\|_\infty \leqq \sqrt{2} d_\infty \quad \text{and} \quad \|F+\Theta_2 H\Theta_1\|_2 \leqq \sqrt{2} d_2. \tag{1.5}$$

In the scalar case with $\Theta_1 = \Theta_2 = 1$, the previous theorem reduces to the following elegant result of Kaftal, Larson, and Weiss [16], which motivated much of our work.

COROLLARY 1.2. *Let $\delta > 0$ and f be any scalar valued functions in L^∞. Then there exists a function h in H^∞ satisfying* (1.1).

To prove Theorem 1.1 we will need an explicit expression for d_∞ and d_2. Because the L^2 norm defines an inner product, we can easily compute d_2 by relaying on an argument used in Wiener filtering. To this end, let $[G]_c$ be the causal part of a function G in $L^\infty(\mathscr{F}_1, \mathscr{F}_2)$, that is, if $G = \sum_{-\infty}^{\infty} G_n e^{\text{int}}$ is the Fourier series expansion of G, then

$$[G]_c = \sum_{n=0}^{\infty} G_n e^{\text{int}}.$$

We claim that

$$d_2 = d_2(F) = \|F - \Theta_2[\Theta_2^* F \Theta_1^*]_c \Theta_1\|_2. \tag{1.6}$$

By the projection theorem, it is sufficient to show that $F - \Theta_2[\Theta_2^* F\Theta_1^*]_c\Theta_1$ is orthogonal to $\Theta_2 H^\infty(\mathscr{F}_1, \mathscr{F}_2)\Theta_1$ with respect to the L^2 inner product $(\cdot, \cdot)_2$. To verify (1.6), notice that for any H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ we have

$$\begin{aligned} (F - \Theta_2[\Theta_2^* F\Theta_1^*]_c\Theta_1, \Theta_2 H\Theta_1)_2 &= \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(\Theta_1^* H^*(\Theta_2^* F - [\Theta_2^* F\Theta_1^*]_c\Theta_1))\, dt \\ &= \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}((\Theta_2^* F - [\Theta_2^* F\Theta_1^*]_c\Theta_1)\Theta_1^* H^*)\, dt \\ &= \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}((\Theta_2^* F\Theta_1^* - [\Theta_2^* F\Theta_1^*]_c) H^*)\, dt = 0. \end{aligned}$$

The first and third equalities follow from the fact that Θ_2 is inner and Θ_1 is $*$-inner, respectively. The second equality follows from the fact that $\operatorname{tr}(BC) = \operatorname{tr}(CB)$, where

C maps $\mathscr{E}_1$ into $\mathscr{F}_2$ and B maps $\mathscr{F}_2$ into $\mathscr{E}_1$. The above analysis shows that $F-\Theta_2[\Theta_2^*F\Theta_1^*]_c\Theta_1$ is orthogonal to $\Theta_2H^\infty(\mathscr{F}_1, \mathscr{F}_2)\Theta_1$. Therefore, by the projection theorem, (1.6) holds.

2. The computation of d_∞. In this section, following Young [23], we will use the commutant lifting theorem to show that d_∞ equals the norm of a certain intertwining operator $\Lambda(F)$. As a byproduct of this computation, we will also show that the H^∞ optimization problem in (1.3) defines a game whose value is the norm of $\Lambda(F)$. Finally, it is noted that d_∞ can also be computed by converting the H^∞ optimization problem in (1.3) to a four block problem [9]. However, for our purposes this conversion is not necessary.

Throughout, we will follow the standard notation for Hilbert spaces in [10] and [22]. For example, $L^2(\mathscr{E})$ is the Hilbert space of all square integrable Lebesgue measurable functions in $[0, 2\pi)$ with values in $\mathscr{E}$. The Hardy space $H^2(\mathscr{E})$ is the subspace of $L^2(\mathscr{E})$ identified with the set of all analytic functions in the open unit disc with values in $\mathscr{E}$, whose Fourier series coefficients are square summable. Now let $\mathscr{H}_1$ and $\mathscr{H}_2$ be the subspaces defined by

$$\mathscr{H}_1 = L^2(\mathscr{E}_1)\ominus\Theta_1^*K^2(\mathscr{F}_1) \quad \text{and} \quad \mathscr{H}_2 = L^2(\mathscr{E}_2)\ominus\Theta_2H^2(\mathscr{F}_2),$$

where $K^2(\mathscr{F})$ is the orthogonal complement of $H^2(\mathscr{F})$ in $L^2(\mathscr{F})$. Let $\Lambda(F)$ be the operator from $\mathscr{H}_1$ into $\mathscr{H}_2$, with symbol F, defined by $\Lambda(F)f = P_2M_Ff$, where f is in $\mathscr{H}_1$ and P_2 is the orthogonal projection onto $\mathscr{H}_2$. Here M_F is the multiplication operator from $L^2(\mathscr{E}_1)$ to $L^2(\mathscr{E}_2)$ defined by $(M_Fg)(e^{it}) = Fg(e^{it})$, where g is in $L^2(\mathscr{E}_1)$. We need the following useful observation.

LEMMA 2.1. *Let W be a function in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$. Then $\Lambda(W) = 0$ if and only if W admits a factorization of the form $W = \Theta_2H\Theta_1$, where H is in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$. In this case W is in $H^\infty(\mathscr{E}_1, \mathscr{E}_2)$.*

Proof. Assume that $\Lambda(W) = 0$. Then $W\mathscr{H}_1 \subseteq \Theta_2H^2(\mathscr{F}_2)$. Since $H^2(\mathscr{E}_1) \subseteq \mathscr{H}_1$, this implies that $WH^2(\mathscr{E}_1) \subseteq \Theta_2H^2(\mathscr{F}_2)$. Therefore, $W = \Theta_2R$, where R is a function in $H^\infty(\mathscr{E}_1, \mathscr{F}_2)$; see Corollary IX.2.2 in [10]. Now we have

$$\Theta_2R(L^2(\mathscr{E}_1)\ominus\Theta_1^*K^2(\mathscr{F}_1)) = W\mathscr{H}_1 \subseteq \Theta_2H^2(\mathscr{F}_2).$$

Hence R maps $L^2(\mathscr{E}_1)\ominus\Theta_1^*K^2(\mathscr{F}_1)$ into $H^2(\mathscr{F}_2)$, or equivalently, its adjoint R^* maps $K^2(\mathscr{F}_2)$ into $\Theta_1^*K^2(\mathscr{F}_1)$, that is, $R^*K^2(\mathscr{F}_2) \subseteq \Theta_1^*K^2(\mathscr{F}_1)$. As before, this readily implies that $R^* = \Theta_1^*H^*$ for some H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$. Therefore, $W = \Theta_2H\Theta_1$. On the other hand, notice that $\Theta_1\mathscr{H}_1$ is orthogonal to $K^2(\mathscr{F}_1)$, or equivalently, $\Theta_1\mathscr{H}_1$ is contained in $H^2(\mathscr{F}_1)$. So if $W = \Theta_2H\Theta_1$, then using $\Theta_1\mathscr{H}_1 \subseteq H^2(\mathscr{F}_1)$ it follows that $\Lambda(W) = 0$. This completes the proof.

The following result is a classical application of the commutant lifting theorem (see [23]).

THEOREM 2.2. *Let F be a function in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$. Then*

$$d_\infty = \|\Lambda(F)\| = \inf\{\|F + \Theta_2H\Theta_1\|_\infty : H \in H^\infty(\mathscr{F}_1, \mathscr{F}_2)\}. \tag{2.1}$$

Moreover, there exists an optimal H_* in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying

$$\|F + \Theta_2H_*\Theta_1\|_\infty = \|\Lambda(F)\| = d_\infty.$$

Proof. Since $\Lambda(\Theta_2H\Theta_1) = 0$ for all H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$, we have

$$\|\Lambda(F)\| = \|\Lambda(F + \Theta_2H\Theta_1)\| = \|P_2(F + \Theta_2H\Theta_1)\,|\,\mathscr{H}_1\| \leqq \|F + \Theta_2H\Theta_1\|_\infty.$$

Therefore,

$$\|\Lambda(F)\| \leqq \inf\{\|F + \Theta_2H\Theta_1\|_\infty : H \in H^\infty(\mathscr{F}_1, \mathscr{F}_2)\}. \tag{2.2}$$

Now let V_1 and V_2 be the bilaterial shifts (multiplication by e^{it}) on $L^2(\mathscr{E}_1)$ and $L^2(\mathscr{E}_2)$, respectively. Clearly $\mathscr{H}_1$ is an invariant subspace for V_1, and $\mathscr{H}_2$ is an invariant subspace for V_2^*. Let T_1 on $\mathscr{H}_1$ be the isometry and T_2 on $\mathscr{H}_2$ be the co-isometry defined by $T_1 = V_1|\mathscr{H}_1$ and $T_2 = P_2V_2|\mathscr{H}_2$. Obviously, V_2 is an isometric lifting of T_2, that is, $P_2V_2 = T_2P_2$. Using $P_2V_2 = T_2P_2$ and $V_2M_F = M_FV_1$ it is easy to verify that $T_2\Lambda(F) = \Lambda(F)T_1$. By the commutant lifting theorem (see Theorem VII.1.2 in [10], [21], [22]), there exists an operator B mapping $L^2(\mathscr{E}_1)$ into $L^2(\mathscr{E}_2)$ satisfying $\|B\| = \|\Lambda(F)\|$ and $P_2B|\mathscr{H}_1 = \Lambda(F)$ and $V_2B = BV_1$. Since B intertwines the bilateral shifts, $B = M_G$, where G is a function in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ and $\|G\|_\infty = \|B\|$; see [10], [22]. So there exists a G in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ satisfying $\|G\|_\infty = \|\Lambda(F)\|$ and $\Lambda(F) = \Lambda(G)$. Notice that $\Lambda(G - F) = 0$. By Lemma 2.1, we see that $G - F = \Theta_2H_*\Theta_1$, where H_* is a function in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$. In other words, there exists a function H_* in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying

$$\|\Lambda(F)\| = \|F + \Theta_2H_*\Theta_1\|_\infty \geqq \inf\{\|F + \Theta_2H\Theta_1\|_\infty : H \in H^\infty(\mathscr{F}_1, \mathscr{F}_2)\} \geqq \|\Lambda(F)\|.$$

The last inequality follows from (2.2). This readily gives (2.1) and completes the proof.

Remark 2.3. Ball and Helton [3] showed that the Nevanlinna-Pick problem arising in control theory defines a game. Our previous analysis can also be used to show that the two-sided Nehari problem defines a game, that is,

$$(2.3) \qquad d_\infty = \|\Lambda(F)\| = \sup_{\|g\|=1} \inf_H \|(F + \Theta_2H\Theta_1)g\| = \inf_H \sup_{\|g\|=1} \|(F + \Theta_2H\Theta_1)g\|,$$

where the supremum is taken over all unit vectors g in $H^2(\mathscr{E}_1)$ and the infimum is taken over all H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$. Moreover, the norm of $\Lambda(F)$ is the value of the game. To prove (2.3), notice that $\|\Lambda(F)g\| = \|\Lambda(F + \Theta_2H\Theta_1)g\| \leqq \|(F + \Theta_2H\Theta_1)g\|$ gives

$$\begin{aligned}\|\Lambda(F)\| &= \sup_{\|g\|=1} \inf_H \|\Lambda(F)g\| \leqq \sup_{\|g\|=1} \inf_H \|(F + \Theta_2H\Theta_1)g\| \\ &\leqq \inf_H \sup_{\|g\|=1} \|(F + \Theta_2H\Theta_1)g\| \\ &= \inf\{\|F + \Theta_2H\Theta_1\|_\infty : H \in H^\infty(\mathscr{F}_1, \mathscr{F}_2)\} = \|\Lambda(F)\|.\end{aligned}$$

The last equality follows from the previous theorem, which is essentially a corollary of the commutant lifting theorem. Now (2.3) follows from the previous equation.

3. The central contractive intertwining lifting. In this section, we will use the central solution in the Schur representation for the commutant lifting theorem, to construct a contractive intertwining lifting that satisfies a special bound. In the next section we will use this purely geometric result to give a simple proof of Theorem 1.1.

To establish some notation, if $\mathscr{H}$ is a subspace of some Hilbert space $\mathscr{K}$, then $P_\mathscr{H}$ is the orthogonal projection onto the subspace $\mathscr{H}$. If C is a contraction ($\|C\| \leqq 1$), from $\mathscr{E}$ to $\mathscr{F}$, then D_C is the positive square root of $I - C^*C$ and $\mathscr{D}_C$ is the closed range of D_C. Throughout this paper A is a contraction from $\mathscr{H}$ into $\mathscr{H}'$ satisfying $T'A = AT$, where T is an isometry on $\mathscr{H}$, and T' is a contraction on $\mathscr{H}'$. We say that U' on $\mathscr{K}'$ is an *isometric lifting* of T' if U' is an isometry on $\mathscr{K}' \supseteq \mathscr{H}'$ and $P_{\mathscr{H}'}U' = T'P_{\mathscr{H}'}$. An operator B from $\mathscr{H}$ to $\mathscr{K}'$ is called a *contractive intertwining lifting* of A if B is a contraction satisfying $U'B = BT$ and $P_{\mathscr{H}'}B = A$. The commutant lifting theorem states that there always exists a contractive intertwining lifting B of A; see [10], [21], [22]. Furthermore, references [2], [10] provide a complete characterization of all contractive intertwining liftings B of A. This sets the stage for one of the main results of this paper.

THEOREM 3.1. *Let A be a strict contraction* ($\|A\| < 1$) *from $\mathscr{H}$ into $\mathscr{H}'$ satisfying $T'A = AT$, where T is an isometry and T' is a contraction. Let* $\mathscr{L} = \mathscr{H} \ominus T\mathscr{H} = \ker(T^*)$

be the wandering subspace determined by T. Then there exists a contractive intertwining lifting B_0 of A satisfying

$$\|B_0a\| \leqq \frac{\|Aa\|}{\sqrt{1-\|A\|^2}} \quad (\textit{for all } a \in \mathcal{L}). \tag{3.1}$$

In particular, if $\mathcal{L}$ is finite-dimensional, then this B_0 satisfies

$$\|B_0 | \mathcal{L}\|_2 \leqq \frac{\|A | \mathcal{L}\|_2}{\sqrt{1-\|A\|^2}}. \tag{3.2}$$

Proof. Since any isometric lifting U' of T' admits a reducing decomposition of the form $U' = U_1 \oplus U_m$, where U_m is the minimal isometric dilation of T' (see Remark VI.3.3 in [10]), we can assume without loss of generality that U' is the minimal isometric dilation of T'. Furthermore, because all minimal isometric dilations are unitarily equivalent, we can also assume that U' on $\mathcal{K}'$ is the Schäffer-Sz.-Nagy minimal isometric dilation of T', that is,

$$U'(h' \oplus f_0 \oplus f_1 \oplus f_2 \oplus \cdots) = T'h' \oplus D_{T'}h' \oplus f_0 \oplus f_1 \oplus f_2 \oplus \cdots,$$

where $h' \oplus (\bigoplus_0^\infty f_i)$ is in $\mathcal{K}' = \mathcal{H}' \oplus (\bigoplus_0^\infty \mathcal{D}_{T'})$; see § 3 of Chapter 6 in [10] or [22].

Now as in § 4 of Chapter XIV in [10], let ω be the unitary operator from $\mathcal{F} = \overline{D_A T \mathcal{H}}$ onto $\mathcal{F}' = \{D_{T'}Ah \oplus D_A h \colon h \in \mathcal{H}\}^-$ defined by

$$\omega D_A T h = D_{T'} A h \oplus D_A h \qquad (h \in \mathcal{H}).$$

Let P' be the orthogonal projection from $\mathcal{D}_{T'} \oplus \mathcal{D}_A$ onto $\mathcal{D}_{T'}$ and P_A the orthogonal projection from $\mathcal{D}_{T'} \oplus \mathcal{D}_A$ onto $\mathcal{D}_A$. According to equation (4.10) in Chapter XIV of [10], one contractive intertwining lifting B_0 from $\mathcal{H}$ to $\mathcal{K}'$ of A is given by

$$B_0 h = Ah \oplus \left(\bigoplus_{j=0}^{\infty} P' \omega P_{\mathcal{F}} (P_A \omega P_{\mathcal{F}})^j D_A h \right) \qquad (h \in \mathcal{H}), \tag{3.3}$$

where $P_{\mathcal{F}}$ is the orthogonal projection into $\mathcal{F}$. So using $P' = I - P_A$ and h in $\mathcal{H}$ we have

$$\begin{aligned}
\|B_0 h\|^2 &= \|Ah\|^2 + \sum_{j=0}^{\infty} \|P' \omega P_{\mathcal{F}} (P_A \omega P_{\mathcal{F}})^j D_A h\|^2 \\
&= \|Ah\|^2 + \sum_{j=0}^{\infty} (\|\omega P_{\mathcal{F}} (P_A \omega P_{\mathcal{F}})^j D_A h\|^2 - \|P_A \omega P_{\mathcal{F}} (P_A \omega P_{\mathcal{F}})^j D_A h\|^2) \\
&\leqq \|Ah\|^2 + \lim_{n \to \infty} \sum_{j=0}^{n} (\|(P_A \omega P_{\mathcal{F}})^j P_{\mathcal{F}} D_A h\|^2 - \|(P_A \omega P_{\mathcal{F}})^{j+1} D_A h\|^2) \\
&= \|Ah\|^2 + \|P_{\mathcal{F}} D_A h\|^2 - \lim_{n \to \infty} \|(P_A \omega P_{\mathcal{F}})^{n+1} D_A h\|^2 \leqq \|Ah\|^2 + \|P_{\mathcal{F}} D_A h\|^2.
\end{aligned}$$

Therefore, we obtain

$$\|B_0 h\|^2 \leqq \|Ah\|^2 + \|P_{\mathcal{F}} D_A h\|^2 \qquad (h \in \mathcal{H}). \tag{3.4}$$

(Using (3.3) and (3.4) it is easy to verify that B_0 is a contractive intertwining lifting of A.)

To complete the proof we need an explicit expression for the orthogonal projection $P_{\mathcal{F}}$. To this end, let X be the operator on $\mathcal{H}$ defined by $X = D_A T$. Since A is a strict contraction, D_A is invertible. So X is one to one. Moreover, the range of X is closed and equals $\mathcal{F}$. Therefore, X^*X is invertible. In fact, because T is an isometry,

$$X^*X = T^*(I - A^*A)T = (I - T^*A^*AT) = D_{AT}^2.$$

This, and the fact that AT is a strict contraction, clearly shows that X^*X is invertible. We claim that

$$P_{\mathcal{F}} = X(X^*X)^{-1}X^* = D_A T D_{AT}^{-2} T^* D_A. \tag{3.5}$$

To verify this notice that the operator P defined by $P = X(X^*X)^{-1}X^*$ is onto $\mathcal{F}$. Obviously, $P^2 = P$ and $P = P^*$. Therefore, $P = P_{\mathcal{F}}$ is the orthogonal projection onto $\mathcal{F}$.

Let a be in $\mathcal{L} = \ker(T^*)$. Substituting (3.5) into (3.4) and using $T^*a = 0$ we have

$$\begin{aligned} \|B_0 a\|^2 &\leqq \|Aa\|^2 + (P_{\mathcal{F}} D_A a, D_A a) = \|Aa\|^2 + (D_{AT}^{-2} T^* D_A^2 a, T^* D_A^2 a) \\ &= \|Aa\|^2 + (D_{AT}^{-2} T^* A^* A a, T^* A^* A a). \end{aligned} \tag{3.6}$$

Now notice that

$$D_{AT}^2 = I - T^*A^*AT = I - A^*T'^*T'A = I - A^*A + A^*(I - T'^*T')A \geqq D_A^2.$$

Therefore, $D_A^2 \leqq D_{AT}^2$, or equivalently, $D_{AT}^{-2} \leqq D_A^{-2}$. (Recall that if Q and R are two invertible positive operators satisfying $Q \leqq R$, then $R^{-1} \leqq Q^{-1}$.) So now (3.6) becomes

$$\begin{aligned} \|B_0 a\|^2 &\leqq \|Aa\|^2 + (D_A^{-2} T^*A^*Aa, T^*A^*Aa) \\ &\leqq \|Aa\|^2 + \frac{\|T^*A^*Aa\|^2}{1 - \|A\|^2} \leqq \|Aa\|^2 + \frac{\|A\|^2 \|Aa\|^2}{1 - \|A\|^2} = \frac{\|Aa\|^2}{1 - \|A\|^2}. \end{aligned}$$

The second inequality follows from the fact that $(1 - \|A\|^2)I \leqq D_A^2$. This proves (3.1). Finally, (3.2) follows from

$$\|B_0|\mathcal{L}\|_2^2 = \sum \|B_0 \phi_i\|^2 \leqq \sum \frac{\|A\phi_i\|^2}{1 - \|A\|^2} = \frac{\|A|\mathcal{L}\|_2^2}{1 - \|A\|^2},$$

where $\{\phi_i\}$ is an orthonormal basis for $\mathcal{L}$. This completes the proof.

Recall that U on $\mathcal{K}$ is a *unitary extension* of T on $\mathcal{H}$, if $\mathcal{H}$ is an invariant subspace for U and $T = U|\mathcal{H}$. This sets the stage for the following useful result.

COROLLARY 3.2. *Let A be a strict contraction from $\mathcal{H}$ to $\mathcal{H}'$ satisfying $T'A = AT$. Let U' on $\mathcal{K}'$ be a unitary lifting of T' and U on $\mathcal{K}$ a unitary extension of T. Then there exists a contraction Y mapping $\mathcal{K}$ into $\mathcal{K}'$ satisfying $U'Y = YU$ and $A = P_{\mathcal{H}'}Y|\mathcal{H}$, and*

$$\|Ya\| \leqq \frac{\|Aa\|}{\sqrt{1 - \|A\|^2}} \quad (\textit{for all } a \in \ker(T^*)). \tag{3.7}$$

In particular, if $\mathcal{L} = \ker T^*$ is finite-dimensional, then

$$\|Y|\mathcal{L}\|_2 \leqq \frac{\|A|\mathcal{L}\|_2}{\sqrt{1 - \|A\|^2}}. \tag{3.8}$$

Proof. By the previous theorem, there exists a contraction B_0 satisfying $U'B_0 = B_0T$ and $A = P_{\mathcal{H}'}B_0$ and (3.1). Obviously, $T^*B_0^* = B_0^*U'^*$ and U^* is an isometric lifting of T^*, that is, $P_{\mathcal{H}}U^* = T^*P_{\mathcal{H}}$. By the commutant lifting theorem, there exists a contraction Y from $\mathcal{K}$ to $\mathcal{K}'$ satisfying $U^*Y^* = Y^*U'^*$ and $P_{\mathcal{H}}Y^* = B_0^*$. The last equation implies that $B_0 = Y|\mathcal{H}$. Therefore, (3.7) follows from (3.1). Obviously, $A = P_{\mathcal{H}'}B_0 = P_{\mathcal{H}'}Y|\mathcal{H}$. This completes the proof.

4. Proof of Theorem 1.1. In this section we will use our previous analysis, based on the commutant lifting theorem, to prove Theorem 1.1. We begin with the following result, which turns out to be equivalent to Theorem 1.1.

THEOREM 4.1. *Let F be a function in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ and assume that $\Lambda(F)$ is a strict contraction. Then there exists a function H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying*

$$(4.1) \qquad \|F+\Theta_2 H\Theta_1\|_\infty \leqq 1 \quad \text{and} \quad \|F+\Theta_2 H\Theta_1\|_2 \leqq \frac{d_2(F)}{\sqrt{1-d_\infty^2(F)}}.$$

Proof. Let $A=\Lambda(F)$ and $T'=T_2$ and $T=T_1$ and $U'=V_2$ and $U=V_1$, where Λ; T_1, T_2, V_1, and V_2 are all defined in § 2. According to Corollary 3.2, there exists a contraction Y mapping $\mathscr{K}=L^2(\mathscr{E}_1)$ into $\mathscr{K}'=L^2(\mathscr{E}_2)$ satisfying $V_2Y=YV_1$ and $\Lambda(F)=P_2Y|\mathscr{H}_1$, and (3.7) holds. Because Y commutes with the bilateral shifts, there exists a function G in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ satisfying $Y=M_G$ and $\|Y\|=\|G\|_\infty$. Moreover, $\Lambda(F)=A=P_2Y|\mathscr{H}=\Lambda(G)$. So $\Lambda(G-F)=0$. By Lemma 2.1, there exists a function H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying $G=F+\Theta_2H\Theta_1$. Substituting this into (3.7), along with $\Theta_1^*\mathscr{F}_1=\ker T^*$, we have $\|Y\|=\|G\|_\infty=\|F+\Theta_2H\Theta_1\|_\infty \leqq 1$ and

$$\|(F+\Theta_2H\Theta_1)\Theta_1^*e\|^2=\|Y\Theta_1^*e\|^2 \leqq \frac{\|\Lambda(F)\Theta_1^*e\|^2}{1-\|A\|^2} \quad (\text{for all } e\in\mathscr{F}_1).$$

Using $\Lambda(F)\Theta_1^*e=(F\Theta_1^*-\Theta_2[\Theta_2^*F\Theta_1^*]_c)e$, along with the fact that $\|A\|=d_\infty$ and Θ_1^* is isometric, we have for $d_\infty=d_\infty(F)$ that

$$(4.2) \qquad \begin{aligned} \|G\Theta_1^*e\|^2 &= \|(F+\Theta_2H\Theta_1)\Theta_1^*e\|^2 \\ &\leqq \frac{\|(F-\Theta_2[\Theta_2^*F\Theta_1^*]_c\Theta_1)\Theta_1^*e\|^2}{1-d_\infty^2} \quad (\text{for all } e\in\mathscr{F}_1). \end{aligned}$$

Now let $G_2=F-\Theta_2[\Theta_2^*F\Theta_1^*]_c\Theta_1$ and recall that $d_2(F)=\|G_2\|_2$; see (1.6). Equation (4.2), along with the definition of the trace norm, implies that

$$(4.3) \qquad \|G\Theta_1^*\|_2^2 \leqq \frac{\|G_2\Theta_1^*\|_2^2}{1-d_\infty^2}.$$

Let $\Delta_1=I-\Theta_1^*\Theta_1$. Since Θ_1^* is isometric, $G\Delta_1=F\Delta_1=G_2\Delta_1$. In particular, this and $1-d_\infty^2 \leqq 1$ implies that

$$(4.4) \qquad \|G\Delta_1\|_2^2 \leqq \frac{\|G_2\Delta_1\|_2^2}{1-d_\infty^2}.$$

Combining this with (4.3) gives

$$(4.5) \qquad \|G\Theta_1^*\|_2^2+\|G\Delta_1\|_2^2 \leqq \frac{\|G_2\Theta_1^*\|_2^2+\|G_2\Delta_1\|_2^2}{1-d_\infty^2}.$$

Notice that because Θ_1^* is an isometry, Δ_1 is an orthogonal projection. This and the fact that the tr $(M^*M)=$ tr (MM^*) gives

$$\|G\Theta_1^*\|_2^2+\|G\Delta_1\|_2^2=\frac{1}{2\pi}\int_0^{2\pi} (\operatorname{tr}(G\Theta_1^*\Theta_1G^*)+\operatorname{tr}(G\Delta_1\Delta_1^*G^*))\,dt=\|G\|_2^2.$$

A similar calculation shows that $\|G_2\Theta_1^*\|_2^2+\|G_2\Delta_1\|_2^2=\|G_2\|_2^2$. Therefore, (4.5) becomes

$$\|G\|_2^2 \leqq \frac{\|G_2\|_2^2}{1-d_\infty^2}=\frac{d_2^2}{1-d_\infty^2}.$$

This completes the proof.

Proof of Theorem 1.1. Let F be any function in $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ and assume that $\delta > 1$. Let F_0 be the function defined by $F_0 = F/\delta\|\Lambda(F)\|$. Obviously, $\|\Lambda(F_0)\| = 1/\delta < 1$. By the previous theorem there exists a function H_0 in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying

$$\|F_0 + \Theta_2 H_0 \Theta_1\|_\infty \leqq 1 \quad \text{and} \quad \|F_0 + \Theta_2 H_0 \Theta_1\|_2 \leqq \frac{d_2(F_0)}{\sqrt{1 - d_\infty^2(F_0)}}.$$

Using the relations $F = \delta d_\infty(F) F_0$, $d_2(F) = \delta d_\infty(F) d_2(F_0)$, $d_\infty(F_0) = 1/\delta$, and setting $H = \delta d_\infty(F) H_0$ we have

$$\|F + \Theta_2 H \Theta_1\|_\infty \leqq \delta d_\infty(F) \quad \text{and} \quad \|F + \Theta_2 H \Theta_1\|_2 \leqq \frac{\delta d_2(F)}{\sqrt{\delta^2 - 1}}.$$

This is precisely (1.4) in Theorem 1.1. The proof is now complete.

Finally, we emphasize that Theorem 1.1 is equivalent to Theorem 4.1. The previous proof shows that Theorem 1.1 is a consequence of Theorem 4.1. So to show that they are equivalent it is sufficient to show that Theorem 4.1 follows from Theorem 1.1. To this end, assume that Theorem 1.1 is true and that $\Lambda = \Lambda(F)$ is a strict contraction. Now let $\delta = 1/\|\Lambda\|$. Then, by Theorem 1.1 and $d_\infty = \|\Lambda\|$, there exists a function H in $H^\infty(\mathscr{F}_1, \mathscr{F}_2)$ satisfying

$$\|F + \Theta_2 H \Theta_1\|_\infty \leqq \delta d_\infty = 1 \quad \text{and} \quad \|F + \Theta_2 H \Theta_1\|_2 \leqq \frac{\delta d_2}{\sqrt{\delta^2 - 1}} = \frac{d_2}{\sqrt{1 - \|\Lambda\|^2}} = \frac{d_2}{\sqrt{1 - d_\infty^2}}.$$

This readily proves Theorem 4.1. Hence Theorems 1.1 and 4.1 are equivalent.

Obviously, as $\delta \to \infty$ the quantity $\delta/\sqrt{\delta^2 - 1} \to 1$. Therefore, our previous analysis shows that as $\delta \to \infty$ our solution $F + \Theta_2 H \Theta_1$ to (1.4), provided by the central intertwining lifting (3.3) in the Schur representation for the commutant lifting theorem, approaches the L^2 optimal solution $F - \Theta_2[\Theta_2^* F \Theta_1^*]_c \Theta_1$. This sort of phenomenon has also been observed in the engineering literature, where they have noticed that certain H^∞ controllers approach the optimal L^2 controller as a certain parameter approaches infinity [7], [18], [19].

5. A state space solution. State space techniques have played an important role in computing all solutions to the rational Nehari interpolation problem; see [4], [5], [10]-[14]. In this section, we will use state space techniques along with some of the results in [10], [12], to compute our solution to the Nehari version of (1.4) when F is rational, that is, given a rational F and a $\delta > 1$, we will give a state space procedure to construct a function H in $H^\infty(\mathscr{E}_1, \mathscr{E}_2)$ to satisfy

$$\|F + H\|_\infty \leqq \delta d_\infty \quad \text{and} \quad \|F + H\|_2 \leqq \frac{\delta d_2}{\sqrt{\delta^2 - 1}}. \tag{5.1}$$

Here we set $\Theta_2 = I$ and $\Theta_1 = I$ and $\mathscr{E}_1 = \mathscr{F}_1$ and $\mathscr{E}_2 = \mathscr{F}_2$. A computational formula for the general two-sided Nehari problem in (1.4) is somewhat more involved, and will be given elsewhere. In order to compute our H in (5.1), we need to compute the central solution (3.3) in the Schur representation for the commutant lifting theorem considered in the monograph [10], Chapters XIII, XIV, and in the paper [12]. Finally, it is noted that our central solution will correspond to a certain maximal entropy extension. For some nice results on maximal entropy extensions see [6], [8], and [18].

To begin, let $K^\infty(\mathscr{E}_1, \mathscr{E}_2)$ be the subspace of $L^\infty(\mathscr{E}_1, \mathscr{E}_2)$ whose Fourier series expansion contain only negative powers ($n < 0$) of e^{int}. Now let F be a rational function in $K^\infty(\mathscr{E}_1, \mathscr{E}_2)$, that is, $zF(z)$ admits a representation of the form $zF(z) = N(z)/q(z)$, where $N(z)$ is an operator valued polynomial and $q(z)$ is scalar valued polynomial with all of its zeros inside the unit circle. Let $\{A_0, B, C\}$ be a minimal realization of F; that is, A_0 is a matrix on C^n, and B maps $\mathscr{E}_1$ into C^n and C maps C^n into $\mathscr{E}_2$ and $\{A_0, B\}$ is controllable and $\{C, A_0\}$ is observable, and

$$(5.2) \qquad F(z) = C(zI - A_0)^{-1}B,$$

where $z = e^{it}$; see [5], [10], [17] for further details on realization theory. Since F is in $K^\infty(\mathscr{E}_1, \mathscr{E}_2)$, all the eigenvalues of A_0 are in the open unit disc. Notice that since $\Theta_2 = I$ and $\Theta_1 = I$, the operator $\Lambda(F)$ is now the Hankel operator mapping $H^2(\mathscr{E}_1)$ into $K^2(\mathscr{E}_2)$ defined by $\Lambda(F) = P_2 M_F | H^2(\mathscr{E}_1)$, where P_2 is now the orthogonal projection onto $K^2(\mathscr{E}_2)$. It is well known (see [5], [10], [11], [13], [14]) that this Hankel operator Λ admits a decomposition of the form $\Lambda = W_0 W_c^*$, where W_0 is the operator from C^n into $K^2(\mathscr{E}_2)$ defined by $W_0 = C(zI - A_0)^{-1}$ and W_c is the operator from C^n to $H^2(\mathscr{E}_1)$ defined by $W_c = B^*(I - zA_0^*)^{-1}$. The *controllability grammian* P and *observability grammian* Q are defined by

$$(5.3) \qquad P \triangleq W_c^* W_c = \sum_0^\infty A_0^i BB^* A_0^{*i} \quad \text{and} \quad Q \triangleq W_0^* W_0 = \sum_0^\infty A_0^{*i} C^* C A_0^i.$$

Because $\{A_0, B, C\}$ is a minimal realization, both P and Q are strictly positive matrices. Recall that P and Q can be computed by solving the following Lyapunov equations:

$$(5.4) \qquad P = A_0 P A_0^* + BB^* \quad \text{and} \quad Q = A_0^* Q A_0 + C^* C.$$

The minimal realization $\{A_0, B, C\}$, along with its controllability and observability grammians P and Q, will play a key role in our state space solution to (5.1). Finally, recall that $\|\Lambda\|^2 = d_\infty^2(F)$ equals the largest eigenvalue λ^2 of QP; see [5], [10], [11], [13], [14].

For the moment, assume that Λ is a strict contraction, or equivalently, $\lambda = d_\infty = \|\Lambda\| < 1$. According to (9.9) in Chapter XIII of [10], the set of all $F + H$ satisfying $\|F + H\|_\infty \leqq 1$, where H is in $H^\infty(\mathscr{E}_1, \mathscr{E}_2)$, is given by the following transmission type formula:

$$(5.5) \qquad (\Phi_{11} + z\Phi_{12}F_1)(\Phi_{21} + z\Phi_{22}F_1)^{-1},$$

where F_1 is an arbitrary function in the closed unit ball of $H^\infty(\mathscr{E}_1, \mathscr{E}_2)$ and $\Phi_{11}, \Phi_{12}, \Phi_{21}$, and Φ_{22} are specified functions. By consulting § 8 in Chapter XIII of [10], the central solution B_0 in (3.3) is given by (5.5), where $F_1 = 0$, that is, $B_0 = \Phi_{11}\Phi_{21}^{-1}$. Equations (1) and (8) in [12] (where our W_0 is precisely their $\hat{W}_0$) show that

$$(5.6) \qquad \Phi_{11} = W_0(I - PQ)^{-1}B \quad \text{and} \quad \Phi_{21} = W_c(I - QP)^{-1}QB + I.$$

Therefore, $F + H = \Phi_{11}\Phi_{21}^{-1}$ is precisely the Schur representation for the central solution (3.3) in the commutant lifting theorem. So, according to Theorem 4.1, along with the fact that $\|\Lambda\| < 1$, we have for $F + H = \Phi_{11}\Phi_{21}^{-1}$ that

$$(5.7) \qquad \|F + H\|_\infty \leqq 1 \quad \text{and} \quad \|F + H\|_2 \leqq \frac{d_2(F)}{\sqrt{1 - d_\infty^2(F)}}.$$

Finally, it is noted that because F is in $K^\infty(\mathscr{E}_1, \mathscr{E}_2)$ the distance $d_2(F) = \|F\|_2$.

Now let $\{A_0, B, C\}$ be a minimal realization of F, and P its controllability grammian and Q its observability grammian. Let $d_\infty^2 = \lambda^2$ be the largest eigenvalue of QP. Then the function $F_0 = F/\lambda\delta$ defines a Hankel operator $\Lambda(F_0)$ satisfying $\|\Lambda(F_0)\| = 1/\delta$. Obviously, $\{A_0, B, (1/\lambda\delta)C\}$ is a minimal realization of F_0 whose observability grammian is $Q/\lambda^2\delta^2$ and controllability grammian is P. By consulting the proof of Theorem 1.1, we see that if H_0 in $H^\infty(\mathscr{C}_1, \mathscr{C}_2)$ satisfies

$$(5.8) \qquad \|F_0+H_0\|_\infty \leqq 1 \quad \text{and} \quad \|F_0+H_0\|_2 \leqq \frac{d_2(F_0)}{\sqrt{1-d_\infty^2(F_0)}},$$

then $F+H=\lambda\delta(F_0+H_0)$ is a solution to (5.1). So by replacing Q by $Q/\lambda^2\delta^2$ and C by $(1/\lambda\delta)C$ in (5.6) and setting $\Phi_1=\lambda\delta\Phi_{11}$ and $\Phi_2=\Phi_{21}$ the previous analysis readily leads to the following method for computing a function H satisfying (5.1).

Procedure 5.1. Let F be a rational function in $K^\infty(\mathscr{C}_1, \mathscr{C}_2)$ and $\delta>1$. Compute a minimal realization $\{A_0, B, C\}$ of $F(z)$ and its corresponding controllability P and observability Q grammians. Compute $d_\infty^2=\lambda^2$, the largest eigenvalue of QP. Compute

$$(5.9) \qquad \Phi_1=\lambda^2\delta^2 W_0(\lambda^2\delta^2 I-PQ)^{-1}B \quad \text{and} \quad \Phi_2=W_c(\lambda^2\delta^2 I-QP)^{-1}QB+I.$$

Then $F+H=\Phi_1\Phi_2^{-1}$ satisfies (5.1).

The results in § 9 of Chapter XIV in [10] can also be used to give another formula to compute our function H in $H^\infty(\mathscr{C}_1, \mathscr{C}_2)$ satisfying (5.1). To see this, for the moment assume that $\|\Lambda(F)\|=d_\infty<1$. Then according to equation (9.8) in Chapter XIV of [10], the set of all $F+H$ satisfying $\|F+H\|_\infty \leqq 1$ is given by the following scattering type formula:

$$(5.10) \qquad F+H=F+\Psi_{22}+\Psi_{21}(I-F_1\Psi_{11})^{-1}F_1\Psi_{12},$$

where F_1 is an arbitrary function in the closed unit ball of $H^\infty(\mathscr{C}_1, \mathscr{C}_2)$ and $\Psi_{11}, \Psi_{12}, \Psi_{21}$, and Ψ_{22} are all specified functions. By the results in Chapter XIV of [10], the central solution B_0 in (3.3) is given by $F_1=0$, that is, $B_0=F+\Psi_{22}$. Reference [12] shows that

$$(5.11) \qquad \Psi_{22}(z)=-CP(I-zA_1)^{-1}(I-A_0^*QA_0P)^{-1}A_0^*QB,$$

where A_1 is the matrix on C^n defined by

$$(5.12) \qquad A_1=(I-A_0^*QA_0P)^{-1}A_0^*(I-QP).$$

Reference [10] also proves that all the eigenvalues of A_1 are in the open unit circle. Therefore, the central solution B_0 in (3.3) is given by $B_0=F+\Psi_{22}$. Now, proceeding exactly as above, we can use formulas (5.11) and (5.12) to obtain the following method for computing our function H satisfying (5.1).

Procedure 5.2. Let F be a rational function in $K^\infty(\mathscr{C}_1, \mathscr{C}_2)$ and $\delta>1$. Compute a minimal realization $\{A_0, B, C\}$ of $F(z)$ and its corresponding controllability P and observability grammian Q. Compute $d_\infty^2=\lambda^2$, the largest eigenvalue of QP. Compute

$$A_2=(\lambda^2\delta^2 I-A_0^*QA_0P)^{-1}A_0^*(\lambda^2\delta^2 I-QP)$$

and

$$(5.13) \qquad \Psi=-CP(I-zA_2)^{-1}(\lambda^2\delta^2 I-A_0^*QA_0P)^{-1}A_0^*QB.$$

Then $F+\Psi$ satisfies (5.1).

Finally, it is noted that if we let $\delta\to\infty$ in (5.9) or (5.13), then, as expected, our $F+H\to F$ the L^2 optimal solution.

Acknowledgment. The authors thank Allen Tannenbaum for some clarification included in §§ 1 and 2.

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Infinite Hankel block matrices and related extension problem*, Izv. Akad. Nauk Armyan SSR, 6 (1971), pp. 87-112. In Russian; Amer. Math. Soc. Transl., 111 (1978), pp. 133-156.

[2] Gr. Arsene, Z. Ceausescu, and C. Foias, *On intertwining dilations* VIII, J. Operator Theory, 4 (1980), pp. 55-91.

[3] J. A. Ball and J. W. Helton, *H^∞ control for nonlinear plants: connections with differential games*, Proc. 28th IEEE Conference on Decision and Control, Tampa, Florida, Dec. 13-15, 1989, pp. 956-962.

[4] J. A. Ball and A. C. M. Ran, *Optimal Hankel norm approximations and Wiener-Hopf Factorizations* I: *The canonical case*, SIAM J. Control Optim., 25 (1987), pp. 362-382.

[5] J. A. Ball, I. Gohberg, and L. Rodman, *Interpolation for Rational Matrix Functions*, Birkhäuser Verlag, Basel, 1990.

[6] T. Constantinescu, *A maximum entropy principle for contractive intertwining dilations*, in Operators in Indefinite Metric Spaces, Scattering Theory and Other Topics; *Operator Theory: Advances and Applications*, 10th International Conference on Operator Theory, Bucharest, Romania, 24 (1985), pp. 69-85.

[7] J. C. Doyle, K. Glover, P. Khargonekar, and B. A. Frances, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. on Automat. Control, 34 (1989), pp. 831-847.

[8] H. Dym and I. Gohberg, *A maximum entropy principle for contractive interpolants*, J. Funct. Anal., 65 (1986), pp. 83-125.

[9] A. Feintuck and B. Frances, *Distance formulas for operator algebras arising in optimal control problems*, in Topics in Operator Theory and Interpolation; Operator Theory Advances and Applications, I. Gohberg, ed., 29 (1989), pp. 151-170.

[10] C. Foias and A. E. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Birkhäuser Verlag, Basel, 1990.

[11] B. A. Francis, *A Course in H_∞ Control Theory*, Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, New York, 1987.

[12] A. E. Frazho and D. L. Augenstein, *A remark on two solutions to the rational Nehari interpolation problem*, Integral Equations Operators Theory, 14 (1991), pp. 299-303.

[13] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. of Control, 39 (1984), pp. 1115-1193.

[14] I. Gohberg, M. A. Kaashoek, and F. van Schagen, *Rational contractive and unitary interpolants in realized form*, Integral Equations Operator Theory, 11 (1988), pp. 105-127.

[15] I. Gohberg, M. A. Kaashoek, and H. J. Woerdeman, *The band method for positive and strictly contractive extension problems: An alternative version and new applications*, Integral Equations Operator Theory, 12 (1989), pp. 343-382.

[16] V. Kaftal, D. Larson, and G. Weiss, *Quasitriangular subalgebras of semifinite von Neumann algebras are closed*, J. Funct. Anal., to appear.

[17] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[18] D. Mustafa and K. Glover, *Minimum Entropy H_∞ Control*, Lecture Notes in Control and Inform. Sci. 146, Springer-Verlag, New York, 1990.

[19] K. M. Nagpal and P. P. Khargonekar, *Filtering and Smoothing in an H^∞ setting*, IEEE Trans. on Automat. Control, 36 (1991), pp. 152-166.

[20] M. A. Rotea and P. P. Khargonekar, *H^2-optimal control with an H^∞-constraint: The state feedback case*, Automatica, 27 (1991), pp. 307-316.

[21] B. Sz.-Nagy and C. Foias, *Dilatation des commutants d'opérateurs*, C. R. Acad. Sci. Paris, Sér. A, 266 (1968), pp. 493-495.

[22] ———, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[23] N. J. Young, *An algorithm for the super-optimal sensitivity-minimising controller*, Proc. Workshop on New Perspectives in Industrial Control System Design Using H_∞ Methods, Oxford University Press, London, 1986.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 995-1003, July 1992

NONUNIFORM SAMPLING OF BANDLIMITED FUNCTIONS OF POLYNOMIAL GROWTH*

GILBERT G. WALTER†

Abstract. A version of the Shannon sampling theorem for nonuniform sampling points and signals with infinite energy is given. It is valid for signals of polynomial growth whose Fourier transform is a generalized function with compact support. It involves the theory of nonharmonic Fourier series.

Key words. sampling theorem, band limited signals, generalized functions, nonharmonic Fourier series

AMS(MOS) subject classifications. 41A05, 40G05

1. Introduction. The Shannon sampling theorem for bandlimited signals is the name given to the formula

$$f(t)=\sum_{n=-\infty}^{\infty} f(t_n)S_n(t), \qquad t\in\mathbb{R}^1, \tag{1.1}$$

where the t_n are the sampling points and the $S_n(t)$ the sampling functions. The latter have the properties that $S_n(t_k)=\delta_{nk}$, $k=0, \pm1, \cdots$. The t_n may be uniform, $t_n=nT$, and were in Shannon's original work [8]. In this case the $S_n(t)$ are given by

$$S_n(t)=\frac{\sin \sigma(t-nT)}{\sigma(t-nT)}, \qquad \sigma=\pi/T, \tag{1.2}$$

and the series (1.1) converges to $f(t)$ for $f\in L^2(\mathbb{R}^1)$ provided its Fourier transform $F(\omega)$ has support in $[-\sigma, \sigma]$.

The t_n may also be nonuniform, i.e., not equally spaced. In this case the $S_n(t)$ have a more complicated form but (1.1) still converges under the same hypotheses, provided the t_n are "close" to nT [1].

The various versions of this theorem have widespread applications. They are used whenever it is necessary to reconstitute an analog signal from sampled values or from a digital signal. This occurs, e.g., in compact disc recordings, in multiplexed signals, in optics, and in tomography. Several surveys of the literature of both the theory and applications have been given [1], [4].

Most of the results refer to finite energy (i.e., L^2) signals. Yet many signals of practical importance are not finite energy. For example, a pure musical note has infinite energy. However, it can also be recovered from its sampled values. A number of results concerning such signals have been obtained. The earliest were due to Campbell [2] and Pfaffelhuber [7]. These results were extended by several authors [3], [6], [9], [10], [5]; however, in no case did they consider nonuniform sampling, which is important for some of the applications.

In this work we shall extend the results in [9] and [10] to this case of nonuniform sampling. Our signals will be bandlimited and of at most polynomial growth on the real axis. Usually, we shall have to add another condition to obtain the needed uniqueness. In § 2, we introduce some notation and preliminary results. In § 3 we obtain a sampling theorem involving additional factors in the series. In § 4, the form

* Received by the editors August 28, 1989; accepted for publication (in revised form) September 18, 1991. This research was done while visiting the University of South Florida, Tampa, Florida 33620.

† Department of Mathematical Sciences, University of Wisconsin at Milwaukee, Box 413, Milwaukee, Wisconsin 53201.

of (1.1) is retained, but a type of summability is used rather than ordinary convergence. In § 5 the series is treated from the point of view of convergence in a class of generalized functions.

2. Background and preliminaries. An appropriate space in which to study bandlimited functions of polynomial growth (BLPG) is the space $B_r(\sigma)$, $r \geqq 0$. This space consists of those functions in $L^2((1+t^2)^{-r}) = L_r^2$ whose Fourier transforms have support on $[-\sigma, \sigma]$. Each of our BLPG functions must belong to some $B_r(\sigma)$.

In [9] it was shown that a uniform sampling theorem for $f \in B_{2N}(\sigma)$ could be given. It had the form

$$f(t) = P(t) \sin \sigma t + \sum_{n=-\infty}^{\infty} f(nT) \left[\frac{t^2+1}{(nT)^2+1}\right]^N \frac{\sin \sigma(t-nT)}{\sigma(t-nT)}, \tag{2.1}$$

where $P(t)$ is an appropriate polynomial of degree $2N-1$. This added term is necessary because of the lack of uniqueness of sampling sequences in $B_{2N}(\sigma)$. Two functions which differ by $P(t) \sin \sigma t$ will have the same sampling sequence.

2.1. Equivalence classes. In order to obtain uniqueness (and to eliminate the term $P(t) \sin \sigma t$ in (2.1)) we shall use equivalence classes of functions in $B_{2N}(\sigma)$. We denote these by $[f]$;

$$[f] := \{g \in B_{2N}(\sigma) \mid g(nT) = f(nT), n = 0, \pm 1, \cdots\},$$

with a norm defined as usual by

$$\|[f]\| = \inf_{g \in [f]} \|g\|_{-2N},$$

where $\| \quad \|_{-2N}$ denotes the norm in L_{2N}^2. The set $B'_{2N}(\sigma)$ of all such $[f]$ is a Hilbert space with this norm. Its sampling expansion has partial sums given by

$$\begin{aligned} f_m(t) &= \sum_{n=-m}^{m} g(nT) \left[\frac{t^2+1}{n^2T^2+1}\right]^N \frac{\sin \sigma(t-nT)}{\sigma(t-nT)} \\ &= \sum_{n=-m}^{m} g(nT) S_n^{\sigma,N}(t), \end{aligned} \tag{2.2}$$

where $g \in [f]$. It was shown in [9] that the partial sums of (2.1) converge in the sense of L_{2N}^2. Hence, it follows that

$$[f_m] \to [f]$$

in the norm of $B'_{2N}(\sigma)$, since

$$\begin{aligned} \|[f]-[f_m]\| &= \|[f-f_m]\| \\ &= \inf_{g \in [f-f_m]} \|g\|_{-2N} \\ &\leqq \|f(t) - P(t) \sin \sigma t - f_m(t)\|_{-2N}. \end{aligned}$$

There is exactly one element of each $[f]$ such that $f(t)/(t^2+1)^N$ is an entire function, i.e., has no singularity at $t = \pm i$. By restricting ourselves to such f we obtain an isometric copy $B_{2N}^0(\sigma)$ of $B'_{2N}(\sigma)$, which is a subspace of $B_{2N}(\sigma)$.

2.2. The space $B_{2N}^0(\sigma)$. The sampling theorem in [9] may be interpreted in $B_{2N}^0(\sigma)$ as the following.

PROPOSITION 2.1. *The sampling sequence* $\{S_n^{\sigma,N}(t)\}$ *in* (2.2) *is a complete orthogonal system in* $B_{2N}^0(\sigma)$. *Parseval's equality may be expressed as*

$$\|f\|_{-2N}^2 = \sum_{m=-\infty}^{\infty} |f(nT)|^2(n^2T^2+1)^{-2N}. \tag{2.3}$$

This follows from the orthogonality of $S_n^{\sigma,0}(t) = \sin\sigma(t-nT)/\sigma(t-nT)$ in $L^2(\mathbb{R}^1)$ and its completeness in $B_0(\sigma)$. Indeed, we have

$$\begin{aligned}\|f\|_{-2N}^2 &= \int_{-\infty}^{\infty} |f(t)|^2(t^2+1)^{-2N}\,dt\\ &= \left\|\frac{f(t)}{(t^2+1)^N}\right\|_0^2 = \sum_{n=-\infty}^{\infty}\left|\frac{f(nT)}{(n^2T^2+1)^N}\right|^2\\ &= \sum_{n=-\infty}^{\infty} |\langle f, S_n^{\sigma,N}\rangle_{-2N}|^2(n^2T^2+1)^{-2N}.\end{aligned}$$

As is $B_0(\sigma)$, the space $B_{2N}^0(\sigma)$ is a reproducing kernel Hilbert space with reproducing kernel

$$k(t,u) = (t^2+1)^N(u^2+1)^N\frac{\sin\sigma(t-u)}{\sigma(t-u)}, \tag{2.4}$$

which follows from the same sorts of considerations. This kernel may be used to find the projection of $f \in B_{2N}(\sigma)$ onto $B_{2N}^0(\sigma)$,

$$(\mathcal{P}f)(t) = \int_{-\infty}^{\infty} k(t,u)f(u)(u^2+1)^{-2N}\,du. \tag{2.5}$$

Because of the orthogonality of the $S_n^{\sigma,N}$ and the fact that

$$\langle f, S_n^{\sigma,N}\rangle_{-2N} = f(nT)(n^2T^2+1)^{-2N},$$

$(\mathcal{P}f)(t)$ is the same as $f(t)-P(t)\sin\sigma t$ in (2.1). □

The space $B_{2N}^0(\sigma)$ will be the setting for our nonuniform sampling.

2.3. Frames. If the sampling points t_n are no longer uniform, then the sequence $\{S_n^{\sigma,N}\}$ is no longer orthogonal. However, if it is close to orthogonal, it may satisfy an inequality (which is a substitute for Parseval's equality) of the form

$$0 < A\|f\|^2 \leqq \Sigma|\langle f, S_n\rangle|^2 \leqq B\|f\|^2. \tag{2.6}$$

Here f belongs to some Hilbert space H, and A and B are constant. Such sequences $\{S_n\}$ are called "frames" and share many of the properties of orthogonal sequences. In particular [11, p. 186] each $f \in H$ has an associated moment sequence $a_n = \langle g, S_n\rangle$ for some $g \in H$ such that

$$f = \sum_n a_n S_n. \tag{2.7}$$

For $B_0(\sigma)$, it is well known that

$$S_n(t) = \frac{\sin\sigma(t-t_n)}{\sigma(t-t_n)}, \qquad n = 0, \pm1, \pm2, \cdots, t_{-n} = -t_n,\ t, t_n \in \mathbb{R}, \tag{2.8}$$

is a frame provided that $|t_n - n| \leqq L < \frac{1}{4}$ for $\sigma = \pi$. (Kadec's $\frac{1}{4}$ Theorem, [11]). In fact $\{S_n\}$ is an "exact" frame, i.e., no proper subset of $\{S_n\}$ is itself a frame. In this case there is a biorthogonal sequence $\{g_n\}$ with $\{S_n\}$ such that

$$f = \Sigma\langle f, g_n\rangle S_n = \Sigma\langle f, S_n\rangle g_n \tag{2.9}$$

[11, p. 188]. In the case of $B_0(\sigma)$, we can say a little more. The series (2.7) is not the sampling series for f. Rather, it is the other series in (2.9) which gives us a sampling theorem of the form

$$f(t)=\Sigma f(t_n)g_n(t). \tag{2.10}$$

This $\{g_n\}$ may be given [11, p. 149] by the entire functions

$$g_n(t)=\frac{g(t)}{g'(t_n)(t-t_n)}, \tag{2.11}$$

where

$$g(t)=t\prod_{n=1}^{\infty}\left(1-\frac{t^2}{t_n^2}\right). \tag{2.12}$$

Each $g_n\in B_0(\pi)$, but $g(t)\notin B_0(\pi)$. Rather, it is an element of $B_1(\pi)$ just as $\sin \pi t$ is.

3. Nonuniform sampling in $B_{2N}^0(\sigma)$. We shall find a frame in $B_{2N}^0(\sigma)$ and then shall try to emulate the procedure used in $B_0(\sigma)$ to obtain a sampling theorem. We consider only the case $\sigma=\pi$, since the general case follows by a change of scale.

LEMMA 3.1. *Let $\{t_n\}_{-\infty}^{\infty}$ be a sequence of complex numbers such that*

(i) $t_n=-t_{-n}$, $n=0, \pm1, \pm2, \cdots$,

(ii) $\sup_n |\text{Re } t_n-n|<\frac{1}{4}$,

(iii) $|\text{Im } t_n|\leqq C<\infty$;

and let $\{r_n\}$ and $\{h_n\}$ be given by

$$r_n(t)=(t^2+1)^N\frac{\sin \pi(t-t_n)}{\pi(t-t_n)}, \qquad n=0, \pm1, \pm2, \cdots, t\in\mathbb{R}, \tag{3.1}$$

$$h_n(t)=(t^2+1)^N g_n(t), \qquad n=0, \pm1, \pm2, \cdots, t\in\mathbb{R}, \tag{3.2}$$

where g_n is given by (2.11). *Then $\{r_n\}$ and $\{h_n\}$ are both exact frames in $B_{2N}^0(\pi)$ and*

$$\langle r_n, h_m\rangle_{-2N}=\delta_{nm}.$$

Proof. For a sequence $\{t_n\}$ satisfying the hypothesis, the functions $\{e^{it_n\omega}\}$ constitute a frame in $L^2[-\pi,\pi]$, [11, p. 196]. Hence, the Fourier transform of $e^{it_n\omega}\chi_\pi(\omega)$, $\chi_\pi(\omega)$, the characteristic function of $[-\pi,\pi]$, given by

$$S_n(t)=\frac{\sin \pi(t-t_n)}{\pi(t-t_n)},$$

forms a frame in $B_0(\pi)$ since the Fourier transform is an isometry from $L^2(\mathbb{R})$ into itself.

Now let $f\in B_{2N}^0(\pi)$. Then

$$\langle f, r_n\rangle_{-2N}=\int\frac{f(t)}{(t^2+1)^N}S_n(t)\,dt$$
$$=\left\langle\frac{f(t)}{(t^2+1)^N}, S_n\right\rangle_0,$$

and $f(t)(t^2+1)^{-N}\in B_0(\pi)$. Hence, there are constants A and C such that

$$A\left\|\frac{f(t)}{(t^2+1)^N}\right\|_0^2\leqq\sum_n|\langle f, r_n\rangle_{-2N}|^2\leqq C\left\|\frac{f(t)}{(t^2+1)^N}\right\|_0^2 \tag{3.3}$$

by the frame inequality (2.6), since $\{S_n\}$ is a frame in $B_0(\pi)$. But we have

$$\left\|\frac{f(t)}{(t^2+1)^N}\right\|_0^2=\|f\|_{-2N}^2$$

so that $\{r_n\}$ is a frame in $B^0_{2N}(\pi)$. Similarly, since $\{S_n\}$ is exact in $B_0(\pi)$, $\{r_n\}$ is exact in $B^0_{2N}(\pi)$.

The biorthogonality of $\{h_n\}$ with $\{r_n\}$ follows similarly, as does the fact that $\{h_n\}$ is also a frame. □

COROLLARY 3.2. *Let* $f\in B^0_{2N}(\pi)t_n\in\mathbb{R}$; *then*

$$f(t)=\sum_{n=-\infty}^{\infty} f(t_n)g_n(t)\left(\frac{t^2+1}{t_n^2+1}\right)^N,\qquad t\in\mathbb{R},\tag{3.4}$$

with convergence in the sense of L_{2N}.

The proof follows from the fact that

$$\langle f, r_n\rangle_{-2N}=\frac{f(t_n)}{(t_n^2+1)^N}.$$

However, we wish to obtain a theorem similar to (2.1) for $f\in B_{2N}(\pi)$. The correction term $P(t)\sin\sigma t$ will not work in this case since it does not necessarily vanish at the sampling points. It may be replaced by $P(t)g(t)$, where $g(t)$ is given by (2.12), which does work.

COROLLARY 3.3. *Let* $f\in B_{2N}(\pi)$, $t_n\in\mathbb{R}$; *then there is a polynomial* $P(t)$ *of degree* $\leqq 2N-1$ *such that*

$$f(t)=P(t)g(t)+\sum_{n=-\infty}^{\infty} f(t_n)g_n(t)\left(\frac{t^2+1}{t_n^2+1}\right)^N,\tag{3.5}$$

where the series converges in the sense of L_{2N} *and uniformly on compact subsets of the complex plain.*

The polynomial $P(t)$ is chosen such that $f(t)-P(t)g(t)$ belongs to $B^0_{2N}(\pi)$. This will occur if

$$f^{(j)}(\pm i)=\frac{d^j}{dt^j}(P(t)g(t))\big|_{t=\pm i},\qquad j=0,1,\cdots,N-1.\tag{3.6}$$

In practical cases the sampling points t_n will be real and $f(t)$ will be known only for real values of t. In such a case $P(t)$ may be chosen by using

$$\inf_{\substack{a_0,\cdots,a_{N-1}\\ b_0,\cdots,b_{N-1}}}\int\left|f(t)-\sum_{j=0}^{N-1}(a_j-b_jt)(t^2+1)^jg(t)\right|^2(t^2+1)^{-2N}\,dt,\tag{3.7}$$

i.e., the projection of f onto the space spanned by polynomial multiples of degree $2N-1$ of $g(t)$, which belongs to $B^0_{2N}(\pi)$.

The uniform convergence of (3.5) on bounded sets follows as in § 2, by using the reproducing kernel of $B^0_{2N}(\pi)$.

4. Summability of sampling series. The results of the last section (Corollary 3.3), while generalizing the original sampling theorem, lack its elegance and simplicity. In [9], the extra factor in the series was removed by considering (C,α) summability as a substitute for ordinary convergence. In this section we shall see that a similar result holds for nonuniform sampling as well.

A series $\Sigma_{n=-\infty}^{\infty}u_n$ is said to be (C,α) summable to s if

$$\lim_{K\to\infty}\sum_{n=-K}^{K}C^\alpha_{K,n}u_n=s,\qquad C^\alpha_{K,n}=\frac{\binom{K-|n|+\alpha}{\alpha}}{\binom{K+\alpha}{\alpha}}.$$

For Fourier series the (C, α) means are given by

$$\begin{aligned}\sigma_K^\alpha(\omega) &= \sum_{n=-K}^{K} C_{K,n}^\alpha a_n e^{i\omega n}\\ &= \int_{-\pi}^{\pi} F(x)\left\{\frac{1}{2\pi}\sum_{n=-K}^{K} C_{K,n}^\alpha e^{in(\omega-x)}\right\} dx \qquad (4.1)\\ &= \int_{-\pi}^{\pi} F(x) H_K^\alpha(\omega-x)\,dx,\end{aligned}$$

where H_K^α is the kernel of (C, α) summability for Fourier series. In [9] it was shown that if supp $F \subset (-\pi, \pi)$ and if the Fourier transform $f(t) \in B_{2N}(\pi)$, then the sampling series is (C, α) summable to $f(t)$ for $\alpha > 2N$. We shall attempt to do the same for nonuniform sampling.

We shall first consider the function $e^{i\omega t}$ and show that its derivatives have uniformly convergent nonharmonic Fourier series on $-\pi+\varepsilon < \omega < \pi - \varepsilon$. Let $\varphi(\omega)$ be a $C^\infty(\mathbb{R})$ function with support on $[-\pi, \pi]$ such that $\varphi(\omega) = 1$, $|\omega| < \pi - \varepsilon$.

LEMMA 4.1. *Let $\theta_t^p(\omega) = D^p(e^{i\omega t}\varphi(\omega))$ with expansions*

$$(4.2) \qquad \theta_t^p(\omega) = \sum_{n=-\infty}^{\infty} a_n(t)\, e^{it_n\omega} = \sum_{n=-\infty}^{\infty} b_n(t)\, e^{in\omega}$$

with convergence of both series in $L^2(-\pi, \pi)$; then

$$(4.3) \qquad d_K^{\alpha,p}(t,\omega) = \sum_{n=-K}^{K} (a_n(t)\, e^{it_n\omega} - b_n(t)\, e^{in\omega}) C_{K,n}^\alpha, \qquad \alpha \geqq 0,$$

converges to zero uniformly for t on bounded sets and $|\omega| \leqq \pi - \varepsilon$ as $K \to \infty$.

For $\alpha = 0$, the proof is similar to that in [11, p. 198]. We need only prove the uniform convergence for t in addition. Using the notation there, we find that (4.3) may be given by

$$(4.4) \qquad d_K^{\alpha,p}(t,\omega) = \sum_{k=0}^{\infty} \langle \psi_k(t,\omega), (\omega^k - x^k) D_K(\omega - x)\rangle_0,$$

where $D_K(\omega)$ is the Dirichlet kernel and ψ_k the function

$$(4.5) \qquad \psi_k(t,\omega) = \sum_{n=-\infty}^{\infty} a_n(t) \frac{i^k(t_n-n)^k}{k!} e^{in\omega}.$$

Since $|t_n - n| \leqq L < \frac{1}{4}$, the norm of ψ_k in $L^2(-\pi, \pi)$,

$$\begin{aligned}\|\psi_k\|^2 &\leqq \frac{L^{2k}}{(k!)^2}\sum_{n=-\infty}^{\infty} |a_n(t)|^2 \leqq C\frac{L^{2k}}{(k!)^2}\int_{-\pi}^{\pi} |\theta_t^p(\omega)|^2\,d\omega\\ &= \frac{L^{2k}}{(k!)^2} a^2(t),\end{aligned}$$

where $a(t)$ is a continuous function of polynomial growth on $\mathbb{R}^1$. Then each of the terms in (4.4) is dominated by

$$|\langle \psi_k(t,\omega), (\omega^k - x^k) D_K(\omega - x)\rangle| \leqq C a(t) \frac{(\pi L)^k}{k!}$$

as in [11, p. 200]. From this the uniform convergence of the expression (4.3) to zero follows easily for $\alpha = 0$. For other values of α we use the fact that (C, α) summability is a regular summability method for $\alpha > 0$, i.e., all convergent series are summable to the same value. □

We now assume that $f(t)\in B_{2N}(\pi)$ and $F(\omega)$ has its support on $[-\pi+\varepsilon, \pi-\varepsilon]$. Then we may express F as

$$F(\omega)=D^{2N}G(\omega)+\sum_{j=0}^{2N-1} c_j\delta^{(j)}(\omega), \tag{4.6}$$

where $G\in L^2(-\pi, \pi)$ with support on $[-\pi+\varepsilon, \pi-\varepsilon]$ and c_j, $j=0, 1, \cdots, 2N-1$ are constant [9]. It follows that $f(t)$ may be given by

$$f(t)=\frac{1}{2\pi}(-1)^N t^{2N}\int_{-\pi}^{\pi} e^{-i\omega t}G(\omega)\,d\omega+\frac{1}{2\pi}\sum_{j=0}^{2N-1}(it)^j c_j. \tag{4.7}$$

Let us denote by $\sigma_{K,t}$ and $\sigma^*_{K,t}$ the Kth (C, α) mean of the Fourier series and of the nonharmonic Fourier series of $e^{i\omega t}\varphi(\omega)$, respectively. For $p=0$ we have by (4.3),

$$d^{\alpha,0}_K(t, \omega)=\sigma^*_{K,t}(\omega)-\sigma_{K,t}(\omega),$$

which is C^∞ and hence

$$\langle D^{2N}G, d^{\alpha,0}_K(t, \omega)\rangle=\langle G, D^{2N}d^{\alpha,0}_K(t, \omega)\rangle. \tag{4.8}$$

By the uniqueness theorems for the two types of Fourier series, we may deduce that $D^{2N}d^\alpha_K$ is their difference for $D^{2N}(e^{i\omega t}\varphi(\omega))$, i.e.,

$$d^{\alpha,2N}_K(t, \omega)=D^{2N}d^{\alpha,0}_K(t, \omega).$$

Hence, (4.8) becomes, by the lemma,

$$\begin{aligned}\langle D^{2N}G, d^{\alpha,0}_K(t, \omega)\rangle&=\langle G, d^{\alpha,2N}_K(t, \omega)\rangle\to\langle G, \theta^{2N}_t\rangle\\&=\int_{-\pi+\varepsilon}^{\pi-\varepsilon} G(\omega)D^{2N}(e^{-i\omega t}\varphi(\omega))\,d\omega\\&=(-it)^{2N}\int_{-\pi}^{\pi} G(\omega)\,e^{-i\omega t}\,d\omega \quad \text{as } K\to\infty.\end{aligned} \tag{4.9}$$

The terms $\Sigma\, c_j\delta^j$ are treated similarly. Thus we conclude that

$$\begin{aligned}&\langle F, \sigma^*_{K,t}\rangle-\langle F, \sigma_{K,t}\rangle\\&\quad=\langle F, d^{\alpha,0}_K(t, \omega)\rangle\to 0 \quad \text{as } K\to\infty.\end{aligned}$$

But $\langle F, \sigma^*_{K,t}\rangle$ and $\langle F, \sigma_{K,t}\rangle$ are, respectively, the (C, α) means of the series (1.1) and the standard sampling theorem $(t_n=n)$. Since for $\alpha>2N$ the latter converges to $f(t)$ [9], it follows that the former converges as well. We have proved the following.

THEOREM 4.2. *Let $f\in B_{2N}(\pi)$ have a Fourier transform with support in $[-\pi+\varepsilon, \pi-\varepsilon]$ and let $\alpha>2N$; let $\{t_n\}$ satisfy $t_n=-t_{-n}$ and $|t_n-n|\leqq L<\frac{1}{4}$; $t_n\in\mathbb{R}$, $n\in\mathbb{Z}$; then*

$$f(t)=\sum_{n=-\infty}^{\infty} f(t_n)g_n(t)$$

in the sense of (C, α) summability uniformly on bounded subsets of $\mathbb{R}$.

5. Convergence in the sense of generalized functions. The results in §§ 3 and 4 both require additional restrictions on $f(t)$, beyond that it be π bandlimited of polynomial growth. However, there is a sense in which (1.1) holds for all such $f(t)$. It has been shown [7], [5] that (1.1) converges in the sense of Z'_π [12], the dual of the Paley-Wiener space Z_π. This is not a contradiction since the counter example sin πt is equivalent to zero in Z'_π. (The same is true of cos πt, but its coefficients are not zero.)

In order to extend these results to nonuniform sampling points, we must first consider the nonharmonic Fourier series of distributions $F(\omega)\in D'_\pi$ (D_π is the set of $\varphi\in D(\mathbb{R})$ which vanish for $|\omega|>\pi$ with the induced topology. It is the Fourier transform of Z_π.) Since $F(\omega)$ is completely arbitrary for $|\omega|>\pi$, we take it to be zero there.

LEMMA 5.1. *Let $\{t_n\}$ be a sequence satisfying the conditions of Lemma 3.1; let $\varphi\in D_\pi$; then the nonharmonic Fourier series of φ,*

$$\varphi(\omega)\sim\sum a_n e^{it_n\omega}, \tag{5.1}$$

converges to a function $\varphi^*(\omega)$ whose restriction to $(-\pi,\pi)$ is the same as that of $\varphi(\omega)$. This convergence is in the sense of E. (The series and all its derivatives converge uniformly.)

Proof. Since $\varphi\in L^2(-\pi,\pi)$ the series (5.1) converges in the sense of $L^2(-\pi,\pi)$ to $\varphi(\omega)$ and $\{a_n\}\in l^2$. The same is true of $\varphi'(\omega)$ whose coefficients we denote by $\{a'_n\}$, with $a'_0=0$. Hence, we have

$$\varphi(\omega)=\int_{-\pi}^{\omega}\varphi'=\sum_{n\pm 0}a'_n\frac{(e^{it_n\omega}-e^{-i\pi t_n})}{it_n}$$

$$=\sum_{n\pm 0}\frac{a'_n e^{it_n\omega}}{it_n}-\sum_{n\pm 0}\frac{a'_n e^{-i\pi t_n}}{it_n}$$

in which both series converge, the latter to a constant. By using the biorthogonality of $\{G_n(\omega)\}$ and $\{e^{it_n\omega}\}$, we find that

$$a'_n=it_na_n,\qquad n\neq 0,$$

and by iterating the procedure

$$a_n^{(p)}=(it_n)^pa_n,\qquad n\neq 0. \tag{5.2}$$

Here $a_n^{(p)}$ are the coefficients of $\varphi^{(p)}$ and form a sequence in l^2. Thus $\{a_n\}$ is rapidly decreasing and the series (5.1) converges in the sense of E to some $\varphi^*(\omega)$. □

We now turn to generalized functions $F(\omega)$ with support in $[-\pi,\pi]$. We shall require that F be strongly integrable over $[-\pi,\pi]$ (see [10]). This requires that F be the pth derivative of a continuous function G in some neighborhood of π (respectively, $-\pi$) such that

$$\frac{G(\omega)}{(\omega-\pi)^{p-1}}\to 0\quad\text{as }\omega\to\pi\quad\text{(respectively, }-\pi). \tag{5.3}$$

Since F has compact support it belongs to E' and

$$\langle F,\varphi\rangle=\langle F,\varphi^*\rangle=\left\langle F,\sum_n a_n e^{it_n\omega}\right\rangle=\sum_n a_n\langle F,e^{it_n\omega}\rangle. \tag{5.4}$$

The inverse Fourier transform of such $F\in E'$ is given by

$$f(t)=\frac{1}{2\pi}\langle F,e^{it\omega}\rangle.$$

Hence, (5.4) becomes

$$\langle F,\varphi\rangle=\sum_n a_n 2\pi f(t_n). \tag{5.5}$$

But $\langle F, \varphi\rangle = 2\pi\langle f, \tilde{\varphi}\rangle$, where $\tilde{\varphi}$ is the inverse Fourier transform of φ. Also a_n is given by

$$a_n = \frac{1}{2\pi}\langle G_n, \varphi\rangle = \langle g_n, \tilde{\varphi}\rangle,$$

and (5.5) becomes

$$\langle f, \tilde{\varphi}\rangle = \langle \Sigma f(t_n) g_n, \tilde{\varphi}\rangle. \tag{5.6}$$

Thus, except for the uniqueness, we have proved the following.

THEOREM 5.2. *Let f be a π bandlimited function of polynomial growth; let $\{t_n\}$ satisfy the conditions of Lemma* 3.1; *then $f(t)$ has the sampling expansion*

$$f(t) = \Sigma f(t_n) g_n(t),$$

convergent in the sense of Z'_π. It is unique if $F(\omega)$ is strongly integrable over $[-\pi, \pi]$.

The uniqueness does not always follow since nonzero sampling series may converge to zero in Z'_π. This happens for any $f \in Z'_\pi$ such that $\langle f, \psi\rangle = 0$ for all $\psi \in Z_\pi$, e.g., $f(t) = \cos \pi t$ or $f(t) = \sin \pi t$.

If $F(\omega)$ is strongly integrable over $[-\pi, \pi]$ and $\langle F, \psi\rangle = 0$ for all $\psi \in D_\pi$, then F has support on $\pm\pi$. Such generalized functions have the form

$$\sum_{j=0}^{m} b_j \delta^{(j)}(\omega + \pi) + c_j \delta^{(j)}(\omega - \pi),$$

which are strongly integrable only if $b_j = c_j = 0$, $j = 0, 1, \cdots, m$. Thus $F \equiv 0$, and uniqueness follows. □

REFERENCES

[1] P. BUTZER, W. SPLETTSTÖβER, AND R. STENS, *The sampling theorem and linear prediction in signal analysis*, Jahresber. Deutsch. Math.-Verein., 90 (1988), pp. 1-70.

[2] L. CAMPBELL, *A comparison of the sampling theorems of Kramer and Whittaker*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 117-130.

[3] R. F. HOSKINS AND J. DESOUSA PINTO, *Sampling expansions for functions bandlimited in the distributional sense*, SIAM J. Appl. Math., 44 (1984), pp. 605-610.

[4] A. JERRI, *The Shannon sampling theorem—its various extensions and applications: a tutorial review*, Proc. IEEE, 11 (1977), pp. 1565-1596.

[5] R. J. KRECZNER, *Generalized cardinal series*, Ph.D. thesis, University of Wisconsin-Milwaukee, Milwaukee, WI, 1988.

[6] A. LEE, *A note on the Campbell sampling theorem*, SIAM J. Appl. Math., 41 (1981), pp. 553-557.

[7] E. PFAFFELHUBER, *Sampling series for bandlimited generalized functions*, IEEE Trans. Inform. Theory, IT-17 (1971), pp. 650-654.

[8] C. SHANNON, *Communication in the presence of noise*, Proc. IRE, 37 (1949), pp. 10-21.

[9] G. G. WALTER, *Sampling bandlimited functions of polynomial growth*, SIAM J. Math. Anal., 19 (1988), pp. 1198-1201.

[10] ———, *Abel summability for a distributional sampling theorem*, in Generalized functions, convergence structures and their applications, Stankovic, Pap, Pilipovíc, and Vladimirov, eds., Plenum, New York, 1988, pp. 349-357.

[11] R. M. YOUNG, *An introduction to nonharmonic Fourier series*, Academic Press, New York, 1980.

[12] A. H. ZEMANIAN, *Distribution theory and transform analysis*, McGraw-Hill, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 1004-1014, July 1992

015

DISCRETE DISCRETE WAVELETS*

GILBERT G. WALTER†

Abstract. A discrete family of wavelets consisting of discrete functionals in a Sobolev space is studied. It is shown that they form a complete orthonormal system in H^{-s}, $s>\frac{1}{2}$ generated by a single "mother functional." Closed form expressions are derived in certain cases.

Key words. orthonormal basis, Sobolev spaces, generalized functions, wavelets

AMS(MOS) subject classifications. 42C10, 46F10

1. Introduction. A recurring problem in signal analysis is to approximate an analog signal by a digital signal in order to use the more versatile tools of digital signal processing. That is, a real valued function $f(t)$ on $\mathbb{R}$ must be approximated in some sense by (the generalized function)

$$f^*(t)=\sum_{k=-\infty}^{\infty} a_k\delta(t-t_k). \tag{1.1}$$

Furthermore $f(t)$ must be recoverable from $f^*(t)$. This is possible when $f(t)$ is a bandlimited function, i.e., when its Fourier transform

$$\hat{f}(w)=\int_{-\infty}^{\infty} e^{-2\pi iwt}f(t)\,dt$$

has compact support, say in the unit interval $[-\frac{1}{2},\frac{1}{2}]$. Then the sampling theorem says that if f^* is taken to be

$$f^*(t)=\sum_{k=-\infty}^{\infty} f(k)\delta(t-k),$$

$f(t)$ may be recovered by taking the projection of $f^*(t)$ into a subspace of an appropriate Sobolev space (see [6], [1]).

However, there appears to be no simple way of expanding more general functions in sequences of δ's. This is particularly true since $\delta(t-a)$ and $\delta(t-b)$ are not orthogonal for $a\neq b$ in the Sobolev space H^{-s}, for $s>\frac{1}{2}$ to which they belong. Indeed the inner product in H^{-s} is

$$\langle f,g\rangle_{-s}=\int_{-\infty}^{\infty}\hat{f}(w)\overline{\hat{g}(w)}(w^2+1)^{-s}\,dw \tag{1.2}$$

and hence

$$\begin{aligned}\langle \delta_a,\delta_b\rangle_{-s}&=\int_{-\infty}^{\infty} e^{-2\pi iaw}\,e^{2\pi ibw}(w^2+1)^{-s}\,dw\\&=\hat{h}(a-b),\end{aligned}$$

where $h(w)=(w^2+1)^{-s}$.

* Received by the editors December 26, 1989; accepted for publication (in revised form) September 18, 1991.

† Department of Mathematical Sciences, P.O. Box 413, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201.

Fortunately there is a framework in which this expansion can be formulated. It involves the theory of wavelet expansions for the affine group.

1.1. Wavelets. The "wavelets of constant shape," or simply "wavelets," were first proposed by J. Morlet et al. [5] as a tool for analysis of seismic data. They were shown by Grossman and Morlet [3] to involve a square integrable representation of the affine group. A representation of an $L^2(\mathbb{R})$ function f was given in terms of its wavelet expansion

$$f \sim \langle h^{(u,v)}, f\rangle = \phi(u, v),$$

where $h^{(u,v)}(x) = 2^{u/2}h(2^u x - v)$ is an appropriate function in $L^2(\mathbb{R})$. This representation was in terms of integrals of $\phi(u, v)$.

A discrete version was also introduced by Grossman [2], [3] which involved the theory of frames. This was further refined by Meyer [4] who showed how to construct discrete orthognal systems of wavelets. These, in the case of the real line, were of the form

$$\varphi_{m,k}(x) = 2^{m/2}h(2^m x - k),$$

where h is a "mother function" in $L^2(\mathbb{R})$, which is continuous and piecewise linear except at half integers. He also succeeded in constructing systems in which h was in C^m and even C_0^∞.

In this work we shall emulate some aspects of his approach, but with H^{-s} spaces and with δ's instead of functions. The resulting theory is even simpler and will result in a discrete "mother functional" h of the form

$$h(x) = \sum_{k=-\infty}^{\infty} a_k \delta(2x - 1 - k).$$

For H^{-1}, h will be shown to be given by a particularly simple formula

$$h(x) = a\delta(2x-1) - b(\delta(2x) + \delta(2x-2)),$$

where a and b are constants.

2. Wavelets in H^{-s}, $s > \frac{1}{2}$. In this section we present the general theory for any real s. We shall defer detailed calculations to the next section in which we shall obtain a closed form expression for the mother functional in some cases.

2.1. Continuous wavelet expansions. The continuous wavelet expansions in H^{-s} may be easily based on those for $L^2(\mathbb{R})$ [3]. Let $g \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ such that $\int_{-\infty}^{\infty} g = 0$. The wavelet transform of $f \in L^2$ with respect to the wavelets based on g is

$$\phi(u, v) = \int 2^{u/2}\bar{g}(2^u x - v) f(x)\, dx \tag{2.1}$$

and the representation of f is

$$f(x) = C \iint 2^{u/2} g(2^u x - v)\phi(u, v)\, du\, dv, \tag{2.2}$$

where C is constant depending only on g.

We illustrate the extension of H^{-s} for $s = 1$. Let h denote the conjugate Fourier transform of the weight $(w^2+1)^{-1}$, $h(x) = \pi e^{-2\pi|x|}$. Then for $f \in H^{-1}$, we replace (2.1) by

$$\phi(u, v) = \langle f, g^{(u,v)}\rangle_{-1} \tag{2.3}$$

where $g^{(u,v)}(x) = 2^{u/2} g(2^u x - v)$.

But $\langle f, g\rangle_{-1} = \langle f*h, g\rangle_0$ when $g \in L^2(\mathbb{R})$, and hence (2.3) becomes (2.1) with f replaced by $f*h$. Then the representation theorem gives us

$$(f*h)(x) = C \iint g^{(u,v)}(x)\phi(u, v)\, du\, dv. \tag{2.4}$$

Since h is the Green's function for $1-(D/2\pi)^2$, we may recover f by operating on both sides of (2.4) with this operator.

$$f(x) = \left[1-\left(\frac{D}{2\pi}\right)^2\right] C \iint g^{(u,v)}(x)\phi(u, v)\, du\, dv.$$

We shall not pursue this further since our main interest is in discrete rather than continuous wavelets.

2.2. Some subspaces of H^{-s}. We imitate Meyer's [4] procedure for L^2 in H^{-s} and define B_0 to be the subspace

$$B_0 = \{f \in H^{-1} | \operatorname{supp} f \subset \mathbb{Z}\}.$$

The subspace B_j is then defined to be the dilation

$$B_j = \{f \in H^{-1} | f(2^{-j}\cdot) \in B_0\}.$$

We have a characterization of B_0 given by the following.

PROPOSITION 2.1. *$f \in B_0$ if and only if*

$$f(x) = \sum_{k=-\infty}^{\infty} a_k \delta(x-k), \qquad \{a_k\} \in l^2. \tag{2.5}$$

Proof. If f has the form given, then it clearly has support on $\mathbb{Z}$ and its Fourier transform is

$$\hat{f}(w) = \sum_k a_k e^{-2\pi i w k},$$

which is periodic and in L^2_{loc}. Hence $\hat{f} \in L^2[(w^2+1)^{-s}]$ for all $s > \frac{1}{2}$, and hence $f \in B_0 \cap H^{-s}$.

If $f \in B_0$, then f is a tempered distribution in S' of order $\leqq 1$ and hence has the form of (2.5) with convergence in the sense of S'. The Fourier transform must be in $L^2(0, 1)$, and hence the Fourier coefficient sequence $\{a_k\} \in l^2$.

It should be observed that by this argument $B_0 \subset H^{-s}$ for all $s > \frac{1}{2}$. Also it follows that

$$B_0 \subset B_1 \subset \cdots \subset B_j \subset \cdots \subset H^{-s}$$

and

$$\overline{\bigcup B_j} = H^{-s}$$

since trigonometric polynomials are dense in $L^2[(w^2+1)^{-s}]$.

We define C_{j+1} to be the orthogonal complement of B_j in B_{j+1} in the sense of H^{-s}. We shall then look for an element $\varphi_j \in C_{j+1}$ whose translates

$$\varphi_{jk}(x) = \varphi_j(x - 2^{-j}k)$$

constitute an orthogonal basis of C_{j+1}. Since

$$H^{-s} = B_0 \oplus \sum_{j=1}^{\infty} \oplus C_j,$$

it follows that $\{\varphi_{jk}\}$ will be an orthonormal system in H^{-s}, which, together with a basis for B_0, will be a basis of H^{-s}.

2.3. An orthonormal basis for B_0. Let $\varphi \in B_0$, $\varphi(x) = \sum_{k=-\infty}^{\infty} a_k \delta(x-k)$; we wish to choose φ such that translates of φ are orthogonal. Then

$$\delta_{0k} = \langle \varphi, \varphi_k \rangle_{-s} = \int |\hat{\varphi}(w)|^2 \, e^{-2\pi i w k} (w^2+1)^{-s} \, dw. \tag{2.6}$$

Since by the Poisson summation formula

$$\sum_{n=-\infty}^{\infty} f(w+n) = \sum_{n=-\infty}^{\infty} \hat{f}(n) \, e^{2\pi i n w},$$

it follows for $f(w) = |\hat{\varphi}(w)|^2 (w^2+1)^{-s}$ and $\hat{f}(k) = \delta_{0k}$ (2.6) that

$$\sum_{n=-\infty}^{\infty} |\hat{\varphi}(w+n)|^2 ((w+n)^2+1)^{-s} = \sum_{k=-\infty}^{\infty} \delta_{0k} \, e^{2\pi i k w} = 1. \tag{2.7}$$

Clearly $\hat{\varphi}(w) = \sum_{k=-\infty}^{\infty} a_k \, e^{-2\pi i k w}$ is 1-periodic and hence (2.7) becomes

$$|\hat{\varphi}(w)|^2 \beta_s(w) = 1,$$

where

$$\beta_s(w) = \sum_{n=-\infty}^{\infty} ((w+n)^2+1)^{-s}.$$

Since $\beta_s(w) > 0$ for all w, it follows that one solution to (2.6) is

$$\hat{\varphi}_0(w) = \beta_s^{-1/2}(w) \tag{2.8}$$

and the general solution is

$$\hat{\varphi}(w) = \chi(w) \hat{\varphi}_0(w),$$

where $|\chi(w)| = 1$, $\chi \in L^2_{\text{loc}}$, and χ is 1-periodic.

The particular solution φ_0 may be characterized up to a constant $e^{i\lambda}$, as the one with minimum $\|x\varphi\|^2_{-s}$ among those for which this norm is finite. It clearly is for φ_0 since

$$\hat{\varphi}_0'(w) = -\tfrac{1}{2} \beta_s^{-3/2}(w) \beta_s'(w),$$

both of whose factors are continuous and periodic. Then

$$\begin{aligned} \|x\varphi\|^2_{-s} &= \int_{-\infty}^{\infty} |\hat{\varphi}'(w)|^2 (w^2+1)^{-s} \, dw \\ &= \int_{-\infty}^{\infty} |\chi'(w)|^2 |\hat{\varphi}_0(w)|^2 (w^2+1)^{-s} \, dw \\ &\quad + \int_{-\infty}^{\infty} |\chi(w)|^2 |\hat{\varphi}_0'(w)|^2 (w^2+1)^{-s} \, dw \\ &\quad + 2 \operatorname{Re} \int_{-\infty}^{\infty} \hat{\varphi}_0 \hat{\varphi}_0' \chi \overline{\chi'} (w^2+1)^{-s} \, dw. \end{aligned}$$

Since $\chi \overline{\chi'}$ is purely imaginary, this last term is zero and hence

$$\|x\varphi\|^2_{-s} \geqq \|x\varphi_0\|^2_{-s}.$$

We must still show that $\{\varphi_0(x-k)\}$ is a basis for B_0, which we do with the same argument. Indeed, let $f \in B_0$

$$\langle f, T^k\varphi_0\rangle_{-s} = 0, \qquad k = 0, \pm 1, \pm 2, \cdots,$$

where T is the shift operator $T\varphi(x) = \varphi(x-1)$. Then

$$0 = \int \frac{\hat{f}(w)\hat{\varphi}_0(w)\, e^{2\pi i w k}}{(w^2+1)^s}\, dw, \qquad k = 0, \pm 1, \cdots,$$

and by the Poisson summation formula again,

$$\hat{f}(w)\hat{\varphi}_0(w) \sum_{n=-\infty}^{\infty} \frac{1}{((w^2+n)^2+1)^s} = 0, \tag{2.9}$$

and hence $\hat{f}(w) = 0$. This shows completeness. Hence we have proved the following.

PROPOSITION 2.2. *The translates* $\{T^k\varphi_0\}$, φ_0 *given by* (2.8), *form an orthonormal basis of* B_0.

2.4. An orthonormal basis for C_1. In order to find a basis for C_1 consisting of translates of a fixed ψ, we look for $\psi \in B_1$ such that

(i) $\langle \psi, \delta_k\rangle_{-s} = 0$, $k = 0, \pm 1, \cdots$, which will ensure that ψ is orthogonal to B_0 and hence in C_1;

(ii) $\langle \psi, \psi_k\rangle_{-s} = \delta_{0k}$, the orthogonality conditon;

(iii) If $f \in B_1$ satisfies $\langle f, \delta_k\rangle_{-s} = \langle f, \psi_k\rangle_{-s} = 0$, $k = 0, \pm 1, \cdots$, then $f = 0$, the completeness condition.

These three conditions are expressed, respectively, as

$$\text{(i}'\text{)} \quad \int_{-\infty}^{\infty} \hat{\psi}(w)\, e^{i2\pi w k}(w^2+1)^{-s}\, dw = 0, \qquad k = 0, \pm 1, \cdots.$$

or by the Poisson summation formula

$$\sum_{n=-\infty}^{\infty} \hat{\psi}(w+n)((w+n)^2+1)^{-s} = 0, \ w \in \mathbb{R};$$

$$\text{(ii}'\text{)} \quad \sum_{n=-\infty}^{\infty} |\hat{\psi}(w+n)|^2((w+n)^2+1)^{-s} = 1, \ w \in \mathbb{R};$$

$$\text{(iii}'\text{)} \quad \sum_{n=-\infty}^{\infty} \hat{f}(w+n)((w+n)^2+1)^{-s} = 0, \ w \in \mathbb{R}; \quad \text{and}$$

$$\sum_{n=-\infty}^{\infty} \hat{f}(w+n)\overline{\hat{\psi}(w+n)((w+n)^2+1)^{-s}} = 0, \ w \in \mathbb{R}.$$

Since both ψ and f are in B_1, their Fourier transforms are 2-periodic. Hence we separate each of the series into its even and odd terms. The first one is

$$\sum_n \hat{\psi}(w+2n)((w+2n)^2+1)^{-s} + \sum_n \hat{\psi}(w+2n+1)((w+2n+1)^2+1)^{-s}$$
$$= \hat{\psi}(w)\sum_n ((w+2n)^2+1)^{-s} + \hat{\psi}(w+1)\sum_n ((w+2n+1)^2+1)^{-s},$$

which, if we denote $\sum_n ((w+2n)^2+1)^{-s} = A_s(w)$, becomes

$$\hat{\psi}(w)A_s(w) + \hat{\psi}(w+1)A_s(w+1) = 0. \tag{2.10}$$

Similarly (ii′) and (iii′) become

$$(2.11)\qquad |\hat{\psi}(w)|^2A_s(w)+(\hat{\psi}(w+1))^2A_s(w+1)=1,$$

$$(2.12)\qquad \hat{f}(w)A_s(w)+\hat{f}(w+1)A_s(w+1)=0,$$

and

$$(2.13)\qquad \hat{f}(w)\overline{\hat{\psi}(w)A_s(w)}+\hat{f}(w+1)\overline{\hat{\psi}(w+1)A_s(w+1)}=0.$$

We can solve (2.10) and (2.11) simultaneously for $|\hat{\psi}(w)|^2$,

$$(2.14)\qquad |\hat{\psi}(w)|^2=\frac{A_s(w+1)}{A_s(w)(A_s(w+1)+A_s(w))}.$$

Unfortunately the positive square root of (2.14) does not satisfy (2.10). In fact, if it is substituted in (2.10), the left side will be

$$\frac{A_s^{1/2}(w+1)A_s^{1/2}(w)}{(A_s(w+1)+A_s(w))^{1/2}},$$

which is positive for all w. We can get a solution by choosing the square root to be

$$(2.15)\qquad \hat{\psi}_1(w)=e^{-i\pi w}|\hat{\psi}(w)|,$$

which satisfies (2.10). However, (2.15) is not unique. We can make it unique by choosing a solution which minimizes $\|(x-\frac{1}{2})\psi(x)\|^2_{-s}$, as we did in the last section for $\|x\varphi\|_{-s}$.

In order to prove the completeness we must show that the only solution to (2.12) and (2.13) is the zero solution. This holds if the determinant of the system

$$d(w)=A_s(w)A_s(w+1)\overline{(\hat{\psi}(w+1)-\hat{\psi}(w))}\neq 0.$$

Since $A_s(w)>0$ and

$$|\hat{\psi}(w)|^2-|\hat{\psi}(w+1)|^2=\left[\frac{A_s(w+1)}{A_s(w)}-\frac{A_s(w)}{A_s(w+1)}\right]\frac{1}{(A_s(w)+A_s(w+1))}$$

by (2.14), it follows that $d(w)=0$ only at those points where $A_s(w)=A_s(w+1)$, which can easily be seen, is a set of measure zero.

Thus we have proved the following.

PROPOSITION 2.3. *There exists an orthonormal basis of C_1 in the topology of H^{-s}, $s>\frac{1}{2}$, given by translates $\{T^k\psi_1\}$, where $\hat{\psi}_1$ is given by* (2.15).

2.5. An orthonormal basis for $C_j, j>1$. Unfortunately in the case of H^{-s}, the dilation map

$$(D_a\varphi)(x)=a^{1/2}\varphi(ax)$$

is not an isometry of H^{-s} into H^{-s} even though it is a homeomorphism. Therefore, the orthonormal basis of C_1 is not necessarily mapped into an orthonormal set in C_j by the dilation operator D_2j, and we shall have to proceed differently.

We return to C_1 but modify the norm slightly to

$$(2.16)\qquad \|\varphi\|^2_{-s,\delta}=\int|\hat{\varphi}(w)|^2(w^2+\delta^2)^{-s}\,dw,\qquad \delta>0,$$

with the corresponding inner product denoted $\langle\varphi, \psi\rangle_{-s,\delta}$. This of course will change the definition of C_1, the orthogonal complement of B_0 in B_1. But B_0 and B_1 will remain the same sets.

The arguments in § 2.4 may be repeated verbatim, requiring only that we replace $A_s(w)$ by

$$A_s(w, \delta)=\sum_n ((w+2n)^2+\delta^2)^{-s}.$$

The corresponding solution to the replacements for (2.10) and (2.11) is

$$\hat{\psi}(w, \delta)=e^{-iw\pi}\left[\frac{A_s(w+1, \delta)}{A_s(w, \delta)(A_s(w+1, \delta)+A_s(w, \delta))}\right]^{1/2}. \tag{2.17}$$

Thus, we see that we again have an orthonormal basis $\{T_k\psi\}$ of the revised space $C_{1,\delta}$ with the norm given by (2.16). We shall use it to construct an orthonormal basis of C_{j+1} (original definition).

The conditions (i), (ii), and (iii) required for an orthonormal basis of C_{j+1} with the norm of H^{-s} may be translated into conditions on C_1 with a modified norm. Indeed we see that for

$$\varphi \in C_1, \qquad D_{2^j}\varphi \in C_{j+1}$$

and

$$\|D_{2^j}\varphi\|^2_{-s}=\|2^{j/2}\varphi(2^jx)\|^2_{-s}=2^{-2js}\|\varphi\|^2_{-s,2^{-j}}.$$

Since ψ, given by (2.17), has the required properties for $\delta=2^{-j}$, it follows that the system of functions given by

$$2^{j(s+\frac{1}{2})}\psi(2^jx-k, 2^{-j})=\psi_{j+1,k}(x), \quad j=0, 1, \cdots, \quad k=0, \pm 1, \cdots, \tag{2.18}$$

is an orthonormal basis of C_j. This gives us the following.

THEOREM 2.4. *Let ψ_{jk} be given by* (2.18) *for $j=1, 2, \cdots$, $k=0, \pm 1, \cdots$, and by $\psi_{0,k}=\varphi_0(x-k)$, $k=0, \pm 1, \cdots$; then $\{\psi_{jk}\}_{j=0,}^{\infty}\,{}_{k=-\infty}^{\infty}$ is a complete orthonormal basis of H^{-s} for $s>\frac{1}{2}$.*

The proof is immediate since $H^{-s}=B_0\oplus\sum_{j=1}^{\infty}\oplus C_j$.

The results for C_j, $j\geqq 1$ can be extended to nonpositive values of j and lead to the decomposition

$$H^{-s}=\{\delta\}\oplus\sum_{j=-\infty}^{\infty}\oplus C_j$$

since $\bigcap_{j=-\infty}^{\infty} B_j=\{\delta\}$, the space spanned by the δ function.

3. Explicit formulas for H^{-m}, m an integer. In some cases it is possible to sum the series arising in the last section explicitly and hence obtain closed form formulas for $A_s(w, \delta)$ and $\psi(w, \delta)$. We shall do so for integer values for s.

3.1. The case $s=1$. For $s=1$, we can sum this series defining $A_1(w, \delta)$ easily by using the fact that

$$\sum_{n=-\infty}^{\infty}\frac{1}{x+n}=\frac{\pi\cos\pi x}{\sin\pi x} \tag{3.1}$$

with symmetric partial sums being understood. This may be shown using residues or by finding the Fourier series of e^{-ixt} at $t=\pi$. Then

$$A_1(w,\delta)=\sum_n \frac{1}{(w+2n)^2+\delta^2}=\frac{i}{2\delta}\left\{\sum_n \frac{1}{w+2n+i\delta}-\sum_n \frac{1}{w+2n-i\delta}\right\}$$

$$(3.2)\qquad =\frac{\pi i}{4\delta}\;\frac{\cos \pi\left(\dfrac{w+\delta i}{2}\right)}{\sin \pi\left(\dfrac{w+\delta i}{2}\right)}-\frac{\cos \pi\left(\dfrac{w-\delta i}{2}\right)}{\sin \pi\left(\dfrac{w-\delta i}{2}\right)}$$

$$=\frac{\pi i}{2\delta}\,\frac{\sin \pi(-\delta i)}{(\cos \pi(i\delta)-\cos \pi w)}=\frac{\pi}{2\delta}\,\frac{\sinh \pi\delta}{(\cosh \pi\delta-\cos \pi w)}.$$

In order to find $\psi(w,\delta)$ we use (2.17):

$$\hat{\psi}(w,\delta)=e^{-i\pi w}\left\{\frac{A_1(w+1,\delta)}{A_1(w,\delta)(A_1(w+1,\delta)+A_1(w,\delta))}\right\}^{1/2}$$

$$(3.3)\qquad =e^{-i\pi w}\left\{\frac{2\delta/(\cosh \pi\delta+\cos \pi w)}{\dfrac{\pi \sinh \delta\pi}{\cosh \pi\delta-\cos \pi w}\left\{\dfrac{1}{\cosh \pi\delta-\cos \pi w}+\dfrac{1}{\cosh \pi\delta+\cos \pi w}\right\}}\right\}^{1/2}$$

$$=e^{-i\pi w}\left\{\frac{2\delta}{\pi \sinh \delta\pi}\,\frac{(\cosh \pi\delta-\cos \pi w)^2}{2\cosh \pi\delta}\right\}^{1/2}$$

$$=e^{-i\pi w}(\cosh \pi\delta-\cos \pi w)\sqrt{\frac{2\delta}{\pi \sinh 2\pi\delta}}.$$

Hence $\psi(x,\delta)$ may be expressed as

$$(3.4)\qquad \begin{aligned}\psi(x,\delta)&=\sqrt{\frac{2\delta}{\pi \sinh 2\pi\delta}}\left\{(\cosh \pi\delta)\delta\left(x-\frac{1}{2}\right)-\frac{1}{2}\delta(x)-\frac{1}{2}\delta(x-1)\right\}\\ &=b\left\{c\delta\left(x-\frac{1}{2}\right)-\frac{1}{2}\delta(x)-\frac{1}{2}\delta(x-1)\right\},\end{aligned}$$

where $b>0$ and $c>1$ are constants depending on δ. This is the version of "mother functional" in H^{-1}.

The orthonormal basis of B_0 in H^{-1} is given by

$$\hat{\varphi}_0(w)=\beta_1^{-1/2}(w),$$

where

$$\beta_1(w)=\sum_n ((w+n)^2+1)^{-1}=\frac{\pi \sinh 2\pi}{\cosh 2\pi-\cos 2\pi w},$$

$$\hat{\varphi}_0(w)=\left[\frac{\cosh 2\pi-\cos 2\pi w}{\pi \sinh 2\pi}\right]^{1/2}.$$

To find the inverse Fourier transform of $\hat{\varphi}_0$ we use the series expansion

$$(1-x)^{1/2}=\sum_{k=0}^{\infty}\frac{\Gamma(3/2)(-1)^k}{\Gamma(k+1)\Gamma((3/2)-k)}x^k=\sum_{k=0}^{\infty}\gamma_k x^k$$

to express $\hat{\varphi}_0$ as

$$\hat{\varphi}_0(w) = b(1 - c\cos 2w\pi)^{1/2}$$
$$= b\sum_{k=0}^{\infty} \gamma_k c^k \cos^k 2\pi w.$$

Since

$$\int_0^1 \cos^k 2\pi w\, e^{-2\pi i n w}\, dw = \begin{cases} 0, & |n| > k \\ 0, & n+k \text{ odd} \\ \dfrac{1}{2^k}\dbinom{k}{(n+k)/2}, & n+k \quad \text{even,} \end{cases}$$

it follows that the $2n$th Fourier coefficient of $\hat{\varphi}_0(w)$ is

$$c_{2n} = \int_0^1 \hat{\varphi}_0(w)\, e^{-2\pi i 2nw}\, dw = b\sum_{k=|n|}^{\infty} \gamma_{2k} c^{2k} \frac{1}{2^{2k}}\binom{2k}{n+k}$$

while the $(2n+1)$th is

$$c_{2n+1} = \int_0^1 \hat{\varphi}_0(w)\, e^{-2\pi i(2n+1)w}\, dw = b\sum_{k=|n|}^{\infty} \gamma_{2k+1} c^{2k+1}\binom{2k+1}{n+k+1}\frac{1}{2^{2k+1}}.$$

Hence the Fourier series of $\hat{\varphi}_0(w)$ is

$$\hat{\varphi}_0(w) = \sum_{n=-\infty}^{\infty} (c_{2n}\, e^{2\pi i 2nw} + c_{2n+1}\, e^{2\pi i(2n+1)w}).$$

The $\gamma_k < 0$ for $k \geqq 1$, and since all the other factors in each term of the series for c_{2n} and c_{2n+1} are positive, it follows that $c_n < 0$ for all $n \neq 0$. Thus the inverse Fourier transform of $\hat{\varphi}_0(w)$ is

$$\varphi_0(x) = \sum_{n=-\infty}^{\infty} c_n \delta(x-n), \tag{3.5}$$

where none of the terms vanish ($c_0 > 0$).

3.2. Other H^{-m}, $m > 1$. We can find expansion for $A_m(w, \delta)$, $m > 1$ in terms of $A_1(w, \delta)$ and $A_2(w, \delta)$. Indeed we note that $A_s(w, \delta)$ satisfies the recurrence relation

$$\begin{aligned} \frac{d^2}{dw^2}A_s(w,\delta) &= \frac{d^2}{dw^2}\left(\sum \frac{1}{((w+2n)^2+\delta^2)^s}\right) \\ &= s(s+1)\sum((w+2n)^2+\delta^2)^{-s-2}4(w+2n)^2 \\ &\quad -2s\sum((w+2n)^2+\delta^2)^{-s-1} \\ &= 2s(2s+1)A_{s+1}(w,\delta) - 4\delta^2 s(s+1)A_{s+2}(w,\delta). \end{aligned} \tag{3.6}$$

Hence if $A_2(w, \delta)$, in addition to $A_1(w, \delta)$, is known, we can find $A_m(w, \delta)$ for all integers > 0. To find $A_2(w, \delta)$ we use the formula

$$\sum \frac{1}{(x+n)^2} = \frac{\pi^2}{\sin^2 \pi x}, \tag{3.7}$$

which can be obtained by differentiating (3.1). Then we have

$$\begin{aligned}
A_2(w,\delta) &= \sum_n ((w+2n)^2+\delta^2)^{-2} \\
&= \frac{1}{16\delta^2}\left[\sum_n\left(\frac{i}{(w+\delta i)/2+n}\right)^2+\sum_n\left(\frac{i}{(w-\delta i)/2+n}\right)^2\right. \\
&\quad \left.+\sum\frac{2}{(n+(w/2))^2+(\delta/2)^2}\right] \\
&= \frac{-\pi^2}{16\delta^2}\left(\frac{1}{\sin^2\pi(w+\delta i)/2}+\frac{1}{\sin^2\pi(w-\delta i)/2}\right)+\frac{1}{2\delta^2}A_1(w,\delta) \qquad (3.8) \\
&= \frac{-\pi^2}{4\delta^2}\frac{(1-\cosh\pi\delta\cos\pi w)}{(\cosh\pi\delta-\cos\pi w)^2}+\frac{1}{2\delta^2}\frac{\pi}{2\delta}\frac{\sinh\pi\delta}{(\cosh\pi\delta-\cos\pi w)} \\
&= \frac{a_{21}}{\cosh\pi\delta-\cos\pi w}+\frac{a_{22}}{(\cosh\pi\delta-\cos\pi w)^2}.
\end{aligned}$$

The second derivative of $A_1(w, \delta)$ is easily calculated to be

$$\begin{aligned}
\frac{d^2}{dw^2}A_1(w,\delta) &= \frac{\pi^3\sinh\pi\delta}{2\delta}\left[\frac{\cos\pi w}{(\cosh\pi\delta-\cos\pi w)^2}-\frac{2(1-\cos^2\pi w)}{(\cosh\pi\delta-\cos\pi w)^3}\right] \\
&= \frac{C_1}{\cosh\pi\delta-\cos\pi w}+\frac{C_2}{(\cosh\pi\delta-\cos\pi w)^2}+\frac{C_3}{(\cosh\pi\delta-\cos\pi w)^3}.
\end{aligned}$$

Hence by (3.6)

$$\begin{aligned}
A_3(w,\delta) &= \frac{2s+1}{2\delta^2(s+1)}A_2(w,\delta)-\frac{1}{4\delta^2 s(s+1)}\frac{d^2}{dw^2}A_1(w,\delta) \\
&= \frac{a_{31}}{c-z}+\frac{a_{32}}{(c-z)^2}+\frac{a_{33}}{(c-z)^3},
\end{aligned}$$

where

$$c-z=(\cosh\pi\delta-\cos\pi w). \qquad (3.9)$$

The induction step is obvious, so we are able to conclude that

$$A_m(w,\delta)=\sum_{k=1}^m a_{mk}(c-z)^{-k}, \qquad (3.10)$$

where $c-z$ is given by (3.9). Clearly, we also have

$$A_m(w+1,\delta)=\sum_{k=1}^m a_{mk}(c+z)^{-k},$$

and hence

$$|\hat{\psi}(w,\delta)|^2=\frac{\sum_{k=1}^m a_{mk}(c+z)^{-k}}{\sum_{k=1}^m a_{mk}(c-z)^{-k}\{\sum_{k=1}^m a_{mk}[(c-z)^{-k}+(c+z)^{-k}]\}}.$$

The mother functional for H^{-m} will, therefore, again be as in (2.17), where now $|\hat{\psi}(w, \delta)|^2$ is a rational function in $\cos w\pi$. However, the Fourier series of $\hat{\psi}(w, \delta)$ will have an infinite number of terms except in the case of $m=1$. This Fourier series will, nevertheless, be rapidly converging. This follows from the fact that $A_m(w, \delta)>0$ on the real axis and is a holomorphic function in some neighborhood of $\mathbb{R}$ (in fact for $|\mathrm{Im}\, w| < \delta$). The reciprocal $A_m^{-1}(w, \delta)$ will also be holomorphic in some such strip and hence,

$$\frac{A_m(w+1, \delta)}{A_m(w, \delta)\{A_m(w+1)+A_m(w)\}}$$

will be holomorphic as well. Since this is also positive in $\mathbb{R}$, its square root will be holomorphic in a strip also, as will $\hat{\psi}(w, \delta))$. The Fourier coefficients c_n of $\hat{\psi}$, therefore, will behave as

$$c_n = 0(e^{-\varepsilon|n|})$$

for some $\varepsilon > 0$, which will depend on δ.

REFERENCES

[1] P. Butzer, *A survey of the Whittaker-Shannon sampling theorem and some of its extensions*, J. Math. Res. Exposition, 3 (1983), pp. 185-212.

[2] I. Daubechies, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), pp. 961-1005.

[3] A. Grossman and J. Morlet, *Decomposition of Hardy functions into squared integrable wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), pp. 723-736.

[4] Y. Meyer, *The Franklin Wavelets*, 1988, preprint.

[5] J. Morlet, G. Arens, I. Fourgeau and D. Giard, *Wave propagation and sampling theory*, Part II Geophysics, 47 (1982), pp. 203-236.

[6] M. Z. Nashed and G. G. Walter, *General sampling theorems for functions in reproducing kernel Hilbert spaces*, Math. Control Signals Systems, 4 (1991), pp. 373-412.

SIAM J. MATH. ANAL.
Vol. 23, No. 4, pp. 1015–1030, July 1992

SOBOLEV CHARACTERIZATION OF SOLUTIONS OF DILATION EQUATIONS*

TIMO EIROLA†

Abstract. This work studies the smoothness of the solutions of dilation equations, which are encountered in the multiresolution analysis and iterative interpolation processes. Sharp limit of the Sobolev exponent of the solution is given as a function of the spectral radius of an associated finite-dimensional positive operator. In addition, tools are given to get good explicit upper and lower bounds for the exponent.

Key words. wavelets, iterative interpolation, positive operators

AMS(MOS) subject classifications. 39B99, 15A48, 42C15

1. Introduction.

In this paper we study the smoothness of the solutions of the dilation equation

$$\phi(x) = \sum_k h_k \phi(2x - k), \tag{1.1}$$

where h_k are given real coefficients such that $\sum_k h_k = 2$. Such equations arise in iterative interpolation processes for producing continuous curves from discrete data (see [1], [2], [5], [6]) and in multiresolution analysis and in the theory of wavelets [3], [8]. In this work we restrict ourselves to the case where only finitely many of the h_k's are nonzero (say those with $|k| \leq N$). Then [4] the condition on $\{h_k\}$ implies that (1.1) has (up to normalisation) at most one L^1 solution and if it exists it has compact support ($\subset [-N, N]$).

We use the Fourier transform:

$$\hat{f}(\omega) := \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{-i\omega x} f(x) \mathrm{d}x$$

and spaces

$$C^\alpha := \left\{ f \in C^n \mid n = \alpha - \sigma \in \mathbf{N},\ 0 \leq \sigma < 1,\ \sup_{x \neq y} \tfrac{|f^{(n)}(x) - f^{(n)}(y)|}{|x-y|^\sigma} < \infty \right\},$$

$$\mathcal{H}^s_p := \{ f \in L^p(\mathbf{R}) \mid \|f\|_{s,p} := \|[\hat{f}(\omega)(1 + |\omega|^2)^{s/2}\,]\hat{}\ \|_{L_p} < \infty\},$$

where α , $s \geq 0$. For the case $p = 2$, which is mostly considered here, set $\mathcal{H}^s := \mathcal{H}^s_2$.

Assume (1.1) has a solution $\phi \in L^1$. Then for the Fourier transform we have

$$\hat{\phi}(\omega) = p(\omega/2)\hat{\phi}(\omega/2), \tag{1.2}$$

where

$$p(\omega) := \frac{1}{2} \sum_k h_k e^{-ik\omega}. \tag{1.3}$$

*Received by the editors March 20, 1991; accepted for publication (in revised form) November 14, 1991.

†Helsinki University of Technology, Institute of Mathematics, SF-02150 Espoo, Finland (E-mail: timo.eirola@hut.fi).

Instead of (1.1) we will mainly study (1.2), which is somewhat simpler in form and has in any case a meaning since $p(0) = 1$ and p is smooth.

In iterative interpolation equation (1.2) (and (1.1)) comes from the following process (see [5], [6]).

Let initial data $u_0 \in l^2$ be given and $h = 1$. Take an interpolation scheme P which produces new values in the middle points by linear combinations:

$$(1.4) \qquad Pu\left(\left(j+\frac{1}{2}\right)h\right) = \sum_k \beta_k u((j-k)h).$$

In pure interpolation P leaves the old values $u(jh)$ unchanged, but we may consider a more general operator that also replaces the old values of u by another set of linear combinations:

$$(1.5) \qquad Pu(jh) = \sum_k \alpha_k u((j-k)h).$$

Thus P maps $l^2_h \to l^2_{h/2}$, where

$$l^2_h = \left\{ u : h\mathbf{Z} \to \mathbf{R} \mid \|u\|^2_h := h\sum_j |u(jh)|^2 < \infty \right\},$$

interpolating new values in the middle of the intervals according to (1.4) and replacing the old values by the linear combinations (1.5). Now $Pu_0 \in l^2_{1/2}$, and we may apply P again with $h = \frac{1}{2}$, i.e., consider $P^2u_0 \in l^2_{1/4}$. Repeating this gives $P^n u_0 \in l^2_{1/2^n}$—a function defined at points of the form $j/2^n$. The question is: what is $\lim_{n\to\infty} P^n u_0$? Since this is a linear process we may assume that $u_0(0) = 1$ and $u_0(j) = 0$ for $j \neq 0$ and get the general case by linear combinations of the translates of this initial data.

For $u \in l^2_h$ we use the cardinal series interpolation (see [10]) C_h to think of u as a (C^∞) function on $\mathbf{R}$:

$$C_h u(x) = \sum_j u(jh)H(x/h - j), \quad \text{where } H(x) := \frac{\sin(\pi x)}{\pi x}.$$

Since $\{H(\cdot - j)\}_{j\in\mathbf{Z}}$ is an orthonormal family, C_h is an isometry $l^2_h \to L^2(\mathbf{R})$. Set $u_n := P^n u_0 \in l^2_{2^{-n}}$ and $\phi_n(x) := C_{2^{-n}} u_n(x)$. Then

$$\begin{aligned}\hat{\phi}_n(\omega) &= 2^{-n}\sum_j u_n(2^{-n}j)e^{-i\omega j/2^n}\hat{H}(2^{-n}\omega)\\ &= \tfrac{1}{2}\textstyle\sum_k \left[\alpha_k e^{-i2k\omega/2^n} + \beta_k e^{-i(2k+1)\omega/2^n}\right] 2^{1-n}\sum_j u_{n-1}(2^{1-n}j)e^{-i\omega j/2^{n-1}}\hat{H}(\frac{\omega}{2^n})\end{aligned}$$

by using (1.4) and (1.5). By induction, this gives

$$(1.6) \qquad \hat{\phi}_n(\omega) = \hat{H}(2^{-n}\omega)\prod_{j=1}^{n} p(2^{-j}\omega),$$

where

$$(1.7) \qquad p(\omega) := \frac{1}{2}\sum_k \left[\alpha_k e^{-i2k\omega} + \beta_k e^{-i(2k+1)\omega}\right],$$

i.e., (1.3) with $h_{2k} = \alpha_k$ and $h_{2k+1} = \beta_k$. From (1.6) we get

$$\hat{\phi}_{n+1}(\omega) = p(\omega/2)\hat{\phi}_n(\omega/2).$$

Thus assuming $\{\hat{\phi}_n\}_{n\in\mathbf{N}}$ converges suitably, the limit will satisfy (1.2) and consequently (1.1). The limit of the process for general u_0 is then

$$\sum_j u_0(j)\phi(x-j).$$

These interpolation processes are used in computer graphics for generating curves and surfaces from discrete data [1], [2], [5], [6]. It also appears in the analysis of the Picard–Lindelöf iteration (waveform relaxation) for numerical solution of large systems of ordinary differential equations (see [9]). There interest lies in the $\mathcal{H}^s$-stability of the process.

Another instance where (1.1) arises is multiresolution analysis (see [3], [8]). Roughly speaking, there we look for a function ϕ such that

$$V_\phi^0 \subset V_\phi^1 := \{\, f(2\,\cdot) \mid f \in V_\phi^0\},$$

where V_ϕ^0 is the subspace (in $L^2(\mathbf{R})$, say) spanned by the translates $\{\phi(\cdot - k)\}_{k\in\mathbf{Z}}$. Then there exist h_k's such that (1.1) holds. It is a beautiful theory that such ϕ's exist, and there are conditions on h_k's which ensure that the base $\{\phi(\cdot - k)\}_{k\in\mathbf{Z}}$ is orthonormal [3], [8]. Then the functions $\psi(\cdot - k)$, where

$$\psi(x) = \sum_k (-1)^k h_{1-k}\phi(2x-k)$$

form an orthonormal basis for the orthogonal complement of V_ϕ^0 in V_ϕ^1. This ψ is called an orthonormal wavelet.

Remark. Numerical computation of the solution of (1.1) is easy (see [4]). We can apply the iteration ϕ_{n+1} = r.h.s. of (1.1) with ϕ_n, starting from an initial function with compact support, e.g., the characteristic function of an interval. If (1.1) has a continuous solution, a more economical way is to first find the values of ϕ at the integer points (of the compact support)—this is a system of linear equations after fixing one of them. Then we simply use (1.1) to get the values of ϕ at points $j + \frac{1}{2}$, after that at points $\frac{j}{2} + \frac{1}{4}$, etc. This way we directly get final values of ϕ.

In §2 we prove an abstract result that gives $s_\phi = \sup\{\, s \mid \phi \in \mathcal{H}^s\}$ as a function of the spectral radius of a finite dimensional positive operator associated to p of (1.3). The results about this operator that will be needed are proved in §3. There we also give explicit bounds for s_ϕ in terms of a couple of pointwise values of p. In §4 and 5 these tools will be applied to the symmetric iterative interpolation and to the orthormal wavelets of Daubechies [3], respectively. Both of these applications improve the known results.

The main point to the favor of the technique proposed here is that it connects the smoothness of ϕ to the spectral radius of a positive matrix, while the approach of [4] connects it to the joint spectral radius of a pair of matrices which is more difficult to estimate.

2. An abstract result for $\phi \in \mathcal{H}^s$. In this section we prove a result that allows us to compute $\sup\{\, s \mid \phi \in \mathcal{H}^s\}$ from the spectral radius of a positive operator (eventually a matrix).

Let us study the existence and smoothness of the solution of (1.1) by first taking $\phi_0(x) = H(x) := \sin \pi x/\pi x$, then iterating

$$\phi_{n+1}(x) = \sum_k h_k \phi_n(2x - k) \tag{2.1}$$

and examining in which $\mathcal{H}^s$ the sequence $\{\phi_n\}$ converges (note that $\phi_n \in \mathcal{H}^s$ for every $s \geq 0$). From (2.1) we get

$$\hat{\phi}_{n+1}(\omega) = p(\omega/2)\hat{\phi}_n(\omega/2)$$

(see (1.3)), and consequently,

$$\hat{\phi}_n(\omega) = \hat{H}(2^{-n}\omega) \prod_{j=1}^{n} p(2^{-j}\omega). \tag{2.2}$$

Remark. Comparing (1.6) and (2.2) shows that we can define the same iteration either locally on discrete data by (1.4), (1.5) or globally on continuous data by (2.1).

We will study the two terms of the right-hand side of

$$\|\phi\|_s^2 \leq C\left(\left\|\hat{\phi}\right\|_{L^2}^2 + \int_{\mathbf{R}} \left|\hat{\phi}(\omega)\right|^2 |\omega|^{2s} \mathrm{d}\omega\right) \tag{2.3}$$

separately. Fix $s \geq 0$ and put $e_n(\omega) := \left|\hat{\phi}_n(\omega) - \hat{\phi}_{n-1}(\omega)\right|^2 |\omega|^{2s}$. We want to see for how large values of s the estimate $\int_{\mathbf{R}} e_n(\omega)\mathrm{d}\omega \leq C\theta^n$ holds with some $\theta < 1$. We have $e_n(-\omega) = e_n(\omega)$ and

$$e_{n+1}(\omega) = \left|p\left(\frac{\omega}{2}\right)\right|^2 \left|\hat{\phi}_n\left(\frac{\omega}{2}\right) - \hat{\phi}_{n-1}\left(\frac{\omega}{2}\right)\right| |\omega|^{2s} = q\left(\frac{\omega}{2}\right) e_n\left(\frac{\omega}{2}\right),$$

where $q(t) := 4^s |p(t)|^2$. Since $\hat{H}(\omega) = 0$ for $|\omega| > \pi$, we have

$$\int_{-\infty}^{\infty} e_n(\omega)\mathrm{d}\omega = \int_{|\omega| \leq 2^n\pi} e_n(\omega)\mathrm{d}\omega.$$

We will study the convergence of these integrals through the operator T_q:

$$T_q u(t) := q\left(\frac{t}{2}\right) u\left(\frac{t}{2}\right) + q\left(\pi - \frac{t}{2}\right) u\left(\pi - \frac{t}{2}\right) \tag{2.4}$$

in the following way.

LEMMA 2.1. *Let $\{e_n\}$ satisfy*

$$\text{(a)} \quad e_n(-\omega) = e_n(\omega),$$

$$\text{(b)} \quad e_{n+1}(\omega) = q\left(\frac{\omega}{2}\right) e_n\left(\frac{\omega}{2}\right),$$

where q is even and 2π-periodic. Then

$$I_n := \int_{|\omega| \leq 2^n\pi} e_n(\omega)\mathrm{d}\omega = \int_0^{\pi} \kappa_n(t)\mathrm{d}t, \tag{2.5}$$

where $\kappa_1(t) = 2[e_1(t) + e_1(t-2\pi)]$ *and* $\kappa_{n+1} = T_q\kappa_n$.

Proof. From (a) the claim is trivial for $n = 1$. Assume that the lemma holds for n. Then

$$I_{n+1} = \int_{|\omega|\le 2^{n+1}\pi} e_{n+1}(\omega)\mathrm{d}\omega = \int_{|\omega|\le 2^n\pi} [e_{n+1}(\omega - 2^n\pi) + e_{n+1}(\omega + 2^n\pi)]\mathrm{d}\omega.$$

Call the last integrand $\tilde{e}_n(\omega)$. It satisfies (a) and (b) so that applying the lemma with n to $\tilde{e}_n$ gives $I_{n+1} = \int_0^\pi \tilde{\kappa}_n(t)\mathrm{d}t$, where

$$\begin{aligned}\tilde{\kappa}_1(t) &= 2[e_2(t-2\pi) + e_2(t+2\pi) + e_2(t-4\pi) + e_2(t)]\\ &= 2\left[q\left(\frac{t}{2}-\pi\right)\left[e_1\left(\pi-\frac{t}{2}\right) + e_1\left(-\pi-\frac{t}{2}\right)\right]\right.\\ &\qquad \left. + q\left(\frac{t}{2}\right)\left[e_1\left(\frac{t}{2}-2\pi\right) + e_1\left(\frac{t}{2}\right)\right]\right]\\ &= q\left(\frac{t}{2}\right)\kappa_1\left(\frac{t}{2}\right) + q\left(\pi-\frac{t}{2}\right)\kappa_1\left(\pi-\frac{t}{2}\right) = T_q\kappa_1(t),\end{aligned} \tag{2.6}$$

i.e., $I_{n+1} = \int_0^\pi T_q^{n-1}\tilde{\kappa}_1(t)\mathrm{d}t = \int_0^\pi T_q^n\kappa_1(t)\mathrm{d}t$. □

The key idea in studying the integrals $\int_{\mathbf{R}} e_n(\omega)\mathrm{d}\omega$ in terms of $\int_0^\pi \kappa_n(t)\mathrm{d}t$ lies in the fact that the positive operator T_q has better properties (see the next section) than the one that maps e_n to e_{n+1}. The best property here is that if q is in C_d—the $d+1$-dimensional space spanned by $\{\cos(jt)\}_{j=0}^d$—then T_q maps $C_d \to C_d$. Thus we end up with studying the Frobenius–Perron eigenvalue of a positive operator in a finite-dimensional space.

In the next section we will prove (Theorem 3.4 and Proposition 3.5) the following. Assume

(2.7)
(1) $q(t) = 4^s(2 + 2\cos(t))^\nu r(t)$ with r positive on $[0,\pi]$.
(2) u_0 is nonnegative and not identically zero on $[0,\pi]$ with $u_0 = O(t^{2\tilde{\nu}})$ for small $|t|$ for some $\tilde{\nu} \le \nu$.

Set $u_n := T_q^n u_0$. Then for any $\varepsilon > 0$ there exist $C, c > 0$ such that

$$c(4^s\mu - \varepsilon)^n \le \int_0^\pi u_n(t)\mathrm{d}t \le C(4^{\nu-\tilde{\nu}+s}\mu + \varepsilon)^n,$$

where μ is the spectral radius of T_r.

We apply this to the I_n's twice: for $s = 0$ and $s > 0$, corresponding to the two terms of (2.3). Naturally, we want $4^{\nu-\tilde{\nu}+s}\mu < 1$.

Since $\hat{H} = (1/\sqrt{2\pi})\chi_{[-\pi,\pi]}$ we find for $t \in [0,\pi]$,

$$\kappa_1(t) = 2[e_1(t) + e_1(t-2\pi)] = \tfrac{1}{\pi}\left[|p(\tfrac{t}{2}) - 1|^2\,|t|^{2s} + |p(\tfrac{t}{2}-\pi)|^2\,|t-2\pi|^{2s}\right].$$

Rewriting the assumptions (2.7) in terms of p we get the lower bound for the following.

THEOREM 2.2. *Assume $p(t) \ne 0$ for $t \ne \pi$. Let ν_0, ν_1 be the greatest numbers such that for $t \to 0$,*

$$p(t) = 1 + O(t^{\nu_0}), \qquad p(\pi - t) = O(t^{\nu_1}),$$

where

$$p(t) := \frac{1}{2}\sum_k h_k e^{-ikt}.$$

Let μ be the greatest eigenvalue of the positive operator T_r in $C[0,\pi]$:

$$T_r u(t) = r\left(\frac{t}{2}\right) u\left(\frac{t}{2}\right) + r\left(\pi - \frac{t}{2}\right) u\left(\pi - \frac{t}{2}\right),$$

where

$$r(t) = \frac{|p(t)|^2}{(2+2\cos(t))^{\nu_1}}.$$

If $\mu < 4^{\nu_0 - \nu_1}$, then ϕ—the (normalized) solution of the dilation equation (1.1)—satisfies

$$s_\phi := \sup\{\, s \mid \phi \in \mathcal{H}^s\} = \frac{-\log(\mu)}{\log(4)}.$$

Proof. To show the upper bound $s_\phi \le -\log(\mu)/\log(4)$, assume $\phi \ne 0$—a solution of (1.1)—is in $\mathcal{H}^s$. We have

$$\infty > \int_{-\infty}^{\infty} \left|\hat{\phi}(\omega)\right|^2 |\omega|^{2s}\, \mathrm{d}\omega \ge \int_{|\omega|\le 2^n\pi} \left|\hat{\phi}(\omega)\right|^2 |\omega|^{2s}\, \mathrm{d}\omega.$$

Denote the last integrand by $f_n(\omega)$ (independent of n). We are now in the situation of Lemma 2.1: since (1.1) implies $\hat{\phi}(\omega) = p(\omega/2)\hat{\phi}(\omega/2)$ giving further

$$f_{n+1}(\omega) = \left|p\left(\frac{\omega}{2}\right)\right|^2 \left|\hat{\phi}\left(\frac{\omega}{2}\right)\right|^2 |\omega|^{2s} = q\left(\frac{\omega}{2}\right) f_n\left(\frac{\omega}{2}\right).$$

Thus

$$\int_{|\omega|\le 2^n\pi} f_n(\omega)\mathrm{d}\omega = \int_0^{\pi} T^{n-1}\tilde{\kappa}(t)\mathrm{d}t,$$

where

$$\tilde{\kappa}(t) = 2\left[\left|\hat{\phi}(t)\right|^2 |t|^{2s} + \left|\hat{\phi}(t-2\pi)\right|^2 |t-2\pi|^{2s}\right].$$

By Proposition 3.5,

$$\int_{|\omega|\le 2^n\pi} f_n(\omega)\mathrm{d}\omega \ge c(4^s\mu - \varepsilon)^n.$$

Thus $4^s\mu \le 1$. □

Remark. Any question about weakening the assumptions of this theorem is directly forwarded to §3—can we prove something similar about T_r with weaker properties of r? The same is true if we look for $\mathcal{H}_1^s$-convergence. Also, then we come to similar situations with weaker r. Only in the case when p is already a square (a nonnegative polynomial of $\cos(t)$) can we study the $\mathcal{H}_1^s$-convergence directly by these methods (see §4). On the other hand, generalizations to other two-scale equations of type (1.1) (i.e., 2 replaced by a bigger integer) seem reachable by these techniques.

3. Some properties of T_r. Recall the definition of $T_r : C[0,\pi] \to C[0,\pi]$

$$T_r u(t) = r\left(\frac{t}{2}\right) u\left(\frac{t}{2}\right) + r\left(\pi - \frac{t}{2}\right) u\left(\pi - \frac{t}{2}\right). \tag{3.1}$$

First we prove the main convergence result of iterations with T_r when r is a positive polynomial of $\cos(t)$. Then some results for perturbed r are given. Finally, tools for getting upper and lower bounds for the spectral radius of T_r are derived. We will often write T in place of T_r.

Most of the tools used for the next result are standard in the literature on positive operators, but a suitable theorem for the present purpose was not found.

THEOREM 3.1. *Let r be a d-degree polynomial of $\cos(t)$ positive on $[0,\pi]$. Then for any nonnegative nonzero $u_0 \in C^1[0,\pi]$, the sequence $\{T_r^n u_0/\,\|T_r^n u_0\|\}$ converges to the unique positive norm-1 eigenvector of T_r. This eigenvector is a d-degree polynomial of $\cos(t)$ and corresponds to the greatest eigenvalue, which is equal to the spectral radius of T_r.*

Proof. We show first the existence of a positive eigenvector. Let C_d be the $(d+1)$-dimensional space spanned by $\{\cos(jt)\}_{j=0}^{d}$. If

$$r(t) = \sum_{j=0}^{d} \rho_j \cos(jt), \qquad u(t) = \sum_{j=0}^{d} u_j \cos(jt),$$

then

$$\begin{aligned} Tu(2t) &= \sum_{j,k=0}^{d} \rho_j u_k \left[\cos(jt)\cos(kt) + \cos(j(\pi-t))\cos(k(\pi-t))\right] \\ &= \sum_{\substack{j,k=0 \\ j+k=\text{even}}}^{d} \rho_j u_k \left[\cos((j+k)t) + \cos((j-k)t)\right], \end{aligned}$$

i.e., T maps C_d into itself. Let K, respectively, K^o be the cone of nonnegative, respectively, positive functions in $C[0,\pi]$ and K_d, K_d^o their intersections with C_d. All of these are T-invariant.

LEMMA 3.2. *If $0 \neq u \in K$, then there exists $n > 0$ such that $T^n u \in K^o$.*

Proof. u is positive at some point $t = j\pi/2^m$, $0 \leq j \leq 2^m$.

If $j < 2^{m-1}$, then Tu is positive at $j\pi/2^{m-1}$.

If $j \geq 2^{m-1}$, then Tu is positive at $(2 - j/2^{m-1})\pi$.

In either case Tu is positive at some $j\pi/2^{m-1}$. Iterating this gives that $T^m u$ is positive at zero or π. Thus $T^{m+1}u$ is positive at zero. It is also positive on some interval $[0, 2^{-m'}]$. Then $T^{m'}T^{m+1}u$ is positive on $[0,\pi]$. □

Using compactness gives further in C_d that there exists an M such that $T^M K_d \subset K_d^o$. Now the standard theory of positive operators in finite dimensions states that there exists a unique positive norm-1 eigenvector $\tilde{u}$ of T in C_d. Denote by μ the corresponding eigenvalue.

Return now to $C[0,\pi]$ and take $0 \neq u_0 \in K \cap C^1[0,\pi]$. Take n_0 such that $T^{n_0}u_0 \in K^o$. Since K is normal, a theorem of Stetsenko [7, Thm. 9.1] states that

$$\lim_{n\to\infty} \|T^n(T^{n_0}u_0)\|^{1/n} = \rho_T,$$

where ρ_T is the spectral radius of T. Since u_0 might as well have been equal to $\tilde{u}$ we have $\rho_T = \mu$. Take the sequence $\{v_n\} = \{T^n u_0/\,\|T^n u_0\|\}$. We know that $v_n \in K^o$ for $n \geq n_0$ and $\lim_{n\to\infty} \|Tv_n\| = \mu$. Set

$$\beta_n = \frac{\min_t v_n(t)/\tilde{u}(t)}{\max_\tau v_n(\tau)/\tilde{u}(\tau)}.$$

Then

$$\begin{aligned}\beta_{n+1} &= \frac{\min_t \left[r(\tfrac{t}{2})v_n(\tfrac{t}{2}) + r(\pi - \tfrac{t}{2})v_n(\pi - \tfrac{t}{2})\right]/\tilde{u}(t)}{\max_\tau \left[r(\tfrac{\tau}{2})v_n(\tfrac{\tau}{2}) + r(\pi - \tfrac{\tau}{2})v_n(\pi - \tfrac{\tau}{2})\right]/\tilde{u}(\tau)}\\ &\geq \frac{\min_t v_n(t)/\tilde{u}(t)\min_t T\tilde{u}(t)/\tilde{u}(t)}{\max_\tau v_n(\tau)/\tilde{u}(\tau)\max_\tau T\tilde{u}(\tau)/\tilde{u}(\tau)} = \beta_n.\end{aligned}$$

Thus $\{\beta_n\}$ is nondecreasing, and for $n \geq n_0$,

$$\frac{\min_t v_n(t)}{\max_\tau v_n(\tau)} \geq \beta_n \frac{\min_t \tilde{u}(t)}{\max_\tau \tilde{u}(\tau)} \geq c > 0.$$

LEMMA 3.3. *The sequence $\{v_n\}$ is bounded in $C^1[0,\pi]$.*

Proof. Set $R = \max_t |\dot{r}(t)/r(t)|$ and $a = \min_t(r(t)/r(\pi - t))$. For $n \geq n_0$,

$$\begin{aligned}\left\|\frac{\dot{v}_{n+1}}{v_{n+1}}\right\| &= \frac{1}{2}\max_t \left|\frac{\dot{r}(\tfrac{t}{2})v_n(\tfrac{t}{2}) - \dot{r}(\pi - \tfrac{t}{2})v_n(\pi - \tfrac{t}{2})}{r(\tfrac{t}{2})v_n(\tfrac{t}{2}) + r(\pi - \tfrac{t}{2})v_n(\pi - \tfrac{t}{2})}\right.\\ &\qquad\left. + \frac{r(\tfrac{t}{2})\dot{v}_n(\tfrac{t}{2}) - r(\pi - \tfrac{t}{2})\dot{v}_n(\pi - \tfrac{t}{2})}{r(\tfrac{t}{2})v_n(\tfrac{t}{2}) + r(\pi - \tfrac{t}{2})v_n(\pi - \tfrac{t}{2})}\right|\\ &\leq R + \frac{1}{1+ac}\left\|\frac{\dot{v}_n}{v_n}\right\|.\end{aligned}$$

Thus $\|\dot{v}_n/v_n\|$ and consequently $\|\dot{v}_n\|$ is bounded. □

By Lemma 3.3 there exists a convergent subsequence $\{v_{n_j}\} \to v \in K^o$ (with increasing n_j's). Write $w \succeq u$ if $w - u \in K$. Define for $u \in K^o$,

$$\varphi(u) = \max\{\lambda \mid Tu \succeq \lambda u\}. \tag{3.2}$$

φ is clearly continuous in K^o and $\varphi(Tu) \geq \varphi(u)$. We have $\varphi(u) \leq \mu$ since assuming the contrary, $u, \varepsilon > 0$ such that $Tu \succeq (\mu + \varepsilon)u$ gives

$$T^n u \succeq (\mu + \varepsilon)T^{n-1}u \succeq (\mu + \varepsilon)^n u,$$

implying $\rho_T > \mu$. Similarly,

$$u \in K^o,\ n \geq 0 \quad \Rightarrow \quad T^n u - \mu^n u \notin K^o. \tag{3.3}$$

Assume now that $Tv \neq \varphi(v)v$. Then there exists an m such that $T^m(Tv - \varphi(v)v) \in K^o$, i.e., $\varphi(T^m v) > \varphi(v)$. On the other hand,

$$\varphi(T^m v) = \lim_j \varphi(T^m v_{n_j}) \leq \lim_j \varphi(v_{n_{j+m}}) = \varphi(v).$$

Thus $Tv = \varphi(v)v$ and $\varphi(v) = \mu$.

If $v \neq \tilde{u}$, then either

(1) $v - \tilde{u} \in K \cup (-K)$ or

(2) for $w(t) := \max(\tilde{u}(t), v(t))$ holds $w \neq \tilde{u}$ and $w \neq v$.

In case (1), assume, e.g., $v - \tilde{u} \in K$. Since $\tilde{u}(t') = 1$ for some t' and

$$\mu^m(v - \tilde{u})(t') = T^m(v - \tilde{u})(t') > 0$$

for m big enough we have $v(t') > 1$, a contradiction.

In case (2), there exist m and $\varepsilon > 0$ such that

$$T^m(w - v) \succeq \varepsilon w \quad \text{and} \quad T^m(w - \tilde{u}) \succeq \varepsilon w,$$

i.e., $T^m w \succeq (\mu^m + \varepsilon)w$, a contradiction to (3.3). Thus $v = \tilde{u}$. It follows that $\tilde{u}$ is the only limit point of $\{v_n\}$. By compactness (Lemma 3.3), $\{v_n\} \to \tilde{u}$. This ends the proof of Theorem 3.1. □

Next we turn to the case where we have a zero at π:

$$r_\nu(t) = (2 + 2\cos(t))^\nu r(t).$$

THEOREM 3.4. *Let r be as in Theorem 3.1 and r_ν as above with $\nu \geq 0$. If $0 \neq u_0 \in C^1[0,\pi]$ is nonnegative and satisfies $u_0(t) = O(t^{2\nu})$, then the sequence $\{T^n_{r_\nu}u_0 / \|T^n_{r_\nu}u_0\|\}$ converges to*

$$\tilde{u}^\nu(t) = c(1 - \cos(t))^\nu \tilde{u}(t),$$

where $\tilde{u}$ is the positive eigenvector of T_r and c is such that $\|\tilde{u}^\nu\| = 1$. Moreover, $T^{n+1}_{r_\nu}u_0(t)/T^n_{r_\nu}u_0(t) \to \mu$ uniformly.

Proof. Set $u_n = T^n_{r_\nu}u_0$ and $v_n(t) := (1 - \cos(t))^{-\nu}u_n(t)$. Then $v_0 \in C^1[0,\pi]$ and

$$\begin{aligned} v_{n+1}(t) &= (1 - \cos(t))^{-\nu}\left[r_\nu\left(\frac{t}{2}\right)u_n\left(\frac{t}{2}\right) + r_\nu\left(\pi - \frac{t}{2}\right)u_n\left(\pi - \frac{t}{2}\right)\right] \\ &= \left[\frac{2(1 + \cos(t/2))(1 - \cos(t/2))}{(1 - \cos(t))}\right]^\nu \left[r\left(\frac{t}{2}\right)v_n\left(\frac{t}{2}\right) + r\left(\pi - \frac{t}{2}\right)v_n\left(\pi - \frac{t}{2}\right)\right], \end{aligned}$$

i.e., $v_{n+1} = T_r v_n$. Apply Theorem 3.1. □

For the case where u_0 does not allow full division by $(1 - \cos(t))^\nu$, the following will suffice for our present needs.

PROPOSITION 3.5. *Let $0 \neq u_0 \succeq 0$, $u_n = T^n_{r_\nu}u_0$ and μ , $\tilde{u}$ as before but $u_0(t) = O(t^{2\nu_0})$ with $\gamma := \nu - \nu_0 \geq 0$. Then for given $\varepsilon > 0$ there exist $C, c > 0$ such that*

$$c(\mu - \varepsilon)^n \leq \int_0^\pi u_n(t)\mathrm{d}t \leq C(4^\gamma\mu + \varepsilon)^n.$$

Proof. For the lower bound, take $v_0 \neq 0$ such that $u_0 \succeq v_0 \succeq 0$, and $v(t) = O(t^{2\nu})$. Then $T^n_{r_\nu}(u_0 - v_0) \succeq 0$. For given $\varepsilon > 0$ take (Theorem 3.4) n_0 such that for $n \geq n_0$ holds

$$\left|\frac{T^{n+1}_{r_\nu}v_0(t)}{T^n_{r_\nu}v_0(t)} - \mu\right| < \varepsilon$$

and $c > 0$ such that the last inequality below holds for $n \leq n_0$. Then

$$\int_0^\pi u_n(t)\mathrm{d}t \geq \int_0^\pi T^n_{r_\nu}v_0(t)\mathrm{d}t \geq c(\mu - \varepsilon)^n.$$

For the upper bound set $v_0(t) := (1 - \cos(t))^{-\nu_0}u_0(t)$ and $\hat{r} := 4^\gamma r$. Then $\hat{r} \succeq r_\gamma$ and $T_{\hat{r}}v \succeq T_{r_\gamma}v$ for every $v \succeq 0$. Thus, inductively, $T_{\hat{r}}v_0 \succeq T_{r_\gamma}v_0$, $T_{\hat{r}}(T^n_{\hat{r}}v_0) \succeq T_{\hat{r}}(T^n_{r_\gamma}v_0) \succeq T^{n+1}_{r_\gamma}v_0$, and

$$\begin{aligned} \int_0^\pi u_n(t)\mathrm{d}t &= \int_0^\pi T^n_{r_\nu}u_0(t)\mathrm{d}t = \int_0^\pi (1 - \cos(t))^{\nu_0}T^n_{r_\gamma}v_0(t)\mathrm{d}t \\ &\leq 2^{\nu_0}4^{n\gamma}\int_0^\pi T^n_r v_0(t)\mathrm{d}t \leq C(4^\gamma\mu + \varepsilon)^n. \end{aligned}$$

□

Remark. Thus far the results of this section have used the condition that r is a polynomial of $\cos(t)$ only for proving the existence of a positive eigenvector. If this is known by some other means, then any positive $r \in C^1[0,\pi]$ will do. Then, naturally, $\tilde{u}$ is not a polynomial of $\cos(t)$.

Now we make more heavy use of the fact that r is a polynomial of $\cos(t)$and turn to estimate μ. For $\theta \in [0,\pi]$, set $\eta_\theta \in K'$ (the dual cone of K): $\langle \eta_\theta, u\rangle := u(\theta)$. Then

$$T'\eta_\theta = r\left(\frac{\theta}{2}\right)\eta_{\frac{\theta}{2}} + r\left(\pi - \frac{\theta}{2}\right)\eta_{\pi-\frac{\theta}{2}}. \tag{3.4}$$

In C_d we have a better cone than K_d, namely:

$$K^d := \left\{ u \in C_d \mid u(t) = \sum_{j=0}^{d} u_j(1+\cos(t))^j(1-\cos(t))^{d-j},\ u_j \geq 0 \right\}.$$

This is smaller than K_d since $0 \neq u \in K^d$ implies $u(t) > 0$ for $t \in (0,\pi)$. It follows that the η_θ's with $\theta \in (0,\pi)$ are interior points of $(K^d)'$. We also have

$$\begin{aligned} &\text{(a)} \quad u \in K^n,\ v \in K^m \ \Rightarrow\ uv \in K^{n+m},\\ &\text{(b)} \quad n \leq m \ \Rightarrow\ K^n \subset K^m \ \text{ and } \ K^n + K^m \subset K^m. \end{aligned}$$

Property (a) is simple. From $\frac{1}{2}(1-x) + \frac{1}{2}(1+x) = 1$ we get $1 \in K^1$. Then the repeated use of (a) gives $1 \in K^n$ for all $n \geq 0$. Given $u \in K^n$, take $1 \in K^{m-n}$ and use (a) to show (b).

Naturally we also need the following.

LEMMA 3.6. *If $r \in K^d$, then $TK^d \subset K^d$.*

Proof. Set $y := \cos(t/2)$, $x := \cos(t) = 2y^2 - 1$, and let $r, u \in K^d$: $r(t) = \sum_{j=0}^{d} r_j(1+x)^j(1-x)^{d-j}$, $u(t) = \sum_{j=0}^{d} u_j(1+x)^j(1-x)^{d-j}$. Then

$$Tu(t) = \sum_{i,j=0}^{d} r_i u_j[(1+y)^{i+j}(1-y)^{2d-i-j} + (1+y)^{2d-i-j}(1-y)^{i+j}], \tag{3.5}$$

and the lemma follows after noting that for $k \leq d$,

$$\begin{aligned} &(1+y)^k(1-y)^{2d-k} + (1+y)^{2d-k}(1-y)^k \\ &\quad = (1-y^2)^k[(1+y)^{2(d-k)} + (1-y)^{2(d-k)}] \\ &\quad = \left(\tfrac{1+x}{2}\right)^k 2\textstyle\sum_{l=0}^{d-k}\binom{2d-2k}{2l}\left(\tfrac{1-x}{2}\right)^l \in K^d. \end{aligned}$$ □

From the previous proof we also see that $\tilde{u}$ is an interior point of K^d because $\tilde{u}(0)r(0) > 0 \Rightarrow \tilde{u}_0 r_0 > 0$, and the first term of (3.5) for $\tilde{u}$ gives $\tilde{u}_0 r_0 2\sum_{l=0}^{d}\binom{2d}{2l}((1-x)/2)^l$, thus introducing a positive coefficient (after multiplication with $1 \in K^{d-l}$) for each $(1-x)^{d-l}(1+x)^l$. Thus $0 \neq w \in (K^d)'$ implies $\langle w, \tilde{u}\rangle > 0$.

We base the estimates for μ on the following.

LEMMA 3.7. *If $r \in K^d$ and $0 \neq w \in (K^d)', \lambda \in \mathbf{R}$, then*

$$\begin{aligned} &\text{(a)} \quad T'w - \lambda w \in (K^d)' \ \Rightarrow\ \mu \geq \lambda,\\ &\text{(b)} \quad \lambda w - T'w \in (K^d)' \ \Rightarrow\ \mu \leq \lambda. \end{aligned}$$

Proof. (a) $0 \leq \langle T'w - \lambda w, \tilde{u}\rangle = (\mu - \lambda)\langle w, \tilde{u}\rangle$ and $\langle w, \tilde{u}\rangle > 0$. (b) is similar. □

From (3.4) we get

$$T'\eta_{2\pi/3} = r\left(\frac{2\pi}{3}\right)\eta_{2\pi/3} + r\left(\frac{\pi}{3}\right)\eta_{\pi/3}.$$

In the applications of this paper we have $r(2\pi/3) >> r(\pi/3)$ which suggests that $\eta_{2\pi/3}$ is close to the positive eigenvector of T'. Now Lemma 3.7 gives immediately $\mu \geq r(2\pi/3)$.

Note that for $u \in K^d$ and $t \in (0, \pi/2)$ we have:

$$\begin{aligned} u(t) &\leq \sum_{j=0}^{d} u_j \left(\frac{1+\cos(t)}{1+\cos(2t)}\right)^j (1+\cos(2t))^j (1-\cos(2t))^{d-j} \\ &\leq \left(\frac{1+\cos(t)}{1+\cos(2t)}\right)^d u(2t), \end{aligned} \tag{3.6}$$

and similarly $u(\pi - t) \leq ((1+\cos(t))/(1+\cos(2t)))^d u(\pi - 2t)$,

To obtain an upper bound for μ, set $w = \eta_{2\pi/3} + \alpha\eta_{\pi/3}$. Then

$$\lambda w - Tw = (\lambda - r(2\pi/3))\eta_{2\pi/3} + (\alpha\lambda - r(\pi/3))\eta_{\pi/3} - \alpha\left[r(\pi/6)\eta_{\pi/6} + r(5\pi/6)\eta_{5\pi/6}\right].$$

From (3.6) we get for any $u \in K^d$: $u(\pi/6) \leq su(\pi/3)$ and $u(5\pi/6) \leq su(2\pi/3)$, where $s = \left((2+\sqrt{3})/3\right)^d$. Thus $\lambda w - T'w \in (K^d)'$ if

$$\begin{aligned} \lambda - r\left(\frac{2\pi}{3}\right) - \alpha\ s\ r\left(\frac{5\pi}{6}\right) &=0, \\ \alpha\ \lambda - r\left(\frac{\pi}{3}\right) - \alpha\ s\ r\left(\frac{\pi}{6}\right) &=0. \end{aligned}$$

Solving these gives the following.

THEOREM 3.8. *If* $r \in K^d$ *satisfies* $r(2\pi/3) > \left((2+\sqrt{3})/3\right)^d r(\pi/6)$, *then*

$$r\left(\frac{2\pi}{3}\right) \leq \mu \leq r\left(\frac{2\pi}{3}\right)\left(1 + \frac{r(\pi/3)r(5\pi/6)}{\left(3/(2+\sqrt{3})\right)^d r(2\pi/3)^2 - r(\pi/6)r(2\pi/3)}\right).$$

4. Iterative interpolation.

Here we apply the results of §§2 and 3 to the repeated interpolation scheme (1.4) $P : l_h^2 \to l_{h/2}^2$ in the case where P is "identity on the old grid":

$$\begin{aligned} Pu(jh) &= u(jh), \\ Pu\left(\left(j+\frac{1}{2}\right)h\right) &= \sum_k \beta_k u((j-k)h). \end{aligned} \tag{4.1}$$

We restrict ourselves to odd $(2m-1)$-degree symmetric polynomial interpolation, i.e., with $\{\beta_k\}_{k=-m}^{m-1}$ such that if x is a $(2m-1)$-degree polynomial, then

$$\sum_{k=-m}^{m-1} \beta_k x(-k) = x\left(\frac{1}{2}\right). \tag{4.2}$$

The question here is: how smooth is $\lim_{n\to\infty} P^n u_0$? As noted in §1, we may take $u_0(k) = \delta_{0,k}$. From symmetry we get $\beta_{-k} = \beta_{k-1}$, and further from (1.7),

$$p(t) = \frac{1}{2}\left(1 + \sum_{j=0}^{m-1} \beta_j [e^{-it(2j+1)} + e^{it(2j+1)}]\right) = \frac{1}{2}\left(1 + \sum_{j=-m}^{m-1} \beta_j \cos((2j+1)t)\right).$$

The last sum is the interpolation of $\cos(\tau)$ at $\tau = 0$; thus it has to be $1 + O(t^{2m})$. Furthermore, repeated use of

$$\cos(nt) = 2\cos(t)\cos((n-1)t) - \cos((n-2)t)$$

gives

$$p(t) = \frac{1}{2}\left(1 + \cos(t)\sum_{j=0}^{m-1} \alpha_j \sin(t)^{2j}\right). \tag{4.2}$$

On the other hand, a standard integration formula gives

$$\cos(t)\sum_{j=0}^{m-1}(-1)^j\binom{-\frac{1}{2}}{j}\sin(t)^{2j} = 1 - 2c_m\int_0^t \sin(\tau)^{2m-1}\mathrm{d}\tau = 1 + O(t^{2m}), \tag{4.3}$$

where $c_m = 2^{1-2m}((2m-1)!/((m-1)!)^2)$. Since the polynomials of $\sin(t)$ multiplying $\cos(t)$ in (4.2) and (4.3) are of degree $2m-2$ and have $(2m-1)$-fold tangency at $t = 0$, they have to be equal. Hence

$$p(t) = 1 - c_m\int_0^t \sin(\tau)^{2m-1}\mathrm{d}\tau. \tag{4.4}$$

To apply the results of §2, it remains to divide the zero of p at π. Integration by parts yields

$$\begin{aligned} p(t) &= \frac{c_m}{m}(1+\cos(t))^m\sum_{j=0}^{m-1}\gamma_j^m(1+\cos(t))^j(1-\cos(t))^{m-1-j} \\ &= \left(\tfrac{1+\cos(t)}{2}\right)^m\textstyle\sum_{j=0}^{m-1}\binom{m-1+j}{j}\left(\tfrac{1-\cos(t)}{2}\right)^j, \end{aligned}$$

where $\gamma_j^m := \prod_{k=0}^{j}(m-k)/(m+k)$. Thus

$$\begin{aligned} r(t) &= (2+2\cos(t))^{-2m}\,|p(t)|^2 \\ &= 2^{-4m}\left(\textstyle\sum_{j=0}^{m-1}\binom{m-1+j}{j}\left(\tfrac{1-\cos(t)}{2}\right)^j\right)^2 \\ &= 2^{-2m}\frac{c_m^2}{m^2}\left[\textstyle\sum_{j=0}^{m-1}\gamma_j^m(1+\cos(t))^j(1-\cos(t))^{m-1-j}\right]^2 \\ &= 2^{-2m}\frac{c_m^2}{m^2}(1-\cos(t))^{2m-2}\left[\sum_{j=0}^{m-1}\gamma_j^m\left(\tfrac{1+\cos(t)}{1-\cos(t)}\right)^j\right]^2. \end{aligned} \tag{4.5}$$

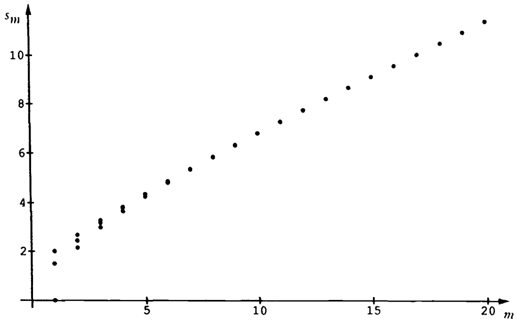


FIG. 1. *The Sobolev exponents s_m with the upper and lower bounds* (4.8).

We use the first form of (4.5) to estimate

$$r(t) \le (2+2\cos(t))^{-2m} \quad \text{for } t \in \left[0, \frac{\pi}{2}\right), \tag{4.6}$$

the third one to see that $r \in K^{2m-2}$, and the fourth to estimate for $t \in (\pi/2, \pi]$:

$$1 \le \frac{r(t)}{2^{-2m}\frac{c_m^2}{m^2}(1-\cos(t))^{2m-2}} \le \left(\frac{1-\cos(t)}{2\cos(t)}\right)^2. \tag{4.7}$$

The form of the second line of (4.5) will be used later in §5. Now $d = 2m-2$ and Theorems 2.1 and 3.8 give the following.

THEOREM 4.1. *For fixed $m \ge 1$ the iteration of $2m-1$-degree symmetric polynomial interpolation converges to a function ϕ_m for which $s_m := \sup\{s \mid \phi_m \in \mathcal{H}^s\} = -\log(\mu)/(2\log(2))$, where μ is the spectral radius of T_r in C_{2m-2}. Further, s_m satisfies* (4.8)

$$\frac{-\log\left(r(\frac{2\pi}{3})\left(1+\frac{r(\pi/3)r(5\pi/6)}{(3/(2+\sqrt{3}))^{2m-2}r(2\pi/3)^2-r(\pi/6)r(2\pi/3)}\right)\right)}{2\log(2)} \le s_m \le \frac{-\log(r(\frac{2\pi}{3}))}{2\log(2)},$$

and $r = r_m$ is given in (4.5).

To obtain asymptotics of s_m, note that $j > \sqrt{(j+1)(j-1)}$ gives

$$\frac{1}{2}\sqrt{m-\frac{1}{2}} \le c_m \le \frac{1}{\sqrt{2}}\sqrt{m-\frac{1}{2}},$$

and this together with (4.7) implies

$$\frac{1}{9}\frac{m-\frac{1}{2}}{m^2}\left(\frac{3}{4}\right)^{2m} \le r\left(\frac{2\pi}{3}\right) \le \frac{1}{2}\frac{m-\frac{1}{2}}{m^2}\left(\frac{3}{4}\right)^{2m}. \tag{4.9}$$

Furthermore, (4.6) and (4.7) give

$$\frac{r(\frac{\pi}{3})r(\frac{5\pi}{6})}{(3/(2+\sqrt{3}))^{2m-2}r(\frac{2\pi}{3})^2 - r(\frac{\pi}{6})r(\frac{2\pi}{3})} \le \frac{\frac{2m^2}{3(m-\frac{1}{2})}\left[4(7+4\sqrt{3})/81\right]^{2m-1}}{1-81m^2\left(\frac{2}{3}\right)^{4m}/((7+4\sqrt{3})(m-\frac{1}{2}))}$$

which tends to zero when $m \to \infty$; hence $\mu_m/(r_m(2\pi/3)) \to 1$. Thus, asymptotically,

$$s_m = \left(2 - \frac{\log(3)}{\log(2)}\right)m + O(\log(m)) \approx 0.415m.$$

For small values of m we can compute the eigenvalue of T_r directly giving, e.g., $s_1 = \frac{3}{2}$, $s_2 = \frac{9}{2} - (\log(5+3\sqrt{17})/2\log(2))$. For Fig. 1 we have computed the maximal eigenvalues of T_r numerically.

Cases $m = 2$ and $m = 3$ were considered also by Deslauriers and Dubuc [5]. They got $\phi_2 \in C^{2-\varepsilon}$ and $\phi_3 \in C^{2.830}$. The case where $m = 2$ is also treated in [4].

As pointed out in the Remark after Theorem 2.1, the $\mathcal{H}_1^s$-convergence can be treated by the present method (and, consequently, the C^s-convergence) when $p(t)$ is a square, which is the case in the present situation. Everything works similarly with $q(t)$ replaced by $2^s p_m(t)$ and $r(t)$ by $\rho_m(t) := (2 + 2\cos(t))^{-m} p_m(t)$ to give

$$\sigma_m := \sup\{\, s \mid \phi_m \in \mathcal{H}_1^s \} = -\log(\mu_m)/\log(2),$$

where μ_m is the spectral radius of T_{ρ_m}. Thus we get the σ_m's in this case for free.

In Table 1 we have the s_m's and σ_m's for different values of m. Note that the values of Deslauriers and Dubuc are sharp.

TABLE 1

m	s_m	σ_m
1	1.5	1.
2	2.4407	2.
3	3.1751	2.8300
4	3.7931	3.5511
5	4.3440	4.1935
6	4.8620	4.7767
7	5.3628	5.3173
8	5.8529	5.8294
9	6.3352	6.3233
10	6.8114	6.8054
11	7.2826	7.2796
12	7.7495	7.7480
13	8.2128	8.2121
14	8.6730	8.6726
15	9.1304	9.1302
16	9.5854	9.5853
17	10.038	10.038
18	10.489	10.489
19	10.938	10.938
20	11.386	11.386

An interesting property of these numbers is that $|s_m - \sigma_m|$ seems to tend to zero. This is also easy to prove from the estimates for the spectral radii. This leads us to expect that for big m the limits might be also in $\mathcal{H}_p^s$ with essentially the same s, but larger p.

5. The orthonormal wavelets of Daubechies. Here we consider solutions of (1.1) with h_k's satisfying a further condition

$$\sum_k h_{k-2i} h_{k-2j} = 2\delta_{i,j},$$

which gives orthogonality for $\phi(\cdot - k)$'s (see [3]). We skip the details and take directly the family of h_k's indexed by $m \geq 1$, which was proposed in [3]. They can be defined from

$$|p_m(t)|^2 = 1 - c_m \int_0^t \sin(\tau)^{2m-1} d\tau. \tag{5.1}$$

Meyer's book [8] contains a proof that the smoothness of ϕ grows linearly with m. Daubechies [3] shows that the growth factor is ≥ 0.1936. Here we give it exactly.

Comparing (5.1) and (4.4) shows that the p_m's here are exactly the square roots of the p_m's of §4 (note that our p equals m_0 of [3]). So it remains just to write it down, as follows.

THEOREM 5.1. *With* (5.1), *the solution* ϕ_m *of* (1.1) *satisfies*

$$s_m := \sup\{\, s \mid \phi_m \in \mathcal{H}^s \} = \frac{-\log(\mu)}{\log(4)},$$

where μ *is the spectral radius of* T_r *in the space of* $(m-1)$*-degree polynomials of* $\cos(t)$ *and*

$$r_m(t) = 2^{-2m} \textstyle\sum_{j=0}^{m-1} \binom{m-1+j}{j} \left(\frac{1-\cos(t)}{2}\right)^j .$$

Moreover, s_m has the estimates

$$(5.2) \qquad \frac{-\log\left(r(\frac{2\pi}{3})\left(1 + \frac{r(\frac{\pi}{3})r(\frac{5\pi}{6})}{(3/(2+\sqrt{3}))^{m-1} r(\frac{2\pi}{3})^2 - r(\frac{\pi}{6})r(\frac{2\pi}{3})}\right)\right)}{2\log(2)} \le s_m \le \frac{-\log(r(\frac{2\pi}{3}))}{2\log(2)} .$$

As in §4, we get, asymptotically,

$$s_m = \left(1 - \frac{\log(3)}{2\log(2)}\right) m + O(\log(m)) \approx 0.2075m.$$

Remark. We see that the asymptotics of s_m is exactly the same as the one that Daubechies [3] gives as the upper bound ever reachable by her method. Also for all m s_m here is exactly half of σ_m of §4.

Again for small values of m the eigenvalue of T_r can be computed directly. This gives, e.g., $s_1 = \frac{1}{2}$, $s_2 = 1$, $s_3 = 3 - (\log(3)/\log(2))$. For Fig. 2 we have computed the maximal eigenvalues of T_r numerically.

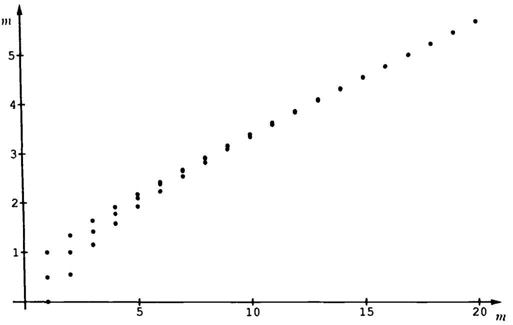


FIG. 2. *The Sobolev exponents* s_m *with the upper and lower bounds* (5.2).

In the next table we list the s_m's together with the σ_m's of Daubechies [3] and Daubechies and Lagarias [4] such that $\phi_m \in C^{\sigma_m}$.

We see that the present results practically imply (via $\mathcal{H}^s \subset C^{s'}$ for $s > s' + \frac{1}{2}$) those of [3] (cases $m \ge 5$), while the results of [4] (cases $2 \le m \le 4$) are in the gap (neither imply nor are implied by the present ones).

TABLE 2

m	s_m	σ_m
1	0.5	
2	1	0.550
3	1.415	1.088
4	1.775	1.618
5	2.096	1.596
6	2.388	1.888
7	2.658	2.158
8	2.914	2.415
9	3.161	2.661
10	3.402	2.902
11	3.639	
12	3.874	
13	4.106	
14	4.336	
15	4.565	
16	4.792	
17	5.019	
18	5.244	
19	5.469	
20	5.693	

6. Conclusions. A positive operator technique has been shown to apply in the study of the smoothness of the solutions of some dilation equations. At present, the positive operator works well only with trigonometric polynomials. Further applicability of this technique depends most of all on possible generalisation of the results of §3.

REFERENCES

[1] A. S. CAVARETTA, W. DAHMEN, AND C.A. MICCHELLI, *Stationary Subdivision*, Mem. Amer. Math. Soc., 453 (1992).

[2] W. DAHMEN AND C. A. MICCHELLI, *On stationary subdivision and the construction of compactly supported wavelets*, in Multivariate Approximation and Interpolation, K. Jetter and W. Haussmann, eds., Birkhäuser, Basel, Switzerland, 1990, pp. 69–90.

[3] INGRID DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[4] INGRID DAUBECHIES AND JEFFREY C. LAGARIAS, *Two-scale difference equations*, SIAM J. Math. Anal., this issue (1992), pp. 1031–1079.

[5] GILLES DESLAURIERS AND SERGE DUBUC, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.

[6] NIRA DYN AND DAVID LEVIN, *Interpolating subdivision schemes for the generation of curves and surfaces*, Internat. Ser. Numer. Math., 94 (1990), pp. 91–106.

[7] M. A. KRASNOSEL'SKIJ, G. M. VAINIKKO, P. P. ZABREIKO, YA. B. RUTITSKIJ, AND V. YA. STETSENKO, *Approximate Solution of Operator Equations*, Wolters-Noordhoff, Groningen, 1972.

[8] YVES MEYER, *Ondelettes*, Hermann, Paris, 1990.

[9] OLAVI NEVANLINNA, *Power bounded prolongations and Picard–Lindelöf iteration*, Numer. Math., 58 (1990), pp. 479–501.

[10] I. J. SCHOENBERG, *Cardinal spline interpolation*, SIAM Regional Conf. Ser. Appl. Math., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1973.

SIAM. J. MATH. ANAL.
Vol. 23, No. 4, pp. 1031-1079, July 1992

017

TWO-SCALE DIFFERENCE EQUATIONS II. LOCAL REGULARITY, INFINITE PRODUCTS OF MATRICES AND FRACTALS*

INGRID DAUBECHIES‡§ AND JEFFREY C. LAGARIAS†

Abstract. This paper studies solutions of the functional equation

$$f(x)=\sum_{n=0}^{N} c_n f(kx-n),$$

where $k\geqq 2$ is an integer, and $\sum_{n=0}^{N} c_n=k$. Part I showed that equations of this type have at most one L^1-solution up to a multiplicative constant, which necessarily has compact support in $[0, N/k-1]$. This paper gives a time-domain representation for such a function $f(x)$ (if it exists) in terms of infinite products of matrices (that vary as x varies). Sufficient conditions are given on $\{c_n\}$ for a continuous nonzero L^1-solution to exist. Additional conditions sufficient to guarantee $f\in C^r$ are also given. The infinite matrix product representations is used to bound from below the degree of regularity of such an L^1-solution and to estimate the Hölder exponent of continuity of the highest-order well-defined derivative of $f(x)$. Such solutions $f(x)$ are often smoother at some points than others. For certain $f(x)$ a hierarchy of fractal sets in $\mathbb{R}$ corresponding to different Hölder exponents of continuity for $f(x)$ is described.

Key words. Hölder continuity, subdivision schemes, wavelets, infinite matrix products

AMS(MOS) subject classifications. 26A15, 26A18, 39A10, 42A05

1. Introduction. This is the second part of a series of two papers concerning functional equations of the type

$$f(x)=\sum_{n=0}^{N} c_n f(kx-n). \tag{1.1}$$

In Part I (Daubechies and Lagarias (1991)) we discussed existence and uniqueness of L^1-solutions. We saw that nontrivial L^1-solutions to (1.1) can only exist if $|\Sigma\, c_n|\geqq k$. If $\Sigma\, c_n=k$, then the solution, if it exists at all, is unique and furthermore it has compact support contained in $[0, N/(k-1)]$. We shall assume that $\Sigma\, c_n=k$ in this paper. Functions satisfying equations of type (1.1) arise in many contexts. Our own motivation lies in the role played by such two-scale difference equations in the construction of orthonormal bases of compactly supported wavelets (see § 6.3 in part I or Daubechies (1988)). Similar equations characterize certain nowhere differentiable functions constructed by De Rham (1956, 1957, 1959). They also play an important role in interpolating subdivision schemes with applications to computer aided design, on which an important body of work exists; papers in this subject treating 2-scale difference equations are, e.g., Dubuc (1986), Micchelli (1986), Dyn, Gregory, and Levin (1987, 1991), Deslauriers and Dubuc (1987, 1989), Micchelli and Prautzsch (1987, 1989), and Dyn and Levin (1989). All these examples satisfy $\Sigma\, c_n=k$.

For the sake of simplicity we shall usually choose $k=2$, even though all our techniques can be applied for general integer values of $k\geqq 2$. We shall also restrict ourselves to real coefficients c_n and, correspondingly, real functions f. Our analysis can, however, also be used for complex c_n, without any changes. We are thus mainly

* Received by the editors November 30, 1988; accepted for publication (in revised form) September 16, 1991.

† AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

‡ Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

§ This author is "Bevoegdverklaard Navorser" at the Belgian National Science Foundation (on leave); also on leave from Department of Theoretical Physics, Vrije Universiteit Brussel, Belgium.

concerned with equations of the type

$$f(x)=\sum_{n=0}^{N} c_n f(2x-n) \tag{1.2}$$

with real c_n and $\Sigma\, c_n=2$.

Not all equations of type (1.2), with $\Sigma\, c_n=2$ have a nontrivial L^1-solution. There exist several possible approaches to the study of existence and smoothness of solutions to (1.2). One approach uses the trigonometric polynomial $p(\xi)=\frac{1}{2}\sum_{n=0}^{N} c_n\, e^{in\xi}$. In particular, if $p(\xi)$ can be factored as

$$p(\xi)=[(1+e^{i\xi})/2]^{L+1}q(\xi) \tag{1.3}$$

and

$$\sup_{\xi\in\mathbb{R}} |q(\xi)|<2^{L+1-m}, \tag{1.4}$$

then there exists a nontrivial L^1-solution f to (1.2), which is, moreover, m times continuously differentiable (Daubechies (1988)). Other sufficient conditions on p guaranteeing existence and smoothness can be found in Deslauriers and Dubuc (1987, 1989). Typically these conditions all achieve regularity of f by imposing decay on its Fourier transform $\hat{f}(\xi)$. These methods work best when $p(\xi)$ is a nonnegative function, as illustrated by Deslauriers and Dubuc (1987), who obtain very sharp information about the regularity of functions f constructed via a symmetric Lagrangian interpolation scheme. For more general examples, lacking the positivity of $p(\xi)$, this analysis leads to less than optimal results (Daubechies (1988), Deslauriers and Dubuc (1989)).

In this paper we use a different technique to study the regularity of solutions of (1.2). It is essentially a "time domain" method, in the sense that it does not involve Fourier transforms at any stage. This time domain approach hinges on the fact that if f satisfies (1.2), then the values $f(x)$ can be easily calculated, recursively, for all dyadic x, i.e., all x of the type $m2^{-j}(m=\mathbb{Z}, j\in\mathbb{N})$, if the values $f(m)$ at the integers are known. Provided suitable conditions are satisfied, we can then show that there exists a continuous extension of these $f(m2^{-j})$ to all of $\mathbb{R}$, thus defining $f(x)$. In fact, an explicit formula for $f(x)$ can be expressed using an infinite product of matrices (depending on x). It is then possible to discuss the Hölder continuity, differentiability, etc., of this extension, which is the desired solution of (1.2).

We explain the matrix technique in detail in § 2. We show how to choose a "correct" initialization $\{f(m);\ m\in\mathbb{Z}\}$ for the iterative spline approximation to f, and we formulate sufficient conditions on the c_n guaranteeing that the process converges to a continuous L^1-solution of (1.2). We also compute lower bounds for the Hölder exponent of the resulting function $f(x)$. In § 3 we show how similar ideas, using the sum rules (1.5) together with some technical conditions, can be used to prove that $f\in C^L$.

The time domain method proposed here leads in many instances to sharper results than the Fourier transform methods mentioned above. One such example is the function ϕ plotted in Fig. 1(a) (or in Fig. 3(a) in Part I). Even when handled with care, the Fourier transform method only establishes that the Hölder exponent of ϕ is at least $.5-\varepsilon$ (see Daubechies (1988)). In fact ϕ is Hölder continuous with exponent $2-\ln(1+\sqrt{3})/\ln 2=.5500...$; the method presented in this paper achieves this (optimal) result. Moreover, the Fourier transform method typically only controls global regularity properties: the detailed, local analysis of the regularity of solutions to (1.2), accessible via our matrix method as explained in § 4, is wholly outside the reach of Fourier transform based techniques. Using the time domain approach we can show, e.g., that the function ϕ in Fig. 1a is almost everywhere differentiable, and that its Hölder exponent .5500... is determined by a dense set of "bad" points which has zero measure. In fact, there exists a whole hierarchy of fractal sets, all with zero measure, corresponding to the Hölder exponent between .5500... and 1. Another interesting example is the basic function associated to the dyadic interpolation scheme first studied in Dubuc (1986) and Dyn, Gregory and Levin (1987). In this case the function $p(\xi)=\sum c_n e^{in\xi}$ is positive, allowing the Fourier transform method to achieve the already optimal result $f\in C^{2-\varepsilon}$ for the global regularity of f. Our time domain method recovers this result (although by a more involved analysis than via the Fourier transform method), but it also establishes that f is almost everywhere twice differentiable, which is a result outside the reach of the Fourier transform method. A detailed discussion of these and other examples can be found in § 5.

When we developed this technique, we were unaware of related and at that time unpublished work of Micchelli and Prautzsch (1989) and Dyn and Levin (1989), to which referees drew our attention. Let us give a short overview of the situation, pointing out the overlap and the differences between our work and theirs. The two papers by Micchelli and Prautzsch (1989) and the two by Dyn and Levin (1989) (hereafter called [MP] and [DL]) are both motivated by the applications of interpolation subdivision schemes to computer aided design (see Micchelli (1986), Micchelli and Prautzsch (1987) and Dyn, Gregory, and Levin (1987, 1989)). The coefficients c_n in an interpolation subdivision scheme typically satisfy $c_{2n+n_0}=\delta_{b0}$ for some n_0, but both [MP] and [DL] are applicable to more general coefficient choices. [MP] focuses on the existence and smoothness of solutions to (1.2), whereas [DL] is more concerned with the convergence of the corresponding interpolation scheme, that is, in the language of § 4 in part I, with existence of smooth solutions *and* convergence of the cascade algorithm to these solutions. It is proved in [DL] that the cascade algorithm converges to a C^L solution only if the coefficients c_n satisfy the $L+1$ "sum rules"

$$\sum_n c_n n^l(-1)^n=0, \qquad l=0,\cdots,L. \tag{1.5}$$

Moment conditions of the type (1.5) also turn up in a different context. If the solution f to (1.2) has the property that the integer translates $f(\cdot-n)$, $n\in\mathbb{Z}$ are all independent, then the moment condition (1.5) is equivalent to saying that all the polynomials of degree at most L can be written as combinations of the $f(\cdot-n)$, and can therefore be obtained exactly by the associated subdivision scheme. See Cavaretta, Dahmen, and Micchelli (1989) for a general multivariate discussion of this aspect. Sum rules of type (1.5) are in fact satisfied by all interesting examples (wavelets as well as interpolating schemes); it is easy to check that they are equivalent to the requirement that $p(\xi)$ can be factored as in (1.3). If we are concerned with only the existence of a reasonably smooth solution to (1.2), without convergence of the cascade algorithm to this solution, then conditions (1.5) are not necessary, and [MP] do not, in fact, use them. The discussion of possible solutions $f(x)$ to (1.2) in [MP] makes use of the same matrices as our approach. In particular, for the choice $k=2$, we deal with two matrices $\mathbf{T}_0$ and $\mathbf{T}_1$, both determined by the c_n, and $f(x)$ is then given by the limit

$$\lim_{n\to\infty}[\mathbf{T}_{d_1(x)}\mathbf{T}_{d_2(x)}\cdots\mathbf{T}_{d_n(x)}\cdots]\begin{bmatrix} f(0)\\ f(1)\\ \vdots\\ f(N-1)\end{bmatrix},$$

where the index $d_j(x)$ is the jth digit in the binary expansion for x (see § 2). A very special structure of the matrices $\mathbf{T}_0, \mathbf{T}_1$ is then required for f to be a C^L-function: both matrices have to have the eigenvalues $1, \frac{1}{2}, \cdots, 2^{-L}$, and the corresponding left eigenvectors $\mathbf{e}_l^0, \mathbf{e}_l^1$ ($l = 0, \cdots, L$) have to define flags of nested subspaces with very particular properties (see § 5 in [MP] or § 2, 3 below). In our paper, we require that the sum rules (1.5) hold; as shown by examples and by [MP], they are not necessary, but they make it possible to write very explicit expressions for the eigenvectors $\mathbf{e}_l^0, \mathbf{e}_l^1$. It turns out that (1.5) forces the $\mathbf{e}_l^0, \mathbf{e}_l^1$ to satisfy many of the conditions in [MP], which explains why we have only one technical condition (as compared to the four conditions in Theorem 5.2 in [MP]). If this condition is satisfied, then a simple spline approximation argument proves that the function f is indeed L times continuously differentiable. This simple argument was partly suggested to us by one of the anonymous referees, who we would like to thank here. In an appendix we shortly sketch our original, longer approach, which does not use splines but which can be generalized to certain higher-dimensional situations where splines would be hard to use.

Another difference between our paper and [MP] is that we also exploit the structure of the matrices $\mathbf{T}_0, \mathbf{T}_1$ to study local regularity properties of f: the importance of the binary expansion of x in determining $f(x)$ means that points with different frequencies of the digits 1 or 0 in their binary expansion correspond to different local Hölder exponents for f. In fact, this feature, first observed experimentally in graphs of orthonormal wavelet bases with compact support, was one of our main motivations for undertaking this study.

2. Continuity and Hölder continuity. If f satisfies (1.2), then the values of $f(x)$ can be computed explicitly, recursively, for all dyadic rationals x, i.e., for all x of the type $m2^{-j}(m \in \mathbb{Z}, j \in \mathbb{N})$, if the values $f(m)$ at the integers m are known. If f is continuous, then the $f(m2^{-j})$ suffice to determine f everywhere; a sequence of convergent spline approximations was constructed in Theorem 4.1 in part I. However, for general c_n and for arbitrarily chosen $f(m)$ this procedure will typically diverge and not lead to a continuous function at all. The following proposition gives a necessary condition that the c_n have to satisfy and specifies restrictions on the choice of the initial values $f(m)$ for the iteration scheme.

PROPOSITION 2.1. *Assume that* $\sum_{n=0}^{N} c_n = 2$. *If f is a nontrivial continuous L^1-solution to* (1.2), *then*:

(1^0) *The* $(N-1)\times(N-1)$ *dimensional matrix* $\mathbf{M}$ *defined by*

$$\mathbf{M}_{ij} = c_{2i-j}, \qquad 1 \leqq i, j \leqq N-1$$

has 1 *as an eigenvalue.*

(2^0) *The* $(N-1)$ *dimensional vector* $(f(1), \cdots, f(N-1))$ *is a right eigenvector of* $\mathbf{M}$ *with eigenvalue* 1, *and* $f(m) = 0$ *for all* $m \leqq 0$ *and* $m \geqq N$.

Proof. By Corollary 2.2 in part I, support $(f) \subset [0, N]$. This implies that $f(m) = 0$ for $m \leqq 0$, $m \geqq N$. The other conclusions follow from Theorem 5.1 in part I. □

In what follows we shall impose that the c_n satisfy the first sum rule in (1.5); this condition is sufficient (but not necessary) for (1^0) to hold. Explicitly,

$$\sum_{n=0}^{N} (-1)^n c_n = 0 \tag{2.1}$$

or, equivalently, since $\sum_n c_n = 2$,

$$\sum_n c_{2n} = \sum_n c_{2n+1} = 1. \tag{2.2}$$

This condition is satisfied in all the practical examples discussed in part I. It also has a very simple interpretation in the cascade algorithm (see § I.4). In this algorithm, the jth level consists of $(2^j-1)N+1$ different coefficients a^j_m, which are computed from the $(j-1)$th level via

$$a^j_{2m}=\sum_k c_{2(m-k)}a^j_k, \qquad a^j_{2m+1}=\sum_k c_{2(m-k)+1}a^j_k$$

(see, e.g., Micchelli and Prautzsch (1987), Dyn, Gregory, and Levin (1989) or Daubechies (1988)). The condition (2.2) states, therefore, that the total weight of the a^{j-1}_l in the computation of a^j_m is independent of m. In the case where the cascade algorithm simplifies to an interpolation subdivision scheme, this "equal opportunity" condition is automatically satisfied, since then

$$c_{l+2n}=\delta_{n0} \quad \text{for some } l \quad (\text{see § I.4}),$$

hence

$$\sum_m = c_{2m+1+l}=\sum_n c_{n+l}-\sum_m c_{2m+l}=\sum_{n=0}^{N} c_n-1=1=\sum c_{2m+l}.$$

If the scaling factor k is an integer larger than 2, then (2.2) has to be rewritten as

$$\sum_n c_{kn}=\sum_n c_{kn+1}=\cdots=\sum c_{kn+(k-1)}=1. \tag{2.2$'$}$$

This is again an "equal opportunity" condition when looked at from the point of view of the cascade algorithm. Note that (2.2′) is equivalent to requiring that $p(\xi)=\sum_{n=0}^N c_n\, e^{in\xi}$ is divisible by $(1+e^{i\xi}+\cdots+e^{i(k-1)\xi})$.

We return to the dyadic case $(k=2)$, and we assume that (2.2) is satisfied. It immediately follows that for all j, $1\leqq j\leqq N-1$,

$$\sum_{i=1}^{N-1}\mathbf{M}_{ij}=\sum_{i=1}^{N-1}c_{2i-j}=1,$$

where $\mathbf{M}_{ij}=c_{2i-j}$ (see above, or § I.5) and where we use the convention $c_l=0$ if $l<0$ or $l>N$. Consequently, 1 is an eigenvalue of $\mathbf{M}$, with left eigenvector $(1,1,\cdots,1)$. The necessary condition in Proposition 2.1 (1^0) is therefore satisfied.

Let us assume that the eigenvalue 1 of $\mathbf{M}$ is nondegenerate. (This will be guaranteed by a technical condition below.) Then there is a unique (up to normalization) right eigenvector $\mathbf{a}$ with eigenvalue 1 for $\mathbf{M}$. This eigenvector cannot be orthogonal to the left eigenvector $(1,1,\cdots,1)$, i.e., $\sum_{i=1}^{N-1}\mathbf{a}_i\neq 0$, so that we can normalize $\mathbf{a}$ to have $\sum_{i=1}^{N-1}\mathbf{a}_i=1$. We then pick the $f(n)$ to be

$$\begin{aligned} f(n)&=\mathbf{a}_n &\quad& n=1,\cdots,N-1\\ &=0 && n\leqq 0 \quad\text{or}\quad n\geqq N. \end{aligned}$$

We define successive spline approximations f_j to f as in § I.4:

(1) $f_0(x)$ is linear on every $[n,n+1]$,
$f_0(n)=f(n)$,

(2) $f_j=V^jf_0$,

where V is the linear operator

$$(Vg)(x)=\sum_{n=0}^N c_n g(2x-n). \tag{2.3}$$

If f is continuous, then the f_j converge to f (see Theorem 4.1 in part I). A priori we have, however, no reason to expect a continuous solution to (1.2); in fact, for many choices of the c_n, even if they satisfy (2.2), f will not be continuous, and the f_j will converge to f only in some distributional sense. We shall impose further conditions on the c_n which will enable us to prove that the f_j constitute a Cauchy sequence in L^∞, which then automatically leads to continuity for f. In order to do this, we introduce a "vector" notation.

Finding solutions to (1.2) is the same as finding fixed points for the linear operator V defined in (2.3). We shall only be concerned here with functions g supported on $[0, N] = \text{support}\,(f)$. For such functions (2.3) can be rewritten in "vector" form. Define the vector valued function $\mathbf{w}\colon [0, 1] \to \mathbb{R}^N$ by

$$\mathbf{w}(x)_n = g(x+n-1), \qquad n = 1, \cdots, N. \tag{2.4}$$

Knowing the function g is equivalent to knowing $\mathbf{w}\colon [0, 1] \to \mathbb{R}^N$. Note that necessarily

$$\mathbf{w}(0)_j = \mathbf{w}(1)_{j-1} \quad \text{for } 2 \leqq j \leqq N. \tag{2.5}$$

Moreover, g is continuous if and only if $\mathbf{w}$ is continuous on $[0, 1]$ and if

$$\mathbf{w}(0)_1 = 0 = \mathbf{w}(1)_N. \tag{2.6}$$

Let us now define the linear operator $\mathbf{V}$ on vector valued functions $\mathbf{w}$ satisfying (2.5), (2.6) by

$$(\mathbf{V}\mathbf{w})(x)_n = (Vg)(x+n-1),$$

where we assume $\mathbf{w}$ and g are linked as in (2.4). Then $\mathbf{V}$ is given explicitly by

$$\mathbf{V}\mathbf{w}(x) = \begin{cases} \mathbf{T}_0\mathbf{w}(2x) & \text{if } x \leqq \frac{1}{2}, \\ \mathbf{T}_1\mathbf{w}(2x-1) & \text{if } x \geqq \frac{1}{2}, \end{cases} \tag{2.7}$$

where $\mathbf{T}_0, \mathbf{T}_1$ are the $N \times N$-matrices defined by

$$\begin{aligned} (\mathbf{T}_0)_{ij} &= c_{2i-j-1}, & 1 \leqq i, j \leqq N, \\ (\mathbf{T}_1)_{ij} &= c_{2i-j}, & 1 \leqq i, j \leqq N, \end{aligned} \tag{2.8}$$

or

$$\mathbf{T}_0 = \begin{bmatrix} c_0 & 0 & 0 & 0 & \cdots & & & 0 & 0 \\ c_2 & c_1 & c_0 & 0 & \cdots & & & 0 & 0 \\ c_4 & c_3 & c_2 & c_1 & c_0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & c_N & c_{N-1} \end{bmatrix},$$

$$\mathbf{T}_1 = \begin{bmatrix} c_1 & c_0 & 0 & 0 & \cdots & & & 0 & 0 \\ c_3 & c_2 & c_1 & c_0 & 0 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots & & 0 & c_N \end{bmatrix}.$$

Note that if we strip either $\mathbf{T}_0$ of its first row and column, or $\mathbf{T}_1$ of its last row and column, then the resulting $(N-1) \times (N-1)$ matrix is exactly $\mathbf{M}$. Moreover, (2.2) implies that the N-dimensional row vector $\mathbf{e}_1 = (1, \cdots, 1)$ is a left eigenvector of both $\mathbf{T}_0$ and $\mathbf{T}_1$, with eigenvalue 1,

$$\mathbf{e}_1 \cdot \mathbf{T}_0 = \mathbf{e}_1 = \mathbf{e}_1 \cdot \mathbf{T}_1.$$

Formula (2.7) can be written more succinctly if we use binary expansions. For any $x \in [0, 1]$ we can write

$$x = \sum_{j=1}^{\infty} d_j 2^{-j}, \quad \text{where } d_j = 0 \text{ or } 1 \quad \text{for all } j.$$

We then define the action of the *shift operator* τ on x by

$$(2.9) \qquad \tau x = \sum_{j=2}^{\infty} d_j 2^{-j+1}.$$

This corresponds to shifting the decimal point in the binary expansion of x one step to the right and dropping the old first decimal. In fact,

$$\tau x = \begin{cases} 2x & \text{if } 0 \leqq x < \frac{1}{2}, \\ 2x - 1 & \text{if } \frac{1}{2} < x \leqq 1. \end{cases}$$

Note that this does not yet define τx for $x = \frac{1}{2}$; we shall come back below to what happens at $x = \frac{1}{2}$. For $x \neq \frac{1}{2}$, (2.3) can therefore be rewritten as

$$(2.10) \qquad \mathbf{V}\mathbf{w}(x) = \mathbf{T}_{d_1(x)}\mathbf{w}(\tau x),$$

where we have introduced the notation $d_j(x)$ for the jth digit in the binary expansion of x. For $x = \frac{1}{2}$ the first digit in the binary expansion (2.9) is not well defined, since we can take either $d_1 = 0$, $d_j = 1$ for $j \geqq 2$ or $d_1 = 1$, $d_j = 0$ for $j \geqq 2$. Extending (2.10) to $x = \frac{1}{2}$ leads, therefore, to two different equations, depending on which choice was made for the binary expansion of $\frac{1}{2}$,

$$\mathbf{w}(\tfrac{1}{2}) = \mathbf{T}_0\mathbf{w}(1) \quad \text{or} \quad \mathbf{w}(\tfrac{1}{2}) = \mathbf{T}_1\mathbf{w}(0).$$

For $\mathbf{w}$ satisfying the restrictions (2.5), (2.6), there is, however, no contradiction between these two equations, since $(\mathbf{T}_1)_{i,j} = (\mathbf{T}_0)_{i+1,j+1}$ for $1 \leqq i, j \leqq N-1$. Equation (2.10) therefore holds true for any $x \in [0, 1]$.

Let us apply all this to the spline functions f_j and define

$$[\mathbf{v}_j(x)]_n = f_j(x + n - 1), \qquad j \in \mathbb{N}, \quad n = 1, \cdots, N.$$

It follows that

$$\begin{aligned} \mathbf{v}_j(x) &= (\mathbf{V}^j\mathbf{v}_0)(x) \\ &= \mathbf{T}_{d_1(x)}\mathbf{T}_{d_2(x)} \cdots \mathbf{T}_{d_j(x)}\mathbf{v}_0(\tau^j x). \end{aligned}$$

Now that we have introduced all the necessary notation, we are in a position to formulate the main result of this section.

We will use the operator norm for matrices defined by $\|\mathbf{T}\| = \sup_{\mathbf{v}\neq 0} \|\mathbf{T}\mathbf{v}\| / \|\mathbf{v}\|$, where $\|\mathbf{v}\|$ is the Euclidean norm of $\mathbf{v}$, $\|\mathbf{v}\|^2 = \sum_{j=1}^{N} \mathbf{v}_j^2$.

THEOREM 2.2. *Assume that the* c_n, $n = 0, \cdots, N$ *satisfy*

$$\sum_n c_{2n} = \sum_n c_{2n+1} = 1.$$

Define the $N \times N$*-matrices* $\mathbf{T}_0$ *and* $\mathbf{T}_1$ *by*

$$(\mathbf{T}_0)_{ij} = c_{2i-j-1}, \qquad (\mathbf{T}_1)_{ij} = c_{2i-j}, \quad 1 \leqq i, j \leqq N.$$

Define E_1 *to be the* $(N-1)$*-dimensional subspace orthogonal to* $\mathbf{e}_1 = (1, \cdots, 1)$, *the common left eigenvector of* $\mathbf{T}_0, \mathbf{T}_1$ *for the eigenvalue* 1. *Assume that there exist* $\lambda < 1$ *and* $C > 0$ *such that, for all* $m \in \mathbb{N}$,

$$(2.11) \qquad \max_{d_j = 0 \text{ or } 1, j = 1, \cdots, m} \|\mathbf{T}_{d_1}\mathbf{T}_{d_2} \cdots \mathbf{T}_{d_m}|_{E_1}\| \leqq C\lambda^m.$$

Then the following hold:

(1) *The eigenvalue* 1 *is of the* $(N-1)\times(N-1)$*-dimensional matrix* $\mathbf{M}$ *defined by* $\mathbf{M}_{ij}=c_{2i-j}$, $1\leqq i, j\leqq N-1$, *is simple and there is an associated right eigenvector* $\mathbf{a}$ *with* $\sum_{i=1}^{N-1}\mathbf{a}_i=1$.

(2) *The vector-valued functions* $\mathbf{v}_j(x)$ *defined above satisfy* $\mathbf{e}_1\cdot\mathbf{v}_j(x)=1$ *for all* $j\in\mathbb{N}$, *all* $x\in[0,1]$.

(3) *The corresponding functions* f_j *converge uniformly to a continuous function* f,

$$\|f_j-f\|_{L^\infty}\leqq C2^{-j|\ln\lambda|/\ln 2}.$$

(4) *The limit function* f *is an* L^1*-solution to* (1.2); *it is normalized so that* $\int_0^N dx\, f(x)=1$, *and it is Hölder continuous,*

$$|f(x)-f(y)|\leqq C|x-y|^\alpha, \tag{2.12}$$

with $\alpha=|\ln\lambda|/\ln 2$.

Proof. (1) The constraint (2.16) automatically implies that 1 is a simple eigenvalue of $\mathbf{T}_0$ and $\mathbf{T}_1$. Indeed, if 1 were not a simple eigenvalue of, e.g., $\mathbf{T}_0$, then there would exist a right eigenvector $\mathbf{e}_1'$ for $\mathbf{T}_0$ in E_1, with eigenvalue 1 (regardless of whether the eigenvalue 1 is degenerate or not, in which last case the matrix $\mathbf{T}_0$, restricted to the invariant subspace for the eigenvalue 1, can be brought in Jordan normal form but not diagonalized). It would then immediately follow that

$$\|\mathbf{T}_0^m|_{E_1}\|\geqq\|\mathbf{T}_0^m\mathbf{e}_1'\|/\|\mathbf{e}_1'\|=1,$$

contradicting (2.11).

(2) We already know that 1 is an eigenvalue of $\mathbf{M}$ and that the $(N-1)$-dimensional vector $(1,\cdots,1)$ is a left eigenvector for this eigenvalue. Since $(\mathbf{T}_0)_{i+1,j+1}=\mathbf{M}_{ij}$, $1\leqq i,j\leqq N-1$ and $(\mathbf{T}_0)_{1j}=0$ for $j\geqq 2$, we find that any eigenvalue of $\mathbf{M}$ is an eigenvalue of $\mathbf{T}_0$, with at least the same multiplicity. Since 1 is a simple eigenvalue of $\mathbf{T}_0$, it is, therefore, also a simple eigenvalue of $\mathbf{M}$. It then follows from arguments presented above that the right eigenvector $\mathbf{a}$ for the eigenvalue 1 of $\mathbf{M}$ can be normalized so that $\sum_{n=1}^{N-1}\mathbf{a}_n=1$. As previously agreed, we then define

$$[\mathbf{v}_0(x)]_n=\mathbf{a}_{n-1}(1-x)+\mathbf{a}_n x,\qquad n=1,\cdots,N,$$

with the convention $\mathbf{a}_0=0=\mathbf{a}_N$.

(3) From the normalization of $\mathbf{a}$, it now follows that, for all $x\in[0,1]$,

$$\mathbf{e}_1\cdot\mathbf{v}_0(x)=\sum_{n=1}^{N}[\mathbf{v}_0(x)]_n=1.$$

We now prove, by induction, that the same is true for all $\mathbf{v}_j$. Suppose $\mathbf{e}_1\cdot\mathbf{v}_j(x)=1$. Then, for all $x\in[0,1]$,

$$\mathbf{e}_1\cdot\mathbf{v}_{j+1}(x)=\mathbf{e}_1\cdot\mathbf{T}_{d_1(x)}\mathbf{v}_j(\tau x)=\mathbf{e}_1\cdot\mathbf{v}_j(\tau x)=1,$$

where we have used $\mathbf{e}_1\cdot\mathbf{T}_0=\mathbf{e}_1=\mathbf{e}_1\cdot\mathbf{T}_1$.

(4) Next we show that the $\mathbf{v}_k(x)$ are uniformly bounded. Since $\mathbf{e}_1\cdot\mathbf{v}_k(x)=1$ for all k, x, it follows that $\mathbf{v}_k(x)-\mathbf{v}_l(x)\in E_1$ for all k, l, x. Hence

$$\begin{aligned}\|\mathbf{v}_{k+1}(x)-\mathbf{v}_{k(x)}\|&=\|\mathbf{T}_{d_1(x)}\cdots\mathbf{T}_{d_k(x)}[\mathbf{v}_1(\tau^k x)-\mathbf{v}_0(\tau^k x)]\|\\&\leqq C\lambda^k\sup_{y\in[0,1]}\|\mathbf{v}_1(y)-\mathbf{v}_0(y)\|.\end{aligned}$$

Consequently,

$$\begin{aligned}\|\mathbf{v}_k(x)\| &\leqq \|\mathbf{v}_0(x)\| + \sum_{j=1}^{k} \|\mathbf{v}_j(x)-\mathbf{v}_{j-1}(x)\| \\ &\leqq \sup_{y\in[0,1]} \|\mathbf{v}_0(y)\| + C(1-\lambda)^{-1} \sup_{y\in[0,1]} \|\mathbf{v}_1(y)-\mathbf{v}_0(y)\|,\end{aligned}$$

so that $\|\mathbf{v}_k(x)\|$ is bounded uniformly in k and in x.

(5) Together with $\mathbf{e}_1 \cdot \mathbf{v}_k(x)=1$, the uniform boundedness of the $\mathbf{v}_k(x)$ implies that the $\mathbf{v}_k$ constitute a Cauchy sequence in L^∞-norm. Indeed, since

$$\begin{aligned}\|\mathbf{v}_{j+k}(x)-\mathbf{v}_j(x)\| &= \|\mathbf{T}_{d_1(x)} \cdots \mathbf{T}_{d_j(x)}[\mathbf{v}_k(x)-\mathbf{v}_0(x)]\| \\ &\leqq 2C\lambda^j \sup_{l,y} \|\mathbf{v}_l(y)\|,\end{aligned}$$

we find that $\sup_{x\in[0,1]} \|\mathbf{v}_{j+k}(x)-\mathbf{v}_j(x)\|$ can be made arbitrarily small by choosing j large enough, independently of k. It follows that there exists a limit,

$$\mathbf{v}(x) = \lim_{j\to\infty} \mathbf{v}_j(x),$$

which is continuous since all the $\mathbf{v}_j$ are continuous and the convergence is uniform in x,

$$\sup_{x\in[0,1]} \|\mathbf{v}(x)-\mathbf{v}_j(x)\| \leqq C\lambda^j.$$

Since every $\mathbf{v}_j$ satisfies (2.5), (2.6), so does $\mathbf{v}$. It follows that the function f defined on $[0, N]$ by

$$f(x) = [\mathbf{v}(x-\lfloor x\rfloor)]_{\lfloor x\rfloor+1}$$

(where $\lfloor x\rfloor$ denotes the largest integer not exceeding x) is continuous, and that

$$\|f-f_j\|_{L^\infty} \leqq C\lambda^j = C2^{-j\alpha} \tag{2.13}$$

with $\alpha = |\ln \lambda|/\ln 2$.

(6) Since $f_j = Vf_{j-1}$, the limit function f satisfies $f = Vf$, i.e., f is a solution to (1.2), which is necessarily L^1 since f is bounded and compactly supported. We have, moreover, $\mathbf{e}_1 \cdot \mathbf{v}(x) = \lim_{j\to\infty} \mathbf{e}_1 \cdot \mathbf{v}_j(x) = 1$, so that

$$\int_0^N dx f(x) = \int_0^1 dx \sum_{n=1}^{N} [\mathbf{v}(x)]_n = 1.$$

(7) The Hölder continuity follows from (2.13) and standard spline results translating estimates on how well a function can be approximated by piecewise polynomials into Hölder estimates on the function itself. (See, e.g., Theorem 6.10 in Schumaker (1981), which uses approximations by piecewise constant functions rather than piecewise linear f_j; an estimate similar to (2.13) can also be proved for piecewise constant approximations to f.) For the sake of convenience, we also give a direct proof by the following short argument. Suppose that $2^{-(j+1)} \leqq y-x \leqq 2^{-j}$. Then there exists $l \in \mathbb{N}$ so that one of the two following alternatives holds: $(l-1)2^{-j} \leqq x \leqq y \leqq l2^{-j}$ or $(l-1)2^{-j} \leqq x \leqq l2^{-j} \leqq y \leqq (l+1)2^{-j}$. We shall only discuss the second case; the first one is similar. We then have

$$\begin{aligned}|f(x)-f(y)| &\leqq |f(x)-f_j(x)| + |f_j(x)-f_j(l2^{-j})| \\ &\quad + |f_j(l2^{-j})-f_j(y)| + |f_j(y)-f(y)| \\ &\leqq 2C2^{-\alpha j} + |f_j(x)-f_j(l2^{-j})| + |f_j(y)-f_j(l2^{-j})|,\end{aligned}$$

by (2.13). Because of the choice of l, there exists $k \in \mathbb{N}$ so that $x' = x - k$ and $l'2^{-j} = l2^{-j} - k$ are both in $[0, 1]$. We can, moreover, choose binary expansions for x' and $l'2^{-j}$ with coinciding first j digits (choose the expansion ending in ones for $l'2^{-j}$, and if x' is dyadic, the expansion ending in zeros for x'). It follows that

$$\begin{aligned}|f_j(x) - f_j(l2^{-j})| &\leqq \|\mathbf{v}_j(x') - \mathbf{v}_j(l'2^{-j})\| \\ &= \|\mathbf{T}_{d_1(x')} \cdots \mathbf{T}_{d_j(x')}[\mathbf{v}_0(\tau^j x') - \mathbf{v}_0(\tau^j(l'2^{-j}))]\| \\ &\leqq C\lambda^j = C2^{-j},\end{aligned}$$

where we have used (2.11), the boundedness of $\mathbf{v}_0$, and $\mathbf{v}_0(u) - \mathbf{v}_0(u') \in E_1$ for all u, u'. Similarly we can bound $|f_j(y) - f_j(l2^{-j})|$; putting it all together leads to

$$|f(x) - f(y)| \leqq C'2^{-\alpha j} \leqq C''|x - y|^{\alpha},$$

which is (2.12). □

Remarks. (1) Note that the argument in point (7) would also work, in principle, if $\lambda < \frac{1}{2}$. In that case $|f(x) - f(y)| \leqq C|x - y|^{1+\varepsilon}$ would follow, with $\varepsilon > 0$, which is only possible if $f \equiv$ constant. Since support $(f) = [0, N]$ and f is continuous, this implies $f \equiv 0$. It follows that λ in (2.11) necessarily satisfies $\lambda \geqq \frac{1}{2}$.

(2) Under the conditions of this theorem, for $0 \leqq x \leqq 1$ all infinite products $\mathbf{T}_{d_1(x)}\mathbf{T}_{d_2(x)}\mathbf{T}_{d_3(x)} \cdots$ of the matrices $\mathbf{T}_0$ and $\mathbf{T}_1$ converge to the limit matrices

$$\mathbf{T}_{d_1(x)}\mathbf{T}_{d_2(x)}\mathbf{T}_{d_3(x)} \cdots = \begin{bmatrix} f(x) & f(x) & \cdots & f(x) \\ f(x+1) & f(x+1) & \cdots & f(x+1) \\ \vdots & \vdots & & \vdots \\ f(x+N-1) & f(x+N-1) & \cdots & f(x+N-1) \end{bmatrix}.$$

Theorem 2.2 is similar to Theorem 5.1 in Micchelli and Prautzsch (1989), with the following differences. On one hand, [MP] are only interested in continuity, and technical condition (b) in their Theorem 5.1 is a little less tight than our (2.11), although both are very similar. On the other hand, they have extra conditions (a), (c), (d), which are here automatically satisfied because we have restricted ourselves to the case where the sum rule (2.2) holds.

To apply Theorem 2.2 we have to verify the technical condition (2.11). It might seem impossible to check in practice, since it involves the norms of infinitely many products of matrices. It turns out, however, that (2.11) can be reduced to a criterion that uses only a finite-time computer search. This is the constant of the next lemma.

LEMMA 2.3. *Define*

$$\lambda_m = \max_{\substack{d_j = 0 \text{ or } 1 \\ j = 1, \cdots, m}} \|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_m}|_{E_1}\|^{1/m}. \tag{2.14}$$

A necessary and sufficient condition for (2.11) *to hold is that*

$$\lambda_m < 1 \qquad \textit{for some } m \in \mathbb{N}. \tag{2.15}$$

Proof. Suppose $\lambda_m < 1$. *Write* $n = qm + r$, with q, $r \in \mathbb{N}$, $0 \leqq r < m$. Then

$$\|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_n}|_{E_1}\| \leqq \lambda_m^{mq}\lambda_r^r \leqq \max(1, \lambda_1, \cdots, \lambda_{m-1}^{m-1})\lambda_m^{-r}\lambda_m^n.$$

This implies (2.11), with $\lambda = \lambda_m < 1$, and $C = \lambda_m^{-m+1}\max(1, \cdots, \lambda_{m-1}^{m-1})$. Conversely, if (2.11) is satisfied, then $\lambda_m \leqq C^{1/m}\lambda$, hence $\lambda_m < 1$ for large enough m. □

In some examples, the smallest value m for which (2.14) holds may still be too large: up to a formidable 2^m matrix norms may have to be checked for every candidate m. In the next section, we shall see some additional tricks to simplify the search.

The technical condition (2.11) or (2.14) on $\mathbf{T}_0, \mathbf{T}_1$ can be interpreted as a "spectral" constraint. In fact, given two matrices $\mathbf{S}_0, \mathbf{S}_1$, we can define (Rota and Strang (1960), Daubechies and Lagarias (1991)) the joint spectral radius for $\mathbf{S}_0, \mathbf{S}_1$ by

$$\hat{\rho}(\mathbf{S}_0, \mathbf{S}_1) = \limsup_{m\to\infty} [\max_{\substack{d_j=0 \text{ or } 1\\ j=1,\cdots,m}} \|\mathbf{S}_{d_1}\cdots \mathbf{S}_{d_m}\|^{1/m}]. \tag{2.16}$$

In the case where the two matrices are identical, it is well known that (2.16) reduces to the spectral radius. We have the following.

LEMMA 2.4. *A necessary and sufficient condition for* (2.11) *to hold is that the joint spectral radius* $\hat{\rho}(\mathbf{T}_0|_{E_1}, \mathbf{T}_1|_{E_1}) < 1$.

Proof. We have $\hat{\rho}(\mathbf{T}_0|_{E_1}, \mathbf{T}_1|_{E_1}) = \limsup_{m\to\infty} \lambda_m$. If $\hat{\rho}(\mathbf{T}_0|_{E_1}, \mathbf{T}_1|_{E_1}) < 1$, then $\lambda_m < 1$ follows for large enough m. On the other hand, the proof of Lemma 2.3 above shows that $\lambda_m < 1$ implies $\lambda_n \leqq C^{1/n}\lambda_m$ for all $n \in \mathbb{N}$, which leads to $\hat{\rho}(\mathbf{T}_0|_{E_1}, \mathbf{T}_1|_{E_1}) < 1$. $\square$

Note that if $\hat{\rho}(\mathbf{T}_0|_{E_1}, \mathbf{T}_1|_{E_1}) = \lambda_0 < 1$, then (2.11) holds for all $\lambda > \lambda_0$, but not necessarily for $\lambda = \lambda_0$.

Remarks. (1) As noted before, the sum rule $\sum c_{2n} = \sum c_{2n+1}$ is not necessary for a continuous solution to exist. An example is the continuous function $f(x) = x/2$ for $0 \leqq x \leqq 2$, $2 - x/2$ for $2 \leqq x \leqq 4$, zero otherwise, which satisfies $f(x) = \frac{1}{2}f(2x) + f(2x-2) + \frac{1}{2}f(2x-4)$, with $\sum c_{2n} = 2$, $\sum c_{2n+1} = 0$. This example has been obtained by "stretching" an equation that does satisfy (2.1): the function $\tilde{f}(x) = f(2x)$ satisfies

$$\tilde{f}(x) = \tfrac{1}{2}\tilde{f}(2x) + \tilde{f}(2x-1) + \tfrac{1}{2}\tilde{f}(2x-2).$$

Consequently, the matrices $\mathbf{T}_0, \mathbf{T}_1$ still have many interesting properties in this case, even though they do not have a common left eigenvector for the eigenvalue 1.

(2) By the same argument as in the remark following Theorem 2.3, we necessarily have $\lambda_m \geqq \frac{1}{2}$ for any choice of $m \in \mathbb{N}$, where λ_m is as defined by (2.14).

(3) For any matrix $\mathbf{T}$ the spectral radius $\rho(\mathbf{T})$,

$$\rho(\mathbf{T}) = \max\{|\mu|;\ \mu \text{ eigenvalue of } \mathbf{T}\},$$

is given by

$$\rho(\mathbf{T}) = \lim_{l\to\infty} \|\mathbf{T}^l\|^{1/l} = \inf_l \|\mathbf{T}^l\|^{1/l}.$$

If we restrict the choices $d_1, \cdots, d_m$ in (2.11) to either all zero or all 1, it follows, therefore, that

$$\rho(\mathbf{T}_0|_{E_1}),\ \rho(\mathbf{T}_1|_{E_1}) \leqq \lambda, \tag{2.17}$$

i.e., all the other eigenvalues of $\mathbf{T}_0, \mathbf{T}_1$ (excluding 1) have absolute value smaller than $\lambda < 1$. The condition $\rho(\mathbf{T}_0|_{E1}), \rho(\mathbf{T}_1|_{E_1}) < 1$ is, however, not sufficient to ensure that (2.11) holds. An example is given by

$$N = 3; \quad c_0 = -.75, \quad c_1 = .2,$$
$$c_2 = 1.75, \quad c_3 = .8.$$

In this case the spectra of $\mathbf{T}_0, \mathbf{T}_1$ are $\{1, -.75, .95\}$ and $\{1, .8, .95\}$, respectively. We have, therefore, $\rho(\mathbf{T}_0|_{E_1}) = \rho(\mathbf{T}_1|_{E_1}) = .95$. On the other hand, the spectrum of $\mathbf{T}_0, \mathbf{T}_1$ is $\{1, .50976\cdots, -1.06226\cdots\}$, implying that $\lambda_m \geqq (1.06226\cdots)^{1/2} > 1$ for any m.

(4) Because of the special structure (2.8) of $\mathbf{T}_0, \mathbf{T}_1$, it can easily be shown that

$$\text{spectrum}\,(\mathbf{T}_0) = \{c_0\} \cup \text{spectrum}\,(\mathbf{M}),$$

$$\text{spectrum}\,(\mathbf{T}_1) = \{c_N\} \cup \text{spectrum}\,(\mathbf{M}).$$

In order for (2.11) to be satisfied, it is therefore necessary that $|c_0|, |c_N| \leqq 1$.

(5) At the end of Deslauriers and Dubuc (1989), which mainly concerns the solutions of an equation of type (1.2) corresponding to a symmetric Lagrangian interpolation scheme, a conjecture is presented concerning generalizations to other, non-Lagrangian, interpolation schemes. Translated in our present terminology, this conjecture reads as follows.

CONJECTURE 2.5. *Assume that* $\sum_n c_{2n} = \sum_n c_{2n+1} = 1$. *Define the bounded operator* $\mathbf{A}$ *on* $l^2(\mathbb{Z})$ *by* $(\mathbf{A}a)_j = \sum_l c_{2j-l}a_l$. *Then there exists a continuous nontrivial solution to* (1.2) *if* 1 *is a nondegenerate eigenvalue of* $\mathbf{A}$ *and if all the other eigenvalues of* $\mathbf{A}$ *have modulus strictly smaller than* 1.

The operator $\mathbf{A}$ does have a rather bothersome spectrum, however; we find that all the complex numbers with modulus strictly smaller than 1 are in the point spectrum of $\mathbf{A}$. For the simple case $c_0 = 1$, $c_1 = c_{-1} = \frac{1}{2}$, $c_n = 0$ for $|n| > 1$, e.g., and $\lambda \in \mathbb{C}$, $|\lambda| < 1$, the sequence a^λ defined by

$$a^\lambda_{-n} = 0, \qquad n \in \mathbb{N}$$

$$a^\lambda_1 = 0, \quad a^\lambda_2 = 1, \quad a^\lambda_{3\cdot 2^m} = -2\lambda^m, \quad a^\lambda_{4\cdot 2^m} = \lambda^m(\lambda+1),$$

$$a^\lambda_n = 0 \quad \text{if } n \geqq 5, \qquad n \neq 3 \cdot 2^m \quad \text{or} \quad 4 \cdot 2^m$$

is clearly in $l^2(\mathbb{Z})$, and satisfies $\mathbf{A}a^\lambda = \lambda a^\lambda$. It follows that the closed unit disk is part of the spectrum of $\mathbf{A}$. In practice, it may be very hard to decide whether 1 is the only element on the unit circle that is not only in the spectrum of $\mathbf{A}$, but also a true eigenvalue of $\mathbf{A}$. For this reason, this conjecture, even if true, does not seem to give an easily checkable criterion for a given set of c_n to lead to a continuous solution of (1.2).

We can prove a result analogous to Conjecture 2.5. Define, for any two matrices $\mathbf{S}_0, \mathbf{S}_1$, their "generalized spectral radius" by Daubechies and Lagarias (1991)

$$\rho(\mathbf{S}_1, \mathbf{S}_2) = \limsup_{n\to\infty} [\max_{\substack{d_j = 0, \text{or } 1 \\ j = 1, \cdots, n}} \rho(\mathbf{S}_{d_1} \cdots \mathbf{S}_{d_n})^{1/n}],$$

where $\rho(\mathbf{S})$ denotes the usual spectral radius. Berger and Wang (1991) prove that

$$\rho(\mathbf{S}_1, \mathbf{S}_2) = \hat{\rho}(\mathbf{S}_1, \mathbf{S}_2).$$

Consequently Theorem 2.2. and Lemma 2.4 give the following.

THEOREM 2.6. *Assume that the* c_n, $n = 0, \cdots, N$ *satisfy* $\sum c_{2n} = 1 = \sum c_{2n+1}$. *Then there exists a continuous nontrivial solution to* (1.2) *if* $\rho(\mathbf{T}_0|_{E_1}, \mathbf{T}_1|_{E_1}) < 1$.

It would be of interest to determine necessary and sufficient conditions on $\{c_n\}$ for the existence of a continuous solution to (2.1) having $\sum c_{2n} = 1 = \sum c_{2n+1}$.

(6) The same analysis can be done for two-scale difference equations having larger (integer) values of k. In general, there will then be k different $(N_0+1) \times (N_0+1)$-matrices $\mathbf{T}_0, \cdots, \mathbf{T}_{k-1}$, where N_0 is the largest integer strictly smaller than $N/(k-1)$. Instead of binary expansions of $x \in [0, 1]$, we take the expansion of x in base k. Otherwise, the proofs carry over without change.

3. Higher-order regularity. If L additional "sum rules" of the type (2.1) are satisfied, then techniques similar to those that proved continuity and Hölder continuity in the preceding section can be used to show that $f \in C^l$, with $l > 1$. The enlarged set of "sum rules" is

$$(3.1)\qquad \sum_{n=0}^{N} (-1)^n n^l c_n = 0 \quad \text{for } l = 0, 1, \cdots L;$$

this is equivalent to requiring that $p(\xi) = \frac{1}{2}\sum_{n=0}^{N} c_n\, e^{in\xi}$ is divisible by $(1+e^{i\xi})^{L+1}$. For $L = 0$, (3.1) reduces to (2.1). We define the vectors $\mathbf{u}_j \in \mathbb{R}^n, j = 1, \cdots, L+1$, by

$$(3.2)\qquad (\mathbf{u}_j)_k = k^{j-1}, \qquad k = 1, \cdots, N.$$

The vector $\mathbf{u}_1$ is a common left eigenvector of $\mathbf{T}_0, \mathbf{T}_1$, with eigenvalue 1. Similarly the $\mathbf{u}_j$ lead us to left eigenvectors of $\mathbf{T}_0$ and $\mathbf{T}_1$ with eigenvalue 2^{-j+1} (see below). This spectral analysis of $\mathbf{T}_0, \mathbf{T}_1$ can then be used to prove the following generalization of Theorem 2.2.

THEOREM 3.1. *Assume that the* c_n, $n = 0, \cdots, N$, *satisfy* $\sum_{n=0}^{N} c_n = 2$ *and*

$$\sum_{n=0}^{N} (-1)^n n^l c_n = 0 \quad \text{for } l = 0, 1, \cdots, L.$$

For every $m = 1, \cdots, L+1$, *define* E_m *to be the subspace of* $\mathbb{R}^N$ *orthogonal to* $U_m = \operatorname{Span}\{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ *where* $\mathbf{u}_j = (1^{j-1}, 2^{j-1}, \cdots, N^{j-1})$. *Assume that there exist* $\frac{1}{2} \leqq \lambda < 1$, $0 \leqq l \leqq L (l \in \mathbb{N})$ *and* $C > 0$ *such that, for all binary sequences* $(d_j)_{j \in \mathbb{N}}$, *and all* $m \in \mathbb{N}$,

$$(3.3)\qquad \|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_m}|_{E_{L+1}}\| \leqq C\lambda^m 2^{-ml}.$$

Then

(1) *There exists a nontrivial continuous* L^1*-solution f for the two-scale equation* (1.2) *associated with the* c_n;

(2) *This solution f is l times continuously differentiable*;

(3) *If* $\lambda > \frac{1}{2}$, *then the lth derivative* $f^{(l)}$ *of f is Hölder continuous, with exponent at least* $|\ln \lambda|/\ln 2$; *if* $\lambda = \frac{1}{2}$, *then the lth derivative* $f^{(l)}$ *of f is almost Lipschitz*: *it satisfies*

$$|f^{(l)}(x+t) - f^{(l)}(t)| \leqq C|t||\ln|t||.$$

Remarks. (1) The restriction $\lambda \leqq \frac{1}{2}$ means only that we pick the largest possible integer $l \leqq L$ for which (3.3) holds with $\lambda < 1$. If $l = L$, then we shall see below that necessarily $\lambda \geqq \frac{1}{2}$; if $l < L$ and $\lambda < \frac{1}{2}$, then we could replace l by $l+1$ and λ by 2λ, and (3.3) would hold for a larger integer l.

(2) The formulation of Theorem 3.1 implicitly assumes that $L+1 < N$. If $L+1 = N$, then $U_{L+1} = \mathbb{R}^N$, $E_{L+1} = \{0\}$, and condition (3.3) becomes meaningless. In the case $L+1 = N$, the $(N+1)$ coefficients c_n are completely determined by the N "sum rules" (3.1) and the requirement $\sum_{n=0}^{N} c_n = 2$. The characteristic determinant of the resulting system of $N+1$ linear equations is different from zero (it can be written as a positive linear combination of positive Vandermonde determinants); the unique solution to the system is $c_n = 2^{-N+1}\binom{N}{n}$, $n = 0, \cdots, N$. For $N = 2$, e.g., we find $c_0 = \frac{1}{2}$, $c_1 = 1$, $c_2 = \frac{1}{2}$. The corresponding function $f(x)$ is given by

$$f(x) = \begin{cases} x & 0 \leqq x \leqq 1, \\ 2-x & 1 \leqq x \leqq 2. \end{cases}$$

This function is Lipschitz but not C^1. This is typical of what happens for larger values of N, when $L+1 = N$: the corresponding function f is the B-spline function of degree L, which is C^{L-1}, and its $(L-1)$-th derivative is Lipschitz, but not everywhere differentiable. The points where $f^{(L-1)}$ fails to be differentiable are $0, 1, \cdots$, N, where the left and right Lth derivatives of f do not coincide. In most of what follows we shall implicitly assume $L+1<N$.

To prove Theorem 3.1, we shall need several technical lemmas. The first one shows how the $\mathbf{u}_j$ are related to left eigenvectors of $\mathbf{T}_0, \mathbf{T}_1$ with eigenvalue 2^{-j+1}.

LEMMA 3.2. *Assume that* (3.1) *holds. Let* $\mathbf{T}_0, \mathbf{T}_1$ *be defined as in* (2.8), $\mathbf{u}_j, j=1, \cdots, L+1$ *as in* (3.2). *Define* $U_j =$ *Linear span* $\{\mathbf{u}_1, \cdots, \mathbf{u}_j\}$, $j=1, \cdots, L+1$. *For all* $j=1, \cdots, L+1, 2^{-j}$ *is an eigenvalue of both* $\mathbf{T}_0, \mathbf{T}_1$. *The corresponding left eigenvectors, denoted by* $\mathbf{e}_j^0, \mathbf{e}_j^1$ *respectively, are both in* U_j. *We fix their normalization by requiring that* $\mathbf{e}_j^0-\mathbf{u}_j$, $\mathbf{e}_j^1-\mathbf{u}_j \in U_{j-1}$. *Then*

$$\mathbf{e}_j^0=\sum_{k=1}^{j}(-1)^{j-k}\binom{j-1}{k-1}\mathbf{e}_k^1 \tag{3.4}$$

or, equivalently,

$$\mathbf{e}_j^1=\sum_{k=1}^{j}\binom{j-1}{k-1}\mathbf{e}_k^0. \tag{3.5}$$

Proof. (1) Define, for $k=0, \cdots, L$,

$$C_k=\sum_n (2n)^k c_{2n}=\sum_n (2n+1)^k c_{2n+1}.$$

We then easily check from the definitions (2.8), (3.2) of $\mathbf{T}_0, \mathbf{T}_1$, and the $\mathbf{u}_j$ that

$$\mathbf{u}_j\cdot \mathbf{T}_1=2^{-j+1}\sum_{n=1}^{j}\binom{j-1}{n-1}C_{j-n}\mathbf{u}_n,$$

$$\mathbf{u}_j\cdot \mathbf{T}_0=2^{-j+1}\sum_{n=1}^{j}\binom{j-1}{n-1}\sum_{l=0}^{j-n}\binom{j-n}{l}C_l\mathbf{u}_n,$$

for $j=1, \cdots, L+1$. It follows that, for every j, $1 \leqq j \leqq L+1$, there are left eigenvectors for both $\mathbf{T}_0, \mathbf{T}_1$, with eigenvalue 2^{-j+1}, in the linear span U_j of $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_j\}$. We denote these eigenvectors by $\mathbf{e}_j^0, \mathbf{e}_j^1$, respectively. Then

$$\mathbf{e}_j^0=\sum_{n=1}^{j}a_{j,n}\mathbf{u}_n, \qquad \mathbf{e}_j^1=\sum_{n=1}^{j}b_{j,n}\mathbf{u}_n, \tag{3.6}$$

where we take $a_{j,j}=1=b_{j,j}$.

(2) From (3.6) and $\mathbf{e}_j^0\cdot \mathbf{T}_0=2^{-j+1}\mathbf{e}_j^0$, $\mathbf{e}_j^1\cdot \mathbf{T}_1=2^{-j+1}\mathbf{e}_j^1$, we derive

$$\begin{aligned} b_{j,n}&=\sum_{j=n}^{j}2^{j-j}\binom{k-1}{n-1}C_{k-n}b_{j,k},\\ a_{j,n}&=\sum_{k=m}^{j}\sum_{i=0}^{k-m}\binom{k-1}{m-1}\left(\frac{k-m}{i}\right)C_i2^{j-k}a_{j,k}. \end{aligned} \tag{3.7}$$

These equations determine $a_{j,n}, b_{j,n}$ recursively, starting from $a_{j,j}, b_{j,j}$. From (3.7) it follows that

$$a_{j,n}=\frac{(j-1)!}{(n-1)!}\alpha_{j-n}, \qquad b_{j,n}=\frac{(j-1)!}{(n-1)!}\beta_{j-n},$$

where $\alpha_0=1=\beta_0$, and the α_k, β_k satisfy the recursions

$$\beta_k=\sum_{i=0}^{k}\frac{1}{(k-i)!}C_{k-i}2^i\beta_i \tag{3.8a}$$

$$\alpha_k=\sum_{i=0}^{k}\sum_{n=i}^{k}2^i\frac{1}{(k-n)!(n-i)!}C_{k-n}\alpha_i \tag{3.8b}$$

(3) Together with $\alpha_0=1=\beta_0$, the recursions (3.8) imply that

$$\alpha_k=\sum_{i=0}^{k}(-1)^i\frac{1}{i!}\beta_{k-i} \tag{3.9}$$

or, equivalently,

$$\beta_k=\sum_{i=0}^{k}\frac{1}{i!}\alpha_{k-i}.$$

One way of checking this is to verify that if the β_i satisfy (3.8a), then the right-hand side of (3.9) satisfies (3.8b). It follows that

$$a_{j,n}=\sum_{k=n}^{j}\binom{j-1}{k-1}(-1)^{j-k}b_{k,n}.$$

This implies (3.4) and (3.5). □

Instead of the piecewise linear splines in § 2, here we shall use piecewise polynomial splines of degree $2l+1$, where l is the same as in (3.3). We shall determine the initial spline function f_0 for the iteration by fixing the values of $f_0^{(k)}(m)$, $0\leqq n\leqq N$, $0\leqq k\leqq l$. They will be defined in terms of the right eigenvectors of $\mathbf{T}_0, \mathbf{T}_1$ for the eigenvalues 1, $\frac{1}{2}, \cdots, 2^{-1}$. We first show that these eigenvalues are all simple.

LEMMA 3.3. *Assume that* (3.1) *and* (3.3) *hold. Then the* $l+1$ *eigenvalues* $1, \frac{1}{2}, \cdots, 2^{-l}$ *of* $\mathbf{T}_0, \mathbf{T}_1$ *are all simple.*

Proof. Since the e_j^0, $h=1,\cdots,L+1$, are left eigenvectors of $\mathbf{T}_0$ with eigenvalue 2^{-j+1}, there exists an appropriate basis transformation $\mathbf{B}$ such that $\mathbf{B}\mathbf{T}_0\mathbf{B}^{-1}$ has the form

$$\left(\begin{array}{cccc|ccc}
1 & 0 & \cdots & 0 & 0 & \cdots & 0\\
0 & \frac{1}{2} & & \vdots & \vdots & & \vdots\\
\vdots & & & 0 & \cdot & & \cdot\\
0 & \cdots & 0 & 2^{-L} & 0 & \cdots & 0\\
 & & \mathbf{A} & & & \mathbf{C} &
\end{array}\right),$$

where $\mathbf{A}$, $\mathbf{C}$ are $(N-L-1)\times(L+1)$- and $(N-L-1)\times(N-L-1)$-dimensional matrices, respectively. The roots of the characteristic equation for $\mathbf{T}_0$ are, therefore, $1, \frac{1}{2}, \cdots, 2^{-L}$ together with the roots of the characteristic equation for $\mathbf{C}$. It is easily checked that $\mathbf{C}=\mathbf{B}|_{E_{L+1}}\mathbf{T}_0|_{E_{L+1}}\mathbf{B}^{-1}|_{\mathbf{B}E_{L+1}}$; the spectral radius $\rho(\mathbf{C})$ of $\mathbf{C}$ can, therefore, be bounded by

$$\begin{aligned}
\rho(\mathbf{C})&=\limsup_{k\to\infty}\|\mathbf{C}^k\|^{1/k}\\
&\leqq\limsup_{k\to\infty}\left[\|\mathbf{B}\|_{E_{L+1}}\|\,\|\mathbf{T}_0^k|_{E_{L+1}}\|\,\|\mathbf{B}^{-1}|_{\mathbf{B}E_{L+1}}\|\right]^{1/k}\\
&\leqq\limsup_{k\to\infty}\left[\|\mathbf{B}|_{E_{L+1}}\|\,\|\mathbf{B}^{-1}|_{\mathbf{B}E_{L+1}}\|C\lambda^k2^{-lk}\right]^{1/k}\\
&=\lambda 2^{-l}<2^{-l}.
\end{aligned}$$

It follows that the eigenvalues $1, \frac{1}{2}, \cdots, 2^{-l}$ of $\mathbf{T}_0$ are also simple. A similar argument applies to $\mathbf{T}_1$. □

Since the eigenvalues $1, \frac{1}{2}, \cdots, 2^{-l}$ of $\mathbf{T}_0$ are also simple, it makes sense to introduce the corresponding right eigenvectors $\tilde{\mathbf{e}}_k^0$, $1\leqq k\leqq l+1$, which are uniquely determined, up to normalization. We fix their normalization by requiring

$$\mathbf{e}_k^0\cdot\tilde{\mathbf{e}}_{k'}^0=\delta_{kk'},\qquad 1\leqq k, k'\leqq l+1.$$

We similarly define right eigenvectors $\tilde{\mathbf{e}}_k^1$ of $\mathbf{T}_1$ corresponding to the eigenvalues $1, \frac{1}{2}, \cdots, 2^{-l}$, with normalization determined by

$$\mathbf{e}_k^1 \cdot \mathbf{e}_{k'}^1 = \delta_{kk'}, \qquad 1 \leqq k, k' \leqq l+1.$$

We now define the piecewise polynomial spline for f_0 of degree $2l+1$ by

$$f_0^{(k)}(n) = \begin{cases} (-1)^k k!(\tilde{\mathbf{e}}_{k+1}^0)_{n+1}, & k=0, \cdots, l, \\ & n=0, \cdots, N-1, \\ 0, & k=0, \cdots, l, \\ & n=N. \end{cases} \tag{3.10}$$

Remark. Because $(\mathbf{T}_0)_{in} = c_0 \delta_{n1}$, the vector $(1, 0, \cdots, 0)$ is a left eigenvector of $\mathbf{T}_0$ with eigenvalue c_0. It follows that $c_0 \notin \{1, \frac{1}{2}, \cdots, 2^{-l}\}$, since $c_0 = 2^{-k}$, $0 \leqq k \leqq l$ would imply the existence of two linearly independent left eigenvectors for 2^{-k}. Consequently, $(\tilde{\mathbf{e}}_{k+1}^0)_1 = (1, 0, \cdots, 0) \cdot \tilde{\mathbf{e}}_{k+1}^0 = 0$ for $0 \leqq k \leqq l$. (We can prove completely analogously that the right eigenvectors $\tilde{\mathbf{e}}_k^1$ of $\mathbf{T}_1$ satisfy $(\tilde{\mathbf{e}}_{k+1}^1)_N = 0$ for $k = 1, \cdots, l$.) For $n = 0$, (3.10), therefore, specializes to

$$f_0^{(k)}(0) = 0, \qquad k = 0, \cdots, l,$$

so that, despite appearances, there is complete symmetry between the constraints at zero and at N, the two ends of the support of f_0.

As in § 2, we again use the "folding" process that associates to a function g: $\mathbb{R} \to \mathbb{R}$, with support $[0, N]$, the vector valued function $\mathbf{w}$, $[0, 1] \to \mathbb{R}^N$ by means of

$$[\mathbf{w}(x)]_n = g(x+n-1), \qquad x \in [0, 1], \quad n = 1, \cdots, N.$$

The function g is l times continuously differentiable if and only if $\mathbf{w}$ is l times continuously differentiable on $[0, 1]$,

$$[\mathbf{w}^{(k)}(0)]_n = [\mathbf{w}^{(k)}(1)]_{n-1}, \qquad 2 \leqq n \leqq N, 0 \leqq k \leqq l, \tag{3.11}$$

$$[w^{(k)}(0)]_1 = 0 = [\mathbf{w}^{(k)}(1)]_N, \qquad 0 \leqq k \leqq l. \tag{3.12}$$

"Folding" f_0 leads to the vector valued function $\mathbf{v}_0$: $[0, 1] \to \mathbb{R}^N$, each component of which is a polynomial of degree $2l+1$; $\mathbf{v}_0$ satisfies (3.11) and (3.12). In particular, for $n = 2, \cdots, N$, $k = 0, \cdots, l$,

$$\begin{aligned} &[\mathbf{v}_0^{(k)}(0)]_n = [\mathbf{v}_0^{(k)}(1)]_{n-1} = (-1)^k k! [\tilde{\mathbf{e}}_{k+1}^0]_n, \\ &[\mathbf{v}_0^{(k)}(0)]_n = [\mathbf{v}_0^{(k)}(1)]_N = 0. \end{aligned} \tag{3.13}$$

Iterating the linear operator V defined by (2.3) on f_0 leads to a sequence f_j of piecewise polynomial splines of degree $2l+1$ and finer knot sets; their "folded" versions $\mathbf{v}_j$ can also be written as

$$\mathbf{v}_j = \mathbf{V}^j \mathbf{v}_0 \quad \text{or} \quad \mathbf{v}_j(x) = \mathbf{T}_{d_1(x)} \mathbf{T}_{d_2(x)} \cdots \mathbf{T}_{d_j(x)} \mathbf{v}_0(\tau^j x)$$

(see § 2). We now have the following lemma.

LEMMA 3.4. *For every $j \in \mathbb{N}$ and $0 \leqq k \leqq \min(2l+1, L)$,*

$$\mathbf{e}_{k+1}^0 \cdot \mathbf{v}_j(x) = (-1)^k x^k. \tag{3.14}$$

Proof. (1) We start by proving the assertion for $j=0$. By the construction of $\mathbf{v}_0$, $\mathbf{e}_k^0 \cdot \mathbf{v}_0(x) = P_k(x)$ is a polynomial on $[0,1]$ of degree $2l+1$. It is completely determined by the values of the $P_k^{(m)}(0)$, $P_k^{(m)}(1)$, $0 \leqq m \leqq l$. It, therefore, suffices to prove

$$P_k^{(m)}(0) = (-1)^k k!\,\delta_{km}, \tag{3.15}$$

$$P_k^{(m)}(1) = \begin{cases} (-1)^k \dfrac{k!}{(k-m)!} & \text{if } m \leqq k, \\ 0 & \text{if } m > k. \end{cases} \tag{3.16}$$

We have (use (3.13))

$$\begin{aligned} P_k^{(m)}(0) &= \mathbf{e}_{k+1}^0 \cdot \mathbf{v}_0^{(m)}(0) = (-1)^m m!\,\mathbf{e}_{k+1}^0 \cdot \tilde{\mathbf{e}}_{m+1}^0 \\ &= (-1)^k k!\,\delta_{km}, \end{aligned}$$

which proves (3.15), and

$$P_k^{(m)}(1) = \mathbf{e}_{k+1}^0 \cdot \mathbf{v}_0^{(m)}(1) = (-1)^m m! \sum_{n=1}^{N-1} (\mathbf{e}_{k+1}^0)_n (\tilde{\mathbf{e}}_{m+1}^0)_{n+1}. \tag{3.17}$$

Due to the special structure of $\mathbf{T}_0, \mathbf{T}_1$, we have

$$(\tilde{\mathbf{e}}_{m+1}^0)_{n+1} = (\tilde{\mathbf{e}}_{m+1}^1)_n.$$

Since $(\tilde{\mathbf{e}}_{m+1}^1)_N = 0$ (see the remark after (3.10)), (3.17) reduces to

$$P_k^{(m)}(1) = (-1)^m m!\mathbf{e}_{k+1}^0 \cdot \tilde{\mathbf{e}}_{m+1}^1,$$

which can be computed with the help of Lemma 3.2:

$$\begin{aligned} \mathbf{e}_{k+1}^0 \cdot \tilde{\mathbf{e}}_{m+1}^1 &= \sum_{r=0}^{k} (-1)^{k-r} \binom{k}{r} \mathbf{e}_{r+1}^1 \cdot \tilde{\mathbf{e}}_{m+1}^1 \\ &= \sum_{r=0}^{k} (-1)^{k-r} \binom{k}{r} \delta_{rm} \\ &= \begin{cases} 0 & \text{if } k < m, \\ (-1)^{k-m} \dbinom{k}{m} & \text{if } k \geqq m. \end{cases} \end{aligned}$$

Hence

$$P_k^{(m)}(1) = \begin{cases} (-1)^k \dfrac{k!}{(k-m)!} & \text{if } k \geqq m \\ 0 & \text{if } k < m, \end{cases}$$

which proves (3.16) and establishes (3.14) for $j=0$.

(2) For higher values of j we proceed by induction. Suppose (3.14) holds for $\mathbf{v}_j$. Then

$$\mathbf{e}_{k+1}^0 \cdot \mathbf{v}_{j+1}(x) = \mathbf{e}_{k+1}^0 \cdot \mathbf{T}_{d_1(x)} \mathbf{v}_j(\tau x). \tag{3.18}$$

If $x \leqq \frac{1}{2}$, then $d_1(x) = 0$ and (3.18) becomes

$$\begin{aligned} \mathbf{e}_{k+1}^0 \cdot \mathbf{v}_{j+1}(x) &= 2^{-k} \mathbf{e}_{k+1}^0 \cdot \mathbf{v}_j(2x) \\ &= 2^{-k}(-1)^k (2x)^k = (-1)^k x^k, \end{aligned}$$

which is (3.14) again for index $j+1$. If $x>\frac{1}{2}$, then we use Lemma 3.2 again:

$$\begin{aligned}\mathbf{e}^0_{k+1}\cdot\mathbf{v}_{j+1}(x)&=\sum_{r=0}^{k}(-1)^{k-r}\binom{k}{r}\mathbf{e}^1_{r+1}\cdot\mathbf{T}_1\mathbf{v}_j(2x-1)\\&=\sum_{r=0}^{k}(-1)^{k-r}\binom{k}{r}2^{-r}\sum_{s=0}^{r}\binom{r}{s}\mathbf{e}^0_{s+1}\cdot\mathbf{v}_j(2x-1)\\&=\sum_{s=0}^{k}(2x-1)^2\binom{k}{s}(-1)^2 2^{-s}\sum_{t=10}^{k-s}\binom{k-s}{t}(-1)^{k-s-t}2^{-t}\\&=(-2)^{-k}\sum_{s=0}^{k}\binom{k}{s}(2x-1)^s\\&=(-2)^{-k}(2x)^k=(-1)^k x^k,\end{aligned}$$

which is again (3.14). This proves the lemma. □

We need one more technical lemma before we attack the proof of Theorem 3.1.

LEMMA 3.5. *Assume that* (3.1) *and* (3.3) *hold. If* $\lambda>\frac{1}{2}$, *then there exists* $C>0$ *so that, for all binary sequences* $(d_j)_{j\in\mathbb{N}}$, *all* $m\in\mathbb{N}$, *and all* k *with* $l\leqq k\leqq L$,

$$\|\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_{k+1}}\|\leqq C\lambda^m 2^{-ml}. \tag{3.19}$$

If $\lambda=\frac{1}{2}$, *then* (3.19) *still holds for* $l+1\leqq k\leqq L$; *for* $k=l$, (3.19) *is replaced by the slightly weaker bound*

$$\|\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_{l+1}}\|\leqq Cm2^{-m(l+1)}. \tag{3.19$'$}$$

Note that we implicitly assume $l<L$; if $l=L$, then the lemma is trivial.

Proof. (1) We prove (3.19) by induction, working from high to low values of k. For $k=L$, (3.19) is (3.3), and we have nothing to prove.

(2) Because of the existence of the left eigenvectors $e^0_m, e^1_m, 1\leqq m\leqq L+1$ for $\mathbf{T}_0,\mathbf{T}_1$, with eigenvalues $1,\frac{1}{2},\cdots 2^{-L}$, and the relationships (3.4) and (3.5) between these eigenvectors, there exists an appropriate basis transformation $\mathbf{B}$ so that $(\mathbf{B}^{-1,t}\mathbf{e}^0_m)_r=\delta_{mr}$, and $\mathbf{B}E_m=\{\mathbf{w};\mathbf{w}_r=0 \text{ for } r\leqq m\}$. The matrices $\mathbf{B}\mathbf{T}_0\mathbf{B}^{-1}, \mathbf{B}\mathbf{T}_1\mathbf{B}^{-1}$ have the form

$$\mathbf{B}\mathbf{T}_0\mathbf{B}^{-1}=\left(\begin{array}{ccccc|c}1&0&\cdots&&0&\\0&\frac{1}{2}&&&\vdots&\\\vdots&\vdots&&&0&0\\0&0&\cdots&0&2^{-L}&\\\hline&&\mathbf{C}_0&&&\mathbf{A}_0\end{array}\right), \tag{3.20}$$

$$\mathbf{B}\mathbf{T}_1\mathbf{B}^{-1}=\left(\begin{array}{cccc|c}1&0&\cdots&0&\\t_{2,1}&\frac{1}{2}&&\vdots&\\\vdots&&&0&0\\t_{L+1,1}&\cdots&t_{L+1,L}&2^{-L}&\\\hline&&\mathbf{C}_1&&\mathbf{A}_1\end{array}\right). \tag{3.21}$$

The matrices $\mathbf{B}|_{E_{k+1}}\mathbf{T}_d|_{E_{k+1}}\mathbf{B}^{-1}|_{\mathbf{B}E_{k+1}}$ are then obtained by deleting the first $k+1$ rows and columns of (3.20), (3.21). Let us denote these $(N-k-1)\times(N-k-1)$-dimensional matrices by $\mathbf{S}^k_d$. It then follows that, for $l\leqq k\leqq L-1$

$$\mathbf{S}^k_d=\left(\begin{array}{c|c}2^{-k-1}&0\ \cdots\ 0\\\hline\mathbf{C}^k_d&\mathbf{S}^{k+1}_d\end{array}\right) \tag{3.22}$$

where $\mathbf{C}^k_0$, $\mathbf{C}^k_1$ are $(N-k-2)$-dimensional column vectors.

(3) Let us now assume that (3.19) holds for index $k+1 \geqq l+1$. Then

$$\begin{aligned}\|\mathbf{S}_{d_1}^{k+1} \cdots \mathbf{S}_{d_m}^{k+1}\| &\leqq \|\mathbf{B}|_{E_{k+2}}\| \, \|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_m}|_{E_{k+2}}\| \, \|\mathbf{B}^{-1}|_{\mathbf{B}E_{k+2}}\| \\ &\leqq C\|\mathbf{B}^{-1}\| \, \|\mathbf{B}\|\lambda^m 2^{-ml} \leqq C_{k+1}\lambda^m 2^{-ml}.\end{aligned}$$

From (3.22) it follows that

$$\text{(3.23)} \quad \mathbf{S}_{d_1}^k \cdots \mathbf{S}_{d_m}^k = \left(\begin{array}{c|c} 2^{-(k+1)m} & 0 \quad 0 \quad \cdots \quad 0 \\ \hline \sum\limits_{r=1}^{m} 2^{-(k+1)(m-r)}\mathbf{S}_{d_1}^{k+1} \quad \cdots \quad \mathbf{S}_{d_{r-1}}^{k+1}\mathbf{C}_{d_r}^k & \mathbf{S}_{d_1}^{k+1} \cdots \mathbf{S}_{d_m}^{k+1} \end{array}\right),$$

where

$$\begin{aligned}&\left\| \sum_{r=1}^{m} 2^{-(k+1)(m-r)}\mathbf{S}_{d_1}^{k+1} \cdots \mathbf{S}_{d_{r-1}}^{k-1}\mathbf{C}_{d_r}^k \right\| \\ &\quad \leqq C_{k+1} \sum_{r=1}^{m} 2^{-(k+1)(m-r)}(\lambda 2^{-l})^{r-1} \cdot \max(\|\mathbf{C}_0^k\|, \|\mathbf{C}_1^k\|) \\ &\quad \leqq C' 2^{-(k+1)(m-1)} \frac{(\lambda 2^{k+1-l})^m - 1}{\lambda 2^{k+1-l} - 1} \\ &\quad \leqq C''\lambda^m 2^{-lm},\end{aligned}$$

where we have assumed that $\lambda 2^{k+1-l} > 1$, i.e., $k \geqq l+1$ or $\lambda > \frac{1}{2}$. We come back below to the case $k = l+1$, $\lambda = \frac{1}{2}$.

It follows that the three pieces in (3.23) are all bounded by $C'''\lambda^m 2^{-lm}$ (since $k \geqq l+1$ and $\lambda \geqq \frac{1}{2}$). There exists, therefore, C_k (independent of m or the d_j) so that

$$\|\mathbf{S}_{d_1}^k \cdots \mathbf{S}_{d_m}^k\| \leqq C_k \lambda^m 2^{-ml}.$$

This proves the induction step if $k \geqq l+1$ or $\lambda > \frac{1}{2}$.

(4) We now treat the case $k = l+1$, $\lambda = \frac{1}{2}$. In this case we find

$$\begin{aligned}\left\| \sum_{r=1}^{m} 2^{-k(m-r)}\mathbf{S}_{d_1}^{k+1} \cdots \mathbf{S}_{d_{r-1}}^{k+1}\mathbf{C}_{d_r}^k \right\| &\leqq C_{k+1} \sum_{r=1}^{m} 2^{-(l+1)(m-r)} 2^{-(l+1)(r-1)} \\ &= C_{k+1} 2^{-(l+1)(m-1)}(m-1).\end{aligned}$$

Therefore, the three pieces in (3.23) can all be bounded by $C' 2^{-(l+1)m} m$, which proves (3.19′). □

We are now ready to attack the proof of Theorem 3.1.

Proof of Theorem 3.1. (1) As a consequence of Lemma 3.4 we have, for $0 \leqq k \leqq l$, and all $j, j' \in \mathbb{N}$,

$$\mathbf{e}_{k+1}^0 \cdot [\mathbf{v}_j(x) - \mathbf{v}_{j'}(x)] = 0,$$

hence $\mathbf{v}_j(x) - \mathbf{v}_{j'}(x) \in E_{l+1}$ for all $x \in [0, 1]$, all $j, j' \in \mathbb{N}$.

(2) It follows that

$$\begin{aligned}\|\mathbf{v}_{j+1}(x) - \mathbf{v}_j(x)\| &= \|\mathbf{T}_{d_1(x)} \cdots \mathbf{T}_{d_j(x)}[\mathbf{v}_1(\tau^j x) - \mathbf{v}_0(\tau^j x)]\| \\ &\leqq C\gamma_\lambda(j)\lambda^j 2^{-lj} \sup_{y \in [0,1]} \|\mathbf{v}_1(y) - \mathbf{v}_0(y)\|,\end{aligned}$$

where

$$\gamma_\lambda(j) = \begin{cases} 1 & \text{if } \lambda > \frac{1}{2}, \\ j & \text{if } \lambda = \frac{1}{2}. \end{cases}$$

Consequently,

$$\begin{aligned}\|\mathbf{v}_j(x)\| &\leqq \|\mathbf{v}_0(x)\| + \sum_{r=0}^{j-1} \|\mathbf{v}_{r+1}(x) - \mathbf{v}_r(x)\| \\ &\leqq C\left[1 + \sum_{r=0}^{j-1} \gamma_\lambda(r)\lambda^r 2^{-lr}\right],\end{aligned}$$

which is bounded uniformly in j and x.

(3) We now use this uniform bound to prove that the $\mathbf{v}_j$ constitute a Cauchy sequence in L^∞. We have

$$\begin{aligned}\|\mathbf{v}_{j+r}(x) - \mathbf{v}_j(x)\| &= \|\mathbf{T}_{d_1(x)} \cdots \mathbf{T}_{d_j(x)}[\mathbf{v}_r(x) - \mathbf{v}_0(x)]\| \\ &\leqq C\gamma_\lambda(j)\lambda^j 2^{-lj},\end{aligned}$$

which can be made arbitrarily small by choosing j large enough, independent of r. The $\mathbf{v}_j$ tend, therefore, to a limit $\mathbf{v}$, which satisfies

$$\|\mathbf{v}(x) - \mathbf{v}_j(x)\| \leqq C\gamma_\lambda(j)\lambda^j 2^{-lj}. \tag{3.24}$$

(4) Since all the $\mathbf{v}_j$ satisfy

$$\begin{aligned}&[\mathbf{v}_j(0)]_{n+1} = [\mathbf{v}_j(1)]_n, \qquad 1 \leqq n \leqq N-1, \\ &[\mathbf{v}_j(0)]_1 = 0 = [\mathbf{v}_j(1)]_N,\end{aligned}$$

so does $\mathbf{v}$. It follows that $\mathbf{v}$ can be "unfolded" into a continuous function f, for which (3.24) translates into

$$\|f - f_j\|_{L^\infty} \leqq C\phi_\lambda(2^{-j}), \tag{3.25}$$

where $\phi_\lambda(t) = t^{l+\alpha}$ if $\lambda > \frac{1}{2}$, $\phi_\lambda(t) = |\log_2 t| t^{l+1}$ if $\lambda = \frac{1}{2}$, with $\alpha = |\ln \lambda|/\ln 2$.

(5) Formula (3.25) tells us again how well f can be approximated by piecewise polynomials, and this can be translated into smoothness estimates on f. This result is no doubt well-known to spline experts, but we could not find in the literature a full proof of the exact result we needed. For convenience's sake, we offer the following proof "from scratch."

Note first that, by Lemma 3.4,

$$\mathbf{e}^0_{k+1} \cdot \mathbf{v}_j^{(l)}(x) = \begin{cases} 0 & \text{if } k < l, \\ (-1)^l l! & \text{if } k = 1. \end{cases}$$

It follows that $\mathbf{v}_j^{(l)}(x) - \mathbf{v}_{j'}^{(l)}(x') \in E_{l+1}$ for all j, j', all x, x'. On the other hand, the recursive definition of the $\mathbf{v}_j$ leads to

$$\mathbf{c}_j^{(l)}(x) = 2^l \mathbf{T}_{d_1(x)} \mathbf{v}_{j-1}^{(l)}(\tau x).$$

Together, these two observations imply

$$\begin{aligned}\|\mathbf{v}_{j+1}^{(l)}(x) - \mathbf{v}_j^{(l)}(x)\| &= 2^{lj} \|\mathbf{T}_{d_1(x)} \cdots \mathbf{T}_{d_j(x)}[\mathbf{v}_1^{(l)}(\tau^j x) - \mathbf{v}_0^{(l)}(\tau^j x)]\| \\ &\leqq C\lambda^j \gamma_\lambda(j) \sup_{y \in [0,1]} [\|\mathbf{v}_1^{(l)}(y)\| + \|\mathbf{v}_0^{(l)}(y)\|],\end{aligned}$$

which can be used, similarly to the argument in (2) above, to prove that the $\mathbf{v}_j^{(l)}(x)$ are bounded uniformly in j and x, and that the $\mathbf{v}_j^{(l)}$ are a Cauchy sequence in L^∞. The limit of the $\mathbf{v}_j^{(l)}$ is necessarily the lth derivative of $\mathbf{v}$. Moreover,

$$\|\mathbf{v}^{(l)}(x) - \mathbf{v}_j^{(l)}(x)\| \leqq C\lambda^j \gamma_\lambda(j),$$

uniformly in x.

The remainder of our argument is similar to (7) in the proof of Theorem 2.3. Take $x \leqq y$ in $[0, N]$ so that $2^{-(j+1)} \leqq y - x \leqq 2^{-j}$ (a slight extension of the argument can be used on a neighborhood of $[0, N]$, so that the results are also true at zero and N). Then there exists $m \in \mathbb{N}$ so that either $(m-1)2^{-j} \leqq x \leqq y \leqq m2^{-j}$ or $(m-1)2^{-j} \leqq x \leqq m2^{-j} \leqq y \leqq (m+1)2^{-j}$. We discuss the second case; the first is similar. Then

$$|f^{(l)}(x) - f^{(l)}(y)| \leqq 2C\lambda^j \gamma_\lambda(j) + |f_j^{(l)}(x) - f_j^{(l)}(m2^{-j})| + |f_j^{(l)}(y) - f_j^{(l)}(m2^{-j})|,$$

by the bound on $\mathbf{v}^{(l)} - \mathbf{v}_j^{(l)}$. There exists n so that $x' = x - n$ and $m'2^{-j} = m2^{-j} - n$ are both in $[0, 1]$; moreover, we can choose binary expansions for x' and $m'2^{-j}$ with the same j first digits. Consequently,

$$\begin{aligned} |f_j^{(l)}(x) - f_j^{(l)}(m2^{-j})| &\leqq \|\mathbf{v}_j^{(l)}(x') - \mathbf{v}_j^{(l)}(m'2^{-j})\| \\ &= 2^{lj} \|\mathbf{T}_{d_1(x)} \cdots \mathbf{T}_{d_j(x)}[\mathbf{v}_0^{(l)}(\tau^j x') - \mathbf{v}_0^{(l)}(\tau^j(m'2^{-j}))]\| \\ &\leqq C\lambda^j \gamma_\lambda(j), \end{aligned}$$

where we have used (3.3), the uniform boundedness of $\mathbf{v}_0^{(l)}(y)$, and $\mathbf{v}_0^{(l)}(y) - \mathbf{v}_0^{(l)}(y') \in E_{l+1}$ for all y, y'. A similar bound holds for $|f_j^{(l)}(y) - f_j^{(l)}(m2^{-j})|$. Putting it all together, we obtain, for $2^{-(j+1)} \leqq |x - y| \leqq 2^{-j}$,

$$|f^{(l)}(x) - f^{(l)}(y)| \leqq C'\lambda^j \gamma_\lambda(j),$$

which translates to

$$|f^{(l)}(x) - f^{(l)}(y)| \leqq C|x - y|^{|\log_2 \lambda|}$$

if $\lambda > \frac{1}{2}$, and to

$$|f^{(l)}(x) - f^{(l)}(y)| \leqq C|\log_2 |x - y|| \, |x - y|$$

if $\lambda = \frac{1}{2}$. □

Remarks. (1) If $l = L$, then we would not need Lemma 3.5, so that the assumption $\lambda \geqq \frac{1}{2}$ would never be used. The argument of (7) would work if (3.3) held for $\lambda < \frac{1}{2}$, but it would lead to a Hölder exponent α larger than 1 for $f^{(l)}$, hence to $f^{(l)} \equiv 0$. This is incompatible with the fact that f is compactly supported, except if $f \equiv 0$. It follows that for matrices $\mathbf{T}_0, \mathbf{T}_1$ constructed as in (2.8), the constant λ in (3.3) is necessarily $\geqq \frac{1}{2}$ if $l = L$.

(2) We chose f_0 so that $\mathbf{v}_0^{(k)}(0) = (-1)^k k! \tilde{\mathbf{e}}_{k+1}^0$ (see (3.13)). Since, for $x < 2^{-j}$,

$$\mathbf{v}_j^{(k)}(x) = 2^{kj} \mathbf{T}_0^j \mathbf{v}_0^{(k)}(2^j x),$$

it follows that

$$\begin{aligned} \mathbf{v}_j^{(k)}(0) &= (-1)^k k! 2^{kj} \mathbf{T}_0^j \tilde{\mathbf{e}}_{k+1}^0 \\ &= (-1)^k k! \tilde{\mathbf{e}}_{k+1}^0, \end{aligned}$$

where we have used $\mathbf{T}_0 \tilde{\mathbf{e}}_{k+1}^0 = 2^{-k} \tilde{\mathbf{e}}_{k+1}^0$. Hence the $f_j^{(k)}(n)$, $0 \leqq k \leqq l$, are independent of j. Because of the bounds (3.25), this implies also, for $0 \leqq k \leqq l$,

$$\begin{aligned} &f^{(k)}(n) = f_j^{(k)}(n) = (-1)^k k! [\tilde{\mathbf{e}}_{k+1}^0]_{n+1}, \qquad 0 \leqq n \leqq N-1, \\ &f^{(k)}(N) = 0. \end{aligned} \tag{3.26}$$

In § I.5 we had already seen that the $f^{(k)}(n)$ were linked to the right eigenvectors of $\mathbf{M}$ for the eigenvalues 2^{-k}, but it wasn't clear how to choose their normalization, explicit in (3.26). Note that $f_0^{(k)}(n) = f^{(k)}(n)$ for $0 \leqq k \leqq l$ implies that $f_j^{(k)}$ converges to $f^{(k)}$ for

$0 \leqq k \leqq l$ (see Theorem 4.2 in part I): not only do the f_j themselves converge, their derivatives up to lth order do as well.

(3) By continuity, Lemma 3.4 carries over to $\mathbf{v}(x)$: for $0 \leqq k \leqq \min(2l+1, L)$,

$$\mathbf{c}^0_{k+1} \cdot \mathbf{v}(x) = (-1)^k x^k. \tag{3.27}$$

The following lemma states that (3.27) holds for all $k \leq L$, even if $2l+1 < L$.

LEMMA 3.6. *For all* $0 \leqq k \leqq L$, $\mathbf{e}^0_{k+1} \cdot \mathbf{v}(x) = (-1)^k x^k$.

Proof. (1) If $2l+1 \geqq L$, then we have nothing to prove. We assume $2l+1 < L$ and restrict ourselves to $k > 2l+1$.

(2) We only need to prove the lemma for dyadic rationals x; by continuity it then holds for all $x \in [0, 1]$. Take, therefore, $x = n2^{-j}$. The proof works by induction on j.

(3) If $j = 0$, then only $n = 0, 1$ lead to $x \in [0, 1]$. Since $\mathbf{v}(x) = \mathbf{T}_{d_1(x)}\mathbf{v}(\tau x)$, we have $\mathbf{v}(0) = \mathbf{T}_0\mathbf{v}(0)$, $\mathbf{v}(1) = \mathbf{T}_1\mathbf{v}(1)$. Consequently,

$$\mathbf{e}^0_{k+1} \cdot \mathbf{v}(0) = \mathbf{e}^0_{k+1} \cdot \mathbf{T}_0\mathbf{v}(0) = 2^{-k}\mathbf{e}^0_{k+1} \cdot \mathbf{v}(0);$$

hence $\mathbf{e}^0_{k+1} \cdot \mathbf{v}(0) = 0$. Similarly, $\mathbf{e}^1_{k+1} \cdot \mathbf{v}(1) = 0$, which implies (use Lemma 3.2)

$$\begin{aligned} \mathbf{e}^0_{k+1} \cdot \mathbf{v}(1) &= \sum_{m=0}^{k} (-1)^{k-m} \binom{k}{m} \mathbf{e}_{m+1} \cdot \mathbf{v}(1) \\ &= (-1)^k \mathbf{e}^1_1 \cdot \mathbf{v}(1) = (-1)^k, \end{aligned}$$

where we have used (3.27) for $k = 0$. This proves $\mathbf{e}^0_{k+1} \cdot \mathbf{v}(x) = (-1)^k x^k$ for all dyadic rations $x = n2^{-j}$ with $j = 0$.

(4) Suppose that the lemma holds for all $x = n2^{-j}$, j fixed, $0 \leqq n \leqslant 2^j$. Take y of the form $r2^{-j-1}$, $r \in \mathbb{N}$, $0 \leqq r \leqq 2^{j+1}$. If $y \leqq \frac{1}{2}$, then

$$\begin{aligned} \mathbf{e}^0_{k+1} \cdot \mathbf{v}(y) &= \mathbf{e}^0_{k+1} \cdot \mathbf{T}_0\mathbf{v}(2y) \\ &= 2^{-k}\mathbf{e}^0_{k+1} \cdot \mathbf{v}(2y) = (-1)^k 2^{-k}(2y)^k \quad \text{(by induction)} \\ &= (-1)^{\mathrm{k}}\mathrm{y}^{\mathrm{k}}. \end{aligned}$$

If $y \geqq \frac{1}{2}$, then

$$\mathbf{e}^0_{k+1} \cdot \mathbf{v}(y) = \mathbf{e}^0_{k+1} \cdot \mathbf{T}_1\mathbf{v}(2y-1).$$

As in the proof of Lemma 3.4, we have

$$\mathbf{e}^0_{k+1}\mathbf{T}_1 = 2^{-k} \sum_{s=1}^{k} \binom{k}{s} (-1)^{k-s} \mathbf{e}^0_{s+1}.$$

Hence

$$\begin{aligned} \mathbf{e}^0_{k+1} \cdot \mathbf{v}(y) &= 2^{-k} \sum_{s=0}^{k} \binom{k}{s} (-1)^{k-s} (1-2y)^s \quad \text{(by induction)} \\ &= 2^{-k}(-2y)^k = (-y)^k. \end{aligned}$$

This proves the lemma. □

As in § 2, the condition (3.3) can be reduced to a condition involving only a finite number of matrix norms or interpreted as a "spectral" constraint:

$$\begin{aligned} (3.3) &\Leftrightarrow \hat{\rho}(\mathbf{T}_0|_{E_{L+1}}, \mathbf{T}_1|_{E_{L+1}}) < 2^{-l}, \\ &\Leftrightarrow \exists m \in \mathbb{N} \quad \text{so that} \\ \gamma^L_m &= \max_{\substack{d_j = 0 \text{ or } 1 \\ j = 1, \cdots, m}} \|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_m}|_{E_{L+1}}\|^{1/m} < 2^{-l}. \end{aligned} \tag{3.28}$$

If (3.28) holds for some m, then (3.3) holds with $\lambda = 2^l \gamma^L_m$.

Even the condition $\gamma_m^L < 2^{-l}$ for some $m \in \mathbb{N}$ may be hard to check in practice: for any fixed m, it is necessary to verify the norms of 2^m matrices. Some simplification may occur if the c_n are symmetric, $c_j = c_{N-j}$, but even then the number of matrices is huge if m is large. In the case of the Lagrangian interpolation scheme with dilation factor 2 and 5 nodes (see § 5), we find that $\gamma_m^1 < \frac{1}{2}$ only if $m \geqq 7$. Because the c_n are symmetric in this case, the number of matrices to check is only 2^6 instead of 2^7, but that still amounts to a large number of computations for a fairly simple example. We can reduce the number of computations significantly by some simple tricks listed in the following proposition.

PROPOSITION 3.7. *The following are all equivalent to* (3.3).

(1^0) *For some* $m \in \mathbb{N}$,

$$\gamma_m^L < 2^{-l}. \tag{3.29}$$

(2^0) *For some* $N \times N$*-matrix* $\mathbf{B}$ *with det* $\mathbf{B} \neq 0$, *and for some* $m \in \mathbb{N}$,

$$\gamma_{m;\mathbf{B}}^L < 2^{-l}, \tag{3.30}$$

where

$$\gamma_{m;\mathbf{B}}^k = \max_{\substack{d_j = 0 \text{ or } 1 \\ j = 1, \cdots, m}} \|\bar{\mathbf{T}}_{d_1} \cdots \bar{\mathbf{T}}_{d_m}|_{\mathbf{B}E_{k+1}}\|^{1/m},$$

with $\bar{\mathbf{T}}_d = \mathbf{B}\mathbf{T}_d\mathbf{B}^{-1}$.

(3^0) *There exists a finite collection of "building blocks"* $D^j, j = 1, \cdots, J$, *with* $D^j = \{d_1^j, d_2^j, \cdots, d_{k_j}^j\}$, *each* $d_n^j = 0$ *or* 1, *which is complete, in the sense that every dyadic sequence can be written as a sequence of blocks* D^j *and for which*

$$\max_{j = 1, \cdots, J} t_{D^j}^L < 2^{-l}, \tag{3.31}$$

where $t_{D^j}^m = \|\mathbf{T}_{d_1^j} \cdots \mathbf{T}_{d_{k_j}^j}|_{E_{m+1}}\|^{1/k_j}$.

Proof. (1) The proof of the equivalence (3.3)⇔(3.29) is similar to the proof of Lemma 2.3.

(2) We prove (3.29)⇔(3.30). Since

$$\|\bar{\mathbf{T}}_{d_1} \cdots \bar{\mathbf{T}}_{d_n}|_{\mathbf{B}E_{L+1}}\| \leqq \|\mathbf{B}\| \, \|\mathbf{B}^{-1}\| \, \|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_n}|_{\mathbf{B}E_{L+1}}\|,$$

we have $\gamma_{n,\mathbf{B}}^L \leqq [\|\mathbf{B}\| \, \|\mathbf{B}^{-1}\|]^{1/n} \gamma_n^L$. By the same argument used in the proof of Lemma 2.7, $\gamma_n^L \leqq C^{1/n} \gamma_m^L$ for some $C > 0$. If $\gamma_m^L < 2^{-l}$, it follows, therefore, that $\gamma_{n;\mathbf{B}}^L < 2^{-l}$ for large enough n, This proves (3.29)⇒(3.30). The converse implication is proved in the same way.

(3) The implication (3.29)⇒(3.31) is obvious: it suffices to take the 2^m "building blocks," each consisting of m entries, with every entry zero or 1.

(4) Finally, we prove (3.31)⇒(3.29), which completes the proof. Suppose (3.31) is satisfied. Define $K = \max\{k_j; j = 1, \cdots, J\}$, $C = \max\{\|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_n}|_{E_{L+1}}\|; n \leqq K, d_j = 0$ or 1 for each $j\}$. Take now any fixed sequence $(d_j)_{j \in \mathbb{N}}$, and $n \in \mathbb{N}$. The n-tuple $(d_1 \cdots d_n)$ can be written as a sequence of building blocks D^j, followed by a stretch of entries that has at most length K,

$$(d_1 \cdots d_n) = (D^{j_1} \cdots D^{j_n} \tilde{D}).$$

If $\alpha = \max t_{D_j}^L$, and $|D^j|$ denotes the number of entries in D^j, then we have, therefore,

$$\begin{aligned} \|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_m}|_{E_{L+1}}\| &\leqq 2^{-l(|D^{jm}| + \cdots + |D^{jm}|)} \alpha^{|D^{j1}| + \cdots + |D^{jm}|} C \\ &\leqq 2^{-lm} \alpha^m C 2^{lK} \alpha^{-K}, \end{aligned}$$

which amounts to (3.3). □

Note that the bounds on γ_m^L, on $\gamma_{m;\mathbf{B}}^L$, or on the $t_{D^j}^L$ all lead to lower bounds for the Hölder exponent of $f^{(L)}$, since (3.29)-(3.31) are equivalent with (3.3), with $2^l\gamma_m^L$, $2^l\gamma_{m;\mathbf{B}}^L$, or $2^l \max t_{D^j}^L$ playing the role of λ.

It is, of course, possible to combine (3.30) and (3.31) and to define $\bar{t}_{D_j}^L$ (for the matrices $\bar{\mathbf{T}}_d = \mathbf{B}\mathbf{T}_d\mathbf{B}^{-1}$). A bound on the $\bar{t}_{D_j}^L$, similar to (3.31), is again equivalent with (3.3).

Remark. The proof of Theorem 3.1 presented here relies on arguments which translate convergence rate to f by spline functions into regularity estimates for f. There exist generalizations of lattice two-scale difference equations to higher dimensions for which it does not seem possible to find appropriate spline functions playing the role of the f_j here (one such generalization is outlined in the Appendix). For this reason we present a sketch of a different proof in the Appendix, a fully detailed version which would be longer and more complicated than the proof in this section, but which does not use spline functions.

4. Local regularity and fractal sets. We assume again that the c_n, $n=0,\cdots,N$ satisfy the $L+1$ sum rules (3.2) (with $L+1<N$), as well as $\sum_{n=0}^N c_n = 2$. If (3.3) is satisfied, then Theorem 3.1 tells us that f is l times continuously differentiable, and that $f^{(l)}$ is Hölder continuous with exponent $|\ln\lambda|/\ln 2$. The proof for the uniform Hölder continuity of $f^{(l-1)}$ uses that, for any point x,

$$\|\mathbf{T}_{d_1(x)}\cdots\mathbf{T}_{d_m(x)}|_{E_{L+1}}\| \leqq 2^{-ml}\lambda^m. \tag{4.1}$$

In some cases, more accurate bounds on the left-hand side of (4.1) can be obtained for some values of x, depending on the relative frequency of the digits 0 and 1 in the binary expansion of x. This can then be used to compute local Hölder exponents, which may be larger than the uniformly valid Hölder exponents for $f^{(l-1)}$. More precisely, we have the following theorem.

THEOREM 4.2. *Assume that the c_n, $n=0,\cdots N$, satisfy the $L+1$ sum rules* (3.2), *and assume that $\sum_{n=0}^N c_n = 2$, with $L<N-1$. Assume that there exist $m\in\mathbb{N}$, and μ_0, $\mu_1<1$ such that*

$$\|\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_{L+1}}\| \leqq 2^{-\lambda}\mu_1^{s_m}\mu_0^{m-s_m} \tag{4.2}$$

for all possible combinations of d_j, $d_j=0$ or $1, j=1,\cdots,m$, with $s_m=\sum_{j=1}^m d_j$.

Let f be the solution f of the two-scale difference equation (1.2) *associated with the c_n, and let $f^{(l)}$ be its lth derivative, which exists and is Hölder continuous, with exponent* $\min(|\ln\mu_1|, |\ln\mu_2|)/\ln 2$, *by Theorem* 3.1. *Take $x\in[0,N]$. Assume that the decimal part of the binary expansion of x satisfies*

$$r_m(x)=\frac{1}{m}\sum_{j=1}^m d_j(x) \text{ tends to a limit } r(x) \text{ as } m\to\infty, \tag{4.3}$$

$$0<r(x)<1. \tag{4.4}$$

Then the following holds:

(1) *For all $\varepsilon>0$, there exist $\delta>0$ and $C<\infty$ such that*

$$|f^{(L)}(x)-f^{(L)}(x+t)| \leqq C|t|^{\alpha(x)-\varepsilon} \quad \text{if } |t|<\delta, \tag{4.5}$$

where $\alpha(x) = -\min\{1, [(1-r(x))|\ln\mu_0| + r(x)|\ln\mu_1|]/\ln 2\}$, *with* $r(x)=r(x-\lfloor x\rfloor)$ *as defined by* (4.4).

(2) *If* $\mu_0 \geqq \frac{1}{2} > \mu_1$ *and* $r(x) > (\ln 2 - |\ln\mu_0|)/|ln\,\mu_1| - |\ln\mu_0|)$, *or if* $\mu_1 \geqq \frac{1}{2} > \mu_0$ *and* $r(x) < (|\ln\mu_0| - \ln 2)/(|\ln\mu_0| - |\ln\mu_1|)$, *then $f^{(l)}$ is differentiable in x.*

To prove this theorem we need the following lemma.

LEMMA 4.2. *Take $x \in [0, 1]$. Assume that (4.3) and (4.4) are satisfied. Then, for all $\varepsilon > 0$, there exists an N_ε so that, for all $m \geqq N_\varepsilon$, the m first digits of the binary expansions of x and $x+t$ are identical whenever $|t| < 2^{-m(1+\varepsilon)-1}$.*

Proof. (1) Fix $\varepsilon > 0$. Choose δ small enough so that

$$[1-r(x)-\delta]^{-1} 2\delta \leqq \varepsilon \quad \text{and} \quad [r(x)-\delta]^{-1} 2\delta \leqq \varepsilon. \tag{4.6}$$

There exists N_ε so that $|r_m(x) - r(x)| \leqq \delta$ for $m \geqq N_\varepsilon$.

(2) Now choose $m \geqq N_\varepsilon$ and $0 \leqq t < 2^{-m(1+\varepsilon)-1}$. If $d_{m+1}(x) = 0$, then it follows from $t < 2^{-m-1}$ that $d_j(x) = d_j(x+t)$ for all $j \leqq m$, and we are done. Suppose that $d_{m+1}(x) = 1 = d_{m+2}(x) = \cdots = d_{m+s}(x)$, and $d_{m+s+1}(x) = 0$. Then the condition (4.6) implies an upper bound on s, since $r_{m+s}(x) \leqq r(x) + \delta$, while $r_m(x) \geqq r(x) - \delta$. Using $r_{m+s}(x) = (m+s)^{-1}[mr_m(x) + s]$ together with these restrictions leads to $s \leqq m2\delta[1-r(x)-\delta]^{-1} \leqq m\varepsilon$. Therefore, $t < 2^{-m(1+\varepsilon)-1} \leqq 2^{-(m+s)-1}$. Since $d_{m+s+1}(x) = 0$, this implies $d_j(x) = d_j(x+t)$ for all $j \leqq m+s$.

(3) The argument for $-2^{-m(1-\varepsilon)-1} < t \leqq 0$ is similar. If $d_{m+1}(x) = 1$, then $d_j(x) = d_j(x+t)$ for all $j < m$ because $|t| < 2^{-m-1}$. If $d_{m+1}(x) = 0 = d_{m+2}(x) = \cdots = d_{m+s}(x)$, and $d_{m+s+1}(x) = 1$, then $s \leqq m2\delta[r(x)-\delta]^{-1} \leqq m\varepsilon$. Hence $t > -2^{-(m+s)-1}$, which implies $d_j(x) = d_j(x+t)$ for all $j \leqq m+s$. □

We now proceed with the proof of Theorem 4.1.

Proof of Theorem 4.1. (1) Since $\mu_0, \mu_1 < 1$, (4.2) implies (3.29), hence (3.3), so that f is l times continuously differentiable. By Lemma 3.6 its "folded" version $\mathbf{v}$ satisfies

$$\mathbf{e}^0_{k+1} \cdot \mathbf{v}^{(l)}(y) = (-1)^l l! \, \delta_{lk} \quad \text{for } 0 \leqq k \leqq l,$$

which implies

$$\mathbf{v}^{(l)}(y) - \mathbf{v}^{(l)}(y') \in E_{l+1} \quad \text{for all } y, y' \in [0, 1].$$

(2) It follows from (4.2) that there exists $C > 0$ so that, for all $p \in \mathbb{N}$,

$$\|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_p}|_{E_{L+1}}\| \leqq C 2^{-lp} \mu_1^{s_p} \mu_0^{p-s_p} \tag{4.7}$$

(by the same argument as in the proof of Lemma 2.3). The same arguments used in the proof of Lemma 3.5 can then be used to derive that, for $p \geqq 1$,

$$\|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_p}|_{E_{l+1}}\| \leqq Cp 2^{-lp} \mu_1^{s_p} \mu_0^{p-s_p}, \tag{4.8}$$

where the constant C may be different from that in (4.7), and where we have introduced the extra p for the case where $\mu_1^{s_n} \mu_0^{n-s_n} = 2^{-n}$ for some n.

(3) Fix $x \in [0, N]$, such that $r(x)$ is well defined and $0 < r(x) < 1$. Note that this excludes all dyadic rationals x, since these have $r(x) = 0$ or 1, depending on which binary expansion is chosen (for dyadic rationals, there are two binary expansions). Fix $\varepsilon > 0$. There exists $\gamma > 0$ so that

$$\begin{aligned} &[(1-r(x)-\gamma/2)|\ln \mu_0| + (r(x)-\gamma/2)|\ln \mu_1|]/(1+\gamma) \\ &\qquad \geqq (1-r(x))|\ln \mu_0| + r(x)|\ln \mu_1| - \varepsilon/2 = \alpha(x) - \varepsilon/2. \end{aligned} \tag{4.9}$$

Define $\delta_1 = 2^{-N_\gamma(1+\gamma)-1}$, where N_γ is chosen as in the proof of Lemma 4.2.

(4) We have $x = n + y$, with $0 < y < 1$, $n = 0, 1, \cdots$, or $N-1$. Take $|t| < \delta = \min(\delta_1, x-n, n+1-x)$. Then $\lfloor x+t \rfloor = \lfloor x \rfloor = n$. It follows that

$$\begin{aligned} |f^{(l)}(x) - f^{(l)}(x+t)| &= |\mathbf{v}^{(l)}_{n+1}(x-n) - \mathbf{v}^{(l)}_{n+1}(x+t-n)| \\ &\leqq \|\mathbf{v}^{(l)}(y) - \mathbf{v}^{(l)}(y+t)\|. \end{aligned}$$

(5) Since $|t|<\delta$, there exist $p \geqq N_\gamma$ so that

$$2^{-(p+1)(1+\gamma)-1} \leqq |t| < 2^{-p(1+\gamma)-1}. \tag{4.10}$$

By Lemma 4.2, $d_j(y)=d_j(y+t)$ for $j \leqq p$. Hence

$$\mathbf{v}^{(l)}(y)-\mathbf{v}^{(l)}(y+t)=2^{pl}\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}[\mathbf{v}^{(l)}(\tau^p y)-\mathbf{v}^{(l)}(\tau^p y+2^p t)].$$

It follows from (4.8) that

$$\|\mathbf{v}^{(l)}(y)-\mathbf{v}^{(l)}(y+t)\| \leqq 2^{pl}C2^{-pl}p\mu_1^{pr_p(y)}\mu_0^{p(1-r_p(y))}. \tag{4.11}$$

(6) From the proof of Lemma 4.2 we have

$$|r_p(y)-r(y)| \leqq \delta,$$

with $[r(y)-\delta]^{-1}2\delta \leqq \gamma$, hence $\delta \leqq r(y)\gamma/2 \leqq \gamma/2$. Substituting this in (4.11) we find

$$\begin{aligned}\|\mathbf{v}^{(l)}(y)-\mathbf{v}^{(l)}(y+t)\| &\leqq Cp\mu_1^{p[r(y)-\gamma/2]}\mu_0^{p[1-r(y)-\gamma/2]}\\ &\leqq C'p2^{-[(p+1)(1+\gamma)]\{[r(y)-\gamma/2]|\ln\mu_1|+[1-r(y)-\gamma/2]|\ln\mu_0|\}/\ln 2}\\ &\leqq C''(1+|\ln|t||)|t|^{\alpha(x)-\varepsilon/2}\\ &\leqq C'''|t|^{\alpha(x)-\varepsilon},\end{aligned}$$

where we have used (4.9). This proves the first part of the theorem.

(7) Suppose now that $\mu_0 \geqq \frac{1}{2} > \mu_1$, and that $r(x) > (\ln 2 - |\ln\mu_0|/|\ln\mu_1| - |\ln\mu_0|)$, i.e., $\mu_1^{r(x)}\mu_0^{1-r(x)} < \frac{1}{2}$. Choose γ so that

$$\mu_1^{r(x)-\gamma/2}\mu_0^{1-r(x)-\gamma/2} \leqq \tfrac{1}{2}(\tfrac{1}{2}+\mu_1^{r(x)}\mu_0^{1-r(x)}) = \lambda(x)/2 < \tfrac{1}{2}$$

and

$$2^{-2\gamma} \geqq \lambda(x).$$

Choose N_γ corresponding to γ, as in Lemma 4.2. Then, as before, for $p \geqq N_\gamma$, $|r_p(y)-r(y)| \leqq \gamma/2$; hence

$$\mu_1^{r_p(x)}\mu_0^{1-r_p(x)} \leqq \lambda(x)/2 < \tfrac{1}{2}.$$

It follows then from (4.2) that there exists $C>0$ so that, for $p \geqq N_\gamma$,

$$\|\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}|_{E_{L+1}}\| \leqq 2^{-p(l+1)}\lambda(x)^p.$$

The same arguments used in the proof of Lemma 3.5 lead to

$$\|\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}|_{E_{l+2}}\| \leqq C2^{-p(l+1)}\lambda(x)^p. \tag{4.12}$$

(We assume $\lambda(x) > \frac{1}{2}$; if necessary we replace $\lambda(x)$ by $\frac{1}{2}+\varepsilon$.) Since $\lambda(x)<1$, this means that the eigenvalue $2^{-(l+1)p}$ of $\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}$ is simple. It follows that $\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}$ has a left eigenvector $\mathbf{e}_{l+2}(p;y)$ and a right eigenvector $\tilde{\mathbf{e}}_{l+2}(p;y)$ for this eigenvalue, both uniquely determined, except for their normalization. By the structure of $\mathbf{T}_0$ and $\mathbf{T}_1$, $\mathbf{e}_{l+2}(p;y)$ is necessarily a linear combination of the $\mathbf{u}_{k+1}$, $0 \leqq k \leqq l+1$, or equivalently, of the $\mathbf{e}_{k+1}^0$, $0 \leqq k \leqq l+1$; we fix its normalization by requiring that the coefficient of $\mathbf{e}_{l+2}^0$ (which is necessarily nonzero) is equal to 1. We also fix the normalization of $\tilde{\mathbf{e}}_{l+2}(p;y)$ by the requirement

$$\mathbf{e}_{l+2}(p;y)\cdot\tilde{\mathbf{e}}_{l+2}(p;y)=1,$$

or equivalently, since $\tilde{\mathbf{e}}_{l+2}(p;y) \in E_{l+1}$, $\mathbf{e}_{l+2}^0\cdot\tilde{\mathbf{e}}_{l+2}(p;y)=1$.

(8) The vectors $\tilde{\mathbf{e}}_{l+2}(p;y)$ are uniformly bounded in p and converge to a limit as $p \to \infty$, as shown by the following argument. We have

$$\begin{aligned}&\tilde{\mathbf{e}}_{l+2}(p;y)-\tilde{\mathbf{e}}_{l+2}(p+1;y)\\ &\quad=2^{(l+1)p}\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}[\tilde{\mathbf{e}}_{l+2}(p;y)-2^{l+1}\mathbf{T}_{d_{p+1}(y)}\tilde{\mathbf{e}}_{l+2}(p+1;y)].\end{aligned}$$

Since $\mathbf{e}^0_{l+2}\cdot \mathbf{T}_d = 2^{-(l+1)}\mathbf{e}^0_{l+2}+$linear combination of $\mathbf{e}^0_{k+1}, 0 \leqq k \leqq l$, we find $\mathbf{e}^0_{l+2}\cdot [\tilde{\mathbf{e}}_{l+2}(p+y) - 2^{l+1}\mathbf{T}_{d_{p+1}(y)}\tilde{\mathbf{e}}_{l+2}(p+1;y)]=0$, i.e., $\tilde{\mathbf{e}}_{l+2}(p;y) - 2^{l+1}\mathbf{T}_{d_{p+1}(y)}\tilde{\mathbf{e}}_{l+2}(p+1;y) \in E_{l+2}$. Hence, by (4.12),

$$(4.13)\qquad \|\tilde{\mathbf{e}}_{l+2}(p;y)-\tilde{\mathbf{e}}_{l+2}(p+1;y)\| \leqq C\lambda(x)^p[\|\tilde{\mathbf{e}}_{l+2}(p;y)\|+\|\tilde{\mathbf{e}}_{l+2}(p+1;y)\|],$$

which implies, for large enough p,

$$\begin{aligned}\|\tilde{\mathbf{e}}_{l+2}(p+1;y)\| &\leq [1+C\lambda(x)^p][1-C\lambda(x)^p]^{-1}\|\tilde{\mathbf{e}}_{l+2}(p;y)\|\\ &\leqq \exp[3C\lambda(x)^p]\|\tilde{\mathbf{e}}_{l+2}(p;y)\|.\end{aligned}$$

Hence, for all $p > p_0$, where p_0 is chosen large enough,

$$\|\tilde{\mathbf{e}}_{l+2}(p;y)\| \leqq \exp[3C\lambda(x)^{p_0}(1-\lambda(x))^{-1}]\|\tilde{\mathbf{e}}_{l+2}(p_0;y)\| \leqq C',$$

and, by (4.13),

$$\|\tilde{\mathbf{e}}_{l+2}(p+r;y)-\tilde{\mathbf{e}}_{l+2}(p;y)\| \leqq 2CC'\lambda(x)^p(1-\lambda(x))^{-1},$$

so that the $\tilde{\mathbf{e}}_{l+2}(p;y)$ form a Cauchy sequence in p, with limit $\tilde{\mathbf{e}}_{l+2}(y)$.

(9) Any $\mathbf{u}$ in E_{l+1} can be written as

$$\mathbf{u} = \tilde{\mathbf{e}}_{l+2}(p;y)(\mathbf{e}^0_{l+2}\cdot\mathbf{u})+\mathbf{u}',$$

where $\mathbf{u}' \in E_{l+2}$, since $\mathbf{e}^0_{l+2}\cdot\mathbf{u}'=0$ and $\|\mathbf{u}'\| \leqq C\|\mathbf{u}\|$, with C independent of p, because the $\tilde{\mathbf{e}}_{l+2}(p;y)$ are bounded uniformly in p.

Choose now $p_1 = \max(p_0, N_\gamma)$, with p_0, N_γ as in (7) and (6). For $|t| < 2^{-p_1(1+\gamma)-1}$ there exists $p \geqq p_1$ so that

$$2^{-(p+1)(1+\gamma)-1} \leqq |t| < 2^{-p(1+\gamma)-1},$$

which implies that x and $x+t$, or y and $y+t$, have the same p first digits in their binary expansion. Then

$$\mathbf{v}^{(l)}(y+t)-\mathbf{v}^{(l)}(y) = 2^{lp}\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}\mathbf{u},$$

with $\mathbf{u} = \mathbf{v}^{(l)}(\tau^p y + 2^p t)-\mathbf{v}^{(l)}(\tau^p y) \in E_{l+1}$. By Lemma 3.6, $\mathbf{e}^0_{l+2}\cdot\mathbf{u} = (-1)^{l+1}(l+1)!2^p t$, so that

$$\begin{aligned}\mathbf{v}^{(l)}(y)-\mathbf{v}^{(l)}(y+t) &= 2^{lp}\mathbf{T}_{d_1(y)}\cdots\mathbf{T}_{d_p(y)}[(-1)^{l+1}(l+1)!2^p t\tilde{\mathbf{e}}_{l+2}(p;y)+\mathbf{u}']\\ &= (-1)^{l+1}(l+1)!\,t\tilde{\mathbf{e}}_{l+2}(p;y)+\mathbf{u}'',\end{aligned}$$

with $\|\mathbf{u}''\| \leqq 2^{lp}C2^{-p(l+1)}\lambda(x)^p$ by (4.12), by the p-incident bound on $\mathbf{u}$, and the boundedness of $\mathbf{v}^{(l)}$. Because $\lambda(x) \leqq 2^{-\gamma}$, it follows that,

$$(4.14)\qquad \begin{aligned}&\left\|\frac{\mathbf{v}^{(l)}(y+t)-\mathbf{v}^{(l)}(y)}{t}-(-1)^{l+1}!\tilde{\mathbf{e}}_{l+2}(y)\right\|\\ &\quad\leqq (l+1)!\,\|\tilde{\mathbf{e}}_{l+2}(p;y)-\tilde{\mathbf{e}}_{l+2}(y)\|+C2^{\gamma+2}2^{-\gamma p}.\end{aligned}$$

As $|t|\to 0, p\to\infty$, and the right-hand side of (4.14) tends to zero, so that $\mathbf{v}$ is $(l+1)$ times differentiable in y (with $(l+1)$-th derivative $(-1)^{l+1}(l+!)!\tilde{\mathbf{e}}_{l+2}(y))$. Hence, f is $(l+1)$ times differentiable in $x = \lfloor x\rfloor + y$, and the theorem is proved. □

Under the conditions of Theorem 4.1, we find that different Hölder exponents α_r correspond to the sets V_r,

$$(4.15)\qquad V_r = \{x \in [0,N];\ r(x) = r(x-\lfloor x\rfloor)=r\}.$$

These sets are fractal sets. Their Haussdorff dimension can be computed explicitly; it is given by the following theorem, conjectured in I. J. Good (1941) and proved in H. G. Eggleston (1949).

THEOREM 4.3. *For $x \in [0, 1]$, $k \in \mathbb{N}$, $k \geqq 2$, and $0 \leqq l \leqq k-1$, define $r_n(x; l, k)$ to be the number of times, divided by n, that the digit* 1 *occurs among the first n digits of the expansion of x in basis k. For any k-tuple $(r_0, r_1, \cdots, r_{k-1}) = r$ with $0 \leqq r_l \leqq 1$ for all $l \leqq k-1$ and $\sum_{l=0}^{k-1} r_l = 1$, define*

$$V_r(k) = \{x[0, 1];\ \lim_{n\to\infty} r_n(x; l, k) = r_l \text{ for } l = 0, \cdots, k-1\}.$$

Then V_r has fractal dimension δ defined by

$$k^{-\delta} = \prod_{l=0}^{k-1} r_l^{r_l}.$$

Remark. Note that the choice $r_l = k^{-1}$, $l = 0, \cdots, k-1$ leads to $\delta = 1$. This was to be expected, since in this case V_r contains all the normal numbers and, therefore, has Haussdorff dimension 1.

Specializing to $k = 2$, we find therefore that V_r, as defined by (4.15), has fractal dimension

$$\delta = \frac{r|\ln r| + (l - r)|\ln(1-r)|}{\ln 2}.$$

COROLLARY 4.4. *We assume the same conditions as in Theorem* 4.1, *with $\frac{1}{2} \leqq \mu_0, \mu_1$. Then, for any α, $\min(|\ln \mu_0|, |\ln \mu_1|) < \alpha < \max(|\ln \mu_0|, |\ln \mu_1|)$, the set S_α of $x \in [0, N]$ on which f is Hölder continuous with exponent at least α is a fractal set with Hausdorff dimension d_α, with*

$$d_\alpha = 1 \quad \text{if } \alpha \leqq \frac{|\ln \mu_0| + |\ln \mu_1|}{2 \ln 2},$$

$$d_\alpha \geqq \frac{r|\ln r| + (1-r)|\ln(1-r)|}{\ln 2} \quad \text{if } \alpha \geqq \frac{|\ln \mu_0| + |\ln \mu_1|}{2 \ln 2},$$

with

$$r = \frac{\alpha \ln 2 - |\ln \mu_0|}{|\ln \mu_1| - |\ln \mu_0|}.$$

Proof. The proof follows immediately from

$$\dim(V_r) = \frac{r|\ln r| + (1-r)|\ln(1-r)}{\ln 2}. \qquad \square$$

As announced in the introduction, we thus find a hierarchy of fractal sets S_α, with decreasing fractal dimension d_α for increasing Hölder exponents α. Similar fractal sets can be introduced if $\mu_0 < \frac{1}{2} \leqq \mu_1$ or $\mu_1 < \frac{1}{2} \leqq \mu_0$.

Remarks. (1) The different Hölder exponents, and the associated hierarchy of fractal sets do not occur as obviously if the coefficients c_n, $n = 0, \cdots, N$ are symmetric, i.e., if $c_n = c_{N-n}$, $n = 0, \cdots, N$. In this case we find indeed that

$$\mathbf{O}\mathbf{T}_0\mathbf{O}^{-1} = \mathbf{T}_1,$$

where $\mathbf{O}$ is the orthogonal matrix defined by

$$\mathbf{O}_{ij} = \delta_{N+1-i-j}.$$

We easily check that the subspaces U_l are invariant under $\mathbf{O}$,

$$\mathbf{O}\mathbf{u}_l = \sum_{n=1}^{l} \binom{l-1}{n-1} (N+1)^{l-n}(-1)^{n+1}\mathbf{u}_n.$$

It follows that the E_k are invariant under $\mathbf{O}^t = \mathbf{O}^{-1} = \mathbf{O}$. ($\mathbf{O}$ is an involution.) Consequently,

$$\mathbf{T}_1|_{V_k} = \mathbf{O}|_{V_k}\mathbf{T}_0|_{V_k}\mathbf{O}^{-1}|_{V_k}$$

for all $l = 1, \cdots, L+1$, and

$$\|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_m}|_{V_l}\| = \|\mathbf{T}_{1-d_1} \cdots \mathbf{T}_{1-d_m}|_{V_l}\|.$$

If an inequality of type (4.2) were to hold for the $\mathbf{T}_d$, this symmetry would immediately imply that the same inequality would hold if μ_0, μ_1 were both replaced by min (μ_0, μ_1). The different Hölder exponents $\alpha(x)$ then all collapse to the uniformly valid exponent $|\ln(\min(\mu_0, \mu_1))|/\ln 2 - \varepsilon$. Some fractal structure can still persist, however. An example of this is given in § 5.3.

(2) The "tricks" in Proposition 3.7 also apply here: in order to verify bounds of the type (4.2) it suffices, e.g., to check a similar bound for the matrices $\bar{\mathbf{T}}_d = \mathbf{B}\mathbf{T}_d\mathbf{B}^{-1}$, restricted to $\mathbf{B}E_{L+1}$, where $\mathbf{B}$ is any invertible matrix. A convenient choice of $\mathbf{B}$ may greatly simplify computations.

(3) In establishing local Hölder exponents and local differentiability we have restricted ourselves to x such that $r(x) = \lim_{n\to\infty} r_n(x)$ is well defined and $0 < r(x) < 1$. In fact, it is possible to handle more general x as well. In fact, Lemma 4.2 tells us that for those x for which $r(x) = \lim_{m\to\infty} r_m(x)$ exists and $0 < r(x) < 1$, there can be no "abnormally long" stretches of $0-s$ or $1-s$. If x is not of this type (it is easy to construct such x; they constitute a set of zero Lebesgue measure, however), then we need to control these stretches in some other way. We show here how this can be done when $1 > \mu_0 > \mu_1 > \frac{1}{2}$. For all $x \in [0, 1]$, $n \in \mathbb{N}$, we define

$$\omega_n^1(x) = \min\{k \in \mathbb{N};\ d_{n-k}(x) = 0\}/n.$$

Then $n\omega_n^1(x)$ is exactly the length of the stretch of digits 1 ending at n; in particular, if $d_n(x) = 0$, then $\omega_n^1(x) = 0$. A detailed analysis shows that, for $0 \leqq t < 2^{-n}$,

$$\|\mathbf{v}^{(l)}(x) - \mathbf{v}^{(l)}(x+t)\| \leqq C[\mu_1^{nr_n(x)}\mu_0^{n(1-r_n(x))} + \mu_1^{n(n_x(x)-\omega_n^1(x))+1}\mu_0^{n(1-r_n(x)+\omega_n^1(x))}].$$

For $-2^{-n} < t \leqq 0$, we find

$$\|\mathbf{v}^{(l)}(x) - \mathbf{v}^{(l)}(x+t)\| \leqq C\mu_1^{nr_n(x)}\mu_0^{n(1-r_n(x))},$$

where ω_n^1 doesn't enter because $\mu_0 > \mu_1$. Now define

$$\bar{r}(x) = \limsup_{n\to\infty} r_n(x),$$

$$\underline{r}(x) = \liminf_{n\to\infty} r_n(x),$$

$$\bar{r}^1(x) = \limsup_{n\to\infty} [r_n(x) - \omega_n^1(x)],$$

$$\underline{r}_1(x) = \liminf_{n\to\infty} [r_n(x) - \omega_n^1(x)].$$

Then it follows that, for small enough t

$$\begin{aligned} \|\mathbf{v}^{(l)}(x) - \mathbf{v}^{(l)}(x+t)\| &\leqq C|t|^{\{\underline{r}_1(x)|\ln\mu_1| + (1-\bar{r}^1(x))|\ln\mu_0|\}/\ln 2} && \text{if } t \geqq 0, \\ &\leqq C|t|^{\{\underline{r}(x)|\ln\mu_1| + \bar{r}(x)|\ln\mu_0|\}/\ln 2} && \text{if } t \leqq 0. \end{aligned}$$

This effectively defines two local Hölder exponents $\alpha_\pm(x)$, one from above and one from below. Similar bounds can be derived if $\mu_0 < \mu_1$; in this case we need to introduce $\omega_n^0(x)$, measuring the stretch of zeros preceding the digit $d_n(x)$.

The situation $\mu_0 \geqq \frac{1}{2} > \mu_1$ or $\mu_1 \geqq \frac{1}{2} > \mu_0$ is a bit more tricky, but can be dealt with in the same way. Note that if $r(x)$ is well defined, then $\bar{r}(x) = \underline{r}(x) = \bar{r}^1(x) = \underline{r}^1(x) = r(x)$.

(4) The two Hölder exponents $\alpha_\pm(x)$ introduced in the preceding remark, one from above and one from below, are particularly interesting in dyadic rational points x. For such points, we have two possible binary expansions, one ending in all zeros, the expansion "from above," which we denote by $d^+(x)$, and one ending in all ones, the expansion "from below," denoted by $d^-(x)$. For the Hölder exponent from above, we have to start from $d^+(x)$. We then have

$$\bar{r}^1(x) = \underline{r}_1(x) = 0$$

leading to $\alpha_+(x) = |\ln \mu_0|/\ln 2$. Similarly, for the Hölder exponent from below, starting from $d^-(x)$, we find

$$\bar{r}(x) = \underline{r}(x) = 1;$$

hence, $\alpha_-(x) = |\ln \mu_1|/\ln 2$.

If $\mu_0 > \frac{1}{2} > \mu_1$, we similarly find that f is left but not right differentiable in every dyadic rational point.

5. Examples. We treat three classes of examples, namely orthonormal wavelets, the de Rham function and generalizations, and the interpolation functions arising in Deslauriers and Dubuc's Lagrangian interpolation scheme or studied in detail by Dyn, Gregory, and Levin.

In many of these examples we shall use the tricks proposed in Proposition 3.7. In particular, we shall try to determine matrices $\mathbf{B}$ so that the $\mathbf{B}\mathbf{T}_d\mathbf{B}^{-1}$ are easier to study than the $\mathbf{T}_d$ themselves. If the c_n satisfy $L+1$ sum rules, then a very particular choice for $\mathbf{B}$ is the following:

$$\mathbf{B}_{ij} = \begin{cases} (i-1)!\dbinom{j-1}{i-1} & \text{for } i \leqq L+1, \\ L!\dbinom{j-i+L}{L} & \text{for } i > L+1, \end{cases}$$

where we use the standard convention that the binomial coefficient $\binom{n_1}{n_2}$ vanishes if $n_2 > n_1$. This is a triangular matrix; for $i \leqq L+1$, the ith row can be written as $\mathbf{u}_i$ – linear combination of $\mathbf{u}_k$, $k < i$. It is easy to check that the inverse matrix $\mathbf{B}^{-1}$ is again a triangular matrix, given by

$$(\mathbf{B}^{-1})_{ij} = \begin{cases} (-1)^{i+j}\dbinom{j-1}{i-1}[(j-1)!]^{-1} & \text{for } j \leqq L+1, \\ (-1)^{i+j}\dbinom{L+1}{i-j+L+1}(L!)^{-1} & \text{for } j > L+1. \end{cases}$$

Note that for $i \leqq L+1$, the ith row of $\mathbf{B}$ is a linear combination of $\mathbf{u}_1, \cdots, \mathbf{u}_i$. Because of the special choice of the first $L+1$ rows of $\mathbf{B}$, and because of the spectral properties of the $\mathbf{T}_d$, with their nested left invariant subspaces spanned by the $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$, $k \leqq L+1$, we find, therefore, that

$$\begin{aligned} (\mathbf{B}\mathbf{T}_d\mathbf{B}^{-1})_{ij} &= 0 \qquad \text{if } i \leqq L+1, j > i, \\ &= 2^{-i+1} \quad \text{if } i = j \leqq L+1. \end{aligned}$$

This means we are in the situation described in the proofs of Lemmas 3.3, 3.5; in order to prove bounds on products of the restrictions $\mathbf{T}_d|_{E_{L+1}}$, it suffices to consider the smaller matrices obtained by deleting the first $L+1$ rows and columns of the $\mathbf{BT}_d\mathbf{B}^{-1}$. These submatrices are completely determined by $\mathbf{T}_d$ and by the last $(N-L-1)$ rows of $\mathbf{B}$ and the last $(N-L-1)$ columns of $\mathbf{B}^{-1}$. The matrix elements in these rows and columns depend only on the difference $j-i$ between column and row index. Since the $(\mathbf{T}_d)_{i,j}=c_{2i-j-1+d}$ depend only on $2i-j$, this property will, therefore, be shared by the submatrices representing $\mathbf{B}|_{E_{L+1}}\mathbf{T}_d|_{E_{L+1}}\mathbf{B}^{-1}|_{E_{L+1}}$. The entries of these submatrices will, in fact, be given exactly by the coefficients of the quotient of $\sum c_n e^{in\xi}$ by $(1+e^{i\xi})^{L+1}$. That is to say, if

$$\frac{1}{2}\sum_n c_n e^{in\xi}=\left(\frac{1+e^{i\xi}}{2}\right)^{L+1}\frac{1}{2}\sum_n q_n e^{in\xi},$$

then the last $(N-L-1)$ rows and columns of the $\mathbf{BT}_d\mathbf{B}^{-1}$ will be given by

$$(\mathbf{BT}_d\mathbf{B}^{-1})_{ij}=2^l q_{2i-j-(N-L)+d}\,.$$

This observation reduces the study to much smaller matrices, obtained by "peeling off" the sum rules; this is the analog, in our matrix notation, of the reduction from an interpolation subdivision scheme S to the "smaller" subdivision schemes $S^{(n)}$ in Dyn and Levin (1989). A similar observation can also be found in Deslauriers and Dubuc (1989).

5.A. Orthonormal wavelets with compact support. An orthonormal basis of wavelets is a family of functions generated from one single function by dilations and translations,

$$\psi_{jk}(x)=2^{-j/2}\psi(2^{-j}x-k),\qquad j,k\in\mathbb{Z}, \tag{5.1}$$

such that the resulting ψ_{jk} constitute an orthonormal basis of $L^2(\mathbb{R})$. A typical construction of such a basis involves the construction of an auxiliary function ϕ such that

$$\phi(x)=\sum c_n\phi(2x-n) \tag{5.2}$$

for some family of coefficients c_n. In order to lead to an orthonormal wavelet basis, the c_n should satisfy the condition

$$|p(\xi)|^2+|p(\xi+\pi)|^2=1, \tag{5.3}$$

where $p(\xi)=\frac{1}{2}\sum_n c_n e^{in\xi}$. If this condition is satisfied, then there exists an L^2-solution to (5.2), and the associated ψ is given by

$$\psi(x)=\sum_n(-1)^n c_{n+1}\phi(2x+n). \tag{5.4}$$

For a thorough discussion of this construction, with proofs, see Mallat (1989) and Meyer (1990). Equation (5.2) is a two-scale difference equation. If there is only a finite number of nonvanishing c_n, then (modulo some convergence conditions) ϕ has finite support, and so has ψ, by (5.4). In Daubechies (1988) the structure of finite families $\{c_n;\ n=0,\cdots,N\}$ satisfying (5.3) was analyzed. It turns out that these c_n are *always* nonsymmetric if we also require ϕ to be continuous. If this continuity requirement is dropped, then there exists one and only one symmetric solution: $c_0=c_1=1$, all other $c_n\equiv 0$, which correspond to $\phi(x)=1$ if $0\leqq x<1$, 0 otherwise, or $\psi(x)=1$ if $0\leqq x<\frac{1}{2}$, -1 if $\frac{1}{2}\leqq x<1$, zero otherwise; the ψ_{jk} then constitute the Haar basis. Apart from this example, the interesting finite families of c_n are asymmetric, and the corresponding ϕ, therefore, have local regularity properties associated with fractal sets, as described in § 4. The simplest examples are

$$(5.5)\qquad \begin{aligned} N=3, \quad c_0&=\frac{1+\sqrt{3}}{4}, \quad c_2=\frac{3-\sqrt{3}}{4},\\ c_1&=\frac{3+\sqrt{3}}{4}, \quad c_3=\frac{1-\sqrt{3}}{4},\end{aligned}$$

$$(5.6)\qquad \begin{aligned} N=5, \quad c_0&=[1+\sqrt{10}+\sqrt{5+2\sqrt{10}}]/16,\\ c_1&=[5+\sqrt{10}+3\sqrt{5+2\sqrt{10}}]/16,\\ c_2&=[10-2\sqrt{10}+2\sqrt{5+2\sqrt{10}}]/16,\\ c_3&=[10-2\sqrt{10}-2\sqrt{5+2\sqrt{10}}]/16,\\ c_4&=[5+\sqrt{10}-3\sqrt{5+2\sqrt{10}}]/16,\\ c_5&=[1+\sqrt{10}-\sqrt{5+2\sqrt{10}}]/16.\end{aligned}$$

Higher-order examples (always corresponding to odd N) cannot be written in closed form; a table with numerical values of the c_n, for N up to 19, is given in Daubechies (1988), as well as the recipe for their computation. For any N, we denote the associated solution of (1.2) by ϕ_N.

For $N=7$, we have for instance

$$(5.7)\qquad \begin{aligned} c_0&=.325803428051, & c_4&=-.264507167369,\\ c_1&=1.01094571509, & c_5&=.043616300474,\\ c_2&=.892200138247, & c_6&=.046503601071,\\ c_3&=-.039575026235, & c_7&=-.014986989330.\end{aligned}$$

(Note that these c_n are larger, by a factor $\sqrt{2}$, than the coefficients $h(n)$ in Daubechies ((1988), Table 1); the $h(n)$ are normalized by $\sum_n h(n)=\sqrt{2}$).

Let us discuss in detail what the analysis of §§ 2-4 leads to when applied to the concrete examples (5.5)-(5.7).

A. $N=3$. The c_n defined in (5.5) satisfy two "sum rules" of type (3.1), corresponding to $l=0$, 1 or $L=1$. It follows that $N-(L+1)=3-2=1$, so that $E_{L+1}=E_2$ is one-dimensional. The matrices $\mathbf{T}_0, \mathbf{T}_1$ are given by

$$\mathbf{T}_0=\begin{pmatrix} \frac{1+\sqrt{3}}{4} & 0 & 0\\ \frac{3-\sqrt{3}}{4} & \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4}\\ 0 & \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4}\end{pmatrix},$$

$$\mathbf{T}_1=\begin{pmatrix} \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4} & 0\\ \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4} & \frac{3+\sqrt{3}}{4}\\ 0 & 0 & \frac{1-\sqrt{3}}{4}\end{pmatrix}.$$

Their eigenvalues are 1, $\frac{1}{2}$, $(1+\sqrt{3})/4$ for $\mathbf{T}_0$, and $1, \frac{1}{2}, (1-\sqrt{3})/4$ for $\mathbf{T}_1$. Since $1>(1+\sqrt{3})/4>\frac{1}{2}>|(1-\sqrt{3})/4|$ we are in the situation described in Theorem 4.1, with

$\mu_0 > \frac{1}{2} > \mu_1$. Because E_{L+1} is one-dimensional, $\mathbf{T}_0|_{E_{L+1}}$ and $\mathbf{T}_1|_{E_{L+1}}$ commute and

$$\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_{L+1}} = \text{multiplication by } 2^{-m}\left(\frac{1+\sqrt{3}}{2}\right)^{m-s_m}\left(\frac{1-\sqrt{3}}{2}\right)^{s_m} \quad \text{on } E_2,$$

for all $m \in \mathbb{N}$. Because we have sharp bounds for the $\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_{L+1}}$, the lower bounds for the Hölder exponents given in Theorem 4.5 are also sharp, except for the extra ε which we always have to introduce in local Hölder exponents.

Since (3.3) is satisfied with $l=0$, $\lambda = (1+\sqrt{3})/4$, it follows from Theorem 3.1 that ϕ_3 is continuous and that ϕ_3 is Hölder continuous with exponent $1 - \ln[(1+\sqrt{3})/2]\ln 2 \cong .5500\cdots$. The following example shows that this estimate is sharp. Applying (1.2) to $x = 2^{-m}$ leads to $\phi_3(2^{-m}) = ((1+\sqrt{3})/4)^m\phi_3(1)$. The values of $\phi_3(1)$, $\phi_3(2)$ are determined by the right eigenvector of $\mathbf{T}_1$ for the eigenvalue 1, which leads to $\phi_3(1) = (1+\sqrt{3})/2$. Hence, for $x_m = 2^{-m} \to 0$,

$$|\phi_3(0) - \phi_3(x_m)| = \left(\frac{1+\sqrt{3}}{2}\right)|x_m|^{-\ln[(1+\sqrt{3})/4]/\ln 2} = \left(\frac{1+\sqrt{3}}{2}\right)|x_m|^{.5500\cdots}.$$

By Theorem 4.1, ϕ_3 is differentiable in all x such that $r(x)$ is well defined and

$$1 > r(x) > \frac{\ln\left(\dfrac{1+2\sqrt{3}}{2}\right)}{\ln\left(\dfrac{1+\sqrt{3}}{2}\right) + \left|\ln\left(\dfrac{\sqrt{3}}{2}\right)\right|} \simeq .2368\cdots.$$

In particular, ϕ_3 is differentiable in all normal numbers, since $r(x) = \frac{1}{2}$ for x normal. Consequently, ϕ_3 is differentiable almost everywhere, as announced earlier. For values of r between .2368 and zero, there exists a corresponding fractal set (see Theorem 4.3) on which ϕ_3 is Hölder continuous with a Hölder exponent, determined by r, between $.5500\cdots$ and 1.

In dyadic rationals x, we find that ϕ_3 is differentiable from below, but is only Hölder continuous from above, with exponent $.5500\cdots$ (see Remark 4 at the end of § 4). This is already illustrated by the behavior of ϕ_3 near $x=0$ (see above); it also accounts for the "jaggedness" of the graph of ϕ_3 at dyadic points (see Fig. 1).

B. $\boldsymbol{N=5}$. We now turn to the properties of ϕ_5. The corresponding c_n, given in (5.6), satisfy three sum rules of type (3.1), corresponding to $L=2$. In this case $N-(L+1) = 5-3 = 2$, i.e., $E_{L+1} = E_3$ is two-dimensional, and obtaining good estimates for $\|\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_{L+1}}\|$ is not as straightforward as in the previous case. Explicit computation shows that

$$(5.8)\qquad \begin{aligned} \text{spectrum}\,\mathbf{T}_0 &= \{1, \tfrac{1}{2}, \tfrac{1}{4}, (\sqrt{5+2\sqrt{10}}+\sqrt{10}+1)/16, (1-\sqrt{10})/8\} \\ &= \{1, \tfrac{1}{2}, \tfrac{1}{4}, .470467\cdots, -.270284\cdots\}, \\ \text{spectrum}\,\mathbf{T}_1 &= \{1, \tfrac{1}{2}, \tfrac{1}{4}, (1-\sqrt{10})/8, (1+\sqrt{10}-\sqrt{5+2\sqrt{10}})/16\} \\ &= \{1, \tfrac{1}{2}, \tfrac{1}{4}, -.270274\cdots, .049817\cdots\}. \end{aligned}$$

The spectral radii of both $\mathbf{T}_0|_{V_3}$ and $\mathbf{T}_1|_{V_3}$ are, therefore, strictly larger than $\frac{1}{4}$, which means that we cannot expect better than a C^1-result of ϕ_5, with some Hölder continuity for ϕ_5'. In order to obtain this much, we need to prove that (3.3) holds for $l=1$, i.e., that $\|\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_3}\| \leqq C2^{-m}\lambda^m$ for some $\lambda \leqq 1$. Straightforward estimates of $\|\mathbf{T}_0|_{E_3}\|$ and $\|\mathbf{T}_1|_{E_3}\|$ are much too large for our goal. To make computations easier, we apply Proposition 3.7 and work with $\mathbf{B}\mathbf{T}_0\mathbf{B}^{-1}$, $\mathbf{B}\mathbf{T}_1\mathbf{B}^{-1}$ instead, for suitably chosen $\mathbf{B}$.

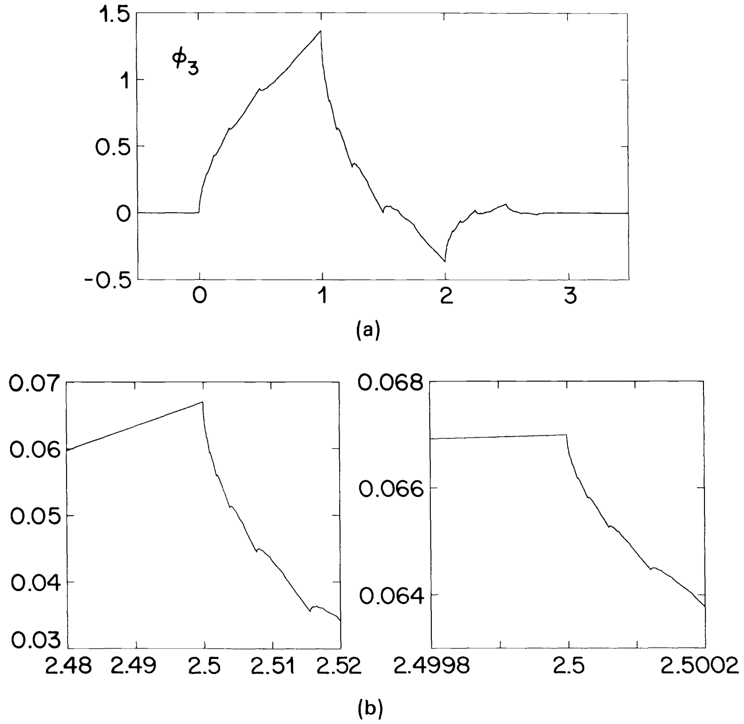


FIG. 1. (a) *The L^1-solution ϕ_3 to the two-scale difference equation $f(x)=[(1+\sqrt{3})f(2x)+(3+\sqrt{3})f(2x-1)+(3-\sqrt{3})f(2x-2)+(1-\sqrt{3})f(2x-3)]/4$. The function ϕ_3 is Hölder continuous with exponent $2-\log_2(1+\sqrt{3})=.5500\cdots$. Moreover, ϕ_3 is almost everywhere differentiable; in dyadic rational points in $[0,3[$ ϕ_3 is left but not right differentiable (see text).* (b) *Two successive blow-ups of ϕ_3 near $x=2.5$, illustrating the self-similar behavior of ϕ_3.*

As a first step, we choose $\mathbf{B}_1$ as outlined at the start of § 5, which in this case means

$$\mathbf{B}_1=\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 2 & 6 & 12 \\ 0 & 0 & 0 & 2 & 6 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

As explained above, the matrices $\mathbf{B}_1\mathbf{T}_d\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3}(d=0$ or $1)$ are simply obtained by stripping $\mathbf{B}_1\mathbf{T}_d\mathbf{B}_d^{-1}$ of their first three columns and rows. This results in

$$\begin{aligned} \mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3}&=\begin{pmatrix} .470467\cdots & 0 \\ .049817\cdots & -.270284\cdots \end{pmatrix}, \\ \mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3}&=\begin{pmatrix} -.270284\cdots & .4770467\cdots \\ 0 & .049817\cdots \end{pmatrix}; \end{aligned} \tag{5.9}$$

as announced earlier, these reduced matrices can also be obtained directly from $(1+e^{i\xi})^{-3}\sum_n c_n e^{in\xi}$. The norms of these matrices are still larger than we would like them. We apply, therefore, the same trick again and multiply (5.9) on the left by $\mathbf{B}_2$, on the right by $\mathbf{B}_2^{-1}$, where $\mathbf{B}_2$ is a conveniently chosen 2×2-matrix. The choice

$$\mathbf{B}_2=\begin{pmatrix} 1 & 0 \\ .33188\cdots & 4.92450\cdots \end{pmatrix}$$

diagonalizes $\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3}$ and reduces $\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3}$ to a symmetric matrix,

$$\mathbf{B}_2(\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3})\mathbf{B}_2^{-1}=\begin{pmatrix} .470467\cdots & 0 \\ 0 & -.270284\cdots \end{pmatrix}$$

and

$$\mathbf{B}_2(\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3})\mathbf{B}_2^{-1}=\begin{pmatrix} -.238644\cdots & -.095535\cdots \\ -.095535\cdots & .018177\cdots \end{pmatrix}.$$

Consequently, the norms of these matrices are given by their spectral radii, which implies

$$\begin{aligned}\|\mathbf{B}_2(\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3})\mathbf{B}_2^{-1}\| &= \rho[\mathbf{B}_2(\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3})\mathbf{B}_2^{-1}] \\ &= \rho(\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3})=\rho(\mathbf{T}_1|_{E_3}),\end{aligned}$$

and similarly

$$\|\mathbf{B}_2(\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_3})\mathbf{B}_2^{-1}\|=\rho(\mathbf{T}_0|_{E_3}).$$

It follows that, for all binary sequences d, and all m,

$$\|\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|_{E_3}\|\leqq C2^{-n}\mu_1^{s_n}\mu_0^{n-s_n},$$

where $s_n=\sum_{j=1}^n d_j$, and $\mu_1=[2\rho(\mathbf{T}_1|_{E_3})]=.540569\cdots$, $\mu_0=[2\rho(\mathbf{T}_0|_{E_3})]=.940934\cdots$. On the other hand, $\mu_d\geqq 2\|\mathbf{T}_d^n|_{E_3}\|^{1/n}\geqq 2\rho(\mathbf{T}_d|_{E_3})(d=0 \text{ or } 1)$. Since $\rho(\mathbf{T}_d|_{E_2})=\rho(\mathbf{T}_d|_{E_3})$ in this case (see (5.8)), the above estimates are, therefore, the sharpest possible for μ_0,μ_1.

Clearly $\frac{1}{2}<\mu_1<\mu_0<1$. It follows that ϕ_5 is continuously differentiable, and that ϕ_5' is Hölder continuous with exponent

$$|\log\mu_0|/\log 2-\varepsilon=.087833\cdots. \tag{5.10}$$

Since the sharpest estimate for μ_1 is strictly larger than $\frac{1}{2}$, ϕ_5 is nowhere twice differentiable; because $\mu_1<\mu_0$, there exists a hierarchy of fractal sets on which ϕ_5' has a larger Hölder exponent than (5.10). In particular, the Hölder exponent of ϕ_5' on the full set of normal numbers is

$$[|\log\mu_0|+|\log\mu_1|]/2\log 2-\varepsilon=.487641\cdots-\varepsilon.$$

The function ϕ_5 is plotted in Fig. 2. At first sight, we have the impression that ϕ_5 has a sharp peak at $x=1$, contradicting its differentiability at this point. Successive blowups of ϕ_5 around $x=1$ show that this peak is not really "sharp" (see Fig. 2(b)).

C. $N=7$. The c_n corresponding to ϕ_7 are given by (5.7); they satisfy (by construction; see Daubechies (1988)) four sum rules of type (3.1), corresponding to $L=3$. It follows that $E_{L+1}=E_4$ is three-dimensional. Explicit computation leads to

$$\text{spectrum } \mathbf{T}_0=\{1,\tfrac{1}{2},\tfrac{1}{4},\tfrac{1}{8},.325803\cdots,-.279620\cdots,.093804\cdots\},$$

$$\text{spectrum } \mathbf{T}_1=\{1,\tfrac{1}{2},\tfrac{1}{4},\tfrac{1}{8},-.279620\cdots,.093804\cdots,-.014986\cdots\}.$$

Since $\rho(\mathbf{T}_0|_{V_4})$, $\rho(\mathbf{T}_1|_{V_4})>\frac{1}{4}$, ϕ_7 is at most once continuously differentiable. In order to prove that ϕ_7 is indeed C^1 and to estimate the Hölder exponent of ϕ_7', we need to

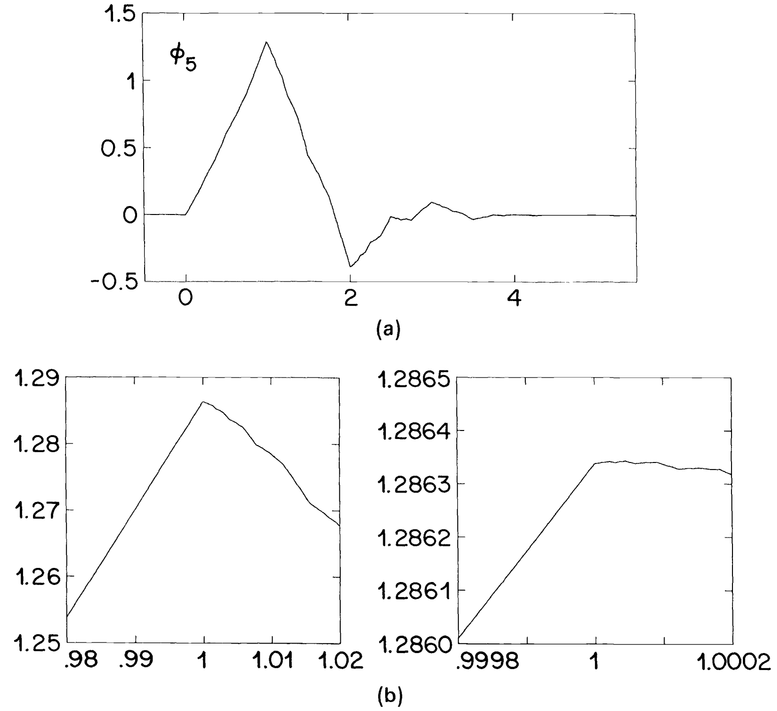


FIG. 2. (a) *The L^1-solution ϕ_5 to the two-scale difference equation $f(x)=\sum_{n=0}^{5} c_n f(2x-n)$, with c_n as given by* (5.6). *The function ϕ_5 is continuously differentiable, despite appearances.* (b) *Two blow-ups of $\phi_5(x)$ around $x=1$, showing that the peak of ϕ_5 at $x=1$ is not really "sharp." We find $\phi_5'(1)=-(\tilde{e}_2^0)_2=1.63845\cdots$.*

find bounds of the type (3.3), with $l=1$, $\lambda<1$. In order to derive such bounds, we shall again study $\|\mathbf{B}\mathbf{T}_\nu\mathbf{B}^{-1}|_{\mathbf{B}E_4}\|$ for suitably chosen, invertible $\mathbf{B}$. We choose $\mathbf{B}_1$ according to the recipe at the start of this section,

$$\mathbf{B}_1=\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 0 & 2 & 6 & 12 & 20 & 30 \\ 0 & 0 & 0 & 6 & 24 & 60 & 120 \\ 0 & 0 & 0 & 0 & 6 & 24 & 60 \\ 0 & 0 & 0 & 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6 \end{pmatrix}.$$

The restriction to $\mathbf{B}_1E_4$ of the matrices $\mathbf{B}_1\mathbf{T}_d\mathbf{B}_1^{-1}$ then simply consists of omitting the first four rows and columns, leading to

$$\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}=\begin{pmatrix} .325803\cdots & 0 & 0 \\ .106451\cdots & -.292267\cdots & .325803 \\ 0 & -014986\cdots & .106451 \end{pmatrix},$$

$$\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}=\begin{pmatrix} -.292267\cdots & .325803\cdots & 0 \\ -.014986\cdots & .106451\cdots & -.292267 \\ 0 & 0 & -.014986 \end{pmatrix}.$$

We recognize again the same characteristic structure as in the matrices $\mathbf{T}_d$ themselves (the ij-element depends only on $2i-j$); the entries are the coefficients of the trigonometric polynomial $2^{-4}(1+e^{i\xi})^{-4}\sum_n c_n e^{in\xi}$. The matrix

$$\mathbf{B}_2=\begin{pmatrix} 1 & 0 & 0 \\ -.144737\cdots & 0.315439\cdots & -8.125871\cdots \\ -.466989\cdots & 2.655918\cdots & -2.241310\cdots \end{pmatrix}$$

diagonalizes $\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}$, and reduces $\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1V_4}$ to

$$\mathbf{B}_2\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1E_4}=\begin{pmatrix} -.238104\cdots & -.034981\cdots & .126825\cdots \\ .034981\cdots & -.002306\cdots & -.004837\cdots \\ .126825\cdots & .080468\cdots & .039608\cdots \end{pmatrix}.$$

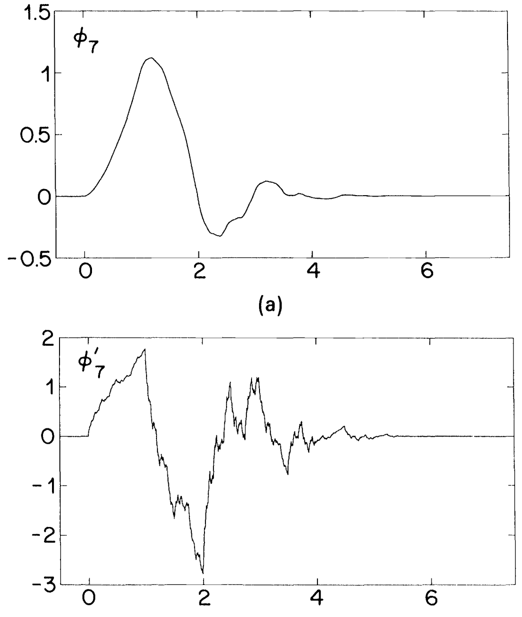
The norm of this last matrix (computed by $\|\mathbf{A}\|=[\rho(\mathbf{A}^t\mathbf{A})]^{1/2}$) is $.296060\cdots$; since $\mathbf{B}_2\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_+E_4}$ is diagonal, its norm is given by the largest eigenvalue of $\mathbf{T}_0|_{E_4}$, which is $.325803\cdots$. We have thus proved that

$$\|\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_m}|E_4\|\leqq C2^{-n}(.651606\cdots)^{n-s_n}(.592120\cdots)^{s_n},$$

which implies that $\phi_7\in C^1$ and that ϕ_7' is Hölder continuous with exponent

$$\frac{|\log(.651606\cdots)|}{\log 2}=.617926\cdots.$$

Since our estimate of μ_0 is sharp, this Hölder exponent is sharp as well. Since $\mu_1<\mu_0$, we again have a hierarchy of fractal sets corresponding to larger Hölder exponents. Note, however, that our estimate for μ_1 is very likely not sharp (the norm of $\mathbf{B}_2\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1E_4}$ is larger than its spectral radius, and it might be possible to sharpen this estimate by other choices of $\mathbf{B}_2$, without losing the sharp estimate on $\mathbf{T}_0$). The range of Hölder exponents might, therefore, be even larger than suggested by our estimates.



(a)

(b)

FIG. 3. (a) *The L^1-solution ϕ_7 to the two-scale difference equation $f(x)=\sum_{n=0}^7 c_nf(2x-n)$, with c_n as given in* (5.7). *The function ϕ_7 is continuously differentiable.* (b) *The derivative ϕ_7' of ϕ_7; ϕ_7' is Hölder continuous with exponent* $.6179\cdots$.

Figures 3(a) and 3(b) give plots of ϕ_7 and ϕ_7', respectively. Both plots have been realized via the cascade algorithm (see part I, § 4).

Remark. Estimates for the Hölder exponents of ϕ_3, ϕ_5, and ϕ_7 were already computed in Daubechies (1988) via a different method, involving the Fourier transform. The table below compares the estimates found here with these earlier results. With the notation $\phi \in C^{n+\alpha}$ for $\phi \in C^n$, $\phi^{(n)}$ Hölder continuous with exponent α ($\alpha \in [0, 1[$), the estimates for $n+\alpha$ compares as follows:

	best estimate in (Daubechies (1988))	as computed here
ϕ_3	$.5-\varepsilon$	$.5500\cdots$
ϕ_5	$.915\cdots-\varepsilon$	$1.0878\cdots$
ϕ_7	$1.275\cdots-\varepsilon$	$1.6179\cdots$

Our present results are significantly better; moreover, they are optimal.

5.B. The de Rham function and generalizations. The de Rham function, as defined in Part I, § 6, is the normalized L^1-solution to the two-scale difference equation

$$\phi(x)=\phi(3x)+\tfrac{1}{3}[\phi(3x-1)+\phi(3x+1)]+\tfrac{2}{3}[\phi(3x-2)+\phi(3x+2)].$$

Since the scaling factor is three instead of two, we are in a slightly different case from before. This example will illustrate, however, that our techniques can be used for all integer scaling factors k, modulo minor adaptations. In general, we have $(k-1)$-matrices $\mathbf{T}_0, \mathbf{T}_1, \cdots, \mathbf{T}_{k-1}$, with matrix elements given by

$$(\mathbf{T}_l)_{ij}=c_{ki-j-N_1-k+1,l}, \qquad 1 \leqq i, j \leqq \frac{N_2+N_1}{k-1},$$

where we assume that the index n of the c_n ranges from $-N_1$ to N_2. In this case, with $k=3$, $c_0=1$, $c_1=c_{-1}=\frac{1}{3}$, $c_2=c_{-2}=\frac{2}{3}$, we have, therefore, three 2×2 matrices,

$$\mathbf{T}_0=\begin{pmatrix}\frac{2}{3} & 0\\ \frac{1}{3} & 1\end{pmatrix} \quad \mathbf{T}_1=\begin{pmatrix}\frac{1}{3} & \frac{2}{3}\\ \frac{2}{3} & \frac{1}{3}\end{pmatrix} \quad \mathbf{T}_2=\begin{pmatrix}1 & \frac{1}{3}\\ 0 & \frac{2}{3}\end{pmatrix}$$

We have $\sum_{n=-2}^{2} c_n=3$, and the c_n satisfy one sum rule:

$$\sum c_{3n}=\sum c_{3n+1}=\sum c_{3n+2}=1. \tag{5.11}$$

The vector (1, 1) is a common left eigenvector for $\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2$, with eigenvalue 1. The remaining eigenvalues of $\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2$ are, respectively, $\frac{2}{3}, -\frac{1}{3}$, and $\frac{2}{3}$. For $x \in [0, 1]$ we shall now use the ternary expansion of x (since $k=3$); as in Theorem 4.3, we define

$$r_n(x, i)=\frac{1}{n}\,\#\{d_j(x)=i; j \leqq n\} \qquad i=0, 1 \text{ or } 2,$$

where $d_j(x)$ are the digits of the ternary expansion of x (each equal to zero, 1 or 2). Then

$$\mathbf{T}_{d_1(x)}\cdots\mathbf{T}_{d_n(x)}|_{E_1}=\text{multiplication by } 3^{-n}2^{n[r_n(x;0)+r_n(x;2)]}(-1)^{nr_n(x;1)}. \tag{5.12}$$

By the extension of Theorem 2.2 to scale factors $k \neq 2$, it follows that ϕ is continuous. Since

$$\|\mathbf{T}(n; d(x))|_{E_1}\| \leqq (\tfrac{2}{3})^n,$$

ϕ is Hölder continuous, with exponent $|\ln(\frac{2}{3})|\ln 3=.36907\cdots$. In fact, this example is so easy (partly due to $c_n \geqq 0$ for all n) that the Hölder continuity can be established without recourse to the matrix analysis presented in this paper; we then find that the Hölder exponent of ϕ is indeed exactly $|\ln(2\frac{2}{3})|/\ln 3$.

From (5.12) it also follows that the local Hölder exponent of ϕ at $y \in [-1, 1]$ is often larger than the uniformly valid exponent $\alpha_{\text{uniform}} = .36907 \cdots$. For $y = n + x$, $n = -1$ or 0, $x \in [0, 1]$, we find that

$$|\phi(y+t) - \phi(y)| \leqq C|t|^{\alpha(y)-\varepsilon},$$

with

$$\alpha(y) = \frac{\ln 3 - (r_0 + r_2) \ln 2}{\ln 3} = \alpha_{\text{uniform}} + r_1(x) \frac{\ln 2}{\ln 3},$$

where we have assumed that $r_1(x) = \lim_{n\to\infty} [r_n(x; 1)]$ exists. Only where $r_1(x) = 0$ is $\alpha(y) = \alpha_{\text{uniform}}$. On the other hand, there exists y such that $r_1(x) = 1$, namely $y \in [0, 2]$ of the type $3^{-j}(l + \frac{1}{2})$; in these points ϕ is Lipschitz. As usual, there exists a hierarchy of fractal sets corresponding to the Hölder exponents between and α_{uniform}; their Haussdorff dimension is given by Theorem 4.3. In particular, for the full set of normal numbers, we find

$$\alpha_{\text{normal}} = \alpha_{\text{uniform}} + \tfrac{1}{3} \frac{\ln 2}{\ln 3} = 1 - \tfrac{2}{3} \frac{\ln 2}{\ln 3} = .57938 \cdots.$$

A straightforward generalization of the de Rham function is obtained by choosing

$$c_0 = 1,\ c_1 = c_{-1} = \tfrac{1}{2} - \gamma,\ c_2 = c_{-2} = \tfrac{1}{2} + \gamma,$$

corresponding to the two-scale difference equation

$$\begin{aligned} \phi^\gamma(x) = \phi^\gamma(3x) &+ (\tfrac{1}{2} - \gamma)[\phi^\gamma(3x-1) + \phi^\gamma(3x+1)] \\ &+ (\tfrac{1}{2} + \gamma)[\phi^\gamma(3x-2) + \phi^\gamma(3x+2)]. \end{aligned}$$

In this case the matrices $\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2$ are

$$\mathbf{T}_0 = \begin{pmatrix} \frac{1}{2}+\gamma & 0 \\ \frac{1}{2}-\gamma & 1 \end{pmatrix}, \mathbf{T}_1 = \begin{pmatrix} \frac{1}{2}-\gamma & \frac{1}{2}+\gamma \\ \frac{1}{2}+\gamma & \frac{1}{2}-\gamma \end{pmatrix}, \mathbf{T}_2 = \begin{pmatrix} 1 & \frac{1}{2}-\gamma \\ 0 & \frac{1}{2}+\gamma \end{pmatrix}.$$

We have again $\sum c_{3n} = \sum c_{3n+1} = \sum c_{3n+2} = 1$, causing $\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2$ to have a common left eigenvector for the eigenvalue 1. The remaining eigenvalues are $\frac{1}{2} + \gamma$, -2γ, $\frac{1}{2} + \gamma$, respectively. It follows that ϕ^γ is continuous only if $|\frac{1}{2} + \gamma| < 1$ and $|2\gamma| < 1$, i.e., if $|\gamma| < \frac{1}{2}$. For $\gamma = \frac{1}{6}$, we obviously revert to the de Rham case; for $\gamma = -\frac{1}{6}$ the function ϕ^γ is piecewise linear, $\phi^{-1/6}(x) = 1 + x$ for $-1 \leqq x \leqq 0$, $1 - x$, for $0 \leqq x \leqq 1$, and zero otherwise; for $\gamma = 0$ the resulting ϕ^0 consists of a copy and its mirror image of the Cantor–Lebesgue function on $[0, 1]$.

The same analysis as before results in Hölder continuity for ϕ^γ, with exponent $\alpha_\gamma = (\min [|\ln (\frac{1}{2} + \gamma)|, |\ln (2\gamma)|])/\ln 3$. For $y = n + x$, $n = -1$ or 0, $x \in [0, 1]$, such that $r_1(x) = \lim_{n\to\infty} [r_n(x; 1)]$ exists, we find a larger local Hölder exponent

$$(5.13) \qquad \begin{aligned} \alpha_\gamma(y) &= \frac{|\ln (\frac{1}{2}+\gamma)|}{\ln 3} + r_1(x) \frac{\ln \dfrac{\frac{1}{2}+\gamma}{|2\gamma|}}{\ln 3} \quad \text{if } \tfrac{1}{2} > \gamma \geqq -\tfrac{1}{6}, \\ & \frac{|\ln |2\gamma||}{\ln 3} + [1 - r_1(x)] \frac{\ln \dfrac{|2\gamma|}{\frac{1}{2}+\gamma}}{\ln 3} \quad \text{if } -\tfrac{1}{6} \geqq \gamma > -\tfrac{1}{2}. \end{aligned}$$

For $\gamma = -\frac{1}{6}$, all these Hölder exponents (local and uniform) collapse to 1, which was to be expected since $\phi^{-1/6}$ is piecewise linear. If, for some γ and some y, $\alpha_\gamma(y)$ as given by (5.13) is larger than 1, then this means that ϕ^γ is differentiable in y. This happens, for instance, if $(\frac{1}{2}+\gamma)^2|2\gamma|<\frac{1}{27}$, and if $r_1(x)=\frac{1}{3}$. In particular, ϕ^γ is differentiable almost everywhere if $-\frac{1}{2}<\gamma<.05921689\cdots$.

Fig. 4 gives the graphs of ϕ^γ for the values $\gamma=\frac{1}{6}$ (de Rham case), $\gamma=\frac{1}{12}$, $\gamma=.05$, $\gamma=0$ and $\gamma=-\frac{1}{12}$.

5.C. Lagrangian interpolation functions and generalizations. In Deslauriers and Dubuc (1989), general symmetric Lagrangian interpolation schemes are defined, for arbitrary integer scaling factor $k>1$, and an arbitrary number of nodes. Deslauriers and Dubuc characterize the regularity of the associated "fundamental functions," which are solutions to two-scale difference equations with scaling factor k,

$$f(x)=f(kx)+\sum_{n=1}^{kN-1} c_n[f(kx-n)+f(kx+n)],$$

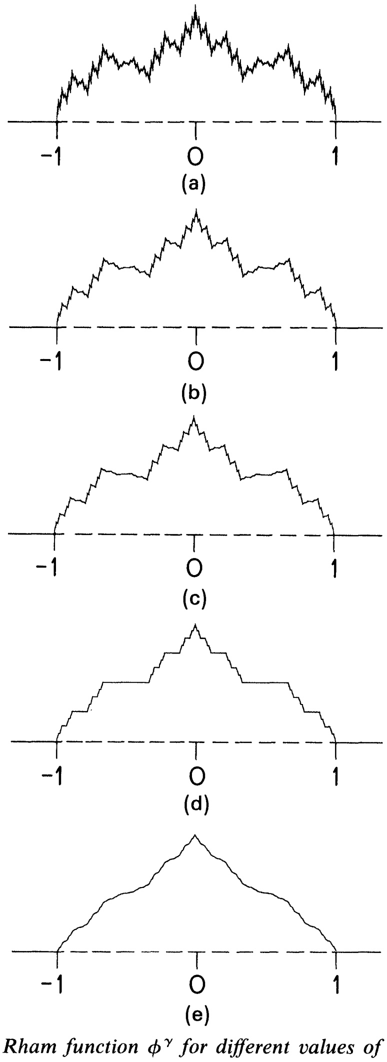


FIG. 4. *The generalized de Rham function* ϕ^γ *for different values of* γ; (a) $\gamma=\frac{1}{6}$ (*the de Rham case*), (b) $\gamma=\frac{1}{12}$, (c) $\gamma=.05$, (d) $\gamma=0$ (*characteristic function of Cantor set*), (e) $\gamma=-\frac{1}{12}$. *In the cases* (c), (d), *and* (e), ϕ^γ *is almost everywhere differentiable.*

where the c_n have the peculiarity that $c_{km} = \delta_{m0}$ (see Part I). The c_n in a Lagrangian symmetric interpolation scheme are completely determined by the requirement that $p(\xi) = k^{-1}[1 + 2\sum_{n=1}^{kN-1} c_n \cos n\xi]$ be divisible by $(1 + e^{i\xi} + \cdots + e^{i(k-1)\xi})^L$, with L as large as possible for the given N (this leads to $L = 2N$). The resulting $p(\xi)$ is positive for real ξ, which is of great help in the analysis of the regularity of f. We have indeed (see part I or Deslauriers and Dubuc (1989)) $\hat{f}(\xi) = \prod_{j=1}^{\infty} p(k^{-j}\xi)$; since $p(\xi) = ((1 - e^{ik\xi})/k(1 - e^{i\xi}))^L q(\xi)$, this leads to $\hat{f}(\xi) = ((1 - e^{i\xi})/\xi)^L \prod_{j=1}^{\infty} q(k^{-j}\xi)$. The quotient $q(\xi)$ is a trigonometric polynomial, $q(\xi) = \sum_n q_n e^{in\xi}$. The regularity of f is given by the largest λ such that $|\xi|^\lambda \hat{f}(\xi)$ is in $L^1(\mathbb{R})$; it turns out that these L^1-norms can be estimated in terms of N and the spectral radius of a finite matrix, constructed from the q_n in the same way as $\mathbf{T}_0, \mathbf{T}_1$ are constructed from the c_n (see § 7, 8 in Deslauriers and Dubuc (1989)). A similar technique was used by one of us, independently, to obtain estimates in the L^1-norm of $|\xi|^\lambda |\hat{\phi}_N(\xi)|$, where ϕ_N are the functions associated to orthonormal wavelet bases (see § 5.A). In that case, however, the function p was not positive, and it was necessary to use the Cauchy–Schwarz inequality to reduce everything to more tractable L^2-estimates, which, however, led to less sharp estimates (see the table at the end of § 5.A).

Let us see now how the present methods, which avoid Fourier transforms, perform on these interpolation functions. We shall restrict ourselves to one of the simpler examples, with $k = 2$ and $N = 2$. In this case the c_n are

$$c_0 = 1, \quad c_1 = c_{-1} = \tfrac{9}{16}, \quad c_3 = c_{-3} = -\tfrac{1}{16}. \tag{5.14}$$

We have two matrices: $\mathbf{T}_0, \mathbf{T}_1$,

$$\mathbf{T}_0 = \begin{pmatrix} -\frac{1}{16} & 0 & 0 & 0 & 0 & 0 \\ \frac{9}{16} & 0 & -\frac{1}{16} & 0 & 0 & 0 \\ \frac{9}{16} & 1 & \frac{9}{16} & 0 & -\frac{1}{16} & 0 \\ -\frac{1}{16} & 0 & \frac{9}{16} & 1 & \frac{9}{16} & 0 \\ 0 & 0 & -\frac{1}{16} & 0 & \frac{9}{16} & 1 \\ 0 & 0 & 0 & 0 & -\frac{1}{16} & 0 \end{pmatrix},$$

$$\mathbf{T}_1 = \begin{pmatrix} 0 & -\frac{1}{16} & 0 & 0 & 0 & 0 \\ 1 & \frac{9}{16} & 0 & -\frac{1}{16} & 0 & 0 \\ 0 & \frac{9}{16} & 1 & \frac{9}{16} & 0 & -\frac{1}{16} \\ 0 & -\frac{1}{16} & 0 & \frac{9}{16} & 1 & \frac{9}{16} \\ 0 & 0 & 0 & -\frac{1}{16} & 0 & \frac{9}{16} \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} \end{pmatrix}.$$

The c_n in (5.14) satisfy four sum rules. Because of the symmetry $c_n = c_{-n}$, we have $\mathbf{T}_1 = \mathbf{O}\mathbf{T}_0\mathbf{O}$, where $\mathbf{O}$ is the involution $\mathbf{O}_{ij} = \delta_{i+j-7}$. Consequently, $\mathbf{T}_1, \mathbf{T}_0$ have the same eigenvalues; we find

$$\text{spectrum}\,(\mathbf{T}_0) = \text{spectrum}\,(\mathbf{T}_1) = \{1, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, -\tfrac{1}{16}\},$$

where the root $\frac{1}{4}$ has multiplicity two in the characteristic polynomial of $\mathbf{T}_0, \mathbf{T}_1$. It follows that $\|\mathbf{T}_0|_{E_3}\| = \|\mathbf{T}_1|_{E_3}\| \geqq \frac{1}{4}$, so that the best we can hope for is a bound of type (3.3) with $l = 1$ and $\lambda = \frac{1}{2}$, corresponding to $|f'(x + t) - f'(x)| \leqq C|t|\,|\log|t||$. It turns out that the restriction of (say) $\mathbf{T}_0$ (the same will be true for $\mathbf{T}_1$) to the two-dimensional space associated with the eigenvalue $\frac{1}{4}$ can be brought to the normal Jordan form, but cannot be diagonalized. There are, therefore, only one left and one right eigenvector of $\mathbf{T}_0$ with eigenvalue $\frac{1}{4}$, and they are orthogonal. This is what causes the extra factor $|\log|t||$ in the "almost Lipschitz" bound on f'. The degeneracy of $\frac{1}{4}$ is lifted, however, as soon as $\mathbf{T}_0, \mathbf{T}_1$ are mixed: if $(d_1, \cdots, d_n)$ does not consist of only $0-s$ or only $1-s$, then the eigenvalue 4^{-n} of $\mathbf{T}_{d_1}\cdots\mathbf{T}_{d_n}$ is nondegenerate. If, however, the sequence d has a tail consisting of only $0-s$ or only $1-s$ (corresponding to a dyadic rational x), then the gap between 4^{-n} and the closest eigenvalue tends to zero as n tends to infinity, and an attempt at the construction of $\tilde{\mathbf{e}}_3(n;\, d(x))$ (see the proof of Theorem 4.1) diverges for $n\to\infty$. If f were twice continuously differentiable, then its second derivative f'' would be given by these limits $\tilde{\mathbf{e}}_3(\infty;\, d(x))$ (see § 4); the divergence of $\tilde{\mathbf{e}}_3(n;\, d(x))$ at dyadic rational x shows, therefore, that f cannot be twice differentiable at dyadic rationals.

To check that (3.3) holds for $l=1, \lambda=\frac{1}{2}$, we shall again compute estimates for $\mathbf{BT}_0\mathbf{B}^{-1}$, $\mathbf{BT}_1\mathbf{B}^{-1}$, with a conveniently chosen $\mathbf{B}$, rather than work with $\mathbf{T}_0|_{E_4}$, $\mathbf{T}_1|_{E_4}$ themselves. The existence of four sum rules suggests that we use for $\mathbf{B}_1$

$$\mathbf{B}_1=\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1\\ 0 & 1 & 2 & 3 & 4 & 5\\ 0 & 0 & 2 & 6 & 12 & 20\\ 0 & 0 & 0 & 6 & 24 & 60\\ 0 & 0 & 0 & 0 & 6 & 24\\ 0 & 0 & 0 & 0 & 0 & 6 \end{pmatrix}. \tag{5.15}$$

The matrices $\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}$, $\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}$ then reduce to 2×2-matrices,

$$\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}=\begin{pmatrix} -\frac{1}{16} & 0\\ -\frac{1}{16} & \frac{1}{4}\end{pmatrix},$$

$$\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}=\begin{pmatrix} \frac{1}{4} & -\frac{1}{16}\\ 0 & -\frac{1}{16}\end{pmatrix}.$$

The matrix

$$\mathbf{B}_2=\begin{pmatrix} 1 & 0\\ \dfrac{\sqrt{6}}{2} & \dfrac{-5\sqrt{6}}{2}\end{pmatrix} \tag{5.16}$$

then diagonalizes $\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}$ and reduces $\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1E_4}$ to a symmetric matrix. It follows that, for $d=0$ or 1,

$$\begin{aligned}\|\mathbf{B}_2\mathbf{B}_1\mathbf{T}_d\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1E_4}\| &= \rho(\mathbf{B}_2\mathbf{B}_1\mathbf{T}_d\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1E_4})\\ &= \rho(\mathbf{T}_d|_{E_4})=\tfrac{1}{4}.\end{aligned}$$

This implies (3.3) with $l=1, \lambda=\frac{1}{2}$. Theorem 3.1 then implies that f is "almost" C^2, in the sense that $f\in C^1$ and that f' is Hölder continuous with exponent $1-\varepsilon$, with $\varepsilon>0$ arbitrarily small; in fact $|f'(x+t)-f'(x)|\leqq Ct|\ln|t||$ for sufficiently small t, a result which was first proved in Dubuc (1986).

Since $\mathbf{T}_1=\mathbf{OT}_0\mathbf{O}^{-1}$, where $\mathbf{O}$ is an orthogonal matrix (see above), we necessarily have $\mu_0=\mu_1$, where μ_0, μ_1 are as defined in § 4. Nevertheless, we can still improve on the estimate $\|\mathbf{T}_{d_1(x)}\cdots\mathbf{T}_{d_n(x)}|_{E_2}\|<Cn2^{-2n}$ for special, nondyadic x. Group the digits in the binary expansion of x together in pairs, $p_n(x)=d_{2n-1}(x)d_{2n}(x)$, $n=1, 2, \cdots$. Assume that the pairs 0 1, 1 0 occur with an asymptotic frequency larger than zero,

$$p(x)=\liminf_{n\to\infty}\frac{1}{n}\,\#\,\{j<n;\, p_j(x)=1\ \ 0 \text{ or } p_j(x)=0\ \ 1\}>0.$$

Then it can be shown that, for large enough n,

$$\|\mathbf{T}(n;\, d(x))|_{E_3}\|\leqq C2^{-2n}\lambda^n, \tag{5.17}$$

where $\lambda < 1$. Moreover, $\tilde{\mathbf{e}}_3(x) = \lim_{n\to\infty} \tilde{\mathbf{e}}_3(n; d(x))$ is well defined in these points x; repeating some of the arguments in § 4 shows, therefore, that f is twice differentiable in x. The estimate (5.17) is a consequence of the fact that $\mathbf{B}_2\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1 E_3}$ and $\mathbf{B}_2\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1 E_3}$ have a common left eigenvector with eigenvalue $\frac{1}{8}$, while also

$$\|\mathbf{B}_2\mathbf{B}_1\mathbf{T}_0\mathbf{T}_1\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}\|_{\mathbf{B}_1 E_4}\|^{1/2} = \|\mathbf{B}_2\mathbf{B}_1\mathbf{T}_1\mathbf{T}_0\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1 E_4}\|^{1/2} = .150156 \cdots.$$

Here $\mathbf{B}_1, \mathbf{B}_2$ are the matrices defined by (5.15), (5.16), respectively. This implies

$$(5.18) \quad \|\mathbf{B}_2\mathbf{B}_1\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_n}\mathbf{B}_2^{-1}|_{\mathbf{B}_1 E_4}\| \leqq C \max\,(8^{-n}, \|\mathbf{B}_2\mathbf{B}_1\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_n}\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1 E_4}\|),$$

where C depends on $\mathbf{B}_1, \mathbf{B}_2$, but not on n. For large n, the right-hand side of (5.18) is bounded by

$$C[(.150156\cdots)^{p(x)}(\tfrac{1}{4})^{1-p(x)}]^n \leqq C4^{-n}(.600624\cdots)^{np(x)}.$$

Together with (5.18), this implies (5.17), with $\lambda = (.600624\cdots)^{p(x)}$.

Note that $p(x) = \frac{1}{2}$ in every normal point x. The above argument proves therefore that f is almost everywhere twice differentiable. Figure 5 gives graphs of both f and f' (both were also plotted in Dubuc (1986)). Note that f looks "smooth," but that f' again exhibits "bumps" which repeat themselves at different scales.

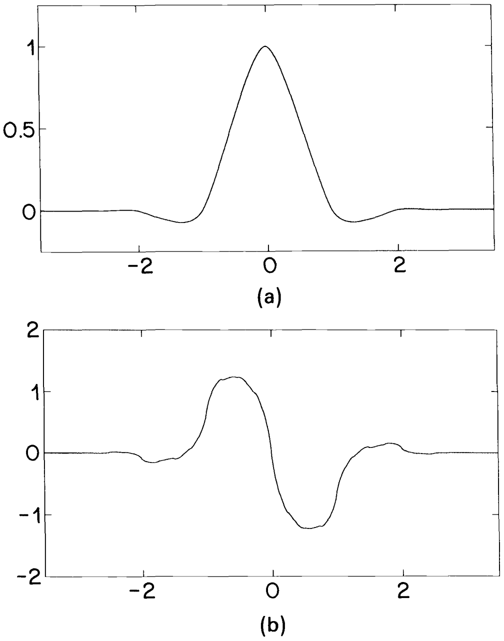


FIG. 5. (a) *The L^1-solution to the two-scale difference equation $f(x) = f(2x) + \frac{9}{16}[f(2x-1) + f(2x+1)] - \frac{1}{16}[f(2x-3) + f(2x+3)]$. Thus function is "almost" C^2 (see text).* (b) *The derivative f' of the function plotted in* (a).

The c_n in (5.14) were chosen so that, with the restrictions $c_n = c_{-n}$, $c_2 = c_{-2} = 0$, the polynomial $p(\xi) = \sum_{n=-3}^{3} c_n e^{in\xi}$ was divisible by the maximum possible number of factors $(1 + e^{i\xi})$. Other examples, still satisfying $c_n = c_{-n}$, $c_2 = c_{-2}$, $c_n = 0$ for $|n| > 3$, were mentioned in Deslauriers and Dubuc (1987). These examples satisfy fewer sum rules, ($p(\xi)$ is divisible by fewer factors $(1 + e^{i\xi})$), and the corresponding $p(\xi)$ is not positive any more. In order to evaluate the regularity of the function f, Deslauriers and Dubuc had to resort to other, less optimal techniques. They prove, e.g., that for

$c_1 = c_{-1} = \frac{1}{2} - a$, $c_3 = c_{-3} = a$, the function f is continuous if $a \in]-\frac{3}{16}, \frac{1}{16}[$. Our present technique leads to better results. For the case $a = -\frac{3}{16}$, e.g., we show that f is continuously differentiable. To do this, we shall use Proposition 3.7 again. Since the c_n satisfy only two sum rules in this case, we choose $\mathbf{B}_1$

$$\mathbf{B}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

$\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1 E_2}$, $\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}|_{\mathbf{B}_1 E_2}$ reduce to 4×4-matrices, with eigenvalues $\frac{3}{8}$, $-\frac{3}{16}$, and $(1 \pm i\sqrt{2})/4$. The matrices

$$\mathbf{B}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2y/9 & y & 0 & -y \\ \dfrac{5-26\sqrt{2}i}{81}z & z & \dfrac{-1+2\sqrt{2}i}{3}z & z \\ \dfrac{5+26\sqrt{2}i}{81}z' & z' & -\dfrac{1+2\sqrt{2}i}{3}z' & \end{pmatrix},$$

where y, z, z' are arbitrary complex parameters, diagonalize $\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}|_{\mathbf{B}_1 E_2}$, so that

$$\|\mathbf{B}_2\mathbf{B}_1\mathbf{T}_0\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_1 E_2}\| = \rho(\mathbf{T}_0|_{E_2}) = |(1 \pm i\sqrt{2})/4| = .433012 \cdots.$$

For the choice $y = .3$, $z = .27 + .08i$, $z' = .27 - .08i$, the norm corresponding to $\mathbf{T}_1$ is

$$\|\mathbf{B}_2\mathbf{B}_1\mathbf{T}_1\mathbf{B}_1^{-1}\mathbf{B}_2^{-1}|_{\mathbf{B}_2 E_2}\| = .604342 \cdots.$$

This is still larger than $\frac{1}{2}$, and, therefore, not sufficient to show that $f \in C^1$. However, if we define

$$t_{d_1 \cdots d_m} = \|\mathbf{T}_{d_1,\mathrm{red}} \cdots \mathbf{T}_{d_m,\mathrm{red}}\|^{1/m},$$

where $\mathbf{T}_{d_{m,\mathrm{red}}} = \mathbf{B}_2\mathbf{B}_1\mathbf{T}_d\mathbf{B}_1^{-1}\mathbf{B}_2^{-2}|_{\mathbf{B}_1 E_2}$, with the parameters in $\mathbf{B}_2$ fixed as above, then we check that

$$\begin{aligned} &t_0,\ t_{10},\ t_{1100},\ t_{110100},\ t_{1101010},\ t_{11010110},\ t_{11010111}, \\ &t_{1101100},\ t_{1101101},\ t_{110111},\ t_{111000},\ t_{111001}, \\ &t_{1110100},\ t_{111010100},\ t_{1110101010},\ t_{11101010110}, \\ &t_{11101010111},\ t_{1110101100},\ t_{1110101101},\ t_{111010111}, \\ &t_{11101100},\ t_{11101101},\ t_{1110111},\ t_{111100}, \\ &t_{1111010},\ t_{11110110},\ t_{11110111},\ t_{111110} \quad \text{and} \quad t_{111111}. \end{aligned} \tag{5.19}$$

are all $<\frac{1}{2}$.

Since these groups of indices constitute a complete set of building blocks (any binary sequence can be decomposed into them), it follows that

$$\|\mathbf{T}_{d_1} \cdots \mathbf{T}_{d_m}|_{E_2}\| \leqq 2^{-m}\lambda^m$$

for some $\lambda < 1$, if m is large enough (see Proposition 3.7). Note that in order to derive (5.19) only 56 matrix norms were calculated (this includes candidates that failed, such as $t_{1101}, \cdots$), even though the longest sequences of digits has 11 elements. Checking all 2^{11} sequences of this length would have been much more cumbersome, and would, in fact, not have been sufficient in order to conclude that (3.3) holds, since some of them still lead to $t_D > \frac{1}{2}$.

From all this it follows that the function f corresponding to the two-scale difference equation with $k=2$, $c_0=1$, $c_0=c_1=\frac{11}{16}$, $c_3=c_{-3}=-\frac{3}{16}$, all other $c_n=0$, is continuously differentiable. It is plotted in Fig. 6, together with its derivative.

For general a, $c_0=1$, $c_1=c_{-1}=\frac{1}{2}-a$, $c_3=c_{-3}=a$, we find that

$$\text{spectrum } (\mathbf{T}_0) = \text{spectrum } (\mathbf{T}_1) = \left\{1, \tfrac{1}{2}, a, -2a, \frac{1 \pm \sqrt{1+16a}}{4}\right\}.$$

We conjecture that

— the associated function f is in C^1 if $-\frac{1}{4} < a < 0$,

— the associated function f is continuous if $-\frac{1}{2} < a < \frac{1}{2}$.

Appendix. In our analysis of lattice two-scale equations of the type

$$f(x) = \sum_{n=0}^{N} c_n f(2x-n), \tag{A.1}$$

the interval $[0, 1]$ and binary expansions play a special role. This is because $[0, 1]$ has the following two properties: $[0, 1]$ and its integer translates tile the real line, and $[0, 1] = \Lambda^{-1}[0, 1] \cup \Lambda^{-1}([0, 1]+1)$, where Λ is multiplication by 2. The interval $[0, 1]$ is, moreover, the unique bounded subset of $\mathbb{R}$ satisfying these two conditions. If the scale factor in (A.1) were k instead of 2, then we would define Λ to be multiplication

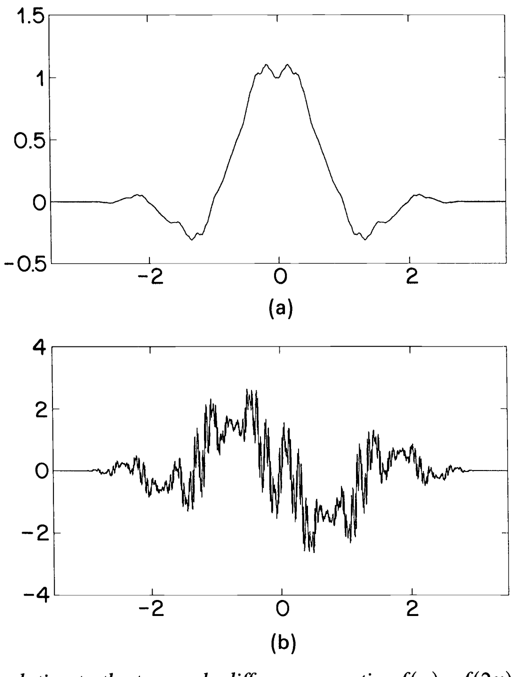


FIG. 6. (a) *The L^1-solution to the two-scale difference equation* $f(x) = f(2x) + \frac{11}{16}[f(2x-1) + f(2x+1)] - \frac{3}{16}[f(2x-3) + f(2x+3)]$. *This function is continuously differentiable* (*see text*). (b) *The derivative of the function plotted in* (a).

by k, and $[0, 1]$ would be tiled by the $\Lambda^{-1}([0,1]+m)$, $0 \leqq m \leqq k-1$, leading to k-adic expansions. All this can be generalized to higher dimensions. We can then write

$$f(x) = \sum_{n \in I} c_n f(\Lambda x - n), \tag{A.2}$$

where $x \in \mathbb{R}^n$, I is a finite subset of $\mathbb{Z}^d$, and Λ is a linear operator on $\mathbb{R}^d$ that preserves $\mathbb{Z}^d$ and has all its singular values strictly larger than 1. For $d=2$, examples of such Λ which have been proposed for applications are

$$\Lambda_0 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \Lambda_1 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad \text{and } \Lambda_2 = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}.$$

Λ_0 is used in many two-dimensional subdivision schemes; since it consists of a simple uniform dilation, the one-dimensional approach used earlier can be transposed without any problems. The role played before by $[0, 1]$ will now be played by $[0, 1]^2$, and binary expansions still do the trick. For the matrices Λ_1 and Λ_2 the situation is more complicated. Λ_1 was first proposed by M. Vetterli (1984) for a subband coding scheme with exact reconstruction for two-dimensional images; presently, several groups are working on the corresponding orthonormal wavelet bases (Gröchenig and Madych (1992), Lawton and Resnikoff (1991)). An interpolation subdivision scheme using Λ_2 is studied in Mongeau (1990); strictly speaking, Mongeau uses

$$\Lambda_2' = \begin{pmatrix} \frac{3}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{3}{2} \end{pmatrix}$$

for which the triangular lattice $\{(m+n/2, n\sqrt{3}/2);\ m, n \in \mathbb{Z}\}$, not $\mathbb{Z}^2$, is invariant. The matrix

$$\mathbf{B} = \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{3}} \end{pmatrix}$$

maps the triangular lattice to $\mathbb{Z}^2$ and can be used to translate Mongeau's results to results for lattice two-scale equations involving Λ_2; we have $\Lambda_2 = \mathbf{B}\Lambda_2'\mathbf{B}^{-1}$.

A large part of our analysis can still be carried out, even for nonuniform dilation matrices such as Λ_1 and Λ_2. Since Λ has only integer entries, and its singular values are all larger than 1, $\det \Lambda = k \in \mathbb{Z}$, with $k > 1$. The role played by $[0, 1]$ for (A.1) will now be played by the unique set Γ defined by

—the collection $\{\Gamma + n\colon n \in \mathbb{Z}^d\}$ tiles $\mathbb{R}^d$,

—Γ itself is tiled by the k elements of $\{\Lambda^{-1}(\Gamma + m);\ m \in \mathbb{Z}^d \cap \Lambda([0, 1[^d))\}$.

For both Λ_1 and Λ_2, the corresponding set Γ is a set with fractal boundary; Γ_1 is the so-called *twin dragon set.* The set Γ will be the "elementary building block" for the support of compactly supported solutions to (A.2), just like $[0, 1]$ was for (A.1). The number N of such building blocks constituting support (f) will be determined by the nonvanishing coefficients c_n in (A.2). We can then "fold" f as follows:

$$[\mathbf{v}(x)]_n = f(x + z_n), \qquad x \in \Gamma,\ z_n \in \mathbb{Z}^d, \quad n = 1, \cdots, N,$$

where $\bigcup_{n=1}^{N}(\Gamma+z_n)\supset$ support (f). Points in $\mathbb{R}^d$ can then be represented by Γ-adic expansions; in particular, for $x\in\Gamma$, we define

$$d_1(x)=\{m\in\mathbb{Z}^d\cap\Lambda([0,1[^d);\ x\in\Lambda^{-1}(\Gamma+m)\},$$
$$d_{j+1}(x)=d_j(\Lambda x-d_1(x)).$$

These expansions are unique, except for x in a set of measure zero (corresponding to the dyadic rationals in the one-dimensional case with $\Lambda=2$). We can again define a shift operator τ on Γ by

$$d_j(\tau x)=d_{j+1}(x)\qquad j=1,2,\cdots,$$

and the equation (A.2) can be rewritten in vector notation as

$$\mathbf{v}(x)=\mathbf{T}_{d_1(x)}\mathbf{v}(\tau x),$$

where the k matrices $\mathbf{T}_l$ are all determined by the coefficients c_n. We can then derive existence, continuity, smoothness, etc., of $\mathbf{v}$ (hence of f) from the spectral properties of the $\mathbf{T}_l$. A first extension to higher dimensions can be found in Mongeau (1990), extending our original approach, which did not use spline functions. It seems hard to use spline functions, which are piecewise polynomial and have to be fitted together correctly on the boundaries of the building blocks, when the building blocks have fractal boundaries as in these examples. Mongeau's work, by using appropriate powers of Λ, reduces everything to pure dilations, and thereby avoids the fractal boundaries, at the price of having to deal with much larger matrices $\mathbf{T}_d$. It is also possible to deal directly with the fractal sets by extending our original, longer, and more complicated proof without spline functions (as shown in Cohen and Daubechies (1991)). Therefore, we would like to outline here the major steps of the proofs of Theorems 2.2 and 3.1 without spline functions. For simplicity, we return for this outline to the one-dimensional case, with $\Lambda=$ multiplication by 2.

What follows is an outline of the proof of Theorem 2.1. The starting point is the equation

$$\begin{aligned}\mathbf{v}(x)&=\mathbf{T}_{d_1(x)}\cdots\mathbf{T}_{d_m(x)}\mathbf{v}(\tau^m x)\\&=\mathbf{T}(m;\,d(x))\mathbf{v}(\tau^m x).\end{aligned}$$

From $\mathbf{v}(0),\mathbf{v}(1)$ (which can be determined from the right eigenvector of $\mathbf{M}$ for the eigenvalue 1—see § I.5), we can thus derive $\mathbf{v}(x)$ for all dyadic rationals x. Existence and continuity of $\mathbf{v}$ will be proved if this definition of $\mathbf{v}$ on a dense set can be continuously extended. This is done in the following steps:

- $\mathbf{e}_1$, the common left eigenvector of $\mathbf{T}_0,\mathbf{T}_1$ for the eigenvalue 1, is a left eigenvector of every $\mathbf{T}(m;x)$, with eigenvalue 1;
- consequently, $\mathbf{e}_1\cdot\mathbf{v}(x)=1$ for every dyadic rational x;
- $\|\mathbf{v}(x)\|$ is uniformly bounded, for all dyadic rationals (this uses the condition (2.11), together with $\mathbf{v}(x)-\mathbf{v}(x')\in E_1$);
- we then use that, for small enough t,

$$\mathbf{v}(x+t)-\mathbf{v}(x)=\mathbf{T}(m;\,d(x))[\mathbf{v}(\tau^m x+2^m t)-\mathbf{v}(\tau^m x)],$$

together with (2.16), to show that the restriction of $\mathbf{v}$ to the dyadic rationals is continuous and satisfies

$$\|\mathbf{v}(x)-\mathbf{v}(y)\|\leqq C\lambda^m$$

if $|x-y|\leqq 2^{-m}$. This suffices to show that $\mathbf{v}$ has a continuous extension to all of $[0,1]$, and to prove the Hölder continuity of this extension.

In the details of the proof, we have to be careful, occasionally, because dyadic rationals have two binary extensions. This does not cause any real problem.

Next we sketch the proof of Theorem 3.1.

- Every $\mathbf{T}(m; d)$ has a left eigenvector $\mathbf{e}_k(m; d)$ for the eigenvalue $2^{-(k-1)}m$, $1 \leqq k \leqq L+1$, which can be written as $\mathbf{e}_k(m; d) = \mathbf{e}_k^0 +$ combination of $\mathbf{e}_{k'}^0$, $k' < k$. These $\mathbf{e}_k(m; d)$ are bounded uniformly in $m \in \mathbb{N}$ and $d \in \{0, 1\}^{\mathbb{N}}$. (This is a purely combinatorial fact, and we do not need (3.3) to prove it.)
- For $k \leqq l+1$, the eigenvalue $2^{-(k-1)m}$ of $\mathbf{T}(m; d)$ is simple; there exists a corresponding right eigenvector $\tilde{\mathbf{e}}_k(m; d)$, uniquely normalized by $\mathbf{e}_k(m; d) \cdot \tilde{\mathbf{e}}_k(m; d) = \mathbf{e}_k^0 \cdot \tilde{\mathbf{e}}_k(m; d) = 1$. These $\tilde{\mathbf{e}}_k(m; d)$ are bounded uniformly in m and d. (To prove this we need (3.3); the proof is analogous to (7) in the proof of Theorem 4.1.)

$$\|\mathbf{T}(m; d)|_{E_{k+1}}\| \leqq C\lambda^m 2^{-ml}, \qquad l+1 \leqq k \leqq L,$$

$$\|\mathbf{T}(m; d)|_{E_{l+1}}\| \leqq C\lambda^m 2^{-ml}\phi_\lambda(m),$$

where

$$\phi_\lambda(m) = 1 \quad \text{if } \lambda > \tfrac{1}{2}, \quad \phi_{1/2}(m) = m,$$

and

$$\|\mathbf{T}(m; d)|_{E_{k+1}}\| \leqq C2^{-mk}, \qquad 0 \leqq k \leqq l-1.$$

(Part of this is Lemma 3.5; the third bound is proved analogously.) By taking $k = 0$, we see that (3.3) implies (2.15), so that $\mathbf{v}$ can be constructed and is continuous by Theorem 2.3.

- The $\tilde{\mathbf{e}}_k(n; d)$ converge to a limit for $n \to \infty$. Moreover, if x is a dyadic rational, then $\lim_{n\to\infty} \tilde{\mathbf{e}}_k(n; d^+(x)) = \lim_{n\to\infty} \tilde{\mathbf{e}}_k(n; d^-(x))$. We can, therefore, define $\tilde{\mathbf{e}}_k(x) = \lim_{n\to\infty} \tilde{\mathbf{e}}_k(n; d(x))$ for all $x \in [0, 1]$, $l \leqq k \leqq l+1$. Finally $\tilde{\mathbf{e}}_k(x)$ is continuous in x.
- For all $x \in [0, 1]$, $k = 0, \cdots, L$,

$$\mathbf{e}_{k+1}^0 \cdot \mathbf{v}(x) = (-1)^k x^k$$

(see the proof of Lemma 3.6).

- With all these ingredients, we prove, e.g., differentiability of $\mathbf{v}$ as follows: for t small enough,

$$\begin{aligned}\mathbf{v}(x+t) - \mathbf{v}(x) &= \mathbf{T}(m; d(x))[\mathbf{v}(\tau^m x + 2^m t) - \mathbf{v}(\tau^m x)] \\ &= 2^{-m}\tilde{\mathbf{e}}_2(m; d(x))\{\mathbf{e}_2(m; d(x)) \cdot [\mathbf{v}(\tau^m x + 2^m t) - \mathbf{v}(\tau^m x)]\} \\ &\quad + \mathbf{T}(m; d(x))\mathbf{r},\end{aligned}$$

where we have used $\mathbf{e}_1(m; d(x)) \cdot [\mathbf{v}(y) - \mathbf{v}(z)] = 0$ and where $\mathbf{r} \in E_2$, $\|\mathbf{r}\|$ bounded independently of x, t, and m. Consequently,

$$\frac{\mathbf{v}(x+t) - \mathbf{v}(t)}{t} = \tilde{\mathbf{e}}_2(m; d(x)) + o(1) \to \tilde{\mathbf{e}}_2(x),$$

and $\mathbf{v}$ is differentiable in x if $l \geqq 1$.

- Higher-order differentiability is established analogously. We find

$$\frac{d^k}{dx^k}\mathbf{v}(x) = (-1)^k k!\tilde{\mathbf{e}}_k(x), \qquad j = 0, \cdots, l.$$

(There is no need to "guess" the ansatz (3.10) with this approach.)

- All these results for $\mathbf{v}$ must then be "unfolded" to state results for f; in principle, problems could occur at the integers. Since $[\tilde{\mathbf{e}}_k(0)]_1 = 0 = [\tilde{\mathbf{e}}_k(1)]_N$ and $[\tilde{\mathbf{e}}_k(0)]_{m+1} = [\tilde{\mathbf{e}}_k(1)]_m$, $m = 1, \cdots\cdots, N-1$, unfolding works without problems.

Here ends our outline of the proof without spline functions.

REFERENCES

M. Berger and Y. Wang (1991), *Bounded semi-groups of matrices*, Linear Algebra Appl., to appear.

A. Cavaretta, W. Dahmen, and C. Micchelli (1989), *Stationary subdivision*, Mem. Amer. Math. Soc., 453, pp. 1-186.

A. Cohen and I. Daubechies (1991), *Non-separable two dimensional wavelet bases*, Rev. Mat. Ibenoamericana, submitted.

I. Daubechies (1988), *Orthonormal bases of wavelets with compact support*, Comm. Pure Appl. Math., 41, pp. 909-996.

I. Daubechies and J. C. Lagarias (1991), *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161, pp. 227-263.

——— (1991), *Two-scale difference equations.* I. *Global regularity of solutions*, SIAM J. Math. Anal., 22, pp. 1388-1410.

G. de Rham (1956), *Sur une courbe plane*, J. Math. Pures Appl., 35, pp. 25-42.

——— (1957), *Sur un exemple de fonction continue sans dérivée*, Enseign. Math., 3, pp. 71-72.

——— (1959), *Sur les courbes limites de polygones obtenue par trisection*, Enseign. Math., 5, pp. 29-43.

G. Deslauriers and S. Dubuc (1987), *Interpolation dyadique*, in Fractals, dimensions non entirères et applications, G. Cherbit, ed., Masson, Paris.

——— (1989), *Symmetric iterative interpolation processes*, Constr. Approx., 5, pp. 49-68.

S. Dubuc (1986), *Interpolation through an iterative scheme*, J. Math. Anal. Appl., 114, pp. 185-204.

N. Dyn, J. A. Gregory, and D. Levin (1987), *A 4-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Design, 4, pp. 257-268.

——— (1991), *Analysis of uniform binary subdivision schemes for curve design*, Constr. Approx. 7, pp. 127-147.

N. Dyn and D. Levin (1989), *Interpolating subdivision schemes for the generation of curves and surfaces*, to appear.

——— (1990), *Uniform subdivision algorithms for curves and surfaces*, Constr. Approx., to appear.

H. G. Eggleston (1949), *The fractional dimension of a set defined by decimal properties*, Quart. J. Math. Oxford Ser., 20, pp. 31-36.

I. J. Good (1941), Proc. Cambridge Phil. Soc., 37, p. 200.

K. Gröchenig and W. R. Madych (1992), *Multiscale analysis, Haar bases and self-similar tilings of* $\mathbb{R}''$, IEEE Trans. Inform. Theory, 38, pp. 556-568.

W. Lawton and H. Resnikoff (1991), *Multidimensional wavelet bases*, SIAM J. Math. Anal., submitted.

S. Mallat (1989), *Multiresolution approximation and wavelet orthonormal bases of* L^2, Trans. Amer. Math. Soc., 315, pp. 69-88.

Y. Meyer (1990), *Ondelettes*, Vol. 1, in Ondelettes et Opérateurs, Hermann, Paris.

C. A. Micchelli (1986), *Subdivision algorithms for curves and surfaces*, Proc. SIGGRAPH, 1986, Dallas, TX.

C. A. Micchelli and H. Prautzsch (1987), *Refinement and subdivision for spaces of integer translates of compactly supported functions* in Numerical Analysis, D. F. Griffith, G. A. Watson, eds., Academic, New York, pp. 192-222.

——— (1987), *Computing curves invariant under halving*, Comput. Aided Geom. Design, 4, pp. 133-140.

——— (1989), *Uniform refinement of curves*, Linear Algebra Appl., 114/115, pp. 841-870.

J. P. Mongeau (1990), *Propriétés de l'interpolation itérative*, Ph.D. thesis, Ecole Polytechnique, Montréal.

G.-C. Rota and W. G. Strang (1960), *A note on the joint spectral radius*, Indag. Math., 22, pp. 379-381.

L. L. Schumaker (1981), *Spline Functions: Basic Theory*, John Wiley, New York.

M. Vetterli (1984), *Multi-dimensional subband coding: some theory and algorithms*, Signal Process., 6, pp. 97-112.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1081-1098, September 1992

QUASI-CONVEX INTEGRANDS AND LOWER SEMICONTINUITY IN L^1*

IRENE FONSECA† AND STEFAN MÜLLER‡

Abstract. In this paper it is shown that, under mild continuity and growth hypotheses, if $f(x, u,.)$ is quasi-convex and if $u_n, u \in W^{1,1}$ are such that $u_n \to u$ in L^1, then

$$\int_\Omega f(x, u(x), \nabla u(x))\, dx \leqq \liminf \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx.$$

The proof relies on a blowup argument in connection with a truncation result that allows one to consider uniformly convergent sequences.

Key words. quasi-convexity, lower semicontinuity

AMS(MOS) subject classification. 49

1. Introduction. Our objective is to discuss lower semicontinuity of the functional

$$I(u) = \int_\Omega f(x, u(x), \nabla u(x))\, dx, \qquad u \in W^{1,1}(\Omega; \mathbb{R}^p),$$

where f is quasi convex in ∇u. Our result is that if

$$(1.1) \qquad u_n, u \in W^{1,1}(\Omega; \mathbb{R}^p) \quad \text{and} \quad u_n \to u \quad \text{in } L^1(\Omega; \mathbb{R}^p),$$

then

$$I(u) \leqq \liminf I(u_n),$$

provided f has linear growth in ∇u and satisfies some technical conditions. Another way to express (1.1) is to say that $u_n, u \in W^{1,1}(\Omega; \mathbb{R}^p)$ and $u_n \to u$ in $\mathcal{D}'(\Omega; \mathbb{R}^p)$, which is much less stringent than assuming, for example, weak convergence in $W^{1,1}(\Omega; \mathbb{R}^p)$. This lower semicontinuity result was obtained by Dal Maso [DM] in the scalar case $p = 1$; in the vector-valued case and for $f = f(A)$ convex, by Ball and Murat [BM] and Reshetnyak [R]; when $p > 1$ and $f = f(x, \nabla u)$ quasi-convex the problem was addressed by Fonseca [Fo] and, independently, by Kinderlehrer [K]. For the case where $f = f(x, u, a)$ and $f(x, u,.)$ is convex, Aviles and Giga [AG] obtained lower semicontinuity results.

The main new tool involved in this paper is a careful truncation technique which, together with a blowup argument, enables us to reduce to the case where the sequence u_n converges uniformly. Murat has informed us that related truncation arguments are used in the context of renormalized solutions to partial differential equations (see, e.g., [BDGM]).

The study of this problem was motivated by the analysis of variational problems for phase transitions and the related question of understanding the relaxation of functionals of the type

$$(1.2) \qquad u \to \int_\Omega f(x, u(x), \nabla u(x))\, dx$$

* Received by the editors August 28, 1991; accepted for publication (in revised form) February 11, 1992.

† Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. The research was partially supported by National Science Foundation grant DMS-9000133.

‡ Department of Mathematics, Carnegie Mellon University, and Institut für Angewandte Mathematik Universität Bonn, Beringstrasse 4, W-5300 Bonn 1, Germany. The research was partially supported by National Science Foundation grant DMS-9002679 and by SFB256 at the University of Bonn.

in spaces admitting discontinuous functions u. An important example consists of the family of singular perturbations

$$E_\varepsilon(u) := \int_\Omega W(u(x))\,dx + \varepsilon^2 \int_\Omega h^2(\nabla u(x))\,dx$$

of the nonconvex energy

$$E(u) := \int_\Omega W(u(x))\,dx,$$

where W has two potential wells at a and b. Depending on the constraints or boundary conditions imposed on the admissible functions, $E(\cdot)$ often admits infinitely many minimizers that are piecewise constant functions of bounded variation, $u \in \{a, b\}$ almost everywhere in Ω. In the search for a reasonable selection criterion the properties of the limits of sequences of minimizers for the perturbed problems (see [FT1], [G1], [G2], [KS], [Mo], [OS]) are studied. The natural notion of convergence for the functional in this context is Γ-convergence, as introduced by De Giorgi [DG] (see [At], [DM], [DD] for more recent expositions).

In the isotropic scalar case, i.e., if $u : \Omega \to \mathbb{R}$ and $h = \|\cdot\|$, using an idea of Modica and Mortola, Modica [Mo] showed that the $\Gamma(L^1)$ limit of the rescaled energies

$$J_\varepsilon(u) := \frac{1}{\varepsilon} E_\varepsilon(u)$$

is given by

$$J_0(u) = \mathscr{F}(u),$$

where

$$\mathscr{F}(u) := \inf_{\{u_n\}} \left\{ \liminf_{n\to+\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x))\,dx \,\middle|\, u_n \in W^{1,1}(\Omega; \mathbb{R}),\ u_n \to u \text{ in } L^1 \right\}$$

is the relaxation in $\mathrm{BV}(\Omega; \mathbb{R})$ of (1.2) and

$$f(x, u, A) = 2\sqrt{W(u)}\, h(A). \tag{1.3}$$

Precisely, if $u \in \{a, b\}$ almost everywhere and if $\{u = a\}$ has finite perimeter in Ω, then

$$\inf_{\{u_\varepsilon\}} \{\liminf J_\varepsilon(u_\varepsilon) \,|\, u_\varepsilon \in W^{1,1}(\Omega; \mathbb{R}),\ u_\varepsilon \to u \text{ in } L^1\} = \mathscr{F}(u).$$

This result was generalized by [OS] to "anisotropic" functions h with linear growth for which h^2 is convex and the integral representation for the relaxation $\mathscr{F}(\cdot)$ was obtained by Dal Maso [DM].

In this work we consider the case were u is vectorial, and we prove lower semicontinuity of (1.2) in L^1, thus obtaining the absolutely continuous part of $\mathscr{F}(u)$ with respect to the N-dimensional Lebesgue measure.

2. Lower semicontinuity in $\mathbf{L}^1$ for quasi-convex integrands. Let $p, N \geqq 1$ and let $M^{p\times N}$ denote the vector space of all $p \times N$ real matrices. Recall that a Borel function $f : M^{p\times N} \to \mathbb{R}$ is said to be *quasi convex* if

$$f(A) \leqq \frac{1}{\operatorname{meas}(D)} \int_D f(A + \nabla\varphi(x))\,dx$$

for all $A \in M^{p\times N}$, for every domain $D \subset \mathbb{R}^N$, and for all $\varphi \in W_0^{1,\infty}(D; \mathbb{R}^p)$. If $|f(A)| \leqq C(1+\|A\|)$ we can easily show by approximation that the inequality holds for all $\varphi \in W_0^{1,1}(D; \mathbb{R}^p)$.

Let $\Omega \subset \mathbb{R}^N$ be an open, bounded domain, and let

$$f: \Omega x \mathbb{R}^p \times M^{p\times N} \to [0, +\infty).$$

We consider the following hypotheses on f:

(H1) f is continuous;

(H2) $f(x, u,.)$ is quasi convex;

(H3) There exists a nonnegative, bounded, continuous function $g: \Omega \times \mathbb{R}^p \to [0, +\infty)$, c, $C > 0$ such that

$$cg(x, u)\|A\| - C \leqq f(x, u, A) \leqq Cg(x, u)(1 + \|A\|)$$

for all $(x, u, A) \in \Omega x \mathbb{R}^p x M^{pxN}$;

(H4) For all $(x_0, u_0) \in \Omega x \mathbb{R}^p$ and for all $\varepsilon > 0$ there exists $\delta > 0$ such that $|x - x_0| + |u - u_0| < \delta$ implies that

$$f(x_0, u, A) - f(x_0, u_0, A) \geqq -\varepsilon(1 + \|A\|),$$

and

$$|f(x_0, u, A) - f(x, u, A)| \leqq \varepsilon(1 + \|A\|).$$

THEOREM 2.1. *Suppose* (H1)-(H4) *hold. If* $u_n, u \in W^{1,1}(\Omega; \mathbb{R}^p)$ *and* $u_n \to u$ *in* $L^1(\Omega; \mathbb{R}^p)$, *then*

$$\int_\Omega f(x, u(x), \nabla u(x))\,dx \leqq \liminf \int_\Omega f(x, u_n(x), \nabla u_n(x))\,dx. \tag{2.1}$$

Remark 2.2. (i) If (H2) is replaced by convexity and if the growth condition (H3) holds, then the hypothesis $(H4)_1$ presents no restriction. This fact will be examined in § 4.

(ii) Lower semicontinuity for functions of the type (1.2) follows from Theorem 2.1. Indeed, suppose that

$$f(x, u, A) = 2\sqrt{W(u)}\,h(A),$$

where h is a nonnegative quasi-convex function and

$$c\|A\| - C \leqq h(A) \leqq C(1 + \|A\|).$$

Set

$$W_M(u) := \min\{M, W(u)\} \quad \text{and} \quad f_M(u, A) := 2\sqrt{W_M(u)}\,h(A).$$

It is clear that f_M satisfies (H1)-(H4) and so, if u_n, $u \in W^{1,1}(\Omega; \mathbb{R}^p)$ are such that $u_n \to u$ in $L^1(\Omega; \mathbb{R}^p)$, then

$$\begin{aligned}\int_\Omega f_M(u(x), \nabla u(x))\,dx &\leqq \liminf \int_\Omega f_M(u_n(x), \nabla u_n(x))\,dx \\ &\leqq \liminf \int_\Omega f(u_n(x), \nabla u_n(x))\,dx.\end{aligned}$$

Letting $M \to +\infty$ and using the monotone convergence theorem, we conclude (2.1).

(iii) As we showed in (ii) the boundedness of g presents no restriction for the examples that we have in mind. This assumption becomes crucial for proving in

Proposition 2.4 that the u_n may be considered to be smooth functions, which in turn allows us to apply in $(2.14)_2$ the change of variables formula (2.3) for Lipschitz functions.

It is possible to remove in (H3) the boundedness constraint imposed on g by using a suitable generalization of the change of variables formula (2.3) for $W^{1,1}$ functions. For the sake of clarity, however, we focus attention on the case where g is bounded.

The main idea of the proof is to use a blowup argument to localize (2.1) (see 2.5) and Step 2 in the proof of Theorem 2.1) and a careful truncation technique for vector-valued functions which allows us to replace L^1 convergence by uniform convergence (see Lemmas 2.6 and Step 3 in the proof of Theorem 2.1). First we recall some auxiliary results.

PROPOSITION 2.3. *If* $f: M^{pxN} \to \mathbb{R}$ *is quasi convex and if* $|f(A)| \leqq C(1+\|A\|)$ *for all* $A \in M^{pxN}$ *and for some constant* $C>0$, *then there exists a constant* $C'=C'(C, N)$ *such that*

$$|f(A)-f(B)| \leqq C'\|A-B\|$$

for all $A, B \in M^{pxN}$.

Proof. This follows from the rank-one convexity of f. We refer the reader to Dacorogna [D, Chap. 4, Lemma 2.2] or Evans [E], for example.

PROPOSITION 2.4. (i) *If Theorem* 2.1 *holds true for* Ω *being a ball, then it holds true for all open, bounded sets* Ω.

(ii) *Let* Ω *be a ball. If* (H1) *and* (H3) *hold and if* u_n, $u \in W^{1,1}(\Omega; \mathbb{R}^p)$ *are such that* $u_n \to u$ *in* $L^1(\Omega; \mathbb{R}^p)$, *then there exist* $\tilde{u}_n \in C_0^\infty(\mathbb{R}^N; \mathbb{R}^p)$ *such that* $\|\tilde{u}_n - u\|_{L^1_{(\Omega)}} \to 0$ *and*

$$\liminf_{n\to+\infty} \int_\Omega f(x, \tilde{u}_n(x), \nabla \tilde{u}_n(x))\, dx = \liminf_{n\to+\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx.$$

Proof. The proof follows essentially the argument by Acerbi and Fusco [AF], and for completeness it is included in § 3.

PROPOSITION 2.5. *Let* $f: M^{pxN} \to \mathbb{R}$ *be a function satisfying* (H1), (H2), *and*

$$0 \leqq f(A) \leqq C(1+\|A\|), \qquad A \in M^{pxN}$$

for some $C>0$. *If* $A_0 \in M^{pxN}$ *and if* $u_n \in W^{1,1}(\Omega; \mathbb{R}^p)$ *are such that* $u_n \to 0$ *in* $L^1(\Omega; \mathbb{R}^p)$ *and* $\{\|\nabla u_n\|_{L^1}\}$ *is bounded, then*

$$\text{meas}\,(\Omega) f(A_0) \leqq \liminf \int_\Omega f(A_0 + \nabla u_n(x))\, dx.$$

Proof. See § 3.

We will also use the following results.

If $u \in W^{1,1}(\Omega; \mathbb{R}^p)$, then for almost everywhere $x_0 \in \Omega$

$$(2.2) \quad \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left\{\frac{1}{\varepsilon^N}\int_{B(x_0,\varepsilon)} |u(x)-u(x_0)-\nabla u(x_0)(x-x_0)|^{N/(N-1)}\, dx\right\}^{(N-1)/N} = 0.$$

If $w \in W^{1,\infty}(\mathbb{R}^N; \mathbb{R})$ and $g \in L^1(\mathbb{R}^N; \mathbb{R})$, then the *change of variables formula* (or *co-area formula*) holds, namely

$$(2.3) \qquad \int_{\mathbb{R}^N} g(x)|\nabla w(x)|\, dx = \int_{-\infty}^{+\infty} \left(\int_{w^{-1}(t)} g(x)\, dH_{N-1}(x)\right) dt.$$

For details see Calderon and Zygmund [CZ], Evans and Gariepy [EG], and Ziemer [Zi]. An easy consequence of (2.3) is the following estimate on level sets of $W^{1,\infty}$ functions.

LEMMA 2.6. *Let* $v \in W^{1,\infty}_{\text{loc}}(\mathbb{R}^N; \mathbb{R}^p)$; *let* $0 < \alpha < \beta < L$, *and let* $C_0 > 0$ *be such that*

$$\int_{\{|v| \leqq L\} \cap B(0,1)} \|\nabla v(x)\| \, dx \leqq C_0.$$

Then

$$\operatorname*{ess\,inf}_{t \in (\alpha,\beta)} tH_{N-1}(\{x \in B(0,1) | \, |v(x)| = t\}) \leqq \frac{C_0}{\ln(\beta/\alpha)}.$$

Proof. Let $B := B(0,1)$, and consider a cutoff function $\varphi \in C_0^\infty(\mathbb{R}^N; \mathbb{R})$ such that $\varphi = 1$ in $B(0,1)$ and its support is contained in $B(0,2)$. Applying the co-area formula (2.3) to

$$w(x) := \varphi(x)|v(x)| \quad \text{and} \quad g(x) := \chi_{[0,L]}(|v(x)|)\chi_B(x),$$

we have

$$\begin{aligned}\int_0^L H_{N-1}(\{x \in B| \, |v(x)| = t\}) \, dt &= \int_{\{|v| \leqq L\} \cap B(0,1)} \|D|v(x)|\| \, dx \\ &\leqq \int_{\{|v| \leqq L\} \cap B(0,1)} \|\nabla v(x)\| \, dx \leqq C_0.\end{aligned}$$

And so, if

$$\operatorname*{ess\,inf}_{t \in (\alpha,\beta)} tH_{N-1}(\{x \in B| \, |v_n(x)| = t\}) = a,$$

then

$$\begin{aligned}C_0 \geqq \int_\alpha^\beta H_{N-1}(\{x \in B| \, |v_n(x)| = t\}) \, dt &\geqq \int_\alpha^\beta \frac{a}{t} \, dt \\ &= a \ln\left(\frac{\beta}{\alpha}\right).\end{aligned}$$

Thus

$$\operatorname*{ess\,inf}_{t \in (\alpha,\beta)} tH_{N-1}(\{x \in B| \, |v_n(x)| = t\}) \leqq \frac{C_0}{\ln(\beta/\alpha)}. \qquad \square$$

Proof of Theorem 2.1. In the sequel, using Proposition 2.4, we assume Ω is a ball and that $u_n \in C_0^\infty(\mathbb{R}^N; \mathbb{R}^p)$. In addition, suppose, without loss of generality, that

$$\liminf_{n \to +\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x)) \, dx = \lim_{n \to +\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x)) \, dx < +\infty.$$

Step 1 (localization). We first reduce the problem to verifying the pointwise inequality (2.5) below. As f is nonnegative, there exists a subsequence such that

$$f(\cdot, u_n(\cdot), \nabla u_n(\cdot)) \to \mu \quad \text{weakly} * \text{in the sense of measures},$$

where μ is a nonnegative finite measure. Using the Radon-Nikodym theorem, we can write μ as a sum of two mutually singular nonnegative measures

$$\mu = \mu_a(x)L_N + \mu_s,$$

where L_N denotes the Lebesgue measure in $\mathbb{R}^N$ and for almost everywhere $x_0 \in \Omega$

$$\mu_a(x_0) = \lim_{\varepsilon \to 0} \frac{\mu(B(x_0, \varepsilon))}{L_N(B(x_0, \varepsilon))} < +\infty. \tag{2.4}$$

We claim that

$$\mu_a(x_0) \geqq f(x_0, u(x_0), \nabla u(x_0)) \quad \text{for a.e. } x_0 \in \Omega. \tag{2.5}$$

Assuming (2.5) momentarily, consider an increasing sequence of smooth cutoff functions φ_k, with $0 \leqq \varphi_k \leqq 1$ and $\sup_k \varphi_k(x) = 1$ in Ω. We obtain

$$\begin{aligned}\lim_{n \to +\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx &\geqq \liminf_{n \to +\infty} \int_\Omega \varphi_k(x) f(x, u_n(x), \nabla u_n(x))\, dx \\ &= \int_\Omega \varphi_k(x)\, d\mu(x) \geqq \int_\Omega \varphi_k(x) \mu_a(x)\, dx \\ &\geqq \int_\Omega \varphi_k(x) f(x, u(x), \nabla u(x))\, dx.\end{aligned}$$

Letting $k \to +\infty$, the result follows now from the monotone convergence theorem. The rest of this section is dedicated to proving claim (2.5).

Step 2 (blowup). We use a blowup argument in connection with (2.2) to derive a lower bound for $\mu_a(x_0)$. Let x_0 be a Lebesgue point for $u, \nabla u$ and such that (2.2) and (2.4) hold, and consider the affine functions

$$u_0(x) := u(x_0) + \nabla u(x_0)x \quad \text{and} \quad w_0(x) := \nabla u(x_0)x.$$

We abbreviate $B := B(0, 1)$, and we consider a subdomain $B' \Subset B$. We claim that there exist sequences $r_n \to 0^+$ and $w_n \in W^{1,\infty}(\mathbb{R}^N; \mathbb{R}^p)$ such that $w_n \to w_0$ in $L^1(B; \mathbb{R}^p)$, and

$$\mu_a(x_0) \geqq \lim_{n \to +\infty} \frac{1}{\text{meas}\,(B)} \int_{B'} f(x_0 + r_n x, u(x_0) + r_n w_n(x), \nabla w_n(x))\, dx. \tag{2.6}$$

Let $\varphi \in C_0(B)$ be a cutoff function such that $0 \leqq \varphi \leqq 1$ and $\varphi(x) = 1$ if $x \in B'$. By (2.4) we have

$$\begin{aligned}\mu_a(x_0) &= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon^N \text{meas}\,(B)} \mu(B(x_0, \varepsilon)) \\ &\geqq \limsup_{\varepsilon \to 0} \frac{1}{\varepsilon^N \text{meas}\,(B)} \int_{B(x_0, \varepsilon)} \varphi\left(\frac{x - x_0}{\varepsilon}\right) d\mu(x) \\ &= \limsup_{\varepsilon \to 0} \lim_{n \to +\infty} \frac{1}{\varepsilon^N \text{meas}\,(B)} \int_{B(x_0, \varepsilon)} \varphi\left(\frac{x - x_0}{\varepsilon}\right) f(x, u_n(x), \nabla u_n(x))\, dx \qquad (2.7) \\ &= \limsup_{\varepsilon \to 0} \lim_{n \to +\infty} \frac{1}{\text{meas}\,(B)} \int_B \varphi(x) f(x_0 + \varepsilon x, u_n(x_0 + \varepsilon x), \nabla u_n(x_0 + \varepsilon x))\, dx \\ &\geqq \limsup_{\varepsilon \to 0} \limsup_{n \to +\infty} \frac{1}{\text{meas}\,(B)} \int_{B'} f(x_0 + \varepsilon x, u(x_0) + \varepsilon w_{n,\varepsilon}(x), \nabla w_{n,\varepsilon}(x))\, dx,\end{aligned}$$

where

$$w_{n,\varepsilon}(x) := \frac{u_n(x_0+\varepsilon x)-u(x_0)}{\varepsilon}$$
$$= \frac{1}{\varepsilon}\,[u_n(x_0+\varepsilon x)-u_0(\varepsilon x)] + w_0(x).$$

By (2.2) and Hölder's inequality

$$\lim_{\varepsilon\to 0}\lim_{n\to+\infty}\|w_{n,\varepsilon}-w_0\|_{L^1(B)} = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_B |u(x_0+\varepsilon x)-u_0(\varepsilon x)|\,dx$$
$$= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon^{N+1}}\int_{B(x_0,\,\varepsilon)} |u(x)-u(x_0)-\nabla u(x_0)(x-x_0)|\,dx$$
$$=0.$$

Now (2.6) is obtained by a standard diagonalization argument. Indeed, choose a sequence $r_k\to 0$, and choose n_k such that

$$\|w_{n_k,r_k}-w_0\|_{L^1(B)} < 1/k + \lim_{n\to+\infty}\|w_{n,r_k}-w_0\|_{L^1(B)}$$

and

$$\frac{1}{\operatorname{meas}(B)}\int_{B'} f(x_0+r_kx, u(x_0)+r_kw_{n_k,r_k}(x), \nabla w_{n_k,r_k}(x))\,dx$$
$$\leqq 1/k + \limsup_{n\to+\infty}\frac{1}{\operatorname{meas}(B)}\int_{B'} f(x_0+r_kx,\, u(x_0)+r_kw_{n,r_k}(x),\, \nabla w_{n,r_k}(x))\,dx.$$

Letting

$$w_k := w_{n_k,r_k},$$

(2.6) follows from (2.7) (a further subsequence may be chosen to ensure that the limit on the right-hand side of (2.6) exists).

Step 3 (truncation). We show that the sequence w_n constructed in Step 2 can be replaced by a uniformly convergent sequence. More precisely, we claim that if $g(x_0, u(x_0))>0$, then there exists a sequence $\tilde{w}_n\in W^{1,\infty}_{\text{loc}}(\mathbb{R}^N;\mathbb{R}^p)$ such that $\|\tilde{w}_n\|_{1,1,B'}\leqq$ Const., $\tilde{w}_n\to w_0$ in $L^\infty(B;\mathbb{R}^p)$, and

$$\mu_a(x_0) \geqq \lim_n \frac{1}{\operatorname{meas}(B)}\int_{B'} f(x_0+r_nx,\, u(x_0)+r_n\tilde{w}_n(x),\, \nabla\tilde{w}_n(x))\,dx. \tag{2.8}$$

Let $0<s<t<1$, and let $\varphi_{s,t}$ be a cutoff function such that $0\leqq\varphi_{s,t}\leqq 1$, $\varphi_{s,t}(\tau)=1$ if $\tau\leqq s$, $\varphi_{s,t}(\tau)=0$ if $\tau\geqq t$, $\|\varphi'_{s,t}\|_\infty\leqq C(t-s)^{-1}$. Set

$$\phi^n_{s,t}(x) := \varphi_{s,t}(|w_n(x)-w_0(x)|)$$

and

$$w^n_{s,t}(x) := w_0(x) + \varphi_{s,t}(|w_n(x)-w_0(x)|)(w_n(x)-w_0(x)).$$

Clearly

$$\|w^n_{s,t}-w_0\|_\infty \leqq t. \tag{2.9}$$

Define

$$h_n(x,s,A) := f(x_0+r_nx,\, u(x_0)+r_ns,\, A),$$

and let $L=\|w_0\|_{L^\infty(B)}+1$. By (H3) and as $g(x_0, u(x_0))>0$, g continuous, there exists n_0 such that for all $n\geqq n_0$, $|s|\leqq L$

$$C(\|A\|+1)\geqq h_n(x, s, A)\geqq c\|A\|-C \tag{2.10}$$

for some $c, C>0$. Also

$$\begin{aligned}\int_{B'} h_n(x, w^n_{s,t}(x), \nabla w^n_{s,t}(x))\,dx &= \int_{B'\cap\{|w_n(x)-w_0(x)|\leqq s\}} h_n(x, w_n(x), \nabla w_n(x))\,dx\\ &+\int_{B'\cap\{s<|w_n(x)-w_0(x)|\leqq t\}} h_n(x, w^n_{s,t}(x), \nabla w^n_{s,t}(x))\,dx\\ &+\int_{B'\cap\{|w_n(x)-w_0(x)|>t\}} h_n(x, w_0(x), \nabla w_0(x))\,dx,\end{aligned} \tag{2.11}$$

and by (2.10) we have

$$-C\leqq h_n(x, w_0(x), \nabla w_0(x))\leqq C,$$

which implies that

$$\int_{B'\cap\{|w_n(x)-w_0(x)|>t\}} h_n(x, w_0(x), \nabla w_0(x))\,dx\leqq C\text{ meas}\,\{x\in B\,|\,|w_n(x)-w_0(x)|>t\}. \tag{2.12}$$

On the other hand, if $s<|w_n(x)-w_0(x)|<t$, then

$$\begin{aligned}\nabla w^n_{s,t}(x)&=\nabla u(x_0)+\varphi_{s,t}(|w_n(x)-w_0(x)|)(\nabla w_n(x)-\nabla u(x_0))\\ &\quad+(w_n(x)-w_0(x))\otimes\varphi'_{s,t}(|w_n(x)-w_0(x)|)\nabla|w_n(x)-w_0(x)|.\end{aligned}$$

Thus, by (2.10), we have

$$\begin{aligned}&\int_{B'\cap\{s<|w_n(x)-w_0(x)|\leqq t\}} h_n(x, w^n_{s,t}(x), \nabla w^n_{s,t}(x))\,dx\\ &\leqq C\int_{\{s<|w_n(x)-w_0(x)|\leqq t\}}(1+\|\nabla w_n(x)-\nabla u(x_0)\|)\,dx\\ &\quad+C\frac{1}{t-s}\int_{B'\cap\{s<|w_n(x)-w_0(x)|\leqq t\}}|w_n(x)-w_0(x)|\,|\nabla|w_n(x)-w_0(x)||\,dx.\end{aligned} \tag{2.13}$$

We remark that for almost all t we have

$$\lim_{s\to t-}\int_{\{s<|w_n(x)-w_0(x)|\leqq t\}}(1+\|\nabla w_n(x)-\nabla u(x_0)\|)\,dx=0, \tag{2.14a}$$

and by the change of variables formula (2.3)

$$\begin{aligned}&\lim_{s\to t-}\frac{1}{t-s}\int_{B'\cap\{s<|w_n(x)-w_0(x)|\leqq t\}}|w_n(x)-w_0(x)|\,|\nabla|w_n(x)-w_0(x)||\,dx\\ &\leqq tH_{N-1}\{x\in B'\,|\,|w_n(x)-w_0(x)|=t\}\end{aligned} \tag{2.14b}$$

for almost every t. Due to (2.10),

$$\begin{aligned}\int_{B'\cap\{|w_n(x)-w_0(x)|\leqq 1\}}|\nabla|w_n(x)-w_0(x)||\,dx&\leqq\int_{B'\cap\{|w_n(x)-w_0(x)|\leqq 1\}}(\|\nabla w_n(x)\|+C)\,dx\\ &\leqq C\int_{B'}[h_n(x, w_n(x), \nabla w_n(x))+1]\,dx\\ &\leqq\text{Const.}\end{aligned}$$

since the latter sequence is convergent. Hence, by Lemma 2.6 there exists $t_n \in [\|w_n - w_0\|_{L^1}^{1/2}, \|w_n - w_0\|_{L^1}^{1/3}]$ such that (2.14) holds (with $t = t_n$), and

$$t_n H_{N-1}\{x \in B' \,|\, |w_n(x) - w_0(x)| = t_n\} \leq \frac{\text{Const.}}{\ln \|w_n - w_0\|_{L^1}^{-1/6}}.$$

According to (2.14), choose $0 < s_n < t_n$ such that

$$\int_{\{s_n < |w_n(x) - w_0(x)| \leq t_n\}} (1 + |\nabla w_n(x) - \nabla u(x_0)|)\, dx = O(1/n),$$

$$\begin{aligned} &\frac{1}{t_n - s_n} \int_{B' \cap \{s_n < |w_n(x) - w_0(x)| \leq t_n\}} |w_n(x) - w_0(x)| \, \|\nabla |w_n(x) - w_0(x)|\, \| \, dx \\ &\quad \leq t_n H_{N-1}\{x \in \Omega \,|\, |w_n(x) - w_0(x)| = t_n\} + O(1/n), \end{aligned}$$

and set

$$\tilde{w}_n(x) := w^n_{s_n, t_n}(x).$$

By (2.9)

$$\|\tilde{w}_n - w_0\|_\infty \leq t_n \to 0,$$

and by (2.6), (2.11)-(2.14) we conclude that

$$\begin{aligned} \mu_a(x_0) &\geq \lim_n \frac{1}{\text{meas}\,(B)} \int_{B'} f(x_0 + r_n x, u(x_0) + r_n w_n(x), \nabla w_n(x))\, dx \\ &\geq \liminf_n \frac{1}{\text{meas}\,(B)} \int_{B' \cap \{|w_n(x) - w_0(x)| \leq s\}} h_n(x, w_n(x), \nabla w_n(x))\, dx \\ &\geq \liminf_n \frac{1}{\text{meas}\,(B)} \left\{ \int_{B'} h_n(x, \tilde{w}_n(x), \nabla \tilde{w}_n(x))\, dx \right. \\ &\quad \left. - O(1/n) - \frac{C}{\ln \|w_n - w_0\|_{L^1}^{-1/6}} - C \,\text{meas}\,\{x \in B \,|\, |w_n(x) - w_0(x)| > t_n\} \right\} \\ &= \liminf_n \frac{1}{\text{meas}\,(B)} \int_{B'} h_n(x, \tilde{w}_n(x), \nabla \tilde{w}_n(x))\, dx, \end{aligned}$$

since $t_n \geq \|w_n - w_0\|_{L^1}^{1/2}$, and thus

$$\text{meas}\,\{x \in B \,|\, |w_n(x) - w_0(x)| > t_n\} \leq \frac{1}{t_n} \|w_n - w_0\|_L^1 \leq \|w_n - w_0\|_{L^1}^{1/2} \to 0.$$

Finally, the bound on $\|\nabla \tilde{w}_n\|_{L^1(B')}$ follows from (2.10).

Step 4 (proof of Claim (2.5)). We want to show that

$$\mu_a(x_0) \geq f(x_0, u(x_0), \nabla u(x_0)) \quad \text{for a.e. } x_0 \in \Omega.$$

Let x_0 be a Lebesgue point for u, ∇u and such that (2.2) and (2.4) hold. If $g(x_0, u(x_0)) = 0$, then (2.5) is satisfied trivially as f is a nonnegative function. If $g(x_0, u(x_0)) > 0$ consider a subdomain $B' \Subset B$ and let $\varepsilon > 0$. By (2.8) and (H4) we have

$$\begin{aligned} \mu_a(x_0) &\geq \lim_n \frac{1}{\text{meas}\,(B)} \int_{B'} f(x_0 + r_n x, u(x_0) + r_n \tilde{w}_n(x), \nabla \tilde{w}_n(x))\, dx \\ &\geq \lim_n \frac{1}{\text{meas}\,(B)} \left\{ \int_{B'} f(x_0, u(x_0), \nabla \tilde{w}_n(x))\, dx - \varepsilon \int_{B'} (1 + \|\nabla \tilde{w}_n(x)\|)\, dx \right\}. \end{aligned}$$

By Proposition 2.5 and taking into account that $\{\nabla \tilde{w}_n\}$ is a sequence bounded in L^1, we deduce that

$$\mu_a(x_0) \geqq \frac{1}{\text{meas}\,(B)} \int_{B'} f(x_0, u(x_0), \nabla u(x_0))\, dx - \varepsilon C.$$

Letting $\varepsilon \to 0$, we conclude (2.5) given the arbitrariness of B'. □

3. Proofs of auxiliary results. In this section we prove Propositions 2.4 and 2.5.

PROPOSITION 2.4. (i) *If Theorem* 2.1 *holds true for* Ω *being a ball, it holds true for all open, bounded sets* Ω.

(ii) *Let* Ω *be a ball. If* (H1) *and* (H3) *hold and if* $u_n, u \in W^{1,1}(\Omega; \mathbb{R}^p)$ *are such that* $u_n \to u$ *in* $L^1(\Omega; \mathbb{R}^p)$, *then there exists* $\tilde{u}_n \in C_0^\infty(\mathbb{R}^N; \mathbb{R}^p)$ *such that* $\|\tilde{u}_n - u\|_{L^1_{(\Omega)}} \to 0$ *and*

$$\liminf_{n\to+\infty} \int_\Omega f(x, \tilde{u}_n(x), \nabla \tilde{u}_n(x))\, dx = \liminf_{n\to+\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx.$$

Proof. (i) As in Acerbi and Fusco [AF], we show that it suffices to prove Theorem 2.1 in the case where Ω is a ball. Indeed, if the result was true whenever the domain is a ball, for an arbitrary open set Ω and using Vitali's Covering Theorem, we can write

$$\Omega = \cup\, (a_i + \varepsilon_i B(0, 1)) \cup E,$$

where meas $(E) = 0$ and $\{a_i + \varepsilon_i B(0, 1)\}$ is a family of mutually disjoint balls. Fixing a positive integer k we have

$$\begin{aligned}\liminf_n \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx &\geqq \sum_{i=1}^k \liminf_n \int_{a_i+\varepsilon_i B(0,1)} f(x, u_n(x), \nabla u_n(x))\, dx \\ &\geqq \sum_{i=1}^k \int_{a_i+\varepsilon_i B(0,1)} f(x, u(x), \nabla u(x))\, dx.\end{aligned}$$

Letting $k \to +\infty$ and using the monotone convergence theorem, we conclude that

$$\int_\Omega f(x, u(x), \nabla u(x))\, dx \leqq \liminf \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx.$$

(ii) As in Acerbi and Fusco [AF], we remark that we can extend $u_n \in W^{1,1}(\Omega; \mathbb{R}^p)$ to $u_n^* \in W^{1,1}(\mathbb{R}^N; \mathbb{R}^p)$. Moreover, as $C_0^\infty(\mathbb{R}^N; \mathbb{R}^p)$ is dense in $W^{1,1}(\mathbb{R}^N; \mathbb{R}^p)$, there exist sequences $v_{n,k} \in C_0^\infty(\mathbb{R}^N; \mathbb{R}^p)$ such that

$$v_{n,k} \to u_n^* \quad \text{in } W^{1,1}(\mathbb{R}^N; \mathbb{R}^p) \tag{3.1}$$

as $k \to +\infty$. Moreover, we may assume that $v_{n,k}$ and $\nabla v_{n,k}$ converge to u_n and ∇u_n, respectively, almost everywhere. We claim that

$$\lim_k \int_\Omega f(x, v_{n,k}(x), \nabla v_{n,k}(x))\, dx = \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx. \tag{3.2}$$

Indeed, by (H3),

$$0 \leqq f(x, u, A) \leqq C(1 + \|A\|),$$

and thus by applying Fatou's lemma to $x \to f(x, v_{n,k}(x), \nabla v_{n,k}(x))$ and $C(1 + \|\nabla v_{n,k}(x)\|) - f(x, v_{n,k}(x), \nabla v_{n,k}(x))$ and by observing that

$$\int_\Omega (1 + \|\nabla v_{n,k}(x)\|)\, dx \to \int_\Omega (1 + \|\nabla u_n(x)\|)\, dx,$$

we have (3.2). Finally, using (3.1) and (3.2), for all n choose k_n such that

$$\|v_{n,k_n}-u_n\|_{L^1}\leqq\frac{1}{n}$$

and

$$\left|\int_\Omega f(x, v_{n,k_n}(x), \nabla v_{n,k_n}(x))\,dx-\int_\Omega f(x, u_n(x), \nabla u_n(x))\,dx\right|\leqq\frac{1}{n}.$$

It is clear that, setting

$$\tilde{u}_n := v_{n,k_n},$$

we have

$$\|\tilde{u}_n-u\|_{L^1_{(\Omega)}}\to 0$$

and

$$\lim_n\int_\Omega f(x, \tilde{u}_n(x), \nabla\tilde{u}_n(x))\,dx=\lim_n\int_\Omega f(x, u_n(x), \nabla u_n(x))\,dx. \qquad \square$$

We next prove Theorem 2.1 in the special case where $f=f(A)$ and u is an affine function. The proof presented here was obtained in Fonseca [Fo] (see Theorem 4.6 and Remark 4.16), and we are now aware of the fact that Marcellini's [Ma] proof for the case of weak convergence in $W^{1,m}$, $m>1$ is essentially the same. Yet another proof has been given by Kinderlehrer [K] who uses a subdivision of Ω in small domains in connection with the Vitali covering argument.

PROPOSITION 2.5. *Let $f: M^{pxN}\to\mathbb{R}$ be a function satisfying* (H1), (H2), *and*

$$0\leqq f(A)\leqq C(1+\|A\|)$$

for some $C>0$. If $A_0\in M^{pxN}$ and if $u_n\in W^{1,1}(\Omega;\mathbb{R}^p)$ are such that $u_n\to 0$ in $L^1(\Omega;\mathbb{R}^p)$ and $\{\|\nabla u_n\|_{L^1}\}$ is bounded, then

$$\text{meas}\,(\Omega)f(A_0)\leqq\liminf\int_\Omega f(A_0+\nabla u_n(x))\,dx.$$

Proof. The proof is taken from [Fo]. Related ideas appear in [DG] and [Ma]. We may assume without loss of generality that

$$\liminf\int_\Omega f(A_0+\nabla u_n(x))\,dx=\lim\int_\Omega f(A_0+\nabla u_n(x))\,dx)<+\infty.$$

Due to the growth condition, $\{\|\nabla u_n\|\}$ is bounded in L^1, and so there exists a subsequence and a finite measure μ in Ω such that

$$\|\nabla u_n\|\to\mu \quad \text{weakly*},$$

i.e., for every $\varphi\in C_0(\Omega)$

$$\int_\Omega \varphi(x)\|\nabla u_n(x)\|\,dx\to\int_\Omega\varphi(x)\,d\mu(x). \tag{3.3}$$

Consider an increasing sequence of subdomains Ω_k such that $\bar{\Omega}_k\Subset\Omega$ and $\Omega=\bigcup\Omega_k$. Let φ^k be a smooth cutoff function such that $0\leqq\varphi^k\leqq 1$, $\varphi^k=1$ in Ω_k, $\varphi^k=0$ in $\Omega\backslash\bar{\Omega}_{k+1}$. Setting

$$u_n^k:=\varphi^k u_n\in W_0^{1,1}(\Omega;\mathbb{R}^p),$$

as f is quasi convex, we have

$$\begin{aligned} f(A_0)\operatorname{meas}(\Omega) &\leqq \int_\Omega f(A_0+\nabla u_n^k(x))\,dx \\ &= \int_{\Omega\setminus\Omega_{k+1}} f(A_0)\,dx + \int_{\Omega_{k+1}\setminus\Omega_k} f(A_0+\nabla u_n^k(x))\,dx \\ &\quad + \int_{\Omega_k} f(A_0+\nabla u_n(x))\,dx, \end{aligned}$$

which implies that

$$f(A_0)\operatorname{meas}(\Omega_{k+1}) \leqq \int_{\Omega_{k+1}\setminus\Omega_k} f(A_0+\nabla u_n^k(x))\,dx + \int_{\Omega_k} f(A_0+\nabla u_n(x))\,dx.$$

As f is nonnegative, we deduce that

$$(3.4)\quad \int_\Omega f(A_0+\nabla u_n(x))\,dx - f(A_0)\operatorname{meas}(\Omega_{k+1}) \geqq -\int_{\Omega_{k+1}\setminus\Omega_k} f(A_0+\nabla u_n^k(x))\,dx.$$

On the other hand,

$$\begin{aligned} \int_{\Omega_{k+1}\setminus\Omega_k} f(A_0+\nabla u_n^k(x))\,dx &\leqq C\int_{\Omega_{k+1}\setminus\Omega_k} (1+\|A_0+\nabla u_n^k(x)\|)\,dx \\ &\leqq C\operatorname{meas}(\Omega_{k+1}\setminus\Omega_k) + C\int_{\Omega_{k+1}\setminus\Omega_k} \|\nabla u_n(x)\|\,dx \\ &\quad + C\int_{\Omega_{k+1}\setminus\Omega_k} |u_n(x)|\,\|\nabla\varphi^k(x)\|\,dx \\ &\leqq C\operatorname{meas}(\Omega_{k+1}\setminus\Omega_k) \\ &\quad + C\int_\Omega (\varphi_{k+1}(x)-\varphi_{k-1}(x))\|\nabla u_n(x)\|\,dx \\ &\quad + C\int_{\Omega_{k+1}\setminus\Omega_k} |u_n(x)|\,\|\nabla\varphi^k(x)\|\,dx. \end{aligned}$$

As $u_n\to 0$ in $L^1(\Omega)$, by (3.3) and (3.4) we obtain

$$\begin{aligned} &\lim_n \int_\Omega f(A_0+\nabla u_n(x))\,dx - f(A_0)\operatorname{meas}(\Omega_{k+1}) \\ &\quad \geqq -C\operatorname{meas}(\Omega_{k+1}\setminus\Omega_k) - C\int_\Omega (\varphi_{k+1}(x)-\varphi_{k-1}(x))\,d\mu(x). \end{aligned}$$

Finally, summing the above inequality for $k=2,\cdots,i$, we have

$$\begin{aligned} &(i-1)\lim_n \int_\Omega f(A_0+\nabla u_n(x))\,dx - f(A_0)\sum_{k=2}^{i}\operatorname{meas}(\Omega_{k+1}) \\ &\quad \geqq -C\sum_{k=2}^{i}\left\{\operatorname{meas}(\Omega_{k+1}\setminus\Omega_k) + \int_\Omega (\varphi_{k+1}(x)-\varphi_{k-1}(x))\,d\mu(x)\right\}. \end{aligned}$$

Dividing by $(i-1)$ we find

$$\lim_n \int_\Omega f(A_0+\nabla u_n(x))\,dx - f(A_0)\frac{1}{i-1}\sum_{k=2}^{i} \text{meas}\,(\Omega_{k+1})$$

$$\geqq -C\frac{1}{i-1}\{\text{meas}\,(\Omega_{i+1})-\text{meas}\,(\Omega_2)$$

$$+\int_\Omega (\varphi_{i+1}(x)+\varphi_i(x)+\varphi_2(x)-\varphi_1(x))\,d\mu(x)\}$$

$$\geqq -C\frac{1}{i-1}\{\text{meas}\,(\Omega)+4\mu(\Omega)\}.$$

Letting $i\to+\infty$, we conclude that

$$\lim_n \int_\Omega f(A_0+\nabla u_n(x))\,dx - f(A_0)\,\text{meas}\,(\Omega)\geqq 0. \qquad \square$$

4. Lower semicontinuity for convex integrands. Suppose that $f:\Omega x\mathbb{R}^p xM^{pxN}\to [0,+\infty)$ satisfies the hypotheses:

(H1) f is continuous;

(H2′) $f(x,u,\cdot)$ is convex;

(H3) there exists a nonnegative, bounded, continuous function $g:\Omega x\mathbb{R}^p\to [0,+\infty)$, c, $C>0$ such that

$$cg(x,u)\|A\|-C\leqq f(x,u,A)\leqq Cg(x,u)(1+\|A\|)$$

for all $(x,u,A)\in\Omega x\mathbb{R}^p xM^{pxN}$;

(H4′) for all $x_0\in\Omega x\mathbb{R}^p$ and for all $\varepsilon>0$ there exists $\delta>0$ such that $|x-x_0|<\delta$ implies that

$$|f(x_0,u,A)-f(x,u,A)|\leqq\varepsilon(1+\|A\|).$$

We obtain the following corollary of Theorem 2.1.

COROLLARY 4.1. *If the assumptions* (H1), (H2′), (H3), *and* (H4′) *hold and if* $u_n, u\in W^{1,1}(\Omega;\mathbb{R}^p)$ *are such that* $u_n\to u$ *in* $L^1(\Omega;\mathbb{R}^p)$, *then*

$$\int_\Omega f(x,u(x),\nabla u(x))\,dx\leqq\liminf\int_\Omega f(x,u_n(x),\nabla u_n(x))\,dx.$$

Clearly, in order to apply Theorem 2.1 it suffices to prove that for convex integrands with linear growth (H4′) reduces to (H4).

PROPOSITION 4.2. *If f satisfies* (H1), (H2′), *and* (H3), *then for all* $(x_0,u_0)\in\Omega x\mathbb{R}^p$ *and for all* $\varepsilon>0$ *there exists* $\delta>0$ *such that*

$$|u-u_0|<\delta \textit{ implies that } f(x_0,u,A)-f(x_0,u_0,A)\geqq-\varepsilon(1+\|A\|).$$

We introduce the recession function f^∞ given by

$$f^\infty(x,u,A):=\sup_{t>0}\frac{f(x,u,tA)-f(x,u,0)}{t}.$$

Note that, for fixed $(x,u,A)\in\Omega x\mathbb{R}^p xM^{pxN}$ and g given by $g(t):=f(x,u,tA)-f(x,u,0)$, g is a convex function with $g(0)=0$, and so

(4.1a) $\qquad t\to g(t)/t$ is increasing;

therefore,

$$\begin{aligned} f^\infty(x, u, A) &= \sup_{t>0} g(t)/t \\ &= \lim \frac{f(x, u, tA)}{t} \quad \text{as } t \to +\infty. \end{aligned} \tag{4.1b}$$

If (H1) and (H3) hold and if $f(x, u, \cdot)$ is convex, then $f^\infty(x, u, \cdot)$ is convex (and hence continuous), homogeneous of degree one, and (see, e.g., Fonseca and Rybka (FR, Lemma 2.3])

$$0 \leqq f^\infty(x, u, A) \leqq Cg(x, u)\|A\|$$

for all $(x, u, A) \in \Omega x \mathbb{R}^p x M^{pxN}$.

The proof of this result is based on the following auxiliary lemmas, where for notational convenience we omit the dependence of f on the variable x.

LEMMA 4.3. *If* (H2′) *and* (H3) *hold, then for all* $u \in \mathbb{R}^p$

$$\lim_{r\to+\infty} \sup_{\|A\|=1} \left| \frac{f(u, rA)}{r} - f^\infty(u, A) \right| = 0.$$

Proof. Fix $u \in \mathbb{R}^p$. By (H2′) and (H3) the functions

$$a \to \frac{f(u, rA)}{r}$$

are Lipschitz continuous uniformly with respect to r. By $(4.1)_2$ these functions converge to $f^\infty(u, \cdot)$ pointwise. By the Ascoli-Arzela theorem, the convergence is uniform on compact sets, and so

$$\lim_{r\to+\infty} \sup_{\|A\|=1} \left| \frac{f(u, rA)}{r} - f^\infty(u, A) \right| = 0. \qquad \square$$

LEMMA 4.4. *If* (H1), (H2′), *and* (H3) *hold, for all* $u_0 \in \mathbb{R}^p$ *and for all* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that*

$$|u - u_0| < \delta \textit{ implies that } f^\infty(u, A) - f^\infty(u_0, A) \geqq -\varepsilon$$

for all matrices $A \in M^{pxN}$ *such that* $\|A\| = 1$.

Proof. Step 1. Assume that $f(u, 0) = 0$ and fix $u_0 \in \mathbb{R}^p$ and $\varepsilon > 0$. By (4.1) and by Lemma 4.3 we may choose $r_0 > 2$ such that

$$0 \leqq f^\infty(u_0, A) - \frac{f(u_0, r_0A)}{r_0} < \frac{\varepsilon}{2}$$

for every A with $\|A\| = 1$. On the other hand, as f is continuous there exists $\delta > 0$ (depending only on ε and r_0) such that

$$|u - u_0| < \delta \text{ implies } \sup_{\|A\|=1} |f(u, r_0A) - f(u_0, r_0A)| < \varepsilon.$$

By (4.1) we have

$$\begin{aligned} f^\infty(u, A) &\geqq \frac{f(u, r_0A)}{r_0} \\ &\geqq \frac{f(u_0, r_0A)}{r_0} - \frac{\varepsilon}{r_0} \\ &\geqq f^\infty(u_0, A) - \frac{\varepsilon}{2} - \frac{\varepsilon}{r_0} \\ &\geqq f^\infty(u_0, A) - \varepsilon. \end{aligned}$$

Step 2. As in the proof of the previous lemma, we set $g(u, A):=f(u, A)-f(u, 0)$, and we apply Step 1. The result follows from the fact that $f^\infty(u, A)=g^\infty(u, A)$. □

Proof of Proposition 4.2. *Step* 1. Assume that $f(u, 0)=0$, and fix $u_0\in\mathbb{R}^p$, $\varepsilon>0$. By (4.1), Lemma 4.3, and by continuity choose $r_0>2$, $\delta>0$ such that

$$0\leqq f^\infty(u_0, A)-\frac{f(u_0, r_0A)}{r_0}<\frac{\varepsilon}{2}$$

for every A with $\|A\|=1$, and

$$|u-u_0|<\delta \text{ implies } \sup_{\|A\|\leqq 1}|f(u, r_0A)-f(u_0, r_0A)|<\varepsilon.$$

Thus, if $|u-u_0|<\delta$ and if $\|A\|\leqq r_0$ we have

$$f(u, A)\geqq f(u_0, A)-\varepsilon\geqq f(u_0, A)-\varepsilon(1+\|A\|), \tag{4.2}$$

and by (4.1) if $A=rB$, $\|B\|=1$, $r>r_0$, then

$$\begin{aligned}\frac{f(u, A)}{\|A\|}=\frac{f(u, rB)}{r}&\geqq\frac{f(u, r_0B)}{r_0}\\ &\geqq\frac{f(u_0, r_0B)}{r_0}-\frac{\varepsilon}{r_0}\\ &\geqq f^\infty(u_0, B)-\frac{\varepsilon}{2}-\frac{\varepsilon}{r_0}\\ &\geqq f^\infty(u_0, B)-\varepsilon.\end{aligned}$$

Finally, as $f^\infty(u, \cdot)$ is homogeneous of degree one, by (4.1) we deduce that

$$f(u, A)\geqq f^\infty(u_0, A)-\varepsilon\|A\|\geqq f(u_0, A)-\varepsilon\|A\|,$$

which, together with (4.2), yields the result.

Step 2. In the general case we apply Step 1 to the function $g(u, A):=f(u, A)-f(u, 0)$ in order to find $\delta>0$ such that

$$|f(u, 0)-f(u_0, 0)|<\frac{\varepsilon}{2} \quad\text{and}\quad g(u, A)\geqq g(u_0, A)-\frac{\varepsilon}{2}(1+\|A\|)$$

whenever $|u-u_0|<\delta$. Hence

$$\begin{aligned}f(u, A)&\geqq f(u, 0)+f(u_0, A)-f(u_0, 0)-\frac{\varepsilon}{2}(1+\|A\|)\\ &\geqq f(u_0, A)-\frac{\varepsilon}{2}-\frac{\varepsilon}{2}(1+\|A\|)\\ &\geqq f(u_0, A)-\varepsilon(1+\|A\|).\end{aligned}$$ □

5. Concluding remarks. The integral representation for the relaxation $\mathcal{F}(\cdot)$ in $BV(\Omega)$ of

$$I(u):=\int_\Omega f(x, u(x), \nabla u(x))\,dx$$

was obtained in the scalar-valued case by Dal Maso [DM], who proved that

$$\begin{aligned}(5.1)\quad \mathcal{F}(u) = \int_\Omega f(x, u(x), \nabla u(x))\,dx + \int_{\Sigma(u)} D(x, u^-(x), u^+(x), \nu(x))\,dH_{N-1}(x)\\ + \int_\Omega f^\infty\left(x, u(x), \frac{dC(u)}{d|C(u)|}(x)\right) d|C(u)|(x),\end{aligned}$$

where H_{N-1} denotes the $N-1$-dimensional Hausdorff measure, and the distributional derivative Du of the function $u \in BV(\Omega; \mathbb{R})$ admits the decomposition into mutually singular Radon measures

$$Du = \nabla u L_N \lfloor \Omega + (u^+ - u^-)\nu H_{N-1} \lfloor \Sigma(u) + C(u).$$

Here L_N is the N-dimensional Lebesgue measure; ∇u denotes the absolutely continuous part of Du, i.e., the Radon-Nidodym derivative of Du with respect to $L_N, \Sigma(u)$ is the jump set of Du with normal ν defined for H_{N-1} almost everywhere $x \in \Omega$, and $C(u)$ is the Cantor part of the derivative (for details we refer the reader to Evans and Gariepy [EG], Federer [Fe], Ziemer [Zi]). In (5.1) f^∞ represents the *recession function* (see § 2), and $D(x, a, b, \nu)$ is given by

$$D(x, a, b, \nu) = \int_a^b f^\infty(x, s, \nu)\,ds.$$

In the isotropic vector-valued case, i.e., if $u : \Omega \to \mathbb{R}^p$ and $h = \|\cdot\|$, Baldo [B] and Fonseca and Tartar [FT1] obtained once again that the Γ-limit $J_0(\cdot)$ of

$$J_\varepsilon(u) := \int_\Omega \frac{1}{\varepsilon} W(u(x))\,dx + \varepsilon \int_\Omega h^2(\nabla u(x))\,dx$$

coincides with the relaxation

$$\mathcal{F}(u) := \inf_{\{u_n\}} \left\{ \liminf_{n \to +\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x))\,dx \,\middle|\, u_n \in W^{1,1}(\Omega; \mathbb{R}^p),\ u_n \to u \text{ in } L^1 \right\},$$

where f is given by (1.3). This result confirms Gurtin's [G1], [G2] conjecture that the "preferred" solution has minimal surface energy (see also [Mo]).

In the anisotropic, vector-valued case and with u subject to the constraint curl $u = 0$, recent work by Kohn and Müller [KM] seems to indicate that the Modica and Mortola inequality

$$J_\varepsilon(u) \geqq \int_\Omega f(x, u(x), \nabla u(x))\,dx$$

with f given by (1.3) is no longer optimal. However, it is clear that

$$u \to \int_\Omega f(x, u(x), \nabla u(x))\,dx$$

still provides a lower bound for the rescaled energies $J_\varepsilon(\cdot)$. In particular, the Γ-limit must be bigger than or equal to $\mathcal{F}(u)$. The issue thus arises to find an integral representation for $\mathcal{F}(u)$ in the vector-valued case.

Fonseca and Rybka [FR] proved that, when $f(x, u, \cdot)$ is convex and if u takes only the values a and b across a plane with normal ν, then

$$\mathcal{F}(u)=\int_\Omega f(x, u(x), 0)\, dx+\int_{\Sigma(u)} K(x, a, b, \nu)\, dH_{N-1}(x),$$

where

$$K(x, a, b, \nu):=\inf\left\{\int_{Q_\nu} f^\infty(x, \xi(y), \nabla\xi(y))\, dy \,\middle|\, \xi\in\mathcal{A}\right\}$$

and

$$\mathcal{A}=\{\xi\in W^{1,1}(Q_\nu;\mathbb{R}^p)|\xi(y)=b \text{ if } y\cdot\nu=\tfrac{1}{2}, \xi(y)=a \text{ if } y.\nu=-\tfrac{1}{2}, \text{ and } \xi \text{ is periodic in the remaining } \nu_1, \cdots, \nu_{N-1} \text{ directions with period } 1\},$$

where $\{\nu_1, \cdots, \nu_{N-1}, \nu=\nu_N\}$ forms an orthonormal basis of $\mathbb{R}^N$ and Q_ν is the cube $\{y\in\mathbb{R}^N||y\cdot\nu_i|<\frac{1}{2}, i=1, \cdots, N\}$. The characterization of the surface energy density K was inspired by the work of Fonseca and Tartar [FT2].

Independently, Ambrosio and Pallara [AP] showed that $\mathcal{F}(\cdot)$ admits an integral representation with the same structure as in (5.1), and this result together with the work of Fonseca and Rybka [FR] provides a complete characterization of $\mathcal{F}(u)$, namely

$$\begin{aligned}\mathcal{F}(u)&=\int_\Omega f(x, u(x), \nabla u(x))\, dx+\int_{\Sigma(u)} K(x, u^-(x), u^+(x), \nu(x))\, dH_{N-1}(x)\\ &\quad+\int_\Omega f^\infty\left(x, u(x), \frac{dC(u)}{d|C(u)|}(x)\right) d|C(u)|(x).\end{aligned}\tag{5.2}$$

To identify the first and the third term on the right-hand side of (5.2) [AP] makes use of the lower semicontinuity results of Aviles and Giga [AG], whose proofs rely on sophisticated tools from geometric measure theory. Also, f has to satisfy linear growth condition from below, i.e.,

$$c\|A\|-C\leqq f(x, u, A)\leqq C(1+\|A\|)\tag{5.3}$$

for some $c, C>0$, preventing a situation as in (1.3). In addition, we remark that the convexity hypothesis on $f(x, u, \cdot)$ may be too restrictive. Indeed, as shown by Acerbi and Fusco [AF], Dacorogna [D], and Morrey [Mr] the $W^{1,1}$-weak lower semicontinuous envelope of the functional (1.2) is the integral of the quasi convexification of the energy density $f(x, u, \cdot)$, and so we expect quasi-convexity as a natural constitutive assumption rather than convexity. This concern is genuine as there are examples of quasi-convex functions with linear growth that are not convex (see Sverák [S] and Zhang [Zh]).

In this work we consider quasi-convex integrands, and we relax (5.3) to include degenerate lower bounds. Under these conditions we provide an analytical proof of the lower semicontinuity of (1.2) in L^1, thus obtaining the first term in the relaxation $\mathcal{F}(u)$. Our method seems to be appropriate to proving the lower semicontinuity of the third term in (5.2) corresponding to the Cantor part of the measure Du, and it might be conjectured that the representation of $\mathcal{F}(u)$ given by (5.2) is still valid for quasi-convex integrands with possibly degenerate lower bounds.

REFERENCES

[AF] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125-145.

[At] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, Boston, MA, 1984.

[AG] P. Aviles and Y. Giga, *Variational integrals on mappings of bounded variation and their lower semicontinuity*, Arch. Rat. Mech. Anal. 115 (1991), pp. 201-255.

[AP] L. Ambrosio and D. Pallara, *Integral representation of relaxed functionals on* BV($\mathbb{R}^n$, $\mathbb{R}^k$) *and polyhedral approximation*, to appear.

[B] S. Baldo, *Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids*, Ann. Inst. H. Poincaré, to appear.

[BM] J. M. Ball and F. Murat, *$W^{1,p}$ quasiconvexity and variational problems for multiple integrals.* J. Funct. Anal., 58 (1984), pp. 225-253.

[BDGM] L. Boccardo, J. I. Diaz, D. Giachetti, and F. Murat, *Existence and regularity of renormalized solutions for some elliptic problems involving derivatives of nonlinear terms.* University of Paris VI, Lab. Anal. Num., 1991, preprint, J. Differential Equations, to appear.

[CZ] A. P. Calderon and A. Zygmund, *Local properties of solutions of elliptic partial differential equations*, Studia Math., 20 (1961), pp. 171-225.

[D] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York, 1989.

[DM] G. Dal Maso, *Integral representation on* BV(Ω) *of Γ-limits of variational integrals*, Manuscripta Math., 30 (1980), pp. 387-416.

[DG] E. De Giorgi, *Sulla convergenza di alcune successioni d'integrali del tipo dell'area*, Rend. Mat., 8 (1975), pp. 277-294.

[DD] E. De Giorgi and G. dal Maso, *Γ-convergence and the calculus of variations*, in Mathematical Theories of Optimization, J. P. Cecconi and T. Zolezzi, eds., Springer Lecture Notes 979, 1983, pp. 121-143.

[E] L. C. Evans, *Weak convergence methods for nonlinear partial differential equations*, CMBS Reg. Conf. Series #74, American Mathematical Society, Providence, RI, 1990.

[EG] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions.* CRC Press, Boca Raton, Ann Arbor, London, 1992.

[Fe] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[Fo] I. Fonseca, *Lower semicontinuity of surface energies*, Proc. Roy. Soc. Edin. Sect. A, 120 (1992), pp. 99-115.

[FR] I. Fonseca and P. Rybka, *Relaxation of multiple integrals in the space $BV(\Omega; \mathbb{R}^p)$*, to appear in Proc. Roy. Soc. Edin. Sect. A.

[FT1] I. Fonseca and L. Tartar, *The gradient theory of phase transitions for systems with two potential wells.* Proc. Roy. Soc. Edin. Sect. A, 111 A (1989), pp. 89-102.

[FT2] ———, *The gradient theory of phase transitions in nonlinear elasticity*, in preparation.

[G1] M. E. Gurtin, *Some remarks concerning phase transitions in* $\mathbb{R}^N$, Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA, 1983.

[G2] ———, *Some results and conjectures in the gradient theory of phase transitions*, in Metastability and Incompletely Posed Problems, S. Antman, J. L. Ericksen, D. Kinderlehrer, and I. Müller, eds., Springer-Verlag, New York, pp. 135-146.

[K] D. Kinderlehrer, private communication.

[KM] R. Kohn and S. Müller, *Surface energy and microstructure in coherent phase transitions*, in preparation.

[KS] R. Kohn and P. Sternberg, *Local minimizers and singular perturbations.* Proc. Roy. Soc. Edin., Sect. A, 111A (1989), p. 69.

[Ma] P. Marcellini, *Approximation of quasiconvex functions, and lower semicontinuity of multiple integrals*, Manuscripta Math., 51 (1985), pp. 1-28.

[Mo] L. Modica, *Gradient theory of phase transitions and minimal interface criterion.* Arch. Rational Mech. Anal., 98 (1987), pp. 123-142.

[Mr] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1986.

[OS] N. C. Owen and P. Sternberg, *Nonconvex variational problems with anisotropic perturbations*, Nonlinear Anal. Theory, Methods and Applications, 16 (1991), pp. 705-719.

[R] Y. G. Reshetnyak, *General theorems on semicontinuity and on convergence with a functional*, Sibirskii Math. J., 8 (1967b), pp. 1051-1069.

[S] V. Sverák, *Quasiconvex functions with subquadratic growth*, 1990, manuscript.

[Zh] K. A. Zhang, *Construction of quasiconvex functions with linear growth at infinity*, to appear.

[Zi] W. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1099–1116, September 1992

STABILITY IN OBSTACLE PROBLEMS FOR THE VON KARMAN PLATE*

E. MIERSEMANN[†] AND H. D. MITTELMANN[‡]

Abstract. The buckling beyond the critical load of a plate governed by the von Karman equations is studied. A variational inequality formulation of the problem is derived. The deflection of the plate is subject to an obstacle and the question of the stability of the state with a nontrivial contact set is considered. A stability criterion characterizing the bound through a Rayleigh quotient is proved in the general case. It is specialized to simply connected plates for which also a stress function is introduced. For a square plate numerical continuation along the variational inequality branch yields solutions whose stability is then checked through evaluation of the stability criterion. Stability bounds for both clamped and simply supported plates are obtained.

Key words. plate buckling, von Karman equations, variational inequality, stability, obstacle problem, continuation method

AMS(MOS) subject classifications. primary 73H05, 35J85, 65H99; secondary 73K10, 35P30, 35B35

1. Introduction.

We consider a bounded domain Ω as base domain for a plate. It is assumed that the admissible deflections away from the base domain are bounded by a given obstacle and that the plate is compressed by a force $P\mathbf{K}(s)$, where $\mathbf{K}$ is a vector in the base domain acting on the boundary $\partial\Omega$ of Ω. Here, P denotes a real parameter. For $P > P_0$, where P_0 is the critical load of the free problem, the plate contacts the obstacle in a certain region. In many cases there exists a critical value $P_{\text{crit}} > P_0$ at which the deflection of the plate switches to another state. In the case of the linear theory of the plate critical loads P_{crit} were calculated in [7] for the circular plate, in [8] for the rectangular plate, in both cases for a constant obstacle.

Recently, earlier results of the authors were extended to nonconstant obstacles; see [9]. The determination of the critical load is based on a stability criterion for variational inequalities. The aim of this paper is to extend these results to the nonlinear theory of the plate governed by the von Karman equations. This nonlinear approach yields a good stability criterion from the mechanical point of view. In contrast to the above mentioned papers, variations of the displacement vector in the base domain are also considered.

In the numerical part of this paper, a rectangular plate with a constant obstacle is calculated. The results will be compared with the calculations for the linear plate problem from [8].

*Received by the editors February 12, 1991; accepted for publication (in revised form) February 19, 1992.

[†] Department of Mathematics, University of Leipzig, Leipzig, Germany. The paper was written while this author was visiting Arizona State University.

[‡] Department of Mathematics, Arizona State University, Tempe, Arizona 85287–1804. The research of this author was supported by Air Force Office of Scientific Research grant AFOSR-90-0080.

2. The physical background. Let $\Omega \subset \mathbf{R}^2$ be the bounded, possibly multiply connected middle surface of a thin elastic plate. We use the following notations:

$$w_{,\alpha} \equiv \frac{\partial w}{\partial x_\alpha}, \quad w_{,\alpha\beta} \equiv \frac{\partial^2 w}{\partial x_\alpha \partial x_\beta}, \quad \alpha, \beta = 1, 2.$$

D bending stiffness, $D = \dfrac{Eh^3}{12(1-\nu^2)}$,

h thickness of the plate,

E modulus of elasticity (Young's modulus),

ν Poisson ratio $(0 < \nu < \frac{1}{2})$.

By $w(x)$ and $v(x)$, $x = (x_1, x_2)$, we denote deflections perpendicular to the base domain and by $u = (u_1(x), u_2(x))$, $r = (r_1(x), r_2(x))$ displacement vectors in the base domain. The bending energy e_1 of the plate is given by (see, for example, [4, p. 50])

$$e_1 = \frac{D}{2} \int_\Omega \left[(\Delta w)^2 + 2(1-\nu)(w_{,12}^2 - w_{,11} w_{,22})\right] dx. \tag{2.1}$$

The stretching energy e_2 is, (cf. [4, p. 63])

$$e_2 = \frac{h}{2} \int_\Omega u_{\alpha\beta} \sigma_{\alpha\beta} \, dx. \tag{2.2}$$

We are using the summation convention. In (2.2), $u_{\alpha\beta}$ is the strain tensor and $\sigma_{\alpha\beta}$ the stress tensor.

According to the von Karman theory, it is assumed that

$$\sigma_{11} = \frac{E}{1-\nu^2}(u_{11} + \nu u_{22}), \quad \sigma_{22} = \frac{E}{1-\nu^2}(u_{22} + \nu u_{11}), \quad \sigma_{12} = \frac{E}{1+\nu} u_{12} \tag{2.3}$$

holds.

Instead of (2.3), we can write

$$\sigma_{\alpha\beta} = a_{\alpha\beta\gamma\sigma} u_{\gamma\sigma} \tag{2.4}$$

with

$$a_{1111} = \frac{E}{1-\nu^2}, \quad a_{1112} = 0, \quad a_{1122} = \frac{\nu E}{1-\nu^2},$$
$$a_{2212} = 0, \quad a_{2222} = \frac{E}{1-\nu^2}, \quad a_{1212} = \frac{E}{1+\nu},$$
$$a_{\alpha\beta\gamma\delta} = a_{\beta\alpha\gamma\delta}, \quad a_{\alpha\beta\gamma\sigma} = a_{\gamma\sigma\alpha\beta}.$$

Set

$$\epsilon_{\alpha\beta} = \tfrac{1}{2}(u_{\alpha,\beta} + u_{\beta,\alpha}).$$

In the nonlinear plate theory that yields the von Karman equations, it is assumed that

$$u_{\alpha\beta} = \epsilon_{\alpha\beta} + \tfrac{1}{2} w_{,\alpha} w_{,\beta} \tag{2.5}$$

holds.

The outer work is defined by

$$A = A_1 + A_2 + A_3, \tag{2.6}$$

where

$$A_1 = \sum_{i=0}^{m} \int_{S_i} u_\alpha f_\alpha^i \, dS;$$

here S_i denotes the $m+1$ disjoint curves of the boundary of Ω, and $f^i = (f_1^i, f_2^i)$ are vector fields defined on S_i.

$$A_2 = \int_\Omega u_\alpha f_\alpha \, dx,$$

where $f = (f_1, f_2)$ is a vector field defined on Ω,

$$A_3 = \int_\Omega wg \, dx;$$

here g denotes a scalar function defined on Ω.

Set $U = (u, w)$ with $u = (u_1, u_2)$, then the total energy is given by

$$e(U) = e_1(U) + e_2(U) - A(U).$$

Now, we assume that the admissible deflections w away from the base domain belong to a convex set V with $0 \in V$ and that the admissible displacements u are in a linear space $\mathcal{L}$. Replacing U by $U + \epsilon R$ with $R = (r, v - w)$, $0 < \epsilon < 1$, $r \in \mathcal{L}$, and $v \in V$, then $U + \epsilon R$ is an admissible vector in $\mathbf{R}^3$. We expand $e(U + \epsilon R)$ in powers of ϵ and obtain

$$e(U + \epsilon R) = e(U) + \epsilon e'(U)(R) + \frac{\epsilon^2}{2} e''(U)(R, R) + O(\epsilon^3). \tag{2.7}$$

Here e', e'', etc., denote (formal) Gateaux derivatives of real functionals.

Concerning the historical background of the following stability criterion, see [3, p. 257].

DEFINITION 2.1. The state U is said to be statically stable if

$$e(U + \epsilon R) > e(U)$$

holds for each fixed $R \neq 0$ and for all ϵ with $0 < \epsilon < \epsilon_0(R)$. This definition and expansion (2.7) imply a variational inequality as a necessary condition for a stable state, namely

$$e'(U)(R) \geq 0 \tag{2.8}$$

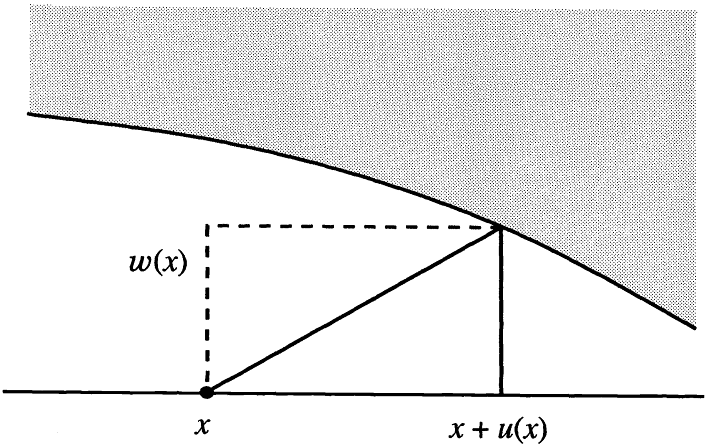


FIG. 1.1. *Plate subject to nonconstant obstacle.*

for all R.

According to the von Karman theory, the displacement vector U is given by

$$U(x) = u_1(x)\mathcal{E}_1 + u_2(x)\mathcal{E}_2 + w(x)\mathcal{E}_3,$$

where u_1, u_2, and w satisfy the system (2.11), (2.12) of §2. Here $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ denote the basis vectors in $\mathbf{R}^3$. If one is interested in the physical problem where the displacements are restricted by an obstacle surface S given by $z = \psi(x)$, then the unilateral condition is characterized by the inequality

$$w(x) \leq \psi(x + u(x));$$

see Fig. 1.1.

This problem coincides with the problem considered in this paper only if $\psi(x) =$ const. For a nonconstant ψ, our problem may be considered merely as an approximation of the physical problem provided $u \cdot \nabla\psi$ is small.

2.1. The first Gateaux derivative. For the convenience of the reader, we derive here the first variation, see, for example, [4, p. 64]. Thus, we recall some more or less known facts. We have

$$e'(U)(R) = e_1'(U)(R) + e_2'(U)(R) - A(R).$$

The definitions of e_1 and A imply

$$e_1'(U)(R) = a(w, v - w),$$

where

$$a(w, v - w) \equiv D\int_\Omega [\Delta w \Delta(v - w) \\ + (1 - \nu)\{2w_{,12}(v - w)_{,12} - w_{,11}(v - w)_{,22} - w_{,22}(v - w)_{,11}\}]\, dx$$

and

$$A(R) = A_1(R) + A_2(R) + A_3(R)$$
$$\equiv \sum_{i=0}^{m} \int_{S_i} r_\alpha f_\alpha^i \, dS + \int_\Omega r_\alpha f_\alpha \, dx + \int_\Omega (v - w) g \, dx.$$

From (2.2) and (2.4) we obtain

$$e_2(U) = h \int_\Omega \psi(u_{\ell j}) \, dx, \tag{2.9}$$

where

$$\psi(u_{\ell j}) \equiv \tfrac{1}{2} u_{\alpha\beta} a_{\alpha\beta\gamma\sigma} u_{\gamma\sigma}.$$

The $u_{\alpha\beta}$ are defined through (2.5). Replacing U in (2.9) by $U + \epsilon R$, setting

$$u_{\alpha\beta}(\epsilon) \equiv \tfrac{1}{2}(u(\epsilon)_{\alpha,\beta} + u(\epsilon)_{\beta,\alpha}) + \tfrac{1}{2} w(\epsilon)_{,\alpha} w(\epsilon)_{,\beta} \tag{2.10}$$

with

$$u(\epsilon) = u + \epsilon r, \qquad w(\epsilon) = w + \epsilon(v - w),$$

we obtain

$$e_2'(U)(R) = h \int_\Omega \sigma_{\alpha\beta} u'_{\alpha\beta}(0) \, dx.$$

Here we have used the relation $\partial\psi/\partial u_{\ell j} = \sigma_{\ell j}$.

Using

$$u'_{\alpha\beta}(0) = \tfrac{1}{2}(r_{\alpha,\beta} + r_{\beta,\alpha}) + \tfrac{1}{2}\{w_{,\alpha}(v - w)_{,\beta} + w_{,\beta}(v - w)_{,\alpha}\}$$

and $\sigma_{\alpha\beta} = \sigma_{\beta\alpha}$, we finally arrive at

$$e_2'(U)(R) = h \int_\Omega \sigma_{\alpha\beta}\{r_{\alpha,\beta} + w_{,\alpha}(v - w)_{,\beta}\} \, dx.$$

Since w belongs to the convex set V and r to a linear space, the variational inequality splits into a coupled system of an inequality and an equation:

$$w \in V : \quad a(w, v - w) + h \int_\Omega \sigma_{\alpha\beta} w_{,\alpha}(v - w)_{,\beta} \, dx - \int_\Omega g(v - w) \, dx \geq 0 \tag{2.11}$$

for all $v \in V$;

$$h \int_\Omega \sigma_{\alpha\beta} r_{\alpha,\beta} \, dx = \sum_{i=0}^{m} \int_{S_i} f_\alpha^i r_\alpha \, dS + \int_\Omega f_\alpha r_\alpha \, dx \tag{2.12}$$

for all $r \in \mathcal{L}$. The symmetric stress tensor $\sigma_{\alpha\beta}$ and its dependence on w is given through (2.3) and (2.5).

For a given $w \in V$, (2.12) defines $\sigma_{\alpha\beta} = \sigma_{\alpha\beta}(w)$. Inserting these $\sigma_{\alpha\beta}$ into (2.11), we see that (2.11) is a nonlinear variational inequality in w and that the displacement vector u does not occur explicitly in (2.11).

For a remark concerning the existence of solutions to (2.11) or (2.12), see the next section.

Remark 2.1. If $\mathcal{L}$ is a convex set as well, with $0 \in \mathcal{L}$, then (2.12) has to be replaced by the associated variational inequality

$$h \int_\Omega \sigma_{\alpha\beta}(s_\alpha - r_\alpha)_{,\beta}\, dx \geq \sum_{j=0}^{m} \int_{S_i} f_\alpha^i (s_\alpha - r_\alpha)\, dS + \int_\Omega f_\alpha (s_\alpha - r_\alpha)\, dx \quad \text{for all } s \in \mathcal{L}.$$

2.2. The second Gateaux derivative. The second derivative of e_1 is given by (see the expansion (2.7))

$$e_1''(U)(R,R) = a(v,v),$$

where $R = (r, v)$; here we replace $v - w$ in (2.10) by v.

Since

$$e_2''(U)(R,R) = h\frac{d^2}{d\epsilon^2}\int_\Omega \psi(u_{\ell j}(\epsilon))\, dx$$

at $\epsilon = 0$, it follows that

$$\begin{aligned} h\frac{d^2}{d\epsilon^2}&\int_\Omega \psi(u_{\ell j}(\epsilon))\, dx \\ &= h\frac{d}{d\epsilon}\int_\Omega \frac{\partial\psi}{\partial u_{\ell j}} u'_{\ell j}(\epsilon)\, dx \\ &= h\int_\Omega \left\{ \frac{\partial\psi}{\partial u_{\ell j}} u''_{\ell j}(\epsilon) + \frac{\partial^2\psi}{\partial u_{\ell j}\partial u_{km}} u'_{\ell j}(\epsilon) u'_{km}(\epsilon) \right\} dx \\ &= h\int_\Omega \sigma_{\alpha\beta} v_{,\alpha} v_{,\beta}\, dx \\ &\quad + h\int_\Omega a_{\alpha\beta\gamma\delta}\left\{\epsilon_{\alpha\beta} + \frac{1}{2}[w_{,\alpha}v_{,\beta} + w_{,\beta}v_{,\alpha}]\right\}\left\{\epsilon_{\gamma\delta} + \frac{1}{2}[w_{,\gamma}v_{,\delta} + w_{,\delta}v_{,\gamma}]\right\} dx \end{aligned}$$

at $\epsilon = 0$, where $\epsilon_{\alpha\beta} = \frac{1}{2}(r_{\alpha,\beta} + r_{\beta,\alpha})$.

3. The stability criterion. Now we fix the function space under consideration. Let V be a closed convex subset of a Sobolev space H with $0 \in V$. In this paper we assume that $H = H_0^2(\Omega)$ for the clamped plate and $H = H_0^1(\Omega) \cap H^2(\Omega)$ for the simply supported plate. Then we set

$$V = \{v \in H; v(x) \leq \psi(x) \text{ on } \Omega\}.$$

Here $\psi \in C^4(\overline{\Omega})$ is given such that $\psi(x) > 0$ holds on $\overline{\Omega}$.

Let $\mathcal{L}$ be a closed linear subspace of $H^1(\Omega) \times H^1(\Omega)$.

Remark 3.1. Set

$$N_3 = \{(a + bx_2, c - bx_1); a, b, c \in \mathbf{R}\}$$

and $N \equiv \mathcal{L} \cap N_3$.

Under the assumption

$$\text{(A)} \qquad A_1(r) + A_2(r) = 0$$

for all $r \in N$, there exists for a given $w \in V$ a solution $u \in \mathcal{L}$ to (2.12) that is uniquely determined up to additive functions belonging to N. Here we have to take into account that (2.12) implies

$$h \int_{\Omega} a_{\alpha\beta\gamma\delta} \left\{ \epsilon_{\alpha\beta} + \frac{1}{2} w_{,\alpha} w_{,\beta} \right\} \eta_{\gamma\delta}\, dx = A_1(r) + A_2(r),$$

where

$$\epsilon_{\alpha\beta} = \tfrac{1}{2}(u_{\alpha,\beta} + u_{\beta,\alpha}), \qquad \eta_{\gamma\delta} = \tfrac{1}{2}(r_{\gamma,\delta} + r_{\delta,\gamma}).$$

From this it follows that

$$\int_{\Omega} a_{\alpha\beta\gamma\delta} w_{,\alpha} w_{,\beta} \eta_{\gamma\delta}\, dx = 0$$

holds for all $r \in N_3$.

The proof of the existence of solutions is based on Korn's inequality (see, for example, [1, Chap. III, §3]) for the three-dimensional case.

For existence results concerning variational inequalities of the type (2.11), see [5].

Define $\mathbf{H} = \mathcal{L} \times H$ and $\mathbf{V} = \mathcal{L} \times V$. Thus, $\mathbf{V}$ is a closed convex subset of $\mathbf{H}$.

Let $U = (u, w) \in \mathbf{V}$ be a solution of the coupled system of the variational inequality (2.11) and of the system of equations (2.12). We recall that $\sigma_{\alpha\beta}(w)$ in (2.11) is defined through (2.12).

In what follows, we are interested in whether U is stable in the sense that

$$e(W) \geq e(U)$$

holds for all $W \in \mathbf{V}$ such that $\|W - U\| < \rho$ for a sufficiently small $\rho > 0$, and equality takes place only for $W = U$. The norm on $\mathbf{H}$ is defined by

$$\|U\|^2 \equiv \|u_1\|^2_{H^1(\Omega)} + \|u_2\|^2_{H^1(\Omega)} + \|w\|^2_{H^2(\Omega)}.$$

For a given w let $\sigma_{\alpha\beta}(w)$ be the solution of (2.12) for $f^i = 0$ and $f = 0$, where $\sigma_{\alpha\beta}$ is defined through (2.3) and (2.5). By $\sigma^{(1)}_{\alpha\beta}$ we denote the solution of (2.12) for $f = 0$, where in (2.12) the $\sigma_{\alpha\beta}$ are defined through (2.3) with $u_{\alpha\beta} \equiv \epsilon_{\alpha\beta}$. That is, we set $w = 0$. Accordingly, we denote by $\sigma^{(2)}_{\alpha\beta}$ the solution with $f^i = 0$ on the right-hand side of (2.12).

Now, we replace f^i in (2.12) by $-\lambda_1 f^i$ and f by $-\lambda_2 f$ and g in (2.11) by $\lambda_3 g$ with real parameters λ_1, λ_2, and λ_3. Using this notation, we can decompose $\sigma_{\alpha\beta}$ of (2.11) as follows:

$$\sigma_{\alpha\beta} = \sigma_{\alpha\beta}(w) - \lambda_1 \sigma^{(1)}_{\alpha\beta} - \lambda_2 \sigma^{(2)}_{\alpha\beta}. \tag{3.1}$$

From the definition of the symmetric stress tensor through (2.12), we obtain ($n^{(i)}$denotes the outer unit normal at S_i)

$$\begin{aligned} \sigma_{\alpha\beta}(w)_{,\beta} &= 0 && \text{in } \Omega,\\ \sigma_{\alpha\beta}(w) n^{(i)}_{\beta} &= 0 && \text{on } S_i,\\ \sigma^{(1)}_{\alpha\beta,\beta} &= 0 && \text{in } \Omega,\\ h\sigma^{(1)}_{\alpha\beta} n^{(i)}_{\beta} &= f^i_{\alpha} && \text{on } S_i,\\ h\sigma^{(2)}_{\alpha\beta,\beta} &= f_{\alpha} && \text{in } \Omega,\\ \sigma^{(2)}_{\alpha\beta} n^{(i)}_{\beta} &= 0 && \text{on } S_i. \end{aligned} \tag{3.2}$$

For the second variation of e_2 we write

$$e_2''(U)(R,R) = e_{21}''(U)(R,R) + e_{22}''(U)(R,R),$$

where we have set

$$\begin{aligned} e_{21}''&(U)(R,R) \\ &= h\int_\Omega \sigma_{\alpha\beta}(w)v_{,\alpha}v_{,\beta}\,dx \\ &\quad + h\int_\Omega a_{\alpha\beta\gamma\delta}\left\{\epsilon_{\alpha\beta} + \frac{1}{2}[w_{,\alpha}v_{,\beta} + w_{,\beta}v_{,\alpha}]\right\}\left\{\epsilon_{\gamma\delta} + \frac{1}{2}[w_{,\gamma}v_{,\delta} + w_{,\delta}v_{,\gamma}]\right\} dx \end{aligned}$$

and

$$e_{22}''(U)(R,R) = -h\lambda_1\int_\Omega \sigma_{\alpha\beta}^{(1)}v_{,\alpha}v_{,\beta}\,dx - h\lambda_2\int_\Omega \sigma_{\alpha\beta}^{(2)}v_{,\alpha}v_{,\beta}\,dx,$$

that is, the right-hand side is independent of U.

For a given $t > 0$ we define

$$\mathbf{V}_t(U) = \{R \in \mathbf{H}; U + tR \in \mathbf{V}\}.$$

Let $R \in \mathbf{V}_t(U)$. Then $e(U+tR)$ is expanded with respect to t:

$$\begin{aligned} e(U+tR) &= e(U) + te'(U)(R) \\ &\quad + \frac{t^2}{2}\{e_1''(U)(R,R) + e_{21}''(U)(R,R) + e_{22}''(U)(R,R)\} \\ &\quad + O(t^3). \end{aligned} \tag{3.3}$$

Let U be a solution of the variational inequality (2.8), or equivalently, of the system (2.11), (2.12), where g in (2.11) is replaced by $\lambda_3 g$ and f^i in (2.12) by $-\lambda_1 f^i$ and f by $-\lambda_2 f$. That is, $w \in V$ is such that

$$\begin{aligned} a(w,v-w) &+ h\int_\Omega \sigma_{\alpha\beta}(w)w_{,\alpha}(v-w)_{,\beta}\,dx \\ &\geq \lambda_1 h\int_\Omega \sigma_{\alpha\beta}^{(1)}w_{,\alpha}(v-w)_{,\beta}\,dx + \lambda_2 h\int_\Omega \sigma_{\alpha\beta}^{(2)}w_{,\alpha}(v-w)_{,\beta}\,dx \\ &\quad + \lambda_3\int_\Omega g(v-w)\,dx \end{aligned}$$

for all $v \in V$, where $\sigma_{\alpha\beta}(w)$, $\sigma_{\alpha\beta}^{(1)}$ and $\sigma_{\alpha\beta}^{(2)}$ are defined through (3.2).

Set $\tilde{\mathbf{H}} = \tilde{\mathcal{L}} \times V$, where $\mathcal{L} = \tilde{\mathcal{L}} \oplus N$ is the orthogonal decomposition with respect to the Hilbert space norm on $H_1 \times H_1$; for notation, see Remark 3.1. Define

$$\tilde{\mathbf{V}} = \mathbf{V} \cap \tilde{\mathbf{H}}$$

and

$$\tilde{\mathbf{V}}_t = \mathbf{V}_t(U) \cap \tilde{\mathbf{H}}.$$

Set

$$e_{21}''(U)(R,R) = e_{211}''(R,R) + e_{212}''(U)(R,R),$$

where we define

$$e''_{211}(R,R) \equiv h \int_\Omega a_{\alpha\beta\gamma\sigma}\epsilon_{\alpha\beta}\epsilon_{\gamma\delta}\,dx.$$

Then, the sum

$$Q(R,R) \equiv e''_1(R,R) + e''_{211}(R,R)$$

is equivalent to the defined Hilbert space norm on $\tilde{\mathbf{H}}$.

We remark that each $W \in \tilde{\mathbf{V}}$, $W \neq U$, may be written as $U + tR$ with $t > 0$, $Q(R,R) = 1$, and $U + tR \in \tilde{\mathbf{V}}$, where $t^2 = Q(W-U, W-U)$ and $R = t^{-1}(W-U)$.

Remark 3.2.

$$Q(U)(R,R) \equiv e''_1(R,R) + e''_{21}(U)(R,R)$$

may also be used as a norm on $\tilde{\mathbf{H}}$, provided $\|U\|$ is not too large. This follows by using the inequality $2ab \leq \epsilon a^2 + \epsilon^{-1}b^2$ for all $\epsilon > 0$.

Define for a given positive constant A

$$\tilde{\mathbf{V}}_{t,A}(U) = \{R \in \tilde{\mathbf{V}}_t(U);\ Q(R,R) \leq 1 \text{ and } e'(U)(R) \leq At\}.$$

Set

$$q(U)(R,R) \equiv -e''_{212}(U)(R,R) - e''_{22}(R,R),$$

and let

$$\Lambda^{-1}_{\tilde{\mathbf{H}}} = \max_{R\in\tilde{\mathbf{H}}\setminus\{0\}} \frac{q(U)(R,R)}{Q(R,R)}.$$

The existence of a maximizer follows from Sobolev embedding theorems.

We make the following hypothesis.

(H) Let $t_n \to 0$, $t_n > 0$, and let $R_n \in \tilde{\mathbf{V}}_{t_n,A}(U)$ be a weakly convergent sequence $R_n \rightharpoonup R$. Then it follows that

$$q(U)(R,R) < 1.$$

THEOREM 3.1. *Suppose that the hypothesis* (H) *is satisfied with a constant A satisfying $2A > \Lambda^{-1}_{\tilde{\mathbf{H}}} - 1$. Then the solution U to the variational inequality $e'(U)(W - U) \geq 0$ for all $W \in \mathbf{V}$ defines a strict local minimum in the sense that there exist positive constants c and ρ such that*

$$e(W) - e(U) \geq c\|W-U\|^2$$

holds for all $W \in \tilde{\mathbf{V}}$ with $\|W - U\| < \rho$.

Proof. The result follows from the expansion

$$(3.4)\quad e(U+tR) = e(U) + te'(U)(R) + \frac{t^2}{2}[Q(R,R) - q(U)(R,R)] + O(t^3), \quad 0 < t \leq t_0.$$

If $R \in \tilde{\mathbf{V}}_t(U)$, $Q(R,R) = 1$ and $e'(U)(R) \geq At$ is satisfied for a constant A not depending on t, then (3.4) implies

$$e(U+tR) - e(U) \geq \frac{t^2}{2}[2A + 1 - \Lambda^{-1}_{\tilde{\mathbf{H}}}] + O(t^3).$$

Now, we consider those $R \in \tilde{\mathbf{V}}_t(U)$ such that $Q(R,R) = 1$ and $e'(U)(R) \leq At$. The expansion (3.4) yields that

$$e(U+tR) - e(U) \geq \frac{t^2}{2}[1 - q(U)(R,R)] + O(t^3)$$

since $e'(U)(R) \geq 0$ holds.

We pose the maximum problem

$$\begin{aligned}\mu_t(U) &\equiv \max_{R \in \tilde{\mathbf{V}}_{t,A}} q(U)(R,R) \\ &= q(U)(R_t, R_t),\end{aligned}$$

where R_t denotes a maximizer. If (H) is satisfied, then the assertion of the theorem follows.

4. A special case. To simplify the matter, we consider a simply connected plate and assume here that $\|U\|$ is small enough such that $Q(U)(R,R)$ defines a norm on $\tilde{\mathbf{H}}$; cf. Remark 3.2. Moreover, we suppose that the solution $U = (u,w)$ of the variational inequality (2.8) satisfies

$$\begin{aligned} w(x) &= \psi(x) \quad \text{on } C = \mathcal{A} \cup \partial\mathcal{A}, \\ w(x) &< \psi(x) \quad \text{on } \Omega \backslash C \end{aligned}$$

for an open set $\mathcal{A}$ with a piecewise smooth $\partial\mathcal{A}$.

We assume that $\lambda_2 = \lambda_3 = 0$ holds. Set $\lambda = \lambda_1/D$,

$$L_\lambda \psi \equiv \Delta^2 \psi - \frac{h}{D}\sigma_{\alpha\beta}(\psi)\psi_{,\alpha\beta} + \lambda h \sigma^{(1)}_{\alpha\beta}\psi_{,\alpha\beta},$$

and

$$A_1 \equiv \frac{\partial}{\partial n}(\Delta\psi - \Delta w)$$

on $\partial\mathcal{A}$; n denotes the outer unit normal on $\partial\mathcal{A}$.

As in [9, Lemma 4.3, Rem. 4.1], we prove the following.

LEMMA 4.1. *Let $L_\lambda\psi = 0$ on $\mathcal{A}$ and $A_1 > 0$ on $\partial\mathcal{A}$. Then the weak limit $R = (r,v)$ of hypothesis* (H) *of* §3 *satisfies $v = \nabla v = 0$ on $\partial\mathcal{A}$.*

Define

$$\mu^{-1} = \max_{R \in \tilde{\mathbf{H}}_0 \backslash \{0\}} \frac{Dh \int_\Omega \sigma^{(1)}_{\alpha\beta} v_{,\alpha} v_{,\beta}\, dx}{Q(U)(R,R)},$$

where

$$\tilde{\mathbf{H}}_0 = \{(r,v) \in \tilde{\mathbf{H}}; v = \nabla v = 0 \quad \text{on } \partial\mathcal{A}\}.$$

As in Theorem 3.1, we can prove the following result.

THEOREM 4.1. *Under the above assumptions, the solution $U = (u,w)$ of the variational inequality* (2.8) *defines a strict local minimum of e if $\lambda < \mu$ is satisfied.*

From the definition of μ it follows that μ is an eigenvalue of the problem: seek $R \in \tilde{\mathbf{H}}_0 \backslash \{0\}$, $R = (r,v)$ such that

$$Q(U)(R,S) = \mu h \int_\Omega \sigma^{(1)}_{\alpha\beta} v_{,\alpha} \phi_{,\beta}\, dx \tag{4.1}$$

for all $S \in \tilde{\mathbf{H}}_0 \backslash \{0\}$, $S = (s, \phi)$, where Q is defined by

$$Q(U)(R,S) \equiv a(v,\phi) + h \int_\Omega \sigma_{\alpha\beta}(w) v_{,\alpha}\phi_{,\beta}\, dx + h \int_\Omega \sigma_{\alpha\beta}(w,v)(s_{\alpha,\beta} + w_{,\alpha}\phi_{,\beta})\, dx$$

with

$$a(v,\phi) \equiv D \int_\Omega [\Delta v \Delta \phi + (1-\nu)\{2v_{,12}\phi_{,12} - v_{,11}\phi_{,22} - v_{,22}\phi_{,11}\}]\, dx.$$

The $\sigma_{\alpha\beta}(w,v)$ are defined through (2.12) with $f_\alpha \equiv 0$, $f_\alpha^{(i)} \equiv 0$, $i = 0, 1, \cdots, m$, and

$$u_{\alpha\beta} = \tfrac{1}{2}(r_{\alpha,\beta} + r_{\beta,\alpha}) + \tfrac{1}{2}(w_{,\alpha}v_{,\beta} + w_{,\beta}v_{,\alpha}).$$

$R = (r, v)$ denotes a solution of the above maximum problem for the Rayleigh quotient. This implies that the $\sigma_{\alpha\beta}$ are bilinear forms satisfying

$$\sigma_{\alpha\beta}(w,v) = \sigma_{\alpha\beta}(v,w) = \sigma_{\beta\alpha}(w,v). \tag{4.2}$$

This follows along the lines of §2.1. The definition of $\sigma_{\alpha\beta}(w)$ and $\sigma_{\alpha\beta}(w,v)$ implies that

$$\sigma_{\alpha\beta}(w + \epsilon v) = \sigma_{\alpha\beta}(w) + \epsilon\sigma_{\alpha\beta}(w,v) + \epsilon^2\sigma_{\alpha\beta}(v)$$

holds.

The equation (4.1) is equivalent to the system

$$\begin{aligned} &a(v,\phi) + h \int_\Omega \sigma_{\alpha\beta}(w) v_{,\alpha}\phi_{,\beta}\, dx + h \int_\Omega \sigma_{\alpha\beta}(w,v) w_{,\alpha}\phi_{,\beta}\, dx \\ &= \mu h \int_\Omega \sigma_{\alpha\beta}^{(1)} v_{,\alpha}\phi_{,\beta}\, dx \end{aligned} \tag{4.3}$$

and

$$\int_\Omega \sigma_{\alpha\beta}(w,v) s_{\alpha,\beta}\, dx = 0 \tag{4.4}$$

for all $s \in \mathcal{L}$, or equivalently,

$$\begin{aligned} \sigma_{\alpha\beta}(w,v)_{,\beta} &= 0 \quad \text{in } \Omega, \\ \sigma_{\alpha\beta}(w,v) n_\beta &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.5}$$

Remark 4.1. In fact, the eigenvalue equation (4.3) is the linearization of the associated equation to the variational inequality (2.11) at w.

4.1. Introduction of the stress function. If the plate is simply connected, then it is well known that a stress function can be introduced and we get the von Karman equations in the nonconstrained case; see, for example, [4, p. 65] or [11, p. 109].

Let $U = (u, w) \in \mathbf{V}$ be a solution of

$$\begin{aligned} &\frac{1}{D} a(w, v - w) + \frac{h}{D} \int_\Omega \sigma_{\alpha\beta}(w) w_{,\alpha}(v - w)_{,\beta}\, dx \\ &\geq \lambda h \int_\Omega \sigma_{\alpha\beta}^{(1)} w_{,\alpha}(v - w)_{,\beta}\, dx \end{aligned} \tag{4.6}$$

for all $v \in V$, where $\sigma^{(1)}_{\alpha\beta}$, $\sigma_{\alpha\beta}(w)$ are defined in §3, and let $R = (r, v) \in \tilde{\mathbf{H}}_0$ be a solution of

$$\begin{aligned}\frac{1}{D}a(v,\phi) + \frac{h}{D}\int_\Omega \sigma_{\alpha\beta}(w)v_{,\alpha}\phi_{,\beta}\,dx + \frac{h}{D}\int_\Omega \sigma_{\alpha\beta}(w,v)w_{,\alpha}\phi_{,\beta}\,dx \\ = \mu h\int_\Omega \sigma^{(1)}_{\alpha\beta}v_{,\alpha}\phi_{,\beta}\,dx\end{aligned} \tag{4.7}$$

for all $(s,\phi) \in \tilde{\mathbf{H}}_0$.

The stress tensor $\sigma_{\alpha\beta}(w,v)$ is defined through (4.4) or (4.5). It follows from this definition that

$$\sigma_{\alpha\beta}(w,w) = 2\sigma_{\alpha\beta}(w) \tag{4.8}$$

holds.

Let $\chi \equiv \chi(w)$, $\chi \in H^2_0(\Omega)$ be the solution of

$$\Delta^2\chi = E(w^2_{,12} - w_{,11}w_{,22}). \tag{4.9}$$

χ is called a stress function. Then we have for $\sigma_{\alpha\beta} \equiv \sigma_{\alpha\beta}(w)$

$$\sigma_{11} = \chi_{,22}, \quad \sigma_{12} = -\chi_{,12}, \quad \sigma_{22} = \chi_{,11}. \tag{4.10}$$

The associated weak equation to (4.9) is given by $\chi \in H^2_0(\Omega)$;

$$\int_\Omega \Delta\chi\Delta\phi\,dx = E\int_\Omega (w_{,1}w_{,22}\phi_{,1} - w_{,1}w_{,2}\phi_{,2})\,dx$$

for all $\phi \in H^2_0(\Omega)$.

The $\sigma_{\alpha\beta}(w,v)$ in (4.7) are defined through (4.10) but with $\chi \in H^2_0(\Omega)$ given by

$$\Delta^2\chi = E(2w_{,12}v_{,12} - w_{,11}v_{,22} - w_{,22}v_{,11}), \tag{4.11}$$

or, equivalently, $\chi \in H^2_0(\Omega)$ such that

$$\int_\Omega \Delta\chi\Delta\phi\,dx = E\int_\Omega \{[w_{,1}v_{,22} + v_{,1}w_{,22}]\phi_{,1} - [w_{,1}v_{,12} + v_{,1}w_{,12}]\phi_{,2}\}\,dx$$

for all $\phi \in H^2_0(\Omega)$.

Let $\mu(\Omega\backslash C)$ be the lowest eigenvalue of (4.7) with v and ϕ defined on $\Omega\backslash C$ and $v = (\partial v/\partial n) = 0$, $\phi = (\partial\phi/\partial n) = 0$ on $\partial\mathcal{A}$ and in the case of a clamped plate on $\partial\Omega$, too. If the plate is simply supported, then the boundary conditions are $v = \phi = 0$ on $\partial\Omega$ and we have to add a free boundary condition on $\partial\Omega$; see, for example, [4, p. 54].

Let $\mu(\mathcal{A})$ be the lowest eigenvalue of (4.7) with $v, \phi \in H^2_0(\mathcal{A})$. Then, $U = (u, w)$ is stable if

$$\lambda < \min\{\mu(\Omega\backslash C), \mu(\mathcal{A})\} \equiv \mu_0$$

holds.

4.2. The rectangular plate. We take $\Omega = [0, a] \times [0, 1]$ and assume that the plate is compressed by the force $f^{(0)} = -\lambda_1 n$, n the outer unit normal. From (3.2) it follows that

$$h\sigma^{(1)}_{\alpha\beta} = \delta_{\alpha\beta},$$

where $\delta_{\alpha\beta}$ denotes Kronecker's symbol. If $v = 0$ on $\partial\Omega$, then (cf., for example, [4, p. 54])

$$a(v, v) = \frac{D}{2}\int_\Omega (\Delta v)^2\, dx - \frac{D}{2}(1 - \nu)\int_{\partial\Omega} \kappa \left(\frac{\partial v}{\partial n}\right)^2 dS.$$

Here κ denotes the curvature that is positive if $\partial\Omega$ is convex with respect to the inner normal at the boundary point under consideration. Thus, in the case of a rectangular plate, we have the variational inequality

$$\begin{aligned}&\int_\Omega \Delta w \Delta(v - w)\, dx + \frac{h}{D}\int_\Omega \sigma_{\alpha\beta}(w) w_{,\alpha}(v - w)_{,\beta}\, dx \\ &\geq \lambda \int_\Omega \nabla w \cdot \nabla(v - w)\, dx\end{aligned} \tag{4.12}$$

for all $v \in V$, where the $\sigma_{\alpha\beta}(w)$ are defined through (4.9) and (4.10).

The eigenvalue equation (4.7) reads now as

$$\Delta^2 v - \frac{h}{D}\sigma_{\alpha\beta}(w) v_{,\alpha\beta} - \frac{h}{D}\sigma_{\alpha\beta}(w, v) w_{,\alpha\beta} = -\mu\Delta v. \tag{4.13}$$

Here the $\sigma_{\alpha\beta}(w, v)$ are defined through (4.11) and (4.10). We have used the relations (4.5) and (3.2).

Further, $\mu(\mathcal{A})$ is the lowest eigenvalue of (4.13) with v defined on $\mathcal{A}$ and $v = (\partial v/\partial n) = 0$ on $\partial\mathcal{A}$, and $\mu(\Omega\backslash C)$ is the lowest eigenvalue of (4.13), where v is defined on $\Omega\backslash C$ and $v = (\partial v/\partial n) = 0$ holds on $\partial\mathcal{A}$, and $v = \Delta v = 0$ on $\partial\Omega$ for the simply supported plate or $v = (\partial v/\partial n) = 0$ on $\partial\Omega$ in the case of the clamped plate.

5. A remark concerning the unconstrained problem.

Here we consider a simply connected plate that is simply supported at the boundary $\partial\Omega$. That is, we have to describe $w = 0$ at $\partial\Omega$ for the deflections w perpendicular to the x-plane.

Let $f^i = 0$, $i = 1, 2, \cdots, m$, $f_\alpha = 0$, and $g = 0$, and set $f^0 = -\lambda_1 n$; n denotes the outer unit normal on $\partial\Omega$. A direct calculation shows that assumption (A) from §3 is satisfied.

In the absence of an obstacle, the system (2.11), (2.12) becomes

$$\Delta^2 w - \frac{h}{D}\sigma_{\alpha\beta}(w) w_{,\alpha\beta} = -\lambda\Delta w \quad \text{in } \Omega, \tag{5.1}$$

$$\Delta^2 \chi = E(w_{,12}^2 - w_{,11}w_{,22}) \quad \text{in } \Omega, \tag{5.2}$$

$$\sigma_{11} = \chi_{,22}, \quad \sigma_{12} = -\chi_{,12}, \quad \sigma_{22} = \chi_{,11}, \tag{5.3}$$

with the boundary conditions

$$\chi = \frac{\partial\chi}{\partial n} = 0 \quad \text{on } \partial\Omega, \tag{5.4}$$

$$w = 0 \quad \text{on } \partial\Omega, \tag{5.5}$$

$$Mw \equiv \frac{\partial^2 w}{\partial n^2} + \nu\kappa \frac{\partial w}{\partial n} = 0 \quad \text{on } \partial\Omega. \tag{5.6}$$

The last one is a free boundary condition; see, for example, [4, p. 54]. In the case of the clamped plate, we have to replace (5.6) by $\partial w/\partial n = 0$ on $\partial\Omega$.

Now, we are interested in the question of whether a solution of (5.1)–(5.6) is stable in the above sense. For this question concerning shells and plates, even when the first eigenvalue is not simple, see [2], [10] and the references therein.

Let λ_0 be the first eigenvalue of the linearized problem to (5.1)–(5.6), that is, of $\Delta^2 w + \lambda \Delta w = 0$ in Ω under the boundary conditions (5.5) and (5.6).

In what follows, we assume that λ_0 is a simple eigenvalue. We recall that, in the case of a simply supported plate, λ_0 is simple, and the associated eigenfunction does not change its sign in Ω provided Ω is convex; see [6].

By a well-known method, we find for the eigenvalue λ of (5.1)–(5.6) the expansion, $|\epsilon| < \epsilon_0$, ϵ_0 sufficiently small,

$$\lambda = \lambda_0 + \epsilon^2 \lambda_2 + O(\epsilon^3), \qquad w = \epsilon w_1 + \epsilon^2 w_2 + O(\epsilon^3), \tag{5.7}$$

where w_1 is an eigenfunction to the linearized problem and λ_2 is defined by

$$\lambda_2 = \frac{\frac{h}{D}\int_\Omega \sigma_{\alpha\beta}(w_1) w_{1,\alpha} w_{1,\beta}\, dx}{\int_\Omega |\nabla w_1|^2\, dx}.$$

Inserting the w from (5.7) into (4.13), we find that

$$v = w_1 + \epsilon^2 v_2 + O(\epsilon^3)$$

and

$$\mu = \lambda_0 + \epsilon^2 \mu_2 + O(\epsilon^3),$$

where μ_2 is given by

$$\mu_2 \int_\Omega |\nabla w_1|^2\, dx = \frac{h}{D}\int_\Omega \sigma_{\alpha\beta}(w_1) w_{1,\alpha} w_{1,\beta}\, dx + \frac{h}{D}\int_\Omega \sigma_{\alpha\beta}(w_1, w_1) w_{1,\alpha} w_{1,\beta}\, dx.$$

Since

$$\int_\Omega \sigma_{\alpha\beta}(w) w_{,\alpha} w_{,\beta}\, dx > 0$$

for $w \not\equiv 0$ (see, for example, [5, Lemma 7.4]) and (4.8) hold, we obtain

$$\mu_2 = \frac{\frac{3h}{D}\int_\Omega \sigma_{\alpha\beta}(w_1) w_{1,\alpha} w_{1,\beta}\, dx}{\int_\Omega |\nabla w_1|^2\, dx}$$

and thus $\lambda_2 < \mu_2$. This means the buckled state is stable, at least for small deflections.

5.1. Continuation from a stable buckled state. Let (w_s, λ_s) be a stable solution of (5.1)–(5.6). Now, we add τg to the right-hand side of (5.1), where τ is a real parameter and g is defined on Ω and sufficiently regular. We are interested in solutions $(w(\tau), \lambda(\tau))$ such that $(w(0), \lambda(0)) = (w_s, \lambda_s)$ holds.

Let $\mu(\tau)$ be the first eigenvalue to (4.13) with $w \equiv w(\tau)$, then, from §4, the stability criterion follows

$$(w(\tau), \lambda(\tau)) \text{ is stable if } \lambda(\tau) < \mu(\tau) \text{ is satisfied.}$$

We say that the stability bound is attained if and only if $\lambda(\tau) = \mu(\tau)$ holds.

6. Numerical results. For the numerical calculation we use the characterization of the eigenvalue μ_0 by the Rayleigh quotient; see §4. Set

$$B(v) = \int_\Omega |\nabla v|^2 \, dx$$

and

$$\begin{aligned} A(r, v) = &\int_\Omega (\Delta v)^2 \, dx + \frac{h}{D} \int_\Omega (\chi_{,22} v_{,1}^2 - 2\chi_{,12} v_{,1} v_{,2} + \chi_{,11} v_{,2}^2) \, dx \\ &+ \frac{h}{D} \int_\Omega \left[\frac{E}{1-\nu^2} (\tilde{\epsilon}_{11}^2 + \tilde{\epsilon}_{22}^2) + \frac{2\nu E}{1-\nu^2} \tilde{\epsilon}_{11} \tilde{\epsilon}_{22} + \frac{4E}{1+\nu} \tilde{\epsilon}_{12}^2 \right] dx, \end{aligned}$$

where

$$\tilde{\epsilon}_{\alpha\beta} = \tfrac{1}{2}(r_{\alpha,\beta} + r_{\beta,\alpha}) + \tfrac{1}{2}(w_{,\alpha} v_{,\beta} + w_{,\beta} v_{,\alpha}).$$

Then, μ_0 is given by

$$\mu_0 = \min_{\substack{(r,v) \in L \\ v \neq 0}} \frac{A(r, v)}{B(v)}, \tag{6.1}$$

where the linear space $\mathcal{L}$ contains all functions from $(H^1 \times H^1) \times H^2$ such that

$$\int_\Omega r_1 \, dx = 0, \quad \int_\Omega r_2 \, dx = 0, \quad \int_\Omega (x_2 r_1 - x_1 r_2 - r_{2,1} + r_{1,2}) \, dx = 0 \tag{6.2}$$

are satisfied for r and

$$\begin{aligned} &v = 0 \quad \text{on } \partial\Omega \quad \left(\text{and } \frac{\partial v}{\partial n} = 0 \text{ in the clamped case} \right), \\ &v = \frac{\partial v}{\partial n} = 0 \quad \text{on } \partial\mathcal{A} \end{aligned}$$

for v.

First, a continuation method has to be used to continue from the first bifurcation point λ_0 along the stable branch of solutions to the von Karman equations (5.1)–(5.5). λ_0 is given as the lowest eigenvalue of the linearized problem. Since the emphasis here is on the determination of the stability bound μ_0, a relatively simple method was used for this continuation, namely the projected relaxation method already used in [9]. Iteratively, the method was applied to (4.12) after first solving (4.9) for χ. Initially, a small multiple of the eigenfunction of the linearized problem was used as a starting guess; subsequently, the solution at a certain λ-value was used as a guess for $\lambda + \delta\lambda$. This corresponds to a zeroth-order predictor step. In general, this should be replaced by a first-order predictor as given in [9]. Also, instead of the relaxation algorithm, a projected Newton method as in §4 of [8] may be preferable. A finite-difference discretization as in [8], [9] was applied to (5.1)–(5.5); see the Appendix.

Second, the variational inequality solutions have to be tested for stability. At each computed point along this part of the branch, the eigenvalue problem (6.1), (6.2) is solved. This problem again was discretized by an analogous finite difference method. Both numerator and denominator in (6.1) are quadratic forms in the values at grid points of the variables v, r_1, and r_2. In order to make sure that the asociated matrices are symmetric, the functionals A, B were discretized, and the orthogonality conditions (6.2) were added through three additional rows and, for symmetry, also columns of the matrix associated with A, while the matrix for B had zero entries in these positions. The inverse iteration method with shift successfully applied in [8] was used to determine μ_0.

In the following we report about some computations for both the simply supported and the clamped square plate subject to a constant obstacle $\psi(x) \equiv d$. A square grid of size $\tilde{h} = 1/(n+1)$ is used, and $\delta\lambda$ is taken as 10(20) in the simply supported (clamped) case. With a bisection method λ is adjusted such that $\lambda = \mu_0(\lambda)$ holds. The physical quantities entering the problem (cf. §2) are chosen as those for a steel plate, $E = 2 \cdot 10^6$ kg/cm^2, $\nu = .25$. The obstacle has the value $d = .05$, and the thickness h of the plate has to satisfy $h < 2d$. A collection of formulae for the discrete problem is given in an appendix.

Computations for $n = 31$ yielded the stability bounds given in Table 6.1. These are the load values λ for which $\lambda = \mu_0(\lambda)$. They show that for the nonlinear theory stability holds for slightly different load values and contact sets compared to the linear theory; cf. [8]. For $w \equiv 0$ in (6.1) this reduces to the linear problem considered in [8] since the r_α are zero in the minimum. Also, formally, (6.1) reduces to this problem for $h/D \to 0$. Here this quotient is on the order of 10^{-2}. More extensive computations, also for nonconstant obstacles, will be done and presented elsewhere.

TABLE 6.1.
Stability bounds for square plate of length 1 *and thickness* $h = .025$

Boundary condition	stab. bound	% contact
simply supported	168.73	37.1
clamped	319.6	23.1

Appendix. The biharmonic operator on a square grid is approximated by the difference stencils (each has to be multiplied by $\tilde{h}^{-4}$)

$$\begin{array}{ccccc} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array}$$

for grid points a distance of $2\tilde{h}$ or more from $\partial\Omega$. For points adjacent to $\partial\Omega$,

$$\begin{array}{ccccc} 1 & -8 & 21 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array} \qquad \begin{array}{ccc} 22 & -8 & 1 \\ -8 & 2 & \\ 1 & & \end{array}$$

are used for clamped, and

$$\begin{array}{ccccc} 1 & -8 & 19 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array} \qquad \begin{array}{ccc} 18 & -8 & 1 \\ -8 & 2 & \\ 1 & & \end{array}$$

are used for simply supported boundary conditions at $\partial\Omega$ or $\partial\mathcal{A}$. The standard five point star (multiplied by $\tilde{h}^{-2}$)

$$\begin{array}{ccc} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{array}$$

is used for the Laplace operator. Analogously, the terms $w_{,\alpha\beta}$ are approximated by the stencils (times $\tilde{h}^{-2}$)

$$\begin{array}{ccc} 1 & -2 & 1 \end{array} \qquad \begin{array}{c} 1 \\ -2 \\ 1 \end{array} \qquad \begin{array}{ccc} -\frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 \\ \frac{1}{4} & 0 & -\frac{1}{4} \end{array}.$$

For the discretization of $B(v)$ in (6.1) the five-point star is used, which is equivalent to replacing B by

$$B_{\tilde{h}}(v) = \tilde{h}^2 \sum_{i,j=1}^{n+1} [(\delta^1_{i,j}v)^2 + (\delta^2_{i,j}v)^2],$$

where $v_{i,j} \hat{=} v(i\tilde{h}, j\tilde{h})$, $\delta^1_{i,j}v = (v_{i,j} - v_{i-1,j})/\tilde{h}$, and $v_{ij} = 0$ for $i, j = 0$, and $n+1$. Analogously,

$$\begin{aligned} \frac{1}{\tilde{h}^2} A_{\tilde{h}}(v, r) &= \sum_{i,j=1}^{n+1} (\delta^{12}_{i,j}v + \delta^{22}_{i,j}v)^2 \\ &\quad + \frac{h}{D}\left[\sum_{i,j=1}^{n+1} \tilde{\sigma}^{(i,j)}_{\alpha\beta}\delta^{\alpha}_{i,j}v\delta^{\beta}_{i,j}v + \frac{E}{1-\nu^2}\sum_{i,j=1}^{n+1}[(\tilde{e}^{(i,j)}_{11})^2 + (\tilde{e}^{(i,j)}_{22})^2]\right. \\ &\qquad \left. + \frac{2\nu E}{1-\nu^2}\sum_{i,j=1}^{n+1} \tilde{e}^{(i,j)}_{11}\tilde{e}^{(i,j)}_{22} + \frac{4E}{1+\nu}\sum_{i,j=1}^{n+1}(\tilde{e}^{(i,j)}_{12})^2\right], \end{aligned}$$

where

$$\delta^{\alpha 2} = \delta^{\alpha}(\delta^{\alpha}), \qquad \sigma_{\alpha\beta} = (-1)^{\alpha+\beta}\chi_{,\alpha\beta},$$
$$\tilde{e}^{(i,j)}_{\alpha\beta} = \tfrac{1}{2}(\delta^{\beta}_{i,j}r_{\alpha} + \delta^{\alpha}_{i,j}r_{\beta} + w^{(i,j)}_{,\alpha}\delta^{\beta}_{i,j}v + w^{(i,j)}_{,\beta}\delta^{\alpha}_{i,j}v),$$

and the summation convention is used with respect to α, β. The $\tilde{\sigma}_{\alpha\beta}$ are averages of the $\sigma_{\alpha\beta}$ in such a way that the matrix of the quadratic form $A_{\tilde{h}}$ is symmetric. The integrals in (6.2) were approximated by

$$\sum_{i,j=1}^{n+1} \tilde{h}^2 r_{1i,j}, \ \sum_{i,j=1}^{n+1} \tilde{h}^2 r_{2i,j},$$
$$\sum_{i,j=1}^{n+1} \tilde{h}^2[r_{1i,j}(j-1)\tilde{h} - r_{2i,j}(i-1)\tilde{h} - \delta^1_{i,j}r_2 + \delta^2_{i,j}r_1].$$

REFERENCES

[1] G. DUVAUT AND J.-L. LIONS, *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.

[2] G. H. KNIGHTLY AND D. SATHER, *Nonlinear buckled states of rectangular plate*, Arch. Rational Mech. Anal., 54 (1974), pp. 356–372.

[3] W. T. KOITER, *Elastic stability and post-buckling behaviour*, in Proc. Symposium on Nonlinear Problems, R. E. Langer, ed., University of Wisconsin Press, Madison, WI, 1963, pp. 257–275.

[4] L. D. LANDAU, AND E. M. LIFSCHITZ, *Elastizitätstheorie*, Akademie-Verlag, Berlin, 1970.

[5] E. MIERSEMANN, *Verzweigungsprobleme für Variationsungleichungen*, Math. Nachr., 65 (1975), pp. 187–209.

[6] ———, *Über positive Lösungen von Eigenwertgleichungen mit Anwendungen auf elliptische Gleichungen zweiter Ordnung und auf ein Beulproblem für die Platte*, Z. Angew. Math. Mech., 59 (1979), pp. 189–194.

[7] E. MIERSEMANN AND H. D. MITTELMANN, *A free boundary problem and stability for the circular plate*, Math. Meth. Appl. Sci., 9 (1987), pp. 240–250.

[8] ———, *A free boundary problem and stability for the rectangular plate*, Math. Meth. Appl. Sci., 12 (1990), pp. 129–138.

[9] ———, *On the stability in obstacle problems with applications to the beam and plate*, Z. Angew. Math. Mech., 71 (1991), pp. 311–321.

[10] D. SATHER, *Branching and stability for nonlinear shells*, in Applications of Methods of Functional Analysis to Problems in Mechanics, Lecture Notes in Math., 503, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[11] E. TREFFTZ, *Mathematische Elastizitätstheorie*, Handbuch der Physik, Band VI, Springer-Verlag, Berlin, 1928, pp. 47–140.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1117-1124, September 1992

003

ON ASYMPTOTICS OF SOLUTIONS OF ELLIPTIC MIXED BOUNDARY VALUE PROBLEMS OF SECOND-ORDER IN DOMAINS WITH VANISHING EDGES*

JACEK BANASIAK†

Abstract. This paper investigates the behavior of variational solutions of second-order elliptic mixed boundary value problems (MBVP) with real coefficients in n-dimensional domains with edges near the points where the edges are vanishing. It is shown that the first coefficient involved in the decomposition of the solution into regular and singular part can be extended continuously in appropriate spaces across such points, thus showing that the standard decomposition formula holds also in such domains.

Key words. elliptic boundary value problems, nonsmooth domains

AMS(MOS) subject classification. 35J25

Introduction. In this paper we investigate the H^2-regularity of variational solutions of the second-order elliptic problem

$$Au=f \text{ in } \Omega, \quad u=\phi \text{ on } \Gamma_D, \quad Bu=\psi \text{ on } \Gamma_N, \tag{0.1}$$

where $\Omega \subset \mathbb{R}^n$ is a bounded domain with boundary Γ consisting of smooth parts Γ_D and Γ_N and a "collision submanifold" Γ_c, which can coincide with an edge of Γ. A and B are, respectively, second- and first-order differential operators defined in some neighborhood of $\bar{\Omega}$. Problem (0.1) has been investigated since the end of the 1960s, starting from the pioneering paper by Kondratiev [9]. It is known that, in general, the solutions to (0.1) have singularities due to presence of Γ_c. One of the constructive methods in dealing with such singularities is to seek a decomposition of the solution into two parts, one of which is regular, whereas the second carries all information about the singularity of solution. Such methods provided practically complete descriptions of two-dimensional problems; see, e.g., [7], [12], [1], [2]. In more dimensions certain difficulties occur. One type of decomposition formula for variational solution to (0.1), [10], specified, for example, to $A=\Delta$, $B=\partial/\partial n$, and either $\pi/2<\alpha\leqq\omega(t)\leqq\beta<\pi$ or $\pi<\gamma\leqq\omega(t)\leqq\delta<3\pi/2$ on Γ_c reads

$$u(x)=u_{\text{reg}}(x)+c_0(t)r^{\pi/2\omega(t)}\sin\frac{\pi v}{2\omega(t)}. \tag{0.2}$$

Above, $x=(r,\vartheta,t)$, $t\in\Gamma_c$, $r(x)=\operatorname{dist}(x,\Gamma_c)$, u_{reg} is a "regular" function, and c_0 is square-summable on Γ_c.

Formulas of that type are not always satisfactory since u_{reg} is usually not sufficiently regular along Γ_c due to poor regularity of c_0; see, e.g., [6]. So, instead of (0.2), the so-called "nontensor product" decomposition [13] was proposed; see, e.g., [8], [5]. Here the "regular" term is indeed regular with respect to all variables. The main disadvantage of such decomposition is that variables, tangential and transversal to Γ_c, are not explicitly separated as in (0.2), except in special cases, discussed in [13], where the authors have shown that for higher regularity of data the tensor product form exists and has required regularity properties.

* Received by the editors November 16, 1990; accepted for publication (in revised form) November 25, 1991.

† Institute of Mathematics, Technical University of Łódź, Poland.

However, as pointed out in [3], e.g., neither method deals with the case when the opening ω between Γ_D and Γ_N is equal to π in certain points of Γ_c. We shall consider this case here, calling it a "vanishing edge." The difficulty arising here is due to the fact that such a domain is not locally diffeomorphic to a dihedral angle.

Our method of dealing with this difficulty is based on a decomposition formula for variational solution u of two-dimensional problems. To use this formula, we reduce (0.1) locally to a model two-dimensional problem with parameter, thus obtaining a family of decomposition formulas, which can be "glued" together thanks to explicit expressions for coefficients of the decomposition. Thus, our technique leads us directly to formulas of type (0.2), and our main result, specified to the quoted example, is that (0.2) holds if $\pi/2<\alpha\leqq\omega(t)\leqq 3\pi/2$.

We decided to confine ourselves only to the first asymptotic in order to show in the clearest way possible the influence of the "vanishing edge" on the first step of regularization of the variational solution. Higher regularization involves a number of both technical and qualitative problems of its own, [10], [12], [3], which, added to geometrical difficulties, see (A6), creates a difficult problem, which was beyond the scope of our investigation.

1. Basic notations and definitions. Let Ω be a bounded open subset of $\mathbb{R}^n$ with the boundary $\Gamma=\Gamma_D\cup\Gamma_N\cup\Gamma_c$, where Γ_D and Γ_N are open $C^{3,1}$-submanifold in Γ such that $\Gamma_D\cap\Gamma_N=\emptyset$ and $\bar{\Gamma}_D\cap\bar{\Gamma}_N=\Gamma_c$. Γ_c is assumed to be $n-2$-dimensional, $C^{3,1}$-submanifold of Γ, and $\Gamma_c=\bigcup_{i=1}^r\Gamma_i$, $1\leqq r<\infty$. We assume that all Γ_i are closed and every point $x\in \operatorname{Int}_r\Gamma_i$ (interior relative to Γ_c) has a neighborhood in $\mathbb{R}^n$, diffeomorphic to an n-dimensional dihedral angle. Some Γ_i can be submanifolds with boundary $(bd\Gamma_i)$, and for $x\in bd\Gamma_i$ we have the $\omega(x)=\pi$. We assume that if $\Gamma_j\cap\Gamma_i\neq\emptyset$, then $\Gamma_i\cap\Gamma_j=bd\Gamma_i\cap bd\Gamma_j$.

In other words, Γ_c can be situated on the edge of the boundary, and this edge can either vanish or change itself from acute to obtuse (or reversely) in some point of the boundary.

We shall denote by ∂^i the ith partial derivative. Let A be defined on a Sobolev space $H^2(\Omega)$ by

$$Au:=-\sum_{i,j=1}^n\partial^j(a_{ij}\partial^i u)+\sum_{i=1}^n a_i\partial^i u+a_0u. \tag{1.1}$$

We assume that A is strongly elliptic with real coefficients and $a_{ij}\in C^{2,1}(\bar{\Omega})$ for $i,j=1,\cdots,n$ and $a_i\in L_\infty(\Omega)$ for $i=0,\cdots,n$. By $\vec{\nu}_A$ we denote a vector field, coinciding with the conormal field on $\bar{\Gamma}_N$ and $\vec{\mu}:=\vec{\nu}_A+b\vec{\tau}$, where $b\in C^{2,1}(\mathbb{R}_n)$ and $\vec{\tau}$ is some vector field, tangent to $\partial\bar{\Omega}_N$. We define $B:=\gamma\partial^\mu=\gamma\sum_{i=1}^n\mu_i\partial^i$.

We consider the following problem: for given $f\in L_2(\Omega)$, find u satisfying

$$\begin{cases}Au=f & \text{in } \Omega,\\ \gamma u=0 & \text{on } \partial\Omega_D,\\ Bu=0 & \text{on } \partial\Omega_N.\end{cases} \tag{1.2}$$

We can assume that (1.2) has a unique solution $u\in H^1(\Omega)$, [1], [7]. We shall investigate the behavior of u in a vicinity of Γ_c, particularly near a "vanishing edge."

2. Geometry of the problem. In the first step we perform a change of variables, which is an analogue of the "flattening" of the boundary. We shall show that the domain with "vanishing edge" is locally diffeomorphic to a domain, called a generalized dihedral angle, where the opening can vary along the edge.

First we introduce some notation. For $\xi := (\xi_1, \cdots, \xi_n) \in \mathbb{R}^n$, $\xi := (\xi_1, \xi') := (\xi_1, \xi_2, \xi'')$, and also $\xi_1^0 := (\xi_1, 0, \cdots, 0)$, $\xi_2^0 := (\xi_1, \xi_2, 0, \cdots, 0)$. $\vec{e}_i$ is a unit coordinate vector of a Cartesian system, and for any two vectors $\vec{a}$, $\vec{b}$, $\omega(\vec{a}, \vec{b})$ denotes the measure of the angle between them. Moreover, $\{\alpha_{ij}\}_{i,j=i}^n$ is the matrix of principal coefficients of A after transformation.

LEMMA 2.1. *For every $x \in \Gamma_c$ there exists an open neighborhood U_x of x and a $C^{1,1}$ diffeomorphism Φ: $U_x \to W^n$ (a unit cube) such that*

(a) $\Phi(U_x \cap \Gamma_N) = \{\xi \in W^n;\ \xi_1 = 0, \xi_2 > 0\}$;

(b) $\Phi(U_x \cap \Gamma_c) = \{\xi \in W^n;\ \xi_1 = \xi_2 = 0\}$;

(c) *For every $(0, 0, \xi'') \in W^n$ there exist $c_1 = c_1(\xi'')$, $c_2 = c_2(\xi'')$ satisfying $c_1^2 + c_2^2 = 1$ such that*

$$\Phi(U_x \cap \Gamma_D) \cap \{\xi \in \mathbb{R}^n;\ \xi'' = \text{const}\} = \{\xi \in W^n;\ c_1\xi_1 - c_2\xi_2 = 0\}.$$

We define $\omega(\xi'')$ by $\cos \omega = -c_1$ *and* $\sin \omega = c_2$:

(d) $\Phi(U_x \cap \Omega) = \{\xi \in W^n;\ 0 < \omega(\vec{e}_2, \vec{\xi}_2^0) < \omega(\xi'')\}$;

(e) $(\mathcal{J}_\Phi \vec{u})(0, \xi') = \vec{m}(\xi'') = (m_1(\xi''), m_2(\xi''), 0, \cdots, 0)$

for $\xi_2 > 0$ and $m_1 \neq 0$, where $\mathcal{J}_\Phi$ denotes the Jacobi matrix of Φ;

(f) $\bar{\alpha}_{11} = \bar{\alpha}_{22} = 1$ *and* $\bar{\alpha}_{12} = \bar{\alpha}_{21} = 0$ *for* $\xi_1 = \xi_2 = 0$.

Remark 2.2. A sketch of the proof and explicit formula for $\omega(\xi'')$ is given in the Appendix. Here we only note that if the edge of Ω vanishes at x, then $\omega(\xi'') = \pi$, where $(0, 0, \xi'') = \Phi(x)$.

3. Regularity of the solution. According to § 2, MBVP (1.2) is locally equivalent to the following problem:

$$(3.1) \qquad \begin{cases} \tilde{A}u = f & \text{in } D, \\ \tilde{B}u = -\partial_1 u + b(t)\partial_2 u = 0 & \text{on } G_N, \\ \gamma u = 0 & \text{on } G_D. \end{cases}$$

where $t := \xi''$ will hereafter denote the parameter along the edge, D is defined by Lemma 2.1(d), $G_N := \Phi(\Gamma_N)$, $G_D = \Phi(\Gamma_D)$. Moreover, $b(t) = m_2(t)/m_1(t)$ (see Lemma 2.1(e)). We put $G_c := \Phi(\Gamma_c)$. The main aim of the paper is to prove Theorem 3.1, which describes the behavior of the solution to (3.1) close to the points, where $\omega(t) = \pi$. Let $\psi(t) = \operatorname{arccot} b(t)$, $0 < \psi < \pi$, $\lambda_m(t) = (-\psi(t) + m\pi)/\omega(t)$, $m = 0, \pm 1, \cdots$, and $\bar{\xi} = (\xi_1, \xi_2)$, and $D^t = D \cap \{\xi;\ t = \text{const}\}$. Then we have the following.

THEOREM 3.1. *Let $u \in H^1(D)$ be a solution of* (3.1), *and $\vartheta(t) = \omega(t) + \psi(t)$ satisfies $\pi < \vartheta(t) \leqq 2\pi$ for $t \in G_c$. Then*

$$u(\bar{\xi}, t) = u_{\text{reg}}(\bar{\xi}, t) + c_1(t) \cdot r^{\lambda_1(t)} \sin(\lambda_1(t)\theta + \Psi(t)),$$

where

$$c_1 \in L_2(\{t \in G_c;\ \vartheta(t) \geqq k\}),$$

$$u_{\text{reg}} \in L_2(\{t \in G_c;\ k \leqq \vartheta(t) \leqq 2\pi\}, H^2(D_t))$$

for every $k > \pi$.

The proof depends on a number of lemmas and is postponed till the end of this section. We start with the following considerations. According to Lemma 2.1(e),

$$(3.2) \qquad \tilde{A}u = \Delta u + \tilde{A}_1 u + \tilde{A}_2 u + \sum_{i=1}^n \tilde{a}_i \partial^i u + \tilde{a}_0 u,$$

where $\tilde{A}_1 u = \sum_{i,j=1}^{2} \partial^i(\tilde{a}_{ij}\partial^j u)$ with $\tilde{a}_{ij}(0,0,\xi'')=0$ for $i,j=1,2$ and $\tilde{A}_2 u = \sum_{i,j\in K} \partial^i(\tilde{a}_{ij}\partial^j u)$, where

$$K=\{(i,j);\, 1\leqq i\leqq n, 3\leqq j\leqq n \text{ or } 3\leqq i\leqq n, 1\leqq j\leqq n\}.$$

Our aim is to show that if u is a H^1-solution of (3.1) (with compact support), then both $\tilde{A}_1 u$ and $\tilde{A}_2 u$ belong to $L_2(D)$. Certainly, the two last summands in (3.2) are square summable; thus, we include them in the right-hand side of the equation.

We recall that we are concentrated upon points with $\omega(t)=\pi$. Other cases have been extensively dealt with in [4], [10], [12]. We use those results without further comments.

Thanks to Lemma 2.1(a)-(d), the boundary of D near the point $\xi=(0,0,t)\in G_c$ such that $\omega(t)=\pi$ can be described by

$$\xi_1 = d\,(\xi') = \begin{cases} 0 & \text{for } \xi_2 \geqq 0,\\ \xi_2 \tan\omega(t) & \text{for } \xi_2<0,\end{cases}$$

where $\xi'=(\xi_2,t)$. We "flatten" D by

$$\zeta_1=\xi_1-d(\xi'),\qquad \zeta'=\xi'. \tag{3.3}$$

This is an invertible transformation with inverse given by

$$\xi_1=\zeta_1+d\,(\zeta'),\qquad \xi'=\zeta'. \tag{3.3'}$$

For derivatives we have the following formulas:

$$\partial^2\zeta_1=\begin{cases}0, & \xi_2>0,\\ -\tan\omega(t), & \xi_2<0,\end{cases}\qquad \partial^i\zeta_1=\begin{cases}0, & \xi_2>0,\\ -\xi_2\cdot\partial^i\omega(t)\cdot\cos^{-2}\omega(t), & \xi_2<0,\end{cases} \tag{3.4}$$

for $i\geqq 3$ and $\partial^j\zeta_i=\delta_{ij}$ for $i\geqq 2, j\geqq 1$, and $i=j=1$. Since $\partial^i\omega(t)\cdot\cos^{-2}\omega(t)$ is continuous in a neighborhood of t with $\omega(t)=\pi$, we see that $\partial^i\zeta_1$ are continuous for $i\geqq 3$ and only $\partial^2\xi_1$ suffers a jump across G_c. Thus, the transformation (3.3) is bi-Lipschitz with Jacobian equal to 1. Since $u\in H^1(D)$, (3.1) is equivalent to

$$\int_D \sum_{i,j=1}^{n} \hat{a}_{ij}\partial^i u\partial^j v\, d\xi = \int_D fv\, d\xi \tag{3.5}$$

for every $v\in H^1(D)$ such that $\gamma u=0$ on G_D, where $\hat{a}_{12}:=\tilde{a}_{12}-b$ and $\hat{a}_{21}:=\tilde{a}_{21}+b$, and $\hat{a}_{ij}:=a_{ij}$ for remaining i, j. If $\bar{u}(\zeta):=u(\xi(\zeta))$, then $\bar{u}$ satisfies the identity

$$\int_{\mathbb{R}_n^+} \sum_{i,j=1}^{n} \bar{a}_{ij}\partial^i\bar{u}\partial^j v\, d\zeta = \int_{\mathbb{R}_n^+} \bar{f}v\, d\zeta \tag{3.6}$$

for every $v\in H^1(\mathbb{R}_n^+)$ satisfying $\gamma v=0$ for $\zeta_1=0$, $\zeta_2<0$ where $\bar{a}_{ij}:=\sum_{k,l=1}^{n}\hat{a}_{kl}\partial^i\zeta_k\partial^j\zeta_l$. By (3.4) a discontinuity of $\bar{a}_{ij}$ can only appear if either i or j equals 2, and then $\bar{a}_{ij}$ suffers a jump across G_c. Therefore, each $\bar{a}_{ij}$ is Lipschitz continuous in the direction of G_c and the following lemma holds.

LEMMA 3.2. *If $\bar{u}$ is a variational solution of* (3.5) *then $\partial^k\partial^1\bar{u}\in L_2(\mathbb{R}_n^+)$ for $(k,l)\in K$.*

Proof. The proof is analogous to Nirenberg's method of differential quotients [7], if we notice that the estimates for $\partial^k\partial^1\bar{u}$ involve only bounds for $\partial^1\bar{a}_{ij}$ $i,j=1,\cdots,n$ [7, Form. (2.2.2.4)]; so, if $k\geqq 1$, $l\geqq 3$, then $\partial^k\partial^1\bar{u}\in L_2(\mathbb{R}_n^+)$ by (3.4). □

Therefore, we see that $\bar{u}$ satisfies the equation

$$\sum_{i,j=1}^{2}\bar{a}_{ij}\partial^i\partial^j\bar{u}=F=\bar{f}-\sum_{i,j\in K}\bar{a}_{ij}\partial^i\partial^j\bar{u}\quad\text{in } \mathbb{R}_n^+, \tag{3.7}$$

where, by Lemma 3.2, $F \in L_2(\mathbb{R}_n^+)$. However, (3.7) is still not satisfactory since the $\tilde{a}_{ij}$ have jumped across G_c. To overcome this difficulty, let us consider (3.7) in the variational form:

$$\int_{\mathbb{R}_n^+} \sum_{i,j=1}^{2} \tilde{a}_{ij}\partial^i \tilde{u} \partial^j \tilde{v}\, d\zeta = \int_{\mathbb{R}_n^+} Fv\, d\zeta \tag{3.8}$$

for v as in (3.6). Applying (3.3′) to the identity (3.8), we see that $u \in H^1(D)$ satisfies

$$\int_D \sum_{i,j=1}^{2} a_{ij}^* \,\partial^i u \partial^j v\, d\xi = \int_D F^* v\, d\xi \tag{3.9}$$

for v as in (3.5), where

$$a_{11}^* = \tilde{a}_{11} - 2\sum_{j=3}^{n} \tilde{a}_{1j}\partial^j d - \sum_{i,j=3}^{n} \tilde{a}_{ij}\partial^i d \partial^j d, \tag{3.10}$$

$$a_{ij}^* = \hat{a}_{ij} - \sum_{k=3}^{n} \tilde{a}_{2k}(\partial^k d) \quad \text{for } i=1, j=2 \quad \text{or} \quad i=2, j=1 \text{ (see (3.5))},$$

$$a_{22}^* = \tilde{a}_{22} \quad \text{and} \quad F^*(\xi) = F(\zeta(\xi)).$$

Thus, coefficients of (3.9) are Lipschitz continuous. Fubini's theorem, applied to (3.9), shows that for almost everywhere $t \in G_c u(\cdot, t)$ is a variational solution of the two-dimensional problem:

$$\int_{D'} \sum_{i,j=1}^{2} a_{ij}^*\partial^i u \partial^j v\, d\xi_1\, d\xi_2 = \int_{D'} F^* v\, d\xi_1\, d\xi_2, \tag{3.11}$$

where $v \in H^1(D')$, $\gamma v = 0$ on $G_D^t := G_D \cap \{\xi;\ t = \text{const}\}$. Now we have the following.

LEMMA 3.3. *Let* $r = (\xi_1^2 + \xi_2^2)^{1/2}$. *If* $u \in H^1(D)$ *satisfies* (3.1), *then* $r\partial^i\partial^j u \in L_2(D)$ *for* $i, j = 1, 2$.

Proof. The proof depends on Kondratiev's estimation for MBVP in a rectilinear angle, [9]; see also [14]. Since $u(\cdot, t)$ satisfies (3.10), we have for almost every t and $i, j = 1, 2$, [9],

$$\int_{D'} r^2 |\partial^i\partial^j u|^2\, d\xi_1\, d\xi_2 \leqq C \int_{D'} ((F^* r)^2 + (\partial^1 u)^2 + (\partial^2 u)^2 + (r^{-1}u)^2)\, d\xi_1\, d\xi_2, \tag{3.12}$$

where the constant C depends on the bound for coefficients and its first derivatives and, therefore, can be made common for all $t \in G_c$. Since $u = 0$ on Γ_D^t, we can use Hardy's inequality to estimate $(r^{-1}u)^2$. Finally, integrating (3.11) along G_c, we obtain the thesis of the lemma. □

The following proposition concerns the tangential regularity of the solution to (3.1), and although it is not used in the sequel, we give it for its own interest.

PROPOSITION 3.4. *If* u *satisfies* (3.1), *then* $\partial^i\partial^j u \in L_2(D)$ *for* $(i, j) \in K$.

Proof. We have $\partial^i\partial^j \bar{u} = \sum_{k,l=1}^{n} (\partial^k\partial^1 u)\partial^i\xi_k\partial^j\xi_l + \sum_{k=1}^{n} \partial^k u \partial^i\partial^j\xi_k$, $i, j = 1, \cdots, n$ where $\xi(\zeta)$ is defined by (3.3′). Since $(i, j) \in K$ and $u \in H^1(D)$, we see by (3.3) that the second summand is square integrable and will be omitted in the sequel. Moreover, $\partial^i\xi_k\partial^j\xi_1 \neq 0$ if and only if either $k = i$ and $l = j$ or $k = l = 1$ or we have a combination of the above conditions. Therefore, for $j \geqq 3$ $\partial^1\partial^j\bar{u} = \partial^1\partial^1 u\partial^j\xi_1 + \partial^1\partial^j u$, and since then $|\partial^j\xi_1| \leqq Mr$ for some $M \geqq 0$, we see by Lemmas 3.2 and 3.3 that $\partial^1\partial^j u \in L_2(D)$. Now let $i > 1$, $j \geqq 3$. Then $\partial^i\partial^i\bar{u} = \partial^1\partial^1 u\partial^i\xi_1\partial^j\xi_1 + \partial^i\partial^1 u\partial^j\xi_1 + \partial^1\partial^j u\partial^i\xi_1 + \partial^i\partial^j u$, and it follows that $|\partial^i\xi_1\partial^j\xi_1| \leqq Mr$. Therefore, application of Lemmas 3.2 and 3.3 and the first part of the proof gives the thesis. □

Now, from (3.2) and (3.10) we see that $a_{11}^* = 1 + \alpha$, where $|\alpha| \leqq Mr$ for some $M > 0$; thus, by Lemma 3.3, $\alpha \partial^1 \partial^1 u \in L_2(D)$. Similar considerations, applied to the remaining a_{ij}^*, show that for almost every $t \in G_c$, $u(\cdot, t)$ is an H^1-solution of the problem

$$\text{(3.13)} \qquad \begin{cases} \Delta u(\cdot, t) = \bar{\bar{F}}(\cdot, t) & \text{in } D^t, \\ B_t u = \partial^1 u(\cdot, t) + b(t)\partial^2 u = 0 & \text{on } G_N^t, \\ \gamma u(\cdot, t) = 0 & \text{on } G_D^t, \end{cases}$$

where $\bar{\bar{F}} = F^* - \sum_{i=1}^2 (a_{ii}^* - 1)\partial^i \partial^i u - (a_{12}^* + b)\partial^1\partial^2 u - (a_{21}^* - b)\partial^2 \partial^1 u \in L_2(D)$.

Therefore, we can use the two-dimensional theory developed in [1], [2], [4]. If we denote

$$\mathfrak{D}_t^1 = \{u \in H^1(D^t);\ \Delta u \in L_2(D^t),\ B_t u = 0 \text{ on } G_N^t,\ \gamma u = 0 \text{ on } G_D^t\},$$

$$\mathfrak{D}_t^2 = \mathfrak{D}_t^1 \cap H^2(D^t),$$

then, in particular, Δ is a topological isomorphism of $\mathfrak{D}_t^1$ (equipped with the graph norm) onto $L_2(D^t)$, and $\mathfrak{D}_t^2$ is a closed subspace of $\mathfrak{D}_t^1$ of finite codimension for any $\omega(t)$. Also, the image of $\mathfrak{D}_\mathrm{t}^2$ is closed in $L_2(D^t)$ since the boundary data are homogeneous. As a consequence, we see that for almost all $t \in G_c$, the solution u of (3.13) has the form

$$\text{(3.14)} \qquad u(\bar{\xi}, t) = u_r(\bar{\xi}, t) + \sum_{0<\lambda_m(t)<1} c_m(t) r^{\lambda_m(t)} \cdot \sin(\lambda(t)\theta + \psi(t)),$$

where $(r(\bar{\xi}), \theta(\bar{\xi}))$ are plane polar coordinates centered at zero and $u_r(\cdot, t) \in H^2(D^t)$. It follows, [1], that (3.14) holds for every ω and ψ. Now we can complete the proof of Theorem 3.1, focusing on the case $\omega(t) = \pi$, since the detailed considerations concerning the case where $\omega(t) \neq \pi$ can be found in [3], [10].

Proof of Theorem 3.1. In such a simple case as (3.13) it is possible to determine c_1 explicitly as in [10], [11]. We omit lengthy calculations that are similar to those in [10]. [11]. Let $w_1(\bar{\xi}, t) = w_1(r, \theta, t) = r^{-\lambda_1(t)} \times \sin(\lambda_1(t)\theta + \psi(t))$, and $\eta \in C_0^\infty(\mathbb{R}_n)$ be a function equal to 1 in the neighborhood of G_c, satisfying the adjoint boundary condition on $\bar{G}_D \cup \bar{G}_N$. Then

$$\text{(3.15)} \qquad c_1 = \frac{1}{\pi - \psi}\left(\int_{D^t} \bar{\bar{F}}\eta w_1 \, d\bar{\xi} + \int_{D^t} u\Delta(\eta w_1)\, d\bar{\xi}\right).$$

We see that the second integral, say $I_2(t)$, has no influence on regularity of c_1 since $\Delta(\eta w_1) = 0$ in a neighborhood of G_c. For the first integral, we have

$$\text{(3.16)} \qquad |I_1(t)| \leqq R(t) \|\bar{\bar{F}}(\cdot, t)\|_{L_2(D^t)} \cdot (\vartheta(t) - \pi)^{-1}$$

for a bounded R; so $c_1 \in L_2(\{t;\ \vartheta(t) \geqq k\})$ for k defined above.

Next, $u_{\mathrm{reg}}(\bar{\xi}, t) = u(\bar{\xi}, t) - c_1(t) \cdot w_1(\bar{\xi}, t)$ and since both summands on the right-hand side belong to $L_2(\{t;\ \vartheta(t) \geqq k\}, \mathfrak{D}_t^1)$, the left-hand side does also. However, $u_{\mathrm{reg}} \in \mathfrak{D}_t^2$ and norms, induced on $\mathfrak{D}_t^2$ from $H^2(D_t)$ and $\mathfrak{D}_\mathrm{t}^1$, are equivalent, [1], so $u_{\mathrm{reg}} \in L_2(\{t;\ \vartheta(t) \geqq k\}, H^2(D_t))$. □

Remark 3.6. Theorem 3.4 can serve as a tool for proving regularity results for u both in $H^s(\Omega)$, $1 < s < 2$, $s \neq 3/2$ and in weighted Sobolev spaces $H^{2,2}(\Omega, r, \varepsilon)$, [10]. However, as we saw, the first asymptotic of MBVP behaves in the same way in the neighborhood of points, where $\omega = \pi$, as it does near other points of G_c (as only $\vartheta \neq \pi$, 2π); thus, we do not go into detail.

Appendix. Here we prove Lemma 2.1 and give an explicit formula for the opening $\omega(t)$ in terms of geometrical properties of the original boundary.

Proof. We obtain Φ as a superposition of four transformations, the first of which is a modification of a normal transformation [15]. Hereafter, we adopt a convention that superscripts I, II, III, and IV will refer to successive steps of our construction. In some neighborhood of $x, \partial\Omega_N$ is a graph of $C^{3,1}$-function ϕ with outward normal given by

$$\text{(A1)} \qquad \vec{\nu} = \vec{n}/|\vec{n}|, \quad \text{where } \vec{n} = (-1, \partial^2\phi, \cdots, \partial^n\phi)$$

and $\lambda = \lambda(x) = [(\vec{\mu} \cdot \vec{\nu})(x)]^{-1/2} = [(\vec{\nu}_A \cdot \vec{\nu})(x)]^{-1/2}$ for $x \in \Gamma_c$. We define new coordinates in the following way:

$$\text{(A2)} \qquad x_1 = \phi(y_2, \cdots, y_n) - \lambda y_1 \mu_1, \quad x_i = y_i - \lambda y_1 \mu_i, \quad i = 2, \cdots, n.$$

Let $y = \Phi^{\mathrm{I}}(x)$, then $\det \mathcal{J}^{\mathrm{I}} = 1/\lambda(\vec{\mu} \cdot \vec{n}) = (\vec{\mu} \cdot \vec{\nu})/|\vec{n}|$ on Γ_c. It is easy to check that $\mathcal{J}^{\mathrm{I}}\vec{\mu} = -\lambda^{-1}\vec{e}_1$ and also that $a^{\mathrm{I}}_{11} = 1$ on $\Phi^{\mathrm{I}}(\Gamma_c)$.

Furthermore, let Γ_c be locally defined as a graph:

$$\Gamma_c = \{x \in \mathbb{R}^n;\ x_1 = \phi(\gamma(x''), x''),\ x_2 = \gamma(x'')\}.$$

It is seen that $\Gamma^{\mathrm{I}}_c = \Phi^{\mathrm{I}}(\Gamma_c \cap U_x) = \{y \in W^n;\ y_1 = 0,\ y_2 = \gamma(y'')\}$. The second transformation, Φ^{II}, will straighten Γ^{I}_c and make a^{II}_{22} equal 1 on $\Gamma^{\mathrm{II}}_c = \Phi^{\mathrm{II}}(\Gamma^{\mathrm{I}}_c)$. We define it as an inverse of

$$\text{(A3)} \qquad y_1 = z_1, \quad y_2 = z_2\beta(z'') = \gamma(z''), \quad y'' = z'' + \beta(z'')z_2,$$

where β is defined below. Let us denote by A^{I}_0 the Laplace–Beltrami operator on $y_1 = 0$ generated by A^{I} and by $\vec{\nu}_0 = \vec{n}_0/|\vec{n}_0|$, $\vec{n}_0 = (1, -\partial^3\gamma, \cdots, -\partial^n\gamma)$. Taking

$$\text{(A4)} \qquad \beta(y') = [(\vec{\nu}_{A^{\mathrm{I}}_0} \cdot \vec{\nu}_0)(y')]^{1/2}, \qquad y' \in \Gamma^{\mathrm{I}}_c,$$

we obtain $a^{\mathrm{II}} = 1$ on $\Gamma^{\mathrm{II}}_c = \{z;\ z_1 = z_2 = 0\}$, and, as defined, Φ^{II} does not change properties achieved in the first step.

In step III we perform a transformation to obtain $a^{\mathrm{III}}_{12} = a^{\mathrm{III}}_{21} = 0$ on Γ^{III}_c. We can take, for example,

$$\text{(A5)} \qquad v_1 = z_1, \quad v_2 = b(z'')z_1 + c(z'')z_2, \quad v'' = z'',$$

where $b = -a^{\mathrm{II}}_{12} \cdot [1 - (a^{\mathrm{II}}_{12})^2]^{-1/2}$ and $c = [1 - (a^{\mathrm{II}}_{12})^2]^{-1/2}$ on Γ^{II}_c. Now $\mathcal{J}^{\mathrm{III}}(-\lambda^{-1}\vec{e}_1) = -\lambda^{-1}(1, b, 0, \cdots, 0)$, and since both λ and b depend only on variables from Γ_c we see that (e) is satisfied after step (III). Clearly $a^{\mathrm{III}}_{11} = a^{\mathrm{III}}_{22} = 1$ for $v_1 = v_2 = 0$.

In the last step we straighten lines $L(v'') = \Phi^{\mathrm{III}}\Phi^{\mathrm{II}}\Phi^{\mathrm{I}}(\partial\Omega_D \cap U_x) \cap \{v;\ v'' = \text{const}\}$ by projecting them onto their tangent lines. In our case $(x \in bd\Gamma_i)$ we can assume that $L(v'')$ is locally graph of a function $v_1 = \chi(v_2, v'')$, and, consequently, define Φ^{IV} in the following way:

$$\text{(A6)} \qquad \xi_1 = \begin{cases} v_1 & \text{for } v_2 \geqq 0, \\ v_1 + v_2\partial^2\chi(0, v'') - \chi(v_2, v'') & \text{for } v_2 < 0, \end{cases}$$
$$\xi' = v'.$$

Since Γ_D is a $C^{3,1}$-manifold and $\mathcal{J}^{\mathrm{IV}} = \mathbb{I}$ along the collision submanifold, Φ^{IV} is a $C^{1,1}$-diffeomorphism, which does not change properties achieved in the first three steps. Therefore, all requirements of the lemma are satisfied by the transformation $\Phi = \Phi^{\mathrm{IV}}\Phi^{\mathrm{III}}\Phi^{\mathrm{II}}\Phi^{\mathrm{I}}$, defined in a suitable neighborhood of x. □

Now we express $\omega(\xi'')$ and $\psi(\xi'')$ in terms of quantities defined on the original boundary. It is clear, (A4), that $\cos\psi(\xi'') = -a_{12}^{\mathrm{II}}(0,0,v'')$, and furthermore $a_{12}^{\mathrm{II}} = \beta^{-1}[(a_{12}^{\mathrm{I}}, \cdots, a_{1n}^{\mathrm{I}}) \cdot \vec{n}_0]$ (where both sides are calculated along Γ_c), and since $a_{1i}^{\mathrm{I}} = -(\vec{\nu}_A \cdot \vec{m}_i)/(\vec{\nu}_A \cdot \vec{n})$ (see (A1)), we have

$$\cos\psi = (\vec{\nu}_A \cdot \vec{n})^{-1} \sum_{i=2}^{n} (\vec{\nu}_A \cdot \vec{m}_i) n_{0i} \quad \text{on } \Gamma_c. \tag{A7}$$

Let $\vec{\tau}_k := (\partial^k \phi, 0, \cdots, 0, 1, 0, \cdots 0)$, where 1 is situated on the kth place. We define $\vec{m} := \mathbb{X}(\lambda\vec{\mu}, \vec{\tau}_3, \cdots, \vec{\tau}_n)$, where $\mathbb{X}$ denotes the vector product of $n-1$ vectors in n-dimensional space. Let $\vec{\tau}_N = \mathcal{J}_\Phi^{-1}\vec{e}_2$ and denote by $\vec{\tau}_D$ the vector, tangent to G_D, lying in the plane, spanned by $\vec{\tau}_N$ and $\vec{\mu}$. It follows then, from (A5), that

$$\cos\omega = \frac{-b(\vec{n}\cdot\vec{\tau}_D) + c(\vec{m}\cdot\vec{\tau}_D)}{((\vec{n}\cdot\vec{\tau}_D)^2 + [c(\vec{m}\cdot\vec{\tau}_D) - b(\vec{n}\cdot\vec{\tau}_D)]^2)^{1/2}}.$$

REFERENCES

[1] J. BANASIAK AND G. F. ROACH, *On mixed boundary value problem of Dirichlet-oblique derivative type in plane domains with piecewise-differentiable boundary*, J. Differential Equations, 79 (1989), pp. 111-131.

[2] ———, *On corner singularities of solutions to mixed boundary value problems for elliptic and parabolic equations*, Proc. Roy. Soc. London Ser. A, 433 (1991), pp. 209-217.

[3] M. COSTABEL AND M. DAUGE, *Edge asymptotics on a skew cylinder*, to appear.

[4] M. DAUGE, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.

[5] ———, *Problèms aux limites elliptiques dans les domains à coins: singularitès aux sommets et le long des arêtes*, C. R. Acad. Sci. Paris Sér. I. Math., 304 (1987), pp. 563-566.

[6] P. GRISVARD, *Singularitès des problèmes aux limites dans des polyedrès*, Sém. Goulaouic-Meyer-Schwartz, 8 (1981-82).

[7] ———, *Elliptic Boundary Value Problems in Non-Smooth Domains*, Pitman, Boston, MA, 1985.

[8] ———, *Edge behaviour of the solution of an elliptic problem*, Math. Nachr., 132 (1987), pp. 281-299.

[9] V. A. KONDRATIEV, *Boundary value problems for elliptic equations in domains with conical and angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 277-313.

[10] A. KUFNER AND A.-M. SANDIG, *Some Applications of Weighted Sobolev Spaces*, Teubner-Texte Math., Leipzig, 1987.

[11] D. LEGUILLON AND E. SANCHEZ-PALENCIA, *Computation of Singular Solutions in Elliptic Problems and Elasticity*, Rech. Math. Appl. 5, John Wiley-Masson, Paris, 1987.

[12] V. G. MAZ'JA AND J. ROSSMAN, *Über die Asymptotik der Lösungen elliptischer Randwertaufgaben in der Umgebung von Kanten*, Math. Nachr., 138 (1988), pp. 27-53.

[13] T. V. PETERSDORF AND E. P. STEPHAN, *Decomposition in edge and corner singularities for the solution of the Dirichlet problem of the Laplacian in a polyhedron*, Math. Nachr., 149 (1990), pp. 71-104.

[14] J. ROSSMAN, *Das Dirichletproblem fur stark elliptische Differentialgleichungen bei denen die rechte Seite f zum Raum $W^{-k}(G)$ gehort*, in Gebieten mit konischen Ecken, Rostock Math. Kolloq., 22 (1983), pp. 13-41.

[15] J. WLOKA, *Partial Differential Equations*, Cambridge University Press, Cambridge, 1987.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1125-1140, September 1992

004

ABRIKOSOV'S VORTICES IN THE CRITICAL COUPLING*

SHENG WANG† AND YISONG YANG‡

Abstract. A necessary and sufficient condition is obtained for the existence of multivortex solutions of the Bogomol'nyi system arising in the abelian Higgs theory defined on a rectangular domain and subject to a 't Hooft type periodic boundary condition. In particular, the number of vortices of a solution is confined by the size of the domain. Such solutions realize the magnetic periodic cell structure in a superconductor predicted by Abrikosov. If the periodic boundary condition is removed, the Bogomol'nyi equations on a bounded domain possess solutions with an arbitrary number of vortices, and these solutions may be used to approximate the unique finite energy solution over the full plane. Moreover, it is shown that, for any given vortex distribution in the plane, the Bogomol'nyi system has a continuous family of gauge-distinct solutions with infinite energy.

Key words. Bogomol'nyi equations, vortex-like solutions, maximum principle, sub- and supersolutions, 't Hooft periodic boundary condition, quantized magnetic flux, superconductivity, gauge transformations

AMS(MOS) subject classifications. 35Q20, 81J05, 81E13

1. Introduction. One of the most significant features of the classical two-dimensional abelian Higgs theory is that it allows a family of vortex-like solutions labelled by topological integers. It is known that the dimensionless Higgs self-coupling parameter, $\lambda > 0$, plays a crucial role in the model, so that, in suitably normalized units, for $\lambda < 1$, vortices attract, while, for $\lambda > 1$, repel, and the system is unstable against decay into separated vortices [10]. In fact, currently, all known finite energy solutions for $\lambda \neq 1$ are rotationally symmetric [18], [4], which can well be viewed as vortices clustered together, and it was conjectured that [11] these field configurations are the only finite energy solutions of the model modulo gauge transformations, although a verification of this statement still remains outstanding. On the other hand, for the critical choice $\lambda = 1$, the situation is entirely different. The numerical study in [10] first suggested that $\lambda = 1$ vortices do not interact, and solutions with finite separations between vortex locations should exist. This conclusion was then supported by an index theorem argument [22]. Finally, the existence of such arbitrary multivortex solutions was proved through a nonconstructive variational approach [11].

The main purpose of the present paper is to obtain the $\lambda = 1$ multivortices of the abelian Higgs theory over a periodic lattice cell realized by a 't Hooft type boundary condition and to approximate the full plane finite energy vortices of Jaffe and Taubes [11] by bounded domain solutions. The former may well be viewed as a compact version of the model. Multivortex solutions of such a nature were first explored in the pioneering work of Abrikosov [1] in the context of the Ginzburg-Landau theory [8] of low transition temperature superconductivity, where it was predicted that, under some circumstances, the magnetic flux may penetrate the sample in the form of vortex-lines with a period structure, and the total flux through a lattice cell is a quantized value proportional to the number of vortices confined. Although this prediction has been confirmed in numerous physical experiments, a mathematical verification has not been worked out in the literature yet. Earlier attempts include, for example, the work

* Received by the editors April 17, 1991; accepted for publication (in revised form) January 21, 1992.

† Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003.

‡ Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131. Current address: Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

of Odeh [17], where a prescribed flux problem was studied so that the order parameter and the magnetic field were periodic. However, his solutions do not give rise to vortex-lines and the quantized flux. We shall show in this paper that the 't Hooft periodic boundary condition [21] can naturally be used to realize Abrikosov's multivortices in the critical case $\lambda = 1$. This appears to be the first rigorous result concerning the existence of a quantized periodic lattice pattern in the classical theory of superconductivity.[1] On the other hand, the latter arises from the problem of constructing vortex solutions in the self-dual limit. Gauge field theories in two dimensions are rather different from those in three and four dimensions. It is well known that in three and four dimensions topologically nontrivial solutions (monopoles and instantons) may be found explicitly [19], [5], [23], [2], while in two dimensions, no explicit vortex-type solutions have ever been obtained. Hence it will be interesting to provide a constructive method to get such solutions. Some recent relevant numerical simulations include [7], [9], [6], [14].

The contents of the paper are as follows. In § 2, a 't Hooft type periodic boundary condition is introduced for the abelian Higgs model. Due to this boundary condition, the magnetic flux will assume a nontrivial value proportional to the total number of vortices confined in the finite rectangular region under consideration. For the critical choice $\lambda = 1$, the energy lower bound can be saturated by the solutions of the Bogomol'nyi equations [5]. These equations are then reduced to a scalar elliptic equation, coupled with a source term characterizing the locations of vortices. In § 3, we solve this elliptic equation via a monotone iteration method. Such a method may provide a useful construction of the Abrikosov solutions of the Bogomol'nyi system over a periodic lattice cell. The condition which ensures the existence of multivortices indicates that the total number of vortices or magnetic flux strength cannot be arbitrarily large, but is bounded by the size of the domain. Section 4 is a simple remark on the magnetic properties of multivortex solutions in the presence of an external field. In § 5, we solve the Bogomol'nyi system over a bounded domain subject to a natural boundary condition imposed on the observables—the amplitude of the Higgs field and the magnetic field. We shall see that solutions with arbitrarily large vortex numbers exist. Such solutions are constructed by an iterative method similar to that in § 3. Two numerical examples will be presented as an illustration. In § 6 we prove that the bounded domain solutions obtained in § 5 can be used to approximate in a global way the full plane finite energy multivortices of Jaffe and Taubes [11]. In § 7 we show that for any distribution of vortices, the Bogomol'nyi system has a continuous family of gauge-distinct solutions of infinite energy. Such an observation may be viewed as a complement to the uniqueness theorem in [11] for finite energy solutions. Section 8 is a brief summary.

2. The abelian Higgs model over a lattice cell. The energy density of the static abelian Higgs vortex model is given by the expression

$$\mathscr{E}(\phi, A) = \frac{1}{4} F_{jk}^2 + \frac{1}{2} |D_j\phi|^2 + \frac{\lambda}{8} (|\phi|^2 - 1)^2,$$

where $A = (A_1, A_2)$ is the gauge potential of the magnetic field $F_{jk} = \partial_j A_k - \partial_k A_j$, ϕ the complex scalar Higgs field, and $D_j = \partial_j - iA_j$ the gauge-covariant differentiation.

[1] However, in physical experiments, quantized periodic vortices are observed mainly in type II superconductors for which $\lambda > 1$. In such a situation there is no Bogomol'nyi reduction and the full second-order Ginzburg-Landau equations have to be considered and, thus, our method here fails.

We assume that the field configuration (ϕ, A) satisfies the following 't Hooft [21] type periodic boundary condition on the boundary of the rectangular region $\Omega = (-L_1, L_1)\times(-L_2, L_2)$ in $\mathbb{R}^2$:

$$\begin{aligned} &\phi(-L_1, x_2)\, e^{i\xi(-L_1,x_2)} = \phi(L_1, x_2)\, e^{i\xi(L_1,x_2)}, && -L_2 < x_2 < L_2,\\ &\phi(x_1, -L_2)\, e^{i\zeta(x_1,-L_2)} = \phi(x_1, L_2)\, e^{i\zeta(x_1,L_2)}, && -L_1 < x_1 < L_1, \end{aligned} \tag{2.1}$$

$$\begin{aligned} &A(-L_1, x_2) + (\nabla\xi)(-L_1, x_2) = A(L_1, x_2) + (\nabla\xi)(L_1, x_2), && -L_2 < x_2 < L_2,\\ &A(x_1, -L_2) + (\nabla\zeta)(x_1, -L_2) = A(x_1, L_2) + (\nabla\zeta)(x_1, L_2), && -L_1 < x_1 < L_1, \end{aligned} \tag{2.2}$$

where ξ and ζ are real phase change variables. The requirement that ϕ be single-valued implies in particular the relation

$$\begin{aligned} &\xi(L_1, L_2^-) - \xi(L_1, -L_2^+) + \xi(-L_1, -L_2^+) - \xi(-L_1, L_2^-)\\ &\quad + \zeta(-L_1^+, L_2) - \zeta(L_1^-, L_2) + \zeta(L_1^-, -L_2) - \zeta(-L_1^+, -L_2) + 2\pi N = 0 \end{aligned} \tag{2.3}$$

with $N \in \mathbb{Z}$. Therefore, from (2.2) and (2.3), we find the total quantized magnetic flux through Ω:

$$\Phi = \int_\Omega F_{12}\, dx = \int_{\partial\Omega} A_j\, dx_j = 2\pi N. \tag{2.4}$$

It is interesting to notice that Φ is independent of the size of Ω.

For $\lambda = 1$, using (2.4), the energy may be rewritten in the form

$$\begin{aligned} E(\phi, A) &\equiv \int_\Omega \mathscr{E}(\phi, A)\, dx\\ &= \pi|N| + \frac{1}{2}\int_\Omega dx\left\{\left|F_{12} \pm \frac{1}{2}(|\phi|^2 - 1)\right|^2 + |D_1\phi \pm iD_2\phi|^2\right\}\\ &\quad \pm \frac{1}{2}\,\mathrm{Im}\left\{\int_\Omega dx\,\{\partial_j(\varepsilon_{jk}\phi^* D_k\phi)\}\right\}, \end{aligned}$$

according to $N = \pm|N|$, where ε_{jk} is the standard skew-symmetric 2-tensor with $\varepsilon_{12} = 1$.

However, the periodic boundary condition (2.1), (2.2) implies

$$\int_\Omega dx\,\{\partial_j(\varepsilon_{jk}\phi^* D_k\phi)\} = \int_{\partial\Omega} \phi^* D_j\phi\, dx_j = 0.$$

As a consequence, there holds the energy lower bound estimate as in the $\mathbb{R}^2$ case (cf. [11]):

$$E(\phi, A) \geqq \pi|N|.$$

This lower bound is saturated if and only if (ϕ, A) is a solution of the self-dual (or anti-self-dual) Bogomol'nyi equations

$$\begin{aligned} D_1\phi \pm iD_2\phi &= 0,\\ F_{12} \pm \tfrac{1}{2}(|\phi|^2 - 1) &= 0, \end{aligned} \qquad x \in \Omega \tag{2.5}$$

subject to the periodic boundary condition (2.1), (2.2) on $\partial\Omega$.

Without loss of generality, let us consider the self-dual ($N > 0$) equations (2.5+). The solutions of the anti-self-dual system (2.5−) may be obtained by taking the "conjugate" $(\phi, A) \to (\phi^*, -A)$ of the solutions of (2.5+).

Let $Z(\phi)$ be the set of zeros of ϕ in Ω. It is convenient to view Ω as a subset of $\mathbb{C}$ and use $z=x_1+ix_2$ to denote a point in Ω. We assume that ϕ does not vanish on the boundary $\partial\Omega$. As was shown in [11], the structure of $(2.5+)_1$ allows us to conclude that $Z(\phi)$ is a finite set, say $Z(\phi)=\{z_1,\cdots,z_k\}$, and that, in a neighborhood of z_j, ϕ has the representation

$$(2.6)\qquad \phi(z)=(z-z_j)^{n_j}h_j(x_1,x_2)$$

so that n_j is a positive integer and h_j is a smooth nonvanishing scalar function. The well-known prescribed vortex problem is that, given $z_1,\cdots,z_k\in\Omega$ and $n_1,\cdots,n_k\in\mathbb{Z}_+$, find a solution of (2.5+) such that $Z(\phi)=\{z_1,\cdots,z_k\}$ and the multiplicity of the zero $z=z_j$ of ϕ is exactly n_j, $j=1,\cdots,k$. The multiplicity n_j is sometimes called the local charge or local vortex number of the solution at $z=z_j$.

For a solution pair (ϕ, A) of (2.5+), since $|\phi|^2$ is periodic, it is easy to see from the method in [11] that $|\phi|^2<1$ everywhere in the periodic cell or otherwise $|\phi|^2\equiv 1$, which means that the solution is gauge-equivalent to the trivial solution $\phi=1$, $A=0$.

3. A construction of multivortex solutions. With the notation $\partial=\frac{1}{2}(\partial_1-i\partial_2)$, $\alpha=A_1+iA_2$, $(2.5+)_1$ yields the relation

$$(3.1)\qquad \alpha=-2i\partial^*\ln\phi,\quad \text{away from } Z(\phi);$$

therefore, outside $Z(\phi)$,

$$F_{12}=-i(\partial\alpha-\partial^*\alpha^*)=-2\partial\partial^*\ln|\phi|^2=-\tfrac{1}{2}\Delta\ln|\phi|^2.$$

As a consequence of the above equation and (2.6), it is seen that $u\equiv\ln|\phi|^2$ satisfies

$$(3.2)\qquad \begin{array}{ll}\Delta u=e^u-1+4\pi\sum_{j=1}^k n_j\delta(z-z_j) & \text{in }\Omega\\ u & \text{is periodic on }\partial\Omega.\end{array}$$

Conversely, if u is a solution of (3.2), then (ϕ, A) is a smooth solution of the Bogomol'nyi system (2.5+) subject to the periodic boundary condition (2.1), (2.2), where

$$\phi(z)=\exp\frac{1}{2}(u(z)+i\theta(z))\quad\text{with } \theta(z)=2\sum_{j=1}^k n_j\arg(z-z_j),$$

and A is determined by the formula (3.1) (see [11]). Moreover, it is easily verified that, with the notation in (2.3), there holds $N=n_1+\cdots+n_k$.

Therefore, to find a solution to the prescribed vortex problem of (2.5+), it suffices to solve (3.2). In our discussion below, it is most convenient to regard (3.2) as an elliptic equation defined over a 2-torus and make no mention of the domain of the equation if there is no risk of confusion. A monotone iteration method will be adopted to construct the solution of (3.2).

The following result is useful for a background subtraction.

LEMMA 3.1. *For any smooth function f satisfying $\int f(x)\,dx=1$, there is a function u_0, which is smooth in the complement of the set $\{z_1,\cdots,z_k\}$, so that*

$$(3.3)\qquad \Delta u_0=-4\pi Nf+4\pi\sum_{j=1}^k n_j\delta(z-z_j),$$

and in a neighborhood of z_j, $u_0-\ln|z-z_j|^{2n_j}$ is smooth. Here $N=n_1+\cdots+n_k$.

For a proof of this lemma, see [3].

With the notation of Lemma 3.1, let $v=u-u_0$. From (3.2) and (3.3), it is seen that v satisfies

$$(3.4)\qquad \Delta v=e^{v+u_0}+(4\pi Nf-1).$$

Remark 3.1. Kazdan and Warner [12] have studied the (3.4)-type equations on a compact Riemannian manifold via constructing sub- and supersolutions. In their case f is a constant. In order to have larger flexibility in implementing numerical computations, we allow f to be an arbitrary function. It will be much easier to decide a solution u_0 of (3.3) when there is no restriction to f. Actually, we may just choose u_0 to be such that $u_0(z) = \ln |z - z_j|^{2n_j}$ in a neighborhood of $z = z_j, j = 1, \cdots, k$ and let $4\pi f(z) = -\Delta u_0(z) + 4\pi \sum n_j \delta(z - z_j)$. Our development below may be viewed as a specialization of that in [12].

LEMMA 3.2. *Assume $N < L_1L_2/\pi$. There are smooth functions $U \geqq V$ such that*

$$\Delta U - e^{U+u_0} - (4\pi Nf - 1) \leqq 0, \tag{3.5}$$

and

$$\Delta V - e^{V+u_0} - (4\pi Nf - 1) \geqq 0. \tag{3.6}$$

Proof. The property of u_0 implies that e^{u_0} is a smooth function and vanishes at $z_j, j = 1, \cdots, k$. It is not hard to see that

$$(\Delta - e^{u_0})U = 4\pi Nf - 1 \tag{3.7}$$

has a solution. In fact, $P \equiv \Delta - e^{u_0} - 1 : W^{2,2} \to L^2$ is bijective and $P^{-1} : L^2 \to L^2$ is compact. Hence $1 + P^{-1}$ is Fredholm of index zero due to the selfadjointness of P. Equation (3.7) may now be rewritten in the form $(1 + P^{-1})U = P^{-1}(4\pi Nf - 1)$. This equation has a solution because the only solution of $(1 + P^{-1})U = 0$ or $(\Delta - e^{u_0})U = 0$ is the trivial one $U = 0$ (the Fredholm alternatives).

Thus we have

$$\Delta U - e^{U+u_0} - (4\pi Nf - 1) \leqq \Delta U - e^{U+u_0} - (4\pi Nf - 1) + e^{u_0}(e^U - U) = 0,$$

which verifies (3.5). U is a supersolution of (3.4).

To get a subsolution of (3.4) that satisfies (3.6), we consider the equation

$$\Delta V = 4\pi Nf - \sigma, \tag{3.8}$$

where σ is a constant. In order to have a solution to (3.8), it is necessary and sufficient that the right-hand side of the equation is of zero integral mean [3]. This results in the condition

$$\sigma = 4\pi N/|\Omega| = \pi N/L_1L_2.$$

From the assumption in the lemma, we have $\sigma < 1$. There holds, due to (3.8), the equality

$$\Delta V - e^{V+u_0} - (4\pi Nf - 1) = ([1 - \sigma] - e^{V+u_0}). \tag{3.9}$$

Choose a solution V of (3.8) so that

$$\sup V \leqq \min \{\ln (1 - \sigma) - \sup u_0, 0\}.$$

Then it is seen clearly that the right-hand side of (3.9) is nonnegative. Hence (3.6) is proved.

Finally, the comparison $U \geqq V$ follows from the inequality $(\Delta - e^{u_0})(U - V) \leqq (\sigma - 1) < 0$ and the maximum principle.

LEMMA 3.3. *For $N < L_1L_2/\pi$, equation* (3.4) *has a unique solution v. This v satisfies the bounds $U \geqq v \geqq V$ and may be obtained in the limit*

$$\lim_{n \to \infty} v_n = v, \tag{3.10}$$

where $\{v_n\}$ is a monotone approximation sequence of v determined through the iterative scheme

$$\begin{aligned} &v_1 = V \quad \text{or} \quad U, \\ &(\Delta - K)v_n = e^{v_{n-1}+u_0} - Kv_{n-1} + (4\pi Nf - 1), \end{aligned} \tag{3.11}$$

with $K \geqq \sup e^{u_0+U}$. Here the limit in (3.10) *is in the space C^k for any positive integer k.*

Proof. Let us start with the assumption $v_1 = V$. We first show by induction that $v_n \leqq U$, $n = 1, 2, \cdots$. In fact, $v_1 \leqq U$ has been established in Lemma 3.2. Suppose $v_n \leqq U$ for some n. Then

$$(\Delta - K)(U - v_{n+1}) \leqq (e^{u_0+U} - K)(U - v_n) \leqq 0.$$

Thus $v_{n+1} \leqq U$ as well.

Next, we prove that

$$V = v_1 \leqq v_2 \leqq \cdots \leqq v_n \leqq \cdots \leqq U.$$

Indeed, from Lemma 3.2 we have $(\Delta - K)(v_2 - v_1) \leqq 0$. Consequently, $v_2 - v_1 \geqq 0$. By induction, if there holds $V = v_1 \leqq v_2 \leqq \cdots \leqq v_n$, then

$$\begin{aligned} (\Delta - K)(v_{n+1} - v_n) &= e^{u_0}(e^{v_n} - e^{v_{n-1}}) - K(v_n - v_{n-1}) \\ &\leqq (e^{u_0+U} - K)(v_n - v_{n-1}) \leqq 0, \end{aligned}$$

which implies $v_{n+1} - v_n \geqq 0$ as expected.

If $v_1 = U$, a similar argument shows that

$$U = v_1 \geqq v_2 \geqq \cdots \geqq v_n \geqq \cdots \geqq V.$$

Thus, in particular, we have proved that $\{v_n\}$ is bounded and convergent pointwise. By virtue of (3.11), it is seen that $\{v_n\}$ is convergent in the $W^{2,2}$ norm. A bootstrap argument then indicates that $\{v_n\}$ converges in any $W^{k,2}$ norm, hence, in any C^k norm. The limit (3.10) is the unique solution of (3.4).

The lemma is proved.

It is easy to observe that $\sigma < 1$ is also a necessary condition to ensure the solvability of (3.4). In fact, integrating both sides of (3.4), we have

$$0 = \int_\Omega e^{v+u_0}\, dx + 4\pi N - |\Omega|,$$

which implies $\sigma < 1$.

We can now state the following.

THEOREM 3.4. *For any $z_1, \cdots, z_k \in \Omega$ and positive integers $n_1, \cdots, n_k$, the system* (2.5+) *subject to the boundary condition* (2.1), (2.2) *has a smooth solution (ϕ, A) such that the zeros of ϕ are exactly $z_1, \cdots, z_k$ with respective multiplicities $n_1, \cdots, n_k$, if and only if $n_1 + \cdots + n_k = N < L_1L_2/\pi$. Moreover, up to gauge transformations, this solution is unique and can be obtained by the iteration scheme* (3.11).

Proof. The proof of existence has been obtained. To get the uniqueness part, let us assume (ϕ, A) and (ϕ', A') are two solutions of (2.5+) so that the zeros of ϕ and ϕ' are $z_1, \cdots, z_k$ with respective multiplicities $n_1, \cdots, n_k$. Consequently, $u = \ln |\phi|^2$ and $u' = \ln |\phi'|^2$ satisfy the same equation (3.2). This in turn implies $u = u'$ or $|\phi|^2 = |\phi'|^2$, namely, the difference between ϕ and ϕ' is a phase variable. Such a phase variable makes (ϕ, A) and (ϕ', A) gauge-equivalent.

Remark 3.2. If $N > 0$, we can apply the maximum principle to (3.2) in the complement of the set $\{z_1, \cdots, z_k\}$ to show that $u < 0$ everywhere, which corresponds to the property $|\phi|^2 < 1$ for a solution pair (ϕ, A) of (2.5+). In the special case where $N = 0$, the only solution of (3.2) is $u \equiv 0$, which corresponds to the trivial solution $\phi = 1$, $A = 0$ of (2.5+).

Remark 3.3. The Euler–Lagrange equations of the energy density $\mathscr{E}(\phi, A)$ in § 2 assume the form

$$\begin{aligned}\partial_k F_{jk} &= \frac{i}{2}(\phi[D_j\phi]^* - \phi^*[D_j\phi]),\\ D_k D_k \phi &= \frac{\lambda}{2}(|\phi|^2 - 1)\phi,\end{aligned} \tag{3.12}$$

which are known to be the two-dimensional Ginzburg–Landau equations. In $\mathbb{R}^2$, Jaffe and Taubes [11] showed that, when $\lambda = 1$, (2.5±) and (3.12) are equivalent systems for finite energy solutions. In our periodic case here, solutions of (2.5±) subject to (2.1), (2.2) are obviously solutions of (3.12) ($\lambda = 1$) under the same boundary condition. However, we do not know whether the two periodic systems are equivalent for $\lambda = 1$. Thus our condition for the existence of a multivortex solution only applies to the Bogomol'nyi equations (2.5±), but not to (3.12).

Remark 3.4. For a solution (ϕ, A) given in Theorem 3.4, the observables $|\phi|^2$ and F_{12} are both periodic on the lattice cell. Such a vortex structure confirms the prediction of Abrikosov on the magnetic response fashion of certain superconducting materials. Our study above has been restricted on a rectangular periodic cell. However, a general parallelogram domain does not render additional difficulties, and the same existence results hold. In this case, for an investigation of the more delicate existence problem of multivortices in the electroweak theory where the gauge group is $SU(2) \times U(1)$, see [20].

Remark 3.5. The vortex model defined by a holomorphic line bundle L over a compact Riemann surface M has been studied in the work of Noguchi [16]. It is concluded that, if the integer-valued first Chern class $c_1(L) = N \geqq 0$, then an N-vortex solution exists if and only if $N < \mathrm{Vol}\,(M)/4\pi$. Our condition in Theorem 3.4 may be interpreted in this topological spirit. The difference is that in Noguchi's solutions the gauge potentials can only be real in *local* unitary frames of L, while in our solutions, the gauge potentials are globally real, additional geometric complications are not raised, and a transparent physical meaning is contained.

4. The effect of an external field. From § 2, we see that the self-dual ($N > 0$) and anti-self-dual ($N < 0$) Bogomol'nyi vortices occupy the same energy level $E = \pi|N|$; therefore, there is a symmetry in the *vacuum* real world. The purpose of this section is to remark that such a symmetry can be broken by switching on an external field.

For simplicity, let us assume the external field is a constant magnetic field, denoted by F_{jk}^{ex}. Under the influence of such a field, the total (Gibbs) energy over the lattice cell Ω is written in the form

$$E(\phi, A) = \int_\Omega \mathscr{E}\, dx - \frac{1}{2}\int_\Omega F_{jk}F_{jk}^{\mathrm{ex}}\, dx.$$

The periodic boundary condition given in (2.1), (2.2) leads to the quantized flux (2.4). Consequently, the energy now becomes

$$E(\phi, A) = \pi(|N| - 2NF_{12}^{\mathrm{ex}}) + \frac{1}{2}\int_\Omega dx \left\{ \left| F_{12} \pm \frac{1}{2}(|\phi|^2 - 1)\right|^2 + |D_1\phi \pm iD_2\phi|^2 \right\}$$

for $N = \pm|N|$. The lower bound

$$E = \pi(|N| - 2NF_{12}^{\mathrm{ex}}) \tag{4.1}$$

is saturated by the solutions of the self-dual and anti-self-dual Bogomol'nyi equations.

First, from (4.1), we observe that, if the external field is sufficiently weak so that $|F_{12}^{\mathrm{ex}}|<\frac{1}{2}$, then the absolute energy minimizers can only have zero vortex number $N=0$. Theorem 3.4 implies that the Higgs field has no zero in Ω, and the solution is gauge-equivalent to the trivial (superconducting vacuum) solution $\phi=1$, $A=0$. Thus there is no field penetration into the lattice cell, and the magnetic screening is complete.

Next, if $|F_{12}^{\mathrm{ex}}|=\frac{1}{2}$, the energy minimum is still $E=0$, which may be attained at the vacuum solutions as well as at the N-vortex solutions satisfying $\operatorname{sgn} N=\operatorname{sgn} F_{12}^{\mathrm{ex}}$. This situation describes a transition phase between the superconducting vacuum and the onset of a coexisting normal-superconducting phase.

Finally, as the external field goes beyond the critical threshold $|F_{12}^{\mathrm{ex}}|=\frac{1}{2}$, namely, $|F_{12}^{\mathrm{ex}}|>\frac{1}{2}$, the lowest energy level is occupied by those N-vortex solutions that satisfy $\operatorname{sgn} N=\operatorname{sgn} F_{12}^{\mathrm{ex}}$ and $|N|=\max\{n\in\mathbb{Z}_+|n<L_1L_2/\pi\}$. In other words, the orientation of the vortices depends on the directon of the external field, and the vortex number should be as large as possible to achieve a maximal magnetic penetration. This is a kind of vortex-line orientation selection phenomenon under the influence of an external field, and the vacuum symmetry is thus broken.

The above discussion illustrates the celebrated Meissner effect in superconductivity theory.

5. Arbitrary number of vortices in a bounded domain. From Theorem 3.4 we see that, under the 't Hooft type periodic boundary condition given in § 2, the total magnetic flux is proportional to the number of vortices living in the lattice cell, and this number is confined by the size of the domain. In this section, we observe that, if the periodic boundary condition is removed, such a restriction will no longer exist. In other words, a bounded region may allow an arbitrary number of vortices. It is conceivable that, at the same time, we will lose the flux quantization property.

Let us consider a boundary value problem of the Bogomol'nyi system. It is natural to impose a boundary condition on the observables $|\phi|^2$ and F_{12}. Thus, for the self-dual Bogomol'nyi equations, we have the boundary value problem

$$\begin{aligned}&D_1\phi+iD_2\phi=0,\\ &F_{12}+\tfrac{1}{2}(|\phi|^2-1)=0,\qquad x\in\Omega;\\ &|\phi|^2=|\phi^0|^2,\qquad F_{12}=F_{12}^0,\quad x\in\partial\Omega,\end{aligned}\tag{5.1}$$

where $\partial\Omega$ is assumed to be sufficiently regular (Lipschitzian, say).

The equation $(5.1)_2$ implies the compatibility condition $F_{12}^0=(1-|\phi^0|^2)/2$ on $\partial\Omega$ for the boundary data. Moreover, physically, it is natural to assume the inequality

$$0<|\phi^0|\leqq 1.\tag{5.2}$$

Since the treatment of (5.1) is similar to (2.5+) and a special case of (5.1) has been studied in [24], our discussion below will be brief.

Let $z_1,\cdots,z_k$ be the zeros of ϕ (or the vortex locations of a solution (ϕ, A) of (5.1)) with respective multiplicities $n_1,\cdots,n_k\in\mathbb{Z}_+$. Then (5.1) is reduced after the substitution $u=\ln|\phi|^2$, $\bar{u}=\ln|\phi^0|^2$ into the problem

$$\begin{aligned}&\Delta u=e^u-1+4\pi\sum_{j=1}^k n_j\delta(z-z_j),\qquad x\in\Omega,\\ &u=\bar{u},\qquad x\in\partial\Omega.\end{aligned}\tag{5.3}$$

To solve (5.3), we borrow the background subtraction function u_0 from [11]:

$$u_0(z)=-\sum_{j=1}^k n_j\ln(1+|z-z_j|^{-2}).\tag{5.4}$$

Thus $v \equiv u - u_0$ will solve the modified problem

$$\begin{aligned} \Delta v &= e^{v+u_0} + (g-1), \qquad x \in \Omega, \\ v &= \bar{u} - u_0, \qquad x \in \partial\Omega, \end{aligned} \tag{5.5}$$

where

$$g(z) = 4 \sum_{j=1}^{k} n_j (1 + |z - z_j|^2)^{-2}.$$

It is easy to verify that $e^{u_0} \leqq 1$ and is smooth. The following lemma gives us a pair of super- and subsolution of (5.5) as for the problem (3.4).

LEMMA 5.1. *There exist functions $U \geqq V$ such that*

$$\Delta U - e^{U+u_0} - (g-1) \leqq 0, \qquad x \in \Omega, \tag{5.6}$$

$$\Delta V - e^{V+u_0} - (g-1) \geqq 0, \qquad x \in \Omega, \tag{5.7}$$

and $U = V = \bar{u} - u_0$, $x \in \partial\Omega$.

Proof. Let U and V be the solutions of the linear equations $\Delta U = g - 1$ and $\Delta V = g$ $(x \in \Omega)$, respectively, with $U = V = \bar{u} - u_0$ on $\partial\Omega$. Then it is self-evident that U satisfies (5.6). To see that V fulfills (5.7), it suffices to achieve the inequality $V + u_0 \leqq 0$ in $\Omega - \{z_1, \cdots, z_k\}$. First, we have $V + u_0 = \bar{u} \leqq 0$ for $x \in \partial\Omega$ due to (5.2). On the other hand, we may choose $\varepsilon > 0$ sufficiently small to make $B_\varepsilon(z_j) = \{z \mid |z - z_j| < \varepsilon\} \subset \Omega$, $j = 1, \cdots, k$, and $V + u_0 \leqq 0$ on $\partial B_\varepsilon(z_j)$, $j = 1, \cdots, k$. Since $\Delta(V + u_0) = 0$ in $\Omega_\varepsilon \equiv \Omega - \bigcup_{j=1}^{k} B_\varepsilon(z_j)$, we have, in Ω_ε, $V + u_0 \leqq 0$, by virtue of the maximum principle. Finally, letting $\varepsilon \to 0$, we reach the desired conclusion.

Using Lemma 5.1, we can find the unique solution of (5.5) in the limit $\lim_{n\to\infty} v_n = v$ as in § 3, where $\{v_n\}$ is constructed through the scheme

$$\begin{aligned} v_1 &= V \quad \text{or} \quad U, \\ (\Delta - K)v_n &= e^{v_{n-1}+u_0} - Kv_{n-1} + (g-1), \qquad x \in \Omega, \\ v_n &= \bar{u} - u_0, \qquad x \in \partial\Omega, \quad n = 2, 3, \cdots, \end{aligned} \tag{5.8}$$

whereas $K \geqq e^M$, $M = \sup_\Omega U$. Such a solution gives rise to a multivortex solution of the problem (5.1) with prescribed vortex locations $z_1, \cdots, z_k$ and local charges $n_1, \cdots, n_k$ through the construction mentioned in § 3 with $u = u_0 + v$.

In the rest of this section, we present two numerical examples of the multivortex solutions of the Bogomol'nyi system (5.1), computed using the iterative scheme (5.8). We shall see from these solutions that the flux may not be quantized as in the periodic case (see (2.4)). Rather, it depends on the locations of vortices.

In our computations shown below, we choose the domain Ω to be a lattice square: $\Omega = (-3, 3) \times (-3, 3)$. According to the conclusion of Theorem 3.4, such a region cannot allow more than two vortices if the t' Hooft type periodic boundary condition introduced in § 2 is imposed. On the other hand, from the discussion of this section, we see that if the boundary data are prescribed to the observables $|\phi|^2$ and F_{12}, then the domain may accommodate as many vortices as one pleases. We shall specify $|\phi^0| = 1$ and $F_{12}^0 = 0$ for simplicity. Such a condition represents a complete boundary magnetic screening. The numerical implementation can be described as follows.

First of all, the interval $(-3, 3)$ is discretized with 150 equidistant grid points, which results in a finite difference mesh for the square domain Ω. The scheme (5.8) is then solved through the standard five-point approximation algorithm for the boundary value problems of elliptic differential equations. As usual, the discrete approximation to v_n in (5.8), and so on, at the mesh point $(x_1(i), x_2(j))$ will be denoted by $v^n_{i,j}$. The constant K in (5.8) is chosen to be $K = 1 + e^M$ with $M = \sup U_{i,j}$. The initial $v^1_{i,j}$ is taken to be $V_{i,j}$. Here U and V are the unique solutions of the equations $\Delta U = g - 1$ and $\Delta V = g(x \in \Omega)$ with $U = V = -u_0$ on $\partial\Omega$ as in the proof of Lemma 5.1. The termination criterion of the iterative scheme (5.8) is set to be

$$|v^n - v^{n-1}| \equiv \max_{i,j} |v^n_{i,j} - v^{n-1}_{i,j}| < 10^{-3}. \tag{5.9}$$

If the accuracy (5.9) is attained at a certain step $n = k$, then the computation will halt and $v^k_{i,j}$ will be recognized as an approximation of the unique solution of (5.5) at the mesh points. Thus, a numerical solution of the boundary value problem (5.1) of the Bogomol'nyi equations is obtained.

Figures 1 and 2 present two numerical solutions with total vortex number $N = 4$.

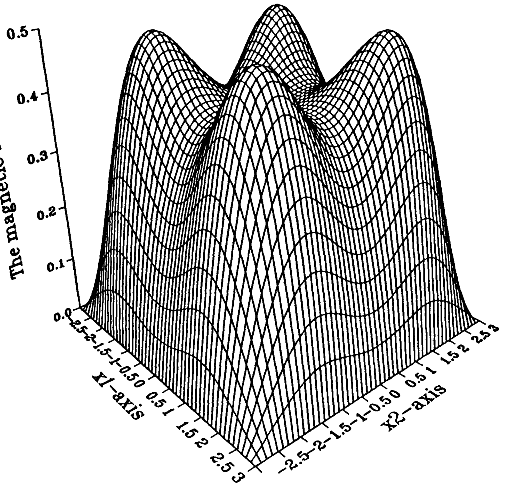


FIG. 1. *A solution of four separated vortices.*

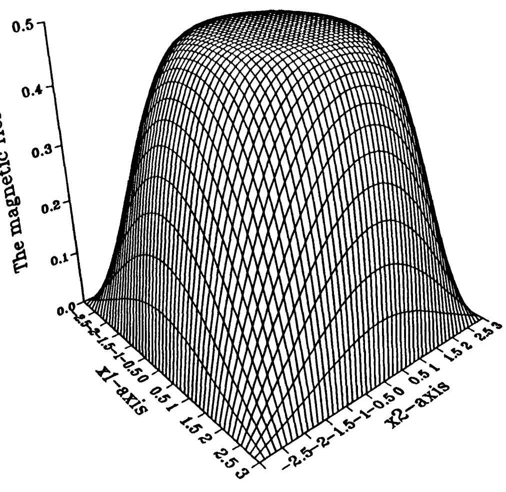


FIG. 2. *A solution of four clustered vortices.*

Figure 1 gives us the behavior of the magnetic field over Ω of a computed solution of four separated vortices located at the points $z_1 = -1.5 - 1.5i$, $z_2 = 1.5 - 1.5i$, $z_3 = 1.5 + 1.5i$, and $z_4 = -1.5 + 1.5i$, with local vortex numbers or charges satisfying $n_1 = n_2 = n_3 = n_4 = 1$. The numerical solution is obtained at $n = 27$ (27 iterations). The magnetic penetration attains its maximal value $F_{12} = 0.5$ at the centers of the vortices as expected. The total normalized flux is $\Phi/2\pi = \int_\Omega F_{12}\, dx/2\pi = 1.774$, which cannot be an integer up to numerical errors.

Figure 2 illustrates the magnetic field distribution in Ω of a numerical solution of four clustered vortices centered at $z = 0$, with the local charge $= N = 4$. In this case, the program takes 31 iterations to achieve the accuracy (5.9). The shape of the graph of F_{12} indicates that the magnetic penetration in a neighborhood of the center of these clustered vortices now gains a relatively larger average value than the separated ones. The total normalized flux for this solution is $\Phi/2\pi = 1.878$, which is greater than that of the solution of separated vortices given in Fig. 1.

Further computer experiments show that the flux depends sensitively on the locations of vortices. For example, a pair of clustered vortices centered at $z_1 = -1.5 - 1.5i$ and $z_2 = 1.5 + 1.5i$ with $n_1 = n_2 = 2$ (another four-vortex solution of (5.1)) yield a normalized flux $\Phi/2\pi = 1.580$.

Remark 5.1. If we choose

$$u_0(z) = \sum_{j=1}^{k} n_j \ln(|z - z_j|^2),$$

then $g = 0$ in (5.5), which gives us a system with a much simpler inhomogeneous term. However, in this case we lose the bound $e^{u_0} \leqq 1$. In fact, for large domains, $\sup_{\bar\Omega} e^{u_0}$ may take large values, and this in turn requires K be sufficiently large to ensure the convergence of the iterative scheme (5.8). Numerical experiments carried out on the square domain just specified indicate that the choice of u_0 according to the above simpler expression results in as much as four times the computing time needed with the choice of u_0 according to (5.4).

6. Approximating a vortex solution in the full plane. In this section, we prove that the bounded domain solutions obtained in § 5 may be used to approximate a vortex solution of the Bogomol'nyi equations on full $\mathbb{R}^2$.

Let (ϕ, A) be a finite energy solution of (2.5+) on $\mathbb{R}^2$ so that $Z(\phi) = \{z_1, \cdots, z_k\}$, the multiplicity of the zero $z = z_j$ of ϕ is $n_j \geqq 1$, $j = 1, \cdots, k$, and $n_1 + \cdots + n_k = N$. The existence and uniqueness of such a solution has been established in Jaffe and Taubes [11]. Moreover, there holds the exponential decay estimate [11]:

$$(6.1) \qquad 0 \leqq 1 - |\phi(z)|^2 \leqq C(\varepsilon) e^{-(1-\varepsilon)|z|}, \qquad z \in \mathbb{R}^2,$$

where $\varepsilon \in (0, 1)$. The function $u = \ln|\phi|^2$ is the unique solution of the equation

$$(6.2) \qquad \Delta u = e^u - 1 + 4\pi \sum_{j=1}^{k} n_j \delta(z - z_j) \quad \text{in } \mathbb{R}^2,$$

which vanishes at infinity [11].

Suppose Ω_0 is a bounded domain in $\mathbb{R}^2$ so that $Z(\phi) \subset \Omega_0$. Choose $u' \in C^\infty(\mathbb{R}^2)$ satisfying $u' = u$ in $\mathbb{R}^2 - \Omega_0$. From (6.1) and using the simple inequality $\ln(1-s) > -2s$ ($s \in (0, \frac{1}{2})$), we have

$$(6.3) \qquad -2C(\varepsilon)\, e^{-(1-\varepsilon)|z|} \leqq u(z) \leqq 0 \quad \text{for } |z| > r_\varepsilon \equiv \frac{\ln(2C(\varepsilon)+1)}{1-\varepsilon}.$$

Let $\{\Omega_n\}$: $\Omega_1 \subset \Omega_2 \subset \cdots \subset \Omega_n \subset \cdots$ be a monotone chain of bounded convex domains whose boundaries are sufficiently regular (piecewise smooth and Lipschitzian, say) and $\bigcup \Omega_n = \mathbb{R}^2$. Denote by u_n the unique solution of the boundary value problem

$$\begin{aligned} &\Delta u = e^u - 1 + 4\pi \sum_{j=1}^{k} n_j \delta(z - z_j) \quad \text{in } \Omega_n, \\ &u = 0 \quad \text{on } \partial\Omega_n, \end{aligned} \tag{6.4}$$

$n = 1, 2, \cdots$. We shall show that $u_n \to u$ as $n \to \infty$.

LEMMA 6.1. *Given $m = 1, 2, \cdots$ and $\{u_n\}_{n \geqq m}$, there holds*

$$u \leqq \cdots \leqq u_n \leqq \cdots \leqq u_m \leqq 0 \quad in\ \Omega_m.$$

Proof. It is easily observed that $u_n \leqq 0$ in Ω_n for each $n = 1, 2, \cdots$. In fact, since u_n behaves likes $n_j \ln |z - z_j|^2$ in a neighborhood of $z = z_j$, we can find an $\eta_0 > 0$ such that for any $0 < \eta < \eta_0$ we have $B_\eta(z_j) \subset \Omega_n$ and $u_n < 0$ on $\partial B_\eta(z_j)$, $j = 1, \cdots, k$. On the other hand, in $\Omega_\eta = \Omega_n - \bigcup_{j=1}^k B_\eta(z_j)$, $\Delta u_n = e^{u_n} - 1$. Hence the condition $u_n = 0$ on $\partial\Omega_n$ and the maximum principle allow us to conclude that $u_n \leqq 0$ in Ω_η, $\forall \eta < \eta_0$.

Next, we show that $u_{n+1} \leqq u_n$ in Ω_n, $n = 1, 2, \cdots$. To see this, we examine the relation $\Delta(u_{n+1} - u_n) = e^{u_{n+1}} - e^{u_n} = e^{\xi(u_{n+1}, u_n)}(u_{n+1} - u_n)$ in Ω_n. Since $(u_{n+1} - u_n)|_{\partial\Omega_n} = u_{n+1}|_{\partial\Omega_n} \leqq 0$, we conclude again from the maximum principle that $u_{n+1} \leqq u_n$ in Ω_n.

The inequality $u \leqq u_n$ in Ω_n, $n = 1, 2, \cdots$ is contained in the above proof. Thus the lemma follows.

LEMMA 6.2. *There holds the bound*

$$\|u_n - u\|_{W^{2,2}(\Omega_n)} \leqq M,$$

where $M > 0$ is a constant independent of $n = 1, 2, \cdots$.

Proof. Since Ω_n is convex, using the standard L^2-estimate in the equation

$$\begin{aligned} &\Delta(u_n - u) = e^{u_n} - e^u \quad \text{in } \Omega_n, \\ &(u_n - u) = -u' \quad \text{on } \partial\Omega_n, \end{aligned} \tag{6.5}$$

we have

$$\|u_n - u\|_{W^{2,2}(\Omega_n)} \leqq C(\|e^{u_n} - e^u\|_{L^2(\Omega_n)} + \|u'\|_{W^{2,2}(\Omega_n)} + \|u_n - u\|_{L^2(\Omega_n)}),$$

where $C > 0$ does not depend on Ω_n.

First of all, since $u' = u$ outside Ω_0, we see from (6.2), (6.3) that $u' \in W^{2,2}(\mathbb{R}^2)$. On the other hand, by virtue of Lemma 6.1, there holds $|u_n - u| \leqq |u|$. So it follows from (6.3) that $\sup_n \|u_n - u\|_{L^2(\Omega_n)}$ is finite. Finally, again from Lemma 6.1, $|e^{u_n} - e^u| \leqq |1 - e^u| = 1 - |\phi|^2$. Hence (6.1) implies that $\sup_n \|e^{u_n} - e^u\|_{L^2(\Omega_n)}$ is finite as well. This proves the lemma.

We are now ready to establish the expected convergence result. For convenience, we understand that $u_n = 0$ in $\mathbb{R}^2 - \Omega_n$. For a function f decaying sufficiently fast at infinity, we define the norm

$$|f|_\mu = \sup_{z \in \mathbb{R}^2} |e^{\mu|z|} f(z)|.$$

LEMMA 6.3. *Given $0 < \mu < 1$, there holds the limit*

$$\lim_{n \to \infty} |u_n - u|_\mu = 0. \tag{6.6}$$

Proof. From Lemma 6.2, we see that, in particular, for $m = 1, 2, \cdots$,

$$\|u_n - u\|_{W^{2,2}(\Omega_m)} \leqq M, \qquad n \geqq m. \tag{6.7}$$

Therefore $u_n - u \to$ some $w_m \in W^{2,2}(\Omega_m)$ weakly and $w_m = w_{m'}$ in Ω_m for $m' \geqq m$ due to Lemma 6.1. Set

$$w = w_m, \qquad z \in \Omega_m, \quad m = 1, 2, \cdots.$$

Then w is well defined in $\mathbb{R}^2$. We can obtain from (6.7) that $w \in W^{2,2}(\mathbb{R}^2)$. This fact implies that w decays to zero at infinity. Moreover, a simple argument applied to (6.5) leads us to conclude that w is smooth and verifies

$$\Delta w = e^{u+w} - e^u.$$

In other words, $u + w$ is also a solution of (6.2). By the uniqueness theorem in [11], we must have $w \equiv 0$.

Let $\varepsilon > 0$ be so small that $\varepsilon < 1 - \mu$. We have by (6.3) and Lemma 6.1

$$\begin{aligned} e^{\mu|z|}|u_n(z) - u(z)| &\leqq e^{\mu|z|}|u(z)| \\ &\leqq 2C(\varepsilon)\, e^{-([1-\varepsilon]-\mu)|z|}, \qquad |z| > r_\varepsilon. \end{aligned} \tag{6.8}$$

On the other hand, since $\cup\Omega_n = \mathbb{R}^2$, for any given $r > 0$ there exists $m \geqq 1$ so that $B_r = \{z \in \mathbb{R}^2 \,|\, |z| \leqq r\} \subset \Omega_m$. We have shown that $u_n \to u$ weakly in $W^{2,2}(\Omega_m)$. From the compact embedding $W^{2,2}(\Omega_m) \to C^0(\bar{\Omega}_m)$, we may conclude that $u_n \to u$ uniformly on B_r. Combining this observation with (6.8) we arrive at the expected limit (6.6).

In particular, $u_n \to u$ uniformly in $\mathbb{R}^2$.

From the sequence $\{u_n\}$ we can construct as in § 3 the corresponding solution pairs $\{(\phi^{(n)}, A^{(n)})\}$ of the Bogomol'nyi equations (2.5+) on Ω_n. It can be shown that $(\phi^{(n)}, A^{(n)}) \to (\phi, A)$ (in a suitable sense). However, since the physically interesting fields are $|\phi|^2$ and F_{12}, here we only discuss the convergence $|\phi^{(n)}|^2 \to |\phi|^2$, $F_{12}^{(n)} = \partial_1 A_2^{(n)} - \partial_2 A_1^{(n)} \to F_{12}$ in detail.

Using Lemma 6.1, we see easily that

$$\begin{aligned} 0 \leqq |\phi^{(n)}|^2 - |\phi|^2 &= e^{u_n} - e^u \\ &\leqq u_n - u. \end{aligned} \tag{6.9}$$

While from $(2.5+)_2$, which relates $\phi^{(n)}$ and ϕ to $F_{12}^{(n)}$ and F_{12}, we get

$$F_{12}^{(n)} - F_{12} = -\tfrac{1}{2}(|\phi^{(n)}|^2 - |\phi|^2). \tag{6.10}$$

As a consequence of (6.9), (6.10), we conclude that $(|\phi^{(n)}|^2, F_{12}^{(n)}) \to (|\phi|^2, F_{12})$ (as $n \to \infty$) in the same sense as the convergence for $u_n \to u$ stated in Lemma 6.3.

We summarize the results we have just obtained as follows.

THEOREM 6.4. *Let* $(\phi^{(n)}, A^{(n)})$ *be the solution of the Bogomol'yni system* (2.5+) *on* Ω_n, *obtained from the unique solution* u_n *of* (6.4), *and* (ϕ, A) *the finite energy solution of* (2.5+) *on full* $\mathbb{R}^2$ *with the same vortex distribution. Then*

$$1 \geqq |\phi^{(1)}|^2 \geqq \cdots \geqq |\phi^{(n)}|^2 \geqq \cdots \geqq |\phi|^2,$$

$$F_{12} \geqq \cdots \geqq F_{12}^{(n)} \geqq \cdots \geqq F_{12}^{(1)},$$

$||\phi^{(n)}|^2 - |\phi|^2|_\mu \to 0$, $|F_{12}^{(n)} - F_{12}|_\mu \to 0$, *and*

$$\Phi^{(n)} = \int_{\Omega_n} F_{12}^{(n)}\, dx \to \Phi = \int_{\mathbb{R}^2} F_{12}\, dx = 2\pi N$$

(*as* $n \to \infty$). *In other words, the total flux of the vortices confined in a bounded domain approaches the quantized flux of the vortex solution on* $\mathbb{R}^2$ *with larger domains giving successively better approximations.*

7. A family of infinite energy solutions. It has been shown in Jaffe and Taubes [11] that, for any vortex distribution in $\mathbb{R}^2$, the Bogomol'nyi system (2.5+) has a unique finite energy solution. The purpose of this section is to note that, when the finite energy condition is removed, such a uniqueness will no longer hold. More precisely, we state and prove the following.

THEOREM 7.1. *For any* $\{z_1, \cdots, z_k\} \subset \mathbb{R}^2$, $n_1, \cdots, n_k \in \mathbb{Z}_+$, *the Bogomol'nyi system (2.5+) has a family of gauge-distinct infinite energy solutions* $\{(\phi^{(\alpha)}, A^{(\alpha)})\}_{0<\alpha<\infty}$ *so that* $Z(\phi^{(\alpha)}) = \{z_1, \cdots, z_k\}$, *the multiplicity of the zero* $z = z_j$ *of* $\phi^{(\alpha)}$ *is* n_j, $j = 1, 2, \cdots$, *and there hold*

$$C_1\, e^{-[1/4]|z|^2}|z|^{2N+\alpha} \leqq |\phi^{(\alpha)}(z)|^2 \leqq C_2\, e^{-[1/4]|z|^2}|z|^{2N+\alpha} \quad \textit{for large } |z|, \tag{7.1}$$

where $C_1, C_2 > 0$ *are constants, which may depend on* α, *and*

$$\int_{\mathbb{R}^2} |\phi^{(\alpha)}|^2\, dx = 2\pi\alpha. \tag{7.2}$$

Proof. Choose a function $w \in C^\infty(\mathbb{R}^2)$ that verifies the property $w(z) = \ln|z|$ when $|z| \geqq 1$. Thus $f = \Delta w$ is of compact support, and

$$\int_{\mathbb{R}^2} f\, dx = \int_{|x|\leqq 2} f\, dx = \int_{|x|\leqq 2} \Delta w\, dx = \int_{|x|=2} \frac{\partial w}{\partial r}\, ds = 2\pi.$$

Consider (6.2). Introduce as before the background subtraction function

$$u_0(z) = \sum_{j=1}^{k} \ln|z - z_j|^{2n_j} + \alpha w(z) - \tfrac{1}{4}|z|^2.$$

The substitution $v = u - u_0$ leads us from (6.2) to the modified equation

$$\Delta v = K(z)\, e^v - \alpha f, \tag{7.3}$$

where

$$0 \leqq K(z) = e^{u_0(z)} = O(e^{-|z|}) \quad \text{for } |z| \to \infty.$$

Equation (7.3) has been well studied in the elegant works of Ni [15] and McOwen [13]. In particular, McOwen showed that, for any α: $0 < \alpha < \infty$, (7.3) has a smooth solution $v^{(\alpha)}$, which approaches a constant at infinity and

$$\int_{\mathbb{R}^2} K\, e^{v^{(\alpha)}}\, dx = \alpha \int_{\mathbb{R}^2} f\, dx = 2\pi\alpha.$$

From the solution $u^{(\alpha)} = u_0 + v^{(\alpha)}$ of (6.2), we can construct as in Jaffe and Taubes [11] or as in § 3 the solution pair $(\phi^{(\alpha)}, A^{(\alpha)})$ of the Bogomol'nyi equations (2.5+) so that $|\phi^{(\alpha)}|^2 = e^{u^{(\alpha)}} = e^{u_0 + v^{(\alpha)}} = K\, e^{v^{(\alpha)}}$. As a consequence, we see immediately that $|\phi^{(\alpha)}|^2$ verifies the desired decay estimate (7.1) and the space-average (7.2). Since (7.2) is invariant under the $U(1)$ gauge transformation

$$\phi \mapsto \phi\, e^{i\xi}, \qquad A \mapsto A + \nabla\xi,$$

different values of α give rise to gauge-distinct solutions. These solutions are necessarily all of infinite energy.

Remark 7.1. If the underlying domain $\mathbb{R}^2$ is replaced by an asymptotically Euclidean Riemann surface, the solutions of the curved space version of the Bogomol'nyi equations are superconducting vortices in a shell geometry or cosmic strings [25]. It is possible to modify the argument used in this section to establish a similar class of infinite energy solutions there.

8. Conclusions. In this paper we have shown that, on a lattice cell $\Omega = (-L_1, L_1) \times (-L_2, L_2)$, when a t' Hooft type periodic boundary condition is imposed, the Bogomol'nyi system arising in the abelian Higgs theory allows an arbitrarily distributed N-vortex solution, if and only if $|N| < L_1 L_2/\pi$. Such a solution exhibits the periodic structure of Abrikosov's mixed state vortices in a superconductor so that the magnetic flux can only take a quantized spectrum of values. On the other hand, when the periodic boundary condition is replaced by a boundary value condition on the observables $|\phi|^2$ and F_{12}, the system possesses a solution for any prescribed vortex locations and given total vortex number. In this case, numerical examples have shown that the flux is no longer quantized as before, but depends sensitively on the locations of vortex-lines. Such solutions can be used to approximate a finite energy full plane vortex solution in a global way, and in the large domain limit the flux approaches the quantized value related to the total vortex number. Moreover, for any prescribed vortex distribution, the Bogomol'nyi system on $\mathbb{R}^2$ has a continuous family of gauge-distinct solutions of infinite energy so that $|\phi|^2$ decays to zero exponentially fast at infinity.

REFERENCES

[1] A. A. ABRIKOSOV, *On the magnetic properties of superconductors of the second group*, Soviet Phys. JETP, 5 (1957), pp. 1174-1182.

[2] M. F. ATIYAH, N. J. HITCHIN, V. G. DRINFELD, AND YU. I. MANIN, *Construction of instantons*, Phys. Lett. A, 65 (1978), pp. 185-187.

[3] T. AUBIN, *Nonlinear Analysis on Manifolds: Monge-Ampére Equations*, Springer-Verlag, New York, Berlin, 1982.

[4] M. S. BERGER AND Y. Y. CHEN, *Symmetric vortices for the Ginzburg-Landau equations and the nonlinear desingularization phenomenon*, J. Funct. Anal., 82 (1989), pp. 259-295.

[5] E. B. BOGOMOL'NYI, *The stability of classical solutions*, Soviet J. Nuclear Phys., 24 (1976), pp. 449-454.

[6] P. H. DAMGAARD AND U. M. HELLER, *Observations of the Meissner effect in the lattice Higgs model*, Phys. Rev. Lett., 60 (1988), pp. 1246-1249.

[7] T. A. DEGRAND AND D. TOUSSAINT, *Topological excitations and Monte Carlo simulation of abelian gauge theory*, Phys. Rev. D, 22 (1980), pp. 2478-2489.

[8] V. L. GINZBURG AND L. D. LANDAU, *On the theory of superconductivity*, in Collected Papers of L. D. Landau, D. ter Haar, ed., Pergamon, New York, 1965, pp. 546-568.

[9] V. GRÖSCH, K. JANSEN, J. JERSÁK, C. B. LANG, T. NEUHAUS, AND C. REBBI, *Monopoles and Dirac sheets in compact $U(1)$ lattice gauge theory*, Phys. Lett. B., 162 (1985), pp. 171-175.

[10] L. JACOBS AND C. REBBI, *Interaction energy of superconducting vortices*, Phys. Rev. B, 19 (1979), pp. 4486-4494.

[11] A. JAFFE AND C. H. TAUBES, *Vortices and Monopoles*, Birkhäuser, Boston, 1980.

[12] J. L. KAZDAN AND F. W. WARNER, *Curvature functions for compact 2-manifolds*, Ann. of Math., 99 (1974), pp. 14-47.

[13] R. C. MCOWEN, *On the equation $\Delta u + Ke^{2u} = f$ and prescribed negative curvature in $\mathbb{R}^2$*, J. Math. Anal. Appl., 103 (1984), pp. 365-370.

[14] K. J. M. MORIARTY, E. MYERS, AND C. REBBI, *Dynamical interactions of superconducting flux vortices*, J. Comput. Phys., 81 (1989), pp. 481-488.

[15] W.-M. NI, *On the elliptic equation $\Delta u + K(x)\, e^{2u} = 0$ and conformal metrics with prescribed Gaussian curvature*, Invent. Math., 66 (1982), pp. 343-352.

[16] M. NOGUCHI, *Yang-Mills-Higgs theory on a compact Riemann surface*, J. Math. Phys., 28 (1987), pp. 2343-2346.

[17] F. ODEH, *Existence and bifurcation theorems for the Ginzburg-Landau equations*, J. Math. Phys., 8 (1967), pp. 2351-2356.

[18] B. J. PLOHR, Ph.D. thesis, Princeton University, Princeton, NJ, 1980.

[19] M. K. PRASAD AND C. M. SOMMERFIELD, *Exact classical solution for the 't Hooft monopole and the Julia-Zee dyon*, Phys. Rev. Lett., 35 (1975), pp. 760-762.

[20] J. SPRUCK AND Y. YANG, *On multivortices in the electroweak theory.* I: *existence of periodic solutions*, Comm. Math. Phys., 144 (1992), pp. 1–16.

[21] G. 'T HOOFT, *A property of electric and magnetic flux in nonabelian gauge theories*, Nuclear Phys. B, 153 (1979), pp. 141–160.

[22] E. J. WEINBERG, *Multivortex solutions of the Ginzburg–Landau equations*, Phys. Rev. D, 19 (1979), pp. 3008–3012.

[23] E. WITTEN, *Some exact multipseudoparticle solutions of the classical Yang–Mills theory*, Phys. Rev. Lett., 38 (1977), pp. 121–124.

[24] Y. YANG, *Boundary value problems of the Ginzburg–Landau equations*, Proc. Roy. Soc. Edinburgh, Sect. A 114 (1990), pp. 355–365.

[25] ———, *Vortices on asymptotically Euclidean Riemann surfaces*, Nonlinear Anal., 15 (1990), pp. 577–596.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1141-1166, September 1992

EXISTENCE AND NONEXISTENCE OF SOLITARY WAVE SOLUTIONS TO HIGHER-ORDER MODEL EVOLUTION EQUATIONS*

SATYANAD KICHENASSAMY† AND PETER J. OLVER†‡

Abstract. The problem of existence of solitary wave solutions to some higher-order model evolution equations arising from water wave theory is discussed. A simple direct method for finding monotone solitary wave solutions is introduced, and by exhibiting explicit necessary and sufficient conditions, it is illustrated that a model admit exact sech^2 solitary wave solutions. Moreover, it is proven that the only fifth-order perturbations of the Korteweg–deVries equation that admit solitary wave solutions reducing to the usual one-soliton solutions in the limit are those admitting families of explicit sech^2 solutions.

Key words. solitary wave, nonlinear evolution equation, water waves, singular perturbation

AMS(MOS) subject classifications. 76B25, 35Q51, 35Q53, 35B25, 76B15

1. Introduction. In the study of equations modeling wave phenomena, one of the fundamental objects of study is the traveling wave solution, meaning a solution of constant form moving with a fixed velocity. The determination of such solutions is accomplished by solving a reduced differential equation in fewer independent variables by one. In particular, the traveling wave solutions for a one-dimensional wave equation are found by solving a connection problem for an associated ordinary differential equation. Of particular interest are three types of traveling waves: the *solitary waves*, which are localized traveling waves, asymptotically zero at large distances, the *periodic waves*, and the *kink waves*, which rise or descend from one asymptotic state to another. All of these are, in the completely integrable case, solitons, coming from the inverse scattering solution to an eigenvalue problem, and dependent on a free parameter. On the other hand, the existence of these types of solutions is not dependent on integrability of the model, or the connection with an inverse scattering transform method of solution, as evidenced by the φ^4 theory; cf. [37], [38], and the examples described here. Except in the simplest instances, it is by no means obvious that such types of traveling wave solutions even exist. In addition, once existence is known, the delicacy of the connection problem to be solved makes their numerical computation rather difficult to effect in an easy, practical manner.

In this paper, we concentrate on the determination of solitary waves, whose importance for fluids came to the forefront with Scott Russell's experimental observation of solitary water waves in the Edinburgh canal [33]. Airy's premature dismissal of these solutions based on a linearized analysis of the free boundary problem necessitated the construction of suitable models exhibiting such solutions. This was accomplished, in the case of long waves over shallow water, through Boussinesq's bidirectional models and, subsequently, the celebrated Korteweg–deVries model, whose solitary wave solutions are explicit sech^2 solutions, which, moreover, have the remarkable soliton property of interacting without change of form. More recently, Amick and Toland, [4], and others, [1], [2], [19], have proved the existence of such waves for the full water wave problem. For small amplitude waves, the Korteweg–deVries solitons do a good job of modeling solitary water waves, [13]. However, the model fails to replicate such important physical phenomena as having a wave of maximal height,

* Received by the editors April 16, 1991; accepted for publication (in revised form) January 17, 1992.

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

‡ The research of this author was supported in part by National Science Foundation grant DMS-89-01600.

originally conjectured by Stokes (cf. [1]) and the breaking of large amplitude waves. Owing to the difficulty of analyzing the water wave problem directly, the construction of suitable models is of great importance. One possible approach is to retain higher-order terms in the Boussinesq perturbation expansion, leading to fifth-order model evolution equations. One of the principal purposes of this paper is to show that there are definite difficulties with this procedure, in that for most of these higher-order models, solitary wave solutions of the appropriate form do not even exist! Indeed, this holds for almost all versions of the models derived from the water wave problem. (An alternative approach would be to employ the two-timing approach advocated by Segur, [42], and others, in which the higher-order terms in the expansion are forced evolution equations governed by the leading order Korteweg-deVries equation. However, it is hard to see how the requisite phenomena of maximal height and breaking would manifest themselves in this approach.)

The present paper is devoted to the analysis of solitary wave solutions to a general class of scalar fifth-order evolution equations; see (2.1) below. We begin by discussing the various models that are included in this class, such as the fifth-order Korteweg-deVries equations, other integrable equations, water wave models, and models from elastic media with microstructure. The third section discusses known results on explicit solitary wave solutions for certain models, numerical results, and a nonexistence result of Amick and McLeod for the critical surface tension water wave model. Next we present a simplified approach to the determination of explicit monotone traveling wave solutions, which reduces the fifth-order evolution equation to a third-order ordinary differential equation. This leads to explicit criteria for the existence of exact sech^2 solitary wave solutions, which imply that these models admit either $0, 1, 2, \infty$, or $\infty+1$ exact sech^2 solitary wave solutions. Here ∞ indicates a one-parameter family of solutions valid for a range of wave speeds, and these particular models are explicitly characterized by a pair of simple algebraic relations on the coefficients. Interestingly, even for fifth-order Korteweg-deVries models, there is the possibility of having more than one solitary wave solution for a given wave speed, leading to unusual "bound state solutions." Finally, we present a nonexistence result that says, in essence, that the only models which are perturbations of the usual Korteweg-deVries equation and that possess solitary wave solutions reducing, in the limit, to Korteweg-deVries solitons are those that have a one-parameter family of explicit sech^2 solitary waves. See Theorem 13 for a precise formulation. Our proof relies on a general method introduced by the first author [24] in a similar study of breather solutions of Klein-Gordon equations, which we outline at the end of § 3. Our result does not completely rule out all solitary wave solutions, but only those which reduce to Korteweg-deVries solitary waves in an appropriate scaling limit; nevertheless, it does demonstrate that "physically relevant" solitary wave solutions do not, in general, exist. This has some interesting implications for perturbation theories, which we discuss in the final section.

2. Higher-order model equations.

We will consider a class of fifth-order model evolution equations of the general form

$$\begin{aligned} &u_t+\mu u_{xxx}+\alpha u_{xxxxx}+\beta u u_{xxx}+\delta u_x u_{xx}+P'(u)u_x \\ &\quad =u_t+[\mu u_{xx}+\alpha u_{xxxx}+\beta u u_{xx}+\gamma u_x^2+P(u)]_x=0. \end{aligned} \tag{2.1}$$

Here α, β, γ, $\delta=2\gamma+\beta$, and μ are assumed to be constants, and $P(u)$ is an analytic function of the dependent variable. Many of these models require that P be a cubic polynomial

$$P(u)=pu+qu^2+ru^3, \tag{2.2}$$

where p, q, r are constants, although this will not be necessary for most of our analysis. (However, *only* these models will admit explicit sech^2 solitary waves.) Note that we can assume without loss of generality that $p=0$ by going to a suitable moving coordinate frame. In the models derived by perturbation expansion, the coefficients in (2.1) will depend on a small parameter, ε, in terms of which p is of order 1, q, μ are of order ε, and α, β, δ, and r of order ε^2.

The general class of equations (2.1) includes many well-known equations that have been studied at length in the literature. If the ε^2 terms are absent, the model (2.1) reduces to the well-known Korteweg-deVries equation

$$u_t+pu_x+\mu u_{xxx}+2quu_x=0, \tag{2.3}$$

which serves to model many different wave phenomena requiring a balance between dispersion and nonlinearity, [33], [46]. Also of note is the modified Korteweg-deVries equation

$$u_t+pu_x+\mu u_{xxx}+3ru^2u_x=0. \tag{2.4}$$

Both the Korteweg-deVries and modified Korteweg-deVries equations are known to be integrable via inverse scattering techniques, [33], [42], [46], the scattering operator for the Korteweg-deVries equation being the well-studied Schrödinger operator $L=D^2+v$, where the potential $v(x,t)$ is a suitable multiple of $u(x,t)$, and $D=d/dx$. In particular, their solitary wave solutions are solitons, and interact without change of form. Their speed is related to the value of the associated spectral parameter (eigenvalue). There are additional integrable models included in the class (2.1). The particular parameter values

$$\beta=\tfrac{5}{3}\kappa\alpha,\quad \delta=\tfrac{10}{3}\kappa\alpha,\quad r=\tfrac{5}{18}\kappa^2\alpha,\quad q=\tfrac{1}{2}\kappa\mu, \tag{2.5}$$

where $\kappa\neq 0$, describe a four-parameter family of integrable fifth-order Korteweg-deVries equations [33], which are soluble by the scattering problem associated with the same Schrödinger operator. (More accurately, the models given by (2.5) are linear combinations of purely fifth-order (corresponding to the parameter α) and third-order (corresponding to the parameter μ) Korteweg-deVries equations.) The Sawada-Kotera equation [41],

$$u_t+u_{xxxxx}+30uu_{xxx}+30u_xu_{xx}+180u^2u_x=0, \tag{2.6}$$

and the Kaup equation [21],

$$u_t+u_{xxxxx}+30uu_{xxx}+75u_xu_{xx}+180u^2u_x=0, \tag{2.7}$$

are also known to be integrable, being associated with the scattering problem for the third-order operator $M=D^3+vD+w$; cf. [21]. For the Sawada-Kotera equation, $v=6u$ and $w=0$, whereas for the Kaup equation $v=6u$ and $w=3u_x$. However, in contrast to the higher-order Korteweg-deVries equations, we cannot add in third-order terms to these equations without destroying their integrability.

Other models of the general form (2.1) that are (almost certainly) not integrable also arise in applications. In [34], [35] the second author proposed certain special cases of the general fifth-order model (2.3) as models for the unidirectional propagation of shallow water waves over a flat surface. (See [29] for extensions which include bottom topography.) These arose from two sources: first as the second-degree correction to the standard Korteweg-deVries model for the undirectional propagation of long waves in shallow water arising in the Boussinesq expansion for the full water wave problem. Second, using a general theory of noncanonical perturbation expansions of

Hamiltonian systems, these types of models appear as "Hamiltonian versions" of the Korteweg-deVries model, incorporating the correct first degree expansions of both the water wave Hamiltonian functional (energy) and the Hamiltonian operator. Indeed, whereas the full water wave problem admits a Hamiltonian structure due to Zakharov [50] and the Korteweg-deVries equation admits two distinct Hamiltonian structures [36], neither of these matches the perturbation expansion of Zakharov's structure. Alternatively, we can verify that the first-order perturbation expansion of the water wave energy functional is *not* conserved under the Korteweg-deVries flow. The Hamiltonian models attempt to rectify these unexpected difficulties. In the water wave models, $u(x, t)$ represents either the surface elevation or the horizontal velocity measured at a fraction $0 \leqq \theta \leqq 1$ of the undisturbed fluid depth. There are two small parameters called α, β in [34], [35], but, to avoid confusion, we denote them here by ε, which measures the ratio of wave amplitude to undisturbed fluid depth, and κ, which measures the square of the ratio of fluid depth to wave length. In the shallow water regime, ε and κ are assumed to have the same order of magnitude. The Bond number, which represents a dimensionless magnitude of surface tension, is denoted by τ. In all models, the leading order (Korteweg-deVries) terms are all the same:

$$p=1, \quad \mu=\kappa\frac{1-3\tau}{6}, \quad q=\frac{3}{4}\varepsilon, \tag{2.8}$$

representing a Korteweg-deVries equation except when the Bond number has the critical value $\tau=\frac{1}{3}$. (See below.) The models differ only in the higher-order terms, which take the following forms:

u = horizontal velocity at depth θ; second-order model

$$\alpha=\kappa^2\frac{19-30\tau-45\tau^2}{360}, \quad \beta=\kappa\varepsilon\frac{5-3\tau}{12}, \quad \delta=\kappa\varepsilon\frac{53-36\theta^2-39\tau}{24}, \quad r=0, \tag{2.9}$$

u = horizontal velocity at depth θ; Hamiltonian model

$$\begin{gathered} \alpha=-\kappa^2\frac{(5-6\theta^2-3\tau)(2-3\theta^2)}{18}, \quad \beta=\kappa\varepsilon\frac{53-66\theta^2-27\tau}{24}, \\ \delta=\kappa\varepsilon\frac{139-168\theta^2-81\tau}{24}, \quad r=-\frac{15}{32}\varepsilon^2, \end{gathered} \tag{2.10}$$

u = surface elevation; second-order model

$$\alpha=\kappa^2\frac{19-30\tau-45\tau^2}{360}, \quad \beta=\kappa\varepsilon\frac{5-6\tau}{12}, \quad \delta=\kappa\varepsilon\frac{23+15\tau}{24}, \quad r=-\frac{1}{8}\varepsilon^2, \tag{2.11}$$

u = surface elevation; Hamiltonian model

$$\alpha=0, \quad \beta=\frac{1-3\tau}{8}\kappa\varepsilon, \quad \delta=\frac{3(1-3\tau)}{8}\kappa\varepsilon, \quad r=\frac{5}{32}\varepsilon^2. \tag{2.12}$$

((2.12) corrects an error in [35, eqn. (4.28)].) It is interesting to note that none of these models is integrable, except the Hamiltonian model (2.10) for the horizontal velocity at the particular "magic depth"

$$\theta=\sqrt{\tfrac{11}{12}-\tfrac{3}{4}\tau}, \tag{2.13}$$

where the model turns out to be a fifth-order Korteweg-deVries equation. (This formula corrects a misprint in reference [35].)

The model

$$u_t+pu_x+\mu u_{xxx}+2quu_x+\alpha u_{xxxxx}=0 \tag{2.14}$$

arises in the study of water waves with surface tension in which the Bond number takes on the critical value $\tau = \frac{1}{3}$, where the Korteweg-deVries model no longer applies; cf. [18]. The particular case $p = \mu = 0$ arises in both magneto-acoustics and nonlinear transmission lines; cf. [31], [49]. The equation

$$u_t + pu_x + \mu u_{xxx} + \alpha u_{xxxxx} - uu_{xxx} - 2u_x u_{xx} = 0 \tag{2.15}$$

was proposed by Benney [6] as one possible model for the interaction of short and long waves. Third-order models of the form

$$u_t + u_x + \mu u_{xxx} + 2quu_x + \beta uu_{xxx} + \delta u_x u_{xx} + 3ru^2 u_x = 0, \tag{2.16}$$

in which $\beta = 2\delta \neq 0$, $r = 0$, were proposed by Kunin [28, § 5.3] in his study of elastic media with microstructure. Note that the Hamiltonian model (2.12) for water waves is of this type, but with $\beta = 3\delta \neq 0$, as are both second-order models (2.9), (2.11) at the particular Bond number $\tau = \frac{2}{15}\sqrt{30} - \frac{1}{3} \cong .3970$, and the Hamiltonian model (2.10) at depths $\theta^2 = \frac{2}{3}$ or $\frac{5}{6} - \frac{1}{2}\tau$. Additional models of the form (2.1) have been derived for weakly nonlinear long waves in a stratified fluid [14] and free surface waves over rotational flows [12].

Incidentally, the theory of Kodama [25] shows that all such fifth-order equations with $\alpha \neq 0$, and $P(u)$ a cubic polynomial, can be recast asymptotically into canonical form as fifth-order Korteweg-deVries equations under an appropriate change of variables. Thus, in a certain sense, all the models (2.1), (2.2) are "approximately integrable," although this remark does not imply much in the way of rigorous results for them.

Very recently Ponce [39] has proved that the initial-value problem for (2.1), (2.2) is locally well posed in any Sobolev space $H^s(\mathbb{R})$ for any $s \geqq 4$. Specifically, Ponce proves the following result.

THEOREM 1. *For any* $u_0 \in H^s(\mathbb{R})$ *with* $s \geqq 4$, *there exists a* $T > 0$ *and a unique strong solution* $u(x, t)$ *in the space* $C([0, T], H^s) \cap L^2[0, T], H^{s+2}_{\mathrm{loc}})$ *of the initial value problem* (2.1) *with* $u(x, 0) = u_0(x)$.

3. Solitary wave solutions. We now review known results concerning solitary and other traveling wave solutions to particular models of the form (2.1). We begin by discussing the known explicit solutions to these equations.

First recall that the Korteweg-deVries equation, modified Korteweg-deVries equation, and the class of fifth-order Korteweg-deVries equations (2.5) all possess explicit sech^2 solitary wave solutions for all wave speeds $c > p = P'(0)$. The amplitude of these waves is proportional to the wave speed. If $q/\mu < 0$, then the solitary wave is a wave of elevation, whereas if $q/\mu > 0$ it is a wave of depression. The Sawada-Kotera equation (2.5) also admits sech^2 solitary wave solutions for all wave speeds $c > 0$; cf. [30]. On the other hand, the Kaup equaton (2.6) has solitary wave solutions of the anomalous form

$$u(x, t) = \frac{2a^2(2\cosh 2\xi + 1)}{(\cosh 2\xi + 2)^2}, \qquad \xi = ax - 16a^5 t. \tag{3.1}$$

Again, these exist for a range of wave speeds $c = 16a^4 > 0$.

For the model (2.14) for water waves at critical surface tension, provided $\alpha\mu < 0$, Yamamoto and Takizawa [48] produced an explicit solitary wave of depression in terms of a sech^4 function:

$$u(x, t) = -\frac{105\mu^2}{338\alpha q}\operatorname{sech}^4\left[\sqrt{-\frac{\mu}{52\alpha}}\left\{x + \left(p + \frac{36\mu^2}{169\alpha}\right)t\right\}\right]. \tag{3.2}$$

This solution was also derived by Hereman et al. [15] using a more systematic procedure, and, much later, also by Huang et al. [16]. This "anomalous" solitary wave solution is quite surprising; it only appears for one particular (positive) wave speed: $c = -36\mu^2/169\alpha$. It is unclear whether this solution has any physical meaning. (Other similar "anomalous" sech^2 solitary wave solutions will be determined for many of the models (2.1), (2.2) in § 6.) Of less direct relevance to our results, but still of interest, Hunter and Scheurle [17] proved the existence of traveling waves to the model (2.7) that bifurcate from Korteweg-deVries solitons, but are no longer decreasing as $|x| \to \infty$, having small but finite amplitude oscillatory tails.

Kawahara, [22], claims to numerically establish the existence of "oscillatory solitary wave" solutions to the model (2.14), and Nagashima [31], [32], in the case $p = \mu = 0$, "establishes" their existence experimentally (!). Also, Zufiria [51], in the context of the water wave problem, while more concerned with periodic traveling wave solutions, does investigate "approximate solitary waves" for this model and concludes that they are not unique. However, Amick and McLeod [3] have, using powerful complex-analytic methods, rigorously proved that the model (2.8) does *not* possess a solitary wave of elevation for $\alpha\mu > 0$, with α sufficiently small. (Note that this result does not exclude the exact solitary wave (3.1). See also Hunter and Scheurle [18] for a less rigorous version.) It appears to be quite difficult to extend this technique to the more general models considered in this paper, especially in view of the fact that, for certain models, solitary wave solutions do exist. Amick and McLeod's result implies that Kawahara and Zufiria's numerical solutions cannot be correct, and we propose an explanation for such numerical results in § 8. Indeed, many numerical procedures for finding such waves are, in our opinion, rather suspect, as most of the nonexistence results are of the "exponentially small" variety, i.e., to all orders in ε a solitary wave can be shown to exist, but one may suspect that exponentially small terms (like $e^{-1/\varepsilon}$) prevent its final establishment. See Byatt-Smith [11], Kruskal and Segur [27], [43], and Troy [44], for other problems of this type. Numerical schemes are hard pressed indeed to discover such exponentially small errors!

In the third-order model (2.16), which includes Kunin's third order models for elastic media and some of the water wave models, the equation for solitary waves can, in certain cases, be integrated directly, and one has the intriguing phenomenon of a *wave of maximal height*, reminiscent of the Stokes phenomenon (although the maximal height waves for these models exhibit cusps rather than corners). Indeed, for the full water wave problem, Amick and Toland [4], have proved the existence of monotone solitary wave solutions of small amplitudes up to a maximal height wave with a 120° corner for the problem in the absence of surface tension. (For large values of surface tension, meaning Bond number $\tau > \frac{1}{3}$, Amick and Kirchgässner [2] and Sachs [40] have proved the existence of monotone solitary wave solutions, while very recent results of Iooss and Kirchgässner [19], and Beale [5] demonstrate the existence of solitary wave solutions with damped oscillatory tails for $0 < \tau < \frac{1}{3}$). See also the papers of Wadati, Ichigawa, and Shimizu [45], and Kawamoto [23] for other types of model equations exhibiting limiting cusp waves. It is an interesting question as to whether any of the fifth-order models exhibit such phenomena. Also, the behavior of large amplitude waves (including the possibility of breaking) in these models is not known.

Finally, we mention papers by Yamamoto and Takizawa [47], [48], and Kano and Nakayama [20], which exhibit other types of traveling wave solutions, including periodic waves and solitary sech^2 waves approaching a nonzero asymptotic value as $x \to \pm\infty$. (These can, of course, always be transformed into "genuine" solitary wave solutions, with zero asymptotic limits, to a different model of the same basic form (2.1) by subtracting a suitable constant from u.)

Our own results include the following existence and nonexistence criteria.

On the one hand, we exhibit explicit conditions for a model of the form (2.1) to possess a sech^2 solitary wave solution. First, for such solutions to exist, $P(u)$ must necessarily be a cubic polynomial, (2.2). Interestingly, the parameter space $(\alpha, \beta, \delta, \mu, p, q, r)$ splits into five regions: three of these are relatively open subregions in which there are, respectively, two, one, or no exact sech^2 solitary wave solutions. In the first and second regions, most models have such a solution for a unique, or precisely two, possible wave speeds, similar to the anomalous sech^4 solution to the model (2.8). Secondly, we prove that there are two algebraic relations that must be satisfied by the coefficients in order for the model to admit a one-parameter family of sech^2 solitary wave solutions for a range of wave speeds. This family includes the higher-order Korteweg-deVries equations, (2.4), the Sawada-Kotera equation (2.6), and the Hamiltonian water wave model (2.10) at the particular depth (2.13), but also many other (presumably nonintegrable) equations as well. This leads to the two further regions, each of codimension 2, in which there is either a one-parameter family of sech^2 solitary wave solutions, or such a family plus a single anomalous sech^2 solitary wave solution. These results reconfirm the idea that solitary waves may arise independently of the model being integrable. Also, since the Kaup equation (2.7) admits a one-parameter family of solitary wave solutions for a range of wave speeds that are not sech^2 solutions, we must exercise a bit of caution in drawing unwarranted conclusions from this result!)

On the other hand, assuming $\mu\alpha q \neq 0$, and introducing a small parameter ε representing the departure of the models from the Korteweg-deVries equation, we prove that the only models that admit solitary wave solutions that are perturbations of the corresponding Korteweg-deVries solitons, and satisfy certain analyticity conditions, are the models that satisfy these same algebraic relations. Thus the only physically relevant solitary wave solutions that can exist are always given by sech^2 functions! In outline, our nonexistence result is proved in two basic steps, similar to earlier work of the first author on the nonexistence of breather solutions to a general class of nonlinear Klein-Gordon equations, including the φ^4 equation and the double sine-Gordon equation [24]. We first establish the existence of "solitary wave tails," i.e., traveling wave solutions that decay exponentially fast at either $+\infty$ or $-\infty$, by proving the convergence of the appropriate formal power series solution. The second step in the proof is to match this solution with a formal asymptotic expansion of the solution starting with the one soliton solution of the Korteweg-deVries equation obtained by omitting the fifth-order terms in the model. We then show that, by analyzing the poles of this solution in the complex plane, the second series cannot converge to a true solution, and so we conclude that such a solitary wave solution does not exist. The details will become clearer in the subsequent discussion.

4. The equation for traveling waves. We begin by recalling the elementary method for reducing the problem of traveling wave solutions to an evolution equation such as (2.1) to a connection problem for an ordinary differential equation. A traveling wave solution is just a solution of the particular form

$$u = u(\xi) = u(x - ct), \tag{4.1}$$

where c is the wave speed and $\xi = x - ct$ is the characteristic variable. Substituting the ansatz (4.1) into (2.1), we are led to look for solutions to the fifth-order ordinary differential equation

$$\alpha u''''' + (\beta u + \mu) u''' + \delta u' u'' + [P(u) - cu]' = 0, \tag{4.2}$$

where the primes indicate derivatives with respect to ξ. Any solution $u(\xi)$ of (4.2) thus provides a traveling wave solution to the original evolution equation (2.1). The ordinary differential equation (4.2) can be integrated once, so we effectively have a fourth-order equation

$$\alpha u'''' + (\beta u + \mu)u'' + \gamma u'^2 + Q(u) = 0, \tag{4.3}$$

where

$$Q(u) = P(u) - cu - d, \tag{4.4}$$

with d being a constant of integration.

Consider the case of a localized traveling wave solution, meaning one that is asymptotically small at large distances, so $u \to 0$ as $\xi \to \pm\infty$. Note that this requires $Q(0) = 0$, which fixes the constant of integration d. As it stands, (4.3) is still invariant under the group of translations in ξ (and so could be integrated once more, [36, § 2.5]) and the discrete reflection $\xi \mapsto -\xi$. One way to get rid of this ambiguity is to assume that the wave has its crest (or trough) at $\xi_0 = 0$, and is symmetric with respect to the crest, which means that u is an even function of ξ. Thus we have a fourth-order boundary value problem on the half line $\{\xi > 0\}$, with boundary conditions

$$u'(0) = u'''(0) = 0, \quad \text{and} \quad u(\xi) \to 0, \quad \xi \to +\infty. \tag{4.5}$$

As it stands, it is by no means obvious how to solve the nonlinear connection problem (4.3), (4.5); in particular, the two boundary conditions at $\xi = 0$ define too small a target to try to aim for with a standard shooting approach. This already strongly indicates that, barring exceptional circumstances, the existence of solitary wave solutions will be rare.

5. An equation for monotone solitary waves. We introduce an effective direct method for determining explicit "monotone" (see Definition 2 below) traveling wave solutions to general one-dimensional evolution equations, reducing the fourth-order boundary value problem (4.3), (4.5) on the half line to a (singular) third-order "initial-value problem." The method could also be used to effectively compute solitary and periodic waves (when they exist) numerically, although we have not tried to implement it. (In fact, the method was originally developed by the second author in a failed attempt to prove general existence results concerning solitary wave solutions to these models!) It draws its inspiration from a paper by Kano and Nakayama [20], in which they showed the existence of explicit periodic solutions involving combinations of elliptic functions to certain particular fifth-order models by proving that a suitable polynomial solution w to the reduced equation could be determined; see also Krishnan [26], where a similar method is applied to systems of Boussinesq type. Our method is much more direct and easier to implement than that of Hereman et al. [15].

DEFINITION 2. A *monotone solitary wave solution* is a localized traveling wave solution, i.e., $u \to 0$ as $\xi \to \pm\infty$, which is monotone on the open intervals $(-\infty, \xi_0)$, (ξ_0, ∞), and symmetric about the point ξ_0. The solitary wave is a *wave of elevation* (*depression*) if u is montone increasing (decreasing) on $(-\infty, \xi_0)$, in which case $u_0 = u(\xi_0)$ is called the *crest* (*trough*). A *monotone periodic wave solution* is a traveling wave solution which is periodic in ξ, is monotone on the intervals between crests and troughs, and is symmetric about any crest or trough. A *monotone kink wave solution* is a traveling wave solution which is monotone on the entire real line and approaches limiting values at large distances, so $u \to u_1$ as $\xi \to -\infty$, and $u \to u_2$ as $\xi \to \infty$.

Rather than try to look directly for the required solution $u(\xi)$, we assume that it can be reconstructed as the solution of the simple first-order ordinary differential equation

$$u'^2 = w(u), \qquad u' \equiv \frac{du}{d\xi}, \tag{5.1}$$

where $w(u)$ is a function to be determined. Clearly, once the function $w(u)$ is known, (5.1) can be solved explicitly for $u(\xi)$ by a simple quadrature:

$$\int_a^u \frac{dv}{\sqrt{w(v)}} = \xi + k. \tag{5.2}$$

Examples of solutions that have this form are the soliton and cnoidal wave solutions of the Korteweg-deVries equation [46, § 13.12], where the function $w(u)$ is a cubic polynomial. In particular, if $u(\xi)$ is a monotone function on a given interval, the function $w(u)$ is defined implicitly by the relation (5.1).

The key is the behavior of the function $w(u)$ near its zeros. A simple zero will correspond to a crest or a trough, while a double zero will provide an asymptotic exponential tail for $u(\xi)$ near ∞ or $-\infty$. Thus, a solitary wave solution will correspond to a positive solution $w(u)$ between a double zero at $u=0$ and a simple zero at the crest or trough u_0. (See Fig. 1). Similarly, a periodic wave solution will correspond to a positive solution $w(u)$ between two consecutive simple zeros, (Fig. 2), while a kink solution has two consecutive double zeros, (Fig. 3). We thus have the following useful criterion for the existence of monotone traveling wave solutions to such models.

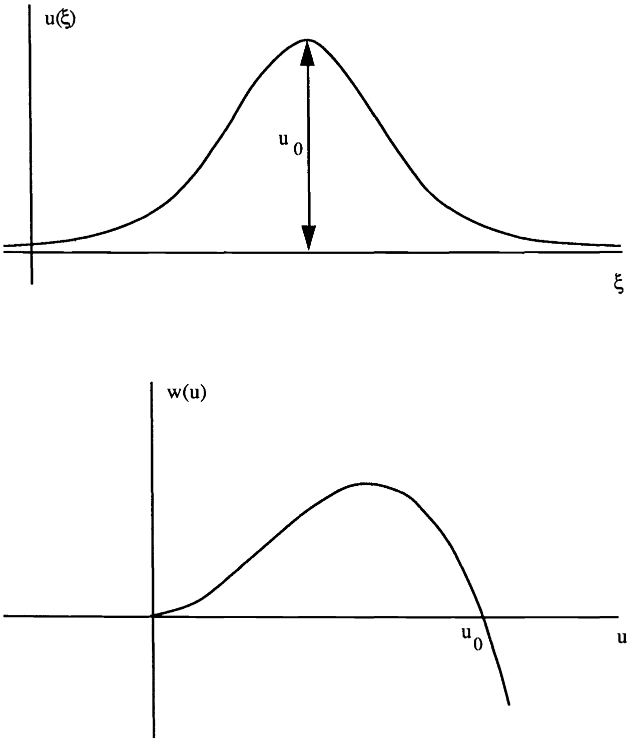


FIG. 1. *Solitary wave solution.*

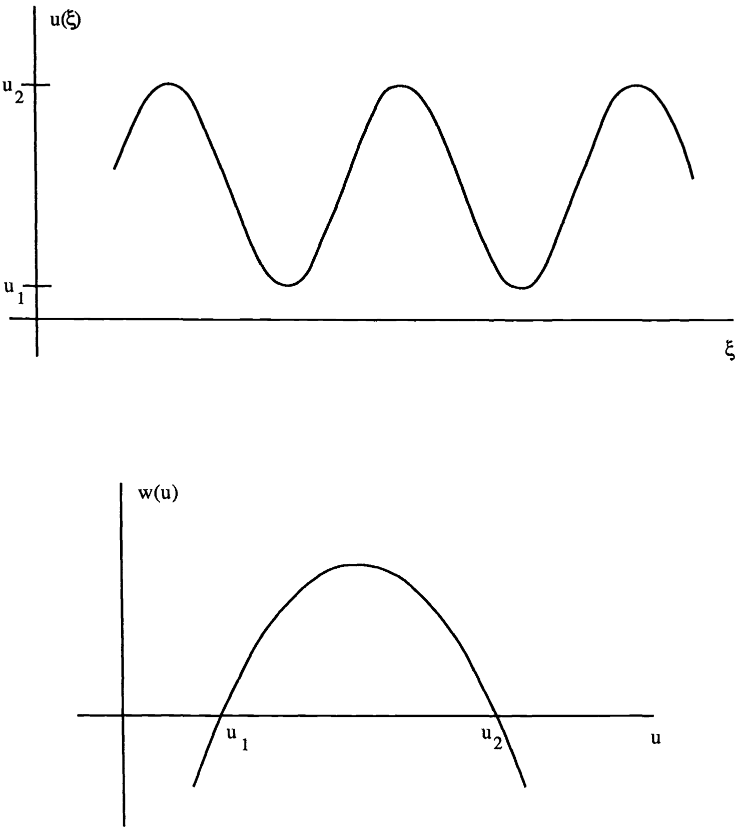


FIG. 2. *Periodic wave solution.*

PROPOSITION 3. *Let $w(u)$ be an analytic function of u, which is positive on the interval $u_0 < u < u_1$, with $w(u_0) = w(u_1) = 0$. Let $u(\xi)$ be the corresponding solution to the first-order ordinary differential equation* (5.1). *If u_0 and u_1 are simple zeros of w, then u is a monotone periodic traveling wave, oscillating between a peak u_1 and a trough u_0. If u_0 is a double zero and u_1 a simple zero of w, then u is a monotone solitary wave of elevation with peak u_1 and asymptotic value u_0 at $\pm\infty$. Conversely, if u_1 is a double zero and u_0 a simple zero of w, then u is a monotone solitary wave of depression with trough u_0 and asymptotic value u_1 at $\pm\infty$. Finally, if u_0 and u_1 are both double zeros of w, then u is a monotone kink wave with asymptotic values u_0, u_1 at $\pm\infty$* (*either going from u_0 to u_1 or the reverse by the reflectional symmetry*).

Using the ansatz (5.1), we substitute into the ordinary differential equation for the traveling wave solution $u(\xi)$, and thereby obtain an ordinary differential equation for the function $w(u)$ of order *one less* than that for u. The goal is then to determine suitable solutions $w(u)$ (if any exist) of this reduced ordinary differential equation. Differentiating our basic equation (5.1), we find that, as long as $u' \neq 0$,

$$\begin{aligned} u'^2 &= w, \\ u'' &= \tfrac{1}{2} w', \\ u'u''' &= \tfrac{1}{2} w w'', \\ u'''' &= \tfrac{1}{2} w w''' + \tfrac{1}{4} w' w'', \end{aligned}$$

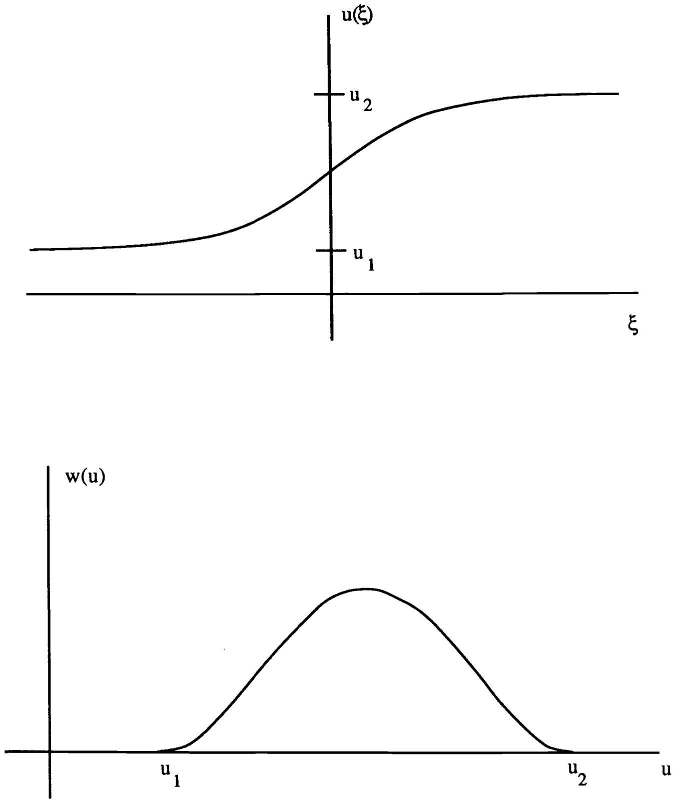


FIG. 3. *Kink wave solution.*

where the primes on w indicate derivatives with respect to u. Substituting into (4.3), we deduce that w must satisfy the third-order ordinary differential equation

$$\frac{\alpha}{4}\{2ww''' + w'w''\} + \frac{1}{2}(\beta u + \mu)w' + \gamma w + Q(u) = 0. \tag{5.3}$$

Any solution to (5.3) will implicitly determine a special traveling wave solution to the original wave equation (2.1) via the integral (5.2). In particular, for a monotone solitary wave solution to the original equation, we need to find a solution $w(u)$ to (5.3) satisfying the initial conditions

$$w(0) = w'(0) = 0, \qquad w''(0) > 0, \tag{5.4}$$

is positive, $w(u) > 0$, for u between 0 and $a \neq 0$, and

$$w(a) = 0, \quad w'(a) \neq 0, \quad w''(a) < \infty. \tag{5.5}$$

In this case a will be the amplitude (crest or trough depending on the sign) of the solitary wave.

6. Exact solitary wave solutions. In certain special cases, we can use the representation (5.1) to easily find exact sech^2 solitary wave solutions to our original evolution equation (2.1). For a solitary wave solution of the specific form

$$u(x, t) = a \operatorname{sech}^2 \lambda(x - ct), \qquad \lambda > 0, \tag{6.1}$$

the corresponding function $w(u)$ must be a cubic polynomial:

$$w(u) = 4\lambda^2\left(u^2 - \frac{1}{a}u^3\right) = \rho u^2 + \sigma u^3, \tag{6.2}$$

where

$$\rho = 4\lambda^2 > 0, \qquad \sigma = -\frac{4\lambda^2}{a} \neq 0 \tag{6.3}$$

are constants to be determined from the equation. Note that since $a = -\rho/\sigma$, we see that $\sigma < 0$ gives a wave of elevation, and $\sigma > 0$ a wave of depression. Substituting (6.2) into (5.3), we first deduce that $Q(u)$ (and hence $P(u)$) must be a cubic polynomial,

$$Q(u) = (p - c)u + qu^2 + ru^3, \tag{6.4}$$

with $Q(0) = 0$; cf. (2.2), (4.4). Moreover, the coefficients ρ and σ must satisfy three polynomial equations:

$$\alpha\rho^2 + \mu\rho + (p - c) = 0, \tag{6.5}$$

$$15\alpha\rho\sigma + 2(\beta + \gamma)\rho + 3\mu\sigma + 2q = 0, \tag{6.6}$$

$$15\alpha\sigma^2 + (3\beta + 2\gamma)\sigma + 2r = 0, \tag{6.7}$$

arising as the coefficients of the powers of u in (5.3). The fact that the solution ρ of the *indicial equation* (6.5) must be positive places certain inequality constraints on the wave speed c depending on the relative signs of the coefficients α, μ. As long as we also have a nonzero solution σ to (6.7), then (6.6) imposes a single compatibility condition on all the coefficients of the evolution equation (2.1) and the wave speed c. As we will see, this implies that there are three open regions in parameter space (coordinated by $\alpha, \beta, \gamma, \mu, p, q, r$), where the model (2.1), (2.2) has precisely 0, 1, or 2 sech^2 solitary wave solutions, for a particular value of the wave speed c.

For a special five-parameter family of models, there is actually a continuum of sech^2 solitary wave solutions for all sufficiently large wave speeds. Note that according to (6.5), ρ will depend on the wave speed c, whereas (6.7) implies that σ does not. Therefore, if the compatibility condition (6.6) is to hold for a range of wave speeds, the coefficient of ρ and the constant term must lead to the same equation for σ. We conclude that the models for which this occurs are those for which

$$(\beta + \gamma)\mu = 5q\alpha \quad \text{and} \quad 15\alpha r = \beta(\beta + \gamma). \tag{6.8}$$

In particular, the four-parameter family of integrable fifth-order Korteweg-deVries equations (2.5), and the Sawada-Kotera equation, (2.6), both satisfy these constraints. However, these do not exhaust all the models satisfying the constraints (6.8); presumably most of the others are not integrable. (Although the Kaup equation, (2.7), has a continuum of solitary wave solutions, they are not of sech^2 type, and so it is in a different class.)

For these particular models, the nature of the sech^2 solitary waves, which comes from an elementary analysis of the conditions for (6.5), (6.7) to admit real solutions ρ, σ, and the resulting signs, is of interest. Since the wave amplitude is given by the formula $a = 3\mu\rho/(2q)$, and $\rho > 0$, if $q\mu > 0$, then the solitary wave is a wave of elevation, whereas if $q\mu < 0$ it is a wave of depression, as in the Korteweg-deVries case (2.3). Substituting into (6.5), we deduce the following quadratic equation relating wave speed and amplitude:

$$c = \frac{4\alpha q^2}{9\mu^2}a^2 + \frac{2}{3}qa + p, \qquad \operatorname{sign} a = \operatorname{sign} q\mu. \tag{6.9}$$

If $\alpha\mu > 0$, then there is a unique solitary wave for each *supercritical* wave speed $c > p$. However, if α and μ have opposite signs, then besides these supercritical sech^2 solutions, there is a nonzero sech^2 solitary wave at the critical wave speed $c = p$, and *two* distinct sech^2 solitary waves for the range of *subcritical* wave speeds between p and $p - \mu^2/(4\alpha)$, reducing to a single wave of amplitude $a^* = -3\mu^2/(4\alpha q)$ at the limiting wave speed $c^* = p - \mu^2/(4\alpha)$. Figures 4 and 5 graph the different possible relationships (6.9) between wave speed and amplitude for the one-parameter family of sech^2 solutions to the models satisfying (6.8).

The elementary observation that a model of the form (2.1) can admit more than one distinct solitary wave solution for a given wave speed does not appear to be well known, even for the integrable fifth-order Korteweg-deVries models. In this particular case, this result can also be detected using the associated scattering problem as follows. The Lax pair for such an equation takes the form $L_t = [B, L]$, where L is the usual second-order Schrödinger operator, and $B = \mu B_3 + \alpha B_5 = \mu L_+^{3/2} + \alpha L_+^{5/2}$ is a linear combination of the third- and fifth-order operators giving the homogeneous third- and fifth-order Korteweg-deVries equations. The eigenvalue for the soliton is constant, and the associated norming constant has the time dependence $m(t)^2 = m(0)^2 \exp[8\mu t\eta^3 - \alpha\eta^5]$. The corresponding wave speed is then $c = (8\mu\eta^3 - \alpha\eta^5)/2\eta$. Thus, we can clearly have ranges of wave speeds for which there are two distinct sech^2 solitons traveling at the same speed. Note that the corresponding two-soliton solution for two such waves represents a bound state with two humps traveling at the same

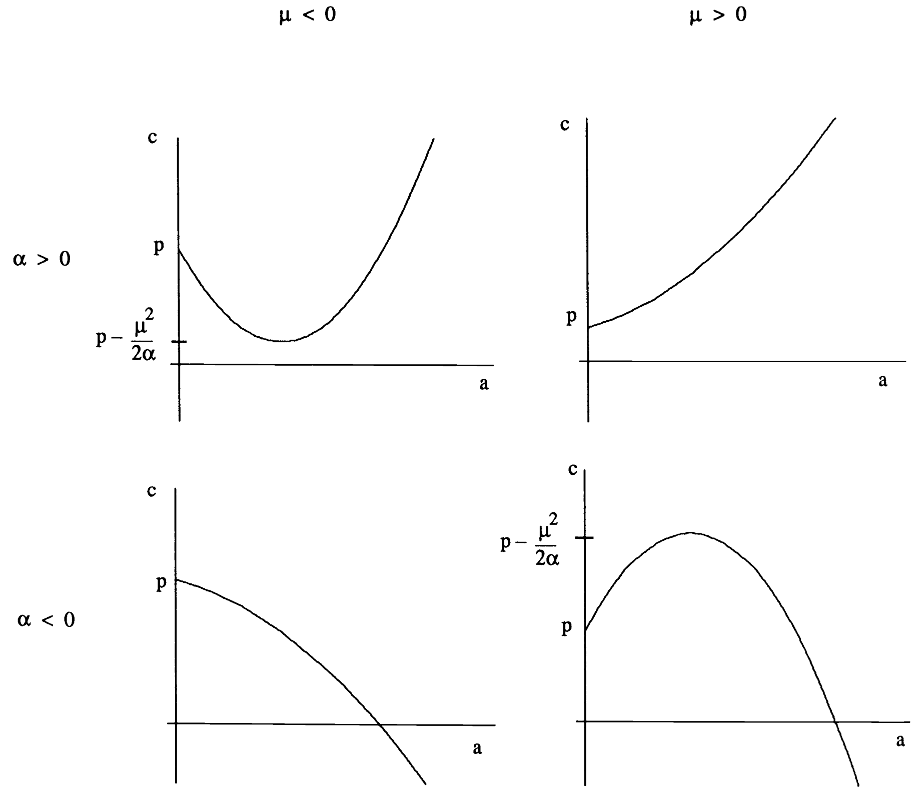


FIG. 4. *Wave speeds and amplitudes for* $q\mu > 0$.

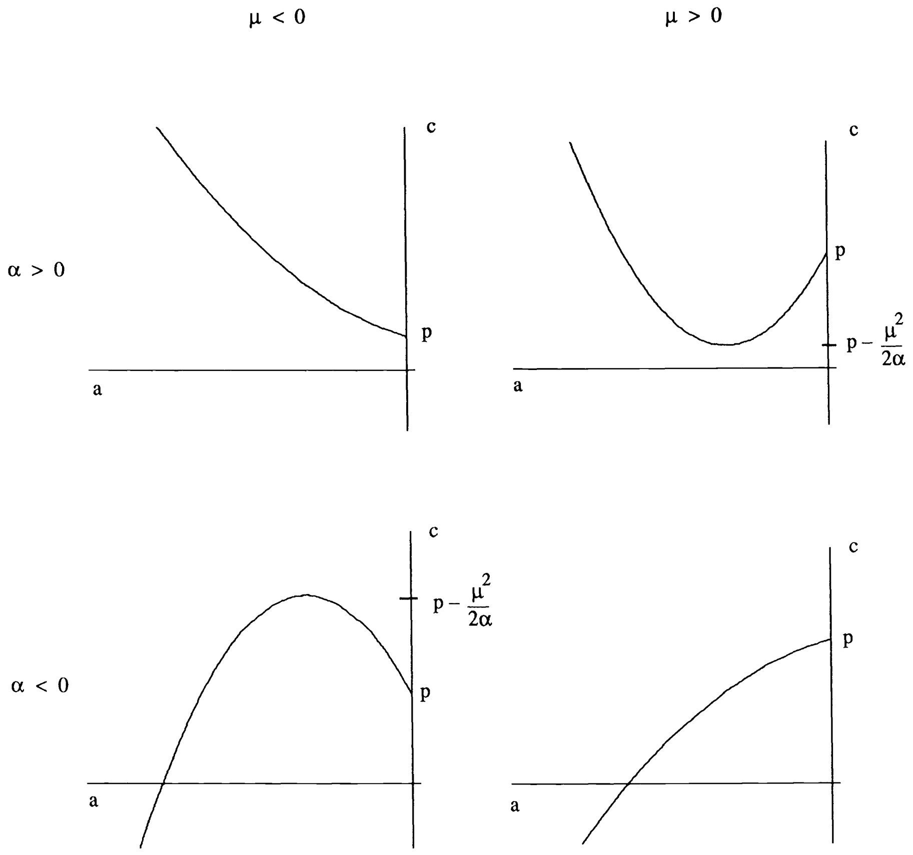


FIG. 5. *Wave speeds and amplitudes for* $q\mu < 0$.

speed. This phenomenon is reminiscent of the construction of bound states to the sine-Gordon equation, consisting of several solitons with phases having real parts with the same speed, the sine-Gordon breather being an example. However, the present property is much stronger, and its appearance for the fifth-order Korteweg-deVries equation is, we believe, a new observation. Note that similarly, we can arrange bound states for linear combinations of higher-order Korteweg-deVries equations to have any number of desired humps traveling in tandem.

Let us summarize our general results completely characterizing models admitting exact sech^2 solitary wave solutions. The different possibilities are: 0, 1, or 2 exact sech^2 solitary wave solutions, a one-parameter family of sech^2 solutions, or a one-parameter family along with a single additional exact sech^2 solution. The first three occur on relatively open subsets of parameter space, whereas the latter two occur on parts of the boundaries between these subsets.

THEOREM 4. *Consider the model evolution equation* (2.1), *assuming* $\alpha \neq 0$. *If* $P(u)$ *is not a cubic polynomial, then the model has no exact* sech^2 *solitary wave solutions. If* $P(u)$ *is given by* (2.2), *then we define*

$$\zeta = (3\beta + 2\gamma)^2 - 120\alpha r, \tag{6.10}$$

so that (6.7) *has* 0, 1, *or* 2 *real roots*

$$\sigma_1, \sigma_2 = \frac{-(3\beta + 2\gamma) \pm \sqrt{\zeta}}{30\alpha}, \tag{6.11}$$

according to whether ζ is negative, zero, or positive. If $r \neq 0$, the real roots are nonzero; if $r=0$, one root, namely $\sigma_1 = -(3\beta+2\gamma)/(15\alpha)$, is nonzero unless $\beta = -\frac{3}{2}\gamma$ also. Then the model (2.1), (2.2) *will have* 0, 1, *or* 2 *exact* sech^2 *solitary wave solutions for each nonzero real root σ_i*, which also satisfies

$$(6.12) \qquad \nu_i = 15\alpha\sigma_i + 2(\beta+\gamma) \neq 0, \qquad \rho_i = \frac{3\mu\sigma_i + 2q}{\nu_i} > 0.$$

Finally, if $(\beta+\gamma)\mu = 5q\alpha$ and $15\alpha r = \beta(\beta+\gamma)$, then the model has a one-parameter family of exact sech^2 *solitary wave solutions valid for a range of wave speeds corresponding to the first root $\sigma_1 = -2(\beta+\gamma)/(15\alpha)$. Moreover, if $\gamma \neq 0$, the second root $\sigma_2 = -\beta/(15\alpha)$ gives rise to a single additional exact* sech^2 *solitary wave solution provided ρ_2, as defined by* (6.12), *is positive.*

Example 5. The only possible water wave model which has a one-parameter family of exact sech^2 solitary wave solutions, i.e., satisfies the conditions (6.8), is the Hamiltonian model (2.10) at the particular depth (2.13). Otherwise, these models all fail to have families of sech^2 solitary wave solutions of the requisite type. However, Theorem 4 implies that many of the water wave models admit one or two anomalous sech^2 solitary wave solutions. The precise numerical values for which the different possibilities occur are rather strange; we will just summarize the results, which were deduced with the help of MATHEMATICA. First, in the case of the second-order depth model (2.9) provided $\alpha \neq 0$, i.e., except for the particular Bond number $\tau = (2\sqrt{30}-5)/15 \cong .3970$, the model admits a single exact sech^2 solitary wave solution unless $3\beta + 2\gamma = 0$, which occurs when $\tau = (73-36\theta^2)/51$. For

$$0 \leqq \tau < \frac{2\sqrt{30}-5}{15} \quad \text{or} \quad \tau > \frac{73-36\theta^2}{51}$$

the anomalous solitary wave is a wave of elevation, while for

$$\frac{2\sqrt{30}-5}{15} < \tau < \frac{73-36\theta^2}{51}$$

it is a wave of depression.

Similarly, for the second-order surface model (2.11) there are one or two exact sech^2 solitary wave solutions provided $\alpha \neq 0$, and $\zeta > 0$, which requires

$$0 \leqq \tau < \frac{4\sqrt{19866}-249}{333} \cong .9453, \qquad \tau \neq \frac{2\sqrt{30}-5}{15} \cong .3970.$$

On the range

$$\frac{\sqrt{85}+5}{30} \cong .4740 < \tau < \frac{\sqrt{23377}-91}{102} \cong .6068,$$

there are two anomalous solitary wave solutions; otherwise, there is just one. In all cases, these are waves of elevation. The Hamiltonian depth model (2.10) also admits exact solitary wave solutions for various ranges of values of the Bond number and depth, but the results are too complicated to warrant inclusion here. We are not sure of the physical significance (if any) of such exact solutions.

7. Existence of solitary wave tails. We now turn to the consideration of more general types of solitary wave solutions. We begin by proving the existence of "solitary wave tails," meaning solutions to the ordinary differential equation (4.3) for traveling waves with the correct asymptotic behavior at $+\infty$. First, let

$$Q(u)=\sum_{m=1}^{\infty} q_m u^m \tag{7.1}$$

be the power series expansion of Q at $u=0$. (Note that $Q(0)=0$ is necessary for the existence of an asymptotically decreasing solution to (4.3).) If $P(u)$ is a cubic of the form (2.2), then

$$q_1=p-c, \quad q_2=q, \quad q_3=r, \quad q_m=0, \quad m>3, \tag{7.2}$$

where c is the wave speed.

Definition 6. A *solitary wave tail* is an exponentially decreasing solution $u(\xi)$ to the equation for traveling waves with asymptotic expansion

$$u(\xi)\sim u_1\, e^{-\theta\xi}+u_2\, e^{-2\theta\xi}+u_3\, e^{-3\theta\xi}+\cdots, \tag{7.3}$$

with $\theta>0$, which converges for ξ sufficiently large.

Of course, we can also discuss solitary wave tails at $\xi=-\infty$, but these are found by using the reflectional symmetry replacing ξ by $-\xi$. We can also consider "oscillatory solitary wave tails," i.e., convergent expansions of the form (7.3) with θ complex and Re $\theta>0$. Our convergence proof will work more or less the same way in this case, but we will just concentrate on the real exponentials for simplicity.

The existence of such an expansion leads to immediate restrictions on the exponent θ and the coefficients in the model. These result from an analysis of the balance equations obtained by substituting (7.3) into (4.3), and equating terms in the various exponentials $e^{-k\theta\xi}$, $k=1,2,3,\cdots$. The first few of these are easily found.

$$e^{-\theta\xi}: \quad (\alpha\theta^4+\mu\theta^2+q_1)u_1=0, \tag{7.4}$$

$$e^{-2\theta\xi}: \quad (16\alpha\theta^4+4\mu\theta^2+q_1)u_2+[(\beta+\gamma)\theta^2+q_2]u_1^2=0, \tag{7.5}$$

$$e^{-3\theta\xi}: \quad (81\alpha\theta^4+9\mu\theta^2+q_1)u_3+[(5\beta+4\gamma)\theta^2+2q_2]u_1u_2+q_3u_1^3=0. \tag{7.6}$$

Since $u_1\neq 0$, (but is otherwise arbitrary), the first balance equation leads immediately to the *indicial equation*

$$\alpha\theta^4+\mu\theta^2+q_1=0. \tag{7.7}$$

The existence of positive real solutions θ to the indicial equation (7.7) places constraints on the coefficients α, μ, q_1 of the linearized model so that exponentially decaying solutions can exist; see Theorem 7 below. Assuming these hold, we eliminate q_1 using (7.7), and the balance equation resulting from the coefficient of $e^{-n\theta\xi}$ takes the form

$$((n^2+1)\alpha\theta^4+\mu\theta^2)u_n=\Psi_n, \tag{7.8}$$

where Ψ_n is a (complicated) polynomial involving the coefficients of the equation and the previous coefficients $u_1,\cdots,u_{n-1}$. Therefore, as long as the *nonresonance* condition

$$(n^2+1)\alpha\theta^2+\mu\neq 0, \qquad n=2,3,\cdots, \tag{7.9}$$

holds for the root θ of the indicial equation, we can solve recursively for all the coefficients u_n, $n=1,2,\cdots$, in the expansion (7.3) and thereby determine a formal solitary wave tail for the equation. Note that if α and μ have the same sign, then the nonresonance condition (7.9) automatically holds. The resonant case is quite intriguing, but we have not investigated it in any detail, and we leave it aside in what follows.

Note in particular, if $u(\xi)=a\operatorname{sech}^2\lambda\xi$, then

$$\theta=-2\lambda, \quad u_1=4a, \quad u_2=-8a, \quad u_3=12a. \tag{7.10}$$

Substituting (7.10) into the three balance equations (7.4), (7.5), (7.7), and using (7.2), (6.3), we recover our earlier three equations (6.5), (6.6), (6.7), relating the equation parameters and the solitary wave parameters a, λ. Thus, we can deduce our earlier parameter restrictions for the existence of sech^2 solitary waves by an alternative procedure based on the asymptotic expansion at ∞. However, in contrast to the earlier direct method, this does not prove that the sech^2 wave is actually a solution to (4.3), since we must also verify the higher-order balance equations. Remarkably, these are all satisfied; see § 8. This observation strongly indicates that only the first three balance equations are important for solitary waves, a fact borne out in the following section.

THEOREM 7. *Consider the model* (2.1), *and let* $Q(u)=P(u)-cu-P(0)$. *If any one of the conditions* (a) $\alpha Q'(0)<0$, (b) $\alpha\mu<0$ *and* $Q'(0)=0$, (c) $\alpha\mu<0$ *and* $4\alpha Q'(0)=\mu^2$, *or* (d) $\alpha=0$ *and* $\mu Q'(0)<0$, *then there exists a unique solitary wave tail* (7.3) *provided the nonresonance condition* (7.9) *holds. If* $0<4\alpha Q'(0)<\mu^2$ *and* $\alpha\mu<0$, *then, again provided the nonresonance condition* (7.9) *holds, there are two solitary wave tails. In all other cases there are no convergent analytic exponentially decreasing solitary wave tails.*

The conditions of Theorem 7 place restrictions on the possible wave speeds c for which there is any possibility of a solitary wave solution decaying exponentially fast to 0 at $\pm\infty$. In the case $\alpha\mu>0$, for a unique asymptotic tail, we need the usual condition that the wave speed be supercritical: $c>p=P'(0)$. (For the water wave models, this gives the standard result that the wave speed of a solitary wave (if it exists) must be larger than 1.) However, if α and μ have opposite signs, there is the possibility of *nonunique* solitary wave tails for some subcritical wave speeds $c<p$. Indeed, this corresponds precisely to what we observed in § 5 for the cases where explicit sech^2 solutions exist.

Proof of Theorem 7. Rather than work with the formal asymptotic expansion for $u(\xi)$ directly, it turns out to be simpler to employ the method introduced in § 5. We let $w(u)=u'^2$ and prove that there is a convergent power series expansion

$$w(u)=\sum_{k=2}^{\infty} w_k u^k = w_2u^2+w_3u^3+\cdots, \tag{7.11}$$

for w at $u=0$, which solves the third-order equation (5.3) with the initial conditions

$$w(0)=w'(0)=0, \qquad w''(0)=2w_2>0. \tag{7.12}$$

It is easy to express the coefficients w_k of w in terms of the coefficients u_i of u; in particular, $w_2=\theta^2$. Clearly, proving the existence of such an analytic solution w will imply that the corresponding solution $u(\xi)$ will have a convergent series expansion (7.3), which is exponentially decreasing as $\xi\to\infty$. Substituting (7.11) into (5.3), we find that the only constant term is $Q(0)$, which must necessarily vanish. The terms involving the first power of u give our by now familiar *indicial equation*

$$\alpha w_2^2+\mu w_2+q_1=0; \tag{7.13}$$

cf. (6.5), (7.7). Assuming that we have a positive solution w_2 to (7.13) (cf. the hypotheses of the theorem), we construct the corresponding power series for w recursively. The coefficient of u^m, $m\geqq 2$, in (5.3) is

$$\sum_{\substack{i+j=m+4\\ i\geqq 3, j\geqq 3}} \frac{\alpha(j-1)(j-2)}{2}[jw_{i-1}w_j+iw_iw_{j-1}]+\frac{\beta m w_m+\mu(m+1)w_{m+1}}{2}+\gamma w_m+q_m=0.$$

Extracting the terms involving w_{m+1} from the sum, we find the recurrence relation

$$w_{m+1}=-\frac{\alpha\sum_{k=3}^{m}k(k-1)(m+k-1)w_kw_{m-k+3}+2(\beta m+2\gamma)w_m+4q_m}{2(m+1)[\alpha w_2(m^2+1)+\mu]}. \tag{7.14}$$

Since $w_2=\theta^2$, the denominator does not vanish owing to the nonresonance condition (7.9), so we can continue to implement the recurrence relation (7.14), and thus construct a formal series solution to (5.3) with the prescribed initial conditions (7.12). We now need to prove convergence, which will follow from the next lemma.

LEMMA 8. *Let $w_2=\theta^2$ be a positive root to the indicial equation* (7.13). *Assume that the nonresonance condition* (7.9) *holds, and let w_m, $m\geqq 3$, satisfy the recurrence relation* (7.14). *Then there exist positive constants A and M such that*

$$|w_m|\leqq\frac{AM^{m-3}}{m^2},\qquad m\geqq 3. \tag{7.15}$$

Proof. Given the convergent power series expansion (7.1) for Q, we know that there exists a number $R>1$ such that the coefficients of the expansion satisfy the inequality

$$|q_m|\leqq R^m\quad\text{for all } m\geqq 1. \tag{7.16}$$

The nonresonance condition implies that there exists a constant $K>0$ such that the inequality

$$m^2+m\leqq 2K|\alpha w_2(m^2+1)+\mu| \tag{7.17}$$

is valid for all $m\geqq 3$. Thus, we have the following estimate on the denominator of (7.14):

$$2(m+1)|\alpha w_2(m^2+1)+\mu m|\geqq\frac{(m+1)^2m}{K}. \tag{7.18}$$

Define the following constants:

$$A=9|w_3|,\qquad M=K\max\left\{\pi^2\alpha A,\frac{2}{3}(|\beta|+|\gamma|),\frac{4R^3}{A},\frac{R}{K}\right\}. \tag{7.19}$$

A straightforward induction, starting at $m=3$, will prove the validity of (7.15). We estimate all of the terms in the numerator of (7.14) in turn. For the summation, we have

$$\begin{aligned}\sum_{k=3}^{m}k(k-1)(m+k-1)|w_k||w_{m-k+3}|&\leqq\sum_{k=3}^{m}\frac{A^2M^{m-3}k(k-1)(m+k-1)}{k^2(m-k+3)^2}\\&\leqq A^2M^{m-3}\sum_{k=3}^{m}\frac{m+k}{(m-k)^2}\\&\leqq A^2M^{m-3}\sum_{j=0}^{m-3}\frac{2m-j}{j^2}\\&\leqq\frac{\pi^2A^2mM^{m-3}}{3}\\&\leqq\frac{AmM^{m-2}}{3K\alpha}.\end{aligned}$$

For the next two terms, we find, since $m\geqq 3$,

$$\begin{aligned}2(|\beta|m+2|\gamma|)|w_m|&\leqq\frac{2A(|\beta|m+2|\gamma|)M^{m-3}}{m^2}\\&\leqq\frac{2Am(|\beta|+|\gamma|)M^{m-3}}{9}\leqq\frac{AmM^{m-2}}{3K},\end{aligned}$$

and, by (7.16),

$$4|q_m| \leqq 4R^m \leqq 4R^3 M^{m-3} \leqq \frac{AmM^{m-2}}{3K},$$

both following from the definition (7.19) of M. Substituting these three estimates and (7.18) into (7.14) easily proves the inductive step for the inequality (7.15).

8. Nonexistence of solitary waves. Having dealt with existence of explicit solitary wave solutions to particular types of the general model (2.1), we now turn our attention to a nonexistence result. We begin by explicitly introducing the small parameter ε into our model, and restrict our attention from the beginning to models in which $P(u)$ is a cubic polynomial. However, this restriction is inessential, and, coupled with the results from Theorem 4, we can deduce that only in this case is there any possibility of suitable solitary wave solutions existing. In the physical models of the form (2.1), (2.2), there is a small parameter ε, relative to which the translation coefficient p has order 1, the Korteweg-deVries terms have coefficients μ, q of order ε, and the fifth-order terms have coefficients α, β, γ (or δ), and r of order ε^2. We also assume that μ, q, and α are all nonzero, so that the model is truly fifth-order, and, moreover, reduces to a Korteweg-deVries equation when the $O(\varepsilon^2)$ terms are neglected. We are interested in the behavior of solutions in the limit $\varepsilon \to 0$, but this is rather trivial without further rescaling since all the terms except the translation will scale out, and everything will reduce to zero. Rather than this, we need to introduce a rescaling of the equation in which the fifth-order terms still have order ε^2, but the translation and Korteweg-deVries terms are of order 1, and compare these solutions in the $\varepsilon \to 0$ limit. In terms of the physical limit, then, we expect the solutions to be order ε^2 perturbations of the corresponding Korteweg-deVries solutions, which are themselves of order ε. Note that, in this limit, the velocity of a Korteweg-deVries soliton has order $c = p + O(\varepsilon^2)$.

We begin with the once-integrated equation for traveling waves (4.3), which, using (2.2), we write in the form

$$(p-c)u + \mu u'' + qu^2 + \alpha u'''' + \beta uu'' + \gamma u'^2 + ru^3 = 0. \tag{8.1}$$

Introduce the scaling

$$\xi = \varepsilon\eta, \quad u = \kappa^2 v, \quad c - p = \kappa^2 s, \tag{8.2}$$

where ε, κ are small parameters, and $s \neq 0$. Rewriting (8.1) for $v = v(\eta)$, we have

$$\varepsilon^2 \mu v'' + \kappa^2(qv^2 - sv) + \varepsilon^4 \alpha v'''' + \kappa^2\varepsilon^2(\beta vv'' + \gamma v'^2) + \kappa^4 rv^3 = 0. \tag{8.3}$$

The condition that the rescaled equation (8.3) possess solutions having the proper expansions in powers of $e^{-\eta}$ at $\eta = +\infty$ is that the rescaled indicial equation

$$s\kappa^2 = \varepsilon^2(\mu + \alpha\varepsilon^2), \tag{8.4}$$

relating the two scaling parameters, hold. This allows us to eliminate κ and rewrite the traveling wave equation in terms of the single small parameter ε:

$$\begin{aligned} &v'' - v + \frac{q}{s}v^2 + \varepsilon^2\left[\frac{\alpha}{\mu}(v'''' - v) + \frac{\alpha q}{s\mu}v^2 + \frac{1}{s}(\beta vv'' + \gamma v'^2) + \frac{\mu r}{s^2}v^3\right] \\ &\quad + \varepsilon^4\left[\frac{\alpha}{s\mu}(\beta vv'' + \gamma v'^2) + 2\frac{r\alpha}{s^2\mu}v^3\right] + \varepsilon^6\frac{\alpha^2 r}{s^2\mu^2}v^3 = 0. \end{aligned} \tag{8.5}$$

PROPOSITION 9. *There exists a formal asymptotic solution to* (8.5) *of the form*

$$v(\varepsilon, \eta) \sim v_0(\eta) + \varepsilon^2 v_1(\eta) + \varepsilon^4 v_2(\eta) + \cdots, \tag{8.6}$$

in which

$$v_0(\eta)=\frac{3s}{2q}\operatorname{sech}^2\frac{\eta}{2}, \tag{8.7}$$

and each $v_j=P_j(v_0)$ *is a polynomial in* $\operatorname{sech}^2(\eta/2)$, *with* $P_j(0)=0$.

Remark. The expansion (8.6) will formally represent the proposed solitary wave solution to the original model reducing to the Korteweg-deVries soliton, (8.7), in the limit $\varepsilon\to 0$. Thus each $v_j(\eta)$ satisfies the condition that it describe a solitary wave; in particular, it is an exponentially decreasing function of $\eta\in\mathbb{R}$. The numerically observed solitary wave solutions [22], [31], [50] can, we believe, be explained by the existence of this nonconvergent formal series. Indeed, a numerical code would be an approximation to a finite truncation of the series (8.6), which would appear to be a numerical approximation to a genuine solitary wave. But owing to the ultimate nonconvergence of the series, the numerically observed solitary wave solution cannot, in fact, be considered to approximate any actual solution to the ordinary differential equation (8.5).

Proof. Note first that (8.7) is the unique even, decaying solution to the zeroth-order equation

$$v_0''-v_0+\frac{q}{s}v_0^2=0. \tag{8.8}$$

To avoid complications in the subsequent formulae, it helps to introduce a further rescaling

$$\zeta=\frac{\eta}{2},\qquad V(\zeta)=\frac{2q}{3s}v(2\zeta), \tag{8.9}$$

in terms of which (8.5) takes the form

$$\begin{aligned}&\tfrac{1}{4}V''-V+\tfrac{3}{2}V^2+\varepsilon^2[\hat{\alpha}\{\tfrac{1}{16}V''''-V+\tfrac{3}{2}V^2\}+\hat{\beta}VV''+\hat{\gamma}V'^2+\hat{r}V^3]\\&\quad+\varepsilon^4\hat{\alpha}[\hat{\beta}VV''+\hat{\gamma}V'^2+2\hat{r}V^3]+\varepsilon^6\hat{\alpha}^2\hat{r}V^3=0,\end{aligned} \tag{8.10}$$

where

$$\hat{\alpha}=\frac{\alpha}{\mu},\quad \hat{\beta}=\frac{3\beta}{8\mu},\quad \hat{\gamma}=\frac{3\gamma}{8\mu},\quad \hat{r}=\frac{9\mu r}{4q^2}. \tag{8.11}$$

The solution $V(\xi)$ will have a formal asymptotic expansion

$$V(\zeta)\sim V_0(\zeta)+\varepsilon^2V_1(\zeta)+\varepsilon^4V_2(\zeta)+\cdots, \tag{8.12}$$

with leading term $V_0(\zeta)=\operatorname{sech}^2\zeta$.

Using the abbreviation $S(\zeta)$ for $\operatorname{sech}^2\zeta$, we group here a few formulae that are elementary, but which will be required in the sequel:

$$S'^2=4S^2(1-S),\qquad S''=4S-6S^2, \tag{8.13}$$

$$\frac{d^2}{d\zeta^2}S^m=mS^m[4m-(4m+2)S]. \tag{8.14}$$

Iterating (8.14) yields

$$\begin{aligned}\frac{d^4}{d\zeta^4}S^m&=16m^4S^m-16m(2m+1)(2m^2+2m+1)S^{m+1}\\&\quad+4m(m+1)(2m+1)(2m+3)S^{m+2}.\end{aligned} \tag{8.15}$$

Consider the particular Schrödinger operator

$$L \equiv -\frac{d^2}{d\zeta^2}+4-12S(\zeta). \tag{8.16}$$

We note that $-12S(\zeta)$ is a three-soliton potential (cf. [33]), so that the spectrum of (8.16) consists of the eigenvalues $\{-5, 0, 3\}$ and a continuous spectrum $\{\lambda \geqq 4\}$; moreover, zero is a simple eigenvalue, with eigenfunction $S'(\zeta)$, which is odd. Thus, L is invertible on even functions in L^2. Also, (8.14) implies

$$L(S^m) = mS^m[4(1-m^2)+(4m^2+2m-12)S]. \tag{8.17}$$

Together, these facts imply the following.

LEMMA 10. *The differential equation*

$$Lf = S^2P(S), \qquad P \text{ a polynomial} \tag{8.18}$$

has a unique even solution which has the form $f = SQ(S)$, where Q is a polynomial.

Now, inserting the expansion (8.12) in (8.10), each coefficient of ε^{2k} results in an equation of the form

$$\tfrac{1}{4}V_k'' - V_k + 3SV_k = F_k(\zeta),$$

or, in view of (8.16)

$$L(V_k) = -4F_k(\zeta). \tag{8.19}$$

One can see by induction that V_k must have the form $SP_k(S)$, where P_k is a polynomial. Indeed, according to Lemma 10, we need only prove that $F_k(\zeta)$ has the form $S^2R_k(S)$, where R_k is a polynomial in S. This results from the following:

(i) The remaining terms in V^2 have the form V_iV_{k-i}, $1 \leqq i \leqq k-1$, and, by the induction hypothesis, each V_i has the form $SP_i(S)$;

(ii) The coefficient of ε^{2k} in the terms ε^2V^2, ε^2V^3, ε^4V^3, and ε^6V^3 is similarly determined from $V_0, \cdots, V_{k-1}$;

(iii) V'^2 is a sum of terms of the form $P(S)'Q(S)'$, and S'^2 has S^2 as a factor by (8.13);

(iv) VV'' has S^2 as a factor by (8.13) again;

(v) (8.15) shows that $\frac{1}{16}V_{\zeta\zeta\zeta\zeta} - V$ also has the form $S^2R(S)$ if $V = SP(S)$.

Therefore, we have proved that there exists a formal series solution to (8.10) of the form

$$V(\zeta) \sim \operatorname{sech}^2 \zeta + \sum_{k=1}^{\infty} \varepsilon^k P_k(\operatorname{sech}^2 \zeta), \tag{8.20}$$

where the P_k are polynomials, $P_k(0) = 0$. This completes the proof of Proposition 9.

PROPOSITION 11. *If the expansion* (8.6) *converges to a holomorphic function in ε and* $\operatorname{sech}^2 \eta/2$ *for $\eta \to \infty$, and ε near zero, then its associated solitary wave tail is a translate of the exponentially decaying tail previously constructed in Lemma* 7.

Proof. By hypothesis, we have a convergent expansion for the tail of the form

$$v(\varepsilon, \eta) = a_1(\varepsilon)\, e^{-\eta} + a_2(\varepsilon)\, e^{-2\eta} + \cdots. \tag{8.21}$$

We must show that $a(\varepsilon) = a_1(\varepsilon)$ never vanishes so that we may replace η by $\eta + \log a(\varepsilon)$ to obtain the series

$$\tilde{v}(\varepsilon, \eta) = e^{-\eta} + b_2(\varepsilon)\, e^{-2\eta} + \cdots, \tag{8.22}$$

which can be compared to the previous form of the tail. To achieve this, we assume $a(\varepsilon_0)=0$ for some ε_0 (possibly complex). Since (8.21) must solve (8.5), the series argument from § 7 immediately shows that in this case, all the coefficients vanish at the point ε_0, $a_k(\varepsilon_0)=0$, and hence $v(\varepsilon_0, \eta)=0$ vanishes for all η. We show that this implies that every ε derivative $(\partial^n v/\partial\varepsilon^n)(\varepsilon_0, \eta)=0$ of v also vanishes at the point ε_0, for all η which, by the holomorphy assumption, ensures $v(\varepsilon, \eta)\equiv 0$, which is impossible since $v_0(\eta)\not\equiv 0$.

Note first that if $v(\varepsilon_0, \eta)$ vanishes for all η, so do all its η-derivatives; therefore, the first ε-derivative $z(\eta)=v_\varepsilon(\varepsilon_0, \eta)$ solves the linear ordinary differential equation

$$z''-z+\varepsilon_0^2\frac{\alpha}{\mu}\{z''''-z\}=0, \tag{8.23}$$

since all the nonlinear terms vanish at ε_0. Moreover, since $v(\varepsilon, \eta)$ is holomorphic, we also have that $z\to 0$ exponentially fast at infinity. But it is easy to see (e.g., by using the Fourier transform) that (8.23) has no nonzero L^2 solutions. Similarly, an easy induction proves that each derivative $z=(\partial^n v/\partial\varepsilon^n)(\varepsilon_0, \eta)$ also solves (8.23), and must, therefore, also be identically zero. This completes the proof and demonstrates the connection between our two series solutions.

Now, by analysis of the analyticity properties of the solutions to our earlier balance equations for the coefficients in the expansion (8.6) we deduce our final nonexistence result.

THEOREM 12. *Suppose* (8.5) *possesses a series solution* (8.6), *which is holomorphic, convergent on a region of the form*

$$|\varepsilon|^2<\left|\frac{\mu}{5\alpha}\right|+\kappa_0, \qquad |e^{-\eta}|<\kappa_1, \tag{8.24}$$

for $\kappa_0, \kappa_1>0$. *Then the equation necessarily satisfies the constraints* (6.8) *and thus has a one-parameter family of exact* sech^2 *solutions.*

Remark. The exact sech^2 solutions are clearly holomorphic in a region of the indicated form (8.24) provided κ_1 is chosen sufficiently small.

Proof. We begin by writing (8.5) in the more convenient form

$$\begin{aligned} &v''+\varepsilon^2\tilde{\alpha}v''''-(1+\varepsilon^2\tilde{\alpha})v \\ &\quad=-\tilde{q}(1+\varepsilon^2\tilde{\alpha})\{v^2+\varepsilon^2[\tilde{\beta}vv''+\tilde{\gamma}v'^2+(1+\varepsilon^2\tilde{\alpha})\tilde{q}\tilde{r}v^3]\}, \end{aligned} \tag{8.24$'$}$$

where

$$\tilde{\alpha}=\frac{\alpha}{\mu}, \quad \tilde{q}=\frac{q}{s}, \quad \tilde{\beta}=\frac{\beta}{q}, \quad \tilde{\gamma}=\frac{\gamma}{q}, \quad \tilde{r}=\frac{\mu r}{q^2}. \tag{8.25}$$

We substitute the expansion (8.21) into (8.24′) to compute the balance equations; cf. (7.4), (7.5), (7.6), for the coefficients a_i. The indicial equation, i.e., the terms in $e^{-\eta}$, are already balanced by design. The terms in $e^{-2\eta}$ lead to the equation

$$3(1+5\varepsilon^2\tilde{\alpha})a_2=-\tilde{q}(1+\varepsilon^2\tilde{\alpha})\{(1+5\varepsilon^2\tilde{\alpha})+\varepsilon^2(\tilde{\beta}+\tilde{\gamma}-5\tilde{\alpha})\}a_1^2. \tag{8.26}$$

Thus, a_2 will have poles at $\varepsilon^2=1/(5\tilde{\alpha})$, contradicting the hypothesis of the theorem, unless $\tilde{\beta}+\tilde{\gamma}=5\tilde{\alpha}$, which, in view of (8.25), is the same as the first condition in (6.8). Assuming this holds, and using (8.26) to solve for a_2, the remaining terms in $e^{-3\eta}$ lead to the further balance equation

$$\begin{aligned} &8(1+10\varepsilon^2\tilde{\alpha})a_3 \\ &\quad=\tfrac{2}{3}\tilde{q}^2(1+\varepsilon^2\tilde{\alpha})^2\{(1+10\varepsilon^2\tilde{\alpha})+\tfrac{1}{2}\varepsilon^2(5\tilde{\beta}+4\tilde{\gamma}-3\tilde{r}-20\tilde{\alpha})\}a_1^3. \end{aligned} \tag{8.27}$$

Thus, a_3 will have poles at $\varepsilon^2 = 1/(10\tilde{\alpha})$, unless $5\tilde{\beta}+4\tilde{\gamma}=3\tilde{r}+20\tilde{\alpha}$, which, in view of the previous condition reduces to $\tilde{\beta}=3\tilde{r}$, and, by (8.25) is the same as the second condition in (6.8); therefore, the expansion will be holomorphic in the indicated domain if and only if the conditions (6.8) hold and the equation admits exact sech^2 solutions. This completes the proof of Theorem 12.

The assumption of analyticity in Theorem 12 parallels that of [24]. It is likely that the constant $\mu/(5\alpha)$ in the domain (8.24) can be replaced by any positive constant $\varepsilon_0 > 0$, as the following argument plausibly indicates. Set, for simplicity,

$$a_1 = \frac{-6}{\tilde{q}(1+\varepsilon^2\tilde{\alpha})}.$$

Then the nth balance equation can, by a simple induction, be shown to take the form

$$(n^2-1)(1+(n^2+1)\varepsilon^2\tilde{\alpha})a_n = \frac{6n+\Phi_n}{\tilde{q}(1+\varepsilon^2\tilde{\alpha})}, \tag{8.28}$$

where each Φ_n is a rational function in ε, with poles at $\varepsilon^2 = -1/((k^2+1)\tilde{\alpha})$, for $k=2,3,\cdots,n-1$, and which vanishes identically if the sech^2 conditions (6.8) hold. In order that the expansion (8.6), and hence the a_i depend analytically on ε in some neighborhood of $\varepsilon=0$, these coefficients cannot have complex poles accumulating at $\varepsilon=0$. Thus, for n sufficiently large, each Φ_n+6n must vanish at $\varepsilon^2=-1/((n^2+1)\tilde{\alpha})$. This infinite collection of polynomial conditions seems highly unlikely in the absence of (6.8). Indeed, we can straightforwardly reduce the size of the domain (8.24) by an involved analysis of the first few of the rational functions Φ_n for n small, perhaps using MATHEMATICA, but we have not tried to implement this.

Note finally that the proof of Theorem 12 can be readily extended to include the case when $P(u)$ is an analytic function, in which case the hypotheses imply that $P(u)$ must be a cubic polynomial also. Indeed, by the above arguments, analyticity of (8.6) in a region (8.24) implies that not only the first three coefficients $p=p_1$, $q=p_2$, $r=p_3$, in the Taylor expansion of $P(u)=\sum p_n u^n$ satisfy (8.6), but, moreover, a simple induction will then show that all remaining coefficients must vanish if the poles in the general recursion relation (8.28) are to cancel, so that $p_n=0$ for $n \geqq 4$. We leave the remaining details to the reader, and conclude this section by summarizing our basic nonexistence result in a convenient unscaled form.

THEOREM 13. *Consider an evolution equation of the form*

$$u_t + [\varepsilon\mu u_{xx} + \varepsilon^2(\alpha u_{xxxx} + \beta u u_{xx} + \gamma u_x^2) + P(u,\varepsilon)]_x = 0, \tag{8.29}$$

where ε is a small parameter, α, β, γ, μ are constants, and P is an analytic function of the form

$$P(u,\varepsilon) = pu + \varepsilon q u^2 + \varepsilon^2 r u^3 + \varepsilon^2 u^4 R(u,\varepsilon), \tag{8.30}$$

where p, q, r are constants, and R is analytic. Assume $q\mu \neq 0$, so that the $O(\varepsilon)$ terms are of Korteweg-deVries type. Then the model has a solitary wave solution of the form $u=u(x-ct,\varepsilon)$ with speed $c=p+\varepsilon^2 s+\cdots$, which has a formal expansion of the form

$$u = \varepsilon\varphi_0[\sqrt{\varepsilon}(x-ct)] + \varepsilon^3\varphi_1[\sqrt{\varepsilon}(x-ct)] + \varepsilon^5\varphi_2[\sqrt{\varepsilon}(x-ct)] + \cdots, \tag{8.31}$$

which reduces to the Korteweg-deVries soliton $\varphi_0(\eta) = \{(3s)/(2q)\}\operatorname{sech}^2 \eta/2$ in the limit. Assume that the expansion (8.31) *converges to an analytic function in a complex domain of the form $|\varepsilon|^2 < |\mu/(5\alpha)| + \kappa$, $\kappa > 0$, $x-ct \gg 0$. Then, necessarily, $R=0$; so $P(u,\varepsilon)$ is a cubic polynomial in u, and the coefficients of* (8.29), (8.30) *are related by the conditions*

$$(\beta+\gamma)\mu = 5q\alpha \quad \textit{and} \quad 15\alpha r = \beta(\beta+\gamma), \tag{8.32}$$

which guarantee the existence of a one-parameter family of exact sech^2 *solitary wave solutions to the model.*

In summary, then, the models (2.1) which admit a one-parameter family of exact sech^2 solitary wave solutions are distinguished by the analyticity properties of their solutions. This result is in direct analogy with those of [24], in which the linear, sine-, and sinh-Gordon equations were distinguished among all one-dimensional Klein-Gordon equations by similar types of analyticity properties. However, our result is more revealing of the general method in that we no longer distinguish, by the smoothness properties of their solutions, just integrable equations, but rather those having particular explicit solutions. The method used here and in [24] is rather general, and is applicable to a wide variety of similar problems.

9. Conclusions and further work. We have been able to prove, under certain reasonable hypotheses, the nonexistence of solitary wave solutions to most fifth-order evolution equations that arise as models for nonlinear water waves. This is very strange, since most of the water wave models, except for the model (2.10) at the particular depth (2.13), where the Hamiltonian model is a fifth-order Korteweg-deVries equation, do not satisfy the requisite conditions (6.8) on the coefficients in the equation. Thus, by trying to do better in modeling real solitary water waves, which are known to exist [4], we, in a sense, do worse. The Korteweg-deVries model does have solitary wave (soliton) solutions that do a reasonably good job approximating solitary water waves [7], [8], [13]. But trying to get a more accurate model by retaining terms in ε^2 leaves us with *no* solitary wave solutions at all! Of course, this is not really an unequivocal problem since presumably the model does do a reasonable job approximating the solitary water waves for times on the order of $1/\varepsilon^2$ (the Kortweg-deVries model being accurate for times on the order of $1/\varepsilon$). Nevertheless, the results of this paper should give one pause in the noncritical application of naïve perturbation expansions as a means for deriving model equations.

This leads us to wonder about the following questions: what happens to initial conditions corresponding to solitary water waves as the time $t \to +\infty$? We expect that small amplitude waves decay by dispersion or radiation, whereas it is plausible that larger waves may even break. Is there a wave of maximal height? How do they behave under collision—specifically do they emerge unscathed as true solitons [33], or is there a small, but nonzero nonelastic effect, as in the BBM equation, [9]? It appears that there is a need for good numerical integration procedures to study these models in more detail. However, these must be long time accurate, and take into account exponentially small effects. For Hamiltonian models, some form of symplectic integrator [10] might be a good bet for investigating these questions. There is a lot of work remaining to be done in this direction.

Acknowledgments. It is a great pleasure to thank Jerry Bona, Stuart Hastings, Bryce McLeod, John Toland, and Bill Troy, all of whom made valuable comments on earlier attempts on this problem.

REFERENCES

[1] C. J. Amick, L. E. Fraenkel, and J. F. Toland, *On the Stokes conjecture for the wave of extreme form*, Acta Math., 148 (1982), pp. 193-214.

[2] C. J. Amick and K. Kirchgässner, *Solitary water waves in the presence of surface tension*, in Dynamical Problems in Continuum Physics, J. L. Bona et al., eds., Springer-Verlag, New York, 1987.

[3] C. J. Amick and J. B. McLeod, *A singular perturbation problem in water waves*, Stability and Appl. Anal. of Continuous Media, to appear.

[4] C. J. AMICK AND J. F. TOLAND, *On solitary waves of finite amplitude*, Arch. Rational Mech. Anal., 76 (1981), pp. 9-95.

[5] J. T. BEALE, *Exact solitary water waves with capillary ripples at infinity*, Comm. Pure Appl. Math., 44 (1991), pp. 211-257.

[6] D. J. BENNEY, *A general theory for interactions between short and long waves*, Stud. Appl. Math., 56 (1977), pp. 81-94.

[7] J. L. BONA, W. G. PRITCHARD, AND L. R. SCOTT, *A comparison of solutions of two model equations for long waves*, in Fluid Dynamics in Astrophysics and Geophysics, N. R. Lebovitz, ed., Lectures in Appl. Math., Vol. 20, American Mathematical Society, Providence, RI, 1983, pp. 235-267.

[8] ———, *An evaluation of a model equation for water waves*, Philos. Trans. Roy. Soc. London, 302 (1981), pp. 457-510.

[9] ———, *Solitary-wave interaction*, Phys. Fluids, 23 (1980), pp. 438-441.

[10] P. J. CHANNEL AND C. SCOVEL, *Symplectic integration of Hamiltonian systems*, Nonlinearity, 3 (1990), pp. 231-259.

[11] J. G. BYATT-SMITH, *On the existence of homoclinic and heteroclinic orbits for differential equations with a small parameter*, Nonlinearity, to appear.

[12] K. W. CHOW, *A second order solution for the solitary wave in a rotational flow*, Phys. Fluids, A1 (1989), pp. 1235-1239.

[13] W. CRAIG, *An existence theory for water waves, and the Boussinesq and Korteweg-deVries scaling limits*, Comm. in Partial Differential Equations, 10 (1985), pp. 787-1003.

[14] J. GEAR AND R. GRIMSHAW, *A second order theory for solitary waves in shallow fluids*, Phys. Fluids, 26 (1983), pp. 14-29.

[15] W. HEREMAN, P. P. BANERJEE, A. KORPEL, G. ASSANTO, A. VAN IMMERZEELE, AND A. MEERPOEL, *Exact solitary wave solutions of nonlinear evolution equations using a direct algebraic method*, J. Phys. A, 19 (1986), pp. 607-628.

[16] G-X. HUANG, S-Y. LUO, AND X-X. DAI, *Exact and explicit solitary wave solutions to a model equation for water waves*, Phys. Lett. A, 139 (1989), pp. 373-374.

[17] J. K. HUNTER AND J. SCHEURLE, *Existence of perturbed solitary wave solutions to a model equation for water waves*, Phys. D, 32 (1988), pp. 253-268.

[18] ———, *Nonexistence of solitary wave solutions of a model equation for water waves*, preprint, 1989.

[19] G. IOOSS AND J. KIRCHGÄSSNER, *Bifurcation d'ondes solitaires en présence d'une faible tension superficielle*, C.R. Acad. Sci. Paris, 311 (1990), pp. 265-268.

[20] K. KANO AND T. NAKAYAMA, *An exact solution of the wave equation* $u_t + uu_x - u_{(5x)} = 0$, J. Phys. Soc. Japan, 50 (1981), pp. 361-362.

[21] D. J. KAUP, *On the inverse scattering problem for cubic eigenvalue problems of the class* $\psi_{xxx} + 6Q\psi_x + 6R\psi = \lambda\psi$, Stud. Appl. Math. 62 (1980), pp. 189-216.

[22] T. KAWAHARA, *Oscillatory solitary waves in dispersive media*, J. Phys. Soc. Japan, 33 (1972), pp. 260-264.

[23] S. KAWAMOTO, *Cusp soliton solutions of the Ito-type coupled nonlinear wave equation*, J. Phys. Soc. Japan, 53 (1984), pp. 1203-1205.

[24] S. KICHENASSAMY, *Breather solutions of the nonlinear wave equation*, Comm. Pure Appl. Math., 44 (1991), pp. 789-818.

[25] Y. KODAMA, *On integrable systems with higher order corrections*, Phys. Lett., 107A (1985), pp. 245-249.

[26] E. V. KRISHNAN, *An exact solution of the classical Boussinesq equation*, J. Phys. Soc. Japan, 51 (1982), pp. 2391-2392.

[27] M. KRUSKAL AND H. SEGUR, *Asymptotics beyond all orders in a model of dendritic crystals*, preprint, 1987.

[28] I. A. KUNIN, *Elastic Media with Microstructure* I, Springer-Verlag, New York, 1982.

[29] T. R. MARCHANT AND N. F. SMYTH, *The extended Korteweg-deVries equation and the resonant flow of a fluid over topography*, J. Fluid Mech., 221 (1990), pp. 263-288.

[30] Y. MATSUNO, *Properties of conservation laws of nonlinear evolution equations*, J. Phys. Soc. Japan, 59 (1990), pp. 3093-3100.

[31] H. NAGASHIMA, *Experiment on solitary waves in the nonlinear transmission line described by the equation* $\partial u/\partial\tau + u\partial u/\partial\xi - \partial^5 u/\partial\xi^5 = 0$, J. Phys. Soc. Japan, 47 (1979), pp. 1387-1388.

[32] H. NAGASHIMA AND M. KUWAHARA, *Computer simulation of solitary waves of the nonlinear wave equation* $u_t + uu_x - u_{(5x)} = 0$, J. Phys. Soc. Japan, 50 (1981), pp. 3792-3800.

[33] A. C. NEWELL, *Solitons in Mathematics and Physics*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 48, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1985.

[34] P. J. OLVER, *Hamiltonian perturbation theory and water waves*, Contemp. Math., 28 (1984), pp. 231-249.

[35] ———, *Hamiltonian and non-Hamiltonian models for water waves*, *in* Trends and Applications of Pure Mathematics to Mechanics, P. G. Ciarlet and M. Roseau, eds., Lecture Notes in Physics No. 195, Springer-Verlag, New York, 1984, pp. 273-290.

[36] P. J. OLVER, *Applications of Lie Groups to Differential Equations*, Graduate Texts in Math. 107, Springer-Verlag, New York, 1986, pp. 140–142.

[37] G. PETIAU, *Sur des fonctions d'ondes d'un type nouveau, solutions d'équatons non linéaires généralisant l'équation des ondes de la mécanique ondulatoire*, Comptes Rendus Acad. Sci. Paris A-B, 244 (1957), pp. 1890–1893.

[38] ———, *Sur un généralisation non linéaire de la mécanique ondulatoire et les propriétés des fonctions d'onde correspondantes*, Nuovo Cimento, 9 (1958), pp. 542–568.

[39] G. PONCE, *Lax pairs and higher order models for water waves*, preprint, 1990.

[40] R. SACHS, *On the existence of small amplitude solitary waves with strong surface tension*, J. Differential Equations, 90 (1991), pp. 31–51.

[41] K. SAWADA AND T. KOTERA, *A method for finding N-soliton solutions of the K.d.V. equation and K.d.V.-like equation*, Progr. Theoret. Phys., 51 (1974), pp. 1355–1367.

[42] H. SEGUR, *Solitons and the inverse scattering transform*, Topics in Ocean Physics, 80 (1982), pp. 235–277.

[43] H. SEGUR AND M. KRUSKAL, *Nonexistence of small amplitude breather solutions in ϕ^4 theory*, Phys. Rev. Lett., 58 (1987), pp. 747–750.

[44] W. C. TROY, *Nonexistence of monotonic solutions in a model of dendritic growth*, Quart. Appl. Math., 48 (1990), pp. 209–215.

[45] M. WADATI, Y. H. ICHIGAWA, AND T. SHIMIZU, *Cusp soliton of a new integrable nonlinear evolution equation*, Progr. Theoret. Phys., 64 (1980), pp. 1959–1967.

[46] G. B. WHITHAM, *Linear and Non-linear Waves*, John Wiley, New York, 1974.

[47] Y. YAMAMOTO AND É. I. TAKIZAWA, *Solutions of nonlinear differential equation* $\partial u/\partial t + (45/2)\delta^2 u^2(\partial u/\partial x) - \partial^5 u/\partial x^5 = 0$, J. Phys. Soc. Japan, 50 (1981), pp. 1055–1056.

[48] ———, *On a solution of nonlinear time-evolution equation of fifth order*, J. Phys. Soc. Japan, 50 (1981), pp. 1421–1422.

[49] K. YOSHIMURA AND S. WATANABE, *Chaotic behaviour of nonlinear evolution equation with fifth order dispersion*, J. Phys. Soc. Japan, 51 (1982), pp. 3028–3035.

[50] V. E. ZAKHAROV, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, J. Appl. Mech. Tech. Phys., 2 (1986), pp. 190–194.

[51] J. A. ZUFIRIA, *Symmetry breaking in periodic and solitary gravity-capillary waves on water of finite depth*, J. Fluid Mech, 184 (1987), pp. 183–206.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1167–1181, September 1992

006

SOBOLEV SPACE METHODS FOR DUAL INTEGRAL EQUATIONS IN AXIALSYMMETRIC SCREEN PROBLEMS*

F. PENZEL[†]

Abstract. The explicit solution of some axialsymmetric scalar screen problems in Sobolev spaces is presented. The well-posedness of the boundary integral equations, which are formulated as dual integral equations, is proved. A reduction for the mixed boundary value problem to a system of singular integral equations with a symbol from the Wiener-algebra is given.

The approach of dual integral equations has a long history [*Mixed Boundary Value Problems in Potential Theory*, North-Holland, Amsterdam, 1966] whereas the new developments reviewed in [E. Meister and F. O. Speck, *Modern Wiener-Hopf methods in diffraction theory*, Proc. Conf. Dundee, 1988, in Ordinary and Partial Differential Equations, B. Sleeman and R. Jarvis, eds., 1989, pp. 130–171] make it possible to handle these equations in Sobolev spaces.

Key words. fractional integrals, potential theory, Laplace equation, Wiener–Hopf equations, dual integral equations

AMS(MOS) subject classifications. 26A33, 31B20, 35J05, 45E10, 45F10

1. Introduction.

In this paper we present the solution of certain dual integral equations by the Wiener–Hopf method. We prove the validity of the method in certain Sobolev spaces. A lot of attempts to solve dual integral equations in distributional spaces has been done before; cf. [11], [26], [28]. Some results about Fredholm properties of certain dual integral equations in L^p spaces are contained in [17]. Nevertheless, the equations considered here do not fall into this class.

During recent years the screen problems considered in [3], [24] were reformulated in Sobolev spaces of locally finite energy $H^1_{\text{loc}}(I\!R^3\backslash\Omega)$ and numerical procedures were given in [6], [7]. This enables us to understand the explicit solution formulas for the Dirichlet and Neumann problem for the Laplacian in the axialsymmetric case in a well-posed operator theoretic setting in Sobolev spaces. For this we have to define some Sobolev spaces of axialsymmetric functions and to characterize their norms by fractional integral operators (see Lemmas 2.1, 2.5, and 2.6).

These results are of independent interest. For example, Lemma 2.1 proves the strong ellipticity of the weakly singular integral operator, which is the boundary integral operator for the Dirichlet problem. This operator was extensively discussed in [27], [6], [7], [16], [29]. The hypersingular integral operator, coming up from the Neumann problem will be handled in Theorem 3.2. Discussions of this integral operator are in [27], [22].

2. Definition of function spaces and operators.

Here we use the following definition of the Fourier transform of a function f:

$$\hat{f}(x) := \int_{I\!R^n} e^{i\xi x} f(\xi) d\xi. \tag{2.1}$$

We start by defining some norms:

$$||f||_{H^s(I\!R^n)} := \left(\int_{I\!R^n} |\hat{f}(x)|^2 (1+|x|)^{2s} dx \right)^{\frac{1}{2}}. \tag{2.2}$$

* Received by the editors July 26, 1989; accepted for publication (in revised form) January 9, 1992.

[†] Fachbereich Mathematik, Technische Hochschule Darmstadt, D-6100 Darmstadt, Germany.

We recall the definition of the following Sobolev spaces that appeared in the solution of boundary value problems from [4]. $\tilde{H}^s(\Omega)$ is the closure of the set of C^∞-functions having compact support in a bounded domain Ω in the norm $||.||_{H^s(I\!R^n)}$. By $H^s(\Omega)$ we denote the closure of the set of functions that are infinitely often differentiable in $I\!R^n$, restricted to Ω in the norm

$$(2.3) \qquad ||f||_{H^s(\Omega)} := \inf_{lf} ||lf||_{H^s(I\!R^n)}.$$

$lf \in H^s(I\!R^n)$ denotes any continuation of f to $I\!R^n$. A lot of methods are available for norms and operators that are homogeneous; therefore, we introduce the following norms in $S'(I\!R^n)$:

$$(2.4) \qquad ||f||_{R^s(I\!R^n)} := \left(\int_{I\!R^n} |\hat{f}(x)|^2 |x|^{2s} dx \right)^{\frac{1}{2}}.$$

Let us define the spaces $R^s(\Omega)$ and $\tilde{R}^s(\Omega)$ analogously to the spaces $H^s(\Omega)$ and $\tilde{H}^s(\Omega)$ by using $||f||_{R^s(I\!R^n)}$ instead of $||f||_{H^s(I\!R^n)}$. Then we can prove our first result.

LEMMA 2.1. *The spaces $\tilde{R}^s(\Omega)$ and $\tilde{H}^s(\Omega)$ are the same set of $S'(I\!R^2)$-distributions for $|s| < 1$. The two norms in these spaces are equivalent. The same assertion holds for the spaces $R^s(\Omega)$ and $H^s(\Omega)$.*

Proof. First we cite from [8] that $\tilde{H}^s(\Omega)$ is a space of S'-distributions having support in Ω. It is known from the book of Gelfand and Shilov [9] that such distributions are derivatives of continuous functions. Therefore, the Fourier transform of these distributions are in C^∞ and the distributions themselves have finite $\tilde{R}^s(\Omega)$-norm for $s > -1$. $\tilde{R}^s(\Omega)$ is a closed subspace of $\tilde{H}^s(\Omega)$; therefore, we can identify the spaces as sets. Because the identity is a continuous and one-to-one mapping from $\tilde{R}^s(\Omega)$ onto $\tilde{H}^s(\Omega)$, we can conclude the stated equivalence of norms. The same holds for the spaces $R^s(\Omega)$ and $H^s(\Omega)$ because it is well known that $H^s(\Omega)$ and $\tilde{H}^{-s}(\Omega)$ are dual spaces with respect to L^2 norm. □

We introduce some Sobolev spaces of axialsymmetric distributions. We restrict ourselves to Sobolev spaces of index s satisfying $|s| < 1$. In this case Lemma 2.1 guarantees the imbedding of our spaces into the usual Sobolev spaces. Here we assume Ω to be the unit disk in $I\!R^2$. For the definition we use functions in $S((0, \infty))$ and have to recognize that they are restrictions of axialsymmetric functions from $S(I\!R^2)$, the space of rapidly decreasing functions from $C^\infty(I\!R^2)$. Here S_0 denotes the Hankel transformation of order zero:

$$(2.5) \qquad S_0 f(x) := \int_0^\infty \xi J_0(\xi x) f(\xi) d\xi.$$

By J_ν we denote the Bessel-function of order ν. If $f \in S(I\!R^2)$ is an axialsymmetric function, then it is well known [12] that the Fourier transform of f is also axialsymmetric and given by the Hankel transform of f of order zero. From now on we identify axialsymmetric functions with their restrictions to $[0, \infty)$ and vice versa. For $|s| < 1$, $\tilde{H}^s_A(\Omega)$ is the closure of the set

$$(2.6) \qquad \{f | f \in C_0^\infty([0, 1)), f^{(2k-1)}(0) = 0 \text{ for k} \in \text{N}\}$$

in the norm $||f||_{\tilde{H}^s_A(\Omega)} := (\int_0^\infty |S_0 l_0 f(x)|^2 x^{2s+1} dx)^{\frac{1}{2}}$, where $l_0 f$ is defined as the extension of f by zero to $[0, \infty)$. The functions from $C_0^\infty([0, 1))$ have to vanish in a neighbourhood of $x = 1$, but they must not vanish at $x = 0$. For $|s| < 1$ $H_A^s(\Omega)$ is defined as the closure of the set

$$\{f|f \in C^\infty([0,1]), f^{(2k-1)}(0) = 0 \text{ for k} \in \mathbb{N}\} \tag{2.7}$$

with respect to the norm $||f||_{\tilde{H}_A^s(\Omega)} := \inf_{lf}(\int_0^\infty |S_0 lf(x)|^2 x^{2s+1}dx)^{\frac{1}{2}}$, where the infimum is taken over all extensions of f on $[0,\infty)$ such that $lf \in S(\mathbb{R}^2)$. For further purposes we need the fractional integral operators defined in [24]:

$$I_{\eta,\alpha}f(x) := \frac{2x^{-2\alpha-2\eta}}{\Gamma(\alpha)}\int_0^x (x^2-u^2)^{\alpha-1}u^{2\eta+1}f(u)du, \tag{2.8}$$

$$K_{\eta,\alpha}f(x) := \frac{2x^{2\eta}}{\Gamma(\alpha)}\int_x^\infty (u^2-x^2)^{\alpha-1}u^{-2\alpha-2\eta+1}f(u)du. \tag{2.9}$$

Γ denotes Euler's Γ-function.

We give two distinct definitions of Mellin transformations:

$$Mf(s) := \int_0^\infty x^{s-1}f(x)dx \tag{2.10}$$

for a complex variable s and

$$M_{\mu,2}f(t) := Mf\left(\frac{\mu}{2}+it\right) \tag{2.11}$$

for a real parameter μ and a real variable t. We introduce weighted Hilbert-spaces:

$$L_\mu{}^2((0,\infty)) := \left\{f|\int_0^\infty |f(x)|^2x^{\mu-1}dx < \infty\right\}. \tag{2.12}$$

Remark 2.1. The space $L_2^2((0,\infty))$ can be identified with the axialsymmetric $L^2(\mathbb{R}^2)$ functions restricted to the half-axis.

We cite from [18].

LEMMA 2.2. *The Mellin transformation $M_{\mu,2}$ is an isomorphism from $L_\mu{}^2((0,\infty))$ onto $L^2(\mathbb{R})$ for all $\mu \in \mathbb{R}$.*

In the next lemma we shall describe equivalent norms in the Sobolev spaces $\tilde{H}_A^s([0,1])$ and $H_A^s([0,1])$ for $|s| < \frac{1}{2}$ by using the Mellin transformation or fractional integral operators. To abbreviate equivalence of norms, we shall use the symbol $\approx$.

LEMMA 2.3. *The following pairs of norms are equivalent in $\tilde{H}_A^s(\Omega)$, respectively, in $H_A^s(\Omega)$ for $|s| = \frac{1}{2}$:*

$$||f||_{H_A^{-\frac{1}{2}}(\Omega)} \approx \inf_{lf}\left(\int_{\mathbb{R}} \frac{|M_{3,2}(lf)(t)|^2}{(1+|t|)}dt\right)^{\frac{1}{2}}, \tag{2.13}$$

$$||f||_{H_A^{\frac{1}{2}}(\Omega)} \approx \inf_{lf}\left(\int_{\mathbb{R}} |M_{1,2}(lf)(t)|^2(1+|t|)dt\right)^{\frac{1}{2}}, \tag{2.14}$$

$$||f||_{\tilde{H}_A^{\frac{1}{2}}(\Omega)} \approx \left(\int_{\mathbb{R}} |M_{1,2}(l_0f)(t)|^2(1+|t|)dt\right)^{\frac{1}{2}}, \tag{2.15}$$

$$||f||_{\tilde{H}_A^{-\frac{1}{2}}(\Omega)} \approx \left(\int_{I\!R} \frac{|M_{3,2}(l_0 f)(t)|^2}{(1+|t|)} dt\right)^{\frac{1}{2}}. \tag{2.16}$$

Proof. Let $f \in C^\infty([0,1]), f^{(2k-1)}(0) = 0$ for $k \in I\!N$, and let $lf \in S(I\!R^2)$ be an extension of f on $(0,\infty)$. The Mellin transformations $M_{3,2}$ and $M_{1,2}$ map this type of functions on functions from $S(I\!R)$,

$$(M_{1,2}(lf))(t) = \int_0^\infty x^{-\frac{1}{2}+it}(lf)(x)dx = \int_{I\!R} e^{i\rho t}e^{\frac{1}{2}\rho}(lf)(e^\rho)d\rho. \tag{2.17}$$

The function $e^{\frac{1}{2}\rho}(lf)(e^\rho)$ is in $S(I\!R)$ because lf and all its derivatives decay exponentially at infinity:

$$e^{\frac{1}{2}\rho}(lf)(e^\rho) \leq \frac{C_k}{1+e^{k\rho}}e^{\frac{\rho}{2}} \quad \text{for } \rho \to \infty \text{ and k} \in \mathbf{N}, \tag{2.18}$$

$$e^{\frac{1}{2}\rho}(lf)(e^\rho) = (lf)(0)O(e^{\frac{\rho}{2}}) \quad \text{for } \rho \to -\infty. \tag{2.19}$$

The proof for $M_{3,2}$ is analogous.

$$\begin{aligned} {||f||^2}_{H_A^{-\frac{1}{2}}(\Omega)} &= \inf_{lf} \int_0^\infty |S_0 lf(x)|^2 dx \\ &= \inf_{lf} \int_{I\!R} \left|(MS_0 lf)\left(\frac{1}{2}+it\right)\right|^2 dt \\ &= \inf_{lf} \int_{I\!R} \left|(MJ_0)\left(\frac{1}{2}+it\right)(Mlf)\left(\frac{3}{2}-it\right)\right|^2 dt \\ &= \inf_{lf} \int_{I\!R} \left|2^{-1/2-it}\Gamma\left(\frac{1}{4}+\frac{it}{2}\right)\Gamma^{-1}\left(\frac{3}{4}-\frac{it}{2}\right)(Mlf)\left(\frac{3}{2}+it\right)\right|^2 dt. \end{aligned} \tag{2.20}$$

The third and the fourth equality can be proved with the aid of formulas from [5].

The asymptotic formula

$$\Gamma(x+iy) \sim |y|^{x-1}, \tag{2.21}$$

which holds for fixed real x and for $y \to \pm\infty$, proves the first equivalence of norms if we take the infimum over all extensions of f to $L_2^2((0,\infty))$.

To prepare the proof of the second equivalence, we use the following formula, which holds for continuously differentiable functions f, whose first derivative vanishes at $x = 0$:

$$S_0 f(x) = -\frac{1}{x}\int_0^\infty \xi J_1(\xi x)f'(\xi)d\xi. \tag{2.22}$$

This we prove by integration by parts on the left-hand side of the equation and by using the relation $(\partial/\partial\xi)((\xi/x)J_1(\xi x)) = \xi J_0(\xi x)$.

Let us assume that f is a function in $C^\infty([0,1])$, satisfying $f'(0) = 0$. Let lf be a smooth extension of f to $(0,\infty)$. Using the relation above and some properties of the

Mellin transformation, which we looked up from [5] , we get the following equation:

(2.23)

$$\begin{aligned}\int_0^\infty & |S_0lf(x)|^2x^2dx\\ &= \int_{\mathbb{R}}\left|(MS_0lf)\left(\frac{3}{2}+it\right)\right|^2dt\\ &= \int_{\mathbb{R}}\left|M\left(\frac{1}{x}\int_0^\infty \xi J_1(\xi x)(lf)'(\xi)d\xi\right)\left(\frac{3}{2}+it\right)\right|^2dt\\ &= \int_{\mathbb{R}}\left|(MJ_1)\left(\frac{1}{2}+it\right)(M(lf)')\left(\frac{3}{2}-it\right)\right|^2dt\\ &= \int_{\mathbb{R}}\left|2^{-\frac{1}{2}+it}\Gamma\left(\frac{3}{4}+\frac{it}{2}\right)\Gamma^{-1}\left(\frac{5}{4}-\frac{it}{2}\right)\left(\frac{1}{2}-it\right)(Mlf)\left(\frac{1}{2}-it\right)\right|^2dt.\end{aligned}$$

We prove the second equivalence of norms substituting t by $-t$ in the last integral, using the above given asymptotic formula for the Γ -function and taking the infimum over all extensions of f to $L_2^2((0,\infty))$. The proofs of the third and the fourth equivalence can be done analogously. We have to use functions f with compact support in $(0,1)$. □

To give an idea which functions are in the Sobolev spaces defined above, we prove a corollary of Lemmas 2.1, 2.2, and 2.3.

COROLLARY 2.1. *The following inclusions hold:*

$$L_{1,2}((0,1))\cap\left\{f\,\middle|\,\inf_{lf\in L_{1,2}(\mathbb{R}^+)}\int_{\mathbb{R}}|M_{1,2}(lf)(t)|^2(1+|t|)dt<\infty\right\}\subset H_A^{\frac{1}{2}}(\Omega),$$

$$L_{3,2}((0,1))\subset \tilde{H}_A^{-\frac{1}{2}}(\Omega). \tag{2.24}$$

Proof. Let $f\in L_{1,2}((0,1))$ have an extension $lf\in L_{1,2}(\mathbb{R}^+)$ such that

$$\int_{\mathbb{R}}|M_{1,2}(lf)(t)|^2(1+|t|)dt<\infty.$$

Then it is possible to approximate lf by functions $lf_n\in C_0^\infty(\mathbb{R}^+)$ in the $L_{1,2}(\mathbb{R}^+)$-sense. By Lemma 2.1, $M_{1,2}(lf_n)$ converges to lf in $L^2(\mathbb{R})$-sense, which implies convergence almost everywhere. If n is large enough, the integrals $\int_{\mathbb{R}}|M_{1,2}(lf_n)(t)|^2(1+|t|)dt$ must also exist; f can be approximated by functions from $C^\infty([0,1])$ in the $H_A^{\frac{1}{2}}(\Omega)$-norm. The proof of the second inclusion is analogous. We have to keep in mind that for functions $f\in L_{3,2}(\mathbb{R}^+)$ we get the estimate

$$\int_{\mathbb{R}}\frac{|M_{3,2}(f)(t)|^2}{(1+|t|)}dt\le ||M_{3,2}f||_{L^2(\mathbb{R})}^2=||f||_{L_{3,2}(\mathbb{R}^+)}^2. \tag{2.25}$$

□

Remark 2.2. The extensions lf in Corollary 2.1 can be restricted to extensions $lf\in C_0^\infty([0,b])$ for $b>1$ because these functions are a dense subset of $L_{\mu,2}(\mathbb{R}^+)$.

Given a set Ω, we denote by χ_Ω the characteristic function of Ω.

LEMMA 2.4. *The operator* $\frac{x}{2}K_{-\frac{1}{2},1}\chi_{[0,1]}$, *which maps* f *onto* $\int_x^1 f(u)du$, *is extendable to an isomorphism from* $\tilde{H}_A^{-\frac{1}{2}}(\Omega)$ *onto* $\tilde{H}_A^{\frac{1}{2}}(\Omega)$. *The inverse is given by*

$-d/dx$. The operator $\frac{x}{2}I_{0,1}\chi_{[0,1]}$, which maps f onto $\int_0^x \frac{u}{x}f(u)du$, is extendable to an isomorphism from $H_A^{-\frac{1}{2}}(\Omega)$ onto $H_A^{\frac{1}{2}}(\Omega)$. The inverse is given by $(Df)(x) := f'(x) + (f(x)/x)$.

Proof. Let $v \in C_0^\infty([0,1)), v^{(2k-1)}(0) = 0$ for k $\in$ N. We prove that $\frac{x}{2}K_{-\frac{1}{2},1}\chi_{[0,1]}$ is an isomorphism:

$$\begin{aligned} \left\|\frac{x}{2}K_{-\frac{1}{2},1}\chi_{[0,1]}v\right\|^2_{\tilde{H}_A^{\frac{1}{2}}} &= \int_{I\!R}\left|\int_0^\infty x^{-\frac{1}{2}+it}\int_x^1 (l_0v)(s)ds\,dx\right|^2 (1+|t|)dt \\ &= \int_{I\!R}\left|\int_0^\infty \frac{x^{\frac{1}{2}+it}}{\frac{1}{2}+it}(l_0v)(x)dx\right|^2 (1+|t|)dt \\ &\approx \int_{I\!R}\left|\int_0^\infty x^{\frac{1}{2}+it}(l_0v)(x)dx\right|^2 \frac{dt}{(1+|t|)} = ||v||_{\tilde{H}_A^{-\frac{1}{2}}(\Omega)}. \end{aligned} \tag{2.26}$$

The formula for the inverse follows from $-(d/dx)\int_x^1 v(s)ds = v(x)$. We shall prove that the operator $\frac{x}{2}I_{0,1}\chi_{[0,1]}$ is the adjoint operator to $\frac{x}{2}K_{-\frac{1}{2},1}\chi_{[0,1]}$. For $f \in C_0^\infty((0,1)), g \in C^\infty([0,1])$ holds:

$$\int_0^1 g(x)\frac{x}{2}(K_{-\frac{1}{2},1}f)(x)dx = \int_0^1 g(x)x\int_x^1 f(u)du\,dx = \int_0^1\int_0^x \frac{u}{x}g(u)du f(x)x\,dx, \tag{2.27}$$

which proves that the operator $\frac{x}{2}I_{0,1}\chi_{[0,1]}$ is adjoint to the operator $\frac{x}{2}K_{-\frac{1}{2},1}\chi_{[0,1]}$ with respect to the natural dual pairing between $\tilde{H}_A^{-\frac{1}{2}}(\Omega)$ and $H_A^{\frac{1}{2}}(\Omega)$. The formula for the inverse is now obvious. □

Remark 2.3. The constant functions are not in the range of the operator $\frac{x}{2}K_{-\frac{1}{2},1}\chi_{[0,1]}$, indeed they are not elements of $\tilde{H}_A^s(\Omega)$ for $s \geq \frac{1}{2}$. The operator D maps the function which is equal to one on the disk onto the function $1/|x|$, which is in $H_A^{-\frac{1}{2}}(\Omega)$. On these facts hinges the injectivity of the operators D and d/dx.

LEMMA 2.5. *The operator $x^{\frac{1}{2}}K_{\gamma,\frac{1}{2}}$ maps $L_2^2((0,1))$ one-to-one onto $\tilde{H}_A^{\frac{1}{2}}(\Omega)$ and the operator $(d/dx)x^{\frac{1}{2}}K_{\gamma,\frac{1}{2}}$ maps $L_2^2((0,1))$ one-to-one onto $\tilde{H}_A^{-\frac{1}{2}}(\Omega)$, if $\gamma > -\frac{1}{2}$.*

Proof. Let $v \in C_0^\infty((0,1))$. The function $g := x^{\frac{1}{2}}K_{\gamma,\frac{1}{2}}v$ behaves like $x^{2\gamma+\frac{1}{2}}$ at $x = 0$ and vanishes in the neighbourhood of $x = 1$. For $\gamma > -\frac{1}{2}$ the function g is in $L_1^2((0,1))$. To estimate the $\tilde{H}_A^{\frac{1}{2}}(\Omega)$ norm of g, we may use the following relation:

$$\begin{aligned} &\left\|\sqrt{1+|t|}M(g)\left(\frac{1}{2}+it\right)\right\|_{L^2(I\!R)} \\ &= \left\|\sqrt{1+|t|}\frac{\Gamma(\frac{1}{2}+\gamma+\frac{it}{2})}{\Gamma(1+\gamma+\frac{it}{2})}(Mv)(1+it)\right\|_{L^2(I\!R)} \approx ||(Mv)(1+it)||_{L^2(I\!R)}. \end{aligned} \tag{2.28}$$

This proves the equivalence of the norms $||x^{\frac{1}{2}}K_{\gamma,\frac{1}{2}}v||_{\tilde{H}_A^{\frac{1}{2}}(\Omega)}$ and $||v||_{L_2^2((0,1))}$. The mapping property of $(d/dx)x^{\frac{1}{2}}K_{\gamma,\frac{1}{2}}$ follows from Lemma 2.4. □

LEMMA 2.6. *The operator $x^{\frac{1}{2}}I_{\eta,\frac{1}{2}}$ maps $L_2^2((0,1))$ one-to-one onto $H_A^{\frac{1}{2}}(\Omega)$ and the operator $Dx^{\frac{1}{2}}I_{\eta,\frac{1}{2}}$ maps $L_2^2((0,1))$ one-to-one onto $H_A^{-\frac{1}{2}}(\Omega)$, if $\eta > -\frac{1}{2}$.*

Proof. Let $v \in C_0^\infty((0,1))$. We calculate the Mellin transform of $h(x) := x^{\frac{1}{2}} I_{\eta,\frac{1}{2}}$ $(l_0 v)(x)$; h is in $C^\infty((0,\infty))$ and vanishes in a neighbourhood of $x = 0$. From [5] we cite

$$\begin{aligned}(Mh)\left(\frac{1}{2}+it\right) &= \frac{2}{\Gamma(\frac{1}{2})}(M((x^2-1)^{-1/2}\chi_{(1,\infty)}))(-2\eta+it)(Ml_0v)(1+it)\\ &= \Gamma\left(\frac{1}{2}+\eta-i\frac{t}{2}\right)\Gamma^{-1}\left(1+\eta-i\frac{t}{2}\right)(Ml_0v)(1+it).\end{aligned} \tag{2.29}$$

This proves, that the first operator in question maps $L_2^2((0,1))$ continuously into $H_A^{1/2}(\Omega)$. Its injectivity is well known [18]. To prove surjectivity, we prove that the inverse operator is densely defined and continuous from $H_A^{\frac{1}{2}}(\Omega)$ into $L_2^2((0,1))$. For this let us assume $h_1 \in C^\infty([0,1]), h_1^{(2k-1)}(0) = 0$, if $k \in I\!N$ is given. Using some formulas from [24], we get the following result:

$$(I_{\eta,\frac{1}{2}}{}^{-1}h_1)(x) = \frac{1}{2}x^{-2\eta-1}\frac{d}{dx}x^{2+2\eta}(I_{\eta+\frac{1}{2},\frac{1}{2}}h_1)(x). \tag{2.30}$$

We extend h_1 by lh_1 to the half-axis such that the support of lh_1 is a bounded interval, say $[0, x_0]$, and obtain

(2.31)

$$\begin{aligned}&M(I_{\eta,\frac{1}{2}}{}^{-1}x^{-\frac{1}{2}}lh_1)(1+it)\\ &= \int_0^\infty x^{-2\eta-1+it}\frac{d}{dx}\frac{2}{\Gamma(\frac{1}{2})}\int_0^x \frac{u^{2\eta+\frac{3}{2}}(lh_1(u))}{\sqrt{x^2-u^2}}du\,dx\\ &= -\frac{2}{\Gamma(\frac{1}{2})}(-2\eta-1+it)\int_0^\infty x^{-2\eta-2+it}\int_0^\infty \chi_{[1,\infty)}\left(\frac{x}{u}\right)\frac{u^{2\eta+\frac{1}{2}}(lh_1(u))}{\sqrt{(\frac{x}{u})^2-1}}du\,dx\\ &= -\frac{2}{\Gamma(\frac{1}{2})}(-1-2\eta+it)M(\chi_{[1,\infty)}(x)(x^2-1)^{-\frac{1}{2}})(-1-2\eta+it)(Mlh_1)\left(\frac{1}{2}+it\right)\\ &= -(-1-2\eta+it)\Gamma\left(1+\eta-i\frac{t}{2}\right)\Gamma^{-1}\left(\frac{3}{2}+\eta-i\frac{t}{2}\right)(Mlh_1)\left(\left(\frac{1}{2}+it\right)\right).\end{aligned}$$

Here we used integration by parts and the relations

$$\left|\int_0^x \frac{u^{2\eta+\frac{3}{2}}(lh_1(u))}{\sqrt{x^2-u^2}}du\right| \le \frac{1}{x}\int_0^{x_0}\frac{u^{2\eta+\frac{3}{2}}|lh_1(u)|}{\sqrt{1-(\frac{u}{x})^2}}du = O\left(\frac{1}{x}\right) \quad \text{for x} \to \infty, \tag{2.32}$$

(2.33)

$$\left|\int_0^x \frac{u^{2\eta+\frac{3}{2}}(lh_1(u))}{\sqrt{x^2-u^2}}du\right| \le ||lh_1||_{L^\infty([0,x_0])}\int_0^1 \frac{(xt)^{2\eta+\frac{3}{2}}}{\sqrt{1-t^2}}dt = O(x^{2\eta+\frac{3}{2}}) \quad \text{for x} \to 0,$$

which ensure the existence of the integrals above and allow us to neglect the boundary terms in the second equality.

Using formulas (2.14), (2.21), we conclude the desired surjectivity of $x^{\frac{1}{2}}I_{\eta,\frac{1}{2}}$. The mapping property of the operator $Dx^{\frac{1}{2}}I_{\eta,\frac{1}{2}}$ follows from Lemma 2.4. □

Remark 2.4. The fractional integral operators $x^{\frac{1}{2}}K_{\gamma,\frac{1}{2}}, x^{-\frac{1}{2}}I_{\eta,\frac{1}{2}}$ map functions with support in $[0,1)$, respectively, $[1,\infty)$ onto functions of the same type. This property holds because their Mellin symbols, which are defined by

$$(x^{\frac{1}{2}}k_{\gamma,\frac{1}{2}})(t) := \frac{\Gamma(\frac{1}{2}+\gamma+i\frac{t}{2})}{\Gamma(1+\gamma+i\frac{t}{2})}, \qquad (x^{\frac{1}{2}}i_{\eta,\frac{1}{2}})(t) := \frac{\Gamma(\frac{1}{2}+\eta-i\frac{t}{2})}{\Gamma(1+\eta-i\frac{t}{2})}, \tag{2.34}$$

are holomorphically extendable to the lower half plane and to the upper half plane, respectively.

3. Formulation and solution of dual integral equations for the Laplacian. The Dirichlet problem for the Laplacian reads

$$\Delta u = 0 \quad \text{in } \mathbf{R}^3\backslash\overline{\Omega}, \quad \mathrm{u(x)} = \mathrm{O}\left(\frac{1}{|\mathrm{x}|}\right) \quad \text{if } |\mathrm{x}| \to \infty, \tag{3.1}$$

$$u = g \quad \text{on } \Omega, \tag{3.2}$$

where we assume g to be in $H_A^{\frac{1}{2}}(\Omega)$. In [27] the following weakly singular integral equation for the jump of the Neumann data was derived:

$$V_\Omega\left[\frac{\partial u}{\partial n}\right](x) := \frac{1}{2\pi}\int_\Omega \frac{1}{|x-y|}\left[\frac{\partial u}{\partial n}\right](y)d\Omega_y = 2g(x). \tag{3.3}$$

In [27] the invertibility of the operator $V_\Omega : \tilde{H}^{\frac{1}{2}}(\Omega) \to H^{1/2}(\Omega)$ is proved and Galerkin methods are investigated.

For the disk Ω and an axialsymmetric function g, equation (3.3) reads

$$\chi_\Omega S_0 \frac{1}{x} S_0 \left[\frac{\partial u}{\partial n}\right] = 2g, \tag{3.4}$$

where we look for $[\partial u/\partial n] \in \tilde{H}_A^{-\frac{1}{2}}(\Omega)$. To point out the relation to dual integral equations, we cite from [24] the following formal derivation of (3.4): The potential ansatz

$$u(\rho,z) = \int_0^\infty B(\xi)e^{-\xi|z|}J_0(\xi\rho)d\xi \tag{3.5}$$

leads to the dual integral equations

$$\int_0^\infty B(\xi)J_0(\xi\rho)d\xi = g(\rho), \quad 0 \le \rho \le 1; \qquad \int_0^\infty \xi B(\xi)J_0(\xi\rho)d\xi = 0, \quad \rho > 1. \tag{3.6}$$

The second equation in (3.6) gives us the information that the Hankel transform of B is supported in $[0,1]$. Using the first equation in (3.6), we end up with equation (3.4), where B denotes the Hankel transform of $[\partial u/\partial n]$.

We have to solve the Wiener–Hopf equation (3.4). This we shall do by "lifting" this equation to a Wiener–Hopf equation in $L_2^2((0,\infty))$. This method is well known for half-space problems; compare [8], [25]. The operators which map $\tilde{H}^{\frac{1}{2}}(\Omega)$ bijectively onto $L^2(\Omega)$ may be constructed by partition of unity [10]. In the axialsymmetric case this can be done more explicitly by use of the fractional integral operators introduced in §2.

THEOREM 3.1. *The explicit solution of equation* (3.4) *is given by*

$$\left[\frac{\partial u}{\partial n}\right] = -\frac{d}{dx}x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}(x^{\frac{1}{2}}I_{-\frac{1}{4},\frac{1}{2}})^{-1}2g. \tag{3.7}$$

Proof. We multiply equation (3.4) by $(x^{\frac{1}{2}}I_{-\frac{1}{4},\frac{1}{2}})^{-1}$ from the left and we use the ansatz $[\partial u/\partial n] = (-(d/dx)x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}})C$ with $C \in L_2^2((0,1))$. Then we end up with an equivalent equation in $L_2^2((0,1))$:

$$(x^{\frac{1}{2}}I_{-\frac{1}{4},\frac{1}{2}})^{-1}\chi_\Omega S_0\frac{1}{x}S_0\chi_\Omega\left(\frac{d}{dx}x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}\right)C = (x^{\frac{1}{2}}I_{-\frac{1}{4},\frac{1}{2}})^{-1}2g. \tag{3.8}$$

The operator on the left-hand side of the last equation is the unit operator in $L_2^2((0,1))$. We prove this by application of the Mellin transformations, assuming $C \in C_0^\infty((0,1))$.

(3.9)

$$\begin{aligned}
&-\left(M_{1,2}S_0\frac{1}{x}S_0\frac{d}{dx}x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}C\right)(t)\\
&\quad= -(M_{1,2}J_0)(t)\left(M\left(\frac{1}{x}S_0\frac{d}{dx}x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}\right)\right)\left(\frac{3}{2}-it\right)\\
&\quad= -(MJ_0)\left(\frac{1}{2}+it\right)(MJ_0)\left(\frac{1}{2}-it\right)\left(M\left(\frac{d}{dx}x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}C\right)\right)\left(\frac{3}{2}+it\right)\\
&\quad= -\frac{\Gamma(\frac{1}{4}+i\frac{t}{2})\Gamma(\frac{1}{4}-i\frac{t}{2})}{2\Gamma(\frac{3}{4}-i\frac{t}{2})\Gamma(\frac{3}{4}+i\frac{t}{2})}\left(M\left(\frac{d}{dx}x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}C\right)\right)\left(\frac{3}{2}+it\right)\\
&\quad= \left(\frac{1}{2}+it\right)\frac{\Gamma(\frac{1}{4}+i\frac{t}{2})\Gamma(\frac{1}{4}-i\frac{t}{2})}{2\Gamma(\frac{3}{4}+i\frac{t}{2})\Gamma(\frac{3}{4}-i\frac{t}{2})}(Mx^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}C)\left(\frac{1}{2}+it\right)\\
&\quad= \left(\frac{1}{4}+i\frac{t}{2}\right)\frac{\Gamma(\frac{1}{4}+i\frac{t}{2})\Gamma(\frac{1}{4}-i\frac{t}{2})}{\Gamma(\frac{5}{4}+i\frac{t}{2})\Gamma(\frac{3}{4}-i\frac{t}{2})}(MC)(1+it)\\
&\quad= \frac{\Gamma(\frac{1}{4}-i\frac{t}{2})}{\Gamma(\frac{3}{4}-i\frac{t}{2})}(MC)(1+it) = (M_{1,2}x^{\frac{1}{2}}I_{-\frac{1}{4},\frac{1}{2}}C)(t).
\end{aligned}$$

So we proved that the functions $x^{\frac{1}{2}}I_{-\frac{1}{4},\frac{1}{2}}C + S_0\frac{1}{x}S_0(d/dx)x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}C$ vanish identically on the whole half axis. □

The Neumann problem for the Laplacian reads

$$\Delta u = 0 \quad \text{in } \mathbf{R}^3 \setminus\overline{\Omega} \quad \mathrm{u(x)} = \mathrm{O}\left(\frac{1}{|\mathrm{x}|}\right) \quad \text{if } |\mathrm{x}| \to \infty, \tag{3.10}$$

$$\frac{\partial u}{\partial n} = h \quad \text{on } \Omega \tag{3.11}$$

where we assume h to be in $H_A^{-\frac{1}{2}}(\Omega)$. In [27] the following hypersingular integral equation for the jump of the Dirichlet data was analyzed:

$$D_\Omega[u](x) := \frac{1}{2\pi}\frac{\partial}{\partial n_x}\int_\Omega [u](y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}d\Omega_y = 2h(x). \tag{3.12}$$

For the disk Ω and an axialsymmetric function h this equation leads to the following Wiener–Hopf equation for $[u] \in \tilde{H}_A^{\frac{1}{2}}(\Omega)$:

$$\chi_\Omega S_0 x S_0 \chi_\Omega [u] = 2h. \tag{3.13}$$

By analogy to Theorem 3.1, which describes the solution of the Dirichlet problem, we can prove the following theorem.

THEOREM 3.2. *The explicit solution of equation* (3.13) *is given by*

$$[u] = x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}} (D x^{\frac{1}{2}} I_{\frac{1}{4},\frac{1}{2}})^{-1} 2h. \tag{3.14}$$

Proof. We multiply (3.13) by $(Dx^{\frac{1}{2}} I_{\frac{1}{4},\frac{1}{2}})^{-1}$ from the left and we use the ansatz $[u] = (x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}})d$ with $d \in L_2^2((0,1))$. Then we obtain an equivalent equation in $L_2^2((0,1))$:

$$(Dx^{\frac{1}{2}} I_{\frac{1}{4},\frac{1}{2}})^{-1} \chi_\Omega S_0 x S_0 \chi_\Omega (x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}}) d = (Dx^{\frac{1}{2}} I_{\frac{1}{4},\frac{1}{2}})^{-1} 2h. \tag{3.15}$$

The operator on the left-hand side of the last equation is the unit operator in $L_2^2((0,1))$. As in the proof of Theorem 3.1, it is sufficient to prove an identity for $d \in C_0^\infty((0,1))$:

$$\chi_\Omega S_0 x S_0 \chi_\Omega x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}} d = \chi_\Omega D x^{\frac{1}{2}} I_{\frac{1}{4},\frac{1}{2}} d. \tag{3.16}$$

To calculate the Mellin transform of $\chi_\Omega S_0 x S_0 \chi_\Omega x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}} d$ we give some comments in advance: The operator $x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}}$ maps $C_0^\infty((0,1))$ into $C_0^\infty((0,1)$; the operator $S_0 x S_0$ maps axialsymmetric functions from $S(\mathbb{R}^2)$ onto distributions from $S'(\mathbb{R}^2)$. Its Mellin symbol is derived by the use of the convolution theorem and the Fourier transform of the $S'(\mathbb{R})$ distribution $e^{3x/2} J_0(e^x)$.

$$\int_0^\infty s^{\frac{1}{2}+it} J_0(s)ds = \int_{\mathbb{R}} e^{(\frac{1}{2}+it)x} J_0(e^x)(e^x)dx = \int_{\mathbb{R}} e^{itx} J_0(e^x)(e^{3x/2})dx. \tag{3.17}$$

The function $e^{3x/2} J_0(e^x)$ grows exponentially as $x \longrightarrow +\infty$; nevertheless, it is an element of $S'(\mathbb{R})$. This follows from the equation

$$e^{3x/2} J_0(e^x) = \frac{1}{2}\left(e^{\frac{x}{2}} J_1(e^x)\right) + \frac{d}{dx}(e^{\frac{x}{2}} J_1(e^x)) \tag{3.18}$$

and the boundedness of $e^{\frac{x}{2}} J_1(e^x)$ at ±∞. Therefore, the Mellin transform $M_{3,2}$ of J_0 may be taken by holomorphic extension from the formula (1) on page 326 of [5]. We calculate the Mellin transform:

$$\begin{aligned}
&M_{3,2}(S_0 x S_0 x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}} d)(t) \\
&= (MJ_0)\left(\frac{3}{2} + it\right)(MJ_0)\left(\frac{3}{2} - it\right) M(x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}} d)\left(\frac{1}{2} + it\right) \\
&= 2\frac{\Gamma(\frac{3}{4} + i\frac{t}{2})\Gamma(\frac{3}{4} - i\frac{t}{2})}{\Gamma(\frac{1}{4} - i\frac{t}{2})\Gamma(\frac{1}{4} + i\frac{t}{2})} M(x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}} d)\left(\frac{1}{2} + it\right) \\
&= 2\frac{\Gamma(\frac{3}{4} - i\frac{t}{2})}{\Gamma(\frac{1}{4} - i\frac{t}{2})}(M(d))(1 + it).
\end{aligned} \tag{3.19}$$

Now we finish the proof:

$$
\begin{aligned}
(M_{3,2}Dx^{\frac{1}{2}}I_{\frac{1}{4},\frac{1}{2}}d)(t) &= \left(M\frac{d}{dx}x^{\frac{3}{2}}I_{\frac{1}{4},\frac{1}{2}}d\right)\left(\frac{1}{2}+it\right)\\
&= -\left(-\frac{1}{2}+it\right)(Mx^{\frac{3}{2}}I_{\frac{1}{4},\frac{1}{2}}d)\left(-\frac{1}{2}+it\right)\\
&= -\left(-\frac{1}{2}+it\right)\frac{\Gamma(\frac{3}{4}-i\frac{t}{2})}{\Gamma(\frac{5}{4}-i\frac{t}{2})}(Md)(1+it)\\
&= -\left(-\frac{1}{2}+it\right)\frac{\Gamma(\frac{3}{4}-i\frac{t}{2})}{(\frac{1}{4}-i\frac{t}{2})\Gamma(\frac{1}{4}-i\frac{t}{2})}(Md)(1+it)\\
&= 2\frac{\Gamma(\frac{3}{4}-i\frac{t}{2})}{\Gamma(\frac{1}{4}-i\frac{t}{2})}M(d)(1+it).
\end{aligned}
\tag{3.20}
$$

These formulas hold in the distributional sense. So we proved that for $d \in C_0^\infty((0,1))$ the distribution $S_0xS_0x^{\frac{1}{2}}K_{-\frac{1}{4},\frac{1}{2}}d - Dx^{\frac{1}{2}}I_{\frac{1}{4},\frac{1}{2}}d$ is equal to the null element of $S'(I\!R^2)$. □

Remark 3.1. Explicit solution formulas for the boundary value problems discussed in §3 for the Helmholtz equation are not known to the author. The methods used here to derive explicit formulas break down because of the nonhomogeneity of the Helmholtz operator. Nevertheless, in a future project we plan to use the methods developed here for investigation of numerical procedures for nonaxialsymmetric cases, for the Helmholtz equation and for time-harmonic crack problems.

4. Mixed boundary value problems. We want to formulate the Dirichlet–Neumann problem for the Laplacian here, which leads to a nontrivial system of Wiener–Hopf equations. This type of boundary value problem appears, for example, in the theory of acoustics; compare [21] and in the theory of subsonic flows, compare [13]. In Rawlins' paper the solution for the half-space case had been given explicitly. One essential step in Rawlins' paper was the derivation of the factorization of a certain matrix. For the discussion of this and related problems, [14], [20] should be consulted. In our case, we reduce the problem to the factorization of a meromorphic nonrational matrix function. This factorization is not explicitly known to us because the matrix does not fall into classes factorized in [1], [19]; therefore, we end up with the regularization of the Wiener–Hopf operator to a system of singular integral equations in $(L^2(I\!R))^2$ with a symbol from the Wiener-algebra.

The Dirichlet–Neumann problem for the Laplacian reads

$$\Delta u = 0 \quad \text{in } \mathrm{R}^3\backslash\overline{\Omega}, \quad \mathrm{u(x)} = \mathrm{O}\left(\frac{1}{\mathrm{|x|}}\right) \quad \text{if } |\mathrm{x}| \to \infty, \tag{4.1}$$

$$u(r,0^-) = g(r), \quad \frac{\partial u}{\partial n}(r,0^+) = h(r) \quad \text{for } r \in (0,1). \tag{4.2}$$

Here we assume $g \in H_A^{\frac{1}{2}}(\Omega)$ and $h \in H_A^{-\frac{1}{2}}(\Omega)$.

The ansatz

$$u(\rho,z) = \int_0^\infty \alpha(\xi)e^{-\xi z}J_0(\xi\rho)d\xi \quad \text{for } z > 0 \tag{4.3}$$

and

$$u(\rho, z) = \int_0^\infty \beta(\xi) e^{\xi z} J_0(\xi\rho) d\xi \quad \text{for } z \leq 0 \tag{4.4}$$

lead to the system of dual integral equations

$$\int_0^\infty \xi\alpha(\xi) J_0(\xi\rho) d\xi = -h(\rho), \qquad \int_0^\infty \beta(\xi) J_0(\xi\rho) d\xi = g(\rho), \quad 0 \leq \rho \leq 1, \tag{4.5}$$

and the transmission conditions imply

$$\int_0^\infty (\alpha(\xi) - \beta(\xi)) J_0(\xi\rho) d\xi = 0, \quad \int_0^\infty \xi(\alpha(\xi) + \beta(\xi)) J_0(\xi\rho) d\xi = 0, \quad \rho > 1. \tag{4.6}$$

The transmission conditions lead to the new ansatz:

$$2\alpha = S_0 \chi_\Omega a + x S_0 \chi_\Omega b, \qquad 2\beta = S_0 \chi_\Omega a - x S_0 \chi_\Omega b. \tag{4.7}$$

This ansatz for a and b reduces the mixed boundary value problem to the following system of Wiener–Hopf equations:

$$\chi_\Omega a + \chi_\Omega S_0 x S_0 \chi_\Omega b = -2h, \qquad \chi_\Omega S_0 \frac{1}{x} S_0 \chi_\Omega a - \chi_\Omega b = 2g. \tag{4.8}$$

Of course, we look for $a \in \tilde{H}_A^{-\frac{1}{2}}(\Omega)$ and for $b \in \tilde{H}_A^{\frac{1}{2}}(\Omega)$.

We "lift" this system of Wiener–Hopf equations to a system in $L_2^2((0,\infty))$: for this we introduce the substitutions

$$a = -\frac{d}{dx} x^{\frac{1}{2}} K_{\frac{1}{4},\frac{1}{2}} c, \quad b = x^{\frac{1}{2}} K_{-\frac{1}{4},\frac{1}{2}} d, \quad 2h = -D x^{\frac{1}{2}} I_{\frac{1}{4},\frac{1}{2}} u, \quad 2g = x^{\frac{1}{2}} I_{-\frac{1}{4},\frac{1}{2}} v. \tag{4.9}$$

Inserting these equations into the system above and using Theorems 3.1 and 3.2, we get the Wiener–Hopf system in $L_2^2((0,1))$ that we looked for:

$$\begin{aligned} -\chi_\Omega (D x^{\frac{1}{2}} I_{\frac{1}{4},\frac{1}{2}})^{-1} \frac{d}{dx} K_{\frac{1}{4},\frac{1}{2}} \chi_\Omega c + \chi_\Omega d &= \chi_\Omega u, \\ \chi_\Omega c - \chi_\Omega (I_{-\frac{1}{4},\frac{1}{2}})^{-1} K_{-\frac{1}{4},\frac{1}{2}} \chi_\Omega d &= \chi_\Omega v. \end{aligned} \tag{4.10}$$

An application of the Mellin transformation leads to

$$\begin{aligned} P\frac{1}{r}(t) M_{2,2} c(1+it) + P M_{2,2} d(1+it) &= P M_{2,2} u(1+it), \\ P M_{2,2} c(1+it) - P r(t) M_{2,2} d(1+it) &= P M_{2,2} v(1+it), \end{aligned} \tag{4.11}$$

where P denotes the projector onto L^2 functions that are boundary values of functions that are holomorphically extendable into the lower half plane, and r denotes the Mellin symbol of the operator $(I_{-\frac{1}{4},\frac{1}{2}})^{-1} K_{-\frac{1}{4},\frac{1}{2}}$.

$$r(t) := \frac{\Gamma(\frac{1}{4} + i\frac{t}{2})\Gamma(\frac{3}{4} - i\frac{t}{2})}{\Gamma(\frac{3}{4} + i\frac{t}{2})\Gamma(\frac{1}{4} - i\frac{t}{2})} = \frac{\sin((\frac{3}{4} + i\frac{t}{2})\pi)}{\sin((\frac{3}{4} - i\frac{t}{2})\pi)}. \tag{4.12}$$

This symbol had already been calculated by Rooney [23]. It is a meromorphic function with all its poles and zeros lying on the line $i\mathbb{R}$. A lengthy but simple calculation proves that the operator $-(Dx^{\frac{1}{2}}I_{\frac{1}{4},\frac{1}{2}})^{-1}\frac{d}{dx}x^{\frac{1}{2}}K_{\frac{1}{4},\frac{1}{2}}$ has the Mellin symbol $\frac{1}{r}$. The matrix

$$\sigma(t) := \begin{pmatrix} \frac{1}{r} & 1 \\ 1 & -r \end{pmatrix} \tag{4.13}$$

is an element of $(L^\infty(I\!R))^{2\times 2}$, having constant determinant. The limit values of r at infinity are given by:

$$\lim_{t\to\pm\infty} r(t) = \pm i. \tag{4.14}$$

It is a matrix of functions that have a finite jump at infinity. The matrix

$$\sigma_1(t) := \begin{pmatrix} 1 & \frac{1}{r} \\ -r & 1 \end{pmatrix} \tag{4.15}$$

is a dissipative matrix in the sense of [2]. Transformation of the results from [2] to the real line allows us to use Banach's fixed point principle for the solution of (4.11). We believe that it is more convenient to give a reduction to a system of singular integral equations with a matrix being continuous at infinity. If we solve this reduced equation by a numerical method, we get an approximate solution of the boundary value problem which has the same singularities like the exact solution. We state the following theorem.

THEOREM 4.1. *The solution of the axialsymmetric mixed boundary value problem* (4.1), (4.2) *is equivalent to the solution of a system of singular integral equations in* $(L^2(I\!R))^2$:

$$P\sigma_0 P u = Pf, \tag{4.16}$$

where the matrix σ_0 *is defined by*

$$\sigma_0(t) = \begin{pmatrix} -r_1(1+\frac{1}{2}(\frac{1}{r}-r)) & -\frac{r_1}{2}(\frac{1}{r}+r) \\ \frac{r_1}{2}(\frac{1}{r}+r) & -r_1(1-\frac{1}{2}(\frac{1}{r}-r)) \end{pmatrix}, \tag{4.17}$$

and r_1 *is defined by*

$$r_1(t) := \frac{\sin((\frac{1}{2}+i\frac{t}{2})\pi)}{\sin((\frac{1}{4}+i\frac{t}{2})\pi)}. \tag{4.18}$$

The matrix $\sigma_0(t)$ *converges to the identity matrix at infinity.*

Proof. We transform (4.11) equivalently by multiplying the matrix σ by the constant matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{4.19}$$

from the right and its inverse from the left to get the new matrix σ_2, defined by

$$\begin{pmatrix} -(1+\frac{1}{2}(\frac{1}{r}-r)) & -\frac{1}{2}(\frac{1}{r}+r)) \\ -\frac{1}{2}(\frac{1}{r}+r)) & (1-\frac{1}{2}(\frac{1}{r}-r)) \end{pmatrix}. \tag{4.20}$$

The function $(r+\frac{1}{r})(t)$ vanishes exponentially for $t \to \pm\infty$. A standard method to reduce the system of singular integral equations with the matrix having a discontinuity at infinity is to fill up the discontinuities by multiplying the matrix with diagonal matrices with entries like $(t \pm i)^\alpha$. Here we multiply the matrix with meromorphic functions that have algebraic behaviour at infinity on the real axis:

$$\sigma_0(t) = \begin{pmatrix} \frac{\Gamma(\frac{3}{4}-i\frac{t}{2})}{\Gamma(\frac{1}{2}-i\frac{t}{2})} & 0 \\ 0 & -\frac{\Gamma(\frac{3}{4}-i\frac{t}{2})}{\Gamma(\frac{1}{2}-i\frac{t}{2})} \end{pmatrix} \sigma_1(t) \begin{pmatrix} \frac{\Gamma(\frac{1}{4}+i\frac{t}{2})}{\Gamma(\frac{1}{2}+i\frac{t}{2})} & 0 \\ 0 & \frac{\Gamma(\frac{1}{4}+i\frac{t}{2})}{\Gamma(\frac{1}{2}+i\frac{t}{2})} \end{pmatrix}. \tag{4.21}$$

We multiplied σ_1 from the left side with a matrix holomorphically extendable into the upper half plane, while the last matrix in the product is holomorphically extendable to the lower half plane. Both matrices have algebraic behaviour at infinity in their half planes of analyticity. By using the well-known formula

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}, \tag{4.22}$$

we get the representation (4.17) for σ_0. By the methods described in [15], the solution of (4.11) may be reduced to (4.16). The diagonal elements of σ_0 converge to +1, which is seen from the formula (4.14) and the following one:

$$\lim_{t\to\pm\infty} \frac{\sin(\alpha + it)}{\sin(\gamma + it)} = e^{\pm i(\alpha-\gamma)\pi}. \tag{4.23}$$

The off-diagonal elements of σ_0 vanish exponentially at infinity on the real axis, which follows from the behaviour of $r + \frac{1}{r}$ and the boundedness of r_1 at infinity. □

REFERENCES

[1] I. D. Abrahams and G. R. Wickham, *On the scattering of sound by two semi-infinite parallel staggered plates* I. *Explicit matrix Wiener-Hopf factorization*, Proc. Roy. Soc. London., Ser. A420 (1988), pp. 131–156.

[2] K. Clancey and I. Gohberg, *Factorization of matrix functions and Singular Integral operators*, Birkhäuser Verlag, Basel, Switzerland, 1981.

[3] E. T. Copson, *On the problem of the electrified disk*, Proc. Edinburgh Math. Soc., 8 (1947), pp. 14–19.

[4] M. Costabel and E. Stephan, *A direct boundary integral equation method for transmission problems*, J. Math. Anal. Appl., 106 (1985), pp. 367–413.

[5] A. Erdelyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Tables of integral transforms*, Volume 1, McGraw-Hill, New York, 1954.

[6] V. Ervin and E. Stephan, *Experimental convergence of boundary element methods for the capacity of the electrified plate*, Boundary Elements, 9 (1987), pp. 167–175.

[7] V. J. Ervin, E. P. Stephan, and S. Abou el-Seoud, *An improved boundary element method for the charge density of a thin electrified plate in* $\mathbb{R}^3$, Math. Methods Appl. Sci., 13 (1990), pp. 291–303.

[8] G. I. Eskin, *Boundary value problems for elliptic pseudodifferential operators*, American Mathematical Society, Providence, RI, 1981.

[9] I. M. Gelfand and G. E. Shilov, *Verallgemeinerte Funktionen* (*Distributionen*), Akademie-Verlag, Berlin, First ed., 1960.

[10] L. Hörmander, *The analysis of linear partial differential operators* III, Springer-Verlag, New York, 1985.

[11] A. C. Mac Bride, *Fractional calculus and integral transforms of generalized functions*, Pitman, London, 1979.

[12] E. Meister, *Integraltransformationen mit Anwendungen auf Probleme der mathematischen Physik*, Verlag Peter Lang, First ed., Frankfurt, 1983.

[13] ———, *Some mixed boundary value problems in the theory of subsonic flow past oscillating profiles*, in Complex analysis and its applications, in Honour of Acad. I.N.Vekua's 70th anniversary, Akad. Nauk SSSR Stekhlov Math. Inst. Izdatelstvo "Nauka," Moscow, 1978, pp. 346–362.

[14] E. MEISTER AND F. O. SPECK, *Modern Wiener–Hopf methods in diffraction theory*, Proc. Conf. Dundee, 1988, in Ordinary and Partial Differential Equations, B. Sleeman and R. Jarvis, eds., 2 (1989), pp. 130–171.

[15] S. G. MICHLIN AND S. PRÖSSDORF, *Singular Integral Operators*, Springer-Verlag, Berlin, 1986.

[16] L. PAIVARINTA AND S. REMPEL, *A deconvolution problem with the kernel* $\frac{1}{|x|}$ *on the plane*, Appl. Anal., 26 (1987), pp. 105–128.

[17] F. PENZEL, *Fredholmeigenschaften dualer Integralgleichungen*, Z. Angew. Math. Mech., 68 (1988), pp. T476–T478.

[18] ———, *On the theory of generalized Abel integral equations*, Integral Equations Operator Theory, 10 (1987), pp. 595–620.

[19] S. PRÖSSDORF AND F. O. SPECK, *A factorization procedure for two by two matrix functions on the circle with two rationally independent entries*, Proc. Roy. Soc. Edinburgh Sect. A, 118 (1990), pp. 119–138.

[20] A. D. RAWLINS, *The explicit Wiener–Hopf factorization of a special matrix*, Z. Angew. Math. Mech., 61 (1981), pp. 527–528.

[21] ———, *The solution of a mixed boundary value problem in the theory of diffraction by a semi-infinite plane*, Proc. Roy. Soc. London Ser. A, 346 (1975), pp. 469–484.

[22] E. Q. RONG, *A new solution for the space crack problem from hypersingular integral equation*, Appl. Anal., 31 (1988), pp. 91–102.

[23] P. G. ROONEY, *On the ranges of certain fractional integrals*, Canad. J. Math., 24 (1972), pp. 1198–1216.

[24] I. N. SNEDDON, *Mixed boundary value problems in potential theory*, North-Holland, Amsterdam, 1966.

[25] F. O. SPECK, *Sommerfeld diffraction problems with first and second kind boundary conditions*, SIAM J. Math. Anal., 20 (1989), pp. 1–12.

[26] R. P. SRIVASTAV, *Dual integral equations with trigonometric kernels and tempered distributions*, SIAM J. Math. Anal., 3 (1972), pp. 413–421.

[27] E. STEPHAN, *Boundary integral equations for screen probems in* $I\!R^3$, Integral Equations Operator Theory, 10 (1987), pp. 236–257.

[28] J. R. WALTON, *The question of uniqueness for dual integral equations of Titchmarsh type*, Proc. Roy. Soc. Edinburgh Sect. A., 76A (1977), pp. 267–282.

[29] P. WOLFE, *An integral operator arising in potential theory*, Appl. Anal., 10 (1980), pp. 71–80.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1182-1188, September 1992

ENERGY INEQUALITIES FOR INTEGRO-PARTIAL DIFFERENTIAL EQUATIONS WITH RIEMANN-LIOUVILLE INTEGRALS*

YASUHIRO FUJITA†

Abstract. This paper presents energy inequalities for the integro-partial differential equations with the Riemann-Liouville integrals. These equations interpolate between the heat equation and the wave equation. This fact is reflected in the energy inequalities so that they correspond to the energy equality for the wave equation. The proofs depend on the Fourier analysis and the probability methods.

Key words. energy inequality, Mittag-Leffler distribution, fractional derivative

AMS(MOS) subject classifications. 45K05, 26A33

1. Introduction.

Let $n \geqq 1$ be an integer and $1 \leqq \alpha \leqq 2$. We study the integro-partial differential equation

$$\text{(IDE)} \qquad u(t,x)=\phi(x)+\frac{t^{\alpha/2}}{\Gamma(1+(\alpha/2))}\psi(x)+\frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\Delta u(s,x)\,ds \qquad (t>0, x\in \mathbf{R}^n),$$

where $\Gamma(x)$ is the gamma function and $\Delta=\sum_{j=1}^n (\partial/\partial x_j)^2$. The integral appeared in (IDE) is the Riemann-Liouville integral of order α defined by

$$I^\alpha f(t)=\frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}f(s)\,ds.$$

The integro-partial differential equation (IDE) interpolates between the heat equation ($\alpha=1$ and $\psi\equiv 0$) and the wave equation ($\alpha=2$); (IDE) is interpreted as the integral form of the formal Cauchy problem $(\partial/\partial t)^\alpha u(t,x)=\Delta u(t,x)$. Several authors studied the qualitative properties of the solution of (IDE) [8], [9], [12], however, about the quantitative properties of it, only $L^p(\mathbf{R}^n)$-decay ($p\geqq 2$) was studied for $\psi\equiv 0$ [10].

The aim of the present paper is to derive energy inequalities for (IDE) ($1\leqq\alpha\leqq 2$). For the solution u_α of (IDE), these energy inequalities deal with the quantity

$$(1) \qquad \mathscr{E}_{m,\alpha}(t)=\|D^{\alpha/2}u_\alpha(t)\|_m^2+\|\nabla u_\alpha(t)\|_m^2$$

and its time average

$$(2) \qquad \mathscr{T}_{m,\alpha}(\lambda)=\lambda^{1/\alpha}\int_0^\infty e^{-t\lambda^{1/\alpha}}\mathscr{E}_{m,\alpha}(t)\,dt \qquad (\lambda>0).$$

Here $\|\cdot\|_m$ is the Sobolev norm of order m and $\|\nabla\phi\|_m^2=\sum_{j=1}^n\|\partial\phi/\partial x_j\|_m^2$; $D^{\alpha/2}$ is the fractional differential operator of order $\alpha/2 (1\leqq\alpha<2)$ and $D^1=(\partial/\partial t)$ (see §2). As far as we know, there exists no paper treating the quantities $\mathscr{E}_{m,\alpha}(t)$ and $\mathscr{T}_{m,\alpha}(\lambda)$ except $\alpha=2$. For $\alpha=2$, the following energy equality is widely known:

$$(3) \qquad \mathscr{E}_{m,2}(t)=\|\psi\|_m^2+\|\nabla\phi\|_m^2 \qquad (t>0).$$

The main results are as follows. In Theorem 1, we derive the energy inequality for $\mathscr{E}_{m,\alpha}(t)$ $(1\leqq\alpha<2)$. This inequality corresponds to the energy equality (3). In

* Received by the editors July 3, 1991; accepted for publication (in revised form) October 27, 1991.

† Department of Mathematics, Faculty of Science, Toyama University, Toyama 930, Japan. The author was partially supported by the Grant-in-Aid for Encouragement of Young Scientists (No. 02740103) by the Ministry of Education, Science, and Culture of Japan.

Theorem 2 we treat the asymptotic behavior of $\mathscr{E}_{m,\alpha}(t)$ as $t\to\infty$ for $1\leqq\alpha<2$. In Theorem 3, we show that $\mathscr{T}_{m,\alpha}(\lambda)$ is a continuous and strictly increasing function of $\alpha\in[1,2]$. Thus the time average $\mathscr{T}_{m,\alpha}(\lambda)$ interpolates monotonically and continuously between $\mathscr{T}_{m,1}(\lambda)$ and $\mathscr{T}_{m,2}(\lambda)$.

There exist many papers about the estimates for the solutions of (Volterra) integro-partial differential equations. (Cf. [3], [4], [5], [11].) As compared with these estimates, our energy inequalities are unique in the sense that they correspond to the energy equality for the wave equation. It seems, however, to be difficult to generalize these inequalities to other equations.

The present paper is organized as follows: we state the main results in §2 and prove them in §3.

2. Main results. Let $H^\infty(\mathbf{R}^n)$ be the Fréchet space consisting of C^∞-functions ϕ such that ϕ and all its derivatives belong to $L^2(\mathbf{R}^n)$; $H^\infty(\mathbf{R}^n)$ is equipped with the sequence of norms $\{\|\cdot\|_m\}_{m=0}^\infty$ defined by

$$\|\phi\|_m=\left\{\int_{\mathbf{R}^n}(1+|\xi|^2)^m|\hat\phi(\xi)|^2\,d\xi\right\}^{1/2},$$

where $\hat\phi$ is the Fourier transform of ϕ in $L^2(\mathbf{R}^n)$:

$$\hat\phi(\xi)=\lim_{A\to\infty}\frac{1}{(2\pi)^{n/2}}\int_{\{|x|\leqq A\}}e^{-ix\cdot\xi}\phi(x)\,dx\quad\text{in } L^2(\mathbf{R}^n).$$

Throughout this paper we assume that ϕ and ψ of (IDE) belong to $H^\infty(\mathbf{R}^n)$.

Definition. The function u in $C([0,\infty)$: $H^\infty(\mathbf{R}^n))$ is said to be a solution of (IDE) if it satisfies (IDE) for every $t>0$ and $x\in\mathbf{R}^n$.

Proposition. *For* $1\leqq\alpha\leqq2$, (IDE) *has a unique solution* u_α.

Now we consider the energy inequalities for (IDE). These inequalities deal with $\mathscr{E}_{m,\alpha}(t)$ and $\mathscr{T}_{m,\alpha}(\lambda)$ defined by (1) and (2), respectively. For $1\leqq\alpha<2$, the fractional differential operator $D^{\alpha/2}$ of order $\alpha/2$ is defined by

$$D^{\alpha/2}f(t)=I^{1-(\alpha/2)}f'(t)=\frac{1}{\Gamma(1-\alpha/2)}\int_0^t(t-s)^{-\alpha/2}f'(s)\,ds.$$

For $\alpha=2$, put $D^{\alpha/2}=D^1=(\partial/\partial t)$. The following inequality for the quantity $\mathscr{E}_{m,\alpha}(t)$ corresponds to the energy equality (3).

Theorem 1. *For* $1\leqq\alpha\leqq2$, *the function* $D^{\alpha/2}u_\alpha$ *is well defined as an element of* $C([0,\infty)$: $H^\infty(\mathbf{R}^n))$. *In addition, for each integer* $m\geqq0$, *we have*

$$\mathscr{E}_{m,\alpha}(t)\leqq\|\psi\|_m^2+\|\nabla\phi\|_m^2,\qquad(t>0).\tag{4}$$

The inequality (4) *reduces to the equality for all* $\phi,\psi\in H^\infty(\mathbf{R}^n)$ *if and only if* $\alpha=2$.

Theorem 2 below treats the asymptotic behavior of $\mathscr{E}_{m,\alpha}(t)$ as $t\to\infty$ for $1\leqq\alpha<2$.

Theorem 2. *Let* $m\geqq0$ *be an integer. Suppose that there exists* $\chi\in H^\infty(\mathbf{R}^n)$ *such that* $\hat\psi(\xi)=|\xi|\hat\chi(\xi)$ almost everywhere. *Then, for each* $1\leqq\alpha<2$, *we have*

$$\lim_{t\to\infty}t^\alpha\mathscr{E}_{m,\alpha}(t)=\frac{1}{\Gamma(1-(\alpha/2))^2}(\|\chi\|_m^2+\|\phi\|_m^2).\tag{5}$$

Remark 1. As an example of ψ satisfying the assumption of Theorem 2, we give $\psi(x)=\sum_{j=1}^n a_j\partial\Psi_j/\partial x_j(x)$ for some constants a_j and $\Psi_j\in H^\infty(\mathbf{R}^n)$ $(1\leqq j\leqq n)$. In this case $\hat\chi(\xi)$ is so chosen that

$$\hat\chi(\xi)=\sum_{j=1}^n a_j\frac{i\xi_j}{|\xi|}\hat\Psi_j(\xi)\quad(\xi\neq0),\qquad=0\quad(\xi=0).$$

Clearly this χ belongs to $H^\infty(\mathbf{R}^n)$.

Next we consider the time average $\mathscr{T}_{m,\alpha}(\lambda)$. As for $\mathscr{E}_{m,\alpha}(t)$, we cannot answer whether it is a monotonic function of α because we know no method comparing $\mathscr{E}_{m,\alpha}(t)$ with $\mathscr{E}_{m,\beta}(t)$ for $1 \leqq \alpha < \beta \leqq 2$. However, as for the time average, the Laplace transforms enable us to show the following.

THEOREM 3. *Let $m \geqq 0$ be an integer and $\lambda > 0$. Unless $\phi = \psi \equiv 0$, then $\mathscr{T}_{m,\alpha}(\lambda)$ is a continuous and strictly increasing function of α. That is, $\mathscr{T}_{m,\alpha}(\lambda)$ is continuous in $\alpha \in [1, 2]$, and the following inequalities hold whenever $1 < \alpha < \beta < 2$:*

$$\mathscr{T}_{m,1}(\lambda) < \mathscr{T}_{m,\alpha}(\lambda) < \mathscr{T}_{m,\beta}(\lambda) < \mathscr{T}_{m,2}(\lambda) \equiv \|\psi\|_m^2 + \|\nabla\phi\|_m^2. \tag{6}$$

Thus, there exists a one-to-one correspondence between the two intervals $(1, 2)$ and $(\mathscr{T}_{m,1}(\lambda), \mathscr{T}_{m,2}(\lambda))$.

Theorem 3 shows that the time average $\mathscr{T}_{m,\alpha}(\lambda)$ interpolates monotonically and continuously between $\mathscr{T}_{m,1}(\lambda)$ and $\mathscr{T}_{m,2}(\lambda)$.

3. Proofs. In the following, let $Y_\alpha(t) = Y_\alpha(t, \omega)$ $(1 \leqq \alpha \leqq 2)$ be the stochastic process on a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ with the Mittag-Leffler distributions of order $\alpha/2$:

$$\mathbf{E}\exp\{-sY_\alpha(t)\} = \sum_{k=0}^{\infty} \frac{(-st^{\alpha/2})^k}{\Gamma(1+(k\alpha/2))}, \qquad \operatorname{Re} s \geqq 0,\, t \geqq 0, \tag{7}$$

where $\mathbf{E}$ stands for the expectation. Then $Y_\alpha(t)$ has the continuous path with probability 1. That is, for almost all $\omega \in \Omega$, the functions $t \to Y_\alpha(t, \omega)$ are continuous for all $t \geqq 0$. These facts were proved in [1] (see also [9]). The use of the stochastic process $Y_\alpha(t)$ enables us to simplify the proofs.

Proof of Proposition. For $t \geqq 0$, put

$$U_\alpha(t, \xi) = \hat{\phi}(\xi)\mathbf{E}\cos(|\xi| Y_\alpha(t)) + \hat{\psi}(\xi)\frac{\mathbf{E}\sin(|\xi| Y_\alpha(t))}{|\xi|}. \tag{8}$$

The function $U_\alpha(t, \xi)$ is defined, except ξ belonging to a null set. Since ϕ and ψ are in $H^\infty(\mathbf{R}^n)$, their Fourier transforms $\hat{\phi}$ and $\hat{\psi}$ belong to $L^1(\mathbf{R}^n)$. Thus, $U_\alpha(t, \xi)$ also belongs to $L^1(\mathbf{R}^n)$ for each $t \geqq 0$ because we have by (7) and (8) for almost every ξ,

$$|U_\alpha(t, \xi)| \leqq |\hat{\phi}(\xi)| + |\hat{\psi}(\xi)|\mathbf{E}Y_\alpha(t) = |\hat{\phi}(\xi)| + \frac{t^{\alpha/2}}{\Gamma(1+(\alpha/2))}|\hat{\psi}(\xi)|. \tag{9}$$

Now define $u_\alpha(t, x) (t \geqq 0, x \in \mathbf{R}^n)$ by

$$u_\alpha(t, x) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbf{R}^n} e^{ix\cdot\xi} U_\alpha(t, \xi)\, d\xi. \tag{10}$$

We show that u_α is a unique solution of (IDE). For each integer $m \geqq 0$, it follows from (9), (10), and Parseval's theorem that

$$\|u_\alpha(t)\|_m^2 \leqq 2\|\phi\|_m^2 + \frac{2t^\alpha}{\Gamma(1+(\alpha/2))^2}\|\psi\|_m^2.$$

Further, since $Y_\alpha(t)$ has the continuous path, we get $\|u_\alpha(t) - u_\alpha(s)\|_m \to 0$ as $|t - s| \to 0$ by (8) and the dominated convergence theorem. Thus, u_α belongs to $C([0, \infty): H^\infty(\mathbf{R}^n))$.

Next we show that u_α satisfies (IDE). By (7), we have for every $\xi \in \mathbf{R}^n$ and $t \geqq 0$,

$$\mathbf{E}\cos(|\xi| Y_\alpha(t)) = \sum_{k=0}^{\infty} \frac{(-|\xi|^2 t^\alpha)^k}{\Gamma(1+k\alpha)} \tag{11}$$

and

$$\mathbf{E}\sin(|\xi|\,Y_\alpha(t)) = |\xi|t^{\alpha/2}\sum_{k=0}^{\infty}\frac{(-|\xi|^2t^\alpha)^k}{\Gamma(1+k\alpha+(\alpha/2))}. \tag{12}$$

Thus we can rewrite (8) as

$$U_\alpha(t,\xi)=\hat{\phi}(\xi)\sum_{k=0}^{\infty}\frac{(-|\xi|^2t^\alpha)^k}{\Gamma(1+k\alpha)}+t^{\alpha/2}\hat{\psi}(\xi)\sum_{k=0}^{\infty}\frac{(-|\xi|^2t^\alpha)^k}{\Gamma(1+k\alpha+\alpha/2)}.$$

It is easy to see that $U_\alpha(\cdot,\xi)$ is a unique solution of the integral equation

$$v(t,\xi)=\hat{\phi}(\xi)+\frac{t^{\alpha/2}}{\Gamma(1+(\alpha/2))}\hat{\psi}(\xi)-\frac{|\xi|^2}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}v(s,\xi)\,ds \tag{13}$$

for almost every $\xi\in\mathbf{R}^n$; therefore, the inverse Fourier transform shows that u_α defined by (10) satisfies (IDE) for every $t>0$ and $x\in\mathbf{R}^n$, so that u_α is a solution of (IDE). The uniqueness of the solution of (IDE) follows from the fact that the integral equation (13) has a unique solution for almost every $\xi\in\mathbf{R}^n$. This completes the proof of the Proposition. □

Remark 2. Let $w(=u_2)$ be the solution of the wave equation. By (8), we have $U_\alpha(t,\xi)=\mathbf{E}\hat{w}(Y_\alpha(t),\xi)$. Thus the solution u_α $(1\leqq\alpha\leqq 2)$ is expressed by

$$u_\alpha(t,x)=\mathbf{E}w(Y_\alpha(t),x).$$

This expression was given by [9] for $n=1$. Another expression was given by [12] for $\psi\equiv 0$.

Proof of Theorem 1. For each $r>0$ we remark that

$$D^{\alpha/2}[1]=0,\qquad D^{\alpha/2}[t^r]=\frac{\Gamma(1+r)}{\Gamma(1+r-(\alpha/2))}t^{r-\alpha/2}.$$

Then, we get, by (11) and (12),

$$\begin{aligned} D^{\alpha/2}[\mathbf{E}\cos(|\xi|\,Y_\alpha(t))] &= \sum_{k=1}^{\infty}\frac{(-|\xi|^2)^k}{\Gamma(1+k\alpha)}D^{\alpha/2}[t^{k\alpha}]\\ &=\sum_{k=1}^{\infty}\frac{(-|\xi|^2)^k}{\Gamma(1+(k-1)\alpha+(\alpha/2))}t^{(k-1)\alpha+(\alpha/2)}\\ &=-|\xi|\mathbf{E}\sin(|\xi|\,Y_\alpha(t)). \end{aligned}\tag{14}$$

The interchange of $D^{\alpha/2}$ and $\sum$ in (14) is permitted, since the third term of (14) converges absolutely. Similarly we get, by (11) and (12),

$$D^{\alpha/2}[\mathbf{E}\sin(|\xi|\,Y_\alpha(t))]=|\xi|\mathbf{E}\cos(|\xi|\,Y_\alpha(t)). \tag{15}$$

We have, by (8), (14), and (15),

$$D^{\alpha/2}U_\alpha(t,\xi)=-|\xi|\hat{\phi}(\xi)\mathbf{E}\sin(|\xi|\,Y_\alpha(t))+\hat{\psi}(\xi)\mathbf{E}\cos(|\xi|\,Y_\alpha(t)).$$

By (10), it is easy to see that $D^{\alpha/2}u_\alpha$ is well defined as an element of $C([0,\infty)\colon H^\infty(\mathbf{R}^n))$, and given by

$$D^{\alpha/2}u_\alpha(t,x)=\frac{1}{(2\pi)^{n/2}}\int_{\mathbf{R}^n}e^{ix\cdot\xi}D^{\alpha/2}U_\alpha(t,\xi)\,d\xi.$$

Similarly

$$\frac{\partial u_\alpha}{\partial x_j}(t,x)=\frac{1}{(2\pi)^{n/2}}\int_{\mathbf{R}^n}e^{ix\cdot\xi}i\xi_jU_\alpha(t,\xi)\,d\xi.$$

It follows from Parseval's theorem that

$$\begin{aligned}\mathscr{E}_{m,\alpha}(t) &\equiv \|D^{\alpha/2}u_\alpha(t)\|_m^2 + \|\nabla u_\alpha(t)\|_m^2 \\ &= \int_{\mathbf{R}^n} [|\xi|^2|\hat{\phi}(\xi)|^2 + |\hat{\psi}(\xi)|^2] \\ &\quad \cdot (1+|\xi|^2)^m [|\mathbf{E}\cos(|\xi|\,Y_\alpha(t))|^2 + |\mathbf{E}\sin(|\xi|\,Y_\alpha(t))|^2]\,d\xi.\end{aligned} \tag{16}$$

Since

$$\begin{aligned}&|\mathbf{E}\cos(|\xi|\,Y_\alpha(t))|^2 + |\mathbf{E}\sin(|\xi|\,Y_\alpha(t))|^2 \\ &\leqq \mathbf{E}\cos^2(|\xi|\,Y_\alpha(t)) + \mathbf{E}\sin^2(|\xi|\,Y_\alpha(t)) = 1 \qquad (t>0,\ \xi\in\mathbf{R}^n),\end{aligned} \tag{17}$$

we obtain the inequality (4). It remains to show that the inequality (4) reduces to the equality for all ϕ, $\psi \in H^\infty(\mathbf{R}^n)$ if and only if $\alpha = 2$. The "if part" is the energy equality (3). The "only if part" is proved as follows. In order that (4) reduces to the equality for all ϕ, $\psi\in H^\infty(\mathbf{R}^n)$, it is necessary that (17) reduces to the equality for all $\xi\in\mathbf{R}^n$, which is clearly equivalent to

$$\cos(|\xi|\,Y_\alpha(t)) = \mathbf{E}\cos(|\xi|\,Y_\alpha(t)), \qquad \sin(|\xi|\,Y_\alpha(t)) = \mathbf{E}\sin(|\xi|\,Y_\alpha(t)) \quad (t\geqq 0)$$

for all $\xi\in\mathbf{R}^n$ with probability 1. Thus,

$$Y_\alpha(t) = \lim_{|\xi|\to 0}\frac{\sin(|\xi|\,Y_\alpha(t))}{|\xi|} = \lim_{|\xi|\to 0}\frac{\mathbf{E}\sin(|\xi|\,Y_\alpha(t))}{|\xi|} = \mathbf{E}Y_\alpha(t) = \frac{t^{\alpha/2}}{\Gamma(1+(\alpha/2))}.$$

To complete the proof of Theorem 1, we need to show that if $Y_\alpha(t) = t^{\alpha/2}/\Gamma(1+(\alpha/2))$ $(t\geqq 0)$ $\mathbf{P}$-a.s., then $\alpha = 2$. This follows from the following lemma. This completes the proof of Theorem 1. □

LEMMA 1. *If there exists a nonrandom continuous function f on* $[0,\infty)$ *such that* $Y_\alpha(t) = f(t)$ *on* $[0,\infty)$ $\mathbf{P}$-a.s., *then* $\alpha = 2$ *and* $f(t) = t$.

Proof. By (7) and the assumption, we have

$$\sum_{k=0}^\infty \frac{(-st^{\alpha/2})^k}{\Gamma(1+(k\alpha/2))} = \mathbf{E}\exp\{-sY_\alpha(t)\} = \exp\{-sf(t)\} = \sum_{k=0}^\infty \frac{(-sf(t))^k}{k!}.$$

Thus

$$\frac{f(t)^k}{k!} = \frac{t^{k\alpha/2}}{\Gamma(1+(k\alpha/2))} \qquad (k=0,1,2,\cdots).$$

Then it is easy to see that $\alpha = 2$, so that $f(t) = t$. This completes the proof of Lemma 1. □

Proof of Theorem 2. Let $1\leqq\alpha<2$. By (11), (12), and [6, chap. 18, § 1 (21), p. 210], we have

$$\lim_{r\to\infty} r^2\,\mathbf{E}\cos(rY_\alpha(1)) = \frac{1}{\Gamma(1-\alpha)} \tag{18}$$

and

$$\lim_{r\to\infty} r\,\mathbf{E}\sin(rY_\alpha(1)) = \frac{1}{\Gamma(1-(\alpha/2))}. \tag{19}$$

For $\alpha = 1$, we interpret that $1/\Gamma(1-\alpha) = 0$. In this case the equality (18) still holds, since $r^2\,\mathbf{E}\cos(rY_1(1)) = r^2\exp\{-r^2\}$. Since $Y_\alpha(t) = t^{\alpha/2}Y_\alpha(1)$ $(t\geqq 0)$ in distribution, we have, by (16) and the assumption of Theorem 2,

$$\begin{aligned}t^\alpha\mathscr{E}_{m,\alpha}(t) &= \int_{\mathbf{R}^n} [|\hat{\chi}(\xi)|^2 + |\hat{\phi}(\xi)|^2](1+|\xi|^2)^m \\ &\quad \cdot [|\xi|^2 t^\alpha |\mathbf{E}\cos(|\xi|t^{\alpha/2}Y_\alpha(1))|^2 + |\xi|^2 t^\alpha |\mathbf{E}\sin(|\xi|t^{\alpha/2}Y_\alpha(1))|^2]\,d\xi.\end{aligned}$$

Then the desired result follows from (18), (19), and the dominated convergence theorem. This completes the proof of Theorem 2. □

To prove Theorem 3, we need a lemma, which was established in [9].

LEMMA 2. *Let $1\leqq\alpha\leqq 2$. For every bounded and continuous function $v(\cdot)$ on $[0,\infty)$, it holds that*

$$\int_0^\infty e^{-\lambda t}\mathbf{E}v(Y_\alpha(t))\,dt=\lambda^{(\alpha/2)-1}\int_0^\infty e^{-y\lambda^{\alpha/2}}v(y)\,dy \qquad (\lambda>0).$$

Proof of Theorem 3. First we show that if $1\leqq\alpha\leqq\beta\leqq 2$, then $\mathcal{T}_{m,\alpha}(\lambda)\leqq\mathcal{T}_{m,\beta}(\lambda)$ for each integer $m\geqq 0$ and $\lambda>0$. Let

$$C_\alpha(t,\xi)=\mathbf{E}\cos(|\xi|Y_\alpha(t)),\qquad S_\alpha(t,\xi)=\mathbf{E}\sin(|\xi|Y_\alpha(t)).$$

By (2) and (16), it holds that

$$\mathcal{T}_{m,\alpha}(\lambda)=\int_{\mathbf{R}^n}(1+|\xi|^2)^m[|\xi|^2|\hat{\phi}(\xi)|^2+|\hat{\psi}(\xi)|^2]A_\alpha(\lambda,\xi)\,d\xi, \tag{20}$$

where

$$A_\alpha(\lambda,\xi)=\lambda^{1/\alpha}\int_0^\infty e^{-t\lambda^{1/\alpha}}[C_\alpha^2(t,\xi)+S_\alpha^2(t,\xi)]\,dt.$$

On the other hand, using Lemma 2, we get, for every $\mu>0$ and $\xi\in\mathbf{R}^n$,

$$\int_0^\infty e^{-\mu t}C_\alpha(t,\xi)\,dt=\int_0^\infty e^{-\mu t}\quad \mathbf{E}C_\beta(Y_{2\alpha/\beta}(t),\xi)\,dt\left(=\frac{\mu^{\alpha-1}}{\mu^\alpha+|\xi|^2}\right),$$

where $Y_{2\alpha/\beta}(t)$ is the stochastic process with the Mittag-Leffler distributions of order α/β (see (7) above). The uniqueness of the Laplace transform leads to

$$C_\alpha(t,\xi)=\mathbf{E}C_\beta(Y_{2\alpha/\beta}(t),\xi).$$

Similarly we get

$$S_\alpha(t,\xi)=\mathbf{E}S_\beta(Y_{2\alpha/\beta}(t),\xi).$$

Thus, by the Cauchy-Schwarz inequality

$$A_\alpha(\lambda,\xi)$$

$$=\lambda^{1/\alpha}\int_0^\infty e^{-t\lambda^{1/\alpha}}[(\mathbf{E}C_\beta(Y_{2\alpha/\beta}(t),\xi))^2+(\mathbf{E}S_\beta(Y_{2\alpha/\beta}(t),\xi))^2]\,dt$$

(21)

$$\leqq\lambda^{1/\alpha}\int_0^\infty e^{-t\lambda^{1/\alpha}}\mathbf{E}[C_\beta^2(Y_{2\alpha/\beta}(t),\xi)+S_\beta^2(Y_{2\alpha/\beta}(t),\xi)]\,dt$$

(22)

$$=\lambda^{1/\alpha}(\lambda^{1/\alpha})^{(\alpha/\beta)-1}\int_0^\infty\exp\{-y(\lambda^{1/\alpha})^{\alpha/\beta}\}[C_\beta^2(y,\xi)+S_\beta^2(y,\xi)]\,dy$$

$$=A_\beta(\lambda,\xi),$$

so that

$$A_\alpha(\lambda,\xi)\leqq A_\beta(\lambda,\xi). \tag{23}$$

Here, in (22), we used Lemma 2. Then the inequality $\mathcal{T}_{m,\alpha}(\lambda)\leqq\mathcal{T}_{m,\beta}(\lambda)$ follows from (20) and (23).

Next, we show that if $\mathcal{T}_{m,\alpha}(\lambda)=\mathcal{T}_{m,\beta}(\lambda)$ $(\alpha \leqq \beta)$, except the trivial case $\phi=\psi\equiv 0$, then $\alpha=\beta$. Here $\lambda>0$ and $m \geqq 0$ are fixed arbitrarily. The case $\mathcal{T}_{m,\alpha}(\lambda)=\mathcal{T}_{m,\beta}(\lambda)$ occurs if and only if (21) reduces to the equality. In order that (21) reduces to the equality, it is necessary that the following equalities hold with probability 1:

$$
\begin{aligned}
C_\beta(Y_{2\alpha/\beta}(t),\xi) &= \mathbf{E}C_\beta(Y_{2\alpha/\beta}(t),\xi) \qquad (t \geqq 0,\ \xi \in \mathbf{R}^n),\\
S_\beta(Y_{2\alpha/\beta}(t),\xi) &= \mathbf{E}S_\beta(Y_{2\alpha/\beta}(t),\xi) \qquad (t \geqq 0,\ \xi \in \mathbf{R}^n).
\end{aligned}
\tag{24}
$$

Remark that

$$
\lim_{|\xi|\to 0}\frac{S_\beta(t,\xi)}{|\xi|}=\lim_{|\xi|\to 0}\frac{\mathbf{E}\sin(|\xi|\,Y_\beta(t))}{|\xi|}=\mathbf{E}\,Y_\beta(t)=\frac{t^{\beta/2}}{\Gamma(1+(\beta/2))}.
$$

Thus we have, by (24),

$$
\frac{(Y_{2\alpha/\beta}(t))^{\beta/2}}{\Gamma(1+(\beta/2))}=\mathbf{E}\frac{(Y_{2\alpha/\beta}(t))^{\beta/2}}{\Gamma(1+(\beta/2))}.
$$

By Lemma 1, we find that $2\alpha/\beta=2$, so that $\alpha=\beta$. This means that if $\alpha<\beta$, then $\mathcal{T}_{m,\alpha}(\lambda)<\mathcal{T}_{m,\beta}(\lambda)$, except the trivial case $\phi=\psi\equiv 0$.

Finally, we show that $\mathcal{T}_{m,\alpha}(\lambda)$ is continuous in $\alpha\in[1,2]$ for every $m\geqq 0$ and $\lambda>0$. For every $t \geqq 0$ and $\xi\in\mathbf{R}^n$, $C_\alpha(t,\xi)$ and $S_\alpha(t,\xi)$ are continuous in $\alpha\in[1,2]$ because the series (11) and (12) converge uniformly in $\alpha\in[1,2]$. Then the desired result follows from (20) and the dominated convergence theorem. This completes the proof of Theorem 3. □

Acknowledgment. The author wishes to thank the referee who has carefully read the manuscript and whose constructive criticisms have been most helpful.

REFERENCES

[1] N. H. Bingham, *Maxima of sums of random variables and suprema of stable processes*, Z. Wahrsch. Geb., 26 (1973), pp. 273-296.

[2] P. L. Butzer and U. Westphal, *An access to fractional differentiation via fractional difference quotients*, in Proc. of the International Conference on Fractional Calculus and Its Applications, New Haven, CT, 1974, Lecture Notes in Math. 457, Springer-Verlag, Berlin, 1975, pp. 116-145.

[3] Ph. Clément and J. A. Nohel, *Abstract linear and nonlinear Volterra equations preserving positivity*, SIAM J. Math. Anal., 10 (1979), pp. 365-388.

[4] ———, *Asymptotic behavior of solutions of nonlinear Volterra equations with completely positive kernels*, SIAM J. Math. Anal., 12 (1981), pp. 514-535.

[5] Ph. Clément and E. Mitidieri, *Qualitative properties of solutions of Volterra equations in Banach spaces*, Israel J. Math., 64 (1988), pp. 1-24.

[6] A. Erdélyi, *Higher Transcendental Functions, Vol.* 3, McGraw-Hill, New York, 1955.

[7] W. Feller, *An Introduction to Probability Theory and its Applications* II, John Wiley, New York, 1966.

[8] Y. Fujita, *Integrodifferential equation which interpolates the heat equation and the wave equation*, Osaka J. Math., 27 (1990), pp. 309-321.

[9] ———, *Integrodifferential equation which interpolates the heat equation and the wave equation* (II), Osaka J. Math., 27 (1990), pp. 797-804.

[10] ———, *Cauchy problems of fractional order and stable processes*, Japan J. Appl. Math., 7 (1990), pp. 459-476.

[11] G. Gripenberg, *Volterra integro-differential equations with accretive nonlinearity*, J. Differential Equations, 60 (1985), pp. 57-79.

[12] W. R. Schneider and W. Wyss, *Fractional diffusion and wave equations*, J. Math. Phys., 30 (1989), pp. 134-144.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1189-1203. September 1992

008

THE GENERALIZED RIEMANN PROBLEM FOR THE MOTION OF ELASTIC STRINGS*

LI TA-TSIEN†, D. SERRE‡, AND ZHANG HAO§

Abstract. It is proven that, except in certain critical cases, the generalized Riemann problem for a nonstrictly hyperbolic system of elastic strings admits a unique local solution in the class of piecewise C^1 functions and in a neighborhood of the origin this solution possesses a structure similar to the similarity solution of the corresponding Riemann problem.

Key words. generalized Riemann problem, elastic strings, nonlinear stability

AMS(MOS) subject classifications. 35L65, 35L67, 73C50

1. Introduction.

Consider the following system for the motion of an elastic string on a plane (cf. [1]-[5])

$$\begin{aligned} &u_t - v_x = 0, \\ &v_t - \left(\frac{\hat{T}(r)}{r}u\right)_x = 0, \end{aligned} \tag{1.1}$$

where $u=(u_1, u_2)^T$, $v=(v_1, v_2)^T$ are unknown vector functions of (t, x), $r=|u|=\sqrt{u_1^2+u_2^2}$, and

$$\hat{T}(r)=\begin{cases} T(r), & r \geqq 1, \\ 0, & 0 \leqq r < 1, \end{cases} \tag{1.2}$$

in which $T(r)$ is a regular and strictly increasing function on $r \geqq 1$ and $T(1)=0$.

In this paper, for the special but important case

$$T(r)=r-1, \tag{1.3}$$

we study the generalized Riemann problem for system (1.1) and prove that, except in certain critical cases, the generalized Riemann problem admits a unique local solution in a class of piecewise continuous and piecewise smooth functions, and in a neighborhood of the origin this solution possesses a structure similar to the similarity solution of the corresponding Riemann problem. Since system (1.1) is strictly hyperbolic only for $r>1$, this shows the nonlinear stability of the solution to the Riemann problem for a system that may degenerate.

The same result can be obtained in a similar way for the motion of an elastic string on a n-dimensional space, $n \geqq 3$.

The case when $T(r)$ is a nonlinear function can be similarly treated: see [8].

2. Preliminaries.

In the domain $0 \leqq r \leqq 1$, system (1.1) simply reduces to

$$\begin{aligned} &u_t - v_x = 0, \\ &v_t = 0. \end{aligned} \tag{2.1}$$

*Received by the editors August 13, 1991; accepted for publication (in revised form) February 14, 1992.

† Department of Mathematics, Fudan University, Shanghai 200433, China; Unité de Mathématiques Pures et Appliquées, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France.

‡ Unité de Mathématiques Pures et Appliquées, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France.

§ Institute of Mathematics, Fudan University, Shanghai 200433, China.

Hence, if the initial data

$$t=0:\ u=u_0(x), \qquad v=v_0(x) \tag{2.2}$$

have a bounded C^1-norm and

$$\operatorname*{Sup}_x r_0(x)=\operatorname*{Sup}_x |u_0(x)|<1. \tag{2.3}$$

We immediately obtain the explicit solution

$$u=u_0(x)+tv_0'(x), \qquad v=v_0(x) \tag{2.4}$$

for $t>0$ suitably small. We point out that system (2.1) is not hyperbolic.

In the domain $r>1$, system (1.1) is strictly hyperbolic. There are four distinct real eigenvalues depending only on r:

$$\lambda_1=:-1<\lambda_2=:-\sqrt{\frac{r-1}{r}}<\lambda_3=:\sqrt{\frac{r-1}{r}}<\lambda_4=:1, \tag{2.5}$$

with the corresponding left eigenvectors

$$\begin{aligned} l_1&=(u,u)=(rp,rp),\\ l_2&=\left(\sqrt{\frac{r-1}{r}}\,w,\ w\right)=(\sqrt{r(r-1)}\,q,\ rq),\\ l_3&=\left(\sqrt{\frac{r-1}{r}}\,w,\ -w\right)=(\sqrt{r(r-1)}\,q,\ -rq),\\ l_4&=(u,-u)=(rp,-rp), \end{aligned} \tag{2.6}$$

and the corresponding right eigenvectors

$$\begin{aligned} r^1&=(u,u)^T=(rp,rp)^T,\\ r^2&=\left(w,\ \sqrt{\frac{r-1}{r}}\,w\right)^T=(rq,\ \sqrt{r(r-1)}\,q)^T,\\ r^3&=\left(-w,\ \sqrt{\frac{r-1}{r}}\,w\right)^T=(-rq,\ \sqrt{r(r-1)}\,q)^T,\\ r^4&=(-u,u)^T=(-rp,rp)^T, \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} &u=rp, \quad p=(p_1,p_2), \quad |p|=1,\\ &w=(-u_2,u_1), \qquad q=(-p_2,p_1). \end{aligned} \tag{2.8}$$

Both the left eigenvectors and the right eigenvectors depend only on u. Moreover, all eigenvalues are linearly degenerate in the sense of P. D. Lax.

On a discontinuous curve $x=x(t)$ ($x(0)=0$), we have the Rankine-Hugoniot's conditions

$$\begin{aligned} &[u]\,dx+[v]\,dt=0,\\ &[v]\,dx+\left[\frac{\hat{T}(r)}{r}u\right]dt=0, \end{aligned} \tag{2.9}$$

where $[u]=u^+-u^-$ is the jump of u, etc.

There are several possibilities.

(1) $0<r^{\pm}\leqq 1$.

Noting (1.2), it follows from (2.9) that $x(t)\equiv 0$ and

$$v^{+}=v^{-}, \tag{2.10}$$

while u may have an arbitrary jump.

(2) $r^{\pm}\geqq 1$ (except $r^{+}=r^{-}=1$).

By means of (1.2), (1.3) we get the following.

(2a) Either

$$r^{+}=r^{-}=:r, \tag{2.11}$$

$$\frac{dx}{dt}=\pm\sqrt{\frac{r-1}{r}}, \tag{2.12}$$

$$[v]=a[p], \tag{2.13}$$

where

$$a=\begin{cases}-\sqrt{r(r-1)}, & \text{for the sign ``+'' in (2.12),}\\ \sqrt{r(r-1)} & \text{for the sign ``$-$'' in (2.12).}\end{cases} \tag{2.14}$$

In this case, $x=x(t)$ is a contact discontinuity of the second or third kind, i.e., corresponding to the second or third (transverse) characteristic family, respectively.

(2b) Or

$$p^{+}=p^{-}=:p, \tag{2.15}$$

$$\frac{dx}{dt}=\pm 1, \tag{2.16}$$

$$[v]=ap, \tag{2.17}$$

where

$$a=\begin{cases}-[r] & \text{for the sign ``+'' in (2.16),}\\ [r] & \text{for the sign ``$-$'' in (2.16),}\end{cases} \tag{2.18}$$

when $r^{\pm}>1$, $x=x(t)$ is a contact discontinuity of the first or fourth kind, i.e., corresponding to the first or fourth (longitudinal) characteristic family, respectively; while, when one of $r^{\pm}$ is equal to 1, $x=x(t)$ is only a lateral contact discontinuity, i.e., a contact discontinuity on only one side.

(3) $r^{+}>1>r^{-}>0$.

In a similar way, we get (2.15), (2.17) with

$$a=\sqrt{(r^{+}-r^{-})(r^{+}-1)} \tag{2.19}$$

and

$$\frac{dx}{dt}=-\sqrt{\frac{r^{+}-1}{r^{+}-r^{-}}}. \tag{2.20}$$

In this case $x=x(t)$ is a lateral shock of the first kind on the right side, which satisfies the P. D. Lax entropy condition

$$\lambda_1(r^{+})=-1<\frac{dx}{dt}<\lambda_2(r^{+})=-\sqrt{\frac{r^{+}-1}{r^{+}}}. \tag{2.21}$$

(4) $r^{-}>1>r^{+}>0$.

Similarly, $x = x(t)$ is a lateral shock of the fourth kind on the left side. On $x = x(t)$ we have (2.15), (2.17) with

$$a = \sqrt{(r^- - r^+)(r^- - 1)} \tag{2.22}$$

and

$$\frac{dx}{dt} = \sqrt{\frac{r^- - 1}{r^- - r^+}}. \tag{2.23}$$

Moreover, the following Lax entropy condition is satisfied:

$$\lambda_3(r^-) = \sqrt{\frac{r^- - 1}{r^-}} < \frac{dx}{dt} < \lambda_4(r^-) = 1. \tag{2.24}$$

3. The Riemann problem. In this section we recall the result (cf. [1]-[4]) on the Riemann problem for system (1.1) with the initial data

$$t = 0\colon\ U = (u, v) = \begin{cases} \hat{U}_l, & x \leqq 0, \\ \hat{U}_r, & x \geqq 0, \end{cases} \tag{3.1}$$

where $\hat{U}_l = (\hat{u}_l, \hat{v}_l)$ and $\hat{U}_r = (\hat{u}_r, \hat{v}_r)$ are constant vectors with $\hat{U}_l \neq \hat{U}_r$.

Case I. Suppose that

$$0 < \hat{r}_l = |\hat{u}_l| < 1, \qquad 0 < \hat{r}_r = |\hat{u}_r| < 1. \tag{3.2}$$

Case IA. If

$$\hat{v}_l = \hat{v}_r, \tag{3.3}$$

then the solution to Riemann problem (1.1), (3.1) is

$$U = (u, v) = \begin{cases} \hat{U}_l, & t \geqq 0, \quad x \leqq 0, \\ \hat{U}_r, & t \geqq 0, \quad x \geqq 0, \end{cases} \tag{3.4}$$

and $x = 0$ is the unique discontinuity.

Case IB. If

$$\hat{v}_l \neq \hat{v}_r, \tag{3.5}$$

then the solution to Riemann problem (1.1), (3.1) can be indicated in Fig. 1, where $\hat{U}_- = (\hat{u}_-, \hat{v}_-)$, $\hat{U}_0 = (\hat{u}_0, \hat{v}_0)$ and $\hat{U}_+ = (\hat{u}_+, \hat{v}_+)$. Moreover,

$$O\hat{A}_1\colon x = -\sqrt{\frac{\hat{r}_- - 1}{\hat{r}_- - \hat{r}_l}}\, t =: \hat{\sigma}_1 t \tag{3.6}$$

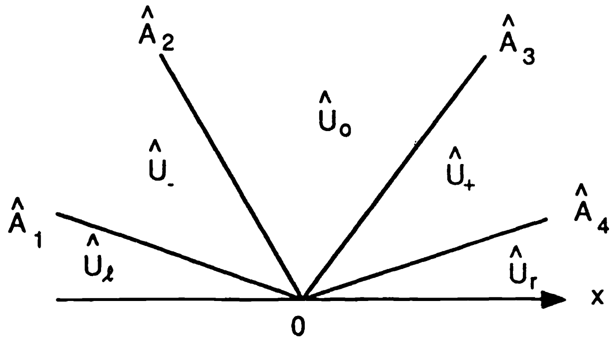


FIG. 1

is a lateral shock of the first kind on the right side, on which we have

$$\hat{p}_- = \hat{p}_l, \tag{3.7}$$

$$\hat{v}_- = \hat{v}_l + \sqrt{(\hat{r}_- - \hat{r}_l)(\hat{r}_- - 1)}\,\hat{p}_l, \tag{3.8}$$

and

$$\hat{r}_- > 1 > \hat{r}_l; \tag{3.9}$$

$$O\hat{A}_2\colon x = -\sqrt{\frac{\hat{r}_0 - 1}{\hat{r}_0}}\,t =: \hat{\sigma}_2 t \tag{3.10}$$

is a contact discontinuity of the second kind, on which we have

$$\hat{r}_0 = \hat{r}_- > 1, \tag{3.11}$$

$$\hat{v}_0 = \hat{v}_- + \sqrt{\hat{r}_0(\hat{r}_0 - 1)}\,(\hat{p}_0 - \hat{p}_-); \tag{3.12}$$

$$O\hat{A}_3\colon x = \sqrt{\frac{\hat{r}_0 - 1}{\hat{r}_0}}\,t =: \hat{\sigma}_3 t \tag{3.13}$$

is a contact discontinuity of the third kind, on which we have

$$\hat{r}_+ = \hat{r}_0 > 1, \tag{3.14}$$

$$\hat{v}_0 = \hat{v}_+ + \sqrt{\hat{r}_0(\hat{r}_0 - 1)}(\hat{p}_+ - \hat{p}_0); \tag{3.15}$$

$$O\hat{A}_4\colon x = \sqrt{\frac{\hat{r}_+ - 1}{\hat{r}_+ - \hat{r}_r}}\,t =: \hat{\sigma}_4 t \tag{3.16}$$

is a lateral shock of the fourth kind on the left side, on which we have

$$\hat{p}_+ = \hat{p}_r, \tag{3.17}$$

$$\hat{v}_+ = \hat{v}_r - \sqrt{(\hat{r}_+ - \hat{r}_r)(\hat{r}_+ - 1)}\,\hat{p}_r \tag{3.18}$$

and

$$\hat{r}_+ > 1 > \hat{r}_r. \tag{3.19}$$

In this case, system (1.1) is strictly hyperbolic only for the solution on the angular domain between $O\hat{A}_1$ and $O\hat{A}_4$.

Case II. Suppose that

$$\hat{r}_l = |\hat{u}_l| > 1, \qquad 0 < \hat{r}_r = |\hat{u}_r| < 1. \tag{3.20}$$

Case IIA. If

$$\hat{v}_l + (1 - \hat{r}_l)\hat{p}_l = \hat{v}_r, \tag{3.21}$$

then the solution to Riemann problem (1.1), (3.1) can be shown in Fig. 2, where $O\hat{A}_1$: $x = -t$ is a lateral contact discontinuity on the left side and $\hat{U}_- = (\hat{u}_-, \hat{v}_-) = (\hat{p}_l, \hat{v}_r)$.

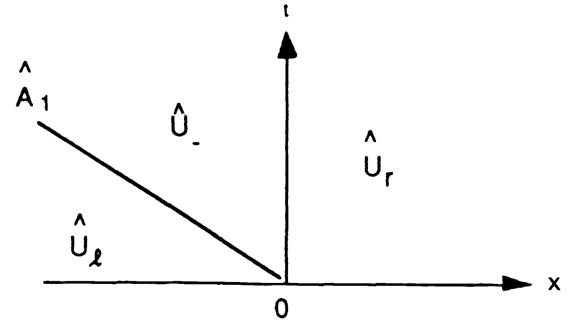


FIG. 2

In this case we have two discontinuities $x=0$ and $x=-t$; moreover, only for the solution on the left side of $O\hat{A}_1$, system (1.1) is strictly hyperbolic.

Case IIB. If

$$\hat{v}_l+(1-\hat{r}_l)\hat{p}_l\neq\hat{v}_r, \tag{3.22}$$

then the solution to the Riemann problem can be still indicated in Fig. 1. Different from Case IB, however,

$$O\hat{A}_1: \quad x=-t \tag{3.23}$$

is a contact discontinuity of first kind, on which we have (3.7),

$$\hat{v}_-=\hat{v}_l+(\hat{r}_--\hat{r}_l)\hat{p}_l \tag{3.24}$$

and

$$\hat{r}_->1. \tag{3.25}$$

In this case, system (1.1) is strictly hyperbolic only on the left side of $O\hat{A}_4$.

Case III. Suppose that

$$0<\hat{r}_l=|\hat{u}_l|<1, \qquad \hat{r}_r=|\hat{u}_r|>1. \tag{3.26}$$

In this case, the situation is completely similar to Case II.

Case IV. Suppose that

$$\hat{r}_l=|\hat{u}_l|>1, \qquad \hat{r}_r=|\hat{u}_r|>1. \tag{3.27}$$

Case IVA. If

$$\hat{v}_l+(1-\hat{r}_l)\hat{p}_l=\hat{v}_r-(1-\hat{r}_r)\hat{p}_r, \tag{3.28}$$

then the solution to the Riemann problem can be shown in Fig. 3, where $O\hat{A}_1$: $x=-t$ and $O\hat{A}_4$: $x=t$ are lateral contact discontinuities on the left side and on the right side, respectively; moreover, $\hat{U}_-=(\hat{u}_-,\hat{v}_-)=(\hat{p}_l,\hat{v}_l+(1-\hat{r}_l)\hat{p}_l)$ and $\hat{U}_+=(\hat{u}_+,\hat{v}_+)=(\hat{p}_r,\hat{v}_r-(1-\hat{r}_r)\hat{p}_r)$.

In this case we have three discontinuities $x=0$ and $x=\pm t$. System (1.1) is strictly hyperbolic only on the left side of $O\hat{A}_1$ and on the right side of $O\hat{A}_4$.

Case IVB. If

$$\hat{v}_l+(1-\hat{r}_l)\hat{p}_l\neq\hat{v}_r-(1-\hat{r}_r)\hat{p}_r, \tag{3.29}$$

then the solution to the Riemann problem can be still indicated in Fig. 1. Different from Case IB, however, not only $O\hat{A}_1$ is a contact discontinuity of the first kind, on which we have (3.7) and (3.23)-(3.25), but also

$$O\hat{A}_4: \ x=t \tag{3.30}$$

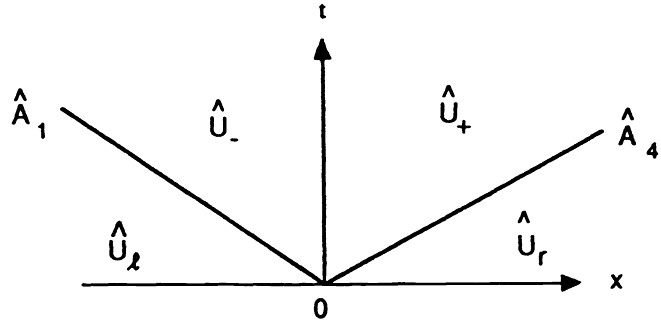


FIG. 3

is a contact discontinuity of fourth kind, on which we have (3.17),

$$\hat{v}_+ = \hat{v}_r - (\hat{r}_+ - \hat{r}_r)\hat{p}_r \tag{3.31}$$

and

$$\hat{r}_+ > 1. \tag{3.32}$$

In this case, system (1.1) is strictly hyperbolic on the whole upper plane $t \geqq 0$.

4. The generalized Riemann problem—Case I. We now consider the generalized Riemann problem for system (1.1) with the following initial data

$$t=0\colon U=(u,v)=\begin{cases} \bar{U}_l(x), & x\leqq 0, \\ \bar{U}_r(x), & x\geqq 0, \end{cases} \tag{4.1}$$

where $\bar{U}_l(x)=(\bar{u}_l(x), \bar{v}_l(x))$ and $\bar{U}_r(x)=(\bar{u}_r(x), \bar{v}_r(x))$ are regular vector functions with bounded C^1 norm, $\bar{r}_l(x)$, $\bar{r}_r(x)>0$, and $\hat{U}_l=(\hat{u}_l, \hat{v}_l) \neq \hat{U}_r=(\hat{u}_r, \hat{v}_r)$, where

$$\hat{U}_l = U_l(0), \qquad \hat{U}_r = U_r(0). \tag{4.2}$$

In this section we first study the following case.

Case I. Suppose that

$$\operatorname*{Sup}_{x\leqq 0} \bar{r}_l(x) = \operatorname*{Sup}_{x\leqq 0} |\bar{u}_l(x)| < 1, \qquad \operatorname*{Sup}_{x\geqq 0} \bar{r}_r(x) = \operatorname*{Sup}_{x\geqq 0} |\bar{u}_r(x)| < 1. \tag{4.3}$$

Case IA. If (3.3) holds, then, by means of (2.4), the solution to the generalized Riemann problem is

$$U=(u,v)=\begin{cases} (\bar{u}_l(x)+t\bar{v}_l'(x), \bar{v}_l(x)), & t\geqq 0 \text{ small}, \quad x\leqq 0, \\ (\bar{u}_r(x)+t\bar{v}_r'(x), \bar{v}_r(x)), & t\geqq 0 \text{ small}, \quad x\geqq 0. \end{cases} \tag{4.4}$$

$x=0$ is the unique discontinuity, and (4.4) possesses a structure similar to (3.4) of the corresponding Riemann problem in a neighborhood of the origin.

Case IB. If (3.5) holds, we still hope to prove that the generalized Riemann problem (1.1), (4.1) admits a unique local solution in a class of piecewise continuous and piecewise smooth functions, and in a neighborhood of the origin this solution has a structure similar to that of the corresponding Riemann problem. In other words, we want to obtain a unique local solution to the generalized Riemann problem (1.1), (4.1) as shown in Fig. 4, where

$$OA_i\colon x=x_i(t) \quad (x_i(0)=0) \quad (i=1,2,3,4) \tag{4.5}$$

are free boundaries.

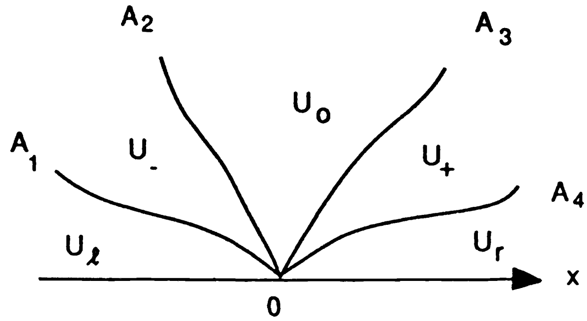


FIG. 4

By (2.4), on the domain

$$D_l(\delta)=\{(t,x)\,|\,0\leqq t\leqq \delta,\ x\leqq x_1(t)\} \tag{4.6}$$

($\delta>0$ small), the solution is known:

$$\begin{aligned} U_l(t,x)&=(u_l(t,x),\ v_l(t,x))=(r_lp_l(t,x),\ v_l(t,x))\\ &=(\bar{u}_l(x)+t\bar{v}_l'(x),\ \bar{v}_l(x)). \end{aligned} \tag{4.7}$$

Similarly, on the domain

$$D_r(\delta)=\{(t,x)|0\leqq t\leqq \delta,\ x\geqq x_4(t)\} \tag{4.8}$$

($\delta>0$ small), the solution is

$$\begin{aligned} U_r(t,x)&=(u_r(t,x),\ v_r(t,x))=(r_rp_r(t,x),\ v_r(t,x))\\ &=(\bar{u}_r(x)+t\bar{v}_r'(x),\ \bar{v}_r(x)). \end{aligned} \tag{4.9}$$

Obviously, we have

$$U_l(0,0)=\hat{U}_l,\qquad U_r(0,0)=\hat{U}_r. \tag{4.10}$$

On the domains

$$D_-(\delta)=\{(t,x)|0\leqq t\leqq \delta,\ x_1(t)\leqq x\leqq x_2(t)\}, \tag{4.11}$$

$$D_0(\delta)=\{(t,x)|0\leqq t\leqq \delta,\ x_2(t)\leqq x\leqq x_3(t)\}, \tag{4.12}$$

$$D_+(\delta)=\{(t,x)|0\leqq t\leqq \delta,\ x_3(t)\leqq x\leqq x_4(t)\} \tag{4.13}$$

($\delta>0$ small), the solution is denoted, respectively, by $U_-(t,x)=(u_-(t,x),\ v_-(t,x))=(r_-p_-(t,x),\ v_-(t,x))$, $U_0(t,x)=(u_0(t,x),\ v_0(t,x))=(r_0p_0(t,x),\ v_0(t,x))$, and $U_+(t,x)=(u_+(t,x),\ v_+(t,x))=(r_+p_+(t,x),\ v_+(t,x))$. All $U_-(t,x)$, $U_0(t,x)$, and $U_+(t,x)$ are unknown regular solutions to system (1.1); moreover,

$$U_-(0,0)=\hat{U}_-,\quad U_0(0,0)=\hat{U}_0,\quad U_+(0,0)=\hat{U}_+, \tag{4.14}$$

where $\hat{U}_-$, $\hat{U}_0$, and $\hat{U}_+$ are furnished by the solution to the corresponding Riemann problem.

Furthermore, $OA_1(x=x_1(t))$ is a lateral shock of the first kind on the right side, on which we have

$$\frac{dx_1(t)}{dt}=-\sqrt{\frac{r_--1}{r_--r_l}},\qquad x_1(0)=0, \tag{4.15}$$

$$p_-=p_l, \tag{4.16}$$

$$v_-=v_l+\sqrt{(r_--r_l)(r_--1)}\,p_l, \tag{4.17}$$

$$r_->1>r_l>0. \tag{4.18}$$

$OA_2(x=x_2(t))$ is a contact discontinuity of the second kind, on which we have

$$\frac{dx_2(t)}{dt}=-\sqrt{\frac{r_0-1}{r_0}},\qquad x_2(0)=0, \tag{4.19}$$

$$r_0=r_->1, \tag{4.20}$$

$$v_0=v_-+\sqrt{r_0(r_0-1)}(p_0-p_-). \tag{4.21}$$

$OA_3(x=x_3(t))$ is a contact discontinuity of the third kind, on which we have

$$\frac{dx_3(t)}{dt}=\sqrt{\frac{r_0-1}{r_0}},\qquad x_3(0)=0, \tag{4.22}$$

$$r_+=r_0>1, \tag{4.23}$$

$$v_0=v_++\sqrt{r_0(r_0-1)}(p_+-p_0). \tag{4.24}$$

$OA_4(x=x_4(t))$ is a lateral shock of the fourth kind on the left side, on which we have

$$\frac{dx_4(t)}{dt}=\sqrt{\frac{r_+-1}{r_4-r_r}},\qquad x_4(0)=0, \tag{4.25}$$

$$p_+=p_r, \tag{4.26}$$

$$v_+=v_r-\sqrt{(r_+-r_r)(r_+-1)}\,p_r, \tag{4.27}$$

$$r_+>1>r_r>0. \tag{4.28}$$

Noticing (4.10) and (4.14), it follows from (4.15), (4.19), (4.22), and (4.25) that

$$\frac{dx_i}{dt}(0)=\hat{\sigma}_i\qquad (i=1,2,3,4), \tag{4.29}$$

where $\hat{\sigma}_i$ $(i=1,2,3,4)$ are also furnished by the solution to the corresponding Riemann problem.

Since $U_l(t,x)$ and $U_r(t,x)$ are known, in order to get the solution we have to solve the free boundary problem (1.1) and (4.14)-(4.28) on the fan-shaped domain $D(\delta)=D_-(\delta)\cup D_0(\delta)\cup D_+(\delta)(\delta>0$ small). According to the result on the Riemann problem, by continuity system (1.1) is strictly hyperbolic on $D(\delta)$ and the inequalities in (4.18), (4.20), (4.23), and (4.28) are always satisfied for $\delta>0$ suitably small.

Let

$$L=\begin{pmatrix} l_1\\ l_2\\ l_3\\ l_4\end{pmatrix} \tag{4.30}$$

be the 4×4 matrix composed of the left eigenvectors, and denote

$$\begin{aligned} V^-&=L(\hat{u}_-)U_-=(V_1^-,V_2^-,V_3^-,V_4^-),\\ V^0&=L(\hat{u}_0)U_0=(V_1^0,V_2^0,V_3^0,V_4^0),\\ V^+&=L(\hat{u}_+)U_+=(V_1^+,V_2^+,V_3^+,V_4^+).\end{aligned} \tag{4.31}$$

Noting that $U_l(t,x)$ and $U_r(t,x)$ are known, we can verify that boundary conditions (4.16), (4.17) on $x=x_1(t)$, (4.20), (4.21) on $x=x_2(t)$, (4.23), (4.24) on $x=x_3(t)$, and (4.26), (4.27) on $x=x_4(t)$ can be rewritten, respectively, as

$$V_i^-=f_i(t,x,V_1^-)\quad (i=2,3,4)\quad \text{on } x=x_1(t), \tag{4.32}$$

$$\left.\begin{aligned} &V_1^-=f_1(V_2^-,V_3^-,V_4^-,V_1^0,V_2^0),\\ &V_i^0=g_i(V_2^-,V_3^-,V_4^-,V_1^0,V_2^0)\quad (i=3,4)\end{aligned}\right\}\quad \text{on } x=x_2(t), \tag{4.33}$$

$$\left.\begin{aligned} &V_i^0=g_i(V_3^0,\,V_4^0,\,V_1^+,\,V_2^+,\,V_3^+)\quad (i=1,2)\\ &V_4^+=h_4(V_3^0,\,V_4^0,\,V_1^+,\,V_2^+,\,V_3^+)\end{aligned}\right\} \text{on } x=x_3(t), \tag{4.34}$$

$$V_i^+=h_i(t,x,V_4^+),\quad (i=1,2,3)\quad \text{on } x=x_4(t). \tag{4.35}$$

At the point $t=x=0$, $V^-=\hat{V}^-=: l(\hat{u}_-)\hat{U}_-$, $V_0=\hat{V}^0=: l(\hat{u}_0)\hat{U}_0$, and $V^+=\hat{V}^+=: l(\hat{u}_+)\hat{U}_+$, we form the following Jacobi matrix:

$$\Theta=\frac{\partial(f_1,\cdots,f_4,g_1,\cdots,g_4,h_1,\cdots,h_4)}{\partial(V_1^-,\cdots,V_4^-,V_1^0,\cdots,V_4^0,V_1^+,\cdots,V_4^+)}. \tag{4.36}$$

Set

$$\begin{aligned}
&\tau_1^-=\frac{\lambda_1(\hat{u}_-)-\hat{\sigma}_1}{\lambda_1(\hat{u}_-)-\hat{\sigma}_2}, && \tau_i^-=\frac{\lambda_i(\hat{u}_-)-\hat{\sigma}_2}{\lambda_i(\hat{u}_-)-\hat{\sigma}_1}\quad (i=2,3,4),\\
&\tau_i^0=\frac{\lambda_i(\hat{u}_0)-\hat{\sigma}_2}{\lambda_i(\hat{u}_0)-\hat{\sigma}_3}\quad (i=1,2), && \tau_i^0=\frac{\lambda_i(\hat{u}_0)-\hat{\sigma}_3}{\lambda_i(\hat{u}_0)-\hat{\sigma}_2}\quad (i=3,4),\\
&\tau_i^+=\frac{\lambda_i(\hat{u}_+)-\hat{\sigma}_3}{\lambda_i(\hat{u}_+)-\hat{\sigma}_4}\quad (i=1,2,3), && \tau_4^+=\frac{\lambda_4(\hat{u}_+)-\hat{\sigma}_4}{\lambda_4(\hat{u}_+)-\hat{\sigma}_3}.
\end{aligned} \tag{4.37}$$

Obviously, we have

$$0\leqq \tau_i^-,\,\tau_i^0,\,\tau_i^+<1\qquad (i=1,2,3,4) \tag{4.38}$$

and

$$\tau_2^-=\tau_2^0=\tau_3^0=\tau_3^+=0. \tag{4.39}$$

Let

$$\Theta_1=\Theta\tau, \tag{4.40}$$

where

$$\tau=\operatorname{diag}(\tau_1^-,\cdots,\tau_4^-,\tau_1^0,\cdots,\tau_4^0,\tau_1^+,\cdots,\tau_4^+\}. \tag{4.41}$$

According to the result in [6], [7], if

$$\|\Theta_1\|_{\min}<1, \tag{4.42}$$

then the free boundary problem under consideration admits a unique piecewise C^1 solution on $D(\delta)$ ($\delta>0$ small), and this solution has the desired structure. Here, for an $n\times n$ matrix $A=(a_{ij})$, define

$$\|A\|=\operatorname*{Max}_{i=1,\cdots,n}\sum_{j=1}^{n}|a_{ij}| \tag{4.43}$$

and

$$\|A\|_{\min}=\operatorname{Inf}\{\|\gamma A\gamma^{-1}\|;\ \gamma=\operatorname{diag}\{\gamma_i\},\ \gamma_i\neq 0,\ i=1,\cdots,n\}. \tag{4.44}$$

Noting that at the point $t=x=0$, $V^-=\hat{V}^-$, $V^0=\hat{V}^0$, and $V^+=\hat{V}^+$, we have

$$\frac{\partial f_2}{\partial V_1^-}=\frac{\partial f_3}{\partial V_1^-}=\frac{\partial h_2}{\partial V_4^+}=\frac{\partial h_3}{\partial V_4^+}=\frac{\partial f_1}{\partial V_2^0}=\frac{\partial g_4}{\partial V_2^0}=\frac{\partial g_1}{\partial V_3^0}=\frac{\partial h_4}{\partial V_3^0}=0. \tag{4.45}$$

By dropping the row (or column) composed of null elements and the corresponding column (or row) in Θ_1, we obtain the following 6×6 matrix:

$$(4.46)\qquad \Theta_2=\begin{bmatrix} 0 & \tau_4^-\dfrac{\partial f_1}{\partial V_4^-} & \tau_1^0\dfrac{\partial f_1}{\partial V_1^0} & 0 & 0 & 0 \\ \tau_1^-\dfrac{\partial f_4}{\partial V_1^-} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tau_4^0\dfrac{\partial g_1}{\partial V_4^0} & \tau_1^+\dfrac{\partial g_1}{\partial V_1^+} & 0 \\ 0 & \tau_4^-\dfrac{\partial g_4}{\partial V_4^-} & \tau_1^0\dfrac{\partial g_4}{\partial V_1^0} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \tau_4^+\dfrac{\partial h_1}{\partial V_4^+} \\ 0 & 0 & 0 & \tau_4^0\dfrac{\partial h_4}{\partial V_4^0} & \tau_1^+\dfrac{\partial h_4}{\partial V_1^+} & 0 \end{bmatrix}.$$

By definition (4.44), it is easy to see (cf. Lemma 5.4 in Chapter 2 of [6]) that (4.42) is equivalent to

$$(4.47)\qquad \|\Theta_2\|_{\min}<1.$$

Moreover, noting that all elements except a nondiagonal element are zero in the second and fifth rows, by definition (4.44), it is easy to see (cf. Lemma 5.5 in Chapter 2 of [6]) that (4.47) is equivalent to

$$(4.48)\qquad \|\Theta_3\|_{\min}<1,$$

where

$$(4.49)\qquad \Theta_3=\begin{pmatrix} \tau_1^-\tau_4^-\dfrac{\partial f_1}{\partial V_4^-}\dfrac{\partial f_4}{\partial V_1^-} & \tau_1^0\dfrac{\partial f_1}{\partial V_1^0} & 0 & 0 \\ 0 & 0 & \tau_4^0\dfrac{\partial g_1}{\partial V_4^0} & \tau_1^+\tau_4^+\dfrac{\partial g_1}{\partial V_1^+}\dfrac{\partial h_1}{\partial V_4^+} \\ \tau_1^-\tau_4^-\dfrac{\partial g_4}{\partial V_4^-}\dfrac{\partial f_4}{\partial V_1^-} & \tau_1^0\dfrac{\partial g_4}{\partial V_1^0} & 0 & 0 \\ 0 & 0 & \tau_4^0\dfrac{\partial h_4}{\partial V_4^0} & \tau_1^+\tau_4^+\dfrac{\partial h_4}{\partial V_1^+}\dfrac{\partial h_1}{\partial V_4^+} \end{pmatrix}.$$

Set

$$(4.50)\qquad e=\sqrt{\frac{\hat r_0-1}{\hat r_0}},\quad f_l=\sqrt{\frac{\hat r_0-1}{\hat r_0-\hat r_l}},\quad f_r=\sqrt{\frac{\hat r_0-1}{\hat r_0-\hat r_r}}$$

and

$$(4.51)\qquad \begin{aligned} C_\pm&=\hat p_0\cdot\hat p_\pm,\\ A_\pm&=2+(1-C_\pm)\frac{(1-e)^2}{2e}. \end{aligned}$$

By (3.9), (3.11), (3.14), and (3.19) we have

$$(4.52)\qquad 0<e<f_l<1,\qquad 0<e<f_r<1$$

and

$$|C_\pm| \leqq 1, \qquad A_\pm \geqq 2. \tag{4.53}$$

Through a direct calculation we can determine the following quantities in Θ_3:

$$\begin{aligned}
&\frac{\partial f_1}{\partial V_4^-} = -\frac{(1-C_-)(1+e)^2}{2A_- e}, \qquad \frac{\partial f_1}{\partial V_1^0} = \frac{2}{A_-},\\
&\frac{\partial g_4}{\partial V_4^-} = \frac{2C_-}{A_-}, \qquad \frac{\partial g_4}{\partial V_1^0} = -\frac{A_- - 2}{A_-},\\
&\frac{\partial g_1}{\partial V_4^0} = -\frac{A_+ - 2}{A_+}; \qquad \frac{\partial g_1}{\partial V_1^+} = \frac{2C_+}{A_+},\\
&\frac{\partial h_4}{\partial V_4^0} = \frac{2}{A_+}, \qquad \frac{\partial h_4}{\partial V_1^+} = -\frac{(1-C_+)(1+e)^2}{2A_+ e},\\
&\frac{\partial f_4}{\partial V_1^-} = -\frac{(1-f_l)^2}{(1+f_l)^2}, \qquad \frac{\partial h_1}{\partial V_4^+} = -\frac{(1-f_r)^2}{(1+f_r)^2}.
\end{aligned} \tag{4.54}$$

Hence, noticing (4.38), (4.52), and (4.53) we obtain

$$\|\Theta_3\| < 1, \tag{4.55}$$

which implies (4.48). In fact, we have

$$\begin{aligned}
\left|\tau_1^- \tau_4^- \frac{\partial f_1}{\partial V_4^-}\frac{\partial f_4}{\partial V_1^-}\right| + \left|\tau_1^0 \frac{\partial f_1}{\partial V_1^0}\right| &< \left|\frac{\partial f_1}{\partial V_4^-}\right| \left|\frac{\partial f_4}{\partial V_1^-}\right| + \left|\frac{\partial f_1}{\partial V_1^0}\right|\\
&= \frac{1}{A_-}\left[(1-C_-)\frac{(1+e)^2}{(1+f_l)^2}\frac{(1-f_l)^2}{2e} + 2\right]\\
&\leqq \frac{1}{A_-}\left[(1-C_-)\frac{(1-f_l)^2}{2} + 2\right]\\
&\leqq \frac{1}{A_-}\left[(1-C_-)\frac{(1-e)^2}{2e} + 2\right] = 1,
\end{aligned}$$

$$\begin{aligned}
\left|\tau_4^0 \frac{\partial g_1}{\partial V_4^0}\right| + \left|\tau_1^+ \tau_4^+ \frac{\partial g_1}{\partial V_1^+}\frac{\partial h_1}{\partial V_4^+}\right| &\leqq \left|\frac{\partial g_1}{\partial V_4^0}\right| + \left|\frac{\partial g_1}{\partial V_1^+}\right| \left|\frac{\partial h_1}{\partial V_4^+}\right|\\
&= \frac{1}{A_+}\left[(1-C_-)\frac{(1-e)^2}{2e} + 2|C_+|\frac{(1-f_r)^2}{(1+f_r)}\right]\\
&< \frac{1}{A_+}\left[(1-C_+)\frac{(1-e)^2}{2e} + 2\right] = 1,
\end{aligned}$$

and similarly

$$\left|\tau_1^- \tau_4^- \frac{\partial g_4}{\partial V_4^-}\frac{\partial f_4}{\partial V_1^-}\right| + \left|\tau_1^0 \frac{\partial g_4}{\partial V_1^0}\right| < 1, \qquad \left|\tau_4^0 \frac{\partial h_4}{\partial V_4^0}\right| + \left|\tau_1^+ \tau_4^+ \frac{\partial h_4}{\partial V_1^+}\frac{\partial h_1}{\partial V_4^+}\right| < 1.$$

Thus, for Case I we reach the desired conclusion mentioned in the Introduction.

5. The generalized Riemann problem—Cases IIB (IIIB) and IVB.

We now consider the following case.

Case II. Suppose that

$$\operatorname*{Inf}_{x \leqq 0} \bar{r}_l(x) = \operatorname*{Inf}_{x \leqq 0} |\bar{u}_l(x)| > 1, \qquad \operatorname*{Sup}_{x \geqq 0} \bar{r}_r(x) = \operatorname*{Sup}_{x \geqq 0} |\bar{u}_r(x)| < 1. \tag{5.1}$$

Case IIB. If (3.22) holds, we still hope to get a result similar to that in § 4. However, it is different from Case IB that in this case

$$OA_1\colon x=x_1(t)\equiv -t \tag{5.2}$$

is a given contact discontinuity of first kind, on which, instead of (4.16)-(4.18), we have

$$p_-=p_l, \tag{5.3}$$

$$v_-=v_l+(r_--r_l)p_l, \tag{5.4}$$

$$r_l,\ r_->1. \tag{5.5}$$

On the left side of OA_1 the solution $U_l(t,x)=(u_l(t,x),\ v_l(t,x))$ can be obtained by solving the Cauchy problem for system (1.1) with the initial data $\bar{U}_l(x)=(\bar{u}_l(x),\ \bar{v}_l(x))$ on $x\leqq 0$. Boundary conditions (5.3), (5.4) on OA_1 can be also rewritten in the form of (4.32), and at the point $t=x=0$, $V_1^-=\hat{V}_1^-$ we have

$$\frac{\partial f_i}{\partial V_1^-}=0 \qquad (i=2,3,4). \tag{5.6}$$

Moreover, we have

$$\tau_1^-=0. \tag{5.7}$$

Thus, the matrix Θ_3 introduced in § 4 reduces to a simpler form, in which the first column is composed of null elements; then we still have (4.55). Case III is completely similar.

Case IV. Suppose that

$$\operatorname*{Inf}_{x\leqq 0}\bar{r}_l(x)=\operatorname*{Inf}_{x\leqq 0}|\bar{u}_l(x)|>1, \qquad \operatorname*{Inf}_{x\geqq 0}\bar{r}_r(x)=\operatorname*{Inf}_{x\geqq 0}|\bar{u}_r(x)|>1. \tag{5.8}$$

Case IVB. If (3.29) holds, then it is different from Case IB that not only OA_1 is a given contact discontinuity of the first kind, on which we have (5.2)-(5.5), but also

$$OA_4\colon x=x_4(t)\equiv t \tag{5.9}$$

is a given contact discontinuity of the fourth kind, on which instead of (4.26)-(4.28) we have

$$p_+=p_r, \tag{5.10}$$

$$v_+=v_r-(r_+-r_r)p_r, \tag{5.11}$$

$$r_r,\ r_+>1. \tag{5.12}$$

On the right side of OA_4 the solution $U_r(t,x)=(u_r(t,x),\ v_r(t,x))$ can be obtained by solving the Cauchy problem for system (1.1) with the initial data $\bar{U}_r(x)=(\bar{u}_r(x),\ \bar{v}_r(x))$ on $x\geqq 0$. Boundary conditions (5.10), (5.11) on OA_4 can be still rewritten in the form of (4.35), and at the point $t=x=0$, $V_4^+=\hat{V}_4^+$ we have

$$\frac{\partial h_i}{\partial V_4^+}=0 \qquad (i=1,2,3). \tag{5.13}$$

Moreover, we have

$$\tau_4^+=0. \tag{5.14}$$

Thus, the matrix Θ_3 introduced in § 4 now reduces to a much simpler form, in which the first and fourth columns are all composed of null elements; then we still get (4.55).

The desired conclusion is then obtained also for Cases IIB (IIIB) and IVB.

6. Remarks on Cases IIA (IIIA) and IVA. In general, the preceding result is no longer true for the critical Case IIA (IIIA) and IVA, unless the initial data (4.1) satisfy certain conditions of compatibility.

We take Case IIA as an example. In this case (3.21) holds, If the generalized Riemann problem under consideration admits a unique local solution that has a structure similar to that given in Fig. 2 in a neighborhood of the origin, then the solution should be shown as in Fig. 5, where

$$OA_1\colon x=-t \tag{6.1}$$

is a lateral contact discontinuity of the first kind on the left side, while $x=0$ is another discontinuity.

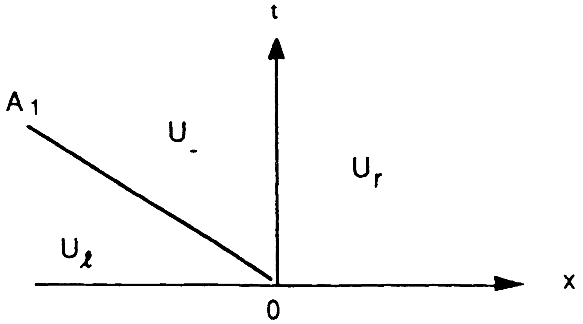


FIG. 5

On the left side of OA_1 the local solution $U_l(t, x)=(u_l(t, x),\ v_l(t, x))$ is still obtained by solving the Cauchy problem for system (1.1) with the initial data $\bar{U}_l(x)$. On the right side of the t-axis the local solution is given by (4.9). Furthermore, on the domain

$$D_-(\delta)=\{(t, x)|0\leqq t\leqq \delta,\ -t\leqq x\leqq 0\}, \tag{6.2}$$

the solution $U_-(t, x)=(u_-(t, x),\ v_-(t, x))$ should satisfy

$$r_-(t, x)=|u_-(t, x)|\leqq 1 \quad \forall (t, x)\in D_-(\delta) \tag{6.3}$$

and the following boundary conditions:

$$v_-=\bar{v}_r \quad \text{on } x=0, \tag{6.4}$$

$$\begin{matrix} u_-=p_l \\ v_-=v_l+(1-r_l)p_l \end{matrix} \quad \text{on } x=-t. \tag{6.5}$$

Then it follows from system (2.1) that on the domain $D_-(\delta)$,

$$\begin{aligned} v_-&=v_-(x)=: v_l(-x, x)+(1-r_l(-x, x))p_l(-x, x), \\ u_-&=u_-(t, x)=: p_l(-x, x)+v'_-(x)(t+x). \end{aligned} \tag{6.6}$$

Thus, in order to satisfy (6.3), certain conditions of compatibility for $\bar{U}_l(x)=(\bar{u}_l(x),\ \bar{v}_l(x))$ should be demanded. In fact, noting that

$$r_-^2(t, x)=1+2p_l(-x, x)\cdot v'_-(x)(t+x)+|v'_-(x)|^2(t+x)^2,$$

by means of system (1.1), it is easy to see that if the initial data $\bar{U}_l(x)$ satisfy the condition

$$\bar{p}_l(0)\cdot \bar{v}'_l(0)-\bar{r}'_l(0)<0, \tag{6.7}$$

then we have (6.3), provided that $\delta>0$ is suitably small.

If (6.7) does not hold, it still might be possible to construct a solution with four waves $x=x_i(t)$ $(i=1,2,3,4)$, some of which are tangent to $x=0$ at the origin. In this situation we guess that the solution to the Riemann problem may be still stable in the L^1 norm.

REFERENCES

[1] B. L. KEYFITZ AND H. C. KRANZER, *A system of non-strictly hyperbolic conservation laws arising in elasticity theory*, Arch. Rational Mech. Anal., 72 (1980), pp. 219-241.

[2] C. CARASSO, M. RASCLE, AND D. SERRE, *Etude d'un modèle hyperbolique en dynamique des câbles*, Modél. Math. Anal. Numér., 19 (1985), pp. 573-599.

[3] M. SHEARER, *The Riemann problem for the planar motion of an elastic string*, J. Differential Equations, 61 (1986), pp. 149-163.

[4] H. GILQUIN AND D. SERRE, *Well-posedness of the Riemann problem; consistency of the Godunov's scheme*, Contemp. Math., 100 (1989), pp. 251-265.

[5] D. SERRE, *Un modèle relaxé pour les câbles inextensibles*, Modél. Math. Anal. Numér., 25 (1991), pp. 465-481.

[6] LI TA-TSIEN AND YU WEN-CI, *Boundary value problems for quasilinear hyperbolic systems*, Duke Univ. Math. Ser. V, Durham, NC, 1985.

[7] ZHAO YAN-CHUN, *Boundary value problems for first order quasilinear hyperbolic systems*, Chinese Ann. Math., 7A (1986), pp. 629-643. (In Chinese.)

[8] LI TA-TSIEN AND PENG YUE-JUN, *Le problème de Riemann généralisé pour une sorte de systèmes des cables*, Portugal. Math., to appear.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1204-1229, September 1992

009

DIMENSIONALITY OF INVARIANT SETS FOR NONAUTONOMOUS PROCESSES*

TEPPER L. GILL† AND W. W. ZACHARY†

Abstract. The existence of global attractors and estimates of their dimensions have been investigated by various authors for a number of dissipative nonlinear partial differential equations which are either autonomous or are subject to time-periodic forcing. In the presence of more general forcing (e.g., almost periodic but not periodic), the usual estimates of the dimensionality of global attractors in terms of uniform (or global) Lyapunov exponents are not valid. This article investigates the estimation of Hausdorff and fractal dimensions of invariant sets corresponding to differential equations of the above type, subject to time-dependent forcing of a quite general class. Working in the framework of skew-product semiflows associated with these equations, the authors consider invariant sets defined in terms of global attractors of semigroups determined by these semiflows. In autonomous situations these invariant sets coincide with the usual global attractors. Upper bounds for the Hausdorff and fractal dimensions of these sets are given in terms of uniform Lyapunov exponents for a large class of dissipative nonlinear partial differential equations with time-dependent forcing terms that include the case of almost periodic functions.

Key words. global attractors, skew-product semiflows, Hausdorff and fractal dimension estimates, time dependent forcing, dissipative nonlinear partial differential equation

AMS(MOS) subject classifications. 35B15, 35B40, 35L15, 58D07, 58D07, 58D25, 38F12

1. Introduction. The existence of global attractors and estimates of their Hausdorff and fractal dimensions have been investigated for numerous dissipative nonlinear partial differential equations (DNLPDE) that are either autonomous [4], [32], [9], [15] or are subject to time-periodic forcing [15]. With suitable conditions on the nonlinearities of such equations on bounded domains, it has been proved that global attractors exist and that they have finite Hausdorff and fractal dimensions. Bounds on these dimensions have been obtained in terms of uniform (or global) Lyapunov exponents (to be distinguished from the corresponding pointwise exponents [1]).

In the presence of more general forcing (e.g., almost periodic but not periodic) the usual dimension estimates of these attractors are not valid because the proofs make essential use of the fact that the solutions of the Cauchy problem for the relevant nonlinear differential equations have a semigroup structure. It is well known that solutions of equations do not have this structure when forcing terms are present that are almost periodic but not periodic, or even for periodic forcing for continuous times. The inadequacy of the customary description of global attractors in nonautonomous situations is evident from recent discussions in the literature. Thus, in the case of periodic forcing, Hale [16] and Haraux [19] have advocated the use of a different definition of global attractor than the usual one in terms of discrete semigroups.

In this article we consider the problem of estimating Hausdorff and fractal dimensions of invariant sets for DNLPDE on bounded domains of a Hilbert space subject to time-dependent nonperiodic forcing, including the case of almost periodic forcing. Our main result is that, for a large class of such equations, the types of estimates of Hausdorff and fractal dimensions of invariant sets usually made in autonomous

* Received by the editors October 17, 1990; accepted for publication (in revised form) January 21, 1992. This research was supported by National Science Foundation grant DMS-8813313, U.S. Air Force Office of Scientific Research contract F49 620-89-C-0079, and U.S. Army Research Office contract DAAL03-89-C-0038.

† Computational Science and Engineering Research Center and Department of Electrical Engineering, Howard University, Washington, DC 20059.

cases can also be carried through in this more general situation. Since the solution-maps of the Cauchy problem do not form semigroups in nonautonomous cases, we use the concept of skew-product semiflow to define appropriate invariant sets. These reduce to the usual global attractors in autonomous cases. Working in the skew-product semiflow framework associated with the given system of equations, we consider invariant sets defined in terms of global attractors of semigroups determined by these semiflows.

One of the principal conditions that we require in order to prove the results indicated above is that, for sufficiently long times, the solution-maps for the DNLPDE are invertible on the invariant set. This is true, in particular, if these maps are invertible on the whole space, and our discussion in § 4 involves a class of nonlinear wave equations and systems of such equations for which this condition is satisfied. There are cases, however, in which the semiflows are not invertible on the whole space but are, nevertheless, defined and invertible on the invariant set. In § 5 we discuss a class of reaction-diffusion equations that have this property. In the cases treated in these two sections, all the hypotheses required for the proof of our dimension estimates are satisfied.

We begin in § 2 by defining skew-product semiflows in the context of abstract DNLPDE and outlining the program that we follow in § 3 in order to obtain our results on Hausdorff and fractal dimension estimates of invariant sets for DNLPDE, and systems of such equations, with time-dependent forcing. Section 4 is devoted to the discussion of a class of nonlinear wave equations with nonlinear dissipation, and in § 5 we give a similar discussion for some reaction-diffusion equations with polynomial growth nonlinearities.

2. Skew-product semiflows. Preliminary remarks. The following discussion of skew-product semiflows is analogous to, but different from, the recent discussion by Raugel and Sell [28] in the context of the Navier–Stokes equations.

Consider a solution ψ of a DNLPDE; i.e., a continuous map from $\mathbb{R}$ to a separable Hilbert space K such that $\psi(t+s)$ represents a solution of the equation at time $t+s(t\in\mathbb{R}^+,\ s\in\mathbb{R})$, corresponding to specified initial data $\psi(s)=\phi\in K$ at time s. We assume that ψ satisfies the following nonlinear stability condition.

Assumption 2.1. For each $R>0$ there exists a positive constant $K(R)$ such that

$$\|\psi(t+s)\|_K \leqq K(R) \quad \text{for all } t\geqq 0 \quad \text{whenever } \|\phi\|_K\leqq R. \tag{2.1}$$

We will consider forcing functions $f\in C_b(\mathbb{R}, K)$, where $C_b(\mathbb{R}, K)$ denotes the Banach space of all bounded continuous functions from $\mathbb{R}$ to K. For $f\in C_b(\mathbb{R}, K)$, we define the *translate of f* by

$$f_\tau(t)\equiv(\sigma(\tau)f)(t)=f(t+\tau), \qquad \tau\in\mathbb{R}. \tag{2.2}$$

Then $f_\tau\in C_b(\mathbb{R}, K)$ and f_τ defines a (two-sided) flow on $C_b(\mathbb{R}, K)$. The *positive hull* $H^+(f)$ of $f\in C_b(\mathbb{R}, K)$ is defined as

$$H^+(f)=\text{Closure}_{C_b(\mathbb{R},K)}\{f_\tau,\ \tau\geqq 0\},$$

and the *hull* $H(f)$ as

$$H(f)=\text{Closure}_{C_b(\mathbb{R},K)}\{f_\tau,\ \tau\in\mathbb{R}\}.$$

Note that $H^+(f)$, $H(f)\subset C_b(\mathbb{R}, K)$ if $f\in C_b(\mathbb{R}, K)$. The *ω-limit set* $\omega(f)$ of $f\in C_b(\mathbb{R}, K)$ is defined by $\omega(f)=\bigcap_{\tau\geqq 0} H^+(f_\tau)$. Note that $\omega(f)$ is an invariant set in $C_b(\mathbb{R}, K)$ relative to the translation group $\{\sigma(\tau),\ \tau\in\mathbb{R}\}$.

To guarantee that $\omega(f)$ is nonempty, we restrict consideration to forcing functions for which $H(f)$ is compact. Then $\omega(f)$ is compact as well as nonempty. We list some cases for which this condition is satisfied (other examples can be found in [28]).

(1) Take $K = L^2(\Omega)$, Ω a bounded subset of $\mathbb{R}^n$, with $f \in L^2(\Omega)$ independent of $t \in \mathbb{R}$. Then $H(f) = \{f\}$.

(2) Let $f \in C_b(\mathbb{R}, K)$ be T-periodic, $f(t+T) = f(t)$ for all $t \in \mathbb{R}$. Then $H(f) = \{f_\sigma, \sigma \in [0, T)\}$.

(3) Let f be asymptotically almost periodic from $\mathbb{R}$ to K; i.e., $f = g + h$ with g almost periodic from $\mathbb{R}$ to K [18], and $\|h(t)\|_K \to 0$ as $t \to \pm\infty$.

DEFINITION 2.1. We will say that f is *admissible* if $H(f)$ is compact in $C_b(\mathbb{R}, K)$.

Given a solution ψ of a DNLPDE defined as in the second paragraph of this section, we define a two-parameter family of maps $W(t, s)$ $(t \in \mathbb{R}^+, s \in \mathbb{R})$ by

$$W(t, s)\phi = \psi(t+s), \qquad \psi(s) = \phi \in K. \tag{2.3}$$

Then we have $W(0, s)\phi = \phi (s \in \mathbb{R}, \phi \in K)$, and

$$W(t+\theta, s)\phi = W(\theta, s+t) \circ W(t, s)\phi \qquad (s \in \mathbb{R}, t, \theta \in \mathbb{R}^+), \tag{2.4}$$

where $\circ$ denotes composition. This is an example of a process in the sense of Dafermos [10].

Given a process W, we define its translate W_τ in an analogous manner to the case for functions in (2.2).

DEFINITION 2.2. Given a process W and $\tau \in \mathbb{R}$, the *τ-translate* of W is the process W_τ defined by

$$W_\tau(t, s)\phi \equiv (\sigma(\tau) W)(t, s)\phi = W(t, \tau+s)\phi, \quad t \in \mathbb{R}^+, \quad s \in \mathbb{R}. \tag{2.5}$$

DEFINITION 2.3. A process W on a Hilbert space K is called *almost periodic* if $\bigcup_{\tau \in \mathbb{R}} W_\tau(t, s)\phi$ is precompact in $C_b(\mathbb{R}, K)$ (as a function of the parameter $s \in \mathbb{R}$) for each $t \in \mathbb{R}^+$ and each $\phi \in K$.

Thus, if W is almost periodic, for any sequence $\{\sigma_n\} \subset \mathbb{R}$ there exists a subsequence $\{\sigma_{n_m}\} \subset \{\sigma_n\}$ and a map $V: \mathbb{R}^+ \times \mathbb{R} \times K \to K$ such that

$$\| W_{\sigma_{n_m}}(t, s)\phi - V(t, s)\phi \|_K \to 0 \quad \text{as } m \to +\infty \tag{2.6}$$

uniformly in $s \in \mathbb{R}$ for each $t \in \mathbb{R}^+$ and each $\phi \in K$.

DEFINITION 2.4. Let $W(t, s)\phi$ be an almost periodic process from $\mathbb{R}^+ \times \mathbb{R} \times K$ to K. The closure in $C_b(\mathbb{R}, K)$ of the set of translates of W relative to the above sense of convergence is called the *hull* of W, denoted by $H(W)$.

We will prove later that W is almost periodic if $H(f)$ is compact, and W depends on f in a Lipschitz continuous manner. See Proposition 3.1.

To define a skew-product structure, we let W denote an almost periodic process corresponding to a globally defined unique solution of a given differential equation as in (2.3), and we define the mappings

$$\pi_s(t)(\phi, V) = (V(t, s)\phi, \sigma(t) V), \quad V \in H(W), \quad \phi \in K, \quad s \in \mathbb{R}, \quad t \in \mathbb{R}^+. \tag{2.7}$$

It is easily shown formally [16, p. 44] that $\{\sigma(t), t \in \mathbb{R}\}$ is a group on $H(W)$ and that $\{\pi_s(t), t \geqq 0\}$, with s fixed in $\mathbb{R}$, is a semigroup on $K \times H(W)$. We will prove later that they are C^0 if certain conditions are satisfied. As we shall see, $\{\pi_s(t), t \geqq 0\}$ (s fixed in $\mathbb{R}$) defines a semiflow on $K \times H(W)$ if global solutions of the DNLPDE exist for each $h \in H(f)$. In addition, we will state sufficient conditions such that there exists a one-to-one correspondence between processes in $H(W)$ corresponding to a given admissible forcing function f and the distinguished almost periodic process W with a forcing function $h \in H(f)$. See equations (2.10)-(2.12) and the associated discussion.

For ordinary differential equations, it has been more conventional to define a skew-product structure in terms of translated forcing functions, rather than in terms of translated processes. Thus, consider the system of equations $u_t = f(t, u)$, where $f(\cdot, \cdot): \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Then, under appropriate continuity conditions, we define a skew-product structure on $\mathbb{R}^n \times H(f)$ by $\tilde{\pi}_s(t)(x, f) = (u(t, s)x, \sigma(t)f)$, where $u(s, s)x = x \in \mathbb{R}^n$. For more details of this approach, we refer the reader to [16], [29], [30].

Let s be fixed in $\mathbb{R}$. A compact set A in $K \times H(W)$ is said to be an *attractor* if it is invariant under the action of $\pi_s(t)$, $\pi_s(t)A = A$ for $t \in \mathbb{R}$, and if there exists an open neighborhood U of A such that $\pi_s(t)$ B converges to A as $t \to +\infty$, where B denotes any bounded subset of U. If these properties remain true when U is replaced by the whole space $K \times H(W)$, then A is called the *maximal* or *global attractor*. At the end of the present section we will state conditions under which the global attractor is independent of s.

Given an almost periodic process W, let ${}_sA_\pi(W)$ denote a global attractor (if one exists) of the semigroup $\{\pi_s(t), t \geqq 0\}$ (s fixed in $\mathbb{R}$) and, following the discussion in [16], consider the set

$$E_s = \{\chi \in K \mid (\chi, V) \in {}_sA_\pi(W),\ V \in H(W)\}. \tag{2.8}$$

Some insight concerning the relevance of the sets E_s to the study of dynamical systems can be gleaned by considering autonomous and periodic processes. For an autonomous process, $W(t, s)$ is independent of s, and $H(W)$ consists of the single process W. We find $A_\pi(W) = E \times \{W\}$ and $\pi(t) = S(t) \times I$, where $S(t)$, I, and E denote the usual solution-map semigroup for an autonomous process, the identity map on processes, and the global attractor for $S(t)$, respectively. For T-periodic processes we have

$$H(W) = \{W_\sigma, \sigma \in [0, T)\}. \tag{2.9}$$

It is known that the set

$$A(W) = \bigcup_{\sigma \in [0, T)} W(\sigma, 0) \bigcap_{m \geqq 0} \text{Closure}_K \left(\bigcup_{n \geqq m} W(nT, 0)B_0 \right),$$

with B_0 a bounded absorbing set, corresponds to a set of the type (2.8) for T-periodic processes [16]. Thus, for autonomous and T-periodic processes, the usual global attractors correspond to sets of the form (2.8).

In the present work we will establish the following results for processes corresponding to solutions of admissible time-dependently forced DNLPDE: (a) proof of existence of sets of the type (2.8), (b) proof that these sets have finite Hausdorff and fractal dimensions, and (c) derivation of upper bounds for these dimensions in terms of uniform Lyapunov exponents. These results will be based on the idea that, for DNLPDE subject to admissible time-dependent nonperiodic forcing, the Hausdorff and fractal dimensions of sets of the form (2.8) can be estimated by consideration of the first variational equations corresponding to equations in the hull of the given equation (or system of equations). Here, an equation with forcing function h is said to belong to the hull of a given equation with forcing function f if $h \in H(f)$.

The fundamental result on the invariance of the sets (2.8) is the following.

LEMMA 2.1 (positive invariance of E_s). *Let W be an almost periodic process on K, and let ${}_sA_\pi(W)$ be a global attractor on $K \times H(W)$ relative to the semigroup $\{\pi_s(t), t \geqq 0\}$. Then, for given $s \in \mathbb{R}$, $\phi \in E_s$ implies that there exists a process $Z \in H(W)$ such that $Z(t, s)\phi \in E_s$ for all $t \geqq 0$.*

For related results, see Hale [16, p. 46] and Dafermos [10], [11]. The first of these references refers only to periodic processes. In that case the situation is simpler than for almost periodic nonperiodic processes because the hull of a periodic process has the special structure (2.9).

We can readily give examples of processes V belonging to the hull of a given almost periodic process W that are not translates of W. Examples of this phenomenon are well known in the case of uniform (Bohr) almost periodic nonperiodic functions [13], and examples of almost periodic nonperiodic processes with the desired property can be constructed in terms of such functions.

As noted in [16] and [19], under reasonable conditions we should be able to prove that, when W is almost periodic, $H(W)$ consists precisely of the set of processes in the hull of the equation (or system of equations) under consideration. Take $V \in H(W)$ with W almost periodic. Then, by Definition 2.3, there exists a sequence $\{\tau_n\} \subset \mathbb{R}$ such that $W_{\tau_n}(t, s)\phi \rightarrow_{n\rightarrow+\infty} V(t, s)\phi$ for each $t \in \mathbb{R}^+$ and each $\phi \in K$ uniformly in $s \in \mathbb{R}$. For the translated process we should have

$$W_{\tau_n}(t, s; f)\phi = W(t, s+\tau_n; f)\phi \tag{2.10}$$

$$= W(t, s; f_{\tau_n})\phi. \tag{2.11}$$

Modulo certain continuity arguments, it would then follow that V is given by

$$V(t, s) = W(t, s; h), \tag{2.12}$$

where $h = \lim_{n\rightarrow+\infty} f_{\tau_n} \in H(f)$ is the uniform limit of the f_{τ_n}. These continuity considerations will be discussed in § 3 for processes corresponding to solutions of a large class of DNLPDE. In view of Lemma 2.1 and the above considerations, it is useful to consider the maps (with s given in $\mathbb{R}$)

$$S_s(t)\phi = W(t, s; h)\phi, \qquad h \in H(f). \tag{2.13}$$

In order to prove existence of sets of the type (2.8) and to obtain estimates for their Hausdorff and fractal dimensions, we will prove the following statements, which are analogous to the corresponding program in autonomous situations but contain some additional requirements.

(1) For a distinguished almost periodic process W related to a globally defined solution of a DNLPDE (or a system of such equations) as in (2.3), $\{\sigma(t),\ t \geqq 0\}$ and $\{\pi_s(t),\ t \geqq 0,\ s \text{ fixed in } \mathbb{R}\}$ are C^0-semigroups in $H(W)$ and $K \times H(W)$, respectively.

(2) For each $t \geqq 0$ and $S \in \mathbb{R}$; $S_s(t)$ exists, is unique, and is differentiable on K for all initial data in K and for all $h \in H(f)$.

(3) Proof of the relations between $H(W)$ and $H(f)$ indicated in (2.10)-(2.12).

(4) Nonlinear stability. (See Assumption 2.1.)

(5) Existence of bounded absorbing sets.

(6) Asymptotic compactness: for all bounded sets $B \subset K$, there exists a compact set $G \subset K$ such that

$$\sup_{\chi \in B} d(W(t, s)\chi, G) \rightarrow 0 \quad \text{as } t \rightarrow +\infty \tag{2.14}$$

for each $s \in \mathbb{R}$, where, for two sets X, $Y \subset K$,

$$d(X, Y) = \sup_{\Phi \in X} \inf_{\chi \in Y} \|\Phi - \chi\|_K. \tag{2.15}$$

(7) For each $h \in H(f)$ and for $t > 0$ sufficiently large, $S_s(t)$ has an inverse on the range of $S_s(t)E_s$ that is Lipschitz continuous.

Points (1) and (3) are, of course, specific to the framework of skew-product semiflows, while the conditions in (2) and (4)-(6) require generalizations of corresponding results already known for autonomous and periodic processes. In particular, proofs are required for all $h \in H(f)$. We will see in § 3 that the validity of conditions (4)-(6) implies the existence of global attractors ${}_sA_\pi(W)$ for the semigroups $\{\pi_s(t),\ t \geqq 0\}$ (s fixed in $\mathbb{R}$).

Condition (7) is a crucial property which allows us to obtain estimates for the Hausdorff and fractal dimensions of the sets (2.8) in an analogous manner to the method used to obtain corresponding estimates in autonomous [32], [4], [7]-[9], [15] and periodic [15] processes. It can be seen to be a natural requirement by the following argument. Consider a semigroup $\{S(t),\ t \geqq 0\}$ corresponding to an autonomous process. Then, as is well known, a global attractor for this process is invariant under the action of this semigroup,

$$S(t)A = A, \qquad t \geqq 0. \tag{2.16}$$

Lemma 2.1 can be thought of as a generalization to almost periodic processes of the positive invariance condition for global attractors in the autonomous case, $S(t)A \subset A (t \geqq 0)$, which is "one-half" of the invariance condition (2.16). However, it is known for autonomous systems that the important condition required for estimates of Hausdorff and fractal dimensions of global attractors is that of negative invariance, $S(t)A \supset A(t \geqq 0)$ [24], [32]. Returning now to the consideration of almost periodic processes, we will see later that condition (7) allows us to transform the positive invariance of E_s under $S_s(t)$, as described by Lemma 2.1 and (2.13), into the negative invariance of E_s under $(S_s(t))^{-1}$.

Actually, while condition (7) only requires that $S_s(t)$ be invertible and that its inverse be Lipschitz continuous on the range of $S_s(t)$, there are a number of DNLPDE's with admissible time-dependent forcing for which these maps are invertible on the whole Hilbert space. We will discuss some equations of this type in § 4. We expect that there are many situations for which the maps $S_s(t)$ are invertible on their range even though they may not be invertible on all of K. In § 5 we will discuss a class of parabolic equations with admissible time-dependent forcing that have this property.

As a result of condition (7), it can be shown that the global attractors ${}_sA_\pi(W)$ are actually independent of s. This follows from the previously noted result that conditions (4)-(6) imply the existence of these attractors and a straightforward modification of [19, Prop. 1.10].

3. General results. In this section our results will be formulated in an abstract manner and then, in the following two sections, examples of DNLPDE's that satisfy our hypotheses will be discussed.

We first establish the continuity of the skew-product semiflow $\{\pi_s(t),\ t \geqq 0\}(s \in \mathbb{R})$ defined in (2.7). For analogous proofs in the case of ordinary differential equations, see [29], [30], [6]. Consider the formulation of § 2 in which K is a separable real Hilbert space and $W(t, s)$ a distinguished process associated as in (2.3) with a solution ψ of a given DNLPDE.

The following proposition establishes (2.10), (2.11), and related results.

PROPOSITION 3.1. *Define a distinguished process $W(t, s)$ in terms of a globally defined uniformly bounded unique solution ψ of a system of* DNLPDE'*s that satisfies* (2.1) *and, in addition, assume that ψ depends on the forcing function f in a Lipschitz-continuous manner: i.e., there exists a positive constant $c(t)$, generally depending upon t, such that for two solutions ψ_1, ψ_2 (with the same initial data) corresponding to two* *forcing terms* f_1, f_2,

$$\|\psi_1(t+s)-\psi_2(t+s)\|_K \leqq c(t) \sup_{t'\in[s,t+s]} \|f_1(t')-f_2(t')\|_K. \tag{3.1}$$

Finally, assume that the forcing functions f_1, f_2 *are admissible. Then*

(a) *For all* $\tau \in \mathbb{R}$,

$$\begin{aligned} W_\tau(t, s; f)\psi(s) &= W(t, \tau+s; f)\psi(\tau+s) \\ &= W(t, s; f_\tau)\psi(s), \end{aligned}$$

and, more generally,

(b) *Given* $V \in H(W)$, *there exists* $h \in H(f)$ *such that*

$$W(t, s; h)\psi(s) = V(t, s; f)\psi(s). \tag{3.2}$$

Conversely, given $h \in H(f)$, *there exists* $V \in H(W)$ *such that* (3.2) *holds.*

Proof. By using the existence and uniqueness of the solutions of the system of differential equations it is easy to establish (a). To prove (b), we first note that it follows from (3.1) and (a) that W is almost periodic if f is admissible. To see this, take a sequence $\{W_{\tau_n}(t, s)\phi\}$ from $\bigcup_{\tau\in\mathbb{R}} W_\tau(t, s)\phi$. Since $H(f)$ is compact in $C_b(\mathbb{R}, K)$, there exists a subsequence $\{\tau_{n_m}\} \subset \{\tau_n\}$ such that $\{f_{\tau_{n_m}}\}$ is convergent, and, therefore, Cauchy; i.e., given $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that l, $m \geqq N$ implies that

$$\|f_{\tau_{n_m}}(t') - f_{\tau_{n_l}}(t')\|_K < \frac{\varepsilon}{c(t)}$$

for all $t' \in [s, t+s]$. From (a) and (3.1) we then obtain $\|W_{\tau_{n_m}}(t, s; f)\phi - W_{\tau_{n_l}}(t, s; f)\phi\|_K < \varepsilon$ for l, $m \geqq N$ and $\varepsilon > 0$ so that $\{W_{\tau_{n_m}}(t, s; f)\phi\}$ is a Cauchy sequence and, therefore, convergent. Thus, from an arbitrary sequence in $\bigcup_\tau W_\tau(t, s; f)\phi$ we have obtained a convergent subsequence from which we infer that $\bigcup_{\tau\in\mathbb{R}} W_\tau(t, s)\phi$ is precompact in $C_b(\mathbb{R}, K)$ for each $t \in \mathbb{R}^+$, $s \in \mathbb{R}$, $\phi \in K$, and admissible f.

Take $V \in H(W)$. Then, since W is almost periodic, there exists a sequence $\{\tau_n\} \subset \mathbb{R}$ such that

$$W_{\tau_n}(t, s; f)\psi(s) \xrightarrow[n\to+\infty]{} V(t, s; f)\psi(s) \tag{3.3}$$

for each $t \in \mathbb{R}^+$ and each $\psi(s) = \phi \in K$, uniformly in $s \in \mathbb{R}$. From (a), this can be expressed as

$$W(t, s; f_{\tau_n})\psi(s) \xrightarrow[n\to+\infty]{} V(t, s; f)\psi(s).$$

To complete the proof, we obtain a relation between the convergence of translates of the forcing functions and translates of processes. Let

$$f_{\tau_n} \xrightarrow[n\to+\infty]{} h \in H(f), \tag{3.4}$$

so that, by (3.1),

$$\|\psi_1(t+s, f_{\tau_n}) - \psi_2(t+s, h)\|_K \leqq c(t) \sup_{t'\varepsilon[s,\, t+s]} \|f_{\tau_n}(t') - h(t')\|_K, \tag{3.5}$$

i.e.,

$$\lim_{n\to+\infty} W(t, s; f_{\tau_n})\psi(s) = W(t, s; h)\psi(s) \tag{3.6}$$

so that, upon comparison with (3.3),

$$W(t, s; h)\psi(s) = V(t, s; f)\psi(s).$$

Conversely, take $h \in H(f)$. Then there exists a sequence $\{\tau_n\} \subset \mathbb{R}$ such that the limit in (3.4) exists in the uniform topology. Then, from (3.6) and (a),

$$\lim_{n\to+\infty} W_{\tau_n}(t, s; f)\psi(s) = W(t, s; h)\psi(s),$$

and, using (3.5), we easily show that this result holds for each $t \in \mathbb{R}^+$ and each $\psi(s) = \phi \in K$ uniformly in $s \in \mathbb{R}$. We see that $W(t, s; h) \in H(W(t, s; f))$, and the proof is complete.

PROPOSITION 3.2. *Assume the hypotheses of Proposition* 3.1 *and further, assume that the translation group* $\{\sigma(\tau), \tau \in \mathbb{R}\}$ *defined in* (2.5) *is continuous on* $H(W)$. *Then, for each fixed* $s \in \mathbb{R}$, $\pi_s(t): K \times H(W) \to K \times H(W)$ *is jointly continuous.*

Proof. From (2.7) and (2.3) we have

$$\pi_s(t)(\phi, W) = (W(t, s; f)\phi, \sigma(t)W) = (\psi(t+s; f, \phi), W_t)$$

for $t \in \mathbb{R}^+$, $s \in \mathbb{R}$, $f \in C_b(\mathbb{R}, K)$, and $\psi(s) = \phi$. Let $\{(\phi_n, W_{\tau_n})\}$ and $\{t_n\}$ denote sequences in $K \times H(W)$ and $\mathbb{R}^+$, respectively, such that $(\phi_n, W_{\tau_n}) \to (\phi, Y)$ and $t_n \to t$ as $n \to +\infty$. We must prove that, for fixed $s \in \mathbb{R}$,

$$\pi_s(t_n)(\phi_n, W_{\tau_n}) \to \pi_s(t)(\phi, Y) \text{ as } n \to +\infty, \quad \text{i.e., } (W_{\tau_n}(t_n, s; f)\phi_n, W_{\tau_n+t_n}) \to (Y(t, s; f)\phi, Y_t)$$

or, using part (a) of Proposition 3.1,

$$(W(t_n, s; f_{\tau_n})\phi_n, W_{\tau_n+t_n}) \to (Y(t, s; f)\phi, Y_t).$$

We have $W_{\tau_n+t_n} \to Y_t$ by the continuity of the translation group (2.5) on $H(W)$, and

$$W(t_n, s; f_{\tau_n})\phi_n \to W(t, s; h)\phi, \qquad h \in H(f)$$

because of the Lipschitz condition (3.1), $\phi_n \to \phi$, and that the solutions ψ of the system of DNLPDE's are uniformly bounded for all $t \in \mathbb{R}^+$ and are uniquely determined by the initial data. The proof is concluded by noting that $W(t, s; h)\phi = Y(t, s; f)\phi$ by Proposition 3.1.

LEMMA 3.1. *Assume the hypotheses of Proposition* 3.1. *and, in addition, assume that the translation group* $\{\sigma(\tau), \tau \in \mathbb{R}\}$ *is continuous and that* $\{S_s(t), t \geqq 0\}$ *possesses a bounded absorbing set* $B_0 \subset K$; *i.e., for each* $h \in H(f)$, $W(t, s; h)\chi \subset B_0$ *whenever* $t \in \mathbb{R}^+$ *is sufficiently large, and* χ *belongs to any bounded subset of* K. *Furthermore, assume that* $\{S_s(t), t \geqq 0\}$ *is asymptotically compact*: *there exists a compact set* $G \subset K$ *that attracts all bounded subsets of* K *under* $S_s(t)$;

$$\lim_{t\to+\infty} \sup_{\psi(s)=\phi\in K} d_K(S_s(t)\psi(s), G) = 0, \tag{3.7}$$

where d_K *is defined in* (2.15). *Then, the skew-product semiflow* (2.7) *has the following properties. For fixed* $s \in \mathbb{R}$,

(a) *There exists a bounded absorbing set in* $K \times H(W)$ *for the semigroup* $\{\pi_s(t), t \geqq 0\}$;

(b) $\pi_s(t)$*-orbits of bounded sets are bounded*;

(c) $\{\pi_s(t), t \geqq 0\}$ *is asymptotically compact*: *there exists a compact set in* $K \times H(W)$ *which attracts all bounded subsets of* $K \times H(W)$.

Proof. Since f is admissible, we see that W is almost periodic as in the proof of Proposition 3.1.

(a) Let $B_1 \times B_2$ be a bounded subset of $K \times H(W)$. Then B_2 has the form

$$B_2 = \{W_{\tau_n}\} \tag{3.8}$$

for a sequence $\Lambda = \{\tau_n\} \subset \mathbb{R}$. By (2.7),

$$\pi_s(t)(\chi, V) = (V(t, s; f)\chi, \sigma(t)V)$$

for $(\chi, V) \in B_1 \times B_2$. So from $V \in B_2$ we infer from (3.8) that $V \in \bigcup_{\tau \in \Gamma} W_\tau$, where Γ denotes some subsequence of Λ. Thus,

$$\pi_s(t)(\chi, V) \in \left(\bigcup_{\tau \in \Gamma} W_\tau(t, s; f)\chi, \sigma(t) \bigcup_{\tau \in \Gamma} W_\tau \right) = \bigcup_{\tau \in \Gamma} (W(t, s; f_\tau)\chi, W_{\tau+t}), \tag{3.9}$$

where we have used Proposition 3.1 and the continuity of $\sigma(t)$ to obtain the last line. By assumption, $\{S_s(t), t \geqq 0\}$ possesses a bounded absorbing set $B_0 \subset K$. Take $\chi \in B_0$. Then a bounded absorbing set for $\{\pi_s(t), t \geqq 0\}$ is $B_0 \times \{W_\tau, \text{all } \tau \in \mathbb{R}\}$.

(b) Again take $(\chi, V) \in B_1 \times B_2$ with B_1 and B_2 as in (a). Then (3.9) holds and, by (2.1), there exists $K(R) > 0$ such that

$$\| W(t, s; f_\tau)\chi \|_K \leqq K(R) \tag{3.10}$$

for all $t \in \mathbb{R}^+$. By Proposition 3.1,

$$\begin{aligned} W_{\tau+t}(p, s; f)\chi &= W(p, s; f_{t+\tau})\chi \\ &= W(p, s+t+\tau; f)\chi \end{aligned}$$

is bounded in $C_b(\mathbb{R}, K)$ with respect to $s+t+\tau \in \mathbb{R}$ for each $p \in \mathbb{R}^+$ and each $\chi \in K$. Combination of this result with (3.10) proves assertion (b).

(c) We again use (3.9) with B_1 and B_2 as in (a) and (b). By hypothesis, there exists a compact set $Y \subset K$ which attracts all bounded subsets of K under $S_s(t)$, and, therefore, under $W(t, s; f_\tau)$ in (3.9). So, $Y \times H(W)$ attracts all bounded sets in $K \times H(W)$ under $\pi_s(t)$. Finally, $H(W)$ is compact (in $C_b(\mathbb{R}, K)$) so that $Y \times H(W)$ is compact in the product topology by Tychonoff's theorem, and the proof is complete.

We now consider the maps (2.13). Assuming that these are Fréchet differentiable on sets of type (2.8) and that the corresponding derivatives $L_s(t, \chi)$ are bounded on $L(K)$, we set up an apparatus for estimating Hausdorff and fractal dimensions of invariant sets in an analogous way to the treatment of autonomous equations [12], [7]-[9], [32], [4].

Define

$${}_s\bar{\omega}_j(t) = \sup_{\chi \in E_s} \omega_j(L_s(t, \chi)), \quad j \in \mathbb{N}, \quad t \in \mathbb{R}^+, \tag{3.11}$$

where

$$\omega_j(L_s(t, \chi)) = \alpha_1(L_s) \cdots \alpha_j(L_s) \quad \text{with}$$

$$\alpha_j(L_s(t, \chi)) = \sup_{\substack{F \subset K \\ \dim F = j}} \inf_{\substack{\eta \in F \\ \|\eta\|_K = 1}} \|L_s(t, \chi)\eta\|_K \qquad (j \in \mathbb{N}).$$

For noninteger cases we set $\omega_d(L) = (\omega_n(L))^{1-s}(\omega_{n+1}(L))^s$ for $d = n+s$, $n = \text{integer} \geqq 1, 0 < s < 1$.

Consider the composition property (2.4) for a process U corresponding to a function h in the hull of a forcing function f for a given DNLPDE:

$$\mathsf{U}(t+p, s, \chi(s); h) = \mathsf{U}(t, p+s, \mathsf{U}(p, s, \chi(s); h); h)$$

or

$$\mathsf{U}(t+p, 0, \chi(0); h_s) = \mathsf{U}(t, 0, \mathsf{U}(p, 0, \chi(0); h_s); h_{p+s})$$

using Proposition 3.2. Alternatively, in terms of the maps (2.13),

$$S_s(t+p)\chi(s) = S_{p+s}(t) \circ S_s(p)\chi(s). \tag{3.12}$$

The relations corresponding to (3.12) for the derivatives $L_s(t, \chi)$ (assuming that they exist) are

$$\begin{aligned} L_s(t+p, \chi(s)) &\equiv S_s'(t+p)(\chi(s)) \\ &= L_{p+s}(t, S_s(p)(\chi(s))) \circ L_s(p, \chi(s)). \end{aligned} \tag{3.13}$$

Corresponding to a DNLPDE, we have a first variational equation (see [32], [4], [9] for the corresponding autonomous situation). Let $\Phi(t)$ denote a solution of the latter equation subject to the initial value $\rho \in K$. Then, as is well known, we have the following relation between $L_s(t, \chi)$ and $\Phi(t)$:

$$L_s(t, \chi(s))\rho = \Phi(t), \qquad \Phi(s) = \rho \in K. \tag{3.14}$$

We assume that the corresponding equation for $\Phi(t)$ is independent of the forcing term of the original DNLPDE. This is true for a large number of DNLPDE's, and we discuss some examples in the following two sections. In such cases, L depends on the subscripts in (3.13) only through its dependence on $S_s(t)$, and we can write

$$L(t+p, \chi(s)) = L(t, S_s(p)(\chi(s))) \circ L(p, \chi(s))$$

in place of (3.13).

We now prove the important subexponential property for the quantities (3.11) for DNLPDE with time-dependent admissible forcing functions.

THEOREM 3.1. *Assume the hypotheses of Proposition* 3.2 *and, in addition, assume that the derivatives L defined in* (3.13) *exist. Then the quantities* (3.11) *satisfy the subexponential condition*

$${}_s\bar{\omega}_j(t+p) \leqq {}_s\bar{\omega}_j(t)\,{}_s\bar{\omega}_j(p); \quad t, p \in \mathbb{R}^+, \quad j \in \mathbb{N},$$

where s is a fixed real number.

Proof. From Lemma 2.1, $\phi \in E_s$ implies that, for given $s \in \mathbb{R}$, there exists a process $V \in H(W)$ such that $V(t, s)\phi \in E_s$ for all $t \geqq 0$. As in the proof of Proposition 3.1, f admissible implies that W is almost periodic, from which we infer that V is almost periodic, so that there exists a sequence $\{\tau_n\} \subset \mathbb{R}$ such that $W_{\tau_n}(t, s)\phi \to_{n \to +\infty} V(t, s)\phi$ for each $t \in \mathbb{R}^+$ and each $\phi \in K$, uniformly in $s \in \mathbb{R}$. Moreover, from Proposition 3.2, we see that $V(t, s)\phi = W(t, s; h)\phi$ with $h \in H(f)$.

Consider the maps (2.13). These satisfy (3.12), and, therefore, also (3.14), if the corresponding derivatives L exist. Thus we obtain

$$\omega_j(L_s(t+p, \phi)) \leqq \omega_j(L_{p+s}(t, S_s(p)(\phi)))\omega_j(L_s(p, \phi))$$

by [32, Cor. 1.1, p. 267]. Following our earlier argument, which centered about (3.14), we see that we can remove the subscripts on L so that

$$\omega_j(L(t+p, \phi)) \leqq \omega_j(L(t, S(p)(\phi)))\omega_j(L(p, \phi)).$$

The proof can now be completed as in the autonomous case.

We have proved that the quantities $\{{}_s\bar{\omega}_j, j \in \mathbb{N}\}$ are subexponential even though the maps (2.13) are not semigroups. This result allows one to prove that the limits

$$\lim_{t \to +\infty} ({}_s\bar{\omega}_j(t))^{1/t} \tag{3.15}$$

exist for each $s \in \mathbb{R}$, and uniform Lyapunov exponents can be defined in an analogous way to autonomous situations [7]-[9], [32]. We will discuss this in more detail at the end of the present section. Before doing this, we obtain estimates of Hausdorff and fractal dimensions of sets of the form (2.8).

Generalizations will be given of [32, Thms. 3.1-3.3, pp. 282-289] to the case of admissible time-dependent forcing.

THEOREM 3.2. *Assume the hypotheses of Proposition* 3.1 *as well as the condition of asymptotic compactness* (3.7), *and consider the maps* (2.13) *corresponding to an admissible forcing function f. In addition, assume that the derivatives L exist. Assume also that*

$$\sup_{t\in[0,1]}\ \sup_{u\in E_s} \|L(t,u)\|_{L(K)} \leqq m < +\infty$$

for some $m>0$ *as well as the condition*

$$\sup_{t\in[0,1]} \omega_d(L(t,u)) < 1 \tag{3.16}$$

for some $d>0$ *and all* $u\in E_s$ *for some fixed* $s\in\mathbb{R}$.

Finally, we assume that, for all $h\in H(f)$, *the maps* $S_s(t)$ *are Lipschitz continuous on* E_s *with inverses that exist as Lipschitz continuous surjective maps from* $S_s(t)E_s$ *onto* E_s *when* $t>0$ *is sufficiently large. Then the Hausdorff dimension of* E_s, $d_H(E_s)$, *is finite and* $d_H(E_s)\leqq d$.

Proof. By hypothesis, the derivatives L of the maps (2.13) exist. From Proposition 3.2, Lemma 3.1, and [16, Thm. 3.7.2] there exists a global attractor $A_\pi(W)$ for $\{\pi_s(t), t\geqq 0\}$ which, by the result stated at the end of § 2, is independent of s. Moreover, E_s is compact. Then, by the method of proof used in autonomous cases [12], [7]-[9], [32], [4], it follows that the set $S_s(t)E_s$ has zero Hausdorff d-measure when $t>0$ is sufficiently large:

$$\mu_H(S_s(t)E_s, d) = 0. \tag{3.17}$$

By hypothesis, for all $h\in H(f)$, $S_s(t)$ is Lipschitz-continuous on E_s with a Lipschitz-continuous inverse when $t>0$ is sufficiently large. Then, since Hausdorff measure has the property $\mu_H(FB,d)\leqq(\mathrm{Lip}\,F)^d\mu_H(B,d)$ for Lipschitz maps F on metric spaces [21], it is clear that $\mu_H(E_s,d)\leqq \mathrm{Lip}\,((S_s(t))^{-1})^d\mu_H(S_s(t)E_s,d)$ so that, using (3.17), $\mu_H(E_s,d)=0$ and $d_H(E_s)\leqq d$.

As with corresponding autonomous situations, estimates of the fractal dimension of sets of the form (2.8) proceed in a similar manner to the estimates of their Hausdorff dimension in the preceding theorem, except that more stringent hypotheses are required.

THEOREM 3.3. *Consider the same hypotheses as in Theorem* 3.2 *with the exception that condition* (3.16) *is replaced by the following condition.*

For some $d=n+s$, $n\in\mathbb{N}$ *with* $n>1$, $s\in(0,1]$, ${}_s\bar{\bar{\omega}}_j({}_s\bar{\bar{\omega}}_{n+1})^{d-j/n+1}<1$ *for* $j=1,\cdots,n$, *where* ${}_s\bar{\bar{\omega}}_j=\sup_{t\in[0,1]}{}_s\bar{\omega}_j(t)$. *Then the fractal dimension of* E_s, $d_F(E_s)$, *is finite and* $d_F(E_s)\leqq d$.

Proof. The proof is similar to that of Theorem 3.2. From the method of proof used in autonomous cases (see, e.g., [32, pp. 284-287]) we obtain a covering of the set $Y_s(t)=S_s(t)E_s$ by a minimum number $n(\varepsilon)$ of K-balls of radius $\varepsilon>0$ when $t>0$ is sufficiently large. The assertion of the theorem follows from an estimate of the capacity of the set $Y_s(t)$ when t is sufficiently large, obtained by the methods indicated above, the compactness of the set E_s, and the result that the capacity of a compact set is invariant under a mapping of the set by a Lipschitz-continuous homeomorphism with a Lipschitz-continuous inverse [27].

With the results of the preceding two theorems in hand, we can define uniform Lyapunov numbers and uniform Lyapunov exponents in an analogous manner to autonomous cases [7]-[9], [32]. Let $\Pi_{j,s}$ denote the respective limits (3.15). They exist because of the subexponential property of the quantities $\{{}_s\bar{\omega}_j\}$. We then define the quantities $\Lambda_{j,s}(j=1,\cdots,m)$ recursively by

$$\Lambda_{1,s}=\Pi_{1,s},\qquad \Lambda_{1,s}\Lambda_{2,s}=\Pi_{2,s},\cdots,\Lambda_{1,s}\cdots\Lambda_{m,s}=\Pi_{m,s}. \tag{3.18}$$

The quantities $\Lambda_{m,s}$ and

$$\mu_{m,s}=\log\Lambda_{m,s},\quad m\geqq 1 \tag{3.19}$$

will be called uniform (or global) Lyapunov numbers and exponents, respectively, for the sets E_s.

Using these exponents, we can give alternative versions of Theorems 3.2 and 3.3.

THEOREM 3.4. *Assume the hypotheses of Theorem* 3.2, *and consider the maps* (2.13) *corresponding to admissible forcing functions. If, for some* $n>1$,

$$\mu_{1,s}+\cdots+\mu_{n+1,s}<0,$$

then $\mu_{n+1,s}<0$, $\mu_{1,s}+\cdots+\mu_{n,s}/|\mu_{n+1,s}|<1$, *and*

(i) $\quad d_H(E_s)\leqq n+(\mu_{1,s}+\cdots+\mu_{n,s})_+/|\mu_{n+1,s}$,

(ii) $\quad d_F(E_s)\leqq(n+1)\max\limits_{1\leqq j\leqq n}\left(1+\dfrac{(\mu_{1,s}+\cdots+\mu_{j,s})+}{|\mu_{1,s}+\cdots+\mu_{n+1,s}|}\right)$,

where $x_+=\max(x,0)$.

The proof is analogous to that of [32, Thm. 3.3, p. 287].

As in autonomous situations, it is convenient to define auxiliary quantities (see (3.21) below) from which Lyapunov exponents and also the Hausdorff and fractal dimensions of invariant sets can be estimated. Thus, if we write the first variational equation corresponding to the mapping (3.14) in the form

$$\frac{d}{dt}\Phi(t)=F'(\psi)\Phi(t), \tag{3.20}$$

then we have, in the m-fold exterior product of the Hilbert space K,

$$\|\Phi_1(t)\Lambda\cdots\Lambda\Phi_m(t)\|_{\Lambda^m K}=\|\rho_1\Lambda\cdots\Lambda\rho_m\|_{\Lambda^m K}\exp\left(\int_0^t\operatorname{Tr}F'(S_s(p)(\phi))\circ Q_m(p)\,dp\right),$$

where $Q_m(p)=Q_m(p,\phi;\rho_1,\cdots,\rho_m)$ is the orthogonal projection onto the subspace spanned by $\Phi_1(p),\cdots,\Phi_m(p)$. We then define

$${}_sq_m=\limsup_{t\to+\infty}{}_sq_m(t) \tag{3.21}$$

with

$${}_sq_m(t)=\sup_{\chi\in E_s}\ \sup_{\substack{\rho_i\in K\\ \|\rho_i\|_K\leqq 1\\ (i=1,\cdots,m)}}\left(\frac{1}{t}\int_0^t\operatorname{Tr}F'(S_s(p)(\chi))\circ Q_m(p)\,dp\right).$$

Then, for a given DNLPDE with time-dependent forcing by admissible functions, we can obtain bounds for the uniform Lyapunov exponents in terms of the ${}_sq_j$. We will discuss this for a class of dissipative nonlinear wave equations in the following section, and for a class of reaction-diffusion equations in § 5.

4. Nonlinear wave equations with nonlinear dissipation. In this section we consider equations of the form

$$u_{tt}+\beta(u_t)-\Delta u+g(u)=f\qquad\text{on }\Omega\times[s,\infty) \tag{4.1}$$

for particular classes of polynomial nonlinearities g, admissible time-dependent forcing terms f, and nonlinear dissipation terms β for fixed $s \in \mathbb{R}$. A number of authors have considered equations of the type (4.1) with various assumptions on f, g, and β. Our results generalize investigations of autonomous [2]-[4], [15], [22], [16], [32] and time-periodic [15] forcing with linear dissipation ($\beta(u_t) = \alpha u_t$, $\alpha \in \mathbb{R}$).

We shall assume that Ω is a connected bounded open subset of $\mathbb{R}^n (n \geqq 3)$ with a smooth (at least C^2) boundary $\partial\Omega$. We consider processes W related to solutions u of (4.1) as in (2.3) and assume Dirichlet boundary conditions

$$u(x, t) = 0 \quad \text{on}\, \partial\Omega \times [s, \infty) \tag{4.2}$$

with initial conditions

$$u(x, s) = u_s(x), \quad u_t(x, s) = \tilde{u}_s(x), \quad x \in \Omega. \tag{4.3}$$

Different linear operators in (4.1) and either Neumann or periodic boundary conditions can also be considered, but we will not discuss them.

For equations of the type (4.1)-(4.3), Haraux [18], [20] proved existence and uniqueness of global solutions and existence of bounded absorbing sets in $H_0^1(\Omega) \times L^2(\Omega)$ for certain classes of nonlinearities g and dissipation terms β. However, it is clear from our discussion in § 3 that more general results are required in order to establish the dimensionality estimates in Theorems 3.2-3.4. In particular, it is necessary that the system be asymptotically compact in the sense of (3.7) or some facsimile thereof. We will establish results of this type in the present section for a class of nonlinear dissipations β.

In [19], Haraux announced asymptotic compactness results for equations of the type (4.1)-(4.3) with almost periodic forcing and weak nonlinear dissipations β whose derivatives are bounded from both above and below. We note that Haraux's result involves a concept of "uniform asymptotic compactness" that is different from our asymptotic compactness conditions (2.14) and (3.7). We will verify that (3.7) and the other hypotheses of Theorems 3.2-3.4 are satisfied for a large class of equations of the type (4.1)-(4.3). We also require the same weakness condition on β' that Haraux uses. It is recognized that this condition is excessively restrictive (see a similar remark in [19]), but this defect is common to all studies of attractors for (4.1) at the present time. We also extend our results to systems of equations analogous to (4.1)-(4.3).

We make the following assumptions concerning the nonlinearities g and β:

(4.4) g is a C^1 mapping from $V_1 \equiv H_0^1(\Omega)$ into $H \equiv L^2(\Omega)$, Fréchet differentiable with differential g', which satisfies

$$|g'(u)| \leqq C_3(1 + |u|^r) \quad \text{a.e. on } \mathbb{R} \tag{4.5}$$

with a constant $C_3 > 0$ and $r = 2$ if $n = 3$, $r = 0$ if $n \geqq 4$. Let $G(r)$ denote the following primitive of g:

$$G(r) \equiv \int_{-\infty}^{r} g(s)\, ds \quad \text{for all } r \in \mathbb{R}. \tag{4.6}$$

Then we also require that

(4.7) for all $s \in \mathbb{R}$, $G(s) \geqq \left(\frac{-\lambda_1}{2} + \omega\right)s^2 - C_4$ for $\omega > 0$, $C_4 \geqq 0$, and for all $s \in \mathbb{R}$;

(4.8) $sg(s) - G(s) \geqq \left(\frac{-\lambda_2}{2} + \delta\right)s^2 - C_5$ for $\delta > 0$, $C_5 \geqq 0$, where λ_1 denotes the smallest eigenvalue of $-\Delta$.

(4.9) β is an odd C^1 mapping from $\mathbb{R}$ to $\mathbb{R}$ which satisfies the following conditions: there exists $\alpha > 0$ and $C_1 \geqq 0$

such that

$$\beta(v)v \geqq \alpha|v|^2 - C_1 \quad \text{for all } v \in \mathbb{R}, \tag{4.10}$$

$$|\beta(v)| \leqq c(1+|v|^p) \quad \text{for all } v \in \mathbb{R} \tag{4.11}$$

with $0 \leqq p \leqq (n+2)/(n-2)$, $n \geqq 3$. Then we have the following.

THEOREM 4.1 ([20], [18]). *Assume that* (4.4), (4.5), *and* (4.7)-(4.11) *hold. Let f, u_s, $\tilde{u}_s$ be given such that (for fixed $s \in \mathbb{R}$) $f \in C_b(\mathbb{R}, H)$, $u_s \in V_1$, $\tilde{u}_s \in H$. Then the problem* (4.1)-(4.3) *has a unique solution u that satisfies $\{u, u_t\} \in C_b([s, \infty), V_1 \times H = B)$. In addition, for all $t \geqq 0$, the mapping $\{u_s, \tilde{u}_s\} \to \{u(t+s), u_t(t+s)\}$ is a homeomorphism from B onto itself. Furthermore, there exists a closed ball in B that is absorbing for* (4.1)-(4.3). *Moreover, if f is admissible, then the above results are valid if f is replaced by any $h \in H(f)$.*

The homeomorphic property of the solution-maps in Theorem 4.1 (not stated by Haraux) is associated with the properties of (4.1) under time-reversal and the fact that β is assumed to be an odd mapping. The proofs of Theorem 4.1 by Haraux use the weaker formulation of (4.5) that $0 \leqq r < \infty$ if $n = 1$ or 2 and $0 \leqq r \leqq 2$ if $n = 3$; but for the proof of Theorem 4.2 (see below) and subsequent results, we require the stronger assumption stated in (4.5).

For the proofs of the results to follow, it will be convenient to follow the customary procedure in the treatment of hyperbolic equations and write (4.1) as a system of first-order equations. Also, in order that exponential decay properties as $t - +\infty$ can be proved for solutions of the linearized equations, we decompose the dissipation term $\beta(u_t)$ into a linear part and a remainder:

$$\beta(u_t) = \gamma u_t + \tilde{\beta}(u_t), \tag{4.12}$$

and we use the renorming technique introduced by Haraux [17] and later generalized in [15] (for a short discussion see [26]). Then we have the following:

$$\psi_t + \Lambda_\varepsilon \psi + \Gamma(\psi) + D(\psi) = F, \tag{4.13}$$

where $\psi = (u, v)$, $v = u_t + \varepsilon u$, $\Gamma(\psi) = (0, g(u))$, $D(\psi) = (0, \tilde{\beta}(u_t))$, $F = (0, f)$, $\Lambda_\varepsilon = \left(\begin{smallmatrix} & \varepsilon & -1 \\ \varepsilon(\varepsilon-\gamma)-\Delta & & \gamma-\varepsilon \end{smallmatrix}\right)$, $0 < \varepsilon \leqq \varepsilon_0$ with $\varepsilon_0 = \min(\gamma/4, \lambda_1/2\gamma)$, and the initial data are expressed in the form

$$\psi(s) \equiv \psi_s = (u_s, v_s), \qquad v_s = \tilde{u}_s + \varepsilon u_s. \tag{4.14}$$

We will prove that the system (4.13), (4.14) satisfies the asymptotic compactness condition (3.7). However, we first need a result analogous to Theorem 4.1 pertaining to the domain of $-\Delta$.

THEOREM 4.2. *Assume that* (4.4), (4.5), (4.7)-(4.10) *hold and, in place of* (4.11), *assume that $\beta(v)$ has a decomposition* (4.12) *with $\tilde{\beta}$ a C^1 mapping with a Fréchet differential $\tilde{\beta}'$ such that, for some constant $C_6 > 0$,*

$$|\tilde{\beta}'(v)| \leqq C_6, \qquad v \in \mathbb{R}. \tag{4.15}$$

Let f, u_s, and $\tilde{u}_s$ be given such that (for fixed $s \in \mathbb{R}$)

$$f, f_t \in C_b(\mathbb{R}, H), \quad u_s \in H^2(\Omega) \cap H_0^1(\Omega) \equiv V_2, \quad \tilde{u}_s \in V_1. \tag{4.16}$$

Then, if

$$C_6 < \min\left(\frac{\varepsilon}{4}+\frac{\gamma}{2}, \frac{1}{8\varepsilon}\left[-1+\sqrt{1+16\varepsilon\left(\frac{\varepsilon}{4}+\frac{\gamma}{2}\right)}\right]\right), \tag{4.17}$$

the solutions of (4.1)-(4.3) *obtained in Theorem* 4.1 *satisfy* $\{u, u_t\} \in C_b([s, \infty), V_2 \times V_1)$. *Furthermore, there exists a closed ball in* $V_2 \times V_1$ *that is absorbing for* (4.13), (4.14). *Moreover, if f is admissible, the above results are also valid if f is replaced by any* $h \in H(f)$.

Proof. Except for the final assertion concerning the extension of the results to any $h \in H(f)$ when f is admissible, the proof is analogous to the proof of Theorem 2.2 in [15]; therefore, we will not give the details but will just note that the proof involves a liberal use of Young and Poincaré inequalities, combined with Gronwall estimates and the continuity of the imbedding

$$V_2 \times V_1 \hookrightarrow V_1 \times L^2. \tag{4.18}$$

The condition (4.17) results from a requirement that the positive parameters ε that occur in Young inequalities of the form

$$ab \leqq \frac{\varepsilon}{2} a^2 + (2\varepsilon)^{-1} b^2, \qquad a, b > 0$$

be chosen in such a way that we may conclude that $(u, u_t) \in C_b([s, \infty), V_2 \times V_1)$ from the appropriate Gronwall estimate. Combining this procedure with the result that $(u, u_t) \in C_b([s, \infty), B)$, which follows from Theorem 4.1, we infer the existence of a bounded absorbing set in $V_2 \times V_1$.

Now we have the following lemma.

LEMMA 4.1. *Assume the hypotheses of Theorem* 4.2 *and in addition that* $f \in C_b(\mathbb{R}, V_1)$. *Then, for all $t \in \mathbb{R}^+$ and every fixed $s \in \mathbb{R}$, the continuous mapping $U_s(t): B \to B$ defined by*

$$U_s(t) = S_s(t) - \exp(-\Lambda_\varepsilon t) \tag{4.19}$$

is uniformly compact; i.e., it satisfies the condition that, for all bounded sets $\tilde{B} \subset B$ and for all $t \in \mathbb{R}^+$, the union $\bigcup_{\tau \geqq t} U_s(\tau)\,\tilde{B}$ is contained in a compact subset of B. Moreover, if f is admissible, this result is true if f is replaced by any $h \in H(f)$.

Proof. From (4.13), (4.14) we have

$$\psi(s+t) = \Sigma_\varepsilon(t)\phi + \int_0^t \Sigma_\varepsilon(t-\sigma)(F(s+\sigma) - \Gamma(\psi(s+\sigma)) - D(\psi(s+\sigma)))\,d\sigma \tag{4.20}$$

with $\psi(s) = \phi$, where $\Sigma_\varepsilon(t) = \exp(-\Lambda_\varepsilon t)$ is the group associated with the corresponding linear equation so that, upon comparison with (4.19),

$$U_s(t+s)\phi = \int_0^t \Sigma_\varepsilon(t-\sigma)(F(s+\sigma) - \Gamma(\psi(s+\sigma)) - D(\psi(s+\sigma)))\,d\sigma. \tag{4.21}$$

We require the exponential decay of the group $\Sigma_\varepsilon(t)$ in the space $V_2 \times V_1$: $\|\Sigma_\varepsilon(t)\|_{L(V_2 \times V_1)} \leqq \exp(-(\varepsilon/2)t)$, $t \geqq 0$. This follows from the combination of an energy estimate in $V_2 \times V_1$ and a Gronwall argument as in [15]. Then we find from (4.21), (4.5), (4.15), Theorem 4.2, and the additional hypothesis that $f \in C_b(\mathbb{R}, V_1)$:

$$\begin{aligned} \|U_s(t+s)\phi\|_{V_2 \times V_1} \leqq \frac{2}{\varepsilon}\left(1 - \exp\left(-\frac{\varepsilon}{2}t\right)\right) & \\ \times \sup_{\sigma \varepsilon [0, t]} (\|f(s+\sigma)\|_{V_1} + C(R)(1 + \|u(s+\sigma)\|_{H^2(\Omega)})^\sigma & \\ + C_6 \|u_t(s+\sigma)\|_{V_1}). & \end{aligned}$$

It follows from Theorem 4.2 that, for all bounded sets $\tilde{B} \subset V_1 \times L^2(\Omega)$, $\bigcup_{\tau \geqq t} U_s(\tau)\tilde{B}$ is contained in a bounded set Y in $V_2 \times V_1$. Then, from the compactness of the imbedding (4.18), Y is compact in $V_1 \times L^2(\Omega)$, and the proof is complete.

PROPOSITION 4.1. *Assume the hypotheses of Lemma* 4.1. *Then, for given* $s \in \mathbb{R}$,

(i) $S_s(t)$ *possesses a bounded absorbing set* $B_0 \subset B$;

(ii) *For all bounded sets* $\tilde{B} \subset B$, *there exists a compact set* $G \subset B$ *such that*

$$\lim_{t \to +\infty} \sup_{\phi \in \tilde{B}} d_B(S_s(t)\phi, G) = 0, \tag{4.22}$$

where d_B *is defined as in* (2.15).

In addition, if f is admissible, (i) *and* (ii) *remain valid when f is replaced by any* $h \in H(f)$.

Proof. The proof is analogous to that of a corresponding result in [15]. (i) follows from Theorem 4.1. To prove (ii), let $\tilde{B}$ be a bounded set in B. By Lemma 4.1, there exists a compact set $G \subset B$ such that $\bigcup_{t \geqq 0} U_s(t+s)\tilde{B} \subset G$. We then establish (4.22) by using the exponential decrease with t of the linear group $\Sigma_\varepsilon(t)$.

We are now prepared to prove the continuity of the translation semigroup $\{\sigma(t), t \geqq 0\}$ relative to the system (4.1)-(4.3).

PROPOSITION 4.2. *Assume the hypotheses of Theorem* 4.1 *and, in addition, that* $f_t \in C_b(\mathbb{R}, H)$ *and that f is time-dependent admissible from* $\mathbb{R}$ *to* H. *Then the translation semigroup* $\{\sigma(t), t \geqq 0\}$ *is continuous from* $H(W)$ *to itself, where W is the almost periodic process associated as in* (2.3) *with the unique solution* $\psi(u, v)$ *of* (4.13), (4.14) *obtained in Theorem* 4.1.

Proof. Since $\psi(t+s)$ is uniformly bounded in B by Theorem 4.1, we may assume that it is contained in a B-ball:

$$\|\psi(t+s)\|_B \leqq K_1(R) \quad \text{if } \|\psi(s)\|_B \leqq R \tag{4.23}$$

for some $R > 0$. From the mean value theorem for Banach spaces, (4.5), and the well-known Sobolev imbedding theorem

$$H_0^1(\Omega) \hookrightarrow L^q(\Omega) \quad \text{if } 2 \leqq q \leqq \frac{2n}{n-2} \quad \text{with } n \geqq 3,$$

we find

$$\|\Gamma(\psi) - \Gamma(\tilde{\psi})\|_B \leqq C(R)\|\psi - \tilde{\psi}\|_B \tag{4.24}$$

with a positive constant $C(R)$, where $\tilde{\psi}$ is another solution of (4.13), (4.14) that also satisfies (4.23). Similarly, using (4.15) and the mean value theorem again, we obtain

$$\|D(\psi) - D(\tilde{\psi})\|_B \leqq 2C_6\|\psi - \tilde{\psi}\|_B. \tag{4.25}$$

Also, there exists a positive constant c such that

$$\|F(t+s+\omega) - F(t+s)\|_B \leqq c\omega$$

uniformly in s and t since $f_t \in C_b(\mathbb{R}, H)$ by hypothesis.

Since $\{\sigma(t), t \geqq 0\}$ is a semigroup, it is sufficient to prove continuity at $t = 0$. The solution ψ of (4.13), (4.14) also satisfies (4.20), and we have the following estimate for the linear group $\Sigma_\varepsilon(t)$ [15, p. 278]:

$$\|\Sigma_\varepsilon(t)\|_{L(B)} \leqq \exp\left(-\frac{\varepsilon}{2}t\right), \quad t \geqq 0, \quad 0 < \varepsilon \leqq \varepsilon_0. \tag{4.26}$$

Then, using the Schwarz inequality, we obtain (with $\omega > 0$)

$$\begin{aligned}
&\|\sigma(\omega)W(t,s)\phi - W(t,s)\phi\|_B \leqq \|\Sigma_\varepsilon(t)(\Sigma_\varepsilon(\omega)-1)\phi\|_B \\
&\quad + \int_0^t \|\Sigma_\varepsilon(t-\sigma)(\Sigma_\varepsilon(\omega)-1)(F(s+\sigma)-\Gamma(\psi(s+\sigma))-D(\psi(s+\sigma)))\|_B\, d\sigma \qquad (4.27)\\
&\quad + \int_t^{t+\omega} \|\Sigma_\varepsilon(t+\omega+\sigma)(F(s+\sigma)-\Gamma(\psi(s+\sigma))-D(\psi(s+\sigma)))\|_B\, d\sigma.
\end{aligned}$$

To estimate the first norm, we use the fact that $\Sigma_\varepsilon(t)$ is uniformly bounded in t by (4.26), so that we can approximate ϕ in the B-norm by a sequence $\{\phi_n\}$ from the domain of Λ_ε, $D(\Lambda_\varepsilon)$, which is a dense subset of B. Given $\delta > 0$, we choose $n_0 \in \mathbb{N}$ such that

$$\|\phi + \phi_n\|_B < \frac{\varepsilon}{2(\varepsilon+6)}\delta \quad \text{when } n \geqq n_0. \qquad (4.28)$$

We have the estimate

$$\|(\Sigma_\varepsilon(t)-1)\chi\|_B \leqq t\|\Lambda_\varepsilon \chi\|_B, \quad t \geqq 0, \quad \chi \in D(\Lambda_\varepsilon),$$

which is a general result for analytic semigroups (cf. [16, p. 71]). Using this result and (4.26), we obtain

$$\|\Sigma_\varepsilon(t)(\Sigma_\varepsilon(\omega)-1)\phi\|_B < \frac{\varepsilon\delta}{\varepsilon+6} + \omega\|\Lambda_\varepsilon\phi_n\|_B (n \geqq n_0, \phi_n \in D(\Lambda_\varepsilon)). \qquad (4.29)$$

Similarly, in addition to the uniform boundedness of $\Sigma_\varepsilon(t)$, we use the fact that $\Gamma(\psi(t+s))$, $D(\psi(t+s))$, and $F(t+s)$ also have this property, which follows from (4.23) and the respective hypotheses (4.5), (4.11), and $f \in C_b(\mathbb{R}, H)$. It immediately follows that the second integral in (4.27) is bounded by $C'\omega$ with C' a positive constant. Finally, to estimate the remaining integral in (4.27), we use the boundedness properties noted above to approximate $F(s+\sigma)$, $\Gamma(\psi(s+\sigma))$, and $D(\psi(s+\sigma))$ by respective sequences $\{F_n(s+\sigma)\}$, $\{\Gamma_n(s+\sigma)\}$, and $\{D_n(s+\sigma)\} \subset D(\Lambda_\varepsilon)$ to obtain a bound analogous to (4.29). By choosing n_0 large enough, we can use the same value of δ as in (4.28) for these approximations. Thus, by collecting the above results we have, with $n \geqq n_0$,

$$\begin{aligned}
&\|\sigma(\omega)W(t,s)\phi - W(t,s)\phi\|_B \\
&\quad < \delta + \omega(C' + \|\Lambda_\varepsilon\phi_n\|_B) + \omega\int_0^t \exp\left(-\frac{\varepsilon}{2}(t-\sigma)\right) \qquad (4.30)\\
&\qquad \cdot \{\|\Lambda_\varepsilon F_n(s+\sigma)\|_B + \|\Lambda_\varepsilon\Gamma_n(s+\sigma)\|_B + \|\Lambda_\varepsilon D_n(s+\sigma)\|_B\}\, d\sigma,
\end{aligned}$$

from which it follows that $\sigma(\omega)$ is continuous at $\omega = 0$ on the distinguished process W.

In order to prove that W is almost periodic, we establish an estimate of the type (3.1). Thus, consider two solutions ψ_1, ψ_2 of (4.13), (4.14) with the same initial datum ϕ, but corresponding to two distinct admissible forcing terms f_1, f_2. Then, from (4.13), (4.14), the B-positivity of $\Lambda_\varepsilon((\chi, \Lambda_\varepsilon\chi) \geqq 0, \chi \in B)$, the Lipschitz estimates (4.24), (4.25), Young's inequality, and a Gronwall estimate, we obtain the inequality

$$\begin{aligned}
&\|\psi_1(t+s) - \psi_2(t+s)\|_B \\
&\quad \leqq c^{-1}\exp\left(\frac{1}{2}\tilde{K}t\right) \sup_{\tau\in[0,t]} \|F_1(\tau+s) - F_2(\tau+s)\|_B,
\end{aligned}$$

with $\tilde{K} = c + 2C(R) + 4c\sigma$. It follows that W is almost periodic since f is admissible by hypothesis.

Take $V \in H(W)$. Then, for any sequence $\{\tau_n\} \subset \mathbb{R}$, there exists a subsequence $\{\tau_{n_m}\} \subset \{\tau_n\}$ such that $W_{\tau_{n_m}}(t, s)\phi \to_{m\to\infty} V(t, s)\phi$ uniformly in $s \in \mathbb{R}$ for all $t \in \mathbb{R}^+$ and for all $\phi \in B$. In order to show that $\sigma(\omega)$ is continuous on $V\phi$, consider the following estimate:

$$\begin{aligned}&\|\sigma(\omega)V(t, s)\phi - V(t, s)\phi\|_B\\ (4.31)\ &\leqq \|\sigma(\omega)W(t, s+\tau_{n_m})\phi - W(t, s+\tau_{n_m})\phi\|_B + \|V(t, s+\omega)\phi - W_{\tau_{n_m}}(t, s+\omega)\phi\|_B\\ &\quad + \|V(t, s)\phi - W_{\tau_{n_m}}(t, s)\phi\|_B.\end{aligned}$$

The first norm can be estimated in the same way as (4.30). Given $\eta > 0$, choose m_0 such that each of the last two norms in (4.31) is less than $\eta/2$ when $m \geqq m_0$. This can be done uniformly in $s \in \mathbb{R}$ for all $t \in \mathbb{R}^+$ and all $\phi \in H$ (see (2.6)). The assertion of the proposition follows.

This proposition, together with the result in Theorem 4.1, establishes the hypotheses of Proposition 3.1 for (4.1)-(4.3).

The first variational equation corresponding to (4.13), (4.14) has the form

$$\Phi_t(t+s) + \Lambda_\varepsilon\Phi(t+s) + \Gamma'(\psi)\Phi(t+s) + D'(\psi)\Phi(t+s) = 0, \tag{4.32}$$

$$\Phi(s) = \rho \in B, \tag{4.33}$$

where $\psi = \psi(t+s)$ is a solution of (4.13), (4.14). We see that $\Phi(t+s)$ is independent of the forcing term in (4.13) as required by one of our hypotheses in § 3.

We proceed to the proof of existence of Fréchet derivatives of the maps (2.13).

THEOREM 4.3. *Let W be a process associated in the usual way* (2.3) *with a solution ψ of* (4.13), (4.14), *assuming our hypotheses* (4.4), (4.5), (4.7)-(4.10), $f \in C_b(\mathbb{R}, .H)$, $u_s \in V_1$, $\tilde{u}_s \in H$ *and, in addition, that a decomposition* (4.12) *holds with $\tilde{\beta}$ satisfying* (4.15). *Furthermore, assume that there exists $\nu \in (0, 1]$ and, for every $R > 0$, also a positive constant $C_2 = C_2(R)$ such that*

$$\|g'(\rho) - g'(\eta)\|_{L(V_1,H)} \leqq C_2\|\rho - \eta\|_{V_1}^{\nu} \tag{4.34}$$

for all $\rho, \eta \in V_1$ with $\|\rho\|_{V_1} \leqq R$, $\|\eta\|_{V_1} \leqq R$ and, for the same value of ν, there exists $\tilde{c}_2 = \tilde{c}_2(R)$ such that

$$\|\tilde{\beta}'(u_t) - \tilde{\beta}'(\tilde{u}_t)\|_{L(H)} \leqq \tilde{c}_2\|u_t - \tilde{u}_t\|_H^{\nu} \tag{4.35}$$

for all $u_t, \tilde{u}_t \in H$ with $\|u_t\|_H \leqq R$, $\|\tilde{u}_t\|_H \leqq R$. Then Fréchet derivatives L of the maps (2.13) *exist and are defined in terms of appropriate solutions of* (4.32), (4.33) *as in* (3.14). *Moreover, if f is admissible, this result is also true when f is replaced by any $h \in H(f)$.*

The proof follows the lines of that given in Appendix B of [15] and makes use of the mean value theorem for Banach spaces, Young's inequality, and a Gronwall argument.

COROLLARY. *For all $t > 0$ and all $s \in \mathbb{R}$, the mapping $S_s(t)$ is of class $C^{1,\nu}$, where ν is the same as in* (4.34), (4.35).

We have now verified that all the hypotheses in § 3 are valid for (4.1)-(4.3). In particular, we note that condition (7) of § 2 is satisfied for this system because the maps $S_s(t)$ are Lipschitz homeomorphisms with Lipschitz-continuous inverses on the entire Hilbert space B.

We now obtain more information concerning the Hausdorff and fractal dimensions of sets of type (2.8) by deriving explicit estimates for the quantities (3.21). We note that (4.32) is of the form (3.20) with $F'(\psi) = -\Lambda_\varepsilon - \Gamma'(\psi) - D'(\psi)$. At a given time $p+s$, let $\{X_j(p+s) = (\rho_j(p+s), z_j(p+s)) \in V_1, j = 1 \cdots, m\}$ denote an orthonormal basis of B spanning an m-dimensional subspace $Q_m(p+s)B$. Then, following an argument in [32, pp. 360-364], we obtain

$$(F'(\psi)\chi_j, \chi_j)_B \leqq -\frac{\varepsilon}{4}(\|\rho_j\|_{V_1}^2 + \|z_j\|_H^2) + \frac{2(\gamma^2+C_6^2)}{\varepsilon}\|\rho_j\|_{V_\eta}^2, \tag{4.36}$$

where we have assumed that the initial data (4.14) belongs to a set of type (2.8), which, according to Theorem 4.2, is a bounded subset of $B_1 = V_2 \times V_1 = D(-\Delta) \times H_0^1(\Omega)$. Then

$$\psi(t+s) = \{u(t+s), u_t(t+s) + \varepsilon u(t+s)\}$$

belongs to a bounded subset of B_1, and $u(t+s)$ belongs to a bounded subset M of $D(-\Delta)$ for all $t \in \mathbb{R}^+$. The quantity η in (4.36) arises from the following additional assumption concerning g':

(4.37) g' is a bounded mapping from $D(-\Delta)$ to $L(V_\eta, H)$ for some $\eta \in [0, 1)$.

It follows [15] that there exists $\eta\varepsilon[0, 1)$ such that g' maps M into a bounded subset of $L(V_\eta, H)$. The constant γ in (4.36) is defined by $\gamma = \sup_{w\in D(-\Delta)} \|g'(w)\|_{L(V_\eta,H)}$. The final form of (4.36) is obtained by two applications of Young's inequality. Then, using the fact that the $\{\chi_j\}(j \in \mathbb{N})$ are orthogonal in B and a lemma in [32], we obtain the following estimates for the quantities in (3.21):

$${}_sq_m(t) \leqq \frac{-m\varepsilon}{4} + \frac{2(\gamma^2+C_6^2)}{\varepsilon}\sum_{j=1}^m \lambda_j^{\eta-1}, \tag{4.38}$$

$${}_sq_m \leqq \frac{-m\varepsilon}{4} + \frac{2(\gamma^2+C_6^2)}{\varepsilon}\sum_{j=1}^m \lambda_j^{\eta-1}, \tag{4.39}$$

where the $\{\lambda_j\}$ are eigenvalues of $-\Delta$. It follows from (3.11), (3.18), (3.19), and the above results that we obtain the following estimates for the uniform Lyapunov exponents.

THEOREM 4.4. *Assume the hypotheses of Theorem* 4.3 *as well as the additional hypothesis* (4.37). *Then*

(i) *The uniform Lyapunov exponents $\mu_{j,s}$ associated with sets of type* (2.8) *are majorized according to*

$$\mu_{1,s} + \cdots + \mu_{j,s} \leqq \frac{-m\varepsilon}{4} + \frac{2(\gamma^2+C_6^2)}{\varepsilon}\sum_{i=1}^j \lambda_i^{\eta-1}, \qquad j \in \mathbb{N}; \tag{4.40}$$

(ii) *The m-dimensional volume element is exponentially decreasing in the Hilbert space B*;

(iii) *We have the upper bounds for the Hausdorff and fractal dimensions of E_s,$d_H(E_s) \leqq m_0$, $d_F(E_s) \leqq \frac{4}{3}m_0$, where m_0 is chosen in such a way that*

$$\sum_{i=1}^{m_0} \lambda_i^{\eta-1} \leqq \frac{\varepsilon^2 m_0}{16(\gamma^2+C_6^2)}. \tag{4.41}$$

Remark 4.1. (a) Note that the bounds (4.38)-(4.40) are uniform in s. (b) There are some differences in the details of the proof of Theorem 4.4 compared with the proof of the corresponding autonomous result that we now point out. (b1) As a consequence of the fact that we include effects of weak nonlinear dissipation, the constant C_6 appears in the bounds (4.38)-(4.41). This has the consequence that the allowed values of m_0, defined to be those for which (4.41) is valid, are different from those for the corresponding linearly damped equation. (b2) The validity of (ii) follows from the estimate

$$\operatorname{Tr} F'(\psi(p+s)) \circ Q_m(p+s) \leqq \frac{-m_0\varepsilon}{4} + \frac{2(\gamma^2+C_6^2)}{\varepsilon} \sum_{j=1}^{m} \lambda_j^{\eta-1},$$

the choice of values of m_0 being those allowed by (4.41).

To conclude our discussion of the system (4.1)-(4.3), we note that the considerations of the present section can be extended to systems of equations of type (4.1). Thus, in place of that equation, we consider the system of equations

$$u_{tt} + \beta(u_t) - Au + g(u) = f, \tag{4.42}$$

obtained from (4.1) by letting u be an l-dimensional vector $u = (u_1, \cdots, u_l)$ and by replacing $-\Delta$ by $-A\Delta$ where A is a symmetric matrix, although we could also consider linear unbounded elliptic selfadjoint operators L with smooth coefficients whose inverses are compact on the Hilbert space $H = (L^2(\Omega))^l$, where Ω is a bounded domain on which we impose conditions analogous to those stated at the beginning of this section in the case of (4.1). We impose Dirichlet boundary conditions (4.2) on each component of u, although Neumann or periodic boundary conditions could also be considered. The nonlinearity $g(u)$ is of "potential type"; i.e., there exists a function $G(u_1, \cdots, u_l)$ (a generalization of the primitive (4.6)) such that $g_i(u) = (\partial/\partial u_i)G(u_1, \cdots, u_l)$ $(i = 1, \cdots, l)$. We assume that the dissipative term β has a representation analogous to (4.12) with the constant γ replaced by a positive-definite matrix. The hypotheses (4.4), (4.5), (4.7)-(4.12), (4.15), (4.16), (4.34), (4.35), and (4.37) are replaced by their obvious vector analogues. Under these conditions, we can verify all the hypotheses required for the proofs of the dimension estimates in Theorems 3.2-3.4, and analogous results to those in Theorem 4.4 can be obtained for the system (4.42). These results generalize the work of Babin and Vishik [3], [4], who proved existence and other properties of global attractors for autonomous equatons of the above type with linear dissipation.

5. Reaction-diffusion equations. In this section we consider the nonlinear PDE,

$$\psi_t - \Delta\psi + g(\psi) = f(x, t) \quad \text{in } \Omega \times [s, \infty), \tag{5.1}$$

$$\psi(x, s) = \phi(x), \quad x \in \Omega, \quad \text{some } s \in \mathbb{R}, \tag{5.2}$$

with Dirichlet boundary conditions

$$\psi(x, t) = 0, \quad x \in \partial\Omega, \quad t \geqq s, \tag{5.3}$$

where Ω denotes a connected open bounded subset of $\mathbb{R}^n$ with a smooth (at least C^2) boundary $\partial\Omega$. We consider the Hilbert space $H = L^2(\Omega)$ and set $V = H_0^1(\Omega)$. Global attractors and estimates of their dimensions for the system (5.1)-(5.3) with suitable restrictions on g have been considered previously without a time-dependent forcing term f by Babin and Vishik [2], [4], Marion [25], and Témam [32].

We assume the following conditions.

There exists a real number $p \geqq 2$ and positive constants c_1, c_2, c_3 such that

$$c_1 s^{2p} - c_3 \leqq s g(s) \leqq c_2 s^{2p} + c_3 \tag{5.4}$$

for all $s \in \mathbb{R}$. There exist positive constants C_4, C_5 such that

$$g'(s) \geqq -C_4, \tag{5.5}$$

$$g'(s) \leqq C_5(1 + s^{2p-2}) \tag{5.6}$$

for all $s \in \mathbb{R}$.

There exists a positive constant c_6 such that

$$|g(s)| \leqq c_6(1+|s|^{2p-1}) \tag{5.7}$$

for all $s \in \mathbb{R}$.

The forcing term satisfies

$$f \in C_b(\mathbb{R}, H). \tag{5.8}$$

The following result gives basic local and global existence results for (5.1)-(5.3).

THEOREM 5.1. *Assume that* (5.4)-(5.8) *hold. Then, for each* $\phi \in H$, *there exists a unique solution* ψ *of* (5.1)-(5.3) *such that* $\psi \in L^2([s, T]; V) \cap L^{2p}([s, T]; L^{2p}(\Omega))$ *for all* $T > s$ *and* $\psi \in C_b([s, \infty); H)$ *for each* $s \in \mathbb{R}$. *The mapping* $\phi \to \psi(t+s)$ *is continuous on* H. *Furthermore, if* $\phi \in V$, *then*

$$\psi \in C_b([s, T]; V) \cap L^2([s, T]; H^2(\Omega)) \quad \textit{for all } T > s.$$

Finally, if f is admissible, the above results remain valid if f is replaced by any $h \in H(f)$.

The proof relies on classical arguments [23] and is an extension of the arguments of Marion [25] and Témam [32] to cases of time-dependent forcing. We omit the details.

Remark 5.1. Hypotheses (5.4)-(5.7) are satisfied for the nonlinearities $g(s) = |s|^{p-2}s (p \geqq 4, s \in \mathbb{R})$ and $g(s) = \alpha s^3 - \beta s (\alpha, \beta > 0)$.

We now prove the existence of bounded absorbing sets for the system (5.1)-(5.3). Since the proof only differs from corresponding considerations in [32] by the inclusion of time-dependent forcing terms, many details will be omitted.

THEOREM 5.2. *Assume the hypotheses of Theorem* 5.1. *Then*

(a) *There exists a closed ball in H that is absorbing for* (5.1)-(5.3);

(b) *If, in addition to the above hypotheses,* $f \in C_b(\mathbb{R}, V)$, *then there exists a closed ball in V that is absorbing for* (5.1)-(5.3);

(c) *If we construct a process W as in* (2.3) *in terms of the unique solution of* (5.1)-(5.3) *discussed in Theorem* 5.1, *the maps* $S_s(t)\phi = W(t, s; f)\phi$, $\phi \in H$ *are uniformly compact for* $t \geqq t_0$ *with* $t_0 > 0$ *sufficiently large. Moreover, if f is admissible, the above results are valid with f replaced by any* $h \in H(f)$;

(d) *Asymptotic compactness holds; i.e., for all bounded sets* $\tilde{B} \subset H = L^2(\Omega)$, *there exists a compact set* $G \subset H$ *such that*

$$\lim_{t \to +\infty} \sup_{\phi \in \tilde{B}} d_H(S_s(t)\phi, G) = 0$$

with d_H *defined as in* (2.15). *If f is admissible, this result is true for all* $h \in H(f)$.

Proof (sketch). (a) By multiplying (5.1) by ψ and using (5.4), the Poincaré inequality for $H_0^1(\Omega)$, and Young's inequality, we obtain an energy inequality in H from which we obtain, by the standard Gronwall lemma,

$$\limsup_{t \to +\infty} \|\psi(t+s)\|_H \leqq \sqrt{\frac{2}{\lambda_1}} \sqrt{C_3|\Omega| + \frac{1}{2\lambda_1} \|f\|_H^2} \equiv R_0, \tag{5.9}$$

where λ_1 denotes the smallest eigenvalue of $-\Delta$ on Ω and $|\Omega|$ is the volume of Ω. Thus, any ball in $H = L^2(\Omega)$ centered at zero with radius $R_0' > R_0$ is an absorbing set for (5.1)-(5.3).

(b) Analogously, to obtain an absorbing ball in $H_0^1(\Omega)$, we obtain a corresponding energy inequality by multiplying (5.1) by $-\Delta\psi$ and using (5.5) and (5.3) in conjunction with Green's theorem, Young's inequality, and the Poincaré inequality [32] $\|\nabla\psi\|_H \leqq C_7(\Omega)\|\Delta\psi\|_H$ for some positive constant C_7 depending on Ω, to obtain for any $\varepsilon_1, \varepsilon_2 > 0$:

$$\frac{d}{dt} \|\nabla\psi\|_H^2 + \left(\frac{2}{C_7^2} - 2C_4 - \varepsilon_1 - \varepsilon_2\right) \|\nabla\psi\|_H^2 \leqq \varepsilon_1^{-1} \|f\|_H^2 + \varepsilon_2^{-1} \|\nabla f\|_H^2. \tag{5.10}$$

First consider the case when $C_7^{-2} > C_4$. Then, setting $\varepsilon_1 = \varepsilon_2 = \frac{1}{2}(C_7^{-2} - C_4)$, a standard Gronwall estimate yields the result

$$\limsup_{t\to+\infty} \|\nabla\psi(t+s)\|_H \leqq \sqrt{2}(C_7^{-2} - C_4)^{-1}(\|f\|_H^2 + \|\nabla f\|_H^2)^{1/2} \equiv R_1,$$

and a similar argument to the discussion following (5.9) yields the result that any ball in $H_0^1(\Omega)$ centered at zero with radius $R_1' > R_1$ is an absorbing set for (5.1)-(5.3) when $C_7^{-2} > C_4$.

For the remaining cases $C_7^{-2} \leqq C_4$, we put $\varepsilon_1 = \varepsilon_2 = C_7^{-2}$ and obtain from (5.10) by a standard Gronwall argument,

$$\begin{aligned} \|\nabla\psi(t+s)\|_H^2 \leqq{}& \exp(2C_4 t)\|\nabla\psi(s)\|_H^2 \\ &+ C_7^2(2C_4)^{-1}(\|f\|_H^2 + \|\nabla f\|_H^2)(\exp(2C_4 t) - 1). \end{aligned} \tag{5.11}$$

Then, from (5.10) and the uniform Gronwall lemma [32, p. 89], for an arbitrary fixed $r > 0$,

$$\|\nabla\psi(t+s+r)\|_H^2 \leqq \left(\frac{\kappa(r)}{r} + C_7^2(\|f\|_H^2 + \|\nabla f\|_H^2)r\right)\exp(2C_4 r) \equiv R_2^2, \tag{5.12}$$

where $\kappa(r) = \frac{1}{2}(1+\lambda_1)\|\psi(s)\|_H^2 + (\lambda_1^{-1} + 2r)(C_3|\Omega| + (2\lambda_1)^{-1}\|f\|_H^2)$. Thus, when $C_7^{-2} \leqq C_4$, (5.12) provides uniform bounds for $\|\nabla\psi(t+s)\|_H^2$ when $t+s \geqq r$ while (5.11) provides uniform bounds when $t+s \leqq r$. Any ball of $V = H_0^1(\Omega)$ centered at zero with radius $R_2' > R_2$ is absorbing for (5.1)-(5.3).

(c) The proof of the uniform compactness of the maps (2.13) now follows as in the proof of the corresponding result for the solution semigroups in autonomous cases [32, p. 86].

(d) The proof of asymptotic compactness is analogous to the corresponding proof for nonlinear wave equations in Propositions 4.1.

This completes the proof of the theorem.

The translation semigroup $\{\sigma(t), t \geqq 0\}$ is continuous relative to the system (5.1)-(5.3).

PROPOSITION 5.1. *Assume the hypotheses of Theorem* 5.1 *and, in addition, that* $f_t \in C_b(\mathbb{R}, H)$ *and that* f *is time-dependent admissible from* $\mathbb{R}$ *to* H. *Then the translation semigroup* $\{\sigma(t),\ t \geqq 0\}$ *is continuous from* $H(W)$ *to itself, where* W *is the process associated as in* (2.3) *with the unique solution* ψ *of* (5.1)-(5.3) *obtained in Theorem* 5.1.

Proof. The proof is analogous to the proof of Proposition 4.2.

This proposition, together with the result in Theorem 5.1, establishes the hypotheses of Propositon 3.1 for the system (5.1)-(5.3).

The first variational equation corresponding to (5.1)-(5.3) has the form

$$\Phi_t(t+s) - \Delta\Phi(t+s) + g'(\psi)\Phi(t+s) = 0, \tag{5.13}$$

$$\Phi(s) = \rho \in H, \tag{5.14}$$

where $\psi = \psi(t+s)$ is a solution of (5.1)-(5.3). We see that $\Phi(t+s)$ is independent of the forcing term in (5.1) as required by our hypothesis in § 3.

We now prove existence of the Fréchet derivatives of the maps (2.13) for the system of equations (5.1)-(5.3), (5.13), (5.14).

THEOREM 5.3. *Let* W *be a process associated in the usual way* (2.3) *with a solution* ψ *of* (5.1)-(5.3) *assuming the hypotheses* (5.4)-(5.8) *as well as the additional condition that there exists* $\nu \in (0, 1)$, *and for* $R > 0$ *there exists* $C_8 = C_8(R) > 0$ *such that*

$$\|g'(\chi) - g'(\tilde\chi)\|_{L(H)} \leqq C_8\|\chi - \tilde\chi\|_H^\nu \tag{5.15}$$

for every $\chi, \tilde\chi \in H$ *such that* $\|\chi\|_H \leqq R$, $\|\tilde\chi\|_H \leqq R$. *Then Fréchet derivatives* L *of the maps*

(2.13) *exist and are defined in terms of appropriate solutions of* (5.13), (5.14) *as in* (3.14). *Moreover, if in addition to satisfying* (5.8) *f is also time-dependent admissible, then these results remain valid if f is replaced by any* $h \in H(f)$.

Proof. The proof is analogous to the proof of Theorem 4.3.

COROLLARY. *For all* $t > 0$ *and all* $s \in \mathbb{R}$, *the mapping* $S_s(t)$ *is of class* $C^{1,\nu}$, *where* ν *is the same as in* (5.15).

With the exception of condition (7) of § 2, we have now verified that all the hypotheses of § 3 are valid for (5.1)-(5.3). The validity of this condition follows from the following result.

THEOREM 5.4. *Assume the hypotheses of Theorem* 5.3, *and let* E_s *denote a set of type* (2.8) *corresponding to a global attractor* ${}_s A_\pi(W)$ *for the distinguished process W. Then, for each* $s \in \mathbb{R}$ *and each* $t \in \mathbb{R}^+$, $S_s(t)$ *is invertible on the range of* $S_s(t)H$, *and the surjective map* $(S_s(t))^{-1}: S_s(t)E_s \to E_s$ *is Lipschitz continuous.*

To prove this result, we need the existence of backward extensions in t of the maps $S_s(t)$. This follows from a generalization to almost periodic processes of some results on periodic processes due to Slemrod [31], which we now discuss.

DEFINITION 5.1. Let V be an almost periodic process on a Banach space B, and consider $\eta \in B$. A function $U(\cdot, \cdot; \eta): \mathbb{R} \times \mathbb{R} \to B$ is said to be an *extension* of the process V from η if

(i) $U(\theta, s; \eta)$ is continuous in θ;

(ii) $U(t+\theta, s; \eta) = Z(t, s+\theta;\ U(\theta, s; \eta))$ for $t \in \mathbb{R}^+$, $\theta, s \in \mathbb{R}$, and some $Z \in H(V)$;

(iii) $U(0, s; \eta) = \eta$.

Then we have the following.

LEMMA 5.1. *Let V be an almost periodic process on B. Assume that the positive orbit* $0^+(s, \chi) = \bigcup_{t \geqq 0} V(t, s; \chi)$ *through* $\chi \in B$ *lies in a compact subset* $A \subseteq B$. *If* η *belongs to the* ω-*limit set* $\omega_s(\chi) = \bigcap_{\delta \geqq 0} \text{Closure}_B \bigcup_{t \geqq \delta} V(t, s; \chi)$ *of the orbit of V originating at* $(\chi, s) \in B \times \mathbb{R}$, *then V possesses an extension* $U(\theta, s; \eta)$ *from* η.

Proof. The proof is similar to that of Slemrod for periodic processes, the essential points of difference being (1) the places in Slemrod's proof where periodicity was used, and (2) the verification of condition (ii) for the extension candidate U. We will, therefore, skip some of the details. There exists a sequence $\{t_n\} \subset \mathbb{R}^+$ such that $t_n \to +\infty$ as $n \to +\infty$ with

$$V(t_n, s; \chi) \to \eta \quad \text{as } n \to +\infty. \tag{5.16}$$

Then choose $a > 0$ and pick $n_0 > 0$ sufficiently large so that $t_n \geqq 2a$ when $n \geqq n_0$. Then, following an argument similar to that of Slemrod, it can be shown that the sequence $\{V(t_n + \tau, s; \chi)\}$ is an equicontinuous family of functions of $\tau \varepsilon [-a, a]$.

From Ascoli's theorem, there exists a subsequence $\{T_m\} \subset \{t_n\}$, such that $V(T_m + \tau, s; \chi)$ converges uniformly to a continuous function of $\tau \in [-a, a]$. Call it $U(\tau, s; \eta)$. That is, we have

$$\|V(T_m + \tau, s; \chi) - U(\tau, s; \eta)\|_B \to 0 \quad \text{as } m \to +\infty \tag{5.17}$$

uniformly in τ for $\tau \in [-a, a]$, and (i) is satisfied.

For $\tau = 0$, $V(T_m, s; \chi) \to \eta$ by (5.16) so that (iii) is satisfied for all $s \in \mathbb{R}$ and it only remains to establish (ii). We have

$$\|V_{T_m}(\theta, s+\tau;\ V(\tau - T_m, s; \chi)) - V_{T_m}(\theta, s+\tau;\ U(\tau, s; \eta))\|_B \to 0$$

by continuity. But

$$V_{T_m}(\theta, s+\tau;\ V(\tau + T_m; s; \chi)) = V(\tau + T_m + \theta, s; \chi) \to U(\tau + \theta, s; \eta),$$

i.e.,

$$U(\tau+\theta, s; \eta) = \lim_{m\to+\infty} V_{T_m}(\theta, s+\tau; U(\tau, s; \eta)) = Z(\theta, s+\tau; U(\tau, s; \eta))$$

for some $Z \in H(V)$. It follows from (5.17) that $U(\tau, s; \eta) \in \omega_s(\chi)$ for all $\tau \in [-a, a]$, and hence for all $\tau \in \mathbb{R}$, and the proof is complete.

Proof of Theorem 5.4. For given $s \in \mathbb{R}$, the existence of $S_s(t)$ for each $t \geqq 0$, and each $h \in H(f)$ follows from Theorem 5.1 and its Lipschitz continuity follows from Theorem 5.3 (see the statement of the corollary to that theorem). Then, using the facts that $E_s \subset \omega_s(\phi)$ (with ϕ the same as in (2.13)) and that $S_s(t)$ possesses backward extensions in t from E_s (which follows from Lemma 5.1), we easily see by a slight generalization of the usual argument [32], [5], [14] that the injectivity of $S_s(t)$ is equivalent to the backward uniqueness property for the system (5.1)-(5.3). In fact, with our hypotheses, the proof of backward uniqueness given by Témam [32] is also valid in the present case. Thus, $(S_s(t))^{-1}$ exists on the range of $S_s(t)H$. By Theorems 5.1 and 5.2 and an argument in the proof of Theorem 3.2, there exists an s-independent global attractor $A_\pi(W)$ for the skew-product semiflow $\{\pi_s(t),\ t \geqq 0\}$ associated with the distinguished process W, and the corresponding set E_s defined by (2.8) is compact. Consider the restriction of $S_s(t)$ to E_s. Since $S_s(t)$ is continuous and E_s is compact, $S_s(t)E_s$ is a compact set. Then, by a result due to Tikonov [4, Lemma 3, p. 98], $(S_s(t))^{-1}$ is continuous as a surjective map $(S_s(t))^{-1}: S_s(t)E_s \to E_s$. From the relation $(S_s(t))^{-1} = S_{s+t}(-t)$ and (3.12), we obtain (with a fixed real number t_0) $(S_s(t))^{-1} = S_{s-t_0}(t_0) \circ (S_{s-t_0}(t+t_0))^{-1}$, from which the Lipschitz continuity of $(S_s(t))^{-1}$ follows due to the continuity of $(S_{s-t_0}(t+t_0))^{-1}$ and the Lipschitz continuity of $S_{s-t_0}(t_0)$. This completes the proof.

We now obtain additional information about the Hausdorff dimension of sets of type (2.8) by estimating the quantities (3.21). We note that (5.13) is of the form (3.20) with $F'(\psi) = \Delta - g'(\psi)$. Then, using the condition (5.5) that $g'(s)$ is bounded from below and the procedure in [32, pp. 299-301], we obtain the following estimates, which are uniform in s:

$$ {}_sq_m(t) \leqq \frac{-C_1''}{2|\Omega|^{2/n}} m^{1+2/n} + C_2''C_4^{1+n/2}|\Omega|, \tag{5.18}$$

$$ {}_sq_m \leqq \frac{-C_1''}{2|\Omega|^{2/n}} m^{1+2/n} + C_2''C_4^{1+n/2}|\Omega|, \tag{5.19}$$

where C_1'' depends only on n and the shape of Ω, and C_2'' depends only on n. If m is sufficiently large so that the right-hand side of (5.19) is negative, then the m-volume element is exponentially decaying in H and $d_H(E_s) \leqq m$. Similar estimates can be obtained for the fractal dimension of E_s by using techniques discussed in [32].

To conclude the paper, we note that the considerations of the present section can be extended to systems of equations of type (5.1). Thus, just as we generated the system of hyperbolic equations (4.42) from the single equation (4.1), we obtain the system of parabolic equations

$$\psi_t - a\Delta\psi + g(\psi) = f(x, t) \tag{5.20}$$

from (5.1) by replacing the scalar ψ by an l-dimensional vector $\psi = (\psi_1, \cdots, \psi_l)$ and by replacing Δ by $a\Delta$ where a is a positive symmetric matrix, although we could also consider certain other linear operators L as in our discussion at the end of § 4 for the dissipative nonlinear wave equations. We consider the Hilbert space $H = (L^2(\Omega))^l$, where Ω is a bounded domain on which we impose conditions analogous to those stated at the beginning of this section in the case of (5.1). We impose Dirichlet boundary conditions (5.3) on each component of ψ, although Neumann or periodic boundary conditions could also be considered. There exists a function $H(\psi_1, \cdots, \psi_l)$ such that $g_i(\psi) = (\partial/\partial\psi_i)H(\psi_1, \cdots, \psi_l)$, $i = 1, \cdots, l$; and the hypotheses (5.2), (5.4)-(5.7), (5.15), and conditions such as $f \in C_b(\mathbb{R}, V)$, $f_t \in C_b(\mathbb{R}, H)$, f is time-dependent admissible, are replaced by their obvious vector analogues. Under these conditions, we can verify all the hypotheses required for the proofs of the dimension estimates in Theorems 3.2-3.4, and analogous results to those in (5.18), (5.19) and the succeeding discussion can be obtained for the system (5.20).

Acknowledgments. We thank G. R. Sell, A. Haraux, and A. V. Babin for informative discussions and express our appreciation to A. Haraux for copies of [19], [20] and to A. V. Babin for a copy of [4]. In addition, we are particularly grateful to G. Sell for comments and suggestions concerning this work, even though we did not incorporate all of his suggestions! We thank the two referees for their constructive suggestions and remarks.

REFERENCES

[1] L. Arnold and V. Wihstutz, *Lyapunov exponents: a survey in Lyapunov Exponents*, L. Arnold and V. Wihstutz, eds., Lecture Notes in Math. 1186, Springer-Verlag, Berlin, New York, 1986, pp. 1-26.

[2] A. V. Babin and M. I. Vishik, *Attractors of partial differential evolution equations and estimates of their dimension*, Russian Math. Surveys, 38 (1983), pp. 151-213.

[3] ———, *Regular attractors of semigroups and evolution equations*, J. Math. Pures Appl., 62 (1983), pp. 441-491.

[4] ———, *Attractors of evolutionary equations*, Izdat. "Nauka," Moscow, 1989. (In Russian.)

[5] C. Bardos and L. Tartar, *Sur l'unicité retrograde des équations paraboliques et quelques questions voisines*, Arch. Rational Mech. Anal., 50 (1973), pp. 10-25.

[6] S.-N. Chow, *Almost periodic differential equations*, Ph.D. thesis, Department of Mathematics, University of Maryland, College Park, MD, 1970.

[7] P. Constantin and C. Foias, *Global Lyapunov exponents, Kaplan-Yorke formulas and the dimension of the attractors for 2D Navier-Stokes equations*, Comm. Pure Appl. Math., 38 (1985), pp. 1-27.

[8] P. Constantin, C. Foias, and R. Témam, *Attractors representing turbulent flows*, Memoirs Amer. Math. Soc., 53 (1985).

[9] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago Press, Chicago, IL, 1988.

[10] C. M. Dafermos, *An invariance principle for compact processes*, J. Differential Equations, 9 (1971), pp. 239-252; Erratum, 10 (1971), pp. 179-180.

[11] ———, *Uniform processes and semicontinuous Liapunov functionals*, J. Differential Equations, 11 (1972), pp. 401-415.

[12] A. Douady and J. Oesterlé, *Dimension de Hausdorff des attracteurs*, C.R. Acad. Sci. Paris, 290, Ser. A (1980), pp. 1135-1138.

[13] A. M. Fink, *Almost periodic differential equations*, Lecture Notes in Math. 377, Springer-Verlag, Berlin, New York, 1974.

[14] J.-M. Ghidaglia, *Some backward uniqueness results*, Nonlinear Anal., Theory, Meth. Appl., 10 (1986), pp. 777-790.

[15] J.-M. Ghidaglia and R. Témam, *Attractors for damped nonlinear hyperbolic equations*, J. Math. Pures Appl., 66 (1987), pp. 273-319.

[16] J. K. Hale, *Asymptotic behavior of dissipative systems*, Math. Surveys Monographs, 25, American Mathematical Society, Providence, RI, 1988.

[17] A. Haraux, *Two remarks on hyperbolic dissipative problems*, in Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J.-L. Lions, eds., Pitman Res. Notes in Math. 122 (1985), pp. 161-179.

[18] ———, *Semi-linear hyperbolic problems in bounded domains*, Math. Rep., 3 (1987), pp. 1-281.

[19] ———, *Attractors of asymptotically compact processes and applications to nonlinear partial differential equations*, Comm. Partial Differential Equations, 13 (1988), pp. 1383-1414.

[20] ———, *Systemes dynamiques, processus et applications aux équations aux dérivées partielles*, Cours de D.E.A. 1988-1989, Université Pierre et Marie Curie, Paris, 1989.

[21] R. HARDT AND L. SIMON, *Seminar on Geometric Measure Theory*, Birkhäuser Verlag, Basel, 1986.

[22] O. A. LADYZHENSKAYA, *On the determination of minimal global attractors for the Navier-Stokes and other partial differential equations*, Russian Math. Surveys, 42 (1987), pp. 27-73.

[23] J.-L. LIONS, *Quelques méthodes de resolution des problemes aux limites non linéaires*, Dunod, Paris, 1969.

[24] J. MALLET-PARET, *Negatively invariant sets of compact maps and an extension of a theorem of Cartwright*, J. Differential Equations, 22 (1976), pp. 331-348.

[25] M. MARION, *Attractors for reaction-diffusion equations: existence and estimate of their dimension*, Appl. Anal., 25 (1987), pp. 101-147.

[26] B. NICOLAENKO, *Inertial manifolds for models of compressible gas dynamics*, in The Connection Between Infinite Dimensional and Finite Dimensional Dynamical Systems, B. Nicolaenko, C. Foias, and R. Témam, eds., Contemp. Math. 99, American Mathematical Society, Providence, RI, 1989, pp. 165-179.

[27] Y. B. PESIN, *Dimension type characteristics for invariant sets of dynamical systems*, Russian Math. Surveys, 43 (1988), pp. 111-151.

[28] G. RAUGEL AND G. R. SELL, *Navier-Stokes equations in thin 3D domains: global regularity of solutions* I, Army High Performance Computing Research Center, preprint 90-4, University of Minnesota, Minneapolis, MN, 1990.

[29] G. R. SELL, *Nonautonomous differential equations and topological dynamics. I. The basic theory*, Trans. Amer. Math. Soc., 127 (1967), pp. 241-262.

[30] ———, *Topological Dynamics and Ordinary Differential Equations*, Van Nostrand Reinhold, London, 1971.

[31] M. SLEMROD, *Asymptotic behavior of periodic dynamical systems on Banach spaces*, Ann. Mat. Pura Appl., 86 (1970), pp. 325-330; Erratum, Ibid., 88 (1971), p. 397.

[32] R. TÉMAM, *Infinite-dimensional dynamical systems in mechanics and physics*, Appl. Math. Sci., 68, Springer-Verlag, Berlin, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1230-1254, September 1992

CHAOTIC DYNAMICS OF QUASI-PERIODICALLY FORCED OSCILLATORS DETECTED BY MELNIKOV'S METHOD*

KAZUYUKI YAGASAKI†

Abstract. Nonlinear oscillators that have the form of quasi-periodic perturbations of planar Hamiltonian systems with homoclinic orbits are studied. For such systems, Melnikov's method permits determination, up to the leading term, whether or not the stable and unstable manifolds of normally hyperbolic invariant tori intersect transversely. In a more general setting it is proven that such intersection results in chaotic dynamics. These chaotic orbits are characterized by a generalization of the Bernoulli shift. An example is given to illustrate the theory. The result is also compared with the results of Wiggins [1988b], Scheurle [1986], and Meyer and Sell [1989].

Key words. chaos, Melnikov method, quasi-periodically forced oscillator, Bernoulli shift

AMS(MOS) subject classifications. 34C35, 58F13, 58F27, 58F30, 70K40, 70K50

1. Introduction.

Chaotic dynamics of periodically forced oscillators have been extensively studied in the past decade (cf. Thompson and Stewart [1986] and Moon [1987]). For many cases, the chaotic dynamics result from transverse intersection between the stable and unstable manifolds of hyperbolic periodic orbits. The Smale-Birkhoff homoclinic theorem provides a mechanism for this type of chaotic dynamics, and these chaotic orbits are characterized by the Bernoulli shift. Using Melnikov's method, we can also detect such intersection in a class of periodically forced systems. See Guckenheimer and Holmes [1983] and Wiggins [1990] for details of these ideas. Using the theory of exponential dichotomies and the shadowing lemma, Palmer [1984] also described Melnikov's method and proved the Smale-Birkhoff homoclinic theorem in the context of periodic differential equations.

Recently, Wiggins [1988b] generalized Melnikov's method to a class of quasi-periodically forced systems. This version of Melnikov's method permits us to detect transverse intersection between the stable and unstable manifolds of normally hyperbolic invariant tori. His result is also applicable to systems with frequencies depending on the state variables and involves many of other versions of Melnikov's method, such as those of Holmes and Marsden [1982a], [1982b], [1983] and Wiggins and Holmes [1987]. It is also important to detect such intersection in three or more degrees of freedom Hamiltonian systems since their existence gives a mechanism for the Arnold diffusion (Arnold [1964], Lichtenberg and Lieberman [1983]). See Holmes and Marsden [1982b] and Wiggins [1988b, § 4.1] for the exposition of the Arnold diffusion in terms of the Melnikov theory.

Furthermore, Wiggins [1988b] proved a generalization of the Smale-Birkhoff homoclinic theorem: if the stable and unstable manifolds of a normally hyperbolic invariant torus intersect transversely in a torus satisfying a certain condition (see § 3), then the Bernoulli shift flow can be imbedded into the dynamics of the quasi-periodically forced system. A similar result had been obtained by Silnikov [1968]. Wiggins [1987], [1988a], [1988b], and Ide and Wiggins [1989] also applied these techniques to several types of quasi-periodically forced oscillators and obtained criteria for the existence of chaos, although their proof was not complete (see § 3).

Yagasaki [1990a], [1990b], [1991a] studied chaotic dynamics of quasi-periodically forced oscillators with weak nonlinearity. He used the averaging method and the

* Received by the editors January 28, 1991; accepted for publication (in revised form) January 9, 1992.

† Department of Mechanical Engineering, Tamagawa University, Machida, Tokyo 194, Japan.

standard Melnikov technique to show that the stable and unstable manifolds of a normally hyperbolic invariant torus intersect transversely in a torus, and then applied a generalization of Smale–Birkhoff homoclinic theorem by Wiggins [1988b] to obtain the regions in parameter space where chaos may occur. See also Yagasaki, Sakata, and Kimura [1990].

Scheurle [1986] and Meyer and Sell [1989] studied almost periodically forced systems extending the idea of Palmer [1984]. Scheurle [1986] showed that the existence of random-like solutions can be detected using a Melnikov type analysis and the shadowing lemma. Meyer and Sell [1989] used the concept of the skew product flow and generalized Melnikov's method, the shadowing lemma, and the Smale–Birkhoff homoclinic theorem to describe a mechanism for chaos in almost periodically forced systems. They showed that the chaotic behavior is also described in terms of the Bernoulli shift.

Consider nonlinear oscillators having the form of quasi-periodic perturbations of planar Hamiltonian systems:

$$\begin{aligned} &\dot{x} = JDH(x) + \varepsilon g(x, \theta), \\ &\dot{\theta} = \omega, \qquad (x, \theta) \in \mathbb{R}^2 \times T^l, \quad \omega \in \mathbb{R}^l, \end{aligned} \tag{1.1}$$

where $H(x)$ is a Hamiltonian function and $T^l = \prod_{i=1}^l S^1$ is an l-torus with $S^1 = \mathbb{R}/2\pi$ the circle of length 2π and

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{1.2}$$

We assume that when $\varepsilon = 0$, (1.1) has a homoclinic orbit $\bar{x}_0(t)$ to a hyperbolic fixed point x_0. When ε is sufficiently small, (1.1) has a normally hyperbolic invariant torus T_ε. Using the version of Melnikov's method due to Wiggins [1988b], we can determine the behavior of the stable and unstable manifolds $W^S(T_\varepsilon)$, $W^u(T_\varepsilon)$ of the normally hyperbolic invariant torus T_ε. More precisely, we can state this result as follows. Let $M(\theta)$ be the *Melnikov function* for $\bar{x}_0(t)$, i.e.,

$$M(\theta) = \int_{-\infty}^{\infty} DH(\bar{x}_0(t)) \cdot g(\bar{x}_0(t), \omega t + \theta)\, dt. \tag{1.3}$$

If there exists a point $\theta = \theta_0 \in T^l$ such that

$$M(\theta_0) = 0, \tag{M1}$$

$$DM(\theta_0) \text{ is of rank one,} \tag{M2}$$

then $W^s(T_\varepsilon)$ and $W^u(T_\varepsilon)$ intersect transversely.

The Smale–Birkhoff homoclinic theorem guarantees the existence of chaos when the stable and unstable manifolds of a hyperbolic periodic orbit intersect transversely; however, the generalization of the Smale–Birkhoff homoclinic theorem due to Wiggins [1988b] does not always do so when the stable and unstable manifolds of a normally hyperbolic invariant torus intersect transversely. In general, there exist wide parameter regions in which conditions (M1) and (M2) are satisfied, but Wiggin's theorem does not apply (see § 7).

In this paper we obtain a more general condition for the existence of chaos in (1.1). Specifically, we prove that if there exists a point $\theta_0 \in T^l$ satisfying (M1) and

$$\sum_{j=1}^{l} \omega_j \frac{\partial}{\partial \theta_j} M(\theta_0) \neq 0, \tag{M3}$$

then there exists an invariant set on which orbits are characterized by a generalization of the Bernoulli shift. Precisely, condition (M3) is different from (M2), but their difference is very little (see § 7). Our condition is also very similar to that of Scheurle [1986] for the existence of random-like solutions in almost periodic systems. Furthermore, if (M3) holds for all points $\theta_0 \in T^l$ satisfying (M1), then we can apply the result of Meyer and Sell [1989]. See Appendix A for their results in the quasi-periodic perturbation case.

This paper is organized as follows. In §§ 2 and 3 we discuss the behavior of the stable and unstable manifolds of normally hyperbolic invariant tori in quasi-periodically forced oscillators, using Melnikov's method. In § 4 we present a generalization of the Bernoulli shift to be used to describe chaotic dynamics of quasi-periodically forced oscillators. In § 5 we study a class of diffeomorphisms instead of ordinary differential equations. This class of diffeomorphisms contains the Poincaré maps for quasi-periodically forced oscillators. We prove that if the stable and unstable manifolds of a normally hyperbolic invariant torus intersect transversely in a certain type of manifold, then there exists an invariant set on which the diffeomorphism is topologically conjugate to the generalized Bernoulli shift. In § 6 we provide a criterion for the existence of chaos in quasi-periodically forced oscillators using Melnikov's method and the result of § 5. In § 7 we give an example to illustrate our theory. Our result is also compared with the previous results of Wiggins [1988b], Scheurle [1986], and Meyer and Sell [1989].

After this paper was written, the author learned of the related work of Beigie, Leonard, and Wiggins [1991a], [1991b] and Stoffer [1988]. Their results also improved the generalization of the Smale–Birkhoff homoclinic theorem by Wiggins [1988b]. In particular, Stoffer [1988] described chaotic dynamics in general nonautonomous systems by the "extended shift map," in which the concept is similar to that of our generalized Bernoulli shift. However, in quasi-periodically forced systems, our description of chaos seems to be more suitable than theirs, since it takes a recurrence property of chaotic attractors into account. See Yagasaki [1991b] for more details.

2. Melnikov's method. In this section we briefly review Melnikov's method for quasi-periodically forced oscillators. This version of Melnikov's method is due to Wiggins [1988b]. Although we only deal here with single-degree-of-freedom systems with constant frequencies, this technique has been developed for multi-degree-of-freedom systems with nonconstant frequencies depending on the state variables. See Wiggins [1988b] for the general theory of this method.

We consider systems of the form

$$\begin{aligned} &\dot{x} = JDH(x) + \varepsilon g(x, \theta) \\ &\dot{\theta} = \omega, \qquad (x, \theta) \in \mathbb{R}^2 \times T^l, \quad \omega \in \mathbb{R}^l, \end{aligned} \tag{2.1}$$

where $T^l = \prod_{i=1}^l S^1$ is an l-torus with $S^1 = \mathbb{R}/2\pi$,

$$H : \bar{U} \to \mathbb{R}^1$$

is C^{r+1} $(r \geqq 2)$, and

$$g : \bar{U} \times T^l \to \mathbb{R}^2$$

is C^r with some open set $\bar{U} \subset \mathbb{R}^2$; J is a 2×2 matrix given by (1.2) and $0 < \varepsilon \ll 1$. For $\varepsilon = 0$, (2.1) becomes a Hamiltonian system

$$\dot{x} = JDH(x), \qquad \dot{\theta} = \omega. \tag{2.2}$$

We make the following assumption on (2.2).

(A1) The x-component of (2.2) has a homoclinic orbit $\bar{x}_0(t)$ to a hyperbolic saddle point x_0. We denote $\Gamma=\{\bar{x}_0(t)\,|\,t\in\mathbb{R}\}$.

In the full phase space $\mathbb{R}^2\times T^l$, (2.2) has a normally hyperbolic invariant l-torus

$$T_0=\{(x_0,\theta)\,|\,\theta\in T^l\}=\{x_0\}\times T^l,$$

whose $l+1$-dimensional stable and unstable manifolds $W^s(T_0)$, $W^u(T_0)$ coincide along the $l+1$-dimensional homoclinic manifold given by

$$W^s(T_0)\cap W^u(T_0)=\{(\bar{x}_0(t),\theta)\,|\,t\in\mathbb{R},\ \theta\in T^l\}=\Gamma\times T^l.$$

See Fig. 1 for the phase space of the unperturbed system.

Here "normal hyperbolicity" means that the expansive and contraction rates of the flow generated by (2.2) normal to T_0 dominate those tangent to T_0. For (2.2) this is clear since x_0 is a hyperbolic fixed point of the first equation of (2.2) so that trajectories approach T_0 exponentially fast in positive or negative time, but the flow on T_0 only indicates a rotation with the frequency vector ω. See Hirsch, Pugh, and Shub [1977] for precise definitions of normal hyperbolicity.

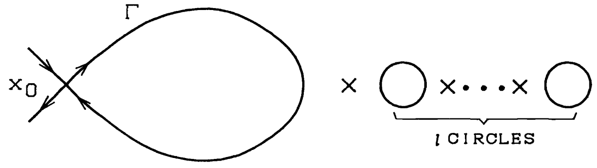


FIG. 1. *The unperturbed phase space of* (2.2).

We will reduce the study of (2.1) to that of the associated Poincaré map. We make the section to the phase space $\mathbb{R}^2\times T^l$ by fixing any element of θ, say θ_i, as follows:

$$\Sigma_i=\{(x,\theta)\in\mathbb{R}^2\times T^l\,|\,\theta_i=0\},\qquad i=1,\cdots,l.$$

Let us denote

$$\bar{\theta}^i=(\theta_1,\cdots,\theta_{i-1},\theta_{i+1},\cdots,\theta_l)\in T^{l-1}$$

and

$$\bar{\omega}^i=(\omega_1,\cdots,\omega_{i-1},\omega_{i+1},\cdots,\omega_l)\in\mathbb{R}^{l-1}.$$

The Poincaré map $P_{\varepsilon,i}:\Sigma_i\to\Sigma_i$ generated by the flow of (2.1) is given by

$$P_{\varepsilon,i}:(x(0),\bar{\theta}^i)\to\left(x\left(\frac{2\pi}{\omega_i}\right),\bar{\theta}^i+\frac{2\pi\bar{\omega}^i}{\omega_i}\right),$$

where $(x(t),\omega t+\theta)$ is a solution of (2.1). We denote the Poincaré map associated with the unperturbed system (2.2) by $P_{0,i}$.

The unperturbed Poincaré map $P_{0,i}$ has a normally hyperbolic invariant $(l-1)$-torus

$$\mathcal{T}_{0,i}=\Sigma_i\cap T_0=\{x_0\}\times T^{l-1},$$

whose l-dimensional stable and unstable manifolds $W^s(\mathcal{T}_{0,i})$, $W^u(\mathcal{T}_{0,i})$ coincide along the l-dimensional homoclinic manifold

$$W^s(\mathcal{T}_{0,i})\cap W^u(\mathcal{T}_{0,i})=\Gamma\times T^{l-1}.$$

See Fig. 2 for the phase space of the unperturbed Poincaré map $P_{0,i}:\Sigma_i\to\Sigma_i$ when $l=2$ and $i=1$. These structures of the unperturbed phase space persist for the perturbed phase space as follows.

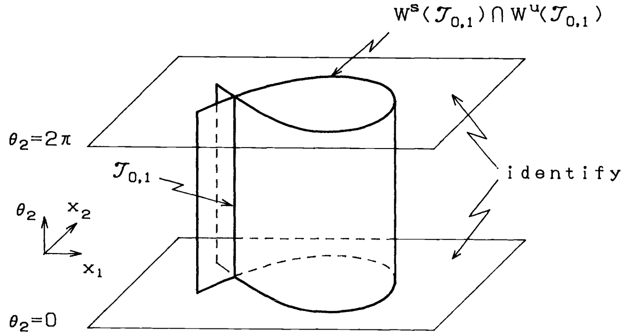


FIG. 2. *The stable and unstable manifolds for the unperturbed Poincaré map $P_{0,1}$ in the case of $l=2$.*

PROPOSITION 2.1. *For ε sufficiently small, the perturbed system* (2.1) *has a C^r normally hyperbolic invariant l-torus T_ε, whose C^r, $l+1$-dimensional local stable and unstable manifolds $W^s_{\text{loc}}(T_\varepsilon)$, $W^u_{\text{loc}}(T_\varepsilon)$ are C^r, ε-close to $W^s(T_0)$ and $W^u(T_0)$, respectively. Equivalently, the Poincaré map $P_{\varepsilon,i}$ of* (2.1) *has a C^r normally hyperbolic invariant $(l-1)$-torus $\mathcal{T}_{\varepsilon,i}$ whose C^r, l-dimensional local stable and unstable manifolds $W^s_{\text{loc}}(\mathcal{T}_{\varepsilon,i})$, $W^u_{\text{loc}}(\mathcal{T}_{\varepsilon,i})$ are C^r, ε-close to $W^s(\mathcal{T}_{0,i})$ and $W^u(\mathcal{T}_{0,i})$, respectively.*

Proof. These are immediate consequences of the invariant manifold theorem. See Hirsch, Pugh, and Shub [1977]. □

The manifolds $W^s(T_\varepsilon)$ and $W^u(T_\varepsilon)$ (or $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$) may not coincide, but can intersect transversely. Computation of the Melnikov function provides information on the behavior of these manifolds.

In Wiggins [1988b], it was shown that the distance between $W^s(T_\varepsilon)$ and $W^u(T_\varepsilon)$ near the point

$$(x, \theta) = (\bar{x}_0(-t_0), \theta_0) \in \mathbb{R}^2 \times T^l$$

is given by

$$d(t_0, \theta_0) = \varepsilon \frac{M(\omega \tau_0 + \theta_0)}{K(t_0)} + \mathcal{O}(\varepsilon^2),$$

where

$$M(\theta) = \int_{-\infty}^{\infty} DH(\bar{x}_0(t)) \cdot g(\bar{x}_0(t), \omega t + \theta)\, dt, \tag{2.3}$$

$K(t_0)$ is a nonzero $\mathcal{O}(1)$ quantity and "$\cdot$" denotes the usual vector dot product. The function $M(\theta)$ is called the *Melnikov function*. We have the following theorem.

THEOREM 2.2. *Suppose that there exists a point $\theta = \theta_0 \in T^l$ such that*

(M1) $$M(\theta_0) = 0,$$

(M2) $$DM(\theta_0) \text{ is of rank one.}$$

Then, for $\varepsilon > 0$ sufficiently small and $t_0 \in \mathbb{R}$, $W^s(T_\varepsilon)$ and $W^u(T_\varepsilon)$ intersect transversely near the point $(\bar{x}(-t_0), \theta_0 - \omega t_0)$. Equivalently, $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely near the point

$$(x, \bar{\theta}^i) = (\bar{x}_0(-t_0), \bar{\theta}^i_0 - \bar{\omega}^i t_0) \in \Sigma_i, \tag{2.4}$$

where $\omega_i t_0 = \theta_{i0} \bmod 2\pi$.

Proof. The first part is a special case of Theorem 4.1.10 of Wiggins [1988b]. The second part is easily proved by noting that $W^s(\mathcal{T}_{\varepsilon,i}) = W^s(T_\varepsilon) \cap \Sigma_i$ and $W^u(\mathcal{T}_{\varepsilon,i}) = W^u(T_\varepsilon) \cap \Sigma_i$. □

Suppose that a point $\theta_0 \in T^l$ satisfies (M1) and (M2). Then, the local implicit function theorem (cf. Chow and Hale [1982]) implies that the zero of the Melnikov function $M(\theta)$ can be continued to an $l-1$-dimensional set in T^l. Hence, from Theorem 2.2 we see that there exists an $l-1$-dimensional manifold $\gamma_{\varepsilon,i}$ in which $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely. We refer to this manifold $\gamma_{\varepsilon,i}$ as a *transverse homoclinic manifold.*

In the next section we will discuss the structure of the transverse homoclinic manifold $\gamma_{\varepsilon,i}$ and describe how $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely, using the Melnikov function $M(\theta)$.

3. Behavior of stable and unstable manifolds. Ide and Wiggins [1989] stated that if $M(\theta)$ has a zero at $\theta = \theta_0$ and $DM(\theta)$ has rank one for all $\theta \in T^l$, then the zero θ_0 can be continued to an $(l-1)$-torus. More generally, suppose that (M2) holds at any point $\theta_0 \in T^l$ satisfying (M1). Then, using the global implicit function theorem (Chow and Hale [1982]) and modifying arguments given in Wiggins [1988b, p. 464], we can take an $(l-1)$-torus as the zero set of $M(\theta)$. We denote this torus by τ_0.

Let us fix the value of i and discuss the phase space for the Poincaré map $P_{\varepsilon,i}: \Sigma_i \to \Sigma_i$. We first consider the case in which condition

$$\text{(M2i)} \qquad \frac{\partial}{\partial \theta_i} M(\theta_0) \neq 0.$$

holds at any point $\theta_0 \in T^l$ satisfying (M1). Note that (M2i) is stronger than (M2). In this case, the global implicit function theorem implies that there exists a C^r function $h: \mathbb{R}^{l-1} \to \mathbb{R}$, such that τ_0 is given by

$$\text{(3.1)} \quad \tau_0 = \{\theta = (\theta_1, \cdots, \theta_l) \in T^l \mid \theta_i = h(\bar{s}^i),\ \theta_j = s_j \bmod 2\pi \text{ for } s_j \in \mathbb{R},\ 1 \leqq j \neq i \leqq l\}.$$

where $\bar{s}^i = (s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_l)$. Moreover, we can assume that for each $j \neq i$, there are two integers j_1 and j_2 such that $j_1 > 0$ and

$$\text{(3.2)} \qquad h(s_1, \cdots, s_j + 2j_1\pi, \cdots, s_l) = h(s_1, \cdots, s_j, \cdots, s_l) + 2j_2\pi,$$

since τ_0 is an $(l-1)$-torus. Choose j_1 as the minimum positive one such that (3.2) holds. Then $j_2 = 0$ implies that $j_1 = 1$, since τ_0 cannot intersect itself.

Let $\gamma_{0,i}$ be the set of all points given by (2.4) with $\theta_0 \in \tau_0$, i.e.,

$$\text{(3.3)} \qquad \gamma_{0,i} = \{(\bar{x}_0(-t_0), \bar{\theta}_0^i - \bar{\omega}^i t_0) \in \Sigma_i \mid (\theta_{10}, \cdots, \theta_{l0}) \in \tau_0,\ \omega_i t_0 = \theta_{i0} \bmod 2\pi\}.$$

By Theorem 2.2. $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely in an $l-1$-dimensional manifold $\gamma_{\varepsilon,i}$ near $\gamma_{0,i}$.

DEFINITION 3.1. Consider the product space $\mathbb{R}^m \times T^l$. Let τ be an l-torus given by

$$\text{(3.4)} \quad \tau = \{(x, \theta) \in \mathbb{R}^m \times T^l \mid x = \tilde{x}(s_1, \cdots, s_l),\ \theta_j = h_j(s_j) \bmod 2\pi,\ s_j \in \mathbb{R}, j = 1, \cdots, l\},$$

where $\tilde{x}: \mathbb{R}^l \to \mathbb{R}^m$ is C^r and 2π-periodic in s_j, $j = 1, \cdots, l$, and $h_j: \mathbb{R} \to \mathbb{R}$, $j = 1, \cdots, l$, are C^r and satisfy $|h_j(2\pi) - h_j(0)| = 2n_j\pi$ for some integer n_j. Choose h_j for $j = 1, \cdots, l$ such that n_j is the minimum nonnegative one, and let $n = (n_1, \cdots, n_l)$. Then we refer to τ as a *torus of n-cycle.*

If $j_2 = 0$ for all $j \neq i$ in (3.2), then $\gamma_{0,i}$ is an $(l-1)$-torus of $(1, \cdots, 1)$-cycle. However, if $j_2 \neq 0$ for some j, then

$$\bar{x}_0(-h(s_1, \cdots, s_j + 2j_1\pi, \cdots, s_l)/\omega_i) \neq \bar{x}_0(-h(s_1, \cdots, s_j, \cdots, s_l)/\omega_i),$$

and hence $\gamma_{0,i}$ is not a torus. Since $\gamma_{\varepsilon,i}$ is ε-close to $\gamma_{0,i}$, $\gamma_{\varepsilon,i}$ is an $(l-1)$-torus of $(1, \cdots, 1)$-cycle if $j_2 = 0$ for all $j \neq i$, but $\gamma_{\varepsilon,i}$ is not a torus otherwise. These situations are shown in Fig. 3(a) and (b) for the Poincaré map $P_{\varepsilon,1}:\Sigma_1\to\Sigma_1$ when $l=2$. (Fig. 3(a) corresponds to the case of $j_2=0$ and Fig. 3(b) to the other case.) When we can take a torus as the transverse homoclinic manifold, as shown in Fig. 3(a), we call the torus a *transverse homoclinic torus.*

We next consider the case in which there is a point $\theta_0\in T^l$ that satisfies (M1) but does not satisfy (M2i), i.e., $M=0$ and $\partial M/\partial\theta_i=0$ at $\theta=\theta_0$. Then, in general, there is not a function h satisfying (3.1) and (3.2). Furthermore, there may exist C^r functions $h_i:\mathbb{R}^{l-1}\to\mathbb{R}$, $h_j:\mathbb{R}\to\mathbb{R}$, $0<j\neq i\leqq l$, of 2π-period in each of the arguments, such that

$$(3.5)\quad \tau_0=\{\theta=(\theta_1,\cdots,\theta_l)\in T^l\mid \theta_i=h_i(\bar{s}^i),\ \theta_j=h_j(s_j) \bmod 2\pi,\ s_j\in\mathbb{R}1\leqq j\neq i\leqq l\}.$$

From (3.3) we see that $\gamma_{\varepsilon,i}$ is an $(l-1)$-torus of $(0,\cdots,0)$-cycle since $\gamma_{\varepsilon,i}$ is ε-close to $\gamma_{0,i}$. See Fig. 3(c).

Suppose that as shown in Fig. 3(a), there exists a transverse homoclinic $(l-1)$-torus of $(1,\cdots,1)$-cycle $\tau_{\varepsilon,i}$. This is the case when the stable and unstable manifolds of normally hyperbolic invariant tori intersect transversely in a class of quasi-periodically forced oscillators with weak nonlinearity (Yagasaki [1990a], [1990b], [1991a]). In this situation, Wiggins [1988b] proved that the dynamics of the Poincaré map normal to the transverse homoclinic torus is chaotic. More precisely, we can state his result as follows.

THEOREM 3.1. *Suppose that $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely in an $(l-1)$-torus of $(1,\cdots,1)$-cycle. Then, for some $k\geqq 1$, $P^k_{\varepsilon,i}$ has an invariant Cantor set*

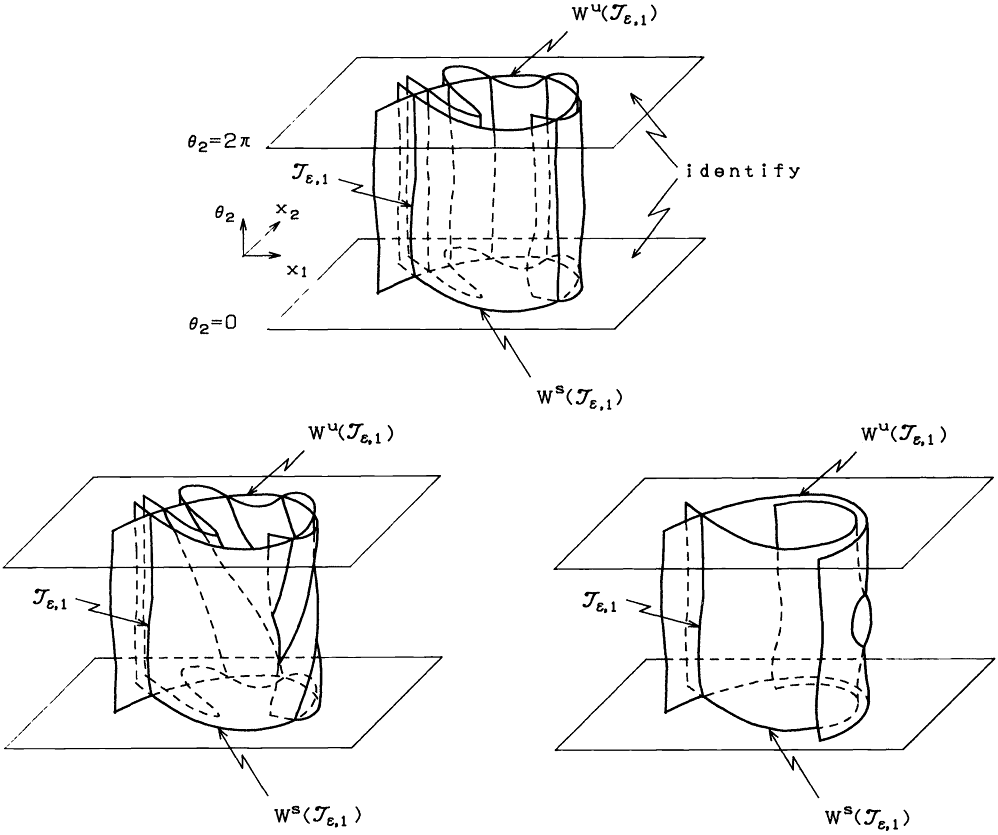

$W^u(\mathcal{T}_{\varepsilon,1})$
$\theta_2=2\pi$
identify
$\mathcal{T}_{\varepsilon,1}$
θ_2
x_2
x_1
$\theta_2=0$
$W^s(\mathcal{T}_{\varepsilon,1})$
$W^u(\mathcal{T}_{\varepsilon,1})$
$\mathcal{T}_{\varepsilon,1}$
$W^s(\mathcal{T}_{\varepsilon,1})$
$W^u(\mathcal{T}_{\varepsilon,1})$
$\mathcal{T}_{\varepsilon,1}$
$W^s(\mathcal{T}_{\varepsilon,1})$

FIG. 3. *Transverse intersection between the stable and unstable manifolds for the Poincaré map $P_{\varepsilon,1}$ in the case of $l=2$.* (a) *$\gamma_{\varepsilon,1}$ is a 1-torus of 1-cycle.* (b) *$\gamma_{\varepsilon,1}$ is not a torus.* (c) *$\gamma_{\varepsilon,1}$ is a 1-torus of 0-cycle.*

of $(l-1)$-tori, Ξ. Moreover, there exists a homeomorphism h taking tori in Ξ to bi-infinite sequences of N symbols such that the following diagram commutes

$$\begin{tikzcd} \Xi \arrow[r, "P^k_{\varepsilon,i}"] \arrow[d, "h"'] & \Xi \arrow[d, "h"] \\ B_N \arrow[r, "\sigma"] & B_N \end{tikzcd}$$

where B_N is the space of bi-infinite sequences of N symbols and $\sigma: B_N \to B_N$ is the shift map.

Proof. This is a special case of Theorem 3.4.1 of Wiggins [1988b]. The assumption that the transverse homoclinic torus τ is of $(1, \cdots, 1)$-cycle is essential in his result, although he did not explicitly state it. In fact, his proof requires that the transverse homoclinic torus τ can be expressed as

$$\tau = \{(x, \bar{\theta}^i) = P^{k'}_{\varepsilon,i}((0, x_2), \bar{\theta}^i_0) \mid \bar{\theta}^i_0 \in T^{l-1}\},$$

with some $x_2 \in \mathbb{R}$ and $k' > 0$ in an adequate coordinate system. This implies that τ is an $(l-1)$-torus of $(1, \cdots, 1)$-cycle. See, e.g., § 4 for the definition of the shift map σ. □

Remark 3.1. The pair (B_N, σ) is referred to as the Bernoulli shift. This is a simple dynamical system displaying stochastic behavior. If the hypothesis of Theorem 3.1 is satisfied, then the dynamics of $P^k_{\varepsilon,i}$ on Ξ is chaotic like the Bernoulli shift. See also § 4 for the Bernoulli shift.

Remark 3.2. The hypothesis of Theorem 3.1 does not always hold even if $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely, i.e., there exists a point $\theta_0 \in T^l$ satisfying (M1) and (M2). In particular, Theorem 3.1 has nothing to say about situations such as those shown in Fig. 3(b) and (c). Wiggins [1987], [1988a], [1988b] and Ide and Wiggins [1989] overlooked this fact.

Using Theorem 2.2, we obtain the following result as a corollary of Theorem 3.1.

THEOREM 3.2. *Suppose that there exists a continuous function $h(\bar{s}^i)$ of period 2π in each of the arguments such that for all $s_j \in \mathbb{R}$, $j \neq i$, (M1) and (M2) hold at $\theta_0 \in T^l$, where $\theta_{i0} = h(\bar{s}^i)$ and $\theta_{j0} = s_j \bmod 2\pi$, $j \neq i$. Then the statements of Theorem 3.1 hold.*

Proof. Suppose that the hypothesis holds. Then, the zero set of the Melnikov function $M(\theta)$ is an $(l-1)$-torus given by

$$\tau_0 = \{\theta_0 \in T^l \mid \theta_{i0} = h(\bar{s}^i),\ \theta_{j0} = s_j \bmod 2\pi \text{ for } s_j \in \mathbb{R}, 0 < j \neq i \leq l\}.$$

It follows from Theorem 2.2 that $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely near

$$\tau_{0,i} = \{(x, \bar{\theta}^i) \in \mathbb{R}^2 \times T^{l-1} \mid x = \bar{x}(-t_0),\ \omega_i t_0 = \theta_{i0},\ \theta_j = \theta_{j0} - \omega_j t_0 \bmod 2\pi \text{ for } j \neq i \text{ and } \theta_0 \in \tau_0\}.$$

We can easily show that $\tau_{0,i}$ is an $(l-1)$-torus of $(1, \cdots, 1)$-cycle. Hence, $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely in an $(l-1)$-torus $\tau_{\varepsilon,i} = \tau_{0,i} + \mathcal{O}(\varepsilon)$ of $(1, \cdots, 1)$-cycle. Applying Theorem 3.1, we obtain the desired result. □

We will give a more comprehensive criterion for the existence of chaos in quasi-periodically forced oscillators in § 6.

4. The generalized Bernoulli shift. In this section we present a generalization of the Bernoulli shift for the precise description of chaos in quasi-periodically forced oscillators. We begin with the definition of the Bernoulli shift.

Let $S_N = \{1, 2, \cdots, N\}$ for $N \geqq 2$. We define $B_N = \prod_{i=-\infty}^{\infty} S_N$, i.e., B_N is a collection of all infinite bisequences of elements of S_N. Thus, if $s \in B_N$, then $s = \{\cdots, s_{-1}, s_0, s_1, \cdots\}$ where $s_i \in S_N$, $i \in \mathbb{Z}$. Let $\sigma: B_N \to B_N$ be the *shift map* defined by $(\sigma(s))_j = s_{j+1}$.

DEFINITION 4.1. The discrete dynamical system (B_N, σ) is called the *Bernoulli shift*, or the *full shift on N symbols.*

It is clear that B_N is invariant by σ. For two sequences $s, s' \in B_N$, define the distance between them by

$$d(s, s') = \sum_{i=-\infty}^{\infty} \frac{1}{2^{|i|}} \frac{|s_i - s'_i|}{1 + |s_i - s'_i|}. \tag{4.1}$$

It is easy to show that $d(\cdot, \cdot)$ is a metric on B_N. Moreover, σ is continuous and B_N is compact. See Wiggins [1988b, § 2.2a, b], [1990, § 4.2] for the details. We also call a sequence $s \in B_N$ an *orbit in* B_N.

The Bernoulli shift contains important features of randomness. In particular, when $N = 2$, it provides a model of a completely random process, coin tossing. The shift map σ has a countable infinity of periodic orbits, an uncountable infinity of nonperiodic orbits, and a dense orbit. The Bernoulli shift is often used to describe chaotic behavior in deterministic dynamical systems. See Guckenheimer and Holmes [1983] and Wiggins [1988b] for such examples; see also Theorem 3.1.

For descriptions of chaos in some dynamical systems, it is necessary to restrict the domain of the shift map σ to a subset of B_N. This is accomplished as follows.

Let A be an $N \times N$ matrix all of whose elements are 0 or 1. The matrix A is called a *transition matrix.* We denote the set of all $N \times N$ transition matrices by M_N.

DEFINITION 4.2. For any $A \in M_N$, let $B_N(A)$ be a subset of B_N given by

$$B_N(A) = \{s \in B_N \mid (A)_{s_i s_{i+1}} = 1 \text{ for all } i\}.$$

The pair $(B_N(A), \sigma)$ is called a *subshift of finite type.* We also say that the transition matrix A is *irreducible* if there is an integer $k > 0$ such that $(A^k)_{ij} \neq 0$ for all $i, j \in S_N$.

It is easy to show that $B_N(A)$ is σ-invariant and compact with the metric (4.1). We call a sequence $s \in B_N(A)$ an *orbit in* $B_N(A)$. If A is irreducible, then for any $i, j \in S_N$ there is an orbit $s \in B_N(A)$ such that $s_0 = i$ and $(\sigma^k(s))_0 = j$ for some $k > 0$. Moreover, the subshift of finite type $(B_N(A), \sigma)$ has such properties as the Bernoulli shift (B_N, σ). For example, there exist a countable infinity of periodic orbits, an uncountable infinity of nonperiodic orbits, and a dense orbit. See Wiggins [1988b, § 2.2c] for the details.

Now we generalize the concept of the Bernoulli shift so that we can describe chaotic dynamics of quasi-periodically forced oscillators.

Let $R_\nu(\theta)$, $\theta = (\theta_1, \cdots, \theta_l) \in T^l$ be a rigid rotation through an angle ν_i in the θ_i direction for $i = 1, \cdots, l$:

$$R_\nu(\theta) = \theta + \nu, \tag{4.2}$$

where $\nu = (\nu_1, \cdots, \nu_l) \in T^l$. Let $\Theta \subset T^l$ be an l-dimensional invariant manifold for R_ν, i.e., $R_\nu(\Theta) = \Theta$. Define a set $\mathscr{B}_N(\nu)$ as follows: if $\xi \in \mathscr{B}_N(\nu)$, then $\xi = \{\cdots, \xi_{-1}, \xi_0, \xi_1, \cdots\}$ where $\xi_i = (s_i, \phi_i)$ with $s_i \in S_N$ and $\phi_i \in \Theta$, $i \in \mathbb{Z}$, such that

$$\phi_{i+1} = R_\nu(\phi_i) \tag{4.3}$$

for $i \in \mathbb{Z}$. We also denote $\xi = \{\cdots, \xi_{-1}, \xi_0, \xi_1, \cdots\} \in \mathscr{B}_N(\nu)$ with $\xi_i = (s_i, \phi_i)$ by (s, ϕ) with $s = \{\cdots, s_{-1}, s_0, s_1, \cdots\}$ and $\phi = \{\cdots, \phi_{-1}, \phi_0, \phi_1, \cdots\}$. Given the metric

$$\tilde{d}(\xi, \xi') = d(s, s') + |\phi_0 - \phi'_0|, \tag{4.4}$$

between $\xi=(s, \phi)$ and $\xi'=(s', \phi')$, $\mathscr{B}_N(\nu)$ becomes a metric space. We can identify $\xi=(s, \phi)\in\mathscr{B}_N(\nu)$ with $(s, \phi_0)\in B_N\times\Theta$, so that $\mathscr{B}_N(\nu)\cong B_N\times\Theta$. We define the shift map $\sigma_\nu: \mathscr{B}_N(\nu)\to\mathscr{B}_N(\nu)$ as

$$(\sigma_\nu(\xi))_i=\xi_{i+1}.$$

It is clear that σ_ν is continuous in $\mathscr{B}_N(\nu)$ with the metric (4.4). Moreover, $\mathscr{B}_N(\nu)$ is compact if Θ is closed. Let $\mathscr{A}=\{A(\theta)\in M_N \mid \theta\in\Theta\}$ be an l-parameter family of transition matrices. For $\theta\in\Theta$, let

$$S_N(\theta)=\{i\in S_N \mid (A(\theta))_{ij}=1 \text{ for some } j\in S_N\}$$

and

$$S'_N(\theta)=\{j\in S_N \mid (A(\theta))_{ij}=1 \text{ for some } i\in S_N\}.$$

DEFINITION 4.3. We say that $\mathscr{A}$ is *consistent with* R_ν if the two following conditions are satisfied:

(1) For any $i\in S_N$ there are $\theta, \theta'\in\Theta$ such that $i\in S_N(\theta)$ and $i\in S'_N(\theta')$;

(2) $S_N(R_\nu(\theta))=S'_N(\theta)$ for any $\theta\in\Theta$.

DEFINITION 4.4. Let $\mathscr{A}$ be consistent with R_ν, and let $\mathscr{B}_N(\mathscr{A}, \nu)$ be a subset of $\mathscr{B}_N(\nu)$ given by

$$\mathscr{B}_N(\mathscr{A}, \nu)=\{\xi=(s, \phi)\in\mathscr{B}_N(\nu) \mid (A(\phi_i))_{s_is_{i+1}}=1 \text{ for all } i\}.$$

We call the pair $(\mathscr{B}_N(\mathscr{A}, \nu), \sigma_\nu)$ the *generalized Bernoulli shift.*

It is clear that $\mathscr{B}_N(\mathscr{A}, \nu)$ is σ_ν-invariant. Let $A(\theta)$ be independent of θ, i.e., $A(\theta)=A=\text{const.}$ Then $\mathscr{B}_N(\mathscr{A}, \nu)\cong B_N(A)\times\Theta$. Moreover, if Θ is closed, then $\mathscr{B}_N(\mathscr{A}, \nu)$ is compact. However, in general, $\mathscr{B}_N(\mathscr{A}, \nu)$ may not be compact even if Θ is closed. In fact, let $l=1$ and $\Theta=S^1$, and suppose that $A(\theta)$ is not continuous but only right continuous at $\theta=\theta_0$, and for some $\varepsilon>0$ there is a pair of integers $i, j\in S_N$ such that $(A(\theta_0))_{ij}=0$ and $(A(\theta))_{ij}=1$ for $\theta\in(\theta_0-\varepsilon, \theta_0)$. Let $\xi^k=(s^k, \phi^k)$, $k=1, 2, \cdots$, be a sequence of elements of $\mathscr{B}_N(\mathscr{A}, \nu)$ such that $s_0^k=i$, $s_1^k=j$, $\theta_0-\varepsilon<\phi_0^k<\theta_0$, and $\lim_{k\to\infty}\phi_0^k=\theta_0$. Then ξ_k converges to an element $\bar{\xi}=(\bar{s}, \bar{\phi})\in\mathscr{B}_N(\nu)$ with $\bar{\phi}_0=\theta_0$, but $\bar{\xi}\notin\mathscr{B}_N(\mathscr{A}, \nu)$ since $(A(\theta_0))_{\bar{s}_0\bar{s}_1}=0$. Thus, $\mathscr{B}_N(\mathscr{A}, \nu)$ is not closed and hence not compact.

We call an element $\xi\in\mathscr{B}_N(\mathscr{A}, \nu)$ an *orbit in* $\mathscr{B}_N(\mathscr{A}, \nu)$. Let us denote

$$A^k(\theta)=A(\theta)A(R_\nu(\theta))\cdots A(R_\nu^{k-1}(\theta)), \qquad k=1, 2, \cdots.$$

DEFINITION 4.5. An l-parameter family of transition matrices

$$\mathscr{A}=\{A(\theta)\in M_N \mid \theta\in\Theta\}$$

is called *irreducible* if for $\theta\in\Theta$ there is an integer $k>0$ such that

$$(A^k(\theta))_{ij}\neq 0$$

for all $i, j\in S_N(\theta)$.

If $\mathscr{A}$ is irreducible, then for any pair $i, j\in S_N(\theta)$ and some $k>0$ there is an orbit $\xi\in\mathscr{B}_N(\mathscr{A}, \nu)$ such that $\xi_0=(i, \theta)$ and $\xi_k=(\sigma_\nu^k(\xi))_0=(j, R_\nu^k(\theta))$. In order to describe the dynamics of the generalized Bernoulli shift, the following definition will be useful.

DEFINITION 4.6. We say that a finite or infinite sequence $\{s_i\}_{i=j_1}^{j_2}$ with $s_i\in S_N$, $i=j_1, \cdots, j_2$, is *admissible* for $\mathscr{B}_N(\mathscr{A}, \nu)$ if there is a point $\theta\in\Theta$ such that

$$(A(R_\nu^i(\theta)))_{s_is_{i+1}}\neq 0, \qquad i=j_1, \cdots, j_2-1. \tag{4.5}$$

In particular, when $s\in B_N$ is admissible for $\mathscr{B}_N(\mathscr{A}, \nu)$, we call the sequence s an *admissible orbit* for $\mathscr{B}_N(\mathscr{A}, \nu)$. We also say that a finite sequence $\{\xi_i\}_{i=j_1}^{j_2}$ with $\xi_i=(s_i, \phi_i)$ is *admissible* for $\mathscr{B}_N(\mathscr{A}, \nu)$ when $\{s_i\}_{i=j_1}^{j_2}$ satisfies (4.5) and $\{\phi_i\}_{i=j_1}^{j_2}$ satisfies (4.3).

Example 4.1. Consider the case in which $S_N=\{1,2\}$, $l=1$, $\nu=1$ and $\Theta=S^1$. Let $\mathscr{A}=\{A(\theta)\in M_N\,|\,\theta\in S^1\}$ be given by

$$A(\theta)=\begin{cases}\begin{pmatrix}1&0\\0&0\end{pmatrix} & \text{for } \theta\in[0,\pi-1),\\[2mm] \begin{pmatrix}0&1\\0&0\end{pmatrix} & \text{for } \theta\in[\pi-1,\pi),\\[2mm] \begin{pmatrix}0&0\\0&1\end{pmatrix} & \text{for } \theta\in[\pi,2\pi-1),\\[2mm] \begin{pmatrix}0&0\\1&0\end{pmatrix} & \text{for } \theta\in[2\pi-1,2\pi).\end{cases}$$

It is easy to show that $\mathscr{A}$ is irreducible. We also see that there is no periodic admissible orbit. If $\xi=(s,\phi)\in\mathscr{B}_N(\mathscr{A},1)$, then $s_i=1$ or 2, depending on whether ϕ_i $(=i+\phi_0 \bmod 2\pi)\in[0,\pi)$ or not, for $i\in\mathbb{Z}$. Hence, every orbit in $\mathscr{B}_N(\mathscr{A},\nu)$ can be completely determined by the value of ϕ_0, and is trivial.

Thus, we will require the following property for $\mathscr{A}$.

Definition 4.7. An l-parameter family of transition matrices

$$\mathscr{A}=\{A(\theta)\in M_N\,|\,\theta\in\Theta\}$$

is called *nontrivial* if card $(S_N(\theta))\geqq 2$ for any $\theta\in\Theta$.

The dynamics of the generalized Bernoulli shift $(\mathscr{B}_N(\mathscr{A},\nu),\sigma_\nu)$ are similar to those of the standard Bernoulli shift (B_N,σ) if $\mathscr{A}$ is nontrivial and irreducible. The generalized Bernoulli shift $(\mathscr{B}_N(\mathscr{A},\nu),\sigma_\nu)$ may have

(1) a countable set of periodic admissible orbits,

(2) an uncountable set of nonperiodic admissible orbits,

(3) a dense admissible orbit in an adequate meaning, although this is not the case in general (see Examples 4.1 and 4.4). We present two examples.

Example 4.2. Let A be an $N\times N$ irreducible transition matrix. Then, $\mathscr{A}=\{A(\theta)=A\,|\,\theta\in\Theta\}$ is irreducible for any rigid rotation R_ν, and nontrivial since $S_N(\theta)=S_N$ for all $\theta\in\Theta$. Since $s\in B_N$ is an admissible orbit for $\mathscr{B}_N(\mathscr{A},\nu)$ if and only if $s\in B_N(A)$, there exist (1), (2), and (3), where "dense" means "dense in $B_N(A)$."

Example 4.3. Let the ith element of ν have the form $\nu_i=2\pi(p_i/q_i)$ with a pair of relatively prime integers p_i, q_i for $i=1,\cdots,l$. Then, R_ν yields a rational flow in Θ and there is an integer k such that R_ν^k is an identical map. In particular, for any $\theta\in\Theta$ there are k different points $\phi_0=\theta,\phi_1,\cdots,\phi_{k-1}\in\Theta$ with $\phi_i=R_\nu(\phi_{i-1})$, $i=1,\cdots,k-1$, and $\phi_0=R_\nu(\phi_{k-1})$. Suppose that $\mathscr{A}=\{A(\theta)\in M_N\,|\,\theta\in\Theta\}$ be irreducible and nontrivial. Let $\mathscr{B}_N^\theta(\mathscr{A},\nu)$ be a subset of $\mathscr{B}_N(\mathscr{A},\nu)$ given by

$$\mathscr{B}_N^\theta(\mathscr{A},\nu)=\{\xi\in\mathscr{B}_N(\mathscr{A},\nu)\,|\,\phi_i=\theta \text{ for some integer } i\}.$$

If $\xi=\{\xi_j\}_{j=-\infty}^{\infty}\in\mathscr{B}_N^\theta(\mathscr{A},\nu)$, then $\xi_i^k=\{\xi_{i+kj}\}_{j=-\infty}^{\infty}\in\mathscr{B}_N^{\phi_i}(\mathscr{A},k\nu)$. Since $R_{k\nu}=R_\nu^k$ is the identical map, $\mathscr{B}_N^{\phi_i}(\mathscr{A},k\nu)=B_N(A^k(\phi_i))\times\prod_{j=-\infty}^{\infty}\{\phi_i\}$. Hence, every orbit in $B_N(A^k(\phi_i))$ is also an admissible orbit for $\mathscr{B}_N^{\phi_i}(\mathscr{A}^k,k\nu)$. Since this statement holds for ϕ_i, $i=0,\cdots,k-1$, and $\mathscr{A}$ is irreducible, there exist (1), (2), and (3) for $\mathscr{B}_N^\theta(\mathscr{A},\nu)$, where "dense" means "dense in the set of all admissible orbits for $\mathscr{B}_N^\theta(\mathscr{A},\nu)$."

We present an example which can be used as a model describing the stochastic behavior in some quasi-periodically forced oscillator as shown in § 5.

Example 4.4. Let $l=1$, $\nu=3$, and $N=3$. Let $\mathscr{A}=\{A(\theta)\in M_N \mid \theta\in S^1\}$ be given by

$$A(\theta)=\begin{cases} \begin{pmatrix} 1 & 1 & 1\\ 1 & 1 & 1\\ 0 & 0 & 0\end{pmatrix} & \text{for } \theta\in\left[1-\dfrac{1}{4}\pi, 1\right],\\[2ex] \begin{pmatrix} 1 & 0 & 1\\ 1 & 0 & 1\\ 1 & 0 & 1\end{pmatrix} & \text{for } \theta\in\left[\dfrac{7}{4}\pi-3, 2\pi-3\right],\\[2ex] \begin{pmatrix} 1 & 1 & 0\\ 1 & 1 & 0\\ 1 & 1 & 0\end{pmatrix} & \text{for } \theta\in\left[\dfrac{7}{4}\pi-2, 2\pi-2\right],\\[2ex] \begin{pmatrix} 1 & 1 & 1\\ 0 & 0 & 0\\ 1 & 1 & 1\end{pmatrix} & \text{for } \theta\in\left[\dfrac{7}{4}\pi, 2\pi\right],\\[2ex] \begin{pmatrix} 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix} & \text{for } \theta\in\left(0, 1-\dfrac{\pi}{4}\right)\cup\left(1, \dfrac{7}{4}\pi-3\right)\cup\left(2\pi-3, \dfrac{7}{4}\pi-2\right)\\ & \qquad\qquad\cup\left(2\pi-2, \dfrac{7}{4}\pi\right). \end{cases}$$

We can easily show that $\mathscr{A}$ is nontrivial and irreducible. Moreover, there is no periodic admissible orbit except $\{\cdots, 1, 1, 1, \cdots\}$.

In general, the generalized Bernoulli shift $(\mathscr{B}_N(\mathscr{A}, \nu), \sigma_\nu)$ exhibits stochastic behavior if $\mathscr{A}$ is nontrivial and irreducible, like the standard Bernoulli shift (B_n, σ): suppose that $\mathscr{A}$ is nontrivial and irreducible, and let $\theta\in\Theta$ be fixed. Then there is an integer $k>0$ such that for all $i, j\in S_N(\theta)$ there is an orbit $\xi\in\mathscr{B}_N(\mathscr{A}, \nu)$ with $\xi_0=(i, \theta)$ and $\xi_k=(j, R_\nu^k(\theta))$. Let $\bar{r}=\text{card }(S_N(\theta))\geqq 2$, and let $S_N(\theta)=\{j_r\in S_N, i=1, \cdots, \bar{r}\}$. Consider an orbit $\xi=\{s, \phi\}$ with $s_0=j_1$ and $\phi_0=\theta$. Then, even though the values of s_i and ϕ_i are known for all $i\leqq 0$, the value of s_k cannot be determined; it can be any of j_r, $r=1, \cdots, \bar{r}$. Thus, if $\mathscr{A}$ is nontrivial and irreducible, then we cannot completely predict the future $(i>0)$ for each orbit in the generalized Bernoulli shift from its past $(i\leqq 0)$, as in the Bernoulli shift.

5. Detection of chaos for diffeomorphisms. In this section we consider a C^r $(r\geqq 2)$ diffeomorphism $f:\mathbb{R}^n\times\mathbb{R}^m\times T^l\to\mathbb{R}^n\times\mathbb{R}^m\times T^l$ of the form

$$f(x, y, \theta)=(g_1(x, y, \theta), g_2(x, y, \theta), R_\nu(\theta)), \qquad (x, y, \theta)\in\mathbb{R}^n\times\mathbb{R}^m\times T^l, \tag{5.1}$$

where $g_1:\mathbb{R}^n\times\mathbb{R}^m\times T^l\to\mathbb{R}^n$, $g_2:\mathbb{R}^n\times\mathbb{R}^m\times T^l\to\mathbb{R}^m$ are C^r and $R_\nu: T^l\to T^l$ is a rigid rotation with $\nu\in T^l$ (see (4.2)). The Poincaré map $P_{\varepsilon,i}$ for (2.1) has the form (5.1) with $m=n=1$. We assume that f has a C^r normally hyperbolic invariant l-torus of $(1, \cdots, 1)$-cycle $\mathscr{T}$ with C^r, $n+l$-dimensional stable manifold $W^s(\mathscr{T})$, and $m+l$-dimensional unstable manifold $W^u(\mathscr{T})$. Furthermore, $\mathscr{T}$ is assumed to be the graph of a function $z: T^l\to\mathbb{R}^n\times\mathbb{R}^m$, i.e., $\mathscr{T}=\{(x, y, \theta)\in\mathbb{R}^n\times\mathbb{R}^m\times T^l \mid (x, y)=z(\theta), \theta\in T^l\}$. These assumptions are satisfied for $P_{\varepsilon,i}$ when ε is sufficiently small.

We begin with a lemma concerning rigid rotations.

LEMMA 5.1. *For any $\varepsilon>0$ there is an integer $k>0$ such that*

$$|R_\nu^k(\theta)-\theta|<\varepsilon \quad \text{for } \theta\in T^l. \tag{5.2}$$

Proof. If for some integers $k_1 > k_2$,

$$R_\nu^{k_1}(\theta) = R_\nu^{k_2}(\theta), \tag{5.3}$$

then we have

$$R_\nu^{k_1-k_2}(\theta) = \theta,$$

which implies (5.2) with $k = k_1 - k_2$. Let us assume that there is not a pair of integers (k_1, k_2) such that (5.3) holds. Then, for $\theta \in T^l$ fixed, $R_\nu^j(\theta)$, $j \in \mathbb{Z}$ are different from each other. The sequence $\{R_\nu^j(\theta), j \in \mathbb{Z}\}$ consists of infinitely many points and hence has an accumulation point since T^l is compact. Thus for any $\varepsilon < 0$ there are two integers $k_1 > k_2$ such that

$$|R_\nu^{k_1}(\theta) - R_\nu^{k_2}(\theta)| < \varepsilon.$$

Since R_ν is an area-preserving map, we obtain (5.2) with $k = k_1 - k_2$. □

Remark 5.1. Lemma 5.1 is a special case of Poincaré's recurrence theorem (see Abraham and Marsden [1978] or Arnold [1989]).

Remark 5.2. If the components of the frequency vector ν are incommensurable, i.e.,

$$(\kappa, \nu) \notin 2\pi\mathbb{Z} \quad \text{for } \kappa \in \mathbb{Z}^l \backslash \{0\}, \tag{5.4}$$

then we have a stronger statement; for any $\varepsilon > 0$ and θ, $\theta' \in T^l$, there is an integer $k > 0$ such that

$$|R_\nu^k(\theta) - \theta'| < \varepsilon, \tag{5.5}$$

i.e., each orbit of R_ν is dense in T^l. See § 3 of Cornfeld, Fomin, and Sinai [1982] for a proof.

From Remark 5.2, we obtain the following.

LEMMA 5.2. *Let $\Theta \subset T^l$ be an l-dimensional manifold. Then we have*

$$\bigcup_{j=0}^{\infty} R_\nu^j(\Theta) = \bigcup_{j=0}^{\infty} R_\nu^{-j}(\Theta) \equiv \tilde{\Theta}. \tag{5.6}$$

Moreover, there is an integer N_0 such that if $N \geqq N_0$, then for any integer j_0,

$$\bigcup_{j=0}^{N} R_\nu^{j+j_0}(\Theta) = \bigcup_{j=0}^{N} R_\nu^{-j+j_0}(\Theta) = \tilde{\Theta}. \tag{5.7}$$

Proof. First, let us assume that ν satisfies (5.4). It follows from Remark 5.2 that we have (5.6) with $\tilde{\Theta} = T^l$. Let $c_\varepsilon(\theta)$ be the l-dimensional cube with sides of length ε centered at θ. For $N > 0$ sufficiently large, Θ contains a cube $c_{4\pi/N}(\theta_0)$ with $\theta_0 \in \Theta$. Let N^l cubes $c_{2\pi/N}(\theta_j)$, $j = 1, \cdots, N^l$, cover T^l. From Remark 5.2 we see that there exist integers $k_j > 0$, $j = 1, \cdots, N^l$, such that

$$|R_\nu^{k_j}(\theta_0) - \theta_j| < \frac{2\pi}{N}.$$

Hence, letting $N_0 = \max k_j$, we obtain (5.7).

Next, let us assume that condition (5.4) does not hold. Then, by permuting the components of ν if necessary, ν can be written as

$$\nu = (\nu^a, \nu^b),$$

where $\nu^a = (\nu_1^a, \cdots, \nu_{l_a}^a)$ satisfies (5.4) with $\nu = \nu^a$ and $l = l_a$, and $\nu^b = (\nu_1^b, \cdots, \nu_{l_b}^b)$ has the form

$$\nu_i^b = \sum_{j=1}^{l_a} \frac{q_{ij}}{p_{ij}} \nu_j^a + 2\pi \frac{q_i}{p_i}, \qquad i = 1, \cdots, l_b,$$

with $p_{ij}, q_{ij}, p_i, q_j \in \mathbb{Z}$, and both (p_{ij}, q_{ij}) and (p_i, q_i) are relatively prime. For $\theta \in T^l$, we write $\theta = (\theta^a, \theta^b)$, where $\theta^a \in T^{l_a}$ and $\theta^b \in T^{l_b}$. Let π^a denote the projection from $T^l = T^{l_a} \times T^{l_b}$ upon T^{l_a}. Let p_0 be the least common multiple for p_i, $i = 1, \cdots, l_b$, and let

$$\Theta_r \equiv R_\nu^r(\Theta) = \left\{ (\theta^a, \theta^b) \,|\, \theta_i^b = \frac{q_{ij}}{p_{ij}}(\theta_j^a - \theta_{j0}^a) + 2\pi r \frac{q_i}{p_i} + \theta_{i0}^b, \right.$$
$$\left. \theta^a \in \pi^a(R_\nu^r(\Theta)), (\theta_0^a, \theta_0^b) \in \Theta \right\}, \qquad r \in \mathbb{Z}.$$

Since the components of ν^a are incommensurable, there is an integer N_1 such that for any $r \in \mathbb{Z}$,

$$\bigcup_{j=0}^{N_1} \pi^a(R_\nu^{jp_0}(\Theta_r)) = \bigcup_{j=0}^{N_1} \pi^a(R_\nu^{-jp_0}(\Theta_r)) = T^{l_a}.$$

On the other hand, for any j, $r \in \mathbb{Z}$,

$$R_\nu^{jp_0}(\Theta_r) = \left\{ (\theta^a, \theta^b) \,|\, \theta_i^b = \frac{q_{ij}}{p_{ij}}(\theta_j^a - \theta_{j0}^a) + 2\pi r \frac{q_i}{p_i} + \theta_{i0}^b, \right.$$
$$\left. \theta^a \in \pi^a(R_\nu^{jp_0}(\Theta_r)), (\theta_0^a, \theta_0^b) \in \Theta \right\},$$

since $2\pi p_0 q_i / p_i = 0 \bmod 2\pi$. Thus we have

$$\bar{\Theta}_r \equiv \bigcup_{j=0}^{N_1} R_\nu^{jp_0}(\Theta_r) = \left\{ (\theta^a, \theta^b) \,|\, \theta_i^b = \frac{q_{ij}}{p_{ij}}(\theta_j^a - \theta_{j0}^a) + 2\pi r \frac{q_i}{p_i} + \theta_{i0}^b, \right.$$
$$\left. \theta^a \in T^{l_a}, (\theta_0^a, \theta_0^b) \in \Theta \right\}, \qquad r = 1, \cdots, p_0 - 1.$$

It is clear that

$$\Theta_r \subset \bigcup_{j=0}^{p_0-1} \bar{\Theta}_j \quad \text{for any } r \in \mathbb{Z}.$$

Hence, we have

$$\bigcup_{j=0}^{\infty} R_\nu^j(\Theta) = \bigcup_{j=0}^{\infty} R_\nu^{-j}(\Theta) = \bigcup_{r=0}^{p_0-1} \bar{\Theta}_r,$$

and

$$\bigcup_{j=0}^{N_1(p_0-1)} R_\nu^j(\Theta) = \bigcup_{j=0}^{N_1(p_0-1)} R_\nu^{-j}(\Theta) = \bigcup_{r=0}^{p_0-1} \bar{\Theta}_r.$$

Letting $\tilde{\Theta} = \bigcup_{r=0}^{p_0-1} \bar{\Theta}_r$ and $N_0 = N_1(p_0 - 1)$, we have (5.6) and (5.7). □

We now state the main theorem of this section. Let π_θ be the projection from $\mathbb{R}^n \times \mathbb{R}^m \times T^l$ upon T^l.

THEOREM 5.3. *Suppose that $W^s(\mathcal{T})$ and $W^u(\mathcal{T})$ intersect transversely in an l-dimensional manifold γ with* $\dim \pi_\theta(\gamma)=l$. *Then there exist an integer $N\geqq 2$, an l-dimensional manifold $\tilde{\Theta}\subset T^l$, and a nontrivial and irreducible l-parameter family of transition matrices $\mathcal{A}=\{A(\theta)\in M_N \mid \theta\in\tilde{\Theta}\}$ such that for some $k\geqq 1$, f^k has an invariant set Λ on which it is topologically conjugate to the generalized Bernoulli shift $\sigma_\nu:\mathcal{B}_N(\mathcal{A},k\nu)\to\mathcal{B}_N(\mathcal{A},k\nu)$.*

Proof. The proof is similar to that of the Smale-Birkhoff homoclinic theorem (see Smale [1963] and Newhouse [1980]).

Without loss of generality, we can assume that $\mathcal{T}$ is given by

$$\mathcal{T}=\{(0,0,\theta)\in\mathbb{R}^n\times\mathbb{R}^m\times T^l \mid \theta\in T^l\},$$

and that in a neighborhood U of $\mathcal{T}$, $W^s(\mathcal{T})$ and $W^u(\mathcal{T})$ are locally straightened, i.e.,

$$W^s(\mathcal{T})\cap U=\{(x,0,\theta)\in U \mid \theta\in T^l\},$$

and

$$W^u(\mathcal{T})\cap U=\{(0,y,\theta)\in U \mid \theta\in T^l\}.$$

See Wiggins [1988b, pp. 322-325].

Let $\Theta=\pi_\theta(\gamma)$, and let

$$\tilde{\Theta}=\bigcup_{j=-\infty}^{\infty} R_\nu^j(\Theta).$$

From Lemma 5.2 we see that there is an integer N_0 such that

$$\bigcup_{j=0}^{N_0} R_\nu^j(\Theta)=\tilde{\Theta}.$$

Since $\tilde{\Theta}$ is invariant by R_ν, $\mathbb{R}^2\times\tilde{\Theta}$ is also invariant by f. Hence, we have only to consider the restriction of f on $\mathbb{R}^2\times\tilde{\Theta}$, which is denoted by $\tilde{f}$. Thus $\tilde{f}$ is a C^r diffeomorphism from $\mathbb{R}^2\times\tilde{\Theta}$ onto $\mathbb{R}^2\times\tilde{\Theta}$. It is clear that $\tilde{\mathcal{T}}=\{(x,y,\theta)\in\mathcal{T} \mid \theta\in\tilde{\Theta}\}$ is a normally hyperbolic invariant manifold for $\tilde{f}$. For $j=0,\cdots,N_0$, let

$$\Theta_j=R_\nu^j(\Theta),$$

and let

$$\gamma_j=\tilde{f}^j(\gamma).$$

Since $\lim_{p\to\infty}\tilde{f}^p\gamma\subset\tilde{\mathcal{T}}$, we can assume that γ is close to $\tilde{\mathcal{T}}$ and hence $\gamma\subset U$. Replacing γ with a subset of γ if necessary, we can assume the following:

(H1) Θ is an open, simply connected, l-dimensional submanifold of T^l, and γ is the graph of a C^r function $x_\gamma:\Theta\to\mathbb{R}^n$, i.e.,

$$\gamma=\{(x,0,\theta)\in\mathbb{R}^n\times\mathbb{R}^m\times T^l \mid x=x_\gamma(\theta),\ \theta\in\Theta\}.$$

(H2) $\gamma_i\cap\gamma_j=\phi$ for $i\neq j$.

It follows from (H1) that there are C^r functions $x_j:\Theta_j\to\mathbb{R}^n, j=0,\cdots,N_0$, such that

$$\gamma_j=\{(x,0,\theta)\in\mathbb{R}^n\times\mathbb{R}^m\times T^l \mid x=x_j(\theta),\ \theta\in\Theta_j\}.$$

We also see that for any $0<N'<N_0$ and $0\leqq j_r\leqq N_0$, $r=1,\cdots,N'$, the intersection $\bigcap_{r=1}^{N'}\Theta_{j_r}$ is a nonempty open subset of $\tilde{\Theta}$ if it is nonempty.

Let $I^s \subset \mathbb{R}^n$ be a closed n-disk such that $D^s = I^s \times \tilde{\Theta}$ is an $n+l$-dimensional manifold in $W^s(\tilde{\mathcal{T}})$, and $\tilde{\mathcal{T}}$, $\bigcup_{j=0}^{N_0} \gamma_j \subset (I^s - \partial I^s) \times \tilde{\Theta}$. For $\delta > 0$ small, $N_\delta = \{(x, y, \theta) \in \mathbb{R}^n \times \mathbb{R}^m \times T^l \mid x \in I^s, y \in [-\delta, \delta], \theta \in \Theta\}$ is a neighborhood of D^s and contains $\tilde{\mathcal{T}}$ and $\bigcup_{j=0}^{N_0} \gamma_j$. See Fig. 4. If δ is sufficiently small, then $\tilde{f}^k(N_\delta)$ accumulates along $W^u(\tilde{\mathcal{T}})$ when $k \to \infty$, as shown in Fig. 5. Hence, we can choose k large such that $N_\delta \cap \tilde{f}^k(N_\delta)$ contains $\gamma_j, j = 1, \cdots, N_0$, and has $N_0 + 2$ connected components containing $\tilde{\mathcal{T}}$ or γ_j, $j = 0, 1, \cdots, N_0$. Let us denote the connected component containing $\tilde{\mathcal{T}}$ by V_1 and those containing γ_j by V_{j+2} for $j = 0, \cdots, N_0$. See Fig. 6.

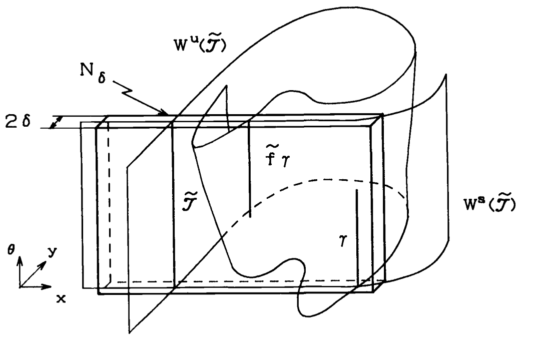


FIG. 4. *The set* N_δ.

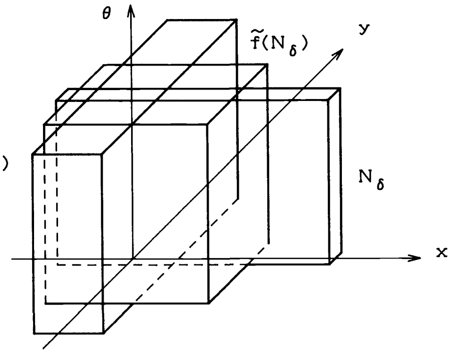


FIG. 5. N_δ, $\tilde{f}(N_\delta)$, *and* $\tilde{f}^2(N_\delta)$.

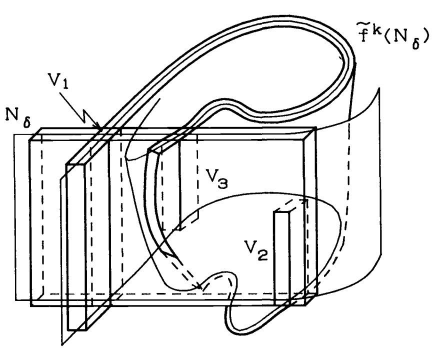


FIG. 6. *Construction of* V_i, $i = 1, 2, 3$.

Let us set $N=N_0+2$. For each $\theta\in\tilde{\Theta}$, define a $N\times N$ transition matrix $A(\theta)$ as follows: if (1) $i=j=1$, (2) $i=1$ and $\theta\in R_{k\nu}^{-1}(\Theta_{j-2})$ for $j\geqq 2$, (3) $j=1$ and $\theta\in\Theta_{i-2}$ for $i\geqq 2$, or (4) $\theta\in\Theta_{i-2}\cap R_{k\nu}^{-1}(\Theta_{j-2})$ for $i,j\geqq 2$, then

$$(A(\theta))_{ij}=1,$$

and, otherwise,

$$(A(\theta))_{ij}=0.$$

We notice that if $l=1$, $N_0=1$, $\tilde{\Theta}=S^1$, $\nu=1$, $k=3$, and $\Theta=(0,(7/4)\pi)$, then we have the one parameter family of transition matrices $\mathscr{A}$ in Example 4.4. Since

$$S_N(\theta)=\{2\leqq i\leqq N \mid \theta\in\Theta_{i-2}\}\cup\{1\}$$

and

$$S'_N(\theta)=\{2\leqq j\leqq N \mid \theta\in R_{k\nu}^{-1}(\Theta_{j-2})\}\cup\{1\},$$

we have $S_N(R_{k\nu}(\theta))=S'_N(\theta)$ and card $(S_N(\theta))\geqq 2$ for any $\theta\in\tilde{\Theta}$. Hence, the l-parameter family of transition matrices $\mathscr{A}=\{A(\theta)\in M_N \mid \theta\in\tilde{\Theta}\}$ is consistent with $R_{k\nu}$ and nontrivial. Let $S_N(\theta)=\{1,j_1,\cdots,j_{N'}\}$ for $\theta\in\tilde{\Theta}$ fixed. Since $\Theta_{j_1\cdots j_{N'}}\equiv\bigcap_{r=1}^{N'}\Theta_{j_r-2}(\ni\theta)$ is a nonempty open subset of $\tilde{\Theta}$ by (H1), it follows from Lemma 5.1 that there is an integer $K>0$ such that $R_{k\nu}^K(\theta')\in\Theta_{j_1\cdots j_{N'}}$. Thus,

$$(A^K(\theta))_{ij}\geqq(A(\theta))_{i1}(A^{K-2}(R_{k\nu}(\theta)))_{11}(A(R_{k\nu}^{K-1}(\theta)))_{1j}\geqq 1,$$

for $i,j=1,j_1,\cdots,j_{N'}$, and hence $\mathscr{A}$ is also irreducible.

Let

$$V_{i,\theta}=\{(x,y,\theta)\in\mathbb{R}^n\times\mathbb{R}^m\times T^l \mid (x,y,\theta)\in V_i\},\qquad \text{for } \theta\in\Theta_i \text{ and } i\in S_N(\theta).$$

By (H2) we can choose δ small such that $V_{i,\theta}\cap V_{j,\theta}=\phi$ for $i\neq j\in S_N(\theta)$. For any admissible sequence $\{\xi_{-j},\cdots,\xi_j\}$ for $\mathscr{B}_N(\mathscr{A},k\nu)$, let

$$V_{\xi_{-j}\cdots\xi_j}=\bigcap_{i=-j}^{j}\tilde{f}^{-ik}(V_{\xi_i}). \tag{5.8}$$

Noting (H1), we can prove that there are constants $C>0$ and $\lambda>1$ such that $V_{\xi_{-j}\cdots\xi_j}$ is an $m+n$-dimensional ball of diameter less than $C\lambda^{-j}$, as in the proof of the Smale-Birkhoff homoclinic theorem (e.g., Newhouse [1980]). Hence, for $\xi\in\mathscr{B}_N(\mathscr{A},k\nu)$, V_ξ is a single point. Let

$$\Lambda=\bigcap_{j\in\mathbb{Z}}\tilde{f}^{jk}\left(\bigcup_{i=1}^{N}V_i\right).$$

We define a map $h:\mathscr{B}_N(\mathscr{A},k\nu)\to\Lambda$ as

$$h(\xi)=V_\xi.$$

Since $V_\xi\neq V_{\xi'}$ for $\xi\neq\xi'\in\mathscr{B}_N(\mathscr{A},k\nu)$, h is a $1-1$ map of $\mathscr{B}_N(\mathscr{A},k\nu)$ onto Λ. It follows from (5.8) that h and h^{-1} are continuous when $\mathscr{B}_N(\mathscr{A},\nu)$ has the metric (4.4). Thus, h is a homeomorphism. Furthermore, $\tilde{f}^{jk}(h(\xi))\in V_{\xi_j}$ for all j, and hence

$$\tilde{f}^{jk}(\tilde{f}^k(h(\xi)))\in V_{\xi_{j+1}}=V_{(\sigma_{k\nu}(\xi))_j}.$$

Thus $\tilde{f}^k(h(\xi))=h(\sigma_{k\nu}(\xi))$ so that h conjugates $\sigma_{k\nu}$ with $\tilde{f}|_\Lambda$. This completes the proof. □

Remark 5.3. Suppose that $W^s(\mathcal{T}_{\varepsilon,i})$ and $W^u(\mathcal{T}_{\varepsilon,i})$ intersect transversely in a torus of $(1, \cdots, 1)$-cycle satisfying (H1). Then we can take $N = 2$ in Theorem 5.3. Moreover,

$$A(\theta) = A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{for all } \theta \in T^l,$$

and $\mathcal{B}_2(\mathcal{A}, \nu) \cong B_2 \times T^l$. Thus, the dynamics of f on Λ can be described by the standard Bernoulli shift, as stated in Theorem 3.1.

Remark 5.4. The Smale-Birkhoff homoclinic theorem was stated in different forms in Smale [1963], Moser [1973, § 3], and Guckenheimer and Holmes [1983, § 5]. Hence, its generalization to a class of maps (5.1) can have some different versions. In Theorem 5.3 we adopted the formulation corresponding to that of Smale [1963].

6. Detection of chaos in quasi-periodically forced oscillators. We now return to the Poincaré map $P_{\varepsilon,i} : \Sigma_i \to \Sigma_i$ of the quasi-periodically forced oscillator (2.1).

Suppose that (M1) and (M2) are satisfied at $\theta_i = \omega_i t_0$, $\bar{\theta}^i = \bar{\omega}^i t_0 + \bar{\theta}_0^i \bmod 2\pi$. Then, the zero of the Melnikov function M can be continued to an $l-1$-dimensional manifold in T^l, as shown in § 2. Moreover, the stable and unstable manifolds $W^s(\mathcal{T}_{\varepsilon,i})$, $W^u(\mathcal{T}_{\varepsilon,i})$ of the normally hyperbolic invariant $(l-1)$-torus $\mathcal{T}_{\varepsilon,i}$ intersect transversely in an $l-1$-dimensional manifold $\gamma_{\varepsilon,i}$ containing a point near

$$(x, \bar{\theta}^i) = (\bar{x}_0(-t_0), \bar{\theta}_0^i) \in \Sigma_i. \tag{6.1}$$

We can denote $\gamma_{\varepsilon,i}$ by

$$\gamma_{\varepsilon,i} = \{\bar{x}_0(-t), \bar{\theta}^i) \mid t = \bar{h}_0(s), \theta_j = \bar{h}_j(s) \bmod 2\pi, s \in \tilde{U}\} + \mathcal{O}(\varepsilon), \tag{6.2}$$

where $\bar{h}_j : \tilde{U} \to \mathbb{R}$, $0 \leqq j(\neq i) \leqq l$ are C^r and satisfy $\bar{h}_0(0) = t_0$ and $\bar{h}_j(0) = \theta_{j0}$ with some neighborhood $\tilde{U}$ of 0 in $\mathbb{R}^{l-1}$. Moreover, (M1) and (M2) are satisfied at $\theta_i = \omega_i \bar{h}_0(s)$, $\theta_j = \bar{h}_j(s) + \omega_j \bar{h}_0(s) \bmod 2\pi$. Thus we have

$$\frac{\partial}{\partial t} M(\tilde{\theta}(t_0, \theta_0)) \frac{\partial}{\partial s_r} \bar{h}_0(0) + \sum_{j \neq i} \frac{\partial}{\partial \theta_j} M(\tilde{\theta}(t_0, \theta_0)) \frac{\partial}{\partial s_r} \bar{h}_j(0) = 0, \qquad r = 1, \cdots, l-1, \tag{6.3}$$

where $s = (s_1, \cdots, s_{l-1})$ and

$$\tilde{\theta}(t, \theta) = (\theta_1 + \omega_1 t, \cdots, \theta_{i-1} + \omega_{i-1} t, \omega_i t, \theta_{i+1} + \omega_{i+1} t, \cdots, \theta_l + \omega_l t).$$

From (6.2) we see that $\gamma_{\varepsilon,i}$ is tangent to $l-1$ vectors

$$t_{\varepsilon,i}^r = \left(\frac{\partial}{\partial s_r} \bar{h}_0(0) JDH(\bar{x}(-t_0)), \frac{\partial}{\partial s_r} \bar{h}_1(0), \cdots, \frac{\partial}{\partial s_r} \bar{h}_{i-1}(0), \right.$$
$$\left. \frac{\partial}{\partial s_r} \bar{h}_{i+1}(0), \cdots, \frac{\partial}{\partial s_r} \bar{h}_l(0) \right) + \mathcal{O}(\varepsilon), \qquad r = 1, \cdots, l-1,$$

at a point near (6.1). Hence, if

$$\frac{\partial}{\partial t} M(\tilde{\theta}(t_0, \theta_0)) = \sum_{j=1}^{l} \omega_j \frac{\partial}{\partial \theta_j} M(\tilde{\theta}(t_0, \theta_0)) \neq 0, \tag{6.4}$$

then $\dim \pi_\theta(\gamma_{\varepsilon,i}) = l-1$ since $DH(\bar{x}(-t_0)) \neq 0$ holds and (6.3) does not hold when $\partial \bar{h}_0(0)/\partial s_r \neq 0$ and $\partial \bar{h}_j(0)/\partial s_r = 0$ for all $j \neq i$.

Noting that (6.4) implies (M2) at $\theta = \tilde{\theta}(t_0, \theta_0)$ and applying Theorem 5.3 to the Poincaré map $P_{\varepsilon,i}$ of (2.1), we obtain the following theorem.

THEOREM 6.1. *Suppose that there exists a point $\theta_0 \in T^l$ satisfying* (M1) *and*

$$\text{(M3)} \qquad \sum_{j=1}^{l} \omega_j \frac{\partial}{\partial \theta_j} M(\theta_0) \neq 0.$$

Then for some $k \geqq 1$, $P^k_{\varepsilon,i}$ has an invariant set Λ on which it is topologically conjugate to the generalized Bernoulli shift on a finite set of symbols.

Remark 6.1. For the existence of chaotic solutions in almost periodically forced systems, Scheurle [1986] obtained conditions similar to those of Theorem 6.1, using a functional analytic method. From the proof of Theorem 5.3 we see that our result also gives a geometrical interpretation of his result, although it is limited to quasi-periodically forced systems. See Theorem A.2 of Appendix A.

7. An illustrative example. We consider a two-frequency perturbation of Duffing's equation

$$\text{(7.1)} \qquad \begin{aligned} &\dot{x} = y, \qquad \dot{y} = x - x^3 + \varepsilon(\gamma_1 \cos \theta_1 + \gamma_2 \cos \theta_2 - \delta y), \\ &\dot{\theta}_1 = \omega_1, \qquad \dot{\theta}_2 = \omega_2, \end{aligned}$$

where δ, γ_i, $\omega_i > 0$, $i = 1, 2$, and $0 < \varepsilon \ll 1$. The phase space of (7.1) is $\mathbb{R}^2 \times T^2$. See Yagasaki [1991a] for more detailed analyses. Chaotic motions near resonant tori are also described there.

When $\varepsilon = 0$, the (x, y)-component of (7.1) reduces to a planar Hamiltonian system

$$\dot{x} = y, \qquad \dot{y} = x - x^3,$$

which has a hyperbolic saddle at $(0, 0)$ and a symmetric pair of homoclinic orbits

$$\text{(7.2)} \qquad (x_\pm(t), y_\pm(t)) = \pm(\sqrt{2} \operatorname{sech} t, -\sqrt{2} \operatorname{sech} t \tanh t).$$

By substituting (7.2) into (2.3) and integrating the resulting equations, the Melnikov functions $M_\pm(t)$ for (7.2) become

$$\text{(7.3)} \qquad \begin{aligned} M_\pm(\theta_1, \theta_2) = &\pm\sqrt{2}\pi\omega_1\gamma_1 \operatorname{sech}\left(\frac{\pi\omega_1}{2}\right) \sin \theta_1 \\ &\pm\sqrt{2}\pi\omega_2\gamma_2 \operatorname{sech}\left(\frac{\pi\omega_2}{2}\right) \sin \theta_2 - \frac{4}{3}\delta. \end{aligned}$$

We see that there exists a point $(\theta_{10}, \theta_{20})$ satisfying (M1) and (M3) if

$$\text{(7.4)} \qquad \delta < \frac{3\sqrt{2}\pi}{4}\left[\omega_1\gamma_1 \operatorname{sech}\left(\frac{\pi\omega_1}{2}\right) + \omega_2\gamma_2 \operatorname{sech}\left(\frac{\pi\omega_2}{2}\right)\right].$$

By Theorem 6.1, (7.4) gives the region in parameter space $(\delta, \gamma_1, \gamma_2, \omega_1, \omega_2)$ where chaos may occur. Note that (7.4) is also the necessary condition for (M1) and (M2) to hold. Thus, in this example, transverse intersection between the stable and unstable manifolds of a normally hyperbolic invariant torus implies the existence of chaotic dynamics.

We next apply the result of Wiggins [1988b] (cf. Theorem 3.2). Suppose that

$$\text{(7.5)} \qquad \delta < \frac{3\sqrt{2}\pi}{4}\left|\omega_1\gamma_1 \operatorname{sech}\left(\frac{\pi\omega_1}{2}\right) - \omega_2\gamma_2 \operatorname{sech}\left(\frac{\pi\omega_2}{2}\right)\right|.$$

Note that if (7.5) holds, then (7.4) also holds. Moreover, let us assume that

$$\omega_1\gamma_1 \operatorname{sech}\left(\frac{\pi\omega_1}{2}\right) > \omega_2\gamma_2 \operatorname{sech}\left(\frac{\pi\omega_2}{2}\right).$$

Then

$$\frac{1}{\omega_1\gamma_1}\cosh\left(\frac{\pi\omega_1}{2}\right)\left[\frac{4\delta}{3\sqrt{2}\,\pi}+\omega_2\gamma_2\,\mathrm{sech}\left(\frac{\pi\omega_2}{2}\right)\right]<1.$$

Hence, we can define a periodic function $h_1(s)$, given by

$$h_1(s)=\sin^{-1}\left(\frac{1}{\omega_1\gamma_1}\cosh\left(\frac{\pi\omega_1}{2}\right)\left[\pm\frac{4\delta}{3\sqrt{2}\pi}-\omega_2\gamma_2\,\mathrm{sech}\left(\frac{\pi\omega_2}{2}\right)\sin s\right]\right).$$

Here the range of $\sin^{-1}$ is $[-\pi/2, \pi/2]$. We see that $h_1(s)$ satisfies the hypothesis of Theorem 3.2 when $i=1$. Similarly, if

$$\omega_1\gamma_1\,\mathrm{sech}\left(\frac{\pi\omega_1}{2}\right)<\omega_2\gamma_2\,\mathrm{sech}\left(\frac{\pi\omega_2}{2}\right),$$

then we can define a periodic function $h_2(s)$, given by

$$h_2(s)=\sin^{-1}\left(\frac{1}{\omega_2\gamma_2}\cosh\left(\frac{\pi\omega_2}{2}\right)\left[\pm\frac{4\delta}{3\sqrt{2}\pi}-\omega_1\gamma_1\,\mathrm{sech}\left(\frac{\pi\omega_1}{2}\right)\sin s\right]\right),$$

which satisfies the hypothesis of Theorem 3.2 when $i=2$.

On the other hand, let us assume that (7.4) holds but (7.5) does not. Then

$$-1<\frac{1}{\omega_i\gamma_i}\cosh\left(\frac{\pi\omega_i}{2}\right)\left[\frac{4\delta}{3\sqrt{2}\pi}-\omega_j\gamma_j\,\mathrm{sech}\left(\frac{\pi\omega_j}{2}\right)\right]<1,\qquad i=1,2,$$

where if $i=1$ then $j=2$ and otherwise $j=1$, and we set

$$a_1=\cos^{-1}\left(\frac{1}{\omega_1\gamma_1}\cosh\left(\frac{\pi\omega_1}{2}\right)\left[\frac{4\delta}{3\sqrt{2}\pi}-\omega_2\gamma_2\,\mathrm{sech}\left(\frac{\pi\omega_2}{2}\right)\right]\right),$$

where the range of $\cos^{-1}$ is $[0, \pi]$. Let

$$h_1'(s)=a_1\sin s\pm\frac{\pi}{2},$$

and let

$$h_2'(s)=\begin{cases}h_2(h_1'(s)), & s\in\left(0,\dfrac{\pi}{2}\right]\cup\left(\dfrac{3}{2}\pi,2\pi\right],\\[2ex] \pm\pi-h_2(h_1'(s)), & s\in\left(\dfrac{\pi}{2},\dfrac{3}{2}\pi\right].\end{cases}$$

Then the zero set of $M_\pm$ is given by

$$\tau_0=\{(\theta_1,\theta_2)\in T^2\mid\theta_1=h_1'(s),\ \theta_2=h_2'(s),\ s\in(0,2\pi]\}.$$

This is a special case of (3.5). Hence, we can only take a 1-torus of 0-cycle as the transverse homoclinic manifold for $P_{\varepsilon,1}$ and $P_{\varepsilon,2}$. Thus (7.5) represents a condition for the existence of chaos given by Theorem 3.2.

Figure 7 shows the regions given by (7.4) and (7.5) in $\delta-\gamma_1$ plane for fixed γ_2, ω_1, and ω_2. In this figure

$$\bar{\gamma}_1=\frac{\omega_2}{\omega_1}\gamma_2\cosh\left(\frac{\pi\omega_1}{2}\right)\mathrm{sech}\left(\frac{\pi\omega_2}{2}\right),$$

$$\bar{\delta}=\frac{3\sqrt{2}}{4}\pi\omega_2\gamma_2\,\mathrm{sech}\left(\frac{\pi\omega_2}{2}\right),$$

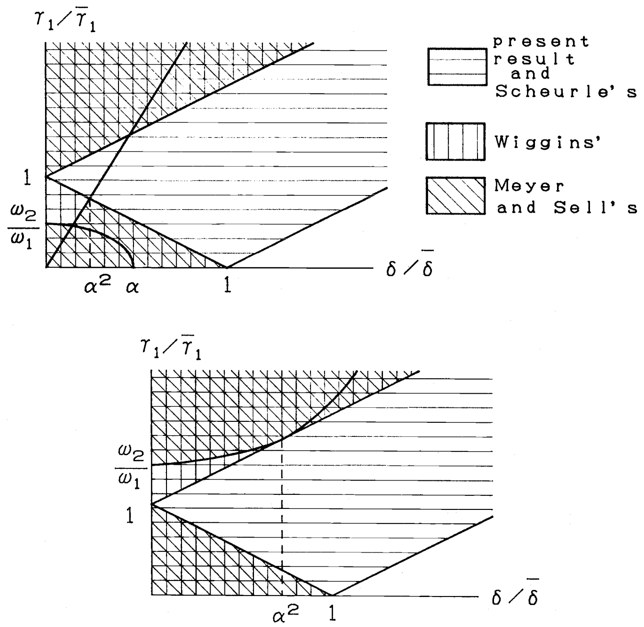


FIG. 7. *Regions in $\delta-\gamma_1$ plane for the existence of chaos in* (7.1), *with fixed γ_2, ω_1, and ω_2.* (a) $\omega_1>\omega_2$. (b) $\omega_1<\omega_2$.

and

$$\alpha=\frac{1}{\omega_1}\sqrt{|\omega_1^2-\omega_2^2|}\,.$$

We also show the regions given by the result of Scheurle [1986] and Meyer and Sell [1989] (see Appendix A). Note that Scheurle [1986] only provided a condition for the existence of random-like solutions (see Remarks A.1), although his condition gives the same regions as ours.

We close this paper with a remark on the extension of our result to a more general class of systems with frequencies depending on the state variables.

Consider systems of the form

$$\begin{aligned}&\dot{x}=JDH(x)+\varepsilon g(x,\theta),\\ &\dot{\theta}=\omega+\varepsilon G(x,\theta),\qquad (x,\theta)\in\mathbb{R}^2\times T^l,\quad \omega\in\mathbb{R}^l,\end{aligned}\tag{7.6}$$

where $G:\mathbb{R}^2\times T^l\to\mathbb{R}^l$ are C^r and $0<\varepsilon\ll 1$. We assume that condition (A1) of § 2 is satisfied, so that there exists a normally hyperbolic invariant torus $T_\varepsilon=T_0+\mathcal{O}(\varepsilon)$ whose stable and unstable manifolds $W^s(T_\varepsilon)$, $W^u(T_\varepsilon)$ are close to the unperturbed homoclinic manifold $\Gamma\times T^l$. In this case, T_ε may be subjected to phase locking.

Let us assume that the invariant torus T_ε is not subjected to phase locking. If G is independent of x, then we can show that transverse intersection between the stable and unstable manifolds of the invariant torus yields chaotic dynamics in (7.6), modifying the arguments given in § 5. It is natural to conjecture that this is the case in the general systems. However, if T_ε is subjected to phase locking, then we cannot immediately determine whether or not chaotic dynamics may occur since we do not have statements similar to those of Lemma 5.2 in general, and hence such arguments as given in § 5 do not apply. In a forthcoming paper (Yagasaki [1991b]) we will consider a perturbation of Duffing's equation with two frequencies depending on the state variables and perform the necessary analyses to detect the existence of chaos.

Appendix A. Other versions of Melnikov's method. Scheurle [1986] and Meyer and Sell [1989] extended Melnikov's method to study almost periodically forced systems. In this appendix we outline their results in the context of quasi-periodically forced oscillators.

Let Z be the zero set of the Melnikov function M, i.e.,

$$Z=\{\theta\in T^l \mid M(\theta)=0\}.$$

Suppose that Z is nonempty and

$$\text{(A.1)}\qquad \frac{d}{dt}M(\omega t+\theta)\Big|_{t=0}=\sum_{j=1}^{l}\omega_j\frac{\partial}{\partial\theta_j}M(\theta)\neq 0,$$

for all $\theta\in Z$. Then $\mathbb{R}^2\times Z$ is a global cross section for the flow of (2.1), and the Poincaré map $\Psi:\mathbb{R}^2\times Z\to\mathbb{R}^2\times Z$ is defined as follows:

$$\Psi:(x(0),\theta_0)\to(x(T),\eta(\theta_0)),$$

where $(x(t),\omega t+\theta_0)$ is a solution of (2.1),

$$\eta: Z\to Z,\qquad \theta_0\to\omega T(\theta_0)+\theta_0,$$

and $T(\theta)>0$ is the least time $t>0$ such that $\omega t+\theta\in Z$. We have the following theorem.

THEOREM A.1. *Suppose that Z is nonempty and* (A.1) *holds for all $\theta\in Z$. Then for ε sufficiently small, Ψ has an invariant set Ω. Moreover, for some integer $n\geqq 2$ and irreducible transient matrix A, $\Psi|_\Omega$ is topologically conjugate to the product map $\sigma\times\eta: B_n(A)\times Z\to B_n(A)\times Z$.*

Proof. See Meyer and Sell [1989]. □

As stated in § 2, (2.1) has a normally hyperbolic invariant l-torus T_ε near $T_0=\{x_0\}\times T^l$. Let $(x_\varepsilon^{\theta_0}(t),\theta^{\theta_0}(t))$ be an orbit on T_ε such that $\theta^0(0)=\theta_0$. Even though not every but only some zeros of M satisfy (A.1), (2.1) has random-like solutions as follows.

THEOREM A.2. *Suppose that there exists a zero $\theta=\theta_0\in T^l$ of M satisfying* (A.1). *Then, for ε sufficiently small,* (2.1) *has a solution $(\bar{x}_\varepsilon^{\theta_0}(t),\theta^{\theta_0}(t))$ with $\bar{x}_\varepsilon^{\theta_0}(t)\not\equiv x_\varepsilon^{\theta_0}(t)$ such that*

$$|\bar{x}_\varepsilon^{\theta_0}(t)-x_\varepsilon^{\theta_0}(t)|\to 0\quad\text{as } |t|\to\infty.$$

Moreover, there exist positive constants $T=T(\varepsilon)$, $\bar{T}=\bar{T}(\varepsilon)$, and $\lambda=\lambda(\varepsilon)$ such that for any interval $I_0\in\mathbb{R}$ with length λ and any sequence of real numbers $\tau_k\geqq 0$ $(k\in\mathbb{Z})$, there exists a sequence of real numbers t_k with $t_0\in I_0$ and

$$T+\tau_k\leqq t_k-t_{k-1}\leqq T+\tau_k+\lambda,$$

such that (2.1) *has a unique solution $(y_\varepsilon^{\theta_0}(t),\theta^{\theta_0}(t))$ satisfying*

$$\text{(A.2)}\qquad |y_\varepsilon^{\theta_0}(t)-\bar{x}_\varepsilon^{\theta_0}(\bar{T}+t-t_{k-1})|\leqq\varepsilon,$$

for $t\in[t_{k-1},t_k]$ and

$$\text{(A.3)}\qquad |y_\varepsilon^{\theta_0}(t)-x_\varepsilon^{\theta_0}(t)|\leqq\varepsilon,$$

for $t\in[t_{k-1}+T,t_k]$. Here we may take $\tau_l=\infty$ for some l.

Proof. The proof follows from Remark 2.9 and Theorem 2.11 of Scheurle [1986]. Note that the hypothesis of Theorem A.2 is equivalent to that of Theorem 6.1. □

Remark A.1. From the proof of Theorem 2.11 of Scheurle [1986], we can take $\bar{T}$ and T such that $\bar{x}_0(t)$ stays outside an ε-neighborhood of x_0 within an interval $I \subset [\bar{T}, \bar{T}+T]$. Hence, from (A.2) and (A.3), we see that the solution $y_\varepsilon^{\theta_0}(t)$ stays inside a neighborhood U of x_0 for some time, leaves U and then reaches U again, as $\bar{x}_\varepsilon^{\theta_0}(t)$ does. These excursions seem to occur quite randomly since $\tau_k \geqq 0$, $k \in \mathbb{Z}$, can be chosen arbitrarily. This implies that $y_\varepsilon^{\theta_0}(t)$ represents a random-like solution. Thus Theorem A.2 gives a condition for the existence of chaotic solutions. However, it has nothing to say about the existence of an invariant set on which the dynamics are characterized by Bernoulli shift or its generalization, in contrast with Theorems 3.2, 6.1, and A.1.

Now we apply Theorem A.1 to the quasi-periodically forced Duffing oscillator (7.1). Note that from Theorem A.2 we have (7.4) as a condition for the existence of chaotic solutions in (7.1).

Let

$$A_i = \frac{3\sqrt{2}\pi}{4}\,\omega_i\gamma_i \operatorname{sech}\left(\frac{\pi\omega_i}{2}\right), \qquad i = 1, 2.$$

From (7.3), the zero $(\theta_{10}, \theta_{20})$ of the Melnikov functions $M_\pm$ is given by

$$\delta = \pm A_1 \sin\theta_{10} \pm A_2 \sin\theta_{20}. \tag{A.4}$$

In this example, (A.1) becomes

$$\omega_1 A_1 \cos\theta_{10} + \omega_2 A_2 \cos\theta_{20} \neq 0.$$

Let us assume that there exists a point $(\theta_{10}, \theta_{20})$ satisfying (A.4) and

$$\omega_1 A_1 \cos\theta_{10} + \omega_2 A_2 \cos\theta_{20} = 0. \tag{A.5}$$

Then the hypothesis of Theorem A.1 does not hold, although the zero set of $M_\pm$ is nonempty. Obviously, $M_\pm$ has a zero if and only if

$$\delta \leqq A_1 + A_2.$$

From (A.4) and (A.5), we obtain

$$(\omega_1^2-\omega_2^2)A_1^2\sin^2\theta_{10} \pm 2\delta\omega_2^2 A_1 \sin\theta_{10} - \omega_1^2A_1^2 + \omega_2^2A_2^2 - \delta^2\omega_2^2 = 0, \tag{A.6a}$$

$$(\omega_2^2-\omega_1^2)A_2^2\sin^2\theta_{20} \pm 2\delta\omega_1^2 A_2 \sin\theta_{20} + \omega_1^2A_1^2 - \omega_2^2A_2^2 - \delta^2\omega_1^2 = 0. \tag{A.6b}$$

We see that (A.6) has a solution $(\theta_{10}, \theta_{20})$ if and only if

$$(\omega_1^2-\omega_2^2)(\omega_1^2A_1^2-\omega_2^2A_2^2)+\delta^2\omega_1^2\omega_2^2 \geqq 0$$

and at least one of the following conditions holds:

(i) $\delta \geqq \max\,((\omega_1^2-\omega_2^2)A_1/\omega_2^2, (\omega_2^2-\omega_1^2)A_2/\omega_1^2)$ and $|A_1-A_2| \leqq \delta \leqq A_1+A_2$,

(ii) $\delta < (\omega_1^2-\omega_2^2)A_1/\omega_2^2$ and $A_1-A_2 \leqq \delta \leqq A_1+A_2$,

(iii) $\delta < (\omega_2^2-\omega_1^2)A_2/\omega_1^2$ and $A_2-A_1 \leqq \delta \leqq A_1+A_2$.

Hence, the hypothesis of Theorem A.1 is satisfied if and only if

$$\delta^2 < \frac{\omega_2^2-\omega_1^2}{\omega_1^2\omega_2^2}(\omega_1^2A_1^2-\omega_2^2A_2^2), \tag{A.7a}$$

$$\max\,((\omega_1^2-\omega_2^2)A_1/\omega_2^2, (\omega_2^2-\omega_1^2)A_2/\omega_1^2) \leqq \delta < |A_1-A_2|, \tag{A.7b}$$

$$\delta < \min\,((\omega_1^2-\omega_2^2)A_1/\omega_2^2, A_1-A_2), \tag{A.7c}$$

or

$$\text{(A.7d)} \qquad \delta < \min\left((\omega_2^2-\omega_1^2)A_2/\omega_1^2,\ A_2-A_1\right).$$

Thus, if at least one of conditions (A.7a-d) holds, then the Poincaré map $\Psi:\mathbb{R}^2\times Z\to\mathbb{R}^2\times Z$ of (7.1) has a chaotic invariant set Ω stated in Theorem A.1. When $\delta=0$, (A.7) becomes

$$\text{(A.8a)} \qquad A_1/A_2 > \max\left(\omega_2/\omega_1, 1\right),$$

or

$$\text{(A.8b)} \qquad A_1/A_2 < \min\left(\omega_2/\omega_1, 1\right),$$

which is the same condition as given for the case of $\delta=0$ in Meyer and Sell [1989].

Acknowledgments. I would like to thank the referees for useful comments and constructive suggestions that improved this work. I also acknowledge Professor S. Wiggins for introducing me to the work of Beigie, Leonard, and Wiggins [1991a], [1991b] and Stoffer [1988].

REFERENCES

R. Abraham and J. E. Marsden (1978), *Foundations of Mechanics*, Second ed., Addison-Wesley, Reading, MA.

V. I. Arnold (1964), *Instability of dynamical systems with many degrees of freedom*, Soviet Math. Dokl., 5, pp. 581-585.

——— (1989), *Mathematical Methods of Classical Mechanics*, Second ed., Springer-Verlag, New York.

D. Beigie, A. Leonard, and S. Wiggins (1991a), *The dynamics associated with the chaotic tangles of two dimensional quasiperiodic vector fields: theory and applications*, in Nonlinear Phenomena in Atmospheric and Oceanic Sciences, G. F. Carnevale and R. Pierrehumbert, eds., Springer-Verlag, New York, to appear.

——— (1991b), *Chaotic transport in the homoclinic and heteroclinic tangle regions of quasiperiodically forced two dimensional dynamical systems*, Nonlinearity, 4, pp. 775-819.

S.-N. Chow and J. K. Hale (1982), *Methods of Bifurcation Theory*, Springer-Verlag, New York.

I. P. Cornfeld, S. V. Fomin and Ya. G. Sinai (1982), *Ergodic Theory*, Springer-Verlag, New York.

J. Guckenheimer and P. J. Holmes (1983), *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York.

M. W. Hirsch, C. C. Pugh, and M. Shub (1977), *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York.

P. J. Holmes and J. E. Marsden (1982a), *Horseshoes in perturbations of Hamiltonian systems with two degrees of freedom*, Comm. Math. Phys., 82, pp. 523-544.

——— (1982b), *Melnikov's method and Arnold diffusion for perturbations of integrable Hamiltonian systems*, J. Math. Phys., 23, pp. 669-675.

——— (1983), *Horseshoes and Arnold diffusion for Hamiltonian systems on Lie groups*, Indiana Univ. Math. J., 32, pp. 273-309.

K. Ide and S. Wiggins (1989), *The bifurcation to homoclinic tori in the quasiperiodically forced Duffing oscillator*, Phys. D, 34, pp. 169-182.

A. J. Lichtenberg and M. A. Lieberman (1983), *Regular and Stochastic Motion*, Springer-Verlag, New York.

K. R. Meyer and G. R. Sell (1989), *Melnikov transforms, Bernoulli bundles, and almost periodic perturbations*, Trans. Amer. Math. Soc., 314, pp. 63-105.

F. C. Moon (1987), *Chaotic Vibrations, an Introduction for Applied Scientists and Engineers*, John Wiley, New York.

J. Moser (1973), *Stable and Random Motions in Dynamical Systems*, Princeton University Press, Princeton, NJ.

S. E. Newhouse (1980), *Lectures on dynamical systems*, in Dynamical Systems, C.I.M.E. Lectures, Bressanone, Italy, June, 1978, Birkhäuser, Boston, MA, pp. 1-114.

K. J. Palmer (1984), *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55, pp. 225-256.

J. Scheurle (1986), *Chaotic solutions of systems with almost periodic forcing*, J. Appl. Math. Phys. (Z. Angew. Math. Phys.), 37, pp. 12-26.

L. P. SILNIKOV (1968), *Structure of the neighborhood of a homoclinic tube of an invariant torus*, Soviet Math. Dokl., 9, pp. 624-628.

S. SMALE (1963), *Diffeomorphisms with many periodic points*, in Differential and Combinatorial Topology, S. S. Cairns, ed., Princeton University Press, Princeton, NJ, pp. 63-80.

D. STOFFER (1988), *Transversal homoclinic points and hyperbolic sets for non-autonomous maps* I & II, J. Appl. Math. Phys. (Z. Angew. Math. Phys.), 39, pp. 518-549 and pp. 783-812.

J. M. T. THOMPSON AND H. B. STEWART (1986), *Nonlinear Dynamics and Chaos*, John Wiley, New York.

S. WIGGINS (1987), *Chaos in the quasiperiodically forced Duffing oscillator*, Phys. Lett. A, 124, pp. 138-142.

——— (1988a), *On the detection and dynamical consequences of orbits homoclinic to hyperbolic periodic orbits and normally hyperbolic invariant tori in a class of ordinary differential equations*, SIAM J. Appl. Math., 48, pp. 262-285.

——— (1988b), *Global Bifurcations and Chaos—Analytical Methods*, Springer-Verlag, New York.

——— (1990), *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York.

S. WIGGINS AND P. J. HOLMES (1987), *Homoclinic orbits in slowly varying oscillators*, SIAM J. Math. Anal., 18, pp. 612-629.

K. YAGASAKI (1990a), *Second-order averaging and chaos in quasiperiodically forced weakly nonlinear oscillators*, Phys. D, 44, pp. 445-458.

——— (1990b), *Chaotic dynamics of a quasiperiodically forced beam*, Trans. ASME J. Appl. Mech., to appear.

——— (1991a), *Chaos in a weakly nonlinear oscillator with parametric and external resonances*, Trans. ASME J. Appl. Mech., 58, pp. 244-250.

——— (1991b), *Chaotic motions near homoclinic manifolds and resonant tori in quasiperiodic perturbations of planar Hamiltonian systems*, in preparation.

——— (1991c), *Homoclinic tangles, phase locking and chaos in a two-frequency perturbation of Duffing's equation*, in preparation.

K. YAGASAKI, M. SAKATA, AND K. KIMURA (1990), *Dynamics of a weakly nonlinear system subjected to combined parametric and external excitation*, Trans. ASME J. Appl. Mech., 57, pp. 209-217.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1255–1268, September 1992

HOMOCLINIC BIFURCATION TO A TRANSITIVE ATTRACTOR OF LORENZ TYPE, II*

CLARK ROBINSON†

Abstract. In this paper it is proven that there is a codimension two bifurcation of a double homoclinic connection of a fixed point with a resonance condition among the eigenvalues to a transitive attractor that is like that of the geometric model of the Lorenz equations. The two key parameters are the variation of the eigenvalues from resonance and the amount that the homoclinic connection is broken. Because of the need to work near resonance of two of the eigenvalues, one of the key steps in the proof is to calculate the Poincaré–Dulac map past a fixed point in this situation. Also indicated is how bifurcation is realized for a specific cubic differential equation introduced by Rychlik, which is closely related to the Lorenz equations.

Key words. attractor, Lorenz, homoclinic bifurcation

AMS(MOS) subject classifications. 58F15, 58F13

1. Statement of main theorem. In our previous paper [8], following the work of Rychlik [10], we proved that there is a homoclinic bifurcation to a transitive attractor. We also discussed the connection with the Lorenz equations and the geometric Lorenz model. In this paper, we reconsider this same theorem, improve the proof of the Poincaré–Dulac map past the fixed point, and emphasize the fact that it is a codimension two bifurcation.

As discussed in §3, both Rychlik and Shilnikov have discussed a bifurcation from a double homoclinic bifurcation to an attractor of Lorenz type, but the assumptions on the eigenvalues of the fixed point and the transversality of the stable and unstable manifolds from which they bifurcate is different than the situation considered in this paper. See [1], [11], and [10].

We start by giving the general assumptions that the parameter must satisfy at the bifurcation value, (A0)–(A5). Then, rather than give conditions that a curve of parameter values must satisfy as in the earlier paper, we describe the part of parameter space near the bifurcation value that has an attractor. This description is given in terms of two key unfolding parameters. In §3, we will verify the assumptions for the cubic equations in R^3 that Rychlik considered.

The zeroth assumption introduces the general assumptions and notation on the the symmetry about the z-axis, and the eigenvalues and eigenvectors for the fixed point at the origin. In this paper we only consider the case when the equations have a symmetry under reflection in the z-axis, $(x, y, z) \to (-x, -y, z)$. In our previous paper, we considered one type of nonsymmetric equations. The general situation for nonsymmetric equations that allows this type of bifurcation has yet to be determined. Assumptions (A1)–(A5) give the conditions at the bifurcation parameter value $\eta = \eta_0$.

(A0) We consider a C^r vector field on R^3 for $r \geq 2$, X, which depends on parameters η with a fixed point at the origin, $Q = (0, 0, 0)$, for all parameter values.

*Received by the editors December 26, 1990; accepted for publication (in revised form) January 20, 1992.

†Department of Mathematics, Northwestern University, Lunt 218B, 2033 Sheridan Road, Evanston, Illinois 60208. The research of this author was supported in part by National Science Foundation grant DMS-8903501-01.

We assume that the vector field is taken to itself under the action of the reflection in the z-axis, $(x, y, z) \to (-x, -y, z)$. Further, we assume that the eigenvalues of $DX(Q)$ are all real with $\lambda_{ss}(\eta) < \lambda_s(\eta) < 0 < \lambda_u(\eta)$, and with respective eigenvectors v_{ss}, v_s, and v_u, and that $v_s = (0, 0, 1)^{tr}$ is along the axis of symmetry.

With these assumptions, there are several invariant manifolds for the fixed point at the origin. We denote the one-dimensional unstable manifold tangent to v_u by $W^u(Q, \eta)$, and the two-dimensional stable manifold tangent to v_s and v_{ss} by $W^s(Q, \eta)$. Next, there is a strong stable manifold tangent to v_{ss}, which we denote by $W^{ss}(Q, \eta)$. This latter manifold is made up of points that converge to Q at an asymptotic rate determined by the eigenvalue λ_{ss}. All of these manifolds are C^r if the vector field is C^r, and are even real analytic if the vector field is real analytic. Finally, there is a two-dimensional manifold tangent to the two most expanding directions, v_u and v_s, which we denote by $W^{us}(Q, \eta)$. This manifold is local in the stable direction but can be extended along the unstable manifold by flowing forward in time. We call this the weak unstable manifold even though it is not expanding in all directions. This manifold is at least C^1 (and C^2 with assumption (A3) on the dominance of the contraction toward $W^{us}(Q, \eta)$ given by $e^{\lambda_{ss}}$, in comparison with the greatest contraction within $W^{us}(Q, \eta)$ given by e^{λ_s}). With this notation we can make the second assumption about the existence of a homoclinic orbit. Because we are considering equations with symmetry, if one branch of $W^u(Q, \eta_0)$ is homoclinic, it follows that both sides are homoclinic. Thus within equations with symmetry it is only a codimension one condition to have a double homoclinic connection.

(A1) There is a bifurcation parameter value η_0 for which there is a *double homoclinic connection* with the unstable manifold of Q contained in the stable manifold but outside the strong stable manifold, $\Gamma \equiv W^u(Q, \eta_0) \subset W^s(Q, \eta_0) \setminus W^{ss}(Q, \eta_0)$. Also, because of the symmetry, both branches of $W^u(Q, \eta_0)$ are contained in the same component of $W^s(Q, \eta_0) \setminus W^{ss}(Q, \eta_0)$.

(A2) For η_0, the two-dimensional weak unstable manifold $W^{us}(Q, \eta_0)$ is *transverse* to the two-dimensional stable manifold $W^s(Q, \eta_0)$ along Γ. This transversality, together with the fact that $\Gamma \subset W^s(Q, \eta_0) \setminus W^{ss}(Q, \eta_0)$, implies that $W^{us}(Q, \eta_0)$ is tangent to itself at Q (when it leaves and returns to Q). Let $P_{\eta_0}(q) = T_q W^{us}(Q, \eta_0)$ for $q \in \Gamma$. If the bundle $\{P_{\eta_0}(q) : q \in \Gamma\}$ is orientable along Γ, we set $\nu = 1$, and if it is nonorientable we set $\nu = -1$. (The orientability is the same over both branches of Γ by the symmetry of the equations.)

This condition is generically satisfied and so does not add a codimension to the bifurcation.

(A3) We assume that for η_0 *the strong stable eigenvalue dominates* the other two eigenvalues in the sense that $\lambda_{ss}(\eta_0) - \lambda_s(\eta_0) + \lambda_u(\eta_0) < 0$ and $\lambda_{ss}(\eta_0) - 2\lambda_s(\eta_0) < 0$.

This is an open condition and does not add a codimension to the bifurcation. The second inequality in (A3) is what assures that the manifold $W^{us}(Q, \eta_0)$ is C^2. It is also redundant with the following resonance assumption (but sometimes we want to assume (A3), but not necessarily assume (A4)).

(A4) There is a one-to-one *resonance* between the unstable and weak stable eigenvalues for η_0: $\lambda_u(\eta_0) + \lambda_s(\eta_0) = 0$, so $1 - |\lambda_s(\eta_0)|/\lambda_u(\eta_0) = 0$.

This resonance condition is a codimension one and gives the second codimension that the bifurcation parameter must satisfy. The final assumption on the bifurcation parameter is the extent to which area is changed within the $P(q)$ directions ("within the attractor directions") during one loop around Γ.

(A5) For this assumption fix $\eta = \eta_0$. Let $\{P_{\eta_0}(q) : q \in \Gamma\}$ be the continuous bundle of planes given in (A2). Let $q(t)$ be a homoclinic orbit along one of the two branches of Γ, and let $\operatorname{div}_2(t)$ be the rate of change of area within the plane field $P_{\eta_0}(q(t))$. Define $C_{\eta_0} > 0$ by

$$\int_{-\infty}^{\infty} \operatorname{div}_2(q(t))dt \equiv \log(C_{\eta_0}).$$

We assume that the *change of area* $C_{\eta_0} < 2$. (Note, because the equations are symmetric, the integral is independent of which branch of Γ is used. Note also that if $\lambda_u(\eta_0) + \lambda_s(\eta_0) \neq 0$, then the integral would be $\pm\infty$ and $C_{\eta_0} = \infty$ or 0.)

It turns out that C_{η_0} has meaning in terms of a one-dimensional Poincaré map, f_{η_0}, from making one trip near one of the branches of Γ. We give a few definitions to explain this fact more fully and which will also be used in the final assumption and statement of the theorem. Let Σ be a fixed transversal to Γ out a short distance along the local stable manifold from Q. Points in a neighborhood V of Γ on $\Sigma \setminus W^s(Q,\eta)$ will return to Σ, defining a Poincaré map

$$F_\eta : V \setminus W^s(Q,\eta) \subset \Sigma \to \Sigma.$$

In the proof, it is shown that assumption (A3) implies that there is an invariant continuous bundle of strong stable directions over Γ, $\{E^{ss}(q) : q \in \Gamma\}$ with $E^{ss}(Q) = \langle v_{ss} \rangle$. These conditions are open so this bundle exists not only over Γ for η_0 but also over a neighborhood of Γ for nearby η. The stable manifold theory then implies that there is an invariant strong stable foliation for these nearby η. Each leaf of the strong stable manifold of a point is one-dimensional, but if we take the union of these for points on the same orbit we get two-dimensional strong stable manifolds of orbits that are transverse to Σ. Projection along the leaves of the strong stable manifolds of orbits defines a map $\pi_\eta : \Sigma \to \Sigma^1$, where Σ^1 can either be thought of as the quotient space or a one-dimensional manifold in Σ passing through Γ and transverse to the strong stable foliation. This projection can be used to define a one-dimensional map $f_\eta : V^1 \setminus \{0\} \subset \Sigma^1 \to \Sigma^1$ by $f_\eta(\pi_\eta w) = \pi_\eta F_\eta(w)$, where $V^1 = \pi_\eta(V)$. We use coordinates on Σ^1 so that zero corresponds to the point on $W^s(Q,\eta)$, and the symmetry gives $f_\eta(-u) = -f_\eta(u)$.

Let $a(\eta)$ be the signed distance, as measured in Σ^1 of the negative branch of $W^u(Q,\eta)$ from $W^s(Q,\eta)$. This is equivalent to defining

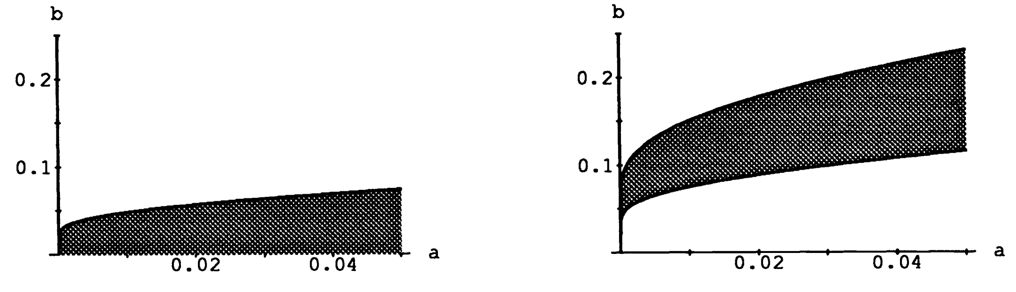
$$a(\eta) = \limsup_{u<0, u\to 0} f_\eta(u).$$

Thus the negative branch of $W^u(Q,\eta)$ intersects the positive side of Σ if $a(\eta) > 0$, and it intersects the negative side if $a(\eta) < 0$. Because of the symmetry, and the distance of the positive branch of $W^u(Q,\eta)$ from $W^s(Q,\eta)$ is -$a(\eta)$. The interval $I_\eta = [-|a(\eta)|, |a(\eta)|]$ corresponds to the interval in Σ^1 between the intersections of the two branches of $W^u(Q,\eta)$ with the transversal after projecting out the strong stable direction. With these definitions, Lemma 1 proves that $f'_{\eta_0}(0) = \nu C_{\eta_0}$. The fact that $C_{\eta_0} < 2$ means there is hope for $f_\eta([0, |a(\eta)|] \subset I_\eta$. See Lemma 2 for details.

Let $b(\eta) = 1 - |\lambda_s(\eta)/\lambda_u(\eta)|$. This quantity measures the extent to which the two eigenvalues are no longer in resonance.

Next, we discuss the relationship between the unfolding parameters $a(\eta)$ and $b(\eta)$ that must be satisfied as the homoclinic connection is broken in order for there to be an attractor. For a small neighborhood $\mathcal{N}$ of η_0 in parameter space, let

$$\mathcal{N}' = \{\eta \in \mathcal{N} : \nu a(\eta) > 0, 0 < f_\eta(a(\eta)) \leq |a(\eta)|, b(\eta) \geq 0, |f'_\eta(\pm a(\eta))| \geq 2^{1/2}\}.$$



(a) (b)

Fig. 1. *Graph of the allowable range of* $b(\eta)$ *versus* $a(\eta)$. (a) *Graph for* $C_{\eta_0} = 1.6$. (*Similar ranges are valid for* $2^{1/2} < C_{\eta_0} < 2$.) (b) *Graph for* $C_{\eta_0} = 1$. (*Similar ranges are valid for* $0 < C_{\eta_0} < 2^{1/2}$. *This is the case we prove occurs for equations* (R′) *and* (Rob).)

Lemma 2 proves that the boundary of $\mathcal{N}'$ is contained in $\partial\mathcal{N}$, γ_1, and γ_2, where the latter two are given by

$$\gamma_1 : \quad \{\eta \in \mathcal{N} : f_\eta(a(\eta)) = \nu a(\eta), b(\eta) > 0\},$$
$$\gamma_2 : \quad \begin{cases} b(\eta) = 0 & \text{if } C_{\eta_0} < 2^{1/2} \\ \{\eta \in \mathcal{N} : b(\eta) > 0, |f'_\eta(\pm a(\eta))| = 2^{1/2}\} & \text{if } C_{\eta_0} > 2^{1/2}. \end{cases}$$

The form of γ_2 when $C_{\eta_0} = 2^{1/2}$ is some combination of the two forms above. See Fig. 1. The condition related to γ_2 implies that for $\eta \in \mathcal{N}'$ $|f'_\eta(u)| > 2^{1/2}$ for all points $u \in \operatorname{interior} I_\eta$. This latter condition is what is used to imply that the attractor is transitive.

In order to insure that $\mathcal{N}' \neq \emptyset$ and $\eta_0 \in \operatorname{closure} \mathcal{N}'$, we need to make the following assumption on the ability to vary the parameters.

(A6) We assume that the parameter space for η is large enough so that $a(\eta)$ and $b(\eta)$ can vary independently for η near η_0.

Finally, we can state the main result.

THEOREM 1. *Assume that vector field in* R^3, *depending on a parameter* η *is* C^2 *and satisfies the above assumptions* (A0)–(A6). *Then, there is a neighborhood* $\mathcal{N}_0 \subset \mathcal{N}$ *in parameter space such that* $\mathcal{N}_0 \cap \mathcal{N}' \neq \emptyset$, $\eta_0 \in$ *closure* $\mathcal{N}'$, *and for* $\eta \in \mathcal{N}_0 \cap \mathcal{N}'$ *the flow for* η *has a topologically transitive attractor. The two pieces* γ_1 *and* γ_2 *of the boundary of* $\mathcal{N}_0 \cap \mathcal{N}'$ *satisfy the following equations*:

$$\gamma_1 : \log(2/C_{\eta_0}) = \limsup\{[(b(\eta)\log(1/|a(\eta)|)] : \eta \to \eta_0, \ \eta \in \gamma_1\},$$
$$\gamma_2 : \begin{cases} \liminf\{[(b(\eta)\log(1/|a(\eta)|)] : \eta \to \eta_0, \ \eta \in \gamma_2\} = \log(2^{1/2}/C_{\eta_0}) & \textit{if } C_{\eta_0} > 2^{1/2} \\ b(\eta) = 0 & \textit{if } C_{\eta_0} < 2^{1/2}. \end{cases}$$

If the value of ν *is* 1 (*respectively,* -1), *then the attractor is orientable* (*respectively, nonorientable*), *and the attractor appears for* $a(\eta)$ *positive* (*respectively, negative*), *i.e., it appears before* (*respectively, after*) *the unstable manifold crosses over the stable manifold. If the vector field is* C^3, *then the resulting one-dimensional Poincaré map* f_η *has an ergodic invariant measure with support equal to the whole interval* I_η *and which is equivalent to Lebesgue on* I_η.

2. Proof of Theorem 1. Throughout this section, we assume that the system satisfies assumptions (A0)–(A3).

The first step is to prove the existence of a $C^{1+\mu}$ for some $0 < \mu < 1$ strong stable foliation in a neighborhood of the homoclinic connection for a perturbation of the flow. (Here $C^{1+\mu}$ means that the first derivative is μ-Hölder.) The details of this argument are carried out in [8] and follow the ideas of the earlier papers [6] and [7]. The sketch of the proof is as follows. The existence of the continuous bundle of planes $\{P_{\eta_0}(q) : q \in \Gamma\}$ can be used to show the existence of a continuous strong stable bundle by using cones that are complementary to the $P_{\eta_0}(q)$. In fact, since a trajectory in a small neighborhood of Γ spends an arbitrarily large proportion of its time in a neighborhood of the fixed point where the eigenvalues give the desired estimates, it is possible to prove the existence of a $C^{1+\mu}$ strong stable bundle, E^{ss}, in a neighborhood of the homoclinic connection. These conditions are open, so for a small perturbation such a strong stable bundle persists in this open set of phase space.

Then by stable manifold theory, there is a $C^{1+\mu}$ foliation whose tangent lines are given by these strong stable bundles. See [6], which uses [3, Thm. 4.8]. A strong stable leaf or manifold at a point ζ_0 of radius r, $W^{ss}_r(\zeta_0, \eta)$ is characterized as being the points ζ within distance r such that the distance between the trajectories for ζ and ζ_0 at time t converges to zero at an exponential rate of almost $\exp(t\lambda_{ss})$.

Once we know that the strong stable manifolds form a $C^{1+\mu}$ foliation, we can form the strong stable manifolds of orbits by taking the union of the strong stable manifolds of points along an orbit:

$$W_r^{ss,\mathrm{orbit}}(\zeta_0, \eta) = \cup\{W_r^{ss}(\zeta, \eta) | \zeta = \varphi_t(\zeta_0, \eta) \text{ for } -r \le t \le r\}.$$

The tangent space to $W_r^{ss,\mathrm{orbit}}(\zeta_0, \eta)$ at ζ_0 is spanned by the strong stable bundle E^{ss} at ζ_0 and the vector field $X_\eta(\zeta_0)$. Thus, these tangent spaces are $C^{1+\mu}$ away from Q. This implies that $\pi_\eta : \Sigma \to \Sigma^1$ is $C^{1+\mu}$.

As a consequence of (A1), the value of $\nu = 1$ (respectively, $\nu = -1$) if the Poincaré map F_η preserves (respectively, reverses) the orientation of the strong stable bundle. Hence, $\nu = 1$ (respectively $\nu = -1$) if f_η preserves (respectively, reverses) the orientation within Σ^1, so f_η is increasing (respectively, decreasing). In terms of the flow, if $\nu = 1$, the "sheets within the attractor" return with either no twist or at least multiples of full twists, so the same side is up. On the other hand, if $\nu = -1$, they return with a half twist plus some multiple of full twists, so the sides are reversed.

We turn now to analyzing the Poincaré map for the flow. We need to find the lowest-order terms for f_η and f'_η in order to show that we have an invariant set and an expansion on it. In the two-dimensional case, if the flow is linearized near a saddle fixed point, then the Poincaré map past the fixed points is given by $g(u) = Cu^E$, where $E = |\lambda_s|/\lambda_u$. In our present situation, we reduce to the two-dimensional case by projecting along the strong stable foliation and use the weaker stable and unstable directions near the fixed point. This quotient space can be thought of as projection onto the invariant surface tangent to $\langle v_u, v_s\rangle$ at Q. The resulting one-dimensional Poincaré map past the fixed point has $\pm Cu^E$ as the lowest-order terms, and the following lemma makes this precise.

LEMMA 1. *Assume* (A0)–(A3) *are satisfied. Let* $E = E_\eta = |\lambda_s(\eta)|/\lambda_u(\eta) = 1 - b(\eta)$. *Let* $J \subset \Sigma^1$ *be a fixed small interval about zero. For* η *in a small neighborhood of* η_0, *the induced one-dimensional Poincaré map* $f_\eta : J \setminus \{0\} \subset \Sigma^1 \to \Sigma^1$ *has continuous derivative on* $J \setminus \{0\}$, *and* f_η *and* f'_η *have the following form:*

$$\begin{aligned} f_\eta(u) &= (-a(\eta) + \nu C_\eta |u|^E)\mathrm{sign}(u) + o(|u|^E), \\ f'_\eta(u) &= \nu E_\eta C_\eta |u|^{E-1} + o(|u|^{E-1}), \end{aligned}$$

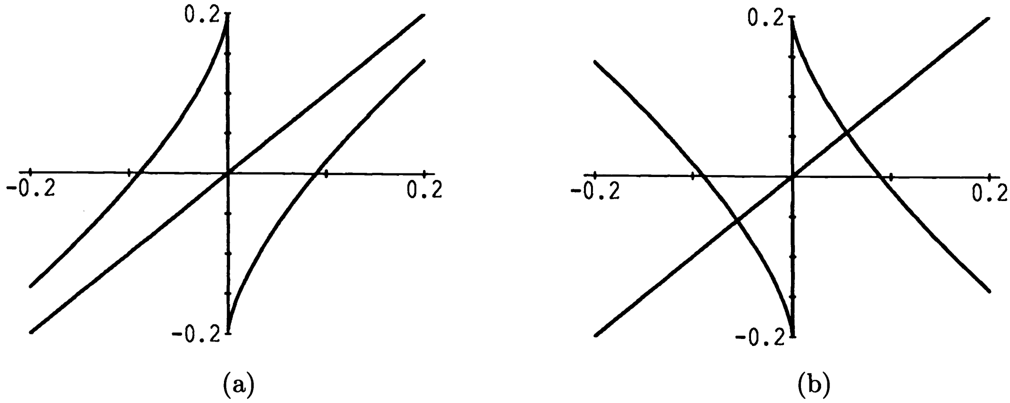


FIG. 2. *Graph of f_η.* (a) $\nu = 1$, $a(\eta) = 0.2$, $E = \frac{2}{3}$, *and* $C = 1$. (b) $\nu = -1$, $a(\eta) = -0.2$, $E = \frac{2}{3}$, *and* $C = 1$.

with $\nu = \pm 1$, depending on whether the strong stable bundle is orientable or not, $C_{\eta_0} \geq 0$ is given by the integral in (A5)*, and C_η depending continuously on η. The remainder terms uniformly go to zero when divided by the terms indicated as u goes to zero uniformly in η. See Fig.* 2*. Note that for η_0, $f_{\eta_0}(u) = \nu C_{\eta_0} u + o(|u|)$, and $f'_\eta(u) = \nu C_{\eta_0} + o(|u|^0)$ so $f'_{\eta_0}(0)$ exists and equals νC_{η_0}. Also, if $E_\eta \leq 1$ (e.g., for $\eta \in \mathcal{N}'$), then the branches of $(f_\eta^{-1})'(u)$ have Hölder extensions at $f_\eta(0\pm)$.*

Remark 1. The proof of this lemma also works in the nonsymmetric case. The delicate part of the argument about the Poincaré–Dulac map past the fixed point is the same. The part of the return map outside the neighborhood depends on the branch of the unstable manifold that is followed. Therefore, the form of the map is

$$f_\eta(u) = \begin{cases} a^+(\eta) + \nu^+ C_\eta^+ |u|^E) + o(|u|^E) & \text{for } u > 0, \\ a^-(\eta) + \nu^- C_\eta^- |u|^E) + o(|u|^E) & \text{for } u < 0, \end{cases}$$

where $a^\pm(\eta)$, $\nu^\pm$, and $C_\eta^\pm$ all depend on whether u is positive or negative.

Remark 2. There are several ways to prove this lemma under a variety of more or less restrictive hypothesis. In [8], an elementary but messy proof was given. The equations are normalized near the fixed point with coordinates u in the unstable direction, v in the strong stable direction, and z in the weaker stable direction. The variable $\rho = uv$, which is introduced, then satisfies the scalar equation

$$\dot\rho = \Lambda_\eta \rho + A_2 \rho^2 + \cdots + A_n \rho^n,$$

where $\Lambda_\eta = \lambda_u(\eta) + \lambda_s(\eta)$. It is then necessary to prove the expansion of the solution $\rho(\tau)$ with initial condition ρ_0 for a time $\tau \approx -\lambda_u^{-1} \log |\rho_0|$, which are uniformly valid for $\Lambda_\eta \geq 0$. (This value of τ is the time it takes to flow past the fixed point to a second transversal.) The method used only obtained estimates on an interval of initial conditions in state space whose length decreases as the parameter $b(\eta) = 1 - E_\eta$ goes to zero. A better proof of this general type was done earlier by Roussarie [9, Thm. F]. He obtained a higher-order expansion and showed that it was valid on an interval of uniform size as the parameter varies, although it involves $x \log(x)$ terms which go to zero slowly in the position variable x. We make further comments below about how to use his theorem to obtain our result. Finally, it is possible to prove the lemma by $C^{1+\alpha}$ linearizing within an invariant two-dimensional surface near the fixed point. Hartman showed that this was possible in two dimensions even with resonance. This proof is also valid on an interval of fixed size and is more in keeping with other arguments used in the proof of the main theorem. We give details of this proof below. We should mention that both C. Chicone and C. Pugh suggested using this approach.

Proof using Hartman's linearization. We want to prove that the one-dimensional Poincaré map is C^1. The two-dimensional Poincaré map F_η can be split into the part G_η past Q from Σ to a second transversal S, and the part H_η from S back to Σ. Since there is a bounded amount of time for the trajectory to pass from S back to Σ, H_η is C^2. All that remains to check is the form of G_η past Q. Hartman's linearization proof works by constructing two invariant C^1 foliations in the stable and unstable directions. The difficulty is that the unstable foliation can only be shown to be C^0 in the directions transverse to the leaves if we consider the flow in three dimensions. To avoid this difficulty, we show there is an invariant two-dimensional manifold $W^{us}(Q,\eta)$ in the weak stable and unstable directions, and we show that the foliations are $C^{1+\beta}$ when restricted to this two-dimensional manifold. The proof of this differentiability is very similar to the one we used above to get the existence of the strong stable foliation. The difference is that everything must be restricted to $W^{us}(Q,\eta)$ in order for the estimates to be true. The estimates for one-dimensional Poincaré map g_η is then determined using the linearization of the flow on $W^{us}(Q,\eta)$ using these two foliations. Then $f_\eta(u) = \pi_\eta H_\eta \circ G_\eta(u)$.

The limit of $f_\eta(u)$ as u approaches zero from below (zero is the point that corresponds to the stable manifold) is $a(\eta)$ because $\pi_\eta H_\eta$ takes the negative branch of the unstable manifold to $a(\eta)$.

By (A3), $\lambda_{ss}(\eta)-(2+\alpha)\lambda_s(\eta) < 0$ for small $\alpha > 0$ and η near η_0. By the invariant manifold theorem of [4], there is a $C^{2+\alpha}$ manifold $W^{us}(Q,\eta)$ tangent to the v_u and v_s directions at Q. Its tangent space E^{us} is $C^{1+\alpha}$. Thus the derivative of the time one map of the flow $D\varphi^1$ is $C^{1+\alpha}$ when restricted to $E^{us}|W^{us}(Q,\eta)$.

Turning to the invariant subbundles, if β is chosen with $0 < \beta \le \alpha$ and $\lambda_s(\eta_0) + \beta\lambda_u(\eta_0) < 0$, then there is a $C^{1+\beta}$ invariant subbundle $E^s \subset E^{us}|W^{us}(Q,\eta)$ for η near η_0. The reason this is true is that $D\varphi^{-1}$ contracts toward E^s within E^{us} by a factor of almost $\exp(\lambda_s(\eta_0) - \lambda_u(\eta_0))$. The Lipschitz constant of φ^1 is about $\exp(\lambda_u(\eta_0))$. To prove the bundle is $C^{1+\beta}$, the results of [4] say we need the product of the first of these two numbers times the $(1+\beta)$ power of the second has to be less than one, so $0 > \lambda_s(\eta_0) - \lambda_u(\eta_0) + (1+\beta)\lambda_u(\eta_0)$. Thus in a neighborhood of Q we will have the correct domination of the contraction rate.

Next, by taking $\beta > 0$ smaller so that $-\lambda_u(\eta_0) - \beta\lambda_s(\eta_0) < 0$, there is a $C^{1+\beta}$ invariant subbundle E^u of $E^{us}|W^{us}(Q,\eta)$ for η near η_0. This follows because the contraction rate toward this bundle by $D\varphi^1$ is about $\exp(-\lambda_u(\eta_0) + \lambda_s(\eta_0))$. The Lipschitz constant of φ^{-1} on $W^{us}(Q,\eta)$ is about $\exp(-\lambda_s(\eta_0))$. Again, the above estimate gives us that the product of the first of these number times $1+\beta$ times the second is less than one. This gives the stated differentiability of the bundle.

Since each of these bundles is $C^{1+\beta}$, there are two $C^{1+\beta}$ foliations tangent to them. These give coordinates on $W^{us}(Q,\eta)$ in terms of coordinates on $W^u(Q,\eta)$ and $W^s(Q,\eta)$. Since these coordinates can $C^{1+\beta}$ linearize the flow the Poincaré–Dulac map past Q in these coordinates is the same as that for the linear flow, $g_\eta(u) = Cu^E$ where $E = E(\eta) = |\lambda_s(\eta)|/\lambda_u(\eta)$. The map in the original coordinates can then be written as $k_\eta \circ g_\eta \circ h_\eta$, where k_η and h_η are the $C^{1+\beta}$ change of coordinates which vary continuously with respect to η. This shows that the remainders go to zero as u goes to zero, uniformly in η.

The connection of C_η with the integral in (A5) is discussed in [8] and uses the formula for the derivative of the Poincaré map in terms of the integral of the divergence given in [2, §28].

The comment about the branches of $(f_\eta^{-1})'(\zeta)$ follows by looking at the flow for negative time past Q. The above proof shows that the derivative of the inverse on either branch has the form

$$(f_\eta^{-1})'(\zeta) = \nu E_\eta C_\eta |\zeta - a(\eta)|^{(1/E)-1} + o(|\zeta - a(\eta)|^{(1/E)-1}),$$

which has a Hölder extension at $\zeta = a(\eta)$. □

Proof using Roussarie's expansion. To use Roussarie's result, we again need to put the equation in normal form, so we need to assume the vector field is C^∞ (or we need to keep track of some higher differentiability, which is determined by the eigenvalues). As is done in [8] and [9], by a "Sternberg differentiable linearization," there is a C^∞ change of coordinates that put the differential equations into the normal form

$$\begin{aligned} \dot v &= \lambda_{ss} v, \\ \dot u &= \lambda_u u, \\ \dot z &= z[\lambda_s + \lambda_u \alpha_2 \rho + \lambda_u \alpha_3 \rho^2 + \cdots], \end{aligned}$$

where $\rho = uz$. The proof of the differentiable linearization is valid uniformly in the parameter by using the form given by Takens in [13], as was mentioned to us by Roussarie. This also shows that the theorem is just as valid in conjugating to a normal form as given above as it is to conjugating to a linear system. All that is necessary is for the two systems to have the same C^∞ jet at the fixed point.

As in the first proof, the only part of the Poincaré map that needs to be checked is the one-dimensional map past Q. Using the normal form, this is calculated using the u and z equations. Reference [9, Thm. F] proves the form of this map to any finite order that is uniformly valid for a range of parameters near resonance. He considers the two variables u and ρ and looks at the two equations

$$\begin{aligned} \dot u &= \lambda_u u, \\ \dot \rho &= \lambda_u [b(\eta)\rho + \alpha_2 \rho^2 + \cdots], \end{aligned}$$

where $b(\eta) = [\lambda_s(\eta) + \lambda_u(\eta)]/\lambda_u(\eta)$ is as before. (He rescales time by $\lambda_u(\eta)$, which is the reason that we have factored out this term.) Next, he defines

$$\omega(b,u) = \begin{cases} \frac{u^{-b}-1}{b} & \text{for } b = b(\eta) \neq 0, \\ -\log(u) & \text{for } b(\eta) = 0. \end{cases}$$

Then he proves that the following expansion is uniformly valid in η with C^2 remainder:

$$g_\eta(u) = Cu^{1-b} + \alpha_2 u^{2-b}\omega(b,u) + \psi_\eta(u),$$

where $\psi_\eta(u)$ is C^2 and C^2 flat at $u = 0$. Remember that $E_\eta = 1 - b(\eta)$. Since $u\omega(b,u) - u\omega(0,u)$ goes to zero uniformly on a fixed interval and $u\omega(0,u) = o(u^0)$, we get the form given in the theorem. A direct calculation shows that

$$g'_\eta(u) = CEu^{E-1} + \alpha_2(1+E)u^{E-1}(u\omega) - \alpha_2 u^{E-1}u^E + \psi'_\eta(u);$$

so $f'_\eta(u)$ has the form that is given in the lemma. The other part of the proof is as is done above in the Hartman's linearization proof. □

The following lemma proves the form of the boundary of $\mathcal{N}'$.

LEMMA 2. *Assume that the systems satisfies* (A0)–(A3).

(a) *In a small neighborhood* $\mathcal{N}_0$ *of* η_0, *for the system for* $\eta \in \mathcal{N}_0$ *to have a transitive attractor with* $|f'_\eta(u)| > 2^{1/2}$ *for all* $u \in \text{interior}\, I_\eta$, *it is necessary for* $\eta \in \mathcal{N}'$. *If we assume* $\eta_0 \in \text{closure}\, \mathcal{N}'$ *so there are transitive attractors arbitrarily near* η_0, *then* (A4) *is true; so* $b(\eta_0) = 0$ *or* $E_{\eta_0} = 1$, *and* (A5) *is true in the modified sense that* $C_{\eta_0} \leq 2$.

(b) *Assume the system satisfies* (A0)–(A6). *Then, the two pieces of the boundary of the set* $\mathcal{N}'$ *are defined by*

$$\begin{aligned}
\gamma_1 :&\quad \{\eta \in \mathcal{N} : f_\eta(a(\eta)) = \nu a(\eta), b(\eta) > 0\},\\
\gamma_2 :&\quad \begin{cases} b(\eta) = 0 & \text{if } C_{\eta_0} < 2^{1/2}\\ \{\eta \in \mathcal{N} : b(\eta), |f'_\eta(\pm a(\eta))| = 2^{1/2}\} & \text{if } C_{\eta_0} > 2^{1/2}.\end{cases}
\end{aligned}$$

Further, they satisfy the following two equations:

$$\begin{aligned}
&\gamma_1 : \log(2/C_{\eta_0}) = \limsup\{[(b(\eta)\log(1/|a(\eta)|)] : \eta \to \eta_0,\ \eta \in \gamma_1\},\\
&\gamma_2 : \begin{cases} \liminf\{[(b(\eta)\log(1/|a(\eta)|)] : \eta \to \eta_0,\ \eta \in \gamma_2\} = \log(2^{1/2}/C_{\eta_0}) & \text{if } C_{\eta_0} > 2^{1/2},\\ b(\eta) = 0 & \text{if } C_{\eta_0} < 2^{1/2}.\end{cases}
\end{aligned}$$

Proof. To obtain an attracting set for the flow that contains the fixed point at the origin, f_η must preserve the interval $I_\eta = [-|a(\eta|, |a(\eta)|] \subset V^1$. For $f_\eta(u) \in I_\eta$ for small u, the form of f_η given in Lemma 1 shows that it is necessary for $\nu a(\eta) > 0$. The invariance of the interval also implies that $f_\eta(\pm a(\eta)) \in I_\eta$; so $f_\eta(a(\eta)) \leq |a(\eta)|$, i.e., that the second intersection of the unstable manifolds must lie between the first intersection of the two branches after projecting out along the strong stable foliation. (The reader can check the well-known result that if $f_\eta(I_\eta) \setminus I_\eta \neq \emptyset$, then there is horseshoe for F_η which is not an attracting set.) Also, for Q to be part of the attracting set, we need zero to be an element of the image, $f_\eta(I_\eta)$, so $0 \leq f_\eta(a(\eta))$. Also, the fact that f_η is transitive on I_η implies that $0 < f_\eta(a(\eta))$.

The condition that the absolute value of the derivative at all points of interior I_η is greater than $2^{1/2}$ certainly implies that it is greater or equal to $2^{1/2}$ at $\pm a(\eta)$. Finally, if $b(\eta) = 1 - E_\eta < 0$, then $|f'_\eta(u)| < 1$ for small u. Thus it is necessary for $b(\eta) \geq 0$. This completes the proof that it is necessary for $\eta \in \mathcal{N}'$.

If $\eta_0 \in \text{closure}\, \mathcal{N}'$, the fact that $b(\eta) \geq 0$ for $\eta \in \mathcal{N}'$ implies that $b(\eta_0) \geq 0$. To get the opposite inequality we consider the $\eta \in \mathcal{N}'$ that approach η_0 and write a for $a(\eta)$. (The reader may find it easier to just consider the case $\nu = 1$.) For these η, $\nu a > 0$; so $\nu f_\eta(\nu a) \leq \nu a$, $\nu f_\eta(\nu a) + \nu a \leq 2\nu a$, $|f_\eta(\nu a) + a| \leq 2|a|$, and

$$\frac{|f_\eta(\nu a) + a|}{|aC_\eta|} \leq \frac{2}{C_\eta}.$$

Because of the form of the lowest-order terms of f_η given in Lemma 1, taking limits as η goes to η_0,

$$\begin{aligned}
\limsup\left\{b(\eta)\log\left(\frac{1}{|a(\eta)|}\right)\right\} &= \limsup\left\{\log\left(\frac{|f_\eta(\nu a(\eta)) + a(\eta)|}{|a(\eta)C_\eta|}\right)\right\}\\
&\leq \limsup\left\{\log\left(\frac{2}{C_\eta}\right)\right\}\\
&= \log\left(\frac{2}{C_{\eta_0}}\right).
\end{aligned}$$

Since $\log(2/C_{\eta_0})$ is a constant and $\log(1/|a(\eta)|)$ goes to infinity, we also get that $b(\eta_0) \leq 0$. Combining with the inequality that $b(\eta_0) \geq 0$, we get that $b(\eta_0) = 0$. Also since $b(\eta) \geq 0$ for $\eta \in \mathcal{N}'$ and $\log(1/|a(\eta)|)$ goes to infinity, we get that $C_{\eta_0} \leq 2$. This completes the proof of part (a).

On one boundary of $\mathcal{N}'$ $f_\eta(a(\eta)) = \nu a(\eta)$ as stated for γ_1, and putting equality in the above limit as $\eta \to \eta_0$ shows that these values satisfy

$$\gamma_1: \quad \log\left(\frac{2}{C_{\eta_0}}\right) = \limsup\left\{(b(\eta)\log\left(\frac{1}{|a(\eta)|}\right) : \eta \to \eta_0, \ \eta \in \gamma_1\right\}$$

as stated.

From the form of $f'_\eta(u)$ given in Lemma 1, it follows that the smallest value of $|f'_\eta(u)|$ occurs for $u = \pm a(\eta)$. If $C_{\eta_0} > 2^{1/2}$, then the calculation below shows that the η with $|f'_\eta(a(\eta))| = 2^{1/2}$ occurs for $b(\eta) > 0$; so it forms the other boundary.

Along the η with $|f'_\eta(a(\eta))| = 2^{1/2}$, the form of the derivative given in Lemma 1 gives the following:

$$\begin{aligned}\liminf\left\{b(\eta)\log\left(\frac{1}{|a(\eta)|)}\right)\right\} &= \liminf\left\{\log E_\eta + b(\eta)\log\left(\frac{1}{|a(\eta)|}\right)\right\} \\ &= \liminf\left\{\log\left(\frac{|f'_\eta(a(\eta))|}{C_\eta}\right)\right\} \\ &= \liminf\left\{\log\left(\frac{2^{1/2}}{C_\eta}\right)\right\} \\ &= \log\left(\frac{2^{1/2}}{C_{\eta_0}}\right).\end{aligned}$$

This gives the form of γ_2 for $C_{\eta_0} > 2^{1/2}$. If $C_{\eta_0} < 2^{1/2}$, then the calculation above shows that $|f'_\eta(a(\eta))| > 2^{1/2}$; so the other boundary is given by $b(\eta) = 0$. This completes the proof of Lemma 2. □

Using Lemma 2, for $\eta \in \mathcal{N}'$, $|f_\eta(u)| > 2^{1/2}$ for $u \in$ interior I_η. This means that f_η satisfies the condition of R. Williams, which implies that f_n is locally eventually onto and so topologically transitive. See [14] and [7]. If the flow is C^3, the fact that $f_\eta^{-1}(u)$ has a Hölder extension at $f_\eta(0\pm)$ means that the the theorem of G. Keller applies, and we get the existence of an invariant measure for the interval map. See [7], which uses the results of [5].

This completes the proof of Theorem 1. □

3. Verification of assumptions for specific equations. The Lorenz equations are given by the equations

$$\text{(L)} \qquad \begin{aligned}\dot{x} &= -\sigma x + \sigma y, \\ \dot{y} &= \rho x - y - xz, \\ \dot{z} &= -\beta z + xy.\end{aligned}$$

They arise from modeling turbulence, and their numerical simulation exhibits chaotic behavior.

Later, Rychlik showed that a slight variation of the Lorenz equation could be proved to possess a transitive attractor immediately after a bifurcation from a double homoclinic connection. The equations that he considers are motivated by noting that by a change of variables the Lorenz equations, (L), can be put in the following form:

$$\text{(R)}\qquad \begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - 2x^3 + \alpha y + \beta x^2 y + \delta x z, \\ \dot{z} &= -\gamma z + x^2, \end{aligned}$$

with $\beta = 0$, [10]. He then adds the $\beta x^2 y$ term, to be able to control the unstable manifold of the origin while keeping $\delta = 0$. He determines the form of the resulting two-dimensional Poincaré map, and proves that it has a $C^{1+\mu}$ strong stable foliation immediately after a homoclinic bifurcation, but not at the bifurcation value. This proof extends the type of proof in [6], where one of the approaches show how to verify that the Poincaré map had a C^1 foliation. We also remark that Rychlik considers a different bifurcation problem than the one considered in this paper and in [8]. In particular, instead of (A2), he assumes that $W^{us}(Q,\eta_0)$ and $W^s(Q,\eta_0)$ are tangent along Γ. (This is his second codimension.) With this assumption, he is able to take $\lambda_u(\eta_0) + \lambda_s(\eta_0) > 0$. Shilnikov has also considered a bifurcation from a double homoclinic connection, which is more like that of Rychlik with $\lambda_u(\eta_0) + \lambda_s(\eta_0) > 0$. He and his coauthors show there is an attractor of Lorenz type in a neighborhood. See [1] and [11].

In our previous paper we verified the assumptions for the set of equations

$$\text{(Rob)}\qquad \begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - 2x^3 + \alpha y + \beta x^2 y - \nu y z, \\ \dot{z} &= -\gamma z + \delta x^2, \end{aligned}$$

with $\nu = \pm 1$. These equations can be obtained from (R) by scaling z to shift the coefficient δ from the xz term in the $\dot{y}$ equation to the x^2 term in the $\dot{z}$ equation. (This change is made because we treat a different type of perturbation problem than Rychlik.) We also change the xz term in the $\dot{y}$ equation to a yz term. Although this makes the equations farther from the Lorenz equations, they are easier to analyze.

In this paper, we return to the equations (R) but scale the equations to shift the coefficient δ to the $\dot{z}$-equation; the αy term is replaced by $-\alpha y$ so we can take $\alpha > 0$, and we allow $-\nu x z$ in the $\dot{y}$ equation with $\nu = \pm 1$:

$$\text{(R}'\text{)}\qquad \begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - 2x^3 - \alpha y + \beta x^2 y - \nu x z, \\ \dot{z} &= -\gamma z + \delta x^2. \end{aligned}$$

We can now state the theorem.

THEOREM 2. *Equations* (R$'$) *satisfy assumptions* (A0)–(A6) *for correctly chosen values of the parameters* (*in particular, for some* $\alpha_0 > 2^{-\frac{1}{2}}$).

Remark 1. It is not clear that the choice of ν in the equations corresponds to the choice in assumption (A2); that is, it determines whether the one-dimensional Poincaré–Dulac map is monotonically increasing or decreasing (or the tangent space of $W^{us}(Q,\eta_0)$ is orientable or nonorientable along Γ). The difficulty is that the proof only proves transversality by an argument that uses the fact that an analytic function that is not identically zero must be nonzero in any interval.

Remark 2. It seems unlikely that any of these results, here [8] or [10], can be applied directly to the actual Lorenz equations (L). The difficulty is finding parameter values that verify both the homoclinic connection condition (A1) and the resonance condition (A4) for the same parameter values. The addition of the extra parameter β is what gives equations (R), (R$'$), or (Rob) the flexibility needed to satisfy these simultaneously.

Proof of Theorem 2. The parameters are $\eta = (\alpha, \beta, \gamma, \delta, \nu)$. The fixed point Q is always the origin. The linearization of the vector field is given by

$$DX(x,y,z) = \begin{pmatrix} 0 & 1 & 0 \\ 1 - 6x^2 + 2\beta xy - \nu z & -\alpha + \beta x^2 & -\nu x \\ 2\delta x & 0 & -\gamma \end{pmatrix}.$$

At the origin, the eigenvalues are $\lambda_{ss} = -\alpha/2 - (1 + \alpha^2/4)^{1/2}$, $\lambda_u = -\alpha/2 + (1 + \alpha^2/4)^{1/2}$, and $\lambda_s = -\gamma$. By picking the parameter $\gamma_0 = \lambda_u = -\alpha_0/2 + (1 + \alpha_0^2/4)^{1/2}$ at the bifurcation, we can insure that $\lambda_s(\eta_0) + \lambda_u(\eta_0) = 0$, giving assumption (A4).

To obtain (A3), we need the combination of all three eigenvalues less than zero: $\lambda_{ss}(\eta_0) - \lambda_s(\eta_0) + \lambda_u(\eta_0) < 0$,

$$\begin{aligned} 0 &> [-\alpha_0/2 - (1 + \alpha_0^2/4)^{1/2}] + 2[-\alpha_0/2 + (1 + \alpha_0^2/4)^{1/2}] \\ &> -3\alpha_0/2 + (1 + \alpha_0^2/4)^{1/2}, \end{aligned}$$

or

$$\alpha_0 > 2^{-1/2}.$$

Thus to obtain a C^1 foliation, it is not possible to take a small perturbation of the integrable case where $\alpha = \beta = \delta = 0$. Given the resonance condition (A4), the second inequality in (A3) follows from the first.

LEMMA 3. *For either $\nu = \pm 1$, there are parameter values $(\alpha_0^\pm, \beta_0^\pm, \gamma_0^\pm, \delta_0^\pm, \nu)$, with $\alpha_0^\pm > 1/(2^{1/2}), \beta_0^\pm > 0$, $\gamma_0^\pm = -(\alpha_0^\pm/2) + [1 + ((\alpha_0^\pm)^2/4)]^{1/2}$, and $\delta_0^\pm > 0$, but near zero, for which the equations have a homoclinic connection and an invariant continuous plane field, satisfying* (A1)–(A5).

Proof. The only difference in verifying this lemma for the equations (R$'$) and (Rob) is in the verification of (A2). We repeat the other steps because they are easy, and we give a different approach to verifying (A2).

First, taking $\alpha = \beta = \delta = 0$, we obtain a Hamiltonian system in the (x,y)-plane with energy $H = (y^2 - x^2 + x^4)/2$; therefore, the origin has a double homoclinic connection.

Now, we increase the value of α to $\alpha_0 > 1/(2)^{1/2}$, keeping $\beta = \delta = 0$, and choosing $\gamma_0 = -\alpha_0/2 + (1 + \alpha_0^2/4)^{1/2}$. The xy-plane is still invariant; so $z = 0$ along the unstable manifold. The $-\alpha y$ term in the $\dot{y}$ equation is a friction term; so the unstable manifold $W^u(Q,\eta)$ stays on the same side of $W^s(Q,\eta)$ and spirals into one of the stable fixed points. Next, increase the value of β. The term $\beta x^2 y$ in the $\dot{y}$ equation is an anti-friction term for $\beta > 0$. For large enough β, $W^u(Q,\eta)$ will cross over to the other side of $W^s(Q,\eta)$; therefore, there is a value of β' that will yield a homoclinic connection. By the symmetry of the equations, it is a double homoclinic connection.

Finally, we perturb δ to $\delta_0 > 0$, but keep the homoclinic connection. The derivative of z is given by $\dot{z} = -\gamma z + \delta x^2$; so z is positive but small along $W^u(Q,\eta)$. Thus the $-\nu xz$ term in $\dot{y}$ is a slight inward or outward push (depending on the point) along $W^u(Q,\eta)$. By adjusting the value of β' to β_0 (or adjusting α_0), we can preserve the homoclinic connection, giving (A1). The choice of α_0 and γ_0 gives (A3) and (A4).

We are left to verify (A2) and (A5). We can prove this directly for $\alpha, \beta \approx 0$. Then we use the analyticity of the manifolds to say that they cannot be identically tangent for all $\alpha_0 > 2^{-1/2}$. Thus there must be values that satisfy this inequality for which the manifold are transverse.

Before verifying (A2) for small α and β, we note more carefully the analyticity of $W^s(Q,\eta)$ and $W^{us}(Q,\eta)$. The fact that $W^s(Q,\eta)$ is analytic and varies analytically with parameters is completely standard. The proof for $W^{us}(Q,\eta)$ is not as standard. To prove that this manifold exists locally it is necessary to make an extension that is equal to the linear map outside a neighborhood of Q (or another such construction). It is not clear how to do this construction analytically. However, the tangent space of $W^{us}(Q,\eta)$ at points of $W^u(Q,\eta)$ is determined without any such extension and so depends analytically on η. By looking at $W^s(Q,\eta)$ and $W^{us}(Q,\eta)$ where they both cross $y = 0$, the angle between these two manifolds is an analytic function of η. If this is nonzero for small α, it cannot be identically equal to zero for $\alpha > 2^{-1/2}$; therefore, we can choose an $\alpha_0 > 2^{-1/2}$ for which these manifolds are transverse.

To verify the transversality given in (A2) for α and β near zero, we let $p(t)$ be the covectors that are perpendicular to $T_qW^{us}(Q)$ at points q on Γ. We write covectors as rows and vectors as columns, so that the pairing between them is just matrix multiplication. Covectors satisfy the adjoint equation to the first variation equation that is satisfied by vectors. Remember that the first variation equation for vectors is given by

$$\dot{v} = DX(q(t))v,$$

where $DX(q(t))$ is given as above, and that it is satisfied by $v(t) = X(q(t))$. The adjoint equation for covectors is obtained by differentiating the equation $C = pv$ with respect to t and obtaining $0 = \dot{p}v + pDX(q(t))v$ for all vectors v, so

$$\dot{p} = -pDX(q(t)).$$

(Note, this is the equation with p written as a row. If p^{tr} is the corresponding column vector, then we get $\dot{p}^{tr} = -[DX(q(t))]^{tr}p^{tr}$.)

We first consider $\delta = 0$, and α and β both positive but near zero and chosen so there is a homoclinic orbit. We take the parameterization of $q(t)$ with $y(0) = 0$. The equations for $\dot{p}_1$ and $\dot{p}_2$ are independent of p_3 and so can be solved independently for a solution $(\overline{p}_1(t), \overline{p}_2(t))$, that is, perpendicular in the (x,y)-plane to the homoclinic orbit. There is then a solution $\overline{p}_3(t)$ and $\overline{p}(t) = (\overline{p}_1(t), \overline{p}_2(t), \overline{p}_3(t))$, so that $(q(t), \overline{p}(t))$ lies on the unstable manifold of $(Q, 0)$ in the space of positions and covectors.

We need to determine properties of $\overline{p}_3(t)$ and $\overline{p}(t)$ as $t \to \infty$. Since $\dot{p}_3 = \nu x(t)\overline{p}_2(t) + \gamma p_3$,

$$\overline{p}_3(t) = e^{\gamma(t-t_0)}\overline{p}_3(t_0) + \nu e^{\gamma t}\int_{t_0}^{t} x(s)\overline{p}_2(s)e^{-\gamma s}\,ds.$$

As $t_0 \to -\infty$, $e^{\gamma(t-t_0)}\overline{p}_3(t_0) \to 0$ because $\overline{p}_3(t_0) \to 0$ at least at a rate of $e^{-|\lambda_{ss}||t_0|}$ (because $|\lambda_{ss}|$ is the most unstable eigenvalue for covectors) and $\gamma - |\lambda_{ss}| < 0$. Thus,

$$\overline{p}_3(t) = \nu e^{\gamma t}\int_{-\infty}^{t} x(s)\overline{p}_2(s)e^{-\gamma s}\,ds.$$

We want to show that $\overline{p}_3(t) \to \nu\infty$, as $t \to \infty$. It is easy to see that the integral converges as $t \to \infty$. For $\alpha, \beta = 0$, $x(-s) = -x(s)$, and $p_2(-s) = -p_2(s)$, so we get

$$\int_0^{\infty} x(s)\overline{p}_2(s)[e^{-\gamma s} - e^{\gamma s}]\,ds,$$

which is positive since $x(s)$ is positive and $p_2(s)$ is negative for positive s. Since this integral is positive for $\alpha, \beta = 0$, it is positive for $\alpha, \beta \approx 0$. Since it is multiplied by $\nu e^{\gamma t}$, $\bar{p}_3(t) \to \nu\infty$ as $t \to \infty$. On the other hand, $(\bar{p}_1(t), \bar{p}_2(t)) \to 0$ (as seen by the eigenvalues at $(Q,0)$). This implies that the tangent plane to $W^{us}(Q,\eta)$ limits on the (x,y)-plane and is transverse to $W^s(Q,\eta)$ for these values of η with $\delta = 0$. Transversality is an open condition and so is true for nearby small $\alpha, \beta, \delta > 0$ (for which there is a homoclinic connection), which proves (A2) for these parameter values. By analyticity, as mentioned above, we get (A2) true for some $\alpha_0 > 2^{-1/2}$.

As argued in [8], the rates of change of area near $t = \pm\infty$ show that the integral in (A5) is $-\infty$ for $\delta = 0$. Thus for small $\delta > 0$, $\log(C_{\eta_0}) << 0$; so $C_{\eta_0} << 1$. This completes the proof of Lemma 3. □

The fact that $a(\eta)$ and $b(\eta)$ can be varied independently follows because increasing β (or decreasing α) makes the manifold $W^u(Q,\eta)$ cross over $W^s(Q,\eta)$; so $a(\eta)$ varies, and varying γ makes $b(\eta)$ vary. This completes the proof of Theorem 2. □

REFERENCES

[1] V. S. Afraimovich, V. V. Bykov, and L. P. Shilnikov, *On attracting structurally unstable limit sets of Lorenz attractor type*, Trudy Moskov Mat. Obshch., 44 (1982), pp. 150–212. (In Russian.)

[2] A. Andronov, E. A. Leontovich, I. I. Gordon, and A. G. Maier, *Theory of Bifurcation of Dynamical Systems on the Plane*, John Wiley, New York, 1973.

[3] M. Hirsch and C. Pugh, *Stable manifolds and hyperbolic sets*, Proc. Symposium in Pure Math. 14 (1970), pp. 133–164, American Mathematical Society, Providence, RI.

[4] M. Hirsch, C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag New York, Berlin, 1977.

[5] G. Keller, *Generalized bounded variation and applications to piecewise monotonic transformations*, Z. Wahrsch. Verw. Gebiete, 69 (1985), pp. 461–478.

[6] C. Robinson, *Differentiability of the stable foliations for the model Lorenz equations*, Dynamical Systems and Turbulence, D. Rand and L.-S. Young, eds., Lecture Notes in Math., 898, Springer-Verlag, Berlin, New York, 1981, pp. 302–315.

[7] ———, *Transitivity and invariant measures for the geometric model of the Lorenz equations*, Ergodic Theory Dynamical Systems, 4 (1984), pp. 605–611.

[8] ———, *Homoclinic bifurcation to a transitive attractor of Lorenz type*, Nonlinearity, 2 (1989), pp. 495–518.

[9] R. Roussarie, *On the number of limit cycles which appear by perturbation of separatrix loop of planar vector fields*, Bol. Soc. Mat., 17 (1986), pp. 67–101.

[10] M. Rychlik, *Lorenz attractors through Sil'nikov-type bifurcation, Part* I, Ergodic Theory Dynamical Systems, 10 (1990), pp. 793–822.

[11] L. P. Shilnikov, *Theory of bifurcation and quasi-hyperbolic attractors*, Uspekhi Mat. Nauk, 36 (1981), pp. 240–241. (In Russian.)

[12] M. Shub, *Global Stability of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1987.

[13] F. Takens, *Partially hyperbolic fixed points*, Topology, 10 (1971), pp. 133–147.

[14] R. Williams, *The structure of Lorenz attractors*, Publ. Math. IHES, 50 (1979), pp. 321–347.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1269-1290, September 1992

SINGULAR PERTURBATIONS OF HOMOCLINIC ORBITS IN $\mathbb{R}^4$*

WIKTOR ECKHAUS†

Abstract. A rigorous asymptotic method is developed for the study of a class of singularly perturbed ODE's which in the limit $\varepsilon \downarrow 0$ have a homoclinic orbit. It is shown that the homoclinic orbit does not survive the perturbation and suffers an "exponentially small splitting."

Key words. singular perturbations, splitting of homoclinic orbits

AMS(MOS) subject classifications. 31A45, 31E20, 76D30

1. Introduction. In this paper we develop an asymptotic method of analysis for the "model problem" arising in the theory of water-waves in the presence of surface tension [2], given by the equation

$$\varepsilon^2 \frac{d^4 y}{dx^4} + \frac{d^2 y}{dx^2} - y + y^2 = 0, \tag{1.1}$$

and for the generalizations

$$\varepsilon^2 \frac{d^4 y}{dx^4} + \frac{d^2 y}{dx^2} - y + y^2 = \varepsilon^2 \mathcal{P}\left(y, \frac{dy}{dx}, \frac{d^2 y}{dx^2}; \varepsilon\right). \tag{1.2}$$

Here ε is a small parameter. The perturbation terms $\mathcal{P}$ will be specified in § 6.

In the limit for $\varepsilon = 0$, we find an integral

$$\left(\frac{dy}{dx}\right)^2 = y^2 - \frac{2}{3} y^3 + c, \tag{1.3}$$

and for $c = 0$ we find homoclinic orbit sketched in Fig. 1.

The solution $y(x)$ of the limit problem that corresponds to the homoclinic orbit tends to zero for $x \to \pm\infty$. In the context of water waves it is a solitary wave. In [2] this question was asked: *do there exist nontrivial solutions of the singularly perturbed model problem* (1.1) *which tend to zero for* $x \to \pm\infty$?

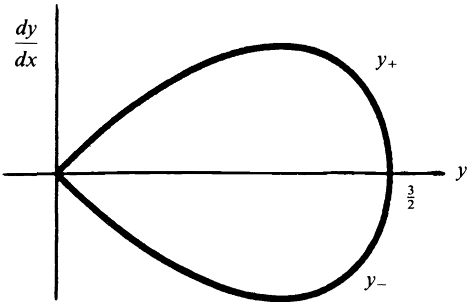


FIG. 1

* Received by the editors May 22, 1991; accepted for publication (in revised form) January 6, 1992.

† Mathematical Institute, Rijksuniversiteit Utrecht, Budapestlaan 6, Postbus 80.010, 3508 TA Utrecht, the Netherlands.

Let us briefly look at the influence of the perturbation term in (1.1) on small solution by considering the linearized equation. We then find

$$y(x) \approx e^{\omega x}, \tag{1.4}$$

$$\omega_{1.2} = \mp 1 + o(\varepsilon), \qquad \omega_{2.3} = \mp \frac{i}{\varepsilon} + o(1). \tag{1.5}$$

The perturbation thus introduces fast oscillations. The existence question can hence be rephrased as follows: *does the homoclinic orbit survive the tendency to rapid oscillations?*

In this perspective, the problem appears to be linked to the phenomenon of "exponentially small splitting of separatrices" [5], [8], which has recently drawn attention. In these publications the break-up of homoclinic orbits under influence of rapid external oscillations is studied. The mathematical analysis is highly nontrivial. This is because the quantity that we attempt to determine, the "splitting distance," is exponentially small. It is quite impressive to learn from the introduction to [5] that Poincaré already knew the phenomenon.

The nonexistence of solitary waves for the model equation (1.1) has recently been proven by Amick and McLeod [1] and by Hammersley and Mazzarino [4]. This has been done as follows.

Consider a half-orbit $y_-(x)$ that tends to zero for $x \to +\infty$, and let $x = 0$ be chosen such that $y'_-(0) = 0$. We can show that for a smooth continuation of $y_-(x)$ into a solitary wave that tends to zero for $x \to -\infty$ we need $y'''_-(0) = 0$. In [1] it has been shown that $y'''_-(0) \neq 0$. This is achieved by a rather subtle excursion into the complex plane, and no order of magnitude has been given. In [4] an ingenious combination of analysis and numerical computation is used. The results show that $y'''_-(0) \neq 0$ and indeed is exponentially small. In fact, the behaviour with ε which we shall find in § 4 if in precise agreement with the results of [4].

The original question for the model equation (1.1) has thus been settled. However, new questions arise. We list a few.

(i) For an analyst it is a challenge to establish the main result for (1.1), including the exponentially small order of magnitude estimate, without recourse to numerical computations.

(ii) What is the behaviour of the smooth continuation of $y_-(x)$ for $x < 0$?

(iii) How are we to deal with more general problems (1.2)?

In this paper we develop a method of analysis by which these questions can be answered. The organization of the paper is as follows.

In the main body of the paper we study the model equation (1.1). The problem is transformed into a second-order differential equation for the function $z(y)$, defined through

$$z(y) := \left(\frac{dy}{dx}\right)^2. \tag{1.6}$$

In § 3 formal asymptotic expansions of $z(y)$ for ε small are constructed and analysed. The proof of validity of these expansions is delayed to § 5. In § 4 we collect and discuss the main results, including the analysis of the "exponentially small splitting."

Section 5 is the mathematical backbone of the paper. We prove the validity of the formal asymptotic expansions by a contraction-mapping argument. This is a nontrivial exercise. We have made an effort to explain and motivate the various steps and tricks of the analysis in some detail. The reason is that the analysis of § 5 is not only a proof of results of § 4, but outlines a method to deal with other problems.

In § 6 we show that the analysis can be extended to large classes of perturbed model problems (1.2). On the other hand, in the companion-paper [3] we use the method of § 5 to answer question (ii), formulated above. We show there that there exist solutions $y(x)$ of (1.1) that can be called "quasi-solitary" waves. They look like solitary waves over very large distances $(-\infty, X(\varepsilon))$, or $(-X(\varepsilon), \infty)$, or $(-X(\varepsilon), X(\varepsilon))$ with $X(\varepsilon)\to\infty$ as $\varepsilon \downarrow 0$ and $y(X(\varepsilon)) = o(\varepsilon^m)$, m arbitrary positive. These solutions are approximated (with an exponentially small error) by solitary waves that have an exponentially small jump in the third derivative at $x=0$.

2. An integral and an equation for the trajectories. The perturbed problem (1.1) also has an integral. We can find it by brute force, multiplying by dy/dx and performing the integration. The perturbation term can be computed integrating by parts three times. The result is

$$\varepsilon^2\left\{\frac{dy}{dx}\frac{d^3y}{dx^3}-\frac{1}{2}\left(\frac{d^2y}{dx^2}\right)^2\right\}=-\frac{1}{2}\left(\frac{dy}{dx}\right)^2+\frac{1}{2}y^2-\frac{1}{3}y^3+c. \tag{2.1}$$

Further reduction follows by introducing y as an independent variable, using formulas such as

$$\frac{d^2y}{dx^2}=\frac{1}{2}\frac{d}{dy}\left(\frac{dy}{dx}\right)^2,\qquad \frac{d^3y}{dx^3}=\frac{1}{2}\left(\frac{dy}{dx}\right)\frac{d^2}{dy^2}\left(\frac{dy}{dx}\right)^2. \tag{2.2}$$

We finally introduce

$$\left(\frac{dy}{dx}\right)^2=z \tag{2.3}$$

and find the equation for trajectories:

$$\varepsilon^2\left\{z\frac{d^2z}{dy^2}-\frac{1}{4}\left(\frac{dz}{dy}\right)^2\right\}+z=y^2-\frac{2}{3}y^3+c. \tag{2.4}$$

One can find the integral in a more direct and elegant way by introducing y as an independent variable in (1.1). Manipulations of the type (2.2) then produce immediately

$$\frac{d}{dy}\left\{\varepsilon^2\left[z\frac{d^2z}{dy^2}-\frac{1}{4}\left(\frac{dz}{dy}\right)^2\right]+z\right\}=2(y-y^2). \tag{2.5}$$

In (2.4) we have reduced the problem from fourth- to second-order. The price is that the equation is highly nonlinear and degenerates at $z=0$. Nevertheless, the equation (2.4) will be basic in all that follows.

Remarks. The integral (2.1) also occurs in [1] and [6], but does not play a predominant role in the analysis. In [4], from (2.1) an equation equivalent to (2.4) is derived and used as a starting point of the theory. We shall comment on this further in § 3.2.

3. Formal approximations. In a search for a homoclinic orbit we put in (2.4) $c=0$. In this section we develop formal approximations for solutions of the basic equation

$$\varepsilon^2\left\{z\frac{d^2z}{dy^2}-\frac{1}{4}\left(\frac{dz}{dy}\right)^2\right\}+z=y^2\left(1-\frac{2}{3}y\right). \tag{3.1}$$

3.1. Straightforward iteration. It seems natural to look at what happens close to the unperturbed orbit

$$z_0=y^2\left(1-\frac{2}{3}y\right). \tag{3.2}$$

We introduce the transformation

$$z(y,\varepsilon)=z_0(y)+\varepsilon^2\rho_1(y,\varepsilon), \tag{3.3}$$

and obtain the equation

$$\varepsilon^2\left\{(z_0+\varepsilon^2\rho_1)\frac{d^2\rho_1}{dy^2}-\frac{1}{2}\frac{dz_0}{dy}\frac{d\rho_1}{dy}+\frac{d^2z_0}{dy^2}\rho_1-\frac{\varepsilon^2}{4}\left(\frac{d\rho_1}{dy}\right)^2\right\}+\rho_1=f_1(y) \tag{3.4}$$

with

$$f_1(y)=-z_0\frac{d^2z_0}{dy^2}+\frac{1}{4}\left(\frac{dz_0}{dy}\right)^2.$$

But now it again is natural to repeat the operation, putting

$$\begin{aligned}\rho_1(y,\varepsilon)&=z_1(y)+\varepsilon^2\rho_2(y,\varepsilon),\\ z_1(y)&=f_1(y).\end{aligned} \tag{3.5}$$

The iteration-procedure can be pursued indefinitely, leading to a formal approximation with a remainder term, in the form

$$z(y,\varepsilon)=\sum_{n=0}^{m-1}\varepsilon^{2n}z_n(y)+\varepsilon^{2m}\rho_m(y,\varepsilon). \tag{3.6}$$

Introducing the definition

$$\Phi_m=\sum_{n=0}^{m-1}\varepsilon^{2n}z_n, \tag{3.7}$$

we find a quite transparent structure:

$$z_n=-\left\{\Phi_n\frac{d^2z_{n-1}}{dy^2}-\frac{1}{2}\frac{d\Phi_{n-1}}{dy}\frac{dz_{n-1}}{dy}+\frac{d^2\Phi_{n-1}}{dy^2}z_{n-1}-\frac{1}{4}\varepsilon^{2(n-1)}\left(\frac{dz_{n-1}}{dy}\right)^2\right\}. \tag{3.8}$$

The remainder term ρ_m satisfies

$$\varepsilon^2\left\{(\Phi_m+\varepsilon^{2m}\rho_m)\frac{d^2\rho_m}{dy^2}-\frac{1}{2}\frac{d\Phi_m}{dy}\frac{d\rho_m}{dy}+\frac{d^2\Phi_m}{dy^2}\rho_m-\frac{1}{4}\varepsilon^{2m}\left(\frac{d\rho_m}{dy}\right)^2\right\}+\rho_m=f_m, \tag{3.9}$$

where f_m is defined in a formal way as z_m in (3.8).

Inspection of the formulas shows that the formal approximation Φ_m has the structure

$$\Phi_m=y^2\left[1-\frac{y}{\alpha_m(\varepsilon)}\right]g_m(y,\varepsilon), \tag{3.10}$$

where g_m is a polynomial in y without zeros on the interval $[0,\alpha_m]$. $\alpha_m(\varepsilon)$ can be computed explicitly. In the first approximations, we find

$$\alpha_m(\varepsilon)=\tfrac{3}{2}[1+\tfrac{1}{4}\varepsilon^2+\cdots]. \tag{3.11}$$

Let us now recall that $z=(dy/dx)^2$. Consider an orbit defined as a union of the curves $(dy/dx)_\pm$, given by the formal approximation

$$\left(\frac{dy}{dx}\right)_\pm=\pm y\sqrt{1-\frac{y}{\alpha_m(\varepsilon)}}\,g_m^{1/2}(y,\varepsilon). \tag{3.12}$$

It is easily verified that the corresponding solution $y(x)$ has continuous derivatives up to the fourth order (and beyond) at $y=\alpha_m(\varepsilon)$. Hence, the formal approximation Φ_m, for any m, has all the desired features of a homoclinic orbit of (1.1). This result is encouraging, but will turn out to be misleading.

3.2. Power series. It is remarkable that we can also compute a formal power series expansion for $z(y)$, which, at the same time, is an asymptotic expansion in ε. To show the structure, a small display of the algebra is needed. We abbreviate

$$Lz := \varepsilon^2\left\{z\frac{d^2z}{dy^2}-\frac{1}{4}\left(\frac{dz}{dy}\right)^2\right\}+z-y^2+\frac{2}{3}y^3 \tag{3.13}$$

and introduce

$$\tilde{\Phi}_m(y)=\alpha y^2-\beta y^3-\sum_{n=1}^{m} a_n y^{n+3}. \tag{3.14}$$

Substitution produces the following result:

$$\begin{aligned} L\tilde{\Phi}_m &= (\varepsilon^2\alpha^2+\alpha-1)y^2+\left[\frac{2}{3}-\beta(1+5\varepsilon^2\alpha)\right]y^3 \\ &\quad -y^4\sum_{n=1}^{m} a_n[1+\varepsilon^2\alpha[(n+3)(n+2)+2]]y^{n-1} \\ &\quad +\varepsilon^2y^4\left\{\frac{15}{4}\beta^2+y\beta\sum_{n=1}^{m} a_n\left[(n+3)\left(n+\frac{1}{4}\right)+6\right]y^{n-1}\right. \\ &\quad \left. +y^2\left[\sum_{n=1}^{m} a_n(n+3)\left(n+\frac{3}{2}\right)y^{n-1}\right]\left[\sum_{n=1}^{m} a_n(n+3)\left(n+\frac{5}{4}\right)y^{n-1}\right]\right\}. \end{aligned} \tag{3.15}$$

We can now determine α, β, and a_n, $n=1,\cdots,m$ by putting all coefficients of y^p, $p=2,\cdots m+3$ on the right-hand side of (3.15) equal to zero.

The first requirement is that α should be the positive root of

$$\varepsilon^2\alpha^2+\alpha-1=0. \tag{3.16}$$

This means that α^2 is the exact value of $\omega_{1,2}^2$ in (1.5). Next we find

$$\beta=\frac{2}{3}\frac{1}{1+5\varepsilon^2\alpha}, \tag{3.17}$$

$$a_1=\varepsilon^2\frac{15}{4}\beta^2\frac{1}{1+\varepsilon^2 14\alpha}. \tag{3.18}$$

Further coefficients, a_n, $n=2,\cdots,m$, can be determined recursively. A general recursion formula is of little use to us. It is sufficient to observe (from the structure of the right-hand side of (3.15)) that

$$a_n=O(\varepsilon^{2n}),\qquad a_n>0. \tag{3.19}$$

We summarize these results as follows:

$$\tilde{\Phi}_m(y)=\alpha y^2-\beta y^3-\sum_{n=1}^{m}\varepsilon^{2n}\bar{a}_n y^{n+3}, \tag{3.20}$$

$$\bar{a}_n=O(1),\qquad \bar{a}_n>0, \tag{3.21}$$

$$L\tilde{\Phi}_m=\varepsilon^{2(m+1)}y^{m+4}\sum_{p=0}^{m}\sigma_p y^p, \tag{3.22}$$

$$\sigma_p=O(1),\qquad \sigma_p>0. \tag{3.23}$$

In a final step we write

$$z = \tilde{\tilde{\Phi}}_m + \varepsilon^{2(m+2)} \tilde{\rho}_m. \tag{3.24}$$

We can easily see that for the remainder $\tilde{\rho}_m$ an equation similar to (3.9) is obtained. On the other hand, the formal approximation $\tilde{\tilde{\Phi}}_m$ is similar in structure to Φ_m, as expressed in (3.10). Working with the power-series expansion may have some advantages. This we shall see when studying the problem of splitting of the homoclinic orbit.

Remarks. In the work of Hammersley and Mazzarino [4] a formal power series of a structure similar to (3.14) (with $m = \infty$) is used as a basic step of the analysis. A recursion formula for the coefficients is given, and convergence results are derived. The authors do not recognize the asymptotic structure of the series for ε small. As already mentioned, Hammersley and Mazzarino use an ingenious combination of analysis and numerical computations. Their final results are obtained at the expense of a very impressive amount of both formula-manipulation and numerics.

4. Main results.

4.1. The solution for $z(y)$. Let us state again our basic equation

$$\varepsilon^2 \left\{ z \frac{d^2 z}{dy^2} - \frac{1}{4} \left(\frac{dz}{dy} \right)^2 \right\} + z = y^2 \left(1 - \frac{2}{3} y \right). \tag{4.1}$$

We look for solutions $z(y)$, which for $y \to 0$ behave as y^2. To this end we introduce the transformation

$$z = y^2 \bar{z} \tag{4.2}$$

and obtain for $\bar{z}$ the equation

$$\varepsilon^2 \left\{ y^2 \left[\bar{z} \frac{d^2 \bar{z}}{dy^2} - \frac{1}{4} \left(\frac{d\bar{z}}{dy} \right)^2 \right] + 3y\bar{z} \frac{d\bar{z}}{dy} \right\} + \bar{z}(1 + \varepsilon^2 \bar{z}) = 1 - \frac{2}{3} y. \tag{4.3}$$

Next we write

$$\bar{z} = \varphi_m + \varepsilon^{2m} \psi_m,$$

where φ_m is the result of formal iteration of § 3.1, or the truncated power series of § 3.2, in both cases with y^2 factored out. For the remainder ψ_m we get the equation

$$\begin{aligned} &\varepsilon^2 y^2 (\varphi_m + \varepsilon^{2m} \psi_m) \frac{d^2 \psi_m}{dy^2} \\ &\quad + \varepsilon^2 \left[-\frac{1}{2} y^2 \left(\varphi'_m + \frac{1}{2} \varepsilon^{2m} \psi'_m \right) + 3y(\varphi_m + \varepsilon^{2m} \psi_m) \right] \frac{d\psi_m}{dy} + \psi \\ &= \bar{f}_m(y) - \varepsilon^2 \{ y^2 \varphi''_m + 3y\varphi'_m + (2\varphi_m + \varepsilon^{2m} \psi_m) \} \psi_m, \end{aligned} \tag{4.4}$$

where $\bar{f}_m(y)$ is a polynomial in y. In the equation (4.4) we use both primes and d/dy to denote the derivative. The reason for this will become clear in § 5, where we shall prove (by a contraction mapping argument) the following.

RESULT. *There exists a unique solution $\psi_m(y; \varepsilon)$, which on an interval $y \in [0, y_0]$, is bounded for $\varepsilon \downarrow 0$, and the same is true for the derivative $\psi'_m(y; \varepsilon)$. y_0 is such that*

$$|\varphi_m(y; \varepsilon)| \geqq c\varepsilon^m$$

with c a constant and $m \geqq 2$.

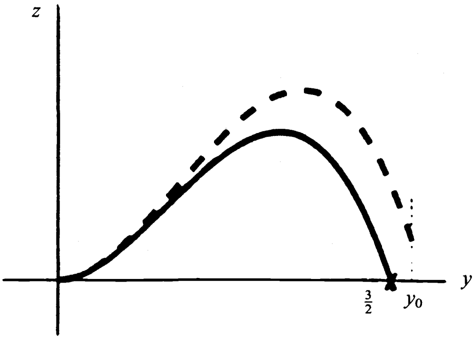


FIG. 2

Remarks. Note that m is an arbitrary integer. Hence we have the existence of a solution $z(y;\,\varepsilon)$ of the basic equation (4.1), which starts out as y^2 and ultimately gets closer to zero than any power of ε. The situation is sketched in Fig. 2.

It is further important to remark that $y_0 > 3/2$. This follows from the result (3.11).

What happens to the continuation of the solution? Obviously $z(y)$ cannot just stop at some positive value. Suppose that the function would like to turn upward and escape to large values. This could only happen at $y > y_0$. At some $z > 0$ we would have $dz/dy = 0$ and $d^2z/dy^2 \geqq 0$, and this leads to contradiction in the equation (4.1). Suppose next that $z \to 0$ as $y \to y_1 > y_0$; however, z' also tends to zero. Since z' comes from negative values, and did not pass through zero, we must still have $z'' \geqq 0$, which again gives contradiction in (4.1). The contradiction remains if we take $y_1 = +\infty$, or assume that z tends to a nonzero positive value as $y \to \infty$.

Hence the continuation of the solution $z(y)$ must reach $z = 0$ with a nonzero slope.

4.2. Splitting of the homoclinic orbit. We define two half-orbits $y_+(x), x \in (-\infty, 0]$, $y_-(x)$, $x \in [0, \infty)$ as solutions of

$$\left(\frac{dy}{dx}\right)_\pm = \pm\sqrt{z(y)}, \tag{4.5}$$

$$y_+(0) = y_-(0) = y_1(\varepsilon). \tag{4.6}$$

It is understood that

$$z(y_1) = 0. \tag{4.7}$$

We wish to investigate the regularity of the union of $y_+(x)$ and $y_-(x)$ at $x = 0$. Since the model equation is of fourth order we need continuity up to the fourth derivative. First we compute

$$\left(\frac{d^2y}{dx^2}\right)_\pm = (\pm)\frac{1}{2}\frac{1}{\sqrt{z}}\frac{dz}{dy}\left(\frac{dy}{dx}\right)_\pm = \frac{1}{2}\frac{dz}{dy}. \tag{4.8}$$

Hence

$$\left(\frac{d^2y}{dx^2}\right)_+ = \left(\frac{d^2y}{dx^2}\right)_- \quad \text{at } x = 0. \tag{4.9}$$

From the equation (1.1), it follows that fourth derivative also match continuously.

Next we compute

$$\left(\frac{d^3y}{dx^3}\right)_\pm = (\pm)\frac{1}{2}\sqrt{z}\frac{d^2z}{dy^2}. \tag{4.10}$$

We would have the continuity of the third derivative if we could show that d^2z/dy^2 is bounded as $y \to y_1$. There is, however, no reason for such assertion. In fact, in our companion paper [3, § 4], we show that

$$z(y) \approx c_1(y_1 - y) + c_2(y_1 - y)^{3/2} + \cdots \quad \text{as } y \uparrow y_1.$$

On the other hand, from (1.1) we can deduce another representation of the third derivative at $x = 0$, as follows.

First we "solve" for the second derivative and obtain

$$\left(\frac{d^2y}{dx^2}\right)_+ = -\frac{1}{\varepsilon} \int_{-\infty}^{x} \sin \frac{1}{\varepsilon}(x - \xi)[y(\xi) - y^2(\xi)]\, d\xi. \tag{4.11}$$

Next, after differentiation, we get

$$\frac{d^3y_+}{dx^3}(0) = I(\varepsilon), \tag{4.12}$$

$$I(\varepsilon) := \frac{1}{\varepsilon^2} \int_{-\infty}^{0} \cos \frac{1}{\varepsilon} \xi \cdot [y(\xi) - y^2(\xi)]\, d\xi. \tag{4.13}$$

The idea is to compute the integral along approximate trajectories, which are obtained by replacing $z(y)$ in (4.5) by its asymptotic approximations.

We define $y^{(m)}(x)$, $x < 0$ through

$$\frac{dy^{(m)}}{dx} = \sqrt{\Phi_m(y; \varepsilon)} \tag{4.14}$$

with Φ_m the asymptotic expansion defined in § 3.1. Furthermore,

$$I^{(m)}(\varepsilon) = \frac{1}{\varepsilon^2} \int_{-\infty}^{0} \cos \frac{1}{\varepsilon} \xi [y^{(m)}(\xi) - (y^{(m)}(\xi))^2]\, d\xi. \tag{4.15}$$

In the appendix, which is due to Temme, we have computed the integrals for $m = 1$ and $m = 2$, which correspond to

$$\Phi_m = \sum_{n=0}^{m-1} \varepsilon^{2n} z_n(y), \qquad m = 1, 2. \tag{4.16}$$

The results can be summarized as follows:

$$I^{(m)}(\varepsilon) = -\frac{\pi}{\varepsilon^5} c_m e^{-\pi/\varepsilon}[1 + o(1)], \tag{4.17}$$

$$c_1 = 6, \qquad c_2 \cong 3.6 c_1. \tag{4.18}$$

We see that a small correction of the trajectories does not result in a small change in the constant c_m. Before further analysis, we shall compare the results with those of Hammersley and Mazzarino [4]. Adjusting the notation of [4] to ours (ε in [4] is ε^2 here, and the quantity computed in [4] is $d^2y_-/dx^3(0)$), we find

$$I(\varepsilon) = -\Omega(\varepsilon) \frac{(1 + \varepsilon^2)^{7/4}}{\varepsilon^5} e^{-\pi/\varepsilon}. \tag{4.19}$$

The function $\Omega(\varepsilon)$ is positive. It has been computed numerically over a wide range of ε. For ε small $\Omega(\varepsilon)$ is of the order of magnitude of 10^2.

It appears that the behaviour with ε given in (4.17) is in precise agreement with (4.19). We may venture that for large m the constants c_m will settle to a definite value, but analytical determination seems beyond possibility. However, we are able to prove that for all m the integral $I^{(m)}$ is negative. This can be done using the asymptotic expansion defined by the power-series of § 3.2. The reasoning is amusing.

We recall some relevant formulas from § 3.2:

$$Lz := \varepsilon^2\left(z\frac{d^2z}{dy^2}-\frac{1}{4}\left(\frac{dz}{dy}\right)^2\right)+z-y^2+\frac{2}{3}y^3, \tag{4.20}$$

$$\tilde{\Phi}_m(y) := \alpha y^2+\beta y^3-\sum_{n=1}^{m}\varepsilon^{2n}\bar{a}_n y^{n+3}. \tag{4.21}$$

The coefficients α, β, and $\bar{a}_n$ can be determined in such way that

$$L\tilde{\Phi}_m = \varepsilon^{2(m+1)}f_m(y), \tag{4.22}$$

$$f_m(y) = y^{m+4}\sum_{p=0}^{m}\sigma_p y^p, \tag{4.23}$$

$$\sigma_p = O(1), \qquad \sigma_p > 0. \tag{4.24}$$

Let us look at $\tilde{y}^{(m)}(x)$, a solution for $x<0$ of the equation

$$\frac{d\tilde{y}^{(m)}}{dx} = \sqrt{\tilde{\Phi}^{(m)}(y^{(m)})}. \tag{4.25}$$

Because of (4.22), $\tilde{y}^{(m)}$ satisfies a perturbed model equation

$$\varepsilon^2\frac{d^4\tilde{y}^{(m)}}{dx^4}+\frac{d^2\tilde{y}^{(m)}}{dx^2} = \tilde{y}^{(m)}-[\tilde{y}^{(m)}]^2+\varepsilon^{2(m+1)}f'_m(y^{(m)}) \tag{4.26}$$

with

$$f'_m(y) = \frac{d}{dy}f_m(y). \tag{4.27}$$

Following steps analogous to (4.11), (4.12), and (4.13), we find

$$\frac{d^3\tilde{y}_+^{(m)}}{dx^3}(0) = \frac{1}{\varepsilon^2}\int_{-\infty}^{0}\cos\frac{1}{\varepsilon}\,\xi\{\tilde{y}^{(m)}-[\tilde{y}^{(m)}]^2+\varepsilon^{2(m+1)}f'_m(\tilde{y}^{(m)})\}\,d\xi. \tag{4.28}$$

However, the quantity on the left-hand side of (4.28) is zero because $\tilde{\Phi}^{(m)}$ is just a finite polynomial, and hence the second derivative which appears in (4.10) exists. We thus find

$$\begin{aligned} I^{(m)}(\varepsilon) &= -\varepsilon^{2m}\int_{-\infty}^{0}\cos\frac{1}{\varepsilon}\,\xi f'_m(\tilde{y}^{(m)}(\xi))\,d\xi \\ &= -\varepsilon^{2m}\sum_{p=0}^{m}\sigma_p(m+4+p)\int_{-\infty}^{0}\cos\frac{1}{\varepsilon}\,\xi[\tilde{y}^{(m)}(\xi)]^{m+3+p}\,d\xi. \end{aligned} \tag{4.29}$$

The next claim is that each of the integrals on the right-hand side of (4.29) is positive. This is because $\tilde{y}^{(m)}(\xi)$ is a monotone function. If we divide the interval of integration into subintervals $[-2\pi/\varepsilon, 0]$, $[-4\pi/\varepsilon, -2\pi/\varepsilon]$, etc., we get positive contributions from each subinterval. Finally, from (4.24), $\sigma_p > 0$.

Hence

$$I^{(m)}(\varepsilon) < 0 \quad \forall m.$$

5. Construction of a contraction mapping. In this section we simplify notations by dropping the subscript m.

5.1. Difficulties of the enterprise. Equation (4.4) is of the following structure:

$$\mathscr{L}(\psi)=\bar{f}(y)+\varepsilon^2 g_1(y,\psi), \tag{5.1}$$

$$\mathscr{L}=\varepsilon^2 A(y,\psi)\frac{d^2}{dy^2}+\varepsilon^2 B(y,\psi,\psi')\frac{d}{dy}+1, \tag{5.2}$$

$$A(y,\psi)=y^2(\varphi+\varepsilon^{2m}\psi), \tag{5.3}$$

$$B(y,\psi,\psi')=-\tfrac{1}{2}y^2(\varphi'+\tfrac{1}{2}\varepsilon^{2m}\psi')+3y(\varphi+\varepsilon^{2m}\psi), \tag{5.4}$$

$$g_1(y,\psi)=-\{y^2\varphi''+3y\varphi'+(2\varphi+\varepsilon^{2m}\psi)\}\psi. \tag{5.5}$$

We wish to prove the existence of a bounded solution. A standard way to proceed would be a decomposition of $\mathscr{L}$ into

$$\mathscr{L}=\mathscr{L}_1+\mathscr{L}_2 \tag{5.6}$$

such that an inverse $\mathscr{L}_1^{-1}$ could be defined and the problem reformulated into

$$\psi=\mathscr{L}_1^{-1}\{-\mathscr{L}_2(\psi)+\bar{f}+\varepsilon^2 g_1\}. \tag{5.7}$$

The right-hand side of (5.7) should define a neat mapping in some Banach space, and we should have sufficient knowledge of $\mathscr{L}_1^{-1}$ to investigate and prove contraction properties of the mapping.

In order to have sufficient control over $\mathscr{L}_1^{-1}$, we shall use for $\mathscr{L}_1$ a linear operator. This means, however, that on the right-hand side of (5.7) nonlinear terms will appear containing first and second derivatives of ψ. It turns out that the second derivative can be eliminated (integrating by parts), but the first derivative cannot be avoided. It is for this reason that we must also study the problem for the derivative ψ'. Differentiating (5.1) we find

$$\tilde{\mathscr{L}}(\psi')=\bar{f}'(y)+\varepsilon^2 g_2(y,\psi,\psi'), \tag{5.8}$$

$$\tilde{\mathscr{L}}=\varepsilon^2 A(y,\psi)\frac{d^2}{dy^2}+\varepsilon^2\tilde{B}(y,\psi,\psi')\frac{d}{dy}+1 \tag{5.9}$$

with $A(y,\psi)$ as in (5.3), and furthermore

$$\tilde{B}=5y(\varphi+\varepsilon^{2m}\psi)+\tfrac{1}{2}y^2(\varphi'+\varepsilon^{2m}\psi'), \tag{5.10}$$

$$\begin{aligned} g_2={}&-(y^2\varphi'''+5y\varphi''+5\varphi')\psi\\ &-\{5(\varphi+\varepsilon^{2m}\psi)+6y(\varphi'+\varepsilon^{2m}\psi')+y^2\varphi''-\tfrac{5}{2}\varepsilon^{2m}\psi'\}\psi'. \end{aligned} \tag{5.11}$$

Decomposing $\tilde{\mathscr{L}}$ with a linear operator we again face difficulties with nonlinear terms containing derivatives in the equivalent of (5.7). The second derivative will be eliminated by a special trick, and the first derivative will be eliminated by integration by parts. This is possible because in (5.8) we do not have a term of the structure $(d\psi'/dy)^2$.

The first task now is to define a suitable operator $\mathscr{L}_1$.

5.2. An intermezzo on "exact WKB functions." Physicists are familiar with the WKB approximation (Wentzel, Kramers, and Brillouin). In the context of equations that are of interest to us, the procedure goes as follows.

Consider a homogeneous equation

$$\varepsilon^2 a \frac{d^2\theta}{dy^2} + \varepsilon^2 b \frac{d\theta}{dy} + \theta = 0, \tag{5.12}$$

where a and b are given functions. We introduce

$$\theta = e^{1/\varepsilon Q}. \tag{5.13}$$

We get

$$aQ'^2 + \varepsilon a Q'' + \varepsilon b Q' + 1 = 0. \tag{5.14}$$

Next we introduce a formal expansion

$$Q = q_0 + \varepsilon q_1 + \varepsilon^2 q_2 + \cdots. \tag{5.15}$$

We readily find that

$$(q_0')^2 = -\frac{1}{a}, \tag{5.16}$$

$$q_1' = \frac{1}{4}\frac{a'}{a} - \frac{1}{2}\frac{b}{a}. \tag{5.17}$$

The result is more transparent if we introduce a function ν defined by

$$\nu' = -\frac{1}{2}\frac{b}{a}\nu. \tag{5.18}$$

This function appears in the well-known Liouville transformation for the equation (5.12). We find, as a formal WKB approximation,

$$\theta_0(y) = \nu(y)[a(y)]^{1/4} \exp\left\{ \mp \frac{i}{\varepsilon} \int^y \frac{d\eta}{\sqrt{a(\eta)}} \right\}. \tag{5.19}$$

If the coefficients $a(y)$, $b(y)$ are sufficiently regular, then we can prove for the linearly independent solutions of (5.12) the result

$$\theta = \theta_0\{1 + o(\varepsilon)\}. \tag{5.20}$$

This is a very nice result because it gives an approximation with a relative precision.

In our case, (5.3), (5.4), (5.10), the coefficients vanish at $y = 0$, and the proof of validity appears difficult. Additional complication arises for the values of y such that $\varphi(y)$ becomes small. However, *we shall not need any proof of validity in the approach that we shall follow now.*

Let us ask this question: *what is the differential equation that is satisfied exactly by the* WKB *approximation*? We answer that question for the linearization of $\tilde{\mathscr{L}}$. In that case we have

$$a = y^2\varphi, \tag{5.21}$$

$$b = 5y\varphi + \tfrac{1}{2}y^2\varphi', \tag{5.22}$$

and we find

$$\theta_0 = \frac{1}{y^2} \exp\left\{ \mp \frac{i}{\varepsilon} \int_y \frac{d\eta}{\eta\sqrt{\varphi}} \right\}. \tag{5.23}$$

Next a straightforward computation shows that

$$\varepsilon^2\left\{a\frac{d^2\theta_0}{dy^2}+b\frac{d\theta_0}{dy}\right\}+\theta_0=\varepsilon^2[4\varphi+y\varphi']\theta_0. \tag{5.24}$$

Hence, θ_0 *satisfies a very mildly perturbed original equation* (5.12).

It is interesting to observe that very similar results hold for the full nonlinear operator $\tilde{\mathscr{L}}$ of (5.9). In fact, everything remains true if in (5.21)-(5.24) the function φ is replaced by $\varphi+\varepsilon^{2m}\psi$. This result seems elegant; however, difficulties do arise if we attempt to use it in the construction of a contraction mapping.

We now define, for further use, the operator $\mathscr{L}_1$ explicitly:

$$\mathscr{L}_1=\varepsilon^2y^2\varphi\frac{d^2}{dy^2}+\varepsilon^2\left[5y\varphi+\frac{1}{2}y^2\varphi'\right]\frac{d}{dy}+[1+\varepsilon^2(4\varphi+y\varphi')]\cdot I. \tag{5.25}$$

The two linearly independent solutions of the homogeneous equation $\mathscr{L}_1\theta_0=0$ are given by (5.23).

Remark 1. We can perform similar analysis on the linearization of the operator $\mathscr{L}$ of the problem for ψ, given in (5.2). The equation satisfied by the corresponding WKB approximations turns out to be less tractable. However, in the decomposition of $\mathscr{L}$ we can also use $\mathscr{L}_1$. This introduces first derivatives in $\mathscr{L}_2$, which are unavoidable anyhow.

Remark 2. In the approach described above we introduce an additional perturbation such that the resulting equation is satisfied exactly by the WKB functions. This author found, to his surprise, no trace of this simple approach in the literature.

5.3. Transformation to integral equations. Let us consider an inhomogeneous problem

$$\mathscr{L}_1(\hat{\psi})=R \tag{5.26}$$

with $\mathscr{L}_1$ given by (5.25). We ask for bounded solutions on some nonempty interval $y\in[0,y_0]$, $y_0>0$.

Using the linearly independent solutions given by (5.23), we find, by elementary manipulations,

$$\hat{\psi}(y)=\frac{1}{\varepsilon^2}\cdot\left(\frac{\varepsilon}{2i}\right)\int_0^y\{\theta_0^{(1)}(\eta)\theta_0^{(2)}(y)-\theta_0^{(1)}(y)\theta_0^{(2)}(\eta)\}\eta^3\frac{1}{\sqrt{\varphi(\eta)}}R(\eta)\,d\eta. \tag{5.27}$$

To get a more explicit form we define

$$\Omega(\eta,y)=\int_y^\eta\frac{d\bar{\eta}}{\bar{\eta}\sqrt{\varphi(\bar{\eta})}} \tag{5.28}$$

and obtain

$$\hat{\psi}(y)=-\frac{1}{\varepsilon}\frac{1}{y^2}\int_0^y\sin\left[\frac{1}{\varepsilon}\Omega(\eta,y)\right]\frac{\eta}{\sqrt{\varphi(\eta)}}R(\eta)\,d\eta. \tag{5.29}$$

Let us first show that $\hat{\psi}(y)$ indeed is bounded as $y\to0$. By elementary estimate we get

$$|\hat{\psi}(y)|\leqq\frac{1}{\varepsilon}\frac{1}{y^2}\int_0^y\frac{\eta}{\sqrt{\varphi(\eta)}}d\eta\sup_{y\in[0,y_0]}|R(y)|. \tag{5.30}$$

The estimate is valid on intervals $y\in[0,y_0]$ such that $\varphi(y)\geqq0$.

From this it follows that

$$|\hat{\psi}(y)|\leqq\frac{c}{\varepsilon}\sup|R|. \tag{5.31}$$

Next we show that if R is differentiable, then $\hat{\psi}(y)$ is $O(1)$ as $\varepsilon \downarrow 0$. For this it is sufficient to observe that

$$-\frac{1}{\varepsilon}\sin\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\cdot\frac{1}{\eta\sqrt{\varphi(\eta)}}=\frac{d}{d\eta}\cos\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]. \tag{5.32}$$

It then follows that

$$\hat{\psi}(y)=R(y)-\frac{1}{y^2}\int_0^y \cos\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{d}{d\eta}[\eta^2 R(\eta)]\,d\eta. \tag{5.33}$$

The integral on the right-hand side of (5.33) can be estimated as in (5.29) and (5.30).

Finally, if $R(\eta)$ is twice differentiable, then

$$\hat{\psi}(y)=R(y)+O(\varepsilon). \tag{5.34}$$

In fact, explicitly,

$$\hat{\psi}(y)=R(y)+\frac{\varepsilon}{y^2}\int_0^y \sin\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{d}{d\eta}\left[\eta\sqrt{\varphi}\frac{d}{d\eta}\eta^2 R\right]d\eta. \tag{5.35}$$

With these results we now turn to the problems for ψ and ψ'.

It will be convenient to have a compact notation for the integral operators that occur in the analysis. We define

$$S_1[R](y)=-\frac{1}{y^2}\int_0^y \sin\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{\eta}{\sqrt{\varphi(\eta)}}R(\eta)\,d\eta, \tag{5.36}$$

$$S_2[R](y)=-\frac{1}{y^2}\int_0^y \cos\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{\eta}{\sqrt{\varphi(\eta)}}R(\eta)\,d\eta. \tag{5.37}$$

For intervals $y\in[0, y_0]$ such that $\varphi(y)\geqq 0$ we have

$$|S_{1,2}[R]|\leqq c\ \mathrm{Sup}\,|R|. \tag{5.38}$$

Furthermore, if $R(\eta)$ is twice differentiable, then

$$\frac{1}{\varepsilon}S_1[R](y)=R(y)+O(\varepsilon). \tag{5.39}$$

5.3.1. The problem for ψ.

Starting from (5.1)-(5.5) and using the definition of $\mathscr{L}_1$ given in (5.25), we find after some arithmetic exercises that

$$\mathscr{L}_1(\psi)=\bar{f}(y)+\varepsilon^2 G(y, \psi, \psi')-\varepsilon^{2(m+1)}y^2\psi\frac{d^2\psi}{dy}, \tag{5.40}$$

where $G(y, \psi, \psi')$ is an expression of the structure

$$G=\beta_0(y)\psi+\gamma_0(y)\psi'+\varepsilon^{2m}[\beta_1(y)\psi+\gamma_1(y)\psi']\psi'. \tag{5.41}$$

The β's and γ's are polynomials in y. We shall use the symbol G generically, to denote expressions of the structure (5.41). Of course, the G's occuring in different terms have, in general, different polynomials.

Using the results the equations (5.40) is transformed to

$$\begin{aligned}\psi&=\frac{1}{\varepsilon}S_1[\bar{f}(\eta)]+\varepsilon S_1[G(\eta, \psi, \psi')]\\&\quad+\varepsilon^{2m+1}\frac{1}{y^2}\int_0^y \sin\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{\eta^3}{\sqrt{\varphi(\eta)}}\psi\frac{d^2\psi}{d\eta^2}\,d\eta.\end{aligned} \tag{5.42}$$

Integrating the last term by parts, we find it to be equal to

$$-\varepsilon^{2m+1}\frac{1}{y^2}\int_0^y \sin\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{d}{d\eta}\left[\frac{\eta^3}{\sqrt{\varphi}}\psi\right]\psi'\,d\eta$$

$$-\varepsilon^{2m}\frac{1}{y^2}\int_0^y \cos\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{\eta^2}{\varphi}\psi\psi'\,d\eta.$$

Separating out terms that are different in structure from $S_1[G]$, we obtain the following result:

$$\psi=\psi_0+\varepsilon S_1[G(\eta, \psi, \psi')]-\varepsilon^{2m}S_2\left[\frac{\eta}{\sqrt{\varphi}}\psi\psi'\right]+\frac{1}{2}\varepsilon^{2m+1}S_1\left[\frac{\eta^2}{\varphi}\varphi'\psi\psi'\right], \tag{5.43}$$

where

$$\psi_0=\frac{1}{\varepsilon}S_1[\bar{f}]=\bar{f}+O(\varepsilon). \tag{5.44}$$

The reason for separating out carefully terms of different structure lies in the fact that we want to pursue the analysis as far as possible into values of y such that $\varphi(y)$ becomes very small.

5.3.2. The problem for ψ'. We now turn to (5.8)-(5.11). The first step is to remove the nonlinearity in the term with second derivative of ψ'. This can be done by multiplying the equation by the factor

$$\frac{\varphi}{\varphi+\varepsilon^{2m}\psi}$$

and by using, whenever convenient, the identity

$$\frac{\varphi}{\varphi+\varepsilon^{2m}\psi}=1-\frac{\varepsilon^{2m}\psi}{\varphi+\varepsilon^{2m}\psi}.$$

The explicit result is

$$\begin{aligned}&\varepsilon^2y^2\varphi\frac{d^2}{dy^2}\psi'+\varepsilon^2\left(5y\varphi+\frac{1}{2}y^2\psi'\right)\frac{d\psi'}{dy}+\psi'\\ &\quad=[1-\varepsilon^{2m}\psi(\varphi+\varepsilon^{2m}\psi)^{-1}](\bar{f}'+\varepsilon^2g_2)+\varepsilon^{2m}\psi(\varphi+\varepsilon^{2m}\psi)^{-1}\psi'\\ &\quad\quad+\varepsilon^{2m+2}(\varphi+\varepsilon^{2m}\psi)^{-1}\frac{1}{2}y^2(\varphi'\psi-\varphi\psi')\frac{d\psi'}{dy}\end{aligned} \tag{5.45}$$

with g_2 given by (5.11). Next we introduce $\mathscr{L}_1$ by definition (5.25), and we use the generic notation G for terms of the structure (5.41). This produces the equation

$$\begin{aligned}\mathscr{L}_1(\psi')&=\bar{f}'+\varepsilon^2G_1(y, \psi, \psi')+\varepsilon^{2m}(\varphi+\varepsilon^{2m}\psi)^{-1}\psi G_2(y, \psi, \psi')\\ &\quad+\varepsilon^{2m+2}(\varphi+\varepsilon^{2m}\psi)^{-1}\frac{1}{4}y^2\left[2\varphi'\psi\frac{d\psi'}{dy}-\varphi\frac{d}{dy}(\psi')^2\right].\end{aligned} \tag{5.46}$$

When inverting the equation (5.46), we have to deal with the term

$$-\varepsilon^{2m+1}\frac{1}{y^2}\frac{1}{4}\int_0^y \sin\left[\frac{1}{\varepsilon}\Omega(\eta, y)\right]\frac{\eta^3}{\sqrt{\varphi}}(\varphi+\varepsilon^{2m}\psi)^{-1}\left[2\varphi'\psi\frac{d\psi'}{d\eta}-\varphi\frac{d}{d\eta}(\psi')^2\right]d\eta.$$

Integrating by parts, we find this term to be equal to

$$\begin{aligned}&-\varepsilon^{2m}S_2\left\{\frac{1}{4}\frac{\eta^2}{\sqrt{\varphi}}(\varphi+\varepsilon^{2m}\psi)^{-1}(2\varphi'\psi-\varphi\psi)\psi\right\}\\&-\varepsilon^{2m+1}S_1\left\{\frac{3}{4}\eta(\varphi+\varepsilon^{2m}\psi)^{-1}(2\varphi'\psi-\varphi\psi')\psi'\right\}\\&+\varepsilon^{2m+1}S_1\left\{\frac{1}{8}\frac{\eta^2}{\varphi}(\varphi+\varepsilon^{2m}\psi)^{-1}[2\varphi'\psi-\varphi\psi']\psi'\right\}\\&+\varepsilon^{4m+1}S_1\left\{\frac{1}{4}\eta^2(\varphi+\varepsilon^{2m}\psi)^{-2}(2\varphi'\psi-\varphi\psi')(\psi')^2\right\}.\end{aligned}$$

The expression looks very complicated. However, all that matters is the structure. We have to extend somewhat the structure defined in (5.41), and introduce for that purpose

$$\tilde{G}(y,\psi,\psi')=\sum_{l+q=1}^{l+q=3}P_{l,q}(y)(\psi)^l(\psi')^q, \tag{5.47}$$

where $P_{l,q}(y)$ are polynomials in y, which depend on ε in a regular way.

Inverting now fully the equation (5.46) and collecting terms of the same structure, we get

$$\begin{aligned}\psi'=\psi_0'&+\varepsilon S_1[\tilde{G}_1(\eta,\psi,\psi')]+\varepsilon^{2m-1}S_1[(\varphi+\varepsilon^{2m}\psi)^{-1}\tilde{G}_2(\eta,\psi,\psi')]\\&+\varepsilon^{2m}S_2[\varphi^{-1/2}(\varphi+\varepsilon^{2m}\psi)^{-1}\tilde{G}_3(\eta,\psi,\psi')]\\&+\varepsilon^{2m+1}S_1[\varphi^{-1}(\varphi+\varepsilon^{2m}\psi)^{-1}\tilde{G}_4(\eta,\psi,\psi')]\\&+\varepsilon^{4m+1}S_1[(\varphi+\varepsilon^{2m}\psi)^{-2}\tilde{G}_5(\eta,\psi,\psi')],\end{aligned} \tag{5.48}$$

where

$$\psi_0'=\frac{1}{\varepsilon}S_1[\bar{f}^1 1]=\bar{f}^1+O(\varepsilon). \tag{5.49}$$

5.4. The contraction mapping. We collect our result in the following form:

$$\psi=\psi_0+\varepsilon S_1[P_1(\eta,\psi,\psi')]+\varepsilon S_2[P_2(\eta,\psi,\psi')], \tag{5.50}$$

$$\psi'=\psi_0'+\varepsilon S_1[\tilde{P}_1(\eta,\psi,\psi')]+\varepsilon S_2[\tilde{P}_2(\eta,\psi,\psi')]. \tag{5.51}$$

The expressions for $P_{1,2}$ and $\tilde{P}_{1,2}$ are explicitly given in (5.43) and (5.48).

Let $\mathcal{T}$ be an interval $y\in[0,y_0]$, and consider pairs of functions $f, g\in C(\mathcal{T})$. We define a mapping T, with components $T^{(1)}$, $T^{(2)}$, by the following formulas:

$$\begin{aligned}T^{(1)}(f,g)&=\psi_0+\varepsilon S_1[P_1(\eta,f,g)]+\varepsilon S_2[P_2(\eta,f,g)],\\T^{(2)}(f,g)&=\psi_0'+\varepsilon S_1[\tilde{P}_1(\eta,f,g)]+\varepsilon S_2[\tilde{P}_2(\eta,f,g)].\end{aligned} \tag{5.52}$$

The operators S_1, S_2 map continuous functions into continuous functions, so we must assure that $P_{1,2}(\eta,f,g)$ and $\tilde{P}_{1,2}(\eta,f,g)$ are continuous functions. A glance at (5.48) shows that for that purpose the interval $\mathcal{T}$ must be restricted so that

$$\varphi(y)\geqq c\varepsilon^{2m-1},\quad c>0\quad \text{for } y\in\mathcal{T}. \tag{5.53}$$

We consider now T as an operator in the Banach space of pairs of continuous functions (f,g), equipped with the usual norm

$$\sup_{\mathcal{T}}|f|+\sup_{\mathcal{T}}|g|.$$

In that space we consider balls $\mathbb{B}_\varepsilon$ that are centered at the pair ψ_0, ψ_0' and have a radius $\varepsilon^{1-\mu}$, with μ positive and arbitrarily small, i.e., the radius is slightly larger than ε. We want T to map the balls into themselves. With the estimate (5.38) it follows that we must have $P_{1,2}(\eta, f, g) = O(1)$ and $\tilde{P}_{1,2}(\eta, f, g) = 0(1)$. Analysis of the different terms in (5.48) produces a further and more severe restriction on the interval $\mathcal{T}$:

$$\varphi(y) \geqq c\varepsilon^m, \quad c > 0 \quad \text{for } y \in \mathcal{T} \tag{5.54}$$

and side condition

$$m \geqq 2. \tag{5.55}$$

It is now straightforward to show that T is a contraction mapping in $\mathbb{B}_\varepsilon$, that is, for any pair of pairs $(f_1, g_1), (f_2, g_2) \in \mathbb{B}_2$ we have

$$|T^{(1,2)}(f_1, g_1) - T^{(1,2)}(f_2, g_2)| \leqq \varepsilon C\{\sup (f_1 - f_2) + \sup (g_1 - g_2)\} \tag{5.56}$$

with C a constant independent of ε.

Let us show how the analysis runs on a representative term of (5.48). We consider

$$\begin{aligned} &\varepsilon|\varepsilon^{2m} S_1[\varphi^{-1}(\varphi + \varepsilon^{2m} f_1)^{-1} \tilde{G}(\eta, f_1, g_1) - \varphi^{-1}(\varphi + \varepsilon^{2m} f_2)^{-1} \tilde{G}(\eta, f_2, g_2)]| \\ &\leqq \varepsilon \frac{\varepsilon^{2m}}{y^2} \int_0^y \frac{\eta}{\sqrt{\varphi}} \frac{(\varphi + \varepsilon^{2m} f_1)^{-1}}{\varphi} |\tilde{G}(\eta, f_1, g_1) - \tilde{G}(\eta, f_2, g_2)|\, d\eta \\ &\quad + \varepsilon \frac{\varepsilon^{2m}}{y^2} \int_0^y \frac{\eta}{\sqrt{\varphi}} \cdot \frac{1}{\varphi} |\tilde{G}(\eta, f_2, g_2)| \cdot \left| \frac{1}{\varphi + \varepsilon^{2m} f_1} - \frac{1}{\varphi + \varepsilon^{2m} f_2} \right| d\eta. \end{aligned}$$

In the first term it is sufficient to remark that $\tilde{G}(\eta, f, g)$ of the structure (5.47) is Lipschitz-continuous with respect to f and g. Hence this term is bounded by

$$\varepsilon \frac{1}{y^2} \int_0^y \frac{\eta}{\sqrt{\varphi}} d\eta \cdot \sup \frac{\varepsilon^{2m}}{(\varphi + \varepsilon^{2m} f_1)\varphi} \cdot C\{\sup |f_1 - f_2| + \sup |g_1 - g_2|\}.$$

Due to condition (5.54) it follows that

$$\sup \frac{\varepsilon^{2m}}{\varphi + \varepsilon^{2m} f_1} \cdot \frac{1}{\varphi} = O(1),$$

so that the term is finally bounded by

$$\varepsilon C\{\sup |f_1 - f_2| + \sup |g_1 - g_2|\}$$

with C a constant independent of ε.

The second term of the inequality that we study is straightforward. The term is bounded by

$$\varepsilon \frac{1}{y^2} \int_0^y \frac{\eta}{\sqrt{\varphi}} d\eta \cdot \sup \frac{\varepsilon^{4m}}{\varphi} (\varphi + \varepsilon^{2m} f_1)^{-1} (\varphi + \varepsilon^{2m} f_2)^{-1} \cdot \sup |f_1 - f_2|.$$

With the condition (5.54) we find a much stronger contraction, i.e., the term is bounded by

$$\varepsilon^{1+m} \cdot C \cdot \sup |f_2 - f_2|$$

with C independent of ε.

The reader can easily be convinced that the term we have considered is indeed representative, that is, that other terms can be analysed by a completely analogous procedure.

Having demonstrated the contraction property of the mapping T, we have proved that there exists a unique solution $\psi(y)$, $\psi'(y)$ of (5.50), (5.51) in an interval $\mathcal{T}$ limited by the condition (5.54) and that, moreover,

$$\begin{aligned} \psi &= \psi_0 + O(\varepsilon), \\ \psi' &= \psi_0' + O(\varepsilon). \end{aligned} \tag{5.57}$$

This concludes the proof of the result announced in § 4 of this paper.

6. Generalizations: perturbed model problems. Let us call the problem (1.1) a model problem and consider perturbations of it, which are of the structure

$$\varepsilon^2 \frac{d^4 y}{dx^4} + \frac{d^2 y}{dx^2} - y + y^2 = \varepsilon^2 \mathcal{P}\left(y, \frac{dy}{dx}, \frac{d^2 y}{dx^2}; \varepsilon\right). \tag{6.1}$$

More precisely, we shall look at perturbations given by

$$\mathcal{P} = X_1(y)\left(\frac{dy}{dx}\right)^2 + yX_2(y)\frac{d^2 y}{dx^2} + X_3(y)\left(\frac{dy}{dx}\right)^2 \frac{d^2 y}{dx^2} + X_4(y)\left(\frac{d^2 y}{dx^2}\right)^2, \tag{6.2}$$

where $X_{1,2,3,4}(y)$ are polynomials in y. These functions may further depend on ε, but in a regular way.

The special structure of the coefficient of the d^2y/dx^2 is motivated by the consideration that a coefficient that would not be zero at $y = 0$ could be absorbed in the left-hand side.

We shall follow in main lines the method of analysis of the proceeding sections when applied to (6.1), (6.2).

Following § 2 we introduce y as an independent variable. The perturbation term becomes

$$\mathcal{P}(y, z, z') = X_1(y)z + \frac{1}{2} yX_2(y)\frac{dz}{dy} + \frac{1}{2} X_3(y)\, z\frac{dz}{dy} + \frac{1}{4} X_4(y)\left(\frac{dz}{dy}\right)^2. \tag{6.3}$$

Instead of the integral (2.4), we shall now have, as a basic equation,

$$\varepsilon^2\left\{z\frac{d^2 z}{dy^2} - \frac{1}{4}\left(\frac{dz}{dy}\right)^2\right\} + z = y^2\left(1 - \frac{2}{3}y\right) + \varepsilon^2 \int_0^y \mathcal{P}(\eta, z, z')\, d\eta. \tag{6.4}$$

It is quite easy to convince ourselves that a straightforward iteration, starting with

$$z_0 = y^2\left(1 - \frac{2}{3}y\right),$$

produces a formal approximation of the structure

$$\Phi_m = y^2\left[1 - \frac{y}{\alpha_m(\varepsilon)}\right] g_m(y, \varepsilon) \tag{6.5}$$

with $g_m(y, \varepsilon)$ a polynomial without zeros for $y \in [1 - (y/\alpha_m(\varepsilon))]$. This can also be seen from the reformulation of the problem

$$z = y^2 \bar{z}, \tag{6.6}$$

which produces the equation

$$\varepsilon^2\left\{y^2\left[\bar{z}\frac{d^2\bar{z}}{dy^2}-\frac{1}{4}\left(\frac{d\bar{z}}{dy}\right)^2\right]+3y\bar{z}\frac{d\bar{z}}{dy}\right\}+\bar{z}(1+\varepsilon^2\bar{z}) \tag{6.7}$$
$$=\left(1-\frac{2}{3}y\right)+\varepsilon^2\frac{1}{y^2}\int_0^y \eta^2\bar{\mathscr{P}}(\eta,\bar{z},\bar{z}')\,d\eta,$$

$$\bar{\mathscr{P}}=\bar{X}_1(y)\bar{z}+\bar{X}_2(y)\bar{z}'+y\bar{X}_3(y)\bar{z}\bar{z}'+y^2\bar{X}_4(y)(\bar{z}^1)^2. \tag{6.8}$$

Clearly, the perturbation term behaves well as $y\to 0$.

We can now introduce

$$\bar{z}=\varphi_m+\varepsilon^{2m}\psi_m \tag{6.9}$$

and formulate the problem for the remainder ψ_m. This will be the equation (4.4) with additional perturbations, which are integrals over functions containing ψ and ψ'.

Suppose now that, as in the model problem, we can prove the existence of a bounded solution $\psi_m(y;\varepsilon)$ on $y\in[0,y_0]$, y_0 being such that

$$|\varphi_m(y;\varepsilon)|\geqq c\varepsilon^m.$$

Can an argument be made on the continuation of the solutions, as at the end of § 4?

Essential for the argument is the sign of the right-hand side of (6.4) in the region of continuation of solution. Let us look at the formal approximation:

$$z=z_0+\varepsilon^2z_1^{(1)}+\varepsilon^2z_1^{(2)}, \tag{6.10}$$

$$z_1^{(1)}=\frac{1}{4}\left(\frac{dz_0}{dy}\right)^2-z_0\frac{d^2z_0}{dy^2}, \tag{6.11}$$

$$z_2^{(2)}=\int_0^y \mathscr{P}(\eta,z_0,z_0')\,d\eta. \tag{6.12}$$

$z_1^{(1)}$ is positive in the ε^2 neighbourhood of $y=3/2$. It follows that the approximation is still valid for y such that $z_0+\varepsilon^2z_1^{(2)}<0$. Hence, the continuation takes place in the region where the right-hand side of (6.4) is negative, and this is sufficient for the line of reasoning as given at the end of § 4.

We now turn to the construction of a contraction mapping. The problem for ψ is not essentially changed by the addition of the perturbation integral. Crucial is the problem for ψ'. In § 5.1 we have formulated that problem by differentiating the equation for the ψ problem. However, we can also differentiate first the equation for $\bar{z}$ and introduce the formal approximation with remainder term (6.9) afterwards.

We obtain in this way the equation

$$\varepsilon^2\left\{y^2\bar{z}\frac{d^2}{dy^2}(\bar{z}')+\left(\frac{1}{2}y^2\bar{z}'+5y\bar{z}\right)\frac{d}{dy}(\bar{z}')\right\} \tag{6.13}$$
$$+(\bar{z}')+\varepsilon^2\left(5\bar{z}\bar{z}'+\frac{5}{2}y(\bar{z}')^2\right)=-\frac{2}{3}+\varepsilon^2\frac{d}{dy}\frac{1}{y^2}\int_0^y \eta^2\bar{\mathscr{P}}(\eta,\bar{z},\bar{z}')\,d\eta.$$

For any pair of continuous functions $\bar{z}$, $\bar{z}'$ the perturbation term is a continuous function of y on an interval containing the origin. Furthermore, and this is the essential point, the perturbation does not introduce any new terms with the derivatives of $\bar{z}'$. This means that the construction of § 5 can be carried out, the only complication being the bookkeeping of more perturbation terms.

We thus arrive at the conclusion that *the main results of* § 4 *hold for the perturbed problem* (6.1), (6.2).

Let us finally look briefly at perturbations that affect the fourth-order derivative. For example, let us consider

$$\varepsilon^2\left\{1+\varepsilon^2 X_0(y)\frac{d^2y}{dx^2}\right\}\frac{d^4y}{dx^2}+\frac{d^2y}{dx^2}-y+y^2=\varepsilon^2\mathcal{P}\left(y,\frac{dy}{dx},\left(\frac{d^2y}{dx}\right)^2\right), \tag{6.14}$$

where $X_0(y)$ again is a polynomial. Moving the new perturbation to the right-hand side will certainly produce severe trouble in the course of the analysis. However, we can get away with a trick similar to the one used in § 5.3.2.

Multiplying by $\{1+\varepsilon^2X_0(d^2y/dx^2)\}^{-1}$ and using, whenever convenient, the identity

$$\left(1+\varepsilon^2X_0\frac{d^2y}{dx^2}\right)^{-1}=1-\varepsilon^2X_0\frac{d^2y}{dx^2}\left\{1+\varepsilon^2X_0\frac{d^2y}{dx^2}\right\}^{-1},$$

we find

$$\varepsilon^2\frac{d^4y}{dx^4}+\frac{d^2y}{dx^2}-y+y^2=\varepsilon^2\left\{1+\varepsilon^2X_0(y)\frac{d^2y}{dx^2}\right\}^{-1}\tilde{\mathcal{P}}\left(y,\frac{dy}{dx},\frac{d^2y}{dx^2}\right), \tag{6.15}$$

where $\tilde{\mathcal{P}}$ is again an expression of the structure (6.2).

The equation (6.15) does not seem to introduce any new fundamental difficulty into the problem, and the main line of analysis of this paper can again be followed. However, bookkeeping may become extremely cumbersome, even on the level of construction of formal approximations. The most efficient way to attack the problem is probably to use (6.14) as a starting point for the construction of formal approximation and (6.15) for the purpose of constructing the proof of validity. It is beyond the ambitions of this paper to perform all the necessary calculations.

Appendix. Computation of oscillatory integrals. (By N. Temme, C. W. I., Amsterdam.) In this appendix we evaluate two integrals that occur in the analysis of § 4.2. First we consider

$$I^{(1)}(\varepsilon)=\frac{1}{\varepsilon^2}\int_{-\infty}^{0}\cos\left(\frac{1}{\varepsilon}\xi\right)[y^{(1)}(\xi)-(y^{(1)}(\xi))^2]\,d\xi, \tag{A.1}$$

where

$$y^{(1)}(x)=\frac{3}{2}\operatorname{sech}^2\left(\frac{1}{2}x\right)=\frac{3}{\cosh x+1}, \tag{A.2}$$

which is a first-order approximation of the reduced equation (1.1), that is, a symmetric solution of the equation

$$\frac{d^2y^{(1)}}{dx^2}=y^{(1)}(1-y^{(1)}),\qquad x\in\mathbb{R}. \tag{A.3}$$

Using (A.3), we can write (A.1) in the form

$$I^{(1)}(\varepsilon)=\frac{1}{\varepsilon^2}\int_{-\infty}^{0}\cos\left(\frac{1}{\varepsilon}\xi\right)\frac{d^2y^{(1)}}{d\xi^2}\,d\xi,$$

and integrating by parts we find

$$I^{(1)}(\varepsilon)=-\frac{1}{\varepsilon^4}\int_{-\infty}^{0}\cos\left(\frac{1}{\varepsilon}\xi\right)y^{(1)}(\xi)\,d\xi=-\frac{1}{2\varepsilon^4}\int_{-\infty}^{\infty}e^{i\xi/\varepsilon}y^{(1)}(\xi)\,d\xi.$$

Integrals of this type can be evaluated by using residues. Here we express the integral in terms of the Euler beta function. The substitutions $\xi = ln\, u$, $u = t/(1-t)$ yield

$$I^{(1)}(\varepsilon) = -\frac{3}{\varepsilon^4}\int_0^\infty \frac{u^{i/\varepsilon}\, du}{(u+1)^2} = -\frac{3}{\varepsilon^4}\int_0^1 t^{i/\varepsilon}(1-t)^{-i/\varepsilon}\, dt = -\frac{3}{\varepsilon^4}\frac{\Gamma(1+i/\varepsilon)\Gamma(1-i/\varepsilon)}{\Gamma(2)}.$$

Using the reflection formula of the gamma function [9],

$$\Gamma(1+z)\Gamma(1-z) = \frac{\pi z}{\sin(\pi z)},$$

we finally obtain

$$I^{(1)}(\varepsilon) = -\frac{3\pi}{\varepsilon^5 \sinh(\pi/\varepsilon)}. \tag{A.4}$$

The second integral to be evaluated is

$$I^{(2)}(\varepsilon) = \frac{1}{\varepsilon^2}\int_{-\infty}^0 \cos\left(\frac{1}{\varepsilon}\xi\right) y^{(2)}(\xi)[1-y^{(2)}(\xi)]\, d\xi \tag{A.5}$$

with

$$y^{(2)}(x) = \frac{3\lambda(1-\varepsilon^2)}{\cosh(x\sqrt{1-\varepsilon^2}) + \lambda(1-5\varepsilon^2)}, \qquad \lambda = \frac{1}{\sqrt{1+5\varepsilon^2+10\varepsilon^4}}. \tag{A.6}$$

The approximation $y^{(2)}$ is obtained by expanding the solution of the z-equation

$$\varepsilon^2\left[zz'' - \frac{1}{4}(z')^2\right] + z = y^2 - \frac{2}{3}y^3, \quad \sqrt{z} = \frac{dy}{dx}, \quad z' = \frac{dz}{dy}, \quad z'' = \frac{d^2z}{dy^2},$$

in a series $z \sim z_0 + \varepsilon^2 z_1 + \varepsilon^4 z_2 + \cdots$, which gives

$$z_0 = y^2 - \tfrac{2}{3}y^3, \qquad z_1 = \tfrac{1}{4}(z_0')^2 - z_0 z_0'' = -\tfrac{1}{3}y^2(3 - 10y + 5y^2).$$

For $y^{(2)}(x)$ we obtain the equation

$$\frac{dy^{(2)}}{dx} = \sqrt{z_0 + \varepsilon^2 z},$$

with the symmetric solution (A.6). Observe that

$$\lim_{\varepsilon \downarrow 0} y^{(2)}(x) = y^{(1)}(x).$$

For the evaluation of (A.5) we need integrals of the type

$$J_k(a,\gamma) = \int_0^\infty \frac{\cos ax}{(\cosh x + \cos\gamma)^k}\, dx = \frac{1}{2}\int_{-\infty}^\infty \frac{e^{iax}}{(\cosh x + \cos\gamma)^k}\, dx, \qquad k = 1, 2, \tag{A.7}$$

where $\gamma \in (-\pi, \pi)$, and a is a complex number in the strip $|\mathrm{Im}\, a| < k$. Let $k = 1$ and take in the second integral $x = ln\, t$; then

$$J_1(a,\gamma) = \int_0^\infty \frac{t^{ia}}{(t+e^{i\gamma})(t+e^{-i\gamma})}\, dt$$

$$= \frac{e^{i\gamma}}{2i\sin\gamma}\int_0^\infty \frac{t^{ia-1}}{t+e^{i\gamma}}\, dt - \frac{e^{-i\gamma}}{2i\sin\gamma}\int_0^\infty \frac{t^{ia-1}}{t+e^{-i\gamma}}\, dt.$$

Note that when a is real, two divergent integrals appear. However, to ensure convergence in both integrals, we temporarily assume that $-1<\operatorname{Im} a<0$. By putting $t=se^{i\gamma}$, $t=se^{-i\gamma}$, respectively, and using standard methods for complex integrals, both integrals can be combined into

$$J_1(a,\gamma)=\frac{e^{-a\gamma}-e^{a\gamma}}{2i\sin\gamma}\int_0^\infty \frac{s^{ia-1}}{s+1}\,ds,$$

which again can be expressed in terms of gamma functions. The result is

$$J_1(a,\gamma)=\frac{\pi\sinh(a\gamma)}{\sin\gamma\sinh(\pi a)},\tag{A.8}$$

derived under the condition $-1<\operatorname{Im} a<0$. Since both of these final expressions for $J_1(a,\gamma)$ and the integral representation of $J_1(a,\gamma)$ in (A.7) are analytic functions of a in the strip $|\operatorname{Im} a|<1$, we conclude that (A.8) also holds in this strip. Especially, it is valid for real values of a.

By splitting $(t+e^{i\gamma})^{-2}(t+e^{-i\gamma})^{-2}$ into partial fractions, the evaluation of $J_2(a,\gamma)$ can be done in the same manner. However, by differentiating $J_1(a,\gamma)$ of (A.7) with respect to γ, we can derive the desired result in a straightforward way; it is easily verified that the conditions for this approach are satisfied (see [9, par. 4.44]). In doing so we obtain

$$J_2(a,\gamma)=\frac{1}{\sin\gamma}\frac{\partial J_1(a,\gamma)}{\partial\gamma}=\frac{\pi}{\sinh(a\gamma)\sin^3\gamma}[a\sin\gamma\cosh(a\gamma)-\cos\gamma\sinh(a\gamma)].$$

We return to the evaluation of (A.5), and we use the integrals J_k of (A.7):

$$I^{(2)}(\varepsilon)=\frac{1}{\varepsilon^2}3\lambda\sqrt{1-\varepsilon^2}\,[J_1(a,\gamma)-3\lambda(1-\varepsilon^2)J_2(a,\gamma)],\tag{A.9}$$

where

$$a=\frac{1}{\varepsilon\sqrt{1-\varepsilon^2}}\qquad \cos\gamma=\lambda(1-5\varepsilon^2).\tag{A.10}$$

We prescribe that γ is positive and tending to zero as $\varepsilon\to0$. Furthermore, $\sin\gamma=\varepsilon\lambda\sqrt{15(1-\varepsilon^2)}$. It follows that

$$I^{(2)}(\varepsilon)=\frac{1}{\varepsilon^2}\frac{9\lambda^2\pi(1-\varepsilon^2)^{3/2}}{\sin^3\gamma\sinh(\pi a)}[\lambda\sinh(a\gamma)-a\sin\gamma\cosh(a\gamma)].\tag{A.11}$$

It is interesting to compare this result with (A.4), since we may expect that $I^{(2)}(\varepsilon)\sim I^{(1)}(\varepsilon)$ as $\varepsilon\to0$. Recalling that both a and γ depend on ε, we first consider a fixed (i.e., independent of ε). Letting $\varepsilon\to0$ (and hence $\gamma\to0$), we have

$$\frac{\lambda\sinh(a\gamma)-a\sin\gamma\cosh(a\gamma)}{\sin^3\gamma}\sim-\frac{1}{3}a^3+\frac{1}{6}a.\tag{A.12}$$

Substituting this in (A.11), we obtain

$$I^{(2)}(\varepsilon)\sim-\frac{3a^3\pi(1-\frac{1}{2}a^{-2})}{\varepsilon^2\sinh(\pi a)},\tag{A.13}$$

which indeed resembles (A.4), when $a=1/\varepsilon$, which, of course, is not allowed in (A.12). When we consider (A.11) with the true value of a given in (A.10), that is, with $a\sin\gamma=\lambda\sqrt{15}$, the representation of $I^{(2)}$ becomes

$$I^{(2)}(\varepsilon)=\frac{9\lambda^3\pi(1-\varepsilon^2)^{3/2}}{\varepsilon^2\sin^3\gamma\sinh(\pi a)}[\sinh(a\gamma)-\sqrt{15}\cosh(a\gamma)],\tag{A.14}$$

with $a\gamma \sim \sqrt{15}$ as $\varepsilon \to 0$. It follows that the ratio $I^{(2)}(\varepsilon)/I^{(1)}(\varepsilon)$ does not tend to unity when $\varepsilon \to 0$. The limit of the ratio is about 3.6.

Acknowledgments. The subject of this paper was suggested to me by Klaus Kichgässner in a casual conversation. I am grateful to him for this source of mathematical excitement and for the lengthy discussions of my work later on. After the main body of the paper was completed Klaus sent me the preprint of Amick and McLeod [1]. In an exchange of letters, Charles Amick drew my attention to the paper of Hammersley and Mazzarino [4]. I have attempted to link, whenever appropriate, certain steps of my analysis to the contents of these papers, through remarks added in the text.

I gratefully acknowledge the contribution of Nico Temme, of the C.W.I., Amsterdam, who is responsible for the analysis given in the Appendix.

REFERENCES

[1] C. J. AMICK AND J. B. MCLEOD, *A singular perturbation problem in water waves*, Stab. Appl. Anal. of Cont. Media., (1992) pp. 127-148.

[2] C. J. AMICK AND K. KIRCHGÄSSNER, *A theory of solitary water-waves in the presence of surface tension*, Arch. Ration Mech. Anal. (1989), pp. 1-49.

[3] W. ECKHAUS, *On water waves at Froude number slightly higher than one and Bond number less than 1/3*, Z. Agnew. Math. Phys. (1992), pp. 254-269.

[4] J. M. HAMMERSLEY AND G. MAZZARINO, *Computational aspects of some autonomous differential equations*, Proc. Roy. Soc. London Ser. A, 424 (1989).

[5] P. HOLMES, J. MARSDEN, AND J. SCHEURLE, *Exponentially small splittings of separatrices with applications to* KAM *theory and degenerate bifurcations*, Contemp. Math., 81 (1988), pp. 213-243.

[6] J. K. HUNTER AND J. SCHEURLE, *Existence of perturbed solitary wave solutions to a model equation for water waves*, Phys. D, 32 (1988), pp. 253-268.

[7] K. KIRCHGÄSSNER, *Nonlinearly resonant surface waves and homoclinic bifurcations*, Adv. Appl. Mech., 26 (1988), pp. 135-181.

[8] J. SCHEURLE, *Chaos in a rapidly forced pendulum equation*, Contemp. Math., 97 (1980), pp. 411-419.

[9] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, London, 1927.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1291-1304, September 1992

013

UNFOLDING A POINT OF DEGENERATE HOPF BIFURCATION IN AN ENZYME-CATALYZED REACTION MODEL*

BRIAN HASSARD† AND KATIE JIANG‡

Abstract. A point of degenerate Hopf bifurcation in an enzyme-catalyzed model previously studied by Doedel [2] is rigorously analyzed by using techniques of singularity theory and interval analysis. A computation using interval analysis proves the existence of a point of degenerate Hopf bifurcation, which is a smooth function of additional parameters in the model system. Singularity theory as developed by Golubitsky and Langford [*J. Differential Equations*, 3 (1981), pp. 375-415.] and Golubitsky and Schaeffer [*Singularities and Groups in Bifurcation Theory* I, Springer-Verlag, New York, 1985.] is then used to construct universal unfoldings of the degeneracy, to completely characterize the families of small amplitude periodic solutions that arise for parameters near the degenerate values. Computations using interval analysis are employed in this proof. Excellent agreement is found between the bifurcation theoretic unfolding and (numerical) continuation results using pseudoarclength continuation.

Key words. singularity, Hopf bifurcation, enzyme model

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

1. Introduction. In this paper we study a mathematical model of an enzymatically active system. Enzymes are molecules that catalyze the biochemical reactions in the metabolic pathways of living organisms. The complexity of living cells is such that it is difficult to model the whole system where so many phenomena take place, including enzyme reactions and various types of transport, e.g., by diffusion, electrical migration, and convection. In the model we consider, the enzymes are physically confined in one compartment, i.e., the model is an "immobilized" enzyme system. Such models are used to study the interaction of enzyme reaction and diffusion in a well-defined context. The interaction produces oscillations.

The concentration of two chemical species in a single compartment can be modeled by a system of two ordinary differential equations that describes the change of these concentrations in the presence of an enzyme-catalyzed reaction inside the compartment with transport from an outside reservoir called the S-A system:

$$\begin{aligned}\frac{ds}{dt}&=(s_0-s)-\rho R(s,a),\\ \frac{da}{dt}&=\alpha(a_0-a)-\rho R(s,a).\end{aligned}\tag{1}$$

Here s and a denote the concentrations of the two chemical species S and A inside the compartment. S stands for "substrate" and A for "activator" though both are substrates in a reaction where they are consumed. This reaction is catalyzed by an enzyme with reaction rate proportional to

$$R(s,a)=\frac{sa}{(1+s+\kappa s^2)},\qquad \kappa>0.\tag{2}$$

* Received by the editors February 5, 1990; accepted for publication (in revised form) January 20, 1992.

† Department of Mathematics, State University of New York at Buffalo, Buffalo, New York, 14214.

‡ Department of Science and Mathematics, GMI Engineering and Management Institute, 1700 West 3rd Avenue, Flint, Michigan 48504.

The parameter ρ is a ratio of characteristic times, $\rho = \theta_T/\theta_R$, where θ_R is the characteristic time of diffusion of s from the outside reservoir into the compartment and θ_T is the characteristic time of the enzyme reaction; the terms $(s_0 - s)$ and $\alpha(a_0 - a)$ describe the transport from the outside reservoir where the concentrations are held at a constant s_0 and a_0. The parameter α represents the ratio of diffusion coefficients between the reservoir and the compartment, the parameters s_0, a_0, ρ, α, and κ are all dimensionless and positive as are the state variables s and a.

Doedel [2] used purely numerical techniques to demonstrate the birth of isolas of periodic solutions in the S-A system as the two parameters ρ and s_0 are varied. With ρ as the bifurcation parameter, his results suggest that there is a value s_0^d such that for s_0 slightly less than s_0^d, a branch of periodic solutions arises from the stationary branch in a Hopf bifurcation, and returns to the stationary branch at a second point of Hopf bifurcation. While for s_0 slightly greater than s_0^d, there is an isola (an isolated branch of periodic solutions) and no Hopf bifurcation from the stationary branch. We shall show that for $s_0 = s_0^d$, the S-A system exhibits degenerate Hopf bifurcation.

The purpose of the present paper is to analyze the birth of isolas in the S-A system by means of singularity theory and to compare the approximate bifurcation diagrams obtained from unfolding the point of degenerate Hopf bifurcation with "actual" diagrams obtained using Doedel's code AUTO [2].

The technique we apply involves first a nonlinear change of variables taking the S-A system into Poincare-Birkhoff normal form [6], followed by application of Lyapunov-Schmidt reduction as was done in [3], [5]. The degenerate Hopf bifurcation is classified using singularity theory [4], [5]. We then unfold the degeneracy with respect to the parameter s_0.

2. Location of the point of degenerate Hopf bifurcation.

2.1. Selection of s_* as bifurcation parameter. We choose to use s_*, the s-component of the stationary solution, as bifurcation parameter rather than ρ as in [2]. Then a_*, the a-component of the stationary solution, is uniquely defined by

$$a_* = a_0 + (s_* - s_0)/\alpha,$$

and the parameter ρ is replaced by

$$\rho_* = (s_0 - s_*)/R(s_*, a_*).$$

The advantage in selecting s_* as the bifurcation parameter is that the stationary solution is simple and uniquely defined as a function of s_*, whereas with ρ as the bifurcation parameter, there may be up to three distinct stationary solutions. We note that as a function of s_*, ρ_* has poles at 0 and $s_0 - \alpha a_0$. Since $a_* = a_0 + (s_* - s_0)/\alpha$, while s_* and a_* are equilibrium values of concentrations of chemical species, for physical meaning both $s_* > 0$, and $a_* > 0$; thus $s_* > \max(0, s_0 - \alpha a_0)$, which avoids the poles of ρ_*. Also, for physical reasons $\rho > 0$. Since $\kappa > 0$, $1 + s_* + \kappa s_*^2 > 0$ and $R(s_*, a_*) > 0$ for all $s_* > 0$. The inequality $\rho_* > 0$, therefore, further restricts $s_* < s_0$.

2.2. Linear stability analysis for stationary solution. The Jacboian matrix for the S-A system at the stationary solution (s_*, a_*) is

$$\left.\begin{bmatrix} -1 - \rho_* \partial R/\partial s & -\rho_* \partial R/\partial a \\ -\rho_* \partial R/\partial s & -\alpha - \rho_* \partial R/\partial a \end{bmatrix}\right|_{(s,a)=(s_*,a_*)}, \tag{3}$$

where $\partial R/\partial s = a(1 - \kappa s^2)/(1 + s + \kappa s^2)^2$ and $\partial R/\partial a = s/(1 + s + \kappa s^2)$. The linear stability of (s_*, a_*) is then determined by the real parts of the eigenvalues λ of this matrix, which satisfy the characteristic equation

$$\lambda^2 + (1 + \alpha + \rho_* \partial R/\partial s + \rho_* \partial R/\partial a)\lambda + (\alpha + \rho_* \partial R/\partial a + \alpha \rho_* \partial R/\partial s) = 0$$

We are primarily interested in the case of complex eigenvalues $\lambda = \sigma \pm i\omega$, where

$$\sigma = -\tfrac{1}{2}(1 + \alpha + \rho_* \partial R/\partial s + \rho_* \partial R/\partial a),$$
$$\omega^2 = (\alpha + \rho_* \partial R/\partial a + \alpha \rho_* \partial R/\partial s) - \sigma^2 > 0.$$

At points of Hopf bifurcation, critical values s_*^c of the bifurcation parameter must satisfy $\sigma(s_*^c, s_0) = 0$, where

$$\begin{aligned}\sigma(s_*, s_0) &= -(1 + \alpha + (s_0 - s_*)(1/a_* + (1 - \kappa s_*^2)/s_*(1 + s_* + \kappa s_*^2)))/2\\ &= (A(s_*)s_0^2 + B(s_*)s_0 + C(s_*))/D(s_*, s_0),\end{aligned}$$

where we define

$$\begin{aligned}A(s_*) &= \kappa s_*^2 - 1,\\ B(s_*) &= -3\kappa s_*^3 - (a_0\alpha\kappa + 1)s_*^2 + s_* + a_0\alpha,\\ C(s_*) &= 2\kappa s_*^4 + (1 + (a_0\alpha^2 + 2a_0\alpha)\kappa)s_*^3 + (a_0\alpha^2 + a_0\alpha)s_*^2 + a_0\alpha^2 s_*,\\ D(s_*, s_0) &= 2[(\kappa s_*^3 + s_*^2 + s_*)s_0 - \kappa s_*^4 - (a_0\alpha\kappa + 1)s_*^3 - (a_0\alpha + 1)s_*^2 - a_0\alpha s_*].\end{aligned}$$

Figures 4.5 and 4.6 of [2] suggest the existence of values (s_*^d, s_0^d) (d for degenerate) such that for s_0 slightly less than s_0^d, two distinct Hopf bifurcations occur at points $s_*^{(1)}, s_*^{(2)}$ close to s_*^d and obey $s_*^{(1)} < s_*^d < s_*^{(2)}$, that is,

$$\sigma(s_*^{(1)}(s_0), s_0) = \sigma(s_*^{(2)}(s_0), s_0) = 0,$$

and as $s_0 \to s_0^d$, the points $s_*^{(1)}(s_0)$ and $s_*^{(2)}(s_0)$ coalesce.

Assuming $\sigma(s_*, s_0)$ is smooth in a neighborhood of the supposed degenerate point (s_*^d, s_0^d), we therefore expect both $\sigma = \partial\sigma/\partial s_* = 0$ at (s_*^d, s_0^d), that is,

$$A(s_*^d)(s_0^d)^2 + B(s_*^d)s_0^d + C(s_*^d) = A'(s_*^d)(s_0^d)^2 + B'(s_*^d)s_0^d + C'(s_*^d) = 0.$$

In Lemmas 1 and 2 we shall prove the existence of such a degenerate point (s_*^d, s_0^d), and justify the smoothness assumption by showing $D(s_*^d, s_0^d) \neq 0$.

The present work depends on numerical computations of degenerate value s_*^d of the bifurcation parameter, and of various other expressions at this point. We state the numerical results in the form of lemmas. In Lemma 1, $s_*^d(500, 0.2, 0.1)$ is given to a relatively high degree of precision because there will be loss of precision in the subsequent computations leading to the coefficient a_1 in Lemma 3. For the same reason, the rectangle R_0 is chosen small: we return to this subject in the discussion.

LEMMA 1. *There is a smooth function* $s_*^d(a_0, \alpha, \kappa)$ *defined for all* (a_0, α, κ) *in the rectangle* $R_0 = \{|a_0 - 500| \leqq 10^{-14},\ |\alpha - 0.2| \leqq 10^{-16},\ |\kappa - 0.1| \leqq 10^{-16}\}$ *and obeying* $|s_*^d - 25.508771186| \leqq 10^{-9}$, *such that* $s_* = s_*^d(a_0, \alpha, \kappa)$ *is a simple zero of the polynomial equation*

$$P(s_*; a_0, \alpha, \kappa) = A'(A'C - AC')^2 - B'(A'B - AB')(A'C - AC') + C'(A'B - AB')^2 = 0.$$

Furthermore, $s_*^d(500, 0.2, 0.1)$ *is in the interval* $25.50877118629661337716 \pm 0.5 \times 10^{-20}$.

Proof. MACSYMA was used to generate analytical expressions for the coefficients of the powers of s_* in $P(s_*; a_0, \alpha, \kappa)$; see Appendix B. P is the numerator of the rational function obtained by substituting $-(A'C - AC')/(A'B - AB')$ for s_0 in the expression $A'(s_0)^2 + B's_0 + C'$. At $a_0 = 500$, $\alpha = 0.2$, $\kappa = 0.1$, in particular

$$\begin{aligned}P(s_*; 500, 0.2, 0.1) = {}&0.00004s_*^{11} - 0.0104s_*^{10} + 0.2436s_*^9 + 0.316s_*^8 - 10.78s_*^7 - 9.2s_*^6\\ &+ 178.28s_*^5 - 114s_*^4 - 3616s_*^3 - 5920s_*^2 - 480s_*,\end{aligned}$$

where each coefficient is exact.

The version of the Bairstow algorithm as implemented by Aberth in Precision BASIC (program POLY [1]) then was used to compute all zeros of $P(s_*; 500, 0.2, 0.1)$. Fixed decimal format with 20 decimal digits was specified. The program found four complex zeros and seven real zeros, including the zero $s_* = 25.50877118629661337716$ correctly rounded to 20 decimal places. Since 11 distinct zeros were found and the polynomial $P(s_*)$ is of degree 11, the zero given above is necessarily simple. Since Aberth's program, when it succeeds, gives correctly rounded results, the zero $s_*^d(500, 0.2, 0.1)$ is known to lie in the stated interval.

To show the existence of the smooth function $s_*^d(a_0, \alpha, \kappa)$, we evaluated $P'(s_*, a_0, \alpha, \kappa)$ at the ranged values $s_* = 25.508771186 \pm 10^{-9}$, $a_0 = 500 \pm 10^{-14}$, $\alpha = 0.2 \pm 10^{-16}$, and $\kappa = 0.1 \pm 10^{-16}$, and found that $P' < 0$, i.e., both ends of the interval representing P', were negative. We then evaluated $P(25.508771185;\ a_0, \alpha, \kappa)$ and $P(25.508771187;\ a_0, \alpha, \kappa)$ for the same ranged values of a_0, α, and κ, and found that P changes sign. Thus for all (a_0, α, κ) in R_0, P has a simple zero $s_*^d(a_0, \alpha, \kappa)$ obeying $|s_*^d - 25.508771186| \leqq 10^{-9}$. Smoothness of $s_*^d(a_0, \alpha, \kappa)$ follows from the implicit function theorem. □

LEMMA 2. *There is a smooth function $s_0^d(a_0, \alpha, \kappa)$ defined for all (a_0, α, κ) in R_0 and obeying $|s_0^d - 110.474757| \leqq 3 \times 10^{-6}$ such that $(s_*^d(a_0, \alpha, \kappa), s_0^d(a_0, \alpha, \kappa))$ solves the pair of equations $\sigma(s_*, s_0) = \partial\sigma/\partial s_*(s_*, s_0) = 0$. At (s_*^d, s_0^d), $\omega(s_*^d, s_0^d)$ satisfies $|\omega - 0.92965287| \leqq 9.5 \times 10^{-7}$. Furthermore, $s_0^d(500, 0.2, 0.1)$ is in the interval* $110.47475708580432554 \pm 2.9 \times 10^{-17}$.

Proof. MACSYMA was used to generate analytical expressions for $A'C - AC'$, $A'B - AB'$, $s_0^d = -(A'C - AC')/(A'B - AB')$, and D, which were then evaluated using a PBASIC program at the ranged value s_*^d stated in Lemma 1. Output from this program gave

$$-(A'C - AC') = 1.3402397625337886555 2 \times 10^6 \pm 3.0 \times 10^{-14},$$

$$A'B - AB' = 1.2131638012952059610 5 \times 10^4 \pm 3.0 \times 10^{-16},$$

$$s_0^d = 110.47475708580432554 4 \pm 2.9 \times 10^{-17},$$

$$D = -7.0240576831843718 61 \times 10^4 \pm 2.4 \times 10^{-13}.$$

In particular, $s_0^d(500, 0.2, 0.1)$ is well defined. The polynomial equation of degree 11 in Lemma 1 divided by $(A'B - AB')^2$ at s_*^d becomes

$$A'(s_*^d)(s_0^d)^2 + B'(s_*^d)s_0^d + C'(s_*^d) = 0,$$

which implies

$$AA'(s_0^d)^2 + AB's_0^d + AC' = 0.$$

By the definition of s_0^d,

$$AB's_0^d + AC' = A'Bs_0^d + A'C;$$

thus,

$$AA'(s_0^d)^2 + A'Bs_0^d + A'C = 0.$$

Since $A'(s_*^d) = 2\kappa s_*^d \neq 0$,

$$A(s_*^d)(s_0^d)^2 + B(s_*^d)s_0^d + C(s_*^d) = 0.$$

But $\sigma(s_*, s_0)$ is of the form

$$(A(s_*)s_0^2 + B(s_*)s_0 + C(s_*))/D(s_*, s_0),$$

where from above, $D(s_*^d, s_0^d) \neq 0$, so $\sigma(s_*^d, s_0^d)$ and $\partial\sigma/\partial s_*(s_*^d, s_0^d)$ are both well defined: $\sigma(s_*^d, s_0^d) = 0$ and $\partial\sigma/\partial s_*(s_*^d, s_0^d) = 0$ then follow immediately.

To show existence of the smooth function $s_0^d(a_0, \alpha, \kappa)$, we evaluated the analytical expressions for $A'C - AC'$, $A'B - AB'$, s_0^d, and D at the ranged values $s_* = 25.508771186 \pm 10^{-9}$, $a_0 = 500 \pm 10^{-14}$, $\alpha = 0.2 \pm 10^{-16}$, and $\kappa = 0.1 \pm 10^{-16}$ as in Lemma 1. We found that $A'B - AB' > 0$, so $s_0^d(a_0, \alpha, \kappa)$ is well defined and smooth for all (a_0, α, κ) in R_0. We found the ranged value $110.474757 \pm 3 \times 10^{-6}$ for s_0^d, so $|s_0^d(a_0, \alpha, \kappa) - 110.474757| \leqq 3 \times 10^{-6}$ for all (a_0, α, κ) in R_0. Also, we found $D < 0$, which implies as above that $\sigma = \partial\sigma/\partial s_* = 0$ at $(s_*^d(a_0, \alpha, \kappa), s_0^d(a_0, \alpha, \kappa))$.

The formula $\omega^2 = (1-\alpha)((s_0 - s_*)/a_*) - \alpha^2$ was derived from the equations for σ and ω^2 in § 2.2 and the condition $\sigma = 0$. A separate PBASIC program was written to evaluate this formula at the ranged values of a_0, α, s_0^d, and s_*^d given above. Output from this program gave $\omega = 0.92965287 \pm 9.5 \times 10^{-7}$. □

3. Nonlinear analysis.

3.1. Transformation to Poincaré–Birkhoff normal form. First we perform a linear change of variables such that in the new variables y_1, y_2, the Jacobian of the S-A system is in real canonical form. Define

$$P = \begin{bmatrix} 1 & 0 \\ -(1+\sigma+\rho_*\partial R/\partial s)/(\rho_*\partial R/\partial a) & \omega/(\rho_*\partial R/\partial a) \end{bmatrix}.$$

The columns of P are just $\operatorname{Re}(v)$ and $-\operatorname{Im}(v)$, where v is the eigenvector of (3) corresponding to $\lambda = \sigma + i\omega$. Let

$$\begin{bmatrix} s \\ a \end{bmatrix} = \begin{bmatrix} s_* \\ a_* \end{bmatrix} + P \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

In terms of y_1, y_2, the S-A system becomes

$$\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = P^{-1} F\left(\begin{bmatrix} s_* \\ a_* \end{bmatrix} + P \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}; s_*, s_0\right),$$

where

$$F\left(\begin{bmatrix} s \\ a \end{bmatrix}; s_*, s_0\right) = \begin{bmatrix} s_0 - s - \rho_* R(s, a) \\ \alpha(a_0 - a) - \rho_* R(s, a) \end{bmatrix}.$$

For this new system, the Jacobian at $y_1 = y_2 = 0$ is

$$\begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix}.$$

It is convenient to let $z = y_1 + iy_2$, $\mu = s_* - s_*^d$, and $\nu = s_0 - s_0^d$ and write the system in the form

$$\dot{z} = (\sigma + i\omega)z + g(z, \bar{z}, \mu, \nu).$$

The derivatives of g at $z = 0$ are related to those of f at $y_1 = y_2 = 0$ by

$$\begin{aligned} g_{jk} &= (\partial/\partial z)^j (\partial/\partial \bar{z})^k g(z, \bar{z}, \mu, \nu)|_{z=\bar{z}=0} \\ &= \left(\frac{1}{2}\left(\frac{\partial}{\partial y_1} - i\frac{\partial}{\partial y_2}\right)\right)^j \left(\frac{1}{2}\left(\frac{\partial}{\partial y_1} + i\frac{\partial}{\partial y_2}\right)\right)^k (f_1(y_1, y_2, \mu, \nu) + if_2(y_1, y_2, \mu, \nu)). \end{aligned}$$

Next, we perform a nonlinear change of variables. Let

$$z=\xi+\chi_{20}\xi^2/2+\chi_{11}\xi\bar{\xi}+\chi_{02}\bar{\xi}^2/2+\chi_{30}\xi^3/6+\chi_{12}\xi^2\bar{\xi}+\chi_{03}\bar{\xi}^3/6,$$

where

$$\begin{aligned}
&\chi_{20}=g_{20}/\lambda,\\
&\chi_{11}=g_{11}/\bar{\lambda},\\
&\chi_{02}=g_{02}/(2\bar{\lambda}-\lambda),\\
&\chi_{30}=3(g_{20}\chi_{20}+g_{11}\bar{\chi}_{02}+g_{30}/3)/2\lambda,\\
&\chi_{12}=(g_{20}\chi_{02}+g_{11}(\bar{\chi}_{20}+2\chi_{11})+2g_{02}\bar{\chi}_{11}+g_{12})/2\bar{\lambda},\\
&\chi_{03}=3(g_{11}\chi_{02}+g_{02}\bar{\chi}_{20}+g_{03}/3)/(3\bar{\lambda}-\lambda).
\end{aligned}$$

In terms of new (complex) variable ξ, the S-A system assumes the Poincaré-Birkhoff (P-B) normal form

$$\dot{\xi}=(\sigma+i\omega)\xi+c(\xi,\bar{\xi}),$$

where

$$c(\xi,\bar{\xi})=c_1(\mu,\nu)\xi^2\bar{\xi}+0(\xi^4)$$

and

$$c_1(\mu,\nu)=[g_{20}g_{11}(2\lambda+\bar{\lambda})/|\lambda|^2+2|g_{11}|^2/\lambda+|g_{02}|^2/(2\lambda-\bar{\lambda})+g_{21}]/2.$$

(The coefficients χ_{20}, χ_{11}, and χ_{02} were chosen to make the quadratic terms of $c(\xi,\bar{\xi})$ vanish, and χ_{30}, χ_{12}, and χ_{03} were chosen to make the coefficients of ξ^3, $\xi\bar{\xi}^2$, and $\bar{\xi}^3$ in $c(\xi,\bar{\xi})$ all vanish. The evaluation of c_1 does not involve χ_{30}, χ_{12}, or χ_{03}.)

3.2. Lyapunov–Schmidt reduction. Letting $\xi=\xi_1+i\xi_2$ and $c_1=a_1+ib_1$, where all of ξ_1, ξ_2, a_1 and b_1 are real, the P-B normal form becomes

$$\begin{bmatrix}\dot{\xi}_1\\ \dot{\xi}_2\end{bmatrix}=\begin{bmatrix}\sigma & -\omega\\ \omega & \sigma\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}+\begin{bmatrix}a_1 & -b_1\\ b_1 & a_1\end{bmatrix}\begin{bmatrix}\xi_1(\xi_1^2+\xi_2^2)\\ \xi_2(\xi_1^2+\xi_2^2)\end{bmatrix}+0(\xi^4).$$

When Lyapunov-Schmidt reduction is applied to this form, the results are simple. There are periodic solutions corresponding to solutions of a bifurcation equation

$$r(u,\mu,\nu)x=0,$$

where $u=x^2$, $r(0,0,0)=0$, and the derivatives of r at $(u,\mu,\nu)=(0,0,0)$ are as follows

$$r_u(0,0,0)=-a_1,\quad r_\mu(0,0,0)=\frac{-\partial\sigma}{\partial s_*},\quad r_\nu(0,0,0)=\frac{-\partial\sigma}{\partial s_0},$$

$$r_{\mu\mu}(0,0,0)=\frac{-\partial^2\sigma}{\partial^2 s_*}.$$

The numerical evaluation of these derivatives is the subject of Lemma 3.

LEMMA 3. *For all* (a_0, α, κ) *in* R_0, *at* $(s_*, s_0) = (s^d_*(a_0, \alpha, \kappa), s^d_0(a_0, \alpha, \kappa))$ *the derivatives* $\partial\sigma/\partial s_0$ *and* $\partial^2\sigma/\partial^2 s_*$ *obey* $|\partial\sigma/\partial s_0 + 0.0305303| \leqq 2\times 10^{-6}$, $|\partial^2\sigma/\partial^2 s_* + 0.00369926| \leqq 3\times 10^{-7}$, *and the coefficient* a_1 *obeys* $|a_1 - 0.000212| \leqq 2\times 10^{-5}$. *Furthermore, for* $(a_0, \alpha, \kappa) = (500, 0.2, 0.1)$, *at* $(s_*, s_0) = (s^d_*(a_0, \alpha, \kappa), s^d_0(a_0, \alpha, \kappa))$,

$$\frac{\partial\sigma}{\partial s_0} = -0.030530330090811963\,58 \pm 1.8\times 10^{-19},$$

$$\frac{\partial^2\sigma}{\partial^2 s_*} = -0.003699263408188179\,71 \pm 1.1\times 10^{-19}$$

$$a_1 = 0.0002126442280682 \pm 2.4\times 10^{-15}.$$

Proof. Differentiating the identity

$$D(s_*, s_0)\sigma = A(s_*)s_0^2 + B(s_*)s_0 + C(s_*)$$

with respect to s_0 gives

$$D(s_*, s_0)\partial\sigma/\partial s_0 + \left(\frac{\partial D}{\partial s_0}\right)\sigma = 2s_0 A(s_*) + B(s_*).$$

At $(s_*, s_0) = (s^d_*, s^d_0)$, $\sigma(s^d_*, s^d_0) = 0$, and so

$$\frac{\partial\sigma}{\partial s_0}(s^d_*, s^d_0) = \frac{2s^d_0 A(s^d_*) + B(s^d_*)}{D(s^d_*, s^d_0)}.$$

Similarly, differentiating the identity twice with respect to s_* and using

$$\sigma(s^d_*, s^d_0) = \frac{\partial\sigma}{\partial s_*}(s^d_*, s^d_0) = 0,$$

gives the formula

$$\frac{\partial^2\sigma}{\partial^2 s_0}(s^d_*, s^d_0) = \frac{A''(s^d_*)(s^d_0)^2 + B''(s^d_*)s^d_0 + C''(s^d_*)}{D(s^d_*, s^d_0)}.$$

MACSYMA was used to generate analytical expressions for $\partial\sigma/\partial s_0$ and $\partial^2\sigma/\partial^2 s_*$ at (s^d_*, s^d_0) based upon these two formulas. An analytical formula for $a_1(0,0) = \operatorname{Re} c_1(0,0)$ was derived as follows: expressions for the partial derivatives $f^i_{jk} = \partial^2 f^i/\partial y_j \partial y_k$ and $f^i_{jkm} = \partial^3 f^i/\partial y_j \partial y_k \partial y_m$, $i, j, k, m = 1, 2$ were obtained; in terms of these derivatives,

$$\begin{aligned} g_{11} &= [f^1_{11} + i f^2_{11}]/4, \\ g_{02} &= [f^1_{11} - 2f^2_{12} + i(f^2_{11} + 2f^1_{12})]/4, \\ g_{20} &= [f^1_{11} + 2f^2_{12} + i(f^2_{11} - 2f^1_{12})]/4, \\ g_{21} &= [f^1_{111} + f^2_{112} + i(f^2_{111} - f^1_{112})]/8. \end{aligned}$$

At $(s_*, s_0) = (s^d_*, s^d_0)$, the eigenvalue $\lambda = \sigma + i\omega$ is pure imaginary, and the formula for c_1 simplifies, giving

$$c_1(0,0) = \frac{i}{2\omega_0}\left(g_{20}g_{11} - 2|g_{11}|^2 - \frac{1}{3}|g_{02}|^2\right) + \tfrac{1}{2}g_{21},$$

where $\omega_0 = (\rho_*\partial R/\partial a + \alpha\rho_*\partial R/\partial s)^{1/2}$. Then

$$\begin{aligned} a_1(0,0) = \operatorname{Re} c_1(0,0) &= \frac{1}{2\omega_0}(-\operatorname{Im}(g_{20}g_{11})) + \frac{1}{2}\operatorname{Re} g_{21} \\ &= \frac{-1}{2\omega_0}(\operatorname{Re} g_{20} \operatorname{Im} g_{11} + \operatorname{Im} g_{20} \operatorname{Re} g_{11}) + \tfrac{1}{2}\operatorname{Re} g_{21}. \end{aligned}$$

The formulas leading to a_1 are given in more detail in Appendix B.

A PBASIC program was then used to evaluate the above formulas, first for $(a_0, \alpha, \kappa) = (500, 0.2, 0.1)$ and then for all (a_0, α, κ) in R_0, as in Lemmas 1 and 2. The bounds for $\partial\sigma/\partial s_0, \partial^2\sigma/\partial^2 s_*$, and a_1 stated above were output from this program. □

3.3. Unfolding the degeneracy. From Lemmas 1, 2, and 3 we get the following.

THEOREM. *For each (a_0, α, κ) in R_0, for $s_0 = s_0^d(a_0, \alpha, \kappa)$ the* S-A *system has a degenerate Hopf bifurcation at $s_* = s_*^d(a_0, \alpha, \kappa)$. The form $-X^3 + \Lambda^2 X$ is the normal form for the degeneracy. There is a neighborhood of (s_*^d, s_0^d) such that if $s_0 < s_0^d$ in this neighborhood, there exist Hopf bifurcations at points $s_*^{(1)}, s_*^{(2)}$ obeying $s_*^{(1)} < s_*^d < s_*^{(2)}$. If $s_0 = s_0^d$, there is a degenerate Hopf bifurcation at $s_* = s_*^d$ in which two distinct families of periodic solution arise. If $s_0 > s_0^d$, then for an interval of values of s_*, there is a branch of periodic solutions, locally isolated from the stationary solution. All the local periodic solutions are unstable.*

Proof. By Lemma 3, the bifurcation equation

$$g(x, \lambda, \nu) = r(u, \lambda, \nu)x = 0$$

satisfies

$$r(0,0,0) = -\sigma = 0,$$

$$r_\lambda(0,0,0) = \frac{-\partial\sigma}{\partial s_*}(s_*^d, s_0^d) = 0, \qquad r_\nu(0,0,0) = \frac{-\partial\sigma}{\partial s_0}(s_*^d, s_0^d) > 0,$$

$$r_u(0,0,0) = -a_1(0,0) < 0, \qquad r_{\lambda\lambda}(0,0,0) = \frac{-\partial^2\sigma}{\partial s_4^2}(s_*^d, s_0^d) > 0.$$

From the solution of the recognition problem for Z_2-symmetric bifurcation problems [5, p. 257], it follows that for $\nu = 0$, the singularity at $x = 0$, $\lambda = 0$ is strongly Z_2-equivalent to the normal form $-X^3 + \lambda^2 X$. According to [5, p. 275], a universal unfolding for this normal form is

$$G(X, \Lambda, A) = R(X^2, \Lambda, A)X,$$

where $R(U, \Lambda, A) = -U^2 + \Lambda^2 + A$ and A is the unfolding parameter.

By the strong Z_2-equivalence of $g(x, \lambda, 0)$ and $G(X, \lambda, 0)$ there exist smooth functions $S_0(x, \lambda)$ and $x_0(x, \lambda)$ such that $S_0(-x, \lambda) = S_0(x, \lambda)$, $x_0(-x, \lambda) = -x_0(x, \lambda)$, $S_0(0,0) > 0$, $(\partial x_0/\partial x)(0,0) > 0$, and $G(x, \lambda, 0) = S_0(x, \lambda)g(x_0(x, \lambda), \lambda, 0)$; therefore, $S_0(x, \lambda)g(x_0(x, \lambda), \lambda, \nu)$ is an unfolding of the singularity $G(x, \lambda, 0)$, and as such, may be factored through the universal unfolding $G(X, \Lambda, A)$. That is, there exist smooth functions $S_1(x, \lambda, \nu)$, $X_1(x, \lambda, \nu)$, $\Lambda(\lambda, \nu)$, such that

$$S_0(x, \lambda)g(x_0(x, \lambda), \lambda, \nu) = S_1(x, \lambda, \nu)G(X_1(x, \lambda, \nu), \Lambda(\lambda, \nu), A(\nu)),$$

where $S_1(x, \lambda, 0) = 1$, $X_1(x, \lambda, 0) = x$, $\Lambda(\lambda, 0) = \lambda$, $A(0) = 0$, $S_1(-x, \lambda, \nu) = S_1(x, \lambda, \nu)$, and $X_1(-x, \lambda, \nu) = -X_1(x, \lambda, \nu)$. The next task is to show that the map $\nu \to A(\nu)$ is locally invertible.

Let $x_1(x, \lambda)$ denote the inverse of the map $x \to x_0(x, \lambda)$; let $X_2(x, \lambda, \nu) = X_1(x_1(x, \lambda), \lambda, \nu)$, and let $S_2(x, \lambda, \nu) = S_1(x_1(x, \lambda), \lambda, \nu)/S_0(x_1(x, \lambda), \lambda)$. Since $S_2(x, \lambda, \nu)$ is even in x and is smooth, it may be written as $S_2(x, \lambda, \nu) = S_3(u, \lambda, \nu)$, where $u = x^2$, and S_3 is smooth [5, p. 248]. Similarly, $X_2(x, \lambda, \nu)$ may be written as $X_2(x, \lambda, \nu) = \chi(x^2, \lambda, \nu)x$, where χ is smooth. Then

$$g(x, \lambda, \nu) = S_3(u, \lambda, \nu)G(\chi(u, \lambda, \nu)x, \Lambda(\lambda, \nu), A(\nu)),$$

which implies that

$$r(u, \lambda, \nu) = S(u, \lambda, \nu)R(\chi^2(u, \lambda, \nu)u, \Lambda(\lambda, \nu), A(\nu)),$$

where $S(u, \lambda, \nu) = S_3(u, \lambda, \nu)\chi(u, \lambda, \nu)$.

On taking first- and second-order partial derivatives of this equation for $r(u, \lambda, \nu)$ at $u = \lambda = \nu = 0$, algebraic equations relating partial derivatives of r, S, R, χ, Λ, and A are obtained. We find that

$$S(0,0,0) = \frac{r_{\lambda\lambda}}{2} > 0,$$

$$\chi(0,0,0) = \frac{-2r_u}{r_{\lambda\lambda}} > 0,$$

$$A_\nu(0) = \frac{2r_\nu}{r_{\lambda\lambda}} > 0,$$

so that the map $\nu \to A(\nu)$ is locally invertible; $A > 0$ corresponds to $\nu > 0$ ($s_0 > s_0^d$), $A = 0$ corresponds to $\nu = 0$ ($s_0 = s_0^d$), and $A < 0$ corresponds to $\nu < 0$ ($s_0 < s_0^d$). There are then three cases to consider.

(i) If $s_0 < s_0^d$ and $\nu = s_0 - s_0^d$ is sufficiently small, then the only small positive solutions of the bifurcation equation $g = 0$ are images under the inverse of the map $(x, \lambda) \to (\chi(x^2, \lambda, \nu)x, \Lambda(\lambda, \nu))$ of solutions of $R(X^2, \Lambda, A) = -X^2 + \Lambda^2 + A = 0$. Since $A = A(\nu) < 0$, $R(0, \Lambda, A) = 0$ has two roots $\Lambda_1 < \Lambda_2$. For $\Lambda < \Lambda_1$ and for $\Lambda > \Lambda_2$, $R(X^2, \Lambda, A) = 0$ has exactly one positive solution; but for $\Lambda_1 < \Lambda < \Lambda_2$, there are no (real) solutions. Let λ_1 and λ_2 correspond to Λ_1 and Λ_2, respectively. Then $r(0, \lambda_1, \nu) = 0$, $r_\lambda(0, \lambda_1, \nu) < 0$, $r_u(0, \lambda_1, \nu) < 0$, $r(0, \lambda_2, \nu) = 0$, $r_\lambda(0, \lambda_2, \nu) >$, $r_u(0, \lambda_2, \nu) < 0$, which shows the existence of points $s_*^{(1)} = s_*^d + \lambda_1$ and $s_*^{(2)} = s_*^d + \lambda_2$ of Hopf bifurcation.

(ii) If $s_0 = s_0^d$, then $\nu = 0$ and $A = 0$. For each $\Lambda < 0$, $X = -\Lambda$ is the unique solution of the equation $R = 0$, and for each $\Lambda > 0$, $X = \Lambda$ is the unique solution. These solutions correspond to two families of solutions of $g(x, \lambda, 0) = 0$, which meet at the point $(x, \lambda) = (0, 0)$.

(iii) If $s_0 > s_0^d$, then $\nu > 0$ and $A > 0$. The equation $R(X^2, \Lambda, A) = 0$ has just the branch of solutions $X = (\Lambda^2 + A)^{1/2}$, Λ arbitrary. Therefore, $g(x, \lambda, \nu) = 0$ has a local branch of periodic solutions, locally isolated from the stationary solution.

By Theorem 4.1 of [5, p. 360], a periodic solution of the ordinary differential equation corresponding to a small positive solution x of $g(x, \lambda, \nu) = 0$ is unstable if $g_x(x, \lambda, \nu) < 0$. By direct computation, for sufficiently small (x, λ, ν),

$$\operatorname{sgn} g_x(x, \lambda, \nu) = \operatorname{sgn} G_X(X, \Lambda, A) = \operatorname{sgn} R_U(X^2, A) = -1,$$

so all of the small periodic solutions found are necessarily unstable. □

3.4. Comparison of unfolding with numerical results from AUTO. Figures 1a and 1b are bifurcation diagrams containing curves that represent the families of periodic solutions of the S-A system for $s_0 = s_0^d - 0.01$ (curves 1 and 1′), s_0^d (curves 2 and 2′), $s_0^d + 0.01$ (curves 3 and 3′); all for $(a_0, \alpha, \kappa) = (500, 0.2, 0.1)$. The vertical axis is the peak-to-peak value

$$s_{pp} = \max_t s(t) - \min_t s(t)$$

of solutions $s(t; s_*, s_0)$, and the horizontal axis is the bifurcation parameter s_*. Curves 1, 2, and 3 were obtained by pseudoarclength continuation with Doedel's code AUTO.

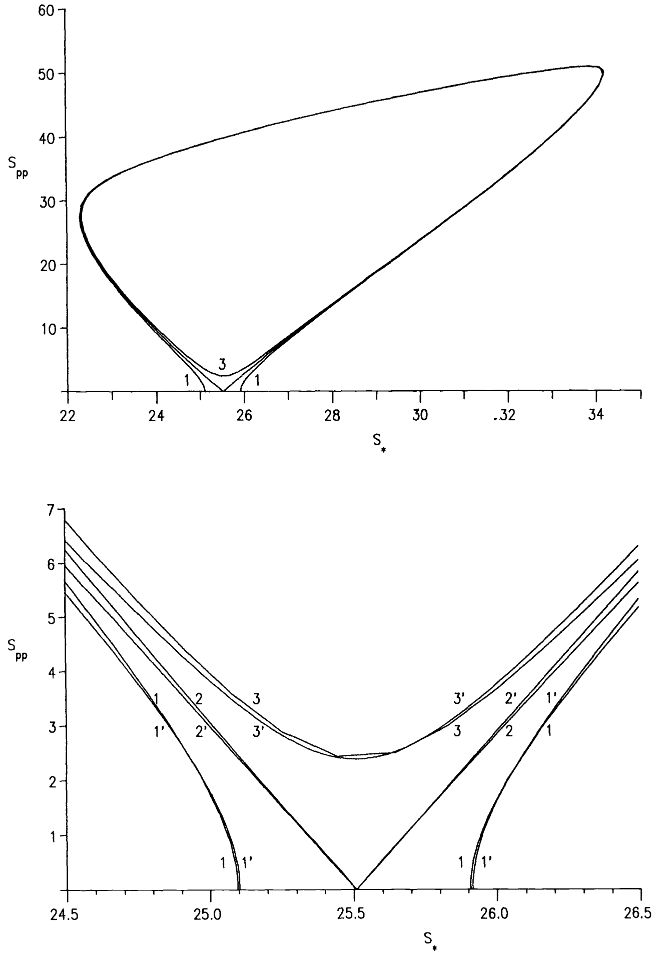


FIG. 1. *Bifurcation diagram of peak-to-peak value for s solution component versus bifurcation parameter* s_*, *for* $s_0 = s_0^d - 0.01$ (*curves* 1, 1′), $s_0 = s_0^d$ (*curves* 2, 2′), *and* $s_0 = s_0^d + 0.01$ (*curves* 3, 3′). *Curves* 1, 2, *and* 3 *are from pseudoarclength continuation* (AUTO) *Curves* 1′, 2′, *and* 3′ *are from unfolding.*

Curve 1 was computed by using AUTO to detect Hopf bifurcation from the stationary branch, and then to generate the Hopf branch by continuation from the left (smaller s_* value) bifurcation point.

Limit point continuation in the two parameters (s_*, s_0) was used to compute the left limit points (points of vertical tangency) in curves 2 and 3, starting from the solution corresponding to the left limit point in curve 1. One parameter continuation in the parameter s_* was then used to compute the other points on curves 2 and 3: since AUTO was observed to cycle when computing these curves, it was modified to stop once the value of s_* had changed direction twice. (AUTO was also modified so as to output the peak-to-peak value of s as a measure of the amplitude.)

Some angularity is visible in curves 1, 2, and 3 because discrete sets of periodic solutions are used to approximate smooth branches of solutions.

Curve 1, as computed, does not return precisely to the right point of Hopf bifurcation, so the data was "touched up" by including results of computing the branch starting from the right bifurcation point. Similarly, as computed, curve 2 does not touch the horizontal axis; rather, it approaches the axis and then jumps to the "outgoing" set of periodic solutions. This data was "touched up" by inclusion of the stationary solution for $s_* = s_*^d$.

In Fig. 1b, curves 1′, 2′, and 3′ were obtained by plotting the zero set of

$$\frac{\partial^2\sigma}{\partial^2 s_*}(s_*^d, s_0^d)\frac{(s_* - s_*^d)^2}{2} + a_1(0,0)\left(\frac{s_{pp}}{2}\right)^2 + \frac{\partial\sigma}{\partial s_0}(s_*^d, s_0^d)(s_0 - s_0^d).$$

This is a Taylor expansion of the factor $r(x^2, \lambda, \nu)$ of the bifurcation equation, retaining terms that correspond to terms in the factor $R(X^2, \Lambda, A)$ of the universal unfolding, and with x approximated by $s_{pp}/2$. The Lyapunov-Schmidt reduction approximates the periodic solutions $\xi(t)$ of the P-B form as $x\,e^{it'} + O(x^2)$, where t' is a scaled time variable. From this approximation it follows that $s_{pp} = 2x + O(x^2)$.

The agreement between the bifurcation theoretic curves 1′, 2′, 3′ and the numeric curves 1, 2, 3 is excellent. Curves 1′, 2′, and 3′ are symmetric about $s_* = s_*^d$. Curves 1, 2, and 3 are approximately symmetric, with the deviations from symmetry becoming larger further away from $(s_*, s_{pp}) = (s_*^d, 0)$. Curves 1 and 1′ each indicate Hopf bifurcations at approximately the same values $s_*^{(1)}$ and $s_*^{(2)}$. Curve 2′ consists of the two tangents to Curve 2. On curves 3 and 3′, the minimum value of s_{pp} is approximately the same. All of these features support the correctness of the values computed in Lemma 3.

4. Discussion. The main point of the present paper is to establish rigorously the existence of certain families of periodic solutions of the S-A system, both for a specific set of parameter values (a_0, α, κ) and for a nontrivial region R_0 of such values. This was accomplished by using symbolic manipulation to perform algebraic manipulations more complicated than we would normally perform by hand, and interval analysis to perform numerical computations "precisely" in the sense of Aberth's monograph [1]. We had hoped to obtain results valid for a larger region in the parameters. For $\alpha = 0.2$, and $\kappa = 0.1$, we have performed traditional numerical computations that indicate that $a_1 = 0$ for $a_0 \approx 330$, i.e., a higher-order degeneracy is present in the model. In future work, we intend to use a continuation argument to extend Lemmas 1 and 2 to a region including this degeneracy, and to construct an unfolding of this degeneracy as well. For application of a continuation argument to be practical, however, the individual overlapping regions in parameter space must be larger than the present region R_0, which means that a different type of interval arithmetic must be employed.

The loss of precision in the computed quantities as we proceed from Lemma 1 through to the theorem is normal for computations using the type of interval arithmetic implemented within PBASIC. In PBASIC, floating point centenary (base 100) arithmetic is used. An arbitrary number of centenary digits is used for the mantissa of the midpoint, but only a single digit is used for the range; this digit is a bound on the error in the last mantissa digit. Whenever the range of a computed expressions exceeds the capacity of a single centenary digit, the number of mantissa digits is decreased and subsequent computations are in reduced precision. This design economizes on storage space for intervals and on computational effort. The design works well to find individual computable numbers (such as $s_*^d(500, 0.2, 0.1)$) to "arbitrary" precision: to compensate for loss of precision during a computation, a higher degree of precision is used at the start. The design is less satisfactory when the object of a computation is to establish inequalities such as $a_1(s_*^d, s_0^d) > 0$ for all (a_0, α, κ) in R_0. In this case, compensating for loss of precision during the computation requires additional restrictions on the parameters is the reason that the region R_0 in the present work is so small.

To minimize such restrictions on the parameters, a type of interval arithmetic must be used that exhibits less loss of precision during the computation. We expect that this could be accomplished by representing both the midpoint and range of the intervals as arbitrary precision numbers.

To contrast the present work with more traditional numerical computations, we note that values for s_*^d, s_0^d, and a_1 at $(a_0, \alpha, \kappa) = (500, 0.2, 0.1)$ could have been calculated to arbitrary precision using either MACSYMA or Mathematica. By recomputing with successively higher precision arithmetic, digits that stop changing as the precision of the arithmetic is increased would be taken as "correct." This procedure for roundoff error estimation does not, however, establish true error bounds. Although relatively unsophisticated in some respects, the language PBASIC used in the present study does provide arbitrary precision *interval* arithmetic as required to make mathematically rigorous statements.

The work [3] outlines several alternate derivations of Hopf bifurcation theory, and presents formulas for bifurcation coefficients derived by application of Lyapunov-Schmidt reduction without preliminary transformation to P-B normal form. To relate our $a_1(s_*^d, s_0^d)$ to the coefficient p_{100} of [3], consider the slopes of the right and left tangent lines to curve 2 of Fig. 1 for $s_* = s_*^d$, $s_{pp} = 0$. These slopes are $\pm (-2(\partial^2\sigma/\partial s_*^2)/a_1)^{1/2}$. Based on results in [3], these same slopes are $\pm|c^1|(-2(\partial^2\sigma/\partial s_*^2)/P_{100})^{1/2}$, where c^1 the first component of the eigenvector c. (In terms of the approximate periodic solution in [3], $s_{pp} = 2\bar{x}|c^1| + 0(\bar{x}^2)$, where $\bar{x}$ is the coordinate on the kernel.) Because the geometry of curve 2 is independent of the particular derivation of bifurcation coefficients, it follows that $a_1 = p_{100}/|c^1|^2$. It should also be possible to establish this algebraically, but the geometric argument is simpler. We verified the relationship numerically. In this application, the choice of technique in deriving Hopf bifurcation seems to be a matter of taste.

Appendix A. Terms from singularity theory.

Z_2-equivalence. Let $g(x, \lambda)$ and $h(x, \lambda)$ be bifurcation problems with Z_{2-} symmetry. Then g and h are Z_2-equivalent if

$$h(x, \lambda) = S(x, \lambda) g(X(x, \lambda), \Lambda(\lambda)),$$

where the triple (S, X, Λ) is an equivalence transformation (i.e., S is nonzero and positive and (X, Λ): $(x, \lambda) \rightarrow (X(x, \lambda), \Lambda(\lambda))$ is a local diffeomorphism that preserves the orientation of x and λ; that is, $X_x(x, \lambda) > 0$ and $\Lambda'(\lambda) > 0$) such that X is odd in x and S is even in x. If this relation holds with $\Lambda(\lambda) = \lambda$, then g and h are strongly Z_2-equivalent.

Factors through. Let $G(x, \lambda, \alpha)$ and $H(x, \lambda, \beta)$ be unfolding of a germ g. We say that H factors through G if there exist smooth mappings S, X, Λ, and A such that

$$H(x, \lambda, \beta) = S(x, \lambda, \beta) G(X(x, \lambda, \beta), \Lambda(\lambda, \beta), A(\beta)),$$

where for $\beta = 0$ the following hold: $S(x, \lambda, 0) = 1$, $X(x, \lambda, 0) = x$, $\Lambda(\lambda, 0) = \lambda$, and $A(0) = 0$.

Unfolding. Let g be in $\varepsilon_{x,\lambda}$ [5, p. 56]; a k-parameter unfolding of g is a germ $G \in \varepsilon_{x,\lambda,\alpha}$, where $\alpha = (\alpha_1, \cdots, \alpha_k) \in R^k$, such that for $\alpha = 0$, $G(x, \lambda, 0) = g(x, \lambda)$. Here G is a germ in all variables: x, λ, $\alpha_1, \cdots \alpha_k$. Thus G is defined and C^∞ on a neighborhood of zero in R^{k+2}.

Universal unfolding, codimension. An unfolding G of g is versal if every other unfolding of g factors through G. A versal unfolding of g depending on the minimum number of parameters possible is called universal. The minimum number is called the codimension of g.

Appendix B: formulas. The polynomial $P(s;\, a_0, \alpha, \kappa)$ is given by the expression

$$\begin{aligned}
P(s;\, a_0, \alpha, \kappa) = {} & 4\kappa^5 s^{11} + (-16a_0\alpha^2\kappa^5 - 8a_0\alpha\kappa^5 + 8\kappa^4)s^{10} \\
& + (-2a_0^2\alpha^4\kappa^5 - 2a_0^2\alpha^3\kappa^5 + 4a_0^2\alpha^2\kappa^5 - 16a_0\alpha^2\kappa^4 - 8a_0\alpha\kappa^4 \\
& \quad - 4\kappa^4 + 4\kappa^3)s^9 + (48a_0\alpha^2\kappa^4 + 24a_0\alpha\kappa^4 - 20\kappa^3)s^8 \\
& + (16a_0^2\alpha^4\kappa^4 + 16a_0^2\alpha^3\kappa^4 - 20a_0^2\alpha^2\kappa^4 + 80a_0\alpha^2\kappa^3 + 40a_0\alpha\kappa^3 \\
& \quad - 20\kappa^3 - 20\kappa^2)s^7 + (8a_0^2\alpha^4\kappa^3 + 8a_0^2\alpha^3\kappa^3 - 4a_0^2\alpha^2\kappa^3 \\
& \quad + 16a_0\alpha^2\kappa^3 + 8a_0\alpha\kappa^3 + 16a_0\alpha^2\kappa^2 + 8a_0\alpha\kappa^2 - 32\kappa^2 - 4\kappa)s^6 \\
& + (-28a_0^2\alpha^4\kappa^3 - 28a_0^2\alpha^3\kappa^3 + 28a_0^2\alpha^2\kappa^3 - 48a_0\alpha^2\kappa^2 - 24a_0\alpha\kappa^2 \\
& \quad - 12\kappa^2 - 8\kappa)s^5 + (-32a_0^2\alpha^4\kappa^2 - 32a_0^2\alpha^3\kappa^2 + 8a_0^2\alpha^2\kappa^2 \\
& \quad - 48a_0\alpha^2\kappa^2 - 24a_0\alpha\kappa^2 - 16a_0\alpha^2\kappa - 8a_0\alpha\kappa - 4\kappa)s^4 \\
& + (-16a_0^2\alpha^4\kappa^2 - 16a_0^2\alpha^3\kappa^2 - 12a_0^2\alpha^2\kappa^2 - 8a_0^2\alpha^4\kappa - 8a_0^2\alpha^3\kappa \\
& \quad - 16a_0\alpha^2\kappa - 8a_0\alpha\kappa)s^3 + (-8a_0^2\alpha^4\kappa - 8a_0^2\alpha^3\kappa - 4a_0^2\alpha^2\kappa)s^2 \\
& + (-2a_0^2\alpha^4\kappa - 2a_0^2\alpha^3\kappa)s.
\end{aligned}$$

The formula for the coefficient a_1 is as follows: at $(s, s_0) = (s_*^d, s_0^d)$, form

$$\begin{aligned}
a_* &= a_0 + (1/\alpha)(s - s_0), \\
R &= sa_*/(1 + s + \kappa s^2), \\
R_s &= a_*(1 - \kappa s^2)/(1 + s + \kappa s^2)^2, \\
R_a &= s/(1 + s + \kappa s^2), \\
\rho_* &= (s_0 - s)/R, \\
\omega_0 &= (\alpha + \rho_* R_a + \alpha\rho_* R_s)^{1/2}, \\
\partial s/\partial y_1 &= 1, \\
\partial a/\partial y_1 &= -(1 + \rho_* R_s)/(\rho_* R_a), \\
\partial a/\partial y_2 &= \omega_0/(\rho_* R_a), \\
p_3 &= \rho_* R_a/\omega_0, \\
p_4 &= (1 + \rho_* R_s)/\omega_0, \\
p &= p_3 + p_4, \\
R_{sa} &= (1 - \kappa s^2)/(1 + s + \kappa s^2)^2, \\
R_{ss} &= a_*(-2 - 6\kappa s + 2\kappa^2 s^3)/(1 + s + \kappa s^2)^3, \\
R_{sss} &= a_*(6 - 6\kappa + 24\kappa s + 36\kappa^2 s^2 - 6\kappa^3 s^4)/(1 + s + \kappa s^2)^4, \text{ and} \\
R_{ssa} &= R_{ss}/a_*.
\end{aligned}$$

The partial derivatives of f are then given by the formulas

$$
\begin{aligned}
&f_{111}=(-\rho_*)(R_{ss}(\partial s/\partial y_1)^2+2R_{sa}(\partial a/\partial y_1)(\partial s/\partial y_1)),\\
&f_{112}=(-\rho_*)R_{sa}(\partial a/\partial y_2)(\partial s/\partial y_1),\\
&f_{211}=pf_{111},\\
&f_{212}=pf_{112},\\
&f_{1111}=(-\rho_*)(R_{sss}(\partial s/\partial y_1)^3+3R_{ssa}(\partial a/\partial y_1)(\partial s/\partial y_1)^2),\\
&f_{1112}=(-\rho_*)R_{ssa}(\partial a/\partial y_2)(\partial s/\partial y_1)^2,\\
&f_{2111}=pf_{1111}, \text{ and}\\
&f_{2112}=pf_{1112}.
\end{aligned}
$$

The coefficient a_1 is then given by

$$a_1=-\frac{1}{2\omega_0}(\operatorname{Re} g_{20} \operatorname{Im} g_{11}+\operatorname{Im} g_{20} \operatorname{Re} g_{11})+\frac{1}{2} \operatorname{Re} g_{21},$$

where

$$
\begin{aligned}
&\operatorname{Re} g_{11}=f_{111}/4,\\
&\operatorname{Im} g_{11}=f_{211}/4,\\
&\operatorname{Re} g_{20}=(f_{111}+2f_{212})/4,\\
&\operatorname{Im} g_{20}=(f_{211}-2f_{112})/4, \text{ and}\\
&\operatorname{Re} g_{21}=(f_{1111}+f_{2112})/8.
\end{aligned}
$$

REFERENCES

[1] O. Aberth, *Precise Numerical Analysis*, Wm. C. Brown, Dubuque, Iowa, 1988.

[2] E. Doedel, AUTO: *Software for continuation and bifurcation problems in Ordinary Differential Equations*, Monograph, Applied Mathematics, California Institute of Technology, Pasadena, CA, May 1986.

[3] W. W. Farr, C. Li, I. S. Labouriau, and W. F. Langford, *Degenerate Hopf bifurcation formulas and Hilbert's 16th problem*, SIAM J. Math. Anal., 20 (1989), pp. 13-30.

[4] M. Golubitsky and W. F. Langford, *Classification and unfoldings of degenerate Hopf bifurcations*, J. Differential Equations, 41, 3 (1981), pp. 375-415.

[5] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory* I, Springer-Verlag, New York, 1985.

[6] B. D. Hassard, N. D. Kazarinoff, and Y-H Wan, *Theory and Applications of Hopf Bifurcation*, Cambridge University Press, London, 1981.

SIAM J. MATH ANAL.
Vol. 23, No. 5, pp. 1305-1308, September 1992

SOME PROPERTIES OF THE FIRST EIGENVALUE OF THE LAPLACE OPERATOR ON THE SPHERICAL BANDS IN S^2*

CHAO-LIANG SHEN† AND CHUNG-TSUN SHIEH†

Abstract. On the unit sphere in the three-dimensional Euclidean space, among all spherical bands with a given area $2\pi A$, the spherical band which is symmetric to the equator has the largest first Dirichlet eigenvalue.

Key words. first eigenvalue, Laplace operator, spherical bands

AMS(MOS) subject classifications. 35B05, 35P99

1. Introduction. In this paper the subject under investigation is the first eigenvalue of the vibrating membrane equation with Dirichlet boundary condition over those spherical bands with parallel boundary circles on the unit sphere S^2 in $\mathbb{R}^3$. We prove that among all spherical bands with a given area $2\pi A$, where $0<A<2$, the one which is symmetric to the equator has the largest eigenvalue. We also find a monotonicity property of the first eigenvalue when the spherical bands move on the unit sphere. These results are stated and proved in Theorem 1 of the next section.

2. Main results. We express points on the unit sphere S^2 by the Euler coordinate (θ, φ), where $0 \leqq \theta \leqq 2\pi$, $0 \leqq \varphi \leqq \pi$. For $0 \leqq \xi < \eta \leqq \pi$, $0<A<2$, if the *spherical band* bounded by $\varphi = \xi$ and $\varphi = \eta$ has area $2\pi A$, then η can be expressed in terms of ξ and A as follows:

$$\eta = \cos^{-1}(\cos \xi - A).$$

We denote this η by $f(\xi, A)$, or $f(\xi)$ if A is fixed. From now on we shall fix A.

Let Δ_{S^2} denote the Laplace operator on S^2. For $0 \leqq \xi < \pi$, let $B(\xi)$ denote the spherical band on S^2 with area $2\pi A$ and bounded by $\varphi = \xi$ and $\varphi = f(\xi)$. Let $\lambda_1(\xi)$ denote the first eigenvalue of the following eigenvalue problem:

$$\begin{aligned} &\Delta_{S^2} u + \lambda u = 0 \quad \text{in } B(\xi), \\ (1) \qquad & u = 0 \quad \text{on } \partial B(\xi), \end{aligned}$$

where $\partial B(\xi) = \{(\theta, \varphi)\colon 0 \leqq \theta \leqq 2\pi,\ \varphi = \xi \text{ or } f(\xi)\}$ denotes the boundary of $B(\xi)$, and

$$\Delta_{S^2} u(\theta, \varphi) = \frac{1}{\sin \varphi}\left[\frac{\partial}{\partial \varphi}(\sin \varphi u_\varphi) + \frac{\partial}{\partial \theta}\left(\frac{1}{\sin \varphi} u_\theta\right)\right].$$

Since $B(\xi)$ is rotationally symmetric in θ-variable, and the first eigenvalue of (1) is nondegenerate, we see that the first eigenfunction u_1 of (1) is independent of the variable θ, and $(\lambda_1(\xi), u_1)$ is the first eigenpair of the following eigenvalue problem:

$$\begin{aligned} &[\sin \varphi v'(\varphi)]' + \lambda \sin \varphi v(\varphi) = 0, \qquad \xi < \varphi < f(\xi), \\ (2) \qquad & v(\xi) = v(f(\xi)) = 0. \end{aligned}$$

* Received by the editors June 25, 1990; accepted for publication (in revised form) January 21, 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.

Recall that $f(\xi)=\cos^{-1}(\cos\xi-A)$, where A is fixed, $0<A<2$. For $\xi\leqq\varphi\leqq f(\xi)$, introducing the new variable

$$a=a(\varphi)=\int_\xi^\varphi \sin t\,dt,$$

and letting

$$w(a)=v(\varphi),$$

(2) can be written as follows:

$$\begin{aligned}&\{[1-(\cos\xi-a)^2]w'(a)\}'+\lambda w(a)=0, \qquad 0<a<A,\\ &w(0)=w(A)=0.\end{aligned}\tag{3}$$

We shall use the notation $\lambda_1(\xi)$ to denote the first eigenvalue of (3) and $w(\xi,a)$ to denote the positive first eigenfunction of (3) that is *normalized*, i.e., $\int_0^A w^2(\xi,a)\,da=1$. Then it follows from the results in § 2.6, Chapter VI of [2] that $\lambda_1(\xi)$ is continuous on the interval $[0,\pi-\cos^{-1}(1-A)]$. Furthermore, by [1, Lemma 3.15], both $w(\xi,a)$ and $dw(\xi,a)/da$ are real analytic in ξ; hence the function $\lambda_1(\xi)$ is real analytic in ξ. Now we present our main results in the following theorem.

THEOREM 1. *As a function of* ξ *on the interval* $[0,\pi-\cos^{-1}(1-A)]$, *the first eigenvalue* $\lambda_1(\xi)$ *of* (1) *attains its maximum at* $\xi=\cos^{-1}(A/2)$, *i.e., when* $B(\xi)$ *is the spherical band symmetric to the equator. Furthermore, the function* $\lambda_1(\xi)$ *is monotonically increasing on* $0\leqq\xi\leqq\cos^{-1}(A/2)$.

Proof. By the equatorical symmetry of the spherical bands $B(\xi)$ and $B(\pi-\cos^{-1}(\cos\xi-A))$, we have $\lambda_1(\xi)=\lambda_1(\pi-\cos^{-1}(\cos\xi-A))$. Thus $\lambda_1'(\cos^{-1}(A/2))=0$, where $\lambda_1'(\xi)=d\lambda_1(\xi)/d\xi$. For the proof of Theorem 1 we need to compute $\lambda_1'(\xi)$. Let $w'(\xi,a)$ denote $dw(\xi,a)/da$, and recall that $\int_0^A w^2(\xi,a)\,da=1$.

By using $w(\xi+\Delta\xi,a)$ (respectively, $w(\xi,a)$) as a testing function to estimate $\lambda_1(\xi)$ (respectively, $\lambda_1(\xi+\Delta\xi)$), it follows from the Rayleigh quotient of (3) and the minimum principle that we have the following inequalities:

$$\int_0^A [1-(\cos\xi-a)^2][w'(\xi+\Delta\xi,a)]^2\,da\geqq\lambda_1(\xi),\tag{4}$$

$$\int_0^A [1-(\cos(\xi+\Delta\xi)-a)^2][w'(\xi,a)]^2\,da\geqq\lambda_1(\xi+\Delta\xi).\tag{5}$$

By (3)-(5), we have the following inequalities:

$$\begin{aligned}&\int_0^A\{[1-(\cos(\xi+\Delta\xi)-a)^2]-[1-(\cos\xi-a)^2]\}[w'(\xi+\Delta\xi,a)]^2\,da\\ &\leqq\lambda_1(\xi+\Delta\xi)-\lambda_1(\xi)\\ &\leqq\int_0^A\{[1-(\cos(\xi+\Delta\xi)-a)^2]-[1-(\cos\xi-a)^2]\}[w'(\xi,a)]^2\,da.\end{aligned}\tag{6}$$

The inequalities in (6) imply

$$\begin{aligned}&2[(\sin\xi)\Delta\xi]\int_0^A(\cos\xi-a)[w'(\xi+\Delta\xi,a)]^2\,da+o(\Delta\xi)\\ &\leqq\lambda_1(\xi+\Delta\xi)-\lambda_1(\xi)\\ &\leqq 2[(\sin\xi)\Delta\xi]\int_0^A(\cos\xi-a)[w'(\xi,a)]^2\,da+o(\Delta\xi).\end{aligned}\tag{7}$$

By (7) and [1, Lemma 3.15] we have

$$\lambda_1'(\xi)=2\sin\xi\int_0^A(\cos\xi-a)[w'(\xi,a)]^2\,da. \tag{8}$$

Thus

$$\lambda_1'\left(\cos^{-1}\left(\frac{A}{2}\right)\right)=2\sin\left(\cos^{-1}\left(\frac{A}{2}\right)\right)\int_0^A\left(\frac{A}{2}-a\right)\left[w'\left(\cos^{-1}\left(\frac{A}{2}\right),a\right)\right]^2da=0,$$

which implies

$$\int_0^A\frac{A}{2}\left[w'\left(\cos^{-1}\left(\frac{A}{2}\right),a\right)\right]^2da=\int_0^A a\left[w'\left(\cos^{-1}\left(\frac{A}{2}\right),a\right)\right]^2da. \tag{9}$$

Let

$$H(\xi)=\int_0^A[1-(\cos\xi-a)^2]\left[w'\left(\cos^{-1}\left(\frac{A}{2}\right),a\right)\right]^2da.$$

Then by the assumption $\int_0^A[w(\cos^{-1}(A/2),a)]^2\,da=1$, we have $H(\cos^{-1}(A/2))=\lambda_1(\cos^{-1}(A/2))$ and

$$\lambda_1(\xi)\leqq H(\xi),\qquad 0\leqq\xi\leqq\cos^{-1}\left(\frac{A}{2}\right). \tag{10}$$

We also have

$$H'(\xi)=2\sin\xi\int_0^A(\cos\xi-a)\left[w'\left(\cos^{-1}\left(\frac{A}{2}\right),a\right)\right]^2da. \tag{11}$$

By (9) and (11), for $0\leqq\xi\leqq\cos^{-1}(A/2)$, we have

$$H'(\xi)\geqq 0 \tag{12}$$

since $\cos\xi\geqq A/2$. By (10) and (12), we have

$$\lambda_1(\xi)\leqq H\left(\cos^{-1}\left(\frac{A}{2}\right)\right)=\lambda_1\left(\cos^{-1}\left(\frac{A}{2}\right)\right).$$

This proves the first part of Theorem 1.

To prove the second part of Theorem 1, by the result of the first part and the fact that $\lambda_1(0)$ is the minimum of $\lambda_1(\xi)$ (see [3]), it is sufficient to prove that $\lambda_1(\xi)$ does not have any critical point in the open interval $0<\xi<\cos^{-1}(A/2)$. We prove it by contradiction. Assume $0<\xi_0<\cos^{-1}(A/2)$ such that

$$\lambda_1'(\xi_0)=2\sin\xi_0\int_0^A(\cos\xi_0-a)[w'(\xi_0,a)]^2\,da=0. \tag{13}$$

Since $0<A<2$, $\cos\xi$ is strictly decreasing in the interval $0\leqq\xi\leqq\cos^{-1}(A/2)$. Thus by (13) we have

$$\int_0^A(\cos\xi-a)[w'(\xi_0,a)]^2\,da\leqq 0 \tag{14}$$

for $\xi_0\leqq\xi\leqq\cos^{-1}(A/2)$, where the equality holds only when $\xi=\xi_0$. Define

$$H_0(\xi)=\int_0^A[1-(\cos\xi-a)^2][w'(\xi_0,a)]^2\,da.$$

Then it follows from (14) that $H_0(\xi)$ is strictly decreasing in the interval $[\xi_0, \cos^{-1}(A/2)]$, i.e., $H_0(\xi_0) > H_0(\cos^{-1}(A/2))$ if $\xi_0 \neq \cos^{-1}(A/2)$. But, as $H_0(\xi_0) = \lambda_1(\xi_0)$, $H_0(\cos^{-1}(A/2)) \geqq \lambda_1(\cos^{-1}(A/2))$, we have $\lambda_1(\xi_0) > \lambda_1(\cos^{-1}(A/2))$, which is absurd. Thus $\lambda_1(\xi)$ is strictly increasing in the interval $(0, \cos^{-1}(A/2))$. The proof of Theorem 1 is complete. □

The proof of Theorem 1 generalizes to certain surfaces of revolution. Let $f(x)$ be a smooth strictly increasing function on the interval $[0, \infty)$. Suppose $f'(0) = 0$. Let S_f denote the surface obtained by revolving f with respect to the y-axis. Let A be a fixed positive number. For each $\xi \geqq 0$, define $F(\xi, A)$ by the identity

$$\int_{\xi}^{F(\xi,A)} x\sqrt{1+[f'(x)]^2}\,dx = A,$$

and let $S_f(\xi)$ denote the band on S_f bounded by $x = \xi$ and $x = F(\xi, A)$. Let $\mu_1(\xi)$ denote the first Dirichlet eigenvalue of the Laplace operator on the surface $S_f(\xi)$. Then the proof of Theorem 1 can be generalized to show that $\mu_1(\xi)$ is an increasing function of ξ. We leave the proof to the reader.

Acknowledgment. The authors wish to thank one of the referees of this paper, whose suggestions made this paper shorter than its original form.

REFERENCES

[1] M. BERGER, *Sur les premières valeurs propres des variétés riemanniennes*, Compositio Math., 26 (1973), pp. 129–149.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. I, Interscience, New York, 1953.

[3] J. PEETRE, *A generalization of Courant's nodal domain theorem*, Math. Scand., 5 (1957), pp. 15–20.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1309–1331, September 1992

015

SUPERRESOLUTION VIA SPARSITY CONSTRAINTS*

DAVID L. DONOHO[†]

Abstract. Consider the problem of recovering a measure μ supported on a lattice of span Δ, when measurements are only available concerning the Fourier Transform $\hat{\mu}(\omega)$ at frequencies $|\omega| \leq \Omega$. If Ω is much smaller than the Nyquist frequency π/Δ and the measurements are noisy, then, in general, stable recovery of μ is impossible. In this paper it is shown that if, in addition, we know that the measure μ satisfies certain sparsity constraints, then stable recovery is possible. Say that a set has *Rayleigh index* less than or equal to R if in any interval of length $4\pi/\Omega \cdot R$ there are at most R elements. Indeed, if the (unknown) support of μ is known, a priori, to have *Rayleigh index* at most R, then stable recovery is possible with a stability coefficient that grows at most like Δ^{-2R-1} as $\Delta \to 0$. This result validates certain practical efforts, in spectroscopy, seismic prospecting, and astronomy, to provide superresolution by imposing support limitations in reconstruction. The results amount to inequalities for interpolation of entire functions of exponential type from values at special point sets which are irregular, yet internally balanced, uniformly discrete, and of uniform density 1.

Key words. inverse problems, spectroscopy, diffraction-limited imaging, Rayleigh criterion, Nyquist rate, superresolution, nonlinear recovery, entire functions of exponential type, interpolation, balayage

AMS(MOS) subject classifications. 42A70, 30D15, 94A12

1. Introduction.

Let $\mu = \sum_{k=-\infty}^{\infty} \alpha_k \delta_{k\Delta}$ be a signed measure supported on the lattice $\{k\Delta\}_{k=-\infty}^{\infty}$, with signed mass α_k attached to the point $k\Delta$. We think of the lattice span Δ as a small number $\Delta \ll 1$. The measure μ may be interpreted as a caricature of certain scientifically interesting objects: for example, a polarized spectrum in a spectroscopy problem; or, in exploration seismology and in medical ultrasound, as the sequence of reflectivities of a layered medium with layers of constant width Δ.

Suppose we obtain noisy measurements on μ in the frequency domain, with frequency cutoff Ω:

$$y(\omega) = \hat{\mu}(\omega) + z(\omega), \qquad |\omega| \leq \Omega. \tag{1}$$

Here $\hat{\mu}(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} d\mu(t) = \sum_{-\infty}^{\infty} \alpha_k e^{-i\omega k\Delta}$ is the Fourier transform of μ, and $z(\omega)$ represents noise. Our objective is to recover μ (or, equivalently, the coefficients α_k) from the data (1).

To make the phrase "recovery of μ" precise, we adapt some notions from the theory of *optimal recovery* (compare Micchelli and Rivlin [21]). We suppose first of all that the noise z can be any function in $L_2[-\Omega, \Omega]$ satisfying $||z||_{L_2[-\Omega,\Omega]} \leq \epsilon$. Let $\tilde{\mu} = \tilde{\mu}(y)$ be a method of recovery. We measure the recovery error with respect to the quadratic "Wiener" norm $||\tilde{\mu}(y) - \mu||_2$, where, for a discrete signed measure ν, we define $||\nu||_2 \equiv (\sum_{t\in\mathrm{supp}(\nu)} |\nu(\{t\})|^2)^{1/2}$. We record our a priori information about μ by saying that $\mu \in \mathcal{M}$, where $\mathcal{M}$ is a class of measures. For example, if our prior information is that μ is a lattice measure, i.e., a member of the set $\mathcal{L}(\Delta) = \{\mu : \mu = \sum_{k=-\infty}^{\infty} \alpha_k \delta_{k\Delta}\}$, then we set $\mathcal{M} = \mathcal{L}(\Delta)$. In general, we measure

*Received by the editors March 13, 1991; accepted for publication (in revised form) January 31, 1992. This work was supported by National Science Foundation grant DMS 88-10192, by NASA contract NCA2-0488, and by grants from Schlumberger-DOll Research, Western Geophysical, and from A.T.T. Foundation.

[†] Department of Statistics, Stanford University, Stanford, California 94305-4065.

the difficulty of recovery with a priori information $\mathcal{M}$ by the minimax error over $\mathcal{M}$:

$$E^*(\epsilon; \mathcal{M}, \Omega) = \inf_{\tilde{\mu}} \sup_{\mu \in \mathcal{M}} \sup_{||z||_{L_2[-\Omega,\Omega]} \leq \epsilon} ||\tilde{\mu}(y) - \mu||_2.$$

In particular, we say that stable recovery of μ is possible, with a priori information $\mathcal{M}$, if

$$E^*(\epsilon; \mathcal{M}, \Omega) \leq \text{Const} \cdot \epsilon,$$

which indicates the stability property

$$\text{Error} \leq \text{Constant} \cdot \text{Noise Level}.$$

1.1. The stability threshold $\Omega \geq \pi/\Delta$. π/Δ is the usual Nyquist frequency for samples taken on the lattice $\{k\Delta\}_{k=-\infty}^{\infty}$. This frequency pops up in our problem: if the frequency cutoff Ω exceeds π/Δ then stable recovery of μ is possible. Indeed, motivated by the Fourier inversion formula

$$\alpha_k = \frac{\Delta}{2\pi} \int_{-\pi/\Delta}^{\pi/\Delta} e^{i\omega k\Delta} \hat{\mu}(\omega) d\omega \tag{2}$$

one is led immediately to the rule

$$\tilde{\alpha}_k = \frac{\Delta}{2\pi} \int_{-\pi/\Delta}^{\pi/\Delta} e^{i\omega k\Delta} y(\omega) d\omega. \tag{3}$$

Parseval's relation implies that the reconstruction formula $\tilde{\mu} = \sum_{k=-\infty}^{\infty} \tilde{\alpha}_k \delta_{k\Delta}$ has error

$$||\tilde{\mu}(y) - \mu||_2^2 = \frac{\Delta}{2\pi} \int_{-\pi/\Delta}^{\pi/\Delta} |z(\omega)|^2 d\omega \leq \frac{\Delta}{2\pi} \epsilon^2.$$

Consequently, if $\mathcal{M} = \mathcal{L}(\Delta)$ we get

$$E^*(\epsilon; \mathcal{M}, \Omega) \leq \sqrt{\frac{\Delta}{2\pi}} \cdot \epsilon,$$

and an extra argument shows that in fact equality holds. Hence, in the case $\Omega \geq \pi/\Delta$, stable recovery is possible, and in fact optimal recovery requires only the simplest of linear reconstruction formulas.

The case $\Omega \ll \pi/\Delta$ is more interesting. In this case data on $\hat{\mu}(\omega)$ are not available on the whole range [−Nyquist, Nyquist], and formulas like (3) are not immediately applicable. Indeed, formula (3) suggests that, if $\Omega < \pi/\Delta$, reconstruction will require some sort of process of *extrapolation* of the noisy measurements inside $[-\Omega, \Omega]$ to produce quasi measurements over the whole of the fundamental interval $[-\pi/\Delta, \pi/\Delta]$.

However, such extrapolation is evidently impossible in the absence of special prior information. Indeed, if μ can be any lattice measure, then (α_k) can be any square summable sequence. We could therefore let $\hat{\mu}(\omega)$ be a nonzero function, periodic of period $2\pi/\Delta$, belonging to L^2 on the fundamental interval $(-\pi/\Delta, \pi/\Delta)$, and vanishing on $[-\Omega, \Omega]$. The sequence (α_k) obtained from (2) would give a nonzero lattice measure μ whose transform agrees, over the low frequency band $|\omega| \leq \Omega$, with the zero measure. Even at noise level $\epsilon = 0$, our observations could not distinguish this μ from the zero measure, nor from its sign-reversal $-\mu$. Consequently, if $\Omega < \pi/\Delta$,

$$\begin{aligned} E^*(\epsilon, \mathcal{L}(\Delta), \Omega) &\geq \sup\{\|\mu - \tilde{\mu}\|_2 : \mu, \tilde{\mu} \in \mathcal{L}(\Delta), \hat{\mu}(\omega) = \hat{\tilde{\mu}}(\omega), |\omega| \leq \Omega\} \\ &\geq \sup\{\|\mu\|_2 : \mu \in \mathcal{L}(\Delta), \hat{\mu}(\omega) = 0, |\omega| \leq \Omega\} \\ &= +\infty. \end{aligned}$$

Stable recovery is not possible under the condition $\Omega < \pi/\Delta$ if all we know a priori is the lattice constraint $\mu \in \mathcal{L}(\Delta)$.

1.2. The Rayleigh threshold $\Omega \geq \pi/\Delta$. A mathematically equivalent reformulation of our problem occurs in the theory of optics. The frequency-domain data (1) are equivalent to the spatial-domain data

$$(4) \qquad Y(t) = (K_\Omega \star \mu)(t) + Z(t), \qquad t \in (-\infty, \infty),$$

where $K_\Omega(t)$ is the sinc-Kernel $\hat{\sin}(\Omega t)/(\pi t)$ and $\star$ denotes convolution; and Z is a bandlimited noise with Fourier transform

$$\hat{Z}(\omega) = \begin{cases} z(\omega) & |\omega| \leq \Omega \\ 0 & \text{else.} \end{cases}$$

Hence, one observes not the measure μ directly, but instead a noisy version which is blurred by convolution with the kernel K_Ω. In this form, Y is a noisy diffraction-limited image of μ, a superposition of point-sources.

The study of diffraction-limited imaging for such superpositions of point sources goes back a long way. Lord Rayleigh studied it, and formulated a "resolution limit" [28, pp. 33–35]: if a measure μ consists of two point sources of equal strength separated by a distance Δ, a visual inspection of the $Y(t)$ curve will suggest the presence of two point sources provided $\Delta \geq 1.22\pi/\Omega$ and of one point source provided that $\Delta < 1.22\pi/\Omega$. Rayleigh's constant 1.22 is rather arbitrary, and Rayleigh's argument could, with minor modifications, yield instead the constant 1.0. This replacement would lead to the criterion: *pointlike sources separated by at least Δ can be resolved into separate sources, using data diffraction-limited by K_Ω, provided $\pi/\Delta \leq \Omega$.*

This modified Rayleigh limit coincides with the threshold $\pi/\Delta \leq \Omega$ for stable recovery mentioned earlier. In this sense, if we *were* able to recover stably the lattice measure μ from data satisfying $\Omega \ll \pi/\Delta$, we would have exceeded Rayleigh's resolution limit. Therefore, the problem of stably recovering μ from noisy data with parameters in the range $\Omega \ll \pi/\Delta$ below the Rayleigh threshold may be called the problem of *superresolution.*

1.3. Empirical superresolution via sparsity. Despite the mathematical fact that stable recovery of the class $\mathcal{L}(\Delta)$ is impossible when the data satisfy $\Omega \ll \pi/\Delta$, there has been considerable effort to develop superresolving algorithms for specific problems. The idea is essentially that additional a priori information about the support of the measure μ should be exploited in the recovery process.

1. Högbom [11] and others, working in radio astronomy, have developed the method CLEAN, which involves finding a small set of delta-functions $\tilde{\mu} = \sum \alpha_k \delta_{t_k}$ such that $K \star \tilde{\mu}$ (where K is a sinc-like kernel) nearly reproduces the original measurements. As Schwarz [27] says, "... some extra information about the brightness distribution must be used. The CLEAN method is designed for the case that the brightness distribution contains only a few sources at well-separated, small regions, i.e., *the brightness distribution is essentially empty.*"

2. Papoulis and Chamzas [23] have proposed a nonlinear iterative method which assumes implicitly that the underlying measure is sparse, attempts to identify adaptively the regions where coefficients might be nonzero, and recover an object supported only in those regions. They describe an application in medical ultrasound [24]. They point out that the Rayleigh limit is exceeded, in some examples, by their method, and that the actual limit of resolution depends on noise and signal in some yet-to-be determined fashion.

3. Working in seismic prospecting, Levy and Fullagar [15], Santosa and Symes [26], and Walker and Ulrych [33] describe methods which attempt to exploit the fact that the underlying object is a "sparse spike train" to recover wideband data from measurements over a limited frequency range. The Levy–Fullagar and Santosa–Symes work exploits special support properties of l_1-norm penalized reconstruction—namely, that for large values of the multiplier attached to the penalty, the algorithm tends to employ very few nonzero elements in the reconstruction. Walker and Ulrych exploit a method based on low-order autoregressive extrapolation of the Fourier data away from the measured frequency band. The low-order autoregressive model for the Fourier transform may be justified by an assumption that few elements in the spatial domain representation of the object are nonzero. Wang [34] recently introduced a method which constrains the reconstruction so that in any segment of a certain length there are only a few nonzero elements.

4. Working in Fourier transform spectroscopy, Kawata, Minami, and Minami [14], [20] and Mammone [18], also exploit parsimony. Kawata et al. use low-order autoregressive extrapolation away from the measured frequency band, and Mammone uses parametric linear programming to get reconstructions which nearly reproduce the data with minimal numbers of nonzero elements. In later work Minami, Kawata, and Minami refined their technique by using the singular value decomposition to improve the choice of order in autoregressive extrapolation.

5. Working in NMR spectroscopy, Barkhuisen et al. [1] use autoregressive extrapolation, combined with singular value decomposition (LPSVD); Newman [22] proposes the use of l_1-norm penalized reconstruction. Tang and Norris [29] and Mazzeo et al. [19] divide the signal into segments and treat individual segments by sparsity-enhancing methods (e.g., LPSVD) mentioned above.

All these researchers seek to recover wideband objects from narrow-band data; all proceed by in some way imposing sparsity limitations on the recovered object; and all have achieved successes in certain computational experiments. Implicitly or explicitly, these successes amount to a claim that a certain sparsity of the unknown object enables recovery.

At first glance, the computational work just mentioned seems to conflict with the Rayleigh and stability criteria developed above. In fact there is no conflict, since the Rayleigh and the stability criteria do not seek to describe the impact of sparsity.

1.4. Theoretical results. We now develop theory that sheds light on the possibilities, and difficulties, of superresolution via sparsity constraints. We will show that, if the support of μ, though unknown, is known to be sufficiently sparse, then even in the case $\Omega \ll \pi/\Delta$, stable recovery is possible. On the other hand, the quantitative degree of stability might be disappointingly poor if we must recover objects that possess a high degree of complexity.

We do not exhibit a practical method for achieving stable recovery, but instead exhibit inequalities which show that, in the sense of the theory of optimal recovery, the object admits of stable reconstruction. Any stable reconstruction scheme is necessarily highly nonlinear. It would be interesting to know whether stability of the kind we establish below holds for the nonlinear methods [1], [11], [14], [15], [20], [22], [26], [33], [24] mentioned above.

Before developing stability results, we discuss uniqueness. Let S be a discrete set. Following Beurling [3], we define the upper uniform density

$$u.u.d.(S) = \lim_{r\to\infty} r^{-1} \sup_t \#(S \cap [t, t+r))$$

and the lower uniform density

$$l.u.d.(S) = \lim_{r\to\infty} r^{-1} \inf_t \#(S \cap [t, t+r)).$$

Both limits exist.

THEOREM 1.1. (a) *Let* $\mathcal{M}_{<1}(\Delta)$ *denote the class of finite signed lattice measures* $\mu \in \mathcal{L}(\Delta)$ *which have density* $u.u.d.(\mathrm{supp}(\mu)) < 1$. *If* $\Omega \geq 2\pi$, μ *is uniquely characterized among* $\mathcal{M}_{<1}(\Delta)$ *by the transform* $\hat{\mu}(\omega)$, $|\omega| \leq \Omega$.

(b) *Let* S_1 *and* S_2 *be any two disjoint sets with* $l.u.d.(S_i) > 1$, $i = 1, 2$. *Let* μ_1 *be any finite signed measure supported on* S_1. *There exists* μ_2 *with support* S_2 *such that* $\hat{\mu}_1(\omega) = \hat{\mu}_2(\omega)$ *for* $|\omega| \leq \pi$, *yet* $\mu_1 \neq \mu_2$.

(c) *There exist disjoint equispaced sets* S_i, *with* $l.u.d.(S_i) = u.u.d.(S_i) = 1$, $i = 1, 2$, *and measures* μ_i *supported in the* S_i, *such that* $\hat{\mu}_1(\omega) = \hat{\mu}_2(\omega)$ *for* $|\omega| \leq 2\pi - \delta$, $\delta > 0$, *yet* $\mu_1 \neq \mu_2$.

The proof, which relies on Beurling's theory of interpolation and balayage [3], [2], is given in §7.

We conclude that $u.u.d. < 1$ and $\Omega \geq 2\pi$ ensures uniqueness; $l.u.d. > 1$ and $\Omega \leq \pi$ ensures nonuniqueness; and $l.u.d. = u.u.d. = 1$ and $\Omega \in (\pi, 2\pi)$ may, in some cases, lead to nonuniqueness. Hence in searching for stability results, we confine attention to the case $u.u.d. < 1$ and $\Omega \geq 2\pi$.

By rescaling, this pair of conditions is equivalent to the single condition $u.u.d. < \Omega/2\pi$. Now compare this uniqueness condition with the modified Rayleigh criterion $\Delta > \pi/\Omega$. For a typical nonsparsely supported measure $\mu \in \mathcal{L}(\Delta)$, $u.u.d.(\mathrm{supp}(\mu)) = \Delta^{-1}$; hence Rayleigh's criterion is comparable to $u.u.d. < \Omega/\pi$. Our uniqueness criterion therefore demands exactly twice the frequency-domain measurement band (or half the spatial domain density) as the Rayleigh criterion. Our stability estimates will demand at least four times as much as the Rayleigh criterion.

We now return to the scaling convention $u.u.d. < 1$.

DEFINITION 1. Let S be a discrete set of upper uniform density less than one. The *Rayleigh index* of S is

$$R^*(S) = \min\left\{R : R \geq \sup_t \#(S \cap [t, t+R))\right\}.$$

The class of lattice measures μ with $(\alpha_k) \in l_1$ and Rayleigh index $R^*(\mathrm{supp}(\mu)) \leq R$ will be denoted $\mathcal{S}(R, \Delta)$.

The Rayleigh index measures, for sets which have on average, no more than one element per unit cell, the maximum number that can be clustered very closely together. We aim to show that the clustering of many elements together in one cell makes superresolution difficult; we will see that the degree of clustering, as measured by the Rayleigh index, enters directly into our bounds on the stability coefficient.

DEFINITION 2. The *modulus of continuity* for the recovery of measures in $\mathcal{S}(R,\Delta)$ is

$$\Lambda(\epsilon;\mathcal{S}(R,\Delta),\Omega) = \sup\{||\mu_1-\mu_2||_2 : \mu_i \in \mathcal{S}(R,\Delta),\\ ||\hat{\mu}_1-\hat{\mu}_2||_{L_2[-\Omega,\Omega]} \leq \epsilon\}.$$

This modulus of continuity measures the extent to which two lattice measures, both satisfying the sparsity condition $R^*(\text{supp}(\mu_i)) \leq R$, can differ, if the bandlimited data $\{\hat{\mu}_i(\omega), |\omega| \leq \Omega\}$ differ by at most ϵ in L_2-norm. Its relevance comes from the following.

LEMMA 1.2.

$$E^*(\epsilon,\mathcal{S}(R,\Delta),\Omega) \leq \Lambda(2\epsilon,\mathcal{S}(R,\Delta),\Omega). \tag{5}$$

The proof is given in §7. (For arguments relating a modulus of continuity to a minimax error in other contexts, see [21], [25], [30], [31], [6].)

Our main result bounds the modulus of continuity directly.

THEOREM 1.3. *Let* $\Omega > 4\pi$.

$$\Lambda(\epsilon,\mathcal{S}(R,\Delta),\Omega) \leq \Delta^{-2R-1} \cdot \beta(R,\Omega) \cdot \epsilon, \qquad \epsilon > 0. \tag{6}$$

β *is a positive finite constant defined below.*

The raw materials on which this result depends are developed in the body of the paper, §§2–6 below. They are assembled to give a formal proof in §7.

The following lower bound shows that our upper bound is nearly sharp. It is proved in §7 below.

THEOREM 1.4. *Let* $\Delta_0 \in (0,1)$. *If* $\Delta < \Delta_0$ *then*

$$\Lambda(\epsilon,\mathcal{S}(R,\Delta),\Omega) \geq \Delta^{-2R+1} \cdot b(R,\Omega,\Delta_0) \cdot \epsilon, \qquad \epsilon > 0. \tag{7}$$

$b(R,\Omega,\Delta_0)$ *is a positive finite constant defined below.*

1.5. Interpretation of the theory. To interpret these results, we introduce some terminology. We speak of the stability coefficient as the *noise amplification* factor in the relation

Reconstruction Error = Noise Amplification · Noise in Data.

We speak of the ratio Ω/Δ as the *superresolution factor.* Our results indicate that *noise amplification increases polynomially with the superresolution factor.* Hence, to achieve reconstructions with a fixed degree of reconstruction error requires data of increasingly low noise level as the superresolution factor increases. Moreover, the rate at which the noise requirement imposes itself is directly tied to the Rayleigh index, and hence it may be extremely difficult to recover an object with a high degree of clumping or irregularity.

Hence superresolution is possible if the object is known to contain, on average, less than one pointlike event per cell of size $4\pi/\Omega$; but this may require extremely precise data, particularly if we cannot rule out the possibility that a few cells contain many more than one pointlike event. These relations, rather than the Rayleigh criterion, determine the ultimate limits of resolution.

2. Balanced (R, Δ)-sets.

DEFINITION 3. A *balanced* (R, Δ)*-set* is a countable set of points $\{t_k\}$ on the real line which may be obtained from the union of two bilateral sequences $(u_i)_{i=-\infty}^{\infty}$ and $(v_i)_{i=-\infty}^{\infty}$ satisfying these conditions:

$$\begin{array}{lrll}
(8) & \text{(Internal Symmetry)} & u_i = i + \delta_i, & i \in \mathbf{Z}, \\
 & & v_i = i - \delta_i, & i \in \mathbf{Z}, \\
(9) & \text{(Uniform Density)} & |\delta_i| \le R, & i \in \mathbf{Z}, \\
(10) & \text{(Uniform Discreteness)} & |t_k - t_l| \ge \Delta, & k \ne l, \\
(11) & \text{(Unicity)} & u_i \ne u_j, & i \ne j.
\end{array}$$

We note that no assumption is made that $v_i \ne v_j$ for $i \ne j$. Nor is there an assumption that $u_i \ne v_j$ for every i, j. Consequently, the multiplicity

$$m_k = \#\{i : u_i = t_k\} + \#\{i : v_i = t_k\}$$

may be greater than 1, in fact, as large as $2R + 2$. Also, no assumption is made that the points t_k belong to a lattice, although this is not excluded, either.

These four conditions describe a set which is allowed to be locally irregular, yet must be globally regular. A long interval, of length N, say, contains roughly $2 \cdot N$ elements of the set $\{t_k\}$, counting with multiplicities m_k. The internal symmetry condition is also important; it implies that even though the points are not equispaced, they may be arranged in pairs whose centers of gravity are equispaced.

Obviously, balanced (R, Δ)-sets are quite special, and do not occur "naturally"; our interest is in sets which can be "filled out," by the addition of new elements, to become balanced (R, Δ)-sets.

DEFINITION 4. $\{s_k\}$ is a pre-(R, Δ)-set if it is a subsequence of the (u_i) sequence associated to an (R, Δ)-set. A measure μ is an (R, Δ)-measure if its support $\{s_k\}$ is a pre-(R, Δ)-set.

We prove the following in §8.

LEMMA 2.1. *The measures in* $\mathcal{S}(R, \Delta)$ *are all* (R, Δ)*-measures.*

(Elizabeth Gassiat, of Université de Paris-Sud, has shown the author in personal correspondence that the above lemma may be improved, with this conclusion: *the measures in* $\mathcal{S}(R, \Delta)$ *are all* $(R/2, \Delta)$ *measures.* Her work, which is used further below, answers a question raised in the preprint of this article. Her argument is given in [10].)

Our introduction of balanced (R, Δ)-sets is geared to the development of certain interpolation schemes based on entire functions. Entire functions with real zeros have zeros which are roughly equispaced (compare [5], [12], [13]) and possess a certain symmetry (compare the discussion of Lindelöf's theorem in [13]). The conditions (8)–(11) serve to guarantee that there is an entire function of exponential type 2π with $\{t_k\}$ as its set of zeros. We prove the following in §8.

LEMMA 2.2. *Let* $\{t_k\}$ *be a balanced* (R, Δ)*-set. Define*

$$G_n(t) = \frac{\Pi_{-n}^{n}(t - u_i)(t - v_i)}{\Pi_{-n}^{n}(t - i)^2} \sin^2(\pi t).$$

Then (G_n) *is a sequence of entire functions of exponential type* 2π*, uniformly bounded on the real axis. This sequence converges uniformly on compacts to a limit function* G*,* *entire of exponential type 2π. The $\{t_k\}$ are the zeros of the function G; the multiplicity of t_k is m_k.*

The representation as a limit of the sequence (G_n) seems nonstandard; we use it for the purpose of bounding $||G||_\infty$ and related quantities. We record now several important bounds, all of which only involve the parameters R and Δ rather than the detailed properties of the set $\{t_k\}$. The proofs are given in §8.

LEMMA 2.3. *Let $\{t_k\}$ be a balanced (R, Δ)-set. Then*

$$\sup_t |G(t)| \leq A_1(R)$$

where $A_1(R)$ is defined below.

LEMMA 2.4. *Let $\{t_k\}$ be a balanced (R, Δ)-set. Let G have a zero at t_0 of multiplicity m_0. Then*

$$\sup_{|t-t_0|\leq 1} \frac{|G(t)|}{|t-t_0|^{m_0}} \leq A_2(R)$$

where $A_2(R)$ is defined below.

LEMMA 2.5. *Let $\{t_k\}$ be a balanced (R, Δ)-set. Let G have a zero of multiplicity m_k at t_k. Let*

$$g_k = \lim_{t\to t_k} G(t)/(t-t_k)^{m_k}.$$

Then

$$g_k \geq A_3(R)\Delta^{2R+1},$$

where $A_3(R)$ is a strictly positive constant defined below.

These bounds on the entire function G, although stated as lemmas, are in fact the "hard analysis" on which our main result depends; in succeeding sections we reduce the question of superresolution to these inequalities by the use of "soft analysis."

3. Superresolution for (R, Δ)-sets. Let $B_q(\Omega)$, $1 \leq q \leq \infty$, denote the space of entire functions of exponential type Ω belonging to L_q on the real axis [4], [5], [16], [17]. For a sigma-finite signed measure ν, define

$$||\nu||_{p,\Omega} = \sup\left\{\int f d\nu : f \in B_q(\Omega), ||f||_q \leq 1\right\}$$

where $1/p + 1/q = 1$ as usual. In the case where $p = 2$, Parseval's relation implies that

$$||\nu||_{2,\Omega}^2 = \frac{1}{2\pi}||\hat{\nu}||_{L_2[-\Omega,\Omega]}^2.$$

Also, for ν a discrete measure, let

$$||\nu||_p = \left(\sum_{t\in\text{supp}(\nu)} |\nu(\{t\})|^p\right)^{1/p}.$$

Thus, for example, the case where $p = 1$ gives the total variation norm of ν.

We call any inequality of the form

$$||\nu||_p \leq C_p(R,\Delta,\Omega)||\nu||_{p,\Omega}, \tag{12}$$

when valid for all finite signed (R,Δ)-measures ν, a *superresolution inequality.* Besides the case $p = 2$, the main case of interest to us is the case $p = 1$, or

$$\text{Variation}(\nu) \leq C_1(R,\Delta,\Omega)||\nu||_{1,\Omega}.$$

The rationale for calling (12) a superresolution inequality is that, ordinarily, bounds on the norms $||\nu||_p$ would seem to require knowledge of the transform $\hat{\nu}(\omega)$ at all frequencies up to the Nyquist π/Δ; but the norm $||\nu||_{p,\Omega}$ involves only knowledge of frequencies in the smaller band, since, by Parseval, we have

$$\int_{-\infty}^{\infty} f(t)d\nu(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} \hat{f}(\omega)\hat{\nu}(\omega)d\omega$$

for $f \in B_q(\Omega)$, $1 \leq q < \infty$.

4. Duality with interpolation. Consider the following interpolation problem for the sequence $(u_i)_{i=-\infty}^{\infty}$ associated with a balanced (R,Δ)-set. *Given constants* $(c_i)_{i=-\infty}^{\infty}$, $(c_i) \in l_q$, *find a function* $f \in B_q(\Omega)$ *satisfying*

$$f(u_i) = c_i, \qquad i \in \mathbf{Z}. \tag{13}$$

Suppose that for *every* sequence $(c_i) \in l_q$, and *every* (u_i) associated to a balanced (R,Δ)-set, we have a solution to (13) satisfying

$$||f||_{L_q} \leq K_q(R,\Delta,\Omega)||c||_{l_q}, \tag{14}$$

where K_q does not depend on the details of the (u_i), but only on the parameters R and Δ. We claim that then, if p and q are conjugate indices $1/p + 1/q = 1$,

$$K_q(R,\Delta,\Omega) \geq C_p(R,\Delta,\Omega). \tag{15}$$

This expresses a certain duality between the superresolution inequality (12) and the interpolation problem (13).

To prove (15), let ν be any (R,Δ)-measure. Then, by definition, $\text{supp}(\nu) \subset \{u_i\}$ for some sequence $(u_i)_{i=-\infty}^{\infty}$ associated with a balanced (R,Δ)-set. Let $(c_i)_{i=-\infty}^{\infty}$ be aligned with $(\nu\{u_i\})$ in the usual sense that

$$c_i = \lambda\,\text{sgn}(\nu\{u_i\})\,|\nu\{u_i\}|^{p-1},$$

with the scalar λ chosen to make $||c||_{l_q} = 1$. Then

$$\begin{aligned}\sum_{i=-\infty}^{\infty} c_i\nu\{u_i\} &= \left(\sum |\nu\{u_i\}|^p\right)^{1/p}\\ &= ||\nu||_p,\end{aligned}$$

as $\text{supp}(\nu) \subset \{u_i\}$. Now, if f solves the interpolation problem (13), we have

$$\int f d\nu = \sum_{i=-\infty}^{\infty} c_i\nu\{u_i\} = ||\nu||_p,$$

and as $f \in B_q(\Omega)$,

$$\int f d\nu \leq ||f||_{L_q} ||\nu||_{p,\Omega}.$$

Thus, $||\nu||_p \leq ||f||_{L_q} ||\nu||_{p,\Omega}$, and so by (14),

$$||\nu||_p \leq K_q(R, \Delta, \Omega) ||\nu||_{p,\Omega}.$$

Relation (15) follows.

As a result of (15) we now turn our attention to problems of interpolation in $B_q(\Omega)$.

5. Pointwise bounds on interpolation. As indicated in §2, a balanced (R, Δ)-set generates a function G, entire of exponential type 2π. By convention, G has a zero of multiplicity m_k at t_k; defining as before

$$g_k = \lim_{t \to t_k} G(t)/(t - t_k)^{m_k}, \tag{16}$$

the function

$$\xi_k(t) = \frac{G(t)}{g_k \cdot (t - t_k)^{m_k}} \tag{17}$$

satisfies formally

$$\xi_k(t_l) = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases} \tag{18}$$

Actually, ξ_k is a well-defined entire function of type 2π which belongs to L_2 on the real axis. Therefore, formally, the "Lagrange interpolation series"

$$f = \sum_k d_k \xi_k \tag{19}$$

gives a solution of the interpolation problem

$$f(t_k) = d_k, \qquad k \in \mathbf{Z}. \tag{20}$$

However, regularity conditions would be needed on (ξ_k) in order to be sure that such sums converge and define elements of $B_q(\Omega)$. (For discussion of Lagrange interpolation for sets where the corresponding function G has only simple zeros ($m_k \equiv 1$), see [3], [35].)

Regularity is easier to establish if we mollify the ξ_k. We record, without proof, the following essentially obvious technical fact.

LEMMA 5.1. *Let $\eta > 0$. There exists a smooth function $h(x)$ satisfying*

(a) *$\hat{h}(\omega)$ is a smooth function supported in $(-\eta, \eta)$;*

(b) *$h \geq 0$;*

(c) *$h(0) = 1$;*

(d) *$h(x) \leq C(\eta)/(1 + x^2)$ for some positive finite constant $C(\eta)$ and all x.*

The mollified functions $\tilde{\xi}_k = \xi_k(t) h(t - t_k)$ again formally satisfy the Lagrange interpolation conditions

$$\tilde{\xi}_k(t_l) = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases}$$

However, they also have good decay at ∞, and so belong to $B_1(2\pi + \eta)$.

The interpolation problem (20) at the (t_k) is somewhat more general than the problem (13) at the (u_i), so it is convenient to work with a subsequence of the $(\tilde{\xi}_k)$. By (11), each u_i occurs exactly once in the set $\{t_k\}$. Hence, a one-one mapping $k(i)$ exists so that $t_{k(i)} = u_i$. Define the function

$$\psi_i = \tilde{\xi}_{k(i)} = \xi_{k(i)} \cdot h(\cdot - t_{k(i)}).$$

Then $(\psi_i)_{i=-\infty}^{\infty}$ is a sequence of functions in $B_1(2\pi+\eta)$, satisfying the formal Lagrange interpolation conditions

$$\psi_i(u_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Let $\phi(t) = (1 + t^2)^{-1}$ and $\phi_i(t) = \phi(t - i)$. It turns out that the functions ψ_i are effectively not worse behaved than ϕ_i.

THEOREM 5.2.

$$|\psi_i(t)| \leq A\phi_i(t), \qquad t \in (-\infty, \infty)$$

where

$$A(R, \Delta, \eta) = \Delta^{-2R-1}\alpha(R, \eta)$$

and $\alpha(R, \eta)$ is a positive, finite constant specified below.

The theorem follows from earlier estimates on G and g_k, on the mollifier h. To see how, put for short $k = k(i)$. Property (d) of the mollifier h gives $h(\cdot - u_i)/\phi(\cdot - u_i) \leq C(\eta)$. Hence

$$\sup_t \frac{|\psi_i(t)|}{|\phi(t - u_i)|} \leq C(\eta) \sup_t \frac{|G(t)|}{g_k|t - t_k|^{m_k}}.$$

Now

$$\begin{aligned} \sup_t \frac{|G(t)|}{g_k|t - t_k|^{m_k}} &\leq g_k^{-1} \max \left[\sup_t |G(t)|, \sup_{|t-t_k|\leq 1} \frac{|G(t)|}{|t - t_k|^{m_k}} \right] \\ &\leq \Delta^{-2R-1} A_3(R)^{-1} \max[A_1(R), A_2(R)] \\ &= A_5(R)\Delta^{-2R-1}, \end{aligned}$$

say. It follows that

$$|\psi_i| \leq C(\eta) A_5(R) \Delta^{-2R-1} \phi(\cdot - u_i).$$

Note that, by (9), $|u_i - i| \leq R$. Simple algebra gives the following lemma.

LEMMA 5.3. *If $|\delta| \leq R$,*

$$\phi(t \pm \delta) \leq A_4(R)\phi(t), \qquad t \in (-\infty, \infty)$$

where

$$A_4(R) = (R + 1)^2.$$

It follows that $\phi(\cdot - u_i) \leq A_4(R)\phi(\cdot - i)$. Hence

$$|\psi_i| \leq \alpha(R, \eta)\Delta^{-2R-1}\phi_i,$$

say, where

$$\alpha(R, \eta) = C(\eta) \cdot A_5(R) \cdot A_4(R).$$

This completes the proof of Theorem 5.2.

6. Operator norms for interpolation. Let $\mathcal{I}(c) = \sum_{i=-\infty}^{\infty} c_i \psi_i$ denote the Lagrange interpolation operator formally "solving" the interpolation problem (13). We now show that this operator is a bounded linear operator from l_q into $L_q(\mathbf{R})$, and so $\mathcal{I}$ rigorously solves the interpolation problem.

LEMMA 6.1. (1,1)-Boundedness. *Suppose that* $|\psi_i| \le A\phi_i$, $i \in \mathbf{Z}$. *Then*

$$\left\| \sum_i c_i \psi_i \right\|_{L_1(\mathbf{R})} \le A \cdot \pi \cdot \|c\|_{l_1}.$$

Proof.

$$\begin{aligned}\left\| \sum_i c_i \psi_i \right\|_{L_1(\mathbf{R})} &\le \sum_i |c_i| \, \|\psi_i\|_{L_1(\mathbf{R})} \\ &\le \sum_i |c_i| \cdot \sup_i \|\psi_i\|_{L_1(\mathbf{R})} \\ &\le \|c\|_{l_1} A \|\phi\|_{L_1(\mathbf{R})}.\end{aligned}$$

The result follows from $\|\phi\|_{L_1(\mathbf{R})} = \int_{-\infty}^{\infty} (1+t^2)^{-1} dt = \pi$.

LEMMA 6.2. (∞,∞)-Boundedness. *Suppose that* $|\psi_i| \le A\phi_i$, $i \in \mathbf{Z}$. *Then*

$$\left\| \sum_i c_i \psi_i \right\|_{L_\infty(\mathbf{R})} \le A \left(\frac{2\pi}{1 - e^{-2\pi}} \right) \|c\|_{l_\infty}.$$

Proof. Using positivity of ϕ,

$$\begin{aligned}\left\| \sum_i c_i \psi_i \right\|_{L_\infty(\mathbf{R})} &\le \sup_t \sum_i |c_i| \, |\psi_i(t)| \\ &\le A \sup_t \sum_i |c_i| \, \phi_i(t) \\ &\le A \|c\|_{l_\infty} \sup_t \sum_i \phi_i(t).\end{aligned}$$

Using the Poisson summation formula [35, p. 105], the formula $\hat{\phi}(\lambda) = \pi e^{-|\lambda|}$, and the relation $|\hat{\phi}_i(\omega)| = \hat{\phi}(\omega)$,

$$\begin{aligned}\sum_i \phi_i(t) &= \sum_j \hat{\phi}_i(2\pi j) \\ &\le \sum_j |\hat{\phi}_i(2\pi j)| = \sum_j \hat{\phi}(2\pi j) \\ &\le 2 \sum_0^{\infty} \pi e^{-2\pi j} = \frac{2\pi}{1 - e^{-2\pi}}.\end{aligned}$$

These lemmas combine to prove that the operator $\mathcal{I}(c) = \sum_i c_i \psi_i$ gives a bounded mapping from l_q into $L_q(\mathbf{R})$ for each $q \in [1, \infty]$. Indeed, with $\theta = 1/q$ we get, from the Riesz–Thorin Interpolation theorem [36, p. 95],

$$\left\| \sum_i c_i \psi_i \right\|_{L_q(\mathbf{R})} \le (A\pi)^\theta \left(A \frac{2\pi}{1 - e^{-2\pi}} \right)^{1-\theta} \|c\|_{l_q}.$$

Combining these bounds with Theorem 5.2 gives the following theorem.

THEOREM 6.3. *Let $\Omega > 2\pi + \eta$. Then for each $q \in [1, \infty]$ the interpolation problem* (13) *has a solution in $B_q(\Omega)$ satisfying the interpolation inequality* (14) *with*

$$K_q(R, \Delta, \Omega) = \Delta^{-2R-1}\kappa(R, \Omega),$$

where

$$\kappa(R, \Omega) = \frac{2\pi}{1 - e^{-2\pi}} \cdot \alpha(R, \eta).$$

7. Proofs for §1.

7.1. Main result. Suppose that μ_1, μ_2 are lattice measures with support of Rayleigh index $\leq R$. For a set A, let $A/2$ denote the dilation $\{t : (2t) \in A\}$. Define the measure ν by

$$\nu(A) = (\mu_1 - \mu_2)\left(\frac{A}{2}\right).$$

Note that ν is supported in the lattice of span $\Delta' = 2\Delta$, and that

$$\begin{aligned} u.u.d.(\mathrm{supp}(\nu)) &= \tfrac{1}{2}\, u.u.d.(\mathrm{supp}(\mu_1) \cup \mathrm{supp}(\mu_2)) \\ &\leq \tfrac{1}{2}\left(u.u.d.(\mathrm{supp}(\mu_1)) + u.u.d.(\mathrm{supp}(\mu_2))\right) < 1. \end{aligned}$$

Moreover, $R^*(\mathrm{supp}(\nu)) \leq R^*(\mathrm{supp}(\mu_1)) + R^*(\mathrm{supp}(\mu_2))$. Hence, $\nu \in \mathcal{S}(2R, \Delta')$. Applying E. Gassiat's improvement of Lemma 2.1, we conclude that ν is an (R, Δ')-measure.

Applying the superresolution inequality (12), the interpolation inequality (14), and Theorem 6.3, we have for any $\Omega' > 2\pi$,

$$\begin{aligned} ||\nu||_2 &\leq C_2(R, \Delta', \Omega')||\nu||_{2,\Omega'} \\ &\leq K_2(R, \Delta', \Omega')||\nu||_{2,\Omega'} \\ &\leq (\Delta')^{-2R-1}\, \kappa(R, \Omega')||\nu||_{2,\Omega'}. \end{aligned}$$

Now we observe that $||\nu||_2 = ||\mu_1 - \mu_2||_2$. Also $\hat{\nu}(\omega) = (\hat{\mu}_1 - \hat{\mu}_2)(2\omega)$. Hence, making the particular choice $\Omega' = \Omega/2$ we have

$$\begin{aligned} ||\nu||_{2,\Omega'} &= \sqrt{2}||\mu_1 - \mu_2||_{2,\Omega} \\ &= \frac{1}{\sqrt{\pi}}||\hat{\mu}_1 - \hat{\mu}_2||_{L_2[-\Omega,\Omega]}. \end{aligned}$$

Upon combining these relations, (6) follows once we set

$$\beta(R, \Omega) = 2^{-2R-1} \cdot \kappa(R, \Omega/2)/\sqrt{\pi}.$$

Remark. If we used Lemma 2.1, rather than Gassiat's improvement, the argument above would give a proof of a result like Theorem 1.3, only with a factor Δ^{-4R-1} rather than Δ^{-2R-1}. This was the result given in the preprint of this article.

7.2. Proof of Theorem 1.1. Part a. Let $\mu_1 \neq \mu_2$ be two distinct measures, and s be a point of $S = \mathrm{supp}(\mu_1) \cup \mathrm{supp}(\mu_2)$ such that $d \equiv \mu_1\{s\} - \mu_2\{s\} \neq 0$. As the μ_i are lattice measures, S is a *separated* set: any two distinct elements of S differ by at least some strictly positive amount. Moreover, $u.u.d.(S) < 2$. Hence by Theorem 1 of Beurling [3], for all sufficiently small $\delta > 0$, there exists a function in $B_\infty(2\pi - \delta)$ solving the interpolation problem

$$f(t) = \begin{cases} 0, & t \in S\backslash\{s\}, \\ \mathrm{sgn}(d), & t = s. \end{cases}$$

Pick a mollifier $h \in B_1(\delta)$ satisfying (i) $h \geq 0$, (ii) $h(0) = 1$. Set

$$\psi(t) = f(t)\, h(t - s).$$

Then, by construction,

$$\int \psi\, d(\mu_1 - \mu_2) = |\mu_1\{s\} - \mu_2\{s\}| > 0.$$

On the other hand, as $\psi \in B_1(2\pi)$, $\hat{\psi}(\omega)$ exists, and by Parseval,

$$\int \psi\, d(\mu_1 - \mu_2) = \frac{1}{2\pi} \int_{-2\pi}^{2\pi} \hat{\psi}(\omega)(\hat{\mu}_1(\omega) - \hat{\mu}_2(\omega)) d\omega.$$

As the integrand of the right side is continuous and as the integral on the left side is strictly positive, we conclude that

$$\hat{\mu}_1(\omega) \neq \hat{\mu}_2(\omega) \quad \text{for some } \omega \in [-2\pi, 2\pi].$$

Part b. This is an application of Beurling's theory of balayage [2]. Say that S *admits balayage* if the restriction to $[-\Omega, \Omega]$ of any Fourier transform of a finite signed measure ν supported in $\mathbf{R}$ can be represented as the restriction to $[-\Omega, \Omega]$ of the Fourier transform of a finite signed measure ν' supported entirely in S, and if the ratio $||\nu'||_1/||\nu||_1$ is bounded above independently of ν. Beurling shows that a necessary and sufficient condition for S to admit balayage is $l.u.d.(S) > \pi/\Omega$; see Theorem 5 [2, p. 346].

To apply this, simply set $\nu = \mu_1$ and $S = S_2$. By assumption, $l.u.d.(S_2) > 1 > \pi/\Omega$ and so we may perform balayage. Let $\mu_2 = \nu'$ be the result. The measure μ_2 is supported on S_2, which by assumption is disjoint from S_1, and yet $\hat{\mu}_1(\omega) = \hat{\mu}_2(\omega)$ for $|\omega| \leq \pi$.

Part c. Let $\Delta = \frac{1}{2}$ and let $f(\omega)$ be a nonzero, smooth function, periodic of period 4π, vanishing on $[-2\pi + \delta, 2\pi - \delta]$, and satisfying the Hermitian symmetry $f(-\omega) = \overline{f}(\omega)$. Set

$$\alpha_k = \frac{\Delta}{2\pi} \int_{-\pi/\Delta}^{\pi/\Delta} e^{i\omega k\Delta} f(\omega) d\omega$$

and define

$$\mu_1 = \sum_{k \text{ odd}} \alpha_k \delta_{k\Delta}$$

and

$$\mu_2 = -\sum_{k \text{ even}} \alpha_k \delta_{k\Delta}.$$

This defines a pair of finite signed measures; the first supported at the half-integers; the second at the integers. These measures have Fourier transforms and

$$(\hat{\mu}_1(\omega) - \hat{\mu}_2(\omega)) = f(\omega) = 0, \qquad |\omega| \leq 2\pi - \delta$$

and so the two measures, although supported disjointly, have Fourier transforms which agree throughout the band $|\omega| \leq 2\pi - \delta$. In addition, note that $l.u.d.(\text{supp}(\mu_i)) = u.u.d.(\text{supp}(\mu_i)) = 1$.

7.3. Proof of Lemma 1.2. In this proof, let $N(y, \epsilon) = N(y, \epsilon; R, \Delta, \Omega) = \{\nu : ||\hat{\nu} - y||_{L_2[-\Omega,\Omega]} \leq \epsilon, \nu \in \mathcal{S}(R, \Delta)\}$ denote the set of all lattice measures with Rayleigh index R which are within an ϵ-distance of y. This set contains μ by assumption. Hence it is nonempty. Consider any "feasible reconstruction" rule $\tilde{\mu}(y)$, i.e., any rule with

$$\tilde{\mu}(y) \in N(y, \epsilon).$$

Such a rule selects, from among all measures which could have generated the data, one which satisfies the assumed sparsity. (It is possible to select from the set $N(y, \epsilon)$ so that $\tilde{\mu}(y)$ is a measurable function of y in the topology generated from $L_2[-\Omega, \Omega]$-norm balls.)

The triangle inequality implies that any such $\tilde{\mu}$ formally has

$$||\hat{\mu} - \hat{\tilde{\mu}}(y)||_{L_2[-\Omega,\Omega]} \leq ||\hat{\mu} - y||_{L_2[-\Omega,\Omega]} + ||y - \hat{\tilde{\mu}}(y)||_{L_2[-\Omega,\Omega]} \leq 2 \cdot \epsilon.$$

Hence, by definition of Λ,

$$||\mu - \tilde{\mu}(y)||_2 \leq \Lambda(2 \cdot \epsilon).$$

The lemma follows.

Remark. The idea of "feasible reconstruction," i.e., of selecting any reconstruction matching known constraints on signal and noise, while of practical value in other contexts [32] is not necessarily practical here, because "projection" onto the set of sparse objects is not a contraction, and so certifiably convergent iterative algorithms are lacking, notwithstanding [23].

7.4. Proof of Theorem 1.4. We make a simple computation. Let $\nu_{r,h} = \sum_{k=0}^{r} (-1)^k C(r,k) \delta_{kh}$, where $C(r,k)$ is the standard combinatorial factor

$$C(r,k) = \frac{r!}{k!(r-k)!}.$$

Then $||\nu_{r,h}||_2 = (\sum_{k=0}^{r} C(r,k)^2)^{1/2}$ independent of h. On the other hand, we recognize that $\nu_{r,h} \star f = D_h^r f$, where D_h^r is the rth-order finite difference operator of span h. From the fact that

$$h^{-r} D_h^r f \rightarrow_{L_2(\mathbf{R})} f^{(r)} \quad \text{as } h \rightarrow 0,$$

for every smooth f of compact support, we get

$$h^{-2r}\int_{-\Omega}^{\Omega}|\hat{\nu}_{r,h}(\omega)|^2 d\omega \to \int_{-\Omega}^{\Omega}\omega^{2r}d\omega = \frac{\Omega^{2r+1}}{r+1/2}.$$

Hence, as $h \to 0$,

$$||\nu_{r,h}||_{2,\Omega} \sim h^r \frac{\Omega^{r+1/2}}{\sqrt{\pi\,(2r+1)}}. \tag{21}$$

Let now

$$\mu_1 = \eta \sum_{\substack{k \text{ odd} \\ 0<k<2R}} C(2R-1,k)\delta_{k\Delta}$$

and

$$\mu_2 = \eta \sum_{\substack{k \text{ even} \\ 0\le k<2R}} C(2R-1,k)\delta_{k\Delta}.$$

Then the μ_i belong to $\mathcal{S}(R,\Delta)$ and $\mu_2-\mu_1 = \eta\,\nu_{2R-1,\Delta}$. Hence if we choose η so that $||\mu_1-\mu_2||_{2,\Omega} = \epsilon$, then

$$\Lambda(\epsilon) \ge ||\mu_1-\mu_2||_2. \tag{22}$$

Now, evidently,

$$||\mu_1-\mu_2||_2 = \frac{||\nu_{2R-1,\Delta}||_2}{||\nu_{2R-1,\Delta}||_{2,\Omega}}\,||\mu_1-\mu_2||_{2,\Omega}.$$

Define

$$b(R,\Omega,\Delta_0) = \inf_{\Delta<\Delta_0}\frac{||\nu_{2R-1,\Delta}||_2}{||\nu_{2R-1,\Delta}||_{2,\Omega}}\,\Delta^{2R-1}.$$

By (21), $b(R,\Omega,\Delta_0) > 0$ and, as $\Delta_0 \to 0$,

$$b(R,\Omega,\Delta_0) \to \frac{\Omega^{2R-1/2}}{\sqrt{\pi\,(4R-1)}}.$$

Hence if $\Delta < \Delta_0$,

$$||\mu_1-\mu_2||_2 \ge \Delta^{-2R+1}\,b(R,\Omega,\Delta_0)\,||\mu_1-\mu_2||_{2,\Omega}$$

and so, by (22) the theorem follows.

8. Proofs for §2.

8.1. Proof of Lemma 2.1. Let $S = \mathrm{supp}(\mu) = \{s_k\}$. Partition the line into intervals $I_m = [mR+\frac{1}{2},(m+1)R+\frac{1}{2})$, $m = \cdots,-1,0,1,\cdots$. In I_m there are by assumption $r_m \le R$ elements of S. We pair up these elements of $S\cap I_m$ with integer elements of I_m from right to left (say). For an integer i paired with an element $s \in S$ in this way, set $u_i = s$. Some integer elements of I_m may remain unpaired after this step; we simply set u_i to be the closest element of the lattice $\{k\Delta\}$ to i which has not been previously assigned to a u_i and which does belong to I_m. In this way we get a bilateral sequence $(u_i)_{i=-\infty}^{\infty}$.

Defining $v_i = 2i - u_i$, gives (8). Since for $i \in I_m$, $u_i \in I_m$, we have $|u_i - i| \leq R$, hence (9). Now any two distinct points in $\{u_i\} \cup \{v_i\}$ are separated by at least the lattice span Δ, hence (10). Also, by our pairing convention, we never assign a u_i to a value which has been previously assigned. Thus all four properties (8)–(11) are verified.

8.2. Proof of Lemma 2.2. If $f \in B_\infty(\Omega)$ has a zero at t_0, then $g(t) = f(t)(t - s_0)/(t - t_0)$ is again in $B_\infty(\Omega)$. Compare, for example, [35, pp. 126–129]. Now, $\sin^2(\pi t) \in B_\infty(2\pi)$; hence, we have $G_n \in B_\infty(2\pi)$ for all $n \geq 1$. Moreover, as $\sin^2(\pi z)$ has only real zeros, $G_n(z)$ has only real zeros.

We show that $\{G_n\}$ forms a Normal Family. Indeed, by the Theorem on page 47 of Koosis [13], an entire function of exponential type with only real zeros satisfies, for $\Im(z) > 0$

$$\log|f(z)| = A_+\Im(z) + \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\Im(z)}{|z-t|^2} \log|f(t)|dt, \tag{23}$$

where

$$A_+ = \limsup_{y\to\infty} \frac{\log|f(iy)|}{y}.$$

A parallel relation, employing

$$A_- = \limsup_{y\to-\infty} \frac{\log|f(iy)|}{|y|},$$

holds in $\Im(z) < 0$. We note that for $f = G_n$, $A_+ = A_- = 2\pi$. Indeed, $G_n(z) = Q_n(z)\sin^2(\pi z)$, where Q_n is a rational function of degree (n, n) which satisfies

$$Q_n(z) \to 1 \quad \text{as } |z| \to \infty$$

because the numerator and denominator polynomials both have coefficient 1 on the highest order term z^n. Moreover,

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{|\Im(z)|}{|z-t|^2} dt = 1$$

for each nonreal z, and so the second term on the right side of (23) is never larger than $\log||f||_{L_\infty(\mathbf{R})}$. Combining these facts we have

$$\log|G_n(z)| \leq 2\pi|z| + \log||G_n||_{L_\infty(\mathbf{R})}. \tag{24}$$

The proof of Lemma 2.3 below shows that $||G_n||_{L_\infty(\mathbf{R})} \leq A_1(R)$ for all n. Hence (24) shows that $\{G_n(z)\}$ is uniformly bounded in each bounded region of the complex plane. Hence, $\{G_n\}$ is a normal family of entire functions. By Montel's theorem, an entire function G exists as a cluster point of $\{G_n\}$; it must obey, because of (24),

$$\log|G(z)| \leq 2\pi|z| + \log A_1(R)$$

and so G must be entire of exponential type 2π. However, a calculation based on the definition of G_n shows that for each fixed t, $G_n(t)$ converges to a definite limit as $n \to \infty$ so there can only be one cluster point of G_n, and hence the sequence actually has G as a proper limit. The limit function G must satisfy the same relation (23) as the G_n, so G has only real zeros. The following lemma allows us to see that these zeros are at the points t_k and have the same multiplicity as the m_k.

LEMMA 8.1.

$$(n!)^4 \cdot \frac{\sin^2(\pi t)}{\prod_{i=-n}^{n}(t-i)^2} \to \pi^2 \quad \textit{as } n \to \infty$$

uniformly on compact sets of t in $(-\infty, \infty)$.

8.3. Proof of Lemma 2.3. We now arrive at the key estimates. It is enough to show that for large n, $||G_n||_\infty \leq A_1(R)$, as G is the limit of the G_n. We may assume, without loss of generality, that $t \in [-\frac{1}{2}, \frac{1}{2}]$ and $n \gg R$.

$$|G_n(t)| = \frac{\prod_{i=-n}^{n} |i+\delta_i - t|\,|i-\delta_i-t|}{\prod_{i=-n}^{n} |i-t|^2} \sin^2(\pi t).$$

Now, for $|i| > R$, the inequality of the Arithmetic-Geometric mean gives

$$|i+\delta_i-t|\,|i-\delta_i-t| \leq |i-t|^2.$$

On the other hand, for $|i| \leq R$,

$$|i+\delta_i-t|\,|i-\delta_i-t| \leq (2R+1)^2.$$

We also note that

$$\sup_{t\in[-1/2,1/2]} \prod_{\substack{|i|\leq R \\ i\neq 0}} |t-i|^{-2} = \prod_{i=1}^{R} |i^2 - 1/4|^{-2}.$$

Combining these estimates,

$$\begin{aligned}
|G_n(t)| &\leq \frac{\prod_{i=-R}^{R} |i+\delta_i-t|\,|i-\delta_i-t|}{\prod_{i=-R}^{R} |i-t|^2} \sin^2(\pi t) \\
&\leq ((2R+1)^2)^{2R+1} \sup_{t\in[-1/2,1/2]} \prod_{\substack{|i|\leq R \\ i\neq 0}} |t-i|^{-2} \sup_t \frac{\sin^2(\pi t)}{t^2} \\
&= ((2R+1)^2)^{2R+1} \prod_{i=1}^{R} |i^2-1/4|^{-2}\, \pi^2 \\
&= A_1(R).
\end{aligned}$$

8.4. Proof of Lemma 2.4. Without loss of generality, let $k = 0$, let $t_0 \in [-\frac{1}{2}, \frac{1}{2})$ and let $i_0 = i_0(t)$ denote an integer closest to t. Let $a_i = 1$ if $u_i \neq t_0$, and zero otherwise; and let $b_i = 1$ if $v_i \neq t_0$, and zero otherwise. Pick $n \gg 2R$.

$$\frac{|G_n(t)|}{|t - t_0|^{m_0}} = \frac{\prod_{i=-n}^{n} |i + \delta_i - t|^{a_i} \, |i - \delta_i - t|^{b_i}}{\prod_{i=-n}^{n} |i - t|^2} \sin^2(\pi t).$$

Note that if $|i| > 2R$ then necessarily $a_i = b_i = 1$. We again have

$$|i + \delta_i - t| \, |i - \delta_i - t| \leq |i - t|^2 \quad \text{for } |i| > 2R,$$

and so

$$\frac{|G_n(t)|}{|t - t_0|^{m_0}} \leq \frac{\prod_{|i| \leq 2R} |i + \delta_i - t|^{a_i} \, |i - \delta_i - t|^{b_i}}{\prod_{\substack{|i| \leq 2R \\ i \neq i_0}} |t - i|^{-2}} \cdot \frac{\sin^2(\pi t)}{|t - i_0|^2}.$$

Now

$$\sup_{t \in [-1/2, 1/2]} \prod_{\substack{|i| \leq 2R \\ i \neq i_0}} |t - i|^{-2} \leq \prod_{i=1}^{2R-1} |i^2 - 1/4|^{-2}$$

and also,

$$|i + \delta_i - t|^{a_i} \, |i - \delta_i - t|^{b_i} \leq (3R + 1)^2 \quad \text{for } |i| \leq 2R.$$

Hence

$$\begin{aligned} \frac{|G_n(t)|}{|t - t_0|^{m_0}} &\leq ((3R + 1)^2)^{4R+1} \prod_{i=1}^{2R-1} |i^2 - 1/4|^{-2} \, \pi^2 \\ &= A_2(R). \end{aligned}$$

8.5. Proof of Lemma 2.5. Without loss of generality, let $k = 0$, and suppose that $|t_0| \leq \frac{1}{2}$. Set

$$G_n(t) = \prod_{-n}^{n} (t - u_i)(t - v_i) \, H_n(t).$$

By Lemma 8.1, $(n!)^4 H_n(t_0) \to \pi^2$ as $n \to \infty$. Let n be so large that $(n!)^4 H_n(t_0) \geq \pi^2/2$. Then set $x_i = |u_i - t_0|$ and $y_i = |v_i - t_0|$. Define a_i and b_i as in the previous section. The convention $0^0 = 1$ will simplify notation below.

Now

$$|g_{0,n}| = |H_n(t_0)| \cdot \prod_{-n}^{n} x_i^{a_i} y_i^{b_i}$$

and so, by our choice of n,

$$|g_{0,n}| \geq \frac{\pi^2}{2} \cdot \frac{\prod_{-n}^{n} x_i^{a_i} y_i^{b_i}}{\prod_{i=1}^{n} i^4}. \tag{25}$$

We need two estimates. First, for a constant $E_0(R)$,

$$\prod_{i=-R}^{R} x_i^{a_i} y_i^{b_i} \geq E_0(R)\, \Delta^{2R+1}. \tag{26}$$

To see this, note that at most $2R+2$ of the x_i and y_i can be less than $\frac{1}{2}$. Indeed, if $i \neq 0$ at most one member of each pair (u_i, v_i) can be in the interval $[-1,1]$; and if $|i| > R$, neither member of a pair can be in the interval. On the other hand, by hypothesis, at least one of the (x_i) or (y_i) must be zero, as t_0 belongs to $\{u_i\} \cup \{v_i\}$. Therefore, the product on the left side of (26) contains at most $2R+1$ terms of size less than $\frac{1}{2}$. Such terms, by (10), must be at least Δ in size. Hence,

$$\prod_{i=-R}^{R} x_i^{a_i} y_i^{b_i} \geq \Delta^{2R+1} \left(\frac{1}{2}\right)^{2R+1}. \tag{27}$$

Thus (26) holds, with $E_0(R) = 2^{-2R-1}$. (Much larger values for E_0 may established with more effort.)

Our second estimate concerns the terms omitted by (26). We wish to show that

$$\prod_{R+1}^{n} \frac{x_i^{a_i} y_i^{b_i} x_{-i}^{a_{-i}} y_{-i}^{b_{-i}}}{i^4} \geq E_1(R) > 0. \tag{28}$$

The key point is that for $|i| > R$, $a_i = b_i = 1$. Now

$$\begin{aligned}
\frac{x_i y_i x_{-i} y_{-i}}{i^4} &= \frac{|i+\delta_i - t_0|\,|i-\delta_i-t_0|\,|-i+\delta_{-i}-t_0|\,|-i-\delta_{-i}-t_0|}{i^4} \\
&= \frac{|(i-t_0)^2 - \delta_i^2|\,|(-i-t_0)^2-\delta_{-i}^2|}{i^4} \\
&= \left|\left(1-\frac{t_0}{i}\right)^2 - \left(\frac{\delta_i}{i}\right)^2\right| \left|\left(1+\frac{t_0}{i}\right)^2 - \left(\frac{\delta_{-i}}{i}\right)^2\right| \\
&\geq \left|\left(1-\frac{1}{2i}\right)^2 - \left(\frac{R}{i}\right)^2\right| \left|\left(1+\frac{1}{2i}\right)^2 - \left(\frac{R}{i}\right)^2\right| \\
&= e_i,
\end{aligned}$$

say (the inequality step is justified by additional calculations, which we omit). Now $e_i \geq [((2R+1)/(2R+2))^2 - (R/(R+1))^2]^2 > 0$. A little algebra shows that for $i > 2R$, $e_i > (1-(2R/i)^2)$. Hence, defining

$$E_1(R) = \prod_{i=R+1}^{\infty} e_i,$$

we have

$$E_1(R) \geq \prod_{i=R+1}^{2R} e_i \cdot \prod_{i=2R+1}^{\infty} \left(1 - \left(\frac{2R}{i}\right)^2\right) > 0,$$

and (28) holds. Again, this is only a very crude estimate; much larger values of $E_1(R)$ can be established.

Combining the inequalities (25), (26), and (28), we have

$$|g_{0,n}| \geq \Delta^{2R+1} \cdot \frac{\pi^2}{2} \cdot E_0(R) \cdot E_1(R) \cdot (R!)^{-4},$$

which, by a limiting process, yields the sought-for inequality

$$|g_0| \geq \Delta^{2R+1} \cdot A_3(R)$$

with the (very crude) value

$$A_3(R) = \frac{\pi^2}{2} \cdot E_0(R) \cdot E_1(R) \cdot (R!)^{-4}.$$

9. Discussion.

A few final remarks may clarify issues raised by the above.

9.1. The optimal exponent.

Theorems 1.3 and 1.4 together suggest that for a certain exponent $e(R)$ we have, for $\Omega > \Omega_0$, and all small Δ

$$E^*(\epsilon, R, \Delta, \Omega) \asymp \Delta^{-e(R)} \cdot \text{Const}(R, \Omega) \cdot \epsilon.$$

If such a relation holds, we must have, by Theorems 1.3 and 1.4,

$$2R - 1 \leq e(R) \leq 2R + 1.$$

What is the value of $e(R)$?

9.2. Relation to other work.

A number of papers treat the problem of recovering a signal from data which are missing information about a whole band of frequencies, by exploiting support limitations: compare results in [26], [7], [8], [25].

The closest result in those papers to the present one may be stated as follows [8, §6]. Suppose that (x_t) is a discrete-time signal, and that we have noisy information $\hat{y}(\omega) = \hat{x}(\omega) + \hat{z}(\omega)$ about the Fourier transform of x; only now the *high* frequencies $|\omega| \in [\frac{\pi}{m}, \pi]$ are observed, m an integer greater than 1. Then, despite the missing information about the *low* frequency band $[-\frac{\pi}{m}, \frac{\pi}{m}]$, we can stably recover (x_t)—provided that in every interval of length m, a fraction less than $1/\pi$ of the samples are nonzero.

The present paper covers the complementary case, where information for the *low* frequency interval $[-\frac{\pi}{m}, \frac{\pi}{m}]$ would be observed, and information for the *high* frequency band $|\omega| \in [\frac{\pi}{m}, \pi]$ would be missing. For stable recovery of (x_t), Theorem 1.4 would require that for each interval of length $4mR$, fewer than R samples are nonzero.

This considerably more restrictive sparsity condition expresses the fact that the problem of missing *high* frequencies is much more ill-posed than the problem of missing *low* frequencies.

In another direction, one might compare the interpolation results developed here with other work on interpolation of entire functions. Both in Boas [4] and in Duffin and Schaeffer [9] there is discussion of sets which are uniformly discrete and of uniform density 1, so that conditions (9) and (10) hold. However, the internal symmetry condition (8) seems different from earlier work and figures significantly in the key Lemmas 2.3, 2.4, and 2.5. It would be interesting to know whether this internal symmetry condition is in some sense necessary.

9.3. Removing the lattice constraint. We consider it plausible that a more general family of superresolution inequalities holds, in which the lattice constraint $\mu \in \mathcal{L}(\Delta)$ is removed. Such an inequality would be of the form

$$||\nu||_{p,\lambda\Omega} \leq A_p(\lambda, \Omega, R)||\nu||_{p,\Omega},$$

where $\lambda > 1$ is the superresolution factor, and the inequality is suppposed valid for all finite signed measures ν with support of Rayleigh index R. Such an inequality would express superresolution by the fact that the norm on a larger frequency band would be controlled by the norm on a smaller frequency band. For comparison, if $\mu \in \mathcal{L}(\Delta)$, and $\lambda\Omega = \pi/\Delta$, then

$$||\nu||_{2,\lambda\Omega}^2 = \frac{1}{2\pi}||\nu_2||_2^2,$$

so we get the inequality proved in this paper as a special case. Evidently λ plays the role of Δ^{-1}; it seems plausible that the stability coefficient A_p would grow with λ like $\lambda^{e(R)}$, where $e(R)$ is the optimal exponent in the lattice case.

Acknowledgments. The author is on leave from the University of California, Berkeley. The author thanks especially E. Gassiat, as well as D. Dacunha-Castelle, J. C. Hoch, R. L. Parker, P. B. Stark, A. Stern, and Kong Wang for interesting discussions and correspondence.

REFERENCES

[1] H. Barkhuisen, R. de Beer, W. M. M. J. Bovée, and D. van Ormondt, *Retrieval of frequencies, amplitudes, damping factors, and phases from time-domain signals using a linear least-squares procedure*, J. Magnetic Resonance, 61 (1985), pp. 465–481.

[2] A. Beurling, IV, *Balayage of Fourier–Stieltjes transforms*, Mittag-Leffler Lectures on Harmonic Analysis 1977–78, in Collected Works of Arne Beurling: Volume 2, Harmonic Analysis, pp. 343–350. Birkhauser, Boston, MA, 1989.

[3] ———, V. *Interpolation for an interval on R^1. 1. A density theorem*, Mittag-Leffler Lectures on Harmonic Analysis 1977-78. in Collected Works of Arne Beurling: Volume 2 Harmonic Analysis, pp. 351–359, Birkhauser, Boston, MA, 1989.

[4] R. P. Boas, Jr., *Entire functions bounded on a line*, Duke Math. J., 6 (1940), pp. 148–169.

[5] ———, *Entire Functions*, Academic Press, New York, 1952.

[6] D. L. Donoho, *Statistical estimation and optimal recovery*, Ann. Statist., 1992, to appear.

[7] D. L. Donoho and P. B. Stark, *Uncertainty principles and signal recovery*, SIAM J. Appl. Math., 49 (1989), pp. 906–931.

[8] D. L. Donoho and B. F. Logan, *Signal recovery and the large sieve*, SIAM J. Appl. Math., 52 (1992), pp. 577-591.

[9] R. J. Duffin and A. C. Schaeffer, *Power series with bounded coefficients*, Amer. J. Math., 67 (1945), pp. 141–154.

[10] E. Gassiat and D. Donoho, *Super-resolution via positivity constraints*, 1991, manuscript.

[11] J. A. Högbom, *Aperture synthesis with a non-regular distribution of interferometer baselines*, Astronom. Astrophys. Suppl., 15 (1974), pp. 417–426.

[12] A. S. B. Holland, *Introduction to the Theory of Entire Functions*, Academic Press, New York, 1973.

[13] P. Koosis, *The Logarithmic Integral,* Cambridge University Press, Cambridge, 1988.

[14] S. Kawata, K. Minami, and S. Minami, *Superresolution of Fourier transform spectroscopy data by the maximum entropy method*, Appl. Optics, 22 (1983), pp. 3593–3606.

[15] S. Levy and P. K. Fullagar, *Reconstruction of a sparse spike train from a portion of its spectrum, and application to high-resolution deconvolution*, Geophysics, 46 (1981), pp. 1235–1243.

[16] B. F. Logan, *Properties of High-Pass Signals*, Ph.D. thesis, Columbia University, New York, 1965.

[17] ———, *The norm of certain convolution transforms on L_p spaces of entire functions of exponential type*, SIAM J. Math. Anal., 16 (1985), pp. 167–179.

[18] R. J. Mammone, *Spectral extrapolation of constrained signals*, J. Opt. Soc. Amer., 73 (1983), p. 1476.

[19] A. R. Mazzeo, M. A. Delsuc, A. Kumar, and G. C. Levy, *Generalized maximum entropy deconvolution of spectral segments*, J. Magnetic Resonance, 81 (1989), pp. 512–519.

[20] K. Minami, S. Kawata, and S. Minami, *Superresolution of Fourier Transform spectra by autoregressive model fitting with singular value decomposition*, Appl. Optics, 24 (1985), pp. 162–167.

[21] C. A. Micchelli and T. J. Rivlin, *A survey of optimal recovery*, in Optimal Estimation in Approximation Theory, C. A. Micchelli and T. J. Rivlin., eds., Plenum Press, New York, 1977, pp. 1–54.

[22] R. H. Newman, *Maximization of entropy and minimization of area as criteria for* NMR *signal processing*, J. Magnetic Resonance, 79 (1988), pp. 448–460.

[23] A. Papoulis and C. Chamzas, *Detection of hidden periodicities by adaptive extrapolation*, IEEE Trans. Acoust. Speech Signal Process., ASSP-27 (1979), pp. 492–500.

[24] ———, *Improvement of range resolution by spectral extrapolation*, Ultrasonic Imaging, 1 (1979), pp. 121–135.

[25] E. R. Pike, J. G. McWhirter, M. Bertero, and C. De Mol, *Generalized information theory for inverse problems in signal processing*, IEE Proceedings, 131 (1984), p. 660.

[26] F. Santosa and W. W. Symes, *Linear inversion of band-limited reflection seismograms*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1307–1330.

[27] U. J. Schwarz, *Mathematical-statistical description of the iterative beam-removing technique* (*Method* CLEAN), Astronom. and Astrophys., 65 (1978), pp. 345–356.

[28] E. G. Steward, *Fourier Optics: An Introduction*, John Wiley, New York, 1983.

[29] J. Tang and J. R. Norris, LP-ZOOM, *a linear-prediction method for local spectral analysis of* NMR *signals*, J. Magnetic Resonance, 79 (1988), pp. 190–196.

[30] J. H. Traub, G. Wasilkowski, and H. Woźniakowski, *Information, Uncertainty, Complexity*, Academic Press, New York, 1981.

[31] ———, *Information-Based Complexity*, Academic Press, New York, 1988.

[32] H. J. Trussel and M. R. Civanlar, *The feasible solution in signal restoration*, IEEE Trans. Acoust. Speech Signal Process., 32 (1984), pp. 201–212.

[33] C. Walker and T. J. Ulrych, *Autoregressive recovery of the acoustic impedance*, Geophysics, 48 (1983), pp. 1338–1350.

[34] K. Wang, Ph.D. thesis, Mining and Materials Science, University of California, Berkeley, CA.

[35] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[36] A. Zygmund, *Trigonometric Series*, Vol. II., Second ed., Cambridge University Press, Cambridge, 1977.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1332–1341, September 1992

HALFWAY POINTS*

W. A. BEYER† AND BLAIR SWARTZ‡

This paper is dedicated to R. S. Varga.

Abstract. As a half space moving perpendicularly to its border crosses a given density distribution, it may be stopped when it contains exactly half of the total mass of the distribution. Any point in common with the boundaries of all such stopped half spaces is called a "halfway point" for the distribution. One may regard it as a multidimensional extension of the "median" of one-dimensional distributions. A context in which such points arise is discussed, and some characteristics of distributions that have halfway points are considered.

Key words. moment center, center of volume, median

AMS(MOS) subject classification. 28A75

Introduction. Suppose two-dimensional cakes come in various shapes and mass distributions. Suppose we buy only those cakes such that there is a unique point on the cake (that we shall call the cake's *halfway point*) such that a sufficiently long knife cutting through that point divides the cake into two pieces of equal mass, independent of the orientation of the cake. What kinds of shapes and mass distributions of cakes are we limited to buying? What about n-dimensional cakes?

This problem was suggested by work concerning interface reconstruction in computational hydrodynamics and edge reconstruction in image enhancement [8]. We outline these contexts—a more complete statement is given in §1 below and in [8]. Suppose we are given a characteristic function $c(x)$ of an open subset Ω of the plane (say), with the boundary of Ω being sufficiently regular. Blur $c(x)$ using a convolution with a scaled measure distribution function ρ (of total mass one) as in (1.4) below. Call the blurred function $\bar{c}(x)$—it depends on the scale h, on the function ρ, on Ω, and on the location z of the origin for ρ. If the scale h is sufficiently small and z is inside ρ's support (here assumed compact), then $\bar{c}(x)$ will be 1 inside Ω and zero outside, except that there will be a narrow strip containing Ω's boundary on which $\bar{c}$ assumes values strictly between zero and 1. This is the reason that $\bar{c}$ is called a blurring of c.

The paper [8] analyzes three algorithms for approximately reconstructing the boundary of Ω from discrete data about its blurred image. When the scale is small (relative to the boundary's curvature), and when the data are taken at a comparable scale, then algorithms designed to recover linear boundaries exactly will have "second-order" accuracy for curved boundaries (assuming the algorithms also have a certain stability property [8]). Such accuracy is useful for efficient interface reconstruction in numerical hydrodynamics when it involves the "volume fractions" of computational cells in two-dimensional calculations. (Second-order accuracy is even more important for three space dimensions.) And the "sub-pixel-size" accuracy associated with the

*Received by the editors August 5, 1991; accepted for publication (in revised form) February 19, 1992. This research was supported by the U. S. Department of Energy under contract W-7405-ENG.36.

†T Division, MS B284, University of California, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

second-order accurate reconstruction of curved "edges" would also be useful in (two-dimensional) image processing.

Thus, the creation of second-order accurate algorithms depends on understanding the details of the behavior of the blurring of linear boundaries. It turns out that if the point z in the blurring (1.4) is the halfway point of the function ρ, then the blurring has the property that, for the case of an unblurred linear boundary with an arbitrary orientation, x lies exactly on the unblurred boundary if and only if $\bar{c}(x)$ has the value $\frac{1}{2}$. This is the property of the halfway point that gives it significance for us. But we may also regard it as locating a "central" point of the distribution—in particular, as a multidimensional extension of the "median" of one-dimensional distributions. Also, suppose a solid convex body, of uniform density, has a halfway point and displaces twice its weight when fully immersed in a fluid. Then, if allowed to float in the fluid, the sea-level plane will contain the halfway point. Indeed, the body's "surface of buoyancy" as it is called in naval architecture (i.e., the envelope of all *candidates* for the sea-level plane) reduces to the halfway point.

In §2 we explore some characteristics of mass distributions that have halfway points—cf. Propositions A and B there.

1. The blurring of boundaries; cumulative distributions. Motivation for this note arose in the following context. The plane is decomposed into two regions, Ω_0 and Ω_1, separated by their common boundary β. We assume that β has bounded curvature, does not cross itself, and consists of one component only. Associate with this setup the color function (essentially, characteristic function) c, defined by

$$c(x) := \begin{cases} 0, & x \text{ in } \Omega_0, \\ 1, & x \text{ in } \Omega_1, \\ \frac{1}{2}, & x \text{ in } \beta. \end{cases} \tag{1.1}$$

In particular, and somewhat artificially, β is the inverse image of $\frac{1}{2}$ under the map c. The following example of a scaled blurring of c is used in computational fluid dynamics (see, e.g., references in [8]). Let $S(x,h)$ be the square of area h^2, centered on the point x, and with sides parallel to the coordinate axes. Associate with c and S the "average color" function

$$\bar{c}(x) := \iint_{S(x,h)} c(\xi)\, dA(\xi)/h^2, \qquad x \text{ in } \mathbb{R}^2. \tag{1.2}$$

For h sufficiently small (relative to β's curvature), the set $\bar{c}^{-1}(\frac{1}{2})$ is one-dimensional and approximates β. Given a uniform mesh of points $X_i := (i_1h, i_2h)$ of grid-size h—the same h as in S—we can and do consider the use of meshpoint values $\bar{c}(X_i)$ (sometimes called "volume fractions") in approximately reconstructing β. For example, [8] discusses local approximate reconstruction using the locations of a few adjacent "gray" values of the average color $(\bar{c} \in (0,1))$. Reconstruction that is "first-order accurate" in the blurring's "scale" h (i.e., has $O(h)$ accuracy as $h \to 0$) is relatively easy, since the Hausdorff distance from β to the set of gray meshpoints does not exceed $h/\sqrt{2}$. But for $O(h^2)$ accuracy, the localized algorithm must essentially reproduce linear boundaries. To do this it is important to know how the local averaging smears the boundaries of half spaces.

Observe that the level curves of the average color function $\bar{c}_L$, of the color function c_L associated with the $h \times h$ square $S(x,h)$ and the two half planes bounded by a common line L (with unit normal ν pointing into Ω_1), are lines parallel to L. Moreover, $\bar{c}_L(x)$ is only a function $f_\nu(\sigma)$ of the (signed) *scaled* distance $\sigma := s/h$ from x to L, with s, say, as measured along ν. Ambiguity is removed by choosing $s > 0$ in Ω_1. With this and no matter how L is oriented, (1) f_ν is strictly monotone increasing where it is neither zero nor 1; (2) the symmetry of the square S means that $f_\nu = \frac{1}{2}$ exactly when $\sigma = 0$, i.e.,

$$L = \bar{c}_L^{-1}(\tfrac{1}{2}) \quad \text{for arbitrary lines } L; \tag{1.3}$$

and (3) the graph of f_ν is symmetric under reflection through the point $(0, \frac{1}{2})$. f_ν has been called the *cumulative distribution function for ν-normal lines* (with respect to the unit square centered on the origin)—cf. [8], where it is also expressed more analytically.

The blurring (1.2) is a special case of the following. Assume we are given a probability density ρ defined on the plane. Let ρ_z be its translate to an "origin" z, i.e., $\rho_z(x) := \rho(x + z)$. Redefine the "average color" of a color function c to be the scaled convolution

$$\bar{c}(x) := \iint_{\text{plane}} \rho_z\bigl((\xi - x)/h\bigr)\, c(\xi)\, dA(\xi)/h^2\,; \tag{1.4}$$

it bears the origin z and the scale h as parameters.

Consider what needs to be imposed on ρ so that, for arbitrary lines L, the property (1.3) of the associated $\bar{c}_L$ holds true. To fix ideas, suppose ρ were the characteristic function of an equilateral triangle T of unit area. Is there some z so that (1.3) holds? There is for fixed L—it is any z on the line $\hat{L}$ parallel to L that contains half of the area of T on one side and half on the other. But if (1.3) is to be independent of L, then the envelope of such lines must reduce to a point, x_0, that we would then label the *point that divides ρ in half*, or, *the halfway point for ρ* (the phrase "ρ's center of mass" having been preempted by $\iint \xi\, \rho(\xi)\, dA$, and "midpoint" being less dynamic). To see that T has no halfway point, it suffices to observe that, while the three altitudes each bisect T and intersect $\frac{2}{3}$ of the way down an altitude from its corresponding vertex, the corresponding relative distance to the line that both bisects T and is parallel to a side of T is $\sqrt{2}/2$. A fixed affine map taking T onto a given general triangle maps these altitudes onto medians and bisecting lines onto bisecting lines (since its Jacobian is independent of location). It follows that no triangle has a characteristic function that has a halfway point. (That is to say, the envelope of the bisecting lines is a curve, not a point. See, e.g., [2], [3], [4, §253, p. 382], [5, p. 190], or [6, Ex. 3, p. 232].)

For what densities do halfway points exist? In terms of polar coordinates centered on ρ's halfway point, the quantity

$$\int_\alpha^{\alpha+\pi} \int_0^\infty r\, \rho(r, \theta)\, dr\, d\theta$$

must be constant (namely, $\frac{1}{2}$). Differentiating it, we see that x_0 divides ρ in half if and only if (in these same polar coordinates and for all θ) $\int_0^\infty r\, \rho(r, \theta)\, dr = \int_0^\infty r\, \rho(r, \pi + \theta)\, dr$. In particular, a halfway point for ρ, if it exists, need not coincide with ρ's center of mass. The following statements are justified in the next section. An equivalent characterization of a halfway point for ρ (should it exist) is that opposite sectors, bounded by two lines through it, have the same mass. ρ has at most one halfway point.

If ρ is symmetric under reflection through a point x_0, (i.e., if $\rho(x_0 + x) = \rho(x_0 - x)$ for all x), then x_0 divides ρ in half. On the other hand, if ρ is nonzero only on an open set S and has a halfway point x_0 with respect to which S is star-shaped, and if $\rho(x_0 + x) = \rho(x_0 - x)$ when $x_0 - x$ and $x_0 + x$ are both in S (for example, if, on the S-part of each line through x_0, ρ were constant), then both S and ρ are symmetric under reflection through x_0. In particular, x_0 is the halfway point for the (normalized) characteristic function χ_K of a bounded open convex set K if and only if K is symmetric under reflection through x_0.

For $L = L_\nu$ to coincide with the associated $\bar{c}_L^{-1}(\frac{1}{2})$ (and not be just a proper subset), we must also assume, with z being ρ's halfway point, that ρ_z's associated *cumulative distribution function for* ν*-normal lines* [8], namely,

$$f_\nu(\sigma) := \int_{-\sigma}^{\infty} \int_{-\infty}^{\infty} \rho(z + s\nu + t\omega)\, dt\, ds, \quad \omega \cdot \nu = 0, \quad \|\omega\| = \|\nu\| = 1, \tag{1.5}$$

is strictly monotone at $(0, \frac{1}{2})$; indeed, it is useful to impose a wider extension of this—see (2.3) ff. below. And recall that ν points into Ω_1, i.e., towards that side of the line L where $c = 1$, so that $\bar{c}_L(x) = f_\nu(\sigma)$, with $h\sigma$, the signed distance from the line to x as measured along ν, being negative for x in Ω_0 and nonnegative, otherwise.

Examples of probability densities having halfway points include (1) the (normalized) characteristic functions of parallelograms or ellipses; (2) normalized, azimuthally symmetric densities (which are functions only of r in polar coordinates centered on some point, i.e., functions called "radial functions" in recent literature concerning scattered data approximation); and (3) the following two structures: Suppose two circular discs each have mass one and uniform but different densities. Interchange a sector $0 \leq \theta < \alpha$ of one disc with a similar sector of the other, matching up (former) centers and bounding radii, this last as far as is possible. For either of the new structures, its halfway point—namely, the discs' former centers—is not its center of mass.

It is worth noting that, for distributions over *one* space dimension, halfway points *always* exist—they are called "medians" of the distribution. Recall that, as in more dimensions, they needn't coincide with the distribution's mean, but that for one space dimension they needn't be unique.

2. Halfway points in n dimensions. Let g, a nonnegative integrable function on n-dimensional Euclidean space $\mathbb{R}^n$, be not identically zero and sufficiently smooth—we take this to mean that (a) the support of g (i.e., the closure of the set on which g is nonzero) is the closure of its nonempty interior, and (b) the amount of $r^{n-1}g$ on half lines, namely,

$$G(x, \nu) := \int_0^\infty r^{n-1} g(x + r\nu)\, dr,$$

is, for each x in $\mathbb{R}^n$, an integrable function on the unit sphere $\|\nu\| = 1$ in $\mathbb{R}^n$.

We shall say that *the point x_0 in a k-dimensional hyperplane $H_k \subseteq \mathbb{R}^n$ divides g in half on H_k* means that *for any $(k-1)$-dimensional hyperplane $H_{k-1} \subset H_k$ through x_0, the amount of g in H_k on either side of H_{k-1} is the same; more precisely,*

$$\int_{H^+(H_{k-1})} g(x)\, d^k x = \int_{H^-(H_{k-1})} g(x)\, d^k x,$$

where the $H^\pm(H_{k-1})$ are the sides of H_k with respect to H_{k-1}. Such a point will be called *a halfway point for g on H_k*. With this, we have the following.

PROPOSITION A. (1) *Suppose $G(x_0, \cdot)$ is continuous on the unit sphere. Then if for some k, $1 \leq k \leq n$, x_0 is a halfway point for the function $\|x - x_0\|^{n-k} g(x)$ on every k-dimensional hyperplane in $\mathbb{R}^n$ containing x_0, then this is so for all such k.* (2) *If x_0 is a halfway point for g on $\mathbb{R}^n$, $G(x_0, \cdot)$ is continuous on the unit sphere, and H is any $(n-1)$-dimensional hyperplane through x_0, then x_0 is the center of mass of the H-section of g, i.e., of the density distribution $g|_H$.* (3) *A halfway point for g on $\mathbb{R}^n$ is unique for $n > 1$.* (4) *If g is symmetric under reflection through a point x_0, then x_0 divides g in half.* (5) *If g is nonzero only on an open set S and has a halfway point x_0 with respect to which S is star-shaped, and if $g(x_0 + x) = g(x_0 - x)$ when $x_0 - x$ and $x_0 + x$ are both in S, then both S and g are symmetric under reflection through x_0. In particular, x_0 is the halfway point for the characteristic function χ_K of a bounded open convex set K if and only if K is symmetric under reflection through x_0.*

For these results we need a lemma and its corollary.

LEMMA. *That x_0 divides g in half on a k-dimensional hyperplane H_k $(k > 1)$ is equivalent to the following. Let $S_1 \neq S_2$ be $(k-1)$-dimensional hyperplanes in H_k and through x_0. Then the amount of g between opposing sectors (bounded by S_1, S_2, and their intersection) is the same.*

Proof. Sufficiency is clear. (Let the two hyperplanes coalesce.) For necessity, identify four regions in H_k, $X_{\pm\mp}$, by the signs of $\lambda_i - \lambda_i(x_0)$ for two appropriate linear functionals λ_i (i.e., such that S_i is the null space in H_k of $\lambda_i - \lambda_i(x_0)$); and let the four numbers $g_{\pm\mp}$ be the amount of g in $X_{\pm\mp}$. Then, the difference between the two relations,

$$g_{++} + g_{+-} = g_{--} + g_{-+} \quad \text{and} \quad g_{+-} + g_{--} = g_{-+} + g_{++},$$

yields the desired relation $g_{++} = g_{--}$. □

In this Lemma, let $S_1 \cap S_2 =: H_{k-2}$ be fixed, and let S_1 approach $S_2 =: H_{k-1}$. Then we are led to the following.

COROLLARY. *If x_0 divides g in half on H_k, some $k \geq 2$, and $G(x_0, \cdot)$ is continuous, then for all $H_{k-2} \subset H_{k-1}$ both containing x_0,*

$$\int_{H^+(H_{k-2})} \|x - x_0\| \, g(x) \, d^{k-1}x = \int_{H^-(H_{k-2})} \|x - x_0\| \, g(x) \, d^{k-1}x.$$

Proof. Let $S_1 \neq S_2 = H_{k-1}$ be $(k-1)$-dimensional hyperplanes in H_k, each containing x_0, with $S_1 \cap S_2 = H_{k-2}$. Let $W^{\pm}$ be the opposite sectors of the regions bounded by S_1 and S_2 with interior angle ϕ. Then from the Lemma,

$$\int_{W^-} g(x) \, d^k x = \int_{W^+} g(x) \, d^k x.$$

Recall from, e.g., Sommerfeld [7, pp. 227–228] that the volume element in k-dimensional spherical coordinates, namely (for $r > 0$, $-\pi < \theta_1 < \pi$, and $0 < \theta_i < \pi$ when $i > 1$),

$$d^k x = \sin^{k-2}\theta_{k-1} \sin^{k-3}\theta_{k-2} \cdots \sin\theta_2 \, r^{k-1} \, dr \, d\theta_1 \cdots d\theta_{k-1} \tag{2.1}$$

satisfies $d^k x = \sin^{k-2}\theta_{k-1} \, (r \, d^{k-1}x) \, d\theta_{k-1}$. Place the origin at x_0, and place the polar axis normal to H_{k-1} and oriented so that all points in $W^{\pm}$ satisfy $\theta_{k-1} = \pi/2 \pm \psi_{\pm}$, some $\psi_\pm \in [0,\phi]$. Let $C^\pm(\psi)$ be the cone with vertex x_0 that is defined by $\theta_{k-1} =$ constant $= \pi/2 \pm \psi$. By Fubini's theorem,

$$\int_{\pi/2-\phi}^{\pi/2} \left[\int_{W^-\cap C^-(\pi/2-\theta_{k-1})} g(x,\theta_{k-1})\, r\, d^{k-1}x \right] \sin^{k-2}\theta_{k-1}\, d\theta_{k-1}$$
$$= \int_{\pi/2}^{\pi/2+\phi} \left[\int_{W^+\cap C^+(\theta_{k-1}-\pi/2)} g(x,\theta_{k-1})\, r\, d^{k-1}x \right] \sin^{k-2}\theta_{k-1}\, d\theta_{k-1}.$$

The conclusion of the Corollary is obtained by letting $\phi \to 0$ and by using the mean value theorem for integrals, which may be done by the continuity of $G(x_0, \cdot)$. □

To prove (1) in Proposition A, we suppose that for some k, $1 \le k \le n$, x_0 is a halfway point for $\|x - x_0\|^{n-k} g(x)$ on all k-dimensional hyperplanes H_k containing x_0. If $k > 1$, apply the Corollary successively to the functions

$$\|x-x_0\|^{n-k} g(x)\big|_{H_k}, \quad \|x-x_0\|^{n-k+1} g(x)\big|_{H_{k-1}}, \cdots, \|x-x_0\|^{n-2} g(x)\big|_{H_2}.$$

Thus, x_0 is a halfway point for the restriction of $\|x-x_0\|^{n-1} g(x)$ to any line L through x_0, so that $G(x_0,-\nu) = G(x_0,\nu)$ for all ν on the unit sphere. Placing the origin for spherical coordinates at x_0 with the polar axis normal to a given but arbitrary $(n-1)$-dimensional hyperplane containing x_0, we see that x_0 is a halfway point for g on $\mathbb{R}^n$. We may now replace k by n in this argument to complete the proof of (1).

To prove (2) in Proposition A, restrict g to the $(n-1)$-dimensional hyperplane H, and let n-dimensional spherical coordinates be centered on x_0 with polar axis normal to H, so that a line L in H through x_0 is determined by fixing $0 < \theta_1, \cdots, \theta_{n-2} < \pi$ and $\theta_{n-1} = \pi/2$. Then, taking $k = 1$ in part (1), the amount of $r^{n-1} g\big|_L$ on either side of the origin is the same: $\int_{r>0} r^{n-1} g(r,\cdot)\big|_L\, dr = \int_{r<0} r^{n-1} g(r,\cdot)\big|_L\, dr$. It follows that the component of the center of mass of $g\big|_H$ along any unit vector u parallel to H vanishes, since it is proportional to $\int_H g\big|_H r\cos\alpha\, dx^{n-1}$, i.e., (using (2.1) with $k = n-1$), to

$$\int_0^\pi \cdots \int_0^\pi \int_0^\pi \int_0^\infty \big(g(r,\cdot,\pi/2) - g(-r,\cdot,\pi/2)\big)\, r\cos\alpha$$
$$r^{n-2}\sin^{n-3}\theta_{n-2}\sin^{n-4}\theta_{n-3}\cdots\sin\theta_2\, dr\, d\theta_1\cdots d\theta_{n-2},$$

in both of which α, the angle (in $\mathbb{R}^n$) between u and the radius vector, is independent of r.

To prove (3) in Proposition A, suppose $n > 1$ and that $x_0 \ne x_1$ are both halfway points determining a line L; and let x not lie on L. We want to show $g(x) = 0$; for then, x being arbitrary, g would be zero or concentrated on L, cases we do not allow. For this, let $H(0)$ and $H(1)$ be two parallel $(n-1)$-dimensional hyperplanes with (a) x strictly between them, and (b) x_0 in $H(0)$ and x_1 in $H(1)$. Then, the amount of g on the side of $H(0)$ not containing x_1 is the amount of g on the side of $H(1)$ not containing x_0. Hence, the amount of g between $H(0)$ and $H(1)$ is zero. Since x lies strictly between these two hyperplanes, $g(x) = 0$.

The proof of (4) in Proposition A is clear.

To prove (5) in Proposition A, choose a line L through x_0; and let $x_0 - w$ and $x_0 + z$ be the endpoints of the open line segment $L \cap S$, choosing $\|w\| \le \|z\|$. Then, with $a(y)$ the amount of the function $\|x - x_0\|^{n-1} g(x_0 + x)$ between x_0 and $x_0 + y$

on L, $a(w) = a(-w)$. But, since x_0 is the halfway point for g, $a(z) = a(-w)$. So, since $x_0 + w$ and $x_0 + z$ lie on the same side of x_0, the amount of g between $x_0 + w$ and $x_0 + z$ is zero. Since S is star-shaped with respect to x_0, this amount would be positive if $\|w\| < \|z\|$. So, $\|w\| = \|z\|$; and, as L was arbitrary, the boundary of S is symmetric under reflection through x_0. As S is star-shaped with respect to x_0, S itself has the same symmetry. Hence, so does g. It remains to show that if, instead, $S = K$ is bounded, open and convex, and if $g := \chi_K / \int_{\mathbb{R}^n} \chi_K$ has halfway point x_0, then $x_0 \in K$ (for then K is star-shaped with respect to x_0). But if $x_0 \notin K$, then there is an $(n-1)$-dimensional hyperplane H with $x_0 \in H$, but with K completely on one side of H—say, the side in the direction of H's unit normal vector ν. But, then, the "cumulative distribution function f_ν for ν-normal hyperplanes" associated with the density $\rho := g$ ((2.3) below, with $z := 0$) satisfies (a) $f_\nu(0) = \frac{1}{2}$ and (b) $f_\nu(\sigma) = 1$ for $\sigma > 0$. But this contradicts the continuity of f_ν, so $x_0 \in K$ after all. □

In Proposition A, (1) fails to characterize halfway points in some easy cases—for example, $G(x_0, \cdot)$ fails to be continuous for the open, azimuthally symmetric, two-dimensional "butterfly" $\{(r,\theta) : 0 < |r| < 1,\ 0 < \theta < \alpha_0 < \pi/2\}$ or its partly open, unsymmetric sibling, $\{(r,\theta) : 0 < r < 1;\ 0 \le \theta < \alpha_0 \text{ or } \pi < \theta \le \pi + \alpha_0\}$. It might prove amusing to explore the extent that (1) (and the definition of halfway point) could be weakened to: $(\hat{1})$ *If for some k, $1 \le k \le n$, x_0 is a halfway point for the function $\|x - x_0\|^{n-k} g(x)$ on almost every k-dimensional hyperplane in $\mathbb{R}^n$ containing x_0, then this is so for all such k.*

Remark. The conclusion of (2) in Proposition A—namely, that a distribution's halfway point in $\mathbb{R}^n$ is also the center of mass of every $(n-1)$-dimensional (hyper)plane section containing it—is also a consequence of the first of a sequence of results in the hydrostatics of floating bodies. This connection goes as follows (here in three dimensions but with no loss in generality): Suppose one is given a body B (say, a convex body) along with a prescribed fraction γ in $(0,1)$. Let E be the envelope of those planes that divide B into two parts having relative volume γ and $1-\gamma$. (If the body B had specific gravity γ and were floating in some orientation, then its sea-level plane section would be tangent to E independent of that orientation.) The result of particular interest in this context states that *the point of tangency of each of the dividing planes to their envelope E is the (two-dimensional) center of mass of the corresponding plane section of B.* This result is the simplest of a sequence of conclusions concerning the orientation and stability of the equilibrium positions for floating bodies; a sequence generally attributed (by Bouasse [4], Appell [1, §651, p. 196ff], Greenhill [5, p. 160], and Lamb [6, pp. 227ff]) to Bouguer (1746) in part and to Dupin (1814)—although Bouasse [4, §220, p. 324] attributes this particular result instead to Lacroix, one of the early (1797) calculus textbook writers. All justify the result in much the same way that our Lemma and its Corollary were proved (but no author clearly states the hypotheses he assumes). The typical argument goes as follows. Two nearby planes of the type mentioned intersect in a line L; the amount of B in each of the two narrow sectors between the planes on either side of L must be the same, but this is also closely approximated by the product of the small angle between the planes with the plane section's first moment in either of the directions orthogonal to L. Let the two planes and the lines L converge. Then the limiting line contains the section's center of mass along with the limiting plane's point of tangency to the envelope.

A similar argument indicates that the result should extend to higher dimensions and to the division by hyperplanes of relatively general mass distributions into two parts having fixed but not necessarily equal masses. (L, then, is an $(n-2)$-dimensional hyperplane.) We see no difficulties with such extensions in the case of interest here—namely, when the entire envelope reduces to a point. But there do exist counterexamples in more general situations—see, e.g., Beyer and Swartz [2, Ex. 10.3].

To complete this Remark we reconsider part (2) of Proposition A in our three-dimensional context. So we suppose above that $B \subset \mathbb{R}^3$ has a halfway point x_0. Then, taking $\gamma = \frac{1}{2}$, the envelope E reduces to x_0. Thus the result above implies, as desired, that every plane section of B containing x_0 has center of mass x_0.

In the context of boundary reconstruction in n dimensions, the average color of a color function c is the analog of the scaled convolution (1.4):

$$\bar{c}(x) := \int_{\mathbb{R}^n} \rho_z\big((\xi - x)/h\big)\, c(\xi)\, d\xi^n/h^n; \tag{2.2}$$

it again bears the "origin" z (of the given probability distribution ρ) and the "scale" h as parameters. With the z the halfway point for ρ (=: g above) assumed to exist, we would also like to ensure that each $(n-1)$-dimensional hyperplane boundary H *coincides* with the inverse image $\bar{c}_H^{-1}(\frac{1}{2})$ under the average coloring $\bar{c}_H$ of $\mathbb{R}^n$ associated with H, and is not just a proper subset of that inverse image. For this, define, for each unit vector ν (regarded as the normal for an associated $(n-1)$-dimensional hyperplane $H(\nu)$ through the origin), the *cumulative distribution function f_ν of ρ for ν-normal hyperplanes,* namely,

$$f_\nu(\sigma) := \int_{-\sigma}^{\infty} \int_{H(\nu)} \rho(z + s\nu + x_{n-1})\, d^{n-1}x_{n-1}\, ds. \tag{2.3}$$

Then, $H = \bar{c}_H^{-1}(\frac{1}{2})$ for all H if and only if each f_ν is strictly increasing on a neighborhood of $f_\nu^{-1}(\frac{1}{2})$; that is to say, there exists no pair of distinct $(n-1)$-dimensional ν-normal hyperplanes that separate the support of ρ into two pieces of mass $\frac{1}{2}$. In fact, [8] imposes a further restriction on ρ that ensures the even more desirable property that each $df_\nu/d\sigma$ is bounded away from zero where f_ν is bounded away from both zero and one: *If, for some ν and some $\sigma_0 \neq 0$, $\int_{H(\nu)} \rho(z + \sigma_0\nu + x_{n-1})\, dx^{n-1} = 0$, then ρ vanishes on the whole open half space on the side of $H(\nu) + \sigma_0\nu$ away from ρ's halfway point z.*

Two examples of cumulative distributions f_ν (2.3) associated with particular probability densities ρ in n dimensions are presented in [8]. The first is the construction of the univariate nth degree polynomial spline that is the cumulative distribution associated with a (mass one) union of interiorly disjoint tetrahedra (simplices)—of course, such objects need not have halfway points. The second involves the determination of the azimuthally symmetric probability density ρ on the n-ball (of diameter one) whose cumulative distribution function is linear where it is neither constantly zero nor constantly one.

The analog $H = \bar{c}_H^{-1}(\frac{1}{2})$ of (1.3) has led us to the notion of a halfway point for the probability distribution ρ occurring in the scaled blurring (2.2), i.e., to a generalized "median" whose existence imposes a sort of symmetry restriction on ρ. Other properties of the blurring can lead to other sorts of restrictions. For example, the cumulative distribution function f_ν for ν-normal hyperplanes (2.3) specifies the behavior of the average color as the moving density distribution crosses the boundary of a half space from its black (0) side to white (1) side. We could ask that identical behavior be obtained if black is interchanged with white and we cross in the opposite direction. That is to say, we could insist that

$$\text{(2.4)} \qquad \text{for each normal } \nu, \qquad f_{-\nu}(\sigma) = f_\nu(\sigma) \quad \text{for all } \sigma.$$

This is equivalent (using the consequence $f_{-\nu}(\sigma) = 1 - f_\nu(-\sigma)$ of invoking a *probability* density ρ) to insisting that the graph of f_ν be symmetric under reflection through $(0, \frac{1}{2})$:

$$\text{(2.5)} \qquad \text{for each normal } \nu, \qquad \tfrac{1}{2} - f_\nu(-\sigma) = f_\nu(\sigma) - \tfrac{1}{2} \quad \text{for all } \sigma.$$

In particular, this insists that, for each ν, $f_\nu(0) = \frac{1}{2}$; i.e., that the associated density ρ have a halfway point. But (2.5) must restrict ρ even more, since, for either structure constructed at the end of §1, the associated f_ν does not satisfy (2.5) except for ν normal to the line that cuts the structure into two mirror images.

And, indeed, (2.5) is equivalent to ρ being almost everywhere symmetric under reflection through its halfway point.

PROPOSITION B. *Let ρ be a probability density on $\mathbb{R}^n$. Then ρ's cumulative distribution function f_ν (defined by (2.3) with $z = 0$) is symmetric under reflection through $(0, \frac{1}{2})$ for all unit vectors ν—i.e., (2.5) holds—if and only if ρ is symmetric under reflection through the origin almost everywhere. Either condition means that the origin is ρ's halfway point.*

Proof. Such symmetry for ρ is clearly sufficient. On the other hand, suppose that f_ν (2.3) (with $z = 0$) satisfies (2.5). Then the cumulative distribution function $f_\nu^{(R)}$ for the "reflected" density

$$\rho^{(R)}(x) := \rho(-x)$$

is the same as f_ν:

$$\begin{aligned} f_\nu^{(R)}(\sigma) &= \int_{-\sigma}^{\infty} \int_{H(\nu)} \rho^{(R)}(s\nu + x_{n-1})\, d^{n-1}x_{n-1}\, ds \\ &= \int_{-\infty}^{\sigma} \int_{H(\nu)} \rho(t\nu - x_{n-1})\, d^{n-1}x_{n-1}\, dt = 1 - f_\nu(-\sigma) = f_\nu(\sigma) \end{aligned}$$

(the last using (2.5)). Differentiating, now, the amount of $\Delta := \rho^{(R)} - \rho$ on any $(n-1)$-dimensional hyperplane is zero. In other words, the Radon transform of Δ is zero. But it now follows—from a nice argument described to us by L. Shepp—that the n-dimensional Fourier transform $\hat{\Delta}$ of Δ is zero. For with $-\infty < \sigma < \infty$, and since $\|\nu\| = 1$ with ν normal to $H(\nu)$,

$$\begin{aligned} (2\pi)^{n/2}\, (\hat{\Delta})(\sigma\nu) :&= \int_{-\infty}^{\infty} \int_{H(\nu)} e^{i\sigma\nu\cdot(s\nu + x_{n-1})}\, \Delta(s\nu + x_{n-1})\, d^{n-1}x_{n-1}\, ds \\ &= \int_{-\infty}^{\infty} e^{i\sigma s} \left(\int_{H(\nu)} \Delta(s\nu + x_{n-1})\, d^{n-1}x_{n-1} \right) ds. \end{aligned}$$

But the inner integral is the Radon transform of Δ at (s, ν). Thus $\hat{\Delta} = 0$; so $\Delta = 0$ almost everywhere. □

Acknowledgment. We thank R. Hersh, J. Louck, and R. Mjolsness for fruitful conversations.

REFERENCES

[1] P. APPELL, *Équilibre et mouvement des milieux continus*, Traité de mécanique rationnelle, Vol. 3, Third ed., Gauthier-Villars, Paris, 1921.

[2] W. A. BEYER AND B. SWARTZ, *The envelope of the planes that bisect a tetrahedron*, Los Alamos National Laboratory, LA-UR-90-2491, Los Alamos, NM, 1990.

[3] ———, *Bisectors of triangles and tetrahedra*, Amer. Math. Monthly (1993), to appear.

[4] H. BOUASSE, *Hydrostatique: manomètres, baromètres, pompes; équilibre des corps flottants*, Librairie Delagrave, Paris, 1923.

[5] A. G. GREENHILL, *A Treatise on Hydrostatics*, Macmillan, London, 1894.

[6] H. LAMB, *Statics: Including Hydrostatics and the Elements of the Theory of Elasticity*, Third ed., Cambridge University Press, Cambridge, 1928.

[7] A. SOMMERFELD, *Partial Differential Equations in Physics*, Academic Press, New York, 1949.

[8] B. SWARTZ, *The second-order sharpening of blurred smooth borders*, Math. Comp., 52 (1989), pp. 675–714.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1342–1351, September 1992

017

ROTATION INVARIANT SEPARABLE FUNCTIONS ARE GAUSSIAN*

PL. KANNAPPAN[†] AND P. K. SAHOO[‡]

Abstract. In digital image analysis, edge detection, line detection, texture classification, and so forth are basic to image understanding and interpretation. Since a typical image is void of predetermined directions of edge, line, or texture, rotation invariant filters are important for their detection. In designing such filters the concept of rotation invariant separable function is often used. Here it is shown that every rotation invariant separable real valued function of two variables is either Gaussian or identically zero. This justifies the use of Gaussian filters in image processing. Furthermore, while deriving this result, the authors obtained the general solutions of two functional equations considered by Swiatak in 1975. Swiatak found the solutions when one of the unknown functions is continuous at zero. No regularity assumptions were made and all the general solutions that contain the solution obtained in [*Aequationes Math.*, 12 (1975), pp. 39–64] were determined.

Key words. Gaussian function, separable function, functional equation

AMS(MOS) subject classifications. 39B40, 82A15, 68U10

1. Introduction. Let R be the set of reals and $\mathbf{R}_+ = \{x \in \mathbf{R} \mid x > 0\}$ be the set of positive reals. Let $\Omega = \{\theta \in \mathbf{R} \mid \theta \neq n\pi/2,\, n = 0, 1, 2, 3, 4, \cdots\}$. A function $f : \mathbf{R}^n \to \mathbf{R}$ is said to be *Gaussian* if and only if

$$f(x_1, x_2, \cdots, x_n) = k\, e^{A(x_1^2 + x_2^2 + \cdots + x_n^2)}, \qquad (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n, \tag{1.1}$$

where k is a real nonzero constant and $A : \mathbf{R} \to \mathbf{R}$ is a function satisfying

$$A(u+v) = A(u) + A(v) \tag{1.2}$$

for all $u, v \in \mathbf{R}$. In the literature, A is often referred to as an *additive* function. The probability density function of a normal distribution is an example of a Gaussian function. A function $G : \mathbf{R}^2 \to \mathbf{R}$ is said to be *separable* if there exist functions $g, h : \mathbf{R} \to \mathbf{R}$ such that

$$G(x, y) = g(x)\, h(y) \tag{1.3}$$

for all $x, y \in \mathbf{R}$. A function $G : \mathbf{R}^2 \to \mathbf{R}$ is said to be *rotation invariant* if

$$G(x, y) = G(x_\theta, y_\theta) \quad \text{for all } \theta \in \Omega, \tag{1.4}$$

where

$$\begin{pmatrix} x_\theta \\ y_\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

*Received by the editors June 12, 1991; accepted for publication (in revised form) January 21, 1992. This work was supported by grants from the College of Arts and Sciences of the University of Louisville and the Natural Sciences and Engineering Research Council of Canada.

[†] Department of Pure Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.

[‡] Department of Mathematics, University of Louisville, Louisville, Kentucky 40292.

If a function $G : \mathbf{R}^2 \to \mathbf{R}$ is rotation invariant and separable then

$$g(x)\,h(y) = k(x_\theta)\,l(y_\theta) \quad \text{for all } x, y \in \mathbf{R} \quad \text{and} \quad \text{for all } \theta \in \Omega, \tag{1.5}$$

where g, h, k, $l : \mathbf{R} \to \mathbf{R}$. The introduction of functions k and l is just to give a general aspect to the functional equation. It is not imposed by separability. In digital image processing edge detection, line detection, texture classification, and so forth are basic to the problems of image understanding and interpretation. Since a typical image does not have predetermined directions of edge, line, or texture, rotation invariant filters are used for their detection. In designing of such filters (see [2]) the concept of rotation invariant separable function is used. Also, in many problems in optics and quantum mechanics we encounter such functions. For instance, two independent particles in quantum mechanics whose wavefunction is a product of function of individual coordinates is also a product when expressed in the center of mass and relative coordinates. In this paper we show that rotation invariant, separable functions satisfy the functional equations (FE) and its analogue considered in §2 are either Gaussian or zero function. In [4], the authors study bifactorizable quantum wave functions. While proving if two quantum systems are prepared independently and if their center of mass is found to be in a pure state, then each of the component systems is also in a pure state, which in the coordinate representation is a Gaussian wave function, the authors of [4] come across the functional equation (FE). They established their result assuming the functions are infinitely differentiable. We establish this result without assuming any regularity conditions on the unknown functions. Our approach is from the functional equation point of view and it is simple, direct, and interesting.

2. Auxiliary results. We give the motivation for considering the functional equation (FE) in the next section. In this section we determine the most general solution of the functional equations (FE), (GFE), (SE), and (GSE). Since the identically zero function is always a solution of our functional equations, we consider nontrivial (that is, not identically zero) solutions. First we prove the following theorem concerning the equation (FE). We require the following theorem to establish our main result.

THEOREM 1. *Let f, $g : \mathbf{R} \to \mathbf{R}$ satisfy the functional equation*

$$f(x+y) = f(x)\,f(y)\,g(xy), \qquad x,\ y \in \mathbf{R}, \tag{FE}$$

where f and g are not identically zero functions. Then the general solution of (FE) *is given by*

$$\left.\begin{aligned} f(x) &= c\,e^{\frac{1}{2}A_1(x^2)+A_2(x)} \\ g(x) &= \frac{1}{c}\,e^{A_1(x)} \end{aligned}\right\} \qquad x \in \mathbf{R}, \tag{2.1}$$

where $A_i : \mathbf{R} \to \mathbf{R}$ $(i = 1, 2)$ are additive functions and c is an arbitrary nonzero real constant.

Proof. If there exists a real number x_o such that $f(x_o) = 0$, then replacing x by $x - x_o$ and y by x_o in (FE), we obtain

$$f(x) = f(x - x_o)\,f(x_o)\,g((x - x_o)\,x_o) = 0$$

for all $x \in \mathbf{R}$. But f is not identically zero, and hence f is nowhere zero. If $g(x_o) = 0$ at some $x_o \in \mathbf{R}$, then from (FE) with $y = 1$, we get

$$f(x_o + 1) = f(x_o)\,f(1)\,g(x_o) = 0. \tag{2.2}$$

But f is nowhere zero, and hence g is also nowhere zero.

From (FE), we obtain

$$\begin{aligned} f(x+y+z) &= f((x+y)+z) \\ &= f(x+y)\,f(z)\,g((x+y)z) \\ &= f(x)\,f(y)\,f(z)\,g(xy)\,g((x+y)z). \end{aligned}$$

Similarly,

$$f(x+(y+z)) = f(x)\,f(y)\,f(z)\,g(yz)\,g(x(y+z))$$

and

$$f(y+(x+z)) = f(x)\,f(y)\,f(z)\,g(xz)\,g(y(x+z)).$$

Thus from the above equations, we get

$$g(xy)\,g((x+y)z) = g(yz)\,g(x(y+z)) = g(xz)\,g(y(x+z)) \tag{2.3}$$

for all $x, y, z \in \mathbf{R}$. Now we restrict f and g to $\mathbf{R}_+$ (positive reals). It is easy to see that (FE) and (2.3) hold for $x, y, z \in \mathbf{R}_+$. For u, v, $w \in \mathbf{R}_+$, let

$$x = \sqrt{\frac{u\,w}{v}}, \qquad y = \sqrt{\frac{u\,v}{w}}, \qquad z = \sqrt{\frac{v\,w}{u}}. \tag{2.4}$$

Then (2.3) becomes

$$g(u)\,g(v+w) = g(v)\,g(u+w) = g(w)\,g(u+v).$$

Since g is nowhere zero, we rewrite the above equation as

$$\frac{g(v+w)}{g(v)} = \frac{g(u+w)}{g(u)}$$

for all u, v, $w \in \mathbf{R}_+$. Hence for fixed $w \in \mathbf{R}_+$, we get $g(u+w)/g(u) =$ constant for $u \in \mathbf{R}_+$. So we have

$$g(u+w) = g(u)\,h(w), \qquad u,\ v \in \mathbf{R}_+, \tag{2.5}$$

where $h : \mathbf{R}_+ \to \mathbf{R}$. The left side of (2.5) is symmetric in u and w. Using the symmetry of the left side of (2.5), we obtain

$$g(u)\,h(w) = g(w)\,h(u), \qquad u,\ v \in \mathbf{R}_+. \tag{2.6}$$

Letting $w = w_o$ in (2.6), we get

$$h(u) = c\,g(u), \tag{2.7}$$

where $c := h(w_o)/g(w_o)$ is a nonzero real constant. Now (2.7) in (2.5) yields

$$g(u+w) = c\,g(u)\,g(w), \qquad u,\ v \in \mathbf{R}_+. \tag{2.8}$$

Hence (see [1, Thm. 5, p. 29])

$$g(x) = \frac{1}{c}\,e^{A_1(x)}, \qquad x > 0, \tag{2.9}$$

where $A_1 : \mathbf{R} \to \mathbf{R}$ is an additive function.

Inserting (2.9) into (FE), we obtain

$$f(x+y) = f(x)\, f(y)\, \frac{1}{c}\, e^{A_1(xy)}, \qquad x,\, y \in \mathbf{R}_+. \tag{2.10}$$

Define

$$F(x) := \frac{1}{c}\, e^{-\frac{1}{2} A_1(x^2)}\, f(x), \qquad x \in \mathbf{R}_+. \tag{2.11}$$

Then (2.10) and (2.11) yield

$$F(x+y) = F(x)\, F(y), \qquad x, y \in \mathbf{R}_+. \tag{2.12}$$

Hence (see [1, Thm. 5, p. 29])

$$F(x) = e^{A_2(x)}, \qquad x \in \mathbf{R}_+, \tag{2.13}$$

where $A_2 : \mathbf{R} \to \mathbf{R}$ is an additive function. Now from (2.11) and (2.13) we obtain

$$f(x) = c\, e^{\frac{1}{2} A_1(x^2) + A_2(x)}, \qquad x > 0. \tag{2.14}$$

Thus we have shown that f and g have the asserted form (2.1) if $x > 0$. Next we show (2.1) is also the asserted form of f and g for $x \leq 0$.

Let $x < 0$ and choose $y > 0$ such that $x + y > 0$. Then from (FE) and (2.14) we get

$$g(xy)\, f(x) = e^{\frac{1}{2} A_1(x^2) + A_2(x) + A_1(xy)} \tag{2.15}$$

for all $x < 0$ and $x + y > 0$. Now for $x < 0$, choose $z > 0$ such that $x + z > 0$. From (2.3), we have

$$g(xy)\, g(xz + yz) = g(xz)\, g(xy + yz),$$

and since $xz + yz > 0$ and $xy + yz > 0$, using (2.9) in the above equation, we get

$$g(xy)\, e^{-A_1(xy)} = g(xz)\, e^{-A_1(xz)}, \tag{2.16}$$

for all $x < 0$ and $y > -x,\ \ z > -x$. From (2.16), we see that

$$g(xy)\, e^{-A_1(xy)} = r(x),$$

where $r(x)$ is a nowhere zero function of x. Thus

$$g(xy) = r(x)\, e^{A_1(xy)} \tag{2.17}$$

holds for all $x < 0$ and $x + y > 0$. Inserting (2.17) into (2.15), we get

$$f(x) = \frac{1}{r(x)}\, e^{\frac{1}{2} A_1(x^2) + A_2(x)}, \qquad x < 0. \tag{2.18}$$

Next let us choose $x < 0$ and $y < 0$. Using (FE), (2.9), and (2.18), we obtain

$$r(x+y) = r(x)\, r(y)\, c \tag{2.19}$$

for all x, $y < 0$.

Let $f(0) = b$ (a nonzero constant). With $y = 1$, (FE) gives

$$f(x+1) = f(1)\, f(x)\, g(x), \qquad x \in \mathbf{R}. \tag{2.20}$$

Letting $x = -1$ in (2.20) and using (2.18) and (2.14), we get

$$g(-1) = \frac{1}{c}\, b\, r(-1)\, e^{A_1(-1)}. \tag{2.21}$$

Substituting $y = -\frac{1}{x}$ in (FE) for $x > 1$, we get

$$f\left(x - \frac{1}{x}\right) = f(x)\, f\left(-\frac{1}{x}\right)\, g(-1). \tag{2.22}$$

Using (2.14) and (2.18) in (2.22), we obtain

$$g(-1) = r\left(-\frac{1}{x}\right)\, e^{A_1(-1)} \tag{2.23}$$

for all $x > 1$. From (2.21) and (2.23), we see that

$$r(y) = d \qquad \text{(constant)} \tag{2.24}$$

for $-1 < y < 0$. Now we choose x, y in the interval $]-1, 0[$ with $x + y < 0$ and $x + y > -1$. Then (2.19) yields

$$d = \frac{1}{c}.$$

For $x < 0$ and $y \in]-1, 0[$, we obtain from (2.19) and (2.24)

$$r(x+y) = r(x) \tag{2.25}$$

for all $x < 0$ and all $y \in]-1, 0[$. Hence $r(x)$ is a constant function. Again using (2.19), we see that

$$r(x) = \frac{1}{c} \quad \text{for } x < 0. \tag{2.26}$$

Now using (2.26) in (2.18), we get

$$f(x) = c\, e^{\frac{1}{2} A_1(x^2) + A_2(x)}, \qquad x < 0. \tag{2.27}$$

From (2.20), (2.14), and (2.27), we get

$$g(x) = \frac{1}{c}\, e^{A_1(x)}, \qquad x < 0. \tag{2.28}$$

Hence f and g have the form (2.1) for $x < 0$. Now $y = -x$ in (FE) yields $b = c$, that is, $f(0) = c$. This completes the proof of the theorem.

Now a few remarks are in order.

Remark 2. First the solution of (FE) is obtained without assuming any regularity condition whatsoever on f or g. Furthermore, this solution is obtained without using the substitutions $x = 0$ or $y = 0$.

Remark 3. If we assume some regularity condition on f, say f is measurable, or continuous or continuous at a point, say, at zero, then we obtain as in [6], $f(x) = c e^{\frac{1}{2} a x^2 + b x}$ and $g(x) = \frac{1}{c} e^{a x}$. The measurability or continuity on f implies (use (2.20)) the same of g, and (2.18) and (2.12) yield continuous additive functions A_1 and A_2. As for continuity at zero of f, as in [6], we can show that f is continuous everywhere.

Our next remark is the following. We can easily obtain the general solutions of the functional equation

$$\text{(GFE)} \qquad f_1(x+y) = f_2(x)\, f_3(y)\, g(xy), \qquad x,\, y \in \mathbf{R},$$

where f_1, f_2, f_3, g are not identically zero by reducing it to (FE).

COROLLARY 4. *The general nontrivial solution of* (GRE), *where* f_1, f_2, f_3, g : $\mathbf{R} \to \mathbf{R}$ *is given by*

$$\left.\begin{aligned} f_1(x) &= a\, b\, c\, e^{\frac{1}{2} A_1(x^2) + A_2(x)} \\ f_2(x) &= \frac{1}{a} f_1(x) \\ f_3(x) &= \frac{1}{b} f_1(x) \\ g(x) &= \frac{1}{c} e^{A_1(x)} \end{aligned}\right\} \qquad x \in \mathbf{R},$$

where $A_i : \mathbf{R} \to \mathbf{R}$ $(i = 1, 2)$ *are additive functions and* a, b, c *are arbitrary nonzero real constants.*

Proof. The assumption that either f_2 or f_3 is zero at some point would imply f_1 is identically zero. Suppose $g(x_o) = 0$. Then (GFE) first gives $f_1(x_o + 1) = 0$ and then with $x = x_o + 1$, $y = 0$ gives $g(0) = 0$. This would imply f_1 is identically zero. So, all f_1, f_2, f_3, and g are nowhere zero. The reduction of (GRE) to (FE) can be accomplished as follows. By substituting into (GFE), first $y = 0$ and then $x = 0$, we have $f_1(x) = a\, f_2(x)$ and $f_1(y) = b\, f_3(x)$ so that (GFE) becomes

$$\frac{1}{a\,b} f_1(x+y) = \frac{1}{a\,b} f_1(x)\, \frac{1}{a\,b} f_1(y)\, g(xy),$$

which is the functional equation (FE).

Next we prove the following corollary to establish another result in [6]. Again we make no regularity assumptions on the unknown functions. Again, as before we solve it by reducing it to (FE).

COROLLARY 5. *Let* f, $g : \mathbf{R} \to \mathbf{R}$ *satisfy the functional equation*

$$\text{(SE)} \qquad f(x-y) = f(x)\, f(y)\, g(xy) \qquad (x, y \in \mathbf{R}),$$

where f *and* g *are not identically zero. Then the general solution of* (SE) *is given by*

$$\text{(2.29)} \qquad \left.\begin{aligned} f(x) &= c\, e^{\frac{1}{2} A_1(x^2)} \\ g(x) &= \frac{1}{c} e^{A_1(-x)} \end{aligned}\right\} \qquad x \in \mathbf{R},$$

where $A_1 : \mathbf{R} \to \mathbf{R}$ *is an additive function and* c *is a nonzero real constant.*

Proof. It is easy to verify that (2.29) satisfies (SE). We assume f and g are nowhere zero. Then interchanging x with y in (SE), we get

$$f(x-y) = f(x)\,f(y)\,g(xy) = f(y-x) \tag{2.30}$$

for all x, $y \in \mathbf{R}$. Thus f is an even function in $\mathbf{R}$. Replacing y by $-y$ in (SE), we get

$$f(x+y) = f(x)\,f(y)\,g(-xy) \tag{2.31}$$

for all $x\,y \in \mathbf{R}$, which is the same as (FE). Hence from Theorem 1, we get

$$f(x) = c\,e^{\frac{1}{2}A_1(x^2)+A_2(x)} \tag{2.32}$$

and

$$g(x) = \frac{1}{c}\,e^{A_1(-x)}. \tag{2.33}$$

Since f is even, (2.32) yields $A_2 \equiv 0$. Thus (2.32) and (2.33) yield (2.29). This completes the proof.

Remark 4. Note (2.31) is a special case of (GFE). Use Corollary 4 to obtain (2.29).

Finally (FE) can be used to solve another functional equation

$$f_1(x-y) = f_2(x)\,f_3(y)\,g(xy), \qquad x,\, y \in \mathbf{R}, \tag{GSE}$$

where f_1, f_2, f_3, and g are not identically zero functions. We will use this equation in the main theorem in the next section.

COROLLARY 6. *Let* f_1, f_2, f_3, $g : \mathbf{R} \to \mathbf{R}$ *satisfy the functional equation* (GSE). *Then the general nontrivial solution of* (GSE) *is given by*

$$\left.\begin{aligned} f_1(x) &= a\,b\,c\,e^{\frac{1}{2}A_1(x^2)+A_2(x)} \\ f_2(x) &= \frac{1}{a}\,f_1(x) \\ f_3(x) &= \frac{1}{b}\,f_1(-x) \\ g(x) &= \frac{1}{c}\,e^{A_1(-x)} \end{aligned}\right\} \qquad x \in \mathbf{R},$$

where $A_i : \mathbf{R} \to \mathbf{R}$ $(i = 1, 2)$ *are additive functions and* a, b, c *are arbitrary nonzero real constants.*

Proof. As before we can show that none of f_1, f_2, f_3, g is anywhere zero. Putting $y = 0$ in (GSE) and then $x = 0$, we obtain $f_1(x) = a\,f_2(x)$ and $f_1(y) = b\,f_3(-y)$ so that (GSE) can be rewritten as

$$\frac{1}{a\,b}\,f_1(x-y) = \frac{1}{a\,b}\,f_1(x)\,\frac{1}{a\,b}\,f_1(-y)\,g(xy).$$

Now changing y to $-y$ in the above equation, we see that it reduces to (FE) (also to (GFE) and then use Corollary 4).

3. The main result. We prove our main result by first deriving the equation (FE).

THEOREM 7. *If $G : \mathbf{R}^2 \to \mathbf{R}$ is rotation invariant separable function, then G is either identically zero or Gaussian.*

Proof. If G is the zero function, it is obviously rotation invariant and separable. So, we assume G is not identically zero. Suppose $G(x_o, y_o) = 0$. Then by using separability and rotation invariance, we can show that $G(x, y) = 0$ for all x, $y \in \mathbf{R}$. The separability condition (1.3) gives $g(x_o)\,k(y_o) = 0$. Assume $g(x_o) = 0$. Then again (1.3) gives $G(x_o, y) = 0$ for all y. Given any (x_1, y_1), it is possible to find a suitable (x_o, y) or a rotation of it, so that (x_1, y_1) is obtained as the image of (x_o, y) by a suitable rotation. Hence, $G(x_1, y_1) = 0$. So, G is nowhere zero. Then from (1.5) we obtain

$$k(\mu x + \nu y)\, l(-\nu x + \mu y) = g(x)\, h(y) \quad \text{for all } x, y \in \mathbf{R} \tag{3.1}$$

for all (μ, ν) not equal to $(0,1)$, $(1,0)$, $(0,-1)$, $(-1,0)$, where $\mu = \cos\theta$ and $\nu = \sin\theta$. Since G is nowhere zero we conclude that k, l, g, h are also nowhere zero. Letting $x = 0$ in (3.1) we get

$$h(y) = a\, k(\nu y)\, l(\mu y), \tag{3.2}$$

where a is a nonzero constant. Similarly, letting $y = 0$ in (3.1) we get

$$g(x) = b\, k(\mu x)\, l(-\nu x), \tag{3.3}$$

where b is a nonzero constant. Using (3.2) and (3.3) in (3.1), we obtain

$$k(\mu x + \nu y)\, l(-\nu x + \mu y) = a\, b\, k(\mu x)\, l(-\nu x)\, k(\nu y)\, l(\mu y),$$

which is

$$\frac{k(\mu x + \nu y)}{k(\mu x)\, k(\nu y)} = a\, b\, \frac{l(-\nu x)\, l(\mu y)}{l(-\nu x + \mu y)}. \tag{3.4}$$

Replacing x by $\frac{1}{\mu}\, x$ and y by $\frac{1}{\nu}\, y$ in (3.4), we get

$$\frac{k(x + y)}{k(x)\, k(y)} = a\, b\, \frac{l(-\frac{\nu}{\mu}\, x)\, l(\frac{\mu}{\nu}\, y)}{l(-\frac{\nu}{\mu}\, x + \frac{\mu}{\nu}\, y)}. \tag{3.5}$$

Letting $\frac{\mu}{\nu} = x$ any nonzero real (recall $\frac{\mu}{\nu} = \cot\theta$, it is possible to choose $\frac{\mu}{\nu} = x$, any nonzero real) in (3.5), we see that the right side of (3.5) is a function of xy. Hence we have

$$k(x + y) = k(x)\, k(y)\, \phi(xy) \tag{3.6}$$

for all x, $y \in \mathbf{R}$ with $x \neq 0$. By Theorem 1 and Remark 2, we obtain

$$\begin{aligned} k(x) &= c\, e^{\frac{1}{2}\, A_1(x^2) + A_2(x)}, \\ \phi(x) &= \frac{1}{c}\, e^{A_1(x)}. \end{aligned} \tag{3.7}$$

Letting $(\mu, \nu) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ in (3.5), we get

$$\frac{k(x + y)}{k(x)\, k(y)} = a\, b\, \frac{l(-x)\, l(y)}{l(-x + y)}. \tag{3.8}$$

Using (3.7) in (3.8), we see that

$$l(-x+y) = a\,b\,c\,l(-x)\,l(y)\,e^{-A_1(xy)}.$$

Thus, by Corollary 6 (could use Corollary 4 and Corollary 5), we have

$$l(x) = \frac{1}{a\,b\,c}\,e^{\frac{1}{2}A_1(x^2)+A_3(x)}. \tag{3.9}$$

Using (3.7), (3.9), (1.4), (1.5), and (3.1), we obtain

$$G(x,y) = k(\mu x+\nu y)\,l(-\nu x+\mu y) = \frac{1}{a\,b}\,e^{\frac{1}{2}A_1(x^2+y^2)+A_2(\mu x+\nu y)+A_3(-\nu x+\mu y)}. \tag{3.10}$$

Since G is rotation invariant for all $\theta \in \Omega$, G in (3.10) remains the same for different choices of (μ,ν). By choosing $(\mu,\nu) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, we get

$$G(x,y) = \frac{1}{a\,b}\exp\left(\frac{1}{2}A_1(x^2+y^2) + A_2\left(\frac{x+y}{\sqrt{2}}\right) + A_3\left(\frac{-x+y}{\sqrt{2}}\right)\right). \tag{3.11}$$

Similarly, choosing $(\mu,\nu) = (-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$, we see that (3.10) becomes

$$G(x,y) = \frac{1}{a\,b}\exp\left(\frac{1}{2}A_1(x^2+y^2) + A_2\left(\frac{-x-y}{\sqrt{2}}\right) + A_3\left(\frac{x-y}{\sqrt{2}}\right)\right). \tag{3.12}$$

Hence from (3.11) and (3.12), we obtain

$$A_2\left(\frac{x+y}{\sqrt{2}}\right) = A_3\left(\frac{x-y}{\sqrt{2}}\right) \tag{3.13}$$

for all x, $y \in \mathbf{R}$. Now from (3.11) we have

$$G(x,y) = \frac{1}{a\,b}\,e^{\frac{1}{2}A_1(x^2+y^2)}.$$

This completes the proof of the theorem.

We conclude this paper with the following remark. A function $G : \mathbf{R}^2 \to \mathbf{R}$ is circularly symmetric if $G(x,y) = f(\sqrt{x^2+y^2})$ for some real valued function f. It was shown in [5] that circularly symmetric separable functions are Gaussian. Circularly symmetric functions are precisely rotation invariant; and not the converse. In harmonic analysis, such functions are generally called radial functions [3]. Thus our Theorem 7 can also be deduced from [5]. However, here we accomplished our task by solving a quite different and interesting functional equation (FE) and the related ones considered by others in different contexts [4], [6]. Furthermore, this approach allows us to study linear transformation invariant separable functions. For similar results, refer to [7], [8].

Acknowledgments. We are thankful to the referee for suggestions that improved the presentation of this paper.

REFERENCES

[1] J. Aczel and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, Cambridge, 1989.

[2] M. Hashimoto and J. Sklansky, *Multiple-order derivatives for detecting local image characteristics*, Comput. Vision, Graphics & Image Process., 39 (1987), pp. 28–55.

[3] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis*, Springer-Verlag OHG, Berlin, 1963.

[4] A. Mann, M. Revzen, F. C. Khanna, and Y. Takahashi, *Bifactorizable wavefunctions*, J. Phys. A., 24 (1991), pp. 425–431.

[5] P. K. Sahoo, *Circularly symmetric separable functions are Gaussian*, Appl. Math. Lett., 3 (1990), pp. 111–113.

[6] H. Swiatak, *On a class of functional equations with several unknown functions*, Aequationes Math., 12 (1975), pp. 39–64.

[7] J. A. Baker, *On the functional equation* $f(x)\,g(y) = \prod_{i=1}^{n} h_i(a_i x + b_i y)$, Aequationes Math., 11 (1974), pp. 154–162.

[8] R. L. Hall, *On single-product functions with rotational symmetry*, Aequationes Math., 8 (1972), pp. 281–286.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1352–1355, September 1992

018

GLOBAL LINEAR INDEPENDENCE AND FINITELY SUPPORTED DUAL BASIS*

KANG ZHAO†

Abstract. Relating the global linear independence of an arbitrary locally finite family of functions to the basic fact that a linear map is onto if and only if its adjoint is 1-1, it is shown that any locally finite and globally linearly independent family of functions has a dual basis consisting of finitely supported linear functionals.

Key words. global linear independence, locally finite, finitely supported dual basis

AMS(MOS) subject classifications. 41A05, 41A15, 41A63

1. Introduction. Call the (indexed) family Φ of distributions on $\mathbb{R}^s$ *locally finite* in case, for any *test function*, i.e., any compactly supported C^∞-function f, $\langle \varphi, f\rangle = 0$ for all but finitely many $\varphi \in \Phi$. For such a locally finite family, the sum

$$\Phi c := \sum_{\varphi\in\Phi} \varphi c(\varphi) \tag{1.1}$$

is a well-defined distribution, regardless of what the coefficient "sequence" $c \in \mathbf{C}^\Phi$ might be, if we define the sum *pointwise*, i.e., set

$$\langle \Phi c, f\rangle := \left\langle \sum_{\varphi\in\Phi} \varphi c(\varphi), f \right\rangle := \sum_{\varphi\in\Phi} c(\varphi)\langle\varphi, f\rangle.$$

We assume that no confusion will arise from the use of the symbol Φ for both the indexed family and the linear map (1.1) naturally associated with it. Call such a locally finite family Φ *globally linearly independent* in case $\Phi c = 0$ implies that $c = 0$. In such a case, there is a *dual basis* for Φ, i.e., a corresponding collection $(\lambda_\varphi)_{\varphi\in\Phi}$ of linear functionals on $\operatorname{ran}\Phi$, given by the rule that

$$\langle \lambda_\varphi, \Phi c\rangle = c(\varphi).$$

Therefore $(\lambda_\varphi)_{\varphi\in\Phi}$ satisfies that

$$\forall\, \psi \in \Phi, \quad \langle \lambda_\varphi, \psi\rangle = \delta_\varphi(\psi) := \begin{cases} 1 & \text{if } \psi = \varphi; \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

Recently, Ben-Artzi and Ron [BR] published the surprising and useful result that if Φ is locally finite, then the elements of this dual basis are *local*, in the sense that, for each $\varphi \in \Phi$, there exists a ball $B = B_\varphi$ (of finite radius) so that $\langle \Phi c, f\rangle = 0$ for all f with $\operatorname{supp} f \subset B$ implies that $c(\varphi) = 0$.

*Received by the editors June 18, 1991; accepted for publication (in revised form) January 9, 1992.

†Mathematics Department, Univeristy of Wisconsin, Madison, Wisconsin 53706. The work of this author was supported by a research assistantship from the National Science Foundation grant DMS-9000053.

It is the purpose of this paper to show that this result is a direct consequence of the basic fact that the adjoint of a linear map is 1-1 if and only if the map itself is onto. By this fact, the above-mentioned result of [BR] is improved and generalized to the case that Φ is a locally finite family of functions defined on *any* domain X. Their result is recovered by considering distributions as functions on the collection of test functions.

2. The results. In linear algebra, an infinite subset of a linear space is called algebraically linearly independent if any finite subset of it is linearly independent. In approximation theory, a locally finite family Φ of functions defined on $\mathbb{R}^s$ is called globally linearly independent if

$$\sum_{\varphi\in\Phi} \varphi c(\varphi) = 0 \quad \Longrightarrow \quad c = 0.$$

In other words, the (indexed) family Φ is globally linearly independent if and only if the corresponding *map* $\Phi : c \mapsto \sum_{\varphi\in\Phi} \varphi c(\varphi)$ is 1-1. In this section, we derive a condition equivalent to global linear independence of Φ, which gives a geometric meaning to the global linear independence of Φ. We rely on the following basic fact.

PROPOSITION 2.1 ([T, p. 52]). *Let U and V be two linear spaces, with duals U' and V', respectively, and let $M : U \to V$ be a linear map. Then M is onto if and only if the adjoint $M' : V' \to U' : \lambda \mapsto \lambda M$ of M is 1-1.*

Let $\mathbb{F}$ be a field. Denote by $\mathbb{F}^X$ the linear space of all functions from X to $\mathbb{F}$ and by $\mathbb{F}_0^X$ the subspace of $\mathbb{F}^X$ of all finitely supported functions. Here, we say that $g \in \mathbb{F}^X$ is *finitely supported* if the *support* of g, i.e., the set

$$\operatorname{supp} g := \{x \in X : g(x) \neq 0\}$$

is finite. We want to *stress* that we make no assumptions concerning the domain X. In particular, the case that Φ is a collection of distributions in $\mathcal{D}'(\mathbb{R}^s)$ is covered because, for this case, X is the space of all test functions.

If X is finite, then $\mathbb{F}^X$ is known to be its own dual space. For arbitrary X, the following is well known.

PROPOSITION 2.2. *For any domain X, $\mathbb{F}^X$ represents the (algebraic) dual of $\mathbb{F}_0^X$ with respect to the natural pairing $\mathbb{F}_0^X \times \mathbb{F}^X \to \mathbb{F} : (\lambda, g) \mapsto \sum_{x\in X} \lambda(x)g(x)$.*

Let Φ be a *locally finite* family in $\mathbb{F}^X$, i.e., only finitely many entries of Φ are nonzero at any particular $x \in X$. Then the corresponding linear map Φ is well defined. If we think of it as a matrix, its rows indexed by $x \in X$ and its columns indexed by $\varphi \in \Phi$, then its "transposed" is the "matrix" $\Psi := (\varphi(x))_{\varphi\in\Phi;x\in X}$, with its rows indexed by $\varphi \in \Phi$ and its columns indexed by $x \in X$. We cannot hope to apply this "matrix" to an arbitrary $g \in \mathbb{F}^X$. But we can apply it to any $g \in \mathbb{F}_0^X$, since, for any such g,

$$\#\{\varphi \in \Phi : \operatorname{supp}\varphi \cap \operatorname{supp} g \neq \emptyset\} < \infty.$$

Therefore,

$$\Psi : \mathbb{F}_0^X \to \mathbb{F}_0^\Phi : g \mapsto \sum_{x\in\operatorname{supp} g} \Phi_x g(x)$$

is a well-defined linear map, where

$$\Phi_x : \Phi \to \mathbb{F} : \varphi \mapsto \varphi(x)$$

is the "x-column" of Ψ; hence

$$\Psi g : \varphi \mapsto \sum_{x \in \operatorname{supp} g} \varphi(x) g(x).$$

For any $c \in \mathbb{F}^{\Phi}$ and any $g \in \mathbb{F}_0^X$,

$$\langle \Psi' c, g\rangle = \langle c, \Psi g\rangle = \sum_{\varphi \in \Phi} c(\varphi) \left(\sum_{x \in \operatorname{supp} g} \varphi(x) g(x) \right) = \sum_{x \in \operatorname{supp} g} \sum_{\varphi \in \Phi} \varphi(x) c(\varphi)\, g(x).$$

This proves that

$$\Psi' : \mathbb{F}^{\Phi} \to \mathbb{F}^X : c \mapsto \Phi c.$$

Consequently, by Proposition 2.1, Φ is 1-1 if and only if Ψ is onto. Now note that, since Φ is locally finite, $\operatorname{ran} \Psi = \operatorname{span} \Phi_X$, with

$$\Phi_X := \{\Phi_x : x \in X\}.$$

Thus we obtain the following geometric description of the global linear independence of Φ.

THEOREM 2.3. *A locally finite family Φ in $\mathbb{F}^X$ is globally linearly independent if and only if*

$$\operatorname{span} \Phi_X = \mathbb{F}_0^{\Phi},$$

with $\Phi_X = \{\Phi_x : x \in X\}$ and $\Phi_x : \Phi \to \mathbb{F} : \varphi \mapsto \varphi(x)$.

Since $\mathbb{F}_0^{\Phi}$ contains, in particular, the "unit sequence" δ_φ (defined in (1.2)), the following corollary is immediate.

COROLLARY 2.4. *If the locally finite family Φ in $\mathbb{F}^X$ is globally linearly independent, then, for each $\varphi \in \Phi$, there exists $h_\varphi \in \mathbb{F}_0^X$ so that*

$$\sum_{x \in X} h_\varphi(x) \psi(x) = \delta_\varphi(\psi), \qquad \psi \in \Phi.$$

This shows that the dual basis $(\lambda_\varphi)_{\varphi \in \Phi}$ for Φ is *finitely supported* in the sense that each λ_φ has a representation in the form

$$\lambda_\varphi : \operatorname{ran} \Phi \to \mathbb{F} : g \mapsto \sum_{x \in X} h_\varphi(x) g(x)$$

for some *finitely supported* h_φ.

THEOREM 2.5 [BR]. *Let Φ be a collection of distributions in $\mathcal{D}'(\mathbb{R}^s)$. If Φ is locally finite and globally linearly independent, then the elements of its dual basis are local.*

Proof. The local finiteness of the family Φ of distributions implies that it is also locally finite as a family of functions on the set X of all test functions. Therefore, with this choice for X, for each $\varphi \in \Phi$, we can find $h_\varphi \in \mathbb{F}_0^{\Phi}$ so that $\sum_{x \in X} h_\varphi(x) \psi(x) = \delta_\varphi(\psi)$. It follows that

$$B := \cup_{x \in \operatorname{supp} h_\varphi} \operatorname{supp} x$$

is a bounded set. Also, if $\langle \Phi c, f\rangle = 0$ for all f with $\operatorname{supp} f \subset B$, then, in particular, $\langle \Phi c, x\rangle = 0$ for all $x \in \operatorname{supp} h_\varphi$; therefore,

$$0 = \sum_{x \in X} h_\varphi(x) \langle \Phi c, x\rangle = c(\varphi). \qquad \square$$

In the following, we compare the global linear independence with the algebraic linear independence of Φ geometrically. By Theorem 2.3, a locally finite Φ is globally linearly dependent if and only if $\operatorname{span}\Phi_X$ is a proper subspace of $\mathbb{F}_0^\Phi$. We claim that if Φ is algebraically linearly independent, then $\operatorname{span}\Phi_X$ cannot be too *small.*

Equip $\mathbb{F}^\Phi$ with the topology of pointwise convergence as follows. For each $\varphi \in \Phi$, let

$$p_\varphi(c) := |c(\varphi)| \quad \forall\, c \in \mathbb{F}^\Phi.$$

It is clear that $\{p_\varphi : \varphi \in \Phi\}$ is a separating family of seminorms on the linear space $\mathbb{F}^\Phi$. By [R, Thm. 1.37], this family produces a locally convex linear topology for $\mathbb{F}^\Phi$. With this topology, for $c_k \in \mathbb{F}^\Phi$, c_k converges to $c \in \mathbb{F}^\Phi$ as $k \to \infty$ if and only if for each $\varphi \in \Phi$,

$$\lim_{k\to\infty} c_k(\varphi) = c(\varphi).$$

We can verify that $\mathbb{F}_0^\Phi$ is the topological dual of $\mathbb{F}^\Phi$. Furthermore, $\mathbb{F}_0^\Phi$ is a dense subspace of $\mathbb{F}^\Phi$. Thus, $\mathbb{F}_0^\Phi$ is its own topological dual.

PROPOSITION 2.6. *A locally finite $\Phi \subset \mathbb{F}^X$ is algebraically linearly independent if and only if* $\operatorname{span}\Phi_X$ *is dense in* $\mathbb{F}_0^\Phi$.

Proof. The span of Φ_X is not dense in $\mathbb{F}_0^\Phi$ if and only if there exists a nontrivial continuous linear functional $\lambda \in (\mathbb{F}_0^\Phi)^* = \mathbb{F}_0^\Phi$, which vanishes on $\operatorname{span}\Phi_X$, i.e., if and only if

$$\sum_{\varphi \in \operatorname{supp}\lambda} \varphi(x)\lambda(\varphi) = 0 \quad \forall\, x \in X,$$

for some $\lambda \in \mathbb{F}_0^\Phi$, and this holds if and only if Φ is (algebraically) linearly dependent. □

Acknowledgment. The presentation of this paper reflects discussions with Carl de Boor.

REFERENCES

[BR] A. BEN-ARTZI AND A. RON, *On the integer translates of a compactly supported function: dual bases and linear projectors*, SIAM J. Math. Anal., 21 (1990), pp. 1550–1562.

[R] W. RUDIN, *Functional Analysis*, McGraw-Hill Book Company, New York, 1973.

[T] A. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

SIAM J. MATH. ANAL.
Vol. 23, No. 5, pp. 1356–1374, September 1992

019

LOCALIZED GALERKIN ESTIMATES FOR BOUNDARY INTEGRAL EQUATIONS ON LIPSCHITZ DOMAINS*

V. ADOLFSSON†, M. GOLDBERG‡, B. JAWERTH§, AND H. LENNERSTAD¶

Abstract. The Galerkin method is studied for solving the boundary integral equations associated with the Laplace operator on nonsmooth domains. Convergence is established with a condition on the meshsize, which involves the local curvature on certain approximating domains. Error estimates are also proved, and the results are generalized to systems of equations.

Key words. Galerkin methods, nonsmooth, Lipschitz, double layer potential, systems of equations

AMS(MOS) subject classifications. primary 65N30; secondary 65R20, 65N15

1. Introduction. The Dirichlet and Neumann problems for the Laplace operator in a Lipschitz domain $\Omega \subset \mathbb{R}^n$ are of fundamental interest, both from a theoretical and applied point of view. It is well known that these boundary value problems can be solved by using the boundary integral method, involving the appropriate layer potentials. The basic facts concerning this approach are contained in the papers by Dahlberg [D], Jerison and Kenig [JK], Verchota [V], and Dahlberg and Kenig [DK].

To be able to use the boundary integral method on Lipschitz domains numerically, we must first study the solvability properties of discretized versions of the layer potentials in nonsmooth domains. The study of the theoretical foundation for such a numerical implementation was initiated by Dahlberg and Verchota [DV]. They established the convergence of the Galerkin method for solving the integral equations associated with layer potentials of the Laplace operator on Lipschitz domains. In particular, they proved the optimal L^p-solvability of the Galerkin procedure for the Neumann and Dirichlet problems in Lipschitz domains. A first step in their approach is to approximate the domain Ω by appropriate smooth domains Ω_h, $h > 0$, so that, among other things, the Ω_h's have uniformly bounded Lipschitz constant as $h \to 0$. On each Ω_h it is then required that there is a finite-dimensional subset X_h of the bounded functions on $\partial\Omega_h$ satisfying certain quite general, but technical, conditions. The L^p convergence of the Galerkin procedure is then proved under the condition that $\lim_{h\to 0} \rho(h)\kappa(h) = 0$, where $\rho(h)$ corresponds to a uniform meshsize on $\partial\Omega_h$, and κ_h denotes the maximum curvature of $\partial\Omega_h$. So, roughly speaking, the meshsize is determined by the maximum curvature of the approximating domains. (More detail and precise conditions will be presented below.)

The main purpose of this paper is to extend this result in the following respects.

*Received by the editors April 14, 1991; accepted for publication January 6, 1992. The authors were partially supported by ONR grant N00014-90-J-1343.

†Woods Hole Oceanographic Institute, Woods Hole, Massachusetts 02543.

‡157E Oakland Avenue, Columbus, Ohio 43201.

§Department of Mathematics, University of South Carolina, Columbia, South Carolina 29208. This author was also partially supported by National Science Foundation grant DMS 8803585 and Air Force Office of Scientific Research grant 89-0455.

¶Department of Mathematics, Chalmers University of Technology, S-412 96 Goteborg, Sweden.

(1) We will show that we may allow variable meshsize and still obtain the convergence of the Galerkin method. The product of the meshsize and a measure of the local curvature on Ω_h is required to go to zero uniformly.

(2) Let E_h be the error between the solution of the boundary integral equation and the solution of the corresponding discretized equation. We shall prove that we have the error estimate

$$\|E_h\|_{L^p} \le ch\,\|\nabla f\|_{L^p}\,, \qquad 2-\epsilon_M < p < \infty,$$

as long as f is in the Sobolev space L^p_1.

(3) We will also establish corresponding results for certain systems of equations. In particular, we will consider the Dirichlet problem in elastostatics, Stokes' system of hydrostatics, and the boundary value problem modeling the stresses exerted on a body inserted in a Stokes flow.

In the interest of brevity we shall only consider in detail the Dirichlet problem for an unbounded domain $\Omega = \{(x,y) : x \in \mathbb{R}^{n-1},\ y > \Phi(x)\}$, where $\Phi : \mathbb{R}^{n-1} \to \mathbb{R}$ is a Lipschitz function with Lipschitz constant M. Such a domain is usually called a Lipschitz graph domain. The Neumann problem and the bounded case require some minor modifications for which we refer to [DK].

A brief outline of the content of this paper is as follows.

In §2 we write down the precise conditions that the finite element spaces must satisfy. We also make some related, simple observations. In particular, we show that if the domain Ω is sufficiently nice, then we may, at least theoretically, work directly with Ω and avoid the approximating domains Ω_h.

In §3 we give the criteria that the approximating spaces Ω_h must satisfy and establish the convergence of the Galerkin procedure. We assume that we are given approximating domains Ω_h, $h > 0$, and a finite element space X_h on each of these. We then show the invertibility of $\Pi_h A_h$ on X_h, uniformly in $h > 0$, and the convergence in the L^p-norm of the approximate solutions $f_h \in X_h$, solving $\Pi_h A_h f_h = \Pi_h g$, to the solution f of $Af = g$.

The conditions that the finite element spaces X_h, $h > 0$ have to satisfy are not entirely simple. In §4 we discuss a simple and efficient way to construct them.

For the Galerkin method to be of any real practical use, we need some error estimates. This is the subject of §5. As we mentioned above, we prove that if the function and its first derivatives are in L^p, i.e., $f \in L^p_1$, and if the meshsize is $\mathcal{O}(h)$, then the rate of convergence of the approximate solutions f_h to the exact solution f is at least $\mathcal{O}(h)$.

In §6 we discuss systems of equations in Lipschitz domains and the modifications of the scalar case necessary to use the Galerkin method for systems.

2. Finite element spaces.

Let us start by reviewing the boundary integral method in the case of the Dirichlet problem in a nonsmooth domain Ω. Let $\Phi : \mathbb{R}^{n-1} \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant M, and let $\Omega = \{(x,y) : x \in \mathbb{R}^{n-1}, y > \Phi(x)\}$. For f defined on $\partial\Omega$, and $P \in \Omega$, the double layer potential of f is defined by

$$\mathcal{D}f(P) = \mathcal{D}_\Omega f(P) = \frac{1}{\omega_n}\int_{\partial\Omega} f(Q)\frac{\langle P-Q, n(Q)\rangle}{|P-Q|^n}\,d\sigma, \tag{2.1}$$

where $d\sigma$ denotes the surface measure of $\partial\Omega$, ω_n is the surface area of the unit sphere, and $n(Q)$ the inward unit normal to $\partial\Omega$. By formally allowing $P \in \partial\Omega$ in the definition of $\mathcal{D}f$, we obtain a principal value operator that we denote by T_Ω or just T. In local coordinates, the kernel of T is

$$k(x,y) = \frac{1}{\omega_n} \frac{\Phi(x) - \Phi(y) - \nabla\Phi(y) \cdot (x-y)}{(|x-y|^2 + (\Phi(x) - \Phi(y))^2)^{n/2}}. \tag{2.2}$$

The boundary values of $\mathcal{D}f$ from the interior of Ω are given by $Af = (\frac{1}{2}I + T)f$. The boundary integral method uses $\mathcal{D}(A^{-1}g)$ to solve the Dirichlet problem in Ω with boundary values g.

To apply this method two basic facts must be established. First we need to prove the boundedness of A. In the case of Lipschitz domains, this follows from the L^p boundedness of the Cauchy integral on Lipschitz curves, cf. [CMM] and [C], and the method of rotations. We thus have

$$\|Af\|_p \leq C_p \|f\|_p \tag{2.3}$$

for $1 < p < \infty$. We must then prove the invertibility of A. In the case of smooth domains it is easy to see that $A = \frac{1}{2}I + T$ with T compact, which means that the invertibility follows from Fredholm theory. In the case of general Lipschitz domains, T is not compact and Fredholm theory does not apply. Instead, the invertibility follows as a consequence of certain Rellich type identities and estimates using these identities. It follows that there exists an $\epsilon_M > 0$, depending only on the Lipschitz constant M, such that A is invertible on $L^p(\partial\Omega)$ for $2 - \epsilon_M < p < \infty$, and

$$\|Af\|_p \geq c_p \|f\|_p \tag{2.4}$$

for this range of p (see [DK]).

To discuss a discretized version of this and the Galerkin procedure we need certain finite element function spaces X. We shall assume that the space X satisfies the following conditions.

(1) $X \subset L^\infty(\partial\Omega)$, X closed under uniform convergence.

(2) There are pairwise disjoint sets $E_j \subset \partial\Omega$ whose union equals $\partial\Omega$, points $P_j \in E_j$, and positive numbers ρ_j, K, c_0, and ϵ_0 such that each point in $\partial\Omega$ lies in at most K of the sets $B_j = B(P_j, \rho_j) \cap \partial\Omega$, $\rho_j > \epsilon_0$,

$$E_j \subset B_j, \tag{2.5}$$

and

$$\|f\|_{L^\infty(E_j)} \leq \frac{c_0}{\sigma(B_j)} \sup \left| \int_{\partial\Omega} fw \, d\sigma \right|, \tag{2.6}$$

where the supremum is taken over all $w \in X$ with $\|w\|_\infty \leq 1$ and w supported in B_j. Here, $B(P,r)$ denotes the ball in $\mathbb{R}^n$ with center P and radius r.

(3) All constant vectors belong to X.

We shall call such an X a localized space with variable scale. We say that the covering $\{B_j\}$ has the finite intersection property, and we will sometimes refer to (2.6) as the localization property.

The conditions on X imply the existence of a projection operator Π.

PROPOSITION 2.1. *With X and $\partial\Omega$, set $X^p = L^p(\partial\Omega) \cap X$. Then X^p is a closed subspace of $L^p(\partial\Omega)$ for $p \in [1,\infty]$. If Π denotes the L^2 projection onto X^2, then Π has a bounded extension $\Pi : L^p(\partial\Omega) \to X^p$ with $\|\Pi f\|_p \le C\,\|f\|_p$, $1 \le p \le \infty$, with C independent of f.*

Proof. The localization property implies that if $f \in X$, then for $x \in E_j$

$$|f(x)| \le \frac{c_0}{\sigma(B_j)} \int_{B_j} |f|\, d\sigma \le C\rho_j^{-(n-1)/p}\, \|f\|_{L^p(B_j)} \le C\epsilon_0^{-(n-1)/p}\, \|f\|_{L^p(B_j)}\,.$$

From this estimate it follows that if a sequence in X^p converges in the L^p sense, it also converges in the maximum norm. X is closed under uniform convergence, so X^p is a closed subspace of $L^p(\partial\Omega)$. Similarly, if $f \in L^p(\partial\Omega) \cap L^2(\partial\Omega)$ and $w \in X$, supported on B_j, satisfies $\|w\|_\infty \le 1$, then for $x \in E_j$

$$|\Pi f(x)| \le \frac{c_0}{\sigma(B_j)} \left| \int w \Pi f\, d\sigma \right| = \frac{c_0}{\sigma(B_j)} \left| \int w f\, d\sigma \right| \le C\rho_j^{-(n-1)/p}\, \|f\|_{L^p(B_j)}\,.$$

Using the variable scale localization property, we now have that

$$\begin{aligned} \int |\Pi f|^p\, d\sigma &= \sum \int_{E_j} |\Pi f|^p\, d\sigma \le C \sum \rho_j^{n-1}\, \|\Pi f\|^p_{L^\infty(E_j)} \\ &\le C \sum \int_{B_j} |f|^p\, d\sigma \le C \int |f|^p\, d\sigma. \end{aligned}$$

This is the step where we use the finite intersection property of the B_j's. The proof is complete. □

When the domain is sufficiently smooth and bounded, then, as we have remarked earlier, the operator T is compact (cf. [FJR]). Hence, the following elementary observation proves the solvability of the Galerkin equations for a family of finite element spaces X_h, $h > 0$, and corresponding projections Π_h, on such domains. Note, however, that we obtain no estimates on the norms involved.

PROPOSITION 2.2. *Let $A = \frac{1}{2}I + T$, as before, and suppose that Ω is Lipschitz. Suppose $T : L^2(\partial\Omega) \to L^2(\partial\Omega)$ is compact. Let X_h, $h > 0$ be a family of closed subspaces of $L^2(\partial\Omega)$, and let Π_h be orthogonal projections of $L^2(\partial\Omega)$ onto X_h. Further, suppose $\|\Pi_h f - f\|_2 \longrightarrow 0$ as $h \longrightarrow 0$ for $f \in L^2(\partial\Omega)$. Then, there exist $h_0, c_0 > 0$ such that if $0 < h \le h_0$,*

$$\|\Pi_h A \Pi_h f\|_2 \ge c_0\, \|\Pi_h f\|_2 \quad \textit{for } f \in L^2(\partial\Omega), \tag{2.7}$$

and $\Pi_h A : X_h \to X_h$ is invertible.

Proof. Suppose this is not so. Then, there is a sequence $f_k \in L^2(\partial\Omega)$ and $\{h_k\}$, $h_k \longrightarrow 0^+$, with $\|\Pi_{h_k} f_k\|_2 = 1$, and

$$\|\Pi_{h_k} A \Pi_{h_k} f_k\|_2 \le \frac{1}{k}. \tag{2.8}$$

Without loss of generality, by compactness, $T\Pi_{h_k} f_k \longrightarrow g \in L^2(\partial\Omega)$. Now, $\Pi_{h_k} A \Pi_{h_k} f = \frac{1}{2}\Pi_{h_k} f_k + \Pi_{h_k} T \Pi_{h_k} f_k$. Since $T\Pi_{h_k} f_k \longrightarrow g$, using our hypothesis, $\Pi_{h_k} T \Pi_{h_k} f_k \longrightarrow g$, and using (2.8), $\frac{1}{2}\Pi_{h_k} f_k \longrightarrow -g$, so $\Pi_{h_k} f_k \longrightarrow -2g$. Note that $g \not\equiv 0$ since $\|\Pi_{h_k} f_k\|_2 = 1$. Since $T\Pi_{h_k} f_k \longrightarrow g$, and $\Pi_{h_k} f_k \longrightarrow -2g$, and T is bounded, we obtain $-2Tg = g$.

This implies $Ag = 0$, i.e., $g \equiv 0$ since A is one to one. This is impossible and (2.7) thus follows.

The inequality (2.7) shows that $\Pi_h A : X_h \to X_h$ is one to one and has closed range. To show that $\Pi_h A$ is onto, suppose $\langle \Pi_h A g, \phi \rangle = 0$ for all $g \in X_h$, for some $\phi \in X_h$. Taking $g = \Pi_h A^* \phi$, and noting that $\phi = \Pi_h \phi$, we obtain $\Pi_h A^* \Pi_h \phi = 0$. Applying the same proof we used to prove (2.7) to A^* and T^*, we obtain $\phi \equiv 0$. This finishes the proof of Proposition 2.2. □

For general Lipschitz domains this result is of little help, since T is not necessarily compact and also because we do not get much information about h_0 and c_0. It is still an open problem as to whether there is an analogous result for Lipschitz domains which would allow us to establish the invertibility with interesting spaces X_h directly on the boundary of the domain.

Following the approach of Dahlberg and Verchota [DV], we shall approximate the Lipschitz domain Ω with smooth domains and prove a uniform invertibility result. This essentially involves quantifying the compactness of T on each of these domains in a suitable way.

Let us fix a domain Ω and a finite element space X on $\partial\Omega$ that is localized with variable scale. We shall say that an operator T satisfies the local approximation property with constant δ if there are $\phi_j \in X$ such that

$$\sum_j \int_{B_j} |Tf - \phi_j|^p \, d\sigma \leq \delta \, \|f\|_p^p \quad \text{if } f \in L^p(\partial\Omega). \tag{2.9}$$

Observe that the existence of *some* δ in (2.9) follows immediately from the boundedness of T and the finite intersection property of the covering $\{B_j\}$, simply by taking $\phi_j \equiv 0$.

Remark 2.3. The local approximation property is closely connected with compactness. Suppose that for each $h > 0$ we are given a decomposition $\partial\Omega = \cup_j B(P_j, \rho(h)) \cap \partial\Omega$ of balls with a uniform size $\rho(h)$ and uniformly bounded overlap. Assume further that the size tends to zero with h, $\lim_{h\to 0} \rho(h) = 0$. Then it is an easy consequence of the Fréchet–Kolmogorov characterization of compact operators on L^p that T is compact if and only if

$$\int_{|x|>R} |Tf(x)|^p \, dx \to 0 \quad \text{as } R \to \infty$$

uniformly in $f \in L^p$, and there are constants c_j such that

$$\sum_j \int_{B(P_j,\rho(h))\cap\partial\Omega} |Tf - c_j|^p \, d\sigma \to 0 \quad \text{as } h \to 0$$

uniformly in $f \in L^p$.

LEMMA 2.4. *Suppose that $\Omega = \{(x,y) : y > \Phi(x)\}$ is a Lipschitz domain with Lipschitz constant M, and that X is a localized space of variable size. Then, for each p, $2 - \epsilon_M < p < \infty$ there is a δ_0 with the following property: if the operator $T = T_\Omega$ satisfies the local approximation property with constant $\delta < \delta_0$, then*

$$\|\Pi A \Pi f\|_p \geq C \, \|\Pi f\|_p \quad \textit{for } f \in L^p(\partial\Omega).$$

Here C and δ_0 only depend on p, the Lipschitz constant M, and the parameters K and c_0 in the definition of X.

Proof. We let $g = \Pi A f$ and write

$$Af = Af - Tf + \phi_j + Tf - \phi_j. \tag{2.10}$$

Now, $Af - Tf + \phi_j = \frac{1}{2}f + \phi_j \in X$; so, by (2.6), for $x \in E_j$,

$$\begin{aligned} |(Af - Tf + \phi_j)(x)| &\le \frac{c_0}{\sigma(B_j)} \sup_w \left| \int_{\partial\Omega} (Af - Tf + \phi_j) w \, d\sigma \right| \\ &\le cg^*(x) + c \left(\int_{B_j} |Tf - \phi_j|^p \frac{1}{\sigma(B_j)} \right)^{1/p}, \end{aligned}$$

where g^* is the Hardy–Littlewood maximal function of g. Hence,

$$\int_{E_j} |Af - Tf + \phi_j|^p \, d\sigma \le c \int_{E_j} (g^*(x))^p + c \int_{B_j} |Tf - \phi_j|^p,$$

and thus, from (2.10),

$$\int_{\partial\Omega} |Af|^p \le c \int_{\partial\Omega} |g|^p + c \sum_j \int_{B_j} |Tf - \phi_j|^p.$$

Now, from (2.4) and (2.9), we obtain

$$(2.11) \qquad (c_p^p - c\delta) \, \|f\|_p^p \le c \, \|g\|_p^p .$$

Thus, if δ is sufficiently small, i.e., so that $c_p^p - c\delta > 0$, we obtain the desired conclusion. □

Remark 2.5. The localization property (2.6) can be replaced, here and also below, by weaker conditions. For instance, let us consider Lemma 2.4 in the special case $p = 2$ and assume a priori that Π is the (bounded) projection of L^2 on X. By a slight modification of the proof above, we may establish the lemma in this special case, assuming instead that

$$(2.12) \qquad \|f\|_{L^2(E_j)} \le c_0 \sup \left| \int_{\partial\Omega} f w \, d\sigma \right|,$$

where the supremum is taken over all $w \in X$ with $\|w\|_2 \le 1$ and w supported in B_j. Suppose, for example, that X is the closure of the linear span of the (real) functions $\{\Phi_j\}$. Then the following is a sufficient condition for (2.12) to hold:

$$(2.13) \qquad \cup_{E_i \cap \operatorname{supp} \Phi_j \ne \emptyset} \operatorname{supp} \Phi_j \subset B_i.$$

To see this, let $f = \sum_j \lambda_i \Phi_i$. Clearly,

$$\|f\|_{L^2(E_j)} \le \left(\int_{\partial\Omega} \left| \sum_{\operatorname{supp} \Phi_i \cap E_j \ne \emptyset} \lambda_i \Phi_i \right|^2 d\sigma \right)^{1/2} = \sup \left| \int_{\partial\Omega} f w \, d\sigma \right|,$$

since $\operatorname{supp} \sum_{\operatorname{supp} \Phi_i \cap E_j \ne \emptyset} \lambda_i \Phi_i \subset B_j$.

3. The Galerkin method on Lipschitz domains. We will next prove the solvability of the Galerkin equation, i.e., show that $\Pi A : X^p \to X^p$ is invertible. In order to prove this, we first recall the following standard continuation lemma (see [DV] or [DK1]).

LEMMA 3.1. *Let Y be a Banach space and suppose $S_t : Y \to Y$ are bounded linear operators for $0 \le t \le 1$. Suppose, furthermore, that there are positive constants c and C such that the following properties hold.*

(i) *For all $t \in [0,1]$, and all $f \in Y$, $\|S_t f\| \ge c\|f\|$.*
(ii) *For all $t, s \in [0,1]$, $\|S_s - S_t\| \le C|s-t|$.*
(iii) *The operator S_0 is invertible on Y.*

Then S_1 is invertible on Y.

Let $A_t = \frac{1}{2}I + T_t$, $0 \le t \le 1$, where the kernel of T_t is given by (2.2) with Φ changed to $t\Phi$. If T_t satisfies the local approximation property with a sufficiently small δ_0, uniformly in t, then the Continuation Lemma applied to ΠA_t, combined with Lemma 2.4, immediately yields the invertibility of ΠA. (By an abuse of notation, we consider A_t acting on functions defined on $\partial\Omega$.) Next we shall give a condition on Ω that guarantees this uniform approximation property.

Let the surface $\partial\Omega = \{(x,y) : y = \Phi(x)\}$ be given by the graph of the smooth function $\Phi : \mathbb{R}^{n-1} \to \mathbb{R}$. For a point $P = (x, \Phi(x))$ on the surface and $\rho > 0$, we define the function κ by

$$\kappa(P,\rho) = \sup_{|x-y|<\rho} \max_{i,j} \left| \frac{\partial^2 \Phi(y)}{\partial y_i \partial y_j} \right|.$$

The function κ is a measure of the maximal curvature near P. This is clear if we recall that the curvature in a direction $\xi \in \mathbb{R}^{n-1}$, $|\xi| = 1$, is given by

$$k(x,\xi) = \frac{D_{\xi\xi}\Phi(x)}{(1 + (D_\xi \Phi(x))^2)^{3/2}},$$

where $D_\xi \Phi$ is the directional derivative of the function Φ in the direction ξ and $D_{\xi\xi}\Phi(x) = \langle \xi, \nabla_2 \Phi \xi \rangle$ with $\nabla_2 \Phi =$ the Hessian of Φ.

THEOREM 3.2. *Suppose that $\Omega = \{(x,y) : y > \Phi(x)\}$ is a smooth domain and that X is a localized space of variable size. For each p, $2 - \epsilon_M < p < \infty$ there is a $\delta_0 > 0$ and a $\gamma > 2$ with the property that if $\sup_j \kappa(P_j, \gamma\rho_j)\gamma\rho_j < \delta_0$, then $\Pi A : X^p \to X^p$ is invertible. In particular, for every $g \in L^p(\partial\Omega)$ there is a unique $f \in X \cap L^p(\partial\Omega)$, such that*

$$\int_{\partial\Omega} Af\, w\, d\sigma = \int_{\partial\Omega} gw\, d\sigma \tag{3.1}$$

for all $w \in X$ with compact support. Here γ and δ_0 only depend on p, the Lipschitz constant M of the function Φ, and the parameters K and c_0 in the definition of X.

Proof. We claim that for every j there is a constant c_j such that for $\gamma > 2$,

$$|Tf(P) - Tf(P_j)| \le c\left(\kappa(P_j, \gamma\rho_j)\gamma\rho_j + \frac{1}{\gamma}\right) f^*(P), \qquad P \in B_j, \tag{3.2}$$

where the constant $c = c(M)$.

This is easy to show. We write $f = a_j + b_j$, where $a_j = f$ on $B(P_j, \gamma\rho_j) \cap \partial\Omega$, and $a_j = 0$ elsewhere, and let $b_j = f - a_j$. Then, by using Taylor's expansion to second order,

$$|Ta_j(P)| \le \kappa(P_j, \gamma\rho_j) \int_{B(P_j,\gamma\rho_j)\cap\partial\Omega} \frac{1}{|P-Q|^{n-2}} |f(Q)|\, d\sigma \le c\kappa(P_j, \gamma\rho_j)\gamma\rho_j f^*(P).$$

The last inequality follows from the standard fact that if q is a nonnegative, radial, decreasing function, then for a function h we have $\int qh \leq h^*(0) \int q$. We now let $c_j = Tb_j$. Using this fact again, we see that

$$|Tb_j(P) - Tb_j(P_j)| \leq c\rho_j \int_{B(P_j,\gamma\rho_j)^c \cap \partial\Omega} \frac{1}{|P_j - Q|^n} |f(Q)|\, d\sigma \leq \frac{c}{\gamma} f^*(P).$$

This proves the claim.

Note that the operators $T = T_t$, $0 \leq t < 1$, satisfy the same inequality (3.2). Hence, the uniform local approximation property, and the theorem, immediately follow by picking γ sufficiently large. □

As in [DV] we approximate a given Lipschitz domain $\Omega = \{(x,y) : y > \Phi(x)\}$ by certain smooth domains Ω_h. These domains must satisfy the following conditions.

(1) $\Omega_h = \{(x,y) : x \in \mathbb{R}^{n-1},\ y > \Phi_h(x)\}$ for some Lipschitz functions Φ_h with a uniform Lipschitz constant M.

(2) dist $\{\partial\Omega_h, \partial\Omega\} \leq Ch$.

(3) $\kappa(h) \leq C(1/h)$, where $\kappa(h)$ denotes the least upper bound of the absolute values of the curvatures of $\partial\Omega_h$.

(4) The mapping $F_h : \partial\Omega \longrightarrow \partial\Omega_h$ defined by $F_h((x, \Phi(x)) = (x, \Phi_h(x))$ is a Lipschitz diffeomorphism such that the Lipschitz constants of F_h and its inverse G_h are uniformly bounded in h.

(5) The mapping $h \longrightarrow \Phi_h(x)$, $0 < h < 1$, is continuous almost everywhere in x, and

$$\lim_{h \downarrow 0} \frac{\Phi_h(x) - \Phi(x)}{h} \text{ exists a.e. in } x.$$

(6) $|F_h(P) - P| \leq Ch$ for all $P \in \partial\Omega$.

(7) The Jacobian of F_h converges almost everywhere to 1.

Note that these conditions are not independent, and that the conditions are satisfied by the following standard regularization. Let $\eta \in C_0^\infty(\mathbb{R}^{n-1})$ be radial, with $\int \eta\, dx = 1$. Set $\eta_h = (1/h^{n-1})\eta(x/h)$ for $h > 0$. Let $\Phi_h(x) = \eta_h * \Phi(x)$ with corresponding domains $\Omega_h = \{(x,y) : y > \Phi_h(x)\}$. Note that the second derivatives of Φ_h (differentiate η_h with one partial, Φ with the second) are bounded by c/h. Φ_h are Lipschitz, with Lipschitz constant bounded by M, and the others of the seven conditions on $\partial\Omega_h$ are satisfied.

Given the domain $\Omega = \{(x,y) : y > \Phi(x)\}$, we let Ω_h, $h > 0$, be such approximating domains, and for each h we let X_h be a finite element space on Ω_h that is localized with variable size, with the constants K, and c_0 independent of h.

If $\rho_{j,h}$ denotes the ρ_j's corresponding to X_h, we let $\rho_h = \max_j \rho_{j,h}$. Suppose now that $\rho_h \longrightarrow 0$ as $h \longrightarrow 0$. Then it is an easy consequence of the localization property that for $q \in L^p(\partial\Omega)$,

$$\left\| \Pi_h(q \circ F_h^{-1}) - q \circ F_h^{-1} \right\|_p \longrightarrow 0 \quad \text{as } h \longrightarrow 0. \tag{3.3}$$

Let A be the operator giving the interior boundary values of the double layer potential on Ω, and A_h the corresponding operators for Ω_h. As a consequence of the conditions on the Ω_h's we also have

$$\left\| Af - A_h(f \circ F_h^{-1}) \circ F_h \right\|_p \longrightarrow 0 \quad \text{as } h \longrightarrow 0. \tag{3.4}$$

The proof of this is similar to that in [DV, Lemma 2.4], so we only give a very brief outline. By a change of variables, $A_h(f \circ F_h^{-1}) \circ F_h = T_hf(P) + \frac{1}{2}f(P)$, where

$$T_hf(P) = T_hf(P) = \frac{1}{\omega_n}\int_{\partial\Omega} \frac{\langle F_h(P) - F_h(Q), n_h(Q)\rangle}{|F_h(P) - F_h(Q)|^n} J_h(Q)f(Q)\,d\sigma. \tag{3.5}$$

Here $n_h(Q)$ is the unit inward normal to $\partial\Omega_h$ at the point $F_h(Q)$, and J_h is the determinant of the Jacobian of F_h. So we need to show T_hf converges to Tf in $L^p(\partial\Omega)$. This is a rather straightforward exercise about singular integrals, using properties of $\partial\Omega_h$. We add and subtract 1 from $n_h(Q)J_h(Q)$. The difference converges to zero in $L^p(\partial\Omega)$ using uniform boundedness of T_h in $L^p(\partial\Omega)$. For the other term, we estimate the difference of this new singular integral and Tf, truncated on a ball of radius rh around x, by looking at the parts close to P, distance at most rh, far away from P, distance at least Rh, and in the middle. The near part has a factor of r coming out, and the far part a factor of $1/R$. For the middle part we use condition (5), satisfied by the approximating domains, to see that it is bounded uniformly by an L^p function and converges pointwise to zero. Letting h go to zero, to estimate the limsup in terms of r and $1/R$, and then letting r go to zero and R go to ∞, finishes the proof of (3.4).

These observations and Theorem 3.2 easily yield the next result. We shall assume that $\Omega = \{(x,y) : y > \Phi(x)\}$ is a Lipschitz domain, and that Φ has Lipschitz constant M. We let $\Omega_h = \{(x,y) : y > \Phi_h(x)\}$, $h > 0$ be approximating domains as defined above, and for each $h > 0$ we let X_h be a finite element space on Ω_h that is localized with variable size, with the constants L, K, and c_0 independent of h.

THEOREM 3.3. *Suppose that* $\lim_{h\to 0}\rho_h = 0$. *Then, given* p, $2 - \epsilon_M < p < \infty$, *there are* $h_0 > 0$ *and* $\gamma > 1$ *with the property that if*

$$\lim_{h\to 0}\kappa_h = \lim_{h\to 0}\sup_j \kappa(P_{j,h}, \gamma\rho_{j,h})\gamma\rho_{j,h} = 0.$$

Then for every $g \in L^p(\partial\Omega)$ *there is a unique* $f_h \in X_h \cap L^p(\partial\Omega_h)$, *for* $h < h_0$, *such that*

$$\int_{\partial\Omega_h} A_hf_h\, w\, d\sigma = \int_{\partial\Omega_h} g_h w\, d\sigma \tag{3.6}$$

for all $w \in X$ *with compact support. Here* $g_h(P) = g(F_h^{-1}(P))$. *Moreover, if* $f \in L^p(\partial\Omega)$ *is defined by* $Af = g$, *then* $f_h \circ F_h$ *converges to* f *in* $L^p(\partial\Omega)$.

Proof. The proof is immediate. Of course, (3.6) follows directly from Theorem 3.2; in fact, it follows that the Π_hA_h's are uniformly invertible. To finish the proof, it remains to show that $f_h \circ F_h$ converge to f in $L^p(\partial\Omega)$, where $Af = g$.

We have $\Pi_hA_hf_h = \Pi_hg_h$, and, since we have assumed that c_0 does not depend on h, the projections Π_h are uniformly bounded. Hence,

$$\begin{aligned}
&\left\|\Pi_hA_hf_h - \Pi_hA_h\Pi_h(f\circ F_h^{-1})\right\|_p = \left\|\Pi_hg_h - \Pi_hA_h\Pi_h(f\circ F_h^{-1})\right\|_p \\
&\qquad \le C\left\|g_h - A_h\Pi_h(f\circ F_h^{-1})\right\|_p = C\left\|g\circ F_h^{-1} - A_h\Pi_h(f\circ F_h^{-1})\right\|_p \\
&\qquad \le C\left\|g - A_h\Pi_h(f\circ F_h^{-1})\circ F_h\right\|_p = C\left\|Af - A_h\Pi_h(f\circ F_h^{-1})\circ F_h\right\|_p.
\end{aligned} \tag{3.7}$$

Now, using (3.3) and the fact that the A_h's are uniformly bounded on $L^p(\partial\Omega_h)$, as well as the uniform invertibility of the F_h, we obtain from (3.7) that

$$\left\|\Pi_hA_hf_h - \Pi_hA_h\Pi_h(f\circ F_h^{-1})\right\|_p \le C\left\|Af - A_h(f\circ F_h^{-1})\circ F_h\right\|_p + o(1). \tag{3.8}$$

Using (3.4) and the uniform invertibility of the $\Pi_h A_h$, (3.8) yields that

$$\left\| f_h - \Pi_h(f \circ F_h^{-1}) \right\|_p \longrightarrow 0. \tag{3.9}$$

Using (3.3) once more, and composing on the right with F_h, we finally get that

$$\| f_h \circ F_h - f \|_p \longrightarrow 0,$$

which establishes the theorem. □

4. An example of a localized space. In this section we shall describe a simple construction of a localized space of variable size on a given smooth domain.

We start in $\mathbb{R}^{n-1}$ and assume that $\Phi : \mathbb{R}^{n-1} \to \mathbb{R}$ is a smooth function with Lipschitz constant M and with $\kappa = \sup_x \max_{i,j} |\partial^2\Phi(x)/\partial x_i \partial x_j| < \infty$. For each cube $Q \subset \mathbb{R}^{n-1}$, we let

$$K(Q) = \ell(Q) \sup_{y \in Q} \max_{i,j} \left| \frac{\partial^2 \Phi(y)}{\partial y_i \partial y_j} \right|.$$

Fix $\epsilon > 0$ and a $\gamma \geq 1$. For each $x \in \mathbb{R}^{n-1}$, we let $Q(x)$ be the largest dyadic cube Q containing x such that $K(\gamma Q) \leq \epsilon$. Here γQ is the cube with the same center as Q, but with sidelength γ as long. Note that there is always such a nonempty finite cube $Q(x)$, and, in fact, $\inf_x \ell(Q(x)) \geq \epsilon_0 > 0$ since $\kappa < \infty$. There are only countably many different such cubes; let us call them Q_i. These cubes have the following properties:

(i) $\mathbb{R}^{n-1} = \cup_i Q_i$;
(ii) The cubes Q_i are pairwise disjoint (modulo sets of measure zero);
(iii) $\ell(Q_i) \geq \epsilon_0$;
(iv) $K(\gamma Q_i) \leq \epsilon$;
(v) For any fixed $\beta \leq \gamma/4$, the cubes βQ_i have bounded overlap.

Of these, only v requires some explanation. Let Q_i and Q_j be two distinct cubes in this cover such that $\frac{\gamma}{4}Q_i \cap \frac{\gamma}{4}Q_j \neq \emptyset$. We claim that Q_i and Q_j must be of approximately the same size. This would clearly prove v.

Let us assume for instance that Q_j is much smaller than Q_i, $\ell(Q_j)/\ell(Q_i) \leq \frac{1}{8}$, say. Simple geometric considerations then show that Q_j is contained in a dyadic cube $\tilde{Q}_j$ with $\ell(\tilde{Q}_j) = 2\ell(Q_j)$ and such that $\gamma\tilde{Q}_j \subset \gamma Q_i$. This implies that $K(\tilde{Q}_j) \leq K(Q_i) \leq \epsilon$, which contradicts the maximality of Q_j.

Now, given the smooth domain $\Omega = \{(x, y) : x \in \mathbb{R}^{n-1}, \Phi(x) > y\}$ with $\kappa < \infty$, we let G_0 be the projection of $\partial\Omega$ on $\mathbb{R}^{n-1}$, and F_0 its inverse so that

$$G_0((x, \Phi(x))) = x, \qquad F_0(x) = (x, \Phi(x)).$$

For any Q_i in the covering of $\mathbb{R}^{n-1}$ that is obtained by our construction, define $E_i = F_0(Q_i)$, $P_j = F_0(x_j)$, where x_j is the center of Q_j, and define ρ_j to be the radius of the ball (in $\mathbb{R}^{n-1}$) inscribed in βQ_j. We can take any $\beta > \delta\sqrt{M^2+1}$ to satisfy $E_i \subset B_i(P_i, \rho_i)$, where δ is a constant depending only on the dimension $n-1$. Note that each $(x, \Phi(x)) \in \partial\Omega$ lies in only finitely many of the $B(P_i, \rho_i)$. Indeed, $(x, \Phi(x)) \in B(P_i, \rho_i)$ implies $x \in G_0(B(P_i, \rho_i))$. But $G_0(B(P_i, \rho_i)) \subset \beta Q_i$, and x is in at most K of these, as noted in our construction.

To satisfy conditions 1, 2, and 3 for a localized variable scale space X, we can, for example, choose X to be the space of polynomials defined on $\mathbb{R}^{n-1}$, of total degree less than or equal to 1 on each Q. See also [Li]. There are many other choices as well. For instance, we can use certain functions satisfying a refinement relation, and, in particular, each of the functions Φ yielding the compactly supported wavelets of Daubechies [Da]. More specifically, let Φ be a compactly supported function defined on $\mathbb{R}$. For $I = [2^{-\nu}k, 2^{-\nu}(k+1)]$ we let $\Phi_I = 2^{\nu/2}\Phi(2^\nu x - k)$. Φ_I is thus supported on $I + 2 - \nu \operatorname{supp} \Phi$. We let $\mathcal{D}_\nu$ denote the set of dyadic intervals with sidelength 2^ν, and $\mathcal{D}$ the union of the $\mathcal{D}_\nu$'s. We assume that Φ satisfies a refinement condition,

$$\Phi(x) = \sum_j c_k \Phi(2x - k).$$

Let S_ν be the span of the functions Φ_I, $I \in \mathcal{D}_\nu$. We also assume that for some integer $r > 0$, S_ν contains the polynomials of degree $< r$. In particular, the Φ_I's form a partition of unity when the length of I is fixed. Finally, for J an interval in $\mathbb{R}$, we let Λ_J denote the set of all j for which $\Phi(\cdot - j)$ is not identically zero on J. We assume that for all $J \in \mathcal{D}$, the functions $\Phi(\cdot - j)$, $j \in \Lambda_J$, are linearly independent over J (i.e., local linear independence). From these assumptions, it follows that if $\Phi_I = \sum_J c_J \Phi_J$, where $\ell(J) < \ell(I)$, then the support of each Φ_J is contained in the support of Φ_I.

Returning to the S_ν's, we see from the refinement property that they form an increasing chain of linear spaces. Each S_ν is associated to a uniform partition of $\mathbb{R}$ into dyadic intervals of length $2^{-\nu}$. Note that even if we have a nonuniform partition of $\mathbb{R}$ into dyadic intervals, with a lower bound on permissible lengths, there is a natural linear space V arising from Φ associated with this partition. The functions of V still retain local linear independence, and further refinements of this partition generate superspaces of V. We will illustrate this association starting with S_0 and perform one subdivision of one of the cubes. The process of associating natural spaces for further subdivisions is analogous. The final space V obtained, which corresponds to a nonuniform partition, will thus be shown to arise in a natural way.

So, let I_0 be a dyadic interval of length 1. We subdivide I_0 into two intervals, J_1 and J_2, of length $\frac{1}{2}$. Let $I_1, I_2, \cdots, I_m$ be the finite number of intervals such that the support of each of $\Phi_{I_1}, \Phi_{I_2}, \cdots, \Phi_{I_m}$ has nonempty intersection with I_0. From the refinement condition, each Φ_{I_k} has a decomposition in terms of intervals of length $\frac{1}{2}$, and for $|I| = 1$, Φ_I has a decomposition containing Φ_{J_1} or Φ_{J_2} only if I is one of $I_1, I_2, \cdots, I_m$. For any I_k such that Φ_{I_k} has Φ_{J_1} or Φ_{J_2} in its decomposition, let $\Phi^*_{I_k}$ be the function obtained by discarding Φ_{J_1} and Φ_{J_2} from the decomposition and summing the other functions. Replace Φ_{I_k} by $\Phi^*_{I_k}$ in S_0 and also adjoin Φ_{J_1} and Φ_{J_2}. The new linear space is the one we choose for the new partition.

In this context, let us make a remark concerning orthogonality. Suppose that the functions Φ_I, $|I| = 1$, in addition, form an orthonormal set in $L^2(\mathbb{R})$. (This is the case for the Φ that yields the compactly supported wavelets.) Then the above construction leads to three mutually orthogonal sets of functions: the functions Φ_I in S_0 that do not contain the Φ_{J_1} or Φ_{J_2}, and, hence, are not changed, the functions Φ_{J_1} and Φ_{J_2}, and, finally, the $\Phi^*_{I_k}$'s. Clearly, the functions in each of the first two sets are also pairwise orthogonal. If we thus want orthogonality also for the functions corresponding to the nonuniform partition, there only remains to orthogonalize the functions $\Phi^*_{I_k}$, which of course can be done in a standard manner.

Using tensor products, it is possible to obtain spaces adapted to nonuniform dyadic decompositions in higher dimensions.

Suppose now that we consider a Lipschitz domain $\Omega = \{(x, y) : x \in \mathbb{R}^{n-1}, \Phi(x) > y\}$ and the corresponding approximating domains $\Omega_h = \{(x, y) : x \in \mathbb{R}^{n-1}, \Phi_h(x) > y\}$. Let us consider a discrete sequence of these, $\Omega^n = \Omega_{2^{-n}}$, $n \geq 0$. It is easy to arrange that the dyadic cube mesh constructed at one level is a refinement of the dyadic cube mesh for the previous level. To see this, we start with the first level, corresponding to $n = 0$, and carry out the construction of the dyadic mesh described above with $K(Q) = K_0(Q)$, defined with respect to Φ_1. This gives us the collection $\{Q_i^0\}$. On the next level, $j = 1$, we only consider dyadic cubes which are subcubes of the dyadic cubes in $\{Q_i^0\}$. In fact, to ensure that the sidelengths of the cubes in $\{Q_i^n\} \longrightarrow 0$ as $n \longrightarrow \infty$, we also restrict the sidelengths of the cubes we consider. For each $x \in \mathbb{R}^{n-1}$, we let $Q(x)$ be the largest dyadic cube Q, contained in some Q_i^0 and containing x, with $\ell(Q) \leq \frac{1}{2}$ and such that $K_1(\gamma Q) \leq \epsilon$. Here $K_1(Q)$ is defined with respect to $\Phi_{1/2}$. As before, this yields the collection $\{Q_i^1\}$. Proceeding by induction gives us the dyadic meshes $\{Q_i^n\}$, $n \geq 0$.

5. An error estimate. In this section we prove error estimates for the Galerkin procedure of the previous sections. The proof we shall give is similar to that of Lin [Li], where the solvability of the Galerkin equations for functions $f \in L_1^p$ is established. In that work p is assumed to lie between $2-\epsilon$ and $2+\epsilon$, and more involved approximation spaces are used.

Consider, as before, the Lipschitz graph case and the corresponding smooth approximating domains satisfying the seven conditions above. On each $\partial\Omega_h$ we assume that we have a function space X_h, localized of variable scale with parameters K and c_0 independent of h. If, under the hypotheses of Theorem 3.3, the function f defined by $Af = g$ has gradient in L^p, then we have the following estimate of the rate of convergence of the approximate solutions to f.

THEOREM 5.1. *Suppose the hypotheses of Theorem* 3.3 *are satisfied, and, in addition, that* $\rho_h = \max_j \rho_{j,h} = \mathcal{O}(h)$. *For each* $2 - \epsilon_M < p < \infty$ *and* $g \in L^p(\partial\Omega)$, *let* f *and* f_h *denote the unique solutions in* $L^p(\partial\Omega)$ *and in* $X_h \cap L_1^p(\partial\Omega_h)$ *of* $Af = g$ *and* $\Pi_h A_h f_h = \Pi_h g_h$, *respectively. Then*

$$\|f - f_h \circ F_h\|_{L^p(\partial\Omega)} \leq Ch \|\nabla f\|_{L^p(\partial\Omega)} \quad \text{for all } f \in L_1^p(\partial\Omega).$$

Before entering into the details of the proof, we shall give a brief outline and state some lemmas which we need.

The idea is simple. We know that $\|f - f_h \circ F_h^{-1}\|_{L^p(\partial\Omega)} \to 0$ as $h \to 0^+$. The main obstacle in obtaining an estimate for the rate of convergence is that $\nabla\Phi_h \to \nabla\Phi$ almost everywhere, but not uniformly, and, as a consequence it is not immediate that $\|Tf - T_h(f \circ F_h^{-1})\|_{L^p(\partial\Omega)} \longrightarrow 0$ as $h \to 0$. However, $\Phi_h \to \Phi$ uniformly; so, exploiting the smoothness of the function $f \in L_1^p(\partial\Omega)$, we can use Green's theorem to move the derivatives from Φ and Φ_h to f and f_h, or, rather, to

$$f(y, \Phi(y)) \frac{x_i - y_i}{(|x-y|^2 + (\Phi(x) - \Phi(y))^2)^{n/2}}$$

and

$$f(y, \Phi(y)) \frac{x_i - y_i}{(|x-y|^2 + (\Phi_h(x) - \Phi_h(y))^2)^{n/2}}.$$

This yields an estimate $\|Tf - T_h(f \circ F_h^{-1})\|_{L^p(\partial\Omega)} = \mathcal{O}(h)$, which is what we need.

We shall use the following two lemmas.

LEMMA 5.2. *Suppose* $\rho_h = \mathcal{O}(h)$. *For* $2 - \epsilon_M < p < \infty$,

$$\|\Pi_h q \circ F_h^{-1} - q \circ F_h^{-1}\|_{L^p(\partial\Omega_h)} \leq Ch \|\nabla q\|_{L^p(\partial\Omega)} \quad \text{for all } q \in L_1^p(\partial\Omega).$$

Proof. Take $x_0 \in E_{j,h}$. The localization property gives

$$\begin{aligned}&\left\|\Pi_h q \circ F_h^{-1} - q \circ F_h^{-1}(x_0)\right\|_{L^\infty(E_{j,h})}\\&\qquad\leq \sup_w \frac{1}{\sigma(B_{j,h})}\left|\int_{B_{j,h}}(q\circ F_h^{-1}(t) - q \circ F_h^{-1}(x_0))w(t)\,dt\right|,\end{aligned}$$

where the supremum is taken over $w \in X_h$ with $\|w\|_\infty \leq 1$ and support in $B_{j,h} = B(P_j, \rho_{j,h}) \cap \partial\Omega_h$. It is easy to estimate the right-hand side by $C|\nabla q|^*(x_0) \max_j \rho_{j,h} \leq Ch|\nabla q|^*(x_0)$, from which the proposition follows. □

The second lemma ultimately depends on the boundedness of the Cauchy integral on Lipschitz curves [CMM], [C]; cf. [Li] and the references therein.

LEMMA 5.3 (Lin [Li]). *Let*

$$T_\epsilon(y) = \int_{|x-y|>\epsilon} \frac{(f(x) - f(y))g(x)\prod_k(\Psi_i(x) - \Psi_i(y))}{(|x-y|^2 + (\Phi(x) - \Phi(y))^2)^{(n+k)/2}}\,dx.$$

Suppose that $\|\nabla\Phi\|\infty \leq M$ *and that* $k \geq 2$ *is an even integer. Then*

$$\left\|\sup_{\epsilon>0}|T_\epsilon(y)|\right\|_{L^r(\mathbb{R}^{n-1})} \leq C(M)\prod_k \|\nabla\Psi_i\|_{L^\infty(\mathbb{R}^{n-1})}\,\|\nabla f\|_{L^p(\mathbb{R}^{n-1})}\,\|g\|_{L^q(\mathbb{R}^{n-1})},$$

where $\frac{1}{r} = \frac{1}{q} + \frac{1}{p}, 1 < p \leq \infty, 1 < q \leq \infty, 1 < r < \infty$.

Proof of Theorem 5.1. We have

$$\begin{aligned}&\|f - f_h \circ F_h\|_{L^p(\partial\Omega)}\\&\qquad\leq c\left\|f\circ F_h^{-1} - \Pi_h(f\circ F_h^{-1})\right\|_{L^p(\partial\Omega_h)} + \left\|\Pi_h(f\circ F_h^{-1}) - f_h\right\|_{L^p(\partial\Omega_h)}.\end{aligned}$$

Hence, by Lemma 5.2 it is sufficient to estimate $\left\|\Pi_h(f\circ F_h^{-1}) - f_h\right\|_{L^p(\partial\Omega_h)}$. By the uniform invertibility of $\Pi_h A_h$, we see that

$$\begin{aligned}&\left\|f_h - \Pi_h(f\circ F_h^{-1})\right\|_{L^p(\partial\Omega_h)} \leq \|(\Pi_h A_h)^{-1}\|\,\left\|\Pi_h A_h f_h - \Pi_h A_h \Pi_h(f\circ F_h^{-1})\right\|_{L^p(\partial\Omega_h)}\\&\quad\leq C\left\|\Pi_h g_h - \Pi_h A_h \Pi_h(f\circ F_h^{-1})\right\|_{L^p(\partial\Omega_h)} \leq C\left\|g_h - A_h\Pi_h(f\circ F_h^{-1})\right\|_{L^p(\partial\Omega_h)}\\&\quad\leq C\left\|g\circ F_h^{-1} - A_h\Pi_h(f\circ F_h^{-1})\right\|_{L^p(\partial\Omega_h)} \leq C\left\|g - A_h\Pi_h(f\circ F_h^{-1})\circ F_h\right\|_{L^p(\partial\Omega)}\\&\quad= C\left\|Af - A_h\Pi_h(f\circ F_h^{-1})\circ F_h\right\|_{L^p(\partial\Omega)}.\end{aligned}$$

Adding and subtracting $A_h(f\circ F_h^{-1})\circ F_h$ and using the triangle inequality shows that this is less than

$$\begin{aligned}&C\big(\left\|Af - A_h(f\circ F_h^{-1})\circ F_h\right\|_{L^p(\partial\Omega)}\\&\quad+\left\|A_h(f\circ F_h^{-1})\circ F_h - A_h\Pi_h(f\circ F_h^{-1})\circ F_h\right\|_{L^p(\partial\Omega)}\big)\\&\qquad\leq C\big(\left\|Tf - T_h(f\circ F_h^{-1})\circ F_h\right\|_{L^p(\partial\Omega)} + \|A_h\|\,\left\|f\circ F_h^{-1} - \Pi_h(f\circ F_h^{-1})\right\|_{L^p(\partial\Omega)}\big).\end{aligned}$$

Applying Lemma 5.2, there only remains to estimate the first term of the right-hand side.

Let $P = (x, \Phi(x))$ and $F(x) = f(x, \Phi(x))$. Since $T1 = T_h 1 = \frac{1}{2}$, we have $Tf(P) - T_h(f \circ F_h^{-1})(F_h(P)) = \lim_{\epsilon \to 0} \tilde{I}^\epsilon(P)$ with

$$\tilde{I}^\epsilon(P) = \frac{1}{\omega_n} \int_{\mathbb{R}^{n-1} \setminus \tilde{B}(y,\epsilon)} (F(y) - F(x)) \left(\frac{(x-y) \cdot \nabla\Phi(y) - (\Phi(x) - \Phi(y))}{(|x-y|^2 + (\Phi(x) - \Phi(y))^2)^{n/2}} \right. \\ \left. - \frac{(x-y) \cdot \nabla\Phi_h(y) - (\Phi_h(x) - \Phi_h(y))}{(|x-y|^2 + (\Phi_h(x) - \Phi_h(y))^2)^{n/2}} \right) dy.$$

Here $\tilde{B}(x,\epsilon) = \{y \in \mathbb{R}^{n-1} : (y, \Phi(y)) \in B(P,\epsilon) \cap \partial\Omega\}$ is the perpendicular projection of $B(P,\epsilon) \cap \partial\Omega$ on $\mathbb{R}^{n-1}$. Let $I^\epsilon(x)$ be defined as $\tilde{I}^\epsilon(P)$, but with $\tilde{B}(y,\epsilon)$ replaced by a regular ball $b(y,\epsilon) = \{y \in \mathbb{R}^{n-1} : |x-y| \le \epsilon\}$ in $\mathbb{R}^{n-1}$. Since Φ is Lipschitz, each $\tilde{B}(y,\epsilon)$ is enclosed between two regular balls of comparable size, $b(y, \epsilon/(1+M^2)^{1/2}) \subset \tilde{B}(y,\epsilon) \subset b(y,\epsilon)$, and, as a consequence, it easy to see that $\lim_{\epsilon \to 0}(\tilde{I}^\epsilon - I^\epsilon) = 0$ in L^p. Hence, $Tf(P) - T_h(f \circ F_h^{-1})(F_h(P)) = \lim_{\epsilon \to 0} I^\epsilon(x)$. In the remainder of the proof we shall estimate I^ϵ.

Let $L(x,y) = (|y-x|^2 + (\Phi(y) - \Phi(x))^2)^{n/2}$ and denote by $L_h(x,y)$ the corresponding entity for Φ_h. Then, dividing into gradient and nongradient parts, and adding and subtracting the quantities $(F(y) - F(x))(\Phi(x) - \Phi(y))/L(x,y)$ and $(F(y) - F(x))((x-y) \cdot \nabla\Phi_h(y))/L(x,y)$, respectively, we get the decomposition $I_\epsilon(x) = I_1^\epsilon(x) + I_2^\epsilon(x) + I_3^\epsilon(x) + I_4^\epsilon(x) + J^\epsilon(x)$, where

$$I_1^\epsilon(x) = \int_{|y-x| \ge \epsilon} (F(y) - F(x)) \frac{\Phi(y) - \Phi_h(y)}{L(x,y)}\, dy,$$

$$I_2^\epsilon(x) = -\int_{|y-x| \ge \epsilon} (F(y) - F(x)) \frac{\Phi(x) - \Phi_h(x)}{L(x,y)}\, dy,$$

$$I_3^\epsilon(x) = \int_{|y-x| \ge \epsilon} (F(y) - F(x))(\Phi_h(x) - \Phi_h(y)) \left(\frac{1}{L(x,y)} - \frac{1}{L_h(x,y)} \right) dy,$$

$$I_4^\epsilon(x) = \int_{|y-x| \ge \epsilon} (F(y) - F(x))(x-y) \cdot \nabla\Phi_h(y) \left(\frac{1}{L(x,y)} - \frac{1}{L_h(x,y)} \right) dy,$$

and

$$J^\epsilon(x) = -\int_{|y-x| \ge \epsilon} (F(y) - F(x))(x-y) \cdot \nabla(\Phi(y) - \Phi_h(y)) \frac{1}{L(x,y)}\, dy.$$

Using Green's formula, we move the derivative away from $\nabla(\Phi(y) - \Phi_h(y))$ in $J^\epsilon(x)$. In this way, we see that $J^\epsilon(x) = I_5^\epsilon(x) + I_6^\epsilon(x) + I_7^\epsilon(x) + I_8^\epsilon(x)$, where

$$I_5^\epsilon(x) = \int_{|y-x| = \epsilon} \frac{F(y) - F(x)}{L(x,y)} (\Phi(y) - \Phi_h(y))|y-x|\, d\sigma(y),$$

$$I_6^\epsilon(x) = -\int_{|y-x| \ge \epsilon} \frac{\nabla F(y) \cdot (x-y))}{L(x,y)} (\Phi(y) - \Phi_h(y))\, dy,$$

$$I_7^\epsilon(x) = \int_{|y-x| \ge \epsilon} \frac{n(F(y) - F(x))(\Phi(x) - \Phi(y))^2}{L(x,y)^{1+2/n}} (\Phi(y) - \Phi_h(y))\, dy,$$

and

$$I_8^\epsilon(x) = -\int_{|y-x| \ge \epsilon} \frac{n(F(y) - F(x))\nabla\Phi(y) \cdot (x-y)}{L(y,x)^{1+2/n}} (\Phi(x) - \Phi(y))(\Phi(y) - \Phi_h(y))\, dy.$$

Our basic decomposition of I^ϵ is $I^\epsilon = \sum_{k=1}^{8} I_k^\epsilon$.

We estimate the boundary integral $I_5^\epsilon(x)$ first. Clearly,

$$|I_5^\epsilon(x)| \leq C \|\Phi - \Phi_h\|_{L^\infty} \int_{|\omega|=1} |\nabla F(x + \epsilon\omega)| \, d\omega.$$

Hence, by Minkowski's inequality and a change of variable,

$$\|I_5^\epsilon\|_p \leq C \|\Phi - \Phi_h\|_{L^\infty} \|\nabla F\|_p .$$

The estimates for the remaining integrals will follow from Lemma 5.2. For I_1^ϵ and I_2^ϵ a direct application of the lemma with $k = 2$, $r = p$, and $q = \infty$ yields

$$\|I_i^\epsilon\|_p \leq \left\|\sup_{\epsilon>0} |I^{\epsilon_i}|\right\|_p \leq C(M) \|\Phi - \Phi_h\|_\infty \|\nabla F\|_p , \qquad i = 1, 2.$$

To estimate I_6^ϵ, we write the integrand as

$$\frac{\sum_i \frac{\partial F}{\partial x_i}(y)(x_i - y_i)(\Phi(y) - \Phi_h(y)) \times \left(|x - y|^2 + (\Phi(x) - \Phi(y))^2\right)}{\left(|x - y|^2 + (\Phi(x) - \Phi(y))^2\right)^{n/2+1}} .$$

Now, Lemma 5.2 with $k = 2$, $r = q = 2$, and $p = \infty$ readily gives

$$\|I_6^\epsilon\|_p \leq C(M) \|\Phi - \Phi_h\|_\infty \|\nabla F\|_p .$$

In a similar way, we get

$$\|I_i^\epsilon\|_p \leq C(M) \|\Phi - \Phi_h\|_\infty \|\nabla F\|_p , \qquad i = 7, 8.$$

There remains to estimate I_3^ϵ and I_4^ϵ. We let $H(q) = H_{xy}(q) = (|y - x|^2 + q^2)^{-n/2}$, $q \in \mathbb{R}$ so that $H'(q) = -nq(|y - x|^2 + q^2)^{-n/2-1}$. We also define $\Psi_t(x) = \Phi_h(x) + t(\Phi(x) - \Phi_h(x))$. Then

$$\begin{aligned}\frac{1}{L} - \frac{1}{L_h} &= H(\Phi(x) - \Phi(y)) - H(\Phi_h(x) - \Phi_h(y)) \\ &= -n\big(\Phi(x) - \Phi_h(x) - (\Phi(y) - \Phi_h(y)\big) \int_0^1 \frac{\Psi_t(x) - \Psi_t(y)}{L_{\Psi_t}^{1+2/n}} \, dt,\end{aligned}$$

where $L_{\Psi_t} = (|x - y|^2 + (\Psi_t(x) - \Psi_t(y))^2)^{n/2}$. Hence, by Fubini's theorem,

$$\begin{aligned}&I_3^\epsilon(x) \\ &= -n \int_0^1 \int_{|y-x|\geq\epsilon} (F(y) - F(x))(\Phi_h(x) - \Phi_h(y))(\Phi(x) - \Phi_h(x)) \frac{\Psi_t(x) - \Psi_t(y)}{L_{\Psi_t}^{1+2/n}} \, dydt \\ &+ n \int_0^1 \int_{|y-x|\geq\epsilon} (F(y) - F(x))(\Phi_h(x) - \Phi_h(y))(\Phi(y) - \Phi_h(y)) \frac{\Psi_t(x) - \Psi_t(y)}{L_{\Psi_t}^{1+2/n}} \, dydt \\ &= I_3' + I_3''.\end{aligned}$$

Another use of Lemma 5.2 shows that

$$\begin{aligned}\|I_3'\|_p \leq & n\,\|\Phi - \Phi_h\|_\infty \int_0^1 \left\| \sup_{\epsilon>0} \left| \int_{|y-x|\geq\epsilon} (F(y) - F(x))(\Phi_h(x) - \Phi_h(y)) \right.\right. \\ & \left.\left. \times \frac{\Psi_t(x) - \Psi_t(y)}{(|x-y|^2 + (\Psi_t(x) - \Psi_t(y)))^{n/2+1}}\, dy \right|\right\|_p dt \\ \leq & n\,\|\Phi - \Phi_h\|_\infty \int_0^1 C(M)\,\|\nabla\Phi\|_\infty\,\|\nabla F\|_p\,\|1\|_\infty\, dt \\ \leq & C(M)\,\|\Phi - \Phi_h\|_\infty\,\|\nabla F\|_p,\end{aligned}$$

where C(M) only depends on M, since the Lipschitz constant of Ψ_t is less than $3M$.

Similar arguments give the same kind of estimates for I_3'' and I_4^ϵ.

Summing up, we have thus proved that

$$\left\|f_h - \Pi_h(f \circ F_h^{-1})\right\|_{L^p(\partial\Omega_h)} \leq C \left\|\lim_{\epsilon\to 0} I^\epsilon\right\|_p \leq C\,\|\Phi - \Phi_h\|_\infty\,\|\nabla F\|_p,$$

which is the desired inequality. □

6. Systems of equations. In this section we shall discuss certain systems of elliptic equations on a Lipschitz domain $\Omega \subset \mathbb{R}^n$, $n \geq 3$. For simplicity, we shall still assume that $\Omega = \{(x,y) : y = \Phi(x)\}$ for some Lipschitz function Φ. We shall concentrate on the discretized version of the Dirichlet problem for the Lamé systems of linearized elastostatics with $L^2(\partial\Omega)$ data. For the detailed treatment of the continuous case, and background and references, we refer to [DKV] and [FKV].

Let the function $\vec{g} = (g_1, \cdots, g_n) \in L^2(\partial\Omega)$ be given. The Dirichlet problem for the Lamé system is to find a function $\vec{u}$ that solves

$$\mu\Delta\vec{u} + (\lambda + \mu)\nabla(\operatorname{div}\vec{u}) = \vec{0} \tag{6.1}$$

with boundary values $\vec{u}|_{\partial\Omega} = \vec{g}$. Here λ and μ are constants, the so-called Lamé moduli, which satisfy

$$\mu > 0, \qquad \lambda > -2\mu/n.$$

The Kelvin matrix Γ of fundamental solutions of (6.1) is given by

$$\Gamma_{ij}(x) = \frac{A}{\omega_n(n-2)}\frac{\delta_{ij}}{|x|^{n-2}} + \frac{B}{\omega_n}\frac{x_i x_j}{|x|^n} \tag{6.2}$$

where

$$A = \frac{1}{2}\left(\frac{1}{\mu} + \frac{1}{2\mu+\lambda}\right), \qquad B = A = \frac{1}{2}\left(\frac{1}{\mu} - \frac{1}{2\mu+\lambda}\right).$$

To discuss the boundary integral method, we first need the analog of the layer potentials. Corresponding to the single layer potential, we have

$$\mathcal{S}\vec{f}(P) = \int_{\partial\Omega} \Gamma(P-Q)\vec{f}(Q)\, d\sigma(Q). \tag{6.3}$$

There are, in fact, infinitely many conormal derivatives and corresponding double layer potentials associated with the Lamé system in a natural way. Following [DKV], we consider

$$\frac{\partial}{\partial\rho}\vec{u} = \left(\lambda + \frac{2\mu^2}{3\mu+\lambda}\right)(\operatorname{div}\vec{u})N + \mu(\nabla\vec{u})N + \left(\mu - \frac{2\mu^2}{3\mu+\lambda}\right)(\nabla\vec{u}^T)N, \tag{6.4}$$

where N denotes the outward normal. The associated double layer potential is

$$\mathcal{D}\vec{f}(P)=\int_{\partial\Omega}\left(\frac{\partial}{\partial\rho}\Gamma(P-Q)\right)^{T}\vec{f}(Q)\,d\sigma(Q).$$

We let $\mathcal{A}$ denote the operator corresponding to the boundary values of this double layer potential. By formally allowing $P\in\partial\Omega$ in the definition of $\mathcal{D}$, we obtain, as in the scalar case, a principal value operator T that satisfies

$$\mathcal{A}\vec{f}(P)=\tfrac{1}{2}\vec{f}(P)+T\vec{f}(P).$$

The conormal derivative defined by (6.4) has the advantage that the operator T is compact for bounded, smooth surfaces. More specifically, the kernel of T is a sum of terms with kernels of the form

$$\frac{(P-Q)_i(P-Q)_j}{|P-Q|^2}\frac{\langle P-Q,N_Q\rangle}{|P-Q|^n}$$

and

$$\frac{\langle P-Q,N_Q\rangle}{|P-Q|^n}\delta_{ij}.$$

Furthermore, the analog of the norm inequality (2.4) is true for $p=2$ (see [DKV]). As a consequence, the Galerkin procedure of Dahlberg and Verchota and the results above extend, in a straightforward way, to the Dirichlet problem for the Lamé system.

From an application point of view, the so-called traction problem is perhaps more interesting. This problem is the analog of the Neumann problem with the (co)normal derivative defined by

$$\frac{\partial}{\partial\nu}\vec{u}=\lambda(\operatorname{div}\vec{u})N+\mu(\nabla\vec{u}+\nabla\vec{u}^{T})N. \tag{6.5}$$

The boundary values of the corresponding double layer potential can be written $\tilde{A}\vec{f}=\frac{1}{2}\vec{f}+\tilde{T}\vec{f}$. However, in this case the operator $\tilde{T}$ is not compact on smooth boundaries, and the Dahlberg–Verchota procedure does not apply, at least not in a trivial way.

There are many other examples of systems of equations to which the theory applies. One such example is the Stokes system of linear hydrostatics:

$$\begin{cases}\Delta u=\nabla p,\\ \nabla\cdot u=0.\end{cases}$$

We define the conormal derivative

$$\frac{\partial}{\partial\rho}\vec{u}=(\nabla\vec{u}+\nabla\vec{u}^{T})N-Np. \tag{6.6}$$

Again the boundary values of the associated double layer potential equals

$$A\vec{f}=\tfrac{1}{2}\vec{f}+T\vec{f}. \tag{6.7}$$

The operator T is of essentially the same type as before; in particular, it is compact on smooth bounded domains. Also, the norm inequality that replaces (2.4) is proved in [DKV] (cf. also [FKV]). This means that the analog of the above results for the Galerkin method procedure and the Dirichlet problem are true. In addition, the Neumann problem associated with this conormal derivative (cf. [DKV]), which results in the so-called slip condition, can be discretized. Let

$$B\vec{f}=-\tfrac{1}{2}\vec{f}+T^{*}\vec{f},$$

where T^* is the adjoint of the operator in (6.7). A typical example is the following result (cf. Theorem 3.3).

THEOREM 6.1. *Suppose that* $\lim_{h\to 0}\rho_h = 0$. *Then there are* $h_0 > 0$ *and* $\gamma > 1$ *with the property that if*

$$\lim_{h\to 0}\kappa_h = \lim_{h\to 0}\sup_j \kappa(P_{j,h}, \gamma\rho_{j,h})\gamma\rho_{j,h} = 0,$$

then for every $\vec{g} \in L^2(\partial\Omega)$ *there is a unique* $\vec{f_h} \in X_h \cap L^2(\partial\Omega_h)$ *for* $h < h_0$, *such that*

$$\int_{\partial\Omega_h} B_h \vec{f_h}\, w\, d\sigma = \int_{\partial\Omega_h} \vec{g}_h \vec{w}\, d\sigma \tag{6.8}$$

for all $\vec{w} = (w_1, \dots, w_n)$, $w_i \in X$, *with compact support. Here* $\vec{g}_h(P) = \vec{g}(F_h^{-1}(P))$. *Moreover, if* $\vec{f} \in L^2(\partial\Omega)$ *is defined by* $B\vec{f} = \vec{g}$, *then* $\vec{f_h} \circ F_h$ *converges to* $\vec{f}$ *in* $L^2(\partial\Omega)$.

Another interesting example is provided by the system

$$\frac{1}{2}f_i(y) = \tilde{f}_i(y) + \frac{3}{4\pi}\int_{\partial\Omega} \frac{y_i - x_i}{|y-x|^5}\langle y - x, n(y)\rangle \langle y - x, \vec{f}(x)\rangle dS_x, \qquad i = 1, 2, 3.$$

This system describes the stress $\vec{f}$ in a Stokes flow after a body of shape $\Omega \subset \mathbb{R}^3$ has been inserted in a Stokes flow with stress $\tilde{f}$ (cf. [BL]); n denotes the inward normal. In this case, the kernel K has the form

$$K_{i,j} = \frac{\langle x - y, n(y)\rangle}{|x-y|^5}(x_i - y_j)(x_j - y_i),$$

which obviously is compact on bounded, smooth domains.

The boundedness of the operator $\frac{1}{2} + K$ and its inverse with respect to the appropriate subspace of $L^2(\partial\Omega)$ follow from Theorem 4.6 in [DKV]. The Galerkin procedure with variable meshsize thus applies for this system as well.

Acknowledgment. This paper grew out of a problem posed to us by Björn E. J. Dahlberg, and arising in a natural way in [DV]. We have also had several useful discussions with Dahlberg.

REFERENCES

[BL] E. BARTA AND N. LIRON, *Motion of a rigid particle in stokes flow: a second kind boundary integral equation formulation*, preprint.

[C] A. P. CALDERÓN, *Cauchy integrals on Lipschitz curves and related operators*, Proc. Nat. Acad. Sci. U.S.A., 74 (1977), pp. 1324–1327.

[CMM] R. COIFMAN, A. MCINTOSH, AND Y. MEYER, *L'integrale de Cauchy definit un operateur borne sur L^2 pour les courbes lipschitziennes*, Ann. of Math., 116 (1982), pp. 361–387.

[D] B. DAHLBERG, *On estimates of harmonic measure*, Arch. Rational Mech. Anal., 65 (1977), pp. 275–288.

[DK] B. DAHLBERG AND C. KENIG, *Hardy spaces and the L^p-Neumann problem for Laplace's equation in a Lipschitz domain*, Ann. Math., 125 (1987), pp. 437–465.

[DK1] B. DAHLBERG AND C. KENIG, *Harmonic Analysis and Partial Differential Equations*, No. 1985-03/ ISSN 0347-2809, Chalmers University of Technology and the University of Göteborg, Göteborg, Sweeden, 1985.

[DKV] B. DAHLBERG, C. KENIG, AND G. VERCHOTA, *Boundary value problems for the systems of elastostatics in Lipschitz domains*, Duke Math. J., 57 (1988), pp. 795–818.

[DV] B. Dahlberg and G. Verchota, *Galerkin methods for the boundary integral equations of elliptic equations in non-smooth domains*, Contemp. Math., 107 (1990), pp. 39–60.

[Da] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[FJR] E. Fabes, M. Jodeit, and N. Riviere, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., 141 (1978), pp. 165–186.

[FKV] E. Fabes, C. Kenig, and G. Verchota, *The Dirichlet problem for the Stokes system on Lipschitz domains*, Duke Math. J., 57 (1988), pp. 769–793.

[JK] D. Jerison and C. Kenig, *The Neumann problem on Lipschitz domains*, Bull. Amer. Math. Soc., 4 (1981), pp. 203–207.

[L] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, London, Paris, 1969.

[Li] N. Lin, *Galerkin methods for the boundary integral equations of the Dirichlet problem in Lipschitz domains*, preprint.

[V] G. Verchota, *Layer potentials and boundary value problems for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.

[Y] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, 1978.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1375-1392, November 1992

MATHEMATICAL ANALYSIS OF MISCIBLE DISPLACEMENT IN POROUS MEDIUM*

PIERRE FABRIE† AND MICHEL LANGLAIS‡

Abstract. When modeling miscible fluid displacement in porous medium [*Stud. Math. Appl.*, 17 (1986)], [*Mathematical and Computational Method in Seismic Exploration and Reservoir Modeling*, SIAM, pp. 108-127] variation with temperature of such physically relevant parameters as viscosity and thermal conductivity needs to be considered [Goyeau, thesis, Université Bordeaux I, 1988]. The dependence of dispersion with velocity is also important when modeling pollution problems. In so doing, one is led to a system of a nonlinear parabolic equation coupled with a Darcy or a Darcy-Forchheimer type equation; natural boundary conditions are supplemented.

A mathematical analysis of such a problem is performed in a cylindrical domain corresponding to a reservoir. A notion of weak solutions is introduced for the two unknown functions: temperature and velocity. The existence of at least one such solution is proved. This is achieved upon introducing a positive time-lag in the nonlinearities featuring in the energy equation. It leads to a decoupling of the system into a pair of linear problems having at least one solution. A uniform bound for the approximate temperature is derived using a maximum principle from which further energy estimates leading to a compactness property follow. Next, using a local Meyers lemma, a compactness property is derived for the approximate velocity.

The existence of a weak solution is obtained upon letting the time-lag go to zero. Some extra properties of the solution, when more or less restrictive conditions are added, are given.

This paper refers to [*J. Differential Equations*, 90 (1991), pp. 186-202] for a mathematical theory of related stationary problem. See also [*Stud. Math. Appl.*, 17 (1986)] for a comprehensive modeling of this type of displacement.

Key words. miscible displacement, elliptic parabolic system, Darcy, Darcy-Forchheimer equations

AMS(MOS) subject classifications. 76T05, 35K55, 76S05, 35K50

1. Introduction.

For an incompressible but dilatable fluid, the main system frequently used reads

—mass conservation law

$$\varepsilon\frac{\partial\rho}{\partial t}+\nabla\cdot\rho\mathbf{V}=0;$$

—energy conservation law

$$(\rho c)^*\frac{\partial T}{\partial t}+(\rho c)_f\mathbf{V}\cdot\nabla T-\nabla\cdot[(\lambda^*(T)+D(\mathbf{V}))\cdot\nabla T]=0;$$

—momentum equation (taking either forms)

—Darcy's law

$$\text{(i)}\qquad \mu(T)\mathbf{V}+K\cdot(\nabla p+\rho g)=0;$$

—Darcy-Forchheimer's law

$$\text{(ii)}\qquad \frac{\rho}{\varepsilon}\frac{\partial\mathbf{V}}{\partial t}+\mu(T)K^{-1}\mathbf{V}+c_j\rho\sigma(\mathbf{V})+\nabla p+\rho\mathbf{g}=0;$$

* Received by the editors September 10, 1990; accepted for publication (in revised form) March 10, 1992.

† UFR Mathématiques et Informatique, Centre de Recherche en Mathematiques de Bordeaux, URA Centre National de la Recherche Scientifique 226, Université Bordeaux I, 33405 Talence Cedex, France and L.E.P.T.-ENSAM, URA Centre National de la Recherche Scientifique 873, Esplanade des Arts et Métiers, 33405 Talence Cedex, France.

‡ UFR Sciences Humaines Appliquées, UA Centre National de la Recherche Scientifique 226, Université Bordeaux II, 33076 Bordeaux Cedex, France.

the underlying constitutive equations are first

$$\rho = \rho_0(1 - \beta(T - T_0))$$

(this constitutive law expresses that the fluid is incompressible $(\partial\rho/\partial p) = 0$ but dilatable $\rho = \rho(T)$) and next

$$\mu(T) = \lambda_0 e^{\gamma(T-T_0)}, \qquad D_{ij}(\mathbf{V}) = \alpha_L|\mathbf{V}|^{-1}V_iV_j + \alpha_T|\mathbf{V}|^{-1}(|\mathbf{V}|^2\delta_{ij} - V_iV_j).$$

In these relations ε stands for porosity, ρ for density, $\mathbf{V}$ the filtration velocity, p the pressure, and T the temperature. We, respectively, denote by $(\rho c)_f$ and $(\rho c)_s$ the fluid and solid heat capacity, so that the equivalent heat capacity of the medium is given by $(\rho c)_* = \varepsilon(\rho c)_f + (1-\varepsilon)(\rho c)_s$. Next $\lambda^*(T)$ is the thermal conductivity of the medium, D the dispersion tensor, $\mu(T)$ the fluid viscosity, and K the permeability. According to the generalized Boussinesq assumption, the fluid density variation is neglected everywhere in all of the previous equations, except in the term $\rho\mathbf{g}$.

The Darcy model is the one most frequently used. But for large velocities, the Darcy-Forchheimer model is more realistic: see [12]; in this case, for isotropic flow $\boldsymbol{\sigma}(\mathbf{V})$ may take the following form:

$$\boldsymbol{\sigma}(\mathbf{V}) = k^{-1/2}|\mathbf{V}|\mathbf{V},$$

where k is the intrinsic permeability and $K = kId$. For more general flows we may take the ith component of $\boldsymbol{\sigma}(\mathbf{V})$ to be

$$\frac{1}{|\mathbf{V}|}\sigma_{ijkl}V_jV_kV_l.$$

As an example, $\boldsymbol{\sigma}(\mathbf{V}) = J(\mathbf{V})\mathbf{V}$ with

$$J_{ij}(\mathbf{V}) = \alpha|\mathbf{V}|\delta_{ij} - \beta\frac{V_iV_j}{|\mathbf{V}|}(1-\delta_{ij}), \qquad \beta > 0, \quad \alpha > 0.$$

So the dimensionless final problem that is to be analyzed here takes the following form:

—isovolume flow

(1) $$\nabla \cdot \mathbf{V} = 0;$$

—energy equation

(2) $$\frac{\partial\theta}{\partial t} + \mathbf{V}\cdot\nabla\theta - \nabla\cdot[(\lambda(\theta) + D(\mathbf{V}))\nabla\theta] = 0;$$

—momentum equation

—Darcy's law

(3a) $$\mu(\theta)\mathbf{V} + K(\nabla p - Ra^*\beta(\theta)\mathbf{e}_z) = 0;$$

—Darcy-Forchheimer's law

(3b) $$\varepsilon^{-1}DaP_r^{-1}\frac{\partial\mathbf{V}}{\partial t} + K^{-1}\mu(\theta)\mathbf{V} + c_jDa^{1/2}P_r^{-1}\boldsymbol{\sigma}(\mathbf{V}) + \nabla p - Ra^*\beta(\theta)\mathbf{e}_z = 0,$$

where $Ra^* = RaDa$ is the filtration Rayleigh number, Pr the Prandtl number, and Da the Darcy number. Further, $\lambda(\theta)$, $\mu(\theta)$, and $\beta(\theta)$ are dimensionless thermal conductivity, viscosity, and density.

In a cylindrical aquifer $\Omega =]0, H[\times]R_i, R_e[\times]0, 2\pi[$, having height H, external radius R_e, crossed by an injection well of radius R_i, the thermal and hydrodynamical boundary conditions are

$$-(\lambda(\theta)+D(\mathbf{V}))\cdot\nabla\theta\cdot\boldsymbol{\eta}=0 \quad \text{for } z=0 \quad \text{and} \quad z=H, \tag{4}$$

where $\boldsymbol{\eta}$ is the unique outward normal vector to $\partial\Omega$

$$-(\lambda(\theta)+D(\mathbf{V}))\cdot\nabla\theta\cdot\boldsymbol{\eta}+g_i(\theta-h_i)=0 \quad \text{for } r=R_i, \tag{5}$$

$$-(\lambda(\theta)+D(\mathbf{V}))\cdot\nabla\theta\cdot\boldsymbol{\eta}=0 \quad \text{for } r=R_e, \tag{6}$$

$$\mathbf{V}\cdot\boldsymbol{\eta}=0 \quad \text{for } z=0 \quad \text{and} \quad z=H, \tag{7}$$

$$\mathbf{V}\cdot\boldsymbol{\eta}=g_j \quad \text{for } r=R_j \quad j=i \text{ or } e, \tag{8}$$

with

$$\int_{r=R_i} g_i\,d\sigma_i+\int_{r=R_e} g_e\,d\sigma_e=0. \tag{9}$$

2. Assumptions, notations, and weak formulation.

2.1. Assumptions and notations.

Σ_j is the subset of the boundary $\partial\Omega$ defined by

$$\Sigma_j=\{(z, R_j, \zeta), 0<z<H, 0\leqq\zeta<2\pi\}, \quad j=i \quad \text{or} \quad e.$$

Throughout this work, the data will satisfy the following conditions:

$$h_i\in L^\infty(\Sigma_i), \quad g_j\in L^\infty(\Sigma_j), \quad j=i \quad \text{or} \quad e, \tag{A1}$$

(A2) there exists $m_i>0$, $M_i>0$ such that

$$-M_i\leqq g_i\leqq -m_i \quad \text{a.e.,}$$

$$0\leqq g_e \quad \text{a.e.,}$$

$$\int_{\Sigma_i} g_i\,d\sigma_i+\int_{\Sigma_e} g_e\,d\sigma_e=0.$$

The tensors λ, D, and $\boldsymbol{\sigma}$ are continuous on their respective domains. The functions μ and β are continuous over $\mathbb{R}$.

Furthermore, we assume that

$$\begin{aligned}&\forall(a,b)\in\mathbb{R}^2, \quad \exists\alpha_1>0, \quad \forall c\in]a,b[, \quad \alpha_1\leqq\mu(c).\\ \text{(A3)}\quad&\forall(a,b)\in\mathbb{R}^2, \quad \exists\alpha_2>0, \quad \forall c\in]a,b[, \quad \forall\boldsymbol{\zeta}\in\mathbb{R}^3, \quad \alpha_2|\boldsymbol{\zeta}|^2\leqq\lambda(c)\cdot\boldsymbol{\zeta}\cdot\boldsymbol{\zeta},\\ &\exists\alpha_3>0, \quad \forall\boldsymbol{\zeta}\in\mathbb{R}^3, \quad \alpha_3|\boldsymbol{\zeta}|^2\leqq K^{-1}\cdot\boldsymbol{\zeta}\cdot\boldsymbol{\zeta},\end{aligned}$$

$$\forall\mathbf{V}\in\mathbb{R}^3, \quad \forall\boldsymbol{\zeta}\in\mathbb{R}^3, \quad D(\mathbf{V})\cdot\boldsymbol{\zeta}\cdot\boldsymbol{\zeta}\geqq 0, \tag{A4}$$

$$\exists(\alpha_4,\alpha_5)\in\mathbb{R}_+^2, \quad \forall\mathbf{V}\in\mathbb{R}^3, \quad |D(\mathbf{V})|\leqq\alpha_4+\alpha_5|\mathbf{V}|, \tag{A5}$$

$$\begin{aligned}\text{(A6)}\quad&\exists\sigma_1>0, \quad \forall\mathbf{V},\mathbf{U}\in\mathbb{R}^3\times\mathbb{R}^3, \quad (\boldsymbol{\sigma}(\mathbf{U})-\boldsymbol{\sigma}(\mathbf{V}))\cdot(\mathbf{U}-\mathbf{V})\geqq\sigma_1|\mathbf{U}-\mathbf{V}|^3,\\ &\exists\sigma_2>0, \quad \forall\mathbf{V},\mathbf{U}\in\mathbb{R}^3\times\mathbb{R}^3, \quad |\boldsymbol{\sigma}(\mathbf{U})-\boldsymbol{\sigma}(\mathbf{V})|\leqq\sigma_2(|\mathbf{U}|+|\mathbf{V}|)(|\mathbf{U}-\mathbf{V}|).\end{aligned}$$

2.2. Weak formulation.

We first introduce the closure in $L_p(\Omega)$ of the smooth and divergence free functions vanishing on $\partial\Omega$, which we denote $\mathbb{H}_p$.

We begin with the Darcy model.

Assuming (A1), $\cdots$, (A5) to hold, a weak solution for problems (1) to (9) is a triplet $(\theta, \mathbf{V}, p)$ such that for all positive t,

$$\begin{aligned}&\theta\in L^\infty(\mathbb{R}_+\times\Omega)\cap L^2((0,t); H^1(\Omega)),\\ &\mathbf{V}\in L^\infty(\mathbb{R}_+; \mathbb{H}_2),\\ &p\in L^\infty(\mathbb{R}_+; H^1(\Omega)/\mathbb{R}),\end{aligned}$$

verifying for any test function φ lying in $L^2((0, t); W^{1,\infty}(\Omega))$,

$$\begin{aligned}&\int_0^t \left\langle \frac{\partial \theta}{\partial t}, \varphi \right\rangle d\tau + \int_0^t \int_\Omega (\lambda(\theta) + D(\mathbf{V})) \cdot \nabla\theta \cdot \nabla\varphi \, d\omega \, d\tau \\ &+\frac{1}{2}\int_0^t \int_\Omega ((\mathbf{V}\cdot\nabla\theta)\varphi - (\mathbf{V}\cdot\nabla\varphi)\theta) \, d\omega \, d\tau + \frac{1}{2}\int_0^t \int_{\Sigma_i} g_i\theta\varphi \, d\sigma_i \, d\tau \\ &+\frac{1}{2}\int_0^t \int_{\Sigma_e} \theta g_e \varphi \, d\sigma_e \, d\tau - \int_0^t \int_{\Sigma_i} g_i(\theta - h_i)\varphi \, d\sigma_i \, d\tau = 0,\end{aligned} \tag{10}$$

and verifying for any test function lying in $H^1(\Omega)$,

$$\begin{aligned}&\int_\Omega \mu^{-1}(\theta)K\cdot(\nabla p - Ra^*\beta(\theta)\mathbf{e}_z)\cdot\nabla\psi \, d\omega + \int_{\Sigma_i} g_i\psi \, d\sigma_i \\ &\quad + \int_{\Sigma_e} g_e\psi \, d\sigma_e = 0 \quad \text{a.e. in } (0, t);\end{aligned} \tag{11a}$$

the velocity is then given by

$$\mathbf{V} = -\mu^{-1}(\theta)K\cdot(\nabla p - Ra^*\beta(\theta)\mathbf{e}_z). \tag{12}$$

Next we consider the Darcy–Forchheimer model.

The only two modifications are in the momentum equation. The velocity function is given by $\mathbf{V} = \mathbf{U} + \mathbf{Z}$, where $\mathbf{Z}$ is a solution to

$$\begin{aligned}&\nabla\cdot\mathbf{Z} = 0,\\ &\mathbf{Z}\cdot\boldsymbol{\eta} = 0 \quad \text{for } z = 0 \quad \text{and} \quad H,\\ &\mathbf{Z}\cdot\boldsymbol{\eta} = g_j, \quad r = R_j, \quad j = i, e\end{aligned}$$

and seek $\mathbf{U}$ in $L^3((0, t), \mathbb{H}_3)$, satisfying, for each test function $\mathbf{W}$ lying in $L^3((0, t), \mathbb{H}_3)$,

$$\begin{aligned}&\int_0^t \int_\Omega \frac{\partial \mathbf{U}}{\partial t}\mathbf{W} \, d\omega \, d\tau + \int_0^t \int_\Omega (\mu(\theta)K^{-1}\cdot(\mathbf{U}+\mathbf{Z}) \\ &\qquad + c_j Da^{1/2}P_r^{-1}\boldsymbol{\sigma}(\mathbf{U}+\mathbf{Z})\cdot\mathbf{W}) \, d\omega \, d\tau \\ &- Ra^* \int_0^t \int_\Omega \beta(\theta)\mathbf{W}\cdot\mathbf{e}_z \, d\omega \, d\tau = 0.\end{aligned} \tag{11b}$$

Note that here we first compute $\mathbf{V}$; the pressure p appears by using a De Rahm type theorem.

3. Main results.

3.1. Existence of a weak solution. The main result of this work is the following.

THEOREM 1. *Under the assumptions* (A1), $\cdots$, (A5) *for each* θ_0, *in* $L^\infty(\Omega)$, *there is at least one weak solution in the Darcy model such that* $\theta(0) = \theta_0$.

Under the assumption (A1), $\cdots$, (A6) *for each* $(\theta_0, \mathbf{U}_0)$ *in* $L^\infty(\Omega)\times\mathbb{H}_2$, *there is at least one weak solution in the Darcy–Forchheimer model such that* $(\theta(0), \mathbf{U}(0)) = (\theta_0, \mathbf{U}_0)$.

We supply the proof in § 4; it will be a consequence of the following theorem.

THEOREM 2. *Under the assumptions* (A1), $\cdots$, (A4), (A6), *and* (A7),

$$\exists\alpha_6 \in \mathbb{R}_+ \quad \forall\mathbf{V}\in\mathbb{R}^3, \quad |D(\mathbf{V})| \leqq \alpha_6, \tag{A7}$$

there is at least one solution satisfying the following estimates.

There exist positive constants k_l and positive continuous functions f_l independent of D as long as it satisfies (A4), *but depending on the data h_i, g_i, g_e, and θ_0 or (θ_0, U_0) such that*

$$|\theta|_{L^\infty(\mathbb{R}_+\times\Omega)} \leqq k_1, \tag{13}$$

$$\int_0^t |\nabla\theta(\tau)|^2_{L^2(\Omega)}\, d\tau \leqq f_1(t), \tag{14}$$

$$|\mathbf{V}|_{L^\infty(\mathbb{R}_+;\mathbb{H}_2)} \leqq k_2. \tag{15}$$

Moreover, there exists q greater than 2 depending only on k_1 such that for the Darcy model

$$|\mathbf{V}|_{L^\infty(\mathbb{R}_+;\mathbb{H}_q)} \leqq k_3, \tag{16}$$

while for the Darcy-Forchheimer model,

$$\int_0^t |\nabla\mathbf{V}(\tau)|^3_{\mathbb{H}_3}\, d\tau \leqq f_2(t).$$

3.2. Uniqueness results.

THEOREM 3. *Assuming λ, D, μ to be constant and β Lipschitz continuous, then the weak solution is unique (for either models).*

The sketch of the proof follows from Gronwall's inequality as in [4], to which we refer for further details.

Now we give, for the Darcy model, uniqueness theorems assuming some regularity on the data and the solution. We are not actually able to prove that weak solutions have such regularity properties.

THEOREM 4. *Suppose that λ, μ, β, D are Lipschitz continuous. Assume there exist two weak solutions $(\theta_l, \mathbf{V}_l, p_l)_{l=1,2}$ for the Darcy model verifying*

$$\theta_l \in L^2((0,t);\ W^{1,\infty}(\Omega)) \cap L^\infty(\mathbb{R}_+\times\Omega),$$
$$p_l \in L^\infty((0,t);\ W^{1,\infty}(\Omega));$$

then $\theta_1(0) = \theta_2(0)$ imply $\theta_1 = \theta_2$, $p_1 = p_2$, and $\mathbf{V}_1 = \mathbf{V}_2$.

Remark. In fact, we have the following continuous dependence: for two weak solutions $(\theta_l, p_l)_{l=1,2}$ verifying

$$\theta_l \in L^2((0,t);\ W^{1,\infty}(\Omega))$$
$$p_l \in L^\infty((0,t);\ W^{1,\infty}(\Omega)),$$

there exists a nonnegative function $f(t)$ in $L^1_{\text{loc}}(\mathbb{R}_+)$ such that

$$|\theta_1(t) - \theta_2(t)|_{L^2(\Omega)} \leqq c_0 \exp\left(\int_0^t f(\tau)\, d\tau\right) |\theta_1(0) - \theta_2(0)|_{L^2(\Omega)}.$$

THEOREM 5. *Assuming D satisfies* (A7), *suppose, moreover, that λ, μ, β are Lipschitz continuous and there exists a solution (θ_1, p_1) verifying*

$$\theta_1 \in L^2((0,t);\quad W^{1,\infty}(\Omega)) \cap L^\infty((0,t)\times\Omega),$$
$$p_1 \in L^\infty((0,t);\quad W^{1,\infty}(\Omega));$$

then for the Darcy model any weak solution $(\theta, \mathbf{V}, p)$ satisfying $\theta(0) = \theta_1(0)$ is equal to $(\theta_1, \mathbf{V}_1, p_1)$.

3.3. Remark on the large time behavior for the Darcy model. If D satisfies (A7), and if the boundary data h_i is constant on Σ_i, then we have

$$\lim_{t\to\infty} |\theta(t)-h_i|_{L^2(\Omega)}=0, \qquad \lim_{t\to\infty} |V(t)-V_\infty|_{\mathbb{H}_2}=0,$$

where V_∞ is the solution of (11), (12) with $\theta=h_i$.

4. Proof of the main results; Theorems 1 and 2. We supply a detailed proof of Theorem 1 for the Darcy model, which turns out to be less regular than the Darcy-Forchheimer one. For this case, in order to get some smoothness properties for the velocity, it is convenient to choose both pressure and temperature as main unknowns.

The proof for the Darcy-Forchheimer model follows the same lines and is sketched in § 4.3.

4.1. Preliminary results. We first prove Theorem 2. The proof goes through several steps. We use a kind of linearization process by introducing a positive delay which uncoupled the system. This is similar in spirit to the fractional step method in [16]. A similar idea, called retarded mollification, is used in [2]. This idea is also known as the method of Rothe, and it is used in [13] for quasilinear parabolic equations. We next derive a priori estimates for the approximate solutions and pass to the limit upon letting the delay go to zero.

4.1.1. Approximate solution. Let h be a small positive number.

For any measurable function $f:\mathbb{R}\times\Omega\to\mathbb{R}$, set $\tau_h f:\mathbb{R}\times\Omega\to\mathbb{R}$ as $\tau_h f(t,x)=f(t-h,x)$.

Next we introduce the following scheme.

Define first

$$\theta_h(t,x)=\theta_0(x); \quad t\in[-h,0], \quad x\in\Omega;$$

then, assuming θ_h to be known on the interval $[kh,(k+1)h]$, $k\in\mathbb{N}$, compute p_h verifying

$$p_h\in L^\infty([kh,(k+1)h];\, H^1(\Omega)/\mathbb{R})$$

such that for any test function ψ belonging to $H^1(\Omega)/\mathbb{R}$

$$\int_\Omega \mu(\theta_h)^{-1}K\cdot(\nabla p_h - Ra^*\beta(\theta_h)\mathbf{e}_z)\cdot\nabla\psi\,d\omega+\int_{\Sigma_i} g_i\psi\,d\sigma_i+\int_{\Sigma_e} g_e\psi\,d\sigma_e=0, \tag{17}$$

set $\mathbf{V}_h$ as

$$\mathbf{V}_h\in L^\infty([kh,(k+1)h];\,\mathbb{H}_2),$$
$$\mathbf{V}_h=\mu(\theta_h)^{-1}K\cdot(\nabla p_h-Ra^*\beta(\theta_h)\mathbf{e}_z);$$

then define θ_h on the interval $[(k+1)h,(k+2)h]$ verifying

$$\theta_h\in L^\infty([(k+1)h,(k+2)h];\,L^\infty(\Omega))\cap L^2([(k+1)h,(k+2)h];\,H^1(\Omega))$$

and such that for any test function φ belonging to $L^2_{\text{loc}}(\mathbb{R}_+;\,H^1(\Omega))$,

$$\begin{aligned}
&\int_{(k+1)h}^{(k+2)h}\left\langle\frac{\partial\theta_h}{\partial t},\varphi\right\rangle d\tau+\int_{(k+1)h}^{(k+2)h}\int_\Omega(\lambda(\tau_h\theta_h)+D(\tau_h\mathbf{V}_h))\cdot\nabla\theta_h\cdot\nabla\varphi\,d\omega\,d\tau\\
&\quad+\frac{1}{2}\int_{(k+1)h}^{(k+2)h}\int_\Omega(\tau_h\mathbf{V}_h\cdot\nabla\theta_h\varphi-\tau_h\mathbf{V}_h\cdot\nabla\varphi\theta_h)\,d\omega\,d\tau\\
&\quad-\int_{(k+1)h}^{(k+2)h}\int_{\Sigma_i}g_i(\theta-h_i)\varphi\,d\sigma_i\,d\tau\\
&\quad+\frac{1}{2}\int_{(k+1)h}^{(k+2)h}\int_{\Sigma_i}g_i\theta_h\varphi\,d\sigma_i\,d\tau+\frac{1}{2}\int_{(k+1)h}^{(k+2)h}\int_{\Sigma_e}g_e\theta_h\varphi\,d\sigma_e\,d\tau=0,
\end{aligned}\tag{18}$$

where $\theta_h((k+1)h,x)$ is known from the previous step.

LEMMA 1. *The algorithm is consistent and gives a global solution* $(\theta_h, p_h, \mathbf{V}_h)$, *satisfying*

$$\theta_h \in L^\infty(\mathbb{R}_+\times\Omega)\cap L^2_{\text{loc}}(\mathbb{R}_+; H^1(\Omega)),$$
$$\mathbf{V}_h \in L^\infty(\mathbb{R}_+; \mathbb{H}_2),$$
$$p_h \in L^\infty(\mathbb{R}_+; H^1(\Omega)/\mathbb{R});$$

furthermore, there exist constants k_l *and a positive and continuous function* f_1 *independent of* D *as long as it satisfies* (A4), *but depending on the data* h_i, g_i, g_e, θ_0 *such that*

$$|\theta_h|_{L^\infty(\mathbb{R}_+\times\Omega)} \leqq k_1, \tag{19}$$

$$\int_0^t |\nabla\theta_h|^2_{L^2(\Omega)}\,d\tau \leqq f_1(t), \tag{20}$$

$$|\mathbf{V}_h|_{L^\infty(\mathbb{R}_+;\mathbb{H}_2)} \leqq k_2, \tag{21}$$

$$|p_h|_{L^\infty(\mathbb{R}_+;H^1(\Omega)/\mathbb{R})} \leqq k_3. \tag{22}$$

Moreover, there exists a positive and continuous function f_2 *depending on* k_1, f_1, k_2, k_3, g_i, g_e, *and* α_6 (*see* (A7)) *such that*

$$\int_0^t \left|\frac{\partial\theta_h}{\partial t}\right|^2_{(H^1(\Omega)')} d\tau \leqq f_2(t). \tag{23}$$

The lengthy proof of this statement is postponed until § 7.

4.1.2. Regularity results for the velocity. Let $(\theta_n)_{n\geqq 0}$ be a sequence of bounded function over Ω. Define $\mathbf{V}_n$ as the solution of the Darcy equation with data θ_n,

$$\mathbf{V}_n = \mu(\theta_n)^{-1}K\cdot(\nabla\pi_n - Ra^*\beta(\theta_n)\mathbf{e}_z), \tag{24}$$

where π_n is the unique solution in $H^1(\Omega)/\mathbb{R}$ to

$$\nabla\cdot(\mu(\theta_n)^{-1}K\cdot(\nabla\pi_n - Ra^*\beta(\theta_n)\mathbf{e}_z)) = 0, \tag{25}$$

$$\begin{aligned}&\mu(\theta_n)^{-1}K\cdot(\nabla\pi_n - Ra^*\beta(\theta_n)\mathbf{e}_z)\cdot\boldsymbol{\eta} = 0 \quad \text{on } \partial\Omega\backslash\Sigma_i\cup\Sigma_e,\\ &\mu(\theta_n)^{-1}K\cdot(\nabla\pi_n - Ra^*\beta(\theta_n)\mathbf{e}_z)\cdot\boldsymbol{\eta} = g_j \quad \text{on } \Sigma_j \quad j=i \quad \text{or} \quad e.\end{aligned} \tag{26}$$

We are going to prove the following.

LEMMA 2. *Assume the previous sequence* $(\theta_n)_{n\geqq 0}$ *is bounded in* $L^\infty(\Omega)$ *by a constant* M *and converges strongly to some function* θ *in* $L^2(\Omega)$; *then there exists* q *greater than two, depending only on* M *such that* $(\mathbf{V}_n)_{n\geqq 0}$ *converges strongly to* $\mathbf{V}$ *in* $L^q(\Omega)$, *where* $\mathbf{V}$ *is the unique solution of the Darcy equation with data* θ.

Proof. Let G be the unique solution in $H^1(\Omega)/\mathbb{R}$ of the variational problem.

For any test function φ belonging to $H^1(\Omega)/\mathbb{R}$,

$$\int_\Omega \nabla G\cdot\nabla\varphi\,d\omega = \int_{\Sigma_i} g_i\varphi\,d\sigma_i + \int_{\Sigma_e} g_e\varphi\,d\sigma_e,$$

where g_j are as in (A2).

According to Lions-Magenes [9], G belongs to $W^{1,p}(\Omega)$ for any p greater than 1. Thus (25), (26) become

$$\int_\Omega \mu(\theta_n)^{-1}\mathbf{K}\cdot\nabla\pi_n\cdot\nabla\varphi\,d\omega = \int_\Omega (Ra^*\mu(\theta_n)^{-1}\beta(\theta_n)\mathbf{e}_z + \nabla G)\cdot\nabla\varphi\,d\omega. \tag{27}$$

Let $\hat{G}$ be a function of $W^{1,p}(\mathbb{R}^3)$ with compact support whose restriction to Ω coincides with G. Consider, then, $\tilde{\theta}_n$ the extension by zero of θ_n to $\hat{\Omega}$, where $\hat{\Omega}$ is an open domain in $\mathbb{R}^3$ containing $\bar{\Omega}$ and supp $\hat{G}$.

Consider the variational problem.

For each φ of $H^1(\hat{\Omega})/\mathbb{R}$.

$$\int_{\hat{\Omega}} \mu(\tilde{\theta}_n)^{-1} K \cdot \nabla \tilde{\pi}_n \cdot \nabla \varphi \, d\omega = \int_{\Omega} (Ra^* \mu(\tilde{\theta}_n)^{-1} \beta(\tilde{\theta}_n) \mathbf{e}_z + \nabla \hat{G}) \cdot \nabla \varphi \, d\omega. \tag{28}$$

By local Meyers regularity theorem [1], [10] there exists p_0 greater than 2, depending only on the ellipticity constant of $\mu(\tilde{\theta}_n)^{-1}K$ and on $|\mu(\tilde{\theta}_n)^{-1}K|_{L^\infty(\mathbb{R}^3)}$ such that $\tilde{\pi}_n$ belongs to $W^{1,p}_{\text{loc}}(\hat{\Omega})$.

The equation (28) being in divergence form on both sides and the boundary conditions being natural ones, π_n coincides with $\tilde{\pi}_{n|\Omega}$.

Remark. When θ_n does not belong to $\mathscr{C}^0(\bar{\Omega})$, we cannot make sure that π_n belongs to $W^{1,p}(\Omega)$ for any p.

Let π be the solution of the Darcy equation with data θ; taking the difference of the equation for π_n and π, we obtain

$$\begin{aligned} |\nabla(\pi_n - \pi)|_{L^2(\Omega)} &\leqq \alpha_0 |(\mu(\theta_n)^{-1} - \mu(\theta)^{-1}) \nabla \pi|_{L^2(\Omega)} \\ &\quad + \alpha_1 |\beta(\theta_n)\mu(\theta_n)^{-1} - \beta(\theta)\mu(\theta)^{-1}|_{L^2(\Omega)} \end{aligned} \tag{29}$$

for some constant α_0, α_1 independent of n.

By assumption and the Lebesgue dominated convergence theorem, we have that $(\theta_n, \mu(\theta_n), \beta(\theta_n))_{n\geqq 0}$ converges strongly for each finite q to $(\theta, \mu(\theta), \beta(\theta))$.

Choose then q such that $(1/q)+(1/p_0)=\frac{1}{2}$; then from (29) π_n converges to π strongly in $H^1(\Omega)/\mathbb{R}$.

Furthermore, $(\nabla \pi_n)_{n\geqq 0}$ is bounded in $L^{p_0}(\Omega)$; so for each small positive ε, $(\pi_n)_n$ converges to π, strongly in $W^{1,p_0-\varepsilon}(\Omega)/\mathbb{R}$.

By definition of $(\mathbf{V}_n)_{n\geqq 0}$, $(\mathbf{V}_n)_{n\geqq 0}$ is strongly converging to $\mathbf{V}$ in $L^q(\Omega)$ for each q satisfying $2 \leqq q < p_0$.

4.1.3. End of the proof of Theorem 2. When D is bounded, we have built a family of functions $(\theta_h, \mathbf{V}_h, \pi_h)$, verifying the uniform estimates of Lemma 1, solutions of the following.

For any test function φ in $H^1(\Omega)$,

$$\begin{aligned} &\left\langle \frac{\partial \theta_h}{\partial t}, \varphi \right\rangle + \int_\Omega (\lambda(\tau_h \theta_h) + D(\tau_h \mathbf{V}_h)) \cdot \nabla \theta_h \cdot \nabla \varphi \, dx \\ &\quad + \int_\Omega \tau_h \mathbf{V}_h \cdot \nabla \varphi \theta_h \, dx + \int_{\Sigma_i} g_i h_i \varphi \, d\sigma_i \\ &\quad + \int_{\Sigma_e} g_e \theta_h \varphi \, d\sigma_e = 0 \quad \text{a.e. in } (0, t); \end{aligned} \tag{30}$$

for any test function ψ in $H^1(\Omega)$,

$$\begin{aligned} &\int_\Omega \mu(\theta_h)^{-1} K \cdot (\nabla p_h - Ra^* \beta(\theta_h) \mathbf{e}_z) \cdot \nabla \psi \, d\omega \\ &\quad + \int_{\Sigma_i} g_i \psi \, d\sigma_i + \int_{\Sigma_e} g_e \psi \, d\sigma_e = 0 \quad \text{a.e. in } (0, t). \end{aligned} \tag{31}$$

Upon extracting a sequence, we may assume that

$$\begin{array}{lll} \theta_h \to \theta & \text{in } L^\infty((0,t)\times\Omega) & \text{weak star,} \\ \theta_h \to \theta & \text{in } L^2((0,t)\times\Omega) & \text{weakly,} \\ p_h \to p & \text{in } L^\infty((0,t);\, H^1(\Omega)/\mathbb{R}) & \text{weak star,} \\ \dfrac{\partial \theta_h}{\partial t} \to \dfrac{\partial \theta}{\partial t} & \text{in } L^2((0,T);\, H^1(\Omega)') & \text{weakly;} \end{array}$$

using Aubin's lemma,

$$\begin{array}{lll} \theta_h \to \theta & \text{in } L^2((0,t)\times\Omega) & \text{strongly and a.e.,} \\ \theta_h \to \theta & \text{in } L^p((0,t)\times\Omega) & \text{strongly for each finite } p; \end{array}$$

then, according to Lemma 2 there exist r and q greater than two such that

$$\begin{array}{lll} p_h \to p & \text{in } L^r(0,t);\, W^{1,r}(\Omega)/\mathbb{R}) & \text{strongly,} \\ \mathbf{V}_h \to \mathbf{V} & \text{in } \mathbb{L}^q((0,t);\, \mathbb{H}_q) & \text{strongly.} \end{array}$$

Noting that translations are strongly continuous in $L^p(0, T)$ for each finite p, we show that $(\theta, \mathbf{V}, \pi)$ is a solution of the problem satisfying $\theta(0) = \theta_0$.

This completes the proof of Theorem 2.

4.2. Proof of Theorem 1. We approximate a general D satisfying (A4), (A5) by a sequence of bounded operator D^l defined in the following way:

$$D^l_{ij}(\mathbf{V}) = \operatorname{Min}\,(l,\, D_{ij}(\mathbf{V})).$$

Now by Theorem 2 for each l there exists a triplet $(\theta_l, \mathbf{V}_l, \pi_l)$ verifying the following.

For each test function φ in $H^1(\Omega)$.

$$\begin{aligned} &\left\langle \frac{\partial \theta_l}{\partial t}, \varphi \right\rangle + \int_\Omega (\lambda(\theta_l) + D^l(\mathbf{V}_l)) \cdot \nabla \theta_l \cdot \nabla \varphi \, d\omega \\ &\quad + \int_\Omega \mathbf{V}_l \cdot \nabla \varphi \theta_l \, d\omega + \int_{\Sigma_i} g_i h_i \varphi \, d\sigma_i \qquad (32) \\ &\quad + \int_{\Sigma_e} g_e \theta_l \varphi \, d\sigma_e = 0 \quad \text{a.e. in } (0,t); \end{aligned}$$

for each test function ψ in $H^1(\Omega)$,

$$\begin{aligned} &\int_\Omega \mu(\theta_l)^{-1} K \cdot (\nabla \pi_l - Ra^* \beta(\theta_l) \mathbf{e}_z) \cdot \nabla \psi \, d\omega \qquad (33) \\ &\quad + \int_{\Sigma_i} g_i \psi \, d\sigma_i + \int_{\Sigma_e} g_e \psi \, d\sigma_e = 0 \quad \text{a.e. in } (0,t). \end{aligned}$$

The sequence $(\theta_l, \mathbf{V}_l, \pi_l)$ satisfies uniformly the estimates (19), (20), (21), (22) of Lemma 1 because each D^l satisfies the assumption (A4).

The estimate on $(\partial\theta_l/\partial t)$ in $L^2((0,t);\,(H^1(\Omega))')$ may not be uniform; so we are to prove a weaker uniform estimate, namely:

There exists a positive function f, independent of l such that

$$\left| \frac{\partial \theta_l}{\partial t} \right|_{L^2((0,t);(H^3(\Omega))')} \leqq f(t). \qquad (34)$$

Toward this end we remark that $D^l(\mathbf{V}_l)\nabla\theta_l$ is uniformly bounded in $\mathbb{L}^2((0, t); L^1(\Omega))$; so from (32) and by the previous estimates it follows that the sequence $(\partial\theta_l/\partial t)_{l\geqq 0}$ is uniformly bounded in $L^2((0, t); H^3(\Omega)')$.

Then we can apply again Aubin's lemma and conclude as in Theorem 2 to the convergence of a subsequence $(\theta_{l_r}, \mathbf{V}_{l_r}, \pi_{l_r})$ toward a weak solution of the original problem.

Remark. We also obtain that $\partial\theta/\partial t$ belongs to $L^2((0, t); W^{-1,1}(\Omega))$.

4.3. Main estimates for the Darcy–Forchheimer model. Let again $\mathbf{Z}$ be a function of $\mathbb{H}_3$ such that

$$\nabla\cdot\mathbf{Z}=0,$$

$$\mathbf{Z}\cdot\boldsymbol{\eta}=0 \quad \text{for } z=0 \quad \text{and} \quad z=H,$$

$$\mathbf{Z}\cdot\boldsymbol{\eta}=g_j \quad \text{for } r=R_j \quad j=i \text{ or } e.$$

Next take $\mathbf{U}=\mathbf{V}-\mathbf{Z}$; $\mathbf{U}$ verify $\mathbf{U}\cdot\boldsymbol{\eta}=0$ on $\partial\Omega$.

Now, as in the Darcy model we introduce the following scheme.

Define first

$$\theta_h(t, x)=\theta_0(x), \qquad t\in[-h, 0],$$

$$\mathbf{U}_h(t, x)=\mathbf{U}_0(x), \qquad t=0;$$

then assuming θ_h to be known on the interval $[kh, (k+1)h]$, $k\in\mathbb{N}$, compute $\mathbf{U}_h$ verifying

$$\mathbf{U}_h\in L^\infty([kh, (k+1)h]; \mathbb{H}_2)\cap L^3([kh, (k+1)h]; \mathbb{H}_3)$$

such that for any test function $\mathbf{W}$ in $L^3_{\text{loc}}(\mathbb{R}^+; \mathbb{H}_3)$,

$$\begin{aligned}
\varepsilon^{-1}DaP_r^{-1}&\int_{kh}^{(k+1)h}\int_\Omega \frac{\partial\mathbf{U}_h}{\partial t}\cdot\mathbf{W}\,d\omega\,d\tau\\
&+\int_{kh}^{(k+1)h}\int_\Omega K^{-1}\mu(\theta_h)(\mathbf{U}_h+\mathbf{Z})\cdot\mathbf{W}\,d\omega\,d\tau\\
&+c_jDa^{1/2}P_r^{-1}\int_{kh}^{(k+1)h}\int_\Omega \sigma(\mathbf{U}_h+\mathbf{Z})\cdot\mathbf{W}\,d\omega\,d\tau\\
&-Ra^*\int_{kh}^{(k+1)/h}\int_\Omega \beta(\theta_h)e_z\mathbf{W}\,d\omega\,d\tau=0,
\end{aligned}$$

where $U_h(kh, x)$ is known from the previous step.

Set $p_h\in L^{3/2}([kh, (k+1)h]; W^{1,3/2}(\Omega)/\mathbb{R})$ such that for any test function ψ belonging to $W^{1,3/2}(\Omega)/\mathbb{R}$,

$$\begin{aligned}
\int_\Omega \nabla p_h\cdot\nabla\psi\,d\omega = &-\int_\Omega K^{-1}\mu(\theta)(\mathbf{U}_h+\mathbf{Z})\nabla\psi\,d\omega\\
&-c_jDa^{1/2}P_r^{-1}\int_\Omega \boldsymbol{\sigma}(\mathbf{U}_h+\mathbf{Z})\mathbf{W}\,d\omega\,d\tau\\
&+Ra^*\int_\Omega \beta(\theta)e_z\cdot\nabla\psi\,d\omega.
\end{aligned}$$

Then define θ_h on the interval $[(k+1)h, (k+2)h]$ verifying

$$\theta_h\in L^\infty([(k+1)h, (k+2)h]; L^\infty(\Omega))\cap L^2([(k+1)h, (k+2)h]; H^1(\Omega))$$

and such that for any test function φ belonging to $L^2_{\text{loc}}(\mathbb{R}_+;\, H^1(\Omega))$,

$$
\begin{aligned}
&\int_{(k+1)h}^{(k+2)h}\left\langle\frac{\partial\theta_h}{\partial t},\varphi\right\rangle d\tau+\int_{(k+1)h}^{(k+2)h}\int_\Omega(\lambda(\tau_h\theta_h)+D(\tau_h\mathbf{V}_h))\cdot\nabla\theta_h\cdot\nabla\varphi\,d\omega\,d\tau\\
&+\frac{1}{2}\int_{(k+1)h}^{(k+2)h}\int_\Omega(\tau_h\mathbf{V}_h\cdot\nabla\theta_h\varphi-\tau_h\mathbf{V}_h\cdot\nabla\varphi\theta h)\,d\omega\,d\tau\\
&-\int_{(k+1)h}^{(k+2)h}\int_{\Sigma_i}g_i(\theta-h_i)\varphi\,d\sigma_i\,d\tau\\
&+\frac{1}{2}\int_{(k+1)h}^{(k+2)h}\int_{\Sigma_i}g_i\theta_h\varphi\,d\sigma_i\,d\tau+\frac{1}{2}\int_{(k+1)h}^{(k+2)h}\int_{\Sigma_e}g_e\theta_h\varphi\,d\sigma_e\,d\tau=0.
\end{aligned}\tag{18}
$$

LEMMA 3. *The algorithm is consistent and gives a global solution* $(\theta_h, p_h, \mathbf{U}_h)$ *satisfying*

$$
\begin{aligned}
&\theta_h\in L^\infty(\mathbb{R}_+\times\Omega)\cap L^2_{\text{loc}}(\mathbb{R}_+;\, H^1(\Omega)),\\
&\mathbf{U}_h\in L^3_{\text{loc}}(\mathbb{R}_+;\, \mathbb{H}_3)\cap L^\infty(\mathbb{R}_+;\, \mathbb{H}_2),\\
&p_h\in L^{3/2}_{\text{loc}}(\mathbb{R}_+;\, W^{1,3/2}_{\text{loc}}(\Omega)/\mathbb{R});
\end{aligned}
$$

furthermore, there exist constants k_l and positive and continuous functions f_l independent of D as long as it satisfies (A4), *but depending on the data h_i, g_i, g_e, θ_0, U_0 such that*

$$|\theta_h|_{L^\infty(\mathbb{R}_+\times\Omega)}\leqq k_1,$$

$$\int_0^t|\nabla\theta_h|_{L^2(\Omega)}\,d\tau\leqq f_1(t),$$

$$|\mathbf{U}_h|_{L^\infty(\mathbb{R}^+;\mathbb{H}_2)}\leqq k_2,$$

$$\int_0^t|\mathbf{U}_h|^3_{\mathbb{H}_3}\,d\tau\leqq f_2(t),$$

$$\int_0^t|\nabla p_h|^{3/2}_{L^{3/2}(\Omega)}\,d\tau\leqq f_3(t),$$

$$\int_0^t\left|\frac{\partial\mathbf{U}_h}{\partial t}\right|^{3/2}_{\mathbb{H}_{3/2}}d\tau\leqq f_4(t).$$

Moreover, there exists a positive and continuous function f_5 depending on k_1, f_1, k_2, f_2, g_i, g_e, and α_6 (see A7) *such that*

$$\int_0^t\left|\frac{\partial\theta_h}{\partial t}\right|^2_{H^1(\Omega)'}d\tau\leqq f_5(t).$$

The existence of a global solution u_h is a classical consequence of the hypothesis (A6); *the estimate of U_h obtained by taking $W=\mathbf{U}_h$ as a test function is the Darcy-Forchheimer equation, using Young's inequalities.*

5. Uniqueness results.

5.1. Proof of Theorem 4. Let $(\theta_l, \mathbf{V}_l, p_l)$ be two weak solutions corresponding to the same data $(h_i, g_i, g_e, \theta_0)$, $l=1$, or 2.

Set $\theta=\theta_2-\theta_1$, $p=p_2-p_1$, $\mathbf{V}=\mathbf{V}_2-\mathbf{V}_1$.

Taking the difference of the equations for θ_1 and θ_2, we get, after straightforward computations,

$$\begin{aligned}
&\left\langle \frac{\partial\theta}{\partial t}, \varphi \right\rangle_{W^{-1,1}(\Omega), W^{1,\infty}(\Omega)} + \int_\Omega (\lambda(\theta_2)+D(\mathbf{V}_2)) \cdot \nabla\theta \cdot \nabla\varphi \, d\omega \\
&\quad -\frac{1}{2}\int_{\Sigma_i} g_i\theta\varphi \, d\sigma_i + \frac{1}{2}\int_{\Sigma_e} g_e\theta\varphi \, d\sigma_e + \frac{1}{2}\int_\Omega (\mathbf{V}_2 \cdot \nabla\theta\varphi - \mathbf{V}_2 \cdot \nabla\varphi\theta) \, d\omega \\
(35) \qquad &\quad = -\int_\Omega (\lambda(\theta_2)+D(\mathbf{V}_2)) - (\lambda(\theta_1)+D(\mathbf{V}_1)) \cdot \nabla\theta_1 \cdot \nabla\varphi \, d\omega \\
&\quad -\frac{1}{2}\int_\Omega (\mathbf{V} \cdot \nabla\theta_1\varphi - \mathbf{V} \cdot \nabla\varphi\theta_1) \, d\omega.
\end{aligned}$$

We may now take $\varphi = \theta$ as θ belongs to $L^2((0, t); W^{1,\infty}(\Omega)) \cap L^\infty(\mathbb{R}_+ \times \Omega)$ to get the following energy inequality:

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|\theta|^2_{L^2(\Omega)} + c_0|\nabla\theta|^2_{\mathbb{L}^2(\Omega)} - \frac{1}{2}\int_{\Sigma_i} g_i\theta^2 \, d\sigma_i + \frac{1}{2}\int_{\Sigma_e} g_e\theta^2 \, d\sigma_e \\
(36) \qquad &\quad \leqq \int_\Omega (c_1|\theta_2-\theta_1| + c_2|\mathbf{V}_2-\mathbf{V}_1|)|\nabla\theta_1||\nabla\theta| \, d\omega \\
&\quad + \int_\Omega |\mathbf{V}|(|\mathbf{V}\theta_1||\theta| + |\nabla\theta||\theta_1|) \, d\omega.
\end{aligned}$$

Using Hölder inequalities, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|\theta|^2_{L^2(\Omega)} + c_0|\nabla\theta|^2_{\mathbb{L}^2(\Omega)} - \frac{1}{2}\int_{\Sigma_i} g_i\theta^2 \, d\sigma_i + \frac{1}{2}\int_{\Sigma_e} g_e\theta^2 \, d\sigma_\epsilon \\
(37) \qquad &\quad \leqq \frac{c_0}{4}|\nabla\theta|^2_{\mathbb{L}^2(\Omega)} + c_3|\theta|^2_{L^2(\Omega)}|\nabla\theta_1|^2_{\mathbb{L}^\infty(\Omega)} + c_4|\mathbf{V}|^2_{\mathbb{H}_2}|\nabla\theta_1|^2_{\mathbb{L}^\infty(\Omega)} \\
&\quad + |\mathbf{V}|_{\mathbb{H}_2}|\theta|^2_{L^2(\Omega)}|\nabla\theta_1|^2_{\mathbb{L}^\infty(\Omega)} + \frac{c_0}{4}|\nabla\theta|^2_{\mathbb{L}^2(\Omega)} + c_5|\theta_1|^2_{L^\infty(\Omega)}|\mathbf{V}|^2_{\mathbb{H}_2}.
\end{aligned}$$

So there exists a function f_1 belonging to $L^1(0, t)$ such that

$$\begin{aligned}
(38) \qquad &\frac{1}{2}\frac{d}{dt}|\theta|^2_{L^2(\Omega)} + \frac{c_0}{2}|\nabla\theta|^2_{\mathbb{L}^2(\Omega)} - \frac{1}{2}\int_{\Sigma_i} g_i\theta^2 \, d\sigma_i + \frac{1}{2}\int_{\Sigma_e} g_e\theta^2 \, d\sigma_e \\
&\quad \leqq f_1(t)(|\theta|^2_{L^2(\Omega)} + |\mathbf{V}|^2_{\mathbb{H}_2}).
\end{aligned}$$

To obtain an estimate on the velocity $\mathbf{V} = \mathbf{V}_2 - \mathbf{V}_1$, we first prove a bound for the pressure gradient $\nabla p = \nabla p_2 - \nabla p_1$.

Taking the difference of the equations for p_2 and p_1, we get, after some calculations,

$$\begin{aligned}
\int_\Omega \mu(\theta_2)^{-1}K \cdot \nabla p \cdot \nabla\psi \, d\omega &= \int_\Omega (\mu(\theta_2)^{-1} - \mu(\theta_1)^{-1})K \cdot \nabla p_1 \cdot \nabla\psi \, d\omega \\
&\quad + Ra^* \int_\Omega (\beta(\theta_2)\mu(\theta_2)^{-1} - \beta(\theta_1)\mu(\theta_1)^{-1})\mathbf{e}_z \cdot \nabla\psi \, d\omega.
\end{aligned}$$

By taking $\psi = p$, we find

$$c_6|\nabla p|^2_{L^2(\Omega)} \leqq c_7|\nabla p_1|_{\mathbb{L}^\infty(\Omega)}|\theta|_{L^2(\Omega)}|\nabla p|_{\mathbb{L}^2(\Omega)} + c_8|\theta|_{L^2(\Omega)}|\nabla p|_{\mathbb{L}^2(\Omega)}.$$

So there exists a function f_2 belonging to $L^\infty(0, t)$ such that

$$(39) \qquad |\nabla p|_{\mathbb{L}^2(\Omega)} \leqq f_2(t)|\theta|_{L^2(\Omega)}.$$

According to the Darcy law, we get

$$\mathbf{V} = -\mu(\theta_2)^{-1} K \nabla p - (\mu(\theta_2)^{-1} - \mu(\theta_1)^{-1}) K \nabla p_1 \\ + Ra^*(\mu(\theta_2)^{-1}\beta(\theta_2) - \mu(\theta_1)^{-1}\beta(\theta_1))\mathbf{e}_z.$$

So we have for some functions f_3 or f_4 belonging $L^\infty(0, t)$,

(40) $$|\mathbf{V}|_{\mathbb{H}_2} \leqq f_3(t) |\nabla p|_{L^2(\Omega)} + f_4(t) |\theta|_{L^2(\Omega)}.$$

So according to the inequality (39), there exists a function f_5 in $L^\infty(0, t)$,

(41) $$|\mathbf{V}|_{\mathbb{H}_2} \leqq f_5(t) |\theta|_{L^2(\Omega)}.$$

By taking this last inequality in the energy equation (38), we prove the existence of a function f in $L^1(0, t)$ such that

$$\frac{d}{dt}|\theta|^2_{L^2(\Omega)} + c_0 |\nabla\theta|^2_{L^2(\Omega)} - \int_{\Sigma_i} g_i \theta^2 \, d\sigma_i + \int_{\Sigma_e} g_e \theta^2 \, d\sigma_e \leqq f(t)(|\theta|^2_{L^2(\Omega)}).$$

According to the assumption (A2), we derive, by using the Gronwall lemma, the uniqueness of a solution.

5.2. Proof of Theorem 5. When D satisfies (A7), the equality (35) has the form

$$\left\langle \frac{\partial\theta}{\partial t}, \varphi \right\rangle_{W^{-1,1}(\Omega), W^{1,\infty}(\Omega)} + \int_\Omega (\lambda(\theta_2) + D(\mathbf{V}_2)) \cdot \nabla\theta \cdot \nabla\varphi \, d\omega \\ -\frac{1}{2}\int_{\Sigma_i} g_i \theta\varphi \, d\sigma_i + \frac{1}{2}\int_{\Sigma_e} g_e \theta\varphi \, d\sigma_e + \frac{1}{2}\int_\Omega (\mathbf{V}_2 \nabla\theta\varphi - \mathbf{V}_2 \nabla\varphi\theta) \, d\omega \\ = -\int_\Omega (\lambda(\theta_2) - \lambda(\theta_1) + D(\mathbf{V}_2) - D(\mathbf{V}_1)) \cdot \nabla\theta_1 \cdot \nabla\varphi \, d\omega \\ -\frac{1}{2}\int_\Omega ((\mathbf{V} \cdot \nabla\theta_1)\varphi - (\mathbf{V} \cdot \nabla\varphi)\theta_1) \, d\omega.$$

Assuming only that θ_1 belongs to $L^2((0, t), W^{1,\infty}(\Omega))$, this equality remains valid for any test function in $L^2((0, t), H^1(\Omega))$, and by taking $\varphi = \theta$ we get the inequality (38).

The bound for $\mathbf{V}$ follows the same lines as in Theorem 4, and uniqueness is deduced by using Gronwall's lemma.

6. Proof of the large time behavior.

When D satisfies (A7), we may use Theorem 2 in § 4.1. This implies that $\partial\theta/\partial t$ belongs to $L^2((0, t), H^1(\Omega)')$. When h_i is constant, the function $T = \theta - h_i$ satisfies the following equation.

For any test function in $H^1(\Omega)$,

$$\left\langle \frac{\partial T}{\partial t}, \varphi \right\rangle + \int_\Omega (\lambda(\theta) + D(\mathbf{V})) \cdot \nabla T \cdot \nabla\varphi \, d\omega \\ + \int_\Omega (\mathbf{V} \cdot \nabla T)\varphi - \mathbf{V} \cdot \nabla\varphi T \, d\omega - \int_{\Sigma_i} g_i T\varphi \, d\sigma_i = 0.$$

Then taking $\varphi = T$ in the above equation, we get the following inequality:

$$\frac{1}{2}\frac{d}{dt}|T|^2_{L^2(\Omega)} + c_0 |\nabla T|^2_{\mathbb{L}^2(\Omega)} + m_i \int_{\Sigma_i} |T|^2 \, d\sigma_i \leqq 0.$$

Because the two-dimensional Lebesgue measure of Σ_i is positive, Poincaré's inequality gives

$$\frac{d}{dt}|T|^2_{L^2(\Omega)}+\beta|T|^2_{L^2(\Omega)}\leqq 0 \quad \text{for some constant } \beta>0,$$

and, therefore,

$$\lim_{t\to\infty}|T(t)|_{L^2(\Omega)}=0.$$

7. Proof of Lemma 1. The equations we are to solve are linear equations. We omit the index h, which is a fixed positive number throughout this section.

We introduce the bilinear and continuous mapping $\alpha_2(\cdot,\cdot)$ defined on $H^1(\Omega)\times H^1(\Omega)$ by

$$a_2(p,\psi)=\int_\Omega \frac{K}{\mu(\theta)}\nabla p\cdot\nabla\psi\,d\omega.$$

Next $b_2(\theta,\psi)$ is the bilinear and continuous mapping defined on $L^2(\Omega)\times H^1(\Omega)$ by

$$b_2(\theta,\psi)=\int_\Omega \frac{K}{\mu(\theta)}\beta(\theta)\mathbf{e}_z\cdot\nabla\psi\,d\omega.$$

7.1. Solving the Darcy equation.

LEMMA 4. *Assume* (A3) *holds. Let* $\theta\in L^\infty(\mathbb{R}_+;L^\infty(\Omega))$; *there exists a unique p such that*

$$p\in L^\infty(\mathbb{R}_+;H^1(\Omega)/\mathbb{R})$$

and a solution of

$$a_2(p,\psi)+b_2(\theta,\psi)+\int_{\Sigma_1}g_1\psi\,d\sigma_2+\int_{\Sigma_2}g_2\psi\,d\sigma_2=0 \quad \forall\psi\in H^1(\Omega). \tag{42}$$

Furthermore, there is a constant c_0 *depending only on* g_i, g_e, α_1, *and* $|\theta|_{L^\infty(\mathbb{R}_+\times\Omega)}$ *such that* $|\nabla p|_{L^2(\Omega)}\leqq c_0$.

Proof. Because θ is fixed, a_2 is coercive on $H^1(\Omega)/\mathbb{R}$ while b_2 is linear. Hence, existence and uniqueness follows from the Lax-Milgram theorem once we have checked that

$$b_2(\theta,1)+\int_{\Sigma_i}g_i\cdot 1\,d\sigma_i+\int_{\Sigma_e}g_e\cdot 1\,d\sigma_e=0 \qquad \text{(A2)}.$$

Assume now that $\theta\in L^2_{\text{loc}}(\mathbb{R}_+;H^1(\Omega))\cap L^\infty(\mathbb{R}_+;L^\infty(\Omega))$. Then p verifies

$$\nabla\cdot\left(\frac{K}{\mu(\theta)}\cdot\nabla p\right)=Ra^*\nabla\cdot K\cdot\left(\frac{\beta(\theta)}{\mu(\theta)}\mathbf{e}_z\right) \quad \text{in } \mathscr{D}'(\Omega). \tag{43}$$

Now that θ is lying in $L^2_{\text{loc}}(\mathbb{R}^+;H^1(\Omega))\cap L^\infty(\mathbb{R}^+;L^\infty(\Omega))$, the right-hand side of (43) belongs to $L^2_{\text{loc}}(\mathbb{R}^+;L^2(\Omega))$; hence we may define

$$\left(\frac{K}{\mu(\theta)}\nabla p\right)\cdot\boldsymbol{\eta}$$

as the normal trace on $\partial\Omega$ of an $L^2(\Omega)$ function having a square integrable divergence.

We may now use Green's formula to get

$$\left(\frac{K}{\mu(\theta)}\nabla p\right)\cdot \boldsymbol{\eta} = Ra^* \frac{\beta(\theta)}{\mu(\theta)} K\mathbf{e}_z \cdot \boldsymbol{\eta} \quad \text{on } z=0 \quad \text{and} \quad z=H,$$

$$(44)\qquad \begin{aligned} \left(\frac{K}{\mu(\theta)}\nabla p\right)\cdot \boldsymbol{\eta} &= Ra^* \frac{\beta(\theta)}{\mu(\theta)} K\mathbf{e}_z \cdot \boldsymbol{\eta} + g_i; \qquad r = R_i,\\ \left(\frac{K}{\mu(\theta)}\nabla p\right)\cdot \boldsymbol{\eta} &= Ra^* \frac{\beta(\theta)}{\mu(\theta)} K\mathbf{e}_z \cdot \boldsymbol{\eta} + g_e, \qquad r = R_e. \end{aligned}$$

Remark. When K is a diagonal tensor, (44) reduces to

$$(45)\qquad \begin{aligned} \left(\frac{1}{\mu(\theta)} k\nabla p\right)\cdot \boldsymbol{\eta} &= g_i \quad \text{on } r = R_i,\\ \left(\frac{1}{\mu(\theta)} k\nabla p\right)\cdot \boldsymbol{\eta} &= g_e \quad \text{on } r = R_e. \end{aligned}$$

Remark. θ is not smooth enough to make sure that $(K\cdot \nabla p)$ has a normal trace.

7.2. Solving the energy equation. Take $\mathbf{V}$ in $\mathbb{L}^\infty(\mathbb{R}^+; \mathbb{H}_2)$ with $\operatorname{div}\mathbf{V}=0$.
Choose (θ, φ) in $L^\infty(\mathbb{R}_+; L^\infty(\Omega)) \cap L^2_{\text{loc}}(\mathbb{R}_+; H^1(\Omega))$.
We then have

$$\int_\Omega (\mathbf{V}\cdot\nabla\theta)\varphi\, d\omega = -\int_\Omega (\mathbf{V}\cdot\nabla\varphi)\theta + \langle \theta\varphi, \mathbf{V}\cdot\boldsymbol{\eta}\rangle_{H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega)}$$

because any element of $H^1(\Omega)\cap L^\infty(\Omega)$ has a trace in $L^\infty(\partial\Omega)\cap H^{1/2}(\partial\Omega)$, which is actually an algebra.

In particular, we may write

$$\int_\Omega (\mathbf{V}\cdot\nabla\theta)\varphi\, d\omega = \frac{1}{2}\left[\int_\Omega (\mathbf{V}\cdot\nabla\theta)\varphi - (\mathbf{V}\cdot\nabla\varphi)\theta\, d\omega\right] + \frac{1}{2}\langle \theta\varphi, \mathbf{V}\cdot\boldsymbol{\eta}\rangle_{H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega)}.$$

For our problem, this reads

$$(46)\qquad \int_\Omega (\mathbf{V}\cdot\nabla\theta)\varphi\, d\omega = \frac{1}{2}\left[\int_\Omega (\mathbf{V}\cdot\nabla\theta)\varphi - (\mathbf{V}\cdot\nabla\varphi)\theta\, d\omega\right] + \frac{1}{2}\int_{\Sigma_i} \theta\varphi g_i\, d\sigma_i + \frac{1}{2}\int_{\Sigma_e} \theta\varphi g_e\, d\sigma_e.$$

The variational formulation of the energy equation is now

$$(47)\qquad \begin{aligned} &\left\langle \frac{\partial\theta}{\partial t}, \varphi\right\rangle + \int_\Omega \lambda(\theta)\nabla\theta\nabla\varphi\, dx + \int_\Omega D(\mathbf{V})\nabla\theta\nabla\varphi\, dx\\ &+\frac{1}{2}\int_\Omega [(\mathbf{V}\cdot\nabla\theta)\varphi - (\mathbf{V}\cdot\nabla\varphi)\theta]\, dx\\ &-\int_\Omega g_i(\theta - h_i)\varphi\, d\sigma_i + \frac{1}{2}\int_{\Sigma_e} g_e\theta\,\varphi\, d\sigma_e + \frac{1}{2}\int_{\Sigma_i} g_i\theta\varphi\, d\sigma_i = 0. \end{aligned}$$

This equation possesses at least one solution in the setting of § 2.2. To see this, we may approximate **V** by smooth functions, use Galerkin's type techniques, and pass to the limit. We omit this lengthy process. (See also [6] and [8].)

7.3. First estimates in $H^1(\Omega)$. Taking $\varphi=\theta$ in (47), which is possible because $D(\mathbf{V})$ is bounded, yields

$$\frac{1}{2}\frac{d}{dt}|\theta|^2_{L^2(\Omega)}+\int_\Omega(\lambda(\theta)\cdot\nabla\theta)\cdot\nabla\theta\,dx+\int_\Omega(D(\mathbf{V})\cdot\nabla\theta)\cdot\nabla\theta\,dx$$
$$-\int_{\Sigma_i}g_i(\theta-h_i)\theta\,d\sigma_i+\frac{1}{2}\int_{\Sigma_i}g_i\theta^2\,d\sigma_i+\frac{1}{2}\int_{\Sigma_e}g_e\theta^2\,d\sigma_e=0.$$

Thus

$$\frac{1}{2}\frac{d}{dt}|\theta|^2_{L^2(\Omega)}+\alpha_1|\nabla\theta|^2_{\mathbb{L}^2(\Omega)}-\frac{1}{2}\int_{\Sigma_i}g_i\theta^2\,d\sigma_i+\frac{1}{2}\int_{\Sigma_e}g_e\theta^2\,d\sigma_e$$
$$\leqq\int_{\Sigma_i}|g_ih_i||\theta|\,d\sigma_i.$$

Hence, using assumption (A2), we have

$$\frac{1}{2}\frac{d}{dt}|\theta|^2_{L^2(\Omega)}+\alpha_1|\nabla\theta|^2_{\mathbb{L}^2(\Omega)}+\frac{1}{2}m_1\int_{\Sigma_i}\theta^2\,d\sigma_i\leqq\int_{\Sigma_i}|g_ih_i||\theta|\,d\sigma_i. \tag{48}$$

Using Young's inequality on the right-hand side of (48), we find a constant c_1 depending only on m_i such that

$$\frac{1}{2}\frac{d}{dt}|\theta|^2_{L^2(\Omega)}+\alpha_1|\nabla\theta|^2_{\mathbb{L}^2(\Omega)}+\frac{1}{4}m_i\int_{\Sigma_i}\theta^2\,d\sigma_i\leqq c_1\int_{\Sigma_i}g_i^2h_i^2\,d\sigma_i.$$

Therefore, there exists a constant c_2 depending only on $(|g_i|_{L^\infty(\Sigma_i)},|h_i|_{L^2(\Sigma_i)},m_i)$ such that

$$|\theta(t)|^2_{L^2(\Omega)}+2\alpha_1\int_0^t|\nabla\theta(\tau)|^2_{\mathbb{L}^2(\Omega)}\,d\tau\leqq|\theta(0)|^2_{L^2(\Omega)}+c_2t. \tag{49}$$

We may conclude that for each t, θ lies in a bounded set of

$$L^\infty(0,t;L^2(\Omega))\cap L^2(0,t;H^1(\Omega)).$$

7.4. Uniform estimates in L^∞.

LEMMA 5. *Assume θ_0 and h_i are bounded. Then*

$$|\theta|_{L^\infty((0,t)\times\Omega)}\leqq\operatorname{Max}(|\theta_0|_{L^\infty(\Omega)},|h_i|_{L^\infty(\Sigma_i)}).$$

Proof. Take $\nu=\operatorname{Max}(|\theta_0|_{L^\infty(\Omega)},|h_i|_{L^\infty(\Sigma_i)})$. Let $T=\theta+\nu$; T is a solution of

$$\begin{aligned}&\left\langle\frac{\partial T}{\partial t},\varphi\right\rangle+\int_\Omega(\lambda(\theta)+D(\mathbf{V}))\nabla T\nabla\varphi\,dx+\frac{1}{2}\int_\Omega(\mathbf{V}\cdot\nabla T)\varphi-(\mathbf{V}\cdot\nabla\varphi)(T-\nu)\,dx\\&-\int_{\Sigma_i}g_i(T-\nu-h_i)\varphi\,d\sigma_i+\frac{1}{2}\int_{\Sigma_i}g_i(T-\nu)\varphi\,d\sigma_i+\frac{1}{2}\int_{\Sigma_e}g_e(T-\nu)\varphi\,d\sigma_e.\end{aligned}\tag{50}$$

Green's formula yields

$$+\frac{1}{2}\int_\Omega\mathbf{V}\cdot\nabla\varphi\nu\,dx=+\frac{1}{2}\int_{\partial\Omega}\nu\varphi\,\mathbf{V}\cdot\boldsymbol{\eta}\,d\sigma$$
$$=+\frac{1}{2}\int_{\Sigma_i}\nu\cdot\varphi g_i\,d\sigma_i+\frac{1}{2}\int_{\Sigma_e}\nu\cdot\varphi g_e\,d\sigma_e.$$

Thus we have

$$
\begin{aligned}
&\left\langle\frac{\partial T}{\partial t}, \varphi\right\rangle+\int_\Omega(\lambda(T)+D(\mathbf{V}))\cdot\nabla T\cdot\nabla\varphi\,dx\\
(51)\qquad&+\frac{1}{2}\int_\Omega(\mathbf{V}\cdot\nabla T)\varphi-(\mathbf{V}\cdot\nabla\varphi)T\,dx\\
&-\int_{\Sigma_i}g_iT\varphi\,d\sigma_i+\int_{\Sigma_i}g_i(\nu+h_i)\varphi\,d\sigma_i.
\end{aligned}
$$

Taking $\varphi=-T^-$ in this relation and writing $T=T^+-T^-$, we find, after some calculations,

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|T^-|^2_{L^2(\Omega)}+\alpha_1|\nabla T^-|^2_{L^2(\Omega)}+m_1\int_{\Sigma_i}|T^-|^2\,d\sigma_i\\
(52)\qquad&=+\int_{\Sigma_i}g_i(\nu+h_i)T^-\,d\sigma_i.
\end{aligned}
$$

We know that

$$-m_i\leqq g_i\leqq -m_i<0,\qquad -\nu-h_i<0.$$

It follows that the left-hand side of (52) is nonpositive; this implies that $T^-=0$. In a similar fashion, setting $T=\theta-\nu$ we show that $T^+=0$. Lemma 7.2 is proved.

7.5. Estimates on $\partial\theta/\partial t$. In this subsection we rewrite

$$\frac{1}{2}\int_\Omega(\mathbf{V}\cdot\nabla\theta\varphi-\mathbf{V}\cdot\nabla\varphi\theta)\,dx$$

as

$$-\int_\Omega(\mathbf{V}\cdot\nabla\varphi)\theta\,dx+\frac{1}{2}\int_{\Sigma_i}g_i\theta\varphi\,d\sigma_i+\frac{1}{2}\int_{\Sigma_e}g_e\theta\varphi\,d\sigma_e.$$

Hence θ is a solution of

$$
\begin{aligned}
&\left\langle\frac{\partial\theta}{\partial t}, \varphi\right\rangle+\int_\Omega(\lambda(\theta)+D(\mathbf{V}))\nabla\theta\cdot\nabla\varphi\,dx-\int_\Omega\mathbf{V}\cdot\nabla\varphi\theta\,dx\\
(53)\qquad&-\int_{\Sigma_i}g_i(\theta-h_i)\varphi\,d\sigma_i+\int_{\Sigma_i}g_i\theta\varphi\,d\sigma_1+\int_{\Sigma_e}g_e\theta\varphi\,d\sigma_e=0.
\end{aligned}
$$

Because θ is bounded, we get

$$
\begin{aligned}
\left|\left\langle\frac{\partial\theta}{\partial t}, \varphi\right\rangle\right|\leqq&\{|\lambda(\theta)|_{L^\infty(\Omega)}|\nabla\theta|_{L^2(\Omega)}+|D(\mathbf{V})|_{L^\infty(\Omega)}|\nabla\theta|_{L^2(\Omega)}+|\mathbf{V}|_{\mathbb{H}_2}|\theta|_{L^\infty}\}|\nabla\varphi|_{L^2(\Omega)}\\
&+|g_1|_{L^\infty(\Sigma_i)}|h_1|_{L^2(\Sigma_i)}|\varphi|_{L^2(\Sigma_i)}+|g_2|_{L^\infty(\Sigma_e)}|\theta|_{L^2(\Sigma_e)}|\varphi|_{L^2(\Sigma_e)}.
\end{aligned}
$$

Therefore, there is a function $f(t)$ depending only on

$$(k_1, k_2, |g_i|_{L^\infty(\Sigma_i)}, |g_e|_{L^\infty(\Sigma_e)}, |\mathbf{V}|_{\mathbb{H}_2}, |D(\mathbf{V})|_{L^\infty(\Omega)})$$

such that

$$(54)\qquad\left|\frac{\partial\theta}{\partial t}\right|_{L^2(0,t;(H^1(\Omega))')}\leqq f(t).$$

This completes the proof of Lemma 1.

REFERENCES

[1] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic analysis for periodic structures*, North-Holland, Amsterdam, 1978.

[2] L. CAFFARELLI, R. KOHN, AND L. NIRENBERG, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771-831.

[3] G. CHAVENT AND J. JAFFRÉ, *Mathematical models and finite elements for reservoir simulation.* Stud. Math. Appl., 17, North Holland, Amsterdam, 1986.

[4] P. FABRIE, *Solutions fortes et comportement asymptotique pour un modèle de convection naturelle en milieu poreux*, Acta Appl. Math., 7 (1986), pp. 49-77.

[5] P. FABRIE, B. GOYEAU, AND M. LANGLAIS, *Sur un modèle de déplacements non isothermes en milieux poreux*, C. R. Acad. Sci. Paris Sér. I Math., (1990), pp. 707-710.

[6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[7] B. GOYEAU, Thèse de l'Université Bordeaux I, Laboratoire Energetique et Phenomenes de Transfert, Décembre, 1988.

[8] O. A. LADYZENSKAJA, U. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and quasilinear equation of parabolic type*, Transl. Math. Monographs, 23, American Mathematical Society, Providence, RI, 1968.

[9] J. L. LIONS AND E. MAGENES, *Problemi ai limiti non omognei* (V), Ann. Scuola Norm. Sup. Pisa, 16 (1962), pp. 1-44.

[10] N. G. MEYERS, *An L^p-estimate for the gradient of second order elliptic divergence equation*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189-206.

[11] A. MIKELIC, *Mathematical Theory of Stationary Miscible Filtration*, J. Differential Equations, 90 (1991), pp. 186-202.

[12] D. A. NIELD AND D. D. JOSEPH, *Effects of quadratic drag on convection in a saturated porous medium*, Phys. Fluids, 28 (1985), pp. 995-997.

[13] O. A. OLEINIK AND S. N. KRUZHKOV, *Quasilinear second order parabolic equations with many independent variables*, Russian Math. Surveys, 16 (1961), pp. 105-146.

[14] T. F. RUSSEL, M. F. WHEELER AND C. CHIANG, *Large scale simulation of miscible displacement by mixed and characteristic finite element method*, in Mathematical and Computational Method in Seismic Exploration and Reservoir Modeling, Society for Industrial and Applied Mathematics, Philadelphia, 1986, pp. 108-127.

[15] J. SIMON, *Régularité de la solution d'un problème aux limites non linéaires.* Ann. Fac. Sci. Toulouse, III, (1981), pp. 247-274.

[16] R. TEMAM, *Navier-Stokes Equations.* North-Holland, Amsterdam, 1979.

[17] ———, *Navier-Stokes Equations and Nonlinear Analysis*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1393–1416, November 1992

002

ONE-DIMENSIONAL THERMOELASTIC CONTACT WITH A STRESS-DEPENDENT RADIATION CONDITION*

KEVIN T. ANDREWS†, ANDRO MIKELIĆ‡, PETER SHI†,
MEIR SHILLOR†, AND STEVE WRIGHT†

Abstract. A one-dimensional quasistatic thermoelastic contact problem with a stress-dependent boundary condition is considered. The problem models the evolution of the temperature and the displacement of a long thin elastic rod that may come into contact with a rigid obstacle. The mathematical problem is reduced to solving a nonlocal heat equation with a nonlinear and nonlocal boundary condition. This boundary condition contains a heat exchange coefficient that depends on the pressure when there is contact with the obstacle and on the size of the gap when there is no contact. The local existence of a strong solution to the problem and local dependence on the initial-boundary data are proved. In addition, the uniqueness of the solution is established. The proof rests on an abstract result dealing with perturbations of monotone operators, as well as some a priori estimates which permit an application of Schauder's fixed point theorem.

Key words. thermoelastic contact, nonlinear heat transfer coefficient, nonlinear perturbations, monotone operators, nonlinear boundary conditions, Signorini's condition

AMS(MOS) subject classifications. primary 35K60; secondary 73C35, 73T05

1. Introduction.

Thermoelastic problems with contact arise naturally in many industrial processes, particularly in the manufacture of such items as castings, mouldings, pistons, thermostats, etc. In these situations two or more elastic materials are forced into contact with each other as a result of thermal expansion. Predicting the behavior of thermoelastically contacting bodies in such situations is of considerable applied importance. In ball bearings, for example, in cases where the ball and the casing are made of different materials, thermal expansion or contraction may cause the bearing either to lock up or to chatter. There is a considerable engineering literature that deals with such problems. For example, Srinivasan and France [SF] report on erratic performance of duplex heat exchange tubes in experimental breeder reactors and suggest that this may be due to the presence of multiple steady states. Richmond and Huang [RH] have suggested that the growth of a sinusoidal perturbation in an otherwise uniform contact pressure between a solidifying casting and the mould may be responsible for experimentally observed waviness in nominally plane cast surfaces. Barber [Ba2] has shown that such behavior may result from the instability of the thermoelastic contact.

In spite of the obvious applied importance of the subject, there are relatively few theoretical results about general problems of thermoelastic contact. Until recently the existing mathematical models make various restrictive assumptions about how the process behaves. A general variational inequality model was derived in Duvaut and Lions [DL], where evolution problems with unilateral conditions were considered, but it assumes that there is no loss of contact. Duvaut [Du] considered only the static problem with Signorini's contact condition and a smoothed stress-dependent radiation condition.

*Received by the editors February 11, 1991; accepted for publication (in revised form) March 31, 1992.

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

‡Department of Physics, Ruder Bošković Institute, 41001 Zagreb, Yugoslavia.

More recently, the one-dimensional quasistatic problem of thermoelastic contact was considered in a series of papers by Gilbert, Shi, and Shillor [GSS], Shi and Shillor [SS1], [SS2], and Shi, Shillor, and Zou [SSZ]. The problem was formulated as a fully coupled variational inequality in [GSS], and the existence of a strong solution was established. A reformulation of the problem in [SS1] led to a decoupled heat equation with a nonlinear and nonlocal source term. The uniqueness of the solution was proved as well as the fact that the solution converges to a steady state. In both of the papers [GSS], [SS1] the Dirichlet condition was assumed for the temperature at the contacting edge. The more realistic heat exchange condition was considered in [SS2], where the existence of a solution was obtained from an abstract result involving perturbations of monotone operators. The model there was somewhat restrictive in that the heat exchange coefficient k was assumed to be a constant.

It is known from the work of Barber and his collaborators, [Ba1], [Ba2], [BDC], [BZ] as well as [CD], that the heat exchange between the contacting edge and the wall seems to depend in a complicated way on the distance of the edge from the obstacle and on the contact pressure. Thus it seems reasonable that a general model for the process would assume that the heat exchange coefficient k depends on the distance of the edge from the obstacle when there is no contact, and on the contact pressure otherwise. In the latter case it is also assumed that there is a thin heat-resisting layer at the edge and that the resistance to the heat conduction decreases when the magnitude of the contact pressure increases. In this paper we consider such a general model. Some numerical simulations for problems with constant and nonconstant k can be found in [SSZ].

We now describe the remaining sections of this paper. The modelling of k is presented in §2, where the precise mathematical problem is described. Decoupling leads to a heat equation with a contact dependent source term. This section also contains a statement of our main result, Theorem 2.1, which guarantees the local existence of a strong solution to the problem and the local dependence on the initial-boundary data. The theorem also asserts that any solution to the problem is unique. The remaining sections of the paper are devoted to constructing a proof of this theorem. In §3 we prove an abstract result, Theorem 3.1, concerning variational inequalities associated with nonlinear perturbations of monotone operators in Banach spaces. This theorem is used in §4 to solve an auxiliary problem. Finally, the results of §4, together with Schauder's fixed point theorem, are used in §5 to prove Theorem 2.1.

2. The model.

The problem under consideration models a homogeneous elastic rod that is held fixed at one edge. It is free to expand or contract at the other edge as a result of the evolution of its temperature and stress field, but the expansion is limited by the existence of an obstacle, say a rigid wall, which blocks any further expansion once the rod comes into contact with it. We assume that the accelerations in the system are negligible, and thus the problem may be described as quasistatic.

The physical setting is depicted in Fig. 1. We denote the temperature by $\theta = \theta(x,t)$, the displacement by $u = u(x,t)$, and the stress by $\sigma = \sigma(x,t)$, all in nondimensional form. The thermoelastic problem for such a system may be posed (see, e.g., Carlson [Ca] or Day [Da]) as follows: find a pair $\{\theta, u\}$ such that

$$\theta_{xx} = \theta_t + au_{xt} \quad \text{for } 0 < x < 1 \quad \text{and} \quad 0 < t < T, \tag{2.1}$$

$$u_{xx} = a\theta_x \quad \text{for } 0 < x < 1 \quad \text{and} \quad 0 < t < T. \tag{2.2}$$

Here (1.1) and (1.2) are the energy and the elasticity equations, respectively. The *coupling constant* a is related to the physical properties of the material (see [Da],

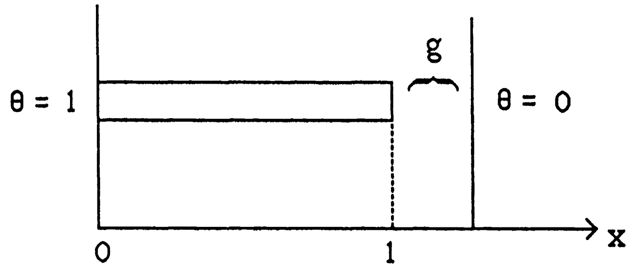


FIG. 1

[SS1], or [GSS]) and is typically a small positive number. We shall assume throughout that $0 < a < 1$. To complete the model, we prescribe the initial condition $\theta(x, 0) = \varphi$ together with the following boundary conditions. At the left edge ($x = 0$) we assume that the temperature is prescribed in the form $\theta = m(t)$ and that the rod is held fixed there, i.e., $u = 0$. At the right edge ($x = 1$) we assume the radiation condition $-\theta_x = k\theta$, which is discussed below. For the displacement at that edge we assume the Signorini contact condition

$$u \leq g, \quad \sigma \leq 0, \quad \text{and} \quad (u - g)\sigma = 0, \tag{2.3}$$

where g is the initial gap between the wall and the free edge. This condition implies that the expansion of the free edge is restricted by the obstacle, i.e., $u \leq g$, and the stress is compressive, i.e., $\sigma \leq 0$. Consequently, if contact occurs, then $u = g$, but if the edge is free then $\sigma = 0$. In one space dimension we have that $\sigma(x, t) = u_x(x, t) - a\theta(x, t)$ (see [Da]). Therefore, the condition may be written in terms of the displacement and temperature as

$$u \leq g, \quad u_x \leq a\theta, \quad \text{and} \quad (u - g)(u_x - a\theta) = 0. \tag{2.4}$$

We turn next to the modelling of the radiation condition, more precisely to k. As was mentioned above, we would like it to depend on the actual gap $g - u$, when there is no contact, and on the pressure σ when contact occurs. It turns out that the natural argument for k is $\eta = g - u + \sigma$. Indeed, in the absence of contact, $\sigma = 0$ and $\eta = g - u > 0$. Thus η measures the distance of the edge from the wall. On the other hand, when contact occurs, $u = g$ and, therefore, $\eta = \sigma \leq 0$, and now η measures the contact pressure. Therefore, we propose the following form for the heat exchange coefficient:

$$k = k(g - u + \sigma) = k(g - u + u_x - a\theta). \tag{2.5}$$

Here $k = k(\eta)$ is a prescribed function of its argument. In this paper we will consider only smooth nonnegative functions k. Primicerio [Pr] has suggested that in some applications it may be appropriate to take k to be a piecewise constant function

$$k(\eta) = \begin{cases} k_0, & \eta \leq 0, \\ k_1, & \eta > 0. \end{cases}$$

Some numerical simulations were performed in [SSZ] using such a choice of k. A more complicated condition (the so-called "imperfect contact" condition) was proposed in Barber [Ba1] (see also [CD]).

We now describe the spaces in which we expect to find solutions. Let $\Omega_T = (0,1) \times (0,T)$ for some $T > 0$. From the form of (2.1) and (2.2) it is natural to seek a solution θ, which lies in the Sobolev space $W_2^{2,1}(\Omega_T)$, which consists of all $L^2(\Omega_T)$-summable functions that possess generalized $L^2(\Omega_T)$-summable second-order space and first-order time derivatives. The norm of this Hilbert space is given by

$$\| \theta \|^2_{W_2^{2,1}(\Omega_T)} = \int_{\Omega_T} (\theta_t^2 + \theta_x^2 + \theta_{xx}^2 + \theta^2)\, dx\, dt. \tag{2.6}$$

Similarly, we seek for u a function that lies in the Sobolev space

$$X = \{u \in H^1(\Omega_T);\ u_{xx}, u_{xt} \in L^2(\Omega_T),\ u(0,t) = 0,\ 0 < t < T\}, \tag{2.7}$$

whose norm is given by

$$\| u \|^2_X = \int_{\Omega_T} (u_t^2 + u_x^2 + u_{xx}^2 + u_{xt}^2) dx\, dt. \tag{2.8}$$

For θ as above, it is known (see, e.g., [LM, p.9]) that the boundary functions $\theta(0,\cdot)$ and $\theta(1,\cdot)$ lie in $H^{3/4}(0,T)$, $\theta_x(1,t)$ in $H^{1/4}(0,T)$, and $\theta(x,0)$ in $H^1(0,1)$. These facts imply that the initial-boundary data must have the degree of regularity described below. The only properties of these fractional Sobolev spaces that we shall need are that they are Hilbert spaces and that they satisfy compact imbedding theorems. For a full treatment of these spaces, we refer the interested reader to [LM]. The definitions of any unexplained spaces or notation may be found in [LSU], but we mention one item in particular here since the notation may be unfamiliar to some. The Banach space $W^1_\infty(R)$ consists of all essentially bounded functions defined on R having a generalized derivative that is also essentially bounded; the norm of this space is the sum of the essential suprema of the function and its derivative.

We can now give a complete description of the initial-boundary value problem that we will consider in this paper.

Given φ in $H^1(0,1)$, m in $H^1(0,T)$, and a nonnegative k in $W^1_\infty(R)$, find θ in $W^{2,1}(\Omega_T)$ and u in X such that

$$\theta_{xx} = \theta_t + au_{xt} \quad \text{in } \Omega_T, \tag{2.9}$$
$$u_{xx} = a\theta_x \quad \text{in } \Omega_T, \tag{2.10}$$
$$\theta(0,t) = m(t), \qquad 0 < t < T, \tag{2.11}$$
$$\theta(x,0) = \varphi(x), \qquad 0 < x < 1, \tag{2.12}$$
$$-\theta_x(1,t) = k\theta(1,t), \qquad 0 < t < T, \tag{2.13}$$
$$u(0,t) = 0, \qquad 0 < t < T, \tag{2.14}$$
$$u(1,t) \le g, \qquad 0 < t < T, \tag{2.15}$$
$$u_x(1,t) \le a\theta(1,t), \qquad 0 < t < T, \tag{2.16}$$
$$[u(1,t) - g][u_x(1,t) - a\theta(1,t)] = 0, \qquad 0 < t < T, \tag{2.17}$$

where $k = k(g - u + u_x - a\theta)$ is given by (2.5). Of course, to find a solution with the desired degree of regularity the data must be compatible, i.e., $\varphi(0) = m(0)$.

It turns out, as was shown in [SS1], [SS2], that the problem (2.9)–(2.17) decouples and can be formulated equivalently in terms of the temperature only. It is this problem that we will solve here. The reformulation is obtained by performing a number of integrations with respect to x, using the boundary conditions. We refer the interested reader to [SS2] for the details. Once the temperature is found, the displacement is given by

$$u(x,t) = a\int_0^x \theta(\xi,t)\,d\xi - x\max\left\{a\int_0^1 \theta(\xi,t)\,d\xi - g, 0\right\}. \tag{2.18}$$

In particular, the displacement of the right edge is

$$u(1,t) = \min\left\{a\int_0^1 \theta(\xi,t)d\xi, g\right\}. \tag{2.19}$$

Also $\sigma(x,t) = \sigma(t)$, as can be seen from (2.10) upon recalling that $\sigma = u_x - a\theta$. It was shown in [SS1], [SS2] that

$$\sigma(t) = \min\left\{g - a\int_0^1 \theta(\xi,t)d\xi, 0\right\}. \tag{2.20}$$

We turn next to consider k and its argument. Using (2.19) and (2.20) we have

$$\begin{aligned}\eta = g - u - \sigma &= g - \min\left\{a\int_0^1 \theta(\xi,t)d\xi, g\right\} + \min\left\{g - a\int_0^1 \theta(\xi,t)d\xi, 0\right\}\\ &= g - a\int_0^1 \theta(\xi,t)d\xi.\end{aligned} \tag{2.21}$$

It is interesting to note that k has a nonlocal argument. We can now state our main result.

THEOREM 2.1. *Given φ in $H^1(0,1)$ and m in $H^1(0,T)$ and a nonnegative k in $W^1_\infty(R)$, there exists a $T_0 \le T$ and a θ in $W_2^{2,1}(\Omega_{T_0})$ which satisfies the conditions*

$$\begin{aligned}(1+a^2)\theta_t - \theta_{xx} &= a\frac{d}{dt}\max\left\{a\int_0^1 \theta(\xi,t)d\xi - g, 0\right\} \quad \text{in } \Omega_{T_0},\\ \theta(0,t) &= m(t) \qquad 0 < t < T_0,\\ -\theta_x(1,t) &= k\left(g - a\int_0^1 \theta(\xi,t)d\xi\right)\theta(1,t) \qquad 0 < t < T_0,\\ \theta(x,0) &= \varphi(x) \qquad 0 < x < 1,\end{aligned} \tag{2.22}$$

provided $0 < a < 1$ and φ and m satisfy the compatibility condition $\varphi(0) = m(0)$. Moreover, if $\varphi_n \to \varphi$ in $H^1(0,1)$ and $m_n \to m$ in $H^1(0,T_0)$, and θ_n and θ are the corresponding solutions, then $\theta_n \to \theta$ in $L^2(\Omega_{T_0})$. Finally, for any T_0, if the solution to (2.22) exists, then it is unique.

As previously indicated, the constant a is typically close to zero, and hence it is possible to satisfy the hypothesis of the theorem. It is shown in [SS2] that if θ satisfies the conditions of Theorem 2.1 and if u is given by (2.18), then the pair $\{\theta, u\}$ solves the problem (2.9)–(2.17). The proof of Theorem 2.1 is given in §5.

3. An abstract existence theorem. In this section we prove an abstract existence result that is used in the proof of Theorem 2.1. The result is of some independent interest and is similar in flavor to results described in Chapter 10 of [BC] in that Fan's lemma plays a key role.

Let Y be a Banach space with dual Y^*, and let K be a convex subset of Y. Let ρ denote a proper convex functional on K, and let κ be a topology on K that is stronger than the norm topology and such that ρ is lower semicontinuous in the κ-topology and ρ-bounded and κ-closed subsets of K are norm compact. We denote the duality pairing between Y and Y^* by $\langle \cdot, \cdot \rangle$. Let Z be a subspace of Y^*. Let A and B be two operators (not necessarily linear) mapping K into Y^* so that the following conditions hold:

(A1) The operators A and B are κ-to-weak* continuous; the range of B is in Z;

(A2) The operator A is monotone, i.e., for any $u, v \in K$ with $u \neq v$ we have that

$$\langle Au - Av, u - v \rangle > 0; \tag{3.1}$$

(A3) There exists a function $F : Z \to R$ and constants $\beta \geq 0, \lambda > 0$ and $\gamma \geq 0$ so that

(i) For each $y^* \in Z$ there is a $y_0 \in K$ such that

$$\langle Ay_0, y - y_0 \rangle \geq \langle y^*, y - y_0 \rangle \quad \forall y \in K, \tag{3.2}$$

and

$$\rho(y_0) \leq F(y^*) + \beta; \tag{3.3}$$

(ii) For each $y \in K$ we have that

$$F(By) \leq \lambda \rho(y) + \gamma; \tag{3.4}$$

(iii) for each $y^* \in Z$ there exists a constant $c_{y^*} \geq 0$ such that

$$F(y^* + w^*) \leq F(w^*) + c_{y^*} \quad \forall w^* \in Z. \tag{3.5}$$

We can now state our abstract existence result.

THEOREM 3.1. *Assume that $\lambda < 1$. Under the assumptions* (A1)–(A3) *we have that for every y^* in Z there exists a $y_0 \in K$ such that*

$$\langle Ay_0 - By_0, y - y_0 \rangle \geq \langle y^*, y - y_0 \rangle \quad \forall y \in K. \tag{3.6}$$

Proof. It suffices to establish the existence of y_0 when $y^* = 0$ since the general case may be reduced to this by replacing A by $A - y^*$. In this reduction we use part (iii) of assumption (A3), and this is the only place where this condition is used in the proof.

Since $\lambda < 1$ and ρ is proper, we can choose $\delta > (\beta + \gamma)/(1 - \lambda) \geq 0$ so that the convex set K_δ, defined by

$$K_\delta = \{ y \in K : \rho(y) \leq \delta \}, \tag{3.7}$$

is nonempty. Notice that the hypotheses imply that K_δ is norm compact. Consider the problem of finding y_1 in K_δ so that

$$\langle Ay_1 - By_1, y - y_1 \rangle \geq 0 \quad \forall y \in K_\delta. \tag{3.8}$$

For each $y \in K_\delta$ we define

$$K(y) = \{ w \in K_\delta : \langle Aw - Bw, y - w \rangle \geq 0 \}.$$

The existence of y_1 in (3.8) is thus equivalent to the assertion that $\bigcap_{y\in K_\delta} K(y) \neq \emptyset$. To establish this latter fact we use Fan's lemma [BC, p. 208]. Note that, for each y, $K(y)$ is nonempty since $y \in K(y)$. Next we show that each $K(y)$ is κ-closed and hence norm compact. Note that κ is first countable since it is stronger than the norm topology. Hence we may check closure using sequences. Let $\{w_n\}$ be a sequence in $K(y)$ that converges in κ and hence in norm to an element w (necessarily in K_δ). Since A and B are κ-to-weak* continuous, it follows that

$$Aw_n - Bw_n \rightarrow Aw - Bw \quad \text{weak* in } Y^*,$$

and hence

$$\begin{aligned} 0 \leq \langle Aw_n - Bw_n, y - w_n\rangle =& \langle Aw_n - Bw_n, y - w\rangle + \langle Aw_n - Bw_n, w - w_n\rangle \\ \rightarrow& \langle Aw - Bw, y - w\rangle. \end{aligned}, \tag{3.9}$$

This shows that each $K(y)$ is κ-closed. It remains to be shown that for each finite subset $\{y_1, \cdots, y_n\}$ of K_δ we have that

$$co\{y_1, \cdots, y_n\} \subset \bigcup_{i=1}^{n} K(y_i).$$

Assume, by way of contradiction, that there exists an element $y = \sum_{i=1}^n \lambda_i y_i$ (with $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$) so that $y \notin K(y_i)$ for each i, i.e.,

$$\langle Ay - By, y_i - y\rangle < 0.$$

But then

$$0 = \langle Ay - By, y - y\rangle = \sum_{i=1}^{n} \lambda_i \langle Ay - By, y_i - y\rangle < 0,$$

a contradiction. Since all conditions of Fan's lemma are satisfied, we have that $\bigcap_{y\in K_\delta} K(y)$ is nonempty. Let y_1 be any point in the intersection. Then by (A3) there is a $y_0 \in K$ so that

$$\langle Ay_0, y - y_0\rangle \geq \langle By_1, y - y_0\rangle \quad \forall y \in K, \tag{3.10}$$

$$\rho(y_0) \leq F(By_1) + \beta, \tag{3.11}$$

$$F(By_1) \leq \lambda\rho(y_1) + \gamma. \tag{3.12}$$

Now (3.11) and (3.12) imply that

$$\rho(y_0) - \beta \leq \lambda\rho(y_1) + \gamma,$$

and hence

$$\rho(y_0) \leq \lambda\rho(y_1) + \gamma + \beta \leq (\lambda)\delta + (1 - \lambda)\delta = \delta,$$

by the choice of δ above. Thus $y_0 \in K_\delta$, and so

$$\langle Ay_1 - By_1, y_0 - y_1\rangle \geq 0. \tag{3.13}$$

On the other hand, substituting $y = y_1$ in (3.10) yields

$$\langle Ay_0 - By_1, y_1 - y_0\rangle \geq 0. \tag{3.14}$$

Combining (3.13) and (3.14) gives

$$\langle Ay_0 - Ay_1, y_1 - y_0\rangle \geq 0$$

or

$$\langle Ay_0 - Ay_1, y_0 - y_1\rangle \leq 0. \tag{3.15}$$

Hence, by condition (A2), we have that $y_0 = y_1$. Consequently, (3.10) may be written as

$$\langle Ay_0 - By_0, y - y_0\rangle \geq 0 \quad \forall y \in K,$$

which is the desired result for $y^* = 0$.

4. A priori estimates and an auxiliary result. In this section we will see how the abstract existence theorem of §3 may be applied to prove the existence of a solution to an auxiliary problem. We seek a solution θ, which lies in the space $W_2^{2,1}(\Omega_T)$. The convex set K will thus be a subset of this space. In order to construct the space Y and the convex functional ρ of the theorem, we need an estimate on the size of solutions to the initial-boundary value problem. This estimate is furnished by Lemma 4.2. This lemma also contains a second estimate, which we will use in §5 to complete the proof of Theorem 2.1. In order to derive this second estimate we need a preliminary lemma, Lemma 4.1, which allows us to control the size of $\theta(1,\cdot)$ in a particular manner. To motivate this result, we recall that there exists a constant c so that $\|\theta(1,\cdot)\|^2_{L^6(0,T)} \leq c\|\theta\|_{W_2^{2,1}(\Omega_T)}$ for all θ in $W_2^{2,1}(\Omega_T)$. However, in general, this constant depends upon T . The purpose of the next lemma is to show that the dependence upon T may be concentrated in the initial-boundary data.

LEMMA 4.1. *Let θ be in $H^1(\Omega_T)$, and suppose that $\theta(0,t) = m(t)$ and $\theta(x,0) = \varphi(x)$ are in $L^\infty(0,T)$ and $L^\infty(0,1)$, respectively. Then*

$$\|\theta(1,\cdot)\|^2_{L^6(0,T)}$$
$$\leq c\Big(\|\theta_t\|^2_{L^2(\Omega_T)} + \|\theta_x\|^2_{L^2(\Omega_T)} + (1+T^2)\,\|\varphi\|^2_{L^\infty(0,1)} + \left(1+T^{1/3}+T\right)\|m\|^2_{L^\infty(0,T)}\Big),$$

where c is a constant that is independent of θ and T.

Proof. We suppose first that $m(t) = 0$. Then by integrating the equality

$$\theta^6(1,t) = 3\theta^3(1,t)\int_0^1 \theta^2(x,t)\theta_x(x,t)dx$$

over [0,T] and then repeatedly using the Cauchy–Schwarz inequality, we obtain

$$\int_0^T \theta^6(1,t)dt$$
$$\leq 3\int_0^T \theta^3(1,t)\int_0^1 |\theta^2(x,t)\theta_x(x,t)|dx\,dt$$
$$\leq 3\left(\int_0^T \theta^6(1,t)dt\right)^{1/2}\left(\int_0^T\left(\int_0^1 |\theta^2(x,t)\theta_x(x,t)|dx\right)^2 dt\right)^{1/2}$$
$$\leq 3\left(\int_0^T \theta^6(1,t)dt\right)^{1/2}\left(\int_0^T\left(\int_0^1 \theta^4(x,t)dx\right)\left(\int_0^1 \theta_x^2(x,t)dx\right)dt\right)^{1/2}$$
$$\leq 3\left(\int_0^T \theta^6(1,t)dt\right)^{1/2}\operatorname*{ess\,sup}_{0<t<T}\left(\int_0^1 \theta^4(x,t)dx\right)^{1/2}\left(\int_0^T\int_0^1 \theta_x^2(x,t)dx\,dt\right)^{1/2}.$$

It follows that

(4.1)
$$\left(\int_0^T \theta^6(1,t)dt\right)^{1/3} \le 3^{2/3} \operatorname*{ess\,sup}_{0<t<T} \left(\int_0^1 \theta^4(x,t)dx\right)^{1/3} \left(\int_0^T \int_0^1 \theta_x^2(x,t)dx\,dt\right)^{1/3}.$$

We now release the assumption that $\theta(0,t)=0$ and apply (4.1) to $\theta - m(t)$ to obtain

(4.2)
$$\begin{aligned} &\|\theta(1,\cdot)\|^2_{L^6(0,T)} \\ &\le 2\left(\|\theta(1,\cdot)-m\|^2_{L^6(0,T)} + \|m\|^2_{L^6(0,T)}\right) \\ &\le c\left(\operatorname*{ess\,sup}_{0<t<T} \|\theta(\cdot,t)-m(t)\|^{4/3}_{L^4(0,1)}\|\theta_x\|^{2/3}_{L^2(\Omega_T)} + \|m\|^2_{L^6(0,T)}\right) \\ &\le c\left(\left(\operatorname*{ess\,sup}_{0<t<T} \|\theta(\cdot,t)\|^{4/3}_{L^4(0,1)} + \|m\|^{4/3}_{L^\infty(0,T)}\right)\|\theta_x\|^{2/3}_{L^2(\Omega_T)} + \|m\|^2_{L^6(0,T)}\right) \\ &\le c\left(\operatorname*{ess\,sup}_{0<t<T} \|\theta(\cdot,t)\|^{4/3}_{L^4(0,1)}\|\theta_x\|^{2/3}_{L^2(\Omega_T)} + \|m\|^2_{L^\infty(0,T)} + \|\theta_x\|^2_{L^2(\Omega_T)} + \|m\|^2_{L^6(0,T)}\right). \end{aligned}$$

Here c denotes a constant that is independent of θ and T but that changes from line to line. We now estimate $\operatorname{ess\,sup}_{0<t<T} \|\theta(\cdot,t)\|^{4/3}_{L^4(0,1)}$ by assuming first that $\theta(x,0)=0$ and then applying the Cauchy–Schwarz inequality to the equality

$$\int_0^1 \theta^4(x,t)dx = 2\int_0^1 \theta^2(x,t)\left(\int_0^t \theta(x,s)\theta_s(x,s)ds\right)dx$$

as before to obtain

$$\begin{aligned} &\int_0^1 \theta^4(x,t)dx \\ &\quad\le 2\left(\int_0^1 \theta^4(x,t)dx\right)^{1/2}\left(\int_0^1\left(\int_0^t \theta(x,s)\theta_s(x,s)ds\right)^2 dx\right)^{1/2} \\ &\quad\le 2\left(\int_0^1 \theta^4(x,t)dx\right)^{1/2}\left(\int_0^1\left(\int_0^t \theta^2(x,s)ds\right)\left(\int_0^t \theta_s^2(x,s)ds\right)dx\right)^{1/2} \\ &\quad\le 2\left(\int_0^1 \theta^4(x,t)dx\right)^{1/2} \operatorname*{ess\,sup}_{0<x<1}\left(\int_0^t \theta^2(x,s)ds\right)^{1/2}\left(\int_0^1\int_0^t \theta_s^2(x,s)ds\,dx\right)^{1/2}. \end{aligned}$$

It follows that

(4.3)
$$\begin{aligned} &\operatorname*{ess\,sup}_{0<t<T}\|\theta(\cdot,t)\|^{4/3}_{L^4(0,1)} \\ &\quad\le 2^{1/3} \operatorname*{ess\,sup}_{0<x<1}\left(\int_0^T \theta^2(x,s)ds\right)^{1/3}\left(\int_0^1\int_0^T \theta_s^2(x,s)ds\,dx\right)^{1/3}. \end{aligned}$$

Again we release the assumption that $\theta(x,0)=0$ and apply (4.3) to $\theta-\varphi$ to obtain

$$
\begin{aligned}
&\operatorname*{ess\,sup}_{0<t<T}\|\theta(\cdot,t)\|^{4/3}_{L^4(0,1)}\\
&\quad\le c\left(\operatorname*{ess\,sup}_{0<t<T}\|\theta(\cdot,t)-\varphi\|^{4/3}_{L^4(0,1)}+\|\varphi\|^{4/3}_{L^4(0,1)}\right)\\
&\quad\le c\left(\operatorname*{ess\,sup}_{0<x<1}\|\theta(x,\cdot)-\varphi\|^{2/3}_{L^2(0,T)}\|\theta_t\|^{2/3}_{L^2(\Omega_T)}+\|\varphi\|^{4/3}_{L^4(0,1)}\right)\\
&\quad\le c\left(\operatorname*{ess\,sup}_{0<x<1}\|\theta(x,\cdot)\|^{2/3}_{L^2(0,T)}\|\theta_t\|^{2/3}_{L^2(\Omega_T)}+T^{2/3}\|\varphi\|^{2/3}_{L^\infty(0,1)}\|\theta_t\|^{2/3}_{L^2(\Omega_T)}+\|\varphi\|^{4/3}_{L^4(0,1)}\right).
\end{aligned}
\tag{4.4}
$$

Using the result (4.4) in (4.2) and then using Young's inequality, we have that

$$
\begin{aligned}
&\|\theta(1,\cdot)\|^2_{L^6(0,T)}\\
&\quad\le c\Bigg(\operatorname*{ess\,sup}_{0<x<1}\|\theta(x,\cdot)\|^{2/3}_{L^2(0,T)}\|\theta_t\|^{2/3}_{L^2(\Omega_T)}\|\theta_x\|^{2/3}_{L^2(\Omega_T)}\\
&\qquad\quad+T^{2/3}\|\varphi\|^{2/3}_{L^\infty(0,1)}\|\theta_t\|^{2/3}_{L^2(\Omega_T)}\|\theta_x\|^{2/3}_{L^2(\Omega_T)}\\
&\qquad\quad+\|\varphi\|^{4/3}_{L^4(0,1)}\|\theta_x\|^{2/3}_{L^2(\Omega_T)}+\|m\|^2_{L^\infty(0,T)}+\|\theta_x\|^2_{L^2(\Omega_T)}+\|m\|^2_{L^6(0,T)}\Bigg)\\
&\quad\le c\Bigg(\operatorname*{ess\,sup}_{0<x<1}\|\theta(x,\cdot)\|^2_{L^2(0,T)}+\|\theta_t\|^2_{L^2(\Omega_T)}+\|\theta_x\|^2_{L^2(\Omega_T)}\\
&\qquad\quad+(1+T^2)\|\varphi\|^2_{L^\infty(0,1)}+\left(1+T^{1/3}\right)\|m\|^2_{L^\infty(0,T)}\Bigg)\\
&\quad\le c\Bigg(\|\theta_t\|^2_{L^2(\Omega_T)}+\|\theta_x\|^2_{L^2(\Omega_T)}+(1+T^2)\|\varphi\|^2_{L^\infty(0,1)}+\left(1+T^{1/3}+T\right)\|m\|^2_{L^\infty(0,T)}\Bigg).
\end{aligned}
$$

This is the desired result.

LEMMA 4.2. *Let f be in $L^2(\Omega_T)$, φ in $H^1(0,1)$, m in $H^1(0,T)$, and let k in $W^1_\infty(R)$ be nonnegative. Suppose θ in $W^{2,1}_2(\Omega_T)$ satisfies the following conditions*:

$$(1+a^2)\theta_t-\theta_{xx}=f\quad in\ \Omega_T,\tag{4.5}$$

$$\theta(x,0)=\varphi(x),\qquad 0<x<1,\tag{4.6}$$

$$\theta(0,t)=m(t),\qquad 0<t<T,\tag{4.7}$$

$$k(t)\theta(1,t)+\theta_x(1,t)=0,\qquad 0<t<T.\tag{4.8}$$

Then

$$
\begin{aligned}
&\frac{3(1+a^2)^2}{4}\|\theta_t\|^2_{L^2(\Omega_T)}+\|\theta_{xx}\|^2_{L^2(\Omega_T)}+\frac{(1+a^2)}{T}(1-4T\|k'\|_{L^\infty(R)})\|\theta_x\|^2_{L^2(\Omega_T)}\\
&\quad\le 3\|f\|^2_{L^2(\Omega_T)}+c\Big(\|m\|^2_{H^1(0,T)}+\|\varphi\|^2_{H^1(0,1)}\Big),
\end{aligned}
\tag{4.9}
$$

where $c=c(a,T,\|k\|_{W^1_\infty(R)})$ is a constant that depends only on a, T, and $\|k\|_{W^1_\infty(R)}$. If instead of (4.8) *we assume that*

$$k(\eta)\theta(1,t)+\theta_x(1,t)=0,\tag{4.10}$$

where $\eta = g - a\int_0^1 \xi(x,t)dx$ and ξ is in $W^{2,1}(\Omega_T)$, then we have

$$\begin{aligned}(4.11)\qquad &(a(1+a^2)^2 - cT^{1/6}\|\xi_t\|_{L^2(\Omega_T)})\|\theta_t\|^2_{L^2(\Omega_T)} + \|\theta_{xx}\|^2_{L^2(\Omega_T)}\\ &+\left(\frac{(1+a^2)}{T} - cT^{1/6}\|\xi_t\|_{L^2(\Omega_T)}\right)\|\theta_x\|^2_{L^2(\Omega_T)}\\ &+(1+a^2)\sup_{0\le\tau\le T}\|\theta_x(\cdot,\tau)\|^2_{L^2(0,1)}\\ &\le \frac{(1+a^2)^2}{a^2}\|f\|^2_{L^2(\Omega_T)} + C\Big(\|m\|^2_{H^1(0,T)} + \|\varphi\|^2_{H^1(0,1)}\Big)\\ &+cT^{1/6}\|\xi_t\|_{L^2(\Omega_T)}\Big((1+T^2)\,\|\varphi\|^2_{L^\infty(0,1)} + \left(1+T^{1/3}+T\right)\|m\|^2_{L^\infty(0,T)}\Big),\end{aligned}$$

where $c = c(a, \|k'\|_{L^\infty(R)})$ is a constant that depends only on a and $\|k'\|_{L^\infty(R)}$, and $C = C(a, T, \|k\|_{W^1_\infty(R)})$ is a constant that depends only on a, T, and $\|k\|_{W^1_\infty(R)}$. Moreover, for fixed a and $\|k\|_{W^1_\infty(R)}$, C is an increasing function of T.

Proof. We first establish the result under the additional assumptions that θ_{xt} and θ_{tt} exist as elements of $L^2(\Omega_T)$. Notice that θ must satisfy

$$(1+a^2)(\theta - m)_t - \theta_{xx} = f - (1+a^2)m' \quad \text{in } \Omega_T.$$

Now squaring both sides of the above and integrating over $\Omega_\tau = (0,1)\times(0,\tau)$ yields

$$\begin{aligned}(4.12)\qquad &(1+a^2)^2\|\theta_t\|^2_{L^2(\Omega_\tau)} - 2(1+a^2)\int_{\Omega_\tau}(\theta-m)_t\theta_{xx} + \|\theta_{xx}\|^2_{L^2(\Omega_\tau)}\\ &= \|f\|^2_{L^2(\Omega_\tau)} - 2(1+a^2)\int_{\Omega_\tau} m'f + 2(1+a^2)^2\int_{\Omega_\tau} m'\theta_t.\end{aligned}$$

The two integrals on the right in (4.12) are easily estimated by Cauchy's inequality with ϵ:

$$(4.13)\qquad \left|-2(1+a^2)\int_{\Omega_\tau} m'f\right| \le 2(1+a^2)^2\|m'\|^2_{L^2(0,\tau)} + \frac{1}{2}\|f\|^2_{L^2(\Omega_\tau)},$$

$$(4.14)\qquad \left|2(1+a^2)^2\int_{\Omega_\tau} m'\theta_t\right| \le 4(1+a^2)^2\|m'\|^2_{L^2(0,\tau)} + \frac{(1+a^2)^2}{4}\|\theta_t\|^2_{L^2(\Omega_\tau)}.$$

Applying integration by parts to the integral on the left of (4.10) and using the boundary conditions, we obtain

$$\begin{aligned}-2(1+a^2)\int_0^\tau\int_0^1 (\theta-m)_t\theta_{xx}dx\,dt &= 2(1+a^2)\int_0^\tau\int_0^1 \theta_x\theta_{xt}dx\,dt\\ &\quad - 2(1+a^2)\int_0^\tau \theta_x(x,t)(\theta-m)_t(x,t)\Big|_0^1 dt\\ (4.15)\qquad &= (1+a^2)\|\theta_x(\cdot,\tau)\|^2_{L^2(0,1)} - (1+a^2)\|\varphi'\|^2_{L^2(0,1)}\\ &\quad - 2(1+a^2)\int_0^\tau k(t)m'(t)\theta(1,t)dt\\ &\quad + 2(1+a^2)\int_0^\tau k(t)\theta(1,t)\theta_t(1,t)dt.\end{aligned}$$

We now estimate the two integrals on the right. The first integral in (4.15) may be estimated directly by Cauchy's inequality with ϵ:

$$\begin{aligned}&\left|2(1+a^2)\int_0^\tau m'(t)k(t)\theta(1,t)dt\right|\\ &\leq 2T(1+a^2)\int_0^\tau (m')^2(t)k(t)dt+\frac{(1+a^2)}{2T}\int_0^\tau k(t)\theta^2(1,t)dt\\ &\leq c\|m\|^2_{H^1(0,\tau)}+\frac{(1+a^2)}{2T}\int_0^\tau k(t)\theta^2(1,t)dt.\end{aligned}\tag{4.16}$$

Here c denotes a constant that depends only on a and $\|k\|_{W^1_\infty(R)}$, and from this point until the end of the proof we will use c to designate such a constant. Applying integration by parts to the second integral in (4.15) yields

$$\begin{aligned}2(1+a^2)\int_0^\tau k(t)\theta(1,t)\theta_t(1,t)dt&=(1+a^2)k(\tau)\theta^2(1,\tau)-(1+a^2)k(0)\varphi^2(0)\\ &\quad-(1+a^2)\int_0^\tau k'(t)\theta^2(1,t)dt.\end{aligned}\tag{4.17}$$

Using the results of (4.13)–(4.17) in (4.12) and rearranging terms leads to

$$\begin{aligned}&\left(\frac{3(1+a^2)^2}{4}\right)\|\theta_t\|^2_{L^2(\Omega_\tau)}+\|\theta_{xx}\|^2_{L^2(\Omega_\tau)}\\ &\quad+(1+a^2)\|\theta_x(\cdot,\tau)\|^2_{L^2(0,1)}+(1+a^2)k(\tau)\theta^2(1,\tau)\\ &\leq\frac{3}{2}\|f\|^2_{L^2(\Omega_\tau)}+c\Big(\|m\|^2_{H^1(0,\tau)}+\|\varphi\|^2_{H^1(0,1)}\Big)\\ &\quad+(1+a^2)\int_0^\tau k'(t)\theta^2(1,t)dt+\frac{(1+a^2)}{2T}\int_0^\tau k(t)\theta^2(1,t)dt.\end{aligned}\tag{4.18}$$

In particular, we have that

$$\begin{aligned}&(1+a^2)\|\theta_x(\cdot,\tau)\|^2_{L^2(0,1)}+(1+a^2)k(\tau)\theta^2(1,\tau)\\ &\leq\frac{3}{2}\|f\|^2_{L^2(\Omega_\tau)}+c\Big(\|m\|^2_{H^1(0,\tau)}+\|\varphi\|^2_{H^1(0,1)}\Big)\\ &\quad+(1+a^2)\int_0^\tau k'(t)\theta^2(1,t)dt+\frac{(1+a^2)}{2T}\int_0^\tau k(t)\theta^2(1,t)dt.\end{aligned}\tag{4.19}$$

Integrating (4.19) with respect to τ over the interval [0,T] and then combining terms yields

$$\begin{aligned}&(1+a^2)\|\theta_x\|^2_{L^2(\Omega_T)}+\frac{(1+a^2)}{2}\int_0^T k(\tau)\theta^2(1,\tau)d\tau\\ &\leq\left(\frac{3}{2}\right)T\|f\|^2_{L^2(\Omega_T)}+c\Big(\|m\|^2_{H^1(0,T)}+\|\varphi\|^2_{H^1(0,1)}\Big)\\ &\quad+(1+a^2)T\|k'\|_{L^\infty(R)}\int_0^T\theta^2(1,t)dt.\end{aligned}\tag{4.20}$$

Using (4.18) with τ=T and (4.20) together we obtain

$$\begin{aligned}\left(\frac{3(1+a^2)^2}{4}\right)\|\theta_t\|^2_{L^2(\Omega_T)} + \|\theta_{xx}\|^2_{L^2(\Omega_T)} + \frac{(1+a^2)}{T}\|\theta_x\|^2_{L^2(\Omega_T)} \\ \le 3\|f\|^2_{L^2(\Omega_T)} + c\Big(\|m\|^2_{H^1(0,T)} + \|\varphi\|^2_{H^1(0,1)}\Big) \\ + 2(1+a^2)\|k'\|_{L^\infty(R)}\int_0^T \theta^2(1,t)dt.\end{aligned} \tag{4.21}$$

Using the elementary estimate $\|\theta(1,\cdot)\|^2_{L^2(0,T)} \le 2\|m\|^2_{L^2(0,T)} + 2\|\theta_x\|^2_{L^2(\Omega_T)}$ in (4.17) and rearranging terms yields the first result.

To obtain the second result, we proceed in a similar way using appropriate choices of ϵ in (4.13) and (4.14) to obtain the estimate

$$\begin{aligned}a(1+a^2)^2\|\theta_t\|^2_{L^2(\Omega_\tau)} + \|\theta_{xx}\|^2_{L^2(\Omega_\tau)} + \frac{(1+a^2)}{T}\|\theta_x\|^2_{L^2(\Omega_\tau)} + (1+a^2)\|\theta_x(\cdot,\tau)\|^2_{L^2(0,1)} \\ \le \frac{(1+a^2)^2}{2a^2}\|f\|^2_{L^2(\Omega_\tau)} + C\Big(\|m\|^2_{H^1(0,\tau)} + \|\varphi\|^2_{H^1(0,1)}\Big) \\ + 2a(1+a^2)\int_0^\tau \left|k'(\eta)\Big(\int_0^1 \xi_t(x,t)dx\Big)\theta^2(1,t)\right|dt.\end{aligned} \tag{4.22}$$

Here $C = C(a,T,\|k\|_{W^1_\infty(R)})$ is a constant that depends only on a, T, and $\|k\|_{W^1_\infty(R)}$, and we will use C to designate such a constant from this point on until the end of the proof. It is also easy to check that, for fixed a and $\|k\|_{W^1_\infty(R)}$, that C is an increasing function of T. It will be apparent in the next section why we wish the coefficients of $\|\theta_t\|^2_{L^2(\Omega_\tau)}$ and $\|f\|^2_{L^2(\Omega_\tau)}$ to take on the above form. We now proceed to estimate the last integral on the right. Using the multi-Hölder inequality and then Lemma 4.1, we obtain

$$\begin{aligned}&\int_0^\tau \left|k'(\eta)\Big(\int_0^1 \xi_t(x,t)dx\Big)\theta^2(1,t)\right|dt \\ &\le \|k'(\eta)\|_{L^6(0,\tau)}\|\xi_t\|_{L^2(\Omega_\tau)}\|\theta(1,\cdot)\|^2_{L^6(0,\tau)} \\ &\le c\tau^{1/6}\|\xi_t\|_{L^2(\Omega_\tau)}\Bigg(\|\theta_t\|^2_{L^2(\Omega_\tau)} + \|\theta_x\|^2_{L^2(\Omega_\tau)} + (1+\tau^2)\,\|\varphi\|^2_{L^\infty(0,1)} \\ &\qquad + \left(1+\tau^{1/3}+\tau\right)\|m\|^2_{L^\infty(0,\tau)}\Bigg).\end{aligned}$$

Here c is a constant that depends only on $\|k'\|_{L^\infty(R)}$. Using this estimate in (4.22) and taking appropriate suprema over all τ in $[0,T]$ leads to

$$\begin{aligned}&a(1+a^2)^2\|\theta_t\|^2_{L^2(\Omega_T)} + \|\theta_{xx}\|^2_{L^2(\Omega_T)} + (1+a^2)\sup_{0\le\tau\le T}\|\theta_x(\cdot,\tau)\|^2_{L^2(0,1)} \\ &\quad + \frac{(1+a^2)}{T}\|\theta_x\|^2_{L^2(\Omega_\tau)} \\ &\le \frac{(1+a^2)^2}{a^2}\|f\|^2_{L^2(\Omega_T)} + C\Big(\|m\|^2_{H^1(0,T)} + \|\varphi\|^2_{H^1(0,1)}\Big) \\ &\quad + cT^{1/6}\|\xi_t\|_{L^2(\Omega_T)}\Big(\|\theta_t\|^2_{L^2(\Omega_T)} + \|\theta_x\|^2_{L^2(\Omega_T)} + (1+T^2)\,\|\varphi\|^2_{L^\infty(0,1)} \\ &\qquad + \left(1+T^{1/3}+T\right)\|m\|^2_{L^\infty(0,\tau)}\Big).\end{aligned}$$

Here c depends upon a as well as $\|k'\|_{L^\infty(R)}$. Rearranging the terms yields (4.11).

To obtain the estimate (4.9) for a general θ we apply the forward averaging operator $(1/h)\int_t^{t+h}(\cdot)d\tau$ to both sides of (4.9)–(4.13) to obtain

$$(4.23) \qquad (1+a^2)(\theta_h)_t - (\theta_h)_{xx} = f_h \quad \text{in } \Omega_{T-h},$$

$$(4.24) \qquad \theta_h(x,0) = \varphi_h(x), \qquad 0 < x < 1,$$

$$(4.25) \qquad \theta_h(0,t) = m_h(t), \qquad 0 < t < T-h,$$

$$(4.26) \qquad (k(t)\theta(1,t))_h + (\theta_h)_x(1,t) = 0, \qquad 0 < t < T-h,$$

where in each case the subscripted variable is the forward average of its unsubscripted counterpart. Note that in (4.26) we may replace $(k(t)\theta(1,t))_h$ by $k(t)\theta_h(1,t)$ and still have the right-hand side converge to zero in the norm of $H^1(0,T)$ since $(k(t)\theta(1,t))_h - k(t)\theta_h(1,t)$ converges to zero in $H^1(0,T)$ as h tends to zero. Since θ_h possesses the additional smoothness properties assumed at the start of the proof, we may apply the previous manipulations to (4.23)–(4.26), and then let h tend to zero to obtain the result. The same technique may be applied to produce (4.11).

The following auxiliary result is a generalization of the main result of [SS2], which treats the case when k is a constant.

THEOREM 4.3. *Let m, φ, and k be as in Lemma 4.2. Then there exists a unique θ in $W_2^{2,1}(\Omega_T)$, which satisfies the conditions*

$$\begin{aligned} (1+a^2)\theta_t - \theta_{xx} &= a\frac{d}{dt}\max\left\{a\int_0^1 \theta(\xi,t)d\xi - g, 0\right\} \quad \text{in } \Omega_T,\\ \theta(x,0) &= \varphi(x), \qquad 0<x<1,\\ \theta(0,t) &= m(t), \qquad 0<t<T,\\ k(t)\theta(1,t) + \theta_x(1,t) &= 0, \qquad 0<t<T. \end{aligned}$$

Proof. We establish the result by applying Theorem 3.1 and begin by defining the various items found in the assumptions of that theorem. We suppose first that k satisfies $k_{\min} = \inf k > 0$. Let

$$\begin{aligned} K &= \left\{(\theta, \theta(1,t)) : \theta \in W_2^{2,1}(\Omega_T),\ \theta(x,0) = \varphi(x),\ \theta(0,t) = m(t)\right\},\\ Y &= L^2(\Omega_T) \oplus L^2(0,T),\\ Z &= L^2(\Omega_T) \oplus 0,\\ \kappa &= \text{ the norm topology of } W_2^{2,1}(\Omega_T), \end{aligned}$$

and

$$\rho(\theta) = \left(\frac{3(1+a^2)^2}{4}\|\theta_t\|^2_{L^2(\Omega_T)} + \|\theta_{xx}\|^2_{L^2(\Omega_T)} + \frac{(1+a^2)}{T}(1 - 4T\|k'\|_{L^\infty(R)})\|\theta_x\|^2_{L^2(\Omega_T)}\right)^{1/2}.$$

We shall identify elements of K with their first components. Note that, for T sufficiently small, ρ is a proper convex functional on K and ρ-bounded subsets of K are Y-compact since ρ-bounded subsets of K are $W_2^{2,1}(\Omega_T)$-norm bounded. Define the operators A, B from K into Y^* by

$$A\theta = \Big((1+a^2)\theta_t - \theta_{xx}, (\theta_x + k(t)\theta)(1,t)\Big)$$

and

$$B\theta = \left(a\frac{d}{dt}\max\left(a\int_0^1 \theta(x,t)dx \;-g,0\right),0\right).$$

We now verify that the hypotheses of Theorem 3.1 are satisfied.

(A1) The operator A is easily seen to be κ-to-weak* continuous since if $\theta_n \to \theta$ in $W_2^{2,1}(\Omega_T)$, then $(\theta_n)_t \to \theta_t$ and $(\theta_n)_{xx} \to \theta_{xx}$ in $L^2(\Omega_T)$, while $\theta_n(1,\cdot) \to \theta(1,\cdot)$ in $L^\infty(0,T)$ and $(\theta_n)_x(1,\cdot) \to \theta_x(1,\cdot)$ in $L^2(0,T)$. The continuity of the operator B has already been demonstrated in [SS2], and its range is obviously in Z.

(A2) To show that A is monotone suppose that θ_1, θ_2 are in K and that $\langle A\theta_1 - A\theta_2, \theta_1 - \theta_2\rangle \leq 0$. Let $z = \theta_1 - \theta_2$. Then

$$\begin{aligned}\langle A\theta_1 - A\theta_2, \theta_1 - \theta_2\rangle &= \int_{\Omega_T} \left[(1+a^2)(\theta_1-\theta_2)_t - (\theta_1-\theta_2)_{xx}\right](\theta_1-\theta_2)\\ &\quad + \int_0^T \left[(\theta_1-\theta_2)_x(1,t) + k(t)(\theta_1(1,t)-\theta_2(1,t))\right](\theta_1-\theta_2)(1,t)dt\\ &= \left(\frac{1}{2}\right)(1+a^2)\|z(\cdot,T)\|^2_{L^2(0,1)} - \int_0^T z_x(1,t)z(1,t)dt\\ &\quad + \|z_x\|^2_{L^2(\Omega_T)} + \int_0^T z_x(1,t)z(1,t)dt + \int_0^T k(t)z^2(1,t)dt\\ &\leq 0.\end{aligned}$$

It follows that $z = 0$ and hence $\langle A\theta_1 - A\theta_2, \theta_1 - \theta_2\rangle > 0$ if $\theta_1 \neq \theta_2$.

(A3) Define $F : Z \to R$ by $F\big((f,0)\big) = \sqrt{3}\|f\|_{L^2(0,T)}$, and let

$$\beta = \left(c\|m\|^2_{H^1(0,T)} + c\|\varphi\|^2_{H^1(0,1)}\right)^{1/2},$$

$$\lambda = \frac{2a^2}{(1+a^2)},$$

and

$$\gamma = 0.$$

Here c is the constant of (4.9). Then the following conditions are satisfied.

(i) The inequalities (3.2) and (3.3) hold since for every pair $(f,0)$ in Z there exists a θ in $W_2^{2,1}(\Omega_T)$ satisfying the standard initial-boundary value problem (4.5)–(4.8) [LM, p. 33], and hence the a priori estimate (4.9).

(ii) The inequality (3.4) holds since

$$\|B\theta\|^2_Z \leq a^4\|\theta_t\|^2_{L^2(0,T)} \leq \frac{4a^4}{3(1+a^2)^2}\rho^2(\theta).$$

(iii) The inequality (3.4) holds simply with $c_{y^*} = \|y^*\|$.

Finally, note that $\lambda < 1$ since $a < 1$. This completes the verification of the hypotheses of Theorem 3.1. Consequently, we obtain, if T satisfies $1 - 4T\|k'\|_{L^\infty} > 0$, a function θ in K that satisfies $\langle A\theta - B\theta, \xi - \theta\rangle \geq 0$ for all ξ in K. If we now put $\xi = \theta \pm \psi$, where ψ is in $C^\infty(\Omega_T)$, with $\psi(x,0) = \psi(0,t) = 0$, then it follows that $\langle A\theta - B\theta, \psi\rangle = 0$ for all such ψ. Hence $A\theta = B\theta$, i.e., θ satisfies all the conclusions of the theorem.

To obtain a solution for a general T we appeal to a standard continuation argument. Choose $T_0 > 0$ so that $1 - 4T_0\|k'\|_{L^\infty(R)} > 0$, and then choose n so that

$nT_0 > T > (n-1)T_0$. Let $T_i = iT_0$ for $i = 0, 1, \cdots n-1$, and let $T_n = T$. Then the preceding argument shows that for each i there is a function θ_{i+1} defined on $\Omega_i = (0,1) \times (T_i, T_{i+1})$ and satisfying the conditions

$$\begin{gathered}
(1+a^2)(\theta_{i+1})_t - (\theta_{i+1})_{xx} = a\frac{d}{dt}\max\left\{a\int_0^1 \theta_{i+1}(\xi,t)d\xi - g, 0\right\} \quad \text{in } \Omega_i, \\
\theta_{i+1}(x,T_i) = \theta_i(x,T_i), \qquad 0 < x < 1, \\
\theta_{i+1}(0,t) = m(t), \qquad T_i < t < T_{i+1}, \\
k(t)\theta_{i+1}(1,t) + (\theta_{i+1})_x(1,t) = 0, \qquad T_i < t < T_{i+1}.
\end{gathered}$$

If we now define θ on Ω_T by $\theta(x,t) = \theta_{i+1}(x,t)$ for (x,t) in $(0,1) \times (T_i, T_{i+1})$, then θ is in $W_2^{2,1}(\Omega_T)$ and satisfies the conclusions of the theorem.

Finally, to obtain a solution for a general $k \geq 0$, let $k_n = k + (1/n)$, and let θ_n be the corresponding solution. We have already seen that if f_n is the first component of $B\theta_n$, then $\|f_n\|_{L^2(\Omega_T)} \leq a^2\|(\theta_n)_t\|_{L^2(\Omega_T)}$. Hence the estimate (4.9) shows that

$$\begin{aligned}
\left(\frac{3(1+a^2)^2}{4} - 3a^4\right) &\|(\theta_n)_t\|^2_{L^2(\Omega_T)} + \|(\theta_n)_{xx}\|^2_{L^2(\Omega_T)} \\
&+ \frac{(1+a^2)}{T}(1 - 4T\|k'\|_{L^\infty(R)})\|(\theta_n)_x\|^2_{L^2(\Omega_T)} \\
&\leq c\left(\|m\|^2_{H^1(0,T)} + \|\varphi\|^2_{H^1(0,1)}\right).
\end{aligned}$$

Note that $3(1+a^2)^2/4 - 3a^4 > 0$ since $a < 1$. Consequently, if $1 - 4T\|k'\|_{L^\infty(R)} > 0$, then an expression equivalent to the $W_2^{2,1}(\Omega_T)$ norm of the θ_n's may be bounded by the initial and boundary data. Hence the θ_n's form a relatively weakly compact set in $W_2^{2,1}(\Omega_T)$, and any weak cluster point θ of the set will satisfy $\theta_x(1,t) = k(t)\theta(1,t)$, and will, therefore, satisfy the conclusion of the theorem. To obtain a solution for a general T we can again use a standard continuation argument. This completes the proof of existence. Because uniqueness may be established using the kinds of manipulations that are found in the next section, we postpone further mention of the proof until then.

5. Existence, uniqueness, and stability.

To complete the proof of the existence of a solution to (2.22) we now prepare to employ Schauder's fixed point theorem. We suppose first that $k_{\min} = \inf k > 0$. For each ξ in $W_2^{2,1}(\Omega_T)$, define $\theta = S\xi$ in $W_2^{2,1}(\Omega_T)$ as a solution to the initial-boundary value problem

$$(1+a^2)\theta_t - \theta_{xx} = a\frac{d}{dt}\max\left\{a\int_0^1 \theta(x,t)dx - g, 0\right\} \quad \text{in } \Omega_T, \tag{5.1}$$

$$\theta(x,0) = \varphi(x), \qquad 0 < x < 1, \tag{5.2}$$

$$\theta(0,t) = m(t), \qquad 0 < t < T, \tag{5.3}$$

$$k\left(g - a\int_0^1 \xi(x,t)dx\right)\theta(1,t) + \theta_x(1,t) = 0, \qquad 0 < t < T. \tag{5.4}$$

Note that the existence of such a θ is guaranteed by Theorem 4.3. Now since

$$\left\|a\frac{d}{dt}\max\left\{a\int_0^1 \theta(x,t)dx - g, 0\right\}\right\|^2_{L^2(0,T)} \leq a^4\|\theta_t\|^2_{L^2(0,T)},$$

the estimate (4.11) implies that

$$\begin{aligned}(5.5)\qquad &(a(1-a)(1+a^2)^2 - cT^{1/6}\|\xi_t\|_{L^2(\Omega_T)})\|\theta_t\|^2_{L^2(\Omega_T)} + \|\theta_{xx}\|^2_{L^2(\Omega_T)} \\ &\quad + \left(\frac{(1+a^2)}{T} - cT^{1/6}\|\xi_t\|_{L^2(\Omega_T)}\right)\|\theta_x\|^2_{L^2(\Omega_T)} \\ &\quad + (1+a^2)\sup_{0\le\tau\le T}\|\theta_x(\cdot,\tau)\|^2_{L^2(0,1)} \\ &\le C\Big(\|m\|^2_{H^1(0,T)} + \|\varphi\|^2_{H^1(0,1)}\Big) \\ &\quad + cT^{1/6}\|\xi_t\|_{L^2(\Omega_T)}\Big(\left(1+T^2\right)\|\varphi\|^2_{L^\infty(0,1)} + \left(1+T^{1/3}+T\right)\|m\|^2_{L^\infty(0,T)}\Big),\end{aligned}$$

where $c = c(a, \|k'\|_{L^\infty(R)})$ is a constant that depends only on a and $\|k'\|_{L^\infty(R)}$, and $C = C(a,T,\|k\|_{W^1_\infty(R)})$ is a constant that depends only on a, T, and $\|k\|_{W^1_\infty(R)}$. We now choose T_0 so that

$$cT_0^{1/6}\left(\frac{2}{a(1-a)(1+a^2)^2}\right)^{1/2}\Big(\left(1+T^2\right)\|\varphi\|^2_{L^\infty(0,1)} + \left(1+T^{1/3}+T\right)\|m\|^2_{L^\infty(0,T)}\Big) < \frac{1}{2},$$

and then choose $M > 1$ so that $M > 2C(a,T,\|k\|_{W^1_\infty(R)})(\|m\|^2_{H^1(0,T)} + \|\varphi\|^2_{H^1(0,1)})$. Note that any smaller value for T_0 will satisfy the above inequality. Consequently, we may choose $0 < T_0 \le \min(1,T)$ so that the following additional conditions hold:

$$a(1-a)(1+a^2)^2 - cT_0^{1/6}\left(\frac{2M}{a(1-a)(1+a^2)^2}\right)^{1/2} > \frac{a(1-a)(1+a^2)^2}{2},$$

$$\frac{(1+a^2)}{T_0} - cT_0^{1/2}\left(\frac{2M}{a(1-a)(1+a^2)^2}\right)^{1/2} > (1+a^2).$$

We now define

$$\begin{aligned}K = \Big\{\xi \in W_2^{2,1}(\Omega_T) :\ &\xi(x,0) = \varphi(x), \xi(0,t) = m(t), \\ &\frac{a(1-a)(1+a^2)^2}{2}\|\xi_t\|^2_{L^2(\Omega_{T_0})} + \|\xi_{xx}\|^2_{L^2(\Omega_{T_0})} + (1+a^2)\|\xi_x\|^2_{L^2(\Omega_{T_0})} \le M\Big\}.\end{aligned}$$

Then, recalling that C is an increasing function of T, (5.5) shows that S maps K into K. Note that K is compact when regarded as a subset of $L^2(\Omega_{T_0})$ since it is bounded in $W_2^{2,1}(\Omega_{T_0})$, and $W_2^{2,1}(\Omega_{T_0})$ embeds compactly into $L^2(\Omega_{T_0})$ [LSU, p. 74]. We now need only show that S is well defined, i.e., that the solution to (5.1)–(5.4) is unique, and that if ξ_n and ξ are in K and $\xi_n \to \xi$ in $L^2(\Omega_{T_0})$, then $S\xi_n \to S\xi$ in $L^2(\Omega_{T_0})$. Since the proofs of these two statements require essentially the same manipulations, we content ourselves with presenting a proof of the latter. Toward this end, let $\theta_n = S\xi_n, \theta = S\xi$, and $w_n = \int_0^t(\theta - \theta_n)$. Note that w_n must satisfy the conditions

$$\begin{aligned}(5.6)\qquad (1+a^2)(w_n)_t - (w_n)_{xx} &= f(\theta,\theta_n), \\ (w_n)(x,0) &= 0, \\ (w_n)(0,t) &= 0, \\ -(w_n)_x(1,t) &= \int_0^t (k_n)(w_n)_s(1,s)ds + \int_0^t (k_n - k_0)\theta(1,s)ds,\end{aligned}$$

where

$$f(\theta,\theta_n)=a\max\left\{a\int_0^1\theta(x,t)dx-g,0\right\}-a\max\left\{a\int_0^1\theta_n(x,t)dx-g,0\right\},$$

$$k_n=k(\eta_n)=k\left(g-a\int_0^1\xi_n(x,s)dx\right),$$

$$k_0=k(\eta)=k\left(g-a\int_0^1\xi(x,s)dx\right).$$

Note that, for each τ satisfying $0<\tau<T$, we have that

$$\|f(\theta,\theta_n)\|^2_{L^2(\Omega_\tau)}\le a^4\|w_t\|^2_{L^2(\Omega_\tau)}.$$

Squaring both sides of (5.6) and integrating over Ω_τ, we obtain

$$\begin{aligned}(1+a^2)^2\|(w_n)_t\|^2_{L^2(\Omega_\tau)}-2(1+a^2)\int_{\Omega_\tau}(w_n)_t(w_n)_{xx}\,dx\,dt+\|(w_n)_{xx}\|^2_{L^2(\Omega_\tau)}&\\ =\|f(\theta,\theta_n)\|^2_{L^2(\Omega_\tau)}&\\ \le a^4\|(w_n)_t\|^2_{L^2(\Omega_\tau)}.&\end{aligned}\tag{5.7}$$

We now proceed, as in Lemma 4.1, to apply integration by parts to the integral on the left of (5.7) and thus obtain

$$\begin{aligned}-2(1+a^2)\int_{\Omega_\tau}(w_n)_t(w_n)_{xx}\,dx\,dt&=(1+a^2)\|(w_n)_x(\cdot,\tau)\|^2_{L^2(0,1)}\\ &\quad+2(1+a^2)\int_0^\tau(w_n)_t(1,t)\int_0^t(k_n)(w_n)_s(1,s)\,ds\,dt\\ &\quad+2(1+a^2)\int_0^\tau(w_n)_t(1,t)\int_0^t(k_n-k_0)\theta(1,s)\,ds\,dt.\end{aligned}\tag{5.8}$$

Applying integration by parts twice to the first integral on the right in (5.8) yields

$$\begin{aligned}&\int_0^\tau(w_n)_s(1,t)\int_0^t(k_n)(w_n)_t(1,s)ds\,dt\\ &\quad=w(1,\tau)\int_0^\tau(k_n)(w_n)_s(1,s)\,ds\\ &\qquad-\int_0^\tau(w_n)(1,t)(k_n)(w_n)_t(1,t)\,dt\\ &\quad=\frac{1}{2}w_n^2(1,\tau)k_n(\tau)\\ &\qquad+aw_n(1,\tau)\int_0^\tau k'(\eta_n)\left(\int_0^1(\xi_n)_s(x,s)dx\right)w_n(1,s)\,ds\\ &\qquad-\frac{1}{2}a\int_0^\tau k'(\eta_n)\left(\int_0^1(\xi_n)_s(x,s)dx\right)w_n^2(1,s)\,ds,\end{aligned}\tag{5.9}$$

where the prime denotes differentiation with respect to the argument. Applying a single integration by parts to the second integral in (5.8) yields

$$\begin{aligned}\int_0^\tau(w_n)_t(1,t)\int_0^t(k_n-k_0)\theta(1,s)ds\,dt&=w_n(1,\tau)\int_0^\tau(k_n-k_0)\theta(1,s)ds\\ &\quad-\int_0^\tau(w_n)(1,t)(k_n-k_0)\theta(1,t)dt.\end{aligned}\tag{5.10}$$

Using the results of (5.8)–(5.10) in (5.7) and rearranging terms, we obtain

$$\begin{aligned}(1+2a^2)\|(w_n)_t\|^2_{L^2(\Omega_\tau)} &+ (1+a^2)\|(w_n)_x(\cdot,\tau)\|^2_{L^2(0,1)}\\ &+ \|(w_n)_{xx}\|^2_{L^2(\Omega_\tau)} + (1+a^2)w_n^2(1,\tau)k_n(\tau)\\ &\leq -2a(1+a^2)w_n(1,\tau)\int_0^\tau k'(\eta_n)\left(\int_0^1(\xi_n)_s(x,s)\,dx\right)w_n(1,s)ds\\ &+ a(1+a^2)\int_0^\tau k'(\eta_n)\left(\int_0^1(\xi_n)_s(x,s)dx\right)w_n^2(1,s)\,ds\\ &- 2(1+a^2)w_n(1,\tau)\int_0^t(k_n-k_0)\theta(1,s)ds\\ &+ 2(1+a^2)\int_0^\tau(w_n)(1,t)(k_n-k_0)\theta(1,t)dt.\end{aligned}\tag{5.11}$$

We now estimate the four integrals on the right in (5.11). The first integral may be estimated using the Cauchy–Schwarz inequality and then Cauchy's inequality with ϵ:

$$\begin{aligned}&\left|2a(1+a^2)w_n(1,\tau)\int_0^\tau k'(\eta_n)\left(\int_0^1(\xi_n)_s(x,s)dx\right)w_n(1,s)ds\right|\\ &\quad\leq 2a(1+a^2)|w_n(1,\tau)|\ \|k'\|_{L^\infty(R)}\|(\xi_n)_s\|_{L^2(\Omega_\tau)}\|w_n(1,\cdot)\|_{L^2(0,\tau)}\\ &\quad\leq \frac{(1+a^2)}{4}w_n^2(1,\tau)k_n(\tau)\\ &\qquad+ 4a^2(1+a^2)k_n(\tau)^{-1}\|k'\|^2_{L^\infty(R)}\|(\xi_n)_s\|^2_{L^2(\Omega_\tau)}\|w_n(1,\cdot)\|^2_{L^2(0,\tau)}.\end{aligned}\tag{5.12}$$

The second integral may be estimated using the Cauchy–Schwarz inequality and then Cauchy's inequality with ϵ:

$$\begin{aligned}&a(1+a^2)\int_0^\tau\left|k'(\eta_n)\left(\int_0^1(\xi_n)_s(x,t)dx\right)w_n^2(1,t)\right|dt\\ &\leq a(1+a^2)\|k'\|_{L^\infty(R)}\|(\xi_n)_t\|_{L^2(\Omega_\tau)}\|(w_n)(1,\cdot)\|_{L^\infty(0,\tau)}\|(w_n)(1,\cdot)\|_{L^2(0,\tau)}\\ &\leq k_{\min}\frac{(1+a^2)}{4}\|(w_n)(1,\cdot)\|^2_{L^\infty(0,\tau)}\\ &\quad+ k_{\min}^{-1}a^2(1+a^2)\|k'\|^2_{L^\infty(R)}\|(\xi_n)_s\|^2_{L^2(\Omega_\tau)}\|(w_n)(1,\cdot)\|^2_{L^2(0,\tau)}\end{aligned}\tag{5.13}$$

The last two integrals are easily estimated using Cauchy's inequality with ϵ:
(5.14)

$$\begin{aligned}\left|2(1+a^2)w_n(1,\tau)\int_0^\tau(k_n-k_0)\theta(1,s)ds\right| &\leq \frac{(1+a^2)}{4}w_n^2(1,\tau)k_n(\tau)\\ &\quad+ 4(1+a^2)k_n(\tau)^{-1}\int_0^\tau(k_n-k_0)^2\theta^2(1,s)ds,\\ \left|2(1+a^2)\int_0^\tau(w_n)(1,t)(k_n-k_0)\theta(1,t)dt\right| &\leq (1+a^2)\int_0^\tau(w_n)^2(1,t)dt\\ &\quad+ (1+a^2)\int_0^\tau(k_n-k_0)^2\theta^2(1,t)dt.\end{aligned}$$

Using the results of (5.12)–(5.14) in (5.11) and rearranging terms we have that

$$\begin{aligned}(1+2a^2)\|(w_n)_t\|^2_{L^2(\Omega_\tau)} &+ (1+a^2)\|(w_n)_x(\cdot,\tau)\|^2_{L^2(0,1)} \\ &+ \|(w_n)_{xx}\|^2_{L^2(\Omega_\tau)} + \frac{(1+a^2)}{2} w_n^2(1,\tau)k_n(\tau) \\ &\le k_{\min}\frac{(1+a^2)}{4}\|(w_n)(1,\cdot)\|^2_{L^\infty(0,\tau)} \\ &+ (1+a^2)\left(5a^2k_{\min}^{-1}\|k'\|^2_{L^\infty(R)}\|(\xi_n)_t\|^2_{L^2(\Omega_\tau)}+1\right)\|(w_n)(1,\cdot)\|^2_{L^2(0,\tau)} \\ &+ 4(1+a^2)(k_{\min}^{-1}+1)\int_0^\tau (k_n-k_0)^2\theta^2(1,s)ds.\end{aligned} \tag{5.15}$$

In particular, we have that

$$\begin{aligned}&\frac{(1+a^2)}{2} w_n^2(1,\tau)k_n(\tau) \\ &\le k_{\min}\frac{(1+a^2)}{4}\|(w_n)(1,\cdot)\|^2_{L^\infty(0,\tau)} \\ &\quad + (1+a^2)\left(5a^2k_{\min}^{-1}\|k'\|^2_{L^\infty(R)}\|(\xi_n)_s\|^2_{L^2(\Omega_\tau)}+1\right)\|(w_n)(1,\cdot)\|^2_{L^2(0,\tau)} \\ &\quad + 4(1+a^2)(k_{\min}^{-1}+1)\int_0^\tau (k_n-k_0)^2\theta^2(1,s)ds \\ &\le k_{\min}\frac{(1+a^2)}{4}\|(w_n)(1,\cdot)\|^2_{L^\infty(0,\nu)} \\ &\quad + (1+a^2)\left(5a^2k_{\min}^{-1}\|k'\|^2_{L^\infty(R)}\|(\xi_n)_s\|^2_{L^2(\Omega_\nu)}+1\right)\|(w_n)(1,\cdot)\|^2_{L^2(0,\nu)} \\ &\quad + 4(1+a^2)(k_{\min}^{-1}+1)\int_0^\nu (k_n-k_0)^2\theta^2(1,s)\,ds\end{aligned} \tag{5.16}$$

for any ν satisfying $\tau \le \nu \le T_0$. It follows that

$$\begin{aligned}&k_{\min}\frac{(1+a^2)}{2}\|(w_n)(1,\cdot)\|^2_{L^\infty(0,\nu)} \\ &\le k_{\min}\frac{(1+a^2)}{4}\|(w_n)(1,\cdot)\|^2_{L^\infty(0,\nu)} \\ &\quad + (1+a^2)\left(5a^2k_{\min}^{-1}\|k'\|^2_{L^\infty(R)}\|(\xi_n)_s\|^2_{L^2(\Omega_\nu)}+1\right)\|(w_n)(1,\cdot)\|^2_{L^2(0,\nu)} \\ &\quad + 4(1+a^2)(k_{\min}^{-1}+1)\int_0^\nu (k_n-k_0)^2\theta^2(1,s)ds,\end{aligned} \tag{5.17}$$

and hence

$$\begin{aligned}(w_n)^2(1,\nu) \le &\left(20a^2k_{\min}^{-2}\|k'\|^2_{L^\infty(R)}\|(\xi_n)_s\|^2_{L^2(\Omega_\nu)}+4k_{\min}^{-1}\right)\|(w_n)(1,\cdot)\|^2_{L^2(0,\nu)} \\ &+ 16k_{\min}^{-1}(k_{\min}^{-1}+1)\int_0^\nu (k_n-k_0)^2\theta^2(1,s)ds.\end{aligned} \tag{5.18}$$

Consequently, by Gronwall's inequality [LSU, p.94], we have that
(5.19)

$$\begin{aligned}\int_0^{T_0} w_n^2(1,\tau)d\tau \le &\exp\left(\int_0^{T_0}\left(20a^2k_{\min}^{-2}\|k'\|^2_{L^\infty(R)}\|(\xi_n)_s\|^2_{L^2(\Omega_\tau)}+4k_{\min}^{-1}\right)d\tau\right) \\ &\times 16k_{\min}^{-1}(k_{\min}^{-1}+1)\left(\int_0^{T_0}\int_0^\tau (k_n-k_0)^2\theta^2(1,s)\,ds\,d\tau\right).\end{aligned}$$

Now

$$|k_n - k_0| \leq a\|k'\|_{L^\infty(R)} \int_0^1 |(\xi_n - \xi)(x,s)|dx, \tag{5.20}$$

and, consequently, (5.19) shows that as $\xi_n \to \xi$ in $L^2(\Omega_{T_0})$ we have that $w_n(1,\cdot) \to 0$ in $L^2(0,T_0)$. Using this fact in (5.17) and then (5.15), we obtain that $(w_n)_t = \theta - \theta_n \to 0$ in $L^2(\Omega_{T_0})$. Consequently, the operator S is continuous when K is viewed with the $L^2(\Omega_{T_0})$-topology and so by Schauder's theorem S must have a fixed point $\theta = S\theta$. This fixed point satisfies all the requirements of the problem (2.22) on Ω_{T_0}. To obtain a solution for a general $k \geq 0$, let $k_n = k + (1/n)$, and let θ_n be the corresponding solution in K. Then the θ_n's form a relatively weakly compact set in $W_2^{2,1}(\Omega_{T_0})$, and any weak cluster point θ of the set will satisfy $\theta_x(1,t) = k(\eta)\theta(1,t)$, and will therefore satisfy the conclusion of the theorem on Ω_{T_0}. This concludes the proof of the existence of a local solution to (2.22).

To show uniqueness of the solution let θ and ξ be two solutions to (2.22) corresponding to the same initial-boundary data, and let $w = \int_0^t(\theta - \xi)$. If we repeat the same manipulations as in (5.6)–(5.11), then we obtain

$$\begin{aligned}
&(1+2a^2)\|w_t\|^2_{L^2(\Omega_\tau)} + (1+a^2)\|w_x(\cdot,\tau)\|^2_{L^2(0,1)} + \|w_{xx}\|^2_{L^2(\Omega_\tau)} + (1+a^2)w^2(1,\tau)k_1(\tau)\\
&\quad \leq -2a(1+a^2)w(1,\tau)\int_0^\tau k'(\eta_\xi)\left(\int_0^1 (\xi)_s(x,s)dx\right)w(1,s)ds\\
&\quad + a(1+a^2)\int_0^\tau k'(\eta_\xi)\left(\int_0^1 (\xi)_s(x,s)dx\right)w^2(1,s)ds\\
&\quad - 2(1+a^2)w(1,\tau)\int_0^t (k_1 - k_0)\theta(1,s)ds\\
&\quad + 2(1+a^2)\int_0^\tau w(1,t)(k_1-k_0)\theta(1,t)dt,
\end{aligned} \tag{5.21}$$

where

$$k_1 = k(\eta_\xi) = k\left(g - a\int_0^1 \xi(x,s)dx\right) \quad \text{and} \quad k_0 = k(\eta_\theta) = k\left(g - a\int_0^1 \theta(x,s)dx\right).$$

We now estimate the four integrals on the right in ways different from what was done in (5.12)–(5.14). The first and second integrals are estimated simply by using the Cauchy–Schwarz inequality:

$$\begin{aligned}
&\left|2a(1+a^2)w(1,\tau)\int_0^\tau k'(\eta_\xi)\left(\int_0^1 \xi_s(x,s)dx\right)w(1,s)ds\right|\\
&\quad \leq 2a(1+a^2)\sup_{0\leq s\leq \tau}|w^2(1,s)|\ \|k'\|_{L^2(R)}\|\xi_s\|_{L^2(\Omega_\tau)}\\
&\quad \leq 2a(1+a^2)\tau^{1/2}\sup_{0\leq s\leq \tau}|w^2(1,s)|\ \|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\tau)},\\
&a(1+a^2)\int_0^\tau \left|k'(\eta_\xi)\left(\int_0^1 \xi_s(x,t)dx\right)w^2(1,t)\right|dt\\
&\quad \leq a(1+a^2)\tau\sup_{0\leq s\leq \tau}|w^2(1,s)|\ \|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\tau)}.
\end{aligned} \tag{5.22}$$

The last two integrals are easily estimated using the Cauchy's inequality with ϵ:

(5.23)

$$\begin{aligned}\left|2(1+a^2)w(1,\tau)\int_0^\tau (k_1-k_0)\theta(1,s)ds\right| \leq &\frac{(1+a^2)}{4}w^2(1,\tau)\\ &+4(1+a^2)\int_0^\tau (k_1-k_0)^2\theta^2(1,s)ds,\\ \left|2(1+a^2)\int_0^\tau w(1,t)(k_1-k_0)\theta(1,t)dt\right| \leq &(1+a^2)\int_0^\tau w^2(1,t)dt\\ &+(1+a^2)\int_0^\tau (k_1-k_0)^2\theta^2(1,t)dt.\end{aligned}$$

Using these facts in (5.21), we have that

$$\begin{aligned}&(1+2a^2)\|w_t\|^2_{L^2(\Omega_\tau)}+(1+a^2)\|w_x(\cdot,\tau)\|^2_{L^2(0,1)}+\|w_{xx}\|^2_{L^2(\Omega_\tau)}\\ &\quad+(1+a^2)w^2(1,\tau)k_1(\tau)\\ &\quad\leq (3a\|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\tau)}+\tau^{1/2})(1+a^2)\tau^{1/2}\sup_{0\leq s\leq\tau}|w^2(1,s)|\\ &\quad+\frac{(1+a^2)}{4}w^2(1,\tau)\\ &\quad+5(1+a^2)\int_0^\tau (k_1-k_0)^2\theta^2(1,s)ds.\end{aligned}\tag{5.24}$$

Now, using the fact that $\|w_x(\cdot,\tau)\|^2_{L^2(0,1)}\geq w^2(1,\tau)$, we have that

$$\begin{aligned}&(1+2a^2)\|w_t\|^2_{L^2(\Omega_\tau)}+\frac{(1+a^2)}{2}\|w_x(\cdot,\tau)\|^2_{L^2(0,1)}\\ &\quad\leq (3a\|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\tau)}+\tau^{1/2})(1+a^2)\tau^{1/2}\sup_{0\leq s\leq\tau}|w^2(1,s)|\\ &\quad+5(1+a^2)\int_0^\tau (k_1-k_0)^2\theta^2(1,s)ds\\ &\quad\leq (3a\|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\tau)}+\nu^{1/2})(1+a^2)\nu^{1/2}\sup_{0\leq s\leq\nu}|w^2(1,s)|\\ &\quad+5(1+a^2)\int_0^\nu (k_1-k_0)^2\theta^2(1,s)ds\end{aligned}\tag{5.25}$$

for any ν satisfying $\tau\leq\nu\leq T$. In particular, we have that

$$\begin{aligned}&(1+a^2)/2\sup\nolimits_{0\leq\tau\leq\nu}\|w_x(\cdot,\tau)\|^2_{L^2(0,1)}\\ &\quad\leq (3a\|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\tau)}+\nu^{1/2})(1+a^2)\nu^{1/2}\sup_{0\leq s\leq\nu}|w^2(1,s)|\\ &\quad+5(1+a^2)\int_0^\nu (k_1-k_0)^2\theta^2(1,s)ds.\end{aligned}\tag{5.26}$$

Using (5.25) and (5.26) and rearranging terms yields

(5.27)

$$\begin{aligned}&(1+2a^2)\|w_t\|^2_{L^2(\Omega_\nu)}\\ &\quad+\left(\frac{(1+a^2)}{2}-2(3a\|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\nu)}+\nu^{1/2})(1+a^2)\nu^{1/2}\right)\sup_{0\leq\tau\leq\nu}\|w_x(\cdot,\tau)\|^2_{L^2(0,1)}\\ &\leq 10(1+a^2)\int_0^\nu (k_1-k_0)^2\theta^2(1,s)ds.\end{aligned}$$

Now, for all ν sufficiently small, we have that

$$\frac{(1+a^2)}{2} - 2(3a\|k'\|_{L^\infty(R)}\|\xi_s\|_{L^2(\Omega_\nu)} + \nu^{1/2})(1+a^2)\nu^{1/2} > 0.$$

Consequently, we have that

$$\begin{aligned} \|w(\cdot,\nu)\|^2_{L^2(0,1)} &\leq (1+a^2)^2\|w_t\|^2_{L^2(\Omega_\nu)} \\ &\leq 10(1+a^2)\int_0^\nu (k_1-k_0)^2\theta^2(1,s)ds \\ &\leq 10(1+a^2)\|\theta(1,\cdot)\|^2_{L^\infty(0,\nu)}\|k'\|^2_{L^\infty(R)}\int_0^\nu\int_0^1 w^2(x,s)\,dx\,ds \end{aligned} \tag{5.28}$$

for all ν less than some T_0. Hence Gronwall's inequality shows that $w=0$, i.e., $\theta=\xi$ on the rectangle $(0,1)\times(0,T_0)$. We may now use a continuation argument to extend the result to all of Ω_T. A similar, but more complicated argument, establishes that if $\varphi_n \to \varphi$ in $H^1(0,1)$ and $m_n \to m$ in $H^1(0,T)$ and if θ_n and θ are the corresponding solutions in $W_2^{2,1}(\Omega_T)$, then $\theta_n \to \theta$ in $L^2(\Omega_{T_0})$ for some T_0 less than T.

REFERENCES

[A] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[BC] C. Baiocchi and A. Capelo, *Variational and Quasivariational Inequalities*, John Wiley, Chichester, UK, 1984.

[Ba1] J. R. Barber, *Contact problems involving a cooled punch*, J. Elasticity, 8 (1978), pp. 409–423.

[Ba2] ———, *Stability of thermoelastic contact*, in Institution of Mechanical Engineers Internat. Conference on Tribology, Mechanical Engineering Publications Ltd, London, 1987, pp. 981–986.

[BDC] J. R. Barber, J. Dundurs, and M. Comninou, *Stability considerations in thermoelastic contact*, J. Appl. Mech., 47 (1980), pp. 871–874.

[BZ] J. R. Barber and R. Zhang, *Transient behaviour and stability for the thermoelastic contact of two rods of dissimilar materials*, Internat. J. Mech. Sci., 30 (1988), pp. 691–704.

[Ca] D. E. Carlson, *Linear thermoelasticity*, in Handbuch der Physik, Vol. VIa/2, S. Flugge, ed., Springer-Verlag, Berlin, 1972, pp. 297–345.

[CD] M. Comninou and J. Dundurs, *On the Barber boundary conditions for thermoelastic contact*, J. Appl. Mech., 46 (1979), pp. 849–853.

[Da] W. A. Day, *Heat Conduction within Linear Thermoelasticity*, Springer-Verlag, New York, 1985.

[Du] G. Duvaut, *Free boundary problems connected with thermoelasticity and unilateral contact*, in Free Boundary Problems: proceedings of a seminar held in Pavia, September–October 1979, Vol. II, Instituto nazionale, Roma, 1980.

[DL] G. Duvaut and J. L. Lions, *Inequations en thermoelasticite et magnetohydrodynamique*, Arch. Rational Mech. Anal., 46 (1972), pp. 241–279.

[GSS] R. P. Gilbert, P. Shi, and M. Shillor, *A quasistatic problem in linear thermoelasticity*, Rend. di Mat., Serie VII, 10 (1990), pp. 785–808.

[LSU] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[LM] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, Vol. II, Springer, New York, 1972.

[Pr] M. Primicerio, private communication.

[RH] O. Richmond and N. C. Huang, *Interface stability during unidirectional solidification of a pure metal*, Proc. 6th Canadian Congress of Appl. Mech. (1977), Vancouver, pp. 453–454.

[SS1] P. Shi and M. Shillor, *Uniqueness and stability of the solution to a thermoelastic contact problem*, Euro. J. Appl. Math., 1 (1990), pp. 371–387.

[SS2] P. Shi and M. Shillor, *A quasistatic contact problem in thermoelasticity with a radiation condition for the temperature*, J. Math. Anal. Appl., to appear.

[SSZ] P. Shi, M. Shillor, and X. L. Zou, *Numerical solutions to one dimensional problems of thermoelastic contact*, Comput. Math. Appl., 22 (1991), pp. 65–78.

[SF] M. G. Srinivasan and D. M. France, *Nonuniqueness in steady state heat transfer in prestressed duplex tubes-analysis and case history*, J. Appl. Mech., 48 (1985), pp. 555–558.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1417-1438, November 1992

003

A DEGENERATE STEFAN-LIKE PROBLEM WITH JOULE'S HEATING*

XIANGSHENG XU†

Abstract. This paper studies the system $(\partial/\partial t)\alpha(u)-\operatorname{div} a(\nabla u)\ni\sigma(u)|\nabla\varphi|^2$, $\operatorname{div}(\sigma(u)\nabla\varphi)=0$ in a bounded domain of $\mathbb{R}^N$ coupled with initial and boundary conditions. Here, α is a maximal monotone graph in $\mathbb{R}$, a a monotone mapping from $\mathbb{R}^N$ to $\mathbb{R}^N$, and σ a positive function on $\mathbb{R}$ with the limit of $\sigma(s)$ as $|s|\to\infty$ being zero. In the generality considered here, the problem may not always have a solution in the sense of distributions. Under certain assumptions on the data, an existence assertion is established for the problem that incorporates the new phenomena involved and, at the same time, retains the main feature of a classical weak solution.

Key words. capacity solutions, degeneracy, compactness

AMS(MOS) subject classifications. primary 35D05, 35K65

1. Introduction. Let Ω be a bounded domain in $\mathbb{R}^N$ with smooth boundary $\partial\Omega$ and T a positive number. Consider the following initial-boundary value problem:

$$\frac{\partial}{\partial t}v-\operatorname{div} a(\nabla u)=\sigma(u)|\nabla\varphi|^2 \quad \text{in } Q_T\equiv\Omega\times(0,T), \tag{1.1a}$$

$$\operatorname{div}(\sigma(u)\nabla\varphi)=0 \quad \text{in } Q_T, \tag{1.1b}$$

$$a(\nabla u)\cdot\nu+f(x,t,u)=0 \quad \text{in } S_T\equiv\partial\Omega\times(0,T), \tag{1.1c}$$

$$\varphi=\varphi_0 \quad \text{in } S_T, \tag{1.1d}$$

$$v=v_0 \quad \text{in } \Omega\times\{0\}, \tag{1.1e}$$

$$v\in\alpha(u) \quad \text{a.e. on } Q_T. \tag{1.1f}$$

Here, α is a maximal monotone graph in $\mathbb{R}$, ν denotes the outward normal to $\partial\Omega$, ∇ (respectively, div) is the gradient (respectively, divergence) in the spatial variable $x\in\mathbb{R}^N$, a is a monotone mapping from $\mathbb{R}^N$ to $\mathbb{R}^N$, and σ is a continuous function on $\mathbb{R}$.

A special case of (1.1) is proposed in [SSX] as a model for the combined processes of heat conduction and electrical conduction in a conductor, which may undergo a change of phase due to the heat generated by the electric current, the so-called Joule heating. An example situation is the process of resistance spot-welding of thin steel sheets [A]. In this case, u is the temperature of the conductor, v the enthalpy density, φ the electrical potential, and $\sigma(u)$ the electrical conductivity. Equations (1.1a) and (1.1f) amount to the employment of the "enthalpy formulation" for the Stefan problem to describe the melting of the conductor as a result of Joule's heating, while equation (1.1b) represents the current conservation. We refer the reader to [SSX] for a detailed discussion on the physical justification of this model.

Mathematical problems related to the combined heat and current flow were considered recently in a number of papers under the title of "the thermistor problem." However, most of the studies only focus on the stationary case; see, e.g., [C2], [C3], [CP], [HRS], [X3], [CF1], [CF2], [CH], [AX], where the effect of various

* Received by the editors April 29, 1991; accepted for publication (in revised form) March 10, 1992. This work was supported in part by National Science Foundation grant DMS-9101382.

† Department of Mathematical Sciences, University of Arkansas, Fayetteville, Arkansas 72701.

assumptions on σ and different boundary conditions on the existence of a solution and regularity properties of the solution is investigated. The time-dependent case was first considered in [C1] for $N=2$. This restriction on the space dimension was later eliminated in [SSX]. However, in both papers it is assumed, among other things, that

$$M_1 \leqq \sigma(s) \leqq M_2 \quad \text{for all } s,$$

where M_1, M_2 are two positive constants. This condition is very crucial to the existence of a classical weak solution. In [X2] σ is only assumed to be positive, but the boundary conditions are linear and no phase change is allowed in the conductor, i.e., $\alpha(s)=s$. In this paper we shall further relax most of the assumptions in [SSX] and [X2], which leads to new mathematical difficulties. Let us make some remarks on the new mathematical aspects of (1.1). For this we need to state our precise assumptions on the data involved first. We shall assume the following:

(A1) The vector field $a(x)=(a_1(x), \cdots, a_N(x))^T$ is continuous on $\mathbb{R}^N$, and satisfies the growth condition

$$|a(x)| \leqq K|x| \quad \text{for } |x| \text{ sufficiently large,} \quad \text{some } K>0.$$

Moreover, $a(x)$ is strongly coercive monotone, i.e.,

$$(a(x)-a(y))\cdot(x-y) \geqq c_0|x-y|^2 \quad \text{for all } x, y \in \mathbb{R}^N$$

for some $c_0>0$.

(A2) The function $f(x, t, \xi)$ is continuous over $\overline{S_T} \times \mathbb{R}$, and satisfies the growth condition

$$|f(x, t, \xi)| \leqq K_0 + K_1|\xi|,$$

where K_0, K_1 are positive constants. Moreover, $\xi \to f(x, t, \xi)$ is monotone at the origin for all $(x, t) \in S_T$, i.e., $f(x, t, \xi)$ sign $\xi \geqq 0$.

(A3) $\sigma(s)$ is continuous on $\mathbb{R}$, and satisfies

$$0 < \sigma(s) \leqq M \quad \text{for all } s \in \mathbb{R}$$

for some $M>0$. Furthermore, $\lim_{|s|\to\infty} \sigma(s)=0$.

(A4) $\alpha(s)$ is given by

$$\alpha(s)=\begin{cases} \alpha_1(s) & \text{if } s>0, \\ [-1, 0] & \text{if } s=0, \\ \alpha_2(s)-1 & \text{if } s<0, \end{cases}$$

where α_1 and α_2 are two known functions that are continuously differentiable on their respective domains, with $\alpha_1(0+)=\alpha_2(0-)=0$. Moreover, for each $\varepsilon>0$ there exist two positive numbers c_1 and c_2 such that

$$c_1 \leqq \alpha'(s) \leqq c_2 \quad \text{for all } s \in \mathbb{R}\backslash(-\varepsilon, \varepsilon).$$

(A5) $\varphi_0 \in L^\infty(0, T; L^\infty(\partial\Omega))$ and there is a function $\overline{\varphi_0}$ in $L^\infty(0, T; W^{1,\infty}(\Omega))$ such that

$$\overline{\varphi_0} = \varphi_0 \quad \text{in } L^\infty(S_T).$$

The initial value v_0 belongs to $L^2(\Omega)$.

Assumption (A4) says that the melting temperature of the conductor is zero. Assumption (A3) arises in the case of metallic conduction; see [C1].

Under these assumptions we see that there are three different types of degeneracy involved in our problem. First, $\alpha'(s)$ may oscillate wildly near zero in the sense that $\alpha'(s)$ may be unbounded above and $\alpha'(s)$ may not be bounded away from zero as $s \to 0$. We mention in passing that results concerning the existence of a solution and regularity properties of the solution for a one-dimensional Stefan-like problem with this feature are established in [X1]. Also, in the references cited there we may find descriptions of several heat conduction phase change processes that give rise to this type of degeneracy in α. Second, (1.1b) does not contain a term involving $(\partial/\partial t)\varphi$. Third, $\sigma(s)$ does not stay away from zero as $|s| \to \infty$. As it is observed in [X2], this condition prevents us from obtaining a solution of (1.1) via the classical weak formulation as it is done in [SSX]. Let us expand on this point a little bit more. In order for us to view the system in the sense of distributions, we should know that φ belongs to $L^2(0, T; W^{1,2}(\Omega))$. This information would be implied by the boundedness of the temperature, which in turn depends upon the regularity of φ. Existing results on the regularity of weak solutions to degenerate parabolic equations of type (1.1a) indicate that there is a gap between the regularity of φ obtained from assuming u is bounded and that needed to yield the boundedness of u. This gap does not seem to be of a technical nature. As a result, the classical notion of a weak solution [SSX] is not appropriate in analyzing (1.1). We conclude that the solution to (1.1) may display new phenomena that cannot be incorporated into the classical notion of a weak solution. It turns out that we are able to employ the notion of a capacity solution developed in [X2] to study (1.1). This notion of a solution is based upon the following observation. Equation (1.1b) degenerates only at points where u is infinity. For each $m > 0$, (1.1b) is uniformly elliptic on the set $E_m \equiv \{(x, t) \in Q_T : |u(x, t)| \leqq m\}$. Thus it is reasonable to expect that $\nabla\varphi$ exists as a vector-valued function on E_m for each m in a certain sense. That is to say, $\nabla\varphi$ in the sense of distributions may not belong to any L^p-space with $p \geqq 1$. Somehow, $\nabla\varphi$ still exists almost everywhere on Q_T as a function. Then we may view $\sigma(u)\nabla\varphi$ as a product of two functions in Q_T. In this way we hope to avoid the above-mentioned gap in our problem. To make our above statements precise, we have the following definition.

Definition. By a capacity solution of (1.1), we mean a quadruplet (u, φ, v, g) such that

(i) $u \in L^2(0, T; W^{1,2}(\Omega))$, $\varphi \in L^\infty(Q_T)$, $v \in L^2(Q_T)$, and $g \in [L^2(0, T; L^2(\Omega))]^N$;

(ii) $v \in \alpha(u)$ almost everywhere on Q_T;

(iii) u, φ, v, g satisfy

$$\begin{aligned}(1.2)\qquad &-\int_{Q_T} v\eta_t + \int_{Q_T} a(\nabla u)\nabla\eta\,dx\,dt + \int_0^T \int_{\partial\Omega} f(x, t, u)\eta\,ds\,dt \\ &= -\int_{Q_T} (\varphi - \overline{\varphi_0})g\nabla\eta\,dx\,dt + \int_{Q_T} g\nabla\overline{\varphi_0}\eta\,dx\,dt + \int_\Omega v_0(x)\eta(x, 0)\,dx\end{aligned}$$

for all $\eta \in H^1(0, T; W^{1,2}(\Omega))$ such that $\eta(x, T) \equiv 0$, and

$$(1.3)\qquad \int_{Q_T} g\nabla\psi\,dx\,dt = 0 \quad \text{for all } \psi \in L^2(0, T; W_0^{1,2}(\Omega));$$

(iv) For each positive integer m and each function θ in $\mathscr{A} \equiv \{b \in C^1(\mathbb{R}): b'(s) \geqq 0$ on $\mathbb{R}$, $b(s) = 0$ on $(-\infty, 0]$, and $b(s) = 1$ on $[1, \infty)\}$ there hold

$$(1.4)\qquad F_\theta^{(m)}\varphi \in L^2(0, T; W^{1,2}(\Omega)),$$

$$F_\theta^{(m)}\varphi = \varphi_0 \quad \text{in } L^2(0, T; L^2(\partial\Omega)), \tag{1.5}$$

$$F_\theta^{(m)}g = \sigma(u)(\nabla(F_\theta^{(m)}\varphi) - \varphi\nabla F_\theta^{(m)}), \tag{1.6}$$

where $F_\theta^{(m)} \equiv 1 - \theta((1/m)|u|)$.

LEMMA 1.1. *Let* (A1)-(A5) *be satisfied. Assume that* (u, φ, v, g) *is a capacity solution of* (1.1). *If* $|u| \leqq M$ *almost everywhere on* Q_T *for some* $M > 0$, *then* $\nabla\varphi \in L^2(0, T; W^{1,2}(\Omega))$, $g = \sigma(u)\nabla\varphi$, *and* (u, φ, v) *is a classical weak solution of* (1.1).

Remark. A classical weak solution of (1.1) can be defined in an obvious way; see, e.g., [SSX]. The proof of this lemma is similar to that given in [X2]. Thus the notion of a capacity solution is a suitable generalization of that of a classical weak solution. If (u, φ, v) is a classical weak solution, then we deduce from (1.3) that for any $\eta \in L^2(0, T; W^{1,2}(\Omega)) \cap L^\infty(Q_T)$,

$$\begin{aligned}\int_{Q_T} \sigma(u)|\nabla\varphi|^2\eta \, dx\, dt &= \int_{Q_T} \sigma(u)\nabla\varphi(\nabla(\varphi - \bar{\varphi}_0)\eta + \nabla\bar{\varphi}_0\eta)\, dx\, dt \\ &= -\int_{Q_T} \sigma(u)\nabla\varphi(\varphi - \bar{\varphi}_0)\nabla\eta\, dx\, dt + \int_{Q_T} \sigma(u)\nabla\varphi\nabla\bar{\varphi}_0\eta\, dx\, dt.\end{aligned} \tag{1.7}$$

Now let (u, φ, v, g) be a capacity solution of (1.1). For each positive integer k, let $E_k = \{(x, t) \in Q_T: |u(x, t)| \leqq k\}$. By the proof of Lemma 1.1 in [X2], we may select a θ from $\mathscr{A}$ and an m so that

$$F_\theta^{(m)}\varphi = \varphi \quad \text{a.e. on } E_k. \tag{1.8}$$

Since $F_\theta^{(m)}\varphi \in L^2(0, T; W^{1,2}(\Omega))$, we may calculate $\nabla(F_\theta^{(m)}(x, t)\varphi(x, t))$ for almost every $(x, t) \in Q_T$ in the sense of Lemma A.2 in [KS, p. 50]. In view of (1.8), it is natural for us to define

$$\nabla\varphi(x, t) = \nabla(F_\theta^{(m)}(x, t)\varphi(x, t)) \quad \text{for } (x, t) \in E_k. \tag{1.9}$$

Since $Q_T \backslash \bigcup_{k=1}^\infty E_k$ is of measure zero, we may evaluate $\nabla\varphi(x, t)$ through (1.9) for almost every $(x, t) \in Q_T$. Thus we say that $\nabla\varphi(x, t)$ exists almost everywhere on Q_T. Clearly, $\nabla\varphi$ so obtained is a measurable function, and it belongs to $[L^2(E_k)]^N$ for each k. Also, we easily obtain from (1.6) that

$$g = \sigma(u)\nabla\varphi \quad \text{a.e. on } Q_T.$$

We may conclude that the difference between a capacity solution and a classical weak solution is that in the classical weak solution the gradient of φ is evaluated in the sense of distributions, while in the capacity solution the same gradient is calculated in the almost everywhere sense. However, in a capacity solution, $\nabla\varphi$ in the sense of distributions may not belong to any L^p-space with $p \geqq 1$. Thus (1.7) only holds for those $\eta \in L^2(0, T; W^{1,2}(\Omega)) \cap L^\infty(Q_T)$ such that $(\varphi - \bar{\varphi}_0)\eta \in L^2(0, T; W_0^{1,2}(\Omega))$. It is easy to see that if $\eta \in L^2(0, T; W^{1,2}(\Omega))$ is such that $\eta = 0$ almost everywhere on $Q_T \backslash E_k$ for some k, then $\eta\varphi \in L^2(0, T; W^{1,2}(\Omega))$. Also, in a capacity solution, the trace $\varphi|_{S_T}$ may not make sense, and thus (1.5) is used to describe the boundary condition for φ. Moreover, the product $\sigma(u)\nabla\varphi$ is taken as a product of two functions in Q_T. Thus through the use of (1.4)-(1.6) we have successfully avoided all possible ambiguity in our problem caused by the degeneracy and at the same time probably retained the best possible regularity information on the solution under the circumstances.

The main result of this paper is that under (A1)-(A5) there is a capacity solution to (1.1).

A capacity solution to (1.1) is constructed as a limit of a sequence of classical weak solutions of the regularized problems. In § 2 we consider the regularized stationary problem, while § 3 is devoted to the study of the regularized time-dependent problem. The results in §§ 2 and 3 are used in § 4 to prove our main existence theorem.

The letter c will be used to denote the generic constant. When distinction among different constants is needed, we add a subscript i to c with $i \in \{0, 1, 2, \cdots\}$.

2. The stationary problem. In this section we consider the following stationary problem:

$$\alpha(u) - \operatorname{div} a(\nabla u) = \sigma(u)|\nabla\varphi|^2 + H(x) \quad \text{in } \Omega, \tag{2.1a}$$

$$\operatorname{div}(\sigma(u)\nabla\varphi) = 0 \quad \text{in } \Omega, \tag{2.1b}$$

$$a(\nabla u) \cdot \nu + f(x, u) = 0 \quad \text{in } \partial\Omega, \tag{2.1c}$$

$$\varphi = \varphi_0 \quad \text{in } \partial\Omega. \tag{2.1d}$$

With respect to the data involved we assume the following.

(B1) α is Lipschitz continuous and strongly coercive monotone, i.e.,

$$C_1|s_1 - s_2|^2 \geqq (\alpha(s_1) - \alpha(s_2))(s_1 - s_2) \geqq C_0|s_1 - s_2|^2 \quad \text{for all } s_1, s_2 \in \mathbb{R},$$

where C_0, C_1 are two positive constants.

(B2) $\sigma(s)$ is continuous and satisfies

$$C_2 \leqq \sigma(s) \leqq C_3 \quad \text{for all } s \in \mathbb{R},$$

where C_2, C_3 are two positive constants with $C_2 \leqq C_3$.

(B3) a is given as in § 1. The function $f(x, \xi)$ satisfies (A2) with (x, t) replaced by x, while the function $H(x)$ belongs to $L^2(\Omega)$. Also, $\varphi_0 \in L^2(\partial\Omega)$, and there is a function $\overline{\varphi_0}$ in $W^{1,\infty}(\Omega)$ such that $\overline{\varphi_0} = \varphi_0$ in $L^2(\partial\Omega)$.

A weak solution to (2.1) is defined as a pair (u, φ) such that

$$u, \varphi \in W^{1,2}(\Omega), \tag{2.2}$$

$$\varphi = \varphi_0 \quad \text{in } L^2(\partial\Omega), \tag{2.3}$$

$$\begin{aligned} &\int_\Omega (\alpha(u)\psi + a(\nabla u)\nabla\psi)\,dx + \int_{\partial\Omega} f(x, u)\psi\,ds \\ &= \int_\Omega (\sigma(u)|\nabla\varphi|^2 + H(x))\psi(x)\,dx \end{aligned} \tag{2.4}$$

for every $\psi(x) \in W^{1,2}(\Omega) \cap L^\infty(\Omega)$, and

$$\int \sigma(u)\nabla\varphi\nabla\eta(x)\,dx = 0 \quad \text{for every } \eta \in W_0^{1,2}(\Omega). \tag{2.5}$$

We immediately have the following.

LEMMA 2.1. *If* (u, φ) *is a weak solution of* (2.1), *then* (2.4) *is equivalent to the following*:

$$\begin{aligned} &\int_\Omega (\alpha(u)\psi + a(\nabla u)\nabla\psi)\,dx + \int_{\partial\Omega} f(x, u)\psi\,ds \\ &= -\int_\Omega \sigma(u)\nabla\varphi\nabla\psi(\varphi - \overline{\varphi_0})\,dx + \int_\Omega \sigma(u)\nabla\varphi\nabla\overline{\varphi_0}\psi\,dx + \int_\Omega H(x)\psi\,dx \end{aligned} \tag{2.6}$$

for every $\psi \in W^{1,2}(\Omega)$.

The proof is similar to that of [HRS, Lemma 1], and we shall omit it here.

LEMMA 2.2. *Under the assumptions* (B1)-(B3) *there is a weak solution to* (2.1).

Proof. We follow the approach adopted in [X3]. For each positive integer k define

$$P_k(x)=\begin{cases} k & \text{if } |x|^2>k, \\ |x|^2 & \text{if } |x|^2\leqq k. \end{cases} \tag{2.7}$$

By an argument similar to that used in the proof of Lemma 2.1 in [X3], we obtain that for each k there exists at least one vector-valued function $u_k=(u_1^{(k)}, u_2^{(k)})\in W^{1,2}(\Omega)\times\{v\in W^{1,2}(\Omega)\colon v|_{\partial\Omega}=\varphi_0\}$ such that

$$\begin{aligned}\int_\Omega (\alpha(u_1^{(k)})\psi(x)+a(\nabla u_1^{(k)})\nabla\psi)\,dx+\int_{\partial\Omega} f(x, u_1^{(k)})\psi(x)\,ds \\ =\int_\Omega (\sigma(u_1^{(k)})P_k(\nabla u_2^{(k)})+H(x))\psi\,dx\end{aligned} \tag{2.8}$$

for all $\varphi\in W^{1,2}(\Omega)$, and

$$\int_\Omega \sigma(u_1^{(k)})\nabla u_2^{(k)}\nabla\eta(x)\,dx=0 \quad \text{for all } \eta\in W_0^{1,2}(\Omega). \tag{2.9}$$

We infer from (2.9) that

$$\sup_\Omega |u_2^{(k)}|\leqq c, \tag{2.10}$$

$$\int_\Omega |\nabla u_2^{(k)}|^2\,dx\leqq c \qquad (k=1,2,\cdots). \tag{2.11}$$

We infer from the proof of Lemma 2.1 in [X3] that we can set $\psi=(u_1^{(k)})^-$, $(u_1^{(k)})^+$ successively in (2.8) to obtain

$$\|u_1^{(k)}\|_{W^{1,2}(\Omega)}\leqq c \qquad (k=1,2,\cdots). \tag{2.12}$$

In view of (2.10), (2.11), and (2.12), we may select a subsequence of $\{k\}$, still denoted by $\{k\}$, so that

$$u_1^{(k)}\to u_1 \quad \text{strongly in } L^2(\Omega) \quad \text{and weakly in } W^{1,2}(\Omega), \tag{2.13}$$

$$u_2^{(k)}\to u_2 \quad \text{strongly in } L^2(\Omega) \quad \text{and weakly in } W^{1,2}(\Omega). \tag{2.14}$$

Setting $\eta=u_2^{(k)}-u_2$ in (2.9) and using (2.14) in the resulting equation yields

$$u_2^{(k)}\to u_2 \quad \text{strongly in } W^{1,2}(\Omega). \tag{2.15}$$

It follows from (2.13) and a result in [M, p.76] that

$$u_1^{(k)}\to u_1 \quad \text{strongly in } L^2(\partial\Omega).$$

We still need to show that

$$\nabla u_1^{(k)}\to\nabla u_1 \quad \text{a.e. on } \Omega. \tag{2.16}$$

For this purpose, fix $\varepsilon, \delta>0$ and define

$$\beta_\varepsilon(s)=\begin{cases} \varepsilon & \text{if } s\geqq\varepsilon, \\ s & \text{if } -\varepsilon<s<\varepsilon, \\ -\varepsilon & \text{if } s\leqq-\varepsilon. \end{cases}$$

According to Egoroff's theorem, there exists a measurable set $E_\delta \subset \Omega$ such that $L^N(\Omega\backslash E_\delta) \leqq \delta$ and $u_1^{(k)} \to u_1$ uniformly on E_δ. Thus if we choose k so large that $|u_1^{(k)} - u_1| \leqq \varepsilon/2$ on E_δ, we may calculate using (2.8) and [B3] that

$$\int_{E_\delta} |\nabla u_1^{(k)} - \nabla u_1|^2 \, dx \leqq c_0 \int_\Omega (a(\nabla u_1^{(k)}) - a(\nabla u_1)) \nabla \beta_\varepsilon(u_1^{(k)} - u_1) \, dx$$

$$= c_0 \bigg(\int_\Omega (\sigma(u_1^{(k)}) P_k(\nabla u_2^{(k)}) + H(x)) \beta_\varepsilon(u_1^{(k)} - u_1) \, dx$$

$$- \int_{\partial\Omega} f(x, u_1^{(k)}) \beta_\varepsilon(u_1^{(k)} - u_1) \, ds - \int_\Omega \alpha(u_1^{(k)}) \beta_\varepsilon(u_1^{(k)} - u_1) \, dx$$

$$- \int_\Omega a(\nabla u_1) \nabla \beta_\varepsilon(u_1^{(k)} - u_1) \, dx \bigg).$$

Now $\beta_\varepsilon(u_1^{(k)} - u_1)$ goes to zero strongly in $L^2(\Omega)$ and weakly in $W^{1,2}(\Omega)$. Also, $P_k(\nabla u_2^{(k)})$ converges strongly to $|\nabla u_2|^2$ in $L^1(\Omega)$ due to (2.15). Consequently,

$$\limsup_{k\to\infty} \int_{E_\delta} |\nabla u_1^{(k)} - \nabla u_1|^2 \, dx \leqq 0.$$

Hence, passing if need be to a further subsequence, we deduce $\nabla u_1^{(k)} \to \nabla u_1$ almost everywhere on E_δ. This is true for each $\delta > 0$, and so (2.16) holds. By (B3), we conclude

$$a(\nabla u_1^{(k)}) \to a(\nabla u_1) \quad \text{weakly in } [L^2(\Omega)]^N.$$

Now we can pass to the limit in (2.8) and (2.9). This completes the proof of the lemma.

Remark. The proof peresented here can easily be modified to cover the case considered in [HRS].

3. The time-dependent problem. In this section we assume that (B1), (B2), (A1), (A2), and (A5) hold. Consider the following problem:

$$\frac{\partial}{\partial t} \alpha(u) - \operatorname{div} a(\nabla u) = \sigma(u) |\nabla \varphi|^2 \quad \text{in } Q_T, \tag{3.1a}$$

$$\operatorname{div}(\sigma(u) \nabla \varphi) = 0 \quad \text{in } Q_T, \tag{3.1b}$$

$$a(\nabla u) \cdot \nu + f(x, t, u) = 0 \quad \text{in } S_T, \tag{3.1c}$$

$$\varphi = \varphi_0 \quad \text{in } S_T, \tag{3.1d}$$

$$\alpha(u) = v_0 \quad \text{in } \Omega \times \{0\}. \tag{3.1e}$$

By a weak solution to (3.1) we mean a pair (u, φ) such that

$$\begin{gathered} u, \varphi \in L^2(0, T; W^{1,2}(\Omega)), \\ \varphi = \varphi_0 \quad \text{in } L^2(0, T; L^2(\partial\Omega)), \end{gathered} \tag{3.2}$$

$$\begin{aligned} &- \int_{Q_T} \alpha(u) \psi_t \, dx \, dt + \int_{Q_T} a(\nabla u) \nabla \psi \, dx \, dt + \int_0^T \int_{\partial\Omega} f(x, t, u) \psi \, ds \, dt \\ &= \int_\Omega v_0 \psi(x, 0) \, dx + \int_{Q_T} \sigma(u) |\nabla \varphi|^2 \psi \, dx \, dt \end{aligned} \tag{3.3}$$

for all $\psi \in H^1(0, T; W^{1,2}(\Omega)) \cap L^\infty(Q_T)$ such that $\psi(x, T) \equiv 0$, and

$$\int_{Q_T} \sigma(u)\nabla\varphi\nabla\eta \, dx \, dt = 0 \tag{3.4}$$

for all $\eta \in L^2(0, T; W_0^{1,2}(\Omega))$.

Let $U = W^{1,2}(\Omega)$ and U^* be the topological dual of U. Define an operator $A: L^2(0, T; U) \to L^2(0, T; U^*)$ by

$$(A(u), v) = \int_{Q_T} a(\nabla u)\nabla v \, dx \, dt + \int_0^T \int_{\partial\Omega} f(x, t, u)v \, ds \, dt, \qquad u, v \in L^2(0, T; U).$$

In view of (A1) and (A2), A is well defined. Let $B(u, v)$ be the operator from $L^2(0, T; U) \times L^2(0, T; U) \cap L^\infty(Q_T)$ to $L^2(0, T; U^*)$ defined by

$$(B(u, v), w) = -\int_{Q_T} \sigma(u)\nabla v(v - \overline{\varphi_0})\nabla w \, dx \, dt + \int_{Q_T} \sigma(u)\nabla v \nabla\overline{\varphi_0} w \, dx \, dt,$$

$$u, w \in L^2(0, T; U), \qquad v \in L^2(0, T; U) \cap L^\infty(Q_T).$$

By [B2] and [A5], $B(u, v)$ is also well defined.

LEMMA 3.1. *If the pair* (u, φ) *is a weak solution of* (3.1), *then* (3.3) *is equivalent to the following*:

$$\frac{\partial}{\partial t}\alpha(u) \in L^2(0, T; U^*), \tag{3.5}$$

$$\frac{\partial}{\partial t}\alpha(u) + A(u) = B(u, \varphi) \quad \text{in } L^2(0, T; U^*). \tag{3.6}$$

For $\psi \in L^2(0, T; U) \cap L^\infty(Q_T)$ we calculate, with the aid of (3.4), that

$$\begin{aligned} \int_{Q_T} \sigma(u)|\nabla\varphi|^2\psi \, dx \, dt &= \int_{Q_T} \sigma(u)\nabla\varphi(\nabla(\varphi - \overline{\varphi_0})\psi + \nabla\overline{\varphi_0}\psi) \, dx \, dt \\ &= -\int_{Q_T} \sigma(u)\nabla\varphi(\varphi - \overline{\varphi_0})\nabla\psi \, dx \, dt \\ &\quad + \int_{Q_T} \sigma(u)\nabla\varphi\nabla\overline{\varphi_0}\psi \, dx \, dt. \end{aligned} \tag{3.7}$$

The lemma can easily be inferrred from using (3.7) in (3.3).

LEMMA 3.2. *Let* (B1), (B2), (A1), *and* (A5) *be satisfied. Assume that* $\alpha^{-1}(v_0) \in W^{1,2}(\Omega)$. *Then there is a weak solution to* (3.1).

Proof. Let k be a positive integer and $\delta = T/k$. Set

$$u_0^{(k)} = \alpha^{-1}(v_0).$$

By Lemma 2.2, we may define a set of k pairs $(u_1^{(k)}, \varphi_1^{(k)}), \cdots, (u_k^{(k)}, \varphi_k^{(k)})$ via the following iteration formula:

$$\frac{\alpha(u_j^{(k)}) - \alpha(u_{j-1}^{(k)})}{\delta} - \operatorname{div} a(\nabla u_j^{(k)}) = \sigma(u_j^{(k)})|\nabla\varphi_j^{(k)}|^2 \quad \text{in } \Omega, \tag{3.8}$$

$$\operatorname{div}(\sigma(u_j^{(k)})\nabla\varphi_j^{(k)}) = 0 \quad \text{in } \Omega, \tag{3.9}$$

$$a(\nabla u_j^{(k)})\nu + f(x, j\delta, u_j^{(k)}) = 0 \quad \text{in } \partial\Omega, \tag{3.10}$$

$$\varphi_j^{(k)} = \frac{1}{\delta}\int_{(j-1)\delta}^{j\delta} \varphi_0(x, t) \, dt \quad \text{on } \partial\Omega, \quad j = 1, 2, \cdots, k. \tag{3.11}$$

Subsequently, set

$$u^{(k)}(x,t)=\begin{cases} u_0^{(k)} & \text{if } t\leqq 0 \\ u_j^{(k)}(x) & \text{if } (j-1)\delta<t\leqq j\delta \end{cases} \quad (j=1,\cdots,k),$$

$$\varphi^{(k)}(x,t)=\begin{cases} \varphi_1^{(k)}(x) & \text{if } t\leqq \delta \\ \varphi_j^{(k)}(x) & \text{if } (j-1)\delta<\leqq j\delta \end{cases} \quad (j=2,\cdots,k).$$

Invoking the weak formulation of problem (3.8)-(3.11), we obtain

$$\text{(3.12)} \qquad \int_{Q_T} \sigma(u^{(k)})\nabla\varphi^{(k)}\nabla\xi\,dx\,dt=0$$

for all $\xi\in L^2(0,T;\, W_0^{1,2}(\Omega))$. We infer from the weak maximum principle that

$$\text{(3.13)} \qquad \sup_{Q_T}|\varphi^{(k)}|\leqq c \qquad (k=1,2,\cdots).$$

Set

$$\overline{\varphi_{0k}}(x,t)=\frac{1}{\delta}\int_{(j-1)\delta}^{j\delta}\overline{\varphi_0}(x,\tau)\,d\tau \quad \text{if } (j-1)\delta<t\leqq j\delta \qquad (j=1,2,\cdots,k).$$

Let $\xi=\varphi^{(k)}-\overline{\varphi_{0k}}$ in (3.12) to obtain

$$\text{(3.14)} \qquad \int_{Q_T}|\nabla\varphi^{(k)}|^2\,dx\,dt\leqq c \qquad (k=1,2,\cdots).$$

Now set

$$J(s)=\int_0^s \alpha^{-1}(\tau)\,d\tau.$$

Without loss of generality, we assume $\alpha(0)=0$. Then the functional $G: L^2(\Omega)\to[0,\infty]$ defined by

$$G(f)=\begin{cases} \int_\Omega J(f)\,dx & \text{if } J(f)\in L^1(\Omega), \\ +\infty & \text{otherwise} \end{cases}$$

is lower semicontinuous on $L^2(\Omega)$.

Now we wish to show that

$$\text{(3.15)} \qquad \sup_{0\leqq t\leqq T}\int_\Omega |u^{(k)}(x,t)|^2\,dx+\int_{Q_T}|\nabla u^{(k)}|^2\,dx\,dt\leqq c \qquad (k=1,2,\cdots).$$

This estimate can easily be inferred from § 4 of [SSX]. We only point out that in our situation there exist two positive numbers c_1, c_2 such that

$$|\alpha^{-1}(s)|\geqq c_1|s|, \qquad |\alpha(s)|\geqq c_2|s| \quad \text{for all } s\in\mathbb{R}.$$

Subsequently,

$$\begin{aligned} G(\alpha(u^{(k)}(x,t)))&=\int_\Omega\int_0^{\alpha(u^{(k)}(x,\tau))}\alpha^{-1}(s)\,ds\,dx \\ &=\int_\Omega\left|\int_0^{\alpha(u^{(k)}(x,t))}\alpha^{-1}(s)\,ds\right|dx\geqq c\int_\Omega|u^{(k)}(x,t)|^2\,dx. \end{aligned}$$

Now define $v^{(k)}$ by

$$v^{(k)}(x,t)=\frac{t-(j-1)\delta}{\delta}\,\alpha(u_j^{(k)})+\frac{j\delta-t}{\delta}\,\alpha(u_{j-1}^{(k)})$$

if $(j-1)\delta < t \leqq j\delta, j=1,\cdots,k$ for $k=1,2,\cdots$. We may infer from Lemma 3.1, (3.8), and (3.9) that

$$\begin{aligned}&\left(\frac{\partial}{\partial t}v^{(k)},\xi\right)+\int_{Q_T}a(\nabla u^{(k)})\nabla\xi\,dx\,dt+\int_0^T\int_{\partial\Omega}f(x,t,u^{(k)})\xi\,ds\,dt\\ &\quad=-\int_{Q_T}\sigma(u^{(k)})\nabla\varphi^{(k)}(\phi^{(k)}-\overline{\varphi_{0k}})\nabla\xi\,dx\,dt\\ &\qquad+\int_{Q_T}\sigma(u^{(k)})\nabla\varphi^{(k)}\nabla\overline{\varphi_{0k}}\xi\,dx\,dt\quad\text{for all }\xi\in L^2(0,T;W^{1,2}(\Omega)),\end{aligned}\tag{3.16}$$

where $(\cdot,\cdot)$ denotes the duality pairing between $L^2(0,T;U^*)$ and $L^2(0,T;W^{1,2}(\Omega))$. Since α is Lipschitz continuous, we derive from (3.15) that

$$\sup_{0\leqq t\leqq T}\int_\Omega|v^{(k)}(x,t)|^2\,dx+\int_0^T\int_\Omega|\nabla v^{(k)}|^2\,dx\,dt\leqq c,\qquad(k=1,\cdots),$$

where c may depend on $\|\nabla\alpha^{-1}(v_0)\|_{L^2(\Omega)}$. This, together with (3.16), (3.13), and (3.14) implies that

$$\left\|\frac{\partial}{\partial t}v^{(k)}\right\|_{L^2(0,T;U^*)}\leqq c\qquad(k=1,\cdots).$$

By Lions-Aubin's theorem, $\{v^{(k)}\}$ is precompact in $L^2(Q_T)$. We estimate

$$\begin{aligned}\int_{Q_T}(v^{(k)}-\alpha(u^{(k)}))^2\,dx\,dt&=\frac{1}{3}\int_{Q_T}(\alpha(u^{(k)}(x,t))-\alpha(u^{(k)}(x,t-\delta)))^2\,dx\,dt\\ &\leqq c\int_{Q_T}(u^{(k)}(x,t)-u^{(k)}(x,t-\delta))^2\,dx\,dt.\end{aligned}\tag{3.17}$$

Using $u_j^{(k)}-u_{j-1}^{(k)}$ as a test function in (3.8), after a sequence of calculations we get

$$c_0\int_0^T\int_\Omega(u^{(k)}(x,t)-u^{(k)}(x,t-\delta))^2\,dx\,dt\leqq c_1\delta^{1/2}\to0\quad\text{as }k\to\infty.$$

This, together with (3.17) implies that $\{\alpha(u^{(k)})\}$ is precompact in $L^2(Q_T)$. Consequently, $\{u^{(k)}\}$ is also precompact in $L^2(0,T;L^2(\Omega))$ due to (B1). Now we can select a subsequence of $\{k\}$, still denoted by $\{k\}$, such that

$$\begin{aligned}u^{(k)}\to u\text{ strongly }&\text{ in }L^2(0,T;L^2(\Omega))\quad\text{and}\\ &\text{weakly in }L^2(0,T;W^{1,2}(\Omega)),\end{aligned}\tag{3.18}$$

$$a(\nabla u^{(k)})\to L(x,t)\quad\text{weakly in }[L^2(Q_T)]^N,\tag{3.19}$$

$$\varphi^{(k)}\to\varphi\quad\text{weakly in }L^2(0,T;W^{1,2}(\Omega)).\tag{3.20}$$

By (3.18) and the trace theorem [CDK], we obtain

$$u^{(k)}\to u\quad\text{strongly in }L^2(0,T;L^2(\partial\Omega)).\tag{3.21}$$

From the definition of $\{\overline{\varphi_{0k}}\}$, we see that

$$\varphi^{(k)}-\overline{\varphi_{0k}}\in L^2(0,T;\ W_0^{1,2}(\Omega))\quad\text{for each } k,$$
$$\overline{\varphi_{0k}}\to\overline{\varphi_0}\quad\text{strongly in } L^2(0,T;\ W^{1,2}(\Omega)).$$

Set $\xi=\varphi^{(k)}-\overline{\varphi_{0k}}-\varphi+\overline{\varphi_0}$ in (3.12) to get

$$\int_{Q_T}\sigma(u^{(k)})\nabla\varphi^{(k)}\nabla(\varphi^{(k)}-\varphi)\,dx\,dt=\int_{Q_T}\sigma(u^{(k)})\nabla\varphi^{(k)}\nabla(\overline{\varphi_{0k}}-\overline{\varphi_0})\,dx\,dt$$

from whence follows

$$\lim_{k\to\infty}\int_{Q_T}|\nabla(\varphi^{(k)}-\varphi)|^2\,dx\,dt=0. \tag{3.22}$$

Note that $(\partial/\partial t)v^{(k)}\to(\partial/\partial t)\alpha(u)$ weakly in $L^2(0,T;\ U^*)$ because of (3.17). We can take $k\to\infty$ in (3.16) to obtain

$$\begin{aligned}&\frac{\partial}{\partial t}\alpha(u)\in L^2(0,T;\ U^*),\\ &\frac{\partial}{\partial t}\alpha(u)+\tilde{A}(u)=B(u,\varphi)\quad\text{in } L^2(0,T;\ U^*),\end{aligned} \tag{3.23}$$

where $B(u,\varphi)$ is given as in Lemma 3.1, and $\tilde{A}(u)\in L^2(0,T;\ U^*)$ is given by

$$(\tilde{A}(u),v)=\int_{Q_T}L\nabla v\,dx\,dt+\int_0^T\int_{\partial\Omega}f(x,t,u)v\,dS\,dt,$$
$$v\in L^2(0,T;\ W^{1,2}(\Omega)).$$

Before we continue, let us cite the following lemma.

LEMMA 3.3. *Let α be a Lipschitz continuous and strongly coercive monotone function on $\mathbb{R}$ and β a Lipschitz continuous function on $\mathbb{R}$ with $\alpha(0)=\beta(0)=0$. Assume that $u\in L^2(0,T;\ W^{1,2}(\Omega))$ and that $(\partial/\partial t)\alpha(u)\in L^2(0,T;\ U^*)$. Then the function $t\to\int_\Omega\int_0^{\alpha(u(x,t))}\beta(\alpha^{-1}(s))\,ds\,dx$ is absolutely continuous on $(0,T)$, and*

$$\frac{d}{dt}\int_\Omega\int_0^{\alpha(u(x,t))}\beta(\alpha^{-1}(s))\,ds\,dx=\left(\frac{\partial}{\partial t}\alpha(u),\beta(u)\right)\qquad\textit{for a.e. } t\in(0,T),$$

where $(\cdot,\cdot)$ denotes the duality pairing between U^* and $W^{1,2}(\Omega)$.

Remark. This lemma is a consequence of Theorem 17 in [BR1], the chain rule, and the fact that $(v,u)=\int_\Omega vu\,dx$ for $u\in W^{1,2}(\Omega)$, $v\in L^\infty(\Omega)\cap U^*$. It is worth pointing out that if v only belongs to $L^1(\Omega)\cap U^*$ the question of whether $(v,u)=\int_\Omega vu\,dx$ is first, raised by Brézis in [BR2], remains open.

Return to the proof of Lemma 3.2. We derive from (3.23) and Lemma 3.3 that

$$\begin{aligned}&\int_\Omega\int_0^{\alpha(u(x,T))}\alpha^{-1}(s)\,ds\,dx-\int_\Omega\int_0^{v_0(x)}\alpha^{-1}(s)\,ds\,dx\\ &\quad+\int_{Q_T}L\nabla u\,dx\,dt+\int_0^T\int_{\partial\Omega}f(x,t,u)u\,dS\,dt\\ &=-\int_{Q_T}\sigma(u)\nabla\varphi(\varphi-\overline{\varphi_0})\nabla u\,dx\,dt+\int_{Q_T}\sigma(u)\nabla\varphi\nabla\overline{\varphi_0}u\,dx\,dt.\end{aligned} \tag{3.24}$$

Here we would like to point out that $\alpha(u)\in C([0,T];L^2(\Omega))$, and thus the first term in (3.24) makes sense. We have

$$
\begin{aligned}
&G(\alpha(u^{(k)}(x,T)))-G(v_0)+\int_{Q_T} a(\nabla u^{(k)})\nabla u^{(k)}\,dx\,dt\\
&\quad+\int_0^T\int_{\partial\Omega} f_k(x,t,u^{(k)})u^{(k)}\,ds\,dt\\
&\quad=-\int_{Q_T}\sigma(u^{(k)})\nabla\varphi^{(k)}(\varphi^{(k)}-\overline{\varphi_0})\nabla u^{(k)}\,dx\,dt\\
&\quad\quad+\int_{Q_T}\sigma(u^{(k)})\nabla\varphi^{(k)}\nabla\overline{\varphi_0}u^{(k)}\,dx\,dt.
\end{aligned}
\tag{3.25}
$$

Remember that $v^{(k)}(x,t)\rightarrow\alpha(u(x,t))$ strongly in $L^2(\Omega)$ for almost every $t\in(0,T]$. Also, $\{(\partial/\partial t)v^{(k)}\}$ is bounded in $L^2(0,T;U^*)$. We easily see that $\alpha(u),v^{(k)}\in C([0,T];L^2(\Omega))$ for each k and

$$
v^{(k)}(x,t)\rightarrow\alpha(u(x,t))\quad\text{weakly in } L^2(\Omega)\quad\text{for each } t \text{ in } [0,T].
$$

In particular, we have

$$
v^{(k)}(x,T)=\alpha(u^{(k)}(x,T))\rightarrow\alpha(u(x,T))\quad\text{weakly in } L^2(\Omega).
$$

Consequently,

$$
\liminf_{k\rightarrow\infty} G(\alpha(u^{(k)}(x,T)))\geqq G(\alpha(u(x,T))).
$$

We easily deduce from (3.22) that

$$
\varphi^{(k)}\rightarrow\varphi\quad\text{strongly in } L^p(Q_T)\quad\text{for each } p\geqq 1.
$$

Thus letting $k\rightarrow\infty$ in (3.25) yields

$$
\begin{aligned}
\limsup_{k\rightarrow\infty}\int_{Q_T} a(\nabla u^{(k)})\nabla u^{(k)}\,dx\,dt&\leqq G(v_0)-G(\alpha(u(x,T)))-\int_0^T\int_{\partial\Omega} f(x,t,u)u\,dS\,dt\\
&\quad-\int_{Q_T}\sigma(u)\nabla\varphi(\varphi-\overline{\varphi_0})\nabla u\,dx\,dt\\
&\quad+\int_{Q_T}\sigma(u)\nabla\varphi\nabla\overline{\varphi_0}u\,dx\,dt\\
&=\int_{Q_T} L\nabla u\,dx\,dt.
\end{aligned}
$$

The last step is due to (3.24).

We estimate from (A1) that

$$
\begin{aligned}
\int_{Q_T}|\nabla u^{(k)}-\nabla u|^2\,dx\,dt&\leqq c\int_{Q_T}(a(\nabla u^{(k)})-a(\nabla u))(\nabla u^{(k)}-\nabla u)\,dx\,dt\\
&=c\left[\int_{Q_T} a(\nabla u^{(k)})\nabla u^{(k)}\,dx\,dt-\int_{Q_T} a(\nabla u^{(k)})\nabla u\,dx\,dt\right.\\
&\quad\left.-\int_{Q_T} a(\nabla u)(\nabla u^{(k)}-\nabla u)\,dx\,dt\right].
\end{aligned}
$$

This implies that

$$\lim_{k\to\infty}\int_{Q_T}|\nabla u^{(k)}-\nabla u|^2\,dx\,dt=0.$$

We immediately obtain, with the aid of (A1) and (3.19), that

$$L=a(\nabla u)\quad \text{a.e. on } Q_T.$$

This completes the proof of Lemma 3.2.

Remark. The arguments presented here can be used to simplify the proof in [SSX].

4. Proof of the main result. Let $J(s)$ be a nonnegative function on $\mathbb{R}$ belonging to $C_0^\infty(\mathbb{R})$ and having the following properties:

(i) $J(s)=0$ if $|s|\geqq 1$; and

(ii) $\int_{\mathbb{R}} J(s)\,ds=1$.

We derive from (A4) that $\alpha^{-1}(s)$ is uniformly continuous on $\mathbb{R}$. Thus the sequence $\{\tilde{\alpha}_k(s)\}$ defined by

$$\tilde{\alpha}_k(s)=\frac{1}{k}\int_{\mathbb{R}} J(k(s-\tau))\alpha^{-1}(\tau)\,d\tau\qquad (k=1,2,\cdots)$$

converges to $\alpha^{-1}(s)$ uniformly on $\mathbb{R}$ as $k\to\infty$. For each k, denote by $\alpha_k(s)$ the inverse of $\tilde{\alpha}_k(s)+(1/k)s-\tilde{\alpha}_k(0)$. Then α_k is Lipschitz continuous and strongly coercive monotone for each fixed k. Now fix a k, and consider the following problem:

$$\frac{\partial}{\partial t}\alpha_k(u)-\operatorname{div} a(\nabla u)=\sigma_k(u)|\nabla\varphi|^2\quad\text{in } Q_T, \tag{4.1}$$

$$\operatorname{div}(\sigma_k(u)\nabla\varphi)=0\quad\text{in } Q_T, \tag{4.2}$$

$$\begin{aligned} a(\nabla u)\cdot\nu+f(x,t,u)&=0\quad\text{in } S_T,\\ \varphi&=\varphi_0\quad\text{in } S_T, \qquad (4.3)\\ \alpha_k(u)&=v_0\quad\text{in } \Omega\times\{0\}, \end{aligned}$$

where $\sigma_k(s)=\sigma(s)+(1/k)$. Without loss of generality, assume $\alpha_k^{-1}(v_0)\in W^{1,2}(\Omega)$ for each k. By Lemma 3.2, for each k there is a pair (u_k,φ_k) such that

$$\varphi_k\in L^\infty(Q_T)\cap L^2(0,T;W^{1,2}(\Omega)),\qquad u_k\in L^2(0,T;W^{1,2}(\Omega)), \tag{4.4}$$

$$\frac{\partial}{\partial t}v_k+A(u_k)=B(u_k,\varphi_k)\quad\text{in } L^2(0,T;U^*), \tag{4.5}$$

$$\int_{Q_T}\sigma_k(u_k)\nabla\varphi_k\nabla\psi\,dx\,dt=0\quad\text{for all } \psi\in L^2(0,T;W_0^{1,2}(\Omega)), \tag{4.6}$$

$$\varphi_k=\varphi_0\quad\text{in } L^2(0,T;L^2(\partial\Omega)), \tag{4.7}$$

$$\alpha_k(u_k)=v_0\quad\text{in } C([0,T);L^2(\Omega)), \tag{4.8}$$

where $v_k=\alpha_k(u_k)$.

Now we proceed to derive a priori estimates on the sequence $\{(u_k,\varphi_k)\}$. Set $\psi=\varphi_k-\overline{\varphi_0}$ in (4.6) to obtain

$$\int_{Q_T}\sigma_k(u_k)|\nabla\varphi_k|^2\,dx\,dt\leqq c\qquad (k=1,2,\cdots). \tag{4.9}$$

By the weak maximum principle,

$$\sup_{Q_T} |\varphi_k| \leqq c \qquad (k=1,2,\cdots). \tag{4.10}$$

We wish to show that

$$\sup_{0\leqq t\leqq T} \int_\Omega u_k^2(x,t)\,dx + \int_{Q_T} |\nabla u_k|^2\,dx\,dt \leqq c \qquad (k=1,2,\cdots). \tag{4.11}$$

For this purpose, fix a positive number M. For each k let s_k be the solution of the following equation:

$$\alpha_k(s) = M. \tag{4.12}$$

By [A4], there is a positive number M_0 so that

$$\liminf_{k\to\infty} s_k > M_0. \tag{4.13}$$

Use $(v_k - M)^+$ as a test function in (4.5) to get

$$\begin{aligned} &\frac{1}{2}\int_\Omega [(v_k(x,T)-M)^+]^2\,dx - \frac{1}{2}\int_\Omega [(v_0-M)^+]^2\,dx \\ &\quad + \int_{(u_k\geqq s_k)} a(\nabla u_k)\alpha_k'(u_k)\nabla u_k\,dx\,dt + \int_0^T \int_{\partial\Omega} f(x,t,u_k)(v_k-M)^+\,ds\,dt \\ &= -\int_{\{u_k\geqq s_k\}} \sigma_k(u_k)\nabla\varphi_k(\varphi_k-\overline{\varphi_0})\alpha_k'(u_k)\nabla u_k\,dx\,dt \\ &\quad + \int_{Q_T} \sigma_k(u_k)\nabla\varphi_k\nabla\overline{\varphi_0}(v_k-M)^+\,dx\,dt. \end{aligned} \tag{4.14}$$

Without loss of generality, assume $a(0)=0$. According to (A4) and (4.3), there exist two positive numbers M_1 and M_2 such that

$$M_1 \leqq \alpha_k'(u_k) \leqq M_2 \quad \text{a.e. on } \{u_k \geqq s_k\} \tag{4.15}$$

at least for k sufficiently large. Also, the term $\int_0^T \int_{\partial\Omega} f(x,t,u_k)(v_k-M)^+\,dS\,dt$ is nonnegative due to (A2). Consequently, we deduce from (4.9), (4.10), and (4.14) that

$$\begin{aligned} &\int_\Omega [(v_k(x,T)-M)^+]^2\,dx + c\int_{\{u_k\geqq s_k\}} |\nabla u_k|^2\,dx\,dt \\ &\leqq c_1 + c_2 \int_{Q_T} [(v_k-M)^+]^2\,dx\,dt. \end{aligned} \tag{4.16}$$

It is not difficult to see from the proof of (4.16) that

$$\frac{1}{2}\int_\Omega [(v_k(x,t)-M)^+]^2\,dx \leqq c_1 + c_2 \int_0^t \int_\Omega [(v_k-M)^+]^2\,dx\,dt$$

for all $t\in(0,T]$.

An application of Gronwall's inequality yields

$$\sup_{0\leqq t\leqq T} \int_\Omega [(v_k(x,t)-M)^+]^2\,dx \leqq c \qquad (k=1,2,\cdots). \tag{4.17}$$

In a similar fashion, we can find a positive number M_3 so that

$$\sup_{0\leqq t\leqq T}\int_\Omega [(v_k(x,t)+M_3)^-]^2\,dx\leqq c \qquad (k=1,2,\cdots). \tag{4.18}$$

Combining (4.17) and (4.18) yields

$$\sup_{0\leqq t\leqq T}\int_\Omega v_k^2(x,t)\,dx\leqq c \qquad (k=1,2,\cdots). \tag{4.19}$$

This, together with [A4], implies

$$\sup_{0\leqq t\leqq T}\int_\Omega u_k^2(x,t)\,dx\leqq c \qquad (k=1,2,\cdots). \tag{4.20}$$

Use u_k as a test function in (4.5), and keep in mind Lemma 3.3, to get

$$\begin{aligned}&\int_\Omega\int_0^{v_k(x,T)}\alpha_k^{-1}(s)\,ds\,dx-\int_\Omega\int_0^{v_0(x)}\alpha_k^{-1}(s)\,ds\,dx+c\int_{Q_T}|\nabla u_k|^2\,dx\,dt\\ &\leqq-\int_{Q_T}\sigma_k(u_k)\nabla\varphi_k(\varphi_k-\overline{\varphi_0})\nabla u_k\,dx\,dt+\int_{Q_T}\sigma_k(u_k)\nabla\varphi_k\nabla\overline{\varphi_0}u_k\,dx\,dt.\end{aligned} \tag{4.21}$$

Since $\alpha_k(0)=0$, the term $\int_0^{v_k(x,T)}\alpha_k^{-1}(s)\,ds=\int_0^{\alpha_k(u_k(x,T))}\alpha_k^{-1}(s)\,ds$ is nonnegative. Also, recall that $\alpha_k^{-1}(s)\to\alpha^{-1}(s)$ uniformly on $\mathbb{R}$ as $k\to\infty$. We have

$$\int_\Omega\int_0^{v_0(x)}\alpha_k^{-1}(s)\,ds\,dx\to\int_\Omega\int_0^{v_0(x)}\alpha^{-1}(s)\,ds\,dx,$$

which is finite by our assumption. We estimate from (4.21), using (4.20), (4.9), and (4.10) that

$$\int_{Q_T}|\nabla u_k|^2\,dx\,dt\leqq c \qquad (k=1,2,\cdots).$$

This completes the proof of (4.11).

It follows from (4.11) and (4.5) that $\{(\partial/\partial t)v_k\}$ is bounded in $L^2(0,T;U^*)$. This, together with (4.19), implies that $\{v_k\}$ is precompact in $L^2(0,T;U^*)$ (see [S]). Thus we may select a subsequence of $\{k\}$, still denoted by $\{k\}$, so that

$$u_k\to u \quad \text{weakly in } L^2(0,T;W^{1,2}(\Omega)), \tag{4.22}$$

$$a(\nabla u_k)\to L \quad \text{weakly in } [L^2(Q_T)]^N, \tag{4.23}$$

$$v_k\to v \quad \text{weakly in } L^2(Q_T) \quad \text{and strongly in } L^2(0,T;U^*), \tag{4.24}$$

$$\sigma_k(v_k)\nabla\varphi_k\to g \quad \text{weakly in } [L^2(Q_T)]^N, \tag{4.25}$$

$$\varphi_k\to\varphi \quad \text{weakly* in } L^\infty(Q_T). \tag{4.26}$$

By (4.22) and (4.24), we have

$$\int_{Q_T}(v_k-v)(u_k-u)\,dx\,dt=\int_0^T(v_k-v,u_k-u)\,dt\to 0 \quad \text{as } k\to\infty,$$

where $(\cdot,\cdot)$ denoted the duality pairing between U^* and $W^{1,2}(\Omega)$.

Consequently,

$$\begin{aligned}\int_{Q_T}&(\alpha_k(u_k)-\alpha_k(u))(u_k-u)\,dx\,dt\\&=\int_{Q_T}(\alpha_k(u_k)-v)(u_k-u)\,dx\,dt\\&\quad+\int_{Q_T}(v-\alpha_k(u))(u_k-u)\,dx\,dt\to 0\quad\text{as } k\to\infty.\end{aligned}$$

Hence, passing if need be to a further subsequence, we deduce

$$0\leqq(\alpha_k(u_k)-\alpha_k(u))(u_k-u)\to 0\quad\text{a.e. on } Q_T. \tag{4.27}$$

Since $\alpha^{-1}(s)$ is a single-valued function, we can derive from (4.27) that

$$u_k\to u\quad\text{a.e. on } Q_T. \tag{4.28}$$

In view of (A4) and (4.28), we have that $v_ku_k\to vu$ almost everywhere on Q_T. By (4.27), $\int_{Q_T} v_ku_k\,dx\,dt\to\int_{Q_T} vu\,dx\,dt$. Then a result of [X4] asserts that $\{v_ku_k\}$ is uniformly integrable. This, together with (A4), implies that $\{u_k^2\}$ is uniformly integrable. We may appeal to Vitali's theorem to conclude that

$$u_k\to u\quad\text{strongly in } L^2(0,T;L^2(\Omega)). \tag{4.29}$$

Now we are in a position to invoke the trace theorem [CDK] to get

$$u_k\to u\quad\text{strongly in } L^2(0,T;L^2(\partial\Omega)). \tag{4.30}$$

We wish to modify the proof presented in [X2, § 3] so that we can establish

$$\varphi_k\to\varphi\quad\text{a.e. on } Q_T. \tag{4.31}$$

For each k let

$$\psi_k=\varphi_k-\overline{\varphi_0}\quad\text{and}\quad\psi=\varphi-\overline{\varphi_0}.$$

Then ψ_k is a solution of the problem

$$-\operatorname{div}(\sigma_k(u_k)\nabla\psi_k)=\operatorname{div}(\sigma_k(u_k)\nabla\overline{\varphi_0})\quad\text{in } Q_T,$$

$$\psi_k=0\quad\text{in } S_T.$$

For $\theta\in\mathscr{A}$, where $\mathscr{A}$ is defined in § 1, and $m\in\{1,2,\cdots\}$, let

$$\phi_k^{(m)}=1-\theta\left(\frac{1}{m}|u_k|\right),\qquad \phi^{(m)}=1-\theta\left(\frac{1}{m}|u|\right)\equiv F_\theta^{(m)}.$$

By the proof presented in [X2, § 3], there hold

$$0\leqq\phi_k^{(m)}\leqq 1,\qquad 0\leqq\phi^{(m)}\leqq 1,$$

$$\int_{Q_T}|\nabla\phi_k^{(m)}|^2\,dx\,dt\leqq\frac{c}{m^2},$$

$$\int_{Q_T}|\nabla\phi^{(m)}|^2\,dx\,dt\leqq\frac{c}{m^2}\qquad(m=1,2,\cdots),$$

and for each fixed m

$$\phi_k^{(m)}\to\phi^{(m)}\quad\text{strongly in } L^2(Q_T)\ \text{ and weakly in } L^2(0,T;W^{1,2}(\Omega)),$$

$$\psi_k\phi_k^{(m)}\to\psi\phi^{(m)}\quad\text{weakly in } L^2(0,T;W_0^{1,2}(\Omega)).$$

Now we are in a position to establish

$$(4.32)\quad \limsup_{k\to\infty}\int_{Q_T}\sigma_k(u_k)|\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})|^2\,dx\,dt\leqq\frac{c}{m}\qquad(m=1,2,\cdots).$$

To this end, we calculate, for $\chi\in L^2(0,T;W_0^{1,2}(\Omega))$, that

$$\begin{aligned}\int_{Q_T}\sigma_k(u_k)\nabla(\psi_k\phi_k^{(m)})\nabla\chi\,dx\,dt=&-\int_{Q_T}\sigma_k(u_k)\nabla\overline{\varphi_0}\nabla(\chi\phi_k^{(m)})\,dx\,dt\\&-\int_{Q_T}\sigma_k(u_k)\chi\nabla\psi_k\nabla\phi_k^{(m)}\,dx\,dt\\&+\int_{Q_T}\sigma_k(u_k)\psi_k\nabla\phi_k^{(m)}\nabla\chi\,dx\,dt.\end{aligned}$$

Note that $\psi_k\phi_k^{(m)},\psi\phi^{(m)}\in L^2(0,T;W_0^{1,2}(\Omega))$. Set $\chi=\psi_k\phi_k^{(m)}-\psi\phi^{(m)}$ in the above equation to obtain

$$\begin{aligned}&\int_{Q_T}\sigma_k(u_k)\nabla(\psi_k\phi_k^{(m)})\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt\\&=-\int_{Q_T}\sigma_k(u_k)\nabla\overline{\varphi_0}\nabla[(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\phi_k^{(m)}]\,dx\,dt\\&\quad-\int_{Q_T}\sigma_k(u_k)(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\nabla\psi_k\nabla\phi_k^{(m)}\,dx\,dt\\&\quad+\int_{Q_T}\sigma_k(u_k)\psi_k\nabla\phi_k^{(m)}\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt\\&\equiv I_1+I_2+I_3.\end{aligned}$$

Each term on the right-hand side is estimated below:

$$\begin{aligned}I_1=&-\int_{Q_T}\sigma_k(u_k)\nabla\overline{\varphi_0}\nabla\phi_k^{(m)}(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt\\&-\int_{Q_T}\sigma_k(u_k)\nabla\overline{\varphi_0}\phi_k^{(m)}\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt\\\leqq&\, c\|\nabla\overline{\varphi_0}\|_{L^2(Q_T)}\|\nabla\phi_k^{(m)}\|_{L^2(Q_T)}\\&-\int_{Q_T}\sigma_k(u_k)\nabla\overline{\varphi_0}\phi_k^{(m)}\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt\\\leqq&\frac{c_1}{m}-\int_{Q_T}\sigma_k(u_k)\nabla\overline{\varphi_0}\phi_k^{(m)}\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt\\\to&\frac{c_1}{m}\quad\text{as } k\to\infty.\end{aligned}$$

I_2 and I_3 can be estimated in the same manner as in [X2, § 3]:

$$\begin{aligned}|I_2|\leqq&\, c\|\sigma_k(u_k)\nabla\psi_k\|_{L^2(Q_T)}\|\nabla\phi_k^{(m)}\|_{L^2(Q_T)}\leqq\frac{c_1}{m},\\|I_3|\leqq&\frac{c}{m}+c_1\frac{1}{m^2}+\frac{c_2}{m}\|\sigma_k(u_k)\nabla(\psi\phi^{(m)})\|_{L^2(Q_T)}\\&\to\frac{c}{m}+c_1\frac{1}{m^2}+\frac{c_2}{m}\|\sigma(u)\nabla(\psi\phi^{(m)})\|_{L^2(Q_T)}\quad\text{as } k\to\infty\leqq\frac{c_3}{m}.\end{aligned}$$

We are ready to estimate

$$\limsup_{k\to\infty}\int_{Q_T}\sigma_k(u_k)|\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})|^2\,dx\,dt$$

$$\leqq\limsup_{k\to\infty}\int_{Q_T}\sigma_k(u_k)\nabla(\psi_k\phi_k^{(m)})\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt$$

$$+\lim_{k\to\infty}-\int_{Q_T}\sigma_k(u_k)\nabla(\psi\phi^{(m)})\nabla(\psi_k\phi_k^{(m)}-\psi\phi^{(m)})\,dx\,dt\leqq\frac{c}{m}.$$

This completes the proof of (4.32).

Choose a bounded domain D in $\mathbb{R}^N$ so that

$$D\supset\bar{\Omega}.$$

We extend each u_k to belong to $L^2(0, T;\ W_0^{1,2}(D))$ in such a manner that there still hold

$$u_k\to u\quad\text{strongly in } L^2(0, T;\ L^2(D)),$$

$$u_k\to u\quad\text{weakly in } L^2(0, T;\ W_0^{1,2}(D)).$$

A careful examination of the proofs of the extension theorems in [AD] clearly indicates that such an extension is possible. We extend each ψ_k to be identically zero outside Q_T. Clearly, we still have $\psi_k\in L^2(0, T;\ W_0^{1,2}(D))$ for each k, and (4.32) still holds with Q_T replaced by $DX(0, T)$.

Whenever $E\subset DX(0, T)$, $1\leqq p<\infty$, define

$$X_p(E)=\inf\left\{\int_{DX(0,T)}|\nabla v|^p\,dx\,dt\colon v\in L^p(0, T;\ W_0^{1,p}(D)),\ v\geqq 1\quad\text{a.e. on } E\right\}.$$

We easily verify that X_p is an outer measure over $DX(0, T)$; see [X2, Lemma 2.1]. We can also prove, as in [X2, Lemma 2.2], the following lemma.

LEMMA 4.1. *Assume that the sequence* $\{f_k\}$ *is bounded in* $L^2(0, T;\ W_0^{1,2}(D))$ *and precompact in* $L^2(0, T;\ L^2(D))$. *Then there exist a subsequence* $\{f_{k_j}\}\subset\{f_k\}$ *and a function* $f\in L^2(0, T;\ W_0^{1,2}(D))$ *such that for each* $1\leqq p<2$ *and each* $\delta>0$, *there exists a set* $E_\delta\subset DX(0, T)$ *with*

$$f_{k_j}\to f\quad\textit{uniformly on } E_\delta\quad\textit{and}\quad X_p(DX(0, T)\backslash E_\delta)\leqq\delta.$$

Once we have Lemma 4.1 and (4.32), we can appeal to the proof presented in [X2, § 3] to conclude that

$$\psi_k\to\psi\quad\text{a.e. on } DX(0, T).$$

This completes the proof of (4.31).

Remark. Here we do not have estimates of the type

$$d_{u,\varepsilon}(h)=\frac{1}{h^\varepsilon}\left(\int_0^{T-h}\int_D(u(x, t+h)-u(x, t))^2\,dx\,dt\right)^{1/2}\leqq c$$

for all $h\in(0, T)$ and some $\varepsilon>0$

for u_n. If we examine the proof of Lemmas 2.1 and 2.2 in [X2], the only purpose such estimates serve is to guarantee that a bounded sequence in $L^2(0, T;\ W_0^{1,2}(D))$ is precompact in $L^2(Q_T)$. Here we already know that $[u_k\}$ is precompact in $L^2(Q_T)$. Thus we may ignore the term $d_{u,\alpha}(h)$ and claim that Lemmas 2.1 and 2.2 in [X2] remain true even when $\varepsilon=0$. However, we are no longer able to use the method employed in [X2, § 2] to prove

$$\nabla u_k \to \nabla u \quad \text{a.e. on } Q_T.$$

For this, we need to adopt a different analysis here. We calculate, using Lemma 3.3, that

$$\begin{aligned}\int_0^T \left(\frac{\partial}{\partial t} v_k, u_k\right) dt &= \int_\Omega \int_0^{v_k(x,T)} \alpha_k^{-1}(s)\, ds\, dx - \int_\Omega \int_0^{v_0(x)} \alpha_k^{-1}(s)\, ds\, dx \\ &= \int_\Omega \int_0^{v_k(x,T)} (\alpha_k^{-1}(s)-\alpha^{-1}(s))\, ds\, dx \\ &\quad - \int_\Omega \int_0^{v_0(x)} (\alpha_k^{-1}(s)-\alpha^{-1}(s))\, ds\, dx \\ &\quad + \int_\Omega \int_0^{v_k(x,T)} \alpha^{-1}(s)\, ds\, dx - \int_\Omega \int_0^{v_0(x)} \alpha^{-1}(s)\, ds\, dx.\end{aligned} \tag{4.33}$$

Recall that $\alpha_k^{-1} \to \alpha^{-1}$ uniformly on $\mathbb{R}$ as $k\to\infty$. Thus the first two terms on the right-hand side of (4.33) converge to zero as $k\to\infty$. Since $\{(\partial/\partial t)v_k\}$ is bounded in $L^2(0, T;\ U^*)$, it is not difficult for us to verify that

$$v_k \in C([0, T];\ L^2(\Omega)),$$
$$v_k(\cdot, t)\to v(\cdot, t) \quad \text{weakly in } L^2(\Omega) \quad \text{for all } t\in[0, T].$$

We deduce from (4.33) that

$$\begin{aligned}\liminf_{k\to\infty} \int_0^T \left(\frac{\partial}{\partial t} v_k, u_k\right) dt &\geqq \int_\Omega \int_0^{v(x,T)} \alpha^{-1}(s)\, ds\, dx \\ &\quad - \int_\Omega \int_0^{v_0(x)} \alpha^{-1}(s)\, ds\, dx.\end{aligned} \tag{4.34}$$

We may take $k\to\infty$ in (4.5) to get

$$\frac{\partial}{\partial t} v + \tilde{A}(u) = \tilde{B}(u, \varphi) \quad \text{in } L^2(0, T;\ U^*), \tag{4.35}$$

where $\tilde{A}(u)$ is given by

$$\begin{aligned}\int_0^T (\tilde{A}(u), v)\, dt = \int_0^T \int_\Omega L\nabla v\, dx\, dt + \int_0^T \int_{\partial\Omega} f(x, t, u) v\, ds\, dt \\ \text{for each } v\in L^2(0, T;\ W^{1,2}(\Omega)),\end{aligned} \tag{4.36}$$

and $\tilde{B}(u, \varphi)$ is defined by

$$\begin{aligned}\int_0^T (\tilde{B}(u, \varphi), v)\, dt = -\int_{Q_T} (\varphi - \overline{\varphi_0}) g \nabla v\, dx\, dt + \int_{Q_T} g\nabla \overline{\varphi_0} v\, dx\, dt \\ \text{for each } v\in L^2(0, T;\ W^{1,2}(\Omega)).\end{aligned} \tag{4.37}$$

A result of [BCS] asserts that

$$v\in\alpha(u) \quad \text{a.e. on } Q_T.$$

Moreover, we have that $(\partial/\partial t)v\in L^2(0, T;\ U^*)$ and that

$$\int_0^T \left(\frac{\partial}{\partial t} v, u\right) dt = \int_\Omega \int_0^{v(x,T)} \alpha^{-1}(s)\, ds\, dx - \int_\Omega \int_0^{v_0(x)} \alpha^{-1}(s)\, ds\, dx. \tag{4.38}$$

Now we use u_k as a test function in (4.5) to get

$$
\begin{aligned}
\int_{Q_T} a(\nabla u_k)\nabla u_k\,dx\,dt = &-\int_0^T \left(\frac{\partial}{\partial t} v_k, u_k\right) dt - \int_0^T \int_{\partial\Omega} f(x,t,u_k)u_k\,ds\,dt \\
&-\int_{Q_T} \sigma_k(u_k)\nabla\varphi_k(\varphi_k - \overline{\varphi_0})\nabla u_k\,dx\,dt \qquad (4.39)\\
&+\int_{Q_T} \sigma_k(u_k)\nabla\varphi_k\nabla\overline{\varphi_0}u_k\,dx\,dt.
\end{aligned}
$$

Pick a θ from $\mathscr{A}$ so that $\theta(s)=0$ on $(-\infty, \frac{1}{2})$. For each m let

$$p(m) = \sup_{|s|\geqq m/2} \sigma(s).$$

We estimate

$$
\begin{aligned}
&\left|\int_{Q_T} \sigma_k(u_k)\nabla\left(\varphi_k\theta\left(\frac{1}{m}|u_k|\right)\right)(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right| \\
&\leqq \left|\int_{Q_T} \sigma_k(u_k)\theta\left(\frac{1}{m}|u_k|\right)\nabla\varphi_k(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right| \\
&\quad+\left|\int_{Q_T} \sigma_k(u_k)\varphi_k\nabla\left(\theta\left(\frac{1}{m}|u_k|\right)\right)(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right| \qquad (4.40)\\
&\leqq\left(\int_{Q_T}\left[\sigma_k(u_k)\theta\left(\frac{1}{m}|u_k|\right)(\varphi_k-\overline{\varphi_0})\right]^2|\nabla\varphi_k|^2\,dx\,dt\right)^{1/2}\|\nabla(u_k-u)\|_{L^2(Q_T)}+\frac{c}{m} \\
&\leqq c_1\sqrt{p(m)+1/k}+\frac{c}{m}.
\end{aligned}
$$

We have that

$$
\begin{aligned}
&\limsup_{k\to\infty}\left|\int_{Q_T} \sigma_k(u_k)\nabla(\varphi_k\phi_k^{(m)})(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right| \\
&\leqq \limsup_{k\to\infty}\left|\int_{Q_T} \sigma_k(u_k)\nabla(\psi_k\phi_k^{(m)})(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right| \qquad (4.41)\\
&\quad+\limsup_{k\to\infty}\left|\int_{Q_T} \sigma_k(u_k)\nabla(\bar{\varphi}_0\phi_k^{(m)})(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right. \\
&\equiv J_1+J_2.
\end{aligned}
$$

Note that $\{\sigma_k(u_k)(\varphi_k-\bar{\varphi})\}$ converges strongly in $L^2(Q_T)$ due to (4.31). Consequently,

$$
\begin{aligned}
J_2 \leqq &\lim_{k\to\infty}\left|\int_{Q_T} \sigma_k(u_k)\phi_k^{(m)}\nabla\overline{\varphi_0}(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right| \\
&+\limsup_{k\to\infty}\left|\int_{Q_T} \sigma_k(u_k)\overline{\varphi_0}\nabla\phi_k^{(m)}(\varphi_k - \overline{\varphi_0})\nabla(u_k-u)\,dx\,dt\right| \\
&\leqq \limsup_{k\to\infty} c\|\nabla\phi_k^{(m)}\|_{L^2(Q_T)}\|\nabla(u_k-u)\|_{L^2(Q_T)} \leqq \frac{c_1}{m}.
\end{aligned}
$$

To estimate J_1, we use (4.32) to obtain

$$J_1 \leqq \limsup_{k\to\infty} \left| \int_{Q_T} \sigma_k(u_k)\nabla(\psi_k\phi_k^{(m)} - \psi\phi^{(m)})(\varphi_k - \overline{\varphi_0})\nabla(u_k - u)\, dx\, dt \right|$$

$$+ \lim_{k\to\infty} \left| \int_{Q_T} \sigma_k(u_k)\nabla(\psi\phi^{(m)})(\varphi_k - \overline{\varphi_0})\nabla(u_k - u)\, dx\, dt \right| \leqq \frac{c}{\sqrt{m}}.$$

We conclude that

$$\limsup_{k\to\infty} \left| \int_{Q_T} \sigma_k(u_k)\nabla(\varphi_k\phi_k^{(m)})(\varphi_k - \overline{\varphi_0})\nabla(u_k - u)\, dx\, dt \right| \leqq \frac{c}{\sqrt{m}}. \tag{4.42}$$

We derive from (4.40) and (4.42) that

$$\begin{aligned} &\limsup_{k\to\infty} - \int_{Q_T} \sigma_k(u_k)\nabla\varphi_k(\varphi_k - \overline{\varphi_0})\nabla u_k\, dx\, dt \\ &\leqq - \int_{Q_T} g(\varphi - \overline{\varphi_0})\nabla u\, dx\, dt \\ &\quad + \limsup_{k\to\infty} - \int_{Q_T} \sigma_k(u_k)\nabla\left(\varphi_k\phi_k^{(m)} + \varphi_k\theta\left(\frac{1}{m}|u_k|\right)\right) \\ &\quad \cdot(\varphi_k - \overline{\varphi_0})\nabla(u_k - u)\, dx\, dt \\ &\leqq - \int_{Q_T} g(\varphi - \overline{\varphi_0})\nabla u\, dx\, dt + c\left(\frac{1}{m} + \frac{1}{\sqrt{m}} + \sqrt{p(m)}\right). \end{aligned} \tag{4.43}$$

By (A3), $\lim_{m\to\infty} p(m) = 0$. Thus we have from (4.43) that

$$\limsup_{k\to\infty} - \int_{Q_T} \sigma_k(u_k)\nabla\varphi_k(\varphi_k - \overline{\varphi_0})\nabla u_k\, dx\, dt \leqq - \int_{Q_T} g(\varphi - \overline{\varphi_0})\nabla u\, dx\, dt. \tag{4.44}$$

Letting $k \to \infty$ in (4.39) and taking into account (4.34), (4.31), (4.30), and (4.44) yields

$$\begin{aligned} \limsup_{k\to\infty} \int_{Q_T} a(\nabla u_k)\nabla u_k\, dx\, dt \leqq &- \int_\Omega \int_0^{v(x,T)} \alpha^{-1}(s)\, ds\, dx + \int_\Omega \int_0^{v_0(x)} \alpha^{-1}(s)\, ds\, dx \\ &- \int_0^T \int_{\partial\Omega} f(x, t, u)u\, ds\, dt \\ &- \int_{Q_T} g(\varphi - \overline{\varphi_0})\nabla u\, dx\, dt + \int_{Q_T} g\nabla\overline{\varphi_0}u\, dx\, dt. \end{aligned}$$

This, together with (4.35) and (4.38), implies

$$\limsup_{k\to\infty} \int_{Q_T} a(\nabla u_k)\nabla u_k\, dx\, dt \leqq \int_{Q_T} L\nabla u\, dx\, dt$$

whence follows

$$\lim_{k\to\infty} \int_{Q_T} |\nabla u_k - \nabla u|^2\, dx\, dt = 0.$$

Consequently,

$$L = a(\nabla u) \quad \text{a.e. } Q_T.$$

Equation (1.2) is a consequence of (4.35). Taking $k \to \infty$ in (4.6) yields (1.3). Equations (1.4), (1.5), and (1.6) can be obtained in the same manner as in [X2]. This completes the proof of our main theorem.

REFERENCES

[A] D. R. ATTHEY, *A finite difference scheme for melting problems*, J. Inst. Maths. Appl., 13 (1974), pp. 353-366.

[AD] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[AX] W. ALLEGRETTO AND H. XIE, *$C^\alpha(\Omega)$ solution of a class of nonlinear degenerate elliptic system arising in the thermistor problem*, preprint.

[BCS] P. BENILAN, M. G. CRANDALL, AND P. SACKS, *Some L^1 existence and dependence results for semilinear elliptic equations under nonlinear boundary conditions*, Appl. Math. Optim., 17 (1988), pp. 203-224.

[BR1] H. BRÉZIS, *Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations*, in Contributions to Nonlinear Analysis, E. Zarantonello, ed., Academic Press, New York, 1971, pp. 101-156.

[BR2] ———, *Integrales convexes dans les espaces de Sobolev*, Israel J. Math., 13 (1972), pp. 9-23.

[CDK] J. R. CANNON, E. DIBENEDETTO, AND G. H. KNIGHTLY, *The bidimensional Stefan problem with convection: the time dependent case*, Comm. Partial Differential Equations, 8 (1983), pp. 1549-1604.

[CF1] X. CHEN AND A. FRIEDMAN, *The thermistor problem for conductivity which vanishes at large temperature*, Quart. Appl. Math., to appear.

[CF2] ———, *The thermistor problem with one-zero conductivity*, IMA preprint series # 793, 1991.

[CH] X. CHEN, *Existence and regularity of solutions of a nonlinear degenerate elliptic system arising from a thermistor problem*, preprint.

[C1] G. CIMATTI, *Existence of weak solutions for the nonstationary problem of the Joule heating of a conductor*, Ann. Mat. Pura Appl., to appear.

[C2] ———, *Remarks on existence and uniqueness for the thermistor problem under mixed boundary conditions*, Quart. Appl. Math., 47 (1989), pp. 117-121.

[C3] ———, *A bound for the temperature in the thermistor's problem*, IMA J. Appl. Math., 40 (1988), pp. 15-22.

[CP] G. CIMATTI AND G. PRODI, *Existence results for a nonlinear elliptic system modeling a temperature dependent electrical resistor*, Ann. Mat. Pura Appl., 63 (1988), pp. 227-236.

[HRS] D. HOWISON, J. R. RODRIGUES, AND M. SHILLOR, *Stationary solutions to the thermistor problem*, J. Math. Anal. Appl., to appear.

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

[M] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.

[O] J. ODEN, *Qualitative Methods in Nonlinear Mechanics*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[SSX] P. SHI, M. SHILLOR, AND X. XU, *Existence of a solution to the Stefan problem with Joule's heating*, J. Differential Equations, to appear.

[S] J. SIMON, *Ecoulement d'un fluide nonhomogéne avec une densité initiale s'annulant*, C.R. Acad. Sci. Paris, 287 (1978), pp. 1009-1012.

[X1] X. XU, *Existence and regularity theorems for a two-phase degenerate Stefan problem with convection*, preprint.

[X2] ———, *An elliptic-parabolic system for degenerate type*, preprint.

[X3] ———, *Existence of a generalized weak solution to the degenerate thermistor problem*, preprint.

[X4] ———, *Existence and convergence theorems for doubly nonlinear partial differential equations of elliptic-parabolic type*, J. Math. Anal. Appl., 150 (1990), pp. 205-223.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1439–1454, November 1992

A FREE BOUNDARY PROBLEM ARISING IN ELECTROPHOTOGRAPHY: SOLUTIONS WITH CONNECTED TONER REGION*

BEI HU† AND LIHE WANG‡

Abstract. A free boundary problem that arises in the development of a photocopy is studied. The electric potential $-u$ satisfies the equation $\Delta u = 1$ in the toner region and $\Delta u = 0$ elsewhere. It is shown that $C^{1+\alpha}$ smoothness of the free boundary would imply the $C^{2+\alpha}$ smoothness of the solution up to both sides of the free boundary. Using this fact, the existence of a solution with connected it is proven that toner region with $\frac{\partial u}{\partial n} = 0$ on the free boundary when the electrical charge length is "small."

Key words. free boundary problems, electrophotography, photocopy, elliptic estimates

AMS(MOS) subject classifications. 35B45, 35R05, 35R35

1. Introduction.

One of the steps in the photocopying process is the development of the electrical image into a visible image. A positively charged toner is brushed on to the electrical image, and a visible dark image is therefore produced. This process is modeled as a nonstandard free boundary problem. (See [2], [3] for more details.)

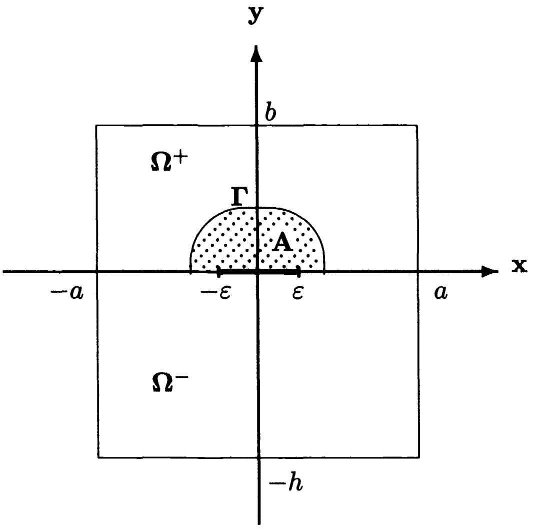


FIG. 1

We set (see Fig. 1)

$$\Omega^+ = \{(x,y);\ |x| < a, 0 < y < b\},$$
$$\Omega^- = \{(x,y);\ |x| < a, -h < y < 0\},$$

* Received by the editors June 17, 1991; accepted for publication (in revised form) March 11, 1992.

† Department of Mathematics, University of Notre Dame, Notre Dame, Indiana 46556. This author was partially supported by National Science Foundation grant DMS–9024986.

‡ Department of Mathematics, Princeton University, Princeton, New Jersey 08544. This author was partially supported by National Science Foundation grant DMS–8905039.

$$I = \{(x,0);\ |x| < \varepsilon\},$$
$$J = \{(x,0);\ |x| < a\},$$
$$\Omega = \{(x,y);\ |x| < a, -h < y < b\} = \Omega^+ \cup J \cup \Omega^-.$$

The problem is formulated as follows. Find the pair (u, A) such that

$$\Delta u = 1 \quad \text{in } A, \tag{1.1}$$
$$\Delta u = 0 \quad \text{in } \Omega \setminus \overline{A}, \tag{1.2}$$
$$u \in C^1(\overline{\Omega} \setminus \overline{I}), \tag{1.3}$$
$$u_y(x, 0+) - u_y(x, 0-) = -\sigma \quad \text{in } I, \tag{1.4}$$

and u satisfies the free boundary condition

$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma, \tag{1.5}$$

where $\Gamma = \partial A \cap \Omega^+$ and $\frac{\partial}{\partial n}$ is the outward normal to A, and also the boundary conditions:

$$u(x, -h) = 0 \qquad -a < x < a, \tag{1.6}$$
$$u(x, b) = M \qquad -a < x < a, \tag{1.7}$$
$$u_x(\pm a, y) = 0 \qquad -h < y < b. \tag{1.8}$$

Here $-u$ is the electrical potential. The jump condition (1.4) on the interval I represents a negative surface charge, and positively charged toner is expected to be attracted to cover the interval I. The domain A is the toner region, and $(\Omega^+ \setminus A)$ is the air region. The condition (1.5) indicates that there is no force to attract or repel the toner on the interface between the toner and air, and, therefore, the toner region reaches its equilibrium. This should be the case when the image is fully developed. Mathematically, the toner region A is a nonempty region where $\Delta u = 1$.

a, b, h, M, σ in (1.1)–(1.8) are positive constants, and it is reasonable to assume that (see [3])

$$M < \sigma h, \qquad b \geq h. \tag{1.9}$$

When $a - \varepsilon$ is small, the problem reduces to a variational inequality; it is proved in [3] that in this case the problem has a unique solution.

When ε is small, it is proved in [3] that the problem is no longer a variational inequality and there are infinitely many solutions with two symmetric components, it is not clear, however, whether such solutions are physical.

The toner is expected to cover the entire interval I, where a black dot should develop. A connected toner region over the interval I will serve such a purpose. It is shown in [3] that there exists a "ε^m-approximate" solution for which the toner set consists of a single component. The tool used there is the topological fixed point theorem. In order to use a topological fixed point theorem, the system (1.1)–(1.4) is solved with the boundary condition (1.6)–(1.8) for each given Γ, and then a new Γ is obtained, roughly speaking, by solving (1.5). The difficulty for finding a real solution with one connected toner region is that the corresponding u will have $\nabla u = 0$ at the point $\Gamma \cap \{x = 0\}$, which makes it very difficult to solve the dynamical system (1.5) for the new $\Gamma = \{(x,y);\ x = x(t),\ y = y(t)\}$. In [3], a $W^{2,p}$ estimate is employed for the Partial Differential Equation (PDE) solution; however, for this dynamical system coupled with the solution of the PDE, it is clear that more regularity is required on the solution of the PDE to use a fixed point theorem.

Here we prove that if ε is small, the problem has a solution with one connected toner region. In §2, we prove an elliptic estimate that is of independent interest. Using the elliptic estimate, we prove in §3 the existence result.

Recently, this problem has been formulated in [6] as an optimization problem. The solution of the optimization problem is an approximate solution of (1.1)–(1.8) in some sense.

2. Elliptic estimate. In this section, we shall establish that $C^{1+\alpha}$ smoothness of the free boundary would imply the $C^{2+\alpha}$ smoothness of the solution up to both sides of the free boundary. Suppose that Γ is given by $y = g(x)$ for $-2 < x < 2$ with $g(0) = 0$. Let $B(s)$ denote the ball of radius s centered at (0,0). Suppose that

$$\Delta u = \theta \chi_E \quad \text{for } (x,y) \in B(2), \quad (0 < \theta \leq 1), \tag{2.1}$$

where χ_E is the characteristic function of E and $E = \{y > g(x)\}$.

THEOREM 2.1. *Suppose that*

$$\sup_{B(2)} |u(x,y)| \leq L, \tag{2.2}$$

and

$$\|g\|_{C^{1+\alpha}(-2,2)} \leq K, \tag{2.3}$$

where $0 < \alpha < 1$. *Then*

$$\|u\|_{C^{2+\alpha}(\overline{E} \cap B(1))} \leq C, \tag{2.4}$$

$$\|u\|_{C^{2+\alpha}((\overline{B(1) \setminus E}))} \leq C, \tag{2.5}$$

where the constant C depends only on L, K, and α.

Remark. The proof below will actually show that the conclusion is also true in n dimensions.

The proof uses harmonic polynomials to approximate the derivatives. Such a technique was used in Caffarelli and Friedman [1]. (See also [4]). It is worth mentioning that because of the discontinuous right-hand side, the space L^1 seems to be the right choice (cf. Lemma 2.4).

We shall divide the proof into several lemmas.

LEMMA 2.2. *If*

$$\Delta u = f \quad \text{in } B(s), \tag{2.6}$$

$$u = 0 \quad \text{on } \partial B(s); \tag{2.7}$$

where $0 < s \leq 1$, *then*

$$\int_{B(s)} |u| \leq \frac{1}{4} \int_{B(s)} |f|. \tag{2.8}$$

Proof.

$$u(x,y) = \int_{B(s)} G(x-\xi, y-\eta) f(\xi,\eta) d\xi d\eta, \tag{2.9}$$

where G is the Green function on $B(s)$. Therefore,

$$\begin{aligned}\int_{B(s)} |u(x,y)|dx\,dy &\le \int_{B(s)} \left(\int_{B(s)} |G(x-\xi, y-\eta)|dx\,dy \right) |f(\xi,\eta)|d\xi d\eta \\ &= \int_{B(s)} \frac{1}{4}(s^2-\xi^2-\eta^2)|f(\xi,\eta)|d\xi d\eta \\ &\le \frac{1}{4}\int_{B(s)} |f(\xi,\eta)|d\xi d\eta. \end{aligned}$$ □

LEMMA 2.3. *Suppose that for some $s \in [\frac{3}{4}, 1]$, we have*

$$\Delta v = 0 \quad \textit{in } B(s) \tag{2.10}$$

$$\int_{\partial B(s)} |v|d\sigma \le L. \tag{2.11}$$

Then there exists a constant C_L depending only on L such that

$$|v(x,y) - P_2[v](x,y)| \le C_L r^3 \quad \textit{for } r = \sqrt{x^2+y^2} \le \tfrac{1}{2}\ , \tag{2.12}$$

where $P_2[v](x,y)$ is the second order polynomial of the Taylor series of v at $(0,0)$.

Proof. Using (2.11) and Poisson formula we conclude that

$$\|D^3 v\|_{L^\infty(B(1/2))} \le C_L, \tag{2.13}$$

from which the lemma follows. □

We now fix L and α. Then we take C_L as in Lemma 2.3, and select a number λ satisfying

$$0 < \lambda \le \frac{1}{2}, \qquad C_L \lambda^{1-\alpha} \le \frac{1}{2\pi}. \tag{2.14}$$

For such a fixed λ, we take ε_0 such that

$$\varepsilon_0 \le \lambda^{4+\alpha}. \tag{2.15}$$

Using the scaling $\overline{u}(x,y) = u(\delta x, \delta y)$ $(\delta = (\varepsilon_0/K)^{1/\alpha})$ if necessary, we may assume without loss of generality that

$$[g']_{C^\alpha(-2,2)} \le \varepsilon_0. \tag{2.16}$$

LEMMA 2.4. *Let the assumptions of Theorem 2.1 be in force, and let (2.16) hold. Then there exists a constant C such that for any $Q \in \Gamma \cap B(1)$, there exists P_Q*

$$\left\| u - \left(P_Q + \frac{\theta}{2}\left((\langle (x,y) - Q, \mathbf{n}_Q \rangle^+ \right)^2 \right) \right\|_{L^1(B_Q(r))} \le Cr^{4+\alpha} \quad \textit{for } 0 < r < \lambda, \tag{2.17}$$

where P_Q is a harmonic polynomial of second degree, and $\mathbf{n}_Q$ is the normal of Γ at Q in the direction of y-axis.

Proof. We may assume without loss of generality that

$$Q = (0,0), \quad g(0) = 0, \quad g'(0) = 0.$$

Then

$$|g(x)| \le \frac{\varepsilon_0}{1+\alpha}|x|^{1+\alpha} \le \varepsilon_0 |x|^{1+\alpha}. \tag{2.18}$$

Set

$$w_1(x,y) = u(x,y) - \frac{\theta}{2}(y^+)^2, \tag{2.19}$$

and define v_1 by

$$\Delta v_1 = 0 \quad \text{in } B(1), \tag{2.20}$$
$$v_1 = w_1 \quad \text{on } \partial B(1). \tag{2.21}$$

Then

$$\Delta(w_1 - v_1) = 0 \quad \text{in } B(1) \setminus \{|y| \ge \varepsilon_0 |x|\}, \tag{2.22}$$
$$|\Delta(w_1 - v_1)| \le 1 \quad \text{in } B(1). \tag{2.23}$$

Therefore, by Lemma 2.2 and (2.15),

$$\|w_1 - v_1\|_{L^1(B(1))} \le \tfrac{1}{4}\text{meas}(B(1) \cap \{|y| \le \varepsilon_0|x|\}) \le \tfrac{1}{2}\varepsilon_0 \le \tfrac{1}{2}\lambda^{4+\alpha}. \tag{2.24}$$

Clearly,

$$\int_{\partial B(1)} |v_1| = \int_{\partial B(1)} |w_1| \le L; \tag{2.25}$$

therefore, by Lemma 2.3,

$$\begin{aligned} &|v_1(x,y) - P_2[v_1](x,y)| \le C_L r^3 \le C_L \lambda^{1-\alpha}\lambda^{2+\alpha} \le \frac{1}{2\pi}\lambda^{2+\alpha} \\ &\quad \text{for } r = \sqrt{x^2+y^2} \le \lambda, \end{aligned} \tag{2.26}$$

which implies

$$\|v_1 - P_2[v_1]\|_{L^1(B(\lambda))} \le \frac{1}{2\pi}\lambda^{2+\alpha}\pi\lambda^2 \le \frac{1}{2}\lambda^{4+\alpha}. \tag{2.27}$$

The inequalities (2.24) and (2.27) imply

$$\|w_1 - P_2[v_1]\|_{L^1(B(\lambda))} \le \lambda^{4+\alpha}. \tag{2.28}$$

Next, define

$$w_2(x,y) = \frac{(w_1 - P_2[v_1])(\lambda x, \lambda y)}{\lambda^{2+\alpha}}, \tag{2.29}$$

and observe that, by (2.28),

$$\|w_2\|_{L^1(B(1))} = \frac{1}{\lambda^{4+\alpha}}\|w_1 - P_2[v_1]\|_{L^1(B(\lambda))} \le 1. \tag{2.30}$$

It follows that there exists $s_0 \in [\frac{3}{4}, 1]$ such that

$$\int_{\partial B(s_0)} |w_2| \le 4 \le L. \tag{2.31}$$

Now, define v_2 by

$$\Delta v_2 = 0 \quad \text{in } B(s_0), \tag{2.32}$$

$$v_2 = w_2 \quad \text{on } \partial B(s_0). \tag{2.33}$$

Then

$$\Delta(w_2 - v_2) = 0 \quad \text{in } B(s_0) \setminus \{|y| \geq \varepsilon_0 \lambda^\alpha |x|\}, \tag{2.34}$$

$$|\Delta(w_2 - v_2)| \leq \frac{1}{\lambda^\alpha} \quad \text{in } B(s_0). \tag{2.35}$$

Therefore, by Lemma 2.2 and (2.15),

$$\|w_2 - v_2\|_{L^1(B(s_0))} \leq \frac{1}{4} \int_{(B(s_0) \cap \{|y| \leq \varepsilon_0 \lambda^\alpha |x|\})} \frac{1}{\lambda^\alpha} \leq \frac{1}{2}\varepsilon_0 \leq \frac{1}{2}\lambda^{4+\alpha}. \tag{2.36}$$

Clearly by Lemma 2.3, using (2.31),

$$\begin{aligned} &|v_2(x,y) - P_2[v_2](x,y)| \leq C_L r^3 \leq C_L \lambda^{1-\alpha}\lambda^{2+\alpha} \leq \frac{1}{2\pi}\lambda^{2+\alpha} \\ &\text{for } r = \sqrt{x^2+y^2} \leq \lambda, \end{aligned} \tag{2.37}$$

and, therefore,

$$\|v_2 - P_2[v_2]\|_{L^1(B(\lambda))} \leq \frac{1}{2\pi}\lambda^{2+\alpha}\pi\lambda^2 \leq \frac{1}{2}\lambda^{4+\alpha}. \tag{2.38}$$

The inequalities (2.36) and (2.38) imply

$$\|w_2 - P_2[v_2]\|_{L^1(B(\lambda))} \leq \lambda^{4+\alpha}. \tag{2.39}$$

Now we inductively define

$$w_n(x,y) = \frac{(w_{n-1} - P_2[v_{n-1}])(\lambda x, \lambda y)}{\lambda^{2+\alpha}}. \tag{2.40}$$

Notice that whenever we scale the domain by λ, we get one more $1/\lambda^\alpha$ factor on the right-hand side of the equation; but that is compensated by the fact that we get one more factor of λ^α for the domain at the same time. Hence we obtain

$$\|w_n - P_2[v_n]\|_{L^1(B(\lambda))} \leq \lambda^{4+\alpha}, \tag{2.41}$$

where $P_2[v_n]$ is a harmonic polynomial of the second degree. It follows from (2.40) and (2.41) that

$$\|w_1 - P_n\|_{L^1(B(\lambda^n))} \leq C(\lambda^n)^{4+\alpha}, \tag{2.42}$$

where $C = 1/\lambda^2$, and

$$P_n = \sum_{k=1}^{n} P_2[v_k]\left(\frac{x}{\lambda^{k-1}}, \frac{y}{\lambda^{k-1}}\right)(\lambda^{k-1})^{2+\alpha}. \tag{2.43}$$

It is obvious that all coefficients of the harmonic polynomials $P_2[v_k]$ are bounded with the bounds depending only on L. If we set

$$P = \lim_{n\to\infty} P_n = \sum_{k=1}^{\infty} P_2[v_k]\left(\frac{x}{\lambda^{k-1}}, \frac{y}{\lambda^{k-1}}\right)(\lambda^{k-1})^{2+\alpha}, \tag{2.44}$$

then (assuming that $\lambda^\alpha \leq \frac{1}{2}$)

$$|\text{0th order coefficients of } (P - P_n)| \leq C \sum_{k=n}^{\infty} (\lambda^{2+\alpha})^k \leq 2C(\lambda^{2+\alpha})^n, \tag{2.45}$$

$$|\text{1st order coefficients of } (P - P_n)| \leq C \sum_{k=n}^{\infty} (\lambda^{1+\alpha})^k \leq 2C(\lambda^{1+\alpha})^n, \tag{2.46}$$

$$|\text{2nd order coefficients of } (P - P_n)| \leq C \sum_{k=n}^{\infty} (\lambda^{\alpha})^k \leq 2C(\lambda^{\alpha})^n. \tag{2.47}$$

So

$$\begin{aligned} &\|P - P_n\|_{L^1(B(\lambda^n))} \\ &\leq C\left[(\lambda^{2+\alpha})^n(\lambda^n)^2 + (\lambda^{1+\alpha})^n\lambda^n(\lambda^n)^2 + (\lambda^\alpha)^n(\lambda^n)^2(\lambda^n)^2\right] \\ &\leq C(\lambda^n)^{4+\alpha}. \end{aligned} \tag{2.48}$$

For each $0 < r < \lambda$, choose n so that $\lambda^{n+1} < r \leq \lambda^n$. Then by (2.48) and (2.41), we obtain

$$\|w_1 - P\|_{L^1(B(r)} \leq Cr^{4+\alpha} \quad \text{for } 0 < r < \lambda, \tag{2.49}$$

where P is a harmonic polynomial of second degree. □

Next, for any point $Q \in B(1)$, take $\pi(Q) \in \Gamma$ such that

$$d(Q, \pi(Q)) = \inf\{|Q - S|;\ S \in \Gamma\}. \tag{2.50}$$

Although Γ is not C^2 and, therefore, $\pi(Q)$ need not be uniquely determined by Q, the map $\pi : B(1) \to \Gamma$ is well defined by axiom of choice.

LEMMA 2.5. *Let the assumptions of Theorem* 2.1 *be in force, and let* (2.16) *hold. Then there exists a constant C such that for any $Q \in B(1)$, there exists P_Q*

$$\left\| u - \left(P_Q + \frac{\theta}{2}\left(\langle (x,y) - \pi(Q), \mathbf{n}_{\pi(Q)}\rangle^+\right)^2\right)\right\|_{L^1(B_Q(r))} \leq Cr^{4+\alpha} \quad \textit{for } 0 < r < \lambda\ , \tag{2.51}$$

where P_Q is a harmonic polynomial of second degree, and $\mathbf{n}_{\pi(Q)}$ is the normal of Γ at $\pi(Q)$ in the direction of y-axis.

Proof. We may assume without loss of generality that

$$\pi(Q) = (0,0), \quad g(0) = 0, \quad g'(0) = 0.$$

Set

$$w(x,y) = u(x,y) - \frac{\theta}{2}(y^+)^2, \tag{2.52}$$

and

$$G = \{x;\ \Delta w(x) \neq 0\}. \tag{2.53}$$

Then it is clear that

$$|G \cap B_Q(r)| \leq \varepsilon_0 r^{2+\alpha} \quad \text{for } 0 < r < 1, \tag{2.54}$$

where $B_Q(r)$ is a ball of radius r centered at Q. Now the remainder of the proof is the same as Lemma 2.4, where we shall use (2.54) instead of Lemma 2.2. □

Lemmas 2.4 and 2.5 give the equalities

$$u(Q) = P_Q(Q), \tag{2.55}$$

$$Du(Q) = DP_Q(Q), \tag{2.56}$$

and if $Q \notin \Gamma$,

$$D^2u(Q) = D^2P_Q(Q) + \frac{\theta}{2} D^2 \left((\langle (x,y) - \pi(Q), \mathbf{n}_{\pi(Q)} \rangle^+)^2 \right|_{(x,y)=Q}, \tag{2.57}$$

which implies that $\|u\|_{W^{2,\infty}(B(1))} \leq C$.

Lemma 2.6. *For any $Q_1, Q_2 \in B(1)$, there holds*

$$\begin{aligned} &\left| P_{Q_1}(x,y) - P_{Q_2}(x,y) + \frac{\theta}{2} \left((\langle (x,y) - \pi(Q_1), \mathbf{n}_{\pi(Q_1)} \rangle^+ \right)^2 \right. \\ &\left. - \frac{\theta}{2} \left((\langle (x,y) - \pi(Q_2), \mathbf{n}_{\pi(Q_2)} \rangle^+ \right)^2 \right| \leq Cr^{2+\alpha} \quad \textit{for} \quad \left| (x,y) - \frac{Q_1+Q_2}{2} \right| \leq \frac{r}{2}, \end{aligned} \tag{2.58}$$

where $r = 3|Q_1 - Q_2|$.

Proof. Let $d_i = d(Q_i, \pi(Q_i))$ for $i = 1, 2$. Without loss of generality, we assume that $d_1 \geq d_2$.

Case 1. $d_2 \geq 2r$.

In this case the balls $B_{Q_1}(r)$ and $B_{Q_2}(r)$ do not intersect and $B_{\frac{1}{2}(Q_1+Q_2)}(r)$ does not intersect Γ. Hence

$$\Delta \left(P_{Q_1} - P_{Q_2} + q \right) = 0 \quad \text{in } B_{\frac{1}{2}(Q_1+Q_2)}(r), \tag{2.59}$$

where

$$q(x,y) = \frac{\theta}{2} \left((\langle (x,y) - \pi(Q_1), \mathbf{n}_{\pi(Q_1)} \rangle^+ \right)^2 - \frac{\theta}{2} \left((\langle (x,y) - \pi(Q_2), \mathbf{n}_{\pi(Q_2)} \rangle^+ \right)^2 .$$

It follows that

$$\|P_{Q_1} - P_{Q_2} + q\|_{L^1(\partial B_{\frac{1}{2}(Q_1+Q_2)}(s_r))} \leq \frac{Cr^{4+\alpha}}{r/4} \leq Cr^{3+\alpha} \tag{2.60}$$

for some $s_r \in \left(\frac{3}{4}r, r\right]$. Now by the Poisson formula

$$\begin{aligned} &\|P_{Q_1} - P_{Q_2} + q\|_{L^\infty(B_{\frac{1}{2}(Q_1+Q_2)}(r/2))} \\ &\leq \frac{C}{r/4} \|P_{Q_1} - P_{Q_2} + q\|_{L^1(\partial B_{\frac{1}{2}(Q_1+Q_2)}(s_r))} \\ &\leq Cr^{2+\alpha}. \end{aligned} \tag{2.61}$$

Case 2. $d_2 < 2r$.

In this case,

$$d_1 = |Q_1 - \pi(Q_1)| \leq |Q_1 - \pi(Q_2)| \leq |Q_1 - Q_2| + d_2 \leq 5r$$

so that

$$|\pi(Q_1) - \pi(Q_2)| \leq d_1 + |Q_1 - Q_2| + d_2 \leq 5r + 3r + 2r = 10r. \tag{2.62}$$

Therefore,

$$\left|\langle \pi(Q_1)-\pi(Q_2), \mathbf{n}_{\pi(Q_2)}\rangle\right| \le Cr^{1+\alpha},$$

$$|\mathbf{n}_{\pi(Q_1)}-\mathbf{n}_{\pi(Q_2)}| = \left|\left[\frac{(-g',1)}{\sqrt{1+(g')^2}}\right]_{\pi(Q_1)}^{\pi(Q_2)}\right| \le Cr^{\alpha}. \tag{2.63}$$

This in turn implies

$$\begin{aligned} |q(x,y)| &\le Cr\left|\langle (x,y)-\pi(Q_1), \mathbf{n}_{\pi(Q_1)}\rangle^+ - \langle (x,y)-\pi(Q_2), \mathbf{n}_{\pi(Q_2)}\rangle^+\right| \\ &\le Cr\left|\langle (x,y)-\pi(Q_1), \mathbf{n}_{\pi(Q_1)}-\mathbf{n}_{\pi(Q_2)}\rangle\right| \\ &\quad +Cr\left|\langle \pi(Q_1)-\pi(Q_2), \mathbf{n}_{\pi(Q_2)}\rangle\right| \\ &\le Cr^{2+\alpha} \end{aligned} \tag{2.64}$$

for $(x,y)\in B(r)$. Similar to the proof of Case 1 from (2.59) to (2.61), we can now estimate $P_{Q_1}-P_{Q_2}$ (which is harmonic). Since the error term q is controlled by (2.64), the proof goes through. □

Proof of Theorem 2.1. Take $Q_1, Q_2 \in B(1)\setminus E$ (in which $\Delta u = 0$). Notice that in this case the second term of (2.57) is zero for both $Q = Q_1$ and $Q = Q_2$; therefore, by Lemma 2.6 and (2.55)–(2.57),

$$\begin{aligned} &\big|\big(u(Q_1)+Du(Q_1)\cdot(X-Q_1)+\tfrac{1}{2}(X-Q_1)^T D^2u(Q_1)(X-Q_1)\big) \\ &\quad -\big(u(Q_2)+Du(Q_2)\cdot(X-Q_2)+\tfrac{1}{2}(X-Q_2)^T D^2u(Q_2)(X-Q_2)\big)\big| \\ &\le C|Q_1-Q_2|^{2+\alpha} \end{aligned} \tag{2.65}$$

for all X with $\left|X-\frac{1}{2}(Q_1+Q_2)\right| \le 3r$. This implies that

$$\|u\|_{C^{2+\alpha}((\overline{B(1)\setminus E}))} \le C.$$

The estimates on $\overline{E}\cap B(1)$ is obtained by considering $-\left(u-\frac{\theta}{4}(x^2+y^2)\right)$. □

3. Existence of a solution with connected toner region. We shall always assume (1.9). Let $\alpha\in(0,1)$ be fixed.

THEOREM 3.1. *The free boundary problem* (1.1)–(1.8) *has a solution* (u, A) *with a connected toner region* A *such that*

$$\Gamma = \partial A \cap \Omega^+ \in C^{1+\alpha}, \tag{3.1}$$

provided ε *is small enough.*

Suppose that u_ε satisfy (1.1)–(1.4) and (1.6)–(1.8). Set

$$\widetilde{u}_\varepsilon(x,y) = \frac{1}{\varepsilon}\left[u_\varepsilon(\varepsilon x, \varepsilon y) - u_\varepsilon(0,0)\right]. \tag{3.2}$$

Then

$$\Delta\widetilde{u}_\varepsilon = \begin{cases} \varepsilon & \text{for } (x,y)\in \frac{1}{\varepsilon}A_\varepsilon \equiv \widetilde{A}_\varepsilon, \\ 0 & \text{for } (x,y)\notin \widetilde{A}_\varepsilon \cup \{(x,0); |x|\le 1\}\ , \end{cases}$$

where A_ε is the region in which $\Delta u_\varepsilon = 1$.

It has been proved in [3, §8] that for any compact $F \subset R^2$

$$\|\widetilde{u}_\varepsilon - \widetilde{v}\|_{W^{2,p}(F)} \le C_{F,p}\,(\text{meas}(A_\varepsilon) + \varepsilon^p)^{1/p} \tag{3.3}$$

for any $p > 2$, where

$$\widetilde{v}(x,y) = -\frac{\sigma}{2\pi}\int_{-1}^{1} \log\sqrt{(x-\xi)^2 + y^2}d\xi + \frac{My}{b+h} - \frac{\sigma}{\pi}. \tag{3.4}$$

The equation $\frac{\partial \widetilde{v}}{\partial n} = 0$ gives the level curves of the harmonic conjugate $z(x,y)$ of $\widetilde{v}$, where

$$\begin{aligned} z(x,y) = \frac{\sigma}{4\pi}\Big\{ & y\log\frac{(1-x)^2+y^2}{(1+x)^2+y^2} - 2(1-x)\arctan\frac{1-x}{y} \\ & + 2(1+x)\arctan\frac{1+x}{y}\Big\} - \frac{M}{b+h}x. \end{aligned} \tag{3.5}$$

There is only one of these curves, $y = \varphi_0(x)$, that passes through y axis. $\varphi_0(x)$ is analytic and it is shown in [3] that

$$\begin{aligned} &\varphi_0(0) = y_0, \qquad \varphi_0'(0) = 0, \\ &\varphi_0'(x) < 0 \quad \text{for } 0 < x \le \bar{x}_0, \\ &\varphi_0''(x) < 0 \quad \text{for } 0 \le x \le \bar{x}_0, \end{aligned} \tag{3.6}$$

where

$$y_0 = \cot\left(\frac{\pi}{\sigma}\frac{M}{b+h}\right), \qquad \bar{x}_0 = \frac{b+h}{M}\frac{\sigma}{2} > 1. \tag{3.7}$$

It is clear that

$$\frac{\partial \widetilde{v}}{\partial n} = 0 \quad \text{on } y = \varphi_0(x). \tag{3.8}$$

We set

$$\begin{aligned} X = \Big\{ & \gamma(x);\quad \gamma \in C^{1+\alpha}[0, \bar{x}_0 + l],\quad \gamma'(0) = 0, \quad \gamma(\bar{x}_0 + l) \le 0, \\ & \gamma(x) \ge \sqrt{\mu^2 - (x-1)^2}, \quad \tfrac{1}{2}y_0 \le \gamma(x) \le \tfrac{3}{2}y_0 \quad \text{for } 0 \le x \le \mu, \\ & \gamma(x) \ge c_0 > 0 \quad \text{for } \mu \le x \le 1, \qquad \|\gamma\|_{C^{1+\alpha}[0,\bar{x}_0+l]} \le K\Big\}, \end{aligned} \tag{3.9}$$

where l, μ, and c_0 are fixed small positive constants, and K is to be determined later.

For each $\gamma \in X$, we extend it to an even function on the interval $[-\bar{x}_0 - l, \bar{x}_0 + l]$ by letting $\gamma(-x) = \gamma(x)$. Denote by $\widetilde{A}_\gamma$ the connected component of the area enclosed by γ and x axis which contains $\{(0,y),\ 0 < y < \gamma(0)\}$. Next, define w by

$$\Delta w = \chi_{\widetilde{A}_\gamma} \quad \text{in } R_{2N}, \tag{3.10}$$

$$w = 0 \quad \text{on } \partial R_{2N}, \tag{3.11}$$

where $R_{2N} = \{|x| < 2N,\ |y| < 2N\}$ $(N/2 > 3y_0/2)$ and $\chi_{\widetilde{A}_\gamma}$ is the characteristic function of $\widetilde{A}_\gamma$.

By $W^{2,p}$ estimates and Sobolev's embedding theorem,

$$\|w\|_{C^{1+\alpha}(R_{2N})} \leq C. \tag{3.12}$$

Here and below we shall use C to denote constants independent of K and C_K to denote the constants depending on K.

Let $V = ([0,1] \times [c_0, N]) \cup \left(([1, \bar{x}_0 + l] \times [0, N]) \setminus \{(x-1)^2 + y^2 < \mu^2\}\right)$. Then by Theorem 2.1,

$$\|w\|_{W^{2,\infty}(R_N \cap V)} \leq C_K. \tag{3.13}$$

Now let u_γ be the solution of (1.1)–(1.4) and (1.6)–(1.8) with $A = \varepsilon \widetilde{A}_\gamma$, and let

$$\widetilde{u}_{\gamma,\varepsilon}(x,y) = \frac{1}{\varepsilon}\left[u_\gamma(\varepsilon x, \varepsilon y) - u_\varepsilon(0,0)\right],$$

and observe that

$$\Delta\left(\widetilde{u}_{\gamma,\varepsilon} - \widetilde{v} - \varepsilon w\right) = 0 \quad \text{in } R_{2N}. \tag{3.14}$$

Therefore, by (3.12) and (3.3) (with $\widetilde{u}_\varepsilon = \widetilde{u}_{\gamma,\varepsilon}$ in (3.3)), we can apply the Schauder's interior estimates to obtain

$$\|\widetilde{u}_{\gamma,\varepsilon} - \widetilde{v} - \varepsilon w\|_{C^3(R_N)} \leq C\varepsilon + C\varepsilon^{2/p}, \tag{3.15}$$

where the constant is independent of K.

By symmetry,

$$\frac{\partial}{\partial x}\left(\widetilde{u}_{\gamma,\varepsilon} - \widetilde{v} - \varepsilon w\right) = 0 \quad \text{on } x = 0, \tag{3.16}$$

and so by (3.15),

$$\frac{\partial}{\partial x}\left(\widetilde{u}_{\gamma,\varepsilon} - \widetilde{v} - \varepsilon w\right) \leq (C\varepsilon + C\varepsilon^{2/p})x \quad \text{for } (x,y) \in R_N. \tag{3.17}$$

Using (3.13) and $\frac{\partial w}{\partial x}(0,y) = 0$ (by symmetry), we get

$$\frac{\partial w}{\partial x} = \frac{\partial w}{\partial x}(x,y) - \frac{\partial w}{\partial x}(0,y) \leq C_K x \quad \text{for } (x,y) \in V. \tag{3.18}$$

Clearly,

$$\widetilde{v}_x = \frac{\sigma}{4\pi}\log\frac{(1-x)^2 + y^2}{(1+x)^2 + y^2} \leq -c_1 x < 0 \quad \text{for } (x,y) \in V. \tag{3.19}$$

It follows that

$$\frac{\partial}{\partial x}\widetilde{u}_{\gamma,\varepsilon} \leq \left(C\varepsilon + C\varepsilon^{2/p} + C_K\varepsilon - c_1\right)x \leq -\frac{1}{2}c_1 x \quad \text{in } V, \tag{3.20}$$

provided $0 < \varepsilon \leq \varepsilon_K$ and ε_K is small enough (depending on K).

Write

$$\widetilde{u}_{\gamma,\varepsilon} \equiv \widetilde{v} + \varepsilon w + \varepsilon^{2/p}\widetilde{w}, \tag{3.21}$$

where $\|\widetilde{w}\|_{C^3(R_N)} \le C$ by (3.15), and

$$\|w_x(\cdot,\gamma(\cdot))\|_{C^{1+\alpha}} \le C_K, \qquad \|w_y(x,\gamma(x))\|_{C^{1+\alpha}} \le C_K,$$

by Theorem 2.1.

Let $w^* = \varepsilon^{1-1/p}w + \varepsilon^{1/p}\widetilde{w}$; then

$$\widetilde{u}_{\gamma,\varepsilon} \equiv \widetilde{v} + \varepsilon^{1/p}w^*. \tag{3.22}$$

If $0 < \varepsilon < \varepsilon_K$ and ε_K is small enough, then

$$\|\zeta_1\|_{C^{1+\alpha}} \le 1, \qquad \|\zeta_2\|_{C^{1+\alpha}} \le 1, \tag{3.23}$$

where $\zeta_1(x) = w^*_x(x,\gamma(x))$ and $\zeta_2(x) = w^*_y(x,\gamma(x))$.

Differentiating (3.19) in x, we obtain

$$\widetilde{v}_{yy} = -\widetilde{v}_{xx} = \frac{\sigma}{4\pi}\left(\frac{2(1-x)}{(1-x)^2+y^2} + \frac{2(1+x)}{(1+x)^2+y^2}\right). \tag{3.24}$$

Therefore,

$$\widetilde{v}_{yy}(0,y) = \frac{\sigma}{\pi}\frac{1}{y^2}. \tag{3.25}$$

It follows that (using also (3.22) and (3.13)),

$$\frac{\partial^2\widetilde{u}_{\gamma,\varepsilon}}{\partial y^2}(0,y) \ge \frac{1}{2}\frac{\sigma}{\pi}\frac{1}{y^2} > 0 \quad \text{in } V, \tag{3.26}$$

provided ε_K is small enough.

Therefore, the equation

$$\frac{\partial\widetilde{u}_{\gamma,\varepsilon}}{\partial y}(0,y) = 0, \qquad \frac{1}{2}y_0 \le y \le \frac{3}{2}y_0 \tag{3.27}$$

has a unique solution $y = y_\gamma$, provided ε_K is small enough. It is obvious that y_γ depends continuously on the $C^{1+\alpha}$ norm of γ.

Now we take $\tau > 0$ and define $\eta = T\gamma$ to be the solution of the ordinary differential equation (ODE)

$$\eta(0) = y_\gamma, \tag{3.28}$$

$$\eta'(x) = \frac{\dfrac{\widetilde{v}_y(x,\eta(x)) - \widetilde{v}_y(x,y_\gamma)}{x+\tau} + \varepsilon^{1/p}\dfrac{\zeta_2(x)-\zeta_2(0)}{x+\tau}}{\dfrac{\widetilde{v}_x(x,\eta(x))}{x} + \varepsilon^{1/p}\left(\dfrac{\zeta_1(x)}{x}\right) * \rho_\tau(x)}, \tag{3.29}$$

where $\rho_\tau(x)$ is a C^∞ mollifier (see [5, Chap. 7]) and the convolution is defined after extending the definition of $\zeta_1(x)/x$ to be $\zeta_1(\overline{x}_0 + l)/(\overline{x}_0 + l)$ for $x > \overline{x}_0 + l$ and $\lim_{x\to 0+}(\zeta_1(x)/x)$ for $x < 0$.

We write the right-hand side of (3.29) as $f_1(x,y)/f_2(x,y)$, where $y = \eta(x)$. Then by analyticity of $\widetilde{v}$, noticing also that $\widetilde{v}_{xy}(0,y) \equiv 0$, we get

$$\left\|\frac{\partial f_2}{\partial y}\right\|_{L^\infty(V)} = \left\|\frac{1}{x}\widetilde{v}_{xy}\right\|_{L^\infty(V)} \le C. \tag{3.30}$$

LEMMA 3.2.

$$\left[\frac{\zeta_1(x)}{x}\right]_{C^\alpha[0,\bar{x}_0+l]} \le 3. \tag{3.31}$$

Proof. Suppose that $0 < x_1 < x_2$. If $x_2 - x_1 \ge x_1$. Then $x_2 = x_2 - x_1 + x_1 \le 2(x_2 - x_1)$. So by (3.23),

$$\begin{aligned}\left|\frac{\zeta_1(x_1)}{x_1} - \frac{\zeta_1(x_2)}{x_2}\right| &\le \left|\frac{\zeta_1(x_1)}{x_1} - \zeta_1'(0)\right| + \left|\frac{\zeta_1(x_2)}{x_2} - \zeta_1'(0)\right| \\ &\le {x_1}^\alpha + {x_2}^\alpha \ \le\ (1+2^\alpha)|x_1 - x_2|^\alpha.\end{aligned}$$

Now we consider the case $x_2 - x_1 < x_1$. Clearly,

$$\zeta_1(x_2) = \zeta_1(x_1) + \zeta_1'(x_1)(x_2 - x_1) + r, \qquad |r| \le |x_2 - x_1|^{1+\alpha}.$$

Therefore,

$$\begin{aligned}\left|\frac{\zeta_1(x_1)}{x_1} - \frac{\zeta_1(x_2)}{x_2}\right| &\le \left|\frac{(x_1 - x_2)\zeta_1(x_1) + x_1\zeta_1'(x_1)(x_2 - x_1)}{x_1x_2}\right| + \frac{|x_2 - x_1|^{1+\alpha}}{x_2} \\ &\le \frac{|x_1 - x_2|}{x_1x_2}|\zeta_1(x_1) - x_1\zeta_1'(x_1)| + \frac{|x_2 - x_1|^{1+\alpha}}{x_2} \\ &\le \frac{|x_1 - x_2|}{x_1x_2}|x_1|^{1+\alpha} + \frac{|x_2 - x_1|^{1+\alpha}}{x_2} \qquad (\text{since } \zeta_1(0) = 0) \\ &\le 2|x_1 - x_2|^\alpha.\end{aligned}$$

□

Since $\widetilde{v}_x(0,y) = 0$, Lemma 3.2 and the analyticity of $\widetilde{v}$ imply that

$$\sup_V |f_2(x,y)| + \sup_{x_1 \ne x_2;\ (x_1,y),(x_2,y)\in V} \frac{|f_2(x_1,y) - f_2(x_2,y)|}{|x_1 - x_2|^\alpha} \le C. \tag{3.32}$$

By (3.19), we have

$$f_2(x,y) \le -\tfrac{1}{2}c_1 < 0, \tag{3.33}$$

provided ε_K is small enough.

Using the analyticity of $\widetilde{v}$, $\widetilde{v}_{yx}(0,y) \equiv 0$, $\zeta_2'(0) = 0$ and an argument similar to Lemma 3.2 for $(\zeta_2(x) - \zeta_2(0))/(x + \tau)$, we conclude

$$\sup_{x\ne y;\ (x_1,y),(x_2,y)\in V} \frac{|f_1(x_1,y) - f_1(x_2,y)|}{|x_1 - x_2|^\alpha} \le C, \tag{3.34}$$

and

$$\sup_{0\le t\le x} |f_1(t,\eta(t))| \le C \sup_{0<t\le x} \frac{|\eta(t) - \eta(0)|}{t}, \tag{3.35}$$

where the constant C is independent of τ, and also

$$\left|\frac{\partial f_1}{\partial y}\right| = \left|\frac{\widetilde{v}_{yy}(x,y)}{x+\tau}\right| \le \frac{C}{x+\tau}. \tag{3.36}$$

Therefore, for each $\tau > 0$, the ODE (3.28), (3.29) is uniquely solvable. It is obvious that for each $\tau > 0$, $\eta \in C^{1+\alpha}$. Next, we shall derive $C^{1+\alpha}$ estimates independent of τ. From now on we shall assume that

$$\text{for the solution } \eta(x), \text{ we have } (x, \eta(x)) \in V. \tag{3.37}$$

What we need to prove is actually the estimates near $x = 0$.

LEMMA 3.3.

$$\lim_{x\to 0} \frac{-x\widetilde{v}_{yy}}{\widetilde{v}_x} = 1. \tag{3.38}$$

Proof.

$$\lim_{x\to 0} \frac{-x\widetilde{v}_{yy}}{\widetilde{v}_x} = \lim_{x\to 0} \frac{-\widetilde{v}_{yy}}{\widetilde{v}_x/x} = \lim_{x\to 0} \frac{-\widetilde{v}_{yy}}{\widetilde{v}_{xx}} = 1.$$ □

The convergence is uniform by the analyticity of $\widetilde{v}$. Thus there exists some small $\mu > 0$ such that

$$\frac{-x\widetilde{v}_{yy}}{\widetilde{v}_x} > 0 \quad \text{for } 0 \le x \le \mu,\ \frac{1}{2}y_0 \le y \le \frac{3}{2}y_0,$$

which implies

$$\frac{(f_1)_y(x,y)}{f_2(x,\eta(x))} = \frac{\dfrac{\widetilde{v}_{yy}(x,y)}{x+\tau}}{f_2(x,\eta(x))} < 0 \quad \text{for } 0 \le x \le \mu,\ \frac{1}{2}y_0 \le y \le \frac{3}{2}y_0. \tag{3.39}$$

Let $\lambda(x) = \eta(x) - \eta(0)$; then

$$\begin{aligned} \lambda'(x) &= \frac{f_1(x,\eta(x))}{f_2(x,\eta(x))} \\ &= \frac{f_1(x,\eta(x)) - f_1(x,\eta(0)) + \varepsilon^{1/p}\dfrac{\zeta_2(x)-\zeta_2(0)}{x+\tau}}{f_2(x,\eta(x))} \\ &\equiv q_1(x)\lambda(x) + q_2(x)\ , \end{aligned} \tag{3.40}$$

where $q_1(x) = (f_1)_y(x,y)/f_2(x,\eta(x)) < 0$ by (3.39), and it is also clear that $q_1(x) \ge -\frac{C}{x+\tau}$.

By the definition of ζ_2 we have $\zeta_2'(0) = 0$; therefore,

$$|q_2(x)| \le C\varepsilon^{1/p}\left|\frac{\zeta_2(x)-\zeta_2(0)}{x+\tau}\right| \le C^* x^\alpha, \tag{3.41}$$

where the constant is independent of K, τ. It follows by comparison (we use $q_1(x) < 0$ here) that

$$|\lambda(x)| \le \int_0^x |q_2(\xi)|d\xi \le C^* x^{1+\alpha}. \tag{3.42}$$

This inequality, together with (3.35), implies that

$$|\eta'(x)| = |\lambda'(x)| \le C^* x^\alpha. \tag{3.43}$$

It follows that

$$\|\eta'\|_{L^\infty[0,\bar{x}_0+l]} \le C^*. \tag{3.44}$$

Next, take $a > 0$, and consider $\widetilde{\lambda}(x) = \eta(x+a) - \eta(x)$ for $x > 0$. Clearly,

$$\begin{aligned}
&\widetilde{\lambda}'(x) \\
&\quad = \frac{f_1(x+a,\eta(x+a))}{f_2(x+a,\eta(x+a))} - \frac{f_1(x,\eta(x))}{f_2(x,\eta(x))} \\
&\quad = \frac{f_1(x+a,\eta(x+a)) - f_1(x+a,\eta(x)) + f_1(x+a,\eta(x)) - f_1(x,\eta(x))}{f_2(x+a,\eta(x+a))} \\
&\qquad + \frac{f_1(x,\eta(x))[f_2(x,\eta(x)) - f_2(x+a,\eta(x+a))]}{f_2(x,\eta(x))f_2(x+a,\eta(x+a))} \\
&\quad \equiv I_1 + I_2.
\end{aligned} \tag{3.45}$$

We estimate I_i, $i = 1,2$ separately. First by (3.32), (3.33), (3.35), and (3.44),

$$|I_2| \le C^* a^\alpha. \tag{3.46}$$

Next, letting

$$I_1 \equiv J_1 + J_2, \tag{3.47}$$

where by (3.34), (3.33),

$$|J_2| = \left| \frac{f_1(x+a,\eta(x)) - f_1(x,\eta(x))}{f_2(x+a,\eta(x+a))} \right| \le C^* a^\alpha. \tag{3.48}$$

Moreover,

$$\begin{aligned}
J_1 &= \frac{f_1(x+a,\eta(x+a)) - f_1(x+a,\eta(x))}{f_2(x+a,\eta(x+a))} \\
&\equiv q(x+a)(\eta(x+a) - \eta(x)) \ ,
\end{aligned} \tag{3.49}$$

where by (3.33), (3.36),

$$|q(x+a)| \le \frac{C^*}{x+a+\tau}. \tag{3.50}$$

By (3.43),

$$|\eta(x+a) - \eta(x)| \le \int_x^{x+a} |\eta'(\xi)| d\xi \le C^*(x+a)^\alpha a \le C^*(x+a)a^\alpha. \tag{3.51}$$

Therefore, (3.49)–(3.51) imply

$$|J_1| \le C^* a^\alpha. \tag{3.52}$$

Combining (3.46)–(3.52), we obtain

$$|\eta'(x+a) - \eta'(x)| = |\widetilde{\lambda}'(x)| \le C^* a^\alpha. \tag{3.53}$$

This proves that if $\varepsilon \in (0, \varepsilon_K)$, then

$$\|\eta\|_{C^{1+\alpha}[0,\bar{x}_0+l]} \le C^{**}, \tag{3.54}$$

where the constant C^{**} is independent of K and τ.

We now choose $K = C^{**}$. Since the curve $\eta = T\gamma$ goes to $y = \varphi_0(x)$ uniformly as $\varepsilon \to 0$, it is easy to choose the remaining constants in the definition of X (as in [3]) so that T maps X into itself.

Obviously T is a continuous map from X to X (using $C^{1+\alpha}$ norm topology). For each $\tau > 0$, the image of T is contained in a bounded set of $C^{1+\beta}$ for any $\beta \in (\alpha, 1)$. Therefore, the closure of the image of T is compact. Hence T has a fixed point, by Schauder's fixed point theorem (see [5]). It is obvious that all the estimates of this section apply to the fixed point $\eta = T\eta$, where the estimates are independent of τ; therefore, we can take a subsequence and pass the limit as $\tau \to 0$ to obtain a solution (u, A) with one connected toner region. This proves Theorem 3.1.

Acknowledgment. We would like to thank the referee for many helpful comments. The referee assisted in revising the exposition. This work was partially carried out while both authors were visiting Institute for Mathematics and Applications at the University of Minnesota. For its hospitality, the authors would like to express their gratitude.

REFERENCES

[1] L. A. Caffarelli and A. Friedman, *The free boundary in the Thomas–Fermi atomic model*, J. Differential Equations, 32 (1979), pp. 335–356.

[2] A. Friedman, *Mathematics in Industrial Problems, Part* 2, IMA Math. Appl. Vol. 24, Springer-Verlag, New York, 1989.

[3] A. Friedman and Bei Hu, *A free boundary problem arising in electrophotography*, J. Nonlinear Anal., TMA, 16 (1991), pp. 729–758.

[4] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems,* Princeton University Press, Princeton, NJ, 1983.

[5] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equation of Second Order,* Second ed., Springer-Verlag, New York, 1983.

[6] V. Barbu and S. Stojanovic, *A variational approach to a free boundary problem arising in electrophotography*, IMA preprint series # 815, May 1991.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1455-1481, November 1992

CONSTRUCTION OF APPROXIMATE INERTIAL MANIFOLDS USING WAVELETS*

OLIVIER GOUBET†

Abstract. In this work, approximate inertial manifolds are constructed for a class of dissipative evolution equations. The innovation is that these manifolds are defined as graphs on orthonormal wavelet bases.

Key words. dissipative nonlinear parabolic equations, approximate inertial manifolds, wavelets

AMS(MOS) subject classifications. 35A35, 35K60, 42C15

1. Introduction. Our aim in this article is to construct approximations of inertial manifolds for dissipative evolution equations by utilization of wavelets. In this way we make the connection between two recent theories: the theory of inertial manifolds that has emerged from the study of dynamical systems and the theory of orthonormal wavelet bases.

First let us have an overview of the inertial manifolds theory that relates to the large time study of dissipative evolution equations.

Let us consider a nonlinear PDE that is dissipative; it means that there exists a *global attractor* for the associated dynamical system, i.e., a compact set that is invariant by the flow of the solutions, and that attracts all the orbits when $t \to +\infty$. Nevertheless, the convergence of the orbits towards the attractor can be arbitrarily slow, and this one can have a complex structure, and even be a fractal (see [18]).

New mathematical tools have been introduced by Foias, Sell, and Temam [7]: the inertial manifolds (IM), which are smooth finite-dimensional manifolds, positively invariant by the flow of the solutions, and which attract all the trajectories with exponential speed. From the physics point of view, IMs model the interaction laws between small and large structures of a turbulent flow, and represent its permanent regime; actually on an IM small eddies are slaved by large ones, and there are similar results, after a transient time, for a trajectory that is not on the IM (see [5], [7]). Nevertheless, until now the existence of IM necessitates a very restrictive property, the spectral gap condition (see [18]).

Hence came the approximate inertial manifolds (AIM). They are smooth finite-dimensional manifolds that attract all the orbits into a thin neighborhood, in a finite time, and with exponential speed. AIMs provide good substitutes for IMs when no existence result for IM is available; moreover, because their equations are rather simple, they make the implementation of numerical algorithms easier (see [3], [4], [5], [19]).

The theory of AIMs, which first developed in the spectral case, has begun to extend beyond. Some nonlinear algorithms have been established for finite elements (see [14]) and finite differences (see [20]). The purpose of this paper, following a suggestion by Temam, is to construct approximate inertial manifolds using the newly developed concept of orthonormal bases of wavelets.

The improvement featured by wavelets with respect to spectral bases (here the trigonometrical system), is to combine good localization properties in space variable,

* Received by the editors March 18, 1991; accepted for publication (in revised form) March 23, 1992. This work was performed while the author was visiting the Institute for Applied Mathematics and Scientific Computing at Indiana University, Bloomington. This research was supported in part by National Science Foundation grant NSF-DMS8802596 and by the Research Fund of Indiana University.

† Laboratoire d'Analyse Numérique, Université de Paris-Sud, 91405 Orsay Cedex, France.

and good localization properties in frequencies (see [13]). Hence by constructing AIMs using wavelets we hope to obtain laws for the slaving of small eddies and high frequencies by large eddies and low frequencies. In this paper we are interested in the space periodic case, and we consider the periodic version of orthonormal wavelet bases of Daubechies, Lemarié, and Meyer (see [2], [11], [12]) as described in [13]. However, wavelets are flexible tools that can be adapted to other domains than the n-dimensional torus (see [10]), and allow us to consider the construction of AIMs in general domains where little information on the spectral bases is available.

Nevertheless, in this paper, for the sake of clarity, we will focus particularly on the one-dimensional spline wavelet bases of [11], referring the reader to the Appendix for multidimensional results and for other wavelet bases.

The paper is organized as follows. In § 2 we briefly recall some results about spline wavelets in the one-dimensional periodic case. In § 3 we give, for the sake of completeness, a proof of a result announced in [13] saying that, if spline wavelets are sufficiently regular, then they provide an unconditional basis for periodic Sobolev spaces. In § 4, we describe a class of nonlinear parabolic PDEs, that includes, for instance, the two-dimensional Navier–Stokes equations, written in the stream function formulation (in order to avoid the difficult problem of the treatment by wavelets of the incompressibility condition), and the Kuramoto–Sivashinsky equation. In § 5 we show how the wavelet expansion of a function can be used to construct several AIMs for the class of evolution equation described above. The method follows [21]; first we define the *induced trajectories*, tools that allow us to estimate the distance, for different topologies, between the orbits and the space spanned by the first k wavelets, ordered in the natural way. These estimates, which hold for large time, are then compared to the ones obtained in the spectral case, for spaces that have the same dimension. The result is that the wavelets provide a flat AIM that provide the same order of approximation as the one obtained using spectral bases. Then we give two examples of nontrivial (nonflat) AIMs that approximate the attractor with higher order than the flat one; once again we match the accuracy obtained in the spectral case. Finally in the Appendix we extend the results of § 3 to the constructions of [2] and of [12] and to the multidimensional case.

Notation. Let $\mathscr{Z}$ (respectively, $\mathscr{R}$, $\mathscr{C}$) be the set of integers (respectively, of real numbers, of complex numbers).

Let $\Pi=\mathscr{R}/\mathscr{Z}$ be the one-dimensional torus. We denote by $C^N(\Pi)$ the space of N times continuously differentiable functions on Π and by $H^s(\Pi)$ the usual periodic Sobolev space.

We denote by $\dot{H}^s(\Pi)$ the space of functions u in $H^s(\Pi)$ such that

$$\int_\Pi u(x)\,dx=0.$$

$\dot{H}^s(\Pi)$ is a Hilbert space when endowed with the scalar product

$$(u,v)_s=\sum_{k\in\mathscr{Z}}|k|^{2s}\hat{u}(k)\overline{\hat{v}(k)},$$

where

$$\hat{u}(k)=\int_\Pi u(x)\,e^{-2i\pi kx}\,dx.$$

Let $|u|_s$ be the corresponding norm,

$$|u|_s=(u,u)_s^{1/2}.$$

When $s=0$ we write $H^0(\Pi)=L^2(\Pi)$.

In the following we will denote by C a constant that only depends on the regularity N of the wavelet, and in § 5 on the data of the equation.

2. Spline wavelet bases of $\dot{L}^2(\Pi)$. We consider the finite-dimensional space $V_j = \{v \in C^N(\Pi);\ v$ is a piecewise polynomial function of degree less than or equal to $N+1$, with nodes at $k/2^j;\ 0 \leqq k < 2^j\}$. Then we have the embeddings

$$V_0 \subset \cdots \subset V_j \subset V_{j+1} \subset \cdots L^2(\Pi).$$

We define

$$W_j = V_{j+1} \cap (V_j)^\perp, \tag{2.1}$$

then we have

$$\dot{L}^2(\Pi) = \bigoplus_{j=0}^{+\infty} W_j, \tag{2.2}$$

the sum being orthogonal.

Let us introduce what are the periodic wavelet bases associated to the W_j's. We first recall the original construction on $\mathscr{R}$; from [11] (see also [1], [13]) we know that, for each integer N, there exists a function ψ_N satisfying

(i)

$$\psi_N \in C^N(\mathscr{R}), \tag{2.3}$$

ψ_N being a piecewise polynomial function of degree less than or equal to $N+1$ with nodes at the half integers.

(ii)

$$\exists \varepsilon_N > 0/ \quad \text{for } m \leqq N+1, \qquad \left| \frac{\partial^m}{\partial x^m} \psi_N(x) \right| \leqq C e^{-\varepsilon_N |x|}. \tag{2.4}$$

(iii) If $m \leqq N+1$

$$\int_{\mathscr{R}} x^m \psi_N(x)\, dx = 0. \tag{2.5}$$

The wavelets, that are derived from ψ_N by a translation and a dilation as below, satisfy

(iv) The family $\{2^{j/2}\psi_N(2^j x - k)\}_{j,k \in \mathscr{Z}}$ is an orthonormal basis of

$$L^2(\mathscr{R}). \tag{2.6}$$

Remark 1. Formula (2.4) shows the exponential decay of the wavelet $2^{j/2}\psi_N(2^j x - k)$ for x away from $(k+\frac{1}{2})/2^j$. Formulae (2.3) and (2.5) describe the localization in frequencies of the wavelet $2^{j/2}\psi_N(2^j x - k)$ around an annulus of radii $C_1 2^j, C_2 2^j$; actually (2.5) means that $|\hat{\psi}_N(\xi)| = O(|\xi|^{N+1})$ when $|\xi| \to 0$, and $|\hat{\psi}_N(\xi)| = O(|\xi|^{-N-1})$ when $|\xi| \to \infty$.

Remark 2. Throughout this paper we shall omit the subscript N on ψ_N.

Then following [13] we define the periodic wavelets as

$$\psi_{j,k}(x) = 2^{j/2} \sum_{l \in \mathscr{Z}} \psi(2^j x + 2^j l - k). \tag{2.7}$$

This periodization transfers to periodic wavelets the localization in frequencies (see Lemma 2 below), and does not deteriorate the localization in space variable too much. Then we have

$$\text{The family } \{\psi_{j,k}\}_{1 \leqq k \leqq 2^j} \text{ is an orthonormal basis of } W_j, \tag{2.8}$$

$$\text{The family } \{\psi_{j,k}\}_{0 \leqq j \leqq +\infty; 1 \leqq k \leqq 2^j} \text{ is an orthonormal basis of } \dot{L}^2(\Pi). \tag{2.9}$$

3. Preliminary results.

3.1. Bernstein inequalities.

PROPOSITION 1. *There exists $C > 0$ such that for any v in V_j*

$$|v|_{N+1} \leqq C 2^{j(N+1)} |v|_0. \tag{3.1}$$

Proof. Let b_N be the Nth fundamental B-spline, defined from the characteristic function χ of $[-\frac{1}{2}, \frac{1}{2}]$ by

$$b_N = \underbrace{\chi * \cdots\cdots * \chi}_{N+1 \text{ times}}. \tag{3.2}$$

For v in V_j,

$$v = \sum_{k=1}^{2^j} \alpha_{j,k} 2^{j/2} b_N(2^j x - k). \tag{3.3}$$

It is well known that

$$c_1 \left(\sum_{k=1}^{2^j} |\alpha_{j,k}|^2 \right) \leqq |v|_0^2 \leqq c_2 \left(\sum_{k=1}^{2^j} |\alpha_{j,k}|^2 \right), \tag{3.4}$$

for some constants c_1, c_2 depending on N.

On the other hand, we have

$$|v|_{N+1}^2 = \sum_{1 \leqq k, p \leqq 2^j} 4^{j(N+1)} \alpha_{j,k} \alpha_{j,p} m_{j,k,p}, \tag{3.5}$$

where

$$m_{j,k,p} = 2^j \int_\Pi \frac{\partial^{N+1}}{\partial x^{N+1}} b_N(2^j x - k) \frac{\partial^{N+1}}{\partial x^{N+1}} b_N(2^j x - p)\, dx. \tag{3.6}$$

We observe that either $m_{j,k,p} = 0$ if $|k - p| > N + 1$ or

$$|m_{j,k,p}| \leqq \left\| \frac{\partial^{N+1}}{\partial x^{N+1}} b_N \right\|_{L^1(\mathscr{R})} \left\| \frac{\partial^{N+1}}{\partial x^{N+1}} b_N \right\|_{L^\infty(\mathscr{R})}. \tag{3.7}$$

It follows

$$|v|_{N+1}^2 \leqq C 4^{j(N+1)} \sum_{|k-p| \leqq N+1} |\alpha_{j,k}| |\alpha_{j,p}|. \tag{3.8}$$

Using

$$\sum_{|k-p| \leqq N+1} |\alpha_{j,k}| |\alpha_{j,p}| \leqq C \sum_{k=1}^{2^j} |\alpha_{j,k}|^2, \tag{3.9}$$

we infer the result from (3.4) and (3.8).

Remark 3. We recall that C is a constant that depends on N, but which is independent of j.

LEMMA 1. *Let r, s, t be real numbers, $r \leqq s \leqq t$. Then for any u in $\dot{H}^t(\Pi)$*

$$|u|_s \leqq |u|_r^{(t-s)/(t-r)} |u|_t^{(s-r)/(t-r)}. \tag{3.10}$$

Proof. This is just a particular case of the interpolation inequality (see [18] and the references therein).

COROLLARY 1. *Let s belong to $[0, N+1]$; then for any v in V_j*

$$|v|_s \leqq C2^{js}|v|_0. \tag{3.11}$$

Proof. Thanks to (3.10),

$$|v|_s \leqq |v|_0^{1-s/(N+1)}|v|_{N+1}^{s/(N+1)}. \tag{3.12}$$

We infer from (3.1) and (3.12)

$$|v|_s \leqq C^{s/(N+1)}2^{js}|v|_0. \tag{3.13}$$

COROLLARY 2. *Let s belong to $[-N-1, 0]$; then for any v in V_j*

$$|v|_0 \leqq C2^{-js}|v|_s. \tag{3.14}$$

Proof. Thanks to (3.10), we have

$$|v|_0 \leqq |v|_s^{(N+1)/(N+1-s)}|v|_{N+1}^{-s/(N+1-s)}. \tag{3.15}$$

Using (3.1)

$$|v|_0 \leqq C^{-s/(N+1-s)}2^{-js(N+1)/(N+1-s)}|v|_0^{-s/(N+1-s)}|v|_s^{(N+1)/(N+1-s)}. \tag{3.16}$$

To conclude we take the $\{(N+1-s)/(N+1)\}$th power of inequality (3.16).

3.2. Poincaré inequalities.

LEMMA 2. *Let f be in $L^1(R)$. Let $g(x)=\sum_{l\in\mathscr{Z}} f(x+l)\in L^1(\Pi)$. Then*

$$\hat{g}(k)=\hat{f}(k). \tag{3.17}$$

(*The left-hand side of the equality represents the* kth *Fourier coefficient of g, the right-hand side denotes the value at the point k of the Fourier transform of f*).

Proof. Thanks to Fubini's theorem,

$$\int_\Pi g(x)\, e^{-2i\pi kx}\, dx = \sum_{l\in\mathscr{Z}} \int_l^{l+1} f(x)\, e^{-2i\pi(k-l)x}\, dx \tag{3.18}$$

provides the result.

PROPOSITION 2. *There exists $C>0$ such that for any w in W_j*

$$|w|_{-N-1} \leqq C2^{-j(N+1)}|w|_0. \tag{3.19}$$

Proof. Let $w\in W_j$; we write

$$w=\sum_{k=1}^{2^j} \alpha_{j,k}\psi_{j,k}. \tag{3.20}$$

We easily infer from (2.7) and (3.17)

$$\hat{\psi}_{j,k}(l)=\frac{1}{2^{j/2}}\,\hat{\psi}\left(\frac{l}{2^j}\right)\exp\left(-2i\pi\frac{kl}{2^j}\right). \tag{3.21}$$

We set

$$m\left(\frac{l}{2^j}\right)=2^{-j/2}\sum_{k=1}^{2^j}\alpha_{j,k}\exp\left(-2i\pi\frac{kl}{2^j}\right). \tag{3.22}$$

This yields

$$\hat{w}(l)=m\left(\frac{l}{2^j}\right)\hat{\psi}\left(\frac{l}{2^j}\right). \tag{3.23}$$

Then, thanks to Parseval's identity,

$$|w|_0^2=\sum_{l\in\mathscr{Z}}|\hat{w}(l)|^2, \tag{3.24}$$

we obtain

$$|w|_0^2=\sum_{l\in\mathscr{Z}}\left|m\left(\frac{l}{2^j}\right)\right|^2\left|\hat{\psi}\left(\frac{l}{2^j}\right)\right|^2, \tag{3.25}$$

we write

$$|w|_0^2=\sum_{k=1-2^{j-1}}^{2^{j-1}}\sum_{l\in\mathscr{Z}}\left|m\left(\frac{k}{2^j}+l\right)\right|^2\left|\hat{\psi}\left(\frac{k}{2^j}+l\right)\right|^2. \tag{3.26}$$

Observing that m is a one-periodic function, we obtain

$$|w|_0^2=\sum_{k=1-2^{j-1}}^{2^{j-1}}\left|m\left(\frac{k}{2^j}\right)\right|^2\left(\sum_{l\in\mathscr{Z}}\left|\hat{\psi}\left(\frac{k}{2^j}+l\right)\right|^2\right). \tag{3.27}$$

Now we need the following.

LEMMA 3. *Let f be in $L^1(\mathscr{R})\cap L^2(\mathscr{R})$ such that $|\hat{f}(z)|\leq C(1+|z|)^{-\alpha}$ with $\alpha>\frac{1}{2}$, and such that the family $\{f(x+l)\}_{l\in\mathscr{Z}}$ is an orthonormal family in $L^2(\mathscr{R})$. Then, for each z in Π*

$$\sum_{l\in\mathscr{Z}}|\hat{f}(z+l)|^2=1. \tag{3.28}$$

Proof. We prove that the Fourier coefficients of the one-periodic function

$$1-\sum_{l\in\mathscr{Z}}|\hat{f}(z+l)|^2$$

are equal to zero.

Lemma 2 yields

$$\int_\Pi\left(\sum_{l\in\mathscr{Z}}|\hat{f}(z+l)|^2\right)e^{-2i\pi kz}\,dz=\int_{\mathscr{R}}|\hat{f}(z)|^2\,e^{-2i\pi kz}\,dz. \tag{3.29}$$

The result follows, using Plancherel's theorem

$$\int_{\mathscr{R}}|\hat{f}(z)|^2\,e^{-2i\pi kz}\,dz=\int_{\mathscr{R}}f(x)f(x-k)\,dx. \tag{3.30}$$

Then, we apply Lemma 3 to (3.27) to obtain

$$|w|_0^2=\sum_{k=1-2^{j-1}}^{2^{j-1}}\left|m\left(\frac{k}{2^j}\right)\right|^2. \tag{3.31}$$

On the other hand, by the same computations as above

$$|w|_{-N-1}^2=\frac{1}{4^{j(N+1)}}\sum_{k=1-2^{j-1}}^{2^{j-1}}\left|m\left(\frac{k}{2^j}\right)\right|^2\left(\sum_{l\in\mathscr{Z}}\left(l+\frac{k}{2^j}\right)^{-2N-2}\left|\hat{\psi}\left(l+\frac{k}{2^j}\right)\right|^2\right). \tag{3.32}$$

Using (3.31), we observe that it is sufficient to show that there exists $C>0$ such that

$$\sum_{l\in\mathscr{Z}}\left|\left(l+\frac{k}{2^j}\right)\right|^{-2N-2}\left|\hat{\psi}\left(l+\frac{k}{2^j}\right)\right|^2\leq C<+\infty \tag{3.33}$$

to conclude the proof of Proposition 2.

Actually, we write

$$\sum_{l\in\mathscr{X}}\left|\left(l+\frac{k}{2^j}\right)\right|^{-2N-2}\left|\hat{\psi}\left(l+\frac{k}{2^j}\right)\right|^2 = \sum_{l\in\mathscr{X}^*}\left|\left(l+\frac{k}{2^j}\right)\right|^{-2N-2}\left|\hat{\psi}\left(l+\frac{k}{2^j}\right)\right|^2+\left|\frac{k}{2^j}\right|^{-2N-2}\left|\hat{\psi}\left(\frac{k}{2^j}\right)\right|^2, \tag{3.34}$$

and we majorize independently the two terms involved in the right-hand side of equality (3.34).

We first observe that for $l\in\mathscr{X}^*$, for $|k|\leqq 2^{j-1}$, we have

$$\left|\frac{k}{2^j}+l\right|^{-2N-2}\leqq 4^{N+1}.$$

Then, using Lemma 3, we obtain

$$\sum_{l\in\mathscr{X}^*}\left(\frac{k}{2^j}+l\right)^{-2N-2}\left|\hat{\psi}\left(l+\frac{k}{2^j}\right)\right|^2\leqq 4^{N+1}\sum_{l\in\mathscr{X}^*}\left|\hat{\psi}\left(l+\frac{k}{2^j}\right)\right|^2\leqq 4^{N+1}. \tag{3.35}$$

Now we prove

$$\left(\frac{k}{2^j}\right)^{-2N-2}\left|\hat{\psi}\left(\frac{k}{2^j}\right)\right|^2\leqq C, \tag{3.36}$$

that is a consequence of (2.5) that implies $|\hat{\psi}(\xi)|=O(|\xi|^{N+1})$ when $\xi\to 0$. This fact concludes the proof of Proposition 2.

Then we also have the following.

COROLLARY 3. *Let s belong to* $[-N-1,0]$; *then for any w in* W_j

$$|w|_s\leqq C2^{js}|w|_0. \tag{3.37}$$

COROLLARY 4. *Let s belong to* $[0,N+1]$; *then for any w in* W_j

$$|w|_0\leqq C2^{-js}|w|_s. \tag{3.38}$$

Proof. We prove (3.37) and (3.38) as we established (3.11) and (3.14), using (3.19) instead of (3.1).

3.3. Conclusion. We summarize §§ 3.1 and 3.2 by the following.

PROPOSITION 3. *There exist* $C_1, C_2>0$ *such that, for each s in* $[-N-1,N+1]$, *for any* w_j *in* W_j, *setting*

$$w_j=\sum_{k=1}^{2^j}\gamma_{j,k}\psi_{j,k},$$

we have

$$C_1|w_j|_s\leqq 2^{js}\left(\sum_{k=1}^{2^j}\gamma_{j,k}^2\right)^{1/2}\leqq C_2|w_j|_s. \tag{3.39}$$

We also have

$$|w_j|_0=\left(\sum_{k=1}^{2^j}\gamma_{j,k}^2\right)^{1/2}. \tag{3.40}$$

Proof. First we observe that (3.40) is a rewriting of (2.8). On the other hand, for w_j as above, for $s>0$,

$$2^{js}\left(\sum_{k=1}^{2^j}\gamma_{j,k}^2\right)^{1/2}\leqq C_2|w_j|_s$$

is established using (3.38) and (3.40). The proofs of the other inequalities are similar using Corollaries 1, 2, 3, and 4.

The following theorem includes the main result of this section.

THEOREM 1. *Let s be in $(-N-1, N+1)$; let u be in $\dot{H}^s(\Pi)$. We set*

$$\gamma_{j,k}=\langle u, \psi_{j,k}\rangle_{\dot{H}^s(\Pi),\dot{H}^{-s}(\Pi)}, \qquad w_j=\sum_{k=1}^{2^j}\gamma_{j,k}\psi_{j,k}.$$

Then, $u=\sum_{j=0}^{+\infty} w_j$, where the sum is unconditionally convergent in $\dot{H}^s(\Pi)$.

Actually, there exist $C_1(s)$, $C_2(s)>0$ such that

$$C_1(s)|u|_s^2 \leqq \sum_{j=0}^{+\infty} |w_j|_s^2 \leqq C_2(s)|u|_s^2. \tag{3.41}$$

Proof. We first prove (3.41) for u regular, such that $u=\sum_{j=0}^{+\infty} w_j$ holds in $\dot{L}^2(\Pi)$, for instance.

Then we conclude by noticing that if $u_n \to u$ in $\dot{H}^s(\Pi)$, then, thanks to Proposition 3, $w_{j,n}\to w_j$ in $\dot{H}^s(\Pi)$ ($w_{j,n}$ defined from u_n in the obvious way).

Now we follow step by step the proof of Theorem 8 [13, Chap. 2]. We write

$$|u|_s^2=\sum_{j=0}^{+\infty}|w_j|_s^2+2\sum_{j<l}(w_j, w_l)_s. \tag{3.42}$$

Choosing ε such that $|s|<N-\varepsilon$, we obtain

$$|(w_j, w_l)_s| \leqq |w_j|_{s+\varepsilon}|w_l|_{s-\varepsilon}, \tag{3.43}$$

and thanks to (3.39)

$$|(w_j, w_l)_s| \leqq C2^{-\varepsilon(l-j)}|w_j|_s|w_l|_s. \tag{3.44}$$

Therefore,

$$\begin{aligned}2\sum_{j<l}|(w_j, w_l)_s| &\leqq C\sum_{j=0}^{+\infty}\left(\sum_{l\neq j}2^{-\varepsilon|l-j|}\right)|w_j|_s^2\\ &\leqq \frac{C}{2^\varepsilon-1}\sum_{j=0}^{+\infty}|w_j|_s^2.\end{aligned} \tag{3.45}$$

This yields

$$|u|_s^2 \leqq C(s)\sum_{j=0}^{+\infty}|w_j|_s^2. \tag{3.46}$$

On the other hand, setting

$$u_J=\sum_{j=0}^{J}4^{js}w_j, \tag{3.47}$$

we have

$$(u, u_J)_0=\sum_{j=0}^{J}4^{js}|w_j|_0^2. \tag{3.48}$$

Thanks to (3.39), we obtain

$$\sum_{j=0}^{J}4^{js}|w_j|_0^2 \geqq C\sum_{j=0}^{J}|w_j|_s^2. \tag{3.49}$$

Then we infer from (3.48), (3.49)

$$\sum_{j=0}^{J} |w_j|_s^2 \leqq C|u|_s |u_J|_{-s}. \tag{3.50}$$

Using the first part of the proof, we obtain

$$|u_J|_{-s}^2 \leqq C(s) \sum_{j=0}^{J} |4^{js} w_j|_{-s}^2. \tag{3.51}$$

This yields, using (3.39),

$$|u_J|_{-s}^2 \leqq C(s) \sum_{j=0}^{J} |w_j|_s^2. \tag{3.52}$$

We then infer from (3.50) and (3.52)

$$\left(\sum_{j=0}^{J} |w_j|_s^2\right)^{1/2} \leqq C(s)|u|_s. \tag{3.53}$$

We let $J \to +\infty$ to end the proof.

We summarize Proposition 3 and Theorem 1 by the following.

COROLLARY 5. *For each s in $(-N-1, N+1)$, there exist $C_1(s), C_2(s) > 0$ such that, if*

$$u = \sum_{j=0}^{+\infty} \sum_{k=1}^{2^j} \gamma_{j,k}\psi_{j,k} \in \dot{H}^s(\Pi),$$

then

$$C_1(s)|u|_s^2 \leqq \sum_{j,k} 4^{js}|\gamma_{j,k}|^2 \leqq C_2(s)|u|_s^2. \tag{3.54}$$

In other words, if $|s| < N+1$, the family $\{\psi_{j,k}\}_{0 \leqq j \leqq +\infty; 1 \leqq k \leqq 2^j}$ is an unconditional basis for $\dot{H}^s(\Pi)$. (For equivalent definitions of an unconditional basis in Hilbert spaces, see, for instance, [10]; see also [13] and the references therein.)

4. A class of evolution equations. Let H be a Hilbert space whose elements are periodic functions. Actually we assume that there exists an integer p such that either

$$H = \dot{H}^p(\Pi^n)$$

or H is a closed subspace of $\dot{H}^p(\Pi^n)$ endowed with its natural topology. The class of evolution equations we shall consider has the form

$$\frac{du}{dt} + \nu Au + Ru + B(u, u) = f, \tag{4.1}$$

where $\nu > 0$, A is the unbounded operator $(-\Delta)^m$ acting on H, whose domain is

$$D(A) = \{u \in H;\ Au \in H\},$$

R is a bounded linear operator from $D(A^{1/2})$ into H, B is a bilinear operator from $D(A^{1/2}) \times D(A^{1/2})$ into $D(A^{-1/2})$, f is in H, and the unknown u maps $\mathcal{R}_+$ into H.

We will consider the initial value problem consisting of (4.1) and of initial condition

$$u(0) = u_0 \in H. \tag{4.2}$$

As usual we denote by $|\cdot|$ and by $(\cdot,\cdot)$, respectively, the norm and the scalar product on H. We set $V=D(A^{1/2})$, $\|\cdot\|$ being the norm on V, and $((\cdot,\cdot))$ being the corresponding scalar product.

We assume that there exists $\alpha, C>0$ such that for any $v\in V$ we have

$$(\nu Au+Ru,u)\geqq\alpha\|u\|^2, \tag{4.3}$$

$$|Ru|\leqq C\|u\|. \tag{4.4}$$

Moreover, we assume that the following properties involving B hold: setting

$$B(u)=B(u,u),\qquad b(u,v,w)=\langle B(u,v),w\rangle_{V',V},$$

for $u, v, w\in V$, we have

$$b(u,v,v)=0, \tag{4.5}$$

and there exists $C_b>0$ such that for any u, v, w in V

$$|b(u,v,w)|\leqq C_b|u|^{1/2}\|u\|^{1/2}\|v\|\,|w|^{1/2}\|w\|^{1/2}. \tag{4.6}$$

We also assume that there exists $C_B>0$ such that for any u, v in $D(A)$

$$|B(u,v)|\leqq C_B|u|^{1/2}|Au|^{1/2}\|v\|, \tag{4.7}$$

$$|B(u,v)|\leqq C_B|u|^{1/2}\|u\|^{1/2}\|v\|^{1/2}|Av|^{1/2}, \tag{4.8}$$

$$|B(u,v)|\leqq C_B\|u\|\,\|v\|\left(1+\mathrm{Log}\left(\frac{|Au|^2}{\lambda_1\|u\|^2}\right)\right)^{1/2}, \tag{4.9}$$

where λ_1 is the smallest eigenvalue of A.

Under these assumptions we recall without proofs the following results that are classical for this class of evolution equations.

THEOREM 2 (Well-posed problem). *There exists a unique solution u of* (4.1) *and* (4.2) *belonging to*

$$C(0,+\infty;H)\cap L^2(0,T;D(A^{1/2}))\quad(\textit{for each } T>0).$$

Moreover, if $u_0\in V$, then

$$u\in C(0,+\infty;V)\cap L^2(0,T;D(A))\quad(\textit{for each } T>0).$$

Proof. See [18] and the references therein.

THEOREM 3 (Dissipativity and absorbing sets). *Let us consider initial data u_0 in* (4.2) *satisfying*

$$|u_0|\leqq R_0.$$

Then there exists a time t_0 that depends on u_0 through R_0, and on the data $\nu, \lambda_1, |f|$ of the equation such that for $t\geqq t_0$

$$|u(t)|\leqq M_0,\qquad \|u(t)\|\leqq M_1, \tag{4.10}$$

where M_0, M_1 are independent of u_0, but dependent on the other data.

Proof. This is related to the existence of an absorbing set in H and V for the dynamical system (4.1); actually $t_0=C_1\log(R_0)+C_2$ is the entrance time in these absorbing sets; see Chapter III, § 2.2, in [18].

THEOREM 4 (Time analyticity). *Let u_0 belong to V; then there exists a domain of $\mathscr{C}$ containing*

$$\begin{aligned}\Gamma(\|u_0\|)=\{\zeta\in\mathscr{C}, \operatorname{Re}\zeta>0,\\ |\operatorname{Im}\zeta|\leqq T_0 \text{ if } \operatorname{Re}\zeta\geqq T_0,\\ |\operatorname{Im}\zeta|\leqq\operatorname{Re}\zeta \text{ if } \operatorname{Re}\zeta\leqq T_0\},\end{aligned}$$

where u can be extended to an analytic map into $D(A)$; here T_0 depends on $\lambda_1, |f|, \|u_0\|, \nu$.

Proof. See [5], [9], [17].

Remark 4. We will use Theorem 4 in the following form: let u_0 belong to H, for $t\geqq t_0$ (t_0 as in Theorem 3), u can be extended to an analytic map from

$$\begin{aligned}\Gamma=\{\zeta\in\mathscr{C}, \operatorname{Re}\zeta>t_0,\\ |\operatorname{Im}\zeta|\leqq T_0 \text{ if } \operatorname{Re}\zeta\geqq t_0+T_0,\\ |\operatorname{Im}\zeta|\leqq\operatorname{Re}\zeta-t_0 \text{ if } \operatorname{Re}\zeta\leqq t_0+T_0\}\end{aligned}$$

into $D(A)$, with T_0 that depends on $\nu, \lambda_1, |f|$ but which is independent of u_0.

We also recall from the proof of Theorem 4 (see [17], for instance) that for $\zeta\in\Gamma$,

$$\|u(\zeta)\|\leqq 2(1+M_1), \tag{4.11}$$

M_1 being as in Theorem 3.

Let us briefly describe two examples of equations satisfying these hypotheses.

Example 1. The stream function formulation of the two-dimensional Navier-Stokes equations.

The usual velocity-pressure formulation of the two-dimensional Navier-Stokes equations reads (see [22]),

$$\frac{\partial u}{\partial t}-\nu\Delta u+(u.\nabla u)+\nabla P=F, \tag{4.12}$$

$$\operatorname{div} u=0, \tag{4.13}$$

where $u=\{u_1, u_2\}$ is the velocity vector, P the pressure; the driving force F is given. These equations are supplemented by space-periodic boundary conditions on Π^2.

Nevertheless, we do not want to address in this paper the difficult question of the treatment by wavelets of the incompressibility condition (4.13). Therefore we rather consider the stream function formulation of the Navier-Stokes equation that reads, setting $u=\operatorname{curl}\Psi$

$$-\frac{\partial\Delta\Psi}{\partial t}+\nu\Delta^2\Psi+\frac{\partial}{\partial x_2}\left(\Delta\Psi\frac{\partial\Psi}{\partial x_1}\right)-\frac{\partial}{\partial x_1}\left(\Delta\Psi\frac{\partial\Psi}{\partial x_2}\right)=\operatorname{curl} F, \tag{4.14}$$

where the unknown Ψ maps $\mathscr{R}_+$ into $\dot{H}^1(\Pi^2)$.

The equation (4.14) has the form (4.1) with $H=\dot{H}^1(\Pi^2), A=-\Delta, D(A)=\dot{H}^3(\Pi^2), R=0$,

$$(B(\Psi_1,\Psi_2),\Psi_3)=\iint_{\Pi^2}\Delta\Psi_1\left(\frac{\partial\Psi_2}{\partial x_2}\frac{\partial\Psi_3}{\partial x_1}-\frac{\partial\Psi_3}{\partial x_2}\frac{\partial\Psi_2}{\partial x_1}\right)dx_1\,dx_2$$

for any Ψ_1, Ψ_2, Ψ_3 in V, and f is defined by

$$(f,\Psi)=\iint_{\Pi^2}F\cdot\operatorname{curl}\Psi\,dx_1\,dx_2.$$

Since the application

$$\operatorname{curl}: \dot{H}^{s+1}(\Pi) \to \{v \in (\dot{H}^s(\Pi))^2; \operatorname{div} v = 0\}$$

$$\Psi \mapsto \left\{\frac{\partial \Psi}{\partial x_2}, -\frac{\partial \Psi}{\partial x_1}\right\}$$

is an isomorphism, all the assumptions (4.3)-(4.9) are satisfied, and are easily derived from similar results that hold for the velocity-pressure formulation of the Navier-Stokes equations (see [18]).

Example 2. The Kuramoto-Sivashinsky equation. This equation reads

$$\frac{\partial v}{\partial t}+\frac{\partial^4 v}{\partial x^4}+\frac{\partial^2 v}{\partial x^2}+v\frac{\partial v}{\partial x}=0, \tag{4.15}$$

where the unknown v maps $\mathscr{R}_+$ into $\dot{L}^2(\Pi)$.

We use the translation method of [15], [16], [18] to transform (4.15) into an equation that has the form (4.1), with assumptions (4.3)-(4.9) satisfied; actually, the difficulty is to prove (4.3), and hence Theorem 3. Following the method of the references above, we restrict ourselves to the case where v is an odd function on $\Pi = [\frac{-1}{2}, \frac{1}{2}]$. Hence

$$H = \{w \in \dot{L}^2(\Pi); \ w \text{ is odd}\},$$

$$A = \frac{\partial^4}{\partial x^4}, \qquad D(A) = \dot{H}^4(\Pi) \cap H.$$

We then set

$$v = u + \phi,$$

where ϕ is an appropriate function in $D(A)$. The new equation for u reads

$$\begin{gathered}\frac{\partial u}{\partial t}+\frac{\partial^4 u}{\partial x^4}+\frac{\partial^2 u}{\partial x^2}+u\frac{\partial \phi}{\partial x}+\phi\frac{\partial u}{\partial x}+u\frac{\partial u}{\partial x}=g(\phi),\\ g(\phi)=-\frac{\partial^4 \phi}{\partial x^4}-\frac{\partial^2 \phi}{\partial x^2}-\phi\frac{\partial \phi}{\partial x}.\end{gathered} \tag{4.16}$$

This equation has the form (4.1), with $\nu = 1$, A as above, $f = g(\phi)$, while $B(u) = u(\partial u/\partial x)$, and

$$R(u) = \frac{\partial^2 u}{\partial x^2} + u\frac{\partial \phi}{\partial x} + \phi\frac{\partial u}{\partial x}.$$

The choice of ϕ such that (4.3) holds is one of the main tasks in [15] and [16]. On the other hand, using Sobolev embeddings and Agmon inequality (see [18]), we easily check that for $u, v \in V$

$$|B(u, v)| \leqq |u|_{L^\infty}\left|\frac{\partial v}{\partial x}\right| \leqq |u|^{1/2}\|u\|^{1/2}|v|^{1/2}\|v\|^{1/2}.$$

This implies (4.6)-(4.9). Finally we observe that (4.5) is not satisfied, but instead we have

$$b(u, u, u) = 0$$

for any $u \in V$. This induces some slight changes in the proofs of Theorem 2 and 3 that still hold (see [18]).

5. Approximate inertial manifolds. Let us consider the dynamical system defined by (4.1). First we recall what is an approximate inertial manifold (AIM).

DEFINITION. A smooth finite-dimensional manifold $\mathcal{M}$ is an AIM of order η (in H) for (4.1) if for any trajectory $u(t)$ there exists t_1 that depends on $|u_0|$ (as in Theorem 3), such that for $t \geqq t_1$

$$\operatorname{dist}_H (u(t), \mathcal{M}) \leqq \eta.$$

Remark 5. Therefore the global attractor is imbedded into an η-neighborhood (in H) of $\mathcal{M}$.

Remark 6. For equivalent definitions of AIMs, and for construction of sequences of AIMs for several dissipative equations, see [3], [4], [5], [21].

Let us describe briefly the aim of this section. We are looking for AIMs defined as graphs of mappings

$$\Phi : PV \rightarrow (\mathrm{Id} - P) V,$$

where P is a suitable linear projector that has finite rank.

The innovation of our work is that, instead of considering a projector onto an eigenspace of A (see the references above), we choose P as the projector onto the space spanned by a finite number of wavelets. As we will see in the next section, it is important to take the wavelets in the natural order, i.e., to consider the functions of W_j before the functions of W_{j+1}. In other words, we choose P as the projector onto a low frequencies wavelet space.

Remark 7. For instance, for Example 1, since $H = \dot{H}^1(\Pi^2)$, we shall use the two-dimensional wavelets derived from the one-dimensional ones by tensor products (see the Appendix). For Example 2, we have to choose wavelets that are odd, since $H = \{v \in \dot{L}^2(\Pi);\ v \text{ is odd}\}$. For that purpose, we proceed as in [13, § VI.11]. We know that the function ψ_N defined in § 2 satisfies

$$\psi_N(1-x) = (-1)^{N+1}\psi_N(x). \tag{5.1}$$

Therefore, N being fixed, a basis for H is given by considering all functions

$$\tilde{\psi}_{j,k}(x) = \frac{1}{\sqrt{2}}\left(\psi_{j,k}\left(x - \tfrac{1}{2}\right) + (-1)^N \psi_{j,k^*}\left(x - \tfrac{1}{2}\right)\right), \tag{5.2}$$

such that $k + k^* = 0[2^j]$, for all $j > 0$ (we recall that $\psi_{j,k}(x)$ is defined in (2.7)). Of course this construction applies to wavelets that enjoy symmetries, as the Littlewood-Paley ones (see [13]).

5.1. Induced trajectories lying in the flat manifold. In the following, for the sake of clarity, we will assume that $n = 1$ and that $H = \dot{H}^p(\Pi)$. The reader could check that only minor changes are needed to extend the results of the next sections to, for instance, the cases described in Examples 1 and 2.

Let V_j (j being fixed in this section) be the flat manifold associated with the (linear) Galerkin approximation of (4.1), (4.2) by periodic piecewise polynomial functions. Being provided $m + p \leqq N + 1$ such that $V_j \subset V$ holds, we define P_j as the orthogonal projector in H onto V_j; let $Q_j = \mathrm{Id}_H - P_j$. We also define $P_{1,j}$ as the orthogonal projector in V onto V_j, and $Q_{1,j} = \mathrm{Id}_V - P_{1,j}$.

Remark 8. In the following we shall omit the subscript j on P_j, Q_j, $P_{1,j}$, and $Q_{1,j}$.

Following the methods developed in [21] we call induced trajectories (lying in the flat manifold) associated with $u(t)$ solution of (4.1), (4.2) the trajectories $y(t)$, $y_1(t)$ defined by

$$y(t) = Pu(t), \qquad y_1(t) = P_1 u(t). \tag{5.3}$$

We also set

$$z(t) = Qu(t), \qquad z_1(t) = Q_1 u(t). \tag{5.4}$$

Remark 9. In contrast with the spectral case we no longer have $y = y_1$.

We recall from the results of § 3 the following proposition.

PROPOSITION 4. *For $p + m < N + 1$, there exists $c > 0$ such that for any z in QV,*

$$|z| \leqq \frac{c}{2^{jm}} \|z\|. \tag{5.5}$$

Proof. For the sake of clarity, and although this is not necessary, we first prove the result for $p = 0$, an then consider the case $p \neq 0$.

First let us assume $p = 0$. We then have $H = \dot{L}^2(\Pi)$, and

$$QH = V_j^{\perp} = \bigoplus_{l \geqq j} W_l,$$

with the notation of §§ 2 and 3. For z in QV, we write

$$z = \sum_{l=j}^{+\infty} \sum_{k=1}^{2^j} \gamma_{l,k} \psi_{l,k},$$

such that

$$|z|^2 = \sum_{l,k} |\gamma_{l,k}|^2.$$

On the other hand, thanks to Corollary 5, we have for $m < N + 1$

$$c(p) \sum_{l,k} 2^{lm} |\gamma_{l,k}|^2 \leqq \|z\|^2.$$

Therefore the two last inequalities yield (5.5).

Let us consider the general case $p \neq 0$. z belongs to QV means that there exists u in $\dot{H}^{m+p}(\Pi)$ such that $z = u - Pu$. Since Pu belongs to $\dot{H}^{N+1}(\Pi)$, we have to assume $m + p \leqq N + 1$ to have $z \in \dot{H}^{m+p}(\Pi)$. Hence

$$(z \in QV) \Leftrightarrow (z \in \dot{H}^{p+m}(\Pi) \quad \text{and} \quad \langle (-\Delta)^p z, \psi_{l,k} \rangle_{\dot{H}^{m-p}, \dot{H}^{p-m}} = 0 \text{ for } l < j).$$

Then we write

$$(-\Delta)^p z = \sum_{l=j}^{+\infty} \sum_{k=1}^{2^l} \gamma_{l,k} \psi_{l,k}, \tag{5.6}$$

this equality holding in $\dot{H}^{m-p}(\Pi)$; moreover, thanks to Corollary 5

$$\sum_{l=j}^{+\infty} \sum_{k=1}^{2^l} |\gamma_{l,k}|^2 4^{l(m-p)} \leqq c(m-p) |(\Delta)^p z|^2_{m-p}. \tag{5.7}$$

We then observe that

$$|(-\Delta)^p z|_{m-p} = |z|_{p+m} = \|z\|, \tag{5.8}$$

which yields

$$C 4^{jm} \sum_{l=j}^{+\infty} \sum_{k=1}^{2^l} 4^{-lp} |\gamma_{l,k}|^2 \leqq \|z\|^2. \tag{5.9}$$

On the other hand, computing the norm in $\dot{H}^{-p}(\Pi)$ of $|(-\Delta)^p z|$, we obtain, thanks to Corollary 5,

$$|z|^2 \leqq C \sum_{l=j}^{+\infty} \sum_{k=1}^{2^l} 4^{-lp} |\gamma_{l,k}|^2. \tag{5.10}$$

Therefore, (5.9) and (5.10) yield (5.5).

We would like to check this kind of inequality for z_1 in $Q_1 V$. For this purpose we need more regularity on the wavelets $\psi_{j,k}$'s.

PROPOSITION 5. *For $2m+p<N+1$, there exists $c>0$ such that for any z_1 in $Q_1 V$*

$$|z_1| \leqq \frac{c}{2^{jm}} \|z_1\|. \tag{5.11}$$

Proof. First we observe that

$$(z_1 \in Q_1 V) \Leftrightarrow (z_1 \in \dot{H}^{p+m}(\Pi) \quad \text{and} \quad \langle (-\Delta)^{p+m} z_1, \psi_{l,k} \rangle_{\dot{H}^{-m-p}, \dot{H}^{p+m}} = 0 \text{ for } l<j).$$

Then we write

$$(-\Delta)^{m+p} z_1 = \sum_{l=j}^{+\infty} \sum_{k=1}^{2^l} \gamma_{l,k} \psi_{l,k}. \tag{5.12}$$

We then apply the method of the proof of Proposition 4 to derive (5.11). The assumption $2m+p<N+1$ comes from the fact that we apply Corollary 5 to estimate the $\dot{H}^{-2m-p}(\Pi)$ norm of $(-\Delta)^{p+m} z_1$ to obtain $|z_1|$.

5.2. Behavior of small eddies.

THEOREM 5. *For $2m+p<N+1$, both $z(t)$, $z_1(t)$ satisfy, for t large enough as in Theorem 3, the following inequalities*:

$$|z(t)|, |z_1(t)|, |z'(t)|, |z_1'(t)| \leqq c \frac{\sqrt{j}}{4^{jm}}, \tag{5.13}$$

$$\|z(t)\|, \|z_1(t)\| \leqq c \frac{\sqrt{j}}{2^{jm}}. \tag{5.14}$$

We reinterpret these inequalities saying that the flat manifold V_j associated to splines of order N is an approximate inertial manifold of order $\sqrt{j}/4^{jm}$ in H and of order $\sqrt{j}/2^{jm}$ in V.

Remark 10. Here we set $z' = dz/dt$, $z_1' = dz_1/dt$.

Remark 11. We match here the accuracy established in the spectral case. For instance, let us consider Example 1, the Navier-Stokes equations. In [5] (see also [21]), it has been proven that if $\tilde{P}_j$ is the orthogonal projector onto the space spanned by the first $(2^j)^2$ (here the dimension space is $n=2$) eigenvectors of A; $\tilde{Q}_j = Id - \tilde{P}_j$; then we have, for t large enough,

$$|\tilde{Q}_j u(t)| \leqq c\delta L^{1/2},$$

$$\|\tilde{Q}_j u(t)\| \leqq c\delta^{1/2} L^{1/2},$$

where $\delta = c/4^{jm}$, L being a logarithmic correction of δ. Here we obtain analogous results, considering a wavelet space that has the same dimension. Therefore, for equations of type (4.1), we can say that wavelets provide a flat AIM that has the same order of approximation as the one obtained using the eigenvectors of A.

Proof of Theorem 5. We take the scalar product in H of (4.1) with z_1; using (4.5) and

$$(Au, z_1) = ((u, z_1)) = \|z_1\|^2, \tag{5.15}$$

we obtain

$$\nu\|z_1\|^2 = (f, z_1) - \left(\frac{du}{dt}, z_1\right) - (Ru, z_1) - b(z_1, y_1, z_1) - b(y_1, y_1, z_1). \tag{5.16}$$

Then (4.9) yields

$$|b(y_1, y_1, z_1)| \leqq C_B\|y_1\|^2\left(1 + \operatorname{Log}\left(\frac{|Ay_1|^2}{\lambda_1\|y_1\|^2}\right)\right)^{1/2}|z_1|. \tag{5.17}$$

Thanks to (3.1), we have

$$\left(1 + \operatorname{Log}\left(\frac{|Ay_1|^2}{\lambda_1\|y_1\|^2}\right)\right)^{1/2} \leqq c\sqrt{j}. \tag{5.18}$$

Hence

$$|b(y_1, y_1, z_1)| \leqq C\sqrt{j}\|y_1\|^2|z_1|. \tag{5.19}$$

On the other hand, we use (4.6) to obtain

$$|b(z_1, y_1, z_1)| \leqq C\|y_1\|\,\|z_1\|\,|z_1|. \tag{5.20}$$

Then we infer from (5.16), (5.19), and (5.20)

$$\nu\|z_1\|^2 \leqq \left(|f| + \left|\frac{du}{dt}\right| + |Ru| + C\sqrt{j}\|y_1\|^2 + C\|y_1\|\,\|z_1\|\right)|z_1|. \tag{5.21}$$

We observe that for t large enough as in Theorem 3 we have both

$$\|y_1\|, \|z_1\| \leqq \|u\| \leqq M_1, \tag{5.22}$$

and that for t large enough as in Theorem 4

$$\left|\frac{du}{dt}\right| \leqq C, \tag{5.23}$$

that is a consequence of the Cauchy formula applied to u in a ball included in Γ, and of (4.11). On the other hand, we derive from (4.4) and (5.22) that $|\mathrm{Ru}| \leqq C$.

We use these facts to obtain

$$\|z_1\|^2 \leqq C\sqrt{j}|z_1|, \tag{5.24}$$

holding for t large enough as above, where C depends on N and on the data ν, λ_1, $|f|$ of the equation, but is independent of j. Then Proposition 5 yields

$$\|z_1\| \leqq C\frac{\sqrt{j}}{2^{jm}}, \tag{5.25}$$

$$|z_1| \leqq C\frac{\sqrt{j}}{4^{jm}}. \tag{5.26}$$

Now we estimate $|z|$ and $\|z\|$; we observe that $z = Qz_1$ and that therefore

$$|z| \leqq |z_1| \leqq C\frac{\sqrt{j}}{4^{jm}} \tag{5.27}$$

holds; but moreover we have the following.

PROPOSITION 6. *For $m+p<N+1$, Q, which is an orthogonal projector in H, maps continuously V into itself, and its norm as a linear operator acting in V is bounded independently of j.*

Proof. For $v \in V$, the following inequalities hold in $\dot{H}^{m-p}(\Pi)$

$$(-\Delta)^p v = \sum_{l=0}^{+\infty} \sum_{k=0}^{2^l} \gamma_{l,k} \psi_{l,k}, \tag{5.28}$$

$$(-\Delta)^p Qv = \sum_{l=j}^{+\infty} \sum_{k=0}^{2^l} \gamma_{l,k} \psi_{l,k}. \tag{5.29}$$

Thanks to Corollary 5 we have

$$\begin{aligned} |(-\Delta)^p Qv|_{m-p} = \|Qv\| &\leqq \frac{1}{c_1(m-p)^{1/2}} \left(\sum_{l=j}^{+\infty} \sum_{k=0}^{2^l} 4^{l(m-p)} |\gamma_{l,k}|^2 \right)^{1/2} \\ &\leqq \frac{1}{c_1(m-p)^{1/2}} \left(\sum_{l=0}^{+\infty} \sum_{k=0}^{2^l} 4^{l(m-p)} |\gamma_{l,k}|^2 \right)^{1/2} \\ &\leqq \left(\frac{c_2(m-p)}{c_1(m-p)} \right)^{1/2} \|v\|, \end{aligned} \tag{5.30}$$

where $c_1(m-p)$, $c_2(m-p)$ are as in Corollary 5.

We apply this result to $z = Qz_1$ to obtain

$$\|z\| \leqq C \frac{\sqrt{j}}{2^{jm}}. \tag{5.31}$$

To end the proof of Theorem 5, it remains to estimate $|z'|$ and $|z_1'|$. For this purpose we observe that z and z_1 are analytic in time in the same domain as u, and then we use Cauchy's formula to get the estimates on z' and z_1' from these on z and z_1. For the reader's convenience we give a complete proof below.

First we need to introduce some notation. Let H_c, V_c, V_c^j, and $D(A^s)_c$ be the complexifications of H, V, V_j, and $D(A^s)$. We recall that if $u_1 + iu_2$ is a typical element of H_c, then we have

$$A(u_1 + iu_2) = Au_1 + iAu_2,$$

$$(u_1 + iu_2, v_1 + iv_2) = (u_1, v_1) + (u_2, v_2) + i[(u_2, v_1) - (u_1, v_2)],$$

and that the multiplication by a complex constant is performed in the natural manner.

We observe that the family $\{\psi_{l,k}\}_{0 \leqq l < j; 1 \leqq k \leqq 2^j}$ is an orthonormal basis of V_c^j in $\dot{L}^2(\Pi)_c$ and that moreover we have the following.

LEMMA 4. The family

$$\{\psi_{l,k}\}_{0 \leqq l < +\infty; 1 \leqq k \leqq 2^l} \text{ is an orthonormal basis of } \dot{L}^2(\Pi)_c, \tag{5.32}$$

and for $(m+p)|s| < N+1$ the family

$$\{\psi_{l,k}\}_{0 \leqq l < +\infty; 1 \leqq k \leqq 2^l} \text{ is an unconditional basis of } D(A^{s/2})_c. \tag{5.33}$$

Proof. The proof is straightforward and left as an exercise to the reader.

Now we observe that y_1 and z_1 can be extended as time analytic functions in the same domain as u. Let Y_1, Z_1, and U be the extensions of y_1, z_1, and u. Then U satisfies, for $\zeta \in \Gamma$,

$$\frac{\partial U}{\partial \zeta} + \nu AU + RU + B(U) = f. \tag{5.34}$$

Taking the scalar product in H_c of (5.34) with Z_1 we obtain

$$\begin{aligned}\nu\|Z_1\|^2 \leqq \left(\left|\frac{\partial U}{\partial \zeta}\right| + |RU|\right)|Z_1| + |b(Y_1, Y_1, Z_1)| \\ + |b(Y_1, Z_1, Z_1)| + |b(Z_1, Y_1, Z_1)| \\ + |b(Z_1, Z_1, Z_1)| + |f||Z_1|.\end{aligned} \tag{5.35}$$

Easy computations yield

$$|b(Z_1, Z_1, Z_1)| \leqq |Z_1| \, \|Z_1\|^2, \tag{5.36}$$

$$|b(Z_1, Y_1, Z_1)| \leqq C|Z_1| \, \|Z_1\| \, \|Y_1\|, \tag{5.37}$$

where C is an absolute constant.

Using (3.1) on Re (Y_1) and Im (Y_1) we obtain

$$|b(Y_1, Y_1, Z_1)| \leqq C\|Y_1\|^2 \sqrt{j}|Z_1|, \tag{5.38}$$

$$|b(Y_1, Z_1, Z_1)| \leqq C\|Y_1\| \, \|Z_1\| \sqrt{j}|Z_1|, \tag{5.39}$$

where C depends on N.

For ζ such that $|\text{Im } \zeta| \leqq T_0/2$ and Re $\zeta \geqq t_0 + 2T_0$ (with t_0 and T_0 as in Remark 4), we apply Cauchy's formula on a ball B centered at ζ, of radius $T_0/2$, to obtain

$$\left|\frac{\partial U}{\partial \zeta}(\zeta)\right| \leq C \sup_{\eta \in B} |U(\eta)|, \tag{5.40}$$

and we infer from (4.11)

$$\left|\frac{\partial U}{\partial \zeta}\right| \leqq C, \tag{5.41}$$

where C depends on the data of the equation through M_1.

We also use (4.11) to majorize $\|Y_1\|$, $\|Z_1\|$, and $|RU|$ (thanks to (4.4)) by $2(1+M_1)$, for ζ as above.

All these facts yield

$$\|Z_1\|^2 \leqq C\sqrt{j}|Z_1|. \tag{5.42}$$

On the other hand, we observe that Lemma 4 provides analogous forms of Poincaré inequalities (5.5) and (5.11) for Z and Z_1. We also infer from Lemma 4 that the orthogonal projector in H_c onto the orthogonal complement of V_c^j is continuous as a linear operator mapping V_c into itself, the corresponding norm being bounded independently of j.

We apply these remarks to (5.42) to obtain, by the same computations as above,

$$|Z|, |Z_1| \leqq C\frac{\sqrt{j}}{4^{jm}}, \tag{5.43}$$

$$\|Z\|, \|Z_1\| \leqq C\frac{\sqrt{j}}{2^{jm}}, \tag{5.44}$$

for ζ belonging to a strip thinner than Γ, for example

$$\left\{\text{Re } \zeta \geqq t_0 + 2T_0, |\text{Im } \zeta| \leqq \frac{T_0}{2}\right\}.$$

To end the proof of Theorem 5, we apply Cauchy's formula on a ball B centered at $t \geqq t_0 + 3T_0$, of radius $T_0/2$ to obtain

$$\left|\frac{dz}{dt}(t)\right| \leqq C \sup_{\eta \in B} |Z(\eta)|, \tag{5.45}$$

where C depends on the data of the equation through M_1. We then infer from (5.43) and (5.45)

$$\left|\frac{dz}{dt}\right| \leqq C \frac{\sqrt{j}}{4^{jm}}. \tag{5.46}$$

An analogous result for z_1' concludes the proof.

5.3. Two nonflat approximate inertial manifolds. Following the methods developed in [21] we provide below two examples of (nonflat) approximate inertial manifolds of higher order than the flat one.

First let us consider the (nonlinear) mapping $\Phi_1 \colon PV \to QV$ defined as follows: for each y in PV, there exists a unique $\Phi_1(y)$ in QV such that

$$\nu((\Phi_1(y), \tilde{z})) = (f - \nu Ay - Ry - B(y), \tilde{z}), \tag{5.47}$$

for any $\tilde{z}$ in QV. Φ_1 is well defined, thanks to a straightforward consequence of the Riesz representation theorem.

Remark 12. Let us notice that the term $(Ay, \tilde{z})$ does not vanish. Actually, unlike the spectral case y and $\tilde{z}$ are orthogonal in H, not in V. This point can be also observed in the nonlinear algorithms described in [14].

Let $\mathcal{M}_1$ be the graph of Φ_1; then we have the following.

PROPOSITION 7. *$\mathcal{M}_1$ is an approximate inertial manifold for* (4.1) *of order $j/8^{jm}$ in H and of order $j/4^{jm}$ in V.*

Remark 13. We match here the accuracy established in the spectral case (see [5], [21]) in the sense of Remark 11.

Proof. We plan to estimate the gap between the trajectory $u(t)$ and its induced trajectory lying in $\mathcal{M}_1$, $y(t) + \Phi_1(y(t))$, where $y(t) = Pu(t)$ as above. We set

$$\chi_1(t) = \Phi_1(y(t)) - z(t). \tag{5.48}$$

We rewrite (5.47) as

$$\nu QA\Phi_1(y) + \nu QAy + QRy + QB(y) = Qf. \tag{5.49}$$

Hence χ_1 satisfies

$$\nu QA\chi_1 + Q(B(y) - B(u)) = \frac{dz}{dt} + Rz. \tag{5.50}$$

We take the scalar product in H of (5.50) with χ_1 to obtain

$$\nu \|\chi_1\|^2 \leqq |b(y, z, \chi_1)| + |b(z, y, \chi_1)| + |b(z, z, \chi_1)| + \left|\frac{dz}{dt}\right| |\chi_1| + |Rz|\,|\chi_1|. \tag{5.51}$$

On the other hand, (4.9) yields

$$|b(y, z, \chi_1)| \leqq C_B \|y\|\,\|z\| \left(1 + \operatorname{Log}\left(\frac{|Ay|^2}{\lambda_1 \|y\|^2}\right)\right)^{1/2} |\chi_1|. \tag{5.52}$$

We infer from (4.10) and (5.30)

$$\|y\| \leqq C\|u\| \leqq CM_1, \tag{5.53}$$

and from (5.14)

$$\|z\| \leqq C\frac{\sqrt{j}}{2^{jm}},$$

these estimates holding for t large enough as in Theorem 5.

We apply (5.5) to obtain

$$|\chi_1| \leqq \frac{C}{2^{jm}} \|\chi_1\|, \tag{5.54}$$

and (3.1) to get

$$\left(1+\operatorname{Log}\left(\frac{|Ay|^2}{\lambda_1\|y\|^2}\right)\right)^{1/2} \leqq C\sqrt{j}. \tag{5.55}$$

Then we finally obtain

$$|b(y, z, \chi_1)| \leqq C\frac{j}{4^{jm}} \|\chi_1\|. \tag{5.56}$$

We also have, using (4.6),

$$|b(z, y, \chi_1)| \leqq C_b|z|^{1/2}\|z\|^{1/2}\|y\| \, |\chi_1|^{1/2}\|\chi_1\|^{1/2}. \tag{5.57}$$

We then infer from (5.13), (5.14), (5.53), (5.54), and (5.57)

$$|b(z, y, \chi_1)| \leqq C\frac{\sqrt{j}}{4^{jm}} \|\chi_1\|, \tag{5.58}$$

for t large enough as above.

Using (4.5) and (4.6) we obtain

$$|b(z, z, \chi_1)| \leqq C_b|z| \, \|z\| \, \|\chi_1\|, \tag{5.59}$$

as well, and thanks to (5.13) and (5.14), we have

$$|b(z, z, \chi_1)| \leqq C\frac{j}{8^{jm}} \|\chi_1\|. \tag{5.60}$$

We also have, thanks to (4.4),

$$|(Rz, \chi_1)| \leqq C\|z\| \, |\chi_1|, \tag{5.61}$$

and thanks to (5.14) and (5.54)

$$|(Rz, \chi_1)| \leqq C\frac{\sqrt{j}}{4^{jm}} \|\chi_1\|. \tag{5.62}$$

Using (5.13) to estimate $|dz/dt|$ we finally obtain

$$\|\chi_1\|^2 \leqq \left[C_1\frac{j}{4^{jm}} + C_2\frac{\sqrt{j}}{4^{jm}} + C_3\frac{j}{8^{jm}} + C_4\frac{\sqrt{j}}{8^{jm}}\right] \|\chi_1\|. \tag{5.63}$$

This yields

$$\|\chi_1\| \leqq C\frac{j}{4^{jm}}, \tag{5.64}$$

and thanks to (5.54),

$$|\chi_1| \leqq C\frac{j}{8^{jm}}, \tag{5.65}$$

holding for t large enough as above.

Now we define $\mathcal{M}_2$ as the graph of the mapping $\Phi_2: PV \to QV$ defined by induction from Φ_1 by:

$$\begin{aligned}\nu((\Phi_2(y), \tilde{z})) = (\nu QA\Phi_1(y) - QR\Phi_1(y) - QB(y, \Phi_1(y))\\ -QB(\Phi_1(y), y), \tilde{z}) \quad \text{for any } \tilde{z} \text{ in } QV.\end{aligned} \tag{5.66}$$

Existence and uniqueness of $\Phi_2(y)$ are consequences of the Riesz representation theorem. We have the following proposition.

PROPOSITION 8. *$\mathcal{M}_2$ is an approximate inertial manifold for* (4.1) *of order* $(j)^{3/2}/16^{jm}$ *in H and of order* $(j)^{3/2}/8^{jm}$ *in V.*

Remark 14. We match here the accuracy established in the spectral case (see [21]) in the sense of Remark 11.

Proof. Setting

$$\chi_2 = \Phi_2(y) - z, \tag{5.67}$$

and using (5.49) we rewrite (5.66) as

$$\begin{aligned}\nu QA\Phi_2(y) + \nu QAy + QB(y) + QRy + QR\Phi_1(y)\\ +QB(y, \Phi_1(y)) + QB(\Phi_1(y), y) = Qf.\end{aligned} \tag{5.68}$$

Hence

$$\nu QA\chi_2 + QR\chi_1 + QB(y, \chi_1) + QB(\chi_1, y) = \frac{dz}{dt} + QB(z), \tag{5.69}$$

where χ_1 is defined as above. We take the scalar product in H of (5.69) with χ_2 to obtain

$$\begin{aligned}\nu\|\chi_2\|^2 \leqq |R\chi_1||\chi_2| + |b(y, \chi_1, \chi_2)| + |b(\chi_1, y, \chi_2)|\\ +|b(z, z, \chi_2)| + \left|\frac{dz}{dt}\right| |\chi_2|.\end{aligned} \tag{5.70}$$

As usual, thanks to (4.9), we obtain

$$|b(y, \chi_1, \chi_2)| \leqq C_B\|y\| \, \|\chi_1\|\left(1 + \operatorname{Log}\left(\frac{|Ay|^2}{\lambda_1\|y\|^2}\right)\right)^{1/2} |\chi_2|. \tag{5.71}$$

From previous estimates (5.53), (5.55), and (5.64) we obtain, for t large enough as above,

$$|b(y, \chi_1, \chi_2)| \leqq C\frac{(j)^{3/2}}{4^{jm}} |\chi_2|, \tag{5.72}$$

and thanks to (5.5)

$$|b(y, \chi_1, \chi_2)| \leqq C\frac{(j)^{3/2}}{8^{jm}} \|\chi_2\|. \tag{5.73}$$

On the other hand, using (4.6)

$$|b(\chi_1, y, \chi_2)| \leqq C_b|\chi_1|^{1/2}\|\chi_1\|^{1/2}\|y\| \, |\chi_2|^{1/2}\|\chi_2\|^{1/2}. \tag{5.74}$$

We infer from (5.5), (5.53), (5.64), and (5.65)

$$(5.75)\qquad \begin{aligned}|b(\chi_1, y, \chi_2)| &\leqq C\frac{j}{(4\sqrt{2})^{jm}}\frac{1}{(\sqrt{2})^{jm}}\|\chi_2\| \\ &\leqq C\frac{j}{8^{jm}}\|\chi_2\|.\end{aligned}$$

By similar computations, using (4.5), (4.6), (5.13), and (5.14)

$$(5.76)\qquad |b(z, z, \chi_2)| \leqq C_b|z|\,\|z\|\,\|\chi_2\| \leqq C\frac{j}{8^{jm}}\|\chi_2\|.$$

We have as well, using (4.4), (5.5), (5.13), and (5.64)

$$(5.77)\qquad |R\chi_1|\,|\chi_2| + \left|\frac{dz}{dt}\right|\,|\chi_2| \leqq C\frac{\sqrt{j}}{8^{jm}}\|\chi_2\|.$$

We summarize (5.70), (5.73), (5.75), (5.76), and (5.77) by

$$(5.78)\qquad \|\chi_2\|^2 \leqq C\frac{(j)^{3/2}}{8^{jm}}\|\chi_2\|,$$

holding for t large enough as above. The Poincaré inequality (5.5) ends the proof.

Appendix. In this Appendix we want first to present a proof of Theorem 1 for the multidimensional periodic wavelet bases built from the one-dimensional ones by tensor products. For the sake of simplicity we present this result for the two-dimensional case. The reader could check that it can be extended without difficulties to the d-dimensional case, $d > 2$.

After that we will end the paper by explaining in a few words how to prove Theorem 1, considering two other important examples of wavelet bases.

In both cases we just have to prove analogous results to Propositions 1 and 2.

We recall from [11], [13] that, for each $N > 0$, there exists a function φ_N satisfying (i) and (ii) such that

$$(A.1)\qquad \int_R \varphi_N(x)\,dx \neq 0,$$

and that, setting

$$(A.2)\qquad \varphi_{j,k}(x) = 2^{j/2}\sum_{l\in\mathscr{Z}} \varphi_N(2^j x + 2^j l - k),$$

the family

$$(A.3)\qquad \{\varphi_{j,k}\}_{1\leqq k\leqq 2^j} \text{ is an orthonormal basis of } V_j.$$

(We dropped for convenience the subscript N on the $\varphi_{j,k}$'s.)

Let us introduce some notation. $H^s(\Pi^2)$ will be the usual periodic Sobolev space on the two-dimensional torus. $\dot{H}^s(\Pi^2)$ will be the set of functions u in $H^s(\Pi^2)$ such that

$$\iint_{\Pi^2} u(x)\,dx_1\,dx_2 = 0.$$

$\dot{H}^s(\Pi^2)$ is a Hilbert space when endowed with the scalar product

$$((u, v))_s = \sum_{k\in\mathscr{Z}^2} |k|^{2s}\hat{u}(k)\overline{\hat{v}(k)},$$

where

$$|k|=(k\cdot k)^{1/2},\ k\cdot l=k_1l_1+k_2l_2,$$

$$\hat{u}(k)=\int\int_{\Pi^2} u(x)\, e^{-2i\pi k\cdot x}\, dx.$$

We denote the corresponding norm

$$\|u\|_s=((u,u))_s^{1/2}.$$

Then let $\mathscr{V}_j$ be the space of functions $v\in L^2(\Pi^2)=H^0(\Pi^2)$ such that both

$$x_1\mapsto v(x_1,x_2), \qquad x_2\mapsto v(x_1,x_2),$$

belong to V_j (i.e., $\mathscr{V}_j=V_j\otimes V_j$). For $\alpha=(\alpha_1,\alpha_2)$ in $\mathscr{A}_j=\{1,\cdots,2^j\}^2$, we set

$$\varphi_\alpha(x)=\varphi_{j,\alpha_1}(x_1)\varphi_{j,\alpha_2}(x_2). \tag{A.4}$$

Then we observe that the family $\{\varphi_\alpha\}_{\alpha\in\mathscr{A}_j}$ is an orthonormal basis of $\mathscr{V}_j$.

Now we are ready to claim the following.

PROPOSITION A.1. *There exists $C>0$ such that for any v in $\mathscr{V}_j$*

$$\|v\|_{N+1}\leqq C2^{j(N+1)}\|v\|_0. \tag{A.5}$$

Remark A.1. As above, C is a constant that depends only on N.

Proof. We use the convexity of the function $\lambda\mapsto\lambda^{N+1}$ to write

$$\begin{aligned}\|u\|_{N+1}^2&=\sum_{k\in\mathscr{Z}^2}(k_1^2+k_1^2)^{N+1}|\hat{u}(k)|^2\\&\leqq 2^N\sum_{k\in\mathscr{Z}^2}(k_1^{2N+2}+k_2^{2N+2})|\hat{u}(k)|^2\\&\leqq 2^N[\|(\partial_1)^{N+1}u\|_0^2+\|(\partial_2)^{N+1}u\|_0^2],\end{aligned} \tag{A.6}$$

setting $(\partial_i)^{N+1}u=\partial^{N+1}u/\partial x_i^{N+1}$; $i=1,2$.

On the other hand, we write for

$$v=\sum_{\alpha\in\mathscr{A}_j}\lambda_\alpha\varphi_\alpha \quad \text{in } \mathscr{V}_j, \tag{A.7}$$

$$\|(\partial_1)^{N+1}v\|_0^2=\sum_{\alpha,\alpha'\in\mathscr{A}_j}\lambda_\alpha\lambda_{\alpha'}(((\partial_1)^{N+1}\varphi_\alpha,(\partial_1)^{N+1}\varphi_{\alpha'}))_0. \tag{A.8}$$

Thanks to Fubini's theorem

$$\begin{aligned}(((\partial_1)^{N+1}\varphi_\alpha,(\partial_1)^{N+1}\varphi_{\alpha'}))_0=&\left(\int_\Pi \varphi_{\alpha_1}^{(N+1)}(x_1)\varphi_{\alpha_1'}^{(N+1)}(x_1)\,dx_1\right)\\&\cdot\left(\int_\Pi \varphi_{\alpha_2}(x_2)\varphi_{\alpha_2'}(x_2)\,dx_2\right),\end{aligned} \tag{A.9}$$

where we set

$$\varphi_{\alpha_1}^{(N+1)}(x_1)=\frac{\partial^{N+1}}{\partial x_1^{N+1}}\varphi_{\alpha_1}(x_1),$$

and where we dropped the subscript j on the φ_{j,α_i}'s to write φ_{α_i}; $i=1,2$.

Using $\int_\Pi \varphi_{\alpha_2}(x_2)\varphi_{\alpha_2'}(x_2)\,dx_2=0$ if $\alpha_2\neq\alpha_2'$ (cf. (A.3)), we obtain

$$\|(\partial_1)^{N+1}v\|_0^2=\sum_{\alpha_2=1}^{2^j}\int_\Pi\left(\sum_{\alpha_1=1}^{2^j}\lambda_\alpha\varphi_{\alpha_1}^{(N+1)}(x_1)\right)^2dx_1. \tag{A.10}$$

We then apply (3.1) to the function $x_1\mapsto\sum_{\alpha_1=1}^{2^j}\lambda_\alpha\varphi_{\alpha_1}^{(N+1)}(x_1)$

$$\int_\Pi\left(\sum_{\alpha_1=1}^{2^j}\lambda_\alpha\varphi_{\alpha_1}^{(N+1)}(x_1)\right)^2dx_1\leqq C4^{j(N+1)}\int_\Pi\left(\sum_{\alpha_1=1}^{2^j}\lambda_\alpha\varphi_{\alpha_1}(x_1)\right)^2dx_1. \tag{A.11}$$

(A.3) yields

$$\int_\Pi\left(\sum_{\alpha_1=1}^{2^j}\lambda_\alpha\varphi_{\alpha_1}(x_1)\right)^2dx_1=\sum_{\alpha_1=1}^{2^j}\lambda_\alpha^2, \tag{A.12}$$

therefore

$$\|(\partial_1)^{N+1}v\|_0^2\leqq C4^{j(N+1)}\left(\sum_{\alpha\in\mathscr{A}_j}\lambda_\alpha^2\right). \tag{A.13}$$

To conclude we recall that the family $\{\varphi_\alpha\}_{\alpha\in\mathscr{A}_j}$ is an orthonormal basis of $\mathscr{V}_j$, then using (A.6), (A.13), and that (A.13) holds for $\|(\partial_2)^{N+1}v\|_0^2$ as well, we obtain

$$\|v\|_{N+1}^2\leqq C4^{j(N+1)}\|v\|_0^2.$$

Now we set

$$\mathscr{W}_j=\mathscr{V}_{j+1}\cap(\mathscr{V}_j)^\perp. \tag{A.14}$$

$\mathscr{W}_j$ can be viewed as the direct sum of three of its subspaces, namely

$$V_j\otimes W_j,\quad W_j\otimes V_j,\quad W_j\otimes W_j.$$

We observe that the family $\{\varphi_{\alpha_1}(x_1)\psi_{\alpha_2}(x_2)\}_{\alpha\in\mathscr{A}_j}$ is an orthonormal basis of

$$V_j\otimes W_j. \tag{A.15}$$

(We dropped the subscript j on ψ_{j,α_2} to write ψ_{α_2}.) We claim the following.

PROPOSITION A.2. *There exists $C>0$ such that for any w in $\mathscr{W}_j$*

$$\|w\|_{-N-1}\leqq C2^{-j(N+1)}\|w\|_0. \tag{A.16}$$

Proof. For w in $\mathscr{W}_j$ we write $w=w_1+w_2+w_3$, where

$$w_1\in V_j\otimes W_j,\quad w_2\in W_j\otimes V_j,\quad w_3\in W_j\otimes W_j.$$

We observe that

$$\|w\|_0^2=\|w_1\|_0^2+\|w_2\|_0^2+\|w_3\|_0^2, \tag{A.17}$$

$$\|w\|_{-N-1}^2\leqq 3(\|w_1\|_{-N-1}^2+\|w_2\|_{-N-1}^2+\|w_3\|_{-N-1}^2). \tag{A.18}$$

It follows that it is sufficient to check (A.16) on both w_1, w_2, w_3. Because the proofs are similar we present below only the proof for w_1.

Let

$$w_1=\sum_{\alpha\in\mathscr{A}_j}\lambda_\alpha\varphi_{\alpha_1}(x_1)\psi_{\alpha_2}(x_2). \tag{A.19}$$

Using (2.7), (3.17), and (A.2) easy computations yield

$$\hat{w}_1(l)=\left(\frac{1}{2^j}\sum_{\alpha\in\mathscr{A}_j}\lambda_\alpha e^{-2i\pi\alpha\cdot l/2^j}\right)\hat{\varphi}\left(\frac{l_1}{2^j}\right)\hat{\psi}\left(\frac{l_2}{2^j}\right). \tag{A.20}$$

Hence

$$\|w_1\|_0^2=\sum_{l\in\mathscr{Z}^2}\left|m\left(\frac{l}{2^j}\right)\right|^2\left|\hat{\varphi}\left(\frac{l_1}{2^j}\right)\right|^2\left|\hat{\psi}\left(\frac{l_2}{2^j}\right)\right|^2, \tag{A.21}$$

setting

$$m\left(\frac{l}{2^j}\right)=\frac{1}{2^j}\sum_{\alpha\in\mathcal{A}_j}\lambda_\alpha\, e^{-2i\pi(l\cdot\alpha/2^j)}. \tag{A.22}$$

m is a $\mathcal{Z}^2$-periodic function. Using this fact, we obtain

$$\|w_1\|_0^2=\sum_{k\in\Gamma}\left|m\left(\frac{k}{2^j}\right)\right|^2\left(\sum_{l_1\in\mathcal{Z}}\left|\hat{\varphi}\left(\frac{k_1}{2^j}+l_1\right)\right|^2\right)\left(\sum_{l_2\in\mathcal{Z}}\left|\hat{\psi}\left(\frac{k_2}{2^j}+l_2\right)\right|^2\right), \tag{A.23}$$

where $\Gamma=\{k\in\mathcal{Z}^2/1-2^{j-1}\leqq k_i\leqq 2^{j-1}, i=1,2\}$.

Thanks to Lemma 3

$$\|w_1\|_0^2=\sum_{k\in\Gamma}\left|m\left(\frac{k}{2^j}\right)\right|^2. \tag{A.24}$$

On the other hand, we write

$$4^{j(N+1)}\|w_1\|_{-N-1}^2=\sum_{l\in\mathcal{Z}_*^2}|\hat{w}_1(l)|^2\left|\frac{l}{2^j}\right|^{-2N-2}. \tag{A.25}$$

By the same computations as above, we obtain

$$\begin{aligned}4^{j(N+1)}\|w_1\|_{-N-1}^2=\sum_{k\in\Gamma}&\left|m\left(\frac{k}{2^j}\right)\right|^2\\ &\cdot\left(\sum_{l\in\mathcal{Z}^2}\left|\frac{k}{2^j}+l\right|^{-2N-2}\left|\hat{\varphi}\left(\frac{k_1}{2^j}+l_1\right)\right|^2\left|\hat{\psi}\left(\frac{k_2}{2^j}+l_2\right)\right|^2\right).\end{aligned} \tag{A.26}$$

We infer from (A.24) and (A.26) that to prove (A.16) for w_1 it is sufficient to majorize the function

$$\left[-\frac{1}{2},\frac{1}{2}\right]^2\to\mathcal{R}_+,\qquad z\mapsto\sum_{l\in\mathcal{Z}^2}|z+l|^{-2N-2}|\hat{\varphi}(z_1+l_1)|^2|\hat{\psi}(z_2+l_2)|^2.$$

Observing that $|z+l|^{-2N-2}\leqq 4^{N+1}$ for z in $[-\frac{1}{2},\frac{1}{2}]^2$ and $l\neq 0$ we obtain

$$\begin{aligned}&\sum_{l\neq 0}|z+l|^{-2N-2}|\hat{\varphi}(z_1+l_1)|^2|\hat{\psi}(z_2+l_2)|^2\\ &\leqq 4^{N+1}\left(\sum_{l_1\in\mathcal{Z}}|\hat{\varphi}(z_1+l_1)|^2\right)\left(\sum_{l_2\in\mathcal{Z}}|\hat{\psi}(z_2+l_2)|^2\right).\end{aligned} \tag{A.27}$$

We then apply Lemma 3 to obtain

$$\sum_{l\neq 0}|z+l|^{-2N-2}|\hat{\varphi}(z_1+l_1)|^2|\hat{\psi}(z_2+l_2)|^2\leqq 4^{N+1}. \tag{A.28}$$

Now we have to majorize

$$|z|^{-2N-2}|\hat{\varphi}(z_1)|^2|\hat{\psi}(z_2)|^2.$$

We infer from (2.5) and (A.1) that

$$|\hat{\varphi}(z_1)|^2=0(1),\qquad |\hat{\psi}(z_2)|^2=0(|z_2|^{2N+2}),$$

when $|z|\to 0$. This fact ends the proof.

Let us recall some results about Daubechies' compactly supported wavelets. (See [2], [13].)

For all $n\geqq 1$, there exist a couple of functions ψ_n, φ_n such that

$$\psi_n,\ \varphi_n\in C^n(\mathcal{R}), \tag{A.29}$$

$$\int x^m\psi_n(x)\,dx=0\quad\text{if } m\leqq n, \tag{A.30}$$

$$\psi_n,\ \varphi_n \text{ are compactly supported.} \tag{A.31}$$

(Actually there exist two constants $c_1, c_2 > 0$ such that the width of their support belongs to $[c_1 n, c_2 n]$.)

(A.32) The family $\{2^{j/2}\psi_n(2^j x - k)\}_{j,k\in\mathscr{Z}}$ is an orthonormal basis of $L^2(\mathscr{R})$.

(A.33) If we denote by $\tilde{W}_j$ the space spanned by the functions $\psi_n(2^j x - k)$; $k \in \mathscr{Z}$, then the family $\{2^{j/2}\varphi_n(2^j x - k)\}_{k\in\mathscr{Z}}$ is an orthonormal basis of $\bigoplus_{l<j} \tilde{W}_l$.

Now we are able to define the periodic Daubechies wavelet bases and to prove Theorem 1 in this case, provided n is large enough with respect to s. Actually, we just have to replace b_N by φ_n in the proof of Proposition 1 and ψ_N by ψ_n in the proof of Proposition 2. The multidimensional results follow.

We now recall the Littlewood-Paley wavelet basis (see [12], [13]). There exist a couple of functions φ, ψ belonging to the Schwartz class $\mathscr{S}(\mathscr{R})$ satisfying, respectively, (A.1), (A.30), for each integer m, (A.32) and (A.33). Moreover, $\hat{\varphi}$ and $\hat{\psi}$ are compactly supported. Then, with the same notation as above, for w in W_j

$$\hat{w}(l) = m(l/2^j)\hat{\psi}(l/2^j), \tag{A.34}$$

where m is a one-periodic function. This yields, for each s in $\mathscr{R}$,

$$\begin{aligned} |w|_s^2 &= \sum_{l\in\mathscr{Z}} |\hat{w}(l)|^2|l|^{2s} \\ &= \sum_{a2^j \leqq l \leqq b2^j} |\hat{w}(l)|^2|l|^{2s}, \end{aligned} \tag{A.35}$$

where a and b are independent of j. This fact yields to Proposition 1 and Proposition 2. We then deduce that the periodic Littlewood-Paley wavelets provide an unconditional basis for all Sobolev spaces $\dot{H}^s(\Pi)$. The multidimensional results follow.

REFERENCES

[1] G. Battle, *A block spin construction of ondelettes, Part* 1: *Lemarié functions*, Comm. Math. Phys., 110 (1987), pp. 601-615.

[2] I. Daubechies, *Orthonormal basis of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909-996.

[3] A. Debussche and M. Marion, *On the construction of families of approximate inertial manifolds*, J. Differential Equations, to appear.

[4] I. Flahaut, *Approximate inertial manifolds for the sine-Gordon equation*, J. Differential Integral Equations, 4 (1991), pp. 1169-1194.

[5] C. Foias, O. Manley, and R. Temam, *Modelling of the interaction of small and large eddies in two-dimensional turbulent flows*, RAIRO Math. Modél. Anal. Numér., 22 (1988), pp. 93-114.

[6] C. Foias, B. Nicolaenko, G. Sell, and R. Temam, *Inertial manifolds for the Kuramoto-Sivashinsky equation and an estimate of their lowest dimension*, J. Math. Pures Appl., 67 (1988), pp. 197-226.

[7] C. Foias, G. Sell, and R. Temam, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309-353.

[8] C. Foias, G. Sell, and E. Titi, *Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations*, J. Dynamics Differential Equations, 1 (1989), pp. 199-244.

[9] C. Foias and R. Temam, *Some analytic and geometric properties of the solutions of the evolution Navier-Stokes equations*, J. Math. Pures Appl., 58 (1979), pp. 339-368.

[10] S. Jaffard and Y. Meyer, *Bases d'ondelettes dans des ouverts de $\mathscr{R}^n$*, J. Math. Pures Appl., 68 (1989), pp. 95-108.

[11] P. G. Lemarié, *Ondelettes à localisation exponentielle*, J. Math. Pures Appl., 67 (1988), pp. 227-236.

[12] P. G. Lemarié and Y. Meyer, *Ondelettes et bases hilbertiennes*, Rev. Mat. Iberoamericana, 2 (1986), pp. 1-18.

[13] Y. Meyer, *Ondelettes et Operateurs* I: *Ondelettes*, Hermann, Paris, 1990.

[14] M. Marion and R. Temam, *Nonlinear Galerkin methods*: *the finite elements case*, Numer. Math., 57 (1990), pp. 205-226.

[15] B. NICOLAENKO, B. SCHEURER, AND R. TEMAM, *Some global dynamical properties of the Kuramoto-Sivashinsky equations: Nonlinear stability and attractors*, Phys. D, 16 (1985), pp. 155-183.

[16] ———, *Some global dynamical properties of a class of pattern formation equations*, Comm. Partial Differential Equations, 14 (1989), pp. 245-297.

[17] K. PROMISLOW, *Time analyticity and Gevrey regularity for solutions of a class of dissipative partial differential equations*, J. Nonlinear Anal.: Theory, Methods Appl., 16 (1991), pp. 959-980.

[18] R. TEMAM, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci., Vol. 68, Springer-Verlag, Berlin, New York 1988.

[19] ———, *Attractors for the Navier-Stokes equations, localization and approximation*, J. Fac. Sci. Univ. Tokyo, Sect. 1A, Math., 36 (1989), pp. 629-647.

[20] ———, *Inertial manifolds and multigrid methods*, SIAM J. Math. Anal., 21 (1990), pp. 154-178.

[21] ———, *Induced trajectories and approximate inertial manifolds*, RAIRO Math. Modél. Anal. Numér., 23 (1989), pp. 541-561.

[22] ———, *Navier-Stokes Equations*, North-Holland, Amsterdam, Third ed., 1984.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1482-1518, November 1992

006

HOMOGENIZATION AND TWO-SCALE CONVERGENCE*

GRÉGOIRE ALLAIRE†

Abstract. Following an idea of G. Nguetseng, the author defines a notion of "two-scale" convergence, which is aimed at a better description of sequences of oscillating functions. Bounded sequences in $L^2(\Omega)$ are proven to be relatively compact with respect to this new type of convergence. A corrector-type theorem (i.e., which permits, in some cases, replacing a sequence by its "two-scale" limit, up to a strongly convergent remainder in $L^2(\Omega)$) is also established. These results are especially useful for the homogenization of partial differential equations with periodically oscillating coefficients. In particular, a new method for proving the convergence of homogenization processes is proposed, which is an alternative to the so-called energy method of Tartar. The power and simplicity of the two-scale convergence method is demonstrated on several examples, including the homogenization of both linear and nonlinear second-order elliptic equations.

Key words. homogenization, two-scale convergence, periodic

AMS(MOS) subject classification. 35B40

Introduction. This paper is devoted to the homogenization of partial differential equations with periodically oscillating coefficients. This type of equation models various physical problems arising in media with a periodic structure. Quite often the size of the period is small compared to the size of a sample of the medium, and, denoting their ratio by ε, an asymptotic analysis, as $\varepsilon \to 0$, is required: namely, starting from a microscopic description of a problem, we seek a macroscopic, or averaged, description. From a mathematical point of view, we have a family of partial differential operators L_ε (with coefficients oscillating with period ε), and a family of solutions u_ε which, for a given domain Ω and source term f, satisfy

$$L_\varepsilon u_\varepsilon = f \quad \text{in } \Omega, \tag{0.1}$$

complemented by appropriate boundary conditions. Assuming that the sequence u_ε converges, in some sense, to a limit u, we look for a so-called homogenized operator $\bar{L}$ such that u is a solution of

$$\bar{L}u = f \quad \text{in } \Omega. \tag{0.2}$$

Passing from (0.1) to (0.2) is the homogenization process. (There is a vast body of literature on that topic; see [10], [40] for an introduction, and additional references.) Although homogenization is not restricted to the case of periodically oscillating operators (cf. the Γ-convergence of DeGiorgi [16], [17], the H-convergence of Tartar [42], [34], or the G-convergence of Spagnolo [41], [49]), we restrict our attention to that particular case. This allows the use of the well-known two-scale asymptotic expansion method [7], [10], [27], [40] in order to find the precise form of the homogenized operator $\bar{L}$. The key to that method is to postulate the following ansatz for u_ε:

$$u_\varepsilon(x) = u_0\left(x, \frac{x}{\varepsilon}\right) + \varepsilon u_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon^2 u_2\left(x, \frac{x}{\varepsilon}\right) + \cdots, \tag{0.3}$$

* Received by the editors November 5, 1991; accepted for publication (in revised form) February 24, 1992.

† Commissariat à l'Energie Atomique, Laboratoire d'Etudes Thermiques des Réacteurs, Service d'Etudes des Réacteurs et de Mathématiques Appliquées, Département de Mécanique et Technologie, Centre d'Etudes Nucléaires de Saclay, F-91191 Gif sur Yvette, Cédex, France.

where each term $u_i(x, y)$ is periodic in y. Then, inserting (0.3) in (0.1) and identifying powers of ε leads to a cascade of equations for each term u_i. In general, averaging with respect to y that for u_0 gives (0.2), and the precise form of $\bar{L}$ is computed with the help of a so-called cell equation in the unit period (see [10], [40] for details). This method is very simple and powerful, but unfortunately is formal since, a priori, the ansatz (0.3) does not hold true. Thus, the two-scale asymptotic expansion method is used only to guess the form of the homogenized operator $\bar{L}$, and other arguments are needed to prove the convergence of the sequence u_ε to u. To this end, the more general and powerful method is the so-called energy method of Tartar [42]. Loosely speaking, it amounts to multiplying equation (0.1) by special test functions (built with the solutions of the cell equation), and passing to the limit as $\varepsilon \to 0$. Although products of weakly convergent sequences are involved, we can actually pass to the limit thanks to some "compensated compactness" phenomenon due to the particular choice of test functions.

Despite its frequent success in the homogenization of many different types of equations, this way of proceeding is not entirely satisfactory. It involves two different steps, the formal derivation of the cell and homogenized equation, and the energy method, which have very little in common. In some cases, it is difficult to work out the energy method (the construction of adequate test functions could be especially tricky). The energy method does not take full advantage of the periodic structure of the problem (in particular, it uses very little information gained with the two-scale asymptotic expansion). The latter point is not surprising since the energy method was not conceived by Tartar for periodic problems, but rather in the more general (and more difficult) context of H-convergence. Thus, there is room for a more efficient homogenization method, dedicated to partial differential equations with periodically oscillating coefficients. The purpose of the present paper is to provide such a method that we call two-scale convergence method.

This new method relies on the following theorem, which was first proved by Nguetseng [36].

THEOREM 0.1. *Let u_ε be a bounded sequence in $L^2(\Omega)$ (Ω being an open set of $\mathbb{R}^N$). There exists a subsequence, still denoted by u_ε, and a function $u_0(x, y) \in L^2(\Omega \times Y)$ ($Y = (0\,; 1)^N$ is the unit cube) such that*

$$\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y u_0(x, y) \psi(x, y)\, dx\, dy \tag{0.4}$$

for any smooth function $\psi(x, y)$, which is Y-periodic in y. Such a sequence u_ε is said to two-scale converge to $u_0(x, y)$.

We provide a simple proof of Theorem 0.1 along with a new corrector result.

THEOREM 0.2. *Let u_ε be a sequence that two-scale converges to $u_0(x, y)$. Then, u_ε weakly converges in $L^2(\Omega)$ to $u(x) = \int_Y u_0(x, y)\, dy$, and we have*

$$\lim_{\varepsilon \to 0} \|u_\varepsilon\|_{L^2(\Omega)} \geqq \|u_0\|_{L^2(\Omega \times Y)} \geqq \|u\|_{L^2(\Omega)}. \tag{0.5}$$

Furthermore, if equality is achieved in the left part of (0.5), *namely,*

$$\lim_{\varepsilon \to 0} \|u_\varepsilon\|_{L^2(\Omega)} = \|u_0\|_{L^2(\Omega \times Y)}, \tag{0.6}$$

and if $u_0(x, y)$ is smooth, then we have

$$\lim_{\varepsilon \to 0} \left\| u_\varepsilon(x) - u_0\left(x, \frac{x}{\varepsilon}\right) \right\|_{L^2(\Omega)} = 0. \tag{0.7}$$

Loosely speaking, Theorem 0.1 is a rigorous justification of the first term in the ansatz (0.3), while Theorem 0.2 gives the condition of a strong convergence to zero of the difference between u_ε and its ansatz. We are now equipped to explain the two-scale convergence method. We multiply equation (0.1) by a test function of the type $\psi(x, x/\varepsilon)$, where $\psi(x, y)$ is a smooth function, Y-periodic in y. After some integration by parts, we pass to the two-scale limit with the help of Theorem 0.1. In the limit, we read off a variational formulation for $u_0(x, y)$. The corresponding partial differential equation is called the two-scale homogenized problem. It is usually of the same type as the original problem (0.1), but it involves two variables x and y. Thus, averaging with respect to y leads to the homogenized problem (0.2). Eventually, so-called corrector results (i.e., strong or pointwise convergences) can be obtained by the application of Theorem 0.2.

We emphasize that the two-scale convergence method is self-contained, i.e., in a single process we find the homogenized equation and we prove convergence. This is in contrast with the former "usual" homogenization process (as described above) which is divided in two steps: first, find the homogenized and cell equations by means of asymptotic expansions; second, prove convergence with the energy method. Another interesting feature of the two-scale convergence method is the introduction of the two-scale homogenized problem. It turns out that it is a well-posed system of equations, which are a combination of the usual homogenized and cell equations. Indeed, if it is expected that the periodic oscillations in the operator L_ε generate only the same type of oscillations in the solution u_ε, the sequence u_ε is completely characterized by its two-scale limit $u_0(x, y)$. Thus, starting from a well-posed problem for u_ε, we should obtain in the limit a well-posed problem of the same type for u_0. However, this is not always the case for the usual macroscopic homogenized equation (the solution of which is $u(x) = \int_Y u_0(x, y)\, dy$). When averaging the two-scale homogenized problem with respect to y, its "nice" form can disappear, and, rather, we could obtain integro-differential terms (corresponding to memory effects), nonlocal terms, or nonexplicit equations. There are many such examples in the literature (see [5], [29], [32], [46], where "classical" methods are used, and [2], [3], [37], where two-scale convergence is applied). In these cases, the two-scale homogenized problem explains and simplifies the complicated form of the macroscopic limit equation, thanks to the additional microscopic variable y, which plays the role of a hidden variable.

Since Theorem 0.1 proves the existence of the first term in the ansatz (0.3), the two-scale convergence method appears as the mathematically rigorous version of the, intuitive and formal, two-scale asymptotic expansion method [7], [10], [27], [40]. The key of the success for such a method is to consider only periodic homogenization problems. This amounts to restricting the class of possible oscillations of the solutions to purely periodic ones. Working with the relatively small class of periodic oscillations allows us to obtain the representation formula (0.4) for weak limits of solutions. For general types of oscillations, a result like (0.4) seems to be out of reach (the main obstacle being how to choose the test functions). On the other hand, periodic homogenization can be cast into the framework of quasi-periodic, or almost-periodic (in the sense of Besicovitch) homogenization (see, e.g., [28], [38]), since periodic functions are a very special subclass of quasi-, or almost-, periodic functions. In this case, test functions can also be written $\psi(x, x/\varepsilon)$, where $\psi(x, y)$ is quasi-, or almost-, periodic in y. However, we do not know if Theorem 0.1 can be generalized to such test functions or if a new convergence method can thus be obtained.

The paper is organized as follows. Section 1 is devoted to the proof of Theorems 0.1 and 0.2, and other related results. In § 2, we show precisely how the two-scale convergence method works on the homogenization of linear second-order elliptic equations (this is the favorite model problem in homogenization; see, e.g., Chapter 1 in [10]). We do this in a fixed domain Ω, but also in a periodically perforated domain Ω_ε (a porous medium), obtained by removing from Ω infinitely many small holes of size ε (their number is of order ε^{-N}), which support a Neumann boundary condition. Two-scale convergence is particularly well adapted to the latter case, and we recover previous results (see [13], [1], [4]) without using any extension techniques. Section 3 generalizes § 2 to the nonlinear case. In the periodic setting, we give a new proof of the Γ-convergence of convex energies (see [31], [16], [17]), and we revisit the homogenization of monotone operators [42]. On the contrary of §§ 2 and 3, § 4 deals with an example of homogenization where typical two-scale phenomena appear. We consider a linear elliptic second-order equation with periodic coefficients taking only two values 1 and ε^2. It models a diffusion process in a medium made of two highly heterogeneous materials. It turns out that the limit diffusion process is of a very special type: the usual homogenized problem is not an explicit partial differential equation. Finally, § 5 is devoted to the proof of a technical lemma used in § 1; more generally, we investigate under which regularity assumptions on a Y-periodic function $\psi(x, y)$ the following convergence holds true:

$$\lim_{\varepsilon\to 0}\int_\Omega \left|\psi\left(x,\frac{x}{\varepsilon}\right)\right| dx=\int_\Omega\int_Y |\psi(x,y)|\,dx\,dy. \tag{0.8}$$

It is easily seen that continuous functions satisfy (0.8). We prove that (0.8) still holds true for functions of $L^1[\Omega; C_\#(Y)]$ or $L^1_\#[Y; C(\bar{\Omega})]$, which are continuous in only one variable, x or y. However, we cannot decrease the regularity of $\psi(x, y)$ too much. Indeed, we construct a counterexample to (0.8) for a function $\psi(x, y)$ of $C[\Omega; L^1_\#(Y)]$, which is not continuous in x for any value of y, but merely continuous in x in the "$L^1(Y)$-mean."

1. Two-scale convergence.

Let us begin this section with a few notations. Throughout this paper Ω is an open set of $\mathbb{R}^N (N \geqq 1)$, and $Y=[0; 1]^N$ is the closed unit cube. As usual, $L^2(\Omega)$ is the Sobolev space of real-valued functions that are measurable and square summable in Ω with respect to the Lebesgue measure. We denote by $C^\infty_\#(Y)$ the space of infinitely differentiable functions in $\mathbb{R}^N$ that are periodic of period Y. Then, $L^2_\#(Y)$ (respectively, $H^1_\#(Y)$) is the completion for the norm of $L^2(Y)$ (respectively, $H^1(Y)$) of $C^\infty_\#(Y)$. Remark that $L^2_\#(Y)$ actually coincides with the space of functions in $L^2(Y)$ extended by Y-periodicity to the whole of $\mathbb{R}^N$.

Let us consider a sequence of functions u_ε in $L^2(\Omega)$ (ε is a sequence of strictly positive numbers which goes to zero). Following the lead of Nguetseng [36], we introduce the following.

Definition 1.1. A sequence of functions u_ε in $L^2(\Omega)$ is said to *two-scale converge* to a limit $u_0(x, y)$ belonging to $L^2(\Omega\times Y)$ if, for any function $\psi(x, y)$ in $D[\Omega; C^\infty_\#(Y)]$, we have

$$\lim_{\varepsilon\to 0}\int_\Omega u_\varepsilon(x)\psi\left(x,\frac{x}{\varepsilon}\right) dx=\int_\Omega\int_Y u_0(x,y)\psi(x,y)\,dx\,dy. \tag{1.1}$$

This new notion of "two-scale convergence" makes sense because of the next compactness theorem.

Theorem 1.2. *From each bounded sequence u_ε in $L^2(\Omega)$, we can extract a subsequence, and there exists a limit $u_0(x, y)\in L^2(\Omega\times Y)$ such that this subsequence two-scale converges to u_0.*

To establish Theorem 1.2, we need the following lemma, the proof of which may be found in § 5.

LEMMA 1.3. *Let $\psi(x, y)$ be a function in $L^2[\Omega; C_\#(Y)]$, i.e., measurable and square summable in $x \in \Omega$, with values in the Banach space of continuous functions, Y-periodic in y. Then, for any positive value of ε, $\psi(x, x/\varepsilon)$ is a measurable function on Ω, and we have*

$$(1.2)\qquad \left\| \psi\left(x, \frac{x}{\varepsilon}\right) \right\|_{L^2(\Omega)} \leqq \|\psi(x, y)\|_{L^2[\Omega;\, C_\#(Y)]} \equiv \left[\int_\Omega \operatorname*{Sup}_{y \in Y} |\psi(x, y)|^2 \, dx \right]^{1/2}$$

and

$$(1.3)\qquad \lim_{\varepsilon \to 0} \int_\Omega \psi\left(x, \frac{x}{\varepsilon}\right)^2 dx = \int_\Omega \int_Y \psi(x, y)^2 \, dx \, dy.$$

DEFINITION 1.4. A function $\psi(x, y)$, Y-periodic in y, and satisfying (1.3), is called an "admissible" test function.

It is well known (and easy to prove) that a continuous function $\psi(x, y)$ on $\Omega \times Y$, Y-periodic in y, satisfies (1.3). However, the situation is not so clear if the regularity of ψ is weakened: in particular, the measurability of $\psi(x, x/\varepsilon)$ is not obvious. To our knowledge, the minimal regularity hypothesis (if any) making of $\psi(x, y)$ an "admissible" test function is not known. In order that the right-hand side of (1.3) makes sense, $\psi(x, y)$ must at least belong to $L^2(\Omega \times Y)$ (in addition to being Y-periodic in y). But, as we shall see in § 5, this is not enough for (1.3) to hold (a counterexample is provided in Proposition 5.8). Loosely speaking, $\psi(x, y)$ turns out to be an "admissible" test function if it is continuous in one of its arguments (as is the case when ψ belongs to $L^2[\Omega; C_\#(Y)]$). For more details, see § 5, which is devoted to the proof of Lemma 1.3 and to the investigation of other regularity assumptions making of ψ an "admissible" test function.

Proof of Theorem 1.2. Let u_ε be a bounded sequence in $L^2(\Omega)$: there exists a positive constant C such that

$$\|u_\varepsilon\|_{L^2(\Omega)} \leqq C.$$

For any function $\psi(x, y) \in L^2[\Omega; C_\#(Y)]$, according to Lemma 1.3, $\psi(x, x/\varepsilon)$ belongs to $L^2(\Omega)$, and the Schwarz inequality yields

$$(1.4)\qquad \left| \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx \right| \leqq C \left\| \psi\left(x, \frac{x}{\varepsilon}\right) \right\|_{L^2(\Omega)} \leqq C \|\psi(x, y)\|_{L^2[\Omega; C_\#(Y)]}.$$

Thus, for fixed ε, the left-hand side of (1.4) turns out to be a bounded linear form on $L^2[\Omega; C_\#(Y)]$. The dual space of $L^2[\Omega; C_\#(Y)]$ can be identified with $L^2[\Omega; M_\#(Y)]$, where $M_\#(Y)$ is the space of Y-periodic Radon measures on Y. By virtue of the Riesz representation theorem, there exists a unique function $\mu_\varepsilon \in L^2[\Omega; M_\#(Y)]$ such that

$$(1.5)\qquad \langle \mu_\varepsilon, \psi \rangle = \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx,$$

where the brackets in the left-hand side of (1.5) denotes the duality product between $L^2[\Omega; C_\#(Y)]$ and its dual. Furthermore, in view of (1.4), the sequence μ_ε is bounded in $L^2[\Omega; M_\#(Y)]$. Since the space $L^2[\Omega; C_\#(Y)]$ is separable (i.e., contains a dense countable family), from any bounded sequence of its dual we can extract a subsequence that converges for the weak* topology. Thus, there exists $\mu_0 \in L^2[\Omega; M_\#(Y)]$ such that, up to a subsequence, and for any $\psi \in L^2[\Omega; C_\#(Y)]$,

$$(1.6)\qquad \langle \mu_\varepsilon, \psi \rangle \to \langle \mu_0, \psi \rangle.$$

By combining (1.5) and (1.6) we obtain, up to a subsequence, and for any $\psi \in L^2[\Omega; C_{\#}(Y)]$,

$$\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx = \langle \mu_0, \psi \rangle. \tag{1.7}$$

From Lemma 1.3 we know that

$$\lim_{\varepsilon \to 0} \left\| \psi\left(x, \frac{x}{\varepsilon}\right) \right\|_{L^2(\Omega)} = \|\psi(x, y)\|_{L^2(\Omega \times Y)}. \tag{1.8}$$

Now, passing to the limit in the first two terms of (1.4) with the help of (1.7) and (1.8), we deduce

$$|\langle \mu_0, \psi \rangle| \leqq C \|\psi\|_{L^2(\Omega \times Y)}.$$

By density of $L^2[\Omega; C_{\#}(Y)]$ in $L^2(\Omega \times Y)$, and by the Riesz representation theorem, μ_0 is identified with a function $u_0 \in L^2(\Omega \times Y)$, i.e.,

$$\langle \mu_0, \psi \rangle = \int_\Omega \int_Y u_0(x, y) \psi(x, y)\, dx\, dy. \tag{1.9}$$

Equalities (1.7) and (1.9) are the desired result. □

Remark 1.5. In the proof of Theorem 1.2, we considered test functions $\psi(x, y)$ in $L^2[\Omega; C_{\#}(Y)]$. Other choices of space of test functions are actually possible. For example, in the case where Ω is bounded, we could have replaced $L^2[\Omega; C_{\#}(Y)]$ by $C[\bar{\Omega}; \mathrm{C}_{\#}\ (\mathrm{Y})]$, or by $L^2_{\#}[Y; C(\bar{\Omega})]$. The main ingredients of the proof would not be affected by this change. All these spaces have in common that they are separable Banach spaces, which is the required property in order to extract a weakly * convergent subsequence from any bounded sequence in their dual. In any case the two-scale limit $u_0(x, y)$ is always the same, whatever the chosen space of test functions (see Remark 1.11).

Before developing further the theory, let us give a few examples of two-scale limits.

(*) For any smooth function $a(x, y)$, being Y-periodic in y, the associated sequence $a_\varepsilon(x) = a(x, x/\varepsilon)$ two-scale converges to $a(x, y)$.

(**) Any sequence u_ε that converges strongly in $L^2(\Omega)$ to a limit $u(x)$, two-scale converges to the same limit $u(x)$.

(***) Any sequence u_ε that admits an asymptotic expansion of the type $u_\varepsilon(x) = u_0(x, x/\varepsilon) + \varepsilon u_1(x, x/\varepsilon) + \varepsilon^2 u_2(x, x/\varepsilon) + \cdots$, where the functions $u_i(x, y)$ are smooth and Y-periodic in y, two-scale converges to the first term of the expansion, namely, $u_0(x, y)$.

In view of the third example we already have a flavour of the main interest of two-scale convergence: even if the above asymptotic expansion does not hold (or is unknown), it is permited to rigorously justify the existence of its first term $u_0(x, y)$. This is very helpful in homogenization theory, where such asymptotic expansions are frequently used in a heuristical way (see [10], [40]). This remark is the key of our two-scale convergence method, as explained in §§ 2, 3, and 4.

The next proposition establishes a link between two-scale and weak L^2-convergences.

PROPOSITION 1.6. *Let u_ε be a sequence of functions in $L^2(\Omega)$, which two-scale converges to a limit $u_0(x, y) \in L^2(\Omega \times Y)$. Then u_ε converges also to $u(x) = \int_Y u_0(x, y)\, dy$ in $L^2(\Omega)$ weakly. Furthermore, we have*

$$\lim_{\varepsilon \to 0} \|u_\varepsilon\|_{L^2(\Omega)} \geqq \|u_0\|_{L^2(\Omega \times Y)} \geqq \|u\|_{L^2(\Omega)}. \tag{1.10}$$

Proof. By taking test functions $\psi(x)$, which depends only on x, in (1.1), we immediately obtain that u_ε weakly converges to $u(x)=\int_Y u_0(x,y)\,dy$ in $L^2(\Omega)$. To obtain (1.10), for $\psi(x,y)\in L^2[\Omega;\, C_\#(Y)]$, we compute

$$\int_\Omega \left[u_\varepsilon(x)-\psi\left(x,\frac{x}{\varepsilon}\right)\right]^2 dx=\int_\Omega u_\varepsilon(x)^2\,dx+\int_\Omega \psi\left(x,\frac{x}{\varepsilon}\right)^2 dx$$

$$-2\int_\Omega u_\varepsilon(x)\psi\left(x,\frac{x}{\varepsilon}\right) dx \geqq 0.$$

Passing to the limit as $\varepsilon\to 0$ yields

$$\lim_{\varepsilon\to 0}\int_\Omega u_\varepsilon(x)^2\,dx \geqq 2\int_\Omega\int_Y u_0(x,y)\psi(x,y)\,dx\,dy-\int_\Omega\int_Y \psi(x,y)^2\,dx\,dy.$$

Then, using a sequence of smooth functions that converges strongly to u_0 in $L^2(\Omega\times Y)$ leads to

$$\lim_{\varepsilon\to 0}\int_\Omega u_\varepsilon(x)^2\,dx\geqq\int_\Omega\int_Y u_0(x,y)^2\,dx\,dy.$$

On the other hand, the Cauchy–Schwarz inequality in Y gives the other inequality in (1.10). □

Remark 1.7. From Proposition 1.6, we see that, for a given bounded sequence in $L^2(\Omega)$, there is more information in its two-scale limit u_0 than in its weak L^2 limit u: u_0 contains some knowledge on the periodic oscillations of u_ε, while u is just the average (with respect to y) of u_0. However, let us emphasize that the two-scale limit captures only the oscillations that are in resonance with those of the test functions $\psi(x, x/\varepsilon)$. Contrary to the example (*) above, the sequence defined by $b_\varepsilon(x)=a(x, x/\varepsilon^2)$ (where $a(x,y)$ is a smooth function, Y-periodic in y) has the same two-scale limit and weak L^2 limit, namely, $\int_Y a(x,y)\,dy$. (This is a consequence of the difference of orders in the speed of oscillations for b_ε and the test functions $\psi(x, x/\varepsilon)$.) In this example, no oscillations are captured because the two-scale limit depends only on the variable x. Remark also here that the independence of the two-scale limit on the "fast" variable y does not imply strong convergence of the sequence in $L^2(\Omega)$.

We claim that there is more information in the two-scale limit of a sequence than in its weak L^2 limit. But does this supplementary knowledge yield some kind of strong convergence? This question is precisely answered by the following theorem.

THEOREM 1.8. *Let u_ε be a sequence of functions in $L^2(\Omega)$ that two-scale converges to a limit $u_0(x,y)\in L^2(\Omega\times Y)$. Assume that*

$$\lim_{\varepsilon\to 0}\|u_\varepsilon\|_{L^2(\Omega)}=\|u_0\|_{L^2(\Omega\times Y)}. \tag{1.11}$$

Then, for any sequence v_ε that two-scale converges to a limit $v_0(x,y)\in L^2(\Omega\times Y)$, we have

$$u_\varepsilon(x)v_\varepsilon(x)\rightharpoonup\int_Y u_0(x,y)v_0(x,y)\,dy \quad \text{in } D'(\Omega). \tag{1.12}$$

Furthermore, if $u_0(x,y)$ belongs to $L^2[\Omega;\, C_\#(Y)]$, we have

$$\lim_{\varepsilon\to 0}\left\|u_\varepsilon(x)-u_0\left(x,\frac{x}{\varepsilon}\right)\right\|_{L^2(\Omega)}=0. \tag{1.13}$$

Remark 1.9. The condition (1.11) can be interpreted as "u_0 contains all the oscillations of the sequence u_ε." Indeed, (1.11) always takes place for a sequence $\psi(x, x/\varepsilon)$, with $\psi(x,y)\in L^2[\Omega;\, C_\#(Y)]$ or, more generally, being an "admissible" test function in the sense of Definition 1.4. The result (1.12) can be defined as a *strong* two-scale convergence for the sequence u_ε; remarkably, it allows to pass to the limit in some product of two weak convergences in $L^2(\Omega)$.

Remark 1.10. As already pointed out before, for a given ε, the function $u_0(x, x/\varepsilon)$ need not be measurable in Ω, if $u_0(x, y)$ merely belongs to $L^2(\Omega\times Y)$. Thus, in order for (1.13) to make sense, some regularity on u_0 is required; more precisely, we restrict ourselves to functions $u_0(x, y)$ in $L^2[\Omega;\ C_\#(Y)]$ (more generally, $u_0(x, y)$ could be any "admissible" test function; see § 5 for details). However, we could wonder if all two-scale limits automatically are "admissible" test functions. Unfortunately, this is not true, and Lemma 1.13 below shows that any function in $L^2(\Omega\times Y)$ is attained as a two-scale limit. In view of the counterexample of Proposition 5.8, it is clear that, in general, a function of $L^2(\Omega\times Y)$ is not "admissible" in the sense of Definition 1.4. Thus, we cannot avoid an assumption on the regularity of u_0 in order to state (1.13).

Finally, we claim that, in the vocabulary of homogenization, (1.13) is a corrector-type result. Indeed, the sequence u_ε is approximated by its two-scale limit $u_0(x, x/\varepsilon)$ up to a strongly convergent reminder in $L^2(\Omega)$. Thus, the weak L^2-convergence of u_ε to its weak limit u is improved by (1.13), and the precise corrector is $u_0(x, x/\varepsilon)-u(x)$.

Proof of Theorem 1.8. Let $\psi_n(x, y)$ be a sequence of smooth functions in $L^2[\Omega;\ C_\#(Y)]$ that converges strongly to $u_0(x, y)$ in $L^2(\Omega\times Y)$. By definition of two-scale convergence for u_ε, and using Lemma 1.3 and assumption (1.11), we obtain

$$\lim_{\varepsilon\to 0}\int_\Omega\left[u_\varepsilon(x)-\psi_n\left(x,\frac{x}{\varepsilon}\right)\right]^2 dx=\int_\Omega\int_Y[u_0(x,y)-\psi_n(x,y)]^2\,dx\,dy. \tag{1.14}$$

Passing to the limit as n goes to infinity, (1.14) yields

$$\lim_{n\to\infty}\lim_{\varepsilon\to 0}\int_\Omega\left[u_\varepsilon(x)-\psi_n\left(x,\frac{x}{\varepsilon}\right)\right]^2 dx=0. \tag{1.15}$$

Let v_ε be a sequence that two-scale converges to a limit $v_0(x, y)$. For any $\phi(x)\in D(\Omega)$, we have

$$\begin{aligned}\int_\Omega\phi(x)u_\varepsilon(x)v_\varepsilon(x)\,dx&=\int_\Omega\phi(x)\psi_n\left(x,\frac{x}{\varepsilon}\right)v_\varepsilon(x)\,dx\\&\quad+\int_\Omega\phi(x)\left[u_\varepsilon(x)-\psi_n\left(x,\frac{x}{\varepsilon}\right)\right]v_\varepsilon(x)\,dx.\end{aligned}$$

Passing to the limit as ε goes to zero (and having in mind that v_ε is a bounded sequence in $L^2(\Omega)$) yields

$$\begin{aligned}&\left|\lim_{\varepsilon\to 0}\int_\Omega\phi(x)u_\varepsilon(x)v_\varepsilon(x)\,dx-\int_\Omega\int_Y\phi(x)\psi_n(x,y)v_0(x,y)\,dx\,dy\right|\\&\quad\leqq C\lim_{\varepsilon\to 0}\left\|u_\varepsilon(x)-\psi_n\left(x,\frac{x}{\varepsilon}\right)\right\|_{L^2(\Omega)}.\end{aligned}$$

Next, passing to the limit when n goes to infinity and using (1.15) leads to (1.12), i.e.,

$$\lim_{\varepsilon\to 0}\int_\Omega\phi(x)u_\varepsilon(x)v_\varepsilon(x)\,dx=\int_\Omega\int_Y\phi(x)u_0(x,y)v_0(x,y)\,dx\,dy.$$

Furthermore, if $u_0(x, y)$ is smooth, say $u_0\in L^2[\Omega;\ C_\#(Y)]$, then (1.14) applies directly with u_0 instead of ψ_n, and it is nothing but (1.13). □

Remark 1.11. As a consequence of Theorem 1.8, we can enlarge the class of test functions $\psi(x, y)$ used in the definition of two-scale convergence. In Definition 1.1, a sequence u_ε two-scale converges to a limit u_0 if

$$(1.16) \qquad \lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x)\psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y u_0(x, y)\psi(x, y)\, dx\, dy$$

for any smooth test function ψ, namely, for $\psi(x, y) \in D[\Omega; C_\#^\infty(Y)]$. The class of test functions has already been considerably enlarged since the compactness Theorem 1.2 is proved for any $\psi(x, y) \in L^2[\Omega; C_\#(Y)]$. In view of Theorem 1.8, the validity of (1.16) is extended to all "admissible" test functions ψ in the sense of Definition 1.4. Indeed, an admissible test function satisfies hypothesis (1.11) in Theorem 1.8, and thus the sequence $\psi(x, x/\varepsilon)$ two-scale converges strongly to $\psi(x, y)$. Retrospectively, the choice of the space $L^2[\Omega; C_\#(Y)]$ in the proof of Theorem 1.2 appears to be purely technical: other choices would have led to the same two-scale limit.

Remark 1.12. Let us conclude this section by some bibliographical comments. As already said, the notion of two-scale convergence and the proof of the compactness Theorem 1.2 go back to Nguetseng [36]. Here we present a new proof of Theorem 1.2, which is simpler than the original one (note in passing that our proof has some similarities with that of Ball [8] for the existence of Young measures). Proposition 1.6 and Theorem 1.8 (concerning corrector results) are new. Recently, a generalization of two-scale convergence to Young measures has been introduced by E [19] in order to handle homogenization of nonlinear hyperbolic conservation laws (see Remark 3.8). Various authors have also developed ideas similar to two-scale convergence: Arbogast, Douglas, and Hornung [6] defined a so-called dilation operator for homogenization problems in porous media, while Mascarenhas [32] introduced a kind of two-scale Γ-convergence in the study of some memory effects in homogenization. All these works can be embedded in the general setting of two-scale convergence.

Now that the basic tools of the two-scale convergence method have been established, we give a few complementary results before explaining how it can be applied to the homogenization of partial differential equations with periodically oscillating coefficients. We first prove that two-scale limits have no extra regularity, as announced in Remark 1.10.

LEMMA 1.13. *Any function $u_0(x, y)$ in $L^2(\Omega \times Y)$ is attained as a two-scale limit.*

Proof. For any function $u_0(x, y) \in L^2(\Omega \times Y)$, we shall construct a bounded sequence u_ε in $L^2(\Omega)$ that two-scale converges to u_0. Let $u_n(x, y)$ be a sequence of smooth, Y-periodic in y functions that converge strongly to u_0 in $L^2(\Omega \times Y)$. Let $[\psi_k(x, y)]_{1 \leqq k \leqq \infty}$ be a dense family of smooth, Y-periodic in y functions in $L^2(\Omega \times Y)$, normalized such that $\|\psi_k\|_{L^2(\Omega \times Y)} = 1$. Obviously, for fixed n, the sequence $u_n(x, x/\varepsilon)$ two-scale converges to $u_n(x, y)$, i.e., for any $\delta > 0$, and for any smooth $\psi(x, y)$, there exists $\varepsilon_0(n, \delta, \psi) > 0$ such that $\varepsilon < \varepsilon_0$ implies

$$\left| \int_\Omega u_n\left(x, \frac{x}{\varepsilon}\right) \psi\left(x, \frac{x}{\varepsilon}\right) dx - \int_\Omega \int_Y u_n(x, y)\psi(x, y)\, dx\, dy \right| \leqq \delta.$$

Now, we extract a diagonal sequence; namely, fixing $\delta_n = \|u_n - u_0\|_{L^2(\Omega \times Y)}$, there exists a sequence of positive numbers $\varepsilon(n)$, which goes to zero as $n \to \infty$ such that

$$\left| \int_\Omega u_n\left(x, \frac{x}{\varepsilon(n)}\right)^2 dx - \int_\Omega \int_Y u_n(x, y)^2\, dx\, dy \right| \leqq \delta_n$$

$$(1.17) \quad \left| \int_\Omega u_n\left(x, \frac{x}{\varepsilon(n)}\right) \psi_k\left(x, \frac{x}{\varepsilon(n)}\right) dx - \int_\Omega \int_Y u_n(x, y)\psi_k(x, y)\, dx\, dy \right| \leqq \delta_n$$

for $1 \leqq k \leqq n$.

Defining the diagonal sequence $u_{\varepsilon(n)}(x) \equiv u_n(x, x/\varepsilon(n))$, and recalling that δ_n is a sequence of positive numbers that goes to zero, it is clear from the first line of (1.17) that the sequence $u_{\varepsilon(n)}$ is bounded in $L^2(\Omega)$. By density of the family $[\psi_k(x, y)]_{1 \leqq k \leqq \infty}$ in $L^2(\Omega \times Y)$, the second line implies that $u_{\varepsilon(n)}$ two-scale converges to u_0. □

So far we have only considered bounded sequences in $L^2(\Omega)$. The next proposition investigates some cases where we have additional bounds on sequences of derivatives.

PROPOSITION 1.14.

(i) *Let u_ε be a bounded sequence in $H^1(\Omega)$ that converges weakly to a limit u in $H^1(\Omega)$. Then, u_ε two-scale converges to $u(x)$, and there exists a function $u_1(x, y)$ in $L^2[\Omega; H^1_\#(Y)/\mathbb{R}]$ such that, up to a subsequence, ∇u_ε two-scale converges to $\nabla_x u(x) + \nabla_y u_1(x, y)$.*

(ii) *Let u_ε and $\varepsilon \nabla u_\varepsilon$ be two bounded sequences in $L^2(\Omega)$. Then, there exists a function $u_0(x, y)$ in $L^2[\Omega; H^1_\#(Y)]$ such that, up to a subsequence, u_ε and $\varepsilon \nabla u_\varepsilon$ two-scale converge to $u_0(x, y)$ and to $\nabla_y u_0(x, y)$, respectively.*

(iii) *Let u_ε be a divergence-free bounded sequence in $[L^2(\Omega)]^N$, which two-scale converges to $u_0(x, y)$ in $[L^2(\Omega \times Y)]^N$. Then, the two-scale limit satisfies* $\operatorname{div}_y u_0(x, y) = 0$ *and* $\int_Y \operatorname{div}_x u_0(x, y)\, dy = 0$.

Proof.

(i) Since u_ε (respectively, ∇u_ε) is bounded in $L^2(\Omega)$ (respectively, $[L^2(\Omega)]^N$), up to a subsequence, it two-scale converges to a limit $u_0(x, y) \in L^2(\Omega \times Y)$ (respectively, $\chi_0(x, y) \in [L^2(\Omega \times Y)]^N$). Thus for any $\psi(x, y) \in D[\Omega; C^\infty_\#(Y)]$ and any $\Psi(x, y) \in D[\Omega; C^\infty_\#(Y)]^N$, we have

$$(1.18) \quad \begin{aligned} &\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y u_0(x, y) \psi(x, y)\, dx\, dy, \\ &\lim_{\varepsilon \to 0} \int_\Omega \nabla u_\varepsilon(x) \cdot \Psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y \chi_0(x, y) \cdot \Psi(x, y)\, dx\, dy. \end{aligned}$$

By integration by parts, we have

$$\varepsilon \int_\Omega \nabla u_\varepsilon(x) \cdot \Psi\left(x, \frac{x}{\varepsilon}\right) dx = -\int_\Omega u_\varepsilon(x) \left[\operatorname{div}_y \Psi\left(x, \frac{x}{\varepsilon}\right) + \varepsilon \operatorname{div}_x \Psi\left(x, \frac{x}{\varepsilon}\right)\right] dx.$$

Passing to the limit in both terms with the help of (1.18) leads to

$$0 = -\int_\Omega \int_Y u_0(x, y) \operatorname{div}_y \Psi(x, y)\, dx\, dy.$$

This implies that $u_0(x, y)$ does not depend on y. Since the average of u_0 is u, we deduce that for any subsequence the two-scale limit reduces to the weak L^2 limit u. Thus, the entire sequence u_ε two-scale converges to $u(x)$. Next, in (1.18) we choose a function Ψ such that $\operatorname{div}_y \Psi(x, y) = 0$. Integrating by parts we obtain

$$\begin{aligned} \lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \operatorname{div}_x \Psi\left(x, \frac{x}{\varepsilon}\right) dx &= -\int_\Omega \int_Y \chi_0(x, y) \cdot \Psi(x, y)\, dx\, dy \\ &= \int_\Omega \int_Y u(x) \operatorname{div}_x \Psi(x, y)\, dx\, dy. \end{aligned}$$

Thus, for any function $\Psi(x, y) \in D[\Omega; C^\infty_\#(Y)]^N$ with $\operatorname{div}_y \Psi(x, y) = 0$, we have

$$\int_\Omega \int_Y [\chi_0(x, y) - \nabla u(x)] \cdot \Psi(x, y)\, dx\, dy = 0. \tag{1.19}$$

Recall that the orthogonal of divergence-free functions are exactly the gradients (see, if necessary, [43] or [47]). This well-known result can be very easily proved in the present context by means of Fourier analysis in Y. Thus, we deduce from (1.19) that there exists a unique function $u_1(x, y)$ in $L^2[\Omega; H^1_\#(Y)/\mathbb{R}]$ such that

$$\chi_0(x, y) = \nabla u(x) + \nabla_y u_1(x, y).$$

(ii) Since u_ε (respectively, $\varepsilon \nabla u_\varepsilon$) is bounded in $L^2(\Omega)$ (respectively, $[L^2(\Omega)]^N$), up to a subsequence, it two-scale converges to a limit $u_0(x, y) \in L^2(\Omega \times Y)$ (respectively, $\chi_0(x, y) \in [L^2(\Omega \times Y)]^N$). Thus for any $\psi(x, y) \in D[\Omega; C^\infty_\#(Y)]$ and any $\Psi(x, y) \in D[\Omega; C^\infty_\#(Y)]^N$, we have

$$\tag{1.20} \begin{aligned} &\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y u_0(x, y) \psi(x, y)\, dx\, dy, \\ &\lim_{\varepsilon \to 0} \int_\Omega \varepsilon \nabla u_\varepsilon(x) \cdot \Psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y \chi_0(x, y) \cdot \Psi(x, y)\, dx\, dy. \end{aligned}$$

Integrating by parts in (1.20), we obtain

$$\begin{aligned} \lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \left[\operatorname{div}_y \Psi\left(x, \frac{x}{\varepsilon}\right) + \varepsilon \operatorname{div}_x \Psi\left(x, \frac{x}{\varepsilon}\right)\right] dx &= -\int_\Omega \int_Y \chi_0(x, y) \cdot \Psi(x, y)\, dx\, dy \\ &= \int_\Omega \int_Y u_0(x, y) \operatorname{div}_y \Psi(x, y)\, dx\, dy. \end{aligned}$$

Disintegrating by parts leads to $\chi_0(x, y) = \nabla_y u_0(x, y)$.

The proof of part (iii) is similar to the previous ones, and is left to the reader. □

Two-scale convergence is not limited to bounded sequences in $L^2(\Omega)$. Our main result, Theorem 1.2, is easily generalized to bounded sequences in $L^p(\Omega)$, with $1 < p \leqq +\infty$. Remark that the case $p = +\infty$ is included, while $p = 1$ is excluded (this is similar to what happens for weak convergence).

COROLLARY 1.15. *Let u_ε be a bounded sequence in $L^p(\Omega)$, with $1 < p \leqq +\infty$. There exists a function $u_0(x, y)$ in $L^p(\Omega \times Y)$ such that, up to a subsequence, u_ε two-scale converges to u_0, i.e., for any function $\psi(x, y) \in D[\Omega; C^\infty_\#(Y)]$, we have*

$$\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y u_0(x, y) \psi(x, y)\, dx\, dy.$$

(*The proof is exactly the same as that of Theorem* 1.2.)

Of course, two-scale convergence is also easily generalized to n-scale convergence, with n any finite integer greater than two. This is a very helpful tool for what is called reiterated homogenization (see [10, Chap. 1, § 8]).

COROLLARY 1.16. *Let u_ε be a bounded sequence in $L^2(\Omega)$. There exists a function $u_0(x, y_1, \cdots, y_{n-1})$ in $L^2(\Omega \times Y^{n-1})$ such that, up to a subsequence, u_ε n-scale converges to u_0, i.e., for any function $\psi(x, y_1, \cdots, y_{n-1}) \in D[\Omega; C^\infty_\#(Y^{n-1})]$, we have*

$$\begin{aligned} &\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}, \cdots, \frac{x}{\varepsilon^{n-1}}\right) dx \\ &\quad = \int_\Omega \int_{Y^{n-1}} u_0(x, y_1, \cdots, y_{n-1}) \psi(x, y_1, \cdots, y_{n-1})\, dx\, dy_1 \cdots dy_{n-1}. \end{aligned}$$

Remark 1.17. In the present paper, the test functions $\psi(x, y)$ are always assumed to be Y-periodic in y. Other choices for the period are possible. For a same sequence u_ε different two-scale limits can arise according to the period chosen for the test functions $y \to \psi(x, y)$, but they are related by a straightforward change of variables.

2. Homogenization of linear second-order elliptic equations. In this section we show how two-scale convergence can be used for the homogenization of linear second-order elliptic equations with periodically oscillating coefficients. We first revisit this favorite model problem of homogenization (see, e.g., [10, Chap. 1, § 6] in a fixed domain Ω, and later on we consider the case of perforated domains Ω_ε (see [13]). Besides recovering previous well-known results from a new point of view, we establish a new form of the limit problem, that we call the two-scale homogenized problem, and which is simply a combination of the usual homogenized problem and the cell problem (see [10], [40] for an introduction to the topic).

Let Ω be a bounded open set of $\mathbb{R}^N$. Let f be a given function in $L^2(\Omega)$. We consider the following linear second-order elliptic equation

$$(2.1) \qquad \begin{aligned} -\operatorname{div}\left(A\left(x, \frac{x}{\varepsilon}\right)\nabla u_\varepsilon\right) &= f \quad \text{in } \Omega, \\ u_\varepsilon &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $A(x, y)$ is a matrix defined on $\Omega \times Y$, Y-periodic in y, such that there exists two positive constants $0 < \alpha \leqq \beta$ satisfying

$$(2.2) \qquad \alpha|\xi|^2 \leqq \sum_{i,j=1}^{N} A_{ij}(x, y)\xi_i\xi_j \leqq \beta|\xi|^2 \quad \text{for any } \xi \in \mathbb{R}^N.$$

Assumption (2.2) implies that the matrix $A(x, y)$ belongs to $[L^\infty(\Omega \times Y)]^{N^2}$, but it doesn't ensure that the function $x \to A(x, x/\varepsilon)$ is measurable, nor that it converges to its average $\int_Y A(x, y)\, dy$ in any suitable topology (see the counterexample of Proposition 5.8). Thus, we also require that $A_{ij}(x, y)$ is an "admissible" test function in the sense of Definition 1.4, namely, $A_{ij}(x, x/\varepsilon)$ is measurable and satisfies

$$(2.3) \qquad \lim_{\varepsilon \to 0} \int_\Omega A_{ij}\left(x, \frac{x}{\varepsilon}\right)^2 dx = \int_\Omega \int_Y A_{ij}(x, y)^2\, dx\, dy.$$

Assumption (2.3) is the weakest possible, but is rather vague. More precise, but also more restrictive, assumptions include, e.g., $A(x, y) \in L^\infty[\Omega; C_\#(Y)]^{N^2}$, $A(x, y) \in L^\infty_\#[Y; C(\bar{\Omega})]^{N^2}$, or $A(x, y) \in C[\Omega; L^\infty_\#(Y)]^{N^2}$ (the latter is the usual assumption in [10]). Under assumptions (2.2), (2.3), equation (2.1) admits a unique solution u_ε in $H^1_0(\Omega)$, which satisfies the a priori estimate

$$(2.4) \qquad \|u_\varepsilon\|_{H^1_0(\Omega)} \leqq C\|f\|_{L^2(\Omega)},$$

where C is a positive constant that depends only on Ω and α, and not on ε. Thus, there exists $u \in H^1_0(\Omega)$ such that, up to a subsequence, u_ε converges weakly to u in $H^1_0(\Omega)$. The homogenization of (2.1) amounts to find a "homogenized" equation that admits the limit u as its unique solution.

Let us briefly recall the usual process of homogenization. In a first step, two-scale asymptotic expansions are used in order to obtain formally the homogenized equation (see, e.g., [10], [40]). In a second step, the convergence of the sequence u_ε to the solution u of the homogenized equation is proved (usually by means of the so-called energy method of Tartar [42]).

The results of the first (heuristic) step are summarized in the following.

DEFINITION 2.1. The homogenized problem is defined as

$$(2.5)\qquad \begin{aligned} -\operatorname{div}[A^*(x)\nabla u(x)]&=f \quad \text{in } \Omega,\\ u&=0 \quad \text{on } \partial\Omega, \end{aligned}$$

where the entries of the matrix A^* are given by

$$(2.6)\qquad A^*_{ij}(x)=\int_Y A(x,y)[\nabla_y w_i(x,y)+e_i]\cdot[\nabla_y w_j(x,y)+e_j]\,dy$$

and, for $1\leqq i\leqq N$, w_i is the solution of the so-called cell problem

$$(2.7)\qquad \begin{aligned} -\operatorname{div}_y[A(x,y)[\nabla_y w_i(x,y)+e_i]]&=0 \quad \text{in } Y,\\ y\to w_i(x,y) &\quad Y\text{-periodic.} \end{aligned}$$

As a result of the second step, we have the following theorem [10, Chap. I, Thm. 6.1].

THEOREM 2.2. *The sequence u_ε of solutions of* (2.1) *converges weakly in $H_0^1(\Omega)$ to the unique solution u of* (2.5).

We are going to recover this last result with the help of two-scale convergence, but we also propose an alternative formulation of the limit problem by introducing the *two-scale homogenized problem*, which is a combination of the usual homogenized equation (2.5) and of the cell equation (2.7).

THEOREM 2.3. *The sequence u_ε of solutions of* (2.1) *converges weakly to $u(x)$ in $H_0^1(\Omega)$, and the sequence ∇u_ε two-scale converges to $\nabla u(x)+\nabla_y u_1(x,y)$, where (u,u_1) is the unique solution in $H_0^1(\Omega)\times L^2[\Omega;H^1_\#(Y)/\mathbb{R}]$ of the following two-scale homogenized system*:

$$(2.8)\qquad \begin{aligned} -\operatorname{div}_y[A(x,y)[\nabla u(x)+\nabla_y u_1(x,y)]]&=0 \quad \text{in } \Omega\times Y,\\ -\operatorname{div}_x\left[\int_Y A(x,y)[\nabla u(x)+\nabla_y u_1(x,y)]\,dy\right]&=f \quad \text{in } \Omega,\\ u(x)&=0 \quad \text{on } \partial\Omega,\\ y\to u_1(x,y) &\quad Y\text{-periodic.} \end{aligned}$$

Furthermore, (2.8) *is equivalent to the usual homogenized and cell equations* (2.5)–(2.7) *through the relation*

$$(2.9)\qquad u_1(x,y)=\sum_{i=1}^{N}\frac{\partial u}{\partial x_i}(x)w_i(x,y).$$

Remark 2.4. The two-scale homogenized problem (2.8) is a system of two equations, two unknowns (u and u_1), where the two space variables x and y (i.e., the macroscopic and microscopic scales) are mixed. Although (2.8) seems to be complicated, it is a well-posed system of equations (cf. its variationial formulation (2.11) below), which is easily shown to have a unique solution. Remark that, here, the two equations of (2.8) can be decoupled in (2.5)–(2.7) (homogenized and cell equations) which are also two well-posed problems. However, we emphasize that this situation is very peculiar to the simple second-order elliptic equation (2.1). For many other types of problems, this decoupling is not possible, or leads to very complicated forms of the homogenized equation, including integro-differential operators and nonexplicit equations. Thus, the homogenized equation does not always belong to a class for which an existence and uniqueness theory is easily available, as opposed to the two-scale homogenized system, which is, in most cases, of the same type as the original problem, but with twice the variables (x and y) and unknowns (u and u_1). The supplementary, microscopic, variable and unknown play the role of "hidden" variables in the vocabulary of mechanics (as remarked by Sanchez-Palencia [40]). Although their presence doubles the size of the limit problem, it greatly simplifies its structure (which could be useful for numerical purposes, too), while eliminating them introduces "strange" effects (like memory or nonlocal effects) in the usual homogenized problem. In short, both formulations ("usual" or two-scale) of the homogenized problem have their pros and cons, and none should be eliminated without second thoughts. Particularly striking examples of the above discussion may be found in § 4, in [2] (a convection-diffusion problem), or in [3] (unsteady Stokes flows in porous media).

Remark 2.5. As stated earlier, the two-scale homogenized problem (2.8) is equivalent to the homogenized system (2.5) and the cell problem (2.7), which are obtained by two-scale asymptotic expansions. This equivalence holds without any assumptions on the symmetry of the matrix A. Recall that, if A is not symmetric, the test functions used in the energy method are not the solutions of (2.7), but that of the dual cell problem (i.e., (2.7), where A is replaced by its transpose ${}^tA$).

Proof of Theorem 2.3. Thanks to the a priori estimate (2.4), there exists a limit u such that, up to a subsequence, u_ε converges weakly to u in $H_0^1(\Omega)$. As a consequence of Proposition 1.14, there exists $u_1(x, y) \in L^2[\Omega; H^1_\#(Y)/\mathbb{R}]$ such that, up to another subsequence, ∇u_ε two-scale converges to $\nabla_x u(x) + \nabla_y u_1(x, y)$. In view of these limits, u_ε is expected to behave as $u(x) + \varepsilon u_1(x, x/\varepsilon)$. This suggests multiplying (2.1) by a test function $\phi(x) + \varepsilon\phi_1(x, x/\varepsilon)$, with $\phi(x) \in D(\Omega)$ and $\phi_1(x, y) \in D[\Omega; C^\infty_\#(Y)]$. This yields

$$\begin{aligned} &\int_\Omega A\left(x, \frac{x}{\varepsilon}\right)\nabla u_\varepsilon\left[\nabla\phi(x) + \nabla_y\phi_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon\nabla_x\phi_1\left(x, \frac{x}{\varepsilon}\right)\right] dx \\ (2.10) \qquad &= \int_\Omega f(x)\left[\phi(x) + \varepsilon\phi_1\left(x, \frac{x}{\varepsilon}\right)\right] dx. \end{aligned}$$

If the matrix $A(x, y)$ is smooth, then the function ${}^tA(x, x/\varepsilon)[\nabla\phi(x) + \nabla_y\phi_1(x, x/\varepsilon)]$ can be considered as a test function in Theorem 1.2, and we pass to the two-scale limit in (2.10). Even if $A(x, y)$ is not smooth, at least, by assumption (2.3), the function ${}^tA(x, x/\varepsilon)[\nabla\phi(x) + \nabla_y\phi_1(x, x/\varepsilon)]$ two-scale converges *strongly* to its limit ${}^tA(x, y)$ $[\nabla\phi(x) + \nabla_y\phi_1(x, y)]$ (i.e., condition (1.11) is satisfied in Theorem 1.8). Thus, using Theorem 1.8, we can still pass to the two-scale limit in (2.10):

$$\begin{aligned} (2.11) \qquad &\int_\Omega\int_Y A(x, y)[\nabla u(x) + \nabla_y u_1(x, y)] \cdot [\nabla\phi(x) + \nabla_y\phi_1(x, y)]\, dx\, dy \\ &= \int_\Omega f(x)\phi(x)\, dx. \end{aligned}$$

By density, (2.11) holds true for any (ϕ, ϕ_1) in $H_0^1(\Omega) \times L^2[\Omega; H^1_\#(Y)/\mathbb{R}]$. An easy integration by parts shows that (2.11) is a variational formulation associated to (2.8). Endowing the Hilbert space $H_0^1(\Omega) \times L^2[\Omega; H^1_\#(Y)/\mathbb{R}]$ with the norm $\|\nabla u(x)\|_{L^2(\Omega)} +$ $\|\nabla_y u_1(x, y)\|_{L^2(\Omega\times Y)}$, we check the conditions of the Lax-Milgram lemma in (2.11). Let us focus on the coercivity in $H_0^1(\Omega) \times L^2[\Omega; H^1_\#(Y)/\mathbb{R}]$ of the bilinear form defined by the left-hand side of (2.11):

$$\begin{aligned} &\int_\Omega\int_Y A(x, y)[\nabla\phi(x) + \nabla_y\phi_1(x, y)] \cdot [\nabla\phi(x) + \nabla_y\phi_1(x, y)]\, dx\, dy \\ &\geqq \alpha \int_\Omega\int_Y |\nabla\phi(x) + \nabla_y\phi_1(x, y)|^2\, dx\, dy \\ &= \alpha\int_\Omega |\nabla\phi(x)|^2\, dx + \alpha\int_\Omega\int_Y |\nabla_y\phi_1(x, y)|^2\, dx\, dy. \end{aligned}$$

Thus, by application of the Lax-Milgram lemma, there exists a unique solution of the two-scale homogenized problem (2.8). Consequently, the entire sequences u_ε and ∇u_ε converge to $u(x)$ and $\nabla u(x)+\nabla_y u_1(x,y)$. At this point, we could content ourselves with (2.8) as a homogenized problem, since its variational formulation (2.11) appears very naturally by application of two-scale convergence. However, it is usually preferable, from a physical or numerical point of view, to eliminate the microscopic variable y (one doesn't want to solve the small scale structure). This is an easy algebra exercise (left to the reader) to average (2.8) with respect to y, and to obtain the equivalent system (2.5)-(2.7), along with formula (2.6) for the homogenized matrix A^*. □

Corrector results are easily obtained with the two-scale convergence method. The next theorem rigorously justifies the two first terms in the usual asymptotic expansion of the solution u_ε (see [10]).

THEOREM 2.6. *Assume that $\nabla_y u_1(x,y)$ is an "admissible" test function in the sense of Definition* 1.4. *Then, the sequence $[\nabla u_\varepsilon(x)-\nabla u(x)-\nabla_y u_1(x, x/\varepsilon)]$ converges strongly to zero in $[L^2(\Omega)]^N$. In particular, if $u_1, \nabla_x u_1$, and $\nabla_y u_1$ are "admissible," then we have*

$$\left[u_\varepsilon(x)-u(x)-\varepsilon u_1\left(x,\frac{x}{\varepsilon}\right)\right]\to 0 \quad \textit{in } H^1(\Omega) \textit{ strongly.}$$

Proof. Let us first remark that the assumption on u_1, being an "admissible" test function, is satisfied as soon as the matrix A is smooth, say $A(x,y)\in C[\Omega; L^\infty_\#(Y)]^{N^2}$, by standard regularity results for the solutions $w_i(x,y)$ of the cell problem (2.7).

Now, using this assumption, we can write

$$\begin{aligned}
&\int_\Omega A\left(x,\frac{x}{\varepsilon}\right)\left[\nabla u_\varepsilon(x)-\nabla u(x)-\nabla_y u_1\left(x,\frac{x}{\varepsilon}\right)\right]^2 dx\\
&=\int_\Omega f(x)u_\varepsilon(x)\,dx+\int_\Omega A\left(x,\frac{x}{\varepsilon}\right)\left[\nabla u(x)+\nabla_y u_1\left(x,\frac{x}{\varepsilon}\right)\right]^2 dx \qquad (2.12)\\
&\quad-\int_\Omega (A+{}^tA)\left(x,\frac{x}{\varepsilon}\right)\nabla u_\varepsilon(x)\cdot\left[\nabla u(x)+\nabla_y u_1\left(x,\frac{x}{\varepsilon}\right)\right]dx.
\end{aligned}$$

Using the coercivity condition (2.2) and passing to the two-scale limit in the right-hand side of (2.12) yields

$$\begin{aligned}
&\alpha\lim_{\varepsilon\to 0}\left\|\nabla u_\varepsilon(x)-\nabla u(x)-\nabla_y u_1\left(x,\frac{x}{\varepsilon}\right)\right\|^2_{L^2(\Omega)} \qquad (2.13)\\
&\leqq\int_\Omega f(x)u(x)\,dx-\int_\Omega\int_Y A(x,y)[\nabla u(x)+\nabla_y u_1(x,y)]^2\,dx\,dy.
\end{aligned}$$

In view of (2.8), the right-hand side of (2.13) is equal to zero, which is the desired result. □

Two-scale convergence can also handle homogenization problems in perforated domains, without requiring any extension lemmas or similar technical ingredients. Let us define a sequence Ω_ε of periodically perforated subdomains of a bounded open set Ω in $\mathbb{R}^N$. The period of Ω_ε is εY^*, where Y^* is a subset of the unit cube $Y=(0;1)^N$, which is called the solid or material part (by opposition to the hole, or void part, $Y-Y^*$). We assume that the material domain E^*, obtained by Y-periodicity from Y^*, is a smooth connected open set in $\mathbb{R}^N$ (remark that no assumptions are made on the void domain $\mathbb{R}^N - E^*$; thus, the holes $Y - Y^*$ may be connected or isolated). Denoting by $\chi(y)$ the characteristic function of E^* (a Y-periodic function), Ω_ε is defined as

$$\Omega_\varepsilon = \left\{ x \in \Omega \Big/ \chi\left(\frac{x}{\varepsilon}\right) = 1 \right\}. \tag{2.14}$$

We consider a linear second-order elliptic equation in Ω_ε,

$$\begin{aligned} -\operatorname{div}\left(A\left(x, \frac{x}{\varepsilon}\right)\nabla u_\varepsilon\right) + u_\varepsilon = f \quad & \text{in } \Omega_\varepsilon, \\ \frac{\partial u_\varepsilon}{\partial v_{A_\varepsilon}} = \left[A\left(x, \frac{x}{\varepsilon}\right)\nabla u_\varepsilon\right] \cdot n = 0 \quad & \text{on } \partial\Omega_\varepsilon - \partial\Omega, \\ u_\varepsilon = 0 \quad & \text{on } \partial\Omega \cap \partial\Omega_\varepsilon, \end{aligned} \tag{2.15}$$

where the matrix A satisfies the same assumptions (2.2), (2.3) as before. From (2.15), we easily deduce the a priori estimates

$$\|u_\varepsilon\|_{L^2(\Omega_\varepsilon)} \leqq C \quad \text{and} \quad \|\nabla u_\varepsilon\|_{L^2(\Omega_\varepsilon)} \leqq C, \tag{2.16}$$

where C is a constant which does not depend on ε. The main difficulty in homogenization in perforated domains is to establish that the sequence u_ε admits a limit u in $H^1(\Omega)$. From (2.16) we cannot extract a convergent subsequence by weak compactness in a given Sobolev space, since each u_ε is defined in a different space $H^1(\Omega_\varepsilon)$, which varies with ε.

Nevertheless, this problem has first been solved by Cioranescu and Saint Jean Paulin [13] in the case of domains perforated with isolated holes (i.e., $Y - Y^*$ is strictly included in Y), while the general case is treated in [1] and [4]. The main result of these three papers is the following theorem.

THEOREM 2.7. *The sequence u_ε of solutions of* (2.15) *"converges" to a limit u, which is the unique solution in $H_0^1(\Omega)$ of the homogenized problem*

$$\begin{aligned} -\operatorname{div}[A^*\nabla u] + \theta u = \theta f \quad & \textit{in } \Omega, \\ u = 0 \quad & \textit{on } \partial\Omega, \end{aligned} \tag{2.17}$$

where θ is the volume fraction of material (i.e., $\theta = \int_Y \chi(y)\,dy = |Y^*|$), and the entries of the matrix A^* are given by

$$A_{ij}^*(x) = \int_{Y^*} A(x, y)[\nabla_y w_i(x, y) + e_i] \cdot [\nabla_y w_j(x, y) + e_j]\,dy, \tag{2.18}$$

and, for $1 \leqq i \leqq N$, w_i is the solution of the cell problem

$$\begin{aligned} -\operatorname{div}_y (A(x, y)[\nabla_y w_i(x, y) + e_i]) = 0 \quad & \textit{in } Y^*, \\ A(x, y)[\nabla_y w_i(x, y) + e_i] \cdot n = 0 \quad & \textit{on } \partial Y^* - \partial Y, \\ y \to w_i(x, y) \quad & Y\textit{-periodic}. \end{aligned} \tag{2.19}$$

Remark 2.8. The convergence of the sequence u_ε is intentionally very "vague" in Theorem 2.7. In view of the a priori estimates (2.16), there is no clear notion of convergence for u_ε, which is defined on a varying set Ω_ε. In the literature this difficulty has been overcome in two different ways. In [13] and [1], an extension of u_ε to the whole domain Ω is constructed, and this extension is proved to converge weakly in $H^1(\Omega)$ to the homogenized limit u. In [4], no sophisticated extensions are used, but a version of the Rellich theorem in perforated domains is established (loosely speaking, the embedding of $H^1(\Omega_\varepsilon)$ in $L^2(\Omega_\varepsilon)$ is compact, uniformly in ε), which allows us to prove that u_ε converges to u in the sense that $\|u_\varepsilon - u\|_{L^2(\Omega)_\varepsilon}$ goes to zero. All these references use classical methods of homogenization (the energy method of Tartar in [13] and [4], and the Γ-convergence of De Giorgi in [1]).

In the next theorem we recover the results of Theorem 2.7, using two-scale convergence. As in [4], we do not use any sophisticated extensions (apart from the trivial extension by zero in the holes $\Omega - \Omega_\varepsilon$), and we give a new interpretation of the "vague" convergence mentioned above.

THEOREM 2.9. *Denote by* $\tilde{}$ *the extension by zero in the domain* $\Omega - \Omega_\varepsilon$. *The sequences* $\tilde{u}_\varepsilon$ *and* $\tilde{\nabla u_\varepsilon}$ *two-scale converge to* $u(x)\chi(y)$ *and* $\chi(y)[\nabla u(x) + \nabla_y u_1(x, y)]$, *respectively, where* (u, u_1) *is the unique solution in* $H_0^1(\Omega) \times L^2[\Omega; H^1_\#(Y^*)/\mathbb{R}]$ *of the following two-scale homogenized system*;

$$\begin{aligned} -\operatorname{div}_y (A(x, y)[\nabla u(x) + \nabla_y u_1(x, y)]) = 0 \quad &\text{in } \Omega \times Y^*, \\ -\operatorname{div}_x \left[\int_{Y^*} A(x, y)[\nabla u(x) + \nabla_y u_1(x, y)]\, dy\right] + \theta u(x) = \theta f(x) \quad &\text{in } \Omega, \\ u(x) = 0 \quad &\text{on } \partial\Omega, \\ y \to u_1(x, y) \quad &Y\text{-periodic} \\ (A(x, y)[\nabla u(x) + \nabla_y u_1(x, y)]) \cdot n = 0 \quad &\text{on } \partial Y^* - \partial Y. \end{aligned} \tag{2.20}$$

Furthermore, (2.20) *is equivalent to the usual homogenized and cell equations* (2.17)-(2.19) *through the relation*

$$u_1(x, y) = \sum_{i=1}^N \frac{\partial u}{\partial x_i}(x) w_i(x, y). \tag{2.21}$$

Proof. In view of (2.16), the two sequences $\tilde{u}_\varepsilon$ and $\tilde{\nabla u_\varepsilon}$ are bounded in $L^2(\Omega)$, and by application of Theorem 1.2 they two-scale converge, up to a subsequence, to $u_0(x, y)$ and $\xi_0(x, y)$, respectively. Since, by definition, $\tilde{u}_\varepsilon$ and $\tilde{\nabla u_\varepsilon}$ are equal to zero in Ω-Ω_ε, their two-scale limit $u_0(x, y)$ and $\xi_0(x, y)$ are also equal to zero if $y \in Y - Y^*$. In order to find the precise form of u_0 and ξ_0 in $\Omega \times Y^*$, we argue as in Proposition 1.14(i). Let $\psi(x, y) \in D[\Omega; C^\infty_\#(Y)]$ and $\Psi(x, y) \in D[\Omega; C^\infty_\#(Y)]^N$ be two functions, equal to zero if $y \in Y - Y^*$ (hence, they belong to $D(\Omega_\varepsilon)$ and $[D(\Omega_\varepsilon)]^N$). We have

$$\begin{aligned} \lim_{\varepsilon \to 0} \int_{\Omega_\varepsilon} u_\varepsilon(x) \psi\left(x, \frac{x}{\varepsilon}\right) dx &= \int_\Omega \int_{Y^*} u_0(x, y) \psi(x, y)\, dx\, dy, \\ \lim_{\varepsilon \to 0} \int_{\Omega_\varepsilon} \nabla u_\varepsilon(x) \cdot \Psi\left(x, \frac{x}{\varepsilon}\right) dx &= \int_\Omega \int_{Y^*} \xi_0(x, y) \cdot \Psi(x, y)\, dx\, dy. \end{aligned} \tag{2.22}$$

By integration by parts, we obtain

$$\varepsilon \int_{\Omega_\varepsilon} \nabla u_\varepsilon(x) \cdot \Psi\left(x, \frac{x}{\varepsilon}\right) dx = -\int_{\Omega_\varepsilon} u_\varepsilon(x) \left[\operatorname{div}_y \Psi\left(x, \frac{x}{\varepsilon}\right) + \varepsilon \operatorname{div}_x \Psi\left(x, \frac{x}{\varepsilon}\right)\right] dx.$$

Passing to the limit in both terms with the help of (2.22) leads to

$$0 = -\int_\Omega \int_{Y^*} u_0(x, y) \operatorname{div}_y \Psi(x, y)\, dx\, dy.$$

This implies that $u_0(x, y)$ does not depend on y in Y^*, i.e., there exists $u(x) \in L^2(\Omega)$ such that

$$u_0(x, y) = u(x)\chi(y).$$

Now, we add to the previous assumptions on $\Psi(x, y)$ the condition $\operatorname{div}_y \Psi(x, y) = 0$. Integrating by parts in Ω_ε gives

$$\int_{\Omega_\varepsilon} \nabla u_\varepsilon(x) \cdot \Psi\left(x, \frac{x}{\varepsilon}\right) dx = -\int_{\Omega_\varepsilon} u_\varepsilon(x) \operatorname{div}_x \Psi\left(x, \frac{x}{\varepsilon}\right) dx. \tag{2.23}$$

Passing to the two-scale limit yields

$$\int_\Omega \int_{Y^*} \xi_0(x, y) \cdot \Psi(x, y)\, dx\, dy = -\int_\Omega \int_{Y^*} u(x) \operatorname{div}_x \Psi(x, y)\, dx\, dy. \tag{2.24}$$

By using Lemma 2.10 below, the right-hand side of (2.24) becomes $\int_\Omega u(x) \operatorname{div}_x \theta(x)\, dx$, while the left-hand side is a linear continuous form in $\theta(x) \in [L^2(\Omega)]^N$. This implies that $u(x) \in H_0^1(\Omega)$. Then, integrating by parts in (2.24) shows that, for any function $\Psi(x, y) \in L^2[\Omega; L^2_\#(Y^*)]^N$ with $\operatorname{div}_y \Psi(x, y) = 0$ and $\Psi(x, y) \cdot n_y = 0$ on $\partial Y^* - \partial Y$, we have

$$\int_\Omega \int_{Y^*} [\xi_0(x, y) - \nabla u(x)] \cdot \Psi(x, y)\, dx\, dy = 0. \tag{2.25}$$

Since the orthogonal of divergence-free functions is exactly the gradients, we deduce from (2.25) that there exists a function $u_1(x, y)$ in $L^2[\Omega; H^1_\#(Y^*)/\mathbb{R}]$ such that $\xi_0(x, y) = \chi(y)[\nabla u(x) + \nabla_y u_1(x, y)]$.

We are now in the position of finding the homogenized equations satisfied by u and u_1. Let us multiply the original equation (2.15) by the test function $\phi(x) + \varepsilon\phi_1(x, x/\varepsilon)$, where $\phi \in D(\Omega)$ and $\phi_1 \in D[\Omega; C^\infty_\#(Y)]$. Integrating by parts and passing to the two-scale limit yields

$$\begin{aligned}\int_\Omega \int_{Y^*} A(x, y)[\nabla u(x) + \nabla_y u_1(x, y)] \cdot [\nabla\phi(x) + \nabla_y \phi_1(x, y)]\, dx\, dy + \theta \int_\Omega u\phi\, dx \\ = \theta \int_\Omega f\phi\, dx.\end{aligned} \tag{2.26}$$

By density, (2.26) holds true for any (ϕ, ϕ_1) in $H_0^1(\Omega) \times L^2[\Omega; H^1_\#(Y^*)/\mathbb{R}]$. An easy integration by parts shows that (2.26) is a variational formulation associated to (2.20). It remains to prove existence and uniqueness in (2.26), and, as in Theorem 2.3, the main point is to show the coercivity of the left-hand side of (2.26). Indeed, it is an easy exercise (left to the reader) to check that $\|\nabla u(x) + \nabla_y u_1(x, y)\|_{L^2(\Omega \times Y^*)}$ is a norm for the Hilbert space $H_0^1(\Omega) \times L^2[\Omega; H^1_\#(Y^*)/\mathbb{R}]$. Remark, however, that this result relies heavily on the assumption on Y^* (namely, the Y-periodic set E^*, with period Y^*, is connected), and even fails if Y^* is strictly included in the unit cell Y. Remark also that here, to the contrary of the situation in Theorem 2.3, the above norm is not equal to $\|\nabla u(x)\|_{L^2(\Omega)} + \|\nabla_y u_1(x, y)\|_{L^2(\Omega \times Y^*)}$. □

LEMMA 2.10. *For any function $\theta(x) \in [L^2(\Omega)]^N$ there exists $\Psi(x, y) \in L^2[\Omega; H^1_\#(Y^*)]^N$ such that*

$$\begin{aligned} &\operatorname{div}_y \Psi(x, y) = 0 \quad \textit{in } Y^*, \\ &\Psi(x, y) = 0 \quad \textit{on } \partial Y^* - \partial Y, \\ &\int_{Y^*} \Psi(x, y)\, dy = \theta(x), \\ &\|\Psi(x, y)\|_{L^2(\Omega;\, H^1_\#(Y^*))^N} \leqq C \|\theta(x)\|_{L^2(\Omega)^N}. \end{aligned} \tag{2.27}$$

Proof. For $1 \leqq i \leqq N$, consider the following Stokes problem:

$$\begin{aligned} \nabla p_i - \Delta v_i = e_i &\quad \text{in } Y^*, \\ \operatorname{div} v_i = 0 &\quad \text{in } Y^*, \\ v_i = 0 &\quad \text{on } \partial Y^* - \partial Y, \\ p_i, v_i &\quad Y\text{-periodic}, \end{aligned}$$

which admits a unique, nonzero solution (p_i, v_i) in $[L^2_\#(Y^*)/\mathbb{R}] \times [H^1_\#(Y^*)]^N$ since we have assumed that E^* (the Y-periodic set obtained from Y^*) is smooth and connected. Denote by A the constant, symmetric, positive definite matrix $(\int_{Y^*} \nabla v_i \cdot \nabla v_j)_{1 \leqq i,j \leqq N}$. Then, for any $\theta(x) \in [L^2(\Omega)]^N$, the function Ψ defined by

$$\Psi(x, y) = \sum_{i=1}^{N} \langle A^{-1}\theta(x), e_i \rangle v_i(y)$$

is easily seen to satisfy all the properties (2.27) since $\int_{Y^*} \nabla v_i \cdot \nabla v_j = \int_{Y^*} v_i \cdot e_j$. □

3. Homogenization of nonlinear operators. In this section we show how two-scale convergence can handle nonlinear homogenization problems. Again, we revisit two well-known model problems in nonlinear homogenization: first, the Γ-convergence of oscillating convex integral functionals, and second, the H-convergence (also known as G-convergence) of oscillating monotone operators. We begin this section by recovering some previous results of De Giorgi, and Marcellini [31], concerning Γ-convergence of convex functionals. Then we recover other results of Tartar [42], about H-convergence of monotone operators, and finally we conclude by giving a few references where generalizations of the two-scale convergence method are applied to the homogenization of nonlinear hyperbolic conservation laws, and nonlinear equations admitting viscosity solutions (see Remark 3.8).

Let Ω be a bounded open set in $\mathbb{R}^N$ and $f(x)$ a given function on Ω. We consider a family of functionals

$$I_\varepsilon(v) = \int_\Omega \left[W\left(\frac{x}{\varepsilon}, \nabla v(x)\right) - f(x)v(x) \right] dx, \tag{3.1}$$

where $v(x)$ is a vector-valued function from Ω into $\mathbb{R}^n$, and the scalar energy $W(y, \lambda)$ satisfies, for some $p > 1$,

$$\begin{aligned} &\text{(i)} && \text{for any } \lambda, \text{ the function } y \to W(y, \lambda) \text{ is measurable and } Y\text{-periodic}, \\ &\text{(ii)} && \text{a.e. in } y, \text{ the function } \lambda \to W(y, \lambda) \text{ is strictly convex and } C^1 \text{ in } \mathbb{R}^{nN}, \\ &\text{(iii)} && 0 \leqq c|\lambda|^p \leqq W(y, \lambda) \leqq C[1 + \lambda|^p] \text{ a.e. in } y, \text{ with } 0 < c < C, \\ &\text{(iv)} && \left| \frac{\partial W}{\partial \lambda}(y, \lambda) \right| \leqq C[1 + |\lambda|^{p-1}] \text{ a.e. in } y. \end{aligned} \tag{3.2}$$

(Actually, assumption (iv) is easily seen to be a consequence of (ii) and (iii), as remarked by Francfort [24].) We also assume that $f(x) \in [L^{p'}(\Omega)]^n$ with $(1/p) + (1/p') = 1$. Since $W(y, \lambda)$ is convex in λ, for fixed ε, there exists a unique $u_\varepsilon(x) \in [W_0^{1,p}(\Omega)]^n$ that achieved the minimum of the functional $I_\varepsilon(v)$ on $[W_0^{1,p}(\Omega)]^n$, i.e.,

$$I_\varepsilon(u_\varepsilon) = \operatorname*{Inf}_{v \in [W_0^{1,p}(\Omega)]^n} \int_\Omega \left[W\left(\frac{x}{\varepsilon}, \nabla v(x)\right) - f(x)v(x) \right] dx. \tag{3.3}$$

The homogenization of the functionals $I_\varepsilon(v)$ amounts to finding an "homogenized" functional $\bar{I}(v)$ such that the sequence of minimizers u_ε converges to a limit u, which is precisely the minimizer of $\bar{I}(v)$. This problem has been solved by Marcellini [31]. His result is the following.

THEOREM 3.1. *There exist a functional $\bar{I}$ and a function u such that*

$$\begin{aligned} &u_\varepsilon \rightharpoonup u \quad \textit{weakly in } [W_0^{1,p}(\Omega)]^n,\\ (3.4)\qquad &I_\varepsilon(u_\varepsilon) \to \bar{I}(u),\\ &\bar{I}(u) = \operatorname*{Inf}_{v\in[W_0^{1,p}(\Omega)]^n} \bar{I}(v). \end{aligned}$$

Furthermore, $\bar{I}$ is given by

$$\bar{I}(v) = \int_\Omega [\bar{W}(\nabla v(x)) - f(x)v(x)]\,dx, \tag{3.5}$$

where the energy $\bar{W}$ is defined by

$$\bar{W}(\lambda) = \operatorname*{Inf}_{w\in[W_\#^{1,p}(Y)]^n} \int_Y W(y, \lambda + \nabla w(y))\,dy. \tag{3.6}$$

Remark 3.2. By definition, $\bar{I}$ is the homogenized functional, and the sequence I_ε is said to Γ-converge to $\bar{I}$. (For more details about the Γ-convergence of De Giorgi, see [16], [17].) In addition, it is easy to see that the energy $\bar{W}$ is also convex and C^1, and satisfies the same growth conditions as W. We emphasize that Theorem 3.1 is restricted to convex energies; the situation is completely different in the nonconvex case (see [12], [33]).

We are going to recover Theorem 3.1 using two-scale convergence, and without any tools form the theory of Γ-convergence.

THEOREM 3.3. *There exists a function $u(x)$ such that the sequence u_ε of solutions of* (3.3) *converges weakly to u in $[W_0^{1,p}(\Omega)]^n$. There also exists a function $u_1(x, y) \in L^p[\Omega; W_\#^{1,p}(Y)]^n$ such that the sequence ∇u_ε two-scale converges to $\nabla_x u(x) + \nabla_y u_1(x, y)$. Furthermore, the homogenized energy is also characterized as*

$$\bar{I}(u) = I(u, u_1) = \operatorname*{Inf}_{\substack{v\in[W_0^{1,p}(\Omega)]^n\\ v_1\in L^p[\Omega;\, W_\#^{1,p}(Y)/\mathbb{R}]^n}} I(v, v_1), \tag{3.7}$$

where $I(v, v_1)$ is the two-scale homogenized functional defined by

$$I(v, v_1) = \int_\Omega \int_Y [W[y, \nabla v(x) + \nabla_y v_1(x, y)] - f(x)v(x)]\,dx\,dy. \tag{3.8}$$

Remark 3.4. Theorem 3.3 furnishes a new characterization of the homogenized problem, which turns out to be a double minimization over two different spaces of functions of two variables x and y. In the quadratic case, this characterization was also proposed by Lions (see his "averaging principle" in the calculus of variations [30, § 5, Chap. 1]). Theorem 3.1 is easily deduced from Theorem 3.3 by averaging the two-scale homogenized functional $I(v, v_1)$ with respect to y to recover the usual homogenized functional $\bar{I}(v)$. The difference between $\bar{I}(v)$ and $I(v, v_1)$ corresponds exactly to the difference in the linear case between the usual and two-scale homogenized problems (see Remark 2.4).

Proof of Theorem 3.3. In view of the growth condition (3.2)(iii) for the energy W, the sequence of minimizers u_ε is bounded in $[W_0^{1,p}(\Omega)]^n$. Thus, there exists a function u such that, up to a subsequence, u_ε converges weakly to u in $[W_0^{1,p}(\Omega)]^n$.

Applying Proposition 1.14 and Corollary 1.15, there also exists a function $u_1(x,y)\in L^p[\Omega;\ W^{1,p}_{\#}(Y)/\mathbb{R}]^n$ such that, up to another subsequence, ∇u_ε two-scale converges to $\nabla_x u(x)+\nabla_y u_1(x,y)$.

In a first step we give a lower bound for $I_\varepsilon(u_\varepsilon)$. Since $W(\cdot,\lambda)$ is convex and differentiable, we have

$$W(\cdot,\lambda)\geqq W(\cdot,\mu)+\left\langle\frac{\partial W}{\partial\lambda}(\cdot,\mu),\lambda-\mu\right\rangle. \tag{3.9}$$

By specifying (3.9), we obtain

$$W\left[\frac{x}{\varepsilon},\nabla u_\varepsilon(x)\right]\geqq W\left[\frac{x}{\varepsilon},\mu\left(x,\frac{x}{\varepsilon}\right)\right]+\left\langle\frac{\partial W}{\partial\lambda}\left[\frac{x}{\varepsilon},\mu\left(x,\frac{x}{\varepsilon}\right)\right],\nabla u_\varepsilon(x)-\mu\left(x,\frac{x}{\varepsilon}\right)\right\rangle. \tag{3.10}$$

For a smooth function $\mu(x,y)\in D[\Omega;\ C^\infty_{\#}(Y)]^n$, we can integrate (3.10) on Ω, and then pass to the two-scale limit in the right-hand side. This leads to

$$\begin{aligned}\lim_{\varepsilon\to0} I_\varepsilon[u_\varepsilon(x)]\geqq&\int_\Omega\int_Y\Big(W[y,\mu(x,y)]-f(x)u(x)\Big)\,dx\,dy\\ &+\int_\Omega\int_Y\left\langle\frac{\partial W}{\partial\lambda}[y,\mu(x,y)],\nabla_x u(x)+\nabla_y u_1(x,y)-\mu(x,y)\right\rangle dx\,dy.\end{aligned} \tag{3.11}$$

Now, we apply (3.11) to a sequence of smooth functions $\mu(x,y)$, Y-periodic in y, which converges to $\nabla_x u(x)+\nabla_y u_1(x,y)$ strongly in $[L^p(\Omega\times Y)]^{nN}$. In view of the growth conditions (3.2)(iii) and (iv) on W and $\partial W/\partial\lambda$, we can pass to the limit in (3.11) and obtain

$$\begin{aligned}\lim_{\varepsilon\to0} I_\varepsilon[u_\varepsilon(x)]&\geqq\int_\Omega\int_Y[W[y,\nabla_x u(x)+\nabla_y u_1(x,y)]-f(x)u(x)]\,dx\,dy\\ &=I(u,u_1).\end{aligned} \tag{3.12}$$

Now, in a second step we establish an upper bound for $I_\varepsilon(u_\varepsilon)$. For $\phi(x)\in[D(\Omega)]^n$ and $\phi_1(x,y)\in D[\Omega;\ C^\infty_{\#}(Y)]^n$, since u_ε is the minimizer, we have

$$I_\varepsilon[u_\varepsilon(x)]\leqq I_\varepsilon\left[\phi(x)+\varepsilon\phi_1\left(x,\frac{x}{\varepsilon}\right)\right]. \tag{3.13}$$

Passing to the two-scale limit in the right-hand side of (3.13) yields

$$\begin{aligned}\lim_{\varepsilon\to0} I_\varepsilon[u_\varepsilon(x)]&\leqq\int_\Omega\int_Y[W[y,\nabla_x\phi(x)+\nabla_y\phi_1(x,y)]-f(x)\phi(x)]\,dx\,dy\\ &=I(\phi,\phi_1).\end{aligned} \tag{3.14}$$

The functional $I(\phi,\phi_1)$ is called the two-scale homogenized functional. By density, we deduce from (3.14) that

$$\lim_{\varepsilon\to0} I_\varepsilon[u_\varepsilon(x)]\leqq \operatorname*{Inf}_{\substack{v\in[W^{1,p}_0(\Omega)]^n\\ v_1\in L^p[\Omega;\,W^{1,p}_{\#}(Y)/\mathbb{R}]^n}} I(v,v_1). \tag{3.15}$$

Combining (3.12) and (3.15) yields

$$\lim_{\varepsilon\to0} I_\varepsilon[u_\varepsilon(x)]=I(u,u_1)=\operatorname*{Inf}_{\substack{v\in[W^{1,p}_0(\Omega)]^n\\ v_1\in L^p[\Omega;\,W^{1,p}_{\#}(Y)/\mathbb{R}]^n}} I(v,v_1). \tag{3.16}$$

Since $W(\cdot, \lambda)$ is a strictly convex energy, there exists a unique minimizer (u, u_1) of (3.16). Thus, the entire sequence u_ε converges weakly to u in $[W_0^{1,p}(\Omega)]^n$, and the entire sequence ∇u_ε two-scale converges to $\nabla_x u(x) + \nabla_y u_1(x, y)$. □

So far, we have considered minimization problems. Instead, we could have solved the corresponding nonlinear Euler equations, satisfied by the minimizers. More generally, we could consider nonlinear second-order elliptic equations, which may not correspond to any energy minimization. Indeed, we are going to generalize Theorem 3.3 to the case of monotone operators, thus recovering previous results of Tartar [42].

Define an operator $a(y, \lambda)$ from $Y \times \mathbb{R}^{nN}$ in $\mathbb{R}^{nN}$ as follows:

$$\text{(3.17)}\quad \begin{array}{ll} \text{(i)} & \text{for any } \lambda, \text{ the function } y \to a(y, \lambda) \text{ is measurable and } Y\text{-periodic,} \\ \text{(ii)} & \text{a.e. in } y, \text{ the function } \lambda \to a(y, \lambda) \text{ is continuous,} \\ \text{(iii)} & 0 \leqq c|\lambda|^p \leqq a(y, \lambda) \cdot \lambda, \text{ for } 0 < c, \text{ and } p > 1, \\ \text{(iv)} & |a(y, \lambda)| \leqq C[1 + |\lambda|^{p-1}], \text{ for } 0 < C. \end{array}$$

Furthermore, the operator a is strictly monotone, i.e.,

$$\text{(3.18)}\qquad [a(y, \lambda) - a(y, \mu)] \cdot (\lambda - \mu) > 0 \quad \text{for any } \lambda \neq \mu.$$

For $f(x) \in [L^{p'}(\Omega)]^n$ (with $(1/p) + (1/p') = 1$), we consider the equation

$$\text{(3.19)}\qquad \begin{aligned} -\operatorname{div} a\left(\frac{x}{\varepsilon}, \nabla u_\varepsilon\right) &= f \quad \text{in } \Omega, \\ u_\varepsilon &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

which admits a unique solution u_ε in $[W_0^{1,p}(\Omega)]^n$.

THEOREM 3.5. *The sequence u_ε of solutions of* (3.19) *converges weakly to a function $u(x)$ in $[W_0^{1,p}(\Omega)]^n$, and the sequence ∇u_ε two-scale converges to $\nabla_x u(x) + \nabla_y u_1(x, y)$, where (u, u_1) is the unique solution in $[W_0^{1,p}(\Omega)]^n \times L^p[\Omega; W_\#^{1,p}(Y)/\mathbb{R}]^n$ of the homogenized problem*

$$\text{(3.20)}\qquad \begin{aligned} -\operatorname{div}_x \left[\int_Y a[y, \nabla u(x) + \nabla_y u_1(x, y)]\, dy\right] &= f \quad \textit{in } \Omega \\ -\operatorname{div}_y a[y, \nabla u(x) + \nabla_y u_1(x, y)] &= 0 \quad \textit{in } Y \\ u &= 0 \quad \textit{on } \partial\Omega \\ y \to u_1(x, y) &\quad Y\text{-periodic.} \end{aligned}$$

Proof. From the growth conditions (3.17), we easily obtain a priori estimates on u_ε, which is bounded in $[W_0^{1,p}(\Omega)]^n$, and $g_\varepsilon = a(x/\varepsilon, \nabla u_\varepsilon)$, which is bounded in $[L^{p'}(\Omega)]^{nN}$. Thus, up to a subsequence, u_ε converges weakly to a limit u in $[W_0^{1,p}(\Omega)]^n$, while ∇u_ε and g_ε two-scale converge to $\nabla u(x) + \nabla_y u_1(x, y)$ and $g_0(x, y)$, respectively. Since $f + \operatorname{div} g_\varepsilon = 0$, arguing as in Proposition 1.14, it is not difficult to check that the two-scale limit g_0 satisfies

$$\text{(3.21)}\qquad \begin{aligned} \operatorname{div}_y g_0(x, y) &= 0 \\ f(x) + \operatorname{div}_x \left[\int_Y g_0(x, y)\, dy\right] &= 0. \end{aligned}$$

The problem is to identify g_0 in terms of a, u, and u_1. To this end, for any positive number t, and any functions ϕ, $\phi_1 \in D[\Omega; C_\#^\infty(Y)]^n$, we introduce a test function defined by

$$\mu_\varepsilon(x) = \nabla\left[u(x) + \varepsilon\phi_1\left(x, \frac{x}{\varepsilon}\right)\right] + t\phi\left(x, \frac{x}{\varepsilon}\right),$$

which two-scale converges to a limit $\mu_0(x,y)=\nabla u(x)+\nabla_y\phi_1(x,y)+t\phi(x,y)$. The monotonicity property (3.18) yields

$$\int_\Omega \left[g_\varepsilon - a\left(\frac{x}{\varepsilon},\mu_\varepsilon\right)\right]\cdot(\nabla u_\varepsilon-\mu_\varepsilon)\,dx \geqq 0,$$

or, equivalently,

$$\int_\Omega \left[-\operatorname{div} g_\varepsilon\cdot u_\varepsilon - a\left(\frac{x}{\varepsilon},\mu_\varepsilon\right)\cdot\nabla u_\varepsilon - g_\varepsilon\cdot\mu_\varepsilon + a\left(\frac{x}{\varepsilon},\mu_\varepsilon\right)\cdot\mu_\varepsilon\right]dx \geqq 0. \tag{3.22}$$

Using (3.19) in the first term of (3.22), and passing to the two-scale limit in all the other terms leads to

$$\int_\Omega\int_Y \left[f\cdot u - a(y,\mu_0)\cdot[\nabla u(x)+\nabla_y u_1(x,y)] - g_0\cdot\mu_0 + a(y,\mu_0)\cdot\mu_0\right]dx\,dy \geqq 0. \tag{3.23}$$

In view of the growth conditions (3.17) on the operator a, we can pass to the limit in (3.23) when considering a sequence of functions $\phi_1(x,y)$ that converges strongly to $u_1(x,y)$ in $[L^p(\Omega;\ W^{1,p}_{\#}(Y))]^n$. Thus, replacing μ_0 by $\nabla u(x)+\nabla_y u_1(x,y)+t\phi(x,y)$ and integrating by parts, (3.23) becomes

$$\begin{aligned}&\int_\Omega\left[f(x)+\operatorname{div}\left(\int_Y g_0(x,y)\,dy\right)\right]\cdot u(x)\,dx+\int_\Omega\int_Y \operatorname{div}_y g_0(x,y)\cdot u_1(x,y)\,dx\,dy\\&+\int_\Omega\int_Y [a[y,\nabla u(x)+\nabla_y u_1(x,y)+t\phi(x,y)]-g_0(x,y)]t\phi(x,y)\,dx\,dy\geqq 0.\end{aligned}\tag{3.24}$$

Thanks to (3.21), the first two terms of (3.24) are equal to zero. Then, dividing by $t>0$, and passing to the limit, as t goes to zero, gives for any function $\phi(x,y)$,

$$\int_\Omega\int_Y [a[y,\nabla u(x)+\nabla_y u_1(x,y)]-g_0(x,y)]\phi(x,y)\,dx\,dy\geqq 0.$$

Thus, we conclude that $g_0(x,y)=a[y,\nabla u(x)+\nabla_y u_1(x,y)]$. Combined with (3.21) it implies that (u,u_1) is a solution of the homogenized system (3.20). Since the operator a is strictly monotone, system (3.20) has a unique solution, and the entire sequence u_ε converges. □

In the case $p=2$, and under the further assumption that the operator a is uniformly monotone, i.e., there exists a positive constant c such that

$$[a(y,\lambda)-a(y,\mu)]\cdot(\lambda-\mu)\geqq c|\lambda-\mu|^2 \quad \text{for any } \lambda,\ \mu, \tag{3.25}$$

we obtain a corrector result similar to Theorem 2.6 in the linear case.

THEOREM 3.6. *Assume that the function $u_1(x,y)$ is smooth. Then, the sequence $u_\varepsilon(x)-u(x)-\varepsilon u_1(x,x/\varepsilon)$ converges strongly to zero in $H^1(\Omega)$.*

Remark 3.7. Corrector results for monotone operators in the general framework of H-convergence have been obtained by Murat [35] (see also [15] in the periodic case). By lack of smoothness for $\nabla u(x)$, the corrector in [35] is not explicit. Here, on the contrary, the corrector is explicitly given as $\nabla_y u_1(x,x/\varepsilon)$. However, we still have to assume that $u_1(x,y)$ is smooth in order to state Theorem 3.6 (more precisely, $\nabla_y u_1(x,x/\varepsilon)$ is required to be, at least, an admissible test function in the sense of Definition 1.4).

Proof of Theorem 3.6. Since $u_1(x, y)$ is assumed to be smooth, we consider the function

$$\mu_\varepsilon(x)=\nabla\left[u(x)+\varepsilon u_1\left(x,\frac{x}{\varepsilon}\right)\right],$$

which two-scale converges to $\mu_0(x, y)=\nabla u(x)+\nabla_y u_1(x, y)$. The monotonicity property (3.25) yields

$$\int_\Omega\left[g_\varepsilon-a\left(\frac{x}{\varepsilon},\mu_\varepsilon\right)\right]\cdot(\nabla u_\varepsilon-\mu_\varepsilon)\,dx\geqq c\int_\Omega|\nabla u_\varepsilon-\mu_\varepsilon|^2\,dx. \tag{3.26}$$

As in the proof of Theorem 3.5, the left-hand side of (3.26) goes to zero, which implies that the sequence $\nabla[u_\varepsilon(x)-u(x)-\varepsilon u_1(x, x/\varepsilon)]$ converges to zero in $[L^2(\Omega)]^N$. □

Remark 3.8. In the literature, homogenization has also been applied to other types of nonlinear equations. A first example is given by certain fully nonlinear, first- or second-order, partial differential equation, which fall within the scope of the theory of viscosity solutions (see the review paper of Crandall, Ishii, and Lions [14]). The key point of the viscosity solutions theory is that it provides a maximum principle that permits comparison between solutions. Based on this fact is the so-called "perturbed test function" method of Evans [22], [23], which provides very elegant proof of convergence for the homogenization of such equations. A perturbed test function is a function of the type $\phi(x)+\varepsilon^i\phi_1(x, x/\varepsilon)$ $(i=1, 2$, depending on the order of the equation), which is, thus very similar to that of the two-scale convergence method. Indeed, the perturbed test function method appears, a posteriori, as the ad hoc version of two-scale convergence in the context of viscosity solutions of nonlinear equations.

A second example is nonlinear hyperbolic conservation laws. To handle homogenization of such equations, E [19] introduced so-called two-scale Young measures, which are a combination of the usual Young measures (introduced for PDEs by Tartar [45]) with two-scale convergence. Combined with DiPerna's method for reducing measure-valued solutions of conservation laws to Dirac masses [18], it allows us to rigorously homogenize nonlinear transport equations, and nonlinear hyperbolic equations with oscillating forcing terms [19], [20]. In the case of linear hyperbolic equations, two-scale convergence has also been applied by Amirat, Hamdache, and Ziani [5] and Hou and Xin [26].

4. Homogenization of a diffusion process in highly heterogeneous media. In § 2 we studied the homogenization of a second-order elliptic equation with varying coefficients $A(x, x/\varepsilon)$. This can be regarded as a stationary diffusion process in a medium made of two materials, if $A(x, x/\varepsilon)$ takes only two different values (of the same order of magnitude). The present section is also devoted to the homogenization of a diffusion process, but the main novelty with respect to § 2 is the high heterogeneity of the two materials: namely, ε being the microscale, the ratio of their diffusion coefficients is taken of order ε^2 (this precise scaling corresponds to an equipartition of the energy in both materials, see Remark 4.9). As we shall see, it changes completely the form of the homogenized problem, which is genuinely of "two-scale" type (see 4.6)). In particular, the elimination of the microscale in the homogenized system does not yield a partial differential equation (see (4.9)).

Let us turn to a brief description of the geometry of the heterogeneous medium. We consider two materials, periodically distributed in a domain Ω (a bounded open set in $\mathbb{R}^N$), with period εY (ε is a small positive number, and $Y=(0\,;1)^N$ is the unit cube). The unit period Y is divided in two complementary parts Y_1 and Y_2, which are occupied by material 1 and material 2, respectively. Let $\chi_1(y)$ (respectively, $\chi_2(y)$)

be the characteristic function of Y_1 (respectively, Y_2), extended by Y-periodicity to the whole $\mathbb{R}^N$. They satisfy

$$\chi_1(y)+\chi_2(y)=1 \quad \text{in } Y.$$

The domain Ω is thus divided in two subdomains Ω_ε^1 and Ω_ε^2 (occupied by materials 1 and 2, respectively), which are defined by

$$\Omega_\varepsilon^1=\left\{x\in\Omega/\chi_1\left(\frac{x}{\varepsilon}\right)=1\right\} \quad \text{and} \quad \Omega_\varepsilon^2=\left\{x\in\Omega/\chi_2\left(\frac{x}{\varepsilon}\right)=1\right\}.$$

We make the fundamental assumption that, in the heterogeneous domain Ω, material 1 is the "matrix," while material 2 can be either "inclusions" or another matrix (like interconnected fibers). More precisely, denoting by E_1 the subset of $\mathbb{R}^N$ obtained by Y-periodicity from Y_1, we assume that E_1 is smooth and connected. On the contrary, no such assumptions are made on E_2 (the Y-periodic set built with Y_2).

Let μ_1 and μ_2 be two positive constants. We define the varying diffusion coefficient μ_ε of the heterogeneous medium Ω by

$$\mu_\varepsilon(x)=\mu_1\chi_1\left(\frac{x}{\varepsilon}\right)+\varepsilon^2\mu_2\chi_2\left(\frac{x}{\varepsilon}\right). \tag{4.1}$$

For a given source term f and positive constant α, we consider the following diffusion process for a scalar u_ε

$$\begin{aligned} -\operatorname{div}[\mu_\varepsilon\nabla u_\varepsilon]+\alpha u_\varepsilon&=f \quad \text{in } \Omega,\\ u_\varepsilon&=0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.2}$$

We implicitly assume in (4.2) the usual transmission condition at the interface of the two materials, namely, u_ε and $\mu_\varepsilon\partial u_\varepsilon/\partial n$ are continuous through $\partial\Omega_\varepsilon^1\cap\partial\Omega_\varepsilon^2$.

Remark 4.1. We emphasize the particular scaling of the coefficients defined in (4.2): the order of magnitude of μ_ε is 1 in material 1 (the "matrix"), and ε^2 in material 2 (the "inclusions" or the "fibers"). This explains why such a medium is called "highly" heterogeneous. (For a motivation of the precise scaling, see Remark 4.9 below.) Problem (4.2) is a simplified version of a system studied by Arbogast, Douglas, and Hornung [6], which models single phase flow in fractured porous media. Its homogenization leads to the so-called double porosity model. In their context, u_ε is the fluid pressure, and μ_ε is the permeability that is much larger in the network of fractures Ω_ε^1 than in the porous rocks Ω_ε^2. Problem (4.2) can also be interpreted as the heat equation. Then, u_ε is the temperature, and μ_ε is the thermal diffusion. (Thus, material 1 is a good conductor, while material 2 is a poor one.) Under additional assumptions on the geometry and the regularity of the source term, problem (4.2) has been studied by Panasenko [39] with the help of the maximum principle (that we do not use here).

Assuming $f\in L^2(\Omega)$, it is well known that there exists a unique solution of (4.2) in $H_0^1(\Omega)$. Multiplying (4.2) by u_ε and integrating by parts leads to

$$\int_\Omega \mu_\varepsilon|\nabla u_\varepsilon|^2+\alpha\int_\Omega u_\varepsilon^2=\int_\Omega fu_\varepsilon. \tag{4.3}$$

Then, if α is strictly positive, the solution u_ε is easily seen to satisfy the a priori estimates

$$\begin{aligned} \|u_\varepsilon\|_{L^2(\Omega)}&\leqq C,\\ \|\nabla u_\varepsilon\|_{L^2(\Omega_\varepsilon^1)}&\leqq C,\\ \|\nabla u_\varepsilon\|_{L^2(\Omega_\varepsilon^2)}&\leqq\frac{C}{\varepsilon}, \end{aligned} \tag{4.4}$$

where C is a positive constant which does not depend on ε.

Remark 4.2. The a priori estimates (4.4) are easily deduced from (4.3) when $\alpha > 0$. Actually they hold true even when $\alpha = 0$, but with a new ingredient, namely, a Poincaré-type inequality. Under the additional assumption that Y_1 is connected in Y, there exists a constant C, which does not depend on ε, and such that, for any $v \in H_0^1(\Omega)$,

$$\|v\|_{L^2(\Omega)} \leqq C[\|\nabla v\|_{L^2(\Omega_\varepsilon^1)} + \varepsilon \|\nabla v\|_{L^2(\Omega_\varepsilon^2)}]. \tag{4.5}$$

Obviously the Poincaré-type inequality (4.5), applied to u_ε, implies (4.4) even for $\alpha = 0$. The proof of (4.5) is rather technical and out of the scope of the present paper. The interested reader is referred to Lemma 3.4 in [4] for a similar proof. Thus, this is only for simplicity that a zero-order term has been introduced in (4.2).

Before stating the main result of the present section, let us define the Hilbert space $H_{0\#}^1(Y_2)$ made of functions of $H_\#^1(Y_2)$, which vanishes on the interface $\partial Y_1 \cap \partial Y_2$.

THEOREM 4.3. *The sequence u_ε of solutions of* (4.2) *two-scale converges to a limit $u(x) + \chi_2(y)v(x, y)$, where (u, v) is the unique solution in $H_0^1(\Omega) \times L^2[\Omega; H_{0\#}^1(Y_2)]$ of the homogenized problem*

$$\begin{aligned} -\mu_1 \operatorname{div}_x [A^*\nabla_x u(x)] + \alpha u(x) &= f(x) - \alpha \int_{Y_2} v(x, y)\, dy \quad \textit{in } \Omega, \\ -\mu_2 \Delta_{yy} v(x, y) + \alpha v(x, y) &= f(x) - \alpha u(x) \quad \textit{in } Y_2, \\ u &= 0 \quad \textit{on } \partial\Omega, \\ v(x, y) &= 0 \quad \textit{on } \partial Y_1 \cap \partial Y_2, \\ y &\to v(x, y) \quad Y\textit{-periodic}, \end{aligned} \tag{4.6}$$

where the entries of the constant matrix A^* are given by

$$A_{ij}^* = \int_{Y_1} [\nabla_y w_i(y) + e_i] \cdot [\nabla_y w_j(y) + e_j]\, dy, \tag{4.7}$$

and, for $1 \leqq i \leqq N$, $w_i(y)$ is the solution of the cell problem

$$\begin{aligned} -\operatorname{div}_y [\nabla_y w_i + e_i] &= 0 \quad \textit{in } Y_1, \\ [\nabla_y w_i + e_i] \cdot n &= 0 \quad \textit{on } \partial Y_1 \cap \partial Y_2, \\ y &\to w_i(y) \quad Y\textit{-periodic.} \end{aligned}$$

Thanks to a separation of variables, the homogenized system (4.6) can be simplified. Denoting by $U(x)$ the weak limit in $L^2(\Omega)$ of the sequence u_ε, we obtain an equation for U. (Let us note in passing that $U(x)$ is not equal to $u(x)$, but rather to $u(x) + \int_{Y_2} v(x, y)\, dy$.)

PROPOSITION 4.4. *Let $w(y)$ be the unique solution in $H_{0\#}^1(Y_2)$ of*

$$\begin{aligned} -\mu_2 \Delta_{yy} w(y) + \alpha w(y) &= 1 \quad \textit{in } Y_2, \\ w(y) &= 0 \quad \textit{on } \partial Y_1 \cap \partial Y_2, \\ y &\to w(y) \quad Y\textit{-periodic.} \end{aligned}$$

Then, $v(x, y) = w(y)[f(x) - \alpha u(x)]$, and $u(x)$ is the unique solution in $H_0^1(\Omega)$ of

$$\begin{aligned} -\mu_1 \operatorname{div}_x [A^*\nabla_x u(x)] + \alpha \left(1 - \alpha \int_{Y_2} w(y)\, dy\right) u(x) &= \left(1 - \alpha \int_{Y_2} w(y)\, dy\right) f(x) \quad \textit{in } \Omega \\ u &= 0 \quad \textit{on } \partial\Omega. \end{aligned} \tag{4.8}$$

Denoting by L^{-1} *the solution operator of* (4.8) *from* $H^{-1}(\Omega)$ *to* $H_0^1(\Omega)$ (*i.e.*, $u(x) = L^{-1}f(x)$), $U(x)$ *can be written as*

$$U(x) = L^{-1}f(x) + \left[\int_{Y_2} w(y)\,dy\right] f(x). \tag{4.9}$$

Remark 4.5. In view of (4.9), $U(x)$ is the solution of a very special diffusion process for which no simple partial differential equation can be found. Of course, if the source term $f(x)$ is smooth, we can apply the operator L to (4.9) and obtain the equation

$$L[U(x)] = f(x) + \left[\int_{Y_2} w(y)\,dy\right] L[f(x)]. \tag{4.10}$$

But (4.10) is only formal, since, a priori, the solution $U(x)$ does not satisfy the required Dirichlet boundary condition. Thus, it seems preferable to write $U(x)$ as the sum of two terms, which are solutions of a more standard problem (4.6). The homogenized problem (4.6) is a system of two coupled equations, one "macroscopic" (in Ω) and the other one "microscopic" (in Y_2): $u(x)$ is the contribution coming from material 1 in Ω_ε^1, and $v(x, y)$ is the additional contribution from material 2 in Ω_ε^2. This is definitely a "two-scale" phenomenon, since in the limit as $\varepsilon \to 0$ (4.6) keeps track of the two different materials on two different scales. This phenomenon allowed Arbogast, Douglas, and Hornung [6] to recover the so-called double porosity model in porous media flows.

The two-scale convergence of u_ε towards $u(x) + \chi_2(y)v(x, y)$ can be improved with the following corrector result.

PROPOSITION 4.6. *Assume that* $v(x, y)$ *is smooth* (*namely, that it is an admissible test function in the sense of Definition* 1.4). *Then we have*

$$\left[u_\varepsilon(x) - u(x) - \chi_2\left(\frac{x}{\varepsilon}\right) v\left(x, \frac{x}{\varepsilon}\right)\right] \to 0 \quad \text{in } L^2(\Omega) \text{ strongly.} \tag{4.11}$$

For the proof of Theorem 4.3 we need the following.

LEMMA 4.7. *There exist functions* $u(x) \in H_0^1(\Omega)$, $v(x, y) \in L^2[\Omega; H_{0\#}^1(Y_2)]$, *and* $u_1(x, y) \in L^2[\Omega; H_\#^1(Y_1)/\mathbb{R}]$ *such that, up to a subsequence,*

$$\begin{pmatrix} u_\varepsilon \\ \chi_1(x/\varepsilon)\nabla u_\varepsilon \\ \varepsilon\chi_2(x/\varepsilon)\nabla u_\varepsilon \end{pmatrix} \text{ two-scale converge to } \begin{pmatrix} u(x) + \chi_2(y)v(x, y) \\ \chi_1(y)[\nabla u(x) + \nabla_y u_1(x, y)] \\ \chi_2(y)\nabla_y v(x, y) \end{pmatrix}. \tag{4.12}$$

Proof. In view of the a priori estimates (4.4), the three sequences in (4.12) admit two-scale limits. Arguing as in Theorem 2.9, it is easily seen that there exist $u(x) \in H_0^1(\Omega)$ and $u_1(x, y) \in L^2[\Omega; H_\#^1(Y_1)/\mathbb{R}]$ such that $\chi_1(x/\varepsilon)u_\varepsilon$ and $\chi_1(x/\varepsilon)\nabla u_\varepsilon$ two-scale converge to $\chi_1(y)u(x)$ and $\chi_1(y)[\nabla u(x) + \nabla_y u_1(x, y)]$. On the other hand, it follows from Proposition 1.14 that there exists a function $u_0(x, y) \in L^2[\Omega; H_\#^1(Y_2)]$ such that $\chi_2(x/\varepsilon)u_\varepsilon$ and $\varepsilon\chi_2(x/\varepsilon)\nabla u_\varepsilon$ two-scale converge to $\chi_2(y)u_0(x, y)$ and $\chi_2(y)\nabla_y u_0(x, y)$. It remains to find the relationship between $u(x)$ and $u_0(x, y)$.

Consider the sequence $\varepsilon\nabla u_\varepsilon$ in the whole domain Ω. For any function $\phi(x, y) \in D[\Omega; C_\#^\infty(Y)]^N$, we know from the above results that

$$\lim_{\varepsilon\to 0} \int_\Omega \varepsilon\nabla u_\varepsilon(x) \cdot \phi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y \chi_2(y)\nabla_y u_0(x, y) \cdot \phi(x, y)\,dx\,dy. \tag{4.13}$$

By integration by parts, the left-hand side of (4.13) is also equal to

$$\lim_{\varepsilon\to 0} -\int_\Omega u_\varepsilon(x)\left[\operatorname{div}_y \phi\left(x,\frac{x}{\varepsilon}\right)+\varepsilon \operatorname{div}_x \phi\left(x,\frac{x}{\varepsilon}\right)\right] dx$$

$$= -\int_\Omega\int_Y [\chi_1(y)u(x)+\chi_2(y)u_0(x,y)] \operatorname{div}_y \phi(x,y)\, dx\, dy.$$

By equality between the two limits, we obtain that $u_0(x,y)=u(x)$ on $\partial Y_1 \cap \partial Y_2$. Thus, there exists $v(x,y)\in L^2[\Omega; H^1_{0\#}(Y_2)]$ such that $u_0(x,y)=u(x)+v(x,y)$. □

Proof of Theorem 4.3. In view of the two-scale limit of the sequence u_ε, we multiply (4.2) by a test function of the form $\phi(x)+\varepsilon\phi_1(x,x/\varepsilon)+\psi(x,x/\varepsilon)$, where $\phi(x)\in D(\Omega)$, $\phi_1(x,y)\in D[\Omega; C^\infty_\#(Y)]$, and $\psi(x,y)\in D[\Omega; C^\infty_\#(Y)]$ with $\psi(x,y)=0$ for $y\in Y_1$. Integrating by parts and passing to the two-scale limit yields

$$\begin{aligned}&\int_\Omega\int_{Y_1} \mu_1[\nabla u(x)+\nabla_y u_1(x,y)]\cdot[\nabla\phi(x)+\nabla_y\phi_1(x,y)]\, dx\, dy\\ &\quad+\int_\Omega\int_{Y_2}\mu_2\nabla_y v(x,y)\cdot\nabla_y\psi(x,y)\, dx\, dy\\ &\quad+\alpha\int_\Omega\int_Y [u(x)+\chi_2(y)v(x,y)]\cdot[\phi(x)+\chi_2(y)\psi(x,y)]\, dx\, dy\\ &=\int_\Omega\int_Y f(x)[\phi(x)+\chi_2(y)\psi(x,y)]\, dx\, dy.\end{aligned}\tag{4.14}$$

By density (4.14) holds true for any $(\phi,\phi_1,\psi)\in H^1_0(\Omega)\times L^2[\Omega; H^1_\#(Y_1)/\mathbb{R}]\times L^2[\Omega; H^1_{0\#}(Y_2)]$. Its left-hand side is easily seen to be coercive on the above functional space; thus (4.14) admits a unique solution (u,u_1,v). Another integration by parts shows that (4.14) is a variational formulation of the following two-scale homogenized system for u, u_1, and v:

$$\begin{aligned}-\mu_1\operatorname{div}_x\left[\int_{Y_1}[\nabla_x u(x)+\nabla_y u_1(x,y)]\, dy\right]+\alpha u(x)&=f(x)-\alpha\int_{Y_2}v(x,y)\, dy \quad \text{in } \Omega,\\ -\operatorname{div}_y[\nabla_x u(x)+\nabla_y u_1(x,y)]&=0 \quad \text{in } Y_1,\\ -\mu_2\Delta_{yy}v(x,y)+\alpha v(x,y)&=f(x)-\alpha u(x) \quad \text{in } Y_2,\\ u&=0 \quad \text{on } \partial\Omega,\\ [\nabla_x u(x)+\nabla_y u_1(x,y)]\cdot n_y&=0 \quad \text{on } \partial Y_1\cap\partial Y_2,\\ y\to u_1(x,y)&\quad Y\text{-periodic},\\ v(x,y)&=0 \quad \text{on } \partial Y_1\cap\partial Y_2,\\ y\to v(x,y)&\quad Y\text{-periodic}.\end{aligned}\tag{4.15}$$

In (4.15), the equation in u_1 can be decoupled from the two other ones, as we did in Theorem 2.9. Then, introducing the matrix A^* defined in (4.7), (4.8), the elimination of u_1 leads to system (4.6). □

Proof of Proposition 4.6. Recall the energy equation (4.3):

$$\int_\Omega \mu_\varepsilon|\nabla u_\varepsilon|^2+\alpha\int_\Omega u_\varepsilon^2=\int_\Omega f u_\varepsilon.\tag{4.3}$$

Passing to the limit in the right-hand side of (4.3), and using the variational formulation (4.14) yields

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_\Omega \mu_\varepsilon|\nabla u_\varepsilon|^2+\alpha\int_\Omega u_\varepsilon^2\\
(4.16)\qquad &=\int_\Omega\int_{Y_1}\mu_1|\nabla u(x)+\nabla_y u_1(x,y)|^2\,dx\,dy\\
&\quad+\int_\Omega\int_{Y_2}\mu_2|\nabla_y v(x,y)|^2\,dx\,dy+\alpha\int_\Omega\int_Y[u(x)+\chi_2(y)v(x,y)]^2\,dx\,dy.
\end{aligned}$$

By application of Proposition 1.6, the limit of each term in the left-hand side of (4.16) is larger than the corresponding two-scale limit in the right-hand side. Thus equality holds for each contribution. In particular, if $\alpha>0$, we have

$$(4.17)\qquad \lim_{\varepsilon\to 0}\int_\Omega u_\varepsilon^2=\int_\Omega\int_Y[u(x)+\chi_2(y)v(x,y)]^2\,dx\,dy.$$

In view of (4.17) and Theorem 1.8, we obtain the desired result (4.11). The result holds true also for $\alpha=0$: first we obtain a corrector result for the gradients $\chi_1(x/\varepsilon)\nabla u_\varepsilon$ and $\varepsilon\chi_2(x/\varepsilon)\nabla u_\varepsilon$, second we use again the Poincaré-type inequality (4.5) to deduce (4.11). □

Remark 4.8. Similarly to the scalar equation (4.2), we could consider a Stokes problem in a domain filled with two fluids having a highly heterogeneous viscosity μ_ε (still defined by (4.1))

$$(4.18)\qquad \begin{aligned}
\nabla p_\varepsilon-\operatorname{div}[\mu_\varepsilon\nabla u_\varepsilon]&=f\quad\text{in }\Omega,\\
\operatorname{div}u_\varepsilon&=0\quad\text{in }\Omega,\\
u_\varepsilon&=0\quad\text{on }\partial\Omega,
\end{aligned}$$

with the usual transmission condition at the interface: u_ε and $p_\varepsilon n-\mu_\varepsilon\partial u_\varepsilon/\partial n$ are continuous through $\partial\Omega_\varepsilon^1\cap\partial\Omega_\varepsilon^2$ (u_ε and p_ε are the velocity and pressure of the fluids). Assuming that Y_2 is a "bubble" strictly included in the period Y, (4.18) can be regarded as a model for bubbly fluids, where the viscosity is much smaller in the bubble than in the surrounding fluid. Because of its simplicity, this model is very academic since the size, the shape, and the periodic arrangement of the bubbles are kept fixed. Nevertheless, in view of Theorem 4.3, the homogenization of (4.18) could be interesting to derive averaged equations for bubbly fluids. Unfortunately, it turns out that the homogenized system can be drastically simplified in the Stokes case. Drawing upon the ideas of [44], Theorem 4.3 can be generalized to the Stokes equation (4.18), and a homogenized system similar to (4.6) is obtained:

$$(4.19)\qquad \begin{aligned}
\nabla p(x)-\mu_1\operatorname{div}_x[A^*\nabla_x u(x)]&=f(x)\quad\text{in }\Omega,\\
\operatorname{div}u(x)&=0\quad\text{in }\Omega,\\
\nabla_y q(x,y)-\mu_2\Delta_{yy}v(x,y)&=f(x)-\nabla p(x)\quad\text{in }Y_2,\\
\operatorname{div}_y v(x,y)&=0\quad\text{in }Y_2,\\
u&=0\quad\text{on }\partial\Omega,\\
v(x,y)&=0\quad\text{on }\partial Y_2,
\end{aligned}$$

where A^* is a given positive fourth-order tensor. Since we assumed that the bubble Y_2 does not touch the faces of Y, they are no periodic boundary condition for $q(x, y)$ and $v(x, y)$. Thus, the unique solution of (4.19) satisfies $v = 0$ and $q = y \cdot [f(x) - \nabla p(x)]$. As the weak limit in $[L^2(\Omega)]^N$ of the sequence u_ε is $u(x) + \int_{Y_2} v(x, y)\, dy$, it coincides with $u(x)$. Thus the homogenized problem can be reduced to the Stokes equation for $u(x)$. In other words, there are no contributions from the bubbles in the limit, and thus no interesting phenomena due to the bubbles appear in the homogenized Stokes equations.

Remark 4.9. We have chosen a very special scaling of the diffusion coefficients in (4.2): the order of magnitude of μ_ε is 1 in material 1, and ε^2 in material 2. Indeed, we could more generally consider a scaling ε^k in material 2, with k any positive real number. Let us motivate our choice of the scaling $k = 2$, and to make things easier, we assume that there is no zero-order term in (4.2), i.e., $\alpha = 0$. Then, it turns out that the value $k = 2$ is the only one (apart from zero) that insures a balance between the energies in material 1 and 2 is the only one (apart from zero) that insures a balance between the energies in material 1 and 2, i.e., as ε goes to zero, both terms $\int_{\Omega_\varepsilon^1} \mu_\varepsilon |\nabla u_\varepsilon|^2$ and $\int_{\Omega_\varepsilon^2} \mu_\varepsilon |\nabla u_\varepsilon|^2$ have the same order of magnitude. Thus, for $k = 2$, the limit problem will exhibit a coupling between material 1 and 2. On the contrary, for $k < 2$ the energy is much larger in material 1 than in 2, and in the limit no contributions from material 2 remains (material 2 behaves as a perfect conductor on the microscopic level). For $k > 2$ the energy is much smaller in material 1 than in 2, and in the limit no contributions from material 1 remains (actually, material 2 is a very poor conductor on the microscopic level, but since the source term is of order one its energy goes to infinity).

In other words, our scaling is the only one which makes of material 1 (respectively, 2) a good conductor on the macroscopic (respectively, microscopic) level, yielding an asymptotic (as ε goes to zero) equipartition of the energies stored in materials 1 and 2.

5. On convergence results for periodically oscillating functions. This section is devoted to the proof of Lemma 1.3, and more generally to the convergence of periodically oscillating functions $\psi(x, x/\varepsilon)$. Although in § 1 the convergence of the sequence $\psi(x, x/\varepsilon)$ was studied in $L^2(\Omega)$, for the sake of clarity we recast Lemma 1.3 and Definition 1.4 in the framework of $L^1(\Omega)$. More precisely, we consider functions of two variables $\psi(x, y)$ ($x \in \Omega$ open set in $\mathbb{R}^N$, $y \in Y$ the unit cube of $\mathbb{R}^N$), periodic of period Y in y, and we investigate the weak convergence of the sequence $\psi(x, x/\varepsilon)$ in $L^1(\Omega)$, as $\varepsilon \to 0$. Recall the analogue of Definition 1.4 obtained by replacing L^2 by L^1.

DEFINITION 5.1. A function $\psi(x, y) \in L^1(\Omega \times Y)$, Y-periodic in y, is called an "admissible" test function if and only if

$$\lim_{\varepsilon \to 0} \int_\Omega \left| \psi\left(x, \frac{x}{\varepsilon}\right) \right| dx = \int_\Omega \int_Y |\psi(x, y)|\, dx\, dy. \tag{5.1}$$

The purpose of this section is to investigate under which assumptions a function $\psi(x, y)$ is admissible in the sense of Definition 5.1. It is easily seen that continuous functions on $\Omega \times Y$ are admissible. However, when less smoothness is assumed on $\psi(x, y)$, the verification of (5.1) is not obvious (first of all, the measurability of $\psi(x, x/\varepsilon)$ is not always clear). In the sequel we propose several regularity assumptions for $\psi(x, y)$ to be admissible (see Lemma 5.2, Corollary 5.4, and Lemma 5.5). They all involve the continuity of ψ in, at least, one of the variables x or y. We emphasize that it is definitely not a necessary condition for (5.1). However, to our knowledge this is the only way to obtain, in general, the measurability of $\psi(x, x/\varepsilon)$, by asserting that $\psi(x, y)$ is a

Caratheodory-type function (for a precise definition, see, e.g., Definition 1.2 of Chapter VIII in [21]). We also emphasize that this question of measurability is not purely technical and futile, but is very much linked to possible counterexamples to (5.1). We actually exhibit a counterexample to (5.1), which clearly indicates that the regularity of $\psi(x, y)$ cannot be decreased too much, even if $\psi(x, x/\varepsilon)$ is measurable (see Proposition 5.8).

Our first and main result is the L^1 equivalent of Lemma 1.3, which we recall for the reader's convenience.

LEMMA 5.2. *Let $\psi(x, y) \in L^1[\Omega; C_\#(Y)]$. Then, for any positive value of ε, $\psi(x, x/\varepsilon)$ is a measurable function on Ω such that*

$$\left\| \psi\left(x, \frac{x}{\varepsilon}\right) \right\|_{L^1(\Omega)} \leqq \|\psi(x, y)\|_{L^1[\Omega; C_\#(Y)]}, \tag{5.2}$$

and $\psi(x, y)$ is an "admissible" test function, i.e., satisfies (5.1).

By definition, $L^1[\Omega; C_\#(Y)]$ is the space of functions, measurable and summable in $x \in \Omega$, with values in the Banach space of continuous functions, Y-periodic in y. More precisely, $L^1[\Omega; C_\#(Y)]$ is a space of *classes* of functions (two functions belong to the same class if they are equal almost everywhere in Ω); however, for simplicity we shall not distinguish a class or any of its representatives. The above definition of $L^1[\Omega; C_\#(Y)]$ is not very explicit, but we also have the following characterization, which implies, in particular, that any function of $L^1[\Omega; C_\#(Y)]$ is of Caratheodory type, i.e., satisfies (i) and (ii).

LEMMA 5.3. *A function $\psi(x, y)$ belongs to $L^1[\Omega; C_\#(Y)]$ if and only if there exists a subset E (independent of y) of measure zero in Ω such that*

(i) *For any $x \in \Omega - E$, the function $y \to \psi(x, y)$ is continuous and Y-periodic*;

(ii) *For any $y \in Y$, the function $x \to \psi(x, y)$ is measurable on Ω*;

(iii) *The function $x \to \operatorname{Sup}_{y \in Y} |\psi(x, y)|$ has a finite $L^1(\Omega)$-norm.*

Proof. We simply sketch the proof that relies on the equivalence between strong and weak measurability for functions with values in a separable Banach space. Recall the following result of functional analysis (see [11, Prop. 10, Chap. IV.5], or Petti's theorem [48, Chap. V]): let $f(x)$ be a function defined on Ω with values in a separable Banach space E, and let ϕ_n be a weak $*$ dense, countable, family of functions in the unit ball of the dual E' of E; the function f is measurable if and only if all the real-valued functions $x \to \langle \phi_n(x), f(x) \rangle_{E', E}$ are measurable.

Applying this result with $E = C_\#(Y)$, and ϕ_n the family of Dirac masses at rational points of Y, yields the result. □

Proof of Lemma 5.2. From Lemma 5.3, we know that $\psi(x, y)$ is a Caratheodory-type function, and this establishes the measurability of $\psi(x, x/\varepsilon)$. Then, inequality (5.2) is a consequence of the definition of the norm $\|\psi(x, y)\|_{L^1[\Omega; C_\#(Y)]} = \int_\Omega \operatorname{Sup}_{y \in Y} |\psi(x, y)|\, dx$. Let us check that $\psi(x, y)$ satisfies (5.1).

For any integer n, we introduce a paving of the unit cube Y made of n^N small cubes Y_i of size n^{-1}. The main properties of this paving are

$$Y = \bigcup_{i=1}^{n^N} Y_i, \quad |Y_i| = \frac{1}{n^N}, \quad |Y_i \cap Y_j| = 0 \quad \text{if } i \neq j. \tag{5.3}$$

Let $\chi_i(y)$ be the characteristic function of the set Y_i extended by Y-periodicity to $\mathbb{R}^N$, and let y_i be a point in Y_i. We approximate any function $\psi(x, y)$ in $L^1[\Omega; C_\#(Y)]$ by a step function in y defined by

$$\psi_n(x,y)=\sum_{i=1}^{n^N}\psi(x,y_i)\chi_i(y). \tag{5.4}$$

We first prove (5.1) for ψ_n, and then show that passing to the limit as n goes to infinity yields the result for ψ. Thanks to Lemma 5.3 the function $x\to\psi(x,y_i)$ belongs to $L^1(\Omega)$, while $\chi_i(x/\varepsilon)$ is in $L^\infty(\Omega)$. Due to the periodicity of χ_i, a well-known result on oscillating functions leads to

$$\lim_{\varepsilon\to 0}\int_\Omega \psi(x,y_i)\chi_i\left(\frac{x}{\varepsilon}\right)dx=\int_\Omega \psi(x,y_i)\,dx\,|Y_i|. \tag{5.5}$$

Summing equalities (5.5) for $i\in[1,\cdots,n^N]$ leads to (5.1) for ψ_n.

It remains to pass to the limit in n. Let us first prove that ψ_n converges to ψ in the strong topology of $L^1[\Omega;\, C_\#(Y)]$. Define

$$\delta_n(x)=\operatorname*{Sup}_{y\in Y}|\psi_n(x,y)-\psi(x,y)|. \tag{5.6}$$

The function $y\to[\psi_n(x,y)\text{-}\psi(x,y)]$ is piecewise continuous in Y almost everywhere in x. Thus, in (5.6) the supremum over $y\in Y$ can be replaced by the supremum over $y\in Y\cap\mathbb{Q}$. This implies that δ_n, being the supremum of a countable family of measurable functions, is measurable, too (see if necessary [11, Chap. IV.5, Thm. 2]). On the other hand, as a result of the continuity in y of ψ, we have

$$\lim_{n\to+\infty}\delta_n(x)=0 \quad \text{a.e. in } \Omega.$$

Furthermore,

$$0\leqq\delta_n(x)\leqq 2\operatorname*{Sup}_{y\in Y}|\psi(x,y)|\in L^1(\Omega).$$

By application of the Lebesgue theorem of dominated convergence, the sequence δ_n strongly converges to zero in $L^1(\Omega)$. Thus ψ_n strongly converges to ψ in $L^1[\Omega;\, C_\#(Y)]$.

Let us estimate the difference

$$\begin{aligned}\left|\int_\Omega\psi\left(x,\frac{x}{\varepsilon}\right)dx-\int_\Omega\int_Y\psi(x,y)\,dx\,dy\right| &\leqq \left|\int_\Omega\psi\left(x,\frac{x}{\varepsilon}\right)dx-\int_\Omega\psi_n\left(x,\frac{x}{\varepsilon}\right)dx\right|\\ &\quad+\left|\int_\Omega\psi_n\left(x,\frac{x}{\varepsilon}\right)dx-\int_\Omega\int_Y\psi_n(x,y)\,dx\,dy\right|\\ &\quad+\left|\int_\Omega\int_Y\psi(x,y)\,dx\,dy-\int_\Omega\int_Y\psi_n(x,y)\,dx\,dy\right|.\end{aligned} \tag{5.7}$$

The first term in the right-hand side of (5.7) is bounded by

$$\int_\Omega\left|\psi\left(x,\frac{x}{\varepsilon}\right)-\psi_n\left(x,\frac{x}{\varepsilon}\right)\right|dx\leqq\int_\Omega\operatorname*{Sup}_{y\in Y}|\psi(x,y)-\psi_n(x,y)|\,dx=\|\psi_n-\psi\|_{L^1[\Omega;C_\#(Y)]}.$$

For fixed n we pass to the limit in (5.7) as $\varepsilon\to 0$:

$$\lim_{\varepsilon\to 0}\left|\int_\Omega\psi\left(x,\frac{x}{\varepsilon}\right)dx-\int_\Omega\int_Y\psi(x,y)\,dx\,dy\right|\leqq 2\|\psi_n-\psi\|_{L^1[\Omega;C_\#(Y)]}. \tag{5.8}$$

Then, we pass to the limit in (5.8) as $n\to\infty$, and we obtain (5.1). □

Reversing the role of x and y (namely, assuming continuity in x and measurability in y), the same proof as that of Lemma 5.2 works also for the following corollary.

COROLLARY 5.4. *Assume that Ω is a bounded open set (its closure $\bar{\Omega}$ is thus compact). Let $\psi(y,x)$ be a function in $L^1_\#[Y; C(\bar{\Omega})]$, i.e., measurable, summable, and Y-periodic in y, with values in the Banach space of continuous functions in $\bar{\Omega}$. Then, for any positive value of ε, $\psi(x/\varepsilon, x)$ is a measurable function on Ω such that*

$$(5.9) \qquad \left\| \psi\left(\frac{x}{\varepsilon}, x\right) \right\|_{L^1(\Omega)} \leqq C(\Omega) \| \psi(y,x) \|_{L^1_\#[Y; C(\bar{\Omega})]},$$

and $\psi(y,x)$ is an "admissible" test function, i.e., $\lim_{\varepsilon\to 0} \int_\Omega |\psi(x/\varepsilon, x)|\, dx = \int_\Omega \int_Y |\psi(y,x)|\, dx\, dy$.

In the literature (see, e.g., [10]) the favorite assumption on $\psi(x,y)$, ensuring it is an admissible test function, is $\psi(x,y) \in C_c[\Omega; L^\infty_\#(Y)]$ (i.e., continuous with compact support in Ω, with values in the Banach space of measurable, essentially bounded, and Y-periodic functions in Y). The next two lemmas are concerned with this situation.

LEMMA 5.5. *Let $\psi(x,y)$ be a function such that there exist a subset $E \subset Y$, of measure zero, independent of x, and a compact subset $K \subset \Omega$ independent of y, satisfying*

(i) *For any $y \in Y - E$, the function $x \to \psi(x,y)$ is continuous, with compact support K;*

(ii) *For any $x \in \Omega$, the function $y \to \psi(x,y)$ is Y-periodic and measurable on Y;*

(iii) *The function $x \to \psi(x,y)$ is continuous on K, uniformly with respect to $y \in Y - E$.*

Then, for any positive value of ε, $\psi(x, x/\varepsilon)$ is a measurable function on Ω, and $\psi(x,y)$ is an admissible test function in the sense of Definition 5.1, *i.e., satisfies*

$$(5.1) \qquad \lim_{\varepsilon\to 0} \int_\Omega \left| \psi\left(x, \frac{x}{\varepsilon}\right) \right| dx = \int_\Omega \int_Y |\psi(x,y)|\, dx\, dy.$$

Before proving Lemma 5.5, let us remark that any function satisfying (i)-(iii) obviously belongs to $C_c[\Omega; L^\infty_\#(Y)]$. The converse is more subtle. Indeed, since $\psi(x,y)$ is an element of $C_c[\Omega; L^\infty_\#(Y)]$, for each $x \in \Omega$, its value $y \to \psi(x,y)$ is a class of functions in $L^\infty_\#(Y)$: picking up a representative for each x and collecting them gives a "representative" of $\psi(x,y)$ in $C_c[\Omega; L^\infty_\#(Y)]$.

LEMMA 5.6. *Let $\psi(x,y)$ be a function in $C_c[\Omega; L^\infty_\#(Y)]$. Then, there exists a "representative" of $\psi(x,y)$ for which properties* (i)-(iii) *in Lemma* 5.5 *hold.*

Proof. Let $\psi(x,y) \in C_c[\Omega; L^\infty_\#(Y)]$. By definition, for any value of $x \in \Omega$, the function $y \to \psi(x,y)$ is measurable on Y, Y-periodic, and there exists a subset $E(x)$ of measure zero in Y such that $\psi(x,y)$ is bounded on $Y - E(x)$. The continuity of $x \to \psi(x,y)$ from Ω in $L^\infty_\#(Y)$ is equivalent to

$$(5.10) \qquad \lim_{\eta\to 0} \operatorname*{Sup}_{y \in Y - [E(x) \cup E(x+\eta)]} |\psi(x+\eta, y) - \psi(x,y)| = 0 \quad \text{for any } x \in \Omega.$$

We emphasize that, a priori, the exceptional set $E(x)$, where the function $y \to \psi(x,y)$ is not defined, depends on x. Nevertheless, thanks to the continuity of $\psi(x,y)$ with respect to the x variable, we are going to exhibit a "representative" of $\psi(x,y)$ for which $E(x)$ is included in a fixed set E of measure zero.

Let $K \subset \Omega$ be the compact support of $x \to \psi(x,y)$. Let $(K_i)_{i=1}^n$ be a sequence of partitions of K (i.e., $\bigcup_{i=1}^n K_i = K$ and $|K_i \cap K_j| = 0$ if $i \neq j$) such that $\lim_{n\to+\infty} \operatorname{Sup}_{1\leqq i\leqq n} \operatorname{diam}(K_i) = 0$. Let $\chi_i(x)$ be the characteristic function of K_i, and x_i a point in K_i. Define the step function $\psi_n(x,y)$ by

$$\psi_n(x,y) = \sum_{i=1}^n \psi(x_i, y) \chi_i(x).$$

By definition of the partitions $(K_i)_{i=1}^n$, and continuity of $x \to \psi(x, y)$ from Ω in $L^\infty_\#(Y)$, we have

$$\lim_{n\to+\infty} \operatorname{Sup}_{x\in K} \|\psi(x, y)-\psi_n(x, y)\|_{L^\infty_\#(Y)}. \tag{5.11}$$

In view of its definition, $\psi_n(x, y)$ is defined and bounded on $\Omega\times(Y-E_n)$, where E_n is a set of measure zero that does not depend on x. Then, the set $E=\bigcup_{n=1}^\infty E_n$ is also of measure zero and does not depend on x. From (5.11) it is easily deduced that $\psi_n(x, y)$ converges pointwise in $\Omega\times(Y-E)$ to a limit $\tilde{\psi}(x, y)$ that is continuous in $x\in\Omega$ uniformly with respect to $y\in Y-E$. As announced, $\tilde{\psi}$ is a "representative" of $\psi(x, y)$, which has the desired properties (i)-(iii). □

Proof of Lemma 5.5. Properties (i) and (ii) imply that $\psi(x, y)$ is a Caratheodory-type function, and thus $\psi(x, x/\varepsilon)$ is measurable on Ω. Using the approximating sequence of step functions $\psi_n(x, y)$ introduced in the proof of Lemma 5.6, and arguing as in Lemma 5.2, leads to (5.1) for ψ. □

In the three previous results, the function $\psi(x, y)$ is assumed to be continuous in, at least, one variable x or y. Of course, it is not a necessary assumption that ψ be an "admissible" test function. For example, if a separation of variables holds, namely, ψ is the product of two functions, each depending on only one variable, we have the following well-known result (for a proof, see, e.g., [9]).

LEMMA 5.7. *Assume that Ω is a bounded open set. Let $\phi_1(x)\in L^p(\Omega)$ and $\phi_2(y)\in L^{p'}_\#(Y)$ with $(1/p)+(1/p')=1$ and $1\leqq p\leqq+\infty$. (In case $p=1$ and $p'=+\infty$, the set Ω can be unbounded.) Then, for any positive value of ε, $\phi_1(x)\phi_2(x/\varepsilon)$ is a measurable function on Ω, and $\phi_1(x)\phi_2(y)$ is an "admissible" test function in the sense of Definition* 5.1.

In general the regularity of ψ cannot be weakened too much: even if $\psi(x, x/\varepsilon)$ is measurable, the function $\psi(x, y)$ may be not "admissible" in the sense of Definition 5.1. Following an idea of Gérard and Murat [25], we are able to construct a counterexample to (5.1) with $\psi(x, y)\in C[\bar{\Omega}; L^1_\#(Y)]$.

PROPOSITION 5.8. *Let $\bar{\Omega}=Y=[0; 1]$. There exists $v(x, y)\in C([0, 1]; L^1_\#[0, 1])$, which is not an "admissible" test function, namely,*

$$\lim_{n\to+\infty} \int_0^1 |v(x, nx)|\,dx \neq \int_0^1\int_0^1 |v(x, y)|\,dx\,dy. \tag{5.12}$$

Remark 5.9. In general, a function $\psi(x, y)\in C[\bar{\Omega}; L^1_\#(Y)]$ is not of Caratheodory type, i.e., is not continuous in x almost everywhere in Y. Thus, the measurability of $\psi(x, x/\varepsilon)$ is usually not guaranteed.

Proof of Proposition 5.8. Let us fix $\bar{\Omega}=Y=[0; 1]$. In the square $[0; 1]^2$, we are going to construct an increasing sequence of measurable subset E_n, which converges to a set E. The desired function $v(x, y)$ will be defined as the characteristic function of E extended by $[0; 1]$-periodicity in y.

For each integer n, we consider the n lines defined in the plane by

$$y=nx-p \quad \text{with}\, p\in\{0, 1, 2, \cdots, n-1\}.$$

Then, we define the set D_n made of all the points (x, y) in $[0; 1]^2$ that are at a distance less than αn^{-3} of one of the lines $y=nx-p$ for $p=0, 1, \cdots, n-1$ (the distance is the usual Euclidean distance, and α is a small strictly positive number). The set D_n is made of n strips of width $2\alpha n^{-3}$ and length of order 1. Next, we define the measurable set $E_n=\bigcup_{p=1}^n D_p$. The sequence E_n is increasing in $[0; 1]^2$, and thus converges to a measurable limit set E. We have a bound on its measure

$$|E| \leqq \sum_{n=1}^{\infty} |D_n| \leqq 4\alpha \sum_{n=1}^{\infty} \frac{1}{n^2}. \tag{5.13}$$

Let $v(x, y)$ be the characteristic function of E extended by $[0; 1]$-periodicity in y. For sufficiently small α, we deduce from (5.13) that $\int_{[0;1]^2} v(x, y)\, dx\, dy = |E| < 1$. Meanwhile, we obviously have $v(x, nx) = 1$ for $x \in [0; 1]$. Thus, the sequence $\int_{[0;1]} v(x, nx)\, dx$ cannot converge to the average of v. To complete the proof it remains to show that $v(x, y)$ belongs to $C([0, 1]; L^1_{\#}[0, 1])$, i.e., for any $x \in [0; 1]$,

$$\lim_{\varepsilon \to 0} \int_{[0;1]} |v(x+\varepsilon, y) - v(x, y)|\, dy = 0. \tag{5.14}$$

(By definition of E, $v(x, y)$ is measurable in $[0; 1]^2$ and is easily seen to be also measurable, at fixed x, in y.) Let $E(x)$ (respectively, $D_n(x)$) be the section of E (respectively, D_n) at fixed abcissa x, i.e.,

$$E(x) = \{y \in [0; 1]/(x, y) \in E\},$$
$$D_n(x) = \{y \in [0; 1]/(x, y) \in D_n\}.$$

Then

$$\int_{[0;1]} |v(x+\varepsilon, y) - v(x, y)|\, dy = |E(x) \cap ([0; 1] - E(x+\varepsilon))| + |E(x+\varepsilon) \cap ([0; 1] - E(x))|.$$

Since $E(x) = \bigcup_{n=1}^{\infty} D_n(x)$, we have

$$|E(x) \cap ([0; 1] - E(x+\varepsilon))| \leqq \sum_{n=1}^{\infty} |D_n(x) \cap ([0; 1] - D_n(x+\varepsilon))|.$$

It is easily seen that $|D_n(x) \cap ([0; 1] - D_n(x+\varepsilon))|$ is constant when x varies in $[0; 1]$. Thus

$$\int_{[0;1]} |v(x+\varepsilon, y) - v(x, y)|\, dy \leqq 2 \sum_{n=1}^{\infty} |D_n(x) \cap ([0; 1] - D_n(x+\varepsilon))|. \tag{5.15}$$

Let us fix $\varepsilon > 0$. Recall that D_n is made of n strips of width $2\alpha n^{-3}$. Denote by l_n (respectively, L_n) the length of the intersection of one strip with the x-axis (respectively, y-axis). It is easily seen that l_n is of order n^{-3}, while L_n is of order n^{-2}. Both points (x, y) and $(x+\varepsilon, y)$ lie in the same strip of D_n if n is smaller than $\varepsilon^{-1/3}$. This suggests to cut the sum in (5.15) in two parts, the first one being

$$\sum_{n=\varepsilon^{-1/3}}^{\infty} |D_n(x) \cap ([0; 1] - D_n(x+\varepsilon))| \leqq \sum_{n=\varepsilon^{-1/3}}^{\infty} |D_n(x)|, \tag{5.16}$$

while the second one is

$$\sum_{n=1}^{\varepsilon^{-1/3}} |D_n(x) \cap ([0; 1] - D_n(x+\varepsilon))|. \tag{5.17}$$

Since $|D_n(x)| = L_n$ is of order n^{-2}, (5.16) is bounded by

$$\sum_{n=\varepsilon^{-1/3}}^{\infty} |D_n(x)| \leqq C \sum_{n=\varepsilon^{-1/3}}^{\infty} \frac{1}{n^2} \leqq C\varepsilon^{1/3}.$$

On the other hand, an easy calculation shows that, for any value of n, $|D_n(x) \cap ([0;1]-D_n(x+\varepsilon))|$ is bounded by $C\varepsilon n$. Thus, (5.17) is bounded by

$$\sum_{n=1}^{\varepsilon^{-1/3}} |D_n(x) \cap ([0;1]-D_n(x+\varepsilon))| \leqq C \sum_{n=1}^{\varepsilon^{-1/3}} \varepsilon n \leqq C\varepsilon^{1/3}.$$

This leads to

$$\int_{[0;1]} |v(x+\varepsilon, y) - v(x, y)| \, dy \leqq C\varepsilon^{1/3},$$

where C is a constant independent of ε. Letting $\varepsilon \to 0$ yields (5.14). □

Acknowledgment. The author wishes to thank F. Murat for stimulating discussions on the topic.

REFERENCES

[1] E. ACERBI, V. CHIADO PIAT, G. DAL MASO, AND D. PERCIVALE, *An extension theorem from connected sets, and homogenization in general periodic domains*, to appear.

[2] G. ALLAIRE, *Homogénéisation et convergence à deux échelles, application à un problème de convection diffusion*, C. R. Acad. Sci. Paris, 312 (1991), pp. 581-586.

[3] ———, *Homogenization of the unsteady Stokes equations in porous media*, in Proceedings of the 1st European Conference on Elliptic and Parabolic Problems, Pont-à-Mousson, June 1991, to appear.

[4] G. ALLAIRE AND F. MURAT, *Homogenization of the Neuman problem with non-isolated holes*, Asymptotic Anal., to appear.

[5] Y. AMIRAT, K. HAMDACHE, AND A. ZIANI, *Homogénéisation non-locale pour des équations dégénérées à coéfficients périodiques*, C. R. Acad. Sci. Paris, 312 (1991), pp. 963-966.

[6] T. ARBOGAST, J. DOUGLAS, AND U. HORNUNG, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823-836.

[7] N. BAKHVALOV AND G. PANASENKO, *Homogenization: averaging processes in periodic media*, Math. Appl., Vol. 36, Kluwer Academic Publishers, Dordrecht, 1990.

[8] J. BALL, *A version of the fundamental theorem for Young measures*, in Pde's and continuum models of phase transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Phys., Vol. 344, Springer-Verlag, New York, 1989.

[9] J. BALL AND F. MURAT, *$W^{1,p}$-quasiconvexity and variational problems for multiple integrals*, J. Funct. Anal., 58 (1984), pp. 225-253.

[10] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[11] N. BOURBAKI, *Eléments de Mathématiques, Intégration, livre* VI, Hermann, Paris, 1965.

[12] A. BRAIDES, *Homogenization of some almost periodic coercive functional*, Rend. Accad. Naz. Sci. XL, 103 (1985), pp. 313-322.

[13] D. CIORANESCU AND J. SAINT JEAN PAULIN, *Homogenization in Open Sets with Holes*, J. Math. Anal. Appl., 71 (1979), pp. 590-607.

[14] M. CRANDALL, H. ISHII, AND P. L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., to appear.

[15] G. DAL MASO AND A. DEFRANCESCHI, *Correctors for the homogenization of monotone operators*, Differential Integral Equations, 3 (1990), pp. 1151-1166.

[16] E. DE GIORGI, *Sulla convergenza di alcune successioni di integrali del tipo dell'area*, Rend. Mat., 8 (1975), pp. 277-294.

[17] ———, *G-operators and Γ-convergence*, Proceedings of the International Congress of Mathematicians Warsazwa, August 1983, PWN Polish Scientific Publishers and North-Holland, Amsterdam, 1984, pp. 1175-1191.

[18] R. DIPERNA, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223-270.

[19] W. E, *Homogenization of linear and nonlinear transport equations*, Comm. Pure Appl. Math., 45 (1992), pp. 301-326.

[20] W. E AND D. SERRE, *Correctors for the homogenization of conservation laws with oscillatory forcing terms*, Asymptotic Anal., 5 (1992), pp. 311-316.

[21] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Paris, 1974.

[22] L. C. EVANS, *The perturbed test function method for viscosity solutions of non-linear partial differential equations*, Proc. Roy. Soc. Edinburgh, to appear.

[23] ———, *Periodic homogenization of certain fully non-linear partial differential equations*, to appear.

[24] G. FRANCFORT, personal communication.

[25] P. GÉRARD AND F. MURAT, personal communication.

[26] T. HOU AND X. XIN, *Homogenization of linear transport equations with oscillatory vector fields*, SIAM J. Appl. Math., 52 (1992), pp. 34-45.

[27] J. B. KELLER, *Darcy's law for flow in porous media and the two-space method*, Lecture Notes in Pure and Appl. Math., 54, Dekker, New York, 1980.

[28] S. KOZLOV, O. OLEINIK, AND V. ZHIKOV, *Homogenization of parabolic operators with almost-periodic coefficients*, Mat. Sbornik, 117 (1982), pp. 69-85.

[29] J. L. LIONS, *Homogénéisation nonlocale*, Proceedings Internat. Meeting on Recent Methods in Non-Linear Analysis, De Giorgi et al., eds, Pitagora, Bologne, 1979, pp. 189-203.

[30] ———, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Science Press, Beijing, Gordon and Breach, New York, 1981.

[31] P. MARCELLINI, *Periodic solutions and homogenization of non linear variational problems*, Ann. Mat. Pura Appl. (4), 117 (1978), pp. 139-152.

[32] L. MASCARENHAS, *Γ-limite d'une fonctionnelle liée à un phénomène de mémoire*, C. R. Acad. Sci. Paris, 313 (1991), pp. 67-70.

[33] S. MULLER, *Homogenization of nonconvex integral functionals and cellular materials*, Arch. Rational Mech. Anal. (1988), pp. 189-212.

[34] F. MURAT, *H-convergence*, Séminaire d'Analyse Fonctionnelle et Numérique de l'Université d'Alger, mimeographed notes, 1978.

[35] ———, *Correctors for monotone problems in non-periodic homogenization*, to appear.

[36] G. NGUETSENG, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608-623.

[37] ———, *Asymptotic analysis for a stiff variational problem arising in mechanics*, SIAM J. Math. Anal., 21 (1990), pp. 1394-1414.

[38] O. OLEINIK AND V. ZHIKOV, *On the homogenization of elliptic operators with almost-periodic coefficients*, Rend. Sem. Mat. Fis. Milano, 52 (1982), pp. 149-166.

[39] G. PANASENKO, *Multicomponent homogenization of processes in strongly nonhomogeneous structures*, Math. USSR Sbornik, 69 (1991), pp. 143-153.

[40] E. SANCHEZ-PALENCIA, *Nonhomogeneous media and vibration theory*, Lecture Notes in Phys. 127, Springer-Verlag, New York, 1980.

[41] S. SPAGNOLO, *Convergence in energy for elliptic operators*, in Numerical Solutions of Partial Differential Equations III Synspade 1975, B. Hubbard, ed., Academic Press, New York, 1976.

[42] L. TARTAR, *Cours Peccot au Collège de France*, partially written by F. Murat in Séminaire d'Analyse Fonctionelle et Numérique de l'Université d'Alger, unpublished.

[43] ———, *Topics in Nonlinear Analysis*, Publications mathématiques d'Orsay 78.13, Université de Paris-Sud, 1978.

[44] ———, *Convergence of the homogenization process*, Appendix of *Nonhomogeneous media and vibration theory*, Lecture Notes in Phys. 127, Springer-Verlag, New York, 1980.

[45] ———, *Compensated compactness and applications to partial differential equations*, Nonlinear analysis and mechanics, Heriot-Watt Symposium IV, Research Notes in Math., 39, R. J. Knops, ed., Pitman, Boston, MA, 1979, pp. 136-212.

[46] ———, *Nonlocal effects induced by homogenization*, in Partial Differential Equations and the Calculus of Variations, Essays in Honor of Ennio De Giorgi, F. Colombini et al., eds., Birkhauser-Verlag, Basel, Switzerland, 1989.

[47] R. TEMAM, *Navier-Stokes Equations*, North-Holland, 1979.

[48] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1964.

[49] V. ZHIKOV, S. KOZLOV, O. OLEINIK, AND K. NGOAN, *Averaging and G-convergence of differential operators*, Russian Math. Surveys, 34 (1979), pp. 69-147.

ENERGY MOMENTS IN TIME AND FREQUENCY FOR TWO-SCALE DIFFERENCE EQUATION SOLUTIONS AND WAVELETS*

LARS F. VILLEMOES†

Abstract. This paper indicates how to find energy moments in direct and Fourier space of a solution to the functional equation $u(x) = \sum_{k=0}^{N-1} 2c_k u(2x-k)$ and shows that the Sobolev regularity of u is determined by the spectral radius of a matrix defined from the coefficients (c_k). The results are applied to compactly supported orthonormal wavelets.

Key words. two-scale difference equations, wavelets, Sobolev regularity

AMS(MOS) subject classifications. 39B32, 42C15, 94A11

1. Introduction.

A two-scale difference equation is a functional equation of the type

$$u(x) = \sum_{k=0}^{N-1} 2c_k u(2x-k), \tag{1.1}$$

where $c_0, c_1, \ldots, c_{N-1}$ are given complex numbers. Such equations arise in many contexts where the concept of changing scale is employed, such as the construction of orthonormal and biorthogonal bases of compactly supported wavelets derived from a multiresolution analysis [D1], [M], [C], [VH], [S], subdivision schemes for computer aided design [CDM], [D], [DD], and subband or pyramid based coding of speech and images [BA], [CMQW], [Ma], [Ri].

In all such applications it is important to control the properties of the solution u from the choice of the coefficients (c_k). We will focus here on the concentration in time and frequency as well as the regularity (smoothness) of u. More specifically, we will explain (in §§6–7) how to find the energy moments in time and frequency,

$$\int_{-\infty}^{+\infty} x^n |u(x)|^2 \, dx \quad \text{and} \quad \int_{-\infty}^{+\infty} \xi^n |\hat{u}(\xi)|^2 \, d\xi, \qquad n = 0, 1, 2, \ldots,$$

when these exist. ($\hat{u}$ denotes the Fourier transform of u.) The number of existing moments in frequency is closely related to Sobolev regularity, and in §9 we present an exact criterion for u to be in the Sobolev space $H^s(\mathbb{R})$ (meaning that $(1+\xi^2)^{s/2}\hat{u} \in L^2(\mathbb{R})$). This optimal regularity estimate relies on the calculation of the spectral radius of a matrix defined from the coefficients (c_k). We describe how this approach actually unifies and improves the Fourier transformation based Hölder regularity estimation methods presented in [D1] and [C]. Finally, §10 is devoted to the study of compactly supported orthonormal wavelets. All results of the paper, except for those of the last

*Received by the editors July 3, 1991; accepted for publication (in revised form) February 4, 1992.

†Mathematical Institute, The Technical University of Denmark, DK 2800 Lyngby, Denmark. Current address: Department of Mathematics, Royal Institute of Technology, S10044 Stockholm, Sweden.

section, have straightforward generalizations to the case where the dilation factor 2 in (1.1) is replaced by any integer $M \geq 2$. For a discussion of more general two-scale difference equations and the history of the subject, we refer to [DL1].

Many of the sharpest results of this paper depend on a technical condition, described in §5, which in some sense requires (c_k) to be nondegenerate. We refer to [C] for a more detailed description of the corresponding degenerate case. A central result which does not depend on the condition of §5, however, is the exact criterion for u to be square integrable, formulated in §4. The role played by *sum rules* is explained in §8, and we show here that maximal Sobolev regularity for a given number of coefficients N is obtained when u is a basic spline.

The starting point of the paper is an existence and uniqueness theorem defining the compactly supported distribution solution to (1.1) that we consider here.

2. The compactly supported solution φ. Assume that u is a compactly supported distribution solving (1.1). Then necessarily,

$$2\mathrm{supp}(u) \subset \mathrm{supp}(u) + [0, N-1],$$

and therefore

$$\mathrm{supp}(u) \subset [0, N-1].$$

The Fourier transform $\hat{u}$ of u can be extended to an entire function, and with the definition

$$m_0(\xi) = \sum_{k=0}^{N-1} c_k e^{-ik\xi}, \tag{2.1}$$

we can write (1.1) in the equivalent form,

$$\hat{u}(2\xi) = m_0(\xi)\hat{u}(\xi). \tag{2.2}$$

Hence, if $\hat{u}(0) \neq 0$, we must have $\sum c_k = m_0(0) = 1$. This is essentially the only interesting case, since a zero of $\hat{u}$ must be of finite order if $u \neq 0$. Thus, solutions to (2.2) with $\hat{u}(0) = 0$ are just finite order derivatives of solutions of the first kind. Now fix the normalization of u to $\hat{u}(0) = 1$. By iterating (2.2) we get

$$\hat{u}(\xi) = \prod_{j=1}^{+\infty} m_0(2^{-j}\xi),$$

where the infinite product must converge uniformly on every compact subset of $\mathbb{C}$. This shows the uniqueness statement of the following.

THEOREM 2.1. *Let $c_0, \ldots, c_{N-1} \in \mathbb{C}$ with $\sum c_k = 1$. Then there is a unique compactly supported distribution solution $u = \varphi$ to* (1.1) *with $\hat{u}(0) = 1$. The support of φ is contained in $[0, N-1]$, and*

$$\hat{\varphi}(\xi) = \prod_{j=1}^{+\infty} m_0(2^{-j}\xi).$$

Proof. By the discussion above and the obvious formal result $\hat{\varphi}(2\xi) = m_0(\xi)\hat{\varphi}(\xi)$, we only have to prove that the infinite product converges to the Fourier transform of a compactly supported distribution. This is a result of Deslauriers and Dubuc [DD], presented with a different proof in [DL1]. For completeness, we sketch the main ideas of both proofs.

Define the distributions u_n, $n = 1, 2, \dots$, by

$$\hat{u}_n(\xi) = \prod_{j=1}^{n} m_0(2^{-j}\xi).$$

Since $m_0(0) = 1$, we have $|m_0(\xi) - 1| \le C|\xi|$ for some positive constant C. This ensures that $\hat{u}_n$ converges uniformly on every compact subset of $\mathbb{C}$ towards an entire function F as $n \to +\infty$. In particular we have, for all n,

$$\sup_{|\xi|\le 2} |\hat{u}_n(\xi)| \le M < +\infty.$$

Using this together with the easy estimate that for some $R \ge 0$ and $B \ge 1$,

$$|m_0(\xi)| \le B e^{R|\mathrm{Im}\,\xi|},$$

we can then prove the fundamental estimate,

$$|\hat{u}_n(\xi)| \le M(1 + |\xi|)^{\log_2 B} e^{R|\mathrm{Im}\,\xi|}, \tag{2.3}$$

by considering the cases $2^m \le |\xi| \le 2^{m+1}$, $m = 1, 2, \dots$, and $|\xi| \le 2$ separately. Letting $n \to +\infty$, we see that the same inequality holds for $F(\xi)$, so by the Paley–Wiener theorem for distributions [R, p. 183], this function is the Fourier transform of a compactly supported distribution φ.

Alternatively, we could have observed that for each n, u_n is a linear combination of Dirac masses supported in $[0, N-1]$. Using (2.3) for real ξ then provides a uniform majorization of polynomial growth, sufficient to ensure the convergence of u_n towards φ in the sense of tempered distributions. □

From this point on, we will always assume $\sum c_k = 1$, define m_0 from (2.1), and let φ denote the solution described in Theorem 2.1. Note that in the case $N = 1$, this solution is just the Dirac mass at $x = 0$.

Example 2.2. (The Haar solution.) Let $N = 2$ and $(c_0, c_1) = (\frac{1}{2}, \frac{1}{2})$. Then $\varphi = \mathbf{1}_{[0,1]}$, the characteristic function of the interval $[0, 1]$. Indeed, except for $x = \frac{1}{2}$,

$$\varphi(x) = \varphi(2x) + \varphi(2x - 1),$$

and φ is integrable with $\int \varphi = \hat{\varphi}(0) = 1$. We call this φ the Haar solution, since the corresponding orthonormal wavelet (see §10) is the well-known Haar function [H].

3. Two useful operators. Let $C(\mathbb{T})$ be the space of 2π-periodic, complex continuous functions on $\mathbb{R}$. Define the linear operators A and A' on $C(\mathbb{T})$ as follows:

$$\begin{aligned} Af(\xi) &= \left|m_0\left(\frac{\xi}{2}\right)\right|^2 f\left(\frac{\xi}{2}\right) + \left|m_0\left(\frac{\xi}{2} + \pi\right)\right|^2 f\left(\frac{\xi}{2} + \pi\right), \\ A'f(\xi) &= 2|m_0(\xi)|^2 f(2\xi). \end{aligned}$$

Here A' is the transpose of A in the following sense.

LEMMA 3.1. *For all $f, g \in C(\mathbb{T})$, we have*

$$\int_{-\pi}^{\pi} Af(\xi)g(\xi)\,d\xi = \int_{-\pi}^{\pi} f(\xi)A'g(\xi)\,d\xi.$$

Proof. Change variable $\xi \to \xi/2$ in the right-hand side integral and use periodicity. □

We can write

$$2|m_0(\xi)|^2 = \sum_{k=1-N}^{N-1} a_0(k)e^{-ik\xi},$$
$$\text{with } a_0(k) = 2(c * \check{\bar{c}})(k) = \sum_l 2c_l\overline{c_{l-k}}.$$

Therefore, A leaves the following space of trigonometric polynomials invariant:

$$S_N = \left\{ P \in C(\mathbb{T}) \mid P(\xi) = \sum_{k=2-N}^{N-2} p(k)e^{-ik\xi}, \quad p(k) \in \mathbb{C} \right\}.$$

(For N=1, define $S_1 = S_2$.) If we assume $c_0c_{N-1} \neq 0$, then $A(S_k) \subset S_k$ exactly when $k \in \{N, N+1\}$, so S_N is the smallest nontrivial space of symmetrically truncated Fourier series which is invariant for A. By identifying the elements of S_N with their coefficients $p \in \mathbb{C}^{2N-3}$, the action of $A : S_N \to S_N$ can be described by a $(2N-3) \times (2N-3)$-matrix with entries $A_{kl} = a_0(2k-l)$, that is,

$$Ap(k) = (a_0 * p)(2k) = \sum_{l=2-N}^{N-2} a_0(2k-l)p(l), \qquad k \in \{2-N, \dots, N-2\}.$$

In the language of signal processing, A is an operator of filtering (convolution with a_0), followed by subsampling by a factor 2. If $|m_0(\xi)|^2 + |m_0(\xi+\pi)|^2 = 1$, this operator belongs the class of transition operators analyzed in [CR].

Remark 3.2. Assume the coefficients (c_k) are real. Then $m_0(-\xi) = \overline{m_0(\xi)}$, and A commutes with the parity operator

$$J : P(\xi) \mapsto P(-\xi).$$

Thus the action of A on a symmetric vector

$$(p(N-2), \dots, p(1), p(0), p(1), \dots, p(N-2))^\top \in \mathbb{C}^{2N-3}$$

can be represented by a $(N-1) \times (N-1)$-matrix A^s acting on

$$(p(0), p(1), \dots, p(N-2))^\top \in \mathbb{C}^{N-1}.$$

Using indices $\{0, 1, \dots, N-2\}$ for the entries of A^s, we find

$$A^s_{kl} = \begin{cases} A_{kl}, & l = 0, \\ A_{kl} + A_{k,-l}, & l > 0. \end{cases}$$

Similarly, the action of A on antisymmetric vectors can be described by a square matrix A^a of dimension $(N-2)$ with entries

$$A^a_{kl} = A_{kl} - A_{k,-l}, \qquad k, l \in \{1, \dots, N-2\}.$$

Also, any eigenvector of A can be chosen to be either symmetric or antisymmetric, and this reduces an eigenvalue problem for A to two problems of dimension $N-1$ and $N-2$, respectively.

Even in the general case of complex coefficients c_k, another kind of reduction in dimension can be obtained from the observation that $|m_0|^2$ is real: the action of A on conjugate symmetric vectors, $(p(-k) = \overline{p(k)})$, can be described by a $(2N-3)\times(2N-3)$-matrix with *real* entries.

Eigenvectors of A corresponding to *real* eigenvalues can be chosen conjugate symmetric, so for problems involving only the real part of the spectrum of A, we can reduce the $2N-3$ dimensions to $2N-3$ real ones.

4. A characterization of the case $\varphi \in L^2(\mathbb{R})$. Before we can define the energy moments in time and frequency of φ

$$\int_{-\infty}^{+\infty} x^n |\varphi(x)|^2 \, dx \quad \text{and} \quad \int_{-\infty}^{+\infty} \xi^n |\hat{\varphi}(\xi)|^2 \, d\xi,$$

we must at least have $\varphi \in L^2(\mathbb{R})$, i.e., that φ can be represented by a measurable function with

$$\|\varphi\|_2^2 = \int_{-\infty}^{+\infty} |\varphi(x)|^2 \, dx < +\infty.$$

If this is the case, all moments in time are well defined since φ has compact support. We postpone the discussion of the existence of moments in frequency to §9.

Assume $\varphi \in L^2(\mathbb{R})$. Then we can define a trigonometric polynomial $P_0 \in S_N$ by its coefficients

$$p_0(k) = \int_{-\infty}^{+\infty} \varphi(x)\overline{\varphi(x-k)} \, dx.$$

In fact, using Plancherel's theorem, we find that for almost every ξ:

$$P_0(\xi) = \sum_{k\in\mathbb{Z}} |\hat{\varphi}(\xi + 2\pi k)|^2.$$

Even better, this infinite sum converges uniformly and unconditionally towards P_0 on $[-\pi, \pi]$, because $x\varphi(x)$ is square integrable [M, p. 29].Observe that P_0 is real and nonnegative, with

$$P_0(0) \geq |\hat{\varphi}(0)|^2 = 1 > 0,$$

and a simple use of $\hat{\varphi}(2\xi) = m_0(\xi)\hat{\varphi}(\xi)$ yields

$$\begin{aligned} P_0(2\xi) &= \sum_{k\in\mathbb{Z}} |m_0(\xi + \pi k)|^2 |\hat{\varphi}(\xi + \pi k)|^2, \\ &= |m_0(\xi)|^2 P_0(\xi) + |m_0(\xi + \pi)|^2 P_0(\xi + \pi), \end{aligned}$$

that is, $AP_0 = P_0$. It turns out that the existence of such a trigonometric polynomial also implies $\varphi \in L^2(\mathbb{R})$:

THEOREM 4.1. *We have $\varphi \in L^2(\mathbb{R})$ if and only if there exists a real nonnegative trigonometric polynomial $P \in S_N$ satisfying $P(0) > 0$ and $AP = P$.*

Proof. We have seen above that if $\varphi \in L^2(\mathbb{R})$, then $P = P_0$ is a trigonometric polynomial with the desired properties.

Conversely, assume we have a P as described, and normalize P to satisfy

$$P(0) = 1.$$

Define for every $n = 0, 1, \ldots,$

$$g_n(\xi) = P(2^{-n}\xi)\left(\prod_{j=1}^{n} |m_0(2^{-j}\xi)|^2\right) \mathbf{1}_{[-\pi,\pi]}(2^{-n}\xi),$$

where $\mathbf{1}_\Omega$ denotes the characteristic function of a set Ω, and we set the product equal to one when $n = 0$. As $n \to +\infty$, g_n converges pointwise to $|\hat{\varphi}|^2$, and as a consequence of Lemma 3.1,

$$\begin{aligned}\int_{-\infty}^{+\infty} g_n(\xi)\, d\xi &= \int_{-\pi}^{\pi} P(\xi) \prod_{j=0}^{n-1} 2|m_0(2^j\xi)|^2\, d\xi,\\ &= \int_{-\pi}^{\pi} P(\xi)(A')^n(1)(\xi)\, d\xi = \int_{-\pi}^{\pi} A^n P(\xi)\, d\xi,\\ &= \int_{-\pi}^{\pi} P(\xi)\, d\xi = 2\pi p(0).\end{aligned}$$

Hence, by Fatou's lemma and Plancherel's theorem, $\|\varphi\|_2^2 \le p(0) < +\infty$. □

The proof of Theorem 4.1 shows that

$$\|\varphi\|_2^2 \le \inf\{p(0) \mid P \in S_N, P \ge 0, AP = P, P(0) = 1\}.$$

Admit for the moment the following result, which is a special case ($n = 0$) of Theorem 7.1.

LEMMA 4.2. *If $\varphi \in L^2(\mathbb{R})$, then $P_0(0) = 1$.*

Then the next corollary is obvious from the fact that $p_0(0) = \|\varphi\|_2^2$.

COROLLARY 4.3. *If $\varphi \in L^2(\mathbb{R})$, then*

$$\|\varphi\|_2^2 = \min\{p(0) \mid P \in S_N, P \ge 0, AP = P, P(0) = 1\}.$$

To illustrate the use of Theorem 4.1 and Corollary 4.3, let us consider a simple example, which follows.

Example 4.4. Consider the case of $N = 3$ coefficients. We will apply Theorem 4.1 to decide which choices lead to $\varphi \in L^2$. Evaluating $AP = P$ at $\xi = 0$ yields

$$|m_0(\pi)|^2 P(\pi) = 0.$$

If $P \in S_3$, $P \ge 0$, and $P(\pi) = 0$, then $P(\xi)$ must be a positive multiple of $\cos^2 \xi/2$. Evaluating $AP = P$ at $\xi = \pi/2$ now gives $m_0(\pi/2) = m_0(-\pi/2) = 0$ and this fixes

$$m_0(\xi) = \frac{1 + e^{-2i\xi}}{2}.$$

Hence, φ is just a dilation of the Haar solution of Example 2.2: $\varphi = \frac{1}{2}\mathbf{1}_{[0,2]} \in L^2$.

If $P(\pi) \neq 0$ we must have $m_0(\pi) = 0$. A parametrization of all the possible choices of coefficients in this case is given by

$$(c_0, c_1, c_2) = \left(\frac{1+re^{i\theta}}{4}, \frac{1}{2}, \frac{1-re^{i\theta}}{4}\right)$$

with $r \geq 0$ and $-\pi \leq \theta < \pi$. The 3×3-matrix representing A has $\lambda = 1$ as eigenvalue for all r, θ and we find that necessarily,

$$p = t\left(\frac{1-r^2+2ir\sin\theta}{4-4ir\sin\theta}, 1, \frac{1-r^2-2ir\sin\theta}{4+4ir\sin\theta}\right)^{\mathsf{T}}.$$

Here $P \geq 0$ implies $t = p(0) = (1/2\pi)\int_{-\pi}^{\pi} P(\xi)\,d\xi \geq 0$ and since we have

$$P(0) = \sum p(k) = t\frac{3-r^2}{2+2r^2\sin^2\theta},$$

we see that $P(0) > 0$ only if $r < \sqrt{3}$. Conversely, it turns out that $P \geq 0$ if $r < \sqrt{3}$. (It suffices to show that $2|p(1)| \leq p(0)$.)

To conclude we have shown that $\varphi \in L^2$ if and only if

$$(c_0, c_1, c_2) \in \left\{\left(\frac{1+z}{4}, \frac{1}{2}, \frac{1-z}{4}\right) \mid |z| < \sqrt{3}\right\} \cup \left\{\left(\frac{1}{2}, 0, \frac{1}{2}\right)\right\}.$$

In the first case, Corollary 4.3 gives the result

$$\int_{-\infty}^{+\infty} |\varphi(x)|^2\,dx = \frac{2+2(\operatorname{Im} z)^2}{3-|z|^2}.$$

5. The Riesz basis property. If $(e_k)_{k\in\mathbb{Z}}$ is a sequence of vectors in a Hilbert space H, such that the collection of finite linear combinations $\sum \alpha_k e_k$ are dense in H, and there exist constants $C_1 > 0$, $C_2 < +\infty$, such that for all these finite sums,

$$C_1 \sum |\alpha_k|^2 \leq \left\|\sum \alpha_k e_k\right\|_H^2 \leq C_2 \sum |\alpha_k|^2,$$

then $(e_k)_{k\in\mathbb{Z}}$ is called a *Riesz basis* for H.

A case where many results of the following sections get sharper, and pathologies in the convergence of subdivision schemes involving the coefficients (c_k) are avoided [C], is when φ has the *Riesz basis property*.

DEFINITION 5.1. We say that φ has the *Riesz basis property*, if $\varphi \in L^2(\mathbb{R})$ and the sequence of functions $\varphi(x-k)$, $k \in \mathbb{Z}$ forms a Riesz basis for the closure of its linear span in $L^2(\mathbb{R})$.

If $\varphi \in L^2$, it turns out that φ has the Riesz basis property if and only if there exist constants $C_1 > 0$, $C_2 < +\infty$, such that

$$C_1 \leq \sum_{k\in\mathbb{Z}} |\hat{\varphi}(\xi + 2\pi k)|^2 \leq C_2.$$

To prove this, use Plancherel's theorem (or see [M, p. 27]). We recognize the sum as $P_0(\xi)$ from §4. Since P_0 is continuous and 2π-periodic the condition above is equivalent to

$$\forall \xi : P_0(\xi) > 0.$$

Cohen [C] and Lawton [L1], [L2] have characterized this situation in different ways. We present some of the results in Theorem 5.3, but first a definition follows.

Definition 5.2. We say that m_0 satisfies *Cohen's criterion* if there is a compact set $K \in \mathbb{R}$ containing a neighborhood of $\xi = 0$ and equal to a finite union of closed intervals, such that

(1) For all but a finite number of $\xi \in [-\pi, \pi]$, there is a unique $\eta \in K$ with $\eta - \xi \in 2\pi\mathbb{Z}$.

(2) $m_0(\xi) \neq 0$ for $\xi \in \bigcup_{j=1}^{+\infty} 2^{-j}K$.

Observe that if m_0 does not vanish on $[-\pi/2, \pi/2]$ then Cohen's criterion is trivially satisfied with $K = [-\pi, \pi]$. If m_0 satisfies Cohen's criterion, we have

$$\hat{\varphi}(\xi) \neq 0 \quad \text{for } \xi \in K,$$

and if $\varphi \in L^2$, this clearly implies $P_0 > 0$. Conversely, if $P_0 > 0$, it is possible to use the uniform convergence on $[-\pi, \pi]$ of the sum defining P_0 to construct a K as in Definition 5.2. This is how to see the equivalence (1) $\Leftrightarrow$ (2) in the following.

Theorem 5.3. *The following three statements are equivalent:*

(1) *φ has the Riesz basis property.*

(2) *$\varphi \in L^2(\mathbb{R})$ and m_0 satisfies Cohen's criterion.*

(3) *Up to scalar multiplication there is a unique solution $P \in S_N$ to $AP = P$, and this solution can be chosen to be strictly positive.*

Proof. We prove only (1) $\Leftrightarrow$ (3) here. By Theorem 4.1 condition (3) implies that $\varphi \in L^2$, and since $AP_0 = P_0$, the unicity statement insures that P_0 is strictly positive. This shows that (3) implies (1).

Conversely, assume (1). Then $P = P_0$ is a strictly positive solution to $AP = P$. If $P(\xi) = \sum p(k)e^{-ik\xi}$ is another solution we find using Lemma 3.1:

$$\begin{aligned}
2\pi p(k) &= \int_{-\pi}^{\pi} P(\xi)e^{ik\xi}\,d\xi = \int_{-\pi}^{\pi} A^j P(\xi)e^{ik\xi}\,d\xi \\
&= \int_{-\pi}^{\pi} P(\xi)(A')^j(e^{ik\cdot})(\xi)\frac{\sum|\hat{\varphi}(\xi + 2\pi k)|^2}{P_0(\xi)}\,d\xi \\
&= \int_{-\infty}^{+\infty} \frac{P(\xi)}{P_0(\xi)}\left(\prod_{l=0}^{j-1} 2|m_0(2^l\xi)|^2\right) e^{ik2^j\xi}|\hat{\varphi}(\xi)|^2\,d\xi \\
&= \int_{-\infty}^{+\infty} \frac{P(2^{-j}\eta)}{P_0(2^{-j}\eta)}|\hat{\varphi}(\eta)|^2 e^{ik\eta}\,d\eta, \qquad (\eta = 2^j\xi).
\end{aligned}$$

This holds for every $j = 1, 2, \ldots$. The last integrand is dominated by $M|\hat{\varphi}|^2$ with $M = \sup_\xi(|P(\xi)|/P_0(\xi))$ which is finite since P_0 attains a strictly positive minimum on $[-\pi, \pi]$. Letting $j \to +\infty$, we conclude that P is a scalar multiple of P_0. □

Example 5.4. If $m_0(\xi) = (1 + e^{-3i\xi})/2$, it turns out that the eigenspace E_1 of $A : S_4 \to S_4$ corresponding to the eigenvalue $\lambda = 1$ is spanned by the functions

$$\{1, \quad (\tfrac{1}{2} + \cos\xi)^2\}.$$

By Theorem 4.1, $\varphi \in L^2$, and we find $\|\varphi\|_2^2 = \min[\frac{1}{3}, 1] = \frac{1}{3}$ from Corollary 4.3, but since $\dim(E_1) = 2$, Theorem 5.3(3) shows that φ does not have the Riesz basis property. In fact, $\varphi = \frac{1}{3}1_{[0,3]}$ and $P_0(\xi) = (\frac{1}{3} + \frac{2}{3}\cos\xi)^2$.

Even if the eigenspace E_1 is one-dimensional, we cannot be sure to be in the Riesz basis case. To see this, consider $m_0(\xi) = (1 + e^{-2i\xi})/2$. Then $\dim(E_1) = 1$ and $P_0(\xi) = \cos^2(\xi/2)$. Again φ is square integrable, but does not have the Riesz basis property since $P_0(\pi) = 0$.

6. Calculating moments in time. Assume $\varphi \in L^2$. For $n = 1, 2, \ldots$, and $k \in \mathbb{Z}$, we can then define the moments in time

$$q_n(k) = \int_{-\infty}^{+\infty} x^n \varphi\left(x + \frac{k}{2}\right) \overline{\varphi\left(x - \frac{k}{2}\right)} \, dx, \tag{6.1}$$

and as a discrete analogue, generalizing a_0 of §3,

$$a_n(k) = 2 \sum_{l \in \mathbb{Z}} \left(l - \frac{k}{2}\right)^n c_l \overline{c_{l-k}}. \tag{6.2}$$

By definition both q_n and a_n are conjugate symmetric, $q_n(k) = 0$ for $|k| > N - 2$, and $a_n(k) = 0$ for $|k| > N - 1$. In §4 we related $q_0 = p_0$ to a_0. In Proposition 6.1, we will generalize these relations to the cases $n = 1, 2, \ldots$ as well. Once q_n is found, we obtain as a special case the energy moment in time,

$$q_n(0) = \int_{-\infty}^{+\infty} x^n |\varphi(x)|^2 \, dx.$$

The extra information contained in $q_n(k)$ for $k \neq 0$ will be needed in §10, were we consider wavelets associated to φ.

In analogy with the definition of A we can define operators $A_0, A_1, A_2, \ldots$, acting on $S_N \simeq \mathbb{C}^{2N-3}$ by

$$A_n q(k) = \sum_l a_n(2k - l) q(l). \tag{6.3}$$

Because of the definition of a_n, these operators have the same symmetry properties as A in Remark 3.2.

PROPOSITION 6.1. *Assume $\varphi \in L^2(\mathbb{R})$, then*

$$q_n = 2^{-n} \sum_{p=0}^{n} \binom{n}{p} A_p q_{n-p} \quad \textit{for } n = 0, 1, 2, \ldots. \tag{6.4}$$

If φ has the Riesz basis property, then the q_n are uniquely determined by (6.4) *and the requirement $\sum_k q_0(k) = 1$.*

Proof. By inserting $\varphi(x) = \sum_k 2c_k \varphi(2x - k)$ into (6.1) we find

$$\begin{aligned}
q_n(k) &= \sum_{l,m} 4 c_l \overline{c_m} \int x^n \varphi(2x + k - l) \overline{\varphi(2x - k - m)} \, dx \\
&= 2^{-n} \sum_{l,m} 2 c_l \overline{c_m} \int \left(y + \frac{l+m}{2}\right)^n \varphi\left(y + k + \frac{m-l}{2}\right) \overline{\varphi\left(y - k + \frac{l-m}{2}\right)} \, dy \\
&\qquad \left(\text{where } y = 2x - \frac{l+m}{2}\right) \\
&= 2^{-n} \sum_{l,m} 2 c_l \overline{c_m} \sum_{p=0}^{n} \binom{n}{p} \left(\frac{l+m}{2}\right)^p q_{n-p}(2k - m - l) \\
&= 2^{-n} \sum_{p=0}^{n} \binom{n}{p} \sum_s \left(\sum_p (l - s/2)^p c_l \overline{c_{l-s}}\right) q_{n-p}(2k - s) \qquad (s = l - m) \\
&= 2^{-n} \sum_{p=0}^{n} \binom{n}{p} A_p q_{n-p}(k),
\end{aligned}$$

which shows (6.4).

Now, assume in addition that φ has the Riesz basis property. Then from Theorem 5.3(3) and Lemma 4.2 we know that $q_0 = p_0$ is uniquely defined by $Aq_0 = q_0$ and $\sum_k q_0(k) = 1$. To find $q_1, q_2, \dots$, we rearrange the terms of (6.4) to get

$$(I - 2^{-n}A)q_n = 2^{-n}\sum_{p=1}^{n}\binom{n}{p}A_p q_{n-p}.$$

If $I - 2^{-n}A$ is invertible, then q_n is uniquely determined from $q_{n-1}, q_{n-2}, \dots, q_0$. This fixes all q_n by iteration, if we can show that the spectral radius $\rho(A)$ of $A : S_N \to S_N$ is less than 2. We will show that $\rho(A) = 1$.

Let $P \in S_N$ be a strictly positive trigonometric polynomial with $AP = P$ (furnished by Theorem 5.3(3)), and equip S_N with the following weighted supremum norm:

$$\|X\|_P = \sup_{\xi\in[-\pi,\pi]} \frac{|X(\xi)|}{P(\xi)}.$$

Let us try to estimate $\|AX\|_P$. For all ξ, we have

$$\begin{aligned}\frac{|AX(2\xi)|}{P(2\xi)} &\leq \frac{|m_0(\xi)|^2P(\xi)}{P(2\xi)}\frac{|X(\xi)|}{P(\xi)} + \frac{|m_0(\xi+\pi)|^2P(\xi+\pi)}{P(2\xi)}\frac{|X(\xi+\pi)|}{P(\xi+\pi)}\\ &\leq \frac{AP(2\xi)}{P(2\xi)}\|X\|_P = \|X\|_P.\end{aligned}$$

Hence, $\|AX\|_P \leq \|X\|_P$ and $\rho(A) \leq \|A\|_P \leq 1$. To see that equality holds, use $X = P$. □

7. Calculating moments in frequency. For each $s \in \mathbb{R}$, define the Sobolev space $H^s(\mathbb{R})$ to be the space of tempered distributions u such that $(1+\xi^2)^{s/2}\hat{u} \in L^2(\mathbb{R})$. Given $n \in \mathbb{N}_0$, assume $\varphi \in H^{n/2}$ and define

$$p_n(k) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \xi^n|\hat{\varphi}(\xi)|^2 e^{ik\xi}\,d\xi. \tag{7.1}$$

Note that since $\hat{\varphi}$ is continuous, $\varphi \in H^{n/2}$ is exactly the condition ensuring that the integral in (7.1) is absolutely convergent. In other words, we ensure that the moment in frequency

$$p_n(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \xi^n|\hat{\varphi}(\xi)|^2\,d\xi,$$

exists.

The definition of p_n is consistent with §4 in the case $n = 0$, and p_n is conjugate symmetric, $p_n(-k) = \overline{p_n(k)}$. For even $n = 2m$, the mth derivative of φ is in L^2 and

$$p_{2m}(k) = \int_{-\infty}^{+\infty} \varphi^{(m)}(x)\overline{\varphi^{(m)}(x-k)}\,dx.$$

Define Φ_n by

$$\hat{\Phi}_n(\xi) = \xi^n|\hat{\varphi}(\xi)|^2. \tag{7.2}$$

If $\varphi \in H^{n/2}$, then $\hat{\Phi}_n$ is integrable so Φ_n itself is continuous and by definition

$$p_n(k) = \Phi_n(k).$$

Since $\Phi_n = (-i)^n\varphi^{(n)} * \check{\overline{\varphi}}$, the support of Φ_n is contained in $[1-N, N-1]$, and by continuity $p_n(k) = 0$ for $|k| > N-2$. Thus $P_n(\xi) = \sum_k p_n(k)e^{-ik\xi}$ defines an element of S_N.

THEOREM 7.1. *Let $n \in \mathbb{N}_0$. If $\varphi \in H^{n/2}$, then*

$$Ap_n = 2^{-n}p_n, \tag{7.3}$$

$$\sum_k k^l p_n(k) = \begin{cases} 0, & l = 0, 1, \ldots, n-1, \\ i^n n!, & l = n. \end{cases} \tag{7.4}$$

Furthermore, p_n is the only solution to (7.3)–(7.4) if φ has the Riesz basis property.

Proof. Observe that $\hat{\Phi}_n(2\xi) = 2^n|m_0(\xi)|^2\hat{\Phi}_n(\xi)$ and for almost every ξ,

$$P_n(\xi) = \sum_{k\in\mathbb{Z}} \hat{\Phi}_n(\xi + 2\pi k).$$

Generalizing the result for P_0 in §4, we obtain

$$P_n(2\xi) = 2^n\big(|m_0(\xi)|^2 P_n(\xi) + |m_0(\xi+\pi)|^2 P_n(\xi+\pi)\big),$$

which is equivalent to (7.3).

To show (7.4) consider the Riemann sums

$$I_n^l(j) = 2^{-j}\sum_k (2^{-j}k)^l \Phi_n(2^{-j}k)$$

for $l = 0, 1, \ldots, n$ and $j \in \mathbb{N}_0$. For $j = 0$ we have

$$I_n^l(0) = \sum_k k^l p_n(k),$$

and since Φ_n is continuous with compact support,

$$I_n^l(j) \to \int_{-\infty}^{+\infty} x^l \Phi_n(x)\,dx \quad \text{for } j \to +\infty.$$

If $a_0(k)$ are the Fourier coefficients of $2|m_0|^2$ as in §3 we have

$$\Phi_n(x) = 2^n \sum_k a_0(k)\Phi_n(2x - k).$$

Inserting this into the definition of I_n^l yields

$$I_n^l(j+1) = 2^{n-l} I_n^l(j) + \sum_{r=0}^{l-1} \gamma(r, j, l, n) I_n^r(j). \tag{7.5}$$

The exact expression for γ is of no importance here. From the inclusion $H^{s'} \subset H^s$ for $s' \geq s$ we see that (7.5) holds with n replaced by any $n' = 0, 1, \ldots, n$. If $n \geq 1$ we have in particular $I_1^0(j+1) = 2I_1^0(j)$ and since the sequence $I_1^0(j)$ converges for $j \to +\infty$ it must be identically zero. By induction, using a similar argument, we obtain in fact

$$I_n^l(j) = 0 \quad \text{for } 0 \leq l < n, \quad j = 0, 1, \ldots.$$

The special case $j = 0$ proves the homogeneous part of (7.4).

For $l = n$ the last term of (7.5) now disappears so $I_n^n(j)$ is independent of j. We thus find

$$\begin{aligned}\sum_k k^n p_n(k) &= I_n^n(0) = \lim_{j\to+\infty} I_n^n(j)\\ &= \int_{-\infty}^{+\infty} x^n \Phi_n(x)\,dx = i^n \hat{\Phi}^{(n)}(0)\\ &= i^n n!\end{aligned}$$

and this concludes the proof of the first part of the theorem.

If φ has the Riesz basis property, then P_0 is a strictly positive trigonometric polynomial and we have just seen that $P_0(0) = 1$. As shown in the case $n = 0$ in the proof of Theorem 5.3, a solution p to $Ap = 2^{-n}p$ must satisfy

$$p(k) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{2^{nj}P(2^{-j}\xi)}{P_0(2^{-j}\xi)}|\hat{\varphi}(\xi)|^2 e^{ik\xi}\,d\xi, \tag{7.6}$$

where $P(\xi) = \sum_k p(k)e^{-ik\xi}$. If p satisfies (7.4) we can write

$$P(\xi) = \left(2\sin\frac{\xi}{2}\right)^n e^{i\xi/2}R(\xi),$$

for some trigonometric polynomial R with $R(0) = 1$. Inserting this into (7.6) we see that as $j \to \infty$, the integrand converges pointwise towards $\xi^n|\hat{\varphi}(\xi)|^2 e^{ik\xi}$ dominated by $C|\xi|^n|\hat{\varphi}(\xi)|^2$ where $C = \sup(|R(\xi)|/P_0(\xi))$. By the assumption $\varphi \in H^{n/2}$, Lebesgue's dominated convergence theorem allows us to conclude that $p(k) = p_n(k)$, thereby showing the last statement of Theorem 7.1. □

Remark. A different proof of Theorem 7.1 has been indicated by Beylkin [B], in the special case where φ is the scaling function of a multiresolution analysis (see §10). Assuming sufficient regularity of φ, we have

$$p_n(k) = i^n \int_{-\infty}^{+\infty} \varphi(x)\overline{\varphi^{(n)}(x-k)}\,dx.$$

Therefore the coefficients $p_n(k)$ arise naturally when the representation of the operator d^n/dx^n in a basis of orthonormal wavelets is sought, for use in the fast numerical algorithm of [BCR].

8. Sum rules. In this section we will describe some simplifications that occur in the analysis of φ when m_0 has a zero at $\xi = \pi$. As we will see in Proposition 8.3, such a zero exists if φ has the Riesz basis property.

DEFINITION 8.1. We say (c_k) satisfies M *sum rules* if $m_0(\xi) = \sum_k c_k e^{-ik\xi}$ has a zero of order at least M at $\xi = \pi$. Equivalently,

$$\sum_{k=0}^{N-1}(-1)^k k^n c_k = 0 \quad \text{for } n = 0, 1, \ldots, M-1.$$

Observe that if (c_k) satisfies M sum rules we have the factorization

$$m_0(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^M \tilde{m}_0(\xi) \tag{8.1}$$

$$\text{with } \tilde{m}_0(\xi) = \sum_{k=0}^{N-M-1} \tilde{c}_k e^{-ik\xi}.$$

Clearly, we can have at most $N-1$ sum rules. Combining Theorem 2.1 with Example 2.2, we see that the factorization (8.1) corresponds to an M-fold convolution with the Haar solution, that is,

$$\varphi = \mathbf{1}_{[0,1]}^{*M} * \widetilde{\varphi}, \tag{8.2}$$

where $\widetilde{\varphi}$ is defined from $\widetilde{m}_0$.

A convolution with $\mathbf{1}_{[0,1]}$ has a very precise effect on the regularity of compactly supported distributions. We describe this in terms of Sobolev regularity in Lemma 8.2, but note that essentially the same proof applies to show the result in terms of Hölder regularity (C^α spaces). This observation could have simplified the higher regularity estimates of [DL2].

LEMMA 8.2. *If $s \in \mathbb{R}$ and u is a compactly supported distribution, then*

$$u \in H^s(\mathbb{R}) \Leftrightarrow \mathbf{1}_{[0,1]} * u \in H^{s+1}(\mathbb{R}).$$

Proof. Define $v = \mathbf{1}_{[0,1]} * u$. If τ_h denotes the operator of translation with h and δ_h is the Dirac mass at $x = h$ we have

$$v' = \mathbf{1}_{[0,1]}{}' * u = (\delta_0 - \delta_1) * u = u - \tau_1 u. \tag{8.3}$$

If $u \in H^s$, we see that $\hat{v}(\xi) = ((1 - e^{-i\xi})/i\xi)\hat{u}(\xi)$, and since the first factor is dominated by $2(1+\xi^2)^{-1/2}$, it follows that $v \in H^{s+1}$. (Note that the boundedness of the support of u is not necessary here.)

Conversely, if $v \in H^{s+1}$ we have $v' \in H^s$ and we will obtain $u \in H^s$ by telescoping (8.3). Indeed we can write for every $n = 1, 2, \ldots,$

$$u - \tau_n u = \sum_{k=0}^{n-1} \tau_k v' \in H^s.$$

For n large enough the supports of u and $\tau_n u$ are disjoint and we can choose a smooth compactly supported test function η, such that $\eta = 1$ on $\mathrm{supp}(u)$ and $\eta = 0$ on $\mathrm{supp}(\tau_n u)$. Since H^s is stable under multiplication by test functions [R, p.199], we conclude that $u = \eta(u - \tau_n u) \in H^s$. □

Since the Dirac mass δ_0 is in H^t if and only if $t < -\frac{1}{2}$, a simple consequence of Lemma 8.2 is that

$$\mathbf{1}_{[0,1]}^{*M} \in H^s \Leftrightarrow s < M - \tfrac{1}{2}. \tag{8.4}$$

In particular, $\varphi = \mathbf{1}_{[0,1]}^{*M}$ is a two-scale difference equation solution with $\varphi \in H^{N-2}$. This can also be seen directly, since φ is a basic spline.

We now turn to the general Riesz basis case. Assume $\varphi \in L^2$ has the Riesz basis property, then P_0 from §4 is strictly positive and

$$P_0(2\xi) = |m_0(\xi)|^2 P_0(\xi) + |m_0(\xi+\pi)|^2 P_0(\xi+\pi).$$

Evaluating at $\xi = 0$ we obtain $m_0(\pi) = 0$, hence (c_k) satisfies one sum rule. With $M = 1$ (8.1)–(8.2) become

$$m_0(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)\widetilde{m}_0(\xi), \qquad \varphi = \mathbf{1}_{[0,1]} * \widetilde{\varphi}.$$

Now if $\varphi \in H^1$, Lemma 8.2 gives $\widetilde{\varphi} \in L^2$. Since $\widetilde{m}_0$ obviously satisfies Cohen's criterion, $\widetilde{\varphi}$ will have the Riesz basis property (Theorem 5.3). Therefore $(\widetilde{c}_k)$ satisfies one sum rule as well, and (c_k) satisfies two.

By induction, we have shown the following.

PROPOSITION 8.3. *Let $m \in \mathbb{N}_0$. If $\varphi \in H^m(\mathbb{R})$ and φ has the Riesz basis property, then (c_k) satisfies $m+1$ sum rules.*

This proposition shows that in the Riesz basis case, φ cannot belong to H^{N-1} and the only possible $\varphi \in H^{N-2}$ is given by

$$\varphi = \mathbf{1}_{[0,1]}^{*(N-1)}, \qquad m_0(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^{N-1}. \tag{8.5}$$

It turns out that this maximal regularity property of the spline solution does not depend on the Riesz basis assumption.

PROPOSITION 8.4. *If $\varphi \in H^{N-2}(\mathbb{R})$, we are in the case* (8.5).

Proof. For $N=1$ we have necessarily $\varphi = \delta_0$ and for $N=2$ the proposition follows from Example 4.4.

Assume therefore $N \geq 3$ and $\varphi \in H^{N-2}$. By downward induction on N we only have to show that (c_k) satisfies one sum rule.

Define the vectors $p_0, p_1, \ldots, p_{2N-4} \in \mathbb{C}^{2N-3}$ as in §7, and let W be the orthogonal complement of the linear span V of $p_1, p_2, \ldots, p_{2N-4}$. From Theorem 7.1 we know that

$$Ap_n = 2^{-n}p_n.$$

Thus, the dimension of V must be $2N-4$ and it follows that W is one-dimensional. If v is left eigenvector for the matrix (A_{kl}) corresponding to the eigenvalue $\lambda = 1$, we must have $v \in W$. On the other hand we have from (7.4) that $(1,1,\ldots,1) \in W$. Since $\dim(W) = 1$, we conclude that v is proportional to $(1,1,\ldots,1)$, hence

$$\sum_k A_{kl} = \sum_k a_0(2k-l) = 1,$$

which implies $|m_0(\pi)|^2 = 0$. □

Assume φ is m times continuously differentiable: $\varphi \in C^m$. Then φ is also in H^m, and Proposition 8.4 gives the lower bound $N \geq m+2$ for the number of coefficients defining φ. In fact, since $\mathbf{1}_{[0,1]}^{*(m+1)}$ is not of class C^m, we have even $N \geq m+3$. This result is well known [DL1], but in the critical case $N = m+3$, there are many different φ of class C^m. In contrast to this, Proposition 8.4 characterizes the spline solution (8.5) as the optimal choice with respect to Sobolev regularity for a given number of coefficients N.

The next lemma describes how the existence of sum rules can reduce the dimension of an eigenvalue problem for A.

LEMMA 8.5. *Let $m_0(\xi) = ((1+e^{-i\xi})/2)\widetilde{m}_0(\xi)$ and define $\widetilde{A}$ from $\widetilde{m}_0$. If $\widetilde{P} \in S_{N-1}$ is an eigenfunction for $\widetilde{A}$ corresponding to the eigenvalue λ and*

$$P(\xi) = \sin^2\left(\frac{\xi}{2}\right)\widetilde{P}(\xi),$$

then $P \in S_N$ is an eigenfunction for A corresponding to the eigenvalue $\lambda/4$.

Proof. Since $\sin\xi = 2\sin(\xi/2)\cos(\xi/2)$, we have

$$\begin{aligned} AP(2\xi) =& \cos^2\left(\frac{\xi}{2}\right)|\widetilde{m}_0(\xi)|^2\sin^2\left(\frac{\xi}{2}\right)\widetilde{P}(\xi) \\ &+ \sin^2\left(\frac{\xi}{2}\right)|\widetilde{m}_0(\xi+\pi)|^2\cos^2\left(\frac{\xi}{2}\right)\widetilde{P}(\xi+\pi) \\ =& \frac{1}{4}\sin^2(\xi)\widetilde{A}\widetilde{P}(2\xi) = \frac{\lambda}{4}P(2\xi). \end{aligned}$$

□

In parallel to this observation we see directly from the definition of Φ_n in §7, that if (c_k) satisfies M sum rules and $\varphi \in H^{n/2}$, then

$$P_n(\xi) = \left(4\sin^2\frac{\xi}{2}\right)^M \widetilde{P}_{n-2M}(\xi)$$

as long as $(n/2) \geq M$ and $\widetilde{P}_{n-2M} \in S_{N-M}$ is defined from $\widetilde{\varphi}$.

In fact, in the presence of M sum rules, any eigenfunction $P \in S_N$ for A corresponding to an eigenvalue $\lambda = 2^{-n}$ where $n < 2M$, must have a zero of order at least n at $\xi = 0$. To see this, let l be the order of the zero of P at $\xi = 0$. Expanding $AP = 2^{-n}P$ in Taylor series around $\xi = 0$ yields

$$(2^{l-n} - 1)\xi^l = \mathrm{o}(\xi^l) + \mathrm{o}(\xi^n).$$

Hence, if $l < n$, dividing by ξ^l leads to a contradicton. By Proposition 8.3, a consequence of this observation is that the homogeneous part of the condition (7.4) is superfluous when φ has the Riesz basis property.

9. Optimal Sobolev regularity estimates. For $m = 0, 1, \ldots$, we can obtain a criterion for $\varphi \in H^m(\mathbb{R})$, generalizing Theorem 4.1, just from the observation that P_{2m} is nonnegative.

THEOREM 9.1. *For $m \in \mathbb{N}_0$ we have $\varphi \in H^m(\mathbb{R})$ if and only if there exists a nonnegative trigonometric polynomial $P \in S_N$, such that $AP = 4^{-m}P$ and*

$$\sum_k k^l p(k) = \begin{cases} 0, & l = 0, 1, \ldots, 2m-1, \\ (-1)^m(2m)!, & l = 2m, \end{cases}$$

Moreover, the minimum over all such P of the central coefficient $p(0)$ equals the moment in frequency $(1/2\pi)\int \xi^{2m}|\hat{\varphi}(\xi)|^2\,d\xi$.

Proof. If $\varphi \in H^m$, then $P = P_{2m}$ will do: the eigenfunction property and corresponding normalizations follow from Theorem 7.1, and since we see from (7.2) that $\hat{\Phi}_{2m} \geq 0$, the nonnegativeness of $P(\xi) = \sum_k \hat{\Phi}_{2m}(\xi + 2\pi k)$ is clear as well.

Conversely, assume P satisfies the stated conditions. Define for $n = 1, 2 \ldots,$

$$g_n(\xi) = 4^{mn}P(2^{-n}\xi)\left(\prod_{j=1}^{n}|m_0(2^{-j}\xi)|^2\right)\mathbf{1}_{[-\pi,\pi]}(2^{-n}\xi).$$

As $n \to +\infty$, g_n converges pointwise to $\xi^{2m}|\hat{\varphi}|^2$, because of the normalization of P. Using Lemma 3.1 as in the proof of Theorem 4.1, we find that the integral of g_n equals $2\pi p(0)$ for all n. Since the g_n are nonnegative, we deduce from Fatou's lemma that

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} \xi^{2m}|\hat{\varphi}(\xi)|^2\,d\xi \leq p(0) < +\infty.$$

By the continuity of $\hat{\varphi}$, this shows that $\varphi \in H^m$.

Finally, since the above integral by definition is equal to $p_{2m}(0)$, the last statement of the theorem is evident. □

If φ has the Riesz basis property, much simpler regularity criteria can be formulated. First, the H^m-criterion of Theorem 9.1 becomes inferior to a simple combination of Proposition 8.3, Lemma 8.2, and the L^2-criterion Theorem 4.1, but we can do even better: we can find an optimal Sobolev regularity estimate s_0, in the sense that $\varphi \in H^s$ if and only if $s < s_0$, by calculating the spectral radius of a certain matrix. This optimal estimate is described in Theorem 9.5. It relies on two results: Lemma 9.2 and Lemma 9.3, which we state and prove first.

In the following, we use the supremum norm on S_N, $\|P\|_\infty = \sup_\xi |P(\xi)|$. Let $\rho(A)$ be the spectral radius of the operator $A : S_N \to S_N$ defined from m_0 as in §3. Since S_N is finite-dimensional, this radius equals the maximum modulus of the set of eigenvalues of A. The first lemma relates $\rho(A)$ and the operator norm $\|A^n\|$ to

$$\sigma_n = \int_{-\pi}^{\pi} \prod_{j=0}^{n-1} 2|m_0(2^j\xi)|^2 \, d\xi. \tag{9.1}$$

LEMMA 9.2. *There is an* $r > 0$ *such that*

$$\forall n \in \mathbb{N}: \quad r\rho(A)^n \le \sigma_n \le 2\pi \|A^n\|.$$

Proof. To show the upper bound on σ_n, observe that

$$\sigma_n = \int_{-\pi}^{\pi} (A')^n(1)(\xi)\, d\xi = \int_{-\pi}^{\pi} A^n(1)(\xi)\, d\xi \le 2\pi \|A^n\| \|1\|_\infty,$$

where we have applied Lemma 3.1 with $f = g = 1$.

Now, let λ be an eigenvalue of A with $|\lambda| = \rho(A)$, let $P \in S_N$ be a corresponding eigenfunction, and consider

$$\gamma_n = \int_{-\pi}^{\pi} P(\xi) \left(\prod_{j=0}^{n-1} 2|m_0(2^j\xi)|^2 \right) \overline{P(2^n\xi)}\, d\xi.$$

By trivial estimates we get

$$|\gamma_n| \le \|P\|_\infty^2 \sigma_n.$$

On the other hand, Lemma 3.1 gives

$$\gamma_n = \int_{-\pi}^{\pi} P(\xi)\overline{(A')^n P(\xi)}\, d\xi = \int_{-\pi}^{\pi} A^n P(\xi)\overline{P(\xi)}\, d\xi = \lambda^n \int_{-\pi}^{\pi} |P(\xi)|^2\, d\xi.$$

Thus, the lower bound on σ_n follows with $r = \|P\|_\infty^{-2} \int_{-\pi}^{\pi} |P(\xi)|^2\, d\xi$. □

The second lemma explains why we consider the sequence σ_n of (9.1).

LEMMA 9.3. *Let* $t > 0$ *and assume that* m_0 *satisfies Cohen's criterion* (*Definition* 5.2). *Then*

$$\varphi \in H^{-t}(\mathbb{R}) \Leftrightarrow \sum_{n=1}^{+\infty} \sigma_n 4^{-tn} < +\infty.$$

Proof. Let $R > 0$ be arbitrary but fixed; we will choose this parameter later. By the continuity of $\hat{\varphi}$, we have

$$\varphi \in H^{-t} \Leftrightarrow \int_{|\xi|>R} |\xi|^{-2t} |\hat{\varphi}(\xi)|^2\, d\xi < +\infty.$$

With the dyadic decomposition $\{|\xi| > R\} = \bigcup_{n=1}^{+\infty} \Omega_n$, where

$$\Omega_n = \{\xi \in \mathbb{R} \mid 2^{n-1}R < |\xi| \leq 2^n R\},$$

we obtain, by simple estimations of $|\xi|^{-2t}$ on each Ω_n,

$$\varphi \in H^{-t} \Leftrightarrow \sum_{n=1}^{+\infty} 4^{-tn} \int_{\Omega_n} |\hat{\varphi}(\xi)|^2 \, d\xi < +\infty.$$

Next, with $I_n = \int_{-2^n R}^{2^n R} |\hat{\varphi}(\xi)|^2 \, d\xi$, we find for each $k = 2, 3, \ldots,$

$$\begin{aligned}\sum_{n=1}^{k} 4^{-tn} \int_{\Omega_n} |\hat{\varphi}(\xi)|^2 \, d\xi &= \sum_{n=1}^{k} 4^{-tn}(I_n - I_{n-1}) \\ &= 4^{-tk} I_k - 4^{-t} I_0 + (1 - 4^{-t}) \sum_{n=1}^{k-1} 4^{-tn} I_n.\end{aligned}$$

Using the fact that $t > 0$, we can thus replace $\int_{\Omega_n} |\hat{\varphi}(\xi)|^2 \, d\xi$ with I_n:

$$\varphi \in H^{-t} \Leftrightarrow \sum_{n=1}^{+\infty} I_n 4^{-tn} < +\infty.$$

To complete the proof of the lemma, we now only have to show that I_n is equivalent to σ_n for $n \longrightarrow +\infty$.

By a change of variable, we can write

$$I_n = \int_{-R}^{R} \left(\prod_{j=0}^{n-1} 2|m_0(2^j \xi)|^2 \right) |\hat{\varphi}(\xi)|^2 \, d\xi.$$

Let K be the compact set of Cohen's criterion. Then $|\hat{\varphi}|^2$ attains a positive minimum $m > 0$ on K. If we choose R equal to a multiple p of π and large enough to have $K \subset [-R, R]$ it follows, with M equal to the maximum of $|\hat{\varphi}|^2$ on $[-R, R]$, that $m\sigma_n \leq I_n \leq pM\sigma_n$. □

Remark 9.4. Note that the convergence of the series in Lemma 9.3 implies $\varphi \in H^{-t}$ even when m_0 does not satisfy Cohen's criterion. (Choose $p = 1$ in the last part of the proof.) However, the condition cannot be removed from the lemma.

Consider the case $N = 3$ with $(c_0, c_1, c_2) = (1, -1, 1)$. Here it turns out that $\rho(A) = 4$, so by Lemma 9.2 we have

$$\sum_{n=1}^{+\infty} \sigma_n 4^{-n} = +\infty.$$

On the other hand, by multiplying $m_0(\xi)$ with $(1 + e^{-i\xi})/2$, we arrive at the first case of Example 5.4. Hence, the result of a convolution of φ with $\mathbf{1}_{[0,1]}$ is an L^2- function. By Lemma 8.2, this is equivalent to $\varphi \in H^{-1}$.

We are now ready to prove the main result of this section.

THEOREM 9.5. *Assume* $m_0(\xi) = ((1+e^{-i\xi})/2)^M \widetilde{m}_0(\xi)$ *where* $\widetilde{m}_0$ *satisfies Cohen's criterion and* $\widetilde{m}_0(\pi) \neq 0$. *If* $\widetilde{A}$ *is defined from* $\widetilde{m}_0$ *and* $s_0 = M - \log_4 \rho(\widetilde{A})$, *then*

$$\varphi \in H^s(\mathbb{R}) \Leftrightarrow s < s_0.$$

Proof. We have $\varphi = \mathbf{1}_{[0,1]}^{*M} * \widetilde{\varphi}$, as in (8.2), and repeated use of Lemma 8.2 gives

$$\varphi \in H^s \Leftrightarrow \widetilde{\varphi} \in H^{s-M}.$$

Recall the spectral radius formula [R, p. 235]

$$\rho(\widetilde{A}) = \lim_{n\to+\infty} \|\widetilde{A}^n\|^{1/n} = \inf_{n\in\mathbb{N}} \|\widetilde{A}^n\|^{1/n}.$$

Since $\widetilde{A}P(0) \geq P(0)$ for all nonnegative $P \in S_{N-M}$, we have $\|\widetilde{A}^n\| \geq 1$ for all n. As a first application of the spectral radius formula, we conclude that

$$\rho(\widetilde{A}) \geq 1.$$

Assume $s < s_0$. With $t = M - s$ we then have $t > \log_4 \rho(\widetilde{A}) \geq 0$, and we have to show that $\widetilde{\varphi} \in H^{-t}$. By Lemmas 9.2 and 9.3 it suffices to show that the power series $\sum_{n=1}^{+\infty} \|\widetilde{A}^n\|(4^{-t})^n$ converges. But this follows from

$$4^{-t} \limsup_{n\to+\infty} \|\widetilde{A}^n\|^{1/n} = 4^{-t}\rho(\widetilde{A}) < 1.$$

Next, we have to show that $\varphi \notin H^{s_0}$. Equivalently, that $\widetilde{\varphi} \notin H^{-t_0}$ where $t_0 = M - s_0 = \log_4 \rho(\widetilde{A})$. Since $\widetilde{\varphi} \in L^2$ would contradict the assumption $\widetilde{m}_0(\pi) \neq 0$, (by Theorem 5.3(2) and Proposition 8.3), we only have to consider the case $t_0 > 0$: from Lemma 9.2 we see that

$$\sum_{n=1}^{+\infty} \sigma_n 4^{-t_0 n} \geq r \sum_{n=1}^{+\infty} \rho(\widetilde{A})^n \rho(\widetilde{A})^{-n} = +\infty.$$

The desired conclusion then follows from Lemma 9.3. □

Example 9.6. Consider the case

$$(c_0, c_1, c_2) = \left(\frac{1+z}{4}, \frac{1}{2}, \frac{1-z}{4}\right), \qquad z \in \mathbb{C}.$$

From Example 4.4 we know that $\varphi \in L^2$ if and only if $|z| < \sqrt{3}$. Using the notation of Theorem 9.5 and excluding the case $z = 0$, we have $M = 1$ and

$$\widetilde{m}_0(\xi) = \frac{1+z}{2} + \frac{1-z}{2}e^{-i\xi}.$$

Since $\widetilde{m}_0$ can at most have one zero in $[-\pi, \pi[$, it is not hard to see that Cohen's criterion is satisfied for all z. The operator $\widetilde{A}$ is given by the 1×1-matrix $(1 + |z|^2)$; hence Theorem 9.5 gives $\varphi \in H^s \Leftrightarrow s < s_0$, where

$$s_0 = 1 - \log_4(1 + |z|^2), \qquad z \neq 0.$$

For $z = 0$ we are in the case $M = 2$ of (8.4), so $s_0 = \frac{3}{2}$. That is, the regularity exhibits a jump from $1 - \epsilon$ in the neighborhood of $z = 0$ for $z \neq 0$ to $\frac{3}{2} - \epsilon$ when $z = 0$.

The result is in agreement with Example 4.4, since $s_0 > 0$ if and only if $|z| < \sqrt{3}$. What is not captured by Theorem 9.5 is the degenerate case $c = (\frac{1}{2}, 0, \frac{1}{2})$ where we know that $\varphi \in L^2$ but φ does not have the Riesz basis property (Example 5.4). Here, Theorem 9.5 would only give the result $s < 0 \Rightarrow \varphi \in H^s$. □

The following estimations of $\rho(A)$ will be useful in order to compare the performance of Theorem 9.5 with other regularity estimates.

PROPOSITION 9.7. *Define for every $n \in \mathbb{N}$*

$$K_n = \left(\sup_{\xi \in [-\pi,\pi]} \prod_{j=0}^{n-1} |m_0(2^j \xi)|^2 \right)^{1/n}.$$

Then we have $\limsup_{n\to+\infty} K_n \le \rho(A) \le 2 \inf_{n\in\mathbb{N}} K_n$.

Proof. By induction we find

$$A^n P(\xi) = \sum_{k=0}^{2^n-1} \left(\prod_{j=1}^{n} |m_0(2^{-j}(\xi + 2\pi k))|^2 \right) P(2^{-n}(\xi + 2\pi k)).$$

Using the triangle inequality and testing $P = 1$ leads then to the result

$$\|A^n\| = \sup_\xi \left(\sum_{k=0}^{2^n-1} \prod_{j=1}^{n} |m_0(2^{-j}(\xi + 2\pi k))|^2 \right).$$

Simple estimates of the right-hand side give

$$K_n \le \|A^n\|^{1/n} \le 2K_n,$$

and the desired result follows from the spectral radius formula. □

Remark 9.8. Inserting $\xi = 2\pi/3$ into the definition of K_n gives the lower bound $K_{2n} \ge |m_0(2\pi/3)m_0(-2\pi/3)|$. ($\xi_0 = 2\pi/3$ is periodic of period 2 for $\xi \mapsto 2\xi$ modulo 2π.) Therefore we get

$$\rho(A) \ge \limsup_{n\to+\infty} K_n \ge \left| m_0\left(\frac{2\pi}{3}\right) m_0\left(-\frac{2\pi}{3}\right) \right|,$$

as a simple consequence of Proposition 9.7

Relation to Hölder regularity estimates. In the appendix of [D1] a method is developed to estimate the regularity of φ from the spectral radius of $\widetilde{A}$. Using the terminology of Theorem 9.5, the result is that

$$\alpha < s_0 - \tfrac{1}{2} \Rightarrow \varphi \in C^\alpha(\mathbb{R}). \tag{9.2}$$

Here we define C^α in the following way: if $\alpha = k + \beta$ with $k \in \mathbb{N}_0$ and $0 \le \beta < 1$, then $f \in C^\alpha$ if f is of class C^k and $f^{(k)}$ is uniformly Hölder continuous with exponent β.

The result (9.2) is also a corollary of Theorem 9.5 and the well-known inclusion

$$H^s(\mathbb{R}) \subset C^\alpha(\mathbb{R}) \quad \text{for } s > \alpha + \tfrac{1}{2}.$$

If C_0^α denotes the space of compactly supported C^α functions, we have an inclusion in the opposite direction [Hö, p. 242],

$$C_0^\alpha(\mathbb{R}) \subset H^s(\mathbb{R}) \quad \text{for } s < \alpha.$$

Therefore, the following Hölder estimates of accuracy $\frac{1}{2}$ hold.

COROLLARY 9.9. *With notation and assumptions as in Theorem* 9.5 *then* $\varphi \in C^\alpha(\mathbb{R})$ *if* $s_0 > \alpha + \frac{1}{2}$ *and* $\varphi \notin C^\alpha(\mathbb{R})$ *if* $s_0 < \alpha$.

Let $\widetilde{K}_n$ be defined from $\widetilde{m}_0$ as in Proposition 9.7. Since $s_0 = M - \log_4 \rho(\widetilde{A})$, we get, for every $n \in \mathbb{N}$,

$$\alpha < M - \log_4 \widetilde{K}_n \Rightarrow \varphi \in C^\alpha(\mathbb{R}). \tag{9.3}$$

This is also obtained in [D1], but as a consequence of the stronger result about the decay of the Fourier transform of φ,

$$|\hat{\varphi}(\xi)| \leq C(1+|\xi|)^{-M+\log_4 \widetilde{K}_n}.$$

Cohen has refined this regularity estimation method starting from the observation that $\widetilde{K}_n$ converges to $\widetilde{K} = \inf_n \widetilde{K}_n$ as $n \to +\infty$. (A consequence of $\widetilde{K}_{n+m}^{n+m} \leq \widetilde{K}_n^n \widetilde{K}_m^m$.) Then (9.3) still holds with $\widetilde{K}_n$ replaced by $\widetilde{K}$ and it is shown in [C], under the technical assumption $|\widetilde{m}_0(\pi)| \geq 1$, that

$$\alpha > M + \tfrac{1}{2} - \log_4 \widetilde{K} \Rightarrow \varphi \notin C^\alpha(\mathbb{R}).$$

In other words, if $\alpha_0 = \sup\{\alpha \mid \varphi \in C^\alpha\}$ and we put $b = \log_4 \widetilde{K}$, then α_0 lies in the interval $[M - b - 1, M - b + \frac{1}{2}]$.

Here we can immediately remove the technical assumption and improve by $\frac{1}{2}$ the upper bound by observing that Proposition 9.7 implies $M - b - \frac{1}{2} \leq s_0 \leq M - b$. Thus we get from Corollary 9.9

$$\alpha_0 \in [M - b - 1, M - b].$$

Finally, we should note that optimal estimates of Hölder regularity have been obtained by Daubechies and Lagarias in [DL2] without making use of the Fourier transform. However, their criteria get very complex as N increases. For purposes where an accuracy of $\frac{1}{2}$ in the estimation of α_0 is sufficient, Corollary 9.9 offers a substantially easier alternative.

10. Application to orthonormal wavelets.

Throughout this section we will assume that

$$|m_0(\xi)|^2 + |m_0(\xi+\pi)|^2 = 1 \tag{10.1}$$

and that m_0 satisfies Cohen's criterion (Definition 5.2). Then $P = 1$ will be the unique solution to $AP = P$ with $P(0) = 1$, so φ is square integrable and the integer translates of φ constitute an orthonormal basis for the closure of their linear span. In this case φ is the scaling function of a multiresolution analysis of $L^2(\mathbb{R})$ [M], [C], and we can associate a compactly supported wavelet ψ to this scaling function as follows:

$$\frac{1}{2}\psi\left(\frac{x}{2}\right) = \sum_k d_k \varphi(x-k), \tag{10.2}$$

$$\text{where} \quad d_k = (-1)^{1-k}\overline{c_{1-k}}. \tag{10.3}$$

If we define $m_1(\xi) = \sum_k d_k e^{-ik\xi}$, then (10.2)–(10.3) can also be written in the form

$$\hat{\psi}(2\xi) = m_1(\xi)\hat{\varphi}(\xi), \tag{10.4}$$

$$\text{with } m_1(\xi) = e^{-i\xi}\,\overline{m_0(\xi+\pi)}. \tag{10.5}$$

Modulo integer translation and multiplication by complex constants of modulus one, ψ is the unique compactly supported L^2-function of unit norm in the span of $\{\varphi(2x-k)\}_{k\in\mathbb{Z}}$, which is orthogonal to all $\varphi(x-k)$. The set of functions

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \qquad j, k \in \mathbb{Z}$$

then defines an orthonormal basis for $L^2(\mathbb{R})$ consisting of compactly supported wavelets [M].

Moments in time. Let us try to apply the results of §6 to find the moment $\int x|\psi(x)|^2\,dx$, i.e., the center of gravity of the squared modulus of the wavelet.

First we define, in analogy with (6.2)–(6.3),

$$b_n(k) = 2\sum_l \left(l - \frac{k}{2}\right)^n d_l \overline{d_{l-k}},$$
$$B_n q(k) = \sum_l b_n(2k-l)q(l).$$

Then by mimicking the proof of Proposition 6.1 we get

$$\int_{-\infty}^{+\infty} x|\psi(x)|^2\,dx = \left(\frac{1}{2}B_0 q_1 + \frac{1}{2}B_1 q_0\right)(0).$$

From (10.1) and the definition of (d_k) it follows that

$$a_0(k) + b_0(k) = 2\delta_{0k},$$
$$a_1(0) + b_1(0) = 1.$$

(Here δ is the Kronecker delta.) Using all this together with Proposition 6.1 for $n = 1$, we obtain

$$\begin{aligned}\int_{-\infty}^{+\infty} x|\psi(x)|^2\,dx &= \left(q_1 - \frac{1}{2}A_0 q_1 + \frac{1}{2}B_1 q_0\right)(0)\\ &= \left(\frac{1}{2}A_1 q_0 + \frac{1}{2}B_1 q_0\right)(0) = \frac{1}{2}(a_1(0) + b_1(0)) \qquad (Q_0 = P_0 = 1)\\ &= \frac{1}{2}.\end{aligned}$$

Hence, we have shown the following amusing result.

PROPOSITION 10.1. *If the orthonormal wavelet ψ is defined from* (10.2) *then*

$$\int_{-\infty}^{+\infty} x|\psi(x)|^2\,dx = \frac{1}{2}.$$

Remark 10.2. This proposition is easily generalized to $\int x|\psi(x)|^2\,dx \in \frac{1}{2} + \mathbb{Z}$ for all orthonormal wavelets which are obtained from a multiresolution analysis and have a reasonable decay at infinity. □

Turning to the second moment in time of ψ, things become more complicated. However, using

$$\int_{-\infty}^{+\infty} x^2|\psi(x)|^2\,dx = \left(\frac{1}{4}B_0 q_2 + \frac{1}{2}B_1 q_1 + \frac{1}{4}B_2 q_0\right)(0)$$

together with simple relations between b_n and a_n for $n = 1, 2$ and Proposition 6.1, we arrive at the formula

$$\begin{aligned}(\Delta x)^2 &= \int_{-\infty}^{+\infty} \left(x - \frac{1}{2}\right)^2 |\psi(x)|^2\, dx \\ &= \frac{1}{2} q_2(0) + \frac{1}{2} a_2(0) + \sum_k a_1(2k+1) q_1(-2k-1).\end{aligned} \tag{10.6}$$

Here $q_2(0)$ and $q_1(2k+1)$ can be found from Proposition 6.1.

The above method was applied in [DV] to choose wavelets with the minimum root mean square duration in time Δx from classes of wavelets with equal $|\hat{\psi}|$.

Moments in frequency and regularity. We have a simple relation between the moments in frequency of φ and ψ. In particular, the regularity of ψ is exactly the same as for φ.

PROPOSITION 10.3. *Define φ and ψ as above. Then*

(1) $\forall s \geq 0$: $\psi \in H^s(\mathbb{R}) \Leftrightarrow \varphi \in H^s(\mathbb{R})$

(2) *Let $n \in \mathbb{N}_0$. If $\psi \in H^{\frac{n}{2}}(\mathbb{R})$, then*

$$\int_{-\infty}^{+\infty} \xi^n |\hat{\psi}(\xi)|^2\, d\xi = (2^{1+n} - 1) \int_{-\infty}^{+\infty} \xi^n |\hat{\varphi}(\xi)|^2\, d\xi. \tag{10.7}$$

Proof. The key ingredient in the proof is the relation

$$|\hat{\psi}(\xi)|^2 = \left|\hat{\varphi}\left(\frac{\xi}{2}\right)\right|^2 - |\hat{\varphi}(\xi)|^2, \tag{10.8}$$

which follows from (10.1) and (10.4)–(10.5). Iterating this result, we obtain for every $k \in \mathbb{N}$:

$$\sum_{j=1}^{k} |\hat{\psi}(2^j \xi)|^2 = |\hat{\varphi}(\xi)|^2 - |\hat{\varphi}(2^k \xi)|^2.$$

The integral of the last term tends to zero as $k \to \infty$, and since the terms of the sum are nonnegative we have pointwise convergence for almost every ξ in

$$\sum_{j=1}^{+\infty} |\hat{\psi}(2^j \xi)|^2 = |\hat{\varphi}(\xi)|^2.$$

By the monotone convergence theorem it follows that

$$\begin{aligned}\int_{-\infty}^{+\infty} |\xi|^{2s} |\hat{\varphi}(\xi)|^2\, d\xi &= \sum_{j=1}^{+\infty} \int_{-\infty}^{+\infty} |\xi|^{2s} |\hat{\psi}(2^j \xi)|^2\, d\xi \\ &= \sum_{j=1}^{+\infty} 2^{-j(2s+1)} \int_{-\infty}^{+\infty} |\xi|^{2s} |\hat{\psi}(\xi)|^2\, d\xi \\ &= (2^{2s+1} - 1)^{-1} \int_{-\infty}^{+\infty} |\xi|^{2s} |\hat{\psi}(\xi)|^2\, d\xi.\end{aligned}$$

This clearly proves (1), and the result (2) is now an easy consequence of (10.8), since (1) guarantees that both integrals in (10.7) are well defined. □

Remark 10.4. The statement (1) about the regularity of ψ could also have been obtained simply by observing from (10.2) that ψ is a finite linear combination of half integer translates of $\varphi(2x)$. Then a slight generalization of the telescoping argument of the proof of Lemma 8.2 would do. Moreover, this approach would also work when (10.1) does not hold, i.e., when we are dealing with compactly supported biorthogonal wavelets as in [VH], [C].

Example 10.5. If m_0 is a solution to (10.1) with M sum rules, i.e.,

$$m_0(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^M \widetilde{m}_0(\xi),$$

then $|\widetilde{m}_0|^2$ must be of the form [D1]

$$|\widetilde{m}_0(\xi)|^2 = \sum_{k=0}^{M-1} \binom{M-1+k}{k} \sin^{2k}\left(\frac{\xi}{2}\right) + \sin^{2M}\left(\frac{\xi}{2}\right) R(\xi), \tag{10.9}$$

where R is a real valued trigonometric polynomial satisfying $R(\xi) + R(\xi+\pi) = 0$:

$$R(\xi) = \sum_{k=1}^{K} \{\alpha_k \cos(2k-1)\xi + \beta_k \sin(2k-1)\xi\}.$$

Conversely, if R is chosen such that the right-hand side of (10.9) is nonnegative, we can find a trigonometric polynomial $\widetilde{m}_0$ solving (10.9), thanks to a lemma of Riesz. (This factorization method is well known in signal analysis [P, p. 231].) The resulting number of coefficients N is equal to $2(M+K)$ if either $\alpha_K \neq 0$ or $\beta_K \neq 0$, and equal to $2M$ if $R = 0$.

Assume R is chosen such that $\widetilde{m}_0$ verifies Cohen's criterion. From Proposition 8.3 we know that $\varphi \in H^m$ implies $M \geq m+1$ and therefore $N \geq 2M \geq 2m+2$.

As a first example, consider the case $M = 2$ and $R = 0$ of (10.9):

$$|\widetilde{m}_0(\xi)|^2 = 1 + 2\sin^2\left(\frac{\xi}{2}\right) = 2 - \cos\xi.$$

Then we have simply $\widetilde{A} = (4)$, so the Sobolev regularity given by Theorem 9.5 is $s_0 = 2 - \log_4 4 = 1$, meaning that

$$\varphi \in H^s \Leftrightarrow s < 1.$$

Clearly, we cannot obtain a better Sobolev regularity for any scaling function φ defined from $N = 4$ coefficients, because if we had $\varphi \in H^1$, the two necessary sum rules would fix $R = 0$. This stands in contrast to the behaviour of the Hölder regularity examined in [D2], where it is shown that the best possible Hölder coefficient is obtained for a choice with $M = 1$ and $R \neq 0$.

For $N = 6$ and $R = 0$ the Sobolev regularity is $s_0 = 3 - \log_4 9 \approx 1.4150$, and numerical experiments suggest that this cannot be improved by allowing $R \neq 0$. Cohen has shown in [C] (using the method described here in Remark 9.8 together with an estimation of $\widetilde{K}_4$) that the regularity of the family (10.9) with $R = 0$ is roughly $N/10$

for large N. Most of the $N/2$ sum rules could therefore very well be unnecessary, and the problem of finding the solution with the highest regularity for a given number of coefficients N is still open. However, if we accept the measure of regularity to be in the Sobolev sense, Theorem 9.5 provides a simpler cost function for this optimization problem than the methods of [DL2].

As mentioned in the previous section, Daubechies developed a spectral radius based regularity estimation method. She applied this to the family (10.9) with $R = 0$, and listed the resulting Hölder regularity estimates for $M = 2, 3, \ldots, 10$ in a table [D1, p. 984]. In fact, adding $\frac{1}{2}$ to the values of this table, we arrive at the corresponding exact Sobolev regularity estimates s_0 of Theorem 9.5. In particular, Corollary 9.9 asserts that the optimal Hölder estimates lies in closed intervals of length $\frac{1}{2}$ with left end points equal to the values of this table.

Acknowledgments. Part of this work was done during a stay at the Department of Mathematics of the Royal Institute of Technology in Stockholm. The author wishes to express his thanks for their warm hospitality. A grant from the Nordic Council of Ministers made the stay possible and is also gratefully acknowledged.

After the completion of this work, the author has learned that part of the results presented here have also been derived by T. Eirola [E].

REFERENCES

[B] G. Beylkin, *On the representation of operators in bases of compactly supported wavelets*, Schlumberger–Doll Research, Ridgefield, CT, preprint.

[BA] P. J. Burt and E. H. Adelson, *The Laplacian pyramid as a compact image code*, IEEE Trans. Comm., 31 (1983), pp. 532–540.

[BCR] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[C] A. Cohen, *Ondelettes, analyses multirésolutions et traitement numérique du signal*, thesis, Universitê Paris IX Dauphine, Dauphine, France, 1990.

[CDM] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc., 93 (1991), Number 453.

[CMQW] R. Coifman, Y. Meyer, S. Quale, and M.V. Wickerhauser, *Signal processing and compression with wavelet packets*, Yale University, New Haven, CT, 1990, preprint.

[CR] J.-P. Conze and A. Raugi, *Fonctions harmonique pour un opérateur de transition et applications*, Bull. Soc. Math. France, 273 (1990), pp. 273–310.

[D] S. Dubuc, *Interpolation through an iterative scheme*, J. Math. Anal. Appl., 114 (1986), pp. 185–204.

[D1] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[D2] ———, *Orthonormal bases of compactly supported wavelets* II. *Variations on a theme*, AT&T Bell Labs., Murray Hill, NJ, preprint.

[DD] G. Deslauriers and S. Dubuc, *Dyadic interpolation*, in Fractals: Non-Integral Dimensions and Applications, G. Cherbit, ed., John Wiley, New York, 1991.

[DL1] I. Daubechies and J. Lagarias, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[DL2] ———, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[DV] C. Dorize and L. F. Villemoes, *Optimizing time-frequency resolution of orthonormal wavelets*, Proc. International Conference on Acoustics, Speech, and Signal Processing, (1991), pp. 2029–2032.

[E] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[H] A. Haar, *Zur Theorie der orthogonalen Funktionensysteme*, Math. Anal., 69 (1910), pp. 331–371.

[Hö] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I., Springer-Verlag, New York, 1983.

[L1] W. M. Lawton, *Tight frames of compactly supported affine wavelets*, J. Math. Phys., 31 (1990), pp. 1898–1901.

[L2] ———, *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys., 32 (1991), pp. 57–61.

[M] Y. Meyer, *Ondelettes et opérateurs* I: *ondelettes*, Hermann, Paris, 1990.

[Ma] S. Mallat, *Review of multifrequency channel decompositions of images and wavelet models*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 2091–2110.

[P] A. Papoulis, *Signal analysis*, McGraw-Hill, New York, 1977.

[R] W. Rudin, *Functional analysis*, McGraw-Hill, New York, 1973.

[Ri] O. Rioul, *A discrete-time multiresolution theory*, IEEE Trans. Signal Process., to appear.

[S] G. Strang, *Wavelets and dilation equations: A brief introduction*, SIAM Rev., 31 (1989), pp. 614–627.

[VH] M. Vetterli and C. Herley, *Wavelets and filter banks: Theory and design*, IEEE Trans. Signal Process., September 1992, to appear.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1544–1576, November 1992

008

SIMPLE REGULARITY CRITERIA FOR SUBDIVISION SCHEMES*

OLIVIER RIOUL†

Abstract. Convergent subdivision schemes arise in several fields of applied mathematics (computer-aided geometric design, fractals, compactly supported wavelets) and signal processing (multiresolution decomposition, filter banks). In this paper, a polynomial description is used to study the existence and Hölder regularity of limit functions of binary subdivision schemes. Sharp regularity estimates are derived; they are optimal in most cases. They can easily be implemented on a computer, and simulations show that the exact regularity order is accurately determined after a few iterations. Connection is made to regularity estimates of solutions to two-scale difference equations as derived by Daubechies and Lagarias, and other known Fourier-based estimates. The former are often optimal, while the latter are optimal only for a subclass of symmetric limit functions.

Key words. subdivision algorithms, Hölder regularity, Sobolev regularity, two-scale difference equations, wavelets

AMS(MOS) subject classifications. 26A15, 26A16, 39B05, 42C15, 46E35, 94A12

1. Introduction.

This paper focuses on the behavior of real-valued discrete sequences u_n $(n \in \mathbf{Z})$ of finite length under repeated action of an operator $\mathcal{G}$ defined as

$$u_n \xrightarrow{\mathcal{G}} v_n = \sum_{k\in\mathbf{Z}} u_k\, g_{n-2k}. \tag{1.1}$$

The fixed sequence g_n that parameterizes $\mathcal{G}$ is called the subdivision *mask* [14], [15]. It plays a central role in the following. Starting from the initial "impulse" sequence

$$\delta_n = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{otherwise,} \end{cases}$$

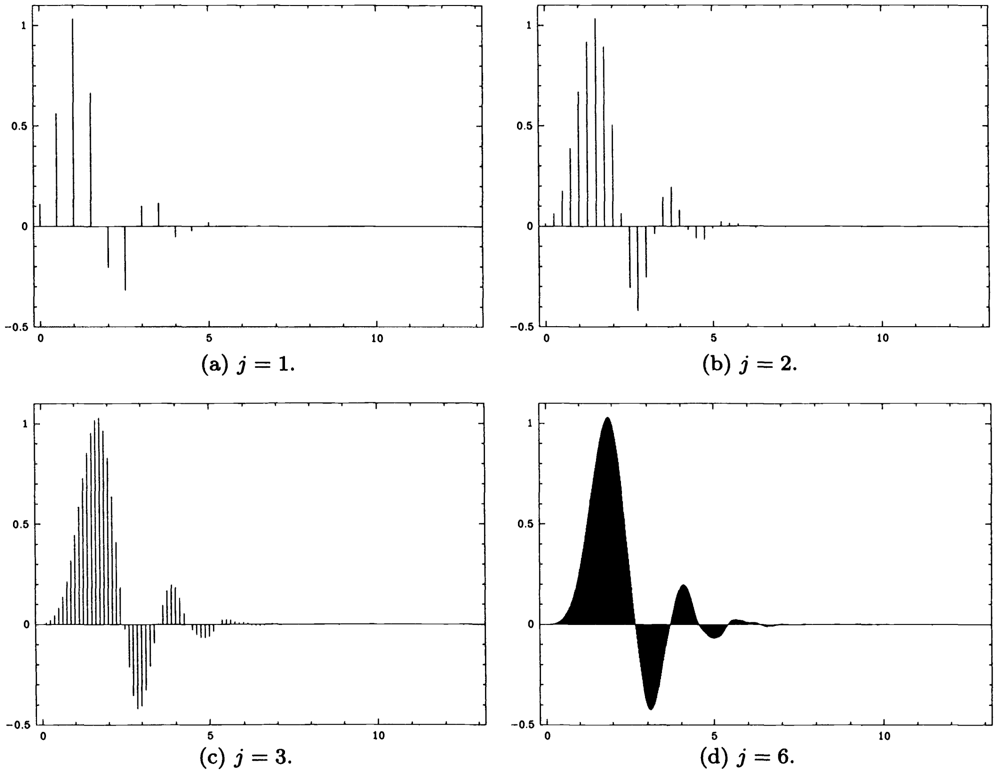
a *binary subdivision scheme* [14], [15], [17] (in one dimension) is an infinite collection of sequences g_n^j $(j \in \mathbf{N})$, defined by iteration as shown.

$$\begin{aligned} g_n^1 &= \mathcal{G}\{\delta_n\} = g_n, \\ g_n^2 &= \mathcal{G}\{g_n^1\}, \\ &\ \vdots \\ g_n^{j+1} &= \mathcal{G}\{g_n^j\}. \qquad (1.2) \\ &\ \vdots \end{aligned}$$

The g_n^j's are fully determined given the mask g_n. Of course other initial sequences can be considered. In addition, this scheme is said to be *interpolatory* [10]–[15] if it satisfies the extra condition $g_{2n} = \delta_n$, which means that all points g_n^j at some level j are carried unchanged to the next level: $g_{2n}^{j+1} = g_n^j$. In this paper we regard interpolatory subvision schemes as a special case to which general results will apply. However, we restrict ourselves to *binary* subdivision schemes, even though the results

* Received by the editors February 25, 1991; accepted for publication (in revised form) January 6, 1992.

† Centre National d'Études des Télécommunications, Centre Paris B, Dept. ETP, 38–40 rue du Général Leclerc, 92131 Issy-Les-Moulineaux, France.



(a) $j = 1$.

(b) $j = 2$.

(c) $j = 3$.

(d) $j = 6$.

FIG. 1. *A binary subdivision scheme converging to a limit function (after* [6]*). The discrete sequences* g_n^j *are plotted as "pulses" against* $n2^{-j}$ *for* $j = 1, 2, 3,$ *and* 6*. At each iteration step the up-scaling operator* (1.1) *is applied, which approximatively doubles the number of coefficients while preserving a global shape. When* $j \to \infty$*, these discrete curves converge to a "nice-looking," regular limit function, compactly supported on* $[0, 13]$.

of this paper easily extend to more general subdivision schemes, for which the number 2 in (1.1) is replaced by any integer $p \geq 2$ [8], [9].

Subdivision schemes arise in several fields of applied mathematics and signal processing. They have been used for curve fitting and to generate fractal or smooth curves and surfaces numerically [10]–[15], [17]. They also play an important role in wavelet theory [1], [3]–[7], [20]–[23], a newly born theory in functional analysis closely related to filter bank theory in signal processing [20], [18], [21], [22]. In all of these applications, the convergence of (1.2) to a function of a continuous variable $\varphi(x)$ as j indefinitely increases is important. It is also important to control several properties of the limit function $\varphi(x)$ from the choice of the mask g_n. For example, whether limit functions $\varphi(x)$ are regular (smooth) or not may be relevant for image coding applications using wavelets [1], [20], and this has motivated the work presented here.

The aim of this paper is to find necessary and sufficient conditions on the mask g_n for the existence and Hölder regularity of the limit function $\varphi(x)$. Figure 1 shows that $\varphi(x)$ can be thought of as a limit of discrete curves g_n^j plotted against $n2^{-j}$. (We then say that the sequences g_n^j "converge" to $\varphi(x)$ as $j \to \infty$.) In addition, we shall often be in the case of *uniform* convergence. Intuitively, this means that the discrete curves g_n^j converge "as a whole" to the limit curve $\varphi(x)$. Section 3 discusses several possible definitions for both types of convergence.

This paper is organized as follows. First, §2 describes binary subdivision schemes

(1.2) using the convenient polynomial notation. Then, various definitions of convergence are discussed (§3), and a basic necessary condition for the existence of a limit function is derived (§4). We show how the values of a limit function can be computed exactly on a computer (§5). The relation between the values of g_n^j and those of $\varphi(x)$ leads us to define "stable" subdivision schemes, to which the results of this paper fully apply (§6). Fortunately, almost all limit functions are stable.

To tackle the regularity problem, we characterize regularity of limit functions in terms of discrete sequences. Continuity is connected to uniform convergence and a necessary and sufficient condition for uniform convergence is derived in §7. Hölder regularity $\dot{C}^\alpha$ $(0 < \alpha \leq 1)$ is expressed by a similar property of the g_n^j's (§8). Finite differences of the g_n^j's play the role of derivatives and N-times continuously differentiable limit functions are, therefore, characterized by uniform convergence of finite differences (§9).

From these equivalences a full characterization of Hölder regularity $\dot{C}^r$ (for all $r > 0$) naturally emerges in terms of discrete sequences (§10). The main result of this paper is an easily implemented, optimal regularity estimate derived in §11. This estimate is then compared to other related work [3]–[12], [23]. A sharp upper bound for Hölder regularity is also derived in §13.

As a general rule, the first parts of the theorems derived in this paper show that a given property of the g_n^j's implies the corresponding regularity property of the limit function $\varphi(x)$. The second parts prove the converse implication, which is useful for proving optimality of regularity estimates and generally assumes the stability condition.

The purpose of this paper is close to the one of Daubechies and Lagarias in [8], [9]. They studied the existence, uniqueness, and regularity of solutions to "two-scale difference equations." We shall see in §5 that the limit function $\varphi(x)$ associated to mask g_n indeed satisfies the following two-scale difference equation.

$$\varphi(x) = \sum_k g_k \, \varphi(2x - k).$$

Although it can be shown [9] that a solution to this equation is not necessarily the limit function of the subdivision scheme with mask g_n, both approaches are closely related. In fact, the study of regularity of solutions to two-scale difference equations can be reduced, after suitable transformation [2], to that of limit functions $\varphi(x)$ of a binary subdivision scheme. However, the contents, formulation, and proofs of this paper differ notably from [8], [9]; Daubechies and Lagarias derive conditions for the existence of $\mathbf{L}^1$-solutions to two-scale difference equations and estimate global and local regularity of solutions that are, in fact, limit functions. This paper concentrates on the determination of *optimal* estimates for global regularity of limit functions, with interpretation in terms of discrete sequences and comparison with Fourier-based techniques. (Local regularity may also be investigated using the methods of this paper [19].)

It was pointed out to the author by one of the referees that the framework of this paper is very close to that of Dyn and Levin [14], [15]. I learned that several results were derived independently in [14], [15] for the study of C^N limit functions (see §§7 and 9).

2. Polynomial notation. Subdivision schemes have been mostly described using matrices or Fourier transforms [6], [8]–[11]. Throughout this paper we often use the polynomial description

$$U(X) = \sum_{n=0}^{L-1} u_n X^n$$

of any causal sequence u_n of length L ($u_n = 0$ for $n < 0$ and $n \geq L$). Since sequences of finite length can always be made causal by shifting, we assume all sequences causal in the following. This notation was adopted in [14], [15], which uses Laurent polynomials for noncausal sequences.

In polynomial notation, the up-scaling operator (1.1) reads

$$U(X) \xrightarrow{\mathcal{G}} V(X) = G(X)\, U(X^2), \tag{2.1}$$

which shows that it can be seen as resulting from two operations.

1. Change X to X^2 in $U(X)$, i.e., insert zeros between every two samples of u_n.
2. Multiply by $G(X)$, i.e., convolve the result with the mask g_n.

In other words, the operator (1.1), (2.1) "smooths" u_n at twice its rate, and (2.1) can be seen as a discrete version of a dilation by two: $f(x) \to f(x/2)$.

Iterating (2.1) gives the polynomial $G^j(X)$, associated to the sequence g_n^j (1.2).

$$G^j(X) = G(X)\, G(X^2)\, G(X^4) \cdots G(X^{2^{j-1}}). \tag{2.2}$$

This equation fully describes binary subdivision schemes in terms of polynomials (see §4 when the initial sequence is not δ_n). It can be rewritten in recursive form in two ways.

$$G^{j+1}(X) = G(X)\, G^j(X^2), \quad \text{i.e.,}\ g_n^{j+1} = \sum_k g_k^j\, g_{n-2k}, \tag{2.3}$$

$$G^{j+1}(X) = G^j(X)\, G(X^{2^j}), \quad \text{i.e.,}\ g_n^{j+1} = \sum_k g_k\, g_{n-2^j k}^j. \tag{2.4}$$

Both are useful in the sequel. Equation (2.3) is simply a rewriting of definition (1.2), while (2.4) links binary subdivision schemes to two-scale difference equations (see §5). We shall also consider (2.2) for polynomials other than $G(X)$. Given any polynomial $U(X)$, $U^j(X)$ (with a superscript index j) is

$$U^j(X) = U(X)\, U(X^2)\, U(X^4) \cdots U(X^{2^{j-1}}). \tag{2.5}$$

In this paper we use l^1 and l^∞-norms of discrete sequences in terms of polynomials,

$$\|U(X)\|_\infty = \max_k |u_k|,$$
$$\|U(X)\|_1 = \sum_k |u_k|,$$

and the following well-known inequality:

$$\|U(X)\, V(X)\|_\infty \leq \|V(X)\|_1 \|U(X)\|_\infty. \tag{2.6}$$

For polynomials with real coefficients, the following useful inequality holds whenever $V(X)$ has no roots *on* the unit circle.

$$\|U(X)\|_\infty \leq c_V\, \|U(X)\, V(X)\|_\infty, \tag{2.7}$$

where c_V is a constant depending only on $V(X)$.

Proof. This is trivially true for infinite sequences when the roots of $V(X)$ lie outside the unit circle; the constant c_V is then the converging l^1-norm of the Laurent series coefficients of $1/V(X)$, which is analytic in the complex-domain region $|X| \leq 1$. Here, since v_n is a sequence of finite length L, the index reversal $n \leftrightarrow L-1-n$ in v_n transforms roots of $V(X)$ inside the unit circle into roots outside the unit circle. Hence (2.7) holds when $V(X)$ has no roots *on* the unit circle. □

3. Definition of convergent subdivision schemes. Various definitions of convergent binary subdivision schemes have been proposed in the literature [6], [10]–[15], [17]. In this paper we restrict ourselves to pointwise or uniform convergence. It is easy to define a limit function in the case of interpolatory subdivision schemes as defined in the introduction: Since for such schemes one has $g_n^{j+1} = g_n^j$, the function $\varphi(x)$ can always be defined on dyadic rationals by

$$\varphi(n2^{-j}) = g_n^j. \tag{3.1}$$

For example, determining a continuous limit function amounts to finding a continuous extension of (3.1) to the real axis [11], [12].

The situation is more complex for general subdivision schemes since the values of g_n^j are not necessarily preserved as j increases. In order to "converge" to a limit function, the sequence g_n^j must be somehow interpolated. The idea is that the resulting sequence of functions of the continuous variable x, indexed by j, converges (pointwise or uniformly) to a limit function $\varphi(x)$ under some conditions on the mask g_n.

In [6], Daubechies chose to interpolate the sequence g_n^j by stepwise constant functions: she defined $\varphi(x)$ as the limit of $\varphi^j(x) = g^j_{\lfloor 2^j x + 1/2 \rfloor}$ as $j \to \infty$. Other kinds of interpolation are possible and yield similar results. Among possible choices are $g^j_{\lfloor 2^j x \rfloor}$, $g^j_{\lceil 2^j x \rceil}$, and the continuous linear interpolation function $\varphi^j_{\mathcal{L}}(x)$ obtained by connecting the points g_n^j by segments as in Figs. 2 and 3. All such interpolation functions $\varphi^j(x)$ agree at the "knots" $n2^{-j}$, i.e., $\varphi^j(n2^{-j}) = g_n^j$. In this paper we use a stronger definition that gives some flexibility on the way the subdivision scheme is interpolated.

DEFINITION 3.1. A binary subdivision scheme g_n^j (1.2) converges (pointwise) to a limit function $\varphi(x)$ if, for any sequence of integers n_j satisfying

$$|n_j 2^{-j} - x| \leq c\,2^{-j} \tag{3.2}$$

(where c is a constant independent of j), we have

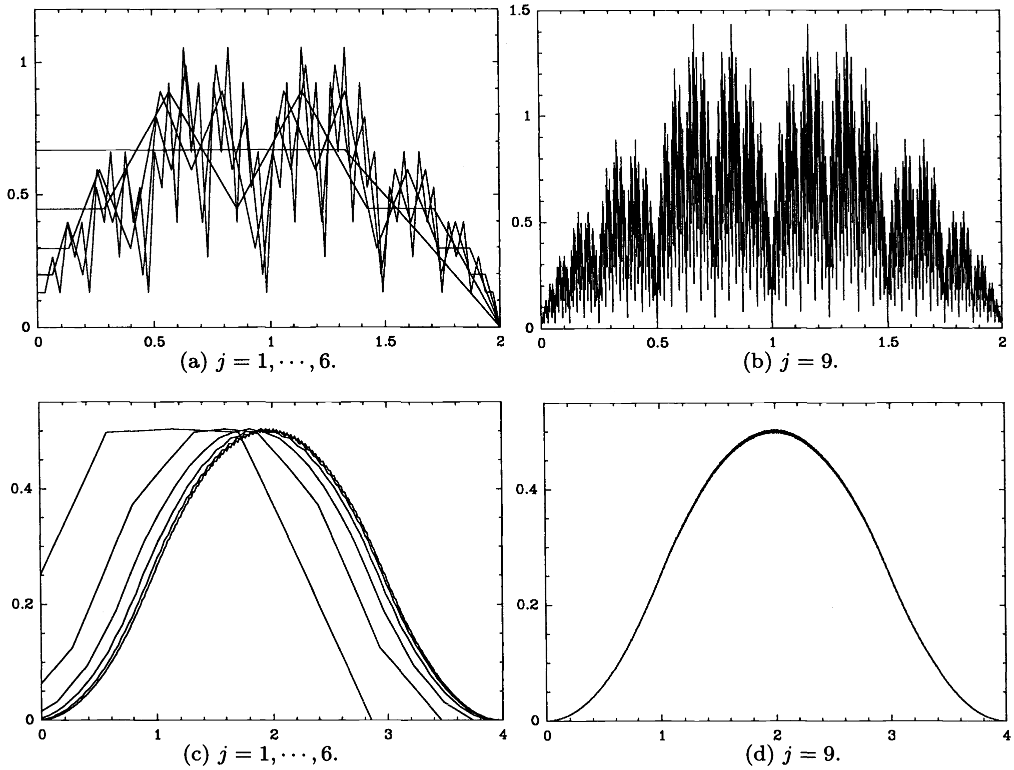
$$\varphi(x) = \lim_{j \to \infty} g^j_{n_j}. \tag{3.3}$$

The convergence is, moreover, *uniform* if

$$\sup_x |\varphi(x) - g^j_{n_j}| \to 0 \quad \text{as } j \to \infty. \tag{3.4}$$

Note that the sequence n_j depends on x, hence $g^j_{n_j}$ can be regarded as a function of x. The flexibility comes from the arbitrary choice of n_j satisfying (3.2)[1]. In particular,

[1] It seems natural to impose the more general condition $n_j 2^{-j} \to x$ as $j \to \infty$ in place of (3.2). But then $n_j - 2^j x$ is allowed to increase indefinitely as $j \to \infty$, and the resulting definition becomes too strong for deriving some of the results of this paper.



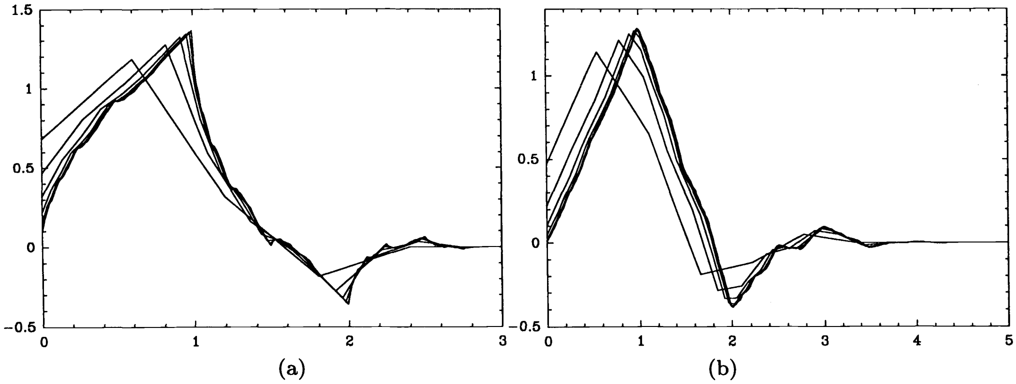
(a) $j = 1, \cdots, 6$.

(b) $j = 9$.

(c) $j = 1, \cdots, 6$.

(d) $j = 9$.

FIG. 2. *Two examples of diverging dyadic up-scaling schemes. Figures* (a), (c) *show six plots of the discrete sequences g^j_n ($j = 1, \cdots, 6$), represented with values joined by segments and plotted against $n2^{-j}$. Figures* (b), (d) *show the obtained curve after* 9 *iterations.* (a), (b) $g_0 = g_1 = g_2 = \frac{2}{3}$, $g_n = 0$ *elsewhere. Here* $G(-1) = \frac{2}{3} \neq 0$. *Note that up-scaling follows a fractal law.* (c), (d) $g_0 = g_4 = 0.5$, $g_1 = g_3 = 0.99$, $g_2 = 1$, $g_n = 0$ *elsewhere, renormalized such that* $G(1) = 2$. *Here* $G(-1) \approx 0.01$ *is so small that divergence is not obvious at the level of the figure. Divergence is here due to oscillations that occur in the graph of g^j_n. Although very small, these oscillations are so rapid that they preclude convergence.*



(a)

(b)

FIG. 3. *Two examples of converging dyadic up-scaling schemes (after* [6]). *The g^j_n's are plotted against $n2^{-j}$ for $j = 1, \cdots, 6$, with coefficients joined by segments, so that the behavior of the "slopes" can be observed.* (a) *The limit function is* $\dot{C}^{0.5500\cdots}$ *and not* C^1; *therefore, slopes are allowed to increase indefinitely near the peaks of the limit function.* (b) *The limit function is* $\dot{C}^{1.0878\cdots}$; *therefore,* C^1. *Slopes are constrained to be bounded, especially near the apparent "peaks" of the limit function.*

the above "stepwise interpolation" examples are recovered by letting $n_j = \lfloor 2^j x + \frac{1}{2} \rfloor$, $\lfloor 2^j x \rfloor$, $\lceil 2^j x \rceil$, respectively.

Convergence of the linear interpolations $\varphi^j_{\mathcal{L}}(x)$ of the g^j_n is also implied by Definition 3.1. This comes from the inequality $|\varphi^j_{\mathcal{L}}(x) - g^j_{n_j}| \leq |g^j_{n_j+1} - g^j_{n_j}|$, which holds for $n_j = \lfloor 2^j x \rfloor$ because $\varphi^j_{\mathcal{L}}$ is monotonous on each interval $[n2^{-j},\ (n+1)2^{-j}]$. From (3.3) this clearly implies $|\varphi^j_{\mathcal{L}}(x) - g^j_{n_j}| \to 0$; hence $|\varphi(x) - \varphi^j_{\mathcal{L}}(x)| \to 0$. Convergence of smoother interpolation functions such as splines are similarly implied by Definition 3.1.

Still another definition of convergence was proposed in [14], [15] by Dyn and Levin. For example, uniform convergence is expressed as the existence of a continuous function $\varphi(x)$ such that $\sup_n |g^j_n - \varphi(n2^{-j})| \to 0$ as $j \to \infty$. Note that this is implied by the uniform convergence of the linear interpolations $\varphi^j_{\mathcal{L}}(x)$; since the $\varphi^j_{\mathcal{L}}(x)$'s are continuous, their uniform limit is continuous and we have $\sup_n |g^j_n - \varphi(n2^{-j})| = \sup_n |\varphi^j(n2^{-j}) - \varphi(n2^{-j})|$, which $\to 0$ as $j \to \infty$.

Therefore, Definition 3.1 implies all the others. In fact, §7 shows that all definitions of uniform convergence presented in this section are equivalent. Since the results of this paper are mostly based on uniform convergence, they remain valid for various frameworks used in other works (in particular [6], [14]).

It is possible, however, to find examples for which *pointwise* convergence holds for one definition and not for another. Consider, for example, $G(X) = X$ for which $g^j_n = 1$ if $n = 2^j - 1$ and zero otherwise. Here pointwise convergence of stepwise—or linear—interpolations holds and we easily find that the limit exists and is $\varphi(x) \equiv 0$. (This is a typical example of a pointwise, nonuniform convergence to a continuous function.) But convergence does not hold for all x in the sense of Definition 3.1 because the scheme diverges for $x = 1$ (take $n_j = 2^j - 1$). Therefore, Definition 3.1 forbids "sharp discontinuities" about which $|g^j_{n_j+1} - g^j_{n_j}|$ does not tend to zero as $j \to \infty$.

The choice $G(X) = 1 + X$ behaves similarly. For $x \neq 0$ and $x \neq 1$, the scheme converges to a limit function equal to 1 for $0 < x < 1$, and zero for $x < 0$ or $x > 1$; however, depending on the choice of the definition of pointwise convergence, it either converges or diverges for $x = 0$ and $x = 1$.

4. Basic properties. Several basic properties and simplifications for the study of convergent binary subdivision schemes follow easily from the description (1.2), (2.2). First note that all functions considered in this paper are compactly supported because the mask g_n is of finite length L. In fact, we easily find that the length of g^j_n is $(2^j - 1)(L - 1) + 1$ by estimating the polynomial degree of (2.2). Therefore, $\varphi(x)$, if it exists, has compact support $[0,\ L - 1]$. This property makes many technical proofs easier.

Second, we can restrict the initial sequence in (1.2) to δ_n. For an arbitrary initial sequence of finite length h_n, (2.2) is simply multiplied by $H(X^{2^j})$:

$$H^j(X) = G^j(X)\, H(X^{2^j}). \tag{4.1}$$

The iterated sequence is, therefore, $h^j_n = \sum_k h_k\, g^j_{n-2^j k}$. From Definition 3.1, the limit function becomes $\psi(x) = \sum_k h_k \varphi(x - k)$ instead of $\varphi(x)$. Moreover, since both functions are compactly supported, $\varphi(x)$ itself can be written as $\varphi(x) = \sum_k (h^{-1})_k \psi(x - k)$, where $(h^{-1})_n$ is the convolutional inverse of h_n, i.e., $\sum_k (h^{-1})_k h_{n-k} = \delta_n$. The convergence and regularity properties of $\varphi(x)$ and $\psi(x)$ are, therefore, the same, and we can restrict ourselves to the study of the g_n^j's and $\varphi(x)$.[2]

In order that $\varphi(x)$ is well defined or does not vanish for all x, the iterated sequences g_n^j should neither diverge nor tend to zero as $j \to \infty$. The following proposition shows that this requires some basic conditions to be fulfilled by mask g_n.

PROPOSITION 4.1. If $\varphi(x) \neq 0$ exists for some $x \in \mathbf{R}$, then

$$\sum_k g_{2k} = \sum_k g_{2k+1} = 1, \quad \text{i.e., } G(1) = 2 \text{ and } G(-1) = 0. \tag{4.2}$$

Proof. The key point is to consider the even and odd-indexed sequences g_{2n}^j and g_{2n+1}^j separately. Let $y = x/2$ and $n_j = n_j(y)$ be a sequence of integers satisfying (3.2) for y. On one hand, from Definition 3.1, the common limit of $g_{2n_j}^j$ and $g_{2n_j+1}^j$ as $j \to \infty$ is $\varphi(2y) = \varphi(x)$. But from (2.3) we also have

$$g_{2n}^j = \sum_k g_{2k}\, g_{n-k}^{j-1},$$

$$g_{2n+1}^j = \sum_k g_{2k+1}\, g_{n-k}^{j-1}.$$

Letting $n = n_j$ and applying Definition 3.1 to the right-hand sides of these equations, we obtain that their respective limits as $j \to \infty$ are $(\sum_k g_{2k})\varphi(2y)$ and $(\sum_k g_{2k+1})\varphi(2y)$. By identification we therefore have

$$\varphi(2y) = \left(\sum_k g_{2k}\right)\varphi(2y) = \left(\sum_k g_{2k+1}\right)\varphi(2y).$$

Dividing the members of this equality by $\varphi(2y) \neq 0$ gives (4.2). □

Condition (4.2) may be interpreted as follows. On one hand, $G(1) = 2$ is just a normalization condition that ensures that the order of magnitude of g_n^j is preserved when $j \to \infty$. On the other hand, the fact that $G(X)$ must have at least one zero at $X = -1$ is a "local" requirement. For example, it ensures that the g_n^j's, for large j, do not rapidly oscillate in n between two different limits, $(\sum_k g_{2k})\varphi(2y)$ and $(\sum_k g_{2k+1})\varphi(2y)$. Figure 2 illustrates this phenomenon on a particular example (see also [20]).

Note that (4.2) is not sufficient to ensure convergence. As an example, consider $G(X) = 1 + X^3$. Here $G^j(X)$ is a polynomial in X^3; therefore, g_n^j vanishes for $n \neq 3k$ ($k \in \mathbf{Z}$), whereas $g_{3k}^j = 1$. It, therefore, cannot converge to a limit function. (Section 7 gives a necessary and sufficient condition for uniform convergence.)

5. Exact computation of limit functions.

Assume that the limit function $\varphi(x)$ of a binary subdivision scheme g_n^j exists for all $x \in \mathbf{R}$. This section derives a simple, easily implementable method for computing the *exact* values of $\varphi(x)$ at dyadic rationals $x = n2^{-j}$, $n \in \mathbf{Z}$, with a finite number of operations. The starting point is the *two-scale difference equation* [8], [9] satisfied by $\varphi(x)$:

$$\varphi(x) = \sum_k g_k\, \varphi(2x - k). \tag{5.1}$$

[2] Note that this restriction works only for initial sequences of *finite* length. If, e.g., $h_n = 1$ for all $n \in \mathbf{Z}$, then using the definition $h_n^{j+1} = \mathcal{G}\{h_n^j\}$ and Proposition 4.1, it easily follows by induction on j that $h_n^j \equiv 1$. Hence $\psi(x) = \sum_k \varphi(x - k) \equiv 1$ is C^∞, whatever the regularity order of $\varphi(x)$.

This equation, which was mentioned in the introduction, is easily derived by using (2.4) for $n = n_j$ (3.2) and applying Definition 3.1.

Now, let

$$\Phi^j(X) = \sum_n \varphi(n2^{-j})X^n \tag{5.2}$$

be the polynomial associated to the sequence $\varphi(n2^{-j})$. Taking $x = n2^{-j-1}$ in the two-scale difference equation yields $\varphi(n2^{-j-1}) = \sum_k g_k\, \varphi(2^{-j}(n - 2^j k))$, i.e., $\Phi^{j+1}(X) = \Phi^j(X)G(X^{2^j})$. By iteration we have

$$\Phi^j(X) = \Phi(X)G^j(X), \tag{5.3}$$

where

$$\Phi(X) = \Phi^0(X) = \sum_n \varphi(n)X^n. \tag{5.4}$$

Equation (5.3) is very useful, since it links the values of the iterated sequences g_n^j to the ones of the limit function $\varphi(n2^{-j})$. The latter are simply obtained from the g_n^j's by convolving them with the sequence $\varphi(n)$, provided that the $\varphi(n)$'s can be predetermined.

There are several methods for precomputing $\varphi(n)$, which is, by definition, the limit of $g_{n2^j}^j$ as $j \to \infty$. First note that we have, from (2.4),

$$g_{n2^{j+1}}^{j+1} = \sum_k g_k\, g_{(2n-k)2^j}^j = \mathcal{G}^*\{g_{n2^j}^j\},$$

where $\mathcal{G}^*$ is the following transposed operator [6], [18] of $\mathcal{G}$ (1.1):

$$u_n \xrightarrow{\mathcal{G}^*} v_n = \sum_{k\in\mathbf{Z}} g_k\, u_{2n-k}. \tag{5.5}$$

Therefore, $\varphi(n)$ can be determined as the limit of $(\mathcal{G}^*)^j\{\delta_n\}$ as $j \to \infty$.

Another method stems from the resulting equality

$$\varphi(n) = \mathcal{G}^*\{\varphi(n)\}. \tag{5.6}$$

The sequence $\varphi(n)$, $n = 0, \cdots, L-1$ (where L is the length of the mask g_n), is here determined, up to normalization, as the eigenvector of the operator $\mathcal{G}^*$ associated to the eigenvalue 1. To obtain a normalization for $\varphi(n)$, rewrite (5.5) under polynomial form

$$V(X^2) = (U(X)G(X) + U(-X)G(-X))/2. \tag{5.7}$$

Since we have, by Proposition 4.1, $G(1) = 2$, and $G(-1) = 0$, it follows that $\mathcal{G}^*$ preserves the sum of sequences; hence $\sum_n \varphi(n) = \sum_n g_{n2^j}^j = \sum_n g_{2n} = 1$, i.e.,

$$\Phi(1) = 1. \tag{5.8}$$

6. Stability. There is an exceptional class of limit functions $\varphi(x)$ for which the regularity estimates derived in this paper will not always be optimal. Optimality, as well as some other results of this paper, will be proven only in the case of "stability," in the sense of the following definition.

DEFINITION 6.1. A binary subdivision scheme converging to a nonzero limit function (or its limit function $\varphi(x) \not\equiv 0$) is *stable* if no root of $\Phi(X)$ (5.4) lies on the unit circle, i.e.,

$$\sum_n \varphi(n)\, e^{in\omega} \neq 0 \quad \text{for all } \omega \in \mathbf{R}.$$

The terminology "stable" comes from (2.6), (2.7) written for $V(X) = \Phi(X)$,

$$c_1 \,\|U(X)\|_\infty \leq \|U(X)\Phi(X)\|_\infty \leq c_2 \,\|U(X)\|_\infty,$$

which means that the filter of impulse response $\varphi(n)$ and its inverse are numerically stable for finite length sequences. The stability condition slightly restricts the choice of the scaling sequence g_n. For example, if the mask length is $L = 4$, "unstable" $\varphi(x)$'s are such that $g_0 = g_3$ and $g_1 = g_2$. All (real-valued) limit functions are stable for lengths up to 3. Note that an interpolatory subdivision scheme is always stable since it has the property that $\varphi(n/2) = g_n$ (see (3.1) for $j = 1$); hence $\Phi(X) \equiv 1$.

In fact, in the rest of this paper, stability can be replaced by the even weaker condition that there exists $x \in \mathbf{R}$ such that

$$\sum_n \varphi(n + x)\, e^{in\omega} \neq 0 \quad \text{for all } \omega \in \mathbf{R} \tag{6.1}$$

(a similar, but stronger stability condition appears in [15]). Condition (6.1) comes from (5.3) where n is replaced by $n + x$, for any fixed number x. That is,

$$\Phi_x^j(X) = \Phi_x(X) G^j(X)$$

where

$$\Phi_x^j(X) = \sum_n \varphi((n + x)2^{-j}) X^n$$

and $\Phi_x(X) = \Phi_x^0(X)$.

Although almost all convergent subdivision schemes are stable, it is easy to construct unstable ones (even with definition (6.1)). As in the preceding section we have the following generalization of (5.6), (5.7).

$$\Phi_x(X^2) = (\Phi_{2x}(X)G(X) + \Phi_{2x}(-X)G(-X))/2.$$

Therefore, any polynomial mask $G(X)$ divisible by $(X^2 - e^{i\omega})$, $\omega \neq 0$, yields unstability since we have $\Phi_x(e^{i\omega}) = 0$.

I conjecture that the converse holds, i.e., stability (in the weak form (6.1)) is equivalent to the condition that $G(X)$ has no pair of opposite zeros $(e^{i\omega/2},\ e^{i(\omega/2+\pi)})$ on the unit circle. If this conjecture is true, then the regularity estimates presented below will be optimal under the simple condition above on the mask coefficients g_n, which is easy to check. When this condition is not satisfied, it is possible to apply a trick as shown at the end of §9 which allows one to consider another, *stable* binary subdivision scheme which has the same regularity properties.

7. Continuous limit functions. The framework of uniform convergence (3.4) is shown to be very convenient in the sequel, and the following theorem shows that all stable continuous limit functions are obtained by uniform convergence. We shall then derive a necessary and sufficient condition for uniform convergence in all cases.

THEOREM 7.1. *Assume that a binary subdivision scheme converges pointwise to a limit function $\varphi(x)$ for all $x \in \mathbf{R}$. If the convergence is uniform, then $\varphi(x)$ is continuous. The converse is true if $\varphi(x)$ is stable.*

Proof. ($\Rightarrow$) In §3 we have seen that uniform convergence (3.4) implies uniform convergence of linear interpolations $\varphi^j_{\mathcal{L}}(x)$ of the g^j_n's to $\varphi(x)$. Since this is a uniformly convergent sequence of compactly supported continuous functions, $\varphi(x)$ is continuous.

($\Leftarrow$) We have

$$\sup_x |\varphi(x) - g^j_{n_j}| \leq \sup_x |\varphi(x) - \varphi(n_j 2^{-j})| + \sup_{n_j} |\varphi(n_j 2^{-j}) - g^j_{n_j}|$$

where n_j is a sequence of integers satisfying (3.2). Since $\varphi(x)$ is compactly supported and continuous, it is uniformly continuous. Therefore $\sup_x |\varphi(x) - \varphi(n_j 2^{-j})| \to 0$. The other term can be written $\sup_n |\varphi(n2^{-j}) - g^j_n| = \|\Phi^j(X) - G^j(X)\|_\infty$. From (5.3) we have $\Phi(X)(\Phi^j(X) - G^j(X)) = (\Phi(X) - 1)\Phi^j(X)$. Since (5.8) holds, $X - 1$ divides $\Phi(X) - 1$ and we can write, using (2.6), $\|\Phi(X)(\Phi^j(X) - G^j(X))\|_\infty \leq c\,\|(X-1)\,\Phi^j(X)\|_\infty$. The latter norm is $\sup_n |\varphi(n2^{-j}) - \varphi((n-1)2^{-j})|$, which tends to zero as $j \to \infty$ because $\varphi(x)$ is uniformly continuous. Now we can use (2.7) with $V(X) = \Phi(X)$ since $\Phi(X)$ is stable. This yields $\|\Phi^j(X) - G^j(X)\| \to 0$ as $j \to \infty$, which ends the proof. □

It is an open problem to find a limit function $\varphi(x)$ for which the convergence is not uniform (in the sense of Definition 3.1). That would imply that $\varphi(x)$ is unstable or discontinuous.

We now derive a necessary and sufficient condition for uniform convergence of a binary subdivision scheme g^j_n to a (continuous) limit function $\varphi(x)$. (By Theorem 7.1, this also gives a necessary and sufficient condition for the continuity of a stable limit function.) We need the following lemma, which will also be useful for deriving an optimal regularity estimate in §11.

LEMMA 7.2. *Assume $G(-1) = 0$ and let $F(X) = G(X)/(1+X)$. The sequence of the first-order differences of g^j_n,*

$$d^j_n = g^j_n - g^j_{n-1},$$

follows a binary subdivision scheme with polynomial mask $F(X)$ and initial sequence's polynomial $1 - X$.

In addition, for any fixed positive integer i, we have

$$\max_n |g^j_{n+1} - g^j_n| \leq c \left(\max_{0 \leq n \leq 2^i - 1} \sum_k |f^i_{n-2^i k}| \right)^{j/i}, \tag{7.1}$$

where f_n is the mask associated to the polynomial $F(X)$ and c is a constant independent of j.

Proof. Let $D^j(X) = (1-X)G^j(X)$ be the polynomial associated to d^j_n. We have

$$\begin{aligned} D^j(X) &= (1-X)(1+X)(1+X^2)\cdots(1+X^{2^{j-1}})\,F^j(X) \\ &= (1-X^{2^j})F^j(X), \end{aligned}$$

which shows, by (4.1), that d_n^j follows the announced subdivision scheme.

Using (2.2) in the above equation, we can write $D^{i+\ell}(X) = F^i(X)\, D^\ell(X^{2^i})$, which also reads $d_n^{i+\ell} = \sum_k f_{n-2^ik}^i d_k^\ell$. Majorating yields

$$\|D^{i+\ell}(X)\|_\infty \le \left(\max_n \sum_k |f_{n-2^ik}^i|\right) \|D^\ell(X)\|_\infty,$$

which, by iteration for $j = \ell + ni$, $0 \le \ell \le i-1$, gives (7.1) where c depends only on the fixed integers i and ℓ. □

THEOREM 7.3. *A binary subdivision scheme g_n^j converges uniformly to a (continuous) limit function if and only if $G(1) = 2$, $G(-1) = 0$ and*

$$\max_n |g_{n+1}^j - g_n^j| \to 0 \quad \text{as } j \to \infty. \tag{7.2}$$

Moreover, there exists $\alpha > 0$ such that

$$\max_n |g_{n+1}^j - g_n^j| \le c\, 2^{-j\alpha}. \tag{7.3}$$

Proof. (⇒) immediately results from Proposition 4.1 and the inequality

$$\max_n |g_{n+1}^j - g_n^j| \le \sup_x |\varphi(x) - g_{n_j+1}^j| + \sup_x |\varphi(x) - g_{n_j}^j|.$$

(⇐) We first prove that (7.3) is implied by conditions $G(1) = 2$, $G(-1) = 0$ and (7.2). First note that from the first part of Lemma 7.2 we have $(1-X)G^j(X) = (1-X^{2^j})F^j(X)$, i.e., $g_n^j - g_{n-1}^j = f_n^j - f_{n-2^j}^j$. Write

$$\begin{aligned} f_n^j &= (f_n^j - f_{n-2^j}^j) + (f_{n-2^j}^j - f_{n-2\cdot 2^j}^j) + \cdots \\ &= (g_n^j - g_{n-1}^j) + (g_{n-2^j}^j - g_{n-2^j-1}^j) + \cdots. \end{aligned}$$

The number of terms in the sums is bounded by L because the length of f_n^j is bounded by 2^jL. From (7.2) each term tends to zero uniformly with respect to n; hence so does f_n^j. Therefore, there exists a (sufficiently large) index i such that $\max_n |f_n^i| \le \varepsilon_i < 1/L$. Now since the number of terms in the sum in (7.1) is bounded by L, the second part of Lemma 7.2 gives (7.3) with $\alpha = -\log_2(L\varepsilon_i)/i > 0$.

To prove the converse part of the theorem, consider $\sup_{n_j} |g_{n_{j+1}}^{j+1} - g_{n_j}^j|$, where n_j satisfies (3.2). This equals $\sup_{n_j} |g_{2n_j+m_j}^{j+1} - g_{n_j}^j|$, where $m_j = n_{j+1} - 2n_j$ is a bounded integer. Now, from (2.3) we can write $g_{2n+m}^{j+1} = \sum_k g_{2k+m}\, g_{n-k}^j$. Therefore, the sequence $g_{2n+m}^{j+1} - g_n^j$ is a convolved version of g_n^j; its associated polynomial can be written in the form $U_m(X)G^j(X)$. But from (4.2), we have $\sum_k g_{2k+m} = 1$ (for all m), and, therefore, $U_m(1) = 0$. Using (2.6), it follows that

$$\sup_{n_j} |g_{2n_j+m_j}^{j+1} - g_{n_j}^j| = \|U_{m_j}(X)G^j(X)\|_\infty \le c'\, \|(1-X)G^j(X)\|_\infty,$$

where c' is a constant (independent of j since m_j is bounded). From (7.3) the latter norm is bounded by $c\,c'\,2^{-j\alpha}$. We, therefore, end up with $\sup_{n_j} |g_{n_{j+1}}^{j+1} - g_{n_j}^j| \le cc'\,2^{-j\alpha}$. Iterating this inequality, we obtain, for any $\ell > 0$,

$$\sup_{n_j} |g_{n_{j+\ell}}^{j+\ell} - g_{n_j}^j| \le c\,c'\,(2^{-(j+\ell-1)\alpha} + \cdots + 2^{-(j+1)\alpha} + 2^{-j\alpha}) \le c''\,2^{-j\alpha}, \tag{7.4}$$

which shows that the sequence of functions $g^j_{n_j(x)}$ is a uniform Cauchy sequence, which converges uniformly to a continuous limit function $\varphi(x)$. □

This theorem has several interesting consequences. First, we shall see in §8 that (7.3), in fact, implies that $\varphi(x)$ is Lipschitz of order α, which is stronger than continuity.[3] Therefore, by Theorem 7.1, a continuous stable limit function is automatically Lipschitz of order α for some $\alpha > 0$.

Second, note that the necessary and sufficient condition is quite weak and intuitive: it is sufficient that the differences $g^j_{n+1} - g^j_n \to 0$ uniformly as $j \to \infty$ to obtain a continuous limit function.[4] In fact, we easily find that (7.2) holds for any definition of uniform convergence presented in §3.1. (For example, any uniformly convergent sequence of interpolating functions $\varphi^j(x)$ of the g^j_n's such that $g^j_n = \varphi^j(n2^{-j})$ clearly gives (7.2).) Since we have seen that Definition 3.1 for uniform convergence implies the others, it follows that *all* these definitions of uniform convergence are equivalent.

In particular, some results derived in this paper have been derived in the framework of Dyn and Levin [14], [15] as well. The necessary and sufficient condition (7.2) appears in [14] for interpolatory subdivision schemes and was first derived in [15, Thm. 3.2] for general binary subdivision schemes—using the (apparently) weaker definition mentioned in §3.1—in a slightly different but equivalent form, namely, $\max_n |f^j_n| \to 0$ as $j \to \infty$. Theorem 7.3 was included here in order that this paper be self-contained, since some material presented in this section is also useful in the sequel.

8. Lipschitz limit functions. In this section we want to characterize Lipschitz limit functions. Recall that $\varphi(x)$ is said to be *Lipschitz of order* α $(0 < \alpha \leq 1)$, $\varphi(x) \in \dot{C}^\alpha$, if we have for all x and $h \in \mathbf{R}$,

$$(8.1) \qquad |\varphi(x+h) - \varphi(x)| \leq c\,|h|^\alpha,$$

where c is a constant. Here, $\varphi(x)$ is compactly supported, and (8.1) needs to be satisfied only for small h's. Since the spaces $\dot{C}^\alpha$, for $0 < \alpha \leq 1$, interpolate between C^0 and C^1, a $\dot{C}^\alpha$-function will be said to be *regular of order* α. There is a slight irritation in that C^1 and $\dot{C}^1$ do not coincide; for example, a linear spline function is $\dot{C}^1$ but not differentiable at its knots.

THEOREM 8.1. *If* $G(1) = 2$, $G(-1) = 0$, *and*

$$(8.2) \qquad \max_n |g^j_{n+1} - g^j_n| \leq c\,2^{-j\alpha}$$

for some $0 < \alpha \leq 1$, *then the binary subdivision scheme converges uniformly to a* $\dot{C}^\alpha$ *limit function. The converse is true if* $\varphi(x)$ *is stable.*

In addition, the more regular the limit, the faster the convergence to this limit:

$$(8.3) \qquad \sup_x |\varphi(x) - g^j_{n_j}| \leq c\,2^{-j\alpha}$$

for any sequence n_j *of integers satisfying* (3.2).

Proof. (⇒) Let us first prove (8.3). Since (8.2) holds, we are in the framework of Theorem 7.3, and (7.4) holds. Letting $\ell \to \infty$ in (7.4) gives (8.3).

[3] Using the same proof as the one of Theorem 7.3, we can show that when (3.2) is replaced by $n_j 2^{-j} \to x$ as $j \to \infty$, uniform convergence requires (7.3) for $\alpha = 1$, which corresponds to almost continuously differentiable functions.

[4] In contrast, the *slopes* $(g^j_{n+1} - g^j_n)/(2^{-j})$ may indefinitely increase (see next section).

We now prove that $\varphi(x)$ is $\dot{C}^\alpha$. Let $n_j = n_j(x)$ satisfy (3.2) (for all $x \in \mathbf{R}$) and consider the inequality

$$\begin{aligned}\sup_x |\varphi(x+h) - \varphi(x)| &\le \sup_x |\varphi(x+h) - g^j_{n_j(x+h)}| \\ &\quad + \sup_x |g^j_{n_j(x+h)} - g^j_{n_j(x)}| + \sup_x |g^j_{n_j(x)} - \varphi(x)|.\end{aligned}$$

By (8.3), the first and third terms in the right-hand side of this inequality are bounded by $c2^{-j\alpha}$. Assume, for example, that $|h| < 1$. If $h > 0$, choose $n_j(x) = \lfloor x2^j \rfloor$, otherwise choose $n_j(x) = \lceil x2^j \rceil$. A simple calculation yields $|n_j(x+h) - n_j(x)| \le |n_j(h)| + \varepsilon$, where $\varepsilon = 0$ or ± 1. Now, let j be such that $2^{-j} \le |h| < 2^{-j+1}$. This gives $|n_j(h)| = 1$; hence we find, from (8.2), that $\sup_x |g^j_{n_j(x+h)} - g^j_{n_j(x)}| \le c2^{-j\alpha}$. Putting all inequalities together yields $\sup_x |\varphi(x+h) - \varphi(x)| \le c'2^{-j\alpha} \le c|h|^\alpha$, i.e., $\varphi(x)$ is $\dot{C}^\alpha$.

($\Leftarrow$) $G(1) = 2$, $G(-1) = 0$ result from Proposition 4.1. Since $\varphi(x)$ is $\dot{C}^\alpha$, we have $|\varphi((n+1)2^{-j}) - \varphi(n2^{-j})| \le c2^{-j\alpha}$, i.e.,

$$\|(1-X)\Phi^j(X)\|_\infty = \|\Phi(X)(1-X)G^j(X)\|_\infty \le c2^{-j\alpha}$$

(the first equality comes from (5.3)). Because $\varphi(x)$ is stable, we can apply (2.7) to obtain the inequality $\|(1-X)G^j(X)\|_\infty \le c'\,2^{-j\alpha}$, which is (8.2). □

This theorem provides an intuitive interpretation of regularity of order $0 \le \alpha < 1$ for binary subdivision schemes: regularity $\dot{C}^\alpha$ holds if and only if the absolute values of the "slopes" $(g^j_{n+1} - g^j_n)/2^{-j}$ of the discrete curves g^j_n's (see next section) grow less than $2^{j(1-\alpha)}$ when j indefinitely increases. For example, if the slopes of g^j_n are always bounded for all j's, then $\varphi(x)$ is $\dot{C}^1$. On the contrary, less regularity allows slopes to increase indefinitely and the resulting limit function, although continuous, may present a "fractal" structure as shown in Fig. 3. Note that in this case, (8.3) means that uniform convergence of the curves g^j_n is slower as slopes increase faster.

As an example, consider the convergence of binary subdivision schemes in the case of positive masks $g_n > 0$, $n = 0, \cdots, L-1$, as studied by Micchelli and Prautzsch in [17]. They found that

$$\sup_{0 \le n-m \le L-2} |g^j_m - g^j_n| \le c\,(1 - \min g_n)^j;$$

hence any binary subdivision scheme with positive mask uniformly converges to a continuous function [17]. Theorem 8.1 immediately applies to show that the limit function is, in fact, $\dot{C}^\alpha$, where $\alpha = -\log_2(1 - \min g_n)$.

Since we have a characterization of regularity for stable $\varphi(x)$'s, it is easy to find a condition on g_n that states an exact regularity order $0 < \alpha < 1$.

COROLLARY 8.2. *Let g_n be a stable binary subdivision scheme such that $G(1) = 2$ and $G(-1) = 0$. If, for $0 < \alpha < 1$,*

$$\max_n |g^j_{n+1} - g^j_n| \text{ decreases as } 2^{-j\alpha} \quad \text{when } j \to \infty, \tag{8.4}$$

then the limit function $\varphi(x)$ is $\dot{C}^\alpha$ but is not $\dot{C}^{\alpha+\varepsilon}$, for any $\varepsilon > 0$.

Proof. This is an immediate consequence of Theorem 8.1. If $\varphi(x)$ were $\dot{C}^{\alpha+\varepsilon}$ (with $\varepsilon > 0$ small enough so that $\alpha + \varepsilon < 1$), we would have $|g^j_{n+1} - g^j_n| \le c2^{-j(\alpha+\varepsilon)}$, which contradicts (8.4). □

Note that Corollary 8.2 does not hold if $\alpha = 1$, since $|g^j_{n+1} - g^j_n|$ cannot decrease faster than 2^{-j} as $j \to \infty$ when $\varphi(x)$ is more regular than C^1 (see §10). Otherwise, intuitively the derivative of $\varphi(x)$ would vanish identically, which would imply $\varphi(x) = 0$ since $\varphi(x)$ is compactly supported.

9. Continuously differentiable limit functions. In this section, we study the derivatives of the limit function $\varphi(x)$. We start by defining finite differences of the g^j_n's, which will be shown to converge to the derivatives of $\varphi(x)$. The first finite difference is

$$\Delta g^j_n = (g^j_n - g^j_{n-1})/2^{-j}, \quad \text{i.e.,} \ \Delta G^j(X) = 2^j(1-X)G^j(X). \tag{9.1}$$

In other words, the Δg^j_n's are the *slopes* of the "discrete curve" g^j_n plotted against $n2^{-j}$ (see Figs. 2 and 3). Finite differences $\Delta^k g^j_n$ of order k are simply obtained by applying k times the difference operator Δ:

$$\Delta^k G^j(X) = 2^{jk}(1-X)^k G^j(X). \tag{9.2}$$

In order to study finite differences $\Delta^k g^j_n$ similarly, as for the g^j_n's, it is convenient to express them as binary subdivision schemes as well, associated to masks other than g_n. The following lemma shows that this is possible when $G(X)$ has enough zeros at $X = -1$.

LEMMA 9.1. *Assume $G(X)$ has at least k zeros at $X = -1$ and define $G_k(X)$ by*

$$G(X) = \left(\frac{1+X}{2}\right)^k G_k(X). \tag{9.3}$$

Then the finite differences $\Delta^k g^j_n$'s follow a binary subdivision scheme with the initial sequence's polynomial $(1-X)^k$ and polynomial mask $G_k(X)$.

Proof. This is an immediate generalization of the first part of Lemma 7.2. From (9.2), (9.3), we have

$$\Delta^k G^j(X) = 2^{jk}(1-X)^k \prod_{i=0}^{j-1} \left(\frac{1+X^{2^i}}{2}\right)^k G^j_k(X),$$

where $G^j_k(X) = (G_k)^j(X)$ is defined by (2.5). Using the identity $(1-Y)(1+Y) = 1 - Y^2$ for $Y = X, X^2, X^4, \cdots$, we obtain

$$\Delta^k G^j(X) = G^j_k(X)(1 - X^{2^j})^k, \tag{9.4}$$

which from (4.1) proves the lemma. □

Using the preceding sections we can extend the results of §7 to higher-order regularity C^N (N-times continuously differentiable functions).

THEOREM 9.2. *If the sequence of the Nth-order finite differences $\Delta^N g^j_{n_j}$ (where n_j satisfies (3.2)) uniformly converges as $j \to \infty$, then $\varphi(x)$ is C^N. The converse is true if $\varphi(x)$ is stable.*

In addition, $\Delta^k g^j_{n_j}$ (where n_j satisfies (3.2)) converges uniformly to $\varphi^{(k)}(x)$, the kth-order derivative of $\varphi(x)$, for $k = 0, \cdots, N$, and $G(X)$ has at least $N+1$ zeros at $X = -1$.

Proof. ($\Rightarrow$) Let us first prove uniform convergence of the kth-order finite differences ($k = 0, \cdots, N$) by backward induction on k. We show that if $\Delta^{k+1} g^j_n$ converges uniformly to some (continuous) function $h(x)$, then $\Delta^k g_n^j$ converges uniformly to the primitive of $h(x)$ defined by

$$H(x) = \int_{-\infty}^{x} h(u)\, du.$$

For simplicity we assume $k = 0$, the proof being identical for $k > 0$.

First we prove that $H(x)$ is compactly supported. The functions $\Delta g^j_{\lfloor x2^j \rfloor}$ are all Riemann-integrable and converge uniformly to the function $h(x)$ of compact support $[0; L-1]$ (where L is the length of g_n); therefore,

$$\int_0^{L-1} \Delta g^j_{\lfloor x2^j \rfloor}\, dx = \sum_n 2^{-j} \Delta g_n^j \text{ tends to } \int_0^{L-1} h(u)\, du \quad \text{as } j \to \infty.$$

But since $\Delta G^j(1) = 0$ (see (9.4)), these integrals vanish, which shows that $H(x)$ is compactly supported.

Now, since $H(x)$ is C^1 and has compact support, it is uniformly continuously differentiable and, therefore, $\sup_x |\Delta g^j_{n_j} - \big(H(n_j 2^{-j}) - H((n_j - 1)2^{-j})\big)/2^{-j}|$ tends to zero as $j \to \infty$, where n_j are integers satisfying (3.2). This can be written

$$\|2^j(1-X)(G^j(X) - \Psi^j(X))\|_\infty \to 0,$$

where $\Psi^j(X)$ is the polynomial associated to the sequence $H(n2^{-j})$. But for any polynomial $U(X)$, we have

$$\|U(X)\|_\infty \le \sum_k |u_k - u_{k-1}| \le d\, \|(1-X)U(X)\|_\infty,$$

where d is the degree of the polynomial $U(X)$. Applying this to $U(X) = G^j(X) - \Psi^j(X)$ of degree $(L-1)(2^j - 1)$, we obtain $\sup_{n_j} |g^j_{n_j} - H(n_j 2^{-j})| = \|G^j(X) - \Psi^j(X)\|_\infty \le (L-1)\|2^j(1-X)(G^j(X) - \Psi^j(X))\|_\infty$, which tends to zero; therefore, g_n^j converges uniformly to $\varphi(x) = H(x)$, and $h(x)$ is the derivative of $\varphi(x)$. By induction it follows that the kth-order finite differences converge uniformly to the kth-order derivatives of $\varphi(x)$ for $0 \le k \le N$.

In particular, the continuous uniform limit of $\Delta^N g_n^j$ is $\varphi^{(N)}(x) \in C^0$. Therefore, $\varphi(x)$ is C^N. The property that $G(X)$ has at least $N+1$ zeros at $X = -1$ follows easily by forward induction on the derivative order k as a consequence of Proposition 4.1 and Lemma 9.1.

($\Leftarrow$) We prove uniform convergence of the kth-order finite differences to the kth-order derivative of $\varphi(x)$ ($k = 0, \cdots, N$), from the assumption that $\varphi(x)$ is stable and C^N, by forward induction on k. For $k = 0$, this is true by Theorem 7.1. It remains to prove that this implies $\sup_x |\varphi^{(k)}(x) - \Delta^k g^j_{n_j}| \to 0$ for $k = 1, \cdots, N$, where n_j satisfies (3.2). For simplicity, assume $k = 1$. The proof is identical for larger k's when one replaces Δ by Δ^k. Define $\Delta\Phi^j(X) = 2^j(1-X)\Phi^j(X)$, where $\Phi^j(X)$ is defined by (5.2), i.e., $\Delta\varphi(n_j 2^{-j}) = 2^j(\varphi(n_j 2^{-j}) - \varphi((n_j - 1)2^{-j}))$. We have

$$\begin{aligned} \sup_x |\varphi'(x) - \Delta g^j_{n_j}| &\le \sup_x |\varphi'(x) - \varphi'(n_j 2^{-j})| \\ &\quad + \sup_x |\varphi'(n_j 2^{-j}) - \Delta\varphi(n_j 2^{-j})| \\ &\quad + \sup_x |\Delta\varphi(n_j 2^{-j}) - \Delta g^j_{n_j}|. \end{aligned} \tag{9.5}$$

The first term in the right-hand side of (9.5) tends to zero as $j \to \infty$ because $\varphi'(x)$ is continuous and compactly supported, hence uniformly continuous. The second term also tends to zero because $\varphi(x)$ is uniformly continuously differentiable on a compact support. Note that this implies

$$\sup_{n_j} |\Delta\varphi(n_j 2^{-j}) - \Delta\varphi((n_j - 1)2^{-j})| = \|(1 - X)\Delta\Phi^j(X)\|_\infty \to 0. \tag{9.6}$$

The third term in the right-hand side of (9.5) can be written as $\|\Delta\Phi^j(X) - \Delta G^j(X)\|_\infty$. But from (5.3) we have $\Phi(X)(\Delta\Phi^j(X) - \Delta G^j(X)) = (\Phi(X) - 1)\Delta\Phi^j(X)$. Since $\Phi(1) = 1$ (5.8), $X - 1$ divides $\Phi(X) - 1$ and we can write, using the norm inequality (2.6), $\|\Phi(X)(\Delta\Phi^j(X) - \Delta G^j(X))\|_\infty \leq c\,\|(X - 1)\Delta\Phi^j(X)\|_\infty$, which tends to zero by (9.6). Now we can use (2.7) with $V(X) = \Phi(X)$ because $\varphi(x)$ is stable. This yields $\|\Delta\Phi^j(X) - \Delta G^j(X)\|_\infty \to 0$ as $j \to \infty$, which ends the proof. □

The direct part of this theorem already appeared in [14], [15]. The converse part also appeared in [14], [15] for interpolatory subdivision schemes (we have seen in §6 that interpolatory subdivision schemes are stable.)

This theorem is useful because it allows us to estimate the regularity of the derivatives of a stable limit function $\varphi(x)$ the same way as for $\varphi(x)$ itself: if $G(X)$ has enough zeros at $X = -1$, the finite differences of the g_n^j's, which converge to the derivatives of $\varphi(x)$, all follow binary subdivision schemes.

Theorem 9.2 also provides an *upper bound* for regularity. Since it is necessary that $G(X)$ has $N + 1$ zeros at $X = -1$ to obtain C^N stable limit functions $\varphi(x)$, the regularity order of $\varphi(x)$ is always bounded by the number of zeros at $X = -1$ in $G(X)$. We shall see that this upper bound may be attained.

However, it is important to note that imposing zeros at $X = -1$ in $G(X)$ does *not* ensure any regularity in general. It does not even ensure convergence, as in the example $G(X) = (1 + X^3)^{N+1}$, which has $N + 1$ zeros at $X = -1$, although g_n^j does not converge for the same reason as for the choice $G(X) = 1 + X^3$ treated in §4. (Section 13 derives a sharp upper bound for regularity.)

Finally, note that the number of zeros of $G(X)$ at $X = -1$ is an upper bound for regularity only for *stable* limit functions. This upper bound may be exceeded for unstable limit functions, as shown in the following example [2], for which the converse part of Theorem 9.2 fails—as well as many other "optimality" results given in the rest of this paper.

Consider the polynomial mask $G(X) = 2^{-N}(1 + X)(1 + X^2)^N$. Setting $U_j(X) = 1 + X + X^2 + \cdots + X^{2^j - 1}$ and applying (2.6) several times give

$$\begin{aligned} \|(1 - X)G^j(X)\|_\infty &\leq 2^{-jN}\|(1 - X^{2^j})\|_1 \|(U_j(X^2))^N\|_\infty \\ &\leq 2^{-jN+1}\|U_j(X)\|_1^{N-1}\|U_j(X)\|_\infty \\ &\leq 2^{-j+1}; \end{aligned}$$

therefore, by Theorem 8.1 the limit function $\varphi(x)$ exists and is $\dot{C}^1$, hence continuous. Theorem 9.2 cannot improve this regularity order since $G(X)$ has only one zero at $X = -1$. However, $\varphi(x)$ is unstable since $1 + X^2$ divides $G(X)$ (see §6), so we might expect higher regularity for $\varphi(x)$.

Now consider another mask $\tilde{G}(X) = 2^{-N}(1 + X)^{N+1}$. It is easy to see that the subdivision scheme $\tilde{g}_n^j$ converges to a $\dot{C}^N$ limit function $\tilde{\varphi}(x)$, i.e., the $(N-1)$th derivative of $\tilde{\varphi}(x)$, for which the mask polynomial is $\tilde{G}_{N-1}(X) = (1+X)^2/2$, is $\dot{C}^1$. This comes from Theorems 8.1 and 9.2 since we have $\|(1 - X)\tilde{G}_{N-1}^j\| \leq$

$2^{-j}\|(1-X^{2^j})\|_1\|U_j(X)\|_\infty \le 2^{-j+1}$. Now, since the two masks are related by $(1+X)^N G(X) = (1+X^2)^N \tilde{G}(X)$, we have by iteration $(1+X)^N G^j(X) = (1+X^{2^j})^N \tilde{G}^j(X)$, i.e.,

$$\sum_{k=0}^{N} \binom{N}{k} g^j_{n-k} = \sum_{k=0}^{N} \binom{N}{k} \tilde{g}^j_{n-2^j k}.$$

Letting $n = n_j$ and $j \to \infty$ gives, by Definition 3.1,

$$\varphi(x) = 2^{-N} \sum_{k=0}^{N} \binom{N}{k} \tilde{\varphi}(x-k),$$

which proves that $\varphi(x)$ is $\dot{C}^N$, hence C^{N-1}, even though $G(X)$ has only one zero at $X=-1$.

This example shows that an unstable binary subdivision scheme may converge to an arbitrary regular limit function while *all* finite differences diverge. Note that since $\tilde{\varphi}(x)$ can also be expressed as a sum of integer translates of $\varphi(x)$ (see the beginning of §4.1), both functions have the same regularity order. It is easy to check that the regularity estimate $\dot{C}^N$ is optimal for $\tilde{\varphi}(x)$ (which is, in fact, the B-spline of degree N [15]); hence it is also optimal for $\varphi(x)$.

Therefore, the argument used in this example has led to an optimal regularity estimate for an *unstable* limit function, while the rest of this paper derives regularity estimates that are optimal for all *stable* limit functions. This example can be easily generalized to the case where unstability is due to the fact that $G(X)$ is divisible by $X^2 - e^{i\omega}$ (see §6). Note that if the conjecture mentioned in §6 is true, then this methods works for arbitrary unstable limit functions (in the sense of (6.1)).

10. Determining the exact Hölder regularity order. Recall the definition of Hölder regularity. The limit function $\varphi(x)$ is regular of order $r = N + \alpha$ $(0 < \alpha \le 1)$, $\varphi(x) \in \dot{C}^r$, if it is C^N and its Nth derivative $\varphi^{(N)}(x)$ is Lipschitz of order α, $\varphi^{(N)}(x) \in \dot{C}^\alpha$, as defined earlier by (8.1). Hölder spaces $\dot{C}^r$ generalize the spaces C^N of N-times continuously differentiable functions. As already mentioned in the case $N=1$, $\dot{C}^N$ contains functions that are not C^N, such as spline functions of degree N. In fact "$\varphi(x)$ is $\dot{C}^N$" can be thought of as "$\varphi(x)$ is almost C^N," since if $\varphi(x)$ is $\dot{C}^{N+\varepsilon}$, for some $\varepsilon > 0$, then $\varphi(x)$ is truly C^N. Other spaces, based on the Fourier transform of $\varphi(x)$, are sometimes used to define a regularity order $r \in R$ as well. They will be considered later in §17.

Using the results of the preceding sections, we can extend the characterization of Lipschitz limit functions $\dot{C}^\alpha$ $(0 < \alpha \le 1)$, derived in §8, to any Hölder regularity order $r > 0$.

THEOREM 10.1. *If $G(1) = 2$, $G(X)$ has at least $N+1$ zeros at $X=-1$ and*

$$\max_n |\Delta^N g^j_{n+1} - \Delta^N g^j_n| \le c\,2^{-j\alpha} \tag{10.1}$$

for some $\alpha > 0$, then $\varphi(x)$ is $\dot{C}^{N+\alpha}$. The converse is true whenever $\varphi(x)$ is stable. Moreover, (10.1) *implies $\alpha \le 1$ (if $\varphi(x) \not\equiv 0$), and*

$$\max_n |\Delta^N g^j_{n+1} - \Delta^N g^j_n| = \|(1-X)\Delta^N G^j(X)\|_\infty \tag{10.2}$$

can be replaced in (10.1) *by any of the following:*

$$\max_n |(g^j_N)_{n+1} - (g^j_N)_n| = \|(1-X)G^j_N(X)\|_\infty, \tag{10.3}$$

$$\max_n |(f_N^j)_n| = \|F_N^j(X)\|_\infty, \tag{10.4}$$

$$\max_{0\le n\le 2^j-1} \sum_k |(f_N^j)_{n+2^jk}|, \tag{10.5}$$

where we have set $G(X) = 2^{-N}(1+X)^N G_N(X) = 2^{-N}(1+X)^{N+1}F_N(X)$. *The iterated polynomials* $G_N^j(X)$, $F_N^j(X)$, *corresponding to the sequences* $(g_N^j)_n$, $(f_N^j)_n$, *are defined by* (2.5).

Proof. ($\Rightarrow$) Assume for the moment that $\alpha \le 1$. Since (10.1) implies, by Theorem 8.1, that $\Delta^N g_n^j$ converges uniformly to a $\dot{C}^\alpha$ function, it follows from Theorem 9.2 that all finite differences $\Delta^k g_n^j$ converge uniformly to $\varphi^{(k)}(x)$, for $k = 0, \cdots, N$. Hence $\varphi(x)$ is $\dot{C}^{N+\alpha}$.

($\Leftarrow$) If $\varphi(x)$ is stable and C^N, then by Theorem 9.2, $\Delta^N g_n^j$ converges uniformly to $\varphi^{(N)}(x) \in \dot{C}^\alpha$. Using (5.3) and the stability of $\varphi(x)$ we have $\|(1-X)\Delta^N G^j(X)\|_\infty \le c\,\|(1-X)\Delta^N\Phi^j(X)\|_\infty$, where $\Delta^N\Phi^j(X) = 2^{jN}(1-X)^N\Phi^j(X)$ corresponds to the coefficients $\Delta^N\varphi(n2^{-j})$. Now, we have

$$\begin{aligned}|\Delta^N\varphi(x) - \Delta^N\varphi(x-2^{-j})| &= 2^j\left|\int_{x-2^{-j}}^{x} (\Delta^{N-1}\varphi'(y) - \Delta^{N-1}\varphi'(y-2^{-j}))\,dy\right| \\ &\le \max_x |\Delta^{N-1}\varphi'(x) - \Delta^{N-1}\varphi'(x-2^{-j})|.\end{aligned}$$

By backward induction on N, it follows that

$$\|(1-X)\Delta^N\Phi^j(X)\|_\infty \le \max_x |\varphi^{(N)}(x) - \varphi^{(N)}(x-2^{-j})| \le c2^{-j\alpha},$$

which proves (10.1).

We now prove that (10.2)–(10.5) are "equivalent" in the following sense. Two sequences u_j and v_j are equivalent if there exist two constants c_1 and c_2, independent of j, such that $c_1v_j \le u_j \le c_2v_j$. From Lemma 9.1, we then have $\Delta^N G^j(X) = (1-X^{2^j})^N G_N^j(X)$. Hence, using the norm inequality (2.6), $\|(1-X)\Delta^N G^j(X)\|_\infty \le 2^N\|(1-X)G_N^j(X)\|_\infty$. Now, since the degree of $(1-X)G_N^j(X)$ is less than 2^jL, where L is the length of the sequence $(g_N)_n$, we also have

$$\begin{aligned}\|(1-X)G_N^j(X)\|_\infty &= \|(1-X^{2^jL})^N(1-X)G_N^j(X)\|_\infty \\ &= \|\left(\frac{1-X^{2^jL}}{1-X^{2^j}}\right)^N (1-X)\Delta^N G^j(X)\|_\infty \\ &\le c_N\,\|(1-X)\Delta^N G^j(X)\|_\infty.\end{aligned}$$

This proves that (10.2) and (10.3) are equivalent. The proof of (10.3)$\Leftrightarrow$(10.4) is very similar, based on the relation $(1-X)G_N^j(X) = (1-X^{2^j})F_N^j(X)$, which comes from Lemma 9.1. The equivalence (10.4)$\Leftrightarrow$(10.5) is obvious.

We finally show that (10.1) implies $\alpha \le 1$. Since $G(1) = 2$, we have $F_N(1) = F_N^j(1) = 1$; therefore, $\|F_N^j(X)\|_\infty \ge 2^{-j}\|F_N^j(X)\|_1 \ge 2^{-j}|F_N^j(1)| = 2^{-j}$, which shows, from (10.1) written with (10.4), that $\alpha \le 1$. □

The "equivalent" sequences (10.2)–(10.5) allow useful flexibility in the formulation of Theorem 10.1. As in §8, the following corollary immediately results from Theorem 10.1.

COROLLARY 10.2. *Let g_n be a stable binary subdivision scheme such that $G(1) = 2$ and $G(X)$ has at least $N+1$ zeros at $X = -1$. If, for $0 < \alpha < 1$,*

$$\max_n |\Delta^N g^j_{n+1} - \Delta^N g^j_n| \quad \text{decreases as } 2^{-j\alpha} \quad \text{when } j \to \infty, \tag{10.6}$$

then the limit function $\varphi(x)$ is $\dot{C}^{N+\alpha}$, but is not $\dot{C}^{N+\alpha+\varepsilon}$, for any $\varepsilon > 0$.

This does not hold for $\alpha = 1$, since by Theorem 10.1, (10.1) implies $\alpha \leq 1$. Of course, in (10.6) we can choose either (10.2), (10.3), (10.4), or (10.5).

Note that the characterization (10.1), or the criterion (10.6), depends on the choice of N. Theorem 10.1 (or Corollary 10.2) therefore allows us to check whether the exact regularity order r (that is, the number such that $\varphi(x)$ is $\dot{C}^r$ but not $\dot{C}^{r+\varepsilon}$, for any $\varepsilon > 0$) falls in the range $N \leq r < N+1$.

Assume, for example, that (10.6) is tested for some $N = N_0$ larger than the *unknown* exact regularity order r. This test necessarily fails, which only ensures that $\varphi(x)$ is not $\dot{C}^{N_0}$. On the other hand, if the value of N is too small, i.e., $N = N_1 < r-1$, then necessarily (10.6) is satisfied with $\alpha = 1$. This shows that $\varphi(x)$ is $\dot{C}^{N_1+1}$, but does not tell whether $\varphi(x)$ is actually more regular or not. In both cases (under or overestimated N's), the criterion (10.6) has to be checked all over again for other values of N to determine r. It is only when it turns out that $N < r \leq N+1$ that the criterion is really optimal and provides $N + \alpha = r$; therefore, the exact regularity order cannot be determined in general unless all possible values of N are tried.

However, if (10.1) can be extended to *negative* values of α, then the exact regularity order r is determined even if N is "too large," i.e., $N+1 \geq r$. That is, even if the criterion (10.1) for regularity order $r > N$ fails, it could be used to characterize lower regularity orders $0 < r \leq N$. In particular, if we use all of the zeros at $X = -1$ in $G(X)$ (i.e., if $G(X)$ has no more than $N+1$ such zeros), then the characterization (10.1), extended to any $\alpha \leq 1$, necessarily provides the exact regularity order r. This extension is provided by the following theorem.

THEOREM 10.3. *Theorem 10.1 and Corollary 10.2 hold for $-N < \alpha \leq 1$, with the following slight restriction. If (10.1) holds for $\alpha = -n$, $n = 0, 1, \cdots, N-1$, then $\varphi(x)$ is only "almost" $\dot{C}^{N-n}$, i.e., its $(N-n-1)$th derivative satisfies*

$$|\varphi^{(N-n-1)}(x+h) - \varphi^{(N-n-1)}(x)| \leq c.|h|| \log |h|| \quad \text{for all } x,\ h \in \mathbf{R}. \tag{10.7}$$

This theorem will be proven if we can simultaneously increase α and decrease N by 1 in (10.1). We, therefore, need the following lemma.

LEMMA 10.4. *Assume that $G(1) = 2$, $G(-1) = 0$, and that $G(X)$ has at least $N+1$ zeros at $X = -1$. The condition*

$$\max_n |\Delta^{N-1} g^j_{n+1} - \Delta^{N-1} g^j_n| \leq c\, 2^{-j(\alpha+1)} \tag{10.8}$$

implies (10.1). The converse implication holds for $\alpha < 0$ only. When $\alpha = 0$, (10.1) implies

$$\max_n |\Delta^{N-1} g^j_{n+1} - \Delta^{N-1} g^j_n| \leq c j 2^{-j}. \tag{10.9}$$

Proof. ($\Rightarrow$) We have

$$2^{-j} \max_n |\Delta^N g^j_n - \Delta^N g^j_{n-1}| = \max_n |\Delta^{N-1} g^j_n - 2\Delta^{N-1} g^j_{n-1} + \Delta^{N-1} g^j_{n-2}|$$

$$\leq \max_n(|\Delta^{N-1}g^j_n - \Delta^{N-1}g^j_{n-1}| \\ +|\Delta^{N-1}g^j_{n-1} - \Delta^{N-1}g^j_{n-2}|);$$

therefore, (10.8) clearly implies (10.1).

($\Leftarrow$) Condition (10.8) implies $1+\alpha \leq 1$ by Theorem 10.1. We, therefore, assume $\alpha \leq 0$ to prove the converse implication. Rewrite (10.1) and (10.8) using (10.4), knowing that $F^j_{N-1}(X) = 2^{-j}F^j_N(X)(1-X^{2^j})/(1-X)$ by Lemma 9.1. We, therefore, have to prove that (10.1), that is, $\|F^j_N(X)\|_\infty \leq c2^{-j\alpha}$ implies (10.8), that is $\|F^j_N(X)(1-X^{2^j})/(1-X)\|_\infty \leq c2^{-j\alpha}$. There is a problem at $X=1$; we, therefore, subtract $F^j_N(1) = F_N(1) = 1$ to $F^j_N(X)$ as shown:

$$\left\|F^j_N(X)\frac{1-X^{2^j}}{1-X}\right\|_\infty \leq \left\|(F^j_N(X)-1)\frac{1-X^{2^j}}{1-X}\right\|_\infty + \left\|\frac{1-X^{2^j}}{1-X}\right\|_\infty .$$

The second term in the right-hand side equals 1. Denote the first one by $\|H^j(X)\|_\infty$. From (2.3) written for $F_N(X)$, we have $F^j_N(X) - 1 = (F^{j-1}_N(X^2)-1) + (F_N(X) - 1)F^{j-1}_N(X^2)$. But since $F_N(1) = 1$, $X-1$ divides $F_N(X)-1$; therefore, $H^j(X) = H^{j-1}(X^2)(1+X) + (X^{2^j}-1)F^{j-1}_N(X^2)(F_N(X)-1)/(X-1)$ and

$$\|H^j(X)\|_\infty \leq \|H^{j-1}(X)\|_\infty + c2^{-(j-1)\alpha}.$$

By induction on j, for $\alpha < 0$, $\|H^j(X)\|_\infty \leq c'2^{-j\alpha}$ follows, which implies (10.8). When $\alpha = 0$, we have $\|H^j(X)\|_\infty \leq c'j$, which implies (10.9). □

Proof of Theorem 10.3. If α is not a negative integer, the generalization of Theorem 10.1 to $-N < \alpha \leq 0$ follows from several applications of Lemma 10.4. When $\alpha = -n$, $n = 0,\cdots,N-1$, by n successive applications of Lemma 10.4, (10.1) implies $\max_n |\Delta^{N-n}g^j_{n+1} - \Delta^{N-n}g^j_n| \leq c$. Applying Lemma 10.4 again, we only obtain $|\Delta^{N-n-1}g^j_{n+1} - \Delta^{N-n-1}g^j_n| \leq cj2^{-j}$. By Theorem 10.1, this implies that $\varphi(x)$ is $\dot{C}^{N-n-\varepsilon}$ (for any $\varepsilon > 0$), but we have a little more: mimicking the proof of the direct part of Theorem 8.1, we have $|\varphi^{(N-n-1)}(x+h) - \varphi^{(N-n-1)}(x)| \leq cj2^{-j}$ for $2^{-j} \leq |h| < 2^{-j+1}$, which gives (10.7). □

11. A practical, optimal Hölder regularity estimate.

Theorem 10.3 already provides an optimal regularity criterion (10.1) (with $-N < \alpha \leq 1$). However, it is not implementable on a computer as written since it needs to be verified for all j's and the order of magnitude of the constant c in (10.1) is unknown. The aim of this section is to transform this criterion into an easily implementable estimate [19] for Hölder regularity, computable with a finite number of operations.

The following theorem assumes some properties and notation we have already met:

- $G(1) = 2$;
- $G(X)$ has at least $N+1$ zeros at $X=-1$;
- $F_N(X)$ (corresponding to the sequence $(f_N)_n$) is, as defined in Theorem 10.1, $G(X)$ "without its $N+1$ zeros at $X=-1$," i.e.,

$$G(X) = \left(\frac{1+X}{2}\right)^N (1+X)\,F_N(X).$$

It generates iterated polynomials $F^j_N(X)$ and sequences $(f^j_N)_n$ by (2.5).

THEOREM 11.1. *With the above notation and assumptions, define the Hölder regularity estimate* $N+\alpha_N^j$ *by*

$$2^{-j\alpha_N^j}=\max_{0\le n\le 2^j-1}\sum_k |(f_N^j)_{n+2^jk}| \tag{11.1}$$

and let $\alpha_N=\sup_j \alpha_N^j$. *The sequence* α_N^j *converges to* $\alpha_N\le 1$ *as* $j\to\infty$. *If there exists* j *such that* $N+\alpha_N^j>0$, *then* $\varphi(x)$ *is* $\dot C^{N+\alpha_N^j}$ *(almost* $\dot C^{N+\alpha_N^j}$ *if* $\alpha_N^j\in -\mathbf{N}$, *see (10.7)); therefore,* $\varphi(x)$ *is* $\dot C^{N+\alpha_N-\varepsilon}$ *for any* $\varepsilon>0$.

In addition, if $\varphi(x)$ *is stable, then the regularity estimate is optimal: If* $\alpha_N\ne 1$, *or if* $\alpha_N=1$, *and* $G(X)$ *has no more than* $N+1$ *zeros at* $X=-1$, *then* $\varphi(x)$ *is* $\dot C^{N+\alpha_N-\varepsilon}$, *but is not* $\dot C^{N+\alpha_N+\varepsilon}$ *for any* $\varepsilon>0$. *Moreover, the rate of convergence of the estimates* $N+\alpha_N^j$ *to the exact regularity order* $N+\alpha_N$ *is given by*

$$0\le \alpha_N-\alpha_N^j\le c/j. \tag{11.2}$$

Proof. From (11.1) and Theorem 10.3 rewritten with (10.5), we have $\alpha_N^j\le 1$ for all j; hence $\alpha_N\le 1$. Now, using the relation $F_N^{i+j}(X)=F_N^j(X)F_N^j(X^{2^j})$ or the matrix formulation (12.1) given in the next section, we easily find that $2^{-(i+j)\alpha_N^{i+j}}\le 2^{-i\alpha_N^i}2^{-j\alpha_N^j}$, i.e.,

$$\alpha_N^{i+j}\ge\frac{i\alpha_N^i+j\alpha_N^j}{i+j}.$$

The following proof of convergence of the α_N^j is due to Cohen [3], [4]: Let $\varepsilon>0$ be an arbitrary small number and J such that $\alpha_N^J\ge\alpha^N-\varepsilon$. For any j, write $j=nJ+i$, $0\le i\le J-1$. Applying the inequality above several times, we find $\alpha_N^j\ge((j-i)\alpha_N^J+i\alpha_N^i)/j$; hence, when j is large enough, $\alpha_N^j\ge\alpha_N-2\varepsilon$, which proves that $\alpha_N^j\to\alpha_N$ as $j\to\infty$.

We now prove the announced regularity order for $\varphi(x)$. Let $G_N(X)=(1+X)F_N(X)$ be as in Theorem 10.1. By Lemma 7.2 applied to $G_N(X)$, we immediately obtain (10.1), written with (10.3) and $\alpha=\alpha_N^i$; therefore, Theorem 10.3 applies with $\alpha=\alpha_N^i$, for any i such that $\alpha_N^i>-N$. The limit function is thus $\dot C^{N+\alpha_N^i}$ (with the restriction (10.7)), and, therefore, $\varphi(x)$ is $\dot C^{N+\alpha_N-\varepsilon}$ for any $\varepsilon>0$.

Now assume that $\varphi(x)$ is stable. From (11.1), the condition (10.1), rewritten with (10.5) is satisfied only when $\alpha\le\liminf_{j\to\infty}\alpha_N^j=\alpha_N$. Now if $\varphi(x)$ were $\dot C^{N+\alpha_N+\varepsilon}$, where $\alpha_N<1$ and $\varepsilon>0$, by Theorem 10.1 (10.1) would hold with $\alpha=\alpha_N+\varepsilon$, which contradicts $\alpha\le\alpha_N$; therefore, if $\varphi(x)$ is stable, $\alpha_N<1$ implies that $\varphi(x)$ is not $\dot C^{N+\alpha_N+\varepsilon}$ for any $\varepsilon>0$. In addition, $\varphi(x)$ cannot be $N+1+\varepsilon$ if $G(X)$ has no more than $N+1$ zeros at $X=-1$ because of Theorem 9.2.

We finally prove (11.2). When $\varphi(x)$ is stable and $\dot C^{N+\alpha_N-\varepsilon}$, by Theorem 10.3, (10.1), written with (10.5), holds for $\alpha=\alpha_N-\varepsilon$. By definition of α_N^j (11.1), we thus have $2^{-j\alpha_N^j}\le c2^{-j(\alpha_N-\varepsilon)}$ for any $\varepsilon>0$, which implies (11.2). □

Of course, we can replace (10.5) in (11.1) by any other equivalent sequence (10.2), (10.3), (10.4). We would still obtain a sequence $N+\beta_N^j$, which converges to the optimal regularity order $N+\alpha_N$, however, $\varphi(x)$ may not be regular of order $N+\beta_N^j$ for any fixed j because β_N^j may be greater than α_N.

Let us make precise the practical outcomes of Theorem 11.1. For a given number of iterations j, and a given N, the computation of $N+\alpha_N^j$—with a finite number of

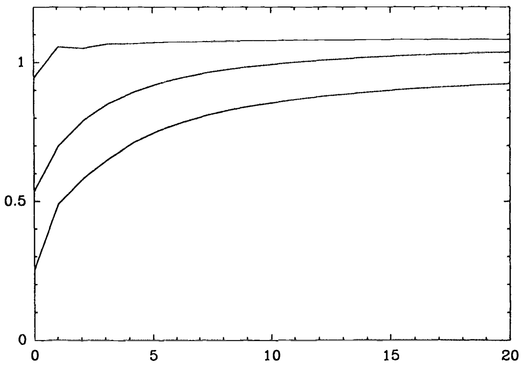


FIG. 4. *Program output of regularity estimates* $N+\alpha_N^j$ (11.1) ($N=0,\ 1,\ 2$) *for* $j=1$ *to* 20 *iterations. The corresponding limit function is the Daubechies "minimum phase" wavelet of length* 5 (*see* §14), *whose exact regularity order is* $r=1.0878\cdots$. *For* $N=0$, *the estimate is bounded by* 1 *and, therefore, does not converge to* r. *For* $N=2$, *the estimate converges fairly rapidly to* r. *After* 20 *iterations we find* $2+a_2^{20}=1.0831\cdots$.

operations—by (11.1) gives a Hölder regularity estimate for $\varphi(x)$ in all cases. Since $\lim_j \alpha_N^j = \sup_j \alpha_N^j$, the estimate is improved when the number of iterations j increases.

Figure 4 shows that N must be chosen large enough because the estimate $N+\alpha_N^j$ is bounded by $N+1$, whereas the exact regularity order of $\varphi(x)$ might be greater than $N+1$. If N is too small, $N+\alpha_N^j$, in fact, necessarily converges to $N+1$. It is therefore recommended that N should be chosen maximal (i.e., such that $G(X)$ has exactly $N+1$ zeros at $X=-1$). In this case Theorem 11.1 ensures that the regularity estimates $N+\alpha_N^j$ converge to $N+\alpha_N$, which, provided that $\varphi(x)$ is stable, gives the exact regularity order of $\varphi(x)$. In practice, Fig. 4 shows that the convergence rate of the estimates α_N^j is fairly high. When the scaling sequence length L is not too large (e.g., $L\le 10$), the exact regularity order is numerically estimated to two decimal places after a few dozen iterations. However, it can be shown [19] that the computational load of an implementation of (11.1) is increasing exponentially with j (increasing j by one roughly doubles the number of operations required to compute (11.1)).

Note that from Theorem 9.2, finite differences $\Delta^k g_n^j$ converge uniformly to the derivatives of a stable limit function $\varphi(x)$ whenever these derivatives exist.

Theorem 11.1 is the main result of this paper. It permits us to estimate sharply Hölder regularity in most cases of interest. (See §9 for the derivation of the optimal regularity estimate on a particular example of an unstable limit function.) The remainder of this paper connects this result to related work on regularity estimates, and illustrates it with examples.

12. Relation to Daubechies and Lagarias estimates. In a recent paper [9], Daubechies and Lagarias determined sharp conditions for Hölder regularity based on matrix products. Although the approach in [9] relies on two-scale difference equations (5.1) rather than on limit functions (3.3), the above results, which were derived independently, are closely related to what can be found in [9]. In fact, (11.1) reads, in matrix notation,

$$2^{-j\alpha_N^j} = \max_{\varepsilon_i=0 \text{ or } 1} \left\| \prod_{i=0}^{j-1} \mathbf{F}_N^{\varepsilon_i} \right\|_1 , \tag{12.1}$$

where the matrices F_N^0 and F_N^1 of size $(L-1)\times(L-1)$ (where L is the length of the sequence $(f_N)_n$) are defined as

$$\mathbf{F}_N^0 = \begin{pmatrix} (f_N)_0 & 0 & 0 & 0 & \cdots \\ (f_N)_2 & (f_N)_1 & (f_N)_0 & 0 & \cdots \\ (f_N)_4 & (f_N)_3 & (f_N)_2 & (f_N)_1 & \cdots \\ (f_N)_6 & (f_N)_5 & (f_N)_4 & (f_N)_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{12.2}$$

$$\mathbf{F}_N^1 = \begin{pmatrix} (f_N)_1 & (f_N)_0 & 0 & 0 & \cdots \\ (f_N)_3 & (f_N)_2 & (f_N)_1 & (f_N)_0 & \cdots \\ (f_N)_5 & (f_N)_4 & (f_N)_3 & (f_N)_2 & \cdots \\ (f_N)_7 & (f_N)_6 & (f_N)_5 & (f_N)_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{12.3}$$

and $\|\mathbf{A}\|_1$ denotes the l^1-norm of a square matrix $\mathbf{A} = ((a_{i,j}))$, i.e.,

$$\|\mathbf{A}\|_1 = \max_i \sum_j |a_{i,j}|.$$

Formulation (12.1) can be proved as follows. Consider the operators of polynomial "biphase decomposition [22]" $\mathbf{D}^\varepsilon$ ($\varepsilon = 0$ or 1), defined by the relation $U(X) = U^0(X^2) + XU^1(X^2)$, where $U^\varepsilon(X) = \mathbf{D}^\varepsilon\{U(X)\}$. Clearly $\mathbf{F}_N^\varepsilon$, seen as an operator acting on polynomials of degree $\leq L-2$, transforms $U(X)$ into $\mathbf{D}^\varepsilon\{F_N(X)U(X)\}$. Applying j times the identity $\mathbf{D}^\varepsilon\{U(X)V(X^2)\} = \mathbf{D}^\varepsilon\{U(X)\}V(X)$ gives the polynomial associated to the sequence $(f_N^j)_{n+2^jk}$ (where $n = \varepsilon_0\varepsilon_1\cdots\varepsilon_{j-1}$ in base 2) as

$$\left(\prod_{i=0}^{j-1}\mathbf{D}^{\varepsilon_i}\right)\{F_N^j(X)\} = \prod_{i=0}^{j-1}(\mathbf{F}^{\varepsilon_i}\{1\}),$$

where the polynomial 1 corresponds to the vector $(1\ 0\ 0\ \cdots\ 0)^t$. Therefore, $(f_N^j)_{n+2^jk}$, seen as a vector indexed by k, is equal to the first column of the matrix product in (12.1). To obtain the other columns, replace the initial polynomial 1 by X^m. This amounts to shifting the value of n in $(f_N^j)_{n+2^jk}$, hence changing the values of the ε_i. But since (11.1) involves the maximum over the values of n, the l^1-norm of the first column of the matrix product can be replaced by the l^1-norm of the whole matrix product, which gives (12.1).

Using (12.1) in place of (11.1) in Theorem 11.1, we easily recover the results on global Hölder regularity derived in [9]. Formulation (12.1) and that used in [9] differ only by some minor details: Daubechies and Lagarias use l^2-norms rather than l^1-norms, and the matrices they consider are a bit larger than (12.2), (12.3) because they correspond to $G(X) = 2^{-N}(1+X)^{N+1}F_N(X)$ rather than $F_N(X)$. Although regularity estimates are not proved to be optimal in general in [9], Daubechies and Lagarias prove optimality for several examples, such as those of §14.

Working with matrices is useful when we want to find optimal regularity estimates "by hand" [9], without implementing (11.1). Unfortunately, it seems difficult to derive a general method for determining the optimal regularity by matrix manipulation. As a result, unlike an implementation of (11.1) on a computer, each example has to be investigated separately and requires fastidious treatment. We here recall for completeness the basic method used in [9].

THEOREM 12.1 (Daubechies and Lagarias [9]). *The following method often provides a sharp Hölder regularity estimate for a limit function* $\varphi(x)$:

- *Compute the eigenvalues of* $\mathbf{F}_N^0$ *and* $\mathbf{F}_N^1$ *and let* ρ^0, ρ^1 *be their respective spectral radii (largest moduli of eigenvalues).*
- *Assume, for example, that* $\rho^0 > \rho^1$. *Compute matrix* $\mathbf{B}$, *whose columns are proportional to the eigenvectors of* $\mathbf{F}_N^0$. *The norm of the diagonal matrix* $\mathbf{B}^{-1}\mathbf{F}_N^0\mathbf{B}$ *is therefore* ρ^0.
- *Parameterize* $\mathbf{B}$ *by* $L-1$ *numbers, one for each column. If we can find a parameterization of* $\mathbf{B}$ *such that*

$$\|\mathbf{B}^{-1}\mathbf{F}_N^1\mathbf{B}\| \leq \rho^0, \tag{12.4}$$

where $\|\cdot\|$ *is any matrix norm, then* $\varphi(x)$ *is regular of order* $N - \log_2 \rho^0$ *(and this Hölder regularity estimate is moreover optimal if* $\varphi(x)$ *is stable).*

Proof. First, specifying $\varepsilon_i = 0$ for all i in (12.1) gives $2^{-j\alpha_N^j} \geq \|(\mathbf{F}_N^0)^j\|$. Let λ be an eigenvalue of $\mathbf{F}_N^0$ and v an associated nonzero eigenvector. We have, on one hand, $\|(\mathbf{F}_N^0)^j v\| \leq \|(\mathbf{F}_N^0)^j\| \cdot \|v\|$, and on the other hand, $\|(\mathbf{F}_N^0)^j v\| = |\lambda|^j \|v\|$. It follows that $(\rho^0)^j = \sup |\lambda|^j \leq \|(\mathbf{F}_N^0)^j\| \leq 2^{-j\alpha_N^j}$. Now, with the change of basis $\mathbf{B}$, we have

$$2^{-j\alpha_N^j} = \max_{\varepsilon_i} \left\| \mathbf{B} \left(\prod_{i=0}^{j-1} \mathbf{B}^{-1}\mathbf{F}_N^{\varepsilon_i}\mathbf{B} \right) \mathbf{B}^{-1} \right\| \leq c \max_{\varepsilon_i} \prod_{i=0}^{j-1} \|\mathbf{B}^{-1}\mathbf{F}_N^{\varepsilon_i}\mathbf{B}\| .$$

But we have $\|\mathbf{B}^{-1}\mathbf{F}_N^0\mathbf{B}\| = \rho^0$ and (12.4); therefore, $2^{-j\alpha_N^j} \leq c\,(\rho^0)^j$ follows. We, therefore, have proved that $(\rho^0)^j \leq 2^{-j\alpha_N^j} \leq c\,(\rho^0)^j$, which implies $\alpha^N = \lim_j \alpha_N^j = -\log_2 \rho^0$. The theorem therefore follows from Theorem 11.1. □

Note that this method is only optimal if (12.4) is met for at least one matrix norm, otherwise the obtained estimate, $N - \log_2 \|\mathbf{B}^{-1}\mathbf{F}_N^1\mathbf{B}\|$, is suboptimal. Whether (12.4) holds for a large class of masks g_n is an open problem [9].

13. A sharp upper bound for regularity. So far we have seen two types of Hölder regularity estimates: One is optimal in (almost) all cases (§11), but many iterations, performed on a computer, are necessary to determine the regularity order accurately. The other (§12) requires the calculation of two spectral radii of matrices, but is sometimes suboptimal. Based on the latter, it is nevertheless possible to obtain a (possibly sharp) *upper bound* for regularity of stable limit functions that only requires the computation of one spectral radius and gives the exact regularity order whenever condition (12.4) is satisfied:

Specifying $\varepsilon_i = 0$ or $\varepsilon_i = 1$ for all i in (12.1) gives $2^{-j\alpha_N^j} \geq \max(\|(\mathbf{F}_N^0)^j\|, \|(\mathbf{F}_N^1)^j\|)$. We have seen that this is greater than $\max((\rho^0)^j, (\rho^1)^j)$; therefore, an upper bound for the Hölder regularity is $N - \log_2 \max(\rho^0, \rho^1)$. By Theorem 12.1, this upper bound is attained for stable limit functions if (12.4) holds.

The computation of this upper bound can be simplified to the search of the spectral radius of only one matrix $\mathbf{F}_N$, defined as the common submatrix of $\mathbf{F}_N^0$ and $\mathbf{F}_N^1$.

$$\mathbf{F}_N^0 = \left(\begin{array}{c|c} (f_N)_0 & 0 \;\cdots\; 0 \\ \hline (f_N)_2 & \\ (f_N)_4 & \mathbf{F}_N \\ \vdots & \end{array} \right) \quad \text{and} \quad \mathbf{F}_N^1 = \left(\begin{array}{c|c} & \vdots \\ \mathbf{F}_N & (f_N)_{L-3} \\ & (f_N)_{L-2} \\ \hline 0 \;\cdots\; 0 & (f_N)_{L-1} \end{array} \right).$$

We have

$$N - \log_2 \max(\rho^0, \rho^1) = N - \log_2 \max(|(f_N)_0|, |(f_N)_{L-1}|, \rho(\mathbf{F}_N)), \tag{13.1}$$

where $\rho(\mathbf{F}_N)$ is the spectral radius of $\mathbf{F}_N$. Therefore the regularity order of a stable limit function is at most (13.1).

A similar upper bound can be computed using the inequality

$$2^{-j\alpha_N^j} \geq \max\left(\sum_k |(f_N^j)_{2^j k}|, \sum_k |(f_N^j)_{2^j k-1}|\right),$$

which yields a fast implementation [19]: the computational load is here linear in j (compare with §11). When $j \to \infty$, this gives an upper bound which may be greater than (13.1) but is still sharp. This result and Theorem 11.1 can be used to compute sharp lower and upper bounds for the Hölder regularity of $\varphi(x)$. Table 1 provides values of these bounds for the examples presented in the next section.

14. Examples: Daubechies orthonormal wavelets. A family of orthonormal wavelets with compact support has been constructed by Daubechies in [6]. The construction is based on binary subdivision schemes. The "mother wavelet" is defined as the limit function $\psi(x)$ of the scheme (1.2) with initial sequence $h_n = (-1)^n g_{L-1-n}$ (where L is the mask length). She showed that the regular functions $2^{-j/2}\psi(2^{-j}x - k)$, defined for all integers j and k, form an orthonormal basis of $\mathbf{L}^2(\mathbf{R})$ if L is even and

$$G(X)\tilde{G}(X) - G(-X)\tilde{G}(-X) = 4X^{L-1}, \tag{14.1}$$

where $\tilde{G}(X)$ is the polynomial associated to the sequence g_{L-1-n}. In [6], $G(X)$ is, moreover, required to have as many zeros at $X = -1$ as possible. This results in several possible solutions for $G(X)$ that have exactly $N + 1 = L/2$ zeros at $X = -1$ [6], [7].

Examples of $G(X)$ in [6] have all zeros outside the unit circle ("minimum phase" choice in the signal processing terminology, since X corresponds to a delay). In [9], the optimal regularities of "minimum phase" Daubechies wavelets $\psi(x)$ for $L = 4$, 6, and 8 are obtained using the method described in the preceding section. It turns out that (12.4) holds for these lengths; therefore, the upper bound (13.1) is attained and actually equals $N - \log_2 |(f_N)_0|$. The estimated regularity of Daubechies "minimum phase" wavelets derived in [6] is, therefore, $-\log_2 |g_0|$ in this case. It can easily be checked that the convergent binary subdivision schemes involved are stable; hence this estimate is optimal. This can be checked directly [9] from Theorem 10.3 by noting that the first "slope" of $\Delta^N g_n^j$ is $|2g_0|^j = 2^{j(1-\alpha)}$, where $\alpha = -\log_2 |g_0|$. Table 1 lists these optimal regularities (for $L \leq 8$), the corresponding outputs of a program implementing (11.1), and upper bounds derived in §13.

There are other solutions g_n, derived for $L \geq 8$ in [7], which, unlike "minimum phase solutions," are close to being symmetric. Table 1 shows that the regularity estimates for these wavelets, based on Theorem 11.1, are lower than those of "minimum phase" wavelets. This will be justified in §17.

15. "Strictly linear phase" symmetric limit functions. In this section we apply the above results to a subclass of scaling sequences that is often encountered. This section is also a prerequisite for comparing Hölder regularity estimates to those determined using Fourier techniques (§17). The subclass considered here consists of

TABLE 1

Some regularity estimates for two types of Daubechies orthonormal wavelets: Minimum phase wavelets [6] *and "more symmetric" ones* [7] (*for mask lengths* $L \geq 8$). *The upper bound fcr Hölder regularity in the right-most column is obtained by adding* $\frac{1}{2}$ *to the optimal Sobolev regularity estimate, derived in* [6, *Appendix*] (*see* §17). *These two apply to all Daubechies wavelets that differ only by their phase. The numbers* r_{20} *are the Hölder regularity estimates* (11.1) *obtained by computer program after* $j = 20$ *iterations. Note that more symmetry decreases regularity in general. For minimum phase wavelets, these estimates converge rapidly to the optimal Hölder regularity estimates* r_∞ *derived in* [9] *by using the method described in* §12. *The upper bounds for both types of wavelets are obtained from* §13. *They are sharper than the "Sobolev" upper bound and in fact give optimal Hölder regularity estimates in the "minimum phase" case for lengths* $L \leq 8$.

L	Optimal Sobolev regularity estimate	More symmetric wavelets		Minimum phase wavelets			Upper bound
		r_{20}	Upper bound	r_{20}	r_∞	Upper bound	
4	0.4999	—	—	0.5500	0.5500	0.5500	0.9999
6	0.9150	—	—	1.0831	1.0878	1.0878	1.4150
8	1.2755	1.3960	1.4026	1.6066	1.6179	1.6179	1.7755
10	1.5967	1.7621	1.7759	1.9424	—	1.9689	2.0967
12	1.8883	2.1019	2.1223	2.1637	—	2.1891	2.3883
14	2.1586	2.4420	2.4681	2.4348	—	2.4604	2.6586
16	2.4147	2.7155	2.7500	2.7358	—	2.7608	2.9147
18	2.6616	2.9977	3.0393	3.0432	—	3.0736	3.1616
20	2.9027	3.2651	3.3110	3.3098	—	3.3813	3.4027

scaling sequences for which either $G(X)$ or $G(X)/(1+X)$ is "strictly linear phase," in the following sense.

DEFINITION 15.1. *A polynomial* $U(X)$ (*or its associated sequence* u_n *of finite length* L) *is* strictly linear phase *if it is symmetric,* $u_n = u_{L-1-n}$, *and if the trigonometric polynomial* $U(e^{i\omega})e^{-i(L-1)\omega/2}$ *does not change sign for any* $w \in R$.

Note that symmetry of u_n implies $U(e^{i\omega})e^{-i(L-1)\omega/2} \in \mathbf{R}$. This condition is called "linear phase" in signal processing [22]. The above definition requires more, namely that no discontinuities of the phase due to a change of sign in $U(e^{i\omega})e^{-i(L-1)\omega/2}$ occur. Therefore, complex zeros of the symmetric polynomial U(X) occur in pairs $(z, 1/\bar{z})$ not only for $|z| \neq 1$, but also *on* the unit circle. That is, roots on the unit circle have even order. It follows that $U(X)$ has an even number of roots, hence L is odd.

If $G(X)$ or $G(X)/(1+X)$ is strictly linear phase, then for N odd (even, respectively), the sequence $(f_N)_n$ in (11.1) is also strictly linear phase. The following theorem shows that in this case, the determination of the exact regularity order of a (stable) limit function $\varphi(x)$ only requires the computation of the spectral radius of one matrix. This is to be compared with §§12 and 13, where it is shown that two matrices are involved in the general case, and the computation of one matrix's spectral radius only provides an upper bound for regularity.

The following regularity estimate has been derived independently, by other means, and on particular examples of strictly linear phase scaling sequences, in [6] and [10] (see §§16 and 17).

THEOREM 15.2. *Assume* $G(1) = 2$, $G(X)$ *has at least* $N+1$ *zeros at* $X = -1$: $G(X) = 2^{-N}(1+X)^{N+1}F_N(X)$, *and* $F_N(X)$ *is strictly linear phase. Define* $(\hat{f}_N)_n = (f_N)_{((L-1)/2)\pm n}$ (*where* L *is the length of* $(f_N)_n$) *and the* $(L-1)/2 \times (L-1)/2$ *matrix*

$\hat{\mathbf{F}}_N$ obtained by "folding" the following $(L-1)/2 \times (L-2)$ matrix

$$\begin{pmatrix} \cdots & (\hat{f}_N)_3 & (\hat{f}_N)_2 & (\hat{f}_N)_1 & (\hat{f}_N)_0 & (\hat{f}_N)_1 & (\hat{f}_N)_2 & (\hat{f}_N)_3 & \cdots \\ \cdots & (\hat{f}_N)_1 & (\hat{f}_N)_0 & (\hat{f}_N)_1 & (\hat{f}_N)_2 & (\hat{f}_N)_3 & (\hat{f}_N)_4 & (\hat{f}_N)_5 & \cdots \\ \cdots & (\hat{f}_N)_1 & (\hat{f}_N)_2 & (\hat{f}_N)_3 & (\hat{f}_N)_4 & (\hat{f}_N)_5 & (\hat{f}_N)_6 & (\hat{f}_N)_7 & \cdots \\ & & & & \vdots & & & & \end{pmatrix}$$

around its middle column, i.e.,

$$\hat{\mathbf{F}}_N = \begin{pmatrix} (\hat{f}_N)_0 & 2(\hat{f}_N)_1 & 2(\hat{f}_N)_2 & \cdots \\ (\hat{f}_N)_2 & (\hat{f}_N)_1 + (\hat{f}_N)_3 & (\hat{f}_N)_0 + (\hat{f}_N)_4 & \cdots \\ (\hat{f}_N)_4 & (\hat{f}_N)_3 + (\hat{f}_N)_5 & (\hat{f}_N)_2 + (\hat{f}_N)_6 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{15.1}$$

Let ρ be its spectral radius. One has $\rho \geq \frac{1}{2}$. If $\rho < 2^N$, then the limit function $\varphi(x)$ is $\dot{C}^{N-\log_2 \rho}$ (almost $\dot{C}^{N-\log_2 \rho}$ in the sense of (10.7) if $\rho \geq 1$ is an integer power of two). In addition, if $\varphi(x)$ is stable, and if either $\rho > \frac{1}{2}$ or $\rho = \frac{1}{2}$ and $G(X)$ has no more than $N+1$ zeros at $X = -1$, then the estimate is optimal: $\varphi(x)$ is not $\dot{C}^{N-\log_2 \rho+\varepsilon}$ for any $\varepsilon > 0$.

Proof. Define $(\hat{f}^j_N)_n = (f^j_N)_{(2^{j-1}-2^{-1})(L-1)+n}$. This noncausal, symmetric sequence is strictly linear phase. We first prove that $\|F^j_N(X)\|_\infty = \max_n |(\hat{f}^j_N)_n| = |(\hat{f}^j_N)_0|$. Using Fourier coefficients, we have $(\hat{f}^j_N)_n = \frac{1}{2\pi}\int_0^{2\pi} \hat{F}^j_N(e^{i\omega})e^{in\omega}\, d\omega$, where $\hat{F}^j_N(e^{i\omega}) = \sum_n (\hat{f}^j_N)_n e^{in\omega} = \pm|\hat{F}^j_N(e^{i\omega})|$. Hence

$$\max_n |(\hat{f}^j_N)_n| \leq \frac{1}{2\pi}\int_0^{2\pi} |\hat{F}^j_N(e^{i\omega})|\, d\omega = |(\hat{f}^j_N)_0|.$$

The theorem, therefore, results from Theorem 10.3 if we prove that $|(\hat{f}^j_N)_0|$ is equivalent to ρ^j as $j \to \infty$. From (2.4) written for $F_N(X)$, we have, for$0 \leq m \leq 2^j - 1$, $(f^{j+1}_N)_{2^{j+1}n+m+2^j} = \sum_k (f_N)_{k+1}(f^j_N)_{2^j(2n-k)+m}$. This means, in matrix notation,

$$((f^{j+1}_N)_{2^{j+1}n+m+2^j})_n = \mathbf{F}^1_N((f^j_N)_{2^j n+m})_n,$$

where $\mathbf{F}^1_N$ is defined by (12.3). Let $m = 2^j - (L-1)/2$ (for sufficiently large j's to ensure $m \geq 0$). The above equation is then rewritten, in terms of the $(\hat{f}^j_N)_n$, as $((\hat{f}^{j+1}_N)_{2^{j+1}(n-(L-3)/2)})_n = \mathbf{F}^1_N((\hat{f}^j_N)_{2^j(n-(L-3)/2)})_n$ By symmetry, this equation can be restricted to $n = 0, \cdots, (L-3)/2$, in which case the action of $\mathbf{F}^1_N$ is exactly that of $\hat{\mathbf{F}}_N$. It follows by induction on j that $|(\hat{f}^j_N)_0|$ is equivalent to $\|(\hat{\mathbf{F}}_N)^j\|_\infty$, hence to ρ^j, when $j \to \infty$. □

16. Examples: Deslauriers and Dubuc interpolatory schemes. Deslauriers and Dubuc [10]–[12] studied the regularity of limit functions of a special family of interpolatory subdivision schemes based on Lagrangian interpolation. Recall that for interpolatory schemes the iterated points g^j_n are carried unchanged at each iteration. Here, we simply insert between $g^j_n = g^{j+1}_{2n}$ and $g^j_{n+1} = g^{j+1}_{2n+2}$ the value g^{j+1}_{2n+1} of the Lagrangian polynomial interpolation of the K consecutive values $g^j_{n+1-K/2}, \cdots,$ $g^j_n, g^j_{n+1}, \cdots, g^j_{n+K/2}$, where K is even. This corresponds to a mask g_n of length

TABLE 2

Optimal Hölder regularity estimates r of interpolatory subdivision schemes of Deslauriers and Dubuc [10], [11], [12] *for several Lagrangian interpolation orders K corresponding to mask lengths $L = 2K - 1$ (see* §16). *These estimates are also optimal in the "Fourier sense," and the numbers $(r-1)/2$ give the optimal Sobolev regularity estimates listed in Table* 1 *(see* §17).

K	L	r	K	L	r
2	3	1.	12	23	4.7767
4	7	2.	14	27	5.3173
6	11	2.8300	16	31	5.8294
8	15	3.5511	18	35	6.3233
10	19	4.1935	20	39	6.8054

$L = 2K - 1$, which reads (when made causal by shifting)

$$\left\{ \begin{array}{l} g_{2n} = \delta_{n-K/2}, \\ g_{2n+1} = L_n(\frac{K-1}{2}), \end{array} \right. \tag{16.1}$$

where $L_n(X)$ is the Lagrangian polynomial $L_n(X) = \prod_{k\neq n}(X-k)/(n-k)$ associated to the interpolation points $k = 0, \cdots, K-1$.

Shensa has shown [21] that $G(X)$ of length $L = 2K - 1$ is exactly $G(X) = G_W(X)\tilde{G}_W(X)$, where $G_W(X)$ is the polynomial mask of Daubechies wavelets of compact support $[0,\ K-1]$ (see §14—this fact will be useful in §17). From §14 it follows that $G(X)$ has exactly K zeros at $X = -1$. Moreover, it is strictly linear phase because $G(e^{i\omega}) = G_W(e^{i\omega})\tilde{G}_W(e^{i\omega}) = |G_W(e^{i\omega})|^2 e^{i(K-1)\omega}$. Thus Theorem 15.2 applies with $N = K - 1$. Moreover, since all interpolatory subdivision schemes are stable (§6), Theorem 15.2 will provide the *exact* regularity order of $\varphi(x)$.

The matrices $\hat{\mathbf{F}}^{K-1}$ (15.1) needed by Theorem 15.2 can be easily determined using the formula

$$(f_{K-1})_n = c \left(\begin{array}{c} K-2 \\ n \end{array} \right)^{-1} \sum_{i=0}^{n} (-1)^i \left(\begin{array}{c} K-1 \\ i \end{array} \right)^2, \qquad n = 0, \cdots, K-2,$$

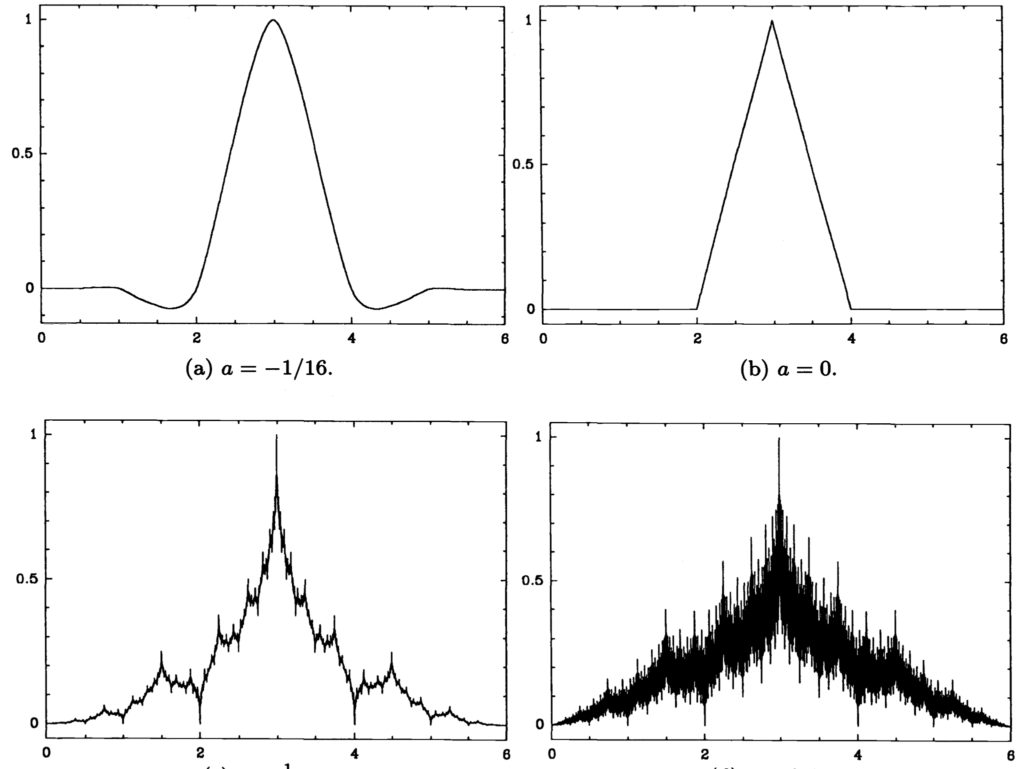
which results from (16.1) after some calculation. Determination of their spectral radii yields to the optimal regularities listed in Table 2. For $L = 7$ (i.e., $K = 4$), using 4 zeros at $X = -1$ in $G(X)$, we find that the limit function is almost $\dot{C}^2$ in the sense of (10.7), which was first proven by Dubuc in [12]. However, when only 2 zeros at $X = -1$ in $G(X)$ are used ($N = 1$), we find that the spectral radius of Theorem 15.2 is $\rho = \frac{1}{2}$, hence the limit function is in fact $\dot{C}^2$.

In [10], Deslauriers and Dubuc extended the study of the previous subdivision scheme for $L = 7$ (i.e., $K = 4$) to the following interpolatory mask (here defined for $n = -3, \cdots, 3$):

$$g_0 = 1, \quad g_{\pm 1} = 1/2 - a, \quad g_{\pm 3} = a, \quad g_n = 0 \quad \text{elsewhere},$$

where $a \in R$. The case $a = -1/16$ corresponds to the previous example, for which the limit function is $\dot{C}^2$.

The simplicity and usefulness of Theorem 15.2 is well illustrated through this example. The mask g_n is easily seen to be strictly linear phase for $-1/16 \leq a \leq \frac{1}{2}$; therefore, Theorem 15.2 applies in this case. (For other values of a we have to use more general theorems such as Theorem 11.1.) Now, for $a \neq -1/16$, $G(X)$ has exactly



(a) $a = -1/16$.

(b) $a = 0$.

(c) $a = \frac{1}{4}$.

(d) $a = 0.4$.

FIG. 5. *Plots of Deslauriers and Dubuc limit functions corresponding to* $g_0 = 1$, $g_{\pm 1} = 0.5 - a$, $g_{\pm 3} = a$, *and* $g_n = 0$ *elsewhere. The successive values of* a *are* $a = -1/16$ *(regularity order* 2*),* $a = 0$ *(regularity order* 1*),* $a = \frac{1}{4}$ *(regularity order* $\log_2(\sqrt{5} - 1) = 0.305\cdots$*) and* $a = 0.4$ *(regularity order* $0.104\cdots$*).*

two zeros at $X = -1$, and we can, therefore, apply Theorem 15.2 with $N = 1$. We have $(\hat{f}_1)_0 = 1 + 4a$, $(\hat{f}_1)_{\pm 1} = -4a$, and $(\hat{f}_1)_{\pm 2} = 2a$; hence

$$\hat{\mathbf{F}}_1 = \begin{pmatrix} 1 + 4a & -8a \\ 2a & -4a \end{pmatrix}.$$

Its spectral radius is $\rho = (1 + \sqrt{1 + 16a})/2$. From Theorem 15.2, the exact regularity order of $\varphi(x)$ is $r = 2 - \log_2(1 + \sqrt{1 + 16a})$, which decreases from 2 to zero when a increases from $-1/16$ to $\frac{1}{2}$. Figure 5 illustrates this through several examples corresponding to various values of a.

17. Comparison with Fourier-based regularity estimates. This paper has developed a direct approach based on the definition of Hölder regularity. But several other approaches for estimating regularity based on the Fourier transform $\hat{\varphi}(\omega)$ of the (compactly supported) limit function $\varphi(x)$ have also been considered [3]–[5], [10], [11], [23]. Note that we have easy access to $\hat{\varphi}(\omega)$ from mask g_n by [3]–[6]

$$\hat{\varphi}(\omega) = \lim_{j \to \infty} G^j(e^{i\omega}). \tag{17.1}$$

The idea is here to estimate the decay of $\hat{\varphi}(\omega)$ as $|\omega| \to \infty$. To do this, several functional spaces (other than $\dot{C}^r$) can be used to interpolate the spaces C^N of N-times continuously differentiable functions. We generally consider one of the following spaces: $\mathbf{H}_1^r$, $\mathbf{H}_2^r$, $\mathbf{H}_\infty^r$, defined by the conditions $|\omega|^r\hat{\varphi}(\omega) \in \mathbf{L}^1, \mathbf{L}^2, \mathbf{L}^\infty$, respectively. (The spaces $\mathbf{H}_2^r$ are the Sobolev spaces of order r.) Estimations of the parameter r for these spaces ensure some Hölder regularity, since we have, for any $\varepsilon > 0$,

$$\mathbf{H}_\infty^{r+1+2\varepsilon} \subset \mathbf{H}_2^{r+1/2+\varepsilon} \subset \mathbf{H}_1^r \subset \dot{C}^r. \tag{17.2}$$

(These inclusions are easily proven. The second one uses the Cauchy–Schwarz inequality and [11] contains a proof of the last one.)

In [6], Daubechies has derived an estimate for $\varphi(x) \in \dot{C}^{r-\varepsilon}$ based on $\mathbf{H}_\infty^{r+1}$. This estimate is easily recovered from the results derived in this this paper. We have, using the notation of Theorem 10.1,

$$\|F_N^j(X)\|_\infty = \max_n |(f_N^j)_n| \le \frac{1}{2\pi}\int_0^{2\pi} |F_N^j(e^{i\omega})|\,d\omega \le \max_{\omega\in\mathbf{R}} |F_N^j(e^{i\omega})|.$$

Define the number β^j such that $2^{-j\beta^j} = \max_{\omega\in\mathbf{R}} |F_N^j(e^{i\omega})|$. Then, by Theorem 10.3 and 11.1, $\varphi(x)$ is $\dot{C}^{N+\beta-\varepsilon}$, where $\beta = \limsup_{j\to\infty}\beta_j$. Cohen [3], [4] has shown that the sequence β^j actually converges to β (the proof is the same as in Theorem 11.1) and that, under some weak conditions on $G(X)$, the optimal regularity order r based on $\mathbf{H}_1^r$ lies between $N+\beta-\varepsilon$ and $N+1+\beta+\varepsilon$. In the case of Daubechies orthonormal wavelets (§14), Cohen and Daubechies [3]–[5] found that β is equivalent to $(\frac{1}{2} - \frac{1}{4}\log_2 3)L \approx 0.10376L$ as $L \to \infty$. It follows (from the following theorem) that the optimal Hölder regularity order of Daubechies orthonormal wavelets is also asymptotically equivalent to $(0.10376\cdots)L$ as $L \to \infty$. However, for small values of L ($L \le 20$), the estimates derived in this paper, listed in Table 1, are much sharper than the asymptotic result of Cohen and Daubechies.

Daubechies has also derived [6, Appendix] other regularity estimates for the special case of her orthonormal wavelets described in §14. It turns out that her estimates are optimal for the Sobolev spaces $\mathbf{H}_2^r$. This is due to Theorem 15.2 and the property, already mentioned in §14, that the polynomial mask $G(X)$ of a Daubechies wavelet is such that $G(X)\tilde{G}(X)$ is the polynomial mask of a Deslauriers and Dubuc interpolatory scheme [21]. We have $G(e^{i\omega})\tilde{G}(e^{i\omega}) = |G(e^{i\omega})|^2$; therefore, from (17.1) the Fourier transform of the limit function of a Deslauriers and Dubuc scheme is $|\hat{\varphi}(\omega)|^2$, where $\hat{\varphi}(\omega)$ is the Fourier transform of the limit function corresponding to the wavelet. The following theorem shows that since the Deslauriers and Dubuc limit functions are strictly linear phase, their optimal Hölder regularity estimates r, provided by Theorem 15.2 and listed in Table 2, are also optimal for the spaces $\mathbf{H}_1^r$. This implies $\varphi(x) \in \mathbf{H}_2^{r/2}$; therefore, $\varphi(x) \in \dot{C}^{(r-1)/2-\varepsilon}$, which is optimal for spaces $\mathbf{H}_2^{r/2}$. This regularity order is exactly the one derived by Daubechies in [6]. Table 1 lists these optimal Sobolev regularity orders for several lengths.

The above discussion shows that if $G(X)\tilde{G}(X)$ is strictly linear phase, then Theorem 15.2 applied on $G(X)\tilde{G}(X)$ provides the optimal Sobolev regularity of the limit function corresponding to the polynomial mask $G(X)$. This result has been derived independently by Daubechies and Cohen [5] using the Littlewood–Paley theory. Recently, Villemoes [23] has shown that this holds more generally under the weak conditions on $G(X)$ of Cohen [3].

But are these "Fourier-optimal" estimates optimal for Hölder regularity? The following theorem shows that the answer is *no*. The basic reason for this is that the exact Hölder regularity order of $\varphi(x)$ depends on the *phase* of $\hat{\varphi}(\omega)$, i.e., on the phase of $G(e^{i\omega})$ by (17.1), whereas Fourier-based regularity estimates only depend on the *modulus* of $\hat{\varphi}(\omega)$ (or $G(e^{i\omega})$). This theorem also shows that in the framework of §15 (the "strictly linear phase" case), optimal Fourier-based estimates are, in fact, also optimal for Hölder regularity. This is natural since the strictly linear phase case corresponds to limit functions that can be made zero-phase by shifting, i.e., $\hat{\varphi}(\omega) \geq 0$.

THEOREM 17.1. *For strictly linear phase masks, optimal regularity estimates based on* $\mathbf{H}_1^r$ *are also optimal for Hölder regularity.*

Optimal regularity estimates based on $\mathbf{H}_2^r$ *are not optimal for Hölder regularity in general. Nonetheless, they are off by* $\frac{1}{2}$ *at most compared to optimal Hölder regularity estimates.*

Proof. We first prove optimality in the strictly linear phase case. From (17.1), the framework of §15 can easily be reduced to the case $\hat{\varphi}(\omega) \geq 0$. Optimality for spaces $\mathbf{H}_1^r$ and $\dot{C}^r$ coincide if we prove that in this case $\varphi(x) \in \dot{C}^\alpha$ implies $\varphi(x) \in \mathbf{H}_1^{\alpha-\varepsilon}$, for any $\varepsilon > 0$. We may restrict to $0 < \alpha \leq 1$, otherwise just consider a derivative of $\varphi(x)$. The integral $I(w) = \int \sin(\omega h/2)|h|^{-1-\alpha+\varepsilon}\,dh$ absolutely converges for $0 < \alpha \leq 1$; making a change of variable yields $I(w) = |w|^{\alpha-\varepsilon} I(1)$; therefore,

$$\begin{aligned}\int \hat{\varphi}(\omega)|w|^{\alpha-\varepsilon}\,dw &= c \iint \hat{\varphi}(\omega)\sin(\omega h/2)|h|^{-1-\alpha+\varepsilon}\,dh d\omega \\ &= c \int (\varphi(h/2) - \varphi(-h/2))|h|^{-1-\alpha+\varepsilon}\,dh\end{aligned}$$

absolutely converges because $\varphi(x)$ is compactly supported and $\dot{C}^\alpha$. This proves that $\varphi(x) \in \mathbf{H}_1^{\alpha-\varepsilon}$.

Table 1 shows that regularity orders of Daubechies orthonormal wavelets that are optimal for Sobolev spaces $\mathbf{H}_2^r$ are not optimal for Hölder regularity.

The fact that optimal Fourier-based estimates are greater than or equal to $r - \frac{1}{2}$, where r is the exact Hölder regularity estimate, results from the well-known inclusion $\dot{C}^r \subset \mathbf{H}_2^{r-\varepsilon}$, which holds for compactly supported functions [16]. □

Trivial extensions of this theorem can be derived for other "Fourier-based" spaces, using inclusions like (17.2).

Note that Table 1 shows that the Hölder regularity estimates of "more symmetric" wavelets are numerically found to be less than those of minimum phase wavelets (the ones that are "nonsymmetric" the most). That is, more symmetry (for the same modulus of $G(e^{i\omega})$) decreases regularity. In addition, both regularity estimates are greater than the optimal Sobolev regularity order that constitutes a lower-bound for the exact Hölder regularity order. In fact, Theorem 17.1 shows that this lower bound is attained for strictly linear phase masks.

Acknowledgments. I would like to thank Pierre Duhamel and Thierry Blu of Centre National d'Études des Télécommunications, Paris, for fruitful discussions and for carefully reading the manuscript, Ingrid Daubechies of AT&T Bell Labs for letting me use her unpublished material [7], [9] and for pointing out the results of Albert Cohen [3]–[5]; and Lars F. Villemoes of the Technical University of Denmark for suggesting reference [16], which led to an improvement of Theorem 17.1.

I am also grateful to the anonymous referees for pointing out several deficiencies, which led to §3, and for suggesting the references of Dyn and Levin.

REFERENCES

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, *Image coding using vector quantization in the wavelet transform domain,* in Proc. 1990 IEEE Internat. Conf. Acoust. Speech Signal Processing, Albuquerque, NM, 1990, pp. 2297–2300.

[2] T. Blu, personal communication, 1991.

[3] A. Cohen, *Ondelettes, analyses multiResolutions et traitement numérique du signal,* Ph.D. thesis, Université Paris IX Dauphine, France, 1990.

[4] ———, *Construction de bases d'ondelettes α-Höldériennes,* Rev. Mat. Iberoamericana, 6 (1990), pp. 91–108.

[5] A. Cohen and I. Daubechies, *Non-separable bidimensional wavelet bases,* Rev. Mat. Iberoamericana, to appear.

[6] I. Daubechies, *Orthonormal bases of compactly supported wavelets,* Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[7] ———, *Orthonormal bases of compactly supported wavelets* II. *Variations on a theme,* SIAM J. Math. Anal., 24 (1993), to appear.

[8] 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.

[9] ———, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals,* SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[10] G. Deslauriers and S. Dubuc, *Interpolation dyadique* in Dimensions Non Entières et Applications, G. Cherbit, ed., Masson, Paris, 1987, pp. 44–56.

[11] ———, *Symmetric iterative interpolation processes,* Constr. Approx. 5 (1989), pp. 49–68.

[12] S. Dubuc, *Interpolation through an iterative scheme,* J. Math. Anal. Appl., 114 (1986), pp. 185–204.

[13] N. Dyn, D. Levin, and J.A. Gregory, *A 4-point interpolatory subdivision scheme for curve design,* Comput. Aided Geom. Design, 4 (1987), pp. 257–268.

[14] N. Dyn and D. Levin, *Interpolating subdivision schemes for the generation of curves and surfaces,* in Multivariate Interpolation and Approximation, W. Haussman and K. Jeller, eds., Birkhäuser, Basel, 1990, pp. 91–106.

[15] N. Dyn, *Subdivision schemes in* CAGD, in Advances in Numerical Analysis II: Wavelets, Subdivision Algorithms and Radial Functions, W. A. Light, ed., Oxford University Press, London, 1991, pp. 36–104.

[16] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, New York, 1983.

[17] C. A. Micchelli and H. Prautzsch, *Refinement and subdivision for spaces of integer translates of a compactly supported function,* in Numerical Analysis, D. F. Griffith and G. A. Watson, eds., Academic Press, New York, 1987, pp. 192–222.

[18] O. Rioul, *A discrete-time multiresolution theory,* IEEE Trans. Signal Proc., to appear.

[19] ———, *Simple, optimal regularity estimates for wavelets,* in Proc. 6th European Signal Processing Conference (EUSIPCO'92), Brussels, 1992.

[20] O. Rioul and M. Vetterli, *Wavelets and signal processing,* IEEE Signal Processing Magazine, 8 (1991), pp. 14–38.

[21] M. J. Shensa, *The discrete wavelet transform: wedding the à trous and Mallat algorithms,* IEEE Trans. Signal Proc., (1992), to appear.

[22] M. Vetterli and C. Herley, *Wavelets and filter banks: relationships and new results,* in Proc. 1990 IEEE Internat. Conf. Acoust. Speech Signal Processing, Albuquerque, NM, 1990, pp. 1723–1726.

[23] L. F. Villemoes, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets,* SIAM J. Math. Anal., (1992), this issue.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1577–1608, November 1992

009

BIFURCATIONS OF NONLINEAR OSCILLATIONS AND FREQUENCY ENTRAINMENT NEAR RESONANCE*

CARMEN CHICONE[†]

Abstract. A unified approach to the Poincaré–Andronov global center bifurcation and the subharmonic Melnikov bifurcation theory is developed using S. P. Diliberto's integration of the variational equations of a two-dimensional system of autonomous ordinary differential equations and a Lyapunov–Schmidt reduction to the implicit function theorem. In addition, the subharmonic Melnikov function is generalized to the case of subharmonic bifurcation from an unperturbed system whose free oscillation is a limit cycle. Thus, results on frequency entrainment are obtained when an external periodic excitation is in resonance with the frequency of the limit cycle. The theory is applied to the subharmonic bifurcations of two coupled van der Pol oscillators running in resonance.

Key words. limit cycles, center bifurcations, subharmonics, Melnikov method, forced oscillations, frequency entrainment

AMS(MOS) subject classifications. 58F14, 58F21, 58F22, 58F30, 34C05, 34C15

1. Introduction.

The subject of this paper is the theory of frequency entrainment for driven nonlinear oscillators when the period of the self-sustained free oscillation is nearly resonance with a periodic external excitation. Our main purpose is to demonstrate a mathematical theorem that is useful in determining the number and position of the subharmonics produced by such an external excitation. Our result is closely related to two well-known theorems: the Poincaré–Andronov theorem on the global center bifurcation [1], [2], [3], [6], and the Melnikov theorem on the global bifurcation of subharmonics from an integrable system [21], [31], [42], [43]. In fact, our theorem can be considered as a generalization of Melnikov's method to cover the case of self-sustained oscillations.

In order to explain the main result, consider a forced oscillation problem of the following type:

$$\dot{x} = \mathbf{f}(x) + \epsilon \mathbf{g}(x,t), \quad x \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

where the unperturbed system

$$\dot{x} = \mathbf{f}(x)$$

with flow $t \mapsto \phi_t$ has a limit cycle Γ of period T as a self-sustained oscillation; the external excitation is periodic of period η, i.e.,

$$\mathbf{g}(x, t+\eta) = \mathbf{g}(x,t),$$

and the period of the external excitation is in resonance with the period of Γ, i.e., there are relatively prime positive integers m and n such that $nT = m\eta$. We are interested in the periodic solutions of the forced system of period $m\eta$, the subharmonics of order m. For this, we look for a curve $\epsilon \mapsto \sigma(\epsilon)$ in the phase plane such that $\sigma(0) \in \Gamma$ and, for sufficiently small ϵ, such that the point $\sigma(\epsilon)$ is the initial value for a subharmonic of order m. When there is such a curve of initial conditions for a family of subharmonics,

* Received by the editors November 28, 1990; accepted for publication (in revised form) February 24, 1992. This research was supported by Air Force Office of Scientific Research and National Science Foundation grants AF-AFOSR-89-0078 and DMS-9022621.

[†] Department of Mathematics, University of Missouri, Columbia, Missouri 65211.

we say $\sigma(0)$ is a subharmonic branch point on Γ. Our main result gives a real valued function $\xi \mapsto \mathcal{C}(\xi)$, for ξ a coordinate on Γ, such that the simple zeros of $\mathcal{C}$ are the subharmonic branch points. The formula for this function is expressed in terms of Euclidean geometrical quantities that we denote as follows: $\|\ \|$ denotes the Euclidean norm, $<, >$ the Euclidean inner product, κ the scalar curvature, div the divergence of a vector field, curl the curl of a vector field, and $\wedge$ the wedge product of two vectors. In fact, the bifurcation function $\mathcal{C}$ is defined by

$$\mathcal{C}(\xi) := [(1-\beta)\mathcal{N} + \alpha\mathcal{M}]\,(m\eta, \xi),$$

where

$$\begin{aligned}
\beta(t) := \beta(t,\xi) &:= \exp\left(\int_0^t \operatorname{div} \mathbf{f}(\phi_s(\xi))\,ds\right),\\
\alpha(t) := \alpha(t,\xi) &:= \int_0^t \left\{\frac{1}{\|\mathbf{f}\|^2}\left[2\kappa\|\mathbf{f}\| - \operatorname{curl}\mathbf{f}\right]\right\}(\phi_\tau(\xi))\beta(\tau)\,d\tau,\\
\mathcal{N}(t,\xi) &:= \int_0^t \left\{\frac{1}{\|\mathbf{f}\|^2}\langle \mathbf{f},\mathbf{g}\rangle - \frac{\alpha(s)}{\beta(s)}\mathbf{f}\wedge\mathbf{g}\right\}(\phi_s(\xi))\,ds,\\
\mathcal{M}(t,\xi) &:= \int_0^t \left\{\frac{1}{\beta(s)}\mathbf{f}\wedge\mathbf{g}\right\}(\phi_s(\xi))\,ds.
\end{aligned}$$

Our main theorem, the limit cycle subharmonic bifurcation theorem, states: *If either $\beta(m\eta,\xi) \neq 1$ or $\alpha(m\eta,\xi) \neq 0$ and if ξ is a simple zero of $\mathcal{C}$, then ξ is a subharmonic branch point.*

If the periodic trajectory Γ of the unperturbed system is not a limit cycle, but rather a periodic trajectory of period T contained in a one parameter family of periodic trajectories of the unperturbed system, then the appropriate bifurcation function is the subharmonic Melnikov function given by $\mathcal{M}$. In order to show that $\mathcal{C}$ reduces to $\mathcal{M}$ in this case, consider a Poincaré section Σ for the unperturbed system that is orthogonal to Γ at a point $\xi_0 \in \Sigma$, and define both the Poincaré return map and the transition time function on Σ. In §2 we will show the derivative of the return map, evaluated at the coordinate on Σ corresponding to ξ_0, is $\beta(T,\xi_0)$, while the derivative of the transition time function is $\alpha(T,\xi_0)$. Thus, for example, if Γ is a member of a one parameter family of periodic orbits, then Γ is not hyperbolic, equivalently, $\beta(nT,\xi) \equiv 1$, and $\mathcal{C}$ reduces to the usual subharmonic Melnikov function. Moreover, in this case, the appropriate nondegeneracy condition of the theorem, $\alpha(m\eta,\xi) \neq 0$, reduces to the condition that the period function for the one parameter family of periodic trajectories of the unperturbed system has a nonzero derivative at Γ; this is the usual nondegeneracy condition for the subharmonic Melnikov theory; cf. [15], [21], [43]. For further results related to period functions, see [4], [9], [11], [12], [14], [16], [17], [33], and [40].

As a typical application, consider two weakly coupled van der Pol oscillators of the form

$$\begin{aligned}
\dot{u} &= v,\\
\dot{v} &= -u + \delta(1-u^2)v,\\
\dot{x} &= \tau y,\\
\dot{y} &= \tau(-x + \delta(1-x^2)y) + \epsilon u,
\end{aligned}$$

where $\delta > 0$, ϵ is a small parameter, and $\tau > 0$ is a rational number. We view the second oscillator as a driven oscillator where the periodic forcing function, $t \mapsto u(t)$,

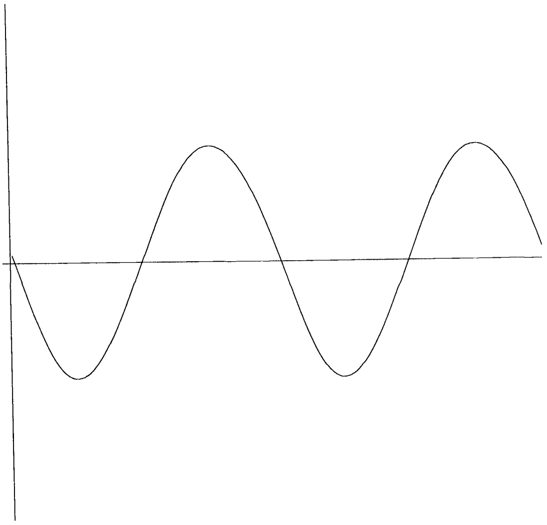


FIG. 1. *Computer generated graph of C vs ξ for weakly coupled van der Pol oscillators* $\dot{u} = v$, $\dot{v} = -u + 0.1(1 - u^2)v$, $\dot{x} = 0.5y$, $\dot{y} = 0.5(-x + 0.1(1 - x^2)y) + \epsilon u$ *with* $2 : 1$ *resonance.*

is the output of the first oscillator. Since τ is rational, the frequency of the free oscillation of the second oscillator is in resonance with the frequency of the external excitation. We ask for the periodic response of the second oscillator when $\epsilon \neq 0$. For background material on forced oscillation problems we refer to [2], [15], [18], [27], [30], [22], [23], [25], [32], [34], [41] for classic treatments and to [21], [26], [36], [42], [43] for some of the latest results. As a typical calculation we fix $\delta = 0.1$ and $\tau = 0.5$. For this example the period of the free oscillation of the second oscillator is twice the period of the "external" force, a $2 : 1$ resonance. A numerical approximation to the graph of $C(\xi)$ is shown in Fig. 1. The graph indicates the existence of four simple zeros of C over one period of the free oscillation and, applying our theorem, we expect, for $|\epsilon|$, sufficiently small, four subharmonics of order two, i.e., four periodic solutions of the forced second oscillator each with period twice the period of the self-sustained oscillation of the first oscillator. Of course, in view of the topology of the circle, two of these subharmonics are stable and two of them are unstable, with the subharmonics corresponding to consecutive zeros of C having opposite stability.

Our treatment of the bifurcation theory of nonlinear oscillations is based on the geometric quadrature of the homogeneous variational equations of a two-dimensional differential system given by Diliberto [20] and a Lyapunov–Schmidt reduction to the implicit function theorem. Using this foundation, we are able to offer a unified approach to the bifurcation theory for plane vector fields that includes the Poincaré–Andronov center bifurcation theorem and the usual subharmonic Melnikov theorem. Since an exposition of these results requires only a minimum of additional effort, we do not offer the most efficient proof of our main result. Rather, we take a slightly longer route to our theorem that allows us to give new proofs of these other classic results. In addition to the unification provided by our implicit function theorem approach to the bifurcation theory, we obtain smooth curves of bifurcating periodic solutions. Thus, the precise positions of the bifurcating solutions can be computed and, perhaps, continued to further global bifurcations. Also, we mention that our method yields proofs of the bifurcation theorems that do not require reduction to a normal form or a change to action angle variables.

The plan of the paper is as follows. In §2 we give a proof of Diliberto's theorem and obtain the geometric interpretation of the functions α and β in terms of the Poincaré map and the transition time function. We also define the functions $\mathcal{N}$ and $\mathcal{M}$ in their natural context as constituents of the solution of a certain inhomogeneous variational equation. In §3 we prove the Poincaré–Andronov center bifurcation theorem. In §4 we prove the subharmonic Melnikov theorem, our main result, the limit cycle subharmonic bifurcation theorem, and we prove some theorems on bifurcation of subharmonics from degenerate families. Also, we connect our theory with the classical perturbation theory for linear systems. In the final section, §5, we give additional applications of the theory.

2. Diliberto's theorem. The fundamental result on which this paper is based is the theorem of Diliberto [20] on the integration of the homogeneous variational equations of a plane autonomous differential equation in terms of geometric quantities along a given trajectory of the system. Here we let $\mathbf{X} = (X_1, X_2)$ denote a smooth plane vector field with flow ϕ_t. The geometric quantities are the curvature, the curl, and the divergence given by

$$\kappa = \|\mathbf{X}\|^{-3}\left(X_1\dot{X}_2 - X_2\dot{X}_1\right), \quad \operatorname{curl}\mathbf{X} = \frac{\partial X_2}{\partial x_1} - \frac{\partial X_1}{\partial x_2}, \quad \operatorname{div}\mathbf{X} = \frac{\partial X_1}{\partial x_1} + \frac{\partial X_2}{\partial x_2},$$

where $\|\mathbf{X}\| = \sqrt{\langle \mathbf{X}, \mathbf{X}\rangle}$ denotes the Euclidean norm. It will also be convenient to define the orthogonal vector field $\mathbf{X}^\perp := (-X_2, X_1)$ as well as the vector field $\mathbf{u}_{\mathbf{X}^\perp}$ parallel to $\mathbf{X}^\perp$ given by

$$\mathbf{u}_{\mathbf{X}^\perp}(p) := \frac{1}{\|\mathbf{X}(p)\|^2}\mathbf{X}^\perp(p).$$

The normalization is chosen so that $\langle \mathbf{X}^\perp, \mathbf{u}_{\mathbf{X}^\perp}\rangle = 1$. Finally, we introduce the wedge product of two vector fields $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (X_1, X_2)$ to be $\mathbf{X} \wedge \mathbf{Y} = X_1Y_2 - X_2Y_1$. In the course of our discussion we make use of the formula

$$\mathbf{X} \wedge \mathbf{Y} = \left\langle \mathbf{X}^\perp, \mathbf{Y}\right\rangle.$$

It allows for a choice between two geometric interpretations of the same quantity, namely, the area given by the wedge product and the projection given by the inner product. This choice is rather arbitrary, but is often made according to tradition.

THEOREM 2.1 (Diliberto's theorem). *If* $\mathbf{X}(p) \neq 0$, *then the linear variational equation along the integral curve* $t \mapsto \phi_t(p)$,

$$\dot{\mathbf{V}} = D\mathbf{X}(\phi_t(p))\mathbf{V},$$

has a fundamental matrix solution $\Phi(t)$, *satisfying* $\det(\Phi(0)) = 1$, *given by*

$$\Phi(t) := [\mathbf{X}(\phi_t(p)), \mathbf{V}(t)],$$

where

$$\mathbf{V}(t) := \alpha(t)\mathbf{X}(\phi_t(p)) + \beta(t)\mathbf{u}_{\mathbf{X}^\perp}(\phi_t(p))$$

and

$$\beta(t) := \beta(t, \mathbf{X}, p) := \exp\left(\int_0^t \operatorname{div}\, \mathbf{X}(\phi_s(p))\, ds\right),$$

$$\alpha(t) := \alpha(t, \mathbf{X}, p) := \int_0^t \left\{\frac{1}{\|\mathbf{X}\|^2}\left[2\kappa\|\mathbf{X}\| - \operatorname{curl}\,\mathbf{X}\right]\right\}(\phi_\tau(p))\beta(\tau)\, d\tau.$$

Moreover, the inverse of this fundamental matrix (partitioned by rows) is given by

$$\Phi^{-1}(t) = \frac{1}{\beta(t)} \begin{bmatrix} -\mathbf{V}^{\perp}(t) \\ \mathbf{X}^{\perp}(\phi_t(p)) \end{bmatrix}.$$

Proof. Define $\gamma(t) := \phi_t(p)$. Since $\mathbf{X}(p) \neq 0$, the function $t \mapsto \mathbf{X}(\gamma(t))$ is a nontrivial solution of the linear variational equation. Next, define

$$\mathbf{P}(t) := \frac{1}{\|\mathbf{X}(\gamma(t))\|} \left[\mathbf{X}(\gamma(t)), \mathbf{X}^{\perp}(\gamma(t))\right]$$

(partitioned by columns) and use the coordinate transformation $\mathbf{U} = \mathbf{P}^{-1}\mathbf{V}$ on the linear variational equation $\dot{\mathbf{V}} = D\mathbf{X}(\gamma(t))\mathbf{V}$ to obtain $\dot{\mathbf{U}} = \mathbf{A}\mathbf{U}$, where

$$\mathbf{A} := \mathbf{P}^{-1}(D\mathbf{X}(\gamma(t)))\mathbf{P} - \mathbf{P}^{-1}\dot{\mathbf{P}}.$$

A somewhat tedious calculation shows the matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} \|\mathbf{X}\|^{-1}\frac{d}{dt}\|\mathbf{X}\| & 2\|\mathbf{X}\|^{-2}(X_1\dot{X}_2 - X_2\dot{X}_1) - \operatorname{curl}\mathbf{X} \\ 0 & -\|\mathbf{X}\|^{-1}\frac{d}{dt}\|\mathbf{X}\| + \operatorname{div}\mathbf{X} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{d}{dt}\ln\|\mathbf{X}\| & 2\|\mathbf{X}\|\kappa - \operatorname{curl}\mathbf{X} \\ 0 & -\frac{d}{dt}\ln\|\mathbf{X}\| + \operatorname{div}\mathbf{X} \end{bmatrix}.$$

Since this system is in triangular form, we can find a simple representation for the general solution of $\dot{\mathbf{U}} = \mathbf{A}\mathbf{U}$. In fact, if we use $\mathbf{e}_1 := (1, 0)$ and $\mathbf{e}_2 := (0, 1)$, then

$$\mathbf{U}(t) = \frac{\|\mathbf{X}(\gamma(t))\|}{\|\mathbf{X}(p)\|}\left\{U_1(0) + U_2(0)\|\mathbf{X}(p)\|^2\alpha(t)\right\}\mathbf{e}_1 + \frac{U_2(0)\|\mathbf{X}(p)\|}{\|\mathbf{X}(\gamma(t))\|}\beta(t)\mathbf{e}_2,$$

where $\mathbf{U} = (U_1, U_2)$. Finally, if we choose the initial conditions so that one solution has initial condition $\mathbf{U}(0) = \|\mathbf{X}(p)\|\mathbf{e}_1$, and a second solution satisfies the initial condition $\|\mathbf{X}(p)\|\mathbf{U}(0) = \mathbf{e}_2$, then we obtain two linearly independent solutions. These solutions form the columns of a fundamental matrix $\tilde{\Phi}$ for $\dot{\mathbf{U}} = \mathbf{A}\mathbf{U}$, while the fundamental matrix in the statement of the theorem is just $\Phi = \mathbf{P}\tilde{\Phi}$. A simple calculation shows that $\det(\Phi(0)) = 1$. The statement about the inverse of the fundamental matrix $\Phi(t)$ is a straightforward deduction, and its proof is omitted. □

Remark. It should be noted that the constant 2 multiplying the curvature in the theorem was omitted in the formulas in Diliberto's original paper.

Diliberto's theorem contains all the important information about the solutions of the linear variational equation along the trajectories of the plane vector field $\mathbf{X}$. We will use the theorem to derive several corollaries that illustrate the importance of the result. The first few of these corollaries can be taken to give the *geometric* meaning of the two functions α and β that are defined in the statement of the theorem. For this we introduce some basic definitions. Let p and q be points in $\mathbb{R}^2$ on the trajectory $t \mapsto \phi_t(p)$ of $\mathbf{X}$ with $q = \phi_T(p)$. We consider two sections for the flow of $\mathbf{X}$, Σ at p, and Δ at q, given by plane curves $s \mapsto \sigma(s)$ and $s \mapsto \delta(s)$ with $\sigma(\xi) = p$. Moreover, we let $\mathbf{Y}$ denote the tangent vector field of σ and $\mathbf{Z}$ the tangent vector field of δ. We use $\widetilde{\mathbf{h}} : \Sigma \to \Delta$ to denote the *transition map.* It assigns to a point $r \in \Sigma$ the point where the trajectory of $\mathbf{X}$ starting from r first meets Δ. We use $\widetilde{T} : \Sigma \to \mathbb{R}$ to denote the *transition time function* that assigns to $r \in \Sigma$ the minimum positive time required for the transition. There will be an open interval $\mathcal{I} \subset \mathbb{R}$ such that for $s \in \mathcal{I}$ both $\widetilde{\mathbf{h}}$ and $\widetilde{T}$ are defined on $\sigma(\mathcal{I})$. We define $\mathcal{J} := \delta^{-1} \circ \widetilde{\mathbf{h}} \circ \sigma(\mathcal{I})$. Then, we can express the transition map and the transition time function in the local coordinates on the two sections defined by the functions σ and δ. It is convenient to give these representations names. In fact, we define the *scalar transition map* $h : \mathcal{I} \to \mathcal{J}$ by the formula

$$h := \delta^{-1} \circ \widetilde{\mathbf{h}} \circ \sigma,$$

and the *scalar transition time function* $T : \mathcal{I} \to \mathbb{R}$ by

$$T := \widetilde{T} \circ \sigma.$$

When we write $\widetilde{\mathbf{h}}'(p)$, we understand this to be the derivative $h'(\xi)$, and when we write $\widetilde{T}'(p)$ we understand this to be the derivative $T'(\xi)$. Finally, we specify two special cases. If p is a periodic point of $\mathbf{X}$, we can take $\Sigma = \Delta$. In this case, the transition map is called the *return map* or the *Poincaré map* on the *Poincaré section* Σ. If, in addition, p is contained in a one parameter family of periodic trajectories of $\mathbf{X}$, we say p is in a *period annulus with* (*Poincaré*) *section* Σ. In this case, the corresponding transition time function is called the *period function.*

The next theorem identifies the functions α and β in Diliberto's theorem geometrically and, in particular, provides formulas for the derivative of the return map and the period function.

THEOREM 2.2. *Let* $\mathbf{X}$ *be a plane vector field with flow* ϕ_t, Σ *a section for the flow at* $p \in \mathbb{R}^2$, *and* Δ *a section for the flow at* $q = \phi_T(p)$. *Also, let* $s \mapsto \sigma(s)$ *be a local coordinate function for* Σ *with tangent vector field* $\mathbf{Y}$ *such that* $\sigma(\xi) = p$, *and let* $s \mapsto \delta(s)$ *be a local coordinate function for* Δ *with tangent vector field* $\mathbf{Z}$. *If* $\widetilde{\mathbf{h}}$ *is the transition map and* $\widetilde{T}$ *is the transition time function, then*

$$\widetilde{\mathbf{h}}'(p) = \frac{\mathbf{X} \wedge \mathbf{Y}(p)}{\mathbf{X} \wedge \mathbf{Z}(\widetilde{\mathbf{h}}(p))}\beta(\widetilde{T}(p), \mathbf{X}, p)$$

and

$$\widetilde{T}'(p) = \frac{\langle \mathbf{Z}, \mathbf{X}\rangle}{\|\mathbf{X}\|^2}(\widetilde{\mathbf{h}}(p))\widetilde{\mathbf{h}}'(p) - \frac{\langle \mathbf{Y}, \mathbf{X}\rangle}{\|\mathbf{X}\|^2}(p) - \mathbf{X} \wedge \mathbf{Y}(p)\alpha(\widetilde{T}(p), \mathbf{X}, p).$$

In particular, if p *is a periodic point of* $\mathbf{X}$, *the derivative of the return map is given by*

$$\widetilde{\mathbf{h}}'(p) = \frac{\mathbf{X} \wedge \mathbf{Y}(p)}{\mathbf{X} \wedge \mathbf{Y}(\widetilde{\mathbf{h}}(p))}\beta(\widetilde{T}(p), \mathbf{X}, p),$$

and, in addition, if p *is contained in a period annulus, the derivative of the period function is given by*

$$\widetilde{T}'(p) = -\mathbf{X} \wedge \mathbf{Y}(p)\alpha(\widetilde{T}(p), \mathbf{X}, p).$$

Proof. By the definition of the transition map and the transition time function we have

$$\widetilde{\mathbf{h}}(\sigma(s)) = \phi_{\widetilde{T}(\sigma(s))}(\sigma(s)),$$

or, in local coordinates,

$$\delta(h(s)) = \phi_{T(s)}(\sigma(s)).$$

After differentiating both sides of the last equation and evaluating at $s = \xi$, we obtain the following formula for the derivative of the transition map:

$$h'(\xi)\mathbf{Z}(q) = \frac{d}{ds}\phi_{T(s)}(\sigma(s))\bigg|_{s=\xi} = D\phi_{T(\xi)}(p)\mathbf{Y}(p) + T'(\xi)\mathbf{X}(q).$$

We will apply Diliberto's theorem to obtain the required formulas. For this, observe that the function $t \mapsto \mathbf{V}(t)$, giving the second column of the Diliberto fundamental matrix of the linear variational equations along the trajectory $t \mapsto \phi_t(p)$ of $\mathbf{X}$, is the unique solution of the following linear variational initial value problem:

$$\dot{\mathbf{V}} = D\mathbf{X}(\phi_t(p))\mathbf{V}, \qquad \mathbf{V}(0) = \mathbf{u}_{\mathbf{X}^\perp}(p),$$

whose solution evaluated at $T(\xi)$ is

$$\mathbf{V}(T(\xi)) = \alpha(T(\xi), \mathbf{X}, p)\mathbf{X}(q) + \beta(T(\xi), \mathbf{X}, p)\mathbf{u}_{\mathbf{X}^\perp}(q).$$

Since $t \mapsto D\phi_t(p)\mathbf{u}_{\mathbf{X}^\perp}(p)$ is a solution of the same initial value problem, we have

$$D\phi_t(p)\mathbf{u}_{\mathbf{X}^\perp}(p) = \mathbf{V}(t),$$

and, of course, we also have

$$D\phi_t(p)\mathbf{X}(p) = \mathbf{X}(\phi_t(p)).$$

Moreover, there are real numbers a, b, c, and d such that

$$\mathbf{Y}(p) = a\mathbf{X}(p) + b\mathbf{u}_{\mathbf{X}^\perp}(p), \qquad \mathbf{Z}(q) = c\mathbf{X}(q) + d\mathbf{u}_{\mathbf{X}^\perp}(q).$$

In fact, these numbers can be expressed in terms of the given vectors as follows:

$$\begin{aligned} a &= \frac{\langle \mathbf{Y}, \mathbf{X}\rangle}{||\mathbf{X}||^2}(p), \qquad & b &= \mathbf{X} \wedge \mathbf{Y}(p), \\ c &= \frac{\langle \mathbf{Z}, \mathbf{X}\rangle}{||\mathbf{X}||^2}(q), \qquad & d &= \mathbf{X} \wedge \mathbf{Z}(q). \end{aligned}$$

Now, we compute

$$D\phi_t(p)\mathbf{Y}(p) = a\mathbf{X}(\phi_t(p)) + b\mathbf{V}(t),$$

and, after the obvious substitutions into the formula for $h'(\xi)$, we obtain

$$(ch'(\xi) - T'(\xi) - a - b\alpha)\,\mathbf{X}(q) + (dh'(\xi) - b\beta)\,\mathbf{u}_{\mathbf{X}^\perp}(q) = 0.$$

The first part of the theorem follows immediately from the last equality and the linear independence of $\mathbf{X}(q)$ and $\mathbf{u}_{\mathbf{X}^\perp}(q)$.

From the definition of the return map we can assume $\Sigma = \Delta, q = \widetilde{\mathbf{h}}(p)$, and $\mathbf{Y} = \mathbf{Z}$. The derivative of the return map in the statement of the theorem is obtained by specializing the formula,

$$h'(\xi) = \frac{b}{d}\beta,$$

just derived for the derivative of the transition map.

For the derivative of the period function we have the same specialization and, in addition, the fact that p lies in a period annulus. The periodicity of p implies $p = q$ and, in turn, $a = c$, while the membership of p in a period annulus implies $\widetilde{\mathbf{h}}'(p) = 1$. Hence, in this case, $ch'(\xi) - a = 0$ and the formula obtained for the derivative of the transition time function reduces to the required formula for the period function,

$$T'(\xi) = -b\alpha. \qquad \square$$

For the majority of situations encountered in practice, when dealing with a spiral flow in the plane, a horizontal line segment can be chosen as a Poincaré section. In this case, the representation of the derivative of the return map and of the period function given in the last theorem takes a particularly simple form. If the horizontal line segment is an interval of the line with equation $y = k$, then, for example, the vector field $\mathbf{Y}$ may be taken to be the unit vector in the (positive) horizontal direction. If X_2 denotes the second component of the vector field $\mathbf{X}$, then, with $\gamma(t) := \phi_t(x, k)$, the representation of the derivative of the return map is

$$h'(x) = \frac{X_2(x,k)}{X_2(h(x),k)} \exp\left(\int_0^{T(x)} \operatorname{div}\mathbf{X}(\gamma(t))\,dt\right),$$

while the representation of the derivative of the period function is

$$T'(x) = X_2(x,k) \int_0^{T(x)} \left\{\frac{1}{\|\mathbf{X}\|^2}\left[2\kappa\|\mathbf{X}\| - \operatorname{curl}\mathbf{X}\right]\right\}(\gamma(\tau)) \exp\left(\int_0^{\tau} \operatorname{div}\mathbf{X}(\gamma(t))\,dt\right) d\tau.$$

These cases are often encountered in the applications.

The geometric identification of the function β is now clear. In fact, generally, if we consider a trajectory $t \mapsto \phi_t(p)$ of $\mathbf{X}$ and two sections Σ at p and Δ at $\phi_\tau(p)$, then $\beta(\tau, \mathbf{X}, p)$ is just the normalized derivative of the transition map. The normalization coefficient is the quotient of the two wedge products given in the theorem. The geometric identification of the function α is more complicated in the general case. But, if the two sections are orthogonal to the trajectory through p, then $\alpha(\tau, \mathbf{X}, p)$ is just the derivative of the transition time function normalized by the multiplicative factor $-\mathbf{X} \wedge \mathbf{Y}(p)$.

The next lemma gives an explicit formula for the solution of the inhomogeneous linear variational equations along a trajectory of $\mathbf{X}$; this formula will be needed in the perturbation theory that will be developed later.

LEMMA 2.3 (variation lemma). *Let* $\mathbf{X} = (X_1, X_2)$ *and* $\mathbf{b} = (b_1, b_2)$ *denote smooth plane vector fields, and let* ϕ_t *denote the flow of* $\mathbf{X}$. *If* $p \in \mathbb{R}^2$ *and* $\mathbf{X}(p) \neq 0$, *then the solution,* $t \mapsto \mathbf{W}(t)$, *of the initial value problem*

$$\dot{\mathbf{W}} = D\mathbf{X}(\phi_t(p))\mathbf{W} + \mathbf{b}(\phi_t(p)), \qquad \mathbf{W}(0) = 0$$

is given by

$$\mathbf{W}(t) = [\mathcal{N}(t) + \alpha(t)\mathcal{M}(t)]\,\mathbf{X}(\phi_t(p)) + [\beta(t)\mathcal{M}(t)]\,\mathbf{u}_{\mathbf{X}^\perp}(\phi_t(p)),$$

where

$$\mathcal{N}(t) := \mathcal{N}(t, \mathbf{X}, \mathbf{b}, p) := \int_0^t \left\{ \frac{1}{||\mathbf{X}||^2} \langle \mathbf{X}, \mathbf{b} \rangle - \frac{\alpha(s)}{\beta(s)} \mathbf{X} \wedge \mathbf{b} \right\} (\phi_s(p))\, ds,$$
$$\mathcal{M}(t) := \mathcal{M}(t, \mathbf{X}, \mathbf{b}, p) := \int_0^t \left\{ \frac{1}{\beta(s)} \mathbf{X} \wedge \mathbf{b} \right\} (\phi_s(p))\, ds,$$

and α, β are defined in the statement of Diliberto's theorem. In addition, if m is a positive integer, p is a point in a period annulus of $\mathbf{X}$ with local section Σ given at p by the integral curve $s \mapsto \sigma(s)$ of $\mathbf{X}^\perp$ such that $\sigma(\xi) = p$, and if T denotes the scalar period function defined on Σ, then

$$\alpha(mT(\xi)) = -\frac{m}{||\mathbf{X}(p)||^2} T'(\xi), \qquad \beta(mT(\xi)) = 1$$

and

$$W(mT(\xi)) = \left[\mathcal{N}(mT(\xi)) - \frac{m}{||\mathbf{X}(p)||^2} T'(\xi) \mathcal{M}(mT(\xi)) \right] \mathbf{X}(p) + \mathcal{M}(mT(\xi)) \mathbf{u}_{\mathbf{X}^\perp}(p).$$

Proof. Using variation of parameters, we have

$$\mathbf{W}(t) = \Phi(t) \int_0^t \Phi^{-1}(s) \mathbf{b}(s)\, ds,$$

where Φ is the Diliberto fundamental matrix. To evaluate this formula for $\mathbf{W}$ we first observe that

$$-\mathbf{V}^\perp(t) = \left\{ -\alpha(t) \mathbf{X}^\perp + \beta(t) \frac{1}{||\mathbf{X}||^2} \mathbf{X} \right\} (\phi_t(p)).$$

Then, using the formulas for the Diliberto fundamental matrix and its inverse given in Diliberto's theorem, we compute

$$\int_0^t \Phi^{-1}(s) \mathbf{b}(\phi_s(p))\, ds = \begin{bmatrix} \mathcal{N}(t) \\ \mathcal{M}(t) \end{bmatrix}.$$

The first statement of the lemma is now immediate.

For the second part of the lemma, we assume p is a periodic point of $\mathbf{X}$ and the coordinate for the section at p is given by the integral curve $s \mapsto \sigma(s)$ of $\mathbf{X}^\perp$ with $\sigma(\xi) = p$. Then, since p belongs to a period annulus, the return map is the identity on Σ, and we conclude from the formula for the return map obtained above that

$$\beta(mT(\xi)) = 1.$$

Moreover, using this fact, together with a straightforward change of variables in the integral representation of $\alpha(mT(\xi))$, or, using the formula for $\widetilde{T}'$, we find

$$\alpha(mT(\xi)) = m\alpha(T(\xi)).$$

Thus, in view of the formula for the derivative of the period function given in the previous theorem,

$$\alpha(mT(\xi)) = -\frac{m}{||\mathbf{X}(p)||^2}T'(\xi).$$

After substitution of these identities into the formula for $\mathbf{W}(t)$ and a simple rearrangement of the terms, the final equality in the statement of the lemma is proved. □

3. The Poincaré–Andronov theorem. In order to state the main result of this section, we consider a plane vector field $(x, y) \mapsto \mathbf{X}(x, y, \epsilon)$ depending on the real small parameter ϵ. In case the phase portrait of the unperturbed system $\mathbf{X}_0(x, y) := \mathbf{X}(x, y, 0)$ contains a period annulus $\mathcal{A}$, we seek to determine if there is a periodic trajectory Γ contained in $\mathcal{A}$ and a continuous family Γ_ϵ of periodic trajectories of $(x, y) \mapsto \mathbf{X}(x, y, \epsilon)$ such that $\Gamma_0 = \Gamma$.

For this bifurcation problem it is convenient to consider the differential equation corresponding to the vector field $(x, y) \mapsto \mathbf{X}(x, y, \epsilon)$ in the form

$$\dot{x} = P(x, y) + \epsilon p(x, y) + O(\epsilon^2), \qquad \dot{y} = Q(x, y) + \epsilon q(x, y) + O(\epsilon^2).$$

Also, we let ϕ_t^ϵ denote the flow of $\mathbf{X}$. We can always arrange the coordinates so that a certain horizontal line segment $\Sigma : y = y_0$ is transverse to the flow of $\mathbf{X}_0$ in $\mathcal{A}$. Then, there is some $\epsilon_0 > 0$ such that $(x, y) \mapsto \mathbf{X}(x, y, \epsilon)$ is transverse to Σ, for all ϵ satisfying $|\epsilon| < \epsilon_0$. We assume Γ is one of the periodic trajectories in $\mathcal{A}$ transverse to Σ, and let ξ denote the usual distance coordinate along Σ. Then, both the scalar transition time function $(\xi, \epsilon) \mapsto T(\xi, \epsilon)$ and the scalar Poincaré return map $(\xi, \epsilon) \mapsto h(\xi, \epsilon)$ are defined on Σ. It is convenient in the analysis to define $\mathbf{X}_0^\perp$ to be the vector field with components $(-Q, P)$ and $\mathbf{H}(\xi, \epsilon)$ to be the vector $(d(\xi, \epsilon), 0)$, where d is the *displacement function* defined by $d(\xi, \epsilon) := h(\xi, \epsilon) - \xi$. We also define the *normalized displacement function* F by

$$F(\xi, \epsilon) = \mathbf{X}_0^\perp(\xi, y_0) \cdot \mathbf{H}(\xi, \epsilon) = -Q(\xi, y_0)d(\xi, \epsilon).$$

There are two basic facts: $F(\xi, \epsilon) = 0$ if and only if the trajectory of $(x, y) \mapsto \mathbf{X}(x, y, \epsilon)$ through (ξ, y_0) is periodic and $F(\xi, 0) \equiv 0$. If there is an $\epsilon_* > 0$ and a continuous function $\beta : (-\epsilon_*, \epsilon_*) \to \Sigma$ such that $F(\beta(\epsilon), \epsilon) \equiv 0$, then, for each ϵ in the domain of β, there is a periodic trajectory Γ_ϵ of the vector field $(x, y) \to \mathbf{X}(x, y, \epsilon)$ passing through the point $(\beta(\epsilon), y_0)$. In this case, we say *a continuous family of periodic trajectories of* $\mathbf{X}$ *emerges from the periodic trajectory* Γ_0.

The next result provides a means to identify the periodic trajectories in a period annulus of the unperturbed system from which a family of limit cycles emerges. It is a version of the theorem given in [1] and is the basic bifurcation theorem in this context. Our approach to the proof of this theorem is somewhat different from the development of the same result in [1]. For example, in [1, §32], the vector field family in which the bifurcation occurs is assumed to be analytic and, in [1, §33], a first integral must be constructed for the unperturbed conservative system. Here, we prove the result by analyzing an appropriate variational equation directly. The theorem has two main components: a reduction and an identification. Here, reduction refers to reducing the problem of the existence of a family of limit cycles to an application of the implicit function theorem, and identification refers to the identification of the appropriate partial derivatives in terms of the components of the vector field $\mathbf{X}$.

THEOREM 3.1. *Let $(x,y) \mapsto \mathbf{X}(x,y,\epsilon)$ denote a vector field with flow $t \mapsto \phi_t^\epsilon$ and corresponding differential equation*

$$\dot{x} = P(x,y) + \epsilon p(x,y) + O(\epsilon^2), \qquad \dot{y} = Q(x,y) + \epsilon q(x,y) + O(\epsilon^2)$$

such that the corresponding unperturbed vector field $\mathbf{X}_0$ given by $(x,y) \to \mathbf{X}(x,y,0)$ has a period annulus with Poincaré section $\Sigma \subset \{(x,y) : y = y_0\}$.

(i) *Reduction. If there is a point $\xi_0 \in \Sigma$ such that the corresponding normalized displacement function F satisfies $F_\epsilon(\xi_0,0) = 0$ and $F_{\xi\epsilon}(\xi_0,0) \neq 0$, then there is a periodic trajectory Γ of $\mathbf{X}_0$ meeting Σ at ξ_0 and a continuous family, Γ_ϵ, of periodic trajectories of $\mathbf{X}$ emerging from Γ. Moreover, for sufficiently small $\epsilon \neq 0$, the periodic trajectory Γ_ϵ is a limit cycle of the vector field $(x,y) \to \mathbf{X}(x,y,\epsilon)$. In fact, if $\epsilon F_{\xi\epsilon}(\xi_0,0)/Q(\xi_0,y_0) > 0$, then Γ_ϵ is asymptotically stable. If $\epsilon F_{\xi\epsilon}(\xi_0,0)/Q(\xi_0,y_0) < 0$, then Γ_ϵ is asymptotically unstable.*

(ii) *Identification. The partial derivative of the normalized displacement function with respect to the bifurcation parameter is given by*

$$F_\epsilon(\xi,0) = \int_0^{T(\xi,0)} (Pq - Qp)\,(\gamma(t)) \exp\left(-\int_0^t \operatorname{div} \mathbf{X}_0(\gamma(s))\,ds\right) dt,$$

where $\gamma(t) := \phi_t^0(\xi, y_0)$ is the integral curve corresponding to the periodic trajectory Γ through (ξ, y_0). In addition, if $F_\epsilon(\xi_0,0) = 0$ for some $\xi_0 \in \Sigma$, then

$$\begin{aligned} F_{\xi\epsilon}(\xi_0,0) = -Q(\xi_0,y_0)\Bigg\{ & \operatorname{div} \mathbf{X}_0(\xi_0,y_0) T_\epsilon(\xi_0,0) \\ & + \int_0^{T(\xi_0,0)} \operatorname{div}(p,q)(\gamma(t))\,dt \\ & + \int_0^{T(\xi_0,0)} \frac{d}{d\epsilon} \operatorname{div} \mathbf{X}_0(\phi_t^\epsilon(\xi_0,y_0))\bigg|_{\epsilon=0} dt \Bigg\}. \end{aligned}$$

(iii) *If $\mathbf{X}_0$ is Hamiltonian* $(\operatorname{div} \mathbf{X}_0 \equiv 0)$, *then for $\xi \in \Sigma$,*

$$F_\epsilon(\xi,0) = \int_0^{T(\xi,0)} (Pq - Qp)(\gamma(t))\,dt = -\int_\Omega \operatorname{div}(p,q)\,dxdy.$$

If, in addition, $F(\xi_0,0) = 0$, then

$$F_{\xi\epsilon}(\xi_0,0) = -Q(\xi_0,y_0) \int_0^{T(\xi_0,0)} \operatorname{div}(p,q)(\gamma(t))\,dt.$$

Proof. Since $F(\xi,0) \equiv 0$,

$$F(\xi,\epsilon) = \epsilon\,(F_\epsilon(\xi,0) + O(\epsilon)) := \epsilon G(\xi,\epsilon).$$

But then, from the hypotheses,

$$G(\xi_0,0) = F_\epsilon(\xi_0,0) = 0 \quad \text{and} \quad G_\xi(\xi_0,0) = F_{\xi\epsilon}(\xi_0,0) \neq 0.$$

The implicit function theorem applied to G implies the existence of the required function $\epsilon \mapsto \beta(\epsilon)$. For the stability of the perturbed limit cycles, just note that the displacement has the form

$$d(\xi,\epsilon) = \epsilon d_\epsilon(\xi,0) + O(\epsilon^2).$$

Thus, for sufficiently small ϵ, if $\epsilon d_{\epsilon\xi}(\xi_0, 0) < 0$, then $\xi \mapsto d(\xi, \epsilon)$ crosses the horizontal line segment Σ with negative slope as ξ increases through ξ_0. It is then immediate from the definition of the displacement that Γ_ϵ is a stable limit cycle. By the same argument, the limit cycle will be unstable when $\epsilon d_{\epsilon\xi}(\xi_0, 0) < 0$. But,

$$\epsilon d_{\epsilon\xi}(\xi_0, 0) = -\epsilon \frac{F_{\epsilon\xi}(\xi_0, 0)}{Q(\xi_0, y_0)},$$

and therefore the statement of the theorem follows.

For the computation of the partial derivative $F_\epsilon(\xi, 0)$, we have

$$F_\epsilon(\xi, 0) = \mathbf{X}_0^\perp(\xi, y_0) \cdot \mathbf{H}_\epsilon(\xi, 0)$$

with $\mathbf{H}_\epsilon(\xi, 0) = (h_\epsilon(\xi, 0), 0)$. Thus, we must compute h_ϵ. For this, consider the integral curve $(x(t, \xi, \epsilon), y(t, \xi, \epsilon))$ of $(x, y) \mapsto \mathbf{X}(x, y, \epsilon)$ starting at the point (ξ, y_0), and let $T(\xi, \epsilon)$ denote the time of first return of this solution to Σ. Clearly, we have

$$x(T(\xi, \epsilon), \xi, \epsilon) = h(\xi, \epsilon), \qquad y(T(\xi, \epsilon), \xi, \epsilon) = y_0,$$

and, after differentiation with respect to ϵ and an evaluation at $\epsilon = 0$, we obtain

$$\begin{aligned} &\dot{x}(T(\xi, 0), \xi, 0)T_\epsilon(\xi, 0) + x_\epsilon(T(\xi, 0), \xi, 0) = h_\epsilon(\xi, 0), \\ &\dot{y}(T(\xi, 0), \xi, 0)T_\epsilon(\xi, 0) + y_\epsilon(T(\xi, 0), \xi, 0) = 0. \end{aligned}$$

Define

$$\mathbf{W}(t) := (x_\epsilon(t, \xi, 0), y_\epsilon(t, \xi, 0)),$$

and then, using the abbreviations $T := T(\xi, 0)$, $T_\epsilon := T_\epsilon(\xi, 0)$, and $\mathbf{H}_\epsilon := \mathbf{H}_\epsilon(\xi, 0)$, we have

$$T_\epsilon \mathbf{X}_0(\gamma(T)) + \mathbf{W}(T) = \mathbf{H}_\epsilon = T_\epsilon \mathbf{X}_0(\xi, y_0) + \mathbf{W}(T).$$

Consequently,

$$F_\epsilon(\xi, 0) = \mathbf{X}_0^\perp(\xi, y_0) \cdot \mathbf{W}(T).$$

In order to compute $\mathbf{W}$ we will solve an appropriate variational initial value problem. Since

$$x(0, \xi, \epsilon) = \xi, \qquad y(0, \xi, \epsilon) = y_0,$$

we have

$$x_\epsilon(0, \xi, \epsilon) = 0, \qquad y_\epsilon(0, \xi, \epsilon) = 0.$$

Thus, it is easy to see that $\mathbf{W}$ is the solution of the the following initial value problem:

$$\dot{\mathbf{W}} = D\mathbf{X}_0(\gamma(t))\mathbf{W} + \mathbf{b}(t), \qquad \mathbf{W}(0) = 0,$$

where $\mathbf{b}(t) = (p(\gamma(t)), q(\gamma(t)))$. An application of the second part of the variation lemma with $m = 1$ implies

$$F_\epsilon(\xi, 0) = \mathbf{X}_0^\perp(\xi, y_0) \cdot \mathbf{W}(T) = \mathcal{M}(T, \mathbf{X}, \mathbf{b}, \gamma(0)).$$

Thus

$$F_\epsilon(\xi,0)=\int_0^T (Pq-Qp)\,(\gamma(t))\exp\left(-\int_0^t \operatorname{div}\mathbf{X}_0(\gamma(\tau))\,d\tau\right)dt$$

as required.

To obtain the representation for $F_{\xi\epsilon}(\xi,0)$ under the hypothesis that $F_\epsilon(\xi_0,0)=0$, we do not compute the mixed partial derivative directly from the representation just obtained for $F_\epsilon(\xi,0)$. Rather, we return to the definition of F and compute the partial derivatives from the formula

$$F(\xi,\epsilon)=\mathbf{X}_0^\perp(\xi,y_0)\cdot\mathbf{H}(\xi,\epsilon).$$

Since $\mathbf{X}_0^\perp$ does not depend on ϵ, we have

$$F_{\xi\epsilon}(\xi,0)=\mathbf{X}_0^\perp\cdot\mathbf{H}_{\xi\epsilon}+\mathbf{X}_{0\xi}^\perp\cdot\mathbf{H}_\epsilon(\xi,0).$$

But, by hypothesis, $\mathbf{H}_\epsilon(\xi_0,0)=0$. So the required derivative is given by

$$F_{\xi\epsilon}(\xi_0,0)=\mathbf{X}_0^\perp(\xi_0,y_0)\cdot\mathbf{H}_{\xi\epsilon}(\xi_0,0).$$

To compute the partial derivatives of $\mathbf{H}$ we use the previously given representation of the scalar return map. In the present case, this takes the form

$$h_\xi(\xi,\epsilon)=\frac{X_2(\xi,y_0,\epsilon)}{X_2(h(\xi,\epsilon),y_0,\epsilon)}\exp\left(\int_0^{T(\xi,\epsilon)}\operatorname{div}\mathbf{X}(\phi_t^\epsilon(\xi,y_0),\epsilon)\,dt\right).$$

Since

$$\mathbf{H}_{\xi\epsilon}(\xi_0,0)=(h_{\epsilon\xi}(\xi_0,0),0),$$

we need only calculate $h_{\epsilon\xi}$. First, using the hypotheses, $h(\xi,0)=\xi$ and $h_\epsilon(\xi,0)=0$, it is easy to verify that the derivative of the first factor of $h_\xi(\xi,\epsilon)$ with respect to ϵ vanishes at $\epsilon=0$, and the value of this factor at $(\xi_0,0)$ is unity. Thus, the required derivative is given by

$$h_{\epsilon\xi}(\xi_0,0)=\exp\left(\int_0^T\operatorname{div}\mathbf{X}_0(\gamma(t))\,dt\right)\left\{\operatorname{div}\mathbf{X}_0(\xi_0,y_0)T_\epsilon(\xi_0,0)\right.$$
$$\left.+\int_0^{T(\xi_0,0)}\frac{d}{d\epsilon}\operatorname{div}\mathbf{X}(\phi_t^\epsilon(\xi_0,y_0),\epsilon)\Big|_{\epsilon=0}dt\right\}.$$

Since the characteristic exponent of Γ vanishes, the exponential term is unity. We also have

$$\frac{d}{d\epsilon}\operatorname{div}\mathbf{X}(\phi_t^\epsilon(\xi_0,y_0),\epsilon)\Big|_{\epsilon=0}=\operatorname{div}(p,q)(\phi_t^0(\xi_0,y_0))+\frac{d}{d\epsilon}\operatorname{div}\mathbf{X}_0(\phi_t^\epsilon(\xi_0,y_0))\Big|_{\epsilon=0},$$

and it follows that

$$h_{\epsilon\xi}(\xi_0,0)=\operatorname{div}\mathbf{X}_0(\xi_0,y_0)T_\epsilon(\xi_0,0)+\int_0^{T(\xi_0,0)}\operatorname{div}(p,q)(\gamma(t))\,dt$$
$$+\int_0^{T(\xi_0,0)}\frac{d}{d\epsilon}\operatorname{div}\mathbf{X}_0(\phi_t^\epsilon(\xi_0,y_0))\Big|_{\epsilon=0}dt$$

as required.

The proof of (iii) is just an application of Green's theorem. □

There is a large literature on the applications of the theorem. In most of these applications the most difficult problem is the computation of the integral for $F_\epsilon(\xi, 0)$ and the determination of its simple zeros. Some examples of such computations can be found in the references [1], [5], [7], [8], [10], [13], [19], [21], [27], [28], [29], [35], [37], [39], [43], [44]. A classic, but very simple application, can be made for the van der Pol oscillator with small damping. For this we have the system

$$\dot{x} = y, \qquad \dot{y} = -x + \epsilon(1 - x^2)y.$$

Here the unperturbed system is linear, and the period annulus fills the entire punctured plane. The positive x-axis is a Poincaré section and, using the theorem, we compute

$$\begin{aligned} F_\epsilon(\xi, 0) &= \int_0^{2\pi} (1 - x^2)y^2 \, dt \\ &= \int_0^{2\pi} (1 - \xi^2 \cos^2 t)\xi^2 \sin^2 t \, dt \\ &= -\frac{\pi}{4}\xi^2(\xi^2 - 4). \end{aligned}$$

The bifurcation function has a unique simple zero at $\xi = 2$, and

$$\epsilon \frac{F_{\epsilon\xi}(2, 0)}{Q(2, 0)} = 2\epsilon\pi.$$

Thus, there is a continuous family of limit cycles emerging from the periodic trajectory $x(t) = 2\cos t \;\; y(t) = -2\sin t$ of the unperturbed system such that the family consists of stable limit cycles for $\epsilon > 0$ and unstable limit cycles for $\epsilon < 0$.

4. Subharmonic bifurcation theory.

In this section we consider bifurcation to periodic solutions in the family E_ϵ of planar differential equations

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon \mathbf{G}(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

where both $\mathbf{f}$ and $\mathbf{G}$ are smooth functions; $\mathbf{G}$ is an η-periodic function, i.e.,

$$\mathbf{G}(\mathbf{x}, t + \eta, \epsilon) = \mathbf{G}(\mathbf{x}, t, \epsilon)$$

for all $\mathbf{x}$, t, and ϵ, and where $\mathbf{G}$ has the form

$$\mathbf{G}(\mathbf{x}, t, \epsilon) = \mathbf{g}(\mathbf{x}, t) + \epsilon \mathbf{g}_R(\mathbf{x}, t, \epsilon)$$

with both $\mathbf{g}$ and $\mathbf{g}_R$ smooth functions of the indicated variables. We are interested in the bifurcation of periodic orbits from periodic trajectories Γ of the unperturbed system E_0 as $|\epsilon|$ increases from zero. For this we make one further assumption. The period of Γ is in $m : n$ *resonance* with the period of the forcing function $\mathbf{G}$, i.e., there are relatively prime positive integers m and n such that the period of Γ is $m\eta/n$. At the end of this section we show how to relax this assumption in the presence of a "detuning."

The idea of the bifurcation theory is to find the periodic trajectories as fixed points of the *parameterized Poincaré map* $\mathbf{P} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$ given by $(\xi, \epsilon) \mapsto \mathbf{x}(m\eta, \xi, \epsilon)$, where $t \mapsto \mathbf{x}(t, \xi, \epsilon)$ is the solution of E_ϵ satisfying the initial condition $\mathbf{x}(0, \xi, \epsilon) = \xi$. In this interpretation, ξ is in the section $\Sigma = \mathbb{R}^2 \times 0$ for the flow considered on the manifold diffeomorphic to $\mathbb{R}^2 \times S^1$ obtained by identification of the time modulo $m\eta$. The basic property of the Poincaré map is that, for fixed ϵ, a solution of E_ϵ that starts at a fixed point of the function $\xi \mapsto \mathbf{P}(\xi, \epsilon)$ is a periodic solution of E_ϵ. For example, suppose $\mathbf{P}(\xi_0, \epsilon) = \xi_0$ and $\mathbf{x}(t, \xi_0, \epsilon)$ satisfies $\mathbf{x}(0, \xi_0, \epsilon) = \xi_0$. If we define $\mathbf{y}(t) := \mathbf{x}(t + m\eta, \xi_0, \epsilon)$, then

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{f}(\mathbf{y}) + \epsilon \mathbf{G}(\mathbf{y}, t + m\eta, \epsilon) \\ &= \mathbf{f}(\mathbf{y}) + \epsilon \mathbf{G}(\mathbf{y}, t, \epsilon). \end{aligned}$$

Thus, $\mathbf{y}$ is a solution of E_ϵ with the initial condition $\mathbf{y}(0) = \mathbf{x}(m\eta, \xi_0, \epsilon) = \xi_0$ and, by uniqueness of the solutions of E_ϵ, we have $\mathbf{x}(t, \xi_0, \epsilon) = \mathbf{y}(t, \xi_0, \epsilon)$. It follows that $\mathbf{x}$ is a periodic function of its first variable. Also, if $\mathbf{x}$ is not a constant solution of E_ϵ, the minimum period of $\mathbf{x}$ must be $m\eta/k$ for some positive integer k.

The Poincaré map is easy to compute on the resonant orbit Γ. Since, for the flow ϕ_t of the unperturbed system, we have

$$\phi_{m\eta}(\xi) = \phi_{n(m\eta/n)}(\xi) = \xi;$$

the function $\xi \mapsto \mathbf{P}(\xi, 0)$ is the identity on Γ. We wish to find conditions on the functions $\mathbf{f}$ and $\mathbf{G}$ such that, for sufficiently small $\epsilon \neq 0$, some of these fixed points remain. In order to state the bifurcation theorems that provide these conditions, we will need a few more definitions. We identify $\Sigma = \mathbb{R}^2 \times 0$ with $\mathbb{R}^2$ and, for $\xi \in \Sigma$, we define the *displacement function* $\delta(\xi, \epsilon) := \mathbf{P}(\xi, \epsilon) - \xi$ together with its *radial projection* $\rho(\xi, \epsilon) := \langle \delta(\xi, \epsilon), \mathbf{f}^\perp(\xi) \rangle$ and its *tangential projection* $\tau(\xi, \epsilon) := \langle \delta(\xi, \epsilon), \mathbf{f}(\xi) \rangle$, where $\langle , \rangle$ denotes the usual inner product on $\mathbb{R}^2$. For Γ a periodic trajectory of the unperturbed system whose period is in resonance with the external periodic force $\mathbf{G}$, we say $\xi \in \Gamma$ is a *subharmonic branch point* if there is an $\epsilon_0 > 0$ and a curve, $\epsilon \mapsto \sigma(\epsilon)$, defined for $|\epsilon| < \epsilon_0$, with image in the section Σ, such that $\sigma(0) = \xi$ and $\delta(\sigma(\epsilon), \epsilon) \equiv 0$. Of course, if $\delta(\sigma(\epsilon), \epsilon) = 0$, then $\sigma(\epsilon) \in \Sigma$ is the initial value for a periodic solution of E_ϵ. When the unperturbed periodic solution Γ is in $m : 1$ resonance with the forcing, the resonance is called *subharmonic of order m* and the perturbed periodic solutions are called subharmonics. Subharmonic resonance of order one is also called *harmonic.* This is the reason for the use of the term "subharmonic" in the definition of subharmonic branch points. However, the bifurcating solutions at a subharmonic branch point may not be, in the strict sense of the term, subharmonics. For example, if $n \neq 1$, a $1 : n$ resonance is called *ultraharmonic* and an $m : n$ resonance is called *ultrasubharmonic.* See [43, pp. 73–78] for a geometric view of these solutions.

As in the Poincaré–Andronov theorem, the bifurcation analysis for subharmonic branch points consists of two main steps: reduction and identification. Here, reduction refers to reducing the problem of the existence of a curve of subharmonics bifurcating from Γ to an application of the implicit function theorem, i.e., to the nonvanishing of a certain partial derivative, while identification refers to finding an explicit formula for this derivative in terms of the functions $\mathbf{f}$ and $\mathbf{G}$. For this we will use throughout the functions defined in §2 and given by

$$\beta(t) := \beta(t, \xi) := \exp\left(\int_0^t \operatorname{div} \mathbf{f}(\phi_s(\xi))\, ds \right),$$

$$\alpha(t) := \alpha(t, \xi) := \int_0^t \left\{ \frac{1}{\|\mathbf{f}\|^2} \left[2\kappa \|\mathbf{f}\| - \operatorname{curl} \mathbf{f} \right] \right\} (\phi_\tau(\xi)) \beta(\tau)\, d\tau,$$

$$\mathcal{N}(\xi) := \int_0^{m\eta} \left\{ \frac{1}{\|\mathbf{f}\|^2} \langle \mathbf{f}, \mathbf{g} \rangle - \frac{\alpha(s)}{\beta(s)} \mathbf{f} \wedge \mathbf{g} \right\} (\phi_s(\xi))\, ds,$$
$$\mathcal{M}(\xi) := \int_0^{m\eta} \left\{ \frac{1}{\beta(s)} \mathbf{f} \wedge \mathbf{g} \right\} (\phi_s(\xi))\, ds.$$

Also, we will compute several times the derivatives of functions of the two variables ξ and ϵ. We will use the convention that derivatives indicated by D, for functions with range in $\mathbb{R}^2$, or by d, for functions with range in $\mathbb{R}$, refer to the derivative with respect to the space variable ξ, while derivatives indicated by a subscripted variable refer to the partial derivative with respect to that variable.

To begin the analysis, we expand the displacement function into its perturbation series

$$\begin{aligned}\delta(\xi, \epsilon) &= \mathbf{P}(\xi, 0) - \xi + \mathbf{P}_\epsilon(\xi, 0)\epsilon + O(\epsilon^2) \\ &= \mathbf{x}(m\eta, \xi, 0) - \xi + \mathbf{x}_\epsilon(m\eta, \xi, 0)\epsilon + O(\epsilon^2),\end{aligned}$$

and we recall that, by the variation lemma, the first-order term of the perturbation series is given by

$$\mathbf{x}_\epsilon = (\mathcal{N} + \alpha\mathcal{M})\mathbf{f} + \beta\mathcal{M}\mathbf{u}_{\mathbf{f}^\perp}.$$

For ξ on the resonant orbit Γ, $\delta(\xi, 0) \equiv 0$, and, consequently,

$$D\delta(\xi, 0)(\mathbf{f}(\xi)) = \frac{d}{dt}\mathbf{P}(\phi_t(\xi), 0)\Big|_{t=0} - \mathbf{f}(\xi) = 0.$$

Thus, $\xi \mapsto D\delta(\xi, 0)$ is not invertible and we cannot use the implicit function theorem directly. However, we can use various forms of the Lyapunov–Schmidt reduction depending on how degenerate the curve Γ is as part of the zero set of the function $\xi \to \delta(\xi, 0)$.

The most degenerate case is the classical case when the unperturbed system is linear and $\delta(\xi, 0) \equiv 0$. Actually, this degeneracy is equivalent to assuming Γ is contained in an isochronous period annulus; cf. [13]. In this case, we have the following proposition.

PROPOSITION 4.1. *Let* $\mathbf{P}$ *denote the parameterized Poincaré map for the system* E_ϵ,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon\mathbf{G}(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

where $\mathbf{G}$ *is* η*-periodic of the form*

$$\mathbf{G}(\mathbf{x}, t, \epsilon) = \mathbf{g}(\mathbf{x}, t) + \epsilon\mathbf{g}_R(x, t, \epsilon),$$

and assume that Γ *is a periodic solution of the unperturbed system that belongs to an isochronous period annulus.*

Reduction. If there are positive integers m and n such that the period of Γ *is equal to* $m\eta/n$ *and if the function* $\xi \mapsto \mathbf{P}_\epsilon(\xi, 0)$ *has a simple zero at* ξ_0, *i.e.,*

$$\mathbf{P}_\epsilon(\xi_0, 0) = 0 \quad \textit{and} \quad \det(D\mathbf{P}_\epsilon(\xi_0, 0)) \neq 0,$$

then ξ_0 *is a subharmonic branch point.*

Identification. The bifurcation function $\xi \mapsto \mathbf{P}_\epsilon(\xi,0)$ *is given by*

$$\mathbf{P}_\epsilon(\xi,0) = \left[\mathcal{N}\mathbf{f} + \frac{1}{||\mathbf{f}||^2}\mathcal{M}\mathbf{f}^\perp\right](\xi).$$

Moreover, in case the unperturbed system is linear with $\mathbf{f}(x_1,x_2) = (\omega x_2, -\omega x_1)$ *and if* $2\pi n/\omega = m\eta$, *then the bifurcation function is given by*

$$\mathbf{P}_\epsilon(\xi,0) = \frac{1}{||\xi||^2}\left(\xi_1 I_2(\xi) + \xi_2 I_1(\xi),\ \xi_2 I_2(\xi) - \xi_1 I_1(\xi)\right),$$

where

$$I_1(\xi) := \int_0^{2\pi n/\omega} x_2 g_1(\mathbf{x},t) - x_1 g_2(\mathbf{x},t)\,dt, \quad I_2(\xi) := \int_0^{2\pi n/\omega} x_1 g_1(\mathbf{x},t) + x_2 g_2(\mathbf{x},t)\,dt.$$

Proof. By the hypotheses, the displacement function, $\delta(\xi,\epsilon) := \mathbf{P}(\xi,\epsilon) - \xi$, for the perturbed system can be represented in the form

$$\delta(\xi,\epsilon) = \epsilon\left[\mathbf{P}_\epsilon(\xi,0) + O(\epsilon)\right].$$

Therefore, the implicit function theorem can be applied to determine when there is an implicit solution of the equation $\mathbf{P}_\epsilon(\xi,0) + O(\epsilon) = 0$ at some point $(\xi_0, 0)$. This proves the reduction statement of the proposition. For the identification we simply observe that Γ is not hyperbolic and that the period function on the period annulus is constant. Then, using the results of §2, we have $\beta(m\eta,\xi) \equiv 1$ and $\alpha(m\eta,\xi) \equiv 0$ for $\xi \in \Gamma$. Thus, from the formula for $\mathbf{x}_\epsilon$ given above, we obtain the desired result.

In case the unperturbed vector field $\mathbf{f}$ is linear, with $\mathbf{f}(x_1,x_2) = (\omega x_2, -\omega x_1)$, the punctured phase plane of the unperturbed system is filled by periodic trajectories, each of which lies on a circle centered at the origin, and the hypotheses of the proposition hold. If, in addition, we assume the period of the external excitation is in resonance with the linear system, i.e., there are positive integers m and n such that $2\pi n/\omega = m\eta$, then the formula for $\mathbf{x}_\epsilon(m\eta,\xi,0)$ reduces to

$$\frac{1}{\omega^2||\xi||^2}\left[\left(\int_0^{m\eta}\langle \mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x},t)\rangle\,dt\right)\mathbf{f}(\xi) + \left(\omega\int_0^{m\eta}\mathbf{f}(\mathbf{x}) \wedge \mathbf{g}(\mathbf{x},t)\,dt\right)\xi\right],$$

where, in components, $\xi = (\xi_1, \xi_2)$ and

$$\mathbf{x}(t) := (x_1(t), x_2(t)) = (\xi_1\cos(\omega t) + \xi_2\sin(\omega t),\ -\xi_1\sin(\omega t) + \xi_2\cos(\omega t)).$$

Moreover, if the components of the external excitation are given by

$$\mathbf{g}(\mathbf{x},t) := (g_1(\mathbf{x},t), g_2(\mathbf{x},t)),$$

and we define

$$I_1(\xi) := \int_0^{2\pi n/\omega} x_2 g_1(\mathbf{x},t) - x_1 g_2(\mathbf{x},t)\,dt, \quad I_2(\xi) := \int_0^{2\pi n/\omega} x_1 g_1(\mathbf{x},t) + x_2 g_2(\mathbf{x},t)\,dt,$$

we then have

$$\mathbf{x}_\epsilon(m\eta,\xi,0) = \frac{1}{||\xi||^2}\left(\xi_1 I_2(\xi) + \xi_2 I_1(\xi),\ \xi_2 I_2(\xi) - \xi_1 I_1(\xi)\right)$$

as required. □

For computational purposes, it is convenient to define $\mathcal{I}(\xi) := (I_1(\xi), I_2(\xi))$, and to observe that

$$\mathbf{P}_\epsilon(\xi, 0) = \begin{pmatrix} \xi_2 & \xi_1 \\ -\xi_1 & \xi_2 \end{pmatrix} \begin{pmatrix} I_1(\xi) \\ I_2(\xi) \end{pmatrix}.$$

Then, the existence of a simple zero of $\mathbf{P}_\epsilon(\xi, 0)$ is easily seen to be equivalent to the existence of a simple zero of $\mathcal{I}$. A final special case, where we take $\omega = 1$ and $g_1(\mathbf{x}, t) \equiv 0$, results in the identification

$$\mathbf{P}_\epsilon(\xi, 0) = \left(-\int_0^{2\pi n} g_2(\mathbf{x}(t), t)\sin t\,dt,\ \int_0^{2\pi n} g_2(\mathbf{x}(t), t)\cos t\,dt\right).$$

Here, the components of $\mathbf{P}_\epsilon(\xi, 0)$ are given by the same formulas obtained from the classical perturbation series methods. See [27, XII, §2] for a version of the classic approach to these formulas and for the computations, using the same formulas, showing the existence of a unique stable harmonic solution of the forced van der Pol oscillator,

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x + \epsilon(1 - x^2)y + \epsilon a \sin t. \end{aligned}$$

When $\xi \to \delta(\xi, 0)$ is not identically zero on some neighborhood of Γ, the bifurcation theory must deal with the zero order terms of the perturbation expansion of the displacement function. The least degenerate case occurs when the kernel $\mathcal{K}$ of the map $D\delta(\xi, 0) : \mathbb{R}^2 \to \mathbb{R}^2$ is one-dimensional. We have already shown $\mathbf{f}(\xi) \in \mathcal{K}$. Thus, this nondegeneracy condition will be satisfied when $\mathbf{f}^\perp(\xi) \notin \mathcal{K}$. It turns out that this nondegeneracy condition is equivalent to having one of the following: either Γ is hyperbolic or, at $\xi \in \Gamma$, the transit time map on a section in the phase plane orthogonal to Γ has a nonzero derivative along the section at ξ. The second possibility is the only way to have nondegeneracy when Γ is in a period annulus. In this case, the transit time reduces to the period function. Of course, if the period function has a zero derivative "at Γ" on a section transverse to Γ, then this derivative will be zero on every such section. When the derivative of the period function vanishes in this way, Γ is called *critical.* However, it is worth noting that the vanishing of the derivative of the transit time along a section intersecting a limit cycle depends on the choice of section.

For the next theorem we assume that the unperturbed system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ has flow ϕ_t and a noncritical periodic trajectory Γ given by $t \mapsto \phi_t(p)$ that lies in a period annulus $\mathcal{A} \subset \Sigma$. We let $T : \mathcal{A} \to \mathbb{R}$ denote the *period function* for the unperturbed system on the period annulus $\mathcal{A}$; it assigns to each $\xi \in \mathcal{A} \subset \Sigma$ the minimum period of the periodic trajectory of the unperturbed system passing through ξ. Also, the *subharmonic Melnikov function* is defined for $\xi \in \mathcal{A}$, in terms of the radial projection of the displacement function ρ, by

$$M(\xi) := \rho_\epsilon(\xi, 0).$$

THEOREM 4.2 (subharmonic bifurcation theorem). *Let E_ϵ denote the parameterized family of differential equations*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon\mathbf{g}(\mathbf{x}, t) + \epsilon^2\mathbf{g}_R(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

such that E_0 has flow $t \mapsto \phi_t$, a period annulus $\mathcal{A}$, and a periodic trajectory $\Gamma \subset \mathcal{A}$ that is in resonance with the η-periodic external force $\mathbf{G}(x,t,\epsilon) := \mathbf{g}(\mathbf{x},t) + \epsilon\mathbf{g}_R(\mathbf{x},t,\epsilon)$, i.e., there are relatively prime natural numbers m and n such that the period of Γ is $m\eta/n$.

(i) *Reduction. If Γ is not critical and $\xi \in \Gamma$ is a simple zero of the subharmonic Melnikov function, i.e., $dT(\mathbf{f}^\perp)(\xi) \neq 0$, $M(\xi) = 0$ and $dM(\mathbf{f})(\xi) \neq 0$, then ξ is a subharmonic branch point.*

(ii) *Identification. The directional derivative of the period function T in the direction $\mathbf{f}^\perp(\xi)$ is given by*

$$dT(\mathbf{f}^\perp)(\xi) = \\ -\|\mathbf{f}(\xi)\|^2 \int_0^{T(\xi)} \left\{ \frac{1}{\|\mathbf{f}\|^2} [2\kappa\|\mathbf{f}\| - \operatorname{curl}\mathbf{f}] \right\} (\phi_s(\xi)) \exp\left(\int_0^s \operatorname{div}\mathbf{f}(\phi_t(\xi))\,dt \right) ds,$$

and the subharmonic Melnikov function is given by

$$M(\xi) = \mathcal{M}(m\eta, \mathbf{f}, \mathbf{g}, \xi) = \int_0^{m\eta} \exp\left\{ -\int_0^t \operatorname{div}\ \mathbf{f}(\phi_s(\xi))\,ds \right\} \mathbf{f}(\phi_t(\xi)) \wedge \mathbf{g}(\phi_t(\xi), t)\,dt.$$

Proof. We have

$$\delta(\xi,\epsilon) = \mathbf{P}(\xi,0) - \xi + \mathbf{P}_\epsilon(\xi,0)\epsilon + O(\epsilon^2).$$

For ξ on the resonant orbit Γ, $\delta(\xi,0) \equiv 0$, and, as we have seen,

$$D\delta(\xi,0)(\mathbf{f}(\xi)) = \frac{d}{dt}\mathbf{P}(\phi_t(\xi),0)\Big|_{t=0} - \mathbf{f}(\xi) = 0.$$

Thus, $\xi \mapsto D\delta(\xi,0)$ is not invertible, and we cannot use the implicit function theorem directly. However, the Lyapunov–Schmidt reduction can be employed. First, we apply the implicit function theorem to the tangential projection τ. For this, let $\xi \in \Gamma$, and compute the directional derivative of $\xi \mapsto \tau(\xi,0)$ in the direction $\mathbf{f}^\perp$ to obtain

$$d\tau(\xi,0)(\mathbf{f}^\perp(\xi)) = \left\langle D\delta(\xi,0)\mathbf{f}^\perp(\xi), \mathbf{f}(\xi) \right\rangle + \left\langle \delta(\xi,0), D\mathbf{f}(\xi)\mathbf{f}^\perp(\xi) \right\rangle.$$

Since $\delta(\xi,0) \equiv 0$ on Γ, the formula for this derivative reduces to

$$d\tau(\xi,0)(\mathbf{f}^\perp(\xi)) = \left\langle D\delta(\xi,0)\mathbf{f}^\perp(\xi), \mathbf{f}(\xi) \right\rangle.$$

In order to compute $D\delta(\xi,0)\mathbf{f}^\perp(\xi)$, recall

$$DP(\xi,0)\mathbf{f}^\perp(\xi) = D\phi_{m\eta}(\xi)\mathbf{f}^\perp(\xi),$$

and let $\mathbf{V}$ be as defined in Diliberto's theorem. Since both

$$t \mapsto D\phi_t(\xi)\frac{1}{\|\mathbf{f}(\xi)\|^2}\mathbf{f}^\perp(\xi) \quad \text{and} \quad t \mapsto \mathbf{V}(t,\mathbf{f},\xi)$$

are solutions of the homogeneous variational equation for the unperturbed system $\dot{x} = \mathbf{f}(x)$ along the trajectory $t \mapsto \phi_t(\xi)$ satisfying the same initial condition, the two functions are equal, and we have

$$D\phi_{m\eta}(\xi)\mathbf{f}^\perp(\xi) = \|\mathbf{f}(\xi)\|^2\mathbf{V}(m\eta) = \|\mathbf{f}(\xi)\|^2 \left\{ \alpha(m\eta)\mathbf{f}(\xi) + \frac{\beta(m\eta)}{\|\mathbf{f}(\xi)\|^2}\mathbf{f}^\perp(\xi) \right\}.$$

In view of the representation of the period function given in the variation lemma and the nonhyperbolicity of Γ, this formula can be expressed more concisely as

$$D\phi_{m\eta}(\xi)\mathbf{f}^{\perp}(\xi) = -\left[m\, dT(\mathbf{f}^{\perp})(\xi)\right]\mathbf{f}(\xi) + \mathbf{f}^{\perp}(\xi),$$

and, in turn, we have a simple expression for the directional derivative of δ :

$$D\delta(\xi,0)\mathbf{f}^{\perp}(\xi) = (D\mathbf{P}(\xi,0) - I)\mathbf{f}^{\perp}(\xi) = -\left[m\, dT(\xi)(\mathbf{f}^{\perp}(\xi))\right]\mathbf{f}(\xi).$$

Taking the inner product with $\mathbf{f}(\xi)$, we find the required directional derivative of τ to be

$$d\tau(\xi,0)(\mathbf{f}^{\perp}(\xi)) = -m\,||\mathbf{f}(\xi)||^2\, dT(\xi)(\mathbf{f}^{\perp}(\xi)) \neq 0.$$

Now, by an application of the implicit function theorem to the function τ, we conclude that there is a smooth two-dimensional surface $\mathcal{S}$ in the (ξ,ϵ)-space passing through the curve $\Gamma \times 0$ such that τ vanishes on $\mathcal{S}$. Since, in addition, for $\xi \in \Gamma$ we have

$$d\tau(\xi,0)\mathbf{f}(\xi) = \langle (D\mathbf{P}(\xi,0) - I)\mathbf{f}(\xi), \mathbf{f}(\xi)\rangle \equiv 0,$$

$\mathcal{S}$ is transverse to the section Σ and $\Gamma \subset \mathcal{S}$.

To complete the reduction we restrict our attention to the manifold $\mathcal{S}$. To be more precise, we consider a neighborhood of the point $(\xi_0,0)$ on Γ. There is a local coordinate chart (U,φ_U) on $\mathcal{S}$ such that U is a product neighborhood in $\mathbb{R}\times\mathbb{R}$ and $\varphi_U : U \to \mathcal{S} \subset \mathbb{R}^2 \times \mathbb{R}$ is a smooth function that can be taken to have the form

$$\varphi_U(\theta,\epsilon) = (\varphi(\theta,\epsilon),\epsilon).$$

Here the image of the function $\theta \mapsto \varphi(\theta,0)$ is contained in Γ, and $\varphi(0,0) = \xi_0$. Now, we can view the restriction of the radial projection ρ to $\mathcal{S}$, $\rho_{\mathcal{S}}$, as the function defined by

$$\rho_{\mathcal{S}}(\theta,\epsilon) = \rho(\varphi_U(\theta,\epsilon)).$$

Since $\rho_{\mathcal{S}}(\theta,0) = \rho(\varphi(\theta,0),0) \equiv 0$, the restriction of ρ, represented locally by its Taylor polynomial with remainder, has the form

$$\rho_{\mathcal{S}}(\theta,\epsilon) = \rho_1(\theta)\epsilon + O(\epsilon^2)$$

on a product neighborhood of the origin in the (θ,ϵ)-space. The first-order Taylor coefficient is given by

$$\rho_1(\theta) = \frac{\partial\rho_{\mathcal{S}}}{\partial\epsilon}(\theta,0) = d\rho(\varphi(\theta,0),0)(\varphi_\epsilon(\theta,0)) + \rho_\epsilon(\varphi(\theta,0),0).$$

But, for $\xi \in \Gamma$, a computation similar to the computation made above for $d\tau$ shows

$$d\rho(\xi,0)(\mathbf{f}^{\perp}(\xi)) = d\rho(\xi,0)(\mathbf{f}(\xi)) = 0.$$

Thus, $d\rho(\xi,0) = 0$ and

$$\rho_1(\theta) = \rho_\epsilon(\varphi(\theta,0),0) = M(\varphi(\theta,0)).$$

The hypothesis $M(\xi_0) = 0$, but $dM(\xi_0)(f(\xi_0)) \neq 0$ implies $\rho_1(0) = 0$ and $\rho_1'(0) \neq 0$. Thus, there is an $\epsilon_0 > 0$ and a smooth function σ_0 defined for $|\epsilon| < \epsilon_0$, with range in the θ-space, such that $\rho_S(\sigma_0(\epsilon), \epsilon) \equiv 0$. If we define $\sigma(\epsilon) := (\varphi(\sigma_0(\epsilon), \epsilon), \epsilon)$, then

$$\rho(\sigma(\epsilon), \epsilon) \equiv \tau(\sigma(\epsilon), \epsilon) \equiv 0,$$

and we have $\delta(\sigma(\epsilon), \epsilon) \equiv 0$ as required to prove the reduction.

For the identification, the derivative of the period function can be computed directly from the variation lemma. In fact,

$$dT(\mathbf{f}^{\perp})(\xi) = -||\mathbf{f}(\xi)||^2 \alpha(T(\xi), \mathbf{f}, \xi).$$

The proof will be complete when we identify the Melnikov function. To do this, we compute $\rho_\epsilon(\xi, 0)$ as follows:

$$\begin{aligned} \rho_\epsilon(\xi, 0) &= \left\langle \delta_\epsilon(\xi, 0), \mathbf{f}^{\perp}(\xi) \right\rangle \\ &= \left\langle \mathbf{P}_\epsilon(\xi, 0), \mathbf{f}^{\perp}(\xi) \right\rangle \\ &= \left\langle \mathbf{x}_\epsilon(m\eta, \xi, 0), \mathbf{f}^{\perp}(\xi) \right\rangle. \end{aligned}$$

But, $\mathbf{x}_\epsilon(t, \xi, 0)$ is the solution of the variational initial value problem

$$\dot{\mathbf{W}} = D\mathbf{f}(\phi_t(\xi))\mathbf{W} + \mathbf{g}(\phi_t(\xi), t), \qquad \mathbf{W}(0) = 0.$$

By the variation lemma, this solution is given by

$$\mathbf{x}_\epsilon(m\eta, \xi, 0) = \beta(\xi)\mathcal{M}(\xi)\mathbf{u}_{\mathbf{f}^{\perp}}(\xi) \qquad (\text{mod } \mathbf{f}).$$

Using the fact that $\beta(\xi) \equiv 1$ and substitution of this expression into the formula for $\rho_\epsilon(\xi, 0)$ we obtain the desired result. □

When Γ is hyperbolic, the result is similar to the last theorem, but the formulas for the partial derivatives of the perturbation series are more complicated. In order to state our bifurcation theorem in this context, we again consider the system E_ϵ given by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon \mathbf{G}(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

where the external excitation G is periodic of period η in its second variable. We assume the unperturbed system, $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, has a limit cycle Γ whose period is in resonance with the period of the forcing function. In fact, we assume the period of Γ is $m\eta/n$ for m and n relatively prime positive integers.

Before stating our theorem, we pause to recall a trivial but important fact. If the limit cycle of the unperturbed system is hyperbolic, it is structurally stable in the class of plane vector fields. When we consider the forced oscillator on the manifold $\mathbb{R}^2 \times S^1$, the three-dimensional system of differential equations has, for $\epsilon = 0$, a normally hyperbolic torus corresponding to the limit cycle. The flow on this torus will be periodic or quasiperiodic depending on whether or not the resonance condition is satisfied. But, in either case, the orbits corresponding to the limit cycle will no longer be structurally stable; the stability of the limit cycle has been "transferred to the torus." In the resonant case, if we consider the appropriate iterate of the Poincaré map, we will have the Poincaré map reduced to the identity on the torus. Thus, it is natural to ask if any of the fixed points of the Poincaré map, corresponding to the points on the limit cycle, survive after perturbation as fixed points or higher order periodic points of the Poincaré map, i.e., if any periodic solutions survive as harmonic or subharmonic solutions of the forced oscillator.

For our bifurcation result, the definitions of the Poincaré section Σ, the parameterized Poincaré map $\mathbf{P}$ and the displacement function δ remain unchanged. We consider $\mathbf{f}$ and $\mathbf{G}$ fixed, with

$$\mathbf{G}(\mathbf{x}, t, \epsilon) = \mathbf{g}(\mathbf{x}, t) + \epsilon \mathbf{g}_R(\mathbf{x}, t, \epsilon),$$

and, for notational convenience, when we refer to the functions α, β, $\mathcal{M}$, and $\mathcal{N}$ with the single argument ξ, we understand that the remaining arguments are fixed at $t = m\eta$, $\mathbf{f}$, and $\mathbf{g}$. In order to obtain our bifurcation result, a correction term must be added to the subharmonic Melnikov function. In fact, the new function we require is defined by

$$\mathcal{C}(\xi) := [(1-\beta)\mathcal{N} + \alpha\mathcal{M}]\,(\xi).$$

We call $\mathcal{C}$ the *subharmonic bifurcation function.* The next theorem applies either when Γ is hyperbolic or when, at the bifurcation point, the derivative of the transit time does not vanish.

THEOREM 4.3 (limit cycle subharmonic bifurcation theorem). *Let E_ϵ denote the parameterized family of differential equations*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon \mathbf{g}(\mathbf{x}, t) + \epsilon^2 \mathbf{g}_R(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

such that E_0 has a limit cycle Γ whose period is in resonance with the η-periodic external force $\mathbf{G}(x, t, \epsilon) := g(x, t) + \epsilon \mathbf{g}_R(x, t, \epsilon)$, i.e., there are relatively prime natural numbers m and n such that the period of Γ is $m\eta/n$. If Γ is hyperbolic and $\xi \in \Gamma$ is a simple zero of the subharmonic bifurcation function $\mathcal{C}$, i.e., $\mathcal{C}(\xi) = 0$ and $d\mathcal{C}(\mathbf{f})(\xi) \neq 0$, then ξ is a subharmonic branch point. Also, if $\xi \in \Gamma$ is a simple zero of the subharmonic bifurcation function and if $\alpha(\xi) \neq 0$, then ξ is a subharmonic branch point.

Proof. With the projections ρ and τ defined exactly as before and with $\xi \in \Gamma$, we have

$$\begin{aligned}
d\tau(\xi,0)(\mathbf{f}(\xi)) &= \langle (D\mathbf{P}(\xi,0) - I)\mathbf{f}(\xi), \mathbf{f}(\xi)\rangle = 0,\\
d\rho(\xi,0)(\mathbf{f}(\xi)) &= \left\langle (D\mathbf{P}(\xi,0) - I)\mathbf{f}(\xi), \mathbf{f}^{\perp}(\xi)\right\rangle = 0,\\
d\tau(\xi,0)(\mathbf{f}^{\perp}(\xi)) &= \left\langle ||\mathbf{f}||^2 \left\{\alpha\mathbf{f} + \frac{\beta}{||\mathbf{f}||^2}\mathbf{f}^{\perp}\right\} - \mathbf{f}^{\perp}, \mathbf{f}\right\rangle(\xi) = \left[||\mathbf{f}||^4\alpha\right](\xi),\\
d\rho(\xi,0)(\mathbf{f}^{\perp}(\xi)) &= \left\langle ||\mathbf{f}||^2 \left\{\alpha\mathbf{f} + \frac{\beta}{||\mathbf{f}||^2}\mathbf{f}^{\perp}\right\} - \mathbf{f}^{\perp}, \mathbf{f}^{\perp}\right\rangle(\xi) = \left[||\mathbf{f}||^2(\beta - 1)\right](\xi).
\end{aligned}$$

Thus, there are two choices. If Γ is hyperbolic, then $\beta \neq 1$, and we can apply the implicit function theorem to the radial projection to obtain a manifold $\mathcal{S}$ transverse to Σ such that ρ is identically zero on $\mathcal{S}$ and such that $\Gamma \subset \mathcal{S}$. If, on the other hand, $\alpha(\xi) \neq 0$, there is a manifold $\mathcal{S}$, defined locally in a neighborhood U of $(\xi, 0)$, such that τ is identically zero on $\mathcal{S}$, $\mathcal{S}$ is transverse to $\Sigma \cap U$, and $\Gamma \cap U \subset \mathcal{S}$. In both cases, the local coordinates are given by

$$(\theta, \epsilon) \mapsto (\varphi(\theta, \epsilon), \epsilon)$$

with $\varphi(\theta, 0) = \xi$. Thus, in the hyperbolic case we can restrict τ to $\mathcal{S}$ and obtain the representation

$$\tau_{\mathcal{S}}(\theta, \epsilon) := \tau_1(\theta)\epsilon + O(\epsilon^2),$$

while in case $\alpha(\xi) \neq 0$, we restrict the projection ρ to $\mathcal{S}$ to obtain

$$\rho_{\mathcal{S}}(\theta, \epsilon) = \rho_1(\theta)\epsilon + O(\epsilon^2).$$

The reduction portion of the theorem is the content of the following propositions. In the hyperbolic case, a simple zero of $\theta \to \tau_1(\theta)$ is a subharmonic branch point, while in the case $\alpha(\xi) \neq 0$, a simple zero of $\theta \to \rho_1(\theta)$ is a subharmonic branch point.

We will complete the proof by identification of the functions τ_1 and ρ_1. For this, note that

$$\begin{aligned}\tau_1(\theta) = \frac{\partial \tau_{\mathcal{S}}}{\partial \epsilon}(\theta, 0) &= d\tau(\varphi(\theta,0),0)(\varphi_\epsilon(\theta,0)) + \tau_\epsilon(\varphi(\theta,0),0)\\ &= d\tau(\xi,0)(\varphi_\epsilon(\theta,0)) + \tau_\epsilon(\xi,0)\end{aligned}$$

and

$$\begin{aligned}\rho_1(\theta) = \frac{\partial \rho_{\mathcal{S}}}{\partial \epsilon}(\theta, 0) &= d\rho(\varphi(\theta,0),0)(\varphi_\epsilon(\theta,0)) + \rho_\epsilon(\varphi(\theta,0),0)\\ &= d\rho(\xi,0)(\varphi_\epsilon(\theta,0)) + \rho_\epsilon(\xi,0).\end{aligned}$$

As in the last computation of the previous theorem, using the variation lemma and remembering that $\beta(\xi)$ may not be unity, we find

$$\tau_\epsilon(\xi,0) = ||\mathbf{f}||^2(\mathcal{N} + \alpha\mathcal{M})(\xi), \qquad \rho_\epsilon(\xi,0) = (\beta\mathcal{M})\,(\xi).$$

To compute the other terms we observe that $\varphi(\theta,\epsilon) \in \Sigma$. Hence, the vector field φ_ϵ can be expressed as a linear combination of $\mathbf{f}$ and $\mathbf{f}^\perp$ evaluated at $\varphi(\theta,\epsilon)$. In fact, there are scalars a and b (perhaps different in the two cases) such that

$$\varphi_\epsilon(\theta,0) = a\mathbf{f}(\xi) + b\mathbf{f}^\perp(\xi).$$

Moreover, since both $d\tau(\xi,0)\mathbf{f}(\xi) = 0$ and $d\rho(\xi,0)\mathbf{f}(\xi) = 0$, we have

$$d\tau(\xi,0)\varphi_\epsilon(\theta,0) = b||\mathbf{f}||^4\alpha(\xi), \qquad d\rho(\xi,0)\varphi_\epsilon(\theta,0) = \left[b(\beta-1)||\mathbf{f}||^2\right](\xi).$$

Now, after substitution, we obtain

$$\tau_1(\theta) = \left[b||\mathbf{f}||^4\alpha + ||\mathbf{f}||^2(\mathcal{N} + \alpha\mathcal{M})\right](\xi), \qquad \rho_1(\theta) = \left[b(\beta-1)||\mathbf{f}||^2 + \beta\mathcal{M}\right](\xi),$$

and it suffices, in each case, to compute b. For this, we recall that on $\mathcal{S}$, in the hyperbolic case,

$$\rho(\varphi(\theta,\epsilon),\epsilon) \equiv 0.$$

So

$$d\rho(\varphi(\theta,0),0)(\varphi_\epsilon(\theta,0)) + \rho_\epsilon(\varphi(\theta,0),0) = 0,$$

and we have

$$b\, d\rho(\xi,0)\mathbf{f}^\perp(\xi) = -\rho_\epsilon(\xi,0).$$

Thus, we can solve for b to obtain

$$b = \frac{\beta \mathcal{M}}{||\mathbf{f}||^2(1-\beta)},$$

and, after substitution into the last formula for τ_1, we compute

$$\tau_1(\theta) = \frac{||\mathbf{f}||^2}{1-\beta}\left[(1-\beta)\mathcal{N} + \alpha\mathcal{M}\right](\theta).$$

In case $\alpha(\xi) \neq 0$, the proof is similar. We find

$$b = -\frac{\tau_\epsilon(\xi, 0)}{||\mathbf{f}(\xi)||^4\alpha(\xi)}$$

and

$$\rho_1(\theta) = \frac{1}{\alpha}\left[(1-\beta)\mathcal{N} + \alpha\mathcal{M}\right](\theta)$$

as required. □

We have just considered subharmonic bifurcation from periodic trajectories of our unperturbed system in the most degenerate case, when the unperturbed system has an isochronous center, and in the least degenerate case, when the kernel of the space derivative of the displacement function at the unperturbed periodic orbit Γ is one-dimensional. A bifurcation theory for the cases of intermediate degeneracy can be carried out quite generally using our methods. However, since computation of the higher-order derivatives that appear in the analysis increase in complexity, we are content to illustrate the analysis in the least degenerate of the remaining cases. For this, it should now be clear that there are two possibilities depending on whether or not Γ is a limit cycle. If Γ is not a limit cycle, it belongs to a period annulus. This means $\beta(\xi) \equiv 1$ for ξ in this period annulus, and the degenerate case is $\alpha(\xi) \equiv 0$ but $d\alpha(\xi)\mathbf{f}^\perp(\xi) \neq 0$ for $\xi \in \Gamma$, i.e., the derivative of the period function vanishes on Γ but its second derivative does not vanish. This case has been treated by different methods when the unperturbed system is Hamiltonian in [38] and in a more abstract setting in [24]. In case Γ is a nonhyperbolic limit cycle, say for $\xi \in \Gamma$, $\beta(\xi) \equiv 1$ but $d\beta(\xi)\mathbf{f}^\perp(\xi) \neq 0$, we already know the bifurcation is not degenerate at a point $\xi \in \Gamma$ where $\alpha(\xi) \neq 0$. So the bifurcation is degenerate when this is not the case, i.e., when Γ is a nonhyperbolic limit cycle of multiplicity 2 (cf. [1, p. 272]), which contains points where the derivative of the transit time map on orthogonal sections vanishes. For these cases there is the possibility that more than one family of subharmonics is found near a subharmonic branch point. We will say $\xi \in \Gamma$ is a *subharmonic branch point with n-branches* if there is an $\epsilon_0 > 0$ and distinct (germs of) curves (at $\epsilon = 0$), $\epsilon \mapsto \sigma_k(\epsilon)$; $k = 1, \cdots, n$, each defined either for $\epsilon_0 < \epsilon \leq 0$ or $0 \leq \epsilon < \epsilon_0$, and each with image in the section Σ, such that $\sigma_k(0) = \xi$ and $\delta(\sigma_k(\epsilon), \epsilon) \equiv 0$. The next theorem gives the result for the case of the period annulus.

THEOREM 4.4 (order 2 subharmonic bifurcation theorem). *Let E_ϵ denote the parameterized family of differential equations*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon\mathbf{g}(\mathbf{x}, t) + \epsilon^2\mathbf{g}_R(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

such that E_0 has a period annulus $\mathcal{A}$ and a periodic trajectory $\Gamma \subset \mathcal{A}$ that is in resonance with the η-periodic external force $\mathbf{G}$, i.e., there are relatively prime natural *numbers m and n such that the period of Γ is $m\eta/n$. If Γ is critical ($\alpha(\xi) \equiv 0$) and if $\xi \in \Gamma$ is a simple zero of the subharmonic Melnikov function, such that*

$$\mathcal{N}(\xi) d\alpha(\xi) \mathbf{f}^{\perp}(\xi) \neq 0,$$

then ξ is a subharmonic branch point with two branches. Moreover, these two branches exist only in the direction of ϵ such that

$$\epsilon \mathcal{N}(\xi) d\alpha(\xi) \mathbf{f}^{\perp}(\xi) < 0.$$

Proof. Fix $\xi \in \Gamma$; let $t \to \phi_t$ denote the flow of $\dot{x} = \mathbf{f}(x)$; let $s \to \psi_s$ denote the flow of $\dot{x} = \mathbf{f}^{\perp}(x)$, and consider the local coordinates defined at $(\xi, 0)$ by the transformation

$$(s, t, \epsilon) \to (\psi_s(\phi_t(\xi)), \epsilon).$$

In these coordinates we have

$$\tau_{\mathrm{loc}}(s, t, \epsilon) := \tau(\psi_s(\phi_t(\xi)), \epsilon).$$

However, for notational convenience, we will write τ for τ_{loc}. Using this convention, we see immediately that $\tau(0, t, 0) \equiv 0$. Also, since

$$\tau_s(s, t, 0) = ||\mathbf{f}(\psi_s(\phi_t(\xi)))||^4 \alpha(\psi_s(\phi_t(\xi))),$$

we have

$$\tau_s(0, t, 0) = ||\mathbf{f}(\phi_t(\xi))||^4 \alpha(\phi_t(\xi)) \equiv 0,$$

and, for each $t \in \mathbb{R}$, we compute

$$\tau_{ss}(0, t, 0) = ||\mathbf{f}(\phi_t(\xi))||^4 d\alpha(\phi_t(\xi)) \mathbf{f}^{\perp}(\phi_t(\xi)) \neq 0.$$

By an application of the (Weierstrass) preparation theorem,

$$\tau(s, t, \epsilon) = \left(a(t, \epsilon) + b(t, \epsilon)s + s^2\right) u(s, t, \epsilon)$$

for functions a, b, and u, of the indicated variables, which satisfy

$$a(t, 0) \equiv 0, \quad b(t, 0) \equiv 0, \quad u(0, t, 0) \neq 0.$$

In particular, there are functions $t \to a_1(t)$ and $t \to b_1(t)$ such that

$$a(t, \epsilon) = a_1(t)\epsilon + O(\epsilon^2), \qquad b(t, \epsilon) = b_1(t)\epsilon + O(\epsilon^2),$$

and we have $\tau(s(t, \epsilon), t, \epsilon) \equiv 0$ for $s(t, \epsilon)$ denoting one of the two roots

$$\frac{-b(t, \epsilon) \pm \sqrt{-4a_1(t)\epsilon + O(\epsilon^2)}}{2}$$

of the Weierstrass polynomial. Now, if $a_1(0) \neq 0$, there is a branched surface $\mathcal{S}$ with exactly two branches along Γ in the direction of ϵ such that $a_1(0)\epsilon < 0$. In fact, for each root there is a locally defined "surface" given by

$$(t, \epsilon) \to (s(t, \epsilon), t, \epsilon)$$

that contains Γ and is such that $\tau(s(t,\epsilon),t,\epsilon)\equiv 0$.

To identify the quantity $a_1(0)$, we first compute

$$\tau_\epsilon(0,0,0)=a_\epsilon(0,0)u(0,0,0)=a_1(0)u(0,0,0).$$

But, in addition, we know that

$$\tau_\epsilon(0,0,0)=\tau_\epsilon(\xi,0)=\left[(\mathcal{N}+\alpha\mathcal{M})||\mathbf{f}||^2\right](\xi)=||\mathbf{f}(\xi)||^2\mathcal{N}(\xi)$$

and

$$||\mathbf{f}(\xi)||^4 d\alpha(\xi)\mathbf{f}^\perp(\xi)=\tau_{ss}(0,0,0)=2u(0,0,0).$$

Thus, after substitution, we obtain

$$a_1(0)=\frac{2\mathcal{N}(\xi)}{||\mathbf{f}(\xi)||^2 d\alpha(\xi)\mathbf{f}^\perp(\xi)},$$

and we see there will be a real branched surface with two branches provided

$$\epsilon\mathcal{N}(\xi)d\alpha(\xi)\mathbf{f}^\perp(\xi)<0.$$

Next, as before, we consider the restriction of the projection ρ to $\mathcal{S}$. However, here there is a slight difference from our previous arguments because the function s is not necessarily smooth at $\epsilon=0$. To overcome this difficulty we must incorporate the Puiseux series for our expansion of s in powers of ϵ. In the quadratic case this is quite simple. In fact, under our hypothesis that $a_1(0)\neq 0$, there exists a function $(t,\zeta)\to s^*(t,\zeta)$, analytic at $(0,0)$, such that

$$s(t,\epsilon)=s^*(t,\sqrt{\epsilon}).$$

When we restrict the projection ρ to the corresponding branch of $\mathcal{S}$, we have a function $(t,\zeta)\to\rho^*(t,\zeta)$, analytic at $(0,0)$, defined by

$$\rho^*(t,\zeta):=\rho_{\mathrm{loc}}(s^*(t,\zeta),t,\zeta^2)$$

with

$$\rho_{\mathcal{S}}(t,\epsilon)=\rho^*(t,\sqrt{\epsilon}).$$

Now, using the definition of ρ,

$$\rho_{\mathrm{loc}}(s,t,\epsilon):=\rho\left(\psi_s\left(\phi_t(\xi)\right),\epsilon\right),$$

but henceforth writing ρ for ρ_{loc}, we obtain from previous computations, $\rho(0,t,0)\equiv 0$, and

$$\rho_s(s,t,0)=d\rho(s,t,0)\mathbf{f}^\perp(s,t)=-||\mathbf{f}(s,t)||^2(1-\beta(s,t))\equiv 0.$$

A calculation using these facts and the chain rule yields $\rho^*_\zeta(t,0)\equiv 0$ and

$$\rho^*_{\zeta\zeta}(t,0)=2\rho_\epsilon(0,t,0)=2\beta\left(\phi_t(\xi)\right)\mathcal{M}\left(\phi_t(\xi)\right).$$

Thus, we have the representation

$$\rho^*(t,\zeta)=\rho_2(t)\zeta^2+O(\zeta^3)$$

with $\rho_2(t) = \mathcal{M}(\phi_t(\xi))$, and we see that if ξ is a simple zero of the Melnikov function, then $t = 0$ is a simple zero of ρ_2, and the implicit function theorem applies to show the existence of a curve $\zeta \to \sigma(\zeta)$, analytic at $\zeta = 0$, and such that $\rho^*(\sigma(\zeta), \zeta) \equiv 0$. It follows that

$$\rho_S(\sigma(\sqrt{\epsilon}), \epsilon) = \rho^*\left(\sigma(\sqrt{\epsilon}), \sqrt{\epsilon}\right) \equiv 0,$$

and, therefore, $t = \sigma(\sqrt{\epsilon})$, i.e.,

$$\epsilon \to \psi_{s(\sigma(\sqrt{\epsilon}),\epsilon)}\left(\phi_{\sigma(\sqrt{\epsilon})}(\xi)\right)$$

is the desired branch of subharmonics at $\xi \in \Gamma$. □

THEOREM 4.5 (order 2 limit cycle subharmonic bifurcation theorem). *Let E_ϵ denote the parameterized family of differential equations*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon\mathbf{g}(\mathbf{x}, t) + \epsilon^2\mathbf{g}_R(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

such that E_0 has a periodic trajectory $\Gamma \subset \mathcal{A}$ that is in resonance with the η-periodic external force **G**, *i.e., there are relatively prime natural numbers m and n such that the period of Γ is $m\eta/n$. If $\xi \in \Gamma$ is such that following three conditions are satisfied:* (i) $\alpha(\xi) = 0$ *and* $\beta(\xi) = 1$, (ii) *either* $\mathcal{M}(\xi)d\beta(\xi)\mathbf{f}^{\perp}(\xi) \neq 0$ *or* $\mathcal{N}(\xi)d\alpha(\xi)\mathbf{f}^{\perp}(\xi) \neq 0$, *and* (iii) $\xi \in \Gamma$ *is a simple zero of the bifurcation function*

$$\mathcal{D} := \mathcal{N}(\xi)d\beta(\xi)\mathbf{f}^{\perp}(\xi) - \mathcal{M}(\xi)d\alpha(\xi)\mathbf{f}^{\perp}(\xi),$$

then ξ is a subharmonic branch point with two branches. Moreover, these two branches exist, in case $\mathcal{M}(\xi)d\beta(\xi)\mathbf{f}^{\perp}(\xi) \neq 0$, only in the direction of ϵ such that

$$\epsilon\mathcal{M}(\xi)d\beta(\xi)\mathbf{f}^{\perp}(\xi) < 0$$

and, in case $\mathcal{N}(\xi)d\alpha(\xi)\mathbf{f}^{\perp}(\xi) \neq 0$, only in the direction of ϵ such that

$$\epsilon\mathcal{N}(\xi)d\alpha(\xi)\mathbf{f}^{\perp}(\xi) < 0.$$

Proof. The proof of this theorem follows exactly the same logic as the proof of the last theorem with only a few complications. The preparation theorem is applied in turn to both projections ρ and τ, but the proof in both cases is the same. Also, using the notation developed in the proof of the last theorem, we must deal with the fact that neither $\rho_{ss}(0,0,0)$ nor $\tau_{ss}(0,0,0)$ must vanish. For example, in the computation of $\tau^*_{\zeta\zeta}$ we obtain

$$\tau^*_{\zeta\zeta}(0,0) = \tau_{ss}(0,0,0)[s^*_\zeta(0,0)]^2 + 2\tau_\epsilon(0,0,0).$$

So, we must compute $s^*_\zeta(0,0)$. However, using the quadratic formula and the definition of s^* it is clear that

$$[s^*_\zeta(0,0)]^2 = -a_1(0),$$

with the quantity $a_1(0)$ computable as before. Thus, by similar, but slightly more complicated computations, the theorem can be proved by computation of the bifurcation derivatives in terms of α, β, $\mathcal{M}$, and $\mathcal{N}$. □

We end this section with an important remark on detuning. For this, consider a parametrized family of differential equations E_ϵ given by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon\mathbf{G}(\mathbf{x}, t, \epsilon), \quad \mathbf{x} \in \mathbb{R}^2, \quad \epsilon \in \mathbb{R},$$

where E_0 has a periodic trajectory Γ. Up to now we have assumed that the period η of the excitation $t \to \mathbf{G}(\mathbf{x}, t, \epsilon)$ is in resonance with the period of Γ. However, this condition can easily be relaxed in our analysis. In fact, we can assume merely that the period of the excitation is given by an expression of the form $\eta + k\epsilon + O(\epsilon^2)$. In this situation the parameter $k \in \mathbb{R}$ is called a *detuning.* When detuning is introduced, we retain our resonance assumption that the period of Γ is equal to $m\eta/n$ and simply reformulate the analysis in terms of an appropriate Poincaré map, namely,

$$\mathbf{P}(\xi, \epsilon) = \mathbf{x}(m\eta + mk\epsilon + O(\epsilon^2), \xi, \epsilon),$$

where $\mathbf{x}(t, \xi, \epsilon)$ is the solution of E_ϵ with $\mathbf{x}(0, \xi, \epsilon) = \xi$. Clearly, all the derivatives of $\mathbf{P}$ with respect to the space variable ξ reduce to previously computed expressions when evaluated at $\epsilon = 0$. On the other hand, we have

$$\begin{aligned}\mathbf{P}_\epsilon(\xi, 0) &= \dot{\mathbf{x}}(m\eta, \xi, 0)mk + \mathbf{x}_\epsilon(m\eta, \xi, 0)\\ &= mk\mathbf{f}(\xi) + [(\mathcal{N} + \alpha\mathcal{M})\mathbf{f} + \beta\mathcal{M}\mathbf{u}_{\mathbf{f}^\perp}](\xi)\\ &= \left[((mk + \mathcal{N}) + \alpha\mathcal{M})\mathbf{f} + \frac{1}{||\mathbf{f}||^2}\beta\mathcal{M}\mathbf{f}^\perp\right](\xi).\end{aligned}$$

Thus, all previous statements of theorems and formulas for derivatives remain valid in the case of a detuning when we replace each occurrence of $\mathcal{N}$ with $mk + \mathcal{N}$.

5. Examples. As a first example to illustrate the use of the limit cycle subharmonic bifurcation theorem, we consider the nonlinear system given by

$$\dot{x} = -y + x(1 - x^2 - y^2) + \epsilon g_1(x, y, t), \qquad \dot{y} = x + y(1 - x^2 - y^2) + \epsilon g_2(x, y, t),$$

where $G(x, y, t) := (g_1(x, y, t), g_2(x, y, t))$ is t-periodic of period $\eta := 2\pi n/m$ for some relatively prime positive integers n and m. Also, we denote the associated vector field of the unperturbed system by $\mathbf{X}$. This vector field is chosen to have a simple representation in polar coordinates,

$$\dot{r} = r(1 - r^2), \qquad \dot{\theta} = 1,$$

and a unique hyperbolic limit cycle Γ of period 2π on the unit circle. In fact the integral curve of $\mathbf{X}$ corresponding to Γ and starting at $\xi := (\xi_1, \xi_2)$ with $|\xi| = 1$ is given by

$$x(t) = \xi_1 \cos t - \xi_2 \sin t, \qquad y(t) = \xi_1 \sin t + \xi_2 \cos t.$$

For this example, we compute $\alpha(t) \equiv 0$ on Γ. Thus, the bifurcation function is

$$\mathcal{C}(\xi) = [(1 - \beta)\mathcal{N}](\xi) = (1 - e^{-2m\eta}) \int_0^{m\eta} xg_2(x, y, t) - yg_1(x, y, t)\, dt.$$

If we specify the forcing function, we can now determine the existence of subharmonic branch points. For example, if

$$g_1(t) = a\cos t + b\sin t, \qquad g_2(t) = c\cos t + d\sin t,$$

we compute

$$\mathcal{C}(\xi) = n\pi\left(1 - e^{-4n\pi}\right)\left((c-b)\xi_1 - (a+d)\xi_2\right),$$

and we see there are generically two subharmonic branch points at the intersection of the unit circle with the line $(c-b)\xi_1 - (a+d)\xi_2 = 0$. For an example where the excitation depends on the space variables, we take

$$g_1(x,y,t) = ax\cos\frac{t}{2}, \qquad g_2(x,y,t) = by\sin\frac{t}{2}$$

and compute, taking $n = 1$,

$$\mathcal{C}(\xi) = \frac{8}{15}(b-a)\left(1 - e^{-4\pi}\right)(\xi_1 - \xi_2)(\xi_1 + \xi_2).$$

Thus, there are four subharmonic branch points corresponding to the intersections of the lines

$$\xi_1 - \xi_2 = 0, \qquad \xi_1 + \xi_2 = 0$$

with the unit circle.

In the above example we are able to give a complete mathematical analysis of the subharmonic response of a system whose free oscillation is a limit cycle when the system is subjected to a resonant periodic external excitation. We are not able at present to give a similar rigorous mathematical analysis for a model equation that arises from a physical problem. However, we have been successful in applying the theory using numerical experiments. To illustrate this, we consider the example mentioned in the introduction of a forced van der Pol oscillator. In fact, we consider the system

$$\begin{aligned}
\dot{u} &= v,\\
\dot{v} &= -u + \delta(1-u^2)v,\\
\dot{x} &= \tau y,\\
\dot{y} &= \tau(-x + \delta(1-x^2)y) + \epsilon u,
\end{aligned}$$

where τ, δ, and ϵ are real parameters. Here, we view

$$\dot{x} = \tau y, \qquad \dot{y} = \tau(-x + \delta(1-x^2)y)$$

as the unperturbed system. It has, for $\delta > 0$ and $\tau > 0$, a stable limit cycle as its free oscillation. If, in addition, τ is a rational number and $\epsilon \neq 0$, then the xy-system is perturbed by a periodic external stimulus provided by the periodic output $t \mapsto u(t)$ of the uv-system, a second van der Pol oscillator running in resonance. To find the number and the positions of the subharmonic branch points where, for sufficiently small ϵ, families of subharmonic solutions of the perturbed oscillator emerge, we must find the simple zeros of the subharmonic bifurcation function $\mathcal{C}$ along the unperturbed limit cycle. Here, we are restricted by the lack of explicit analytic expressions for the solutions of the unperturbed system near its stable limit cycle. Thus, we have resorted to numerical experiments in order to suggest the actual subharmonic response. The graph of the bifurcation function $\mathcal{C}$ can be computed numerically to obtain the subharmonic branch points for various choices of the parameters. A typical graph of this type is depicted in Fig. 1 of the introduction.

Our final example is an application of the order 2 limit cycle subharmonic bifurcation theorem. For this, consider the system E_ϵ given by

$$\dot{x} = y - x(1 - x^2 - y^2)^2 - \epsilon \cos t, \qquad \dot{y} = -x - y(1 - x^2 - y^2)^2 + \epsilon \sin t,$$

which has the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon \mathbf{g}(t).$$

Here, the unperturbed system has a (semistable) multiplicity 2 limit cycle Γ on the unit circle. The corresponding integral curve of E_0 starting at $\xi := (\xi_1, \xi_2)$ is

$$x(t) = \xi_1 \cos t + \xi_2 \sin t, \qquad y(t) = -\xi_1 \sin t + \xi_2 \cos t.$$

To apply the theorem, we define $r := \sqrt{x^2 + y^2}$, and compute

$$\begin{aligned} \operatorname{div} \mathbf{f}(x,y) &= -2(r^2 - 1)(3r^2 - 1), \\ \operatorname{curl} \mathbf{f}(x,y) &= -2, \\ 2\kappa ||\mathbf{f}||(x,y) &= -2\frac{2 - 6r^4 + 8r^6 - 3r^8}{2 - 4r^2 + 6r^4 - 4r^6 + r^8}. \end{aligned}$$

Then, for $\xi \in \Gamma$, we can compute

$$\begin{aligned} \alpha(t,\xi) &\equiv 0, \qquad & d\alpha(\xi)\mathbf{f}^\perp(\xi) &= 0, \\ \beta(\xi) &= 1, \qquad & d\beta(\xi)\mathbf{f}^\perp(\xi) &= -16\pi. \end{aligned}$$

For example, we have $\mathbf{f}^\perp(\xi) = \xi$, and, for any $\zeta \in \mathbb{R}^2$,

$$\beta(\zeta) = \exp\left(\int_0^{2\pi} -2(r^2 - 1)(3r^2 - 1)\, dt\right).$$

Thus,

$$d\beta(\xi)\mathbf{f}^\perp(\xi) = d\beta(\xi)\xi = \frac{d}{ds}\beta((1+s)\xi)\bigg|_{s=0}.$$

After the obvious computation, we find

$$d\beta(\xi)\mathbf{f}^\perp(\xi) = -8\int_0^{2\pi} \frac{dr}{ds}\bigg|_{s=0} dt.$$

But, since

$$\dot{r} = -r^2(1 - r^2)^2,$$

we can easily find the variational equation for r_s on Γ to be

$$\dot{r}_s = 0.$$

Also, since $r(0,s) = ||(1+s)\xi||$, we have $r_s(0,0) = 1$. This means $dr/ds \equiv 1$, and, in turn,

$$d\beta(\xi)\mathbf{f}^\perp(\xi) = -16\pi.$$

Next, we find

$$\mathcal{M}(\xi) = \int_0^{2\pi} y \sin t - x \cos t \, dt = -2\pi \xi_1.$$

Thus, the only zeros of $\mathcal{M}(\xi) = 0$ are on the line $\xi_1 = 0$. In particular, for $\zeta = (\pm 1, 0)$,

$$\mathcal{M}(\zeta) d\beta(\zeta) \mathbf{f}^{\perp}(\zeta) \neq 0.$$

We also have

$$\mathcal{D}(\xi) = -16\pi \mathcal{N}(\zeta) = 16\pi \int_0^{2\pi} (y \cos t + x \sin t)\, dt = 32\pi^2 \xi_2.$$

Thus, $\mathcal{D}$ has a simple zero along Γ at ζ, and, by the theorem, there will be two branches of harmonics at the subharmonic branch point ζ. At $\zeta = (-1, 0)$ these harmonics exist for sufficiently small $\epsilon < 0$, while at $(1, 0)$ they exist for sufficiently small $\epsilon > 0$.

REFERENCES

[1] A. A. Andronov, E. A. Leontovich, I. I. Gordon, and A. G. Maier, *Theory of Bifurcations of Dynamical Systems on a Plane*, John Wiley, New York, 1973.

[2] A. A. Andronov, E. A. Vitt, and S. E. Khaiken, *Theory of Oscillators*, Pergamon Press, Oxford, 1966.

[3] D. V. Anosov and V. I. Arnold, *Dynamical systems* I, in Encyclopaedia of Mathematical Sciences, Vol. 1, Springer-Verlag, Berlin, 1988.

[4] S. Antman, *General solution for plane extensible elasticae having nonlinear stress-strain laws*, Quart. Appl. Math., 26 (1968), pp. 35–47.

[5] V. I. Arnold, *Geometric Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1982.

[6] N. N. Bautin, *On the number of limit cycles which appear with the variation of coefficients from an equilibrium position of focus or center type*, Amer. Math. Soc. Transl., 100 (1954), pp. 1–19.

[7] T. R. Blows and N. G. Lloyd, *The number of limit cycles of certain polynomial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 98 (1984), pp. 215–239.

[8] T. R. Blows and L. M. Perko, *Bifurcation of Limit Cycles from Centers*, University of Northern Arizona, Flagstaff, AZ, preprint, 1990.

[9] C. Chicone, *The monotonicity of the period function for planar Hamiltonian vector fields,* J. Differential Equations, 69 (1987), pp. 310–321.

[10] ———, *On bifurcation of limit cycles from centers,* Lecture Notes in Math., 1455 (1991), pp. 20–43.

[11] C. Chicone and F. Dumortier, *A quadratic system with a non monotonic period function,* Proc. Amer. Math. Soc., 102 (1988), pp. 706–710.

[12] C. Chicone and M. Jacobs, *Bifurcation of critical periods for plane vector fields,* Trans. Amer. Math. Soc., 312 (1989), pp. 433–486.

[13] ———, *Bifurcation of limit cycles from quadratic isochrones,* J. Differential Equations, 91 (1991), pp. 268–327.

[14] ———, *On a computer algebra aided proof in bifurcation theory,* in Computer Aided Proofs in Analysis, K. Meyer and D. Schmidt, eds., IMA Volumes in Mathematics and Its Applications, Springer-Verlag, New York, 28 (1991), pp. 52–70.

[15] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory,* Springer-Verlag, New York, 1982.

[16] S. N. Chow and D. Wang, *On the monotonicity of the period function of some second order equations,* Časopis Pěst. Mat., 111 (1986), pp. 14–25.

[17] S. N. Chow and J. Sanders, *On the number of critical points of the period*, J. Differential Equations, 64 (1986), pp. 51–66.

[18] H. T. Davis, *Introduction to Nonlinear Differential and Integral Equations,* Dover, New York, 1962.

[19] G. Dangelmeyer and J. Guckenheimer, *On a four parameter family of planar vector fields,* Arch. Rational Mech. Anal., 97 (1987), pp. 321–352.

[20] S. P. Diliberto, *On systems of ordinary differential equations,* in Contributions to the Theory of Nonlinear Oscillations, Annals of Mathematics Studies, Vol. 20, Princeton University Press, Princeton, NJ, 1950.

[21] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, Second ed., Springer-Verlag, New York, 1986.

[22] P. Hagadorn, *Non-Linear Oscillations*, Second ed., Clarendon Press, Oxford, 1988.

[23] J. Hale, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[24] J. Hale and P. Taboas, *Bifurcation near degenerate families*, Appl. Anal., 11 (1980), pp. 21–37.

[25] C. Hayashi, *Nonlinear Oscillations in Physical Systems*, McGraw-Hill, New York, 1964.

[26] P. J. Holmes, *Domains of Stability in a Wind Induced Oscillation Problem,* Trans. ASME J. Appl. Mech., 46 (1979), pp. 672–676.

[27] S. Lefschetz, *Differential Equations: Geometric Theory,* Second ed., Dover, New York, 1977.

[28] N. G. Lloyd, *Limit cycles of polynomial systems, some recent developments*, New Directions in Dynamical Systems, LMS Lecture Notes, Series No. 127, Cambridge University Press, Cambridge, 1988, pp. 192–234.

[29] C. Li and C. Rousseau, *Codimension* 2 *symmetric homoclinic bifurcations and applications to* 1 : 2 *resonance*, Université de Montréal, Montréal, Québec, preprint, 1988.

[30] M. W. McLachlan, *Ordinary Non-Linear Differential Equations in Engineering and Physical Sciences*, Second ed., Oxford Univ. Press, Oxford, 1958.

[31] V. K. Melnikov, *On the stability of the center for time periodic perturbations*, Trans. Moscow Math. Soc., 12 (1963), pp. 1–57.

[32] N. Minorsky, *Nonlinear Oscillations,* Van Nostrand, New York, 1962.

[33] W. E. Olmstead and D. J. Mescheloff, *Buckling of a nonlinear elastic rod*, J. Math. Anal. Appl., 46 (1974), pp. 609–634.

[34] A. H. Nayfeh and D. T. Mook, *Nonlinear Oscillations*, John Wiley, New York, 1979.

[35] L. M. Perko, *Global families of limit cycles of planar analytic systems*, Trans. Amer. Math. Soc., 322 (1990), pp. 627–656.

[36] R. H. Rand and P. J. Holmes, *Bifurcation of periodic motions in two weakly coupled Van der Pol oscillators*, Internat. J. Non-Linear Mech., 15 (1980), pp. 387–399.

[37] J. W. Reijn, *A bibliography of the qualitative theory of quadratic systems of differential equations in the plane*, Tech. Report 89-71, Faculty of Technical Mathematics and Informatics, Delft University of Technology, the Netherlands, 1989.

[38] F. Rothe, *The energy-period function and perturbations of Hamiltonian systems in the plane*, in Oscillation, Bifurcation and Chaos, Canadian Mathematical Society, Conference Proceedings, 8 (1987), pp. 621–635.

[39] C. Rousseau, *Bifurcation methods in quadratic systems*, Canadian Mathematical Society Conference Proceedings, 8 (1987), pp. 637–653.

[40] R. Schaaf, *A class of Hamiltonian systems with increasing periods,* J. Reine Angew. Math., 363 (1985), pp. 96–109.

[41] J. J. Stoker, *Nonlinear Vibrations,* John Wiley, New York, 1950.

[42] S. W. Wiggins, *Global Bifurcations and Chaos*, Springer-Verlag, New York, 1988.

[43] ———, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.

[44] H. Żołądek, *Abelian integrals in non-symmetric perturbation of symmetric Hamiltonian vector fields*, Warsaw University, Warsaw, Poland, preprint, 1988.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1609-1622, November 1992

010

LIMIT CYCLES IN A CUBIC SYSTEM WITH A CUSP*

WANG XIAN† AND ROBERT E. KOOIJ‡

Abstract. This paper studies the number of limit cycles in a cubic system with a cusp. By using new results concerning systems of Liénard type, the question of relative position and the maximum number of limit cycles for this system is solved completely.

Key words. cubic system, limit cycles, Liénard system

AMS(MOS) subject classifications. 34C, 58F

1. Introduction. In this paper we give a global analysis of a planar cubic system that arises in the study of three-dimensional viscous flow structures near a plane wall. In [3] Bakker developed a classification strategy to classify two-dimensional viscous flow structures near a plane wall in a systematic way. The investigation of three-dimensional flows is taken up by Bakker and de Winkel [4] and Kooij and Bakker [10]. The classification strategy relies to a great extent on the qualitative theory of differential equations and bifurcation theory. The flow is assumed to be steady, viscous, incompressible, and it satisfies no-slip boundary conditions on the wall. The topology of such a flow is studied on the basis of local solutions of the Navier–Stokes equations. The streamline pattern is represented by the trajectories of a three-dimensional system

$$(*)\qquad \frac{dx}{dt}=\dot{x}=u, \qquad \frac{dy}{dt}=\dot{y}=v, \qquad \frac{dz}{dt}=\dot{z}=w,$$

where t is real time, and u, v, w denote the velocity components in a cartesian reference system. The wall is represented by $z=0$. The local solutions are obtained by performing Taylor expansions of the velocity vector field.

The main objective of the classification strategy developed by Bakker [3] is to give a unified description of all topologically different flow patterns that will arise near a singularity of system $(*)$.

The dynamical behavior of the wall shear stress vector, defined by $\tau=\mu(\partial u/\partial z)_{z=0}$ in the x-direction and $\sigma=\mu(\partial v/\partial z)_{z=0}$ in the y-direction, where μ is the dynamical viscosity of the fluid, is governed by the equation $dy/dx=\lim_{z\to 0}(v(x,y,z)/u(x,y,z))$. The solution curves of this equation are referred to as skin friction lines.

For the type of singularity that is studied by Kooij and Bakker [10], the equation that describes the skin friction lines is equivalent with the following system:

$$\dot{x}=y+ax^2+x^3, \qquad \dot{y}=\lambda x^2-x^3.$$

In this paper we present a qualitative study of this cubic system.

2. Some definitions and theorems concerning Liénard systems. First we shall quote some results from [13]-[15], [17].

Consider the system

$$(1)\qquad \dot{x}=\varphi(y)-F(x), \qquad \dot{y}=-g(x),$$

* Received by the editors July 10, 1991; accepted for publication (in revised form) March 11, 1992.

† Department of Mathematics, Nanjing University, Nanjing, People's Republic of China.

‡ Faculty of Technical Mathematics and Informatics, Delft University of Technology, Delft, the Netherlands.

with $F(x)=\int_0^x f(s)\,ds$, where

$$f(x), g(x)\in C^1_{(\gamma_1,\gamma_2)}, \quad -\infty\leqq\gamma_1<0<\gamma_2\leqq\infty, \quad \varphi(y)\in C^1_{\mathbb{R}}.$$

In (1), $g(x)$ and $\varphi(y)$ may have several zeros, that is, (1) may have several singular points.

Definition 1 [13]. A set of singular points of (1) is called an unstable (stable) singular point system with index +1 if the sum of the indices of the singular points is +1 and there exists a bounded region D, containing this set of singular points, such that from every point on the boundary ∂D only positive (negative) semitrajectories of (1) are leaving D.

Definition 2 [13]. A set of singular points of (1) is called an unstable (stable) singular point-cycle system with index +1 if the sum of the indices of the singular points is +1 and there exists at least one limit cycle or separatrix cycle containing one or several of these singular points in its interior and there exists a bounded region D, containing the set of singular points, such that from every point on the boundary ∂D only positive (negative) semitrajectories of (1) are leaving D.

Theorem 1 [13]. *Consider the special case of system* (1):

$$(2)\qquad \dot{x}=\varphi(y)-E(x), \qquad \dot{y}=-g(x)=-g_0(x)-g_e(x),$$

where $g_0(x)$ and $g_e(x)$ are the odd part and the even part of $g(x)$, respectively. Suppose (2) *satisfies*

(1) $E(0)=0, E(-x)=E(x), y\varphi(y)>0$ $(y\neq 0)$, *and $\varphi(y)$ increasing for $y\in\mathbb{R}$*;

(2) $g(0)=0, xg(x)>0$ for $x\notin[a,0]$ $(\gamma_1<a\leqq 0)$ *and $g_e(x)>0$.*

Then (2) *does not have closed orbits that intersect with $x=a$ and $x=0$ simultaneously.*

Theorem 2 [14]. *Suppose that system* (1) *satisfies*

(1) $y\varphi(y)>0$ $(y\neq 0)$ *and $\varphi(y)$ is increasing for $y\in\mathbb{R}$, and there exist x_0, x_1, Δ, with $0<x_0<x_1<\gamma_2, 0<\Delta\leqq x_1$, such that $F(0)=F(x_1)=0, f(x)<0$ for $x<x_0, f(x)>0$ for $x>x_0$, $g(0)=g(\Delta)=0, xg(x)>0$ $(x\neq 0,\Delta)$*;

(2) *The system of equations*

$$F(u)=F(v), \quad \frac{g(u)}{f(u)}=\frac{g(v)}{f(v)}, \quad \gamma_1<u<0, \quad x_1<v<\gamma_2,$$

has at most one solution;

(3) *The function $f(x)/g(x)$ is increasing for $x\in(x_1,\gamma_2)$.*

Then (1) *has at most one limit cycle, and if it exists it has a negative characteristic exponent.*

Corollary. *If system* (1) *satisfies condition* (1) *in Theorem* 2 *and the function $g(x)/f(x)$ is decreasing for $\gamma_1<x<0$ and $x_1<x<\gamma_2$, then* (1) *has at most one limit cycle, and if it exists it has a negative characteristic exponent.*

Theorem 3 [15]. *Suppose that system* (1) *satisfies*

(1) $y\varphi(y)>0(y\neq 0)$, *$\varphi(y)$ is increasing for $y\in\mathbb{R}$, and there exist $\Delta, x_1, \bar{x}_1, \bar{x}_2, x_2$, with $\gamma_1<x_1<\bar{x}_1<0<\bar{x}_2<x_2<\gamma_2$, $0<\Delta\leqq x_2$, such that $F(x_1)=F(0)=F(x_2)=0, f(x)<0$ for $x\in(\bar{x}_1,\bar{x}_2), f(x)>0$ for $x\notin(\bar{x}_1,\bar{x}_2)$, $g(0)=g(\Delta)=0, xg(x)>0(x\neq 0,\Delta)$*;

(2) $G(x_1)\geqq G(x_2)$, *where $G(x)=\int_0^x g(s)ds$*;

(3) *The system of equations*

$$F(u)=F(v), \quad \frac{g(u)}{f(u)}=\frac{g(v)}{f(v)}, \quad \bar{x}_1<u<0, \quad x_2<v<\gamma_2,$$

has at most one solution;

(4) *The function $f(x)/g(x)$ is increasing for $x \in (x_2, \gamma_2)$.*

Then (1) *has at most one limit cycle, and if it exists it has a negative characteristic exponent.*

Remark. If in system (1) $\varphi(y)=y$, then the last condition of Theorem 3 may be weakened into

(4′) The function $f(x)F(x)/g(x)$ is increasing for (x_2, γ_2).

THEOREM 4 [17]. *Consider the system*

$$\dot{x}=y, \qquad \dot{y}=-g(x)-f(x)y. \tag{3}$$

Suppose there exist $\gamma_1 < c_1 \leqq b_1 < a_1 < 0 < a_2 < b_2 \leqq c_2 < \gamma_2$, such that

(1) *$xg(x)<0$ for $x \in (a_2, b_2) \cup (c_1, a_1)$, $xg(x)>0$ for $x \in (c_2, \gamma_2) \cup (\gamma_1, c_1)$,*

(2) *$f(x) \leqq 0$ for $x \in (b_1, b_2)$, $f(x) \geqq 0$ for $x \notin (b_1, b_2)$,*

(3) *$F(a_2)=F(c_2)$, $F(c_1)=F(a_1)$,*

(4) *The functions $f(x)$, $g(x)/(x-c_2)$ and $(x-c_2)(f(x)/g(x))$ are increasing for $x \in (c_2, \gamma_2)$, the functions $f(x)$, $g(x)/(x-c_1)$ and $(x-c_1)(f(x)/g(x))$ are decreasing for $x \in (\gamma_1, c_1)$.*

Then (3) *has at most two limit cycles surrounding all singular points.*

Remark. It follows from the proof of Theorem 4 that it still can be applied as $c_2=b_2=a_2=0$ or $c_1=b_1=a_1=0$.

3. Some properties of a cubic system with a cusp. The cubic system given in [10] reads

$$\dot{x}=y+ax^2+x^3, \qquad \dot{y}=\lambda x^2-x^3. \tag{4}$$

We can confine ourselves to the case $a>0$ because for $a=0$ it is easy to see that (4) has no closed orbits and for $a<0$ the transformation $(x, y, a, \lambda) \to (-x, -y, -a, -\lambda)$ can be applied to retain system (4) with $a>0$.

For $\lambda \neq 0$, (4) has two singularities, $O(0,0)$ and $A(\lambda, -a\lambda^2-\lambda^3)$, where O is a cusp and A is an antisaddle. A cusp is a nonhyperbolic singularity whose local topology is determined by two tangent separatrices, a stable and an unstable one. By an antisaddle, we mean a node, a focus, or a center. For $\lambda=0$, O and A collapse to form an unstable nilpotent focus ($a<\sqrt{2}$) or a nilpotent point with an elliptic sector ($a \geqq \sqrt{2}$), as can be seen from the classification scheme for nilpotent singularities in Andronov [2].

By comparing the trajectories of (4) for $\lambda=0$, $a<\sqrt{2}$, with the trajectories of the system $\dot{x}=y+ax^2$, $\dot{y}=-x^3$ (which form a family of closed curves), it can be proved that O is globally unstable for $\lambda=0$, $a<\sqrt{2}$; see Appendix A.

It is easy to check that A is a stable (unstable) elementary antisaddle if $\lambda(2a+3\lambda)<0(>0)$ and that for $2a+3\lambda=0$ A is a stable first-order weak focus.

The qualitative behavior of system (4) at infinity is as given in Fig. 1, as can be confirmed by using several blowups; see [6]. In fact, it can be shown that for all values

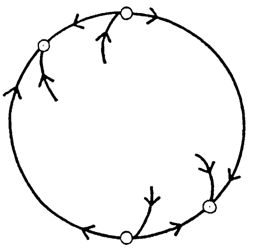


FIG. 1. *Behavior at infinity of system* (4).

of a and λ, the Poincaré equator is attracting. The blowups are described in detail in Appendix B.

The translation of A to the origin and a rescaling of time, $t=-\tau$, reduces (4) to the following Liénard system:

$$\frac{dx}{d\tau}=y-F(x), \qquad \frac{dy}{d\tau}=-g(x), \tag{5}$$

with

$$\begin{aligned}
F(x)&=\lambda(2a+3\lambda)x+(a+3\lambda)x^2+x^3,\\
f(x)&=(x+\lambda)(3x+2a+3\lambda),\\
g(x)&=x(x+\lambda)^2,\\
G(x)&=\frac{x^2}{12}(3x^2+8\lambda x+6\lambda^2).
\end{aligned}$$

4. Statement of the main result. In the following we will call a limit cycle that does not intersect the line $x+\lambda=0$ (in system (5)) small, and a limit cycle that does intersect the line $x+\lambda=0$ (in system (5)) large.

The main result of this paper is the following.

THEOREM 5. (i) *For $\lambda \geqq 0$ system* (5) *has no closed orbits.* (ii) *For fixed $\lambda<0$ there exist $a_1=a_1(\lambda)$, $a_2=a_2(\lambda)$, and $a_3=a_3(\lambda)$ with $-\lambda<a_3\leqq a_2\leqq a_1$ and $-\frac{4}{3}\lambda\leqq a_1<-\frac{3}{2}\lambda$, such that*

(a) *For $a\geqq -\frac{3}{2}\lambda$ system* (5) *has no small limit cycle and exactly one large limit cycle, and it has a negative characteristic exponent*;

(b) *For $a_1<a<-\frac{3}{2}\lambda (a=a_1)$ system* (5) *has exactly one large limit cycle and exactly one small limit cycle* (*one cusploop*); *the small limit cycle has a positive characteristic exponent*; *the large limit cycle is stable*;

(c) *For $a_2<a<a_1$ $(a=a_2)$ system* (5) *has two large limit cycles* (*a unique semistable large limit cycle*) *and no small limit cycles*;

(d) *For $a_3<a<a_2$ system* (5) *has no small limit cycles and the number of large limit cycles is two, one* (*in which case the limit cycle is semistable*), *or zero*;

(e) *For $a_3<a$ $(a=a_3)$ system* (5) *has no closed orbits* (*a unique semistable large limit cycle*).

By a cusploop, we mean a homoclinic orbit that consists of a cusp and its two separatrices that connect. The theorem will be proved completely by six lemmas.

First, let us define the roots of $F(x)=0$ and $f(x)=0$:

$$x_1=-\frac{a+3\lambda+\sqrt{(a-3\lambda)(a+\lambda)}}{2}, \quad x_0=0, \quad x_2=-\frac{a+3\lambda-\sqrt{(a-3\lambda)(a+\lambda)}}{2}$$

and

$$\bar{x}_1=-\tfrac{2}{3}a-\lambda<-\lambda=\bar{x}_2,$$

respectively. The roots of $g(x)$ are $x=0$ and $x=-\lambda$.

LEMMA 1. *For $\lambda\geqq 0$ system* (5) *has no closed orbits.*

Proof. We first consider the case $\lambda=0$; at this time (5) has only one singular point $O(0,0)$ and $xg(x)>0$ $(x\neq 0)$. Since

$$F(u)=F(v), \quad G(u)=G(v), \quad -\infty<u<0, \quad 0<v<\infty,$$

i.e.,

$$u^2+uv+v^2+a(u+v)=0, \quad (u^2+v^2)(u+v)=0, \quad -\infty<u<0, \quad 0<v<\infty$$

does not have a solution, (5) has no closed orbit by Rychkov's theorem [11].

For $\lambda>0$, if $0<a\leqq 3\lambda$, then $F(x)=0$ has no real roots except $x=0$; thus $g(x)F(x)>0$ for $x\in\mathbb{R}\backslash\{0,\lambda\}$. Notice that

$$\mu(x,y)=\tfrac{1}{2}y^2+G(x)=C$$

represents a family of closed curves, and so from

$$\frac{d\mu}{d\tau}=-g(x)F(x)\leqq 0,$$

we know at once that (5) has no closed orbits.

In order to discuss the case $a>3\lambda$, we consider the equivalent system of (5),

$$(6)\qquad \frac{dx}{d\tau}=y,\qquad \frac{dy}{d\tau}=-g(x)-f(x)y.$$

System (6) has two singular points, $B(-\lambda,0)$ and $O(0,0)$. Since at this time $x_1<x_2<\bar{x}_2=-\lambda$, thus $f(x)>0$ for $-\lambda<x<\infty$, it follows that (6) has no closed trajectory surrounding O alone.

Next we consider the system

$$(7)\qquad \frac{dx}{d\tau}=y,\qquad \frac{dy}{d\tau}=-x(x+\lambda)^2-2(a+3\lambda)xy$$

and the equivalent system

$$(8)\qquad \frac{dx}{d\tau}=y-(a+3\lambda)x^2,\qquad \frac{dy}{d\tau}=-x(x^2+\lambda^2)-2\lambda x^2.$$

From Theorem 1 we know that (8) does not have a trajectory intersecting with $x=-\lambda$ and $x=0$ simultaneously; therefore, (7) has no closed trajectory surrounding B, O. Furthermore, we can prove that (8) has no limit cycles surrounding O alone, by using a Dulac function $B(y)=\exp(-(2(a+3\lambda)/\lambda^2)y)$.

Because for (8) O is a stable first-order weak focus, as can be proved using [5, form. 3.4.11], it follows that singular point system B, O of (7) is stable.

Suppose that (6) has a closed trajectory surrounding B, O; then from

$$\begin{aligned}\left.\frac{dy}{dx}\right|_{(6)}&=-\frac{x(x+\lambda)^2}{y}-2(a+3\lambda)x-[3x^2+\lambda(2a+3\lambda)]\\&<-\frac{x(x+\lambda)^2}{y}-2(a+3\lambda)x=\left.\frac{dy}{dx}\right|_{(7)},\end{aligned}$$

it would follow that (7) has at least one limit cycle, but this contradicts the result proved above. Hence (6), that is (5), has no closed trajectory. This completes the proof of Lemma 1.

Next we will discuss the case $\lambda<0$.

LEMMA 2. *For* $0<a\leqq-\lambda$ *system* (5) *has no closed orbits. For* $-\lambda<a<-\frac{4}{3}\lambda$ *system* (5) *has no small limit cycle.*

Proof. The proof of the first statement is the same as for the case $\lambda>0$, $0<a\leqq 3\lambda$ above. After simplifying and by putting $\xi=u+v$, $\eta=uv$, the system of equations

$$(9)\qquad F(u)=F(v),\quad G(u)=G(v),\quad -\infty<u<0,\quad 0<v<-\lambda$$

can be changed into

$$(10)\qquad \begin{aligned}\eta&=\xi^2+(a+3\lambda)\xi+\lambda(2a+3\lambda)=H_1(\xi),\qquad \eta<0,\\ \eta&=\frac{\xi(3\xi^2+8\lambda\xi+6\lambda^2)}{2(3\xi+4\lambda)}=H_2(\xi),\qquad 0<\xi<-\lambda.\end{aligned}$$

The solutions of $H_1(\xi)=H_2(\xi)$ are $\xi_{12}=-a-2\lambda\pm\sqrt{a(3a+4\lambda)/3}$ and $\xi_0=-2\lambda$. Obviously, the numbers ξ_{12} are complex for $3a+4\lambda<0$, and because $\xi_0>-\lambda$, (10) and hence (9) have no solution. The second statement of the lemma follows from Rychkov's theorem [11].

LEMMA 3. *For $a\geqq-\frac{3}{2}\lambda$ system* (5) *has no small limit cycle.*

Proof. If $P(x,y)=y-F(x)$ and $Q(x)=-g(x)$, then for $B(y)=\exp(-(3/\lambda)y)$ we have $\operatorname{div}(BP,BQ)=(B/\lambda)(x+\lambda)(3x^2-\lambda(2a+3\lambda))$. By Dulac's criterion, for $\lambda(2a+3\lambda)<0$, system (5) has no closed orbits that do not intersect $x+\lambda=0$.

LEMMA 4. *For $a<-\frac{3}{2}\lambda$ system* (5) *has at most one small limit cycle, and if it exists it has a positive characteristic exponent.*

Proof. By the change of variables, $t=-\tau$, $y\to-y$, system (5) converts into the system

$$(5')\qquad \frac{dx}{dt}=y-F_1(x),\qquad \frac{dy}{dt}=-g(x),$$

where $F_1(x)=-F(x)$, $f_1(x)=-f(x)$.

It is easy to see that $0<\bar{x}_1<x_1<-\lambda=\bar{x}_2<x_2$, $F_1(0)=F_1(x_1)=0$, $f_1(x)<0$ for $x<\bar{x}_1$, and $f_1(x)>0$ for $x>\bar{x}_1$, $xg(x)>0$ $(x\neq0)$ for $x\in(-\infty,-\lambda)$.

An elementary calculation shows that

$$\frac{d}{dx}\left(\frac{g(x)}{f_1(x)}\right)=-\frac{(x+\lambda)^2}{f_1^2(x)}H(x),$$

with $H(x)=3x^2+2(2a+3\lambda)x+\lambda(2a+3\lambda)$.

The discriminant D of $H(x)$ satisfies $D=8a(2a+3\lambda)<0$, and, therefore, (d/dx)-$(g(x)/f_1(x))<0$ certainly holds for $x\in(-\infty,0)\cup(x_1,-\lambda)$.

The result follows from the corollary of Theorem 2 and the observation that we have used the transformation $t=-\tau$.

LEMMA 5. *For $a\geqq-\frac{3}{2}\lambda$ system* (5) *has exactly one large limit cycle, and if it exists it has a negative characteristic exponent.*

Proof. For $a\geqq-\frac{3}{2}\lambda$, (5) has no closed trajectory surrounding O alone by Lemma 3. It follows that the singular point system O, $A'(-\lambda,-a\lambda^2-\lambda^3)$ is unstable. Notice that for system (5) the singular points at infinity are repelling; so (5) has at least one large stable limit cycle. We will prove the uniqueness of this large limit cycle by using Theorem 3. It is easy to see that $x_1<\bar{x}_1<0<\bar{x}_2=\lambda<x_2$, where $f(x)<0$ for $x\in(\bar{x}_1,\bar{x}_2)$ and $f(x)>0$ for $x\notin(\bar{x}_1,\bar{x}_2)$, and, therefore, condition (1) of Theorem 3 is satisfied; see Fig. 2.

An elementary calculation shows that

$$G(x_1)-G(x_2)=\frac{(a+\lambda)(3a^2+4a\lambda-3\lambda^2)\sqrt{(a+\lambda)(a-3\lambda)}}{12}.$$

The region $S=\{(\lambda,a)\,|\,\lambda(2a+3\lambda)<0,\ a>0\}$ can be divided into

$$S_1=\left\{(\lambda,a)\,\middle|\,-\frac{3a}{2+\sqrt{13}}\leqq\lambda<0,\ a>0\right\}$$

and

$$S_2=\left\{(\lambda,a)\,\middle|\,-\frac{2}{3}a<\lambda\leqq-\frac{3a}{2+\sqrt{13}},\ a>0\right\}.$$

Obviously, for $S_1(S_2)$ the inequality $G(x_1)-G(x_2)\geqq0(\leqq0)$ holds, and, therefore, condition (2) of Theorem 3 holds for S_1.

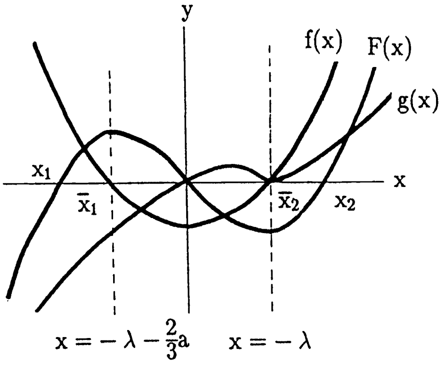


FIG 2. *Relative positions of $f(x)$, $F(x)$, and $g(x)$.*

After simplifying and by putting $\xi = u+v$, $\eta = uv$, the system of equations

$$F(u)=F(v), \quad \frac{f(u)}{g(u)}=\frac{f(v)}{g(v)}, \quad \bar{x}_1<u<0, \quad x_2<v<\infty$$

can be reduced to

$$\eta=\xi^2+(a+3\lambda)\xi+\lambda(2a+3\lambda)=H_1(\xi),$$
$$\eta=-\tfrac{1}{3}(2a+3\lambda)(\xi+\lambda)=H_2(\xi), \qquad \eta<0.$$

It is easy to check that the functions $H_1(\xi)$ and $H_2(\xi)$ have exactly one intersection point with $\eta<0$; see Fig. 3. Therefore, condition (3) of Theorem 3 is also fulfilled.

An elementary calculation shows that

$$\frac{d}{dx}\left(\frac{f(x)F(x)}{g(x)}\right)=\frac{q(x)}{(x+\lambda)^2},$$

where $q(x)=6x^3+(5a+21\lambda)x^2+2\lambda(5a+12\lambda)x-\lambda(a-3\lambda)(2a+3\lambda)$. The numbers $\hat{x}_1=-\lambda$ and $\hat{x}_2=-(5a+12\lambda)/9$ satisfy $q'(\hat{x}_1)=q'(\hat{x}_2)=0$. Furthermore, $\hat{x}_1-\hat{x}_2=(5a+3\lambda)/9>0$ and $q(-\lambda)=-2a\lambda(a+\lambda)>0$. It follows that $q(x)>0$ for $x>x_2>\bar{x}_2=\hat{x}_1$; hence

$$\frac{d}{dx}\left(\frac{f(x)F(x)}{g(x)}\right)>0 \quad \text{for } x\in(x_2,\infty),$$

i.e., condition (4') of Theorem 3 is also fulfilled.

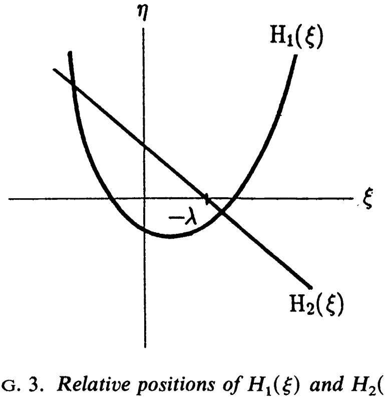


FIG. 3. *Relative positions of $H_1(\xi)$ and $H_2(\xi)$.*

For the region S_1 all conditions of Theorem 3 are satisfied and the lemma is proved. To prove the lemma for region S_2, we will reflect system (5) with respect to the y-axis, yielding

$$\frac{dx}{d\tau}=y-\bar{F}(x), \qquad \frac{dy}{d\tau}=-\bar{g}(x), \tag{11}$$

with $\bar{F}(x)=-F(-x)$ and $\bar{g}(x)=-g(-x)$.

It is easy to see that for system (11) $\bar{G}(x_1)\geqq\bar{G}(x_2)$ in the region S_2. The other conditions of Theorem 3 are also satisfied for system (11), as can be shown in a similar way as above. The lemma is completely proved.

LEMMA 6. *For $-\lambda<a<-\frac{3}{2}\lambda$ system* (4) *has at most two large limit cycles.*

Proof. By a change of variables $t=-\tau$, $y\to-y$, system (4) changes into the system

$$\frac{dx}{d\tau}=y-ax^2-x^3, \qquad \frac{dy}{d\tau}=-x^2(x-\lambda). \tag{4'}$$

Consider the equivalent system of (4'):

$$\frac{dx}{d\tau}=y, \qquad \frac{dy}{d\tau}=-x^2(x-\lambda)-(2ax+3x^2)y. \tag{12}$$

Let $F(x)=x^2(x+a), f(x)=x(2a+3x), g(x)=x^2(x-\lambda)$.

The roots of $F(x)=0$ are $x_1=-a<x_2=0$. The roots of $f(x)=0$ are $\alpha=-\frac{2}{3}a<\beta=0$. The roots of $g(x)=0$ are $c=\lambda<d=0$. For $-\lambda<a<-\frac{3}{2}\lambda$, we have $\frac{3}{2}\lambda<x_1<\lambda$ and $\lambda<\alpha<\frac{2}{3}\lambda$.

Notice that the first three conditions of Theorem 4 are fulfilled with

$$\gamma_1=-\infty, \quad c_1=\lambda, \quad b_1=\alpha, \quad a_2=b_2=c_2=0, \quad \gamma_2=\infty$$

and where a_1 satisfies $F(a_1)=F(\lambda)$ and $b_1<a_1<0$; see Fig. 4.

In the interval $(0,\infty), f'(x)=6x+2a>0, (g(x)/x)'=2x-\lambda>0,$

$$\left(\frac{xf(x)}{g(x)}\right)'=-\frac{2a+3\lambda}{(x-\lambda)^2}>0,$$

so the functions $f(x), g(x)/x$ and $(xf(x))/g(x)$ are increasing for $x\in(0,\infty)$. In the interval $(-\infty,\lambda), f'(x)=6x+2a<3\lambda+(2a+3\lambda)<0, (g(x)/(x-\lambda))'=2x<0,$ $(((x-\lambda)f(x))/g(x))'=-2a/x^2<0$, so the functions $f(x), g(x)/(x-\lambda)$ and $((x-\lambda)f(x))/g(x)$ are decreasing for $x\in(-\infty,\lambda)$. Therefore, from Theorem 4 and its remark,

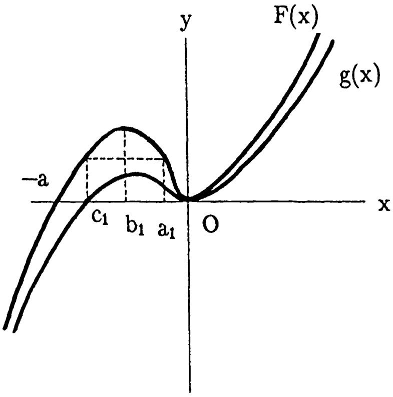


FIG 4. *Relative positions of $F(x)$ and $g(x)$.*

it follows that (12) has at most two large limit cycles surrounding the two singular points O, A. The lemma is completely proved.

Proof of Theorem 5. In the region $K=\{(x,y)\,|\,-\infty<x<-2\lambda, |y|<\infty\}$ system (5), considered as a vector field X_a, is a so-called semicomplete family (mod $x=0$) of rotated vector fields with respect to the parameter "a" because in the region K the following assertions hold except on $x=0$:

—The derivative of the angle between X_0 and X_a with respect to a is strictly negative for $x\neq 0$,

—$\tan\theta\to\pm\infty$ as $a\to\pm\infty$, where θ is the angle of the vector field.

From this we can deduce, see Perko [11], that attracting (repelling) limit cycles expand (shrink) monotonically as a decreases. As a varies, limit cycles appear or disappear in a singular point, in a separatrix cycle or in a semistable limit cycle. It follows from Lemmas 2–4 that there must exist a unique $a_1=a_1(\lambda)$ with $-\frac{4}{3}\lambda\leqq a_1<-\frac{3}{2}\lambda$, such that $(5)_{a=a_1}$ has a cusploop Γ. It is easy to see that Γ is unstable from the inside. We conjecture that when the cusploop Γ exists, it is surrounded by a large stable limit cycle, and hence Γ is also unstable from the outside. However, we were not able to exclude the possibility that for some λ the cusploop is semistable, i.e., $a_2(\lambda)=a_1(\lambda)$.

Thus O, A' form an unstable singular point system ($a\geqq-\frac{3}{2}\lambda$), an unstable singular point-cycle system ($a_1\leqq a<-\frac{3}{2}\lambda$) or O, A' form a stable singular point system ($a<a_1$). The remaining conclusions of Theorem 5 can be deduced at once from Lemmas 2–6.

Remark. Because the parameter a rotates the vector field in the region K only, we were not able to prove the uniqueness of the semistable limit cycle bifurcation set. We conjecture that this bifurcation set is unique (i.e., $a_2=a_3$ in Theorem 5). If it could be shown that (5) has no limit cycles intersecting $x+2\lambda=0$, then the conjecture is proved.

5. Concluding remarks. In this section we give the bifurcation diagram (Fig. 5) and the corresponding phase portraits (Fig. 6) for system (4), using the conjectures stated above. The bifurcation sets in Fig. 5 correspond to Hopf-bifurcation (H), the cusploop (CL), and a semistable limit cycle (SS). The results follow directly from Theorem 5. Notice that in order to transform (4) to (5), the transformation $\tau=-t$ has been used.

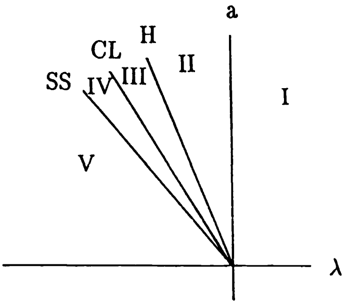
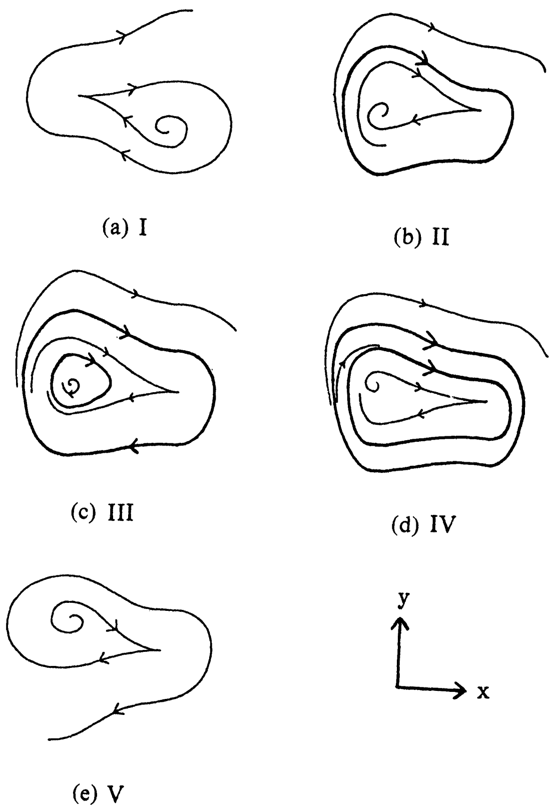


FIG. 5. *The bifurcation diagram.*

If in system (4), a and λ are taken to be small parameters, then (4) can be considered a partial unfolding of a nilpotent focus of codimension 4. We call it a focus of codimension 4, copying some arguments in [8], since normal form theory (see [12]) together with the Tarski–Seidenberg decision theorem (see Appendix [1]) show that



(a) I

(b) II

(c) III

(d) IV

(e) V

FIG. 6. *Phase portraits of system* (4).

this singularity lies on a semialgebraic set of codimension 4. We do not know whether there exists a theorem that asserts that 4 parameters suffice to fully describe a versal unfolding of such a singularity. We conjecture that the following system is a versal unfolding of the nilpotent focus of codimension 4:

$$\dot{x} = y + \mu_1 x + \mu_2 x^2 + x^3, \qquad \dot{y} = \mu_3 x + \lambda x^2 - x^3. \tag{13}$$

Using Theorem 5 and the well-known results concerning the Takens-Bogdanov system, describing the unfolding of a cusp of codimension 2, it follows that system (13) can have three limit cycles and that the distribution of limit cycles as given in Fig. 7 is realizable. Obviously, system (13) can be brought into the equivalent form

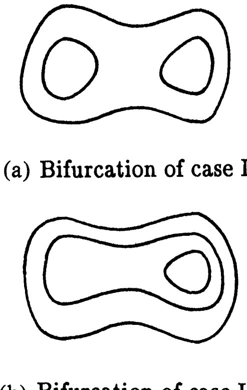
$$\dot{x} = y, \qquad \dot{y} = \lambda_1 x + \lambda_2 x^2 - x^3 + (\nu_1 + \nu_2 x + x^2) y. \tag{14}$$

In [7] Dumortier, Roussarie, and Sotomayor have studied system (14) with ν_2 fixed $(0 < \nu_2 < 2\sqrt{2})$ and considered this system as a versal unfolding of the nilpotent focus of codimension 3. Dangelmeyer and Guckenheimer [5] also obtained some results for system (14).

Appendix A. System (4) with $\lambda = 0$. For $\lambda = 0$ system (4) reduces to

$$\frac{dx}{dt} = y + ax^2 + x^3, \qquad \frac{dy}{dt} = -x^3. \tag{A1}$$

For $a^2 < 2$ the origin $O(0, 0)$ of system (A1) is a nilpotent focus. We will show that it is globally unstable.



(a) Bifurcation of case III

(b) Bifurcation of case IV

FIG. 7. *Limit cycles for system* (14).

Consider the following system:

$$\text{(A2)}\qquad \frac{dx}{dt}=y+ax^2=P(x,y),\qquad \frac{dy}{dt}=-x^3=Q(x).$$

It is obvious that the trajectories of (A2) form a family of closed curves because $P(-x,y)=P(x,y)$, and $Q(-x)=-Q(x)$.

The first integral of (A2) reads

$$V(x,y)=\frac{1}{4}\log(x^4+2ax^2y+2y^2)+\frac{a}{2\omega}\tan^{-1}\left(\frac{ax^2+2y}{\omega x^2}\right)=C,$$

with $\omega=\sqrt{2-a^2}$ and $a^2<2$.

The derivative of $V(x,y)$ along trajectories of (A1) satisfies

$$\left.\frac{dV}{dt}\right|_{(\mathrm{A1})}=\frac{\partial V}{\partial x}(y+ax^2+x^3)-\frac{\partial V}{\partial y}x^3=\frac{x^6}{x^4+2ax^2y+2y^2}\geqq 0\quad \text{for } a^2<2.$$

It follows that the origin $O(0,0)$ of (A1) is globally unstable for $a^2<2$.

Appendix B. Singularities at infinity of system (4). System (4) reads

$$\text{(B1)}\qquad \dot{x}=y+ax^2+x^3,\qquad \dot{y}=\lambda x^2-x^3.$$

It is easy to check that (B1) has no singularity at the end of the x-axis. Therefore, it suffices to use the Poincaré transformation $z=1/y$, $v=x/y$, $dt/d\tau=z^2$, to study the singularities at infinity. System (B1) reduces to

$$\text{(B2)}\qquad \frac{dz}{d\tau}=-\lambda z^2v^2+zv^3,\qquad \frac{dv}{d\tau}=z^2+azv^2+v^3-\lambda zv^3+v^4.$$

The singularities of (B2) on $z=0$ are $I_1(0,0)$ and $I_2(0,-1)$.

First let us study $I_1(0,0)$.

Using a polar blowup, $z=r\cos\varphi$, $v=r\sin\varphi$, $-\pi\leqq\varphi\leqq\pi$, system (B2) changes into

$$\text{(B3)}\qquad \begin{aligned}\dot{r}&=r\cos^2\varphi\sin\varphi+r^2(a\cos\varphi\sin^3\varphi+\sin^4\varphi)+r^3(-\lambda\cos\varphi\sin^2\varphi+\sin^3\varphi),\\ \dot{\varphi}&=\cos^3\varphi+r(a\cos^2\varphi\sin^2\varphi+\cos\varphi\sin^3\varphi).\end{aligned}$$

The singularities of system (B3) on $r=0$ are $A_1(0,-(\pi/2))$ and $A_2(0,\pi/2)$.

Linearization of (B3) near A_1, with $\varphi = -(\pi/2) + \xi$, yields

$$\dot{r} = r^2 - r^3 - r\xi^2 - ar^2\xi - \lambda r^3\xi, \qquad \dot{\xi} = \xi^3 + ar\xi^2 - r\xi. \tag{B4}$$

If in (B4) we make the change of variables $(r, t) \to (-r, -t)$, then we obtain the linearization of (B4) near A_2 (with $\varphi = (\pi/2) + \xi$).

So, in order to study the topological character of A_1 and A_2, it suffices to study the phase portrait of (B4) near the origin both for $r < 0$ and for $r > 0$.

A second blowup, $r = \eta \cos\theta$, $\xi = \eta \sin\theta$, with $-(\pi/2) \leqq \theta \leqq \pi/2$, changes (B4) into

$$\begin{aligned}\dot{\eta} &= \eta(\cos^3\theta - \cos\theta\sin^2\theta) + \eta^2(-\cos^2\theta - a\cos^3\theta\sin\theta + a\cos\theta\sin^3\theta + \sin^4\theta)\\ &\quad -\lambda\eta^3\cos^4\theta\sin\theta,\\ \dot{\theta} &= -2\cos^2\theta\sin\theta + \eta(\cos^3\theta\sin\theta + 2a\cos^2\theta\sin^2\theta + 2\cos\theta\sin^3\theta)\\ &\quad +\lambda\eta^2\cos^3\theta\sin^2\theta.\end{aligned} \tag{B5}$$

The singularities of (B5) on $\eta = 0$ are $C_1(0, -(\pi/2))$, $C_2(0, 0)$, and $C_3(0, \pi/2)$.

Linearization of (B5) near C_1, with $\theta = -(\pi/2) + \delta$, yields

$$\begin{aligned}\dot{\eta} &= \eta^2 - \eta\delta - a\eta^2\delta - \eta^2\delta^2 + \eta\delta^3 + a\eta^2\delta^3 + \lambda\eta^3\delta^4,\\ \dot{\delta} &= -2\eta\delta + 2\delta^2 + 2a\eta\delta^2 - \eta\delta^3 + \lambda\eta^2\delta^3.\end{aligned} \tag{B6}$$

Near C_2 system (B5) has a linear part $\dot{\eta} = \eta$, $\dot{\theta} = -2\theta$, and it follows that C_2 is a hyperbolic saddle.

If in (B6) we make the change of variables $(\delta, a, \lambda) \to (-\delta, -a, -\lambda)$, we obtain the linearization of (B6) near C_3 (with $\theta = (\pi/2) + \delta$).

A third blowup, $\eta = \rho\cos\mu$, $\delta = \rho\sin\mu$, with $0 \leqq \mu \leqq \pi$, changes (B6) into a system with four singularities on $\rho = 0$; $D_1(0, 0)$, $D_2(0, \pi/4)$, $D_3(0, \pi/2)$, and $D_4(0, \pi)$. We do not give the complete system after the blowup, just the linearization near D_i, $i = 1, 2, 3, 4$, with $\mu = \mu_0 + \sigma$:

$$\begin{aligned}&D_1: &&\dot{\rho} = \rho + o(|\rho, \sigma|), &&\dot{\sigma} = -3\sigma + o(|\rho, \sigma|),\\ &D_3: &&\dot{\rho} = 2\rho + o(|\rho, \sigma|), &&\dot{\sigma} = -3\sigma + o(|\rho, \sigma|),\\ &D_4: &&\dot{\rho} = -\rho + o(|\rho, \sigma|), &&\dot{\sigma} = 3\sigma + o(|\rho, \sigma|),\end{aligned}$$

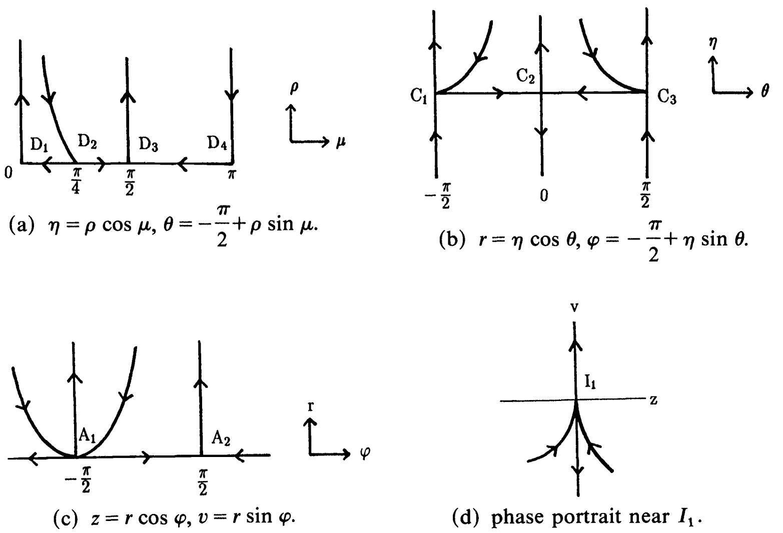
$$\begin{aligned}D_2:\quad &\dot{\rho} = \frac{a}{4}\rho^2 + \frac{\rho\sigma}{\sqrt{2}} - \frac{1}{4\sqrt{2}}\rho^3 + \frac{3}{2}a\rho^2\sigma + 3\sqrt{2}\,\rho\sigma^2 + o(|\rho, \sigma|^3),\\ &\dot{\sigma} = \frac{3}{4}a\rho + \frac{3\sigma}{\sqrt{2}} - \frac{1}{4\sqrt{2}}\rho^2 + \frac{\lambda - a}{8}\rho^3 - \frac{3}{4\sqrt{2}}\rho^2\sigma - \frac{3}{2}a\rho\sigma^2 - \frac{3\sigma^3}{\sqrt{2}} + o(|\rho, \sigma|^3).\end{aligned} \tag{B7}$$

Obviously, D_1, D_3, and D_4 are hyperbolic saddles. D_2 is a semihyperbolic saddle, as can be seen after transforming

$$\omega = \frac{3\sigma}{\sqrt{2}} + \frac{3}{4}a\rho - \frac{1}{4\sqrt{2}}\rho^2$$

in system (B7):

$$\dot{\rho} = -\frac{1}{6\sqrt{2}}\rho^3 + \omega O(|\rho, \omega|) + o(|p|^3), \qquad \dot{\omega} = \frac{3\omega}{\sqrt{2}} + o(|\rho, \omega|).$$



(a) $\eta = \rho \cos \mu$, $\theta = -\frac{\pi}{2} + \rho \sin \mu$.

(b) $r = \eta \cos \theta$, $\varphi = -\frac{\pi}{2} + \eta \sin \theta$.

(c) $z = r \cos \varphi$, $v = r \sin \varphi$.

(d) phase portrait near I_1.

FIG. B1

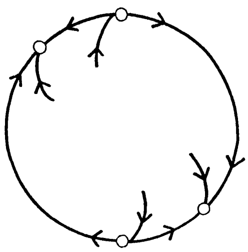


FIG. B2

After three blowups we have obtained only isolated singularities that are hyperbolic or semihyperbolic, so we can determine the topological character of I_1 in system (B2). Notice that the topological character of I_1 is independent of a and λ.

In Fig. B1 the blow-down for the singularity I_1 is sketched. After linearization of system (B1) near I_2 (take $v = -1 + w$), the linear part of (B2) becomes $z' = -z$, $w' = (a + \lambda)z - w$. It follows that I_2 is a stable improper node, independent of a and λ.

In Fig. B2 the topological behavior of system (B1) near infinity is sketched. The Poincaré equator is attracting for all values of a and λ.

REFERENCES

[1] R. Abraham and J. Robbin, *Transversal Mappings and Flows*, Benjamin, New York, 1967.

[2] A. A. Andronov, E. A. Leontovitch, I. I. Gordon, and A. G. Maier, *Qualitative Theory of Second-Order Dynamic Systems*, Halsted, New York, 1973.

[3] P. G. Bakker, *Bifurcations in flow patterns*, in Nonlinear Topics in the Mathematical Sciences, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1991.

[4] P. G. Bakker and M. E. M. de Winkel, *On the topology of three-dimensional separated flow structures and local solutions of the Navier-Stokes equations*, in Topological Fluid Mechanics, Proceedings of the IUTAM Symposium, Cambridge, MA, August 13-18, 1989, pp. 384-394.

[5] G. Dangelmeyer and J. Guckenheimer, *On a four parameter family of planar vector fields*, Arch. Rational Mech. Anal., 97 (1987), pp. 321-352.

[6] F. Dumortier, *Singularities of vector fields on the plane*, J. Differential Equations, 23 (1977), pp. 53-106.

[7] F. Dumortier, R. Roussarie, and J. Sotomayor, *Generic 3-parameter families of plane vector fields, unfoldings of saddle, focus and elliptic singularities with nilpotent linear parts*, Lecture Notes in Math., 1480, Springer-Verlag, New York, 1991.

[8] F. Dumortier and P. Fiddelaers, *Quadratic models for generic local 3-parameter bifurcations on the plane*, Trans. Amer. Math. Soc. (1991), pp. 101-126.

[9] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems and bifurcations of vector fields*, Appl. Math. Sci., 42 (1983), pp. 150-153.

[10] R. E. Kooij and P. G. Bakker, *Three-dimensional viscous flow structures from bifurcation of a degenerate singularity with three zero eigenvalues*, Report LR-572, Faculty of Aerospace Engineering, Delft University of Technology, Delft, the Netherlands, 1989.

[11] L. M. Perko, *Rotated vector fields and the global behaviour of limit cycles for a class of quadratic systems in the plane*, J. Differential Equations, 18 (1975), pp. 63-86.

[12] F. Takens, *Singularities of vector fields*, Publ. Math. IHES, 43 (1974), pp. 47-100.

[13] Wang Xian, *Some remarks on the limit cycles for systems of Liénard type*, Chinese Quart. J. Math., (1990), pp. 1-5.

[14] ———, *On the uniqueness of limit cycles of the system $\dot{x} = \varphi(y) - F(x)$, $\dot{y} = -g(x)$*, J. Nanjing Univ., 26 (1990), pp. 363-372.

[15] ———, *Some cubic Liénard systems with at most one limit cycle*, Ann. Differential Equations, 7 (1991), pp. 94-102.

[16] Ye Yanqian, *Theory of limit cycles*, Transl. Math. Monographs, 66 (1986), pp. 91-102.

[17] Zhou Yurong and Han Maoan, *The conditions of having at most one or two limit cycles surrounding several singular points*, 1990, Acta Math. Sinica, to appear. (In Chinese.)

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1623-1636, November 1992

011

ELIMINATING THE GENERICITY CONDITIONS IN THE SKEW TOEPLITZ OPERATOR ALGORITHM FOR H^{∞}-OPTIMIZATION*

CAIXING GU†

Abstract. In this paper the genericity conditions are eliminated in the skew Toeplitz operator algorithm for H^{∞}-optimization of [*Oper. Theory: Adv. Appl.*, 49 (1988), pp. 21-43], [*Systems Control Lett.*, 11 (1988), pp. 259-264]. This paper gives an explicit formula for a class of weighted sensitivity minimization problems.

Key words. H^{∞}-optimization, skew Toeplitz, Hankel operator, weighted sensitivity

AMS(MOS) subject classifications. 47A20, 93B35, 93C05

1. Introduction. It is well known now that a number of problems in control engineering can be reduced to certain generalized interpolation problems in H^{∞} [1]. Namely, let $w, m \in H^{\infty}$ with w rational and m nonconstant inner. Then the H^{∞}-optimal weighted sensitivity problem amounts to the computation of

$$\rho_0 = \inf \|w - mq\|_{\infty}, \qquad q \in H^{\infty},$$

and find the corresponding q_{opt} that reaches the bound.

This problem has been studied from an operator point of view in a number of papers (see [1]-[4] and the reference therein). Particularly, in [1], [3], [4] Bercovici, Foias, Tannenbaum, and Zames introduced the skew Toeplitz operators and developed several algorithms for the computation of ρ_0 and q_{opt}. The point of this paper is to eliminate the genericity conditions in the algorithm given in [4] (see also [2, § XII]).

To be more precise, recall that ρ_0 is the norm of a certain operator (which is equivalent to the Hankel operator [1]). q_{opt} can be computed from the corresponding singular vector. Namely, $S: H^2 \to H^2$ denotes the unilateral shift (all of our Hardy spaces will be defined on the unit disc D in the standard way) and $P_H: H^2 \to H^2 \ominus mH^2 = H$ denotes the orthogonal projection. Let $T = S(m) = P_H S | H$, then $\rho_0 = \|W(T)\|$: here $W(T) = P_H M_W | H$, M_w is the multiplication operator by w. In [1]-[4], under some specified genericity conditions an algorithm is given to compute ρ_0 and q_{opt} by reducing the problem to the noninvertibility of an $n \times n$ matrix; here $n := \max\{\text{degree } p, \text{degree } q\}$ for relatively prime polynomials p and q such that $w = p/q$. This paper eliminates those genericity conditions, and also reduces the computation of ρ_0 to the noninvertibility of an $(n+l) \times (n+l)$ matrix (see Theorem 3). So the algorithm is now complete (see § 4 for the algorithm).

2. Some basic facts and equalities. Throughout this paper, for a linear bounded operator A on a complex Hilbert space K, we denote its spectrum by $\sigma(A)$, its essential spectrum by $\sigma_e(A)$, and its essential norm by $\|A\|_e$. Recall that the essential spectrum and essential norm are, respectively, the spectrum and norm in the Calkin algebraic (i.e., the quotient of the space of all operators modulo the compact operators). In this

* Received by the editors April 22, 1991; accepted for publication (in revised form) March 31, 1992. This work was supported by a grant from the Research Fund of Indiana University.

† Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

section we present some facts from [2, § XII, 3] that are essential for our work. Let

$$\begin{aligned} T_\rho &= q(T)(\rho^2 - W(T)W(T)^*)q(T)^* \\ &= \rho^2 q(T)q(T)^* - p(T)p(T)^* \\ &= \sum_{0,0}^{n,n} C_{jk}T^jT^{*k}, \end{aligned}$$

where

$$C_{jk} = \rho^2 q_j\bar{q}_k - p_j\bar{p}_k \tag{2.1}$$

and p_j, respectively, q_j are coefficients for z^j in p, respectively, q.

It is clear that T_ρ is not invertible if and only if $\rho^2 I - W(T)W(T)^*$ is not invertible, or equivalently, $\rho^2 \in \sigma(W(T)W(T)^*)$. If $\rho > \|W(T)\|_e$, then the noninvertibility of T_ρ is equivalent to ρ^2 being an eigenvalue of $W(T)W(T)^*$; the largest such $\rho > 0$ is equal to $\inf \|w - mq\|_\infty$, $q \in H^\infty$.

By dropping ρ, we are thus lead to the following problem: given a skew Toeplitz operator [1],

$$T_\rho = \sum_{0,0}^{n,n} C_{jk}T^jT^{*k}$$

with the property that $0 \in \sigma(T_\rho)$ if and only if zero is an eigenvalue of T_ρ (which is equivalent to $\{z, C(z) = 0\} \cap \{z \in \partial D, z \in \sigma(T)\} = \phi$. Here by [8], [9], $\sigma(T) = \{$zeros of m in D and essential singularities of m on $\partial D)$, determine if $0 \in \sigma(T_\rho)$ or not. If $0 \in \sigma(T_\rho)$, find a nonzero y in H such that $T_\rho y = 0$.

Now recall that the unilateral shift on H^2 is the minimal isometric dilation of $T = S(m)$ [5], [6]. This implies that

$$T_\rho = P_H \sum_{0,0}^{n,n} C_{jk}S^jS^{*k} \,|\, H. \tag{2.2}$$

In particular, $T_\rho y = 0$ for some y in H if and only if

$$\sum_{0,0}^{n,n} C_{jk}S^jS^{*k}y = mg \tag{2.3}$$

for some $g \in H^2$. If y_j is the coefficient of z^j in the power series expansion of y, then

$$\begin{aligned} S^jS^{*k}y &= S^j[z^{-k}(y - y_0 - zy, \cdots, -z^{k-1}y_{k-1})] \\ &= z^{j-k}(y - y_0 - zy, \cdots, -z^{k-1}y_{k-1}). \end{aligned} \tag{2.3a}$$

Now let $C(z)$, respectively $C_i(z)$ (for $0 \leqq i < n$) be the polynomials of degree less than or equal to $2n$, respectively, $2n-1$ defined by

$$\begin{aligned} C(z) &= \sum_{0,0}^{n,n} C_{jk}z^{n+j-k} = \sum_{j=0}^{2n} a_jz^j, \\ C_i(z) &= \sum_{i<k\leqq n, 0 \leqq j \leqq n} C_{jk}z^{n+j-k+i} = \sum_{j=1}^{2n-1} a_{ij}z^j, \end{aligned} \tag{2.3b}$$

where C_{jk} is given by (2.1). We notice that for $i \leqq j < n$, we have

$$a_{ij} = a_{j-i}. \tag{2.3c}$$

Substituting (2.3a) into (2.3) and using the definition of $C(z)$ and $C_i(z)$'s, we conclude that $T_\rho y = 0$ if and only if

$$(2.4)\qquad C(z)y(z) - \sum_{i=0}^{n-1} C_i(z)y_i = z^n m(z)q(z)$$

for some $g \in H^2$.

Now we can state the genericity conditions assumed in [4] (see also [2, § XII]).

$$(\mathrm{A})\qquad C(z) \neq 0 \quad \text{for all } z \in \sigma(T) \cup \{0\}.$$

Under the assumption (A), since

$$C(T) = \beta\Pi(T - z_0 I),$$

where $\beta \neq 0$ and z_0 runs over the zeros of $C(z)$ (thus z_0 is not in $\sigma(T)$ so each $T - z_0 I$ is invertible), we deduce that $C(T)$ is invertible. Obviously $C(T)^{-1}$ commutes with T. By Sarason's theorem [10] (see also [2, § IX, 7.4]) there exists $C^{(-1)}$ in H^∞ such that $C^{(-1)}(T) = C(T)^{-1}$. Moreover, using the fact that $(C^{(-1)}C)(T) = I$, we see that there exists an h_1 in H^∞ satisfying

$$(2.5)\qquad C^{(-1)}C = 1 + mh_1.$$

By multiplying (2.4) by $C^{(-1)}$, we obtain

$$(2.6)\qquad y(z) - \sum_{i=0}^{n-1} (C^{(-1)}C_i)(z)y_i = m(z)[z^n C^{(-1)}g(z) - h_1(z)y(z)] = m(z)h(z),$$

where $h(z) \in H^2$.

Applying P_H to the previous equation gives

$$(2.7)\qquad y = \sum_{i=0}^{n-1} y_i P_H C^{(-1)} C_i = \sum_{i=0}^{n-1} y_i \chi_i,$$

where using $P_H C^{(-1)}(S) = C(T)^{-1} P_H$ (see [2] for the definition of $C^{(-1)}(S)$) then

$$(2.7\mathrm{a})\qquad \chi_i = P_H C^{(-1)} C_i = C(T)^{-1} P_H C_i \quad (\text{for } 0 \leqq i < n)$$

does not depend on the particular choice of $C^{(-1)}$ and is obviously in H. Substituting (2.7) into (2.4) produces

$$(2.8)\qquad \sum_{i=0}^{n-1} y_i (C P_H C^{(-1)} C_i - C_i)(z) = z^n m(z) g(z) \qquad (z \in D).$$

Using in addition the lifting property $P_H C(S) P_H = P_H C(S)$, we have $P_H C P_H C^{(-1)} C_i - C_i = P_H C C^{(-1)} C_1 - C_i = (P_H - I)C_i$ in mH^2 $(0 \leqq i < n)$. So by (2.7a) and (2.8) there exists a sequence of functions g_i in H^2 such that

$$(2.9)\qquad C\chi_i - C_i = mg_i \quad (\text{for } 0 \leqq i < n),$$

$$(2.10)\qquad \sum_{i=0}^{n-1} y_i g_i(z) = z^n g(z) \qquad (z \in D).$$

3. Main results. First we will eliminate the condition $C(0) \neq 0$ in A. Namely, we assume $C(z)$ has a zero at zero of order k $(k \leqq n)$ and

$$(\mathrm{A1})\qquad \{z, C(z) = 0\} \cap \sigma(T) = \phi.$$

Let us introduce some notations. For $\alpha \in D$, $f \in H^2$, let $f = \sum_{i=0}^{\infty} f_i (z - \alpha)^i$ be the power series expansion of f at α. Define $\Pi_s(\alpha): H^2 \to H^2$ by

$$\Pi_s(\alpha) f = \sum_{i=0}^{s-1} f_i (z - \alpha)^i = [f_0 \cdots f_{s-1}].$$

It is clear from the definition that for $f, h \in H^2$,

$$\Pi_s(\alpha)f\cdot h = [f_0h_0, f_1h_0+f_0h_1, \cdots] \tag{3.0}$$
$$= \begin{bmatrix} f_0 & & & 0 \\ f_1 & \ddots & & \\ \vdots & \ddots & \ddots & \\ f_{s-1} & \cdots & f_1 & f_0 \end{bmatrix} \begin{bmatrix} h_0 \\ \vdots \\ h_{s-1} \end{bmatrix}.$$

Let $\Pi_n(0)g_i(z) = [g_{i0}, g_{i1}, \cdots, g_{in+1}]$ and $\Pi_n(0)\chi_i(z) = [\chi_{i,0}, \chi_{i,1}, \cdots, \chi_{i,n-1}]$. Notice that under the assumption (A1) all equalities in § 2 are valid. Applying $\Pi_n(0)$ to (2.10) and (2.7), respectively, we get

$$\sum_{i=0}^{n-1} y_i g_{ij} = 0 \qquad (0 \leqq j < n), \tag{3.0a}$$

$$\sum_{i=0}^{n-1} y_i \chi_{ij} = y_j \qquad (0 \leqq j < n). \tag{3.0b}$$

It seems that we get $2n$-equations for n-unknown $[y_0, \cdots, y_{n-1}]$. Soon we will see that some equations are redundant.

Notice by (2.3b, c) that the first n coefficients of polynomials $C_i(z)$ are

$$\Pi_n(0)C_i(z) = \sum_{j=i}^{n-1} a_{ij}z^j = z^i \sum_{j=0}^{n-1-i} a_j z^j \tag{3.1}$$
$$= [0 \cdots 0, a_0, \cdots, a_{n-1-i}] \qquad (0 \leqq i < n).$$

THEOREM 1. *Under the assumption* (A1) *if* $T_\rho y = 0$ *for some nonzero function* y *in* H, *then* y *is given by* (2.7), *where* $[y_0, \cdots, y_{n-1}]$ *is a nonzero solution of*

$$\sum_{i=0}^{n-1} y_i g_{ij} = 0 \qquad (k \leqq j < n) \tag{3.2a}$$

and

$$\sum_{i=0}^{n-1} y_i \chi_{ij} = y_j \qquad (n-k \leqq j < n), \tag{3.2b}$$

or, equivalently,

$$\begin{bmatrix} G_{n-k} \\ H_k \end{bmatrix} [y_0, \cdots, y_{n-1}]' = 0, \tag{3.2}$$

where $[y_0 \cdots y_{n-1}]'$ *is the transpose of vector* $[y_0 \cdots y_{n-1}]$ *and*

$$G_{n-k} = \begin{bmatrix} g_{0k} & \cdots & g_{n-1,k} \\ \vdots & & \vdots \\ g_{0,n-1} & \cdots & g_{n-1,n-1} \end{bmatrix}_{(n-k)\times n},$$

$$H_k = \begin{bmatrix} \chi_{0k} & \cdots & \chi_{n-1,k} \\ \vdots & & \vdots \\ \chi_{0,n-1} & \cdots & \chi_{n-1,n-1} \end{bmatrix} - \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \\ 0 & \cdots & 0 & 1 & \\ \vdots & & \vdots & & \ddots \\ 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}.$$

Conversely, if $[r_0, \cdots, r_{n-1}]$ *is a nonzero solution of* (3.2), *then the function* y *given by* (2.7) *with* $y_i = r_i$ $(0 \leqq i < n)$ *is a nonzero function in* H *satisfying* $T_\rho y = 0$ *and has its first* n *Taylor coefficients given by* $y_i = r_i$ $(0 \leqq i < n)$.

Proof. From the above analysis, it follows that if y is nonzero in H such that $T_\rho y=0$, then y is given by (2.7), and $[y_0,\cdots,y_{n-1}]$ is nonzero satisfying (3.0a) and (3.0b); hence (3.2a) and (3.2b).

Conversely, let $y=\sum r_i\chi_i$. Observe that (A1) implies $a_0=a_1=\cdots=a_{k-1}=0$. By (3.1), $C(z)=z^kd(z)$, $C_i(z)=z^kd_i(z)$ $(0\leqq i<n)$, where $d(z)$, $d_i(z)$ are polynomials. Therefore, (2.9) becomes

$$z^k(d(z)\chi_i-d_i(z))=mg_i \qquad (0\leqq i<n).$$

Thus

$$g_{ij}=0 \quad \text{for } 0\leqq i<n, \quad 0\leqq j\leqq k-1. \tag{3.2c}$$

Hence (3.2a) is equivalent to (3.0a). So by (2.10) and (2.9) we have

$$C(z)y(z)-\sum_{i=0}^{n-1} r_iC_i(z)=\sum_{i=0}^{n-1} r_i(C\chi_i-C_i)=\sum_{i=0}^{n-1} r_img_i=z^nmg. \tag{3.2d}$$

Since (2.4) holds if and only if $T_\rho y=0$, it suffices to prove that the first n Taylor coefficients of y, namely, $y_0, y_1,\cdots,y_{n-1}$ coincide with $r_0, r_1,\cdots,r_{n-1}$. Applying $\Pi_n(0)$ to (3.2d) yields

$$\Pi_n(0)C(z)y(z)-\sum_{i=0}^{n-1} r_i\Pi_n(0)C_i(z)=0.$$

Notice that

$$\sum r_i\Pi_n(0)C_i(z)=\sum_{i=0}^{n-1} r_i[0\cdots 0, a_0,\cdots,a_{n-1-i}]$$
$$=A[r_0,\cdots,r_{n-1}]'$$

where

$$A=\begin{bmatrix} a_0 & & & 0\\ a_1 & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n-1} & \cdots & a_1 & a_0\end{bmatrix}.$$

By (3.0) and (3.1) we obtain

$$A[y_0-r_0,\cdots,y_{n-1}-r_{n-1}]'=0. \tag{3.2e}$$

By (A1), $a_0=a_i=\cdots=a_{k-1}=0$ but $a_k\neq 0$. So by (3.2e), $y_i=r_i$ $(0\leqq i<n-k)$. But $y_i=r_i$ $(n-k\leqq i<n)$ is exactly (3.2b). Thus $y_i=r_i$ $(0\leqq i<n)$. This also shows that $y=0$ if and only if $[y_0,\cdots,y_{n-1}]=0$; therefore, y is nonzero. □

Remark 2. For $k=0$, Theorem 1 reduces to Proposition 3 in [2, § XII, 3]. For the $k>n$ we have $C(z)\equiv 0$ (see § 4 below). Of course in this case, Assumption (A1) is not satisfied, but we can nevertheless characterize ker T_ρ. Namely, now by (2.4) $T_\rho y=0$, $y\in H$, if and only if

$$\sum_{i=0}^{n-1} y_iC_i(z)=z^nm(z)g(z)$$

for some $g(z)\in H^2$. By (3.1), $C_i(z)=z^nd_i(z)$ $(0\leqq i<n)$, where $d_i(z)$ are polynomials of degree at most $n-1$. So we have

$$\sum_{i=0}^{n-1} y_id_i(z)=m(z)g(z). \tag{3.2f}$$

Now let $l(z)=\sum_{i=0}^{n-1} y_id_i(z)$. There are two situations.

(i) If $m(z)$ is a Blaschke product of degree at most $n-1$, namely,

$$m(z)=\prod_{i=1}^{s}\left(\frac{z-\alpha_i}{1-\bar{\alpha}_i z}\right)^{l_i}, \qquad l_1+\cdots+l_s \leqq n-1,$$

then y_i satisfying (3.2f) for some $g\in H^2$ if and only if y_i is a solution of the system

$$\text{(3.2g)} \qquad l^{(j)}(\alpha_i)=0 \qquad 1\leqq i\leqq s, \qquad 0\leqq j<l_i,$$

where $l^{(j)}(\alpha_i)$ means the jth derivative of $l(z)$ evaluated at α_i.

(ii) If $m(z)$ is not a Blaschke product of degree at most $n-1$, namely, either $m(z)$ has a singular part or $m(z)$ has more than $n-1$ roots in D; then (3.2f) is valid for some $g\in H^2$ if and only if $\sum y_i d_i(z)\equiv 0$, i.e.,

$$\text{(3.2h)} \qquad \sum_{i=0}^{n-1} a_{i,j+n}y_i=0 \qquad 0\leqq j<n,$$

where $a_{i,j+n}$ is given by (2.3c).

Notice that in both cases we do not assume $m(0)\neq 0$. Let us denote by Y_1 (respectively, Y_2) the set

$$\text{(3.2i)} \qquad \begin{aligned} Y_1 \text{ (respectively, } Y_2) = \{y\neq 0,\ Y\in H,\ y(z) &\text{ has its first} \\ n\text{-Taylor coefficients } &y_0,\cdots,y_{n-1} \\ \text{satisfying (3.2g)(respectively, (3.2h))}&\}. \end{aligned}$$

From the above discussion, we see that

$$\text{(3.2j)} \qquad \ker T_\rho := \{y\neq 0, y\in H, T_\rho y=0\} = Y_1 \text{ (respectively, } Y_2).$$

Next we consider another nongeneric case, namely, we make the assumption (A2):

$$\text{(A2)} \qquad C(0)\neq 0, \qquad \{z, C(z)=0\}\cap\{\sigma(T)\}=\{\alpha\}.$$

(Recall that by § 2, $\alpha\in D$.)

We need the following facts (for detail see [2, § X.1]).

LEMMA 3. *Let $m(z)$ be an inner function in H^∞ such that*

$$m(z)=\prod_{i=0}^{k}\left(\frac{z-\alpha_i}{1-\bar{\alpha}_i z}\right)^{l_i} m_1(z)=m_0(z)m_1(z), \quad \textit{and} \quad (m_0(z),m_1(z))=1.$$

Let

$$\sigma_{ij}=\left[\frac{(z-\alpha_i)^{l_i-j-1}}{(1-\bar{\alpha}_i z)^{l_i}}\right]\prod_{\substack{s=0\\ s\neq i}}^{k}\left(\frac{z-\alpha_s}{1-\bar{\alpha}_s z}\right)^{l_s} \qquad 0\leqq i\leqq k, \quad 0\leqq j<l_i.$$

Then

$$H(m)=H(m_1)\oplus m_1H(m_0)=H(m_1)\oplus m_i\left\{\sum_{ij} C_{ij}\sigma_{ij},\ C_{ij}\in\mathbb{C}\right\}.$$

Now assume (A2) holds. Let $m(z)=(z-\alpha)/(1-\bar{\alpha}z)^l m_1(z)$, with $m_1(\alpha)\neq 0$ by virtue of (A2), $l\geqq 1$. Let

$$\sigma_j=(z-\alpha)^{l-j-1} \qquad 0\leqq j<l.$$

By Lemma 3.2,

$$H=H(m)=H(m_1)\oplus\left\{\sum_{j=0}^{l-1}\beta_j\frac{m_1}{(1-\bar{\alpha}z)^l}\sigma_j,\ \beta_j\in\mathbb{C}\right\}=H_1\oplus H_0.$$

Let P_{H_i} be the corresponding projection P_{H_i}: $H^2\to H_i$, $i=0,1$.

Let y in H such that $T_\rho y = 0$. The idea is to decompose y into

$$y = P_{H_1}y + P_{H_0}y = P_{H_1}y + \frac{m_1(z)}{(1-\bar{\alpha}z)^l}\sum_{j=0}^{l-1}\beta_j\sigma_j, \tag{3.3}$$

and for $P_{H_1}y$ to apply the techniques in § 2. By (2.4), we have

$$C(z)y(z) - \sum_{i=0}^{n-1} C_i(z)y_i = z^n m_1(z)\left(\frac{z-\alpha}{1-\bar{\alpha}z}\right)^l g(z). \tag{3.4}$$

Since $\sigma(S(m_1)) = \{$zeros of m_1 in D and essential singularities of m_1 on $\partial D\}$, we have $\{z, C(z)=0\} \cap \sigma(S(m_1)) = \phi$ by virtue of (A2). Thus as for (2.5), there exists $C^{(-1)} \in H^\infty$ such that

$$C^{(-1)}C = 1 + m_1h_1. \tag{3.5}$$

Hence

$$\begin{aligned} y(z) - \sum_{i=0}^{n-1} C^{(-1)}C_iy_i &= m_1(z)\left[z^n\left(\frac{z-\alpha}{1-\bar{\alpha}z}\right)^l C^{(-1)}g(z) - h_1(z)y(z)\right] \\ &= m_1(z)h(z). \end{aligned} \tag{3.6}$$

Applying P_{H_1} to the previous equation yields

$$P_{H_1}y(z) = \sum_{i=0}^{n-1} y_iP_{H_1}(C^{(-1)}C_i)(z) = \sum_{i=0}^{n-1} y_i\chi_i, \tag{3.7a}$$

where $\chi_i \in H_1 = H(m_1)$. Substituting (3.7a) into (3.3), we obtain

$$y = \sum_{i=0}^{n-1} y_i\chi_i + \frac{m_1(z)}{(1-\bar{\alpha}z)^l}\sum_{j=0}^{l-1}\beta_j\sigma_j. \tag{3.7}$$

Substituting (3.7) into (3.4) and multiplying both sides by $(1-\bar{\alpha}z)^l$ gives

$$\begin{aligned} &m_1C(z)\left(\sum_{j=0}^{l-1}\beta_j\sigma_j\right) + (1-\bar{\sigma}z)^l\sum_{i=0}^{n-1}(C\chi_i - C_i)(z) \\ &= z^n(z-\alpha)^l m_1(z)g(z) \end{aligned} \tag{3.8}$$

as for (2.9), we have that there exists $g_i \in H^2$ such that

$$C\chi_i - C_i = m_1(z)g_i(z) \qquad (0 \leqq i < n). \tag{3.9}$$

Therefore, by combining (3.8) and (3.9), we have

$$C(z)\sum_{j=0}^{l-1}\beta_j\sigma_j + (1-\bar{\alpha}z)^l\sum_{i=0}^{n-1}y_ig_i = z^n(z-\alpha)^lg(z). \tag{3.10}$$

Let us define $\Pi_l(\alpha)C(z) = [a_0(\alpha), \cdots, a_{l-1}(\alpha)]$ and introduce the matrices $A(\alpha, l\times l, C(z))$ and $M(\alpha, s\times l, f_0, \cdots, f_{s-1})$, where $f_0, \cdots, f_{s-1} \in H^2$,

$$A(\alpha, l\times l, C(z)) = \begin{bmatrix} a_0(\alpha) & & 0 \\ a_1(\alpha) & \ddots & \\ \vdots & \ddots & \ddots \\ a_{l-1}(\alpha) & a_1(\alpha) & a_0(\alpha) \end{bmatrix},$$

$$M(\sigma, l\times s, f_0, \cdots, f_{s-1}) = \begin{bmatrix} \Pi_l(\alpha)f_0 \\ \Pi_l(\alpha)f_{s-1} \end{bmatrix}'.$$

Let $A=A(0, n\times n, C(\cdot))$, $N=M(0, n\times l, \sigma_0, \cdots, \sigma_{l-1})$, and

$$M=M(0, n\times n, (1-\bar{\alpha}z)^l g_0, \cdots, (1-\bar{\alpha}z)^l g_{n-1}).$$

Applying $\Pi_n(0)$ to (3.10) gives

$$\sum_{j=0}^{l-1} \beta_j A[\Pi_n(0)\sigma_j]' + M[y_0, \cdots, y_{n-1}]' = 0.$$

Observe that $\sum_{j=0}^{l-1} \beta_j[\Pi_n(0)\sigma_j]' = N[\beta_0, \cdots, \beta_{l-1}]'$; hence

$$AN[\beta_0, \cdots, \beta_{l-1}]' + M[y_0, \cdots, y_{n-1}]' = 0,$$

that is,

$$[AN, M][\beta_0, \cdots, \beta_{l-1}, y_0, \cdots, y_{n-1}]' = 0. \tag{3.10a}$$

Let $A_\alpha = A(\alpha, l\times l, C(z))$, $N_\alpha = M(\alpha, l\times l, \sigma_0, \cdots, \sigma_{l-1})$, $M_\alpha = M(\alpha, l\times n, (1-\bar{\alpha}z)^l g_0, \cdots, (1-\bar{\alpha}z)^l g_{n-1})$. Similarly, applying $\Pi_l(\alpha)$ to (3.10) produces

$$[A_\alpha N_\alpha, M_\alpha][\beta_0, \cdots, \beta_{l-1}, y_0, \cdots, y_{n-1}]' = 0. \tag{3.10b}$$

Finally, we get the singular system for the skew Toeplitz operator T_ρ,

$$\begin{bmatrix} AN & M \\ A_\alpha N_\alpha & M_\alpha \end{bmatrix} [\beta_0, \cdots, \beta_{l-1}, y_0, \cdots, y_{n-1}]' = 0. \tag{3.11}$$

Remark 4. For the more general case, namely,

$$C(0)\neq 0, \qquad \{z, C(z)=0\}\cap \sigma(T) = \{\alpha_0, \cdots, \alpha_k\},$$

let $m(z)$, σ_{ij} be as in Lemma 2. Let $B(z)=\prod_{i=0}^{k}(1-\bar{\alpha}_i z)^{l_i}$ and $\gamma_{ij}=B(z)\sigma_{ij}$. Obviously, γ_{ij} are polynomials. Reindex γ_{ij} to be $\gamma_0, \cdots, \gamma_{l-1}$; here $l=\sum_{i=0}^{k} l_i$. Let

$$N = M(0, n\times l, \gamma_0, \cdots, \gamma_{l-1}),$$
$$M = M(0, n\times n, B(z)g_0, \cdots, B(z)g_{n-1}),$$
$$N_i = M(\alpha_i, l_i\times l, \gamma_0, \cdots, \gamma_{l-1}),$$
$$M_i = M(\alpha_i, l_i\times n, B(z)g_0, \cdots, B(z)g_{n-1}),$$
$$A_i = A(\alpha_i, l_i\times l_i, C(z)) \qquad (0\leqq i\leqq k).$$

Then, similar to (3.11), we get the $(n+l)\times(n+l)$ matrix of the following form

$$\begin{bmatrix} AN & M \\ A_0N_0 & M_0 \\ \vdots & \vdots \\ A_iN_i & M_i \\ \vdots & \vdots \\ A_kN_k & M_k \end{bmatrix}.$$

For simplicity we shall present the details only for the case when assumption (A2) is valid. Then we are able to present our second theorem.

THEOREM 5. *Under the assumption* (*A*2), *if* $T_\rho y=0$ *for some nonzero function in* H, *then* y *is given by* (3.7), *where* $[\beta_0, \cdots, \beta_{l-1}, y_0, \cdots, y_{n-1}]$ *is a nonzero solution of* (3.11). *Conversely, if* $[\theta_0, \cdots, \theta_{l-1}, \delta_0, \cdots, \delta_{n-1}]$ *is a nonzero solution of* (3.11), *then the function* y *given by* (3.7) *with* $y_i=\delta_i$ $(0\leqq i<n)$ *and* $\beta_j=\theta_j$ $(0\leqq j<l)$ *is a nonzero function in* H *satisfying* $T_\rho y=0$ *and has its first n-Taylor coefficient given by* $y_i=\delta_i$ $(0\leqq i<n)$ (*hence the* β_j*'s in* (3.3) *equal the given* θ_j $(0\leqq j<l)$).

Proof. The first part easily follows from the above discussion. Conversely, if $[\theta_0, \cdots, \theta_{l-1}, \delta_0, \cdots, \delta_{n-1}]$ satisfying (3.11), then (3.10) and (3.8) with β_j and y_k replaced by θ_j and δ_k, respectively, $(0 \leqq j \leqq l-1, 0 \leqq k \leqq n-1)$ are valid with some $g \in H^2$. Introducing

$$y = \sum_{i=0}^{n-1} \delta_i \chi_i + \frac{m_1(z)}{(1-\bar{\alpha}z)^l} \sum_{j=0}^{l-1} \theta_j \alpha_j \tag{3.12}$$

yields

$$\begin{aligned} C(z)y(z) - \sum_{i=0}^{n-1} \delta_i C_i(z) &= z^n \left(\frac{z-\alpha}{1-\bar{\alpha}z}\right)^l m_1(z) g(z) \\ &= z^n m(z) g(z). \end{aligned} \tag{3.12a}$$

Again, like in the proof of Theorem 1, applying $\Pi_n(0)$ to (3.12) we obtain

$$A[y_0 - \delta_0, \cdots, y_{n-1} - \delta_{n-1}]' = 0,$$

but this time by (A2) $a_0 \neq 0$; this implies that A is invertible, hence $y_i = \delta_i$ $(0 \leqq i < n)$. Therefore, $P_{H_1} y = \sum_{i=0}^{n-1} y_i \chi_i = \sum_{i=0}^{n-1} \delta_i y_i$. This in turn shows y is given by (3.3) with $\beta_i = \theta_j$ $(0 \leqq j < l)$, and thus by (3.4), (3.12) implies $T_\rho y = 0$. By (3.7) and the fact that $H = H_1 \oplus H_0$, it is clear that y is a nonzero function in H if and only if $[\beta_0, \cdots, \beta_{l-1}, y_0, \cdots, y_{n-1}]$ is nonzero. This completes the proof. □

Following [3], [4] we shall call the systems (3.2) and (3.11) the singular systems associated to the skew Toeplitz operator T_ρ. Notice that in the more general case when we assume only $C(0) \neq 0$, the system (3.11) must be replaced with that given in the Remark 4.

4. The computation of singular systems. In the singular system (3.11) the matrices N and N_α are easy to compute. In order to complete the algorithm, we have to find the functions g_i $(0 \leqq i < n)$. To illustrate the idea clearly, we introduce the following properties:

(B1). $C(z)$ has a zero at 0 of order k $(k \leqq n)$, and other zeros of $C(z)$ are simple;

(B2). The zeros of $C(z)$ are simple, different from zero.

Condition (B2) is one of the genericity conditions in [3], [4]; the simplicity of the zeros is only a technical simplification there as well as here (see Remark 8 below). Also we notice that $C^\#(z) := z^{2n} C(1/\bar{z}) = C(z)$ since in (2.3b) $C_{jk} = \bar{C}_{kj}$ by (2.1); we have that under the condition (B2), $C(z)$ has $2n$ zeros $z_1, \cdots, z_p, z_{p+1}, \cdots, z_{2n-p}, 1/\bar{z}_1, \cdots, 1/\bar{z}_p$, $|z_i| < 1$ $(1 \leqq i \leqq p)$, and $|z_i| = 1$ $(p+1 \leqq i \leqq 2n-p)$. Under the condition (B1), notice that $C(z)$ is of degree $2n-k$, in addition to zero; $C(z)$ has $2n-2k$ zeros $z_1, \cdots, z_p, z_{p+1}, \cdots, z_{2n-2k-p}, 1/\bar{z}, \cdots, 1/\bar{z}_p$, $|z_i| < 1$ $(1 \leqq i \leqq p)$, and $|z_i| = 1$ $(p+1 \leqq i \leqq 2n-2k-p)$ (see [3], [4] for details).

PROPOSITION 6 (see [3], [2, § XII.3]). *Under the assumptions* (A2) *and* (B2), *the polynomials g_i (of degree at most $2n-1$) $(0 \leqq i < n)$ in* (3.9) *and in* (3.11) *are uniquely determined by the following $2n$ interpolation conditions*:

$$\begin{aligned} g_i(z_s) &= -\frac{C_i(z_s)}{m_1(z_s)} \qquad \begin{pmatrix} 0 \leqq i < n \\ 1 \leqq s \leqq 2n-p \end{pmatrix}, \\ g_i(1/\bar{z}_s) &= -\overline{m_1(z_s)} C_i\left(\frac{1}{\bar{z}_s}\right) \qquad \begin{pmatrix} 0 \leqq i < n \\ 1 \leqq s \leqq p \end{pmatrix}. \end{aligned} \tag{4.1}$$

For the next proposition, we need $\Pi_k(0) m(z) = [m_0 \cdots m_{k-1}]$.

PROPOSITION 7. *Under the assumptions* (A1) *and* (B1), *the polynomials* g_i (*of degree at most* $2n-1$) $(0\leqq i<n)$ *in* (2.9) *and in* (3.2) *are uniquely determined by the following* $2n$ *interpolation conditions*:

$$g_{ij}=0 \qquad (0\leqq i<n,\quad 0\leqq j<k), \tag{4.1a}$$

$$g_{ij}=-\sum_{s}^{2n-1-j}\bar{m}_s a_{i,j+s} \qquad (0\leqq i<n,\quad 2n-k\leqq j\leqq 2n-1), \tag{4.1b}$$

$$\begin{aligned} g_i(z_s)&=-\frac{C_i(z_s)}{m(z_s)} && (0\leqq i<n,\quad 1\leqq s\leqq 2n-2k-p),\\ g_i\left(\frac{1}{\bar{z}_s}\right)&=-\overline{m(z_s)}C_i\left(\frac{1}{\bar{z}_s}\right) && (0\leqq i<n,\quad 1\leqq s\leqq p). \end{aligned} \tag{4.1c}$$

Proof. (4.1a) is from (3.2c). For $z=e^{it}$, by multiplying (2.9) with $\bar{m}(e^{it})$, we get

$$g_i=C\bar{m}\chi_i-\bar{m}C_i \qquad (0\leqq i<n),$$

where $\bar{m}\chi_i\in K_0^2$ is the orthogonal of H^2 in L^2, since $\chi_i\in H=H(m)$. Also, since $C^{\#}(z)=z^{2n}\overline{C(1/\bar{z})}=C(z)=\sum_{j=0}^{2n}a_jz^j$, we have $a_{2n-i}=\bar{a}_i$ $(0\leqq i<n)$; so by (A1), $a_i=0$ $(0\leqq i<k)$ implies $a_{2n-i}=0$ $(0\leqq i<k)$. Thus $C(z)$ is a polynomial of degree $2n-k$. Since $C_i(z)$ is of degree at most $2n-1$, the coefficients of e^{imt} in the Fourier expansion of g_i are all zero for $m\geqq 2n$, and they come from $-\bar{m}C_i(e^{it})$ for $2n-k\leqq m\leqq 2n-1$. By direct computation we have (4.1b). For (4.1c) see [3] and Proposition 3.4 in [2, § XII.3]. The proof is complete. □

Remark 8. The assumption of simplicity for the zeros of $C(z)$ is not essential. Indeed, for instance if z_1 is a zero of multiplicity 2, then from (2.9) $C\chi_i-C_i=mg_i$ we have also that $C'\chi_i'+C\chi_i'-C_i'=m'g_i+mg_i'$, and thus we obtain the two interpolation conditions for g_i at $z=z_1$, namely,

$$g_i(z_1)=-\frac{C_i(z_1)}{m(z_1)}$$

and

$$\begin{aligned} g_i'(z_i)&=[-C_i'(z_1)-m'(z_1)C_i(z_1)]\frac{1}{m(z_1)}\\ &=-\left[-C_i'(z_1)-C_i(z_1)\frac{m'(z_1)}{m(z_1)}\right]\frac{1}{m(z_1)}, \end{aligned}$$

and similarly other two interpolation conditions for g_i at $1/\bar{z}_1$. The general case is handled in a similar way. For details see [3].

Also, by combining the techniques of Theorems 1 and 5, we can handle the most general case. This is the case when

$$\{z,\ C(z)=0\}\cap\sigma(T)=\{0,\alpha_1,\cdots,\alpha_s\}.$$

Remark 9. We would like to mention here a method for constructing $C^{(-1)}(z)$ and $h_1(z)$ in (2.5). More precisely, let $z_1\cdots z_p$ $(p\leqq 2n)$ denote the roots of C. Then

$$C^{(-1)}(z)=\frac{1-\chi(z)m(z)}{C(z)}, \qquad h_1(z)=-\chi(z),$$

where $\chi(z)$ is a polynomial obeying the following interpolation conditions

$$1-\chi(z_i)m(z_i)=0, \qquad 1\leqq i\leqq p.$$

Thus $P_{H(m)}C^{(-1)}(z)C_i(z)=P_H(1-\chi(z)m(z))C_i(z)/C(z)$ $(0\leqq i<n)$. For details see [1] and the references therein.

5. The algorithm. We can now summarize the above discussion and present the computational algorithm for H^∞ optimization.

First compute $\|W\|_\infty=\max\{|W(z)|, z\in D\}$ and $\|W\|_e=\max\{|W(z)|, z\in\sigma_e(T)=\sigma(T)\cap\partial D\}$. Then for any ρ such that $\|W\|_\infty\geqq\rho>\|W\|_e$, compute the zeros of the polynomial $C(z)$. ($C(z)$ is given by (2.3b) and (2.1), hence depending on ρ.)

Case 1. If (A) holds, then compute the Lagrange polynomials $g_0,\cdots,g_{n-1}$ determined by (3.14c) with $k=0$ and by Remark 8, and define (3.2) with $k=0$ accordingly. Retain ρ if (3.2) with $k=0$ has a nonzero solution, that is, if the $n\times n$ matrix $G_n=[g_{ij}]$ is singular. Denote the largest such ρ by ρ_1.

Case 2. If (A1) holds, then compute the $g_0,\cdots,g_{n-1}$ by (3.14a, b, c) and by Remark 8, compute $C^{(-1)}$ and $P_HC^{(-1)}C_i$ $(0\leqq i<n)$ by Remark 8, and hence get the system (3.2). Retain ρ if (3.2) has a nonzero solution. Denote the largest such ρ by ρ_2.

Case 3. If (A2) holds, then obtain $g_0,\cdots,g_{n-1}$ by (4.1) and Remark 8. Compute (3.11) accordingly. Retain ρ if the $(n+l)\times(n+l)$ matrix

$$\begin{bmatrix} AN & M \\ A_\alpha N_\alpha & M_\alpha \end{bmatrix}$$

is singular. Denote the largest such ρ by ρ_3.

Case 4. In all other cases, use Remarks 8 and 9, and adjust accordingly the computations from the previous cases. The largest ρ obtained is denoted by ρ_4.

If none of ρ_i $(1\leqq i\leqq 4)$ exists, then

$$\rho_0=\inf\{\|w-mq\|_\infty\, q\in H^\infty\}=\|w\|_e.$$

If some ρ_i exists, then $\rho_0=\max_{1\leqq i\leqq 4}\rho_i>\|w\|_e$. In this case, we also have [2], [10] q_{opt} is unique.

$$q_{\text{opt}}=\frac{(W(T)-w(z))x(z)}{m(z)x(z)},$$

where $x(z)=W(T)^*y(z)$ and $y(z)$ is any function in H corresponding to the singular value ρ_0, namely, $y(z)$ is given by (2.7) or (3.7) accordingly.

Next we will apply the algorithm to explicitly solve the weighted sensitivity minimization problem for the weight of the form

$$W(z)=\frac{\alpha z+\beta}{\gamma z+\delta},\qquad \alpha\delta-\beta\gamma\neq 0.$$

For background material and similar results by using different methods, see [3], [4], [5]-[7], [11].

$$\begin{aligned} T\rho&=(\gamma T+\delta)(\bar\gamma T^*+\bar\delta)-\frac{1}{\rho^2}(\alpha T+\beta)(\bar\alpha T^*+\bar\beta)\\ &=A+BT+\bar BT^*+CTT^*,\end{aligned}$$

where

$$A=|\delta|^2-\left(\frac{1}{\rho^2}\right)|\beta|^2,\quad B:=\left(\gamma\bar\delta-\left(\frac{1}{\rho^2}\right)\alpha\bar\beta\right),\quad C:=\left(|\gamma|^2-\left(\frac{1}{\rho^2}\right)|\alpha|^2\right).$$

$C(z)=Bz^2+Fz+\bar B$, where $F=A+C$, $C_0(z)=\bar B+Cz$. The zeros of $C(z)$ are $z_1, z_2=(-F\pm\sqrt{F^2-4|B|^2})/2B$. We always assume $|z_1|\leqq|z_2|$. Then $z_2=1/\bar z_1$ if $z_1\neq 0$.

Case 1. (A) holds. If $z_1 \neq z_2$ (that is, if $|F| \neq 2|B|$), then the system (3.2) with $k=0$, $n=1$ has nonzero solution if and only if

$$0 = g_0(0) = \frac{z_2}{z_1 - z_1}\frac{C_0(z_1)}{m(z_1)} + \frac{z_1}{z_2 - z_1}\overline{m(z_1)}C_0(z_2)$$
$$= -\frac{\bar{B}}{B(z_1 - z_2)m(z_1)(1 - |m(z_1)|^2)}\left(\frac{A-C}{2} + \frac{\sqrt{F^2 - 4|B|^2}}{2}\frac{1 + |m(z_1)|^2}{1 - |m(z_1)|^2}\right).$$

If $|F| = 2|B|$, then $z_1 = z_2 = -\text{sign}\,(F)(|B|/B)$, where

$$\text{sign}\, F = \begin{cases} 1 & F > 0, \\ -1 & F < 0. \end{cases}$$

Then the system (3.2) with $k=0$, $n=1$ has a nonzero solution if and only if

$$0 = g_0(0) = g_0(z_1) - g_0'(z_1)z_1$$
$$= -\frac{1}{m(z_1)}\left(C_0(z_1) - Cz_1 + C_0(z_1)z_1\frac{m'(z_1)}{m(z_1)}\right)$$
$$= -\frac{|B|^2}{B^2 m(z_1)}\left(B + (C - (\text{sign}\, F)|B|)\frac{m'(x_1)}{m(z_1)}\right)$$

Case 2. (A1) holds, i.e., $B=0$, $m(0) \neq 0$; hence $C(z) = Fz$, $C_0(z) = Cz = (C/F)C(z)$. So $\chi_0(z) = P_H C^{(-1)}(z)C_0(z) = (C/F)P_H C^{(-1)}(z)C(z) = (C/F)P_H 1 = (C/F)(1 - m(z)\overline{m(0)})$. The system (3.2) reduces to

$$y_0\left(\frac{C}{F}(1 - |\overline{m(0)}|^2) - 1\right) = 0,$$

or, equivalently, (for $y_0 \neq 0$)

$$A + C|m(0)|^2 = 0.$$

If $m(0) \neq 0$; $B=0$, $F=0$, i.e., $C(z) \equiv 0$ (for instance, for $w(z) = (\alpha z + \beta)/(z + \bar{\alpha}/\bar{\beta})$). If $\rho^2 = |\beta^2|$, then $C(z) \equiv 0$. For $w = \alpha z$, if $\rho^2 = |\alpha|^2$, then $C(z) \equiv 0$). Then by $F = A + C = 0$, $T_\rho = A(I - TT^*)$, and $T_\rho y = 0$ for $y \neq 0$ if and only if $(I - TT^*)y = 0$ (since $A = B = 0$ implies $\alpha\delta = \beta\gamma$). We know by [9] that $(I - TT^*) = u_* \otimes u_*$ is a rank one operator, where $u_* = (1 - m(z)\overline{m(0)})$. Therefore, ker T_ρ is the orthogonal of one-dimension space $\{cu_*, c \in \mathbb{C}\}$ in $H(m)$, which is always nonempty when dim $H(m)$ is greater than one.

Case 3. (A2) holds, i.e., $C(z) = Bz^2 + Fz + \bar{B} = B(z - z_1)(z - z_2)$, $0 \leqq |z_1| < 1$, $m(z) = ((z - z_1)/(1 - \bar{z}_1 z))^l m_1(z)$, with $m_1(z_i) \neq 0$ $(i = 1, 2)$. Then we have

$$\sigma_i = (z - z_i)^{l-i-1} \quad 0 \leqq i < l, \qquad A(0, 1 \times 1, C(z)) = \bar{B},$$
$$N = M(0, 1 \times h, \alpha_0, \cdots, \alpha_{i-1}) = [(-z_1)^{l-1}, \cdots, 1],$$

$$A_{z_1} = A(z_1, l \times l, C(z)) = \begin{bmatrix} 0 & & & & 0 \\ 2Bz_1 + Fz_1 & \ddots & & & \\ \bar{B} & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \cdots & B & 2Bz_1 + Fz_1 & 0 \end{bmatrix},$$

$$N_{z_1} = M(z_1, l \times l, \alpha_0, \cdots, \alpha_{l-1}) = \begin{bmatrix} & 0 & \cdots & 1 \\ & \vdots & 1 & \\ & \vdots & \cdot^{\cdot^{\cdot}} & \\ & 1 & 0 & \\ 1 & & & \end{bmatrix},$$

$$M=M(0,1\times 1,(1-\bar z_1 z)^l g_0(z))=g_0(0),$$

$$M_{z_1}=M(z_1,l\times 1,(1-\bar z_1 z)^l g_0(z))=\begin{bmatrix}(1-|z_1|^2)^l g_0(z_1)\\ * \\ \vdots \\ * \end{bmatrix},$$

where $g_0(z)$ is a polynomial of degree 1 given by (3.13); in particular, $g_0(z_1)=-C_0(z_1)/m_1(z_1)$. Therefore, the $(1\times l)\times(1\times l)$ matrix in system (3.11) is

$$\begin{bmatrix} B(-z_1)^{l-1} & ,\cdots, & & \bar B & g_0(0) \\ 0 & ,\cdots, & & 0 & (1-|z_1|^2)^l g_0(z_1) \\ \vdots & & & 2Bz_1+Fz_1 & * \\ \vdots & & \cdot^{\cdot^{\cdot}} & B & * \\ \vdots & & \cdot^{\cdot^{\cdot}} \quad \cdot^{\cdot^{\cdot}} & & \vdots \\ 0 & 2Bz_1+Fz_1 & B & 0 & * \end{bmatrix},$$

whose determinant equals to $\pm(1-|z_1|^2)^l z_1^{l-1}B(2Bz_1+Fz_1)^{l-1}g_0(z_1)$.

Notice by $0<|z_1|<1$, $2Bz_1+Fz_1\neq 0$ (since $2B+F=0$ implies $z_1=z_2$ but $|z_2|=|1/z_1|>1$); hence the system (3.11) has a nonzero solution if and only if $g_0(z_1)=0$, i.e., $C_0(z_1)=0$. So $z_1=-\bar B/C$, but

$$z_1=\frac{-F+\sqrt{F^2-4|B|^2}}{2B}\quad\text{when } F>0,\qquad z_1=\frac{-F-\sqrt{F^2-4|B|^2}}{2B}\quad\text{when } F<0.$$

In both cases we have $AC=|B|^2$, which is true if and only if

$$|\alpha\delta|^2=|\beta\gamma|^2=\bar\alpha\bar\delta\beta\gamma+\overline{\bar\alpha\bar\delta\beta\gamma}.$$

By simple algebraic manipulation, we obtain $\alpha\delta=\beta\gamma$. This is a contradiction.

Case 4. $B=0$, $F\neq 0$, $m(z)=z^l m_1(z)$ with $l\geqq 1$ and $m_1(0)\neq 0$. Then $C(z)=Fz$, $C_0(z)=Cz$ as in Case 2. For this case we have not written down an explicit singular system. As we mentioned in Remark 9, we have to invoke the techniques of Cases 2 and Case 3. The following shows how to do this in principle for the general weight.

Let y be such that $T_\rho y=0$. Observe that (3.7), (3.9), and (3.10) are still valid, namely, we have

$$(*)\qquad y=y_0\chi_0(z)+m_1(z)\sum_{j=0}^{l-1}\beta_j z^{l-j-1},$$

$$(**)\qquad Fz\sum_{j=0}^{l-1}\beta_j z^{l-j-1}+y_0 g_0(z)=z^{l+1}g(z)$$

for some $g(z)\in H^2$, where, by (3.9) and (3.14a, b), $g_0(z)$ is a polynomial of degree 1, $g_0(0)=0$, $g_0'(0)=-\overline{m_1(0)}\cdot C$. Therefore, by $(*)$ and $(**)$, $\beta_j=0$ $(0\leqq j\leqq l-2)$ and

$$F\beta_{l-1}+y_0 g_0'(0)=0,\qquad m_1(0)\beta_{l-1}+y_0(\chi_0(0)-1)=0,$$

where by (3.5) $C^{(-1)}(z)C(z)=1+m_1(z)h_1(z)$. So

$$\chi_0(z)=P_{H(m_1)}C^{(-1)}(z)C_0(z)=\frac{C}{F}P_{H(m_1)}1=\frac{C}{F}(1-m(z)\overline{m(0)}).$$

The above linear equation for (β_{l-1}, y_0) has a nonzero solution if and only if

$$\begin{vmatrix} F & g_0'(0) \\ m_1(0) & \chi_0(0)-1 \end{vmatrix} = A = 0,$$

which plus $B=0$ leads to $\alpha\delta = \gamma\beta$. This is again a contradiction.

In conclusion *for the weight of the form* $w(z)=(\partial z+\beta)/(\gamma z+\delta)$ *in Cases* 3 *and* 4, ker $T_\rho = \{0\}$.

Remark 10. In Case 4 there is a simpler way to show ker $T_\rho = 0$. Recall from [9] that $I - TT^* = 1 \otimes 1$, where 1 is the function with constant value 1; thus $T_\rho = A + CTT^* = A + C(I - 1\otimes 1)$, and for $y \neq 0$, $T_\rho y = 0$ if and only if

$$Fy(z) = C(1\otimes 1)\cdot y(z) = C(y(z), 1)\cdot 1.$$

So by $F \neq 0$, $(y(z), 1) \neq 0$, and $F(y(z), 1) = (Fy(z), 1) = C(y(z), 1)(1, 1)$, i.e., $F = C$, that is again $A = 0$, which is impossible as in Case 4.

REFERENCES

[1] H. Bercovici, C. Foias, and A. Tannenbaum, *On Skew Toeplitz operators*, Oper. Theory: Adv. Appl., 49 (1988), pp. 21-43.

[2] C. Foias and A. E. Frazho, *The commutant lifting approach to interpolation problems*, Oper. Theory: Adv. Appl., 44, Birkhäuser-Verlag, Basel, 1990.

[3] C. Foias, A. Tannenbaum, and G. Zames, *Some explicit formulae for the singular values of certain Hankel operators with factorizable symbol*, SIAM J. Math. Anal., 19 (1988), pp. 1081-1089.

[4] C. Foias and A. Tannenbaum, *Some remarks on optimal interpolation*, Systems Control Lett., 11 (1988), pp. 259-264.

[5] C. Foias, A. Tannenbaum, and G. Zames, *On the H^∞-optimization sensitivity problem for systems with delays*, SIAM J. Control Optim., 25 (1987), pp. 686-705.

[6] ———, *Sensitivity minimization for arbitrary* SISO *distributed plants*, Systems Control Lett., 8 (1987), pp. 189-195.

[7] C. Foias and A. Tannenbaum, *On the Nehari problem for a certain class of L^∞-functions appearing in control theory*, I, J. Funct. Anal., 74 (1987), pp. 146-159; II, J. Funct. Anal., 81 (1988), pp. 207-218.

[8] N. K. Nikolskii, *Treatise on the Shift Operator*, Springer-Verlag, New York, 1986.

[9] B. Sz.-Nagy and C. Foias, *Harmonic analysis of operators on Hilbert space*, North-Holland, Amsterdam, 1970.

[10] D. Sarason, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179-203.

[11] G. Zames, *Feedback and optimal sensitivity*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 302-320.

SIAM J. MATH. ANAL.
Vol. 23, No. 6, pp. 1637-1649, November 1992

ON THE ASYMPTOTICS OF THE JACOBI FUNCTION AND ITS ZEROS*

R. WONG† AND Q.-Q. WANG†

Abstract. Explicit and realistic error bounds are constructed for a one-term and a two-term asymptotic approximation of the Jacobi function $\varphi_\mu^{(\alpha,\beta)}(t)$ as $\mu \to \infty$, uniformly for $t \in (0, \infty)$. A similar result is obtained for the zeros $t_{\mu,k}$ of this function as $\mu \to \infty$, which holds uniformly with respect to unbounded k. Exponentially decaying error bounds are also given for asymptotic approximations of $\varphi_\mu^{(\alpha,\beta)}(t)$ as $t \to \infty$ and of $t_{\mu,k}$ as $k \to \infty$, which are uniform for $\mu \geqq \delta > 0$.

Key words. Jacobi function, uniform asymptotic approximations, error bounds, zeros

AMS(MOS) subject classifications. 33A30, 41A60

1. Introduction. Let α, β, and μ be real numbers with $\mu > 0$ and $\alpha \neq -1, -2, \cdots$. The Jacobi function is defined by

$$(1.1) \qquad \varphi_\mu^{(\alpha,\beta)}(t) = {}_2F_1[\tfrac{1}{2}(\alpha+\beta+1-i\mu), \tfrac{1}{2}(\alpha+\beta+1+i\mu);\ \alpha+1;\ -\sinh^2 t]$$

for $t > 0$, where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function. This function plays an important role in the interactions of special functions and group theory, and for an excellent survey of this topic, we refer to Koornwinder [7].

The Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$ can also be expressed in terms of the hypergeometric function [12, p. 63]

$$(1.2) \qquad \frac{\Gamma(\alpha+1)\Gamma(n+1)}{\Gamma(\alpha+n+1)} P_n^{(\alpha,\beta)}(x) = {}_2F_1\left[-n, n+\alpha+\beta+1;\ \alpha+1; \frac{1}{2}(1-x)\right],$$

and this formula furnishes the extension of the polynomial $P_n^{(\alpha,\beta)}(x)$ to arbitrary values of the degree n. From (1.1) and (1.2), it is evident that

$$(1.3) \qquad \varphi_\mu^{(\alpha,\beta)}(t) = \frac{\Gamma(\alpha+1)\Gamma(\frac{1}{2}(i\mu-\alpha-\beta+1))}{\Gamma(\frac{1}{2}(i\mu+\alpha-\beta+1))} P^{(\alpha,\beta)}_{\frac{1}{2}(i\mu-\alpha-\beta-1)}(\cosh 2t),$$

and for this reason, $\varphi_\mu^{(\alpha,\beta)}(t)$ is called the Jacobi function. This function is also related to the Legendre function [8, (12.05), p. 170]

$$(1.4) \qquad \begin{aligned} P_n^{-m}(z) &= 2^{-m}(z-1)^{m/2}(z+1)^{m/2} \\ &\quad \cdot {}_2F_1(m-n, m+n+1;\ m+1;\ \tfrac{1}{2}-\tfrac{1}{2}z)/\Gamma(\alpha+1), \end{aligned}$$

and the relationship is

$$(1.5) \qquad (\sinh t)^\alpha(\cosh t)^\alpha \varphi_{2\mu}^{(\alpha,\alpha)}(t) = \Gamma(\alpha+1)P^{-\alpha}_{-\frac{1}{2}+i\mu}(\cosh 2t).$$

Recently, asymptotic approximations, complete with error bounds, have been obtained for the Jacobi polynomial $P_n^{(\alpha,\beta)}(\cos\theta)$ as $n \to \infty$, which are uniformly valid for $0 \leqq \theta \leqq \pi/2$; see [4] and [2]. Corresponding results for the Legendre function $P_n^{-m}(\cosh z)$ as $n \to +\infty$, which are uniformly valid for $0 < z < \infty$, can also be found in [8, p. 466] and [10]. The purpose of this paper is to present similar results for the Jacobi function $\varphi_\mu^{(\alpha,\beta)}(t)$ as $\mu \to +\infty$, which is uniform with respect to t in $(0, \infty)$, and

* Received by the editors July 16, 1991; accepted for publication (in revised form) February 24, 1992. This research was partially supported by Natural Sciences and Engineering Research Council of Canada grant A7359.

† Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2.

to use these results to obtain asymptotic approximations for the zeros $t_{\mu,k}$ of this function as $\mu\to\infty$, uniformly with respect to k. Asymptotic expansions of $\varphi_\mu^{(\alpha,\beta)}(t)$ have been given previously by Triméche [13] and Fitouhi and Hamza [3] using differential equation theory. Based on an integral representation of the Jacobi function, Schindler [9] (in the case when $\alpha=\beta$) and Stanton and Tomas [11] have derived asymptotic expansions of $\varphi_\mu^{(\alpha,\beta)}(t)$ as $t\to 0$, which is, in a sense, uniform with respect to μt. However, no error bounds were constructed for these expansions, and no study of the asymptotic behavior of the zeros $t_{\mu,k}$ was made in these investigations. In the present paper, we shall also use the differential equation approach, and our work relies heavily on the results of Olver [8, Chap. 12]. For completeness, we also include an asymptotic approximation of $\varphi_\mu^{(\alpha,\beta)}(t)$ as $t\to\infty$, which is uniform in $\mu\geqq\delta>0$, and a corresponding result for $t_{\mu,k}$ as $k\to\infty$. The problems studied in this paper have been suggested by R. A. Askey.

2. Differential equations. It is known that the Jacobi function $\varphi_\mu^{(\alpha,\beta)}(t)$ is the unique even C^∞-function on **R** which satisfies

$$\begin{aligned} &v''(t)+[(2\alpha+1)\coth t+(2\beta+1)\tanh t]v'(t)\\ &+[\mu^2+(\alpha+\beta+1)^2]v(t)=0 \end{aligned} \tag{2.1}$$

and $v(0)=1$; see [7, p. 2]. If we set

$$u(t)=(\sinh t)^{\alpha+\frac{1}{2}}(\cosh t)^{\beta+\frac{1}{2}}\varphi_\mu^{(\alpha,\beta)}(t), \tag{2.2}$$

then it is easily verified that

$$u''(t)+\left\{\mu^2+\frac{\frac{1}{4}-\alpha^2}{\sinh^2 t}-\frac{\frac{1}{4}-\beta^2}{\cosh^2 t}\right\}u(t)=0. \tag{2.3}$$

When $\alpha>-\frac{1}{2}$, we also have $u(0)=0$.

To apply the asymptotic theory of Olver [8, pp. 438-440], we shall restrict ourselves to the case $\alpha\geqq 0$ and introduce the new variables

$$\begin{aligned} &(-\zeta)^{\frac{1}{2}}=t, \qquad \zeta<0,\\ &W(\zeta)=(-\zeta)^{\frac{1}{4}}u(t). \end{aligned} \tag{2.4}$$

The transformed equation is given by

$$\frac{d^2W}{d\zeta^2}=\left\{\frac{\mu^2}{4\zeta}+\frac{\alpha^2-1}{4\zeta^2}+\frac{\psi(\zeta)}{\zeta}\right\}W(\zeta), \qquad \zeta<0, \tag{2.5}$$

where

$$\psi(\zeta)=\frac{1}{4}\left\{\frac{\frac{1}{4}-\alpha^2}{\zeta}+\left[\frac{\frac{1}{4}-\alpha^2}{\sinh^2(-\zeta)^{\frac{1}{2}}}-\frac{\frac{1}{4}-\beta^2}{\cosh^2(-\zeta)^{\frac{1}{2}}}\right]\right\}. \tag{2.6}$$

Note that $\psi(\zeta)$ is analytic at $\zeta=0$.

3. Asymptotic expansion. For negative ζ, Theorem 4.1 in [8, p. 444] gives two asymptotic solutions to (2.5), one involving the Bessel function $J_\alpha(\mu\sqrt{-\zeta})$ and the other involving $Y_\alpha(\mu\sqrt{-\zeta})$. To identify the function $(-\zeta)^{\frac{1}{4}}u(t)$ in (2.4) with one of these two solutions or a linear combination of them, we note that from (2.2) and (2.4) we have

$$(-\zeta)^{\frac{1}{4}}u(t)\sim(-\zeta)^{(\alpha+1)/2} \tag{3.1}$$

as $\zeta \to 0^-$. Since $J_\alpha(x) \sim (x/2)^\alpha/\Gamma(\alpha+1)$ for x near zero, it can be shown, as in [8, pp. 464 and 466], that

$$\begin{aligned}(3.2)\qquad &(\sinh t)^{\alpha+\frac{1}{2}}(\cosh t)^{\beta+\frac{1}{2}}\varphi_\mu^{(\alpha,\beta)}(t)\\ &= \frac{2^\alpha\Gamma(\alpha+1)}{\mu^\alpha}t^{-\frac{1}{2}}\left\{tJ_\alpha(\mu t)\sum_{s=0}^{n}\frac{A_s(-t^2)}{\mu^{2s}}\right.\\ &\qquad \left.-\frac{t^2}{\mu}J_{\alpha+1}(\mu t)\sum_{s=0}^{n-1}\frac{B_s(-t^2)}{\mu^{2s}}+\varepsilon_{2n+1,1}(\mu,-t^2)\right\}\end{aligned}$$

for $t>0$, where the coefficients $A_s(\zeta)$ and $B_s(\zeta)$ are analytic functions in a region containing $\zeta=0$, and are determined by the recursive formulas

$$(3.3)\qquad B_s(\zeta) = -A_s'(\zeta)+\frac{1}{(-\zeta)^{\frac{1}{2}}}\int_\zeta^0\left\{\psi(v)A_s(v)-\left(\alpha+\frac{1}{2}\right)A_s'(v)\right\}\frac{dv}{(-v)^{\frac{1}{2}}},$$

$$(3.4)\qquad A_{s+1}(\zeta) = \alpha B_s(\zeta)-\zeta B_s'(\zeta)+\int\psi(\zeta)B_s(\zeta)\,d\zeta,$$

with $A_0(\zeta)=1$. The remainder in (3.2) satisfies the estimate

$$\begin{aligned}(3.5)\qquad |\varepsilon_{2n+1,1}(\mu,\zeta)| &\leqq \lambda_3(\alpha)(-\zeta)^{\frac{1}{4}}E_\alpha^{-1}(\mu\sqrt{-\zeta})M_\alpha(\mu\sqrt{-\zeta})\\ &\quad\cdot\exp\left\{\frac{\lambda_2(\alpha)}{\mu}\mathcal{V}_{\zeta,0}(\sqrt{-\zeta}\,B_0)\right\}\frac{\mathcal{V}_{\zeta,0}(\sqrt{-\zeta}\,B_n)}{\mu^{2n+1}},\end{aligned}$$

where the modulus function $M_\alpha(x)$ and the weight function $E_\alpha(x)$ are defined in [8, Chap. 12, § 1.3], $E_\alpha^{-1}(x)=1/E_\alpha(x)$ and $\mathcal{V}_{a,b}(f)$ denotes the total variation of a function $f(x)$ on an interval (a, b). The constants $\lambda_2(\alpha)$ and $\lambda_3(\alpha)$ in (3.5) are given in [8, p. 443]. It is not easy to estimate or compute these constants; some properties and values of these constants can be found in the above mentioned reference.

Note that for each n, expansion (3.2) gives $(2n+1)$ terms, i.e., we always have an odd number of terms. By using the same argument as in [8, Chap. 12], it is possible to derive an expansion that yields an even number of terms, and the result is

$$\begin{aligned}(3.6)\qquad &(\sinh t)^{\alpha+\frac{1}{2}}(\cosh t)^{\beta+\frac{1}{2}}\varphi_\mu^{(\alpha,\beta)}(t)\\ &= \frac{2^\alpha\Gamma(\alpha+1)}{\mu^\alpha}t^{-\frac{1}{2}}\left\{tJ_\alpha(\mu t)\sum_{s=0}^{n-1}\frac{A_s(-t^2)}{\mu^{2s}}\right.\\ &\qquad \left.-\frac{t^2}{\mu}J_{\alpha+1}(\mu t)\sum_{s=0}^{n-1}\frac{B_s(-t^2)}{\mu^{2s}}+\varepsilon_{2n,1}(\mu,-t^2)\right\},\end{aligned}$$

where

$$\begin{aligned}(3.7)\qquad |\varepsilon_{2n,1}(\mu,-t^2)| &\leqq \bar\lambda_3(\alpha)tE_\alpha^{-1}(\mu t)M_\alpha(\mu t)\\ &\quad\cdot\exp\left\{\frac{\lambda_2(\alpha)}{\mu}\mathcal{V}_{0,t}(tB_0)\right\}\frac{\mathcal{V}_{0,t}(A_n)}{\mu^{2n}}\end{aligned}$$

and

$$(3.8)\qquad \bar\lambda_3(\alpha) = \sup_{x\in(0,\infty)}\{\pi xE_\alpha(x)M_\alpha(x)|J_{\alpha+1}(x)|\}.$$

The leading coefficients can be calculated explicitly. For convenience, we put

$$(3.9)\qquad A=\tfrac{1}{4}-\alpha^2,\qquad B=\tfrac{1}{4}-\beta^2.$$

Since $A_0(\zeta)=1$, it follows from (3.3) and (2.6) that

$$B_0(-t^2)=\frac{1}{2t}\int_0^t\{A(-s^{-2}+\operatorname{csch}^2 s)-B\operatorname{sech}^2 s\}\,ds;$$

accordingly,

$$B_0(-t^2)=\frac{1}{2t}\{A(t^{-1}-\coth t)-B\tanh t\}. \tag{3.10}$$

From (3.4), we also have

$$\begin{aligned}A_1(-t^2)=&\tfrac{1}{2}(\alpha+\tfrac{1}{2})t^{-1}\{A(t^{-1}-\coth t)-B\tanh t\}\\&+\tfrac{1}{4}\{A(t^{-2}-\operatorname{csch}^2 t)+B\operatorname{sech}^2 t\}\\&-\tfrac{1}{4}\{A^2(\tfrac{1}{2}t^{-2}-t^{-1}\coth t+\tfrac{1}{2}\operatorname{csch}^2 t)\\&\quad-ABt^{-1}\tanh t-\tfrac{1}{2}B^2\operatorname{sech}^2 t\}+C,\end{aligned} \tag{3.11}$$

where

$$C=\frac{\alpha}{2}\left(\frac{1}{3}A+B\right)-\frac{1}{8}(A^2+2AB+B^2) \tag{3.12}$$

is chosen so that $A_1(0)=0$, a condition needed in deriving (3.2).

4. Estimates for $\mathcal{V}_{0,t}(tB_0)$ and $\mathcal{V}_{0,t}(A_1)$. In order for the error bounds in (3.5) and (3.7) to be of any practical use, the total variations $\mathcal{V}_{0,t}(tB_n)$ and $\mathcal{V}_{0,t}(A_n)$ should be estimated. Here we illustrate only the cases involving B_0 and A_1. First we recall the property

$$\mathcal{V}_{a,b}(f)=\int_a^b|f'(t)|\,dt, \tag{4.1}$$

and observe the relation

$$\mathcal{V}_{\zeta,0}\{\sqrt{-\zeta}\,B_n(\zeta)\}=\mathcal{V}_{0,t}\{tB_n(-t^2)\}.$$

From (3.10), we have

$$\{tB_0(-t^2)\}'=\tfrac{1}{2}\{A(-t^{-2}+\operatorname{csch}^2 t)-B\operatorname{sech}^2 t\}.$$

Since $-t^{-2}+\operatorname{csch}^2 t$ is always negative, it follows from (4.1) that

$$\mathcal{V}_{0,t}(tB_0)\leqq\tfrac{1}{2}\{|A|(\coth t-t^{-1})+|B|\tanh t\}. \tag{4.2}$$

Consequently,

$$\mathcal{V}_{0,t}(tB_0)\leqq\mathcal{V}_{0,\infty}(tB_0)=\tfrac{1}{2}(|A|+|B|). \tag{4.3}$$

On the other hand, since both $t^{-1}\coth t-t^{-2}$ and $t^{-1}\tanh t$ are decreasing in the interval $0<t<\infty$, we have $0\leqq t^{-1}\coth t-t^{-2}\leqq\frac{1}{3}$ and $0\leqq t^{-1}\tanh t\leqq 1$. Hence (4.2) also gives

$$\mathcal{V}_{0,t}(tB_0)\leqq\frac{t}{6}(|A|+3|B|). \tag{4.4}$$

For large values of t, the bound in (4.3) is preferable since it is independent of t. For small values of t, (4.4) is a better estimate since the bound on the right-hand side decreases to zero with t.

Similarly, (3.11) (or (3.4) and (3.10)) gives

$$\{A_1(-t^2)\}' = \tfrac{1}{2}(\tfrac{1}{2}+\alpha)AT_1(t) + \tfrac{1}{4}\{2(\tfrac{1}{2}+\alpha)B - AB\}T_2(t) + \tfrac{1}{4}AT_3(t) + \tfrac{1}{4}A^2T_4(t) + \tfrac{1}{4}(2B+B^2)T_5(t),$$

where

$$T_1(t) = -2t^{-3} + t^{-2}\coth t + t^{-1}\operatorname{csch}^2 t,$$

$$T_2(t) = t^{-2}\tanh t - t^{-1}\operatorname{sech}^2 t,$$

$$T_3(t) = -2t^{-3} + 2\coth t\operatorname{csch}^2 t,$$

$$T_4(t) = t^{-3} - t^{-2}\coth t - t^{-1}\operatorname{csch}^2 t + \coth t\operatorname{csch}^2 \mathrm{t},$$

$$T_5(t) = -\tanh t\operatorname{sech}^2 t.$$

We now show that for each $i = 1, \cdots, 5$, the sign of $T_i(t)$ remains the same for $t \in (0, \infty)$. Clearly, $T_1(t)$ can be written as

$$T_1(t) = \frac{(-2\sinh^2 t + t\sinh t\cosh t + t^2)4}{4t^3\sinh^2 t},$$

and the numerator here is equal to

$$-2(e^{2t} - 2 + e^{-2t}) + t(e^{2t} - e^{-2t}) + (2t)^2.$$

By expanding e^{2t} and e^{-2t} into Maclaurin series and regrouping terms with equal exponents of t, it can be easily verified that this numerator has the power series expansion

$$\sum_{n=2}^{\infty} \frac{-4+2n}{(2n)!}(2t)^{2n},$$

from which it follows immediately that

$$T_1(t) > 0, \qquad t > 0.$$

In a similar manner, it can be proven that for $t > 0$,

$$T_2(t) > 0, \quad T_3(t) < 0, \quad T_4(t) < 0, \quad T_5(t) < 0.$$

Hence

$$\int_0^t |T_1(t)|\,dt = t^{-2} - t^{-1}\coth t + \tfrac{1}{3} \leqq \tfrac{1}{3},$$

$$\int_0^t |T_2(t)|\,dt = -t^{-1}\tanh t + 1 \leqq 1,$$

$$\int_0^t |T_3(t)|\,dt = -t^{-2} + \operatorname{csch}^2 t + \tfrac{1}{3} \leqq \tfrac{1}{3}, \tag{4.5}$$

$$\int_0^t |T_4(t)|\,dt = \tfrac{1}{2}t^{-2} - t^{-1}\coth t + \tfrac{1}{2}\operatorname{csch}^2 t + \tfrac{1}{2} \leqq \tfrac{1}{2},$$

$$\int_0^t |T_5(t)|\,dt = -\tfrac{1}{2}\operatorname{sech}^2 t + \tfrac{1}{2} \leqq \tfrac{1}{2}.$$

From (4.1), it follows that

$$\begin{aligned}\mathscr{V}_{0,t}(A_1) \leqq {} & \tfrac{1}{2}\left|\tfrac{1}{2}+\alpha\right||A|\{t^{-2}-t^{-1}\coth t+\tfrac{1}{3}\} \\ & +\tfrac{1}{4}|2(\tfrac{1}{2}+\alpha)B-AB|\{-t^{-1}\tanh t+1\} \\ & +\tfrac{1}{4}|A|\{-t^{-2}+\operatorname{csch}^2 t+\tfrac{1}{3}\} \\ & +\tfrac{1}{4}A^2\{\tfrac{1}{2}t^{-2}-t^{-1}\coth t+\tfrac{1}{2}\operatorname{csch}^2 t+\tfrac{1}{2}\} \\ & +\tfrac{1}{4}|2B+B^2|\{-\tfrac{1}{2}\operatorname{sech}^2 t+\tfrac{1}{2}\},\end{aligned} \tag{4.6}$$

and

$$\begin{aligned}\mathscr{V}_{0,t}(A_1) & \leqq \mathscr{V}_{0,\infty}(A_1) \\ & =\tfrac{1}{6}(\tfrac{1}{2}+\alpha)|A|+\tfrac{1}{12}|A|+\tfrac{1}{8}A^2 \\ & \quad +\tfrac{1}{4}|2(\tfrac{1}{2}+\alpha)B-AB|+\tfrac{1}{8}|2B+B^2|.\end{aligned} \tag{4.7}$$

It can also be shown that

$$\mathscr{V}_{0,t}(A_1) \leqq t^2 V(\alpha, \beta), \tag{4.8}$$

where

$$\begin{aligned}V(\alpha,\beta)= {} & \tfrac{1}{90}(\tfrac{1}{2}+\alpha)|A|+\tfrac{1}{12}|2(\tfrac{1}{2}+\alpha)B-AB| \\ & +\tfrac{1}{60}|A|+\tfrac{1}{72}A^2+\tfrac{1}{8}|2B+B^2|.\end{aligned} \tag{4.9}$$

To do this, we return to (4.5) and put

$$X_i(t)=\frac{1}{t^2}\int_0^t |T_i(t)|\,dt, \qquad i=1,\cdots,5.$$

By expressing the hyperbolic functions in terms of the exponential function, it can easily be shown that

$$\begin{aligned}X_1'(t)=\frac{t^{-5}}{(e^t-e^{-t})^2}\biggl\{ & -4(e^{2t}-2+e^{-2t})+3t(e^{2t}-e^{-2t}) \\ & +4t^2-\frac{2}{3}t^2(e^{2t}-2+e^{-2t})\biggr\}.\end{aligned}$$

We now apply the argument in § 4 used to show the signs of $T_i(t)$. This leads to a Maclaurin series expansion with all negative coefficients for the quantity inside the above curly bracket. Thus, $X_1(t)$ is a decreasing function in $(0,\infty)$. In a similar manner, it can be verified that for each $i=1,\cdots,5$, $X_i(t)$ is a nonnegative decreasing function in $(0,\infty)$. Furthermore, as $t\to 0^+$, we have

$$X_1(t)\to\tfrac{1}{45}, \qquad X_2(t)\to\tfrac{1}{3}, \qquad X_3(t)\to\tfrac{1}{15},$$

$$X_4(t)=X_1(t)+\tfrac{1}{2}X_3(t)\to\tfrac{1}{18}, \qquad X_5(t)\to\tfrac{1}{2}.$$

From (4.6), it follows that (4.8) is proved.

5. Zeros. The uniform expansions (3.2) and (3.6) can be used to determine the zeros of the Jacobi function $\varphi_\mu^{(\alpha,\beta)}(t)$, and the main tool in this regard is the following result stated in Hethcote [6].

THEOREM A. *In the interval* $[a-\rho, a+\rho]$, *suppose* $f(t)=g(t)+\varepsilon(t)$, *where* $f(t)$ *is continuous*, $g(t)$ *is differentiable*, $g(a)=0$, $m=\min|g'(t)|>0$, *and*

$$E=\max|\varepsilon(t)|<\min\{|g(a-\rho)|, |g(a+\rho)|\}. \tag{5.1}$$

Then there exists a zero c *of* $f(t)$ *in the interval such that* $|c-a|\leqq E/m$.

In (3.2) we now take $n=0$. This gives

$$\begin{aligned}&(\sinh t)^{\alpha+\frac{1}{2}}(\cosh t)^{\beta+\frac{1}{2}}\varphi_\mu^{(\alpha,\beta)}(t)\\ &\quad=\frac{2^\alpha\Gamma(\alpha+1)}{\mu^\alpha}\, t^{\frac{1}{2}}\{J_\alpha(\mu t)+t^{-1}\varepsilon_{1,1}(\mu, -t^2)\}.\end{aligned} \tag{5.2}$$

In terms of the phase function $\theta_\alpha(x)$ defined by

$$\begin{aligned}\theta_\alpha(x)&=-\tfrac{1}{4}\pi, \qquad 0<x\leqq X_\alpha,\\ \theta_\alpha(x)&=\tan^{-1}\left\{\frac{Y_\alpha(x)}{J_\alpha(x)}\right\}, \qquad x\geqq X_\alpha,\end{aligned} \tag{5.3}$$

where X_α denotes the smallest positive root of the equation $J_\alpha(x)+Y_\alpha(x)=0$ and the branch of the inverse tangent is chosen to make $\theta_\alpha(x)$ continuous, the Bessel function can be expressed as

$$J_\alpha(x)=E_\alpha^{-1}(x)M_\alpha(x)\cos\theta_\alpha(x). \tag{5.4}$$

It is known that

$$\theta_\alpha'(x)=\frac{2}{\pi x M_\alpha^2(x)}, \qquad x>X_\alpha, \tag{5.5}$$

and

$$\theta_\alpha(j_{\alpha,s})=(s-\tfrac{1}{2})\pi, \qquad \theta_\alpha(y_{\alpha,s})=(s-1)\pi, \tag{5.6}$$

where $j_{\alpha,s}$ and $y_{\alpha,s}$ are the sth positive zero of $J_\alpha(x)$ and $Y_\alpha(x)$; see [8, p. 437]. To apply the above theorem to (5.2), we take

$$\begin{aligned}f(t)=&\frac{\mu^\alpha}{2^\alpha\Gamma(\alpha+1)}\, t^{-\frac{1}{2}}(\sinh t)^{\alpha+\frac{1}{2}}(\cosh t)^{\beta+\frac{1}{2}}\\ &\cdot E_\alpha(\mu t)M_\alpha^{-1}(\mu t)\varphi_\mu^{(\alpha,\beta)}(t),\end{aligned} \tag{5.7}$$

$$g(t)=\cos\theta_\alpha(\mu t), \tag{5.8}$$

$$\varepsilon(t)=t^{-1}E_\alpha(\mu t)M_\alpha^{-1}(\mu t)\varepsilon_{1,1}(\mu, -t^2). \tag{5.9}$$

From (3.5) with $(-\zeta)^{\frac{1}{2}}=t$ (see (2.4)), it follows that

$$|\varepsilon(t)|\leqq\lambda_3(\alpha)\exp\left\{\frac{\lambda_2(\alpha)}{\mu}\mathcal{V}_{0,t}[tB_0(-t^2)]\right\}\frac{\mathcal{V}_{0,t}[tB_0(-t^2)]}{\mu}, \tag{5.10}$$

which, coupled with (4.3), yields

$$|\varepsilon(t)|\leqq\frac{\lambda_3(\alpha)(|A|+|B|)}{2\mu}\exp\left\{\frac{\lambda_2(\alpha)(|A|+|B|)}{2\mu}\right\}. \tag{5.11}$$

In view of (4.4), the estimate (5.10) also gives

$$|\varepsilon(t)|\leqq\lambda_3(\alpha)\exp\left\{\frac{\lambda_2(\alpha)(|A|+|B|)}{2\mu}\right\}\frac{(|A|+3|B|)}{6}\frac{t}{\mu}. \tag{5.12}$$

Differentiation and use of (5.5) give

$$g'(t) = -\frac{2\mu \sin \theta_\alpha(\mu t)}{\pi \mu t M_\alpha^2(\mu t)}, \qquad \mu t > X_\alpha. \tag{5.13}$$

Since $\lambda_2(\alpha) \geqq \pi x M_\alpha^2(x)$ (see [8, p. 443]), we have

$$|g'(t)| \geqq \frac{2\mu |\sin \theta_\alpha(\mu t)|}{\lambda_2(\alpha)}, \qquad \mu t \geqq X_\alpha. \tag{5.14}$$

According to (5.6), the sth zero of $g(t)$ is $a_s = j_{\alpha,s}/\mu$ and $\sin \theta_\alpha(j_{\alpha,s}) = \pm 1$. We now choose $\rho_s = \mu^{-\frac{3}{2}}$ with μ being sufficiently large so that

$$|\sin \theta_\alpha(\mu t)| \geqq |\sin \theta_\alpha(j_{\alpha,s} \pm \mu^{-\frac{1}{2}})| \geqq \tfrac{1}{2} \tag{5.15}$$

for $a_s - \rho_s \leqq t \leqq a_s + \rho_s$. (Note that by (5.5), $\theta_\alpha(x)$ is an increasing function for $x > X_\alpha$.) Coupling (5.14) and (5.15) gives

$$m = \min |g'(t)| \geqq \frac{\mu}{\lambda_2(\alpha)} > 0, \tag{5.16}$$

the minimum being taken over the interval $[a_s - \rho_s, a_s + \rho_s]$. To verify that condition (5.1) holds, we observe from (5.11) that

$$E = \max |\varepsilon(t)| \leqq \frac{e\lambda_3(\alpha)(|A| + |B|)}{2\mu}, \tag{5.17}$$

if we choose

$$\mu \geqq \tfrac{1}{2}\lambda_2(\alpha)(|A| + |B|). \tag{5.18}$$

From (5.12), we also have

$$E = \max |\varepsilon(t)| \leqq \frac{e\lambda_3(\alpha)(|A| + 3|B|)}{6\mu}\left(\frac{j_{\alpha,s}}{\mu} + \mu^{-\frac{3}{2}}\right),$$

which of course implies

$$E \leqq 0.5473 \lambda_3(\alpha)(|A| + 3|B|)\frac{j_{\alpha,s}}{\mu^2}, \tag{5.19}$$

if (5.18) holds and $\mu \geqq 4$. Here we have made use of the fact that $j_{\alpha,s} \geqq j_{0,1} = 2.40482$ for $\alpha \geqq 0$ and $s \geqq 1$.

On the other hand, since $g(a_s) = 0$ by (5.6), we have

$$g(a_s \pm \rho_s) = \pm g'(\xi_\pm)\rho_s,$$

where $a_s < \xi_+ < a_s + \rho_s$ and $a_s - \rho_s < \xi_- < a_s$. From (5.14) and (5.15) it follows that

$$|g(a_s \pm \rho_s)| \geqq \frac{1}{\lambda_2(\alpha)} \mu^{-\frac{1}{2}}. \tag{5.20}$$

Thus, in view of (5.17), condition (5.1) holds if

$$\mu^{\frac{1}{2}} \geq \frac{e}{2} \lambda_2(\alpha)\lambda_3(\alpha)(|A| + |B|). \tag{5.21}$$

By Theorem A, (5.16), and (5.21), the Jacobi function $\varphi_\mu^{(\alpha,\beta)}(t)$ has a zero $t_{\mu,s}$ satisfying

$$\left| t_{\mu,s} - \frac{j_{\alpha,s}}{\mu} \right| \leqq \frac{e\lambda_2(\alpha)\lambda_3(\alpha)(|A| + |B|)}{2\mu^2}, \tag{5.22}$$

if (5.18) and the second inequality in (5.15) hold. On the other hand, from (5.16) and (5.19), we also have

$$(5.23)\qquad \left|t_{\mu,s}-\frac{j_{\alpha,s}}{\mu}\right| \leqq 0.5473\lambda_2(\alpha)\lambda_3(\alpha)(|A|+3|B|)\frac{j_{\alpha,s}}{\mu^3},$$

if, in addition, (5.21) holds.

Note that since X_α is a zero of $J_\alpha(x)+Y_\alpha(x)$, from (4.3) in [8, p. 443] we have $\lambda_3(\alpha) \geqq \pi X_\alpha M_\alpha^2(X_\alpha)/\sqrt{2}$. Furthermore, since $xM_\alpha^2(x) \geqq 2/\pi$ if $\alpha \geqq \frac{1}{2}$ (see [15, pp. 446 and 447]), it follows that $\lambda_3(\alpha) \geqq \sqrt{2}$. From the integral representation of $J_\alpha^2(x)+Y_\alpha^2(x)$ in [15, p. 444], it is readily seen that $M_\alpha^2(x)$ is an increasing function of α. If $0 \leqq \alpha \leqq \frac{1}{2}$, then $xM_\alpha^2(x)$ is also an increasing function of x; see [15, p. 446]. Since X_α is increasing in α (see [15, p. 508]), $X_\alpha M_\alpha^2(X_\alpha) \geqq X_\alpha M_0^2(X_\alpha) \geqq X_0 M_0^2(X_0)$. Using $X_0 = 0.23$ (see [8, p. 438]), the last quantity is easily computed to be greater or equal to 0.93. Consequently, in both cases, $0 \leqq \alpha \leqq \frac{1}{2}$ and $\alpha \geqq \frac{1}{2}$, we have $\lambda_3(\alpha) \geqq e^{-1}$. Thus, condition (5.18) is included in condition (5.21) for $\mu \geqq 1$.

The condition in the second inequality of (5.15) can also be made more explicit. In view of (5.6), the Mean Value Theorem gives

$$(5.24)\qquad \sin\theta_\alpha(j_{\alpha,s}+\mu^{-\frac{1}{2}}) = \pm 1 + \cos\theta_\alpha(j_{\alpha,s}+\xi)\theta'_\alpha(j_{\alpha,s}+\xi)\mu^{-\frac{1}{2}}$$

for some $0<\xi<\mu^{-\frac{1}{2}}$. Since $j_{\alpha,s}+\xi > X_\alpha$ and $xM_\alpha^2(x) \geqq 2/\pi$ if $\alpha \geqq \frac{1}{2}$, we have, from (5.5), $|\theta'_\alpha(j_{\alpha,s}+\xi)| \leqq 1$ if $\alpha \geqq \frac{1}{2}$. For $0 \leqq \alpha \leqq \frac{1}{2}$, we can repeat an earlier argument (in the previous paragraph) to show that $xM_\alpha^2(x) \geqq X_0 M_\alpha^2(X_0) \geqq 0.93$ for all $x \geqq X_\alpha$. Thus from (5.5) we have $|\theta'_\alpha(j_{\alpha,s}+\xi)| \leqq 0.685$ if $0 \leqq \alpha \leqq \frac{1}{2}$. In both cases, we conclude

$$(5.25)\qquad |\sin\theta_\alpha(j_{\alpha,s}+\mu^{-\frac{1}{2}})| \geqq 1-\mu^{-\frac{1}{2}}.$$

Therefore, (5.15) holds if $\mu^{1/2} \geqq 2$. Summarizing the above results gives the following theorem.

THEOREM 1. *If* $\mu^{\frac{1}{2}} \geqq \max\{2, e\lambda_2(\alpha)\lambda_3(\alpha)(|A|+|B|)/2\}$, *then the zeros* $t_{\mu,s}$ *of the Jacobi function* $\varphi_\mu^{(\alpha,\beta)}$ *satisfy the uniform asymptotic estimates* (5.22) *and* (5.23).

Note that both results (5.22) and (5.23) are needed, since the quantity $j_{\alpha,s}/\mu$ may tend to zero or infinity.

To show that the error estimates in (5.22) and (5.23) are of the correct order, we recall a result given in [8, p. 453, Ex. 7.2], which states that the sth negative zero of the solution in (4.05) of that reference is given by

$$(5.26)\qquad \zeta = -\eta_s + \frac{2\eta_s B_0(-\eta_s)}{\mu^2} + \eta_s^{\frac{1}{2}} O(\mu^{-3}), \qquad \mu \to \infty,$$

uniformly with respect to unbounded s, where $\eta_s = j_{\alpha,s}^2/\mu^2$. Since $t=(-\zeta)^{\frac{1}{2}}$ by (2.4), it is readily seen that

$$(5.27)\qquad \begin{aligned} t_{\mu,s} = \gamma_s - \frac{\gamma_s B_0(-\gamma_s^2)}{\mu^2} &+ \gamma_s B_0^2(-\gamma_s^2) O(\mu^{-4}) \\ &+ B_0(-\gamma_s^2) O(\mu^{-5}) + O(\mu^{-3}), \qquad \mu \to \infty, \end{aligned}$$

uniformly with respect to s, where $\gamma_s = j_{\alpha,s}/\mu$. By the argument used for (4.4), it is easily shown that

$$(5.28)\qquad |B_0(-t^2)| \leqq \tfrac{1}{6}(|A|+3|B|).$$

Hence (5.27) gives

$$(5.29)\qquad t_{\mu,s} = \gamma_s - \frac{\gamma_s B_0(-\gamma_s^2)}{\mu^2} + \gamma_s O(\mu^{-4}) + O(\mu^{-3}).$$

Using the same argument as for (4.3), we also have from (3.10),

$$|tB_0(-t^2)| \leqq \tfrac{1}{2}(|A|+|B|), \qquad t>0. \tag{5.30}$$

In view of (5.28) and (5.30), it is now clear from the second term in (5.29) that the error estimates for the one-term approximation (5.22), (5.23) are of the right asymptotic order.

By using similar arguments as above, error bounds have been constructed for the two-term expansion (5.29). The details are, however, much more complicated, and can be found in [14]. For this case, use has also been made of the results (4.7) and (4.8).

6. Large-*t* behavior. Turning our attention to the asymptotic behavior of $\varphi_\mu^{(\alpha,\beta)}(t)$ as $t \to \infty$, we first recall the following Liouville-Green approximation. Let $f(t)$ be a real, positive, twice continuously differentiable function on a finite or infinite interval (a_1, a_2), and let $g(t)$ be a continuous real or complex function on (a_1, a_2). Theorem 2.2 in [8, p. 196] states that the differential equation

$$\frac{d^2u}{dt^2}+\{f(t)-g(t)\}u=0 \tag{6.1}$$

has two linearly independent, twice continuously differentiable solutions

$$u_1(t)=f^{-\frac{1}{4}}(t)\exp\left\{i\int f^{\frac{1}{2}}(t)\,dt\right\}\{1+\varepsilon_1(t)\}, \tag{6.2}$$

$$u_2(t)=f^{-\frac{1}{4}}(t)\exp\left\{-i\int f^{\frac{1}{2}}(t)\,dt\right\}\{1+\varepsilon_2(t)\}, \tag{6.3}$$

such that

$$|\varepsilon_j(t)| \leqq \exp\{\mathcal{V}_{a,t}(F)\}-1, \qquad j=1,2, \tag{6.4}$$

where a is an arbitrary point in the closure of (a_1, a_2), and

$$F(t)=\int\left\{\frac{1}{f^{\frac{1}{4}}}\frac{d^2}{dt^2}\left(\frac{1}{f^{\frac{1}{4}}}\right)-\frac{g}{f^{\frac{1}{2}}}\right\}dt. \tag{6.5}$$

To apply this result to (2.3), we take $a=\infty$, $f(t)=\mu^2$, and

$$g(t)=-\frac{\frac{1}{4}-\alpha^2}{\sinh^2 t}+\frac{\frac{1}{4}-\beta^2}{\cosh^2 t},$$

and get the two linearly independent solutions

$$u_1(t)=\mu^{-\frac{1}{2}}\,e^{i\mu t}\{1+\varepsilon_1(t)\}, \tag{6.6}$$

$$u_2(t)=\mu^{-\frac{1}{2}}\,e^{-i\mu t}\{1+\varepsilon_2(t)\}, \tag{6.7}$$

with $|\varepsilon_j(t)| \leqq \exp\{\mathcal{V}_{t,\infty}(F)\}-1$. By (6.5),

$$F(t)=\frac{1}{\mu}\int\left\{\frac{\frac{1}{4}-\alpha^2}{\sinh^2 t}-\frac{\frac{1}{4}-\beta^2}{\cosh^2 t}\right\}dt,$$

from which it follows that

$$\begin{aligned}\int_t^\infty |F'(t)|\,dt &\leqq \frac{1}{\mu}\left\{|A|\int_t^\infty \operatorname{csch}^2 t\,dt+|B|\int_t^\infty \operatorname{sech}^2 t\,dt\right\}\\ &\leqq \frac{1}{\mu}\{|A|(\coth t-1)+|B|(1-\tanh t)\}.\end{aligned}$$

Simple estimation gives

$$\mathscr{V}_{t,\infty}(F) \leqq \frac{1}{\mu}\{4|A|\, e^{-2t}+2|B|\, e^{-2t}\}$$

for $t \geqq \frac{1}{2}$. Thus the error terms in (6.6) and (6.7) satisfy

$$|\varepsilon_j(t)| \leqq \exp\left\{\frac{1}{\mu}(4|A|+2|B|)\, e^{-2t}\right\}-1, \qquad j=1,2, \tag{6.8}$$

for $t \geqq \frac{1}{2}$. The general solution $\tilde{u}(t)$ of (2.3) is a linear combination of the two solutions $u_1(t)$ and $u_2(t)$ given in (6.6) and (6.7). Hence

$$\tilde{u}(t)=\frac{1}{\mu^{\frac{1}{2}}}\{c_1\, e^{i\mu t}+c_2\, e^{-i\mu t}+\tilde{\varepsilon}(t)\}, \tag{6.9}$$

where c_1 and c_2 are arbitrary constants and $\tilde{\varepsilon}(t)$ satisfies

$$|\tilde{\varepsilon}(t)| \leqq (|c_1|+|c_2|)\left\{\exp\left[\frac{1}{\mu}(4|A|+2|B|)\, e^{-2t}\right]-1\right\} \tag{6.10}$$

for $t \geqq \frac{1}{2}$. Note that $\tilde{\varepsilon}(t)=O(e^{-2t})$ as $t \to \infty$. Consequently,

$$\tilde{u}(t)=\frac{1}{\mu^{\frac{1}{2}}}\{c_1\, e^{i\mu t}+c_2\, e^{-i\mu t}+O(e^{-2t})\} \tag{6.11}$$

as $t \to \infty$. Let $u(t)$ be the function defined by (2.2), i.e.,

$$u(t)=(\sinh t)^{\alpha+\frac{1}{2}}(\cosh t)^{\beta+\frac{1}{2}}\varphi_\mu^{(\alpha,\beta)}(t). \tag{6.12}$$

From (1.1) and the connection formula (10.16) in [8, p. 167], it can be shown that

$$u(t)=\Gamma(\alpha+1)\{\Lambda(\mu)\, e^{i\mu t}+\overline{\Lambda(\mu)}\, e^{-i\mu t}+O(e^{-2t})\} \tag{6.13}$$

as $t \to \infty$, where

$$\Lambda_\mu \equiv \Lambda_{\alpha,\beta}(\mu) \equiv \frac{\Gamma(i\mu)2^{-i\mu}}{\Gamma[\frac{1}{2}(\alpha+\beta+1+i\mu)]\Gamma[\frac{1}{2}(\alpha-\beta+1+i\mu)]} \tag{6.14}$$

and the bar in (6.13) denotes the complex conjugate; cf. [7, eqns. (2.17) and (2.18)]. Comparing (6.11) and (6.13), we conclude that $u(t)=\tilde{u}(t)$ if we choose

$$c_1=\Gamma(\alpha+1)\mu^{\frac{1}{2}}\Lambda(\mu), \qquad c_2=\Gamma(\alpha+1)\mu^{\frac{1}{2}}\overline{\Lambda(\mu)},$$

and, consequently, we obtain from (6.6) and (6.7).

$$u(t)=2\Gamma(\alpha+1)|\Lambda(\mu)|\{\cos(\mu t+\theta_\mu)+\varepsilon^*(t)\}, \tag{6.15}$$

where $\theta_\mu=\theta_\mu(\alpha,\beta)$ denotes the argument of $\Lambda(\mu)$, i.e.,

$$\Lambda(\mu)=|\Lambda(\mu)|\, e^{i\theta_\mu}$$

and

$$|\varepsilon^*(t)| \leqq \exp\left\{\frac{1}{\mu}(4|A|+2|B|)\, e^{-2t}\right\}-1. \tag{6.16}$$

Note that the error term $\varepsilon^*(t)$ is exponentially decaying in t. Indeed, we have

$$\varepsilon^*(t)=O(e^{-2t}) \tag{6.17}$$

as $t \to \infty$. Direct computation from (6.14) shows that $\theta_\mu(\frac{1}{2},\frac{1}{2})=\theta_\mu(\frac{1}{2},-\frac{1}{2})=-\pi/2$ and $\theta_\mu(-\frac{1}{2},\frac{1}{2})=\theta(-\frac{1}{2},-\frac{1}{2})=0$.

To derive an asymptotic formula for the zeros $t_{\mu,s}$ for large values of s, we will make use of the following corollary of Theorem A stated in [5].

THEOREM B. *In the interval* $[n\pi - \psi - \rho, n\pi - \psi + \rho]$, *where* $\rho < \pi/2$, *suppose* $f(t) = \sin(t+\psi) + \varepsilon(t)$, $f(t)$ *is continuous, and* $E = \max|\varepsilon(t)| < \sin\rho$. *Then there exists a zero* c *of* $f(t)$ *in the interval such that* $|c - (n\pi - \psi)| \leqq E\rho \csc\rho$.

To apply this result, we rewrite (6.15) in the form

$$u(t) = 2\Gamma(\alpha+1)|\Lambda(\mu)|\left\{\sin\left(\mu t + \theta_\mu + \frac{\pi}{2}\right) + \varepsilon^*(t)\right\}, \tag{6.18}$$

and we take $\psi = \theta_\mu + \pi/2$ and $\rho = \pi/4$. For t satisfying $s\pi - \theta_\mu - \frac{3}{4}\pi \leqq \mu t \leqq s\pi - \theta_\mu - \pi/4$, we have from (6.16),

$$|\varepsilon^*(t)| \leqq \exp\left\{\frac{1}{\mu}(4|A|+2|B|)\, e^{-2(s\pi - \theta_\mu - \frac{3}{4}\pi)/\mu}\right\} - 1. \tag{6.19}$$

Note that if $A = B = 0$, then $\varepsilon^*(t) = 0$; cf. [7, eqn. (2.11)]. Since $2 > 1/\ln(1+1/\sqrt{2})$, to have the right-hand side of (6.19) bounded by $\sin(\pi/4) = 1/\sqrt{2}$, we require

$$\left(s\pi - \theta_\mu - \frac{3}{4}\pi\right) \geqq \frac{\mu}{2}\ln\left\{\frac{2}{\mu}(4|A|+2|B|)\right\}. \tag{6.20}$$

Furthermore, since $e^x - 1 \leqq xe^x$, coupling (6.19) and (6.20) gives

$$E = \max|\varepsilon^*(t)| \leqq 1.7072(4|A|+2|B|)\, e^{-2(s\pi - \theta_\mu - \frac{3}{4}\pi)/\mu}\frac{1}{\mu}. \tag{6.21}$$

From Theorem B, it follows that $u(t)$ has a zero $t_{\mu,s}$ such that

$$\left|t_{\mu,s} - \left(s\pi - \theta_\mu - \frac{\pi}{2}\right)\frac{1}{\mu}\right| \leqq 1.8963(4|A|+2|B|)\, e^{-2(s\pi - \theta_\mu - \frac{3}{4}\pi)/\mu}\frac{1}{\mu^2}. \tag{6.22}$$

for all integers s satisfying (6.20). In particular, we have

$$t_{\mu,s} = \left(s\pi - \theta_\mu - \frac{\pi}{2}\right)\frac{1}{\mu} + O(e^{-2\pi s/\mu}) \tag{6.23}$$

as $s \to \infty$. Note that (6.23) agrees with the exact results: $t_{\mu,s} = s\pi/\mu$ when $\alpha = \beta = \frac{1}{2}$, and $t_{\mu,s} = (s - \frac{1}{2})\pi/\mu$ when $\alpha = \beta = -\frac{1}{2}$.

Acknowledgments. The authors would like to thank the referees for a careful reading of this paper and several helpful suggestions, and they are also grateful to one of the referees for pointing out an error in an earlier version.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1965.

[2] P. BARATELLA AND L. GATTESCHI, *The bounds for the error term of an asymptotic approximation of Jacobi polynomials*, in Orthogonal Polynomials and Their Applications, 1986, pp. 203-221; Lecture Notes in Math., 1329, Springer-Verlag, Berlin, New York, 1988.

[3] A. FITOUHI AND M. M. HAMZA, *A uniform expansion for eigenfunction of a singular second order differential operator*, SIAM J. Math. Anal., 21 (1990), pp. 1619-1632.

[4] C. L. FRENZEN AND R. WONG, *A uniform asymptotic expansion of the Jacobi polynomials with error bounds*, Canad. J. Math., 37 (1985), pp. 979-1007.

[5] H. W. HETHCOTE, *Asymptotic Approximations with Error Bounds for Zeros of Airy and Cylindrical Functions*, Ph.D. dissertation, University of Michigan, Ann Arbor, MI, 1968.

[6] H. W. HETHCOTE, *Error bounds for asymptotic approximations of zeros of transcendental functions*, SIAM J. Math. Anal., 1 (1970), pp. 147-152.

[7] T. H. KOORNWINDER, *Jacobi functions and analysis on noncompact semisimple Lie groups*, in Special Functions: Group Theoretical Aspects and Applications, R. A. Askey, T. H. Koornwinder, and W. Schempp, eds., D. Reidel, Dordrecht, Holland, 1984, pp. 1-85.

[8] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[9] S. SCHINDLER, *Some transplantation theorems for the generalized Mehler transforms and related asymptotic expansions*, Trans. Amer. Math. Soc., 155 (1971), pp. 257-291.

[10] P. N. SHIVAKUMAR AND R. WONG, *Error bounds for a uniform asymptotic expansion of the Legendre function* $P_n^{-m}(\cosh z)$, Quart. Appl. Math., 46 (1988), pp. 473-488.

[11] R. J. STANTON AND P. A. TOMAS, *Expansion for spherical functions on noncompact symmetric spaces*, Acta Math., 140 (1978), pp. 251-276.

[12] G. SZEGÖ, *Orthogonal Polynomials*, Colloquium Publications, American Mathematical Society, Providence, RI, 1975.

[13] K. TRIMÉCHE, *Transformation intégrale de Weyl et théorème de Paley-Wiener associés à un opérateur différentièl singulier sur* $(0, \infty)$, J. Math. Pures Appl., 60 (1981), pp. 51-98.

[14] Q.-Q. WANG, *Uniform asymptotic expansions of the Jacobi functions and the Jacobi polynomials*, M. Sc. thesis, University of Manitoba, Winnipeg, Manitoba, 1989.

[15] G. N. WATSON, *A Treatise of the Theory of Bessel Functions*, Cambridge University Press, London, 1944.