SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 1–17, January 1998

THE VARIATIONAL FORMULATION OF THE FOKKER–PLANCK EQUATION*

RICHARD JORDAN†, DAVID KINDERLEHRER‡, AND FELIX OTTO§

In memory of Richard Duffin

Abstract. The Fokker–Planck equation, or forward Kolmogorov equation, describes the evolution of the probability density for a stochastic process associated with an Ito stochastic differential equation. It pertains to a wide variety of time-dependent systems in which randomness plays a role. In this paper, we are concerned with Fokker–Planck equations for which the drift term is given by the gradient of a potential. For a broad class of potentials, we construct a time discrete, iterative variational scheme whose solutions converge to the solution of the Fokker–Planck equation. The major novelty of this iterative scheme is that the time-step is governed by the Wasserstein metric on probability measures. This formulation enables us to reveal an appealing, and previously unexplored, relationship between the Fokker–Planck equation and the associated free energy functional. Namely, we demonstrate that the dynamics may be regarded as a gradient flux, or a steepest descent, for the free energy with respect to the Wasserstein metric.

Key words. Fokker–Planck equation, steepest descent, free energy, Wasserstein metric

AMS subject classifications. 35A15, 35K15, 35Q99, 60J60

PII. S0036141096303359

1. Introduction and overview.

The Fokker–Planck equation plays a central role in statistical physics and in the study of fluctuations in physical and biological systems [7, 22, 23]. It is intimately connected with the theory of stochastic differential equations: a (normalized) solution to a given Fokker–Planck equation represents the probability density for the position (or velocity) of a particle whose motion is described by a corresponding Ito stochastic differential equation (or Langevin equation). We shall restrict our attention in this paper to the case where the drift coefficient is a gradient. The simplest relevant physical setting is that of a particle undergoing diffusion in a potential field [23].

It is known that, under certain conditions on the drift and diffusion coefficients, the stationary solution of a Fokker–Planck equation of the type that we consider here satisfies a variational principle. It minimizes a certain convex free energy functional over an appropriate admissible class of probability densities [12]. This free energy functional satisfies an H-theorem: it decreases in time for any solution of the Fokker–Planck equation [22]. In this work, we shall establish a deeper, and apparently previously unexplored, connection between the free energy functional and the Fokker–Planck dynamics. Specifically, we shall demonstrate that the solution of the

*Received by the editors May 13, 1996; accepted for publication (in revised form) December 9, 1996. The research of all three authors is partially supported by the ARO and the NSF through grants to the Center for Nonlinear Analysis. In addition, the third author is partially supported by the Deutsche Forschungsgemeinschaft (German Science Foundation), and the second author is partially supported by grants NSF/DMS 9505078 and DAAL03-92-0003.

http://www.siam.org/journals/sima/29-1/30335.html

†Center for Nonlinear Analysis, Carnegie Mellon University. Present address: Department of Mathematics, University of Michigan (jordan@math.lsa.umich.edu).

‡Center for Nonlinear Analysis, Carnegie Mellon University (davidk@andrew.cmu.edu).

§Center for Nonlinear Analysis, Carnegie Mellon University and Department of Applied Mathematics, University of Bonn. Present address: Courant Institute of Mathematical Sciences (otto@cims.nyu.edu).

Fokker–Planck equation follows, at each instant in time, the direction of steepest descent of the associated free energy functional.

The notion of a steepest descent, or a gradient flux, makes sense only in context with an appropriate metric. We shall show that the required metric in the case of the Fokker–Planck equation is the Wasserstein metric (defined in section 3) on probability densities. As far as we know, the Wasserstein metric cannot be written as an induced metric for a metric tensor (the space of probability measures with the Wasserstein metric is not a Riemannian manifold). Thus, in order to give meaning to the assertion that the Fokker–Planck equation may be regarded as a steepest descent, or gradient flux, of the free energy functional with respect to this metric, we switch to a discrete time formulation. We develop a discrete, iterative variational scheme whose solutions converge, in a sense to be made precise below, to the solution of the Fokker–Planck equation. The time-step in this iterative scheme is associated with the Wasserstein metric. For a different view on the use of implicit schemes for measures, see [6, 16].

For the purpose of comparison, let us consider the classical diffusion (or heat) equation

$$\frac{\partial \rho(t,x)}{\partial t} = \Delta \rho(t,x)\,, \quad t \in (0,\infty)\,, \quad x \in \mathbb{R}^n\,,$$

which is the Fokker–Planck equation associated with a standard n-dimensional Brownian motion. It is well known (see, for example, [5, 24]) that this equation is the gradient flux of the Dirichlet integral $\frac{1}{2}\int_{R^n} |\nabla \rho|^2 \, dx$ with respect to the $L^2(\mathbb{R}^n)$ metric. The classical discretization is given by the scheme

$$\left.\begin{array}{l} \text{Determine } \rho^{(k)} \text{ that minimizes} \\ \frac{1}{2}\,\|\rho^{(k-1)} - \rho\|^2_{L^2(R^n)} \;+\; \frac{h}{2}\displaystyle\int_{R^n} |\nabla \rho|^2 \, dx \end{array}\right\}$$

over an appropriate class of densities ρ. Here, h is the time step size. On the other hand, we derive as a special case of our results below that the scheme

$$\left.\begin{array}{l} \text{Determine } \rho^{(k)} \text{ that minimizes} \\ \frac{1}{2}\, d(\rho^{(k-1)}, \rho)^2 \;+\; h \displaystyle\int_{\mathbb{R}^n} \rho \log \rho \, dx \\ \text{over all } \rho \in K\,, \end{array}\right\} \tag{1}$$

where K is the set of all probability densities on $\mathbb{R}^n$ having finite second moments, is also a discretization of the diffusion equation when d is the Wasserstein metric. In particular, this allows us to regard the diffusion equation as a steepest descent of the functional $\int_{\mathbb{R}^n} \rho \log \rho \, dx$ with respect to the Wasserstein metric. This reveals a novel link between the diffusion equation and the Gibbs–Boltzmann entropy $(-\int_{R^n} \rho \log \rho \, dx)$ of the density ρ. Furthermore, this formulation allows us to attach a precise interpretation to the conventional notion that diffusion arises from the tendency of the system to maximize entropy.

The connection between the Wasserstein metric and dynamical problems involving dissipation or diffusion (such as strongly overdamped fluid flow or nonlinear diffusion equations) seems to have first been recognized by Otto in [19]. The results in [19] together with our recent research on variational principles of entropy and free energy type for measures [12, 11, 15] provide the impetus for the present investigation. The work in [12] was motivated by the desire to model and characterize metastability

and hysteresis in physical systems. We plan to explore in subsequent research the relevance of the developments in the present paper to the study of such phenomena. Some preliminary results in this direction may be found in [13, 14].

The paper is organized as follows. In section 2, we first introduce the Fokker–Planck equation and briefly discuss its relationship to stochastic differential equations. We then give the precise form of the associated stationary solution and of the free energy functional that this density minimizes. In section 3, the Wasserstein metric is defined, and a brief review of its properties and interpretations is given. The iterative variational scheme is formulated in section 4, and the existence and uniqueness of its solutions are established. The main result of this paper—namely, the convergence of solutions of this scheme (after interpolation) to the solution of the Fokker–Planck equation—is the topic of section 5. There, we state and prove the relevant convergence theorem.

2. The Fokker–Planck equation, stationary solutions, and the free energy functional. We are concerned with Fokker–Planck equations having the form

$$\frac{\partial \rho}{\partial t} = \operatorname{div}\left(\nabla \Psi(x)\rho\right) + \beta^{-1}\Delta\rho\,, \quad \rho(x,0) = \rho^0(x), \tag{2}$$

where the potential $\Psi(x) : \mathbb{R}^n \to [0,\infty)$ is a smooth function, $\beta > 0$ is a given constant, and $\rho^0(x)$ is a probability density on $\mathbb{R}^n$. The solution $\rho(t,x)$ of (2) must, therefore, be a probability density on $\mathbb{R}^n$ for almost every fixed time t. That is, $\rho(t,x) \geq 0$ for almost every $(t,x) \in (0,\infty)\times\mathbb{R}^n$, and $\int_{\mathbb{R}^n} \rho(t,x)\,dx = 1$ for almost every $t \in (0,\infty)$.

It is well known that the Fokker–Planck equation (2) is inherently related to the Ito stochastic differential equation [7, 22, 23]

$$dX(t) = -\nabla\Psi(X(t))dt + \sqrt{2\beta^{-1}}\,dW(t)\,, \quad X(0) = X^0\,. \tag{3}$$

Here, $W(t)$ is a standard n-dimensional Wiener process, and X^0 is an n-dimensional random vector with probability density ρ^0. Equation (3) is a model for the motion of a particle undergoing diffusion in the potential field Ψ. $X(t) \in \mathbb{R}^n$ then represents the position of the particle, and the positive parameter β is proportional to the inverse temperature. This stochastic differential equation arises, for example, as the Smoluchowski–Kramers approximation to the Langevin equation for the motion of a chemically bound particle [23, 4, 17]. In that case, the function Ψ describes the chemical bonding forces, and the term $\sqrt{2\beta^{-1}}\,dW(t)$ represents white noise forces resulting from molecular collisions [23]. The solution $\rho(t,x)$ of the Fokker–Planck equation (2) furnishes the probability density at time t for finding the particle at position x.

If the potential Ψ satisfies appropriate growth conditions, then there is a unique stationary solution $\rho_s(x)$ of the Fokker–Planck equation, and it takes the form of the Gibbs distribution [7, 22]

$$\rho_s(x) = Z^{-1}\exp(-\beta\Psi(x)), \tag{4}$$

where the partition function Z is given by the expression

$$Z = \int_{\mathbb{R}^n} \exp(-\beta\Psi(x))\,dx.$$

Note that, in order for equation (4) to make sense, Ψ must grow rapidly enough to ensure that Z is finite. The probability measure $\rho_s(x)\,dx$, when it exists, is the unique

invariant measure for the Markov process $X(t)$ defined by the stochastic differential equation (3).

It is readily verified (see, for example, [12]) that the Gibbs distribution ρ_s satisfies a variational principle—it minimizes over all probability densities on $\mathbb{R}^n$ the free energy functional

$$F(\rho) = E(\rho) + \beta^{-1} S(\rho), \tag{5}$$

where

$$E(\rho) := \int_{\mathbb{R}^n} \Psi \rho \, dx \tag{6}$$

plays the role of an energy functional, and

$$S(\rho) := \int_{\mathbb{R}^n} \rho \log \rho \, dx \tag{7}$$

is the negative of the Gibbs–Boltzmann entropy functional.

Even when the Gibbs measure is not defined, the free energy (5) of a density $\rho(t, x)$ satisfying the Fokker–Planck equation (2) may be defined, provided that $F(\rho^0)$ is finite. This free energy functional then serves as a Lyapunov function for the Fokker–Planck equation: if $\rho(t, x)$ satisfies (2), then $F(\rho(t, x))$ can only decrease with time [22, 14]. Thus, the free energy functional is an H-function for the dynamics. The developments that follow will enable us to regard the Fokker–Planck dynamics as a gradient flux, or a steepest descent, of the free energy with respect to a particular metric on an appropriate class of probability measures. The requisite metric is the Wasserstein metric on the set of probability measures having finite second moments. We now proceed to define this metric.

3. The Wasserstein metric. The Wasserstein distance of order two, $d(\mu_1, \mu_2)$, between two (Borel) probability measures μ_1 and μ_2 on $\mathbb{R}^n$ is defined by the formula

$$d(\mu_1, \mu_2)^2 = \inf_{p \in \mathcal{P}(\mu_1, \mu_2)} \int_{R^n \times R^n} |x - y|^2 \, p(dxdy), \tag{8}$$

where $\mathcal{P}(\mu_1, \mu_2)$ is the set of all probability measures on $\mathbb{R}^n \times \mathbb{R}^n$ with first marginal μ_1 and second marginal μ_2, and the symbol $|\cdot|$ denotes the usual Euclidean norm on $\mathbb{R}^n$. More precisely, a probability measure p is in $\mathcal{P}(\mu_1, \mu_2)$ if and only if for each Borel subset $A \subset \mathbb{R}^n$ there holds

$$p(A \times \mathbb{R}^n) = \mu_1(A), \;\; p(\mathbb{R}^n \times A) = \mu_2(A).$$

Wasserstein distances of order q with q different from 2 may be analogously defined [10]. Since no confusion should arise in doing so, we shall refer to d in what follows as simply the Wasserstein distance.

It is well known that d defines a metric on the set of probability measures μ on R^n having finite second moments: $\int_{\mathbb{R}^n} |x|^2 \mu(dx) < \infty$ [10, 21]. In particular, d satisfies the triangle inequality on this set. That is, if μ_1, μ_2, and μ_3 are probability measures on $\mathbb{R}^n$ with finite second moments, then

$$d(\mu_1, \mu_3) \leq d(\mu_1, \mu_2) + d(\mu_2, \mu_3)\,. \tag{9}$$

We shall make use of this property at several points later on.

We note that the Wasserstein metric may be equivalently defined by [21]

$$d(\mu_1, \mu_2)^2 = \inf \mathbf{E}|X - Y|^2 \,, \tag{10}$$

where $\mathbf{E}(U)$ denotes the expectation of the random variable U, and the infimum is taken over all random variables X and Y such that X has distribution μ_1 and Y has distribution μ_2. In other words, the infimum is over all possible couplings of the random variables X and Y. Convergence in the metric d is equivalent to the usual weak convergence plus convergence of second moments. This latter assertion may be demonstrated by appealing to the representation (10) and applying the well-known Skorohod theorem from probability theory (see Theorem 29.6 of [1]). We omit the details.

The variational problem (8) is an example of a Monge–Kantorovich mass transference problem with the particular cost function $c(x, y) = |x-y|^2$ [21]. In that context, an infimizer $p^* \in \mathcal{P}(\mu_1, \mu_2)$ is referred to as an optimal transference plan. When μ_1 and μ_2 have finite second moments, the existence of such a p^* for (8) is readily verified by a simple adaptation of our arguments in section 4. For a probabilistic proof that the infimum in (8) is attained when μ_1 and μ_2 have finite second moments, see [10]. Brenier [2] has established the existence of a *one-to-one* optimal transference plan in the case that the measures μ_1 and μ_2 have bounded support and are absolutely continuous with respect to Lebesgue measure. Caffarelli [3] and Gangbo and McCann [8, 9] have recently extended Brenier's results to more general cost functions c and to cases in which the measures do not have bounded support.

If the measures μ_1 and μ_2 are absolutely continuous with respect to the Lebesgue measure, with densities ρ_1 and ρ_2, respectively, we will write $\mathcal{P}(\rho_1, \rho_2)$ for the set of probability measures having first marginal μ_1 and second marginal μ_2. Correspondingly, we will denote by $d(\rho_1, \rho_2)$ the Wasserstein distance between μ_1 and μ_2. This is the situation that we will be concerned with in what follows.

4. The discrete scheme. We shall now construct a time-discrete scheme that is designed to converge in an appropriate sense (to be made precise in the next section) to a solution of the Fokker–Planck equation. The scheme that we shall describe was motivated by a similar scheme developed by Otto in an investigation of pattern formation in magnetic fluids [19]. We shall make the following assumptions concerning the potential Ψ introduced in section 2:

$$\Psi \in C^\infty(\mathbb{R}^n);$$

$$\Psi(x) \;\geq\; 0 \;\text{ for all } x \in \mathbb{R}^n \,; \tag{11}$$

$$|\nabla \Psi(x)| \;\leq\; C\,(\Psi(x) + 1) \;\text{ for all } x \in \mathbb{R}^n \tag{12}$$

for some constant $C < \infty$. Notice that our assumptions on Ψ allow for cases in which $\int_{R^n} \exp(-\beta\Psi)\, dx$ is not defined, so the stationary density ρ_s given by (4) does not exist. These assumptions allow us to treat a wide class of Fokker–Planck equations. In particular, the classical diffusion equation $\frac{\partial \rho}{\partial t} = \beta^{-1}\Delta\rho$, for which $\Psi \equiv$ const., falls into this category. We also introduce the set K of admissible probability densities:

$$K \;:=\; \left\{ \rho \colon \mathbb{R}^n \to [0, \infty) \text{ measurable} \;\middle|\; \int_{\mathbb{R}^n} \rho(x)\, dx = 1 \,, M(\rho) < \infty \right\} ,$$

where

$$M(\rho) = \int_{\mathbb{R}^n} |x|^2\, \rho(x)\, dx \,.$$

With these conventions in hand, we now formulate the iterative discrete scheme:

$$\left.\begin{array}{l} \text{Determine } \rho^{(k)} \text{ that minimizes} \\ \frac{1}{2}\, d(\rho^{(k-1)}, \rho)^2 \;+\; h\, F(\rho) \\ \text{over all } \;\rho \in K\,. \end{array}\right\} \tag{13}$$

Here we use the notation $\rho^{(0)} = \rho^0$. The scheme (13) is the obvious generalization of the scheme (1) set forth in the Introduction for the diffusion equation. We shall now establish existence and uniqueness of the solution to (13).

PROPOSITION 4.1. *Given $\rho^0 \in K$, there exists a unique solution of the scheme (13).*

Proof. Let us first demonstrate that S is well defined as a functional on K with values in $(-\infty, +\infty]$ and that, in addition, there exist $\alpha < 1$ and $C < \infty$ depending only on n such that

$$S(\rho) \;\geq\; -C\,(M(\rho)+1)^{\alpha} \quad \text{for all } \rho \in K\,. \tag{14}$$

Actually, we shall show that (14) is valid for any $\alpha \in (\frac{n}{n+2}, 1)$. For future reference, we prove a somewhat finer estimate. Namely, we demonstrate that there exists a $C < \infty$, depending only on n and α, such that for all $R \geq 0$, and for each $\rho \in K$, there holds

$$\int_{R^n - B_R} |\min\{\rho \log \rho, 0\}|\, dx \;\leq\; C \left(\frac{1}{R^2+1}\right)^{\frac{(2+n)\,\alpha - n}{2}} (M(\rho)+1)^{\alpha}\,, \tag{15}$$

where B_R denotes the ball of radius R centered at the origin in $\mathbb{R}^n$. Indeed, for $\alpha < 1$ there holds

$$|\min\{z \log z, 0\}| \;\leq\; C\, z^{\alpha} \quad \text{for all } z \geq 0.$$

Hence by Hölder's inequality, we obtain

$$\begin{aligned} &\int_{R^n - B_R} |\min\{\rho \log \rho, 0\}|\, dx \\ &\leq C \int_{R^n - B_R} \rho^{\alpha}\, dx \\ &\leq C \left(\int_{R^n - B_R} \left(\frac{1}{|x|^2+1}\right)^{\frac{\alpha}{1-\alpha}}\, dx \right)^{1-\alpha} (M(\rho)+1)^{\alpha}\,. \end{aligned}$$

On the other hand, for $\frac{\alpha}{1-\alpha} > \frac{n}{2}$, we have

$$\int_{R^n - B_R} \left(\frac{1}{|x|^2+1}\right)^{\frac{\alpha}{1-\alpha}}\, dx \;\leq\; C \left(\frac{1}{R^2+1}\right)^{\frac{\alpha}{1-\alpha} - \frac{n}{2}}\,.$$

Let us now prove that for given $\rho^{(k-1)} \in K$, there exists a minimizer $\rho \in K$ of the functional

$$K \ni \rho \mapsto \tfrac{1}{2}\, d(\rho^{(k-1)}, \rho)^2 \;+\; h\, F(\rho)\,. \tag{16}$$

Observe that S is not bounded below on K and hence F is not bounded below on K either. Nevertheless, using the inequality

$$M(\rho_1) \;\le\; 2\,M(\rho_0) \,+ 2\,d(\rho_0,\rho_1)^2 \quad \text{for all } \rho_0,\rho_1 \in K \tag{17}$$

(which immediately follows from the inequality $|y|^2 \le 2|x|^2 + 2|x-y|^2$ and from the definition of d) together with (14) we obtain

$$\begin{aligned} &\tfrac{1}{2}\,d(\rho^{(k-1)},\rho)^2 \;+\; h\,F(\rho) \\ &\overset{(17)}{\ge}\; \tfrac{1}{4}\,M(\rho) \;-\; \tfrac{1}{2}\,M(\rho^{(k-1)}) \;+\; h\,S(\rho) \\ &\overset{(14)}{\ge}\; \tfrac{1}{4}\,M(\rho) \;-\; C\,(M(\rho)+1)^{\alpha} \;-\; \tfrac{1}{2}\,M(\rho^{(k-1)}) \quad \text{for all } \rho \in K\,, \end{aligned} \tag{18}$$

which ensures that (16) is bounded below. Now, let $\{\rho_\nu\}$ be a minimizing sequence for (16). Obviously, we have that

$$\{S(\rho_\nu)\}_\nu \quad \text{is bounded above}\,, \tag{19}$$

and according to (18),

$$\{M(\rho_\nu)\}_\nu \quad \text{is bounded}\,. \tag{20}$$

The latter result, together with (15), implies that

$$\left\{ \int_{\mathbb{R}^n} |\min\{\rho_\nu \,\log \rho_\nu, 0\}|\; dx \right\}_\nu \quad \text{is bounded}\,,$$

which combined with (19) yields that

$$\left\{ \int_{\mathbb{R}^n} \max\{\rho_\nu \,\log \rho_\nu, 0\}\; dx \right\}_\nu \quad \text{is bounded}\,.$$

As $z \mapsto \max\{z\log z, 0\}, z \in [0,\infty)$, has superlinear growth, this result, in conjunction with (20), guarantees the existence of a $\rho^{(k)} \in K$ such that (at least for a subsequence)

$$\rho_\nu \overset{w}{\rightharpoonup} \rho^{(k)} \quad \text{in } L^1(\mathbb{R}^n)\,. \tag{21}$$

Let us now show that

$$S(\rho^{(k)}) \;\le\; \liminf_{\nu\uparrow\infty} S(\rho_\nu)\,. \tag{22}$$

As $[0,\infty) \ni z \mapsto z\log z$ is convex and $[0,\infty) \ni z \mapsto \max\{z\log z, 0\}$ is convex and nonnegative, (21) implies that for any $R < \infty$,

$$\int_{B_R} \rho^{(k)} \,\log \rho^{(k)}\; dx \;\le\; \liminf_{\nu\uparrow\infty} \int_{B_R} \rho_\nu \,\log \rho_\nu \; dx\,, \tag{23}$$

$$\int_{\mathbb{R}^n - B_R} \max\{\rho^{(k)} \log \rho^{(k)}, 0\}\, dx \;\le\; \liminf_{\nu\uparrow\infty} \int_{\mathbb{R}^n - B_R} \max\{\rho_\nu \log \rho_\nu, 0\}\, dx. \tag{24}$$

On the other hand we have according to (15) and (20)

$$\limsup_{R\uparrow\infty} \sup_{\nu\in\mathbb{N}} \int_{\mathbb{R}^n - B_R} |\min\{\rho_\nu \,\log \rho_\nu, 0\}|\; dx \;=\; 0\,. \tag{25}$$

Now observe that for any $R < \infty$, there holds

$$S(\rho^{(k)}) \;\le\; \int_{B_R} \rho^{(k)}\,\log\rho^{(k)}\;dx \;+\; \int_{R^n - B_R} \max\{\rho^{(k)}\,\log\rho^{(k)}, 0\}\;dx\,,$$

which together with (23), (24), and (25) yields (22).

It remains for us to show that

$$E(\rho^{(k)}) \le \liminf_{\nu\uparrow\infty} E(\rho_\nu)\,, \tag{26}$$

$$d(\rho^{(k-1)}, \rho^{(k)})^2 \le \liminf_{\nu\uparrow\infty} d(\rho^{(k-1)}, \rho_\nu)^2\,. \tag{27}$$

Equation (26) follows immediately from (21) and Fatou's lemma. As for (27), we choose $p_\nu \in \mathcal{P}(\rho^{(k-1)}, \rho_\nu)$ satisfying

$$\int_{R^n\times R^n} |x-y|^2\,p_\nu(dx\,dy) \;\le\; d(\rho^{(k-1)}, \rho_\nu)^2 \;+\; \tfrac{1}{\nu}.$$

By (20) the sequence of probability measures $\{\rho_\nu\,dx\}_{\nu\uparrow\infty}$ is tight, or relatively compact with respect to the usual weak convergence in the space of probability measures on $\mathbb{R}^n$ (i.e., convergence tested against bounded continuous functions) [1]. This, together with the fact that the density $\rho^{(k-1)}$ has finite second moment, guarantees that the sequence $\{p_\nu\}_{\nu\uparrow\infty}$ of probability measures on $\mathbb{R}^n\times\mathbb{R}^n$ is tight. Hence, there is a subsequence of $\{p_\nu\}_{\nu\uparrow\infty}$ that converges weakly to some probability measure p. From (21) we deduce that $p \in \mathcal{P}(\rho^{(k-1)}, \rho^{(k)})$. We now could invoke the Skorohod theorem [1] and Fatou's lemma to infer (27) from this weak convergence, but we prefer here to give a more analytic proof. For $R < \infty$ let us select a continuous function $\eta_R\colon \mathbb{R}^n \to [0,1]$ such that

$$\eta_R \;=\; 1 \;\text{ inside of } B_R \quad\text{and}\quad \eta_R \;=\; 0 \;\text{ outside of } B_{2R}\,.$$

We then have

$$\left.\begin{aligned} &\int_{R^n\times R^n} \eta_R(x)\,\eta_R(y)\,|x-y|^2\,p(dx\,dy) \\ &\quad= \lim_{\nu\uparrow\infty} \int_{R^n\times R^n} \eta_R(x)\,\eta_R(y)\,|x-y|^2\,p_\nu(dx\,dy) \\ &\quad\le \liminf_{\nu\uparrow\infty} d(\rho^{(k-1)}, \rho_\nu)^2 \end{aligned}\right\} \tag{28}$$

for each fixed $R < \infty$. On the other hand, using the monotone convergence theorem, we deduce that

$$\begin{aligned} d(\rho^{(k-1)}, \rho^{(k)})^2 &\le \int_{R^n\times R^n} |x-y|^2\,p(dx\,dy) \\ &= \lim_{R\uparrow\infty} \int_{R^n\times R^n} \eta_R(x)\,\eta_R(y)\,|x-y|^2\,p(dx\,dy)\,, \end{aligned}$$

which combined with (28) yields (27).

To conclude the proof of the proposition we establish that the functional (16) has at most one minimizer. This follows from the convexity of K and the strict convexity of (16). The strict convexity of (16) follows from the strict convexity of S, the linearity of E, and the (obvious) convexity over K of the functional $\rho \mapsto d(\rho^{(k-1)}, \rho)^2$. □

Remark. One of the referees has communicated to us the following simple estimate that could be used in place of (14)–(15) in the previous and subsequent analysis: for any $\Omega \subset \mathbb{R}^n$ (in particular, for $\Omega = \mathbb{R}^n - B_R$) and for all $\rho \in K$ there holds

$$\int_\Omega |\min\{\rho \log \rho, 0\}|\, dx \le C \int_\Omega \mathrm{e}^{-\frac{|x|}{2}}\, dx + \epsilon M(\rho) + \frac{1}{4\epsilon} \int_\Omega \rho\, dx \tag{29}$$

for any $\epsilon > 0$. To obtain the inequality (29), select $C > 0$ such that for all $z \in [0,1]$, we have $z|\log z| \le C\sqrt{z}$. Then, defining the sets $\Omega_0 = \Omega \cap \{\rho \le \exp(-|x|)\}$ and $\Omega_1 = \Omega \cap \{\exp(-|x|) < \rho \le 1\}$, we have

$$\begin{aligned}\int_\Omega |\min\{\rho \log \rho, 0\}|\, dx &= \int_{\Omega_0} \rho |(\log \rho)_-|\, dx + \int_{\Omega_1} \rho |(\log \rho)_-|\, dx \\ &\le C \int_\Omega \mathrm{e}^{-\frac{|x|}{2}}\, dx + \int_\Omega |x| \rho\, dx\,.\end{aligned}$$

The desired result (29) then follows from the inequality $|x| \le \epsilon |x|^2 + 1/(4\epsilon)$ for $\epsilon > 0$.

5. Convergence to the solution of the Fokker–Planck equation. We come now to our main result. We shall demonstrate that an appropriate interpolation of the solution to the scheme (13) converges to the unique solution of the Fokker–Planck equation. Specifically, the convergence result that we will prove here is as follows.

THEOREM 5.1. *Let $\rho^0 \in K$ satisfy $F(\rho^0) < \infty$, and for given $h > 0$, let $\{\rho_h^{(k)}\}_{k \in \mathbb{N}}$ be the solution of* (13)*. Define the interpolation $\rho_h \colon (0,\infty) \times \mathbb{R}^n \to [0,\infty)$ by*

$$\rho_h(t) \;=\; \rho_h^{(k)} \quad \textit{for } t \in [k\,h, (k+1)\,h) \;\; \textit{and } \; k \in \mathbb{N} \cup \{0\}.$$

Then as $h \downarrow 0$,

$$\rho_h(t) \;\rightharpoonup\; \rho(t) \quad \textit{weakly in } L^1(\mathbb{R}^n) \quad \textit{for all } \; t \in (0,\infty)\,, \tag{30}$$

where $\rho \in C^\infty((0,\infty) \times \mathbb{R}^n)$ is the unique solution of

$$\frac{\partial \rho}{\partial t} \;=\; \mathrm{div}(\rho \nabla \Psi) \;+\; \beta^{-1} \Delta \rho\,, \tag{31}$$

with initial condition

$$\rho(t) \;\to\; \rho^0 \quad \textit{strongly in } \; L^1(\mathbb{R}^n) \quad \textit{for} \quad t \downarrow 0 \tag{32}$$

and

$$M(\rho),\, E(\rho) \;\in\; L^\infty((0,T)) \quad \textit{for all } \; T < \infty\,. \tag{33}$$

Remark. A finer analysis reveals that

$$\rho_h \;\to\; \rho \quad \textit{strongly in } L^1((0,T) \times \mathbb{R}^n) \quad \textit{for all } \; T < \infty\,.$$

Proof. The proof basically follows along the lines of [19, Proposition 2, Theorem 3]. The crucial step is to recognize that the first variation of (16) with respect to the independent variables indeed yields a time-discrete scheme for (31), as will now be demonstrated. For notational convenience only, we shall set $\beta \equiv 1$ from here on in. As will be evident from the ensuing arguments, our proof works for any positive β. In

fact, it is not difficult to see that, with appropriate modifications to the scheme (13), we can establish an analogous convergence result for time-dependent β.

Let a smooth vector field with bounded support, $\xi \in C_0^\infty(\mathbb{R}^n, \mathbb{R}^n)$, be given, and define the corresponding *flux* $\{\Phi_\tau\}_{\tau \in R}$ by

$$\partial_\tau \, \Phi_\tau \;=\; \xi \circ \Phi_\tau \;\text{ for all } \tau \in \mathbb{R} \quad \text{and} \quad \Phi_0 \;=\; \mathrm{id}\,.$$

For any $\tau \in \mathbb{R}$, let the measure $\rho_\tau(y)\,dy$ be the *push forward* of $\rho^{(k)}(y)\,dy$ under Φ_τ. This means that

$$\int_{R^n} \rho_\tau(y)\,\zeta(y)\,dy \;=\; \int_{R^n} \rho^{(k)}(y)\,\zeta(\Phi_\tau(y))\,dy \quad \text{for all } \zeta \in C_0^0(\mathbb{R}^n)\,. \tag{34}$$

As Φ_τ is invertible, (34) is equivalent to the following relation for the densities:

$$\det \nabla \Phi_\tau \; \rho_\tau \circ \Phi_\tau \;=\; \rho^{(k)}\,. \tag{35}$$

By (16), we have for each $\tau > 0$

$$\tfrac{1}{\tau} \, \Big(\Big(\tfrac{1}{2}\, d(\rho^{(k-1)}, \rho_\tau)^2 + h\, F(\rho_\tau) \Big) - \Big(\tfrac{1}{2}\, d(\rho^{(k-1)}, \rho^{(k)})^2 + h\, F(\rho^{(k)}) \Big) \Big) \;\geq\; 0, \tag{36}$$

which we now investigate in the limit $\tau \downarrow 0$. Because Ψ is nonnegative, equation (34) also holds for $\zeta = \Psi$, i.e.,

$$\int_{R^n} \rho_\tau(y)\,\Psi(y)\,dy \;=\; \int_{R^n} \rho^{(k)}(y)\,\Psi(\Phi_\tau(y))\,dy\,.$$

This yields

$$\tfrac{1}{\tau} \, \Big(E(\rho_\tau) - E(\rho^{(k)}) \Big) \;=\; \int_{R^n} \tfrac{1}{\tau} \, \big(\Psi(\Phi_\tau(y)) - \Psi(y) \big) \; \rho^{(k)}(y)\,dy\,.$$

Observe that the difference quotient under the integral converges uniformly to $\nabla\Psi(y)\cdot\xi(y)$, hence implying that

$$\tfrac{\mathrm{d}}{\mathrm{d}\,\tau} \, [E(\rho_\tau)]_{\tau=0} \;=\; \int_{R^n} \nabla\Psi(y)\cdot\xi(y)\,\rho^{(k)}(y)\,dy\,. \tag{37}$$

Next, we calculate $\frac{\mathrm{d}}{\mathrm{d}\,\tau} \, [S(\rho_\tau)]_{\tau=0}$. Invoking an appropriate approximation argument (for instance approximating log by some function that is bounded below), we obtain

$$\begin{aligned}
&\int_{R^n} \rho_\tau(y) \, \log(\rho_\tau(y))\,dy \\
&\overset{(34)}{=} \int_{R^n} \rho^{(k)}(y) \, \log(\rho_\tau(\Phi_\tau(y)))\,dy \\
&\overset{(35)}{=} \int_{R^n} \rho^{(k)}(y) \, \log\left(\frac{\rho^{(k)}(y)}{\det\nabla\Phi_\tau(y)} \right) \, dy\,.
\end{aligned}$$

Therefore, we have

$$\tfrac{1}{\tau} \, \Big(S(\rho_\tau) - S(\rho^{(k)}) \Big) \;=\; -\int_{R^n} \rho^{(k)}(y) \, \tfrac{1}{\tau} \, \log(\det \nabla\Phi_\tau(y))\,dy\,.$$

Now using

$$\tfrac{\mathrm{d}}{\mathrm{d}\tau}[\det\nabla\Phi_\tau(y)]_{\tau=0} \;=\; \mathrm{div}\xi(y)\,,$$

together with the fact that $\Phi_0 = \mathrm{id}$, we see that the difference quotient under the integral converges uniformly to $\mathrm{div}\xi$, hence implying that

$$\tfrac{\mathrm{d}}{\mathrm{d}\,\tau}\;[S(\rho_\tau)]_{\tau=0} \;=\; -\int_{\mathbb{R}^n}\rho^{(k)}\,\mathrm{div}\xi\;dy\,. \tag{38}$$

Now, let p be optimal in the definition of $d(\rho^{(k-1)},\rho^{(k)})^2$ (see section 3). The formula

$$\int_{R^n\times R^n}\zeta(x,y)\,p_\tau(dx\,dy) = \int_{R^n\times R^n}\zeta(x,\Phi_\tau(y))\,p(dx\,dy)\;,\zeta\in C_0^0(\mathbb{R}^n\times\mathbb{R}^n)$$

then defines a $p_\tau\in\mathcal{P}(\rho^{(k-1)},\rho_\tau)$. Consequently, there holds

$$\begin{aligned}&\tfrac{1}{\tau}\;\Big(\tfrac12\,d(\rho^{(k-1)},\rho_\tau)^2\;-\;\tfrac12\,d(\rho^{(k-1)},\rho^{(k)})^2\Big)\\ &\leq\;\int_{R^n\times R^n}\tfrac{1}{\tau}\;\big(\tfrac12\,|\Phi_\tau(y)-x|^2\;-\tfrac12\,|y-x|^2\big)\;p(dx\,dy)\,,\end{aligned}$$

which implies that

$$\begin{aligned}&\limsup_{\tau\downarrow0}\,\tfrac{1}{\tau}\;\Big(\tfrac12\,d(\rho^{(k-1)},\rho_\tau)^2\;-\;\tfrac12\,d(\rho^{(k-1)},\rho^{(k)})^2\Big)\\ &\leq\;\int_{R^n\times R^n}(y-x)\cdot\xi(y)\,p(dx\,dy)\,.\end{aligned} \tag{39}$$

We now infer from (36), (37), (38), and (39) (and the symmetry in $\xi\to-\xi$) that

$$\left.\begin{aligned}&\int_{R^n\times R^n}(y-x)\cdot\xi(y)\,p(dx\,dy)\;+\;h\int_{R^n}\left(\nabla\Psi\cdot\xi-\mathrm{div}\xi\right)\,\rho^{(k)}\,dy\;=\;0\\ &\text{for all }\xi\in C_0^\infty(\mathbb{R}^n,\mathbb{R}^n)\,.\end{aligned}\right\} \tag{40}$$

Observe that because $p\in\mathcal{P}(\rho^{(k-1)},\rho^{(k)})$, there holds

$$\begin{aligned}&\left|\int_{R^n}(\rho^{(k)}-\rho^{(k-1)})\,\zeta\;dy\;-\;\int_{R^n\times R^n}(y-x)\cdot\nabla\zeta(y)\,p(dx\,dy)\right|\\ &=\;\left|\int_{R^n\times R^n}(\zeta(y)-\zeta(x)\;+\;(x-y)\cdot\nabla\zeta(y))\,p(dx\,dy)\right|\\ &\leq\;\tfrac12\,\sup_{R^n}|\nabla^2\zeta|\,\int_{R^n\times R^n}|y-x|^2\,p(dx\,dy)\\ &=\;\tfrac12\,\sup_{R^n}|\nabla^2\zeta|\,d(\rho^{(k-1)},\rho^{(k)})^2\end{aligned}$$

for all $\zeta\in C_0^\infty(\mathbb{R}^n)$. Choosing $\xi=\nabla\zeta$ in (40) then gives

$$\begin{aligned}&\left|\int_{R^n}\left\{\tfrac{1}{h}\,(\rho^{(k)}-\rho^{(k-1)})\,\zeta+(\nabla\Psi\cdot\nabla\zeta-\Delta\zeta)\;\rho^{(k)}\right\}\,dy\right|\\ &\qquad\leq\tfrac12\sup_{R^n}|\nabla^2\zeta|\;\tfrac{1}{h}\,d(\rho^{(k-1)},\rho^{(k)})^2\;\text{ for all }\;\zeta\in C_0^\infty(\mathbb{R}^n)\,.\end{aligned} \tag{41}$$

We wish now to pass to the limit $h \downarrow 0$. In order to do so we will first establish the following a priori estimates: for any $T < \infty$, there exists a constant $C < \infty$ such that for all $N \in \mathbb{N}$ and all $h \in [0,1]$ with $N\,h \le T$, there holds

$$M(\rho_h^{(N)}) \le C\,, \tag{42}$$

$$\int_{R^n} \max\{\rho_h^{(N)}\,\log\rho_h^{(N)}, 0\}\,dx \le C\,, \tag{43}$$

$$E(\rho_h^{(N)}) \le C\,, \tag{44}$$

$$\sum_{k=1}^{N} d(\rho_h^{(k-1)}, \rho_h^{(k)})^2 \le C\,h\,. \tag{45}$$

Let us verify that the estimate (42) holds. Since $\rho_h^{(k-1)}$ is admissible in the variational principle (13), we have that

$$\tfrac{1}{2}\,d(\rho_h^{(k-1)}, \rho_h^{(k)})^2 \;+\; h\,F(\rho_h^{(k)}) \;\le\; h\,F(\rho_h^{(k-1)})\,,$$

which may be summed over k to give

$$\sum_{k=1}^{N} \tfrac{1}{2\,h}\,d(\rho_h^{(k-1)}, \rho_h^{(k)})^2 \;+\; F(\rho_h^{(N)}) \;\le\; F(\rho^0)\,. \tag{46}$$

As in Proposition 4.1, we must confront the technical difficulty that F is not bounded below. The inequality (42) is established via the following calculations:

$$\begin{aligned}
M(\rho_h^{(N)}) &\overset{(17)}{\le} 2\,d(\rho^0, \rho_h^{(N)})^2 \;+\; 2\,M(\rho^0) \\
&\le\; 2\,N \sum_{k=1}^{N} d(\rho_h^{(k-1)}, \rho_h^{(k)})^2 \;+\; 2\,M(\rho^0) \\
&\overset{(46)}{\le} 4\,h\,N\,\Big(F(\rho^0) - F(\rho_h^{(N)})\Big) \;+\; 2\,M(\rho^0) \\
&\overset{(14)}{\le} 4\,T\,\Big(F(\rho^0) + C\,(M(\rho_h^{(N)}) + 1)^\alpha\Big) \;+\; 2\,M(\rho^0)\,,
\end{aligned}$$

which clearly gives (42). To obtain the second line of the above display, we have made use of the triangle inequality for the Wasserstein metric (see equation (9)) and the Cauchy–Schwarz inequality. The estimates (43), (44), and (45) now follow readily from the bounds (14) and (15), the estimate (42), and the inequality (46), as follows:

$$\begin{aligned}
\int_{R^n} \max\{\rho_h^{(N)}\,\log\rho_h^{(N)}, 0\}\,dx \;&\le\; S(\rho_h^{(N)}) \;+\; \int_{R^n} |\min\{\rho_h^{(N)}\,\log\rho_h^{(N)}, 0\}|\,dx \\
&\overset{(15)}{\le} S(\rho_h^{(N)}) \;+\; C\,(M(\rho_h^{(N)}) + 1)^\alpha \\
&\le\; F(\rho_h^{(N)}) \;+\; C\,(M(\rho_h^{(N)}) + 1)^\alpha \\
&\overset{(46)}{\le} F(\rho^0) \;+\; C\,(M(\rho_h^{(N)}) + 1)^\alpha\,;
\end{aligned}$$

$$\begin{aligned}
E(\rho_h^{(N)}) \;&=\; F(\rho_h^{(N)}) \;-\; S(\rho_h^{(N)}) \\
&\overset{(14)}{\le} F(\rho_h^{(N)}) \;+\; C\,(M(\rho_h^{(N)}) + 1)^\alpha \\
&\overset{(46)}{\le} F(\rho^0) \;+\; C\,(M(\rho_h^{(N)}) + 1)^\alpha\,;
\end{aligned}$$

$$\sum_{k=1}^{N} d(\rho_h^{(k-1)}, \rho_h^{(k)})^2 \overset{(46)}{\leq} 2\,h\,\Big(F(\rho^0)\;-\;F(\rho_h^{(N)})\Big)$$
$$\overset{(14)}{\leq} 2\,h\,\Big(F(\rho^0)\;+\;C\,(M(\rho_h^{(N)})+1)^{\alpha}\Big)\,.$$

Now, owing to the estimates (42) and (43), we may conclude that there exists a measurable $\rho(t,x)$ such that, after extraction of a subsequence,

$$\rho_h \;\rightharpoonup\; \rho \quad \text{weakly in } L^1((0,T)\times\mathbb{R}^n) \quad \text{for all } T<\infty\,. \tag{47}$$

A straightforward analysis reveals that (42), (43), and (44) guarantee that

$$\begin{aligned} &\rho(t)\in K \quad \text{for a.e. } t\in(0,\infty)\,,\\ &M(\rho),\,E(\rho)\in L^\infty((0,T)) \quad \text{for all } T<\infty\,. \end{aligned} \tag{48}$$

Let us now improve upon the convergence in (47) by showing that (30) holds. For a given finite time horizon $T<\infty$, there exists a constant $C<\infty$ such that for all $N,N'\in\mathbb{N}$ and all $h\in[0,1]$ with $N\,h\leq T$, and $N'\,h\leq T$, we have

$$d(\rho_h^{(N')},\rho_h^{(N)})^2 \;\leq\; C\,|N'\,h-N\,h|\,.$$

This result is obtained from (45) by use of the triangle inequality (9) for d and the Cauchy–Schwarz inequality. Furthermore, for all $\rho,\rho'\in K\,, p\in\mathcal{P}(\rho,\rho')$, and $\zeta\in C_0^\infty(\mathbb{R}^n)$, there holds

$$\begin{aligned} \left|\int_{R^n}\zeta\,\rho'\,dx\;-\;\int_{R^n}\zeta\,\rho\,dx\right| &= \left|\int_{R^n\times R^n}(\zeta(x)\;-\;\zeta(y))\,p(dx\,dy)\right|\\ &\leq \sup_{R^n}|\nabla\zeta|\int_{R^n\times R^n}|x-y|\,p(dx\,dy)\\ &\leq \sup_{R^n}|\nabla\zeta|\left(\int_{R^n\times R^n}|x-y|^2\,p(dx\,dy)\right)^{\frac12}\,, \end{aligned}$$

so from the definition of d we obtain

$$\left|\int_{R^n}\zeta\,\rho'\,dx\;-\;\int_{R^n}\zeta\,\rho\,dx\right| \;\leq\; \sup_{R^n}|\nabla\zeta|\,d(\rho,\rho') \;\text{ for } \rho,\rho'\in K \text{ and } \zeta\in C_0^\infty(\mathbb{R}^n)\,.$$

Hence, it follows that

$$\left.\begin{aligned} &\left|\int_{R^n}\zeta\,\rho_h(t')\,dx\;-\;\int_{R^n}\zeta\,\rho_h(t)\,dx\right| \;\leq\; C\sup_{R^n}|\nabla\zeta|\;(|t'-t|+h)^{\frac12}\\ &\text{for all } t,t'\in(0,T),\text{ and } \zeta\in C_0^\infty(\mathbb{R}^n)\,. \end{aligned}\right\} \tag{49}$$

Let $t\in(0,T)$ and $\zeta\in C_0^\infty(\mathbb{R}^n)$ be given, and notice that for any $\delta>0$, we have

$$\begin{aligned} &\left|\int_{R^n}\zeta\,\rho_h(t)\,dx\;-\;\int_{R^n}\zeta\,\rho(t)\,dx\right|\\ &\leq\;\left|\int_{R^n}\zeta\,\rho_h(t)\,dx\;-\;\tfrac{1}{2\,\delta}\int_{t-\delta}^{t+\delta}\int_{R^n}\zeta\,\rho_h(\tau)\,dx\,d\tau\right|\\ &+\;\left|\tfrac{1}{2\,\delta}\int_{t-\delta}^{t+\delta}\int_{R^n}\zeta\,\rho_h(\tau)\,dx\,d\tau\;-\;\tfrac{1}{2\,\delta}\int_{t-\delta}^{t+\delta}\int_{R^n}\zeta\,\rho(\tau)\,dx\,d\tau\right|\\ &+\;\left|\tfrac{1}{2\,\delta}\int_{t-\delta}^{t+\delta}\int_{R^n}\zeta\,\rho(\tau)\,dx\,d\tau\;-\;\int_{R^n}\zeta\,\rho(t)\,dx\right|\,. \end{aligned}$$

According to (49), the first term on the right-hand side of this equation is bounded by

$$C \sup_{R^n} |\nabla\zeta| \, (\delta + h)^{\frac{1}{2}} \, ,$$

and owing to (47), the second term converges to zero as $h \downarrow 0$ for any fixed $\delta > 0$. At this point, let us remark that from the result (47) we may deduce that ρ is smooth on $(0,\infty)\times\mathbb{R}^n$. This is the conclusion of assertion (a) below, which will be proved later. From this smoothness property, we ascertain that the final term on the right-hand side of the above display converges to zero as $\delta \downarrow 0$. Therefore, we have established that

$$\int_{R^n} \zeta \, \rho_h(t) \, dx \; \to \; \int_{R^n} \zeta \, \rho(t) \, dx \quad \text{for all } \zeta \in C_0^\infty(\mathbb{R}^n) \, . \tag{50}$$

However, the estimate (42) guarantees that $M(\rho_h(t))$ is bounded for $h \downarrow 0$. Consequently, (50) holds for any $\zeta \in L^\infty(\mathbb{R}^n)$, and therefore, the convergence result (30) does indeed hold.

It now follows immediately from (41), (45), and (47) that ρ satisfies

$$\left.\begin{aligned} -\int_{(0,\infty)\times R^n} \rho \, (\partial_t\zeta - \nabla\Psi\cdot\nabla\zeta + \Delta\zeta) \; dx \, dt \; &= \; \int_{R^n} \rho^0 \, \zeta(0) \; dx \, , \\ \text{for all } \zeta \in C_0^\infty(\mathbb{R}\times\mathbb{R}^n) \, . & \end{aligned}\right\} \tag{51}$$

In addition, we know that ρ satisfies (33). We now show that

(a) any solution of (51) is smooth on $(0,\infty)\times\mathbb{R}^n$ and satisfies equation (31);
(b) any solution of (51) for which (33) holds satisfies the initial condition (32);
(c) there is at most one smooth solution of (31) which satisfies (32) and (33).

The corresponding arguments are, for the most part, fairly classical.

Let us sketch the proof of the regularity part (a). First observe that (51) implies

$$\begin{aligned} \int_{R^n} \rho(t_1) \, \zeta(t_1) \; dx - \int_{(t_0,t_1)\times R^n} \rho \, (\partial_t\zeta - \nabla\Psi\cdot\nabla\zeta + \Delta\zeta) \; dx \, dt \\ = \int_{R^n} \rho(t_0) \, \zeta(t_0) \; dx \end{aligned} \tag{52}$$

$$\text{for all } \; \zeta \in C_0^\infty(\mathbb{R}\times\mathbb{R}^n) \; \text{ and a.e. } \; 0 \le t_0 < t_1 \, .$$

We fix a function $\eta \in C_0^\infty(\mathbb{R}^n)$ to serve as a cutoff in the spatial variables. It then follows directly from (52) that for each $\zeta \in C_0^\infty(\mathbb{R}\times\mathbb{R}^n)$ and for almost every $0 \le t_0 < t_1$, there holds

$$\left.\begin{aligned} \int_{R^n} \eta \, \rho(t_1) \, \zeta(t_1) \, dx \; - \quad & \int_{(t_0,t_1)\times R^n} \eta \, \rho \, (\partial_t\zeta + \Delta\zeta) \; dx \, dt \\ = \; & \int_{(t_0,t_1)\times R^n} \rho \; (\Delta\eta - \nabla\Psi\cdot\nabla\eta) \; \zeta \, dx \, dt \\ + \; & \int_{(t_0,t_1)\times R^n} \rho \; (2\,\nabla\eta - \eta\nabla\Psi)\cdot\nabla\zeta \, dx \, dt \\ + \quad & \int_{R^n} \eta \, \rho(t_0) \, \zeta(t_0) \, dx \, . \end{aligned}\right\} \tag{53}$$

Notice that for fixed $(t_1, x_1) \in (0,\infty)\times\mathbb{R}^n$ and for each $\delta > 0$, the function

$$\zeta_\delta(t,x) \; = \; G(t_1 + \delta - t, x - x_1)$$

is an admissible test function in (53). Here G is the heat kernel

$$G(t,x) \;=\; t^{-\frac{n}{2}}\, g(t^{-\frac{1}{2}}x) \quad \text{with} \quad g(x) \;=\; (2\pi)^{-\frac{n}{2}} \exp(-\tfrac{1}{2}\,|x|^2). \tag{54}$$

Inserting ζ_δ into (53) and taking the limit $\delta \downarrow 0$, we obtain the equation

$$\left.\begin{aligned}(\rho\,\eta)(t_1) = &\int_{t_0}^{t_1} [\rho(t)\,(\Delta\eta - \nabla\Psi{\cdot}\nabla\eta)] * G(t_1-t)\,dt\\ +\;&\int_{t_0}^{t_1} [\rho(t)\,(2\,\nabla\eta - \eta\,\nabla\Psi)] * \nabla G(t_1-t)\,dt\\ &+ (\rho\,\eta)(t_0) * G(t_1-t_0) \quad \text{for a.e. } 0 \le t_0 < t_1\,,\end{aligned}\right\} \tag{55}$$

where $*$ denotes convolution in the x-variables. From (55), we extract the following estimate in the L^p-norm:

$$\begin{aligned}\|(\rho\,\eta)(t_1)\|_{L^p} = &\int_{t_0}^{t_1} \|\rho(t)\,(\Delta\eta - \nabla\Psi{\cdot}\nabla\eta)\|_{L^1}\, \|G(t_1-t)\|_{L^p}\,dt\\ +\;&\int_{t_0}^{t_1} \|\rho(t)\,(2\,\nabla\eta - \eta\,\nabla\Psi)\|_{L^1}\, \|\nabla G(t_1-t)\|_{L^p}\,dt\\ &+ \|(\rho\,\eta)(t_0)\|_{L^1}\, \|G(t_1-t_0)\|_{L^p} \quad \text{for a.e. } 0 \le t_0 < t_1\,.\end{aligned}$$

Now observe that

$$\begin{aligned}\|G(t)\|_{L^p} &= t^{(\frac{1}{p}-1)\,\frac{n}{2}}\, \|g\|_{L^p}\,,\\ \|\nabla G(t)\|_{L^p}, &= t^{\frac{1}{p}\,\frac{n}{2}-\frac{n+1}{2}}\, \|\nabla g\|_{L^p}\,,\end{aligned}$$

which leads to

$$\begin{aligned}&\|(\rho\,\eta)(t_1)\|_{L^p}\\ &= \operatorname*{ess\,sup}_{t\in(t_0,t_1)} \|\rho(t)\,(\Delta\eta - \nabla\Psi{\cdot}\nabla\eta)\|_{L^1} \int_0^{t_1-t_0} t^{(\frac{1}{p}-1)\,\frac{n}{2}}\, \|g\|_{L^p}\,dt\\ &+ \operatorname*{ess\,sup}_{t\in(t_0,t_1)} \|\rho(t)\,(2\,\nabla\eta - \eta\,\nabla\Psi)\|_{L^1} \int_0^{t_1-t_0} t^{\frac{1}{p}\,\frac{n}{2}-\frac{n+1}{2}}\, \|\nabla g\|_{L^p}\,dt\\ &+ \|(\rho\,\eta)(t_0)\|_{L^1}\, \|G(t_1-t_0)\|_{L^p} \quad \text{for a.e. } 0 \le t_0 < t_1\,.\end{aligned}$$

For $p < \frac{n}{n-1}$ the t-integrals are finite, from which we deduce that

$$\rho \;\in\; L^p_{\mathrm{loc}}((0,\infty){\times}\mathbb{R}^n)\,.$$

We now appeal to the L^p-estimates [18, section 3, (3.1), and (3.2)] for the potentials in (55) to conclude by the usual bootstrap arguments that any derivative of ρ is in $L^p_{\mathrm{loc}}((0,\infty){\times}\mathbb{R}^n)$, from which we obtain the stated regularity condition (a).

We now prove assertion (b). Using (55) with $t_0 = 0$, and proceeding as above, we obtain

$$\begin{aligned}&\|(\rho\,\eta)(t_1) \;-\; (\rho^0\,\eta) * G(t_1)\|_{L^1}\\ &= \operatorname*{ess\,sup}_{t\in(0,t_1)} \|\rho(t)\,(\Delta\eta - \nabla\Psi{\cdot}\nabla\eta)\|_{L^1} \int_0^{t_1} \|g\|_{L^1}\,dt\\ &+ \operatorname*{ess\,sup}_{t\in(0,t_1)} \|\rho(t)\,(2\,\nabla\eta - \eta\,\nabla\Psi)\|_{L^1} \int_0^{t_1} t^{-\frac{1}{2}}\, \|\nabla g\|_{L^1}\,dt \quad \text{for all } t_1 > 0\end{aligned}$$

and therefore,

$$(\rho\,\eta)(t) \;-\; (\rho^0\,\eta) * G(t) \;\to\; 0 \quad \text{in } L^1(\mathbb{R}^n) \quad \text{for } t \downarrow 0\,.$$

On the other hand, we have

$$(\rho^0\,\eta) * G(t) \;\to\; \rho^0\,\eta \quad \text{in } L^1(\mathbb{R}^n) \quad \text{for } t \downarrow 0\,,$$

which leads to

$$(\rho\,\eta)(t) \;\to\; \rho^0\,\eta \quad \text{in } L^1(\mathbb{R}^n) \quad \text{for } t \downarrow 0\,.$$

From this result, together with the boundedness of $\{M(\rho(t))\}_{t\downarrow 0}$, we infer that (32) is satisfied.

Finally, we prove the uniqueness result (c) using a well-known method from the theory of elliptic–parabolic equations (see, for instance, [20]). Let ρ_1, ρ_2 be solutions of (32) which are smooth on $(0,\infty)\times\mathbb{R}^n$ and satisfy (32), (33). Their difference ρ satisfies the equation

$$\frac{\partial \rho}{\partial t} \;-\; \operatorname{div}\left[\rho\,\nabla\Psi + \nabla\rho\right] \;=\; 0.$$

We multiply this equation for ρ by $\phi_\delta'(\rho)$, where the family $\{\phi_\delta\}_{\delta\downarrow 0}$ is a convex and smooth approximation to the modulus function. For example, we could take

$$\phi_\delta(z) \;=\; (z^2+\delta^2)^{\frac{1}{2}}\,.$$

This procedure yields the inequality

$$\begin{aligned}
&\partial_t[\phi_\delta(\rho)] \;-\; \operatorname{div}\left[\phi_\delta(\rho)\,\nabla\Psi + \nabla[\phi_\delta(\rho)]\right] \\
&= \;-\phi_\delta''(\rho)\,|\nabla\rho|^2 \;+\; \left(\phi_\delta'(\rho)\,\rho - \phi_\delta(\rho)\right)\,\Delta\Psi \\
&\le \;\left(\phi_\delta'(\rho)\,\rho - \phi_\delta(\rho)\right)\,\Delta\Psi,
\end{aligned}$$

which we then multiply by a nonnegative spatial cutoff function $\eta \in C_0^\infty(\mathbb{R}^n)$ and integrate over $\mathbb{R}^n$ to obtain

$$\begin{aligned}
\frac{d}{dt}\left[\int_{R^n} \phi_\delta(\rho(t))\,\eta\,dx\right] + \int_{R^n} \phi_\delta(\rho(t))\,\left(\nabla\Psi\cdot\nabla\eta - \Delta\eta\right)\,dx \\
\le \int_{R^n} \left(\phi_\delta'(\rho)\,\rho - \phi_\delta(\rho)\right)\,\Delta\Psi\,\eta\,dx\,.
\end{aligned}$$

Integrating over $(0,t)$ for given $t \in (0,\infty)$, we obtain with help of (32)

$$\begin{aligned}
\int_{R^n} \phi_\delta(\rho(t))\,\eta\,dx \;+ \int_{(0,t)\times R^n} \phi_\delta(\rho(t))\,\left(\nabla\Psi\cdot\nabla\eta - \Delta\eta\right)\,dx\,dt \\
\le \int_{(0,t)\times R^n} \left(\phi_\delta'(\rho)\,\rho - \phi_\delta(\rho)\right)\,\Delta\Psi\,\eta\,dx\,dt\,.
\end{aligned}$$

Letting δ tend to zero yields

$$\int_{R^n} |\rho(t)|\,\eta\,dx \;+ \int_{(0,t)\times R^n} |\rho(t)|\,\left(\nabla\Psi\cdot\nabla\eta - \Delta\eta\right)\,dx\,dt \;\le\; 0\,. \tag{56}$$

According to (12) and (33), ρ and $\rho\nabla\Psi$ are integrable on the entire $\mathbb{R}^n$. Hence, if we replace η in (56) by a function η_R satisfying

$$\eta_R(x) \;=\; \eta_1\left(\frac{x}{R}\right)\,, \quad \text{where } \eta_1(x) = 1 \text{ for } |x| \le 1\,,\;\; \eta_1(x) = 0 \text{ for } |x| \ge 2\,,$$

and let R tend to infinity, we obtain $\int_{R^n} |\rho(t)|\,dx = 0$. This produces the desired uniqueness result. □

Acknowledgments. The authors would like to thank the anonymous referees for valuable suggestions and comments.

REFERENCES

[1] P. Billingsley, *Probability and Measure*, John Wiley, New York, 1986.
[2] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.
[3] L. A. Caffarelli, *Allocation maps with general cost functions*, in Partial Differential Equations and Applications, P. Marcellini, G. G. Talenti, and E. Vesintini, eds., Lecture Notes in Pure and Applied Mathematics 177, Marcel Dekker, New York, NY, 1996, pp. 29–35.
[4] S. Chandrasekhar, *Stochastic problems in physics and astronomy*, Rev. Mod. Phys., 15 (1942), pp. 1–89.
[5] R. Courant, K. Friedrichs, and H. Lewy, *Über die partiellen Differenzgleichungen der mathematischen Physik*, Math. Ann., 100 (1928), pp. 1–74.
[6] S. Demoulini, *Young measure solutions for a nonlinear parabolic equation of forward–backward type*, SIAM J. Math. Anal., 27 (1996), pp. 376–403.
[7] C. W. Gardiner, *Handbook of stochastic methods*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 1985.
[8] W. Gangbo and R. J. McCann, *Optimal maps in Monge's mass transport problems*, C. R. Acad. Sci. Paris, 321 (1995), pp. 1653–1658.
[9] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.
[10] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions*, Michigan Math. J., 31 (1984), pp. 231–240.
[11] R. Jordan, *A statistical equilibrium model of coherent structures in magnetohydrodynamics*, Nonlinearity, 8 (1995), pp. 585–613.
[12] R. Jordan and D. Kinderlehrer, *An extended variational principle*, in Partial Differential Equations and Applications, P. Marcellini, G. G. Talenti, and E. Vesintini, eds., Lecture Notes in Pure and Applied Mathematics 177, Marcel Dekker, New York, NY, 1996, pp. 187–200.
[13] R. Jordan, D. Kinderlehrer, and F. Otto, *Free energy and the Fokker–Planck equation*, Physica D., to appear.
[14] R. Jordan, D. Kinderlehrer, and F. Otto, *The route to stability through the Fokker–Planck dynamics*, Proc. First U.S.–China Conference on Differential Equations and Applications, to appear.
[15] R. Jordan and B. Turkington, *Ideal magnetofluid turbulence in two dimensions*, J. Stat. Phys., 87 (1997), pp. 661–695.
[16] D. Kinderlehrer and P. Pedregal, *Weak convergence of integrands and the Young measure representation*, SIAM J. Math. Anal., 23 (1992), pp. 1–19.
[17] H. A. Kramers, *Brownian motion in a field of force and the diffusion model of chemical reactions*, Physica, 7 (1940), pp. 284–304.
[18] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi–Linear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.
[19] F. Otto, *Dynamics of labyrinthine pattern formation in magnetic fluids: A mean–field theory*, Archive Rat. Mech. Anal., to appear.
[20] F. Otto, *L^1–contraction and uniqueness for quasilinear elliptic–parabolic equations*, J. Differential Equations, 131 (1996), pp. 20–38.
[21] S. T. Rachev, *Probability metrics and the stability of stochastic models*, John Wiley, New York, 1991.
[22] H. Risken, *The Fokker-Planck equation: Methods of solution and applications*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 1989.
[23] Z. Schuss, *Singular perturbation methods in stochastic differential equations of mathematical physics*, SIAM Rev., 22 (1980), pp. 119–155.
[24] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wardsworth & Brooks/Cole, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 18–29, January 1998

002

STABILITY OF A RELAXATION MODEL WITH A NONCONVEX FLUX*

HAILIANG LIU†, JINGHUA WANG‡, AND TONG YANG§

Abstract. In this paper, we study the nonlinear stability of travelling wave solutions with shock profile for a relaxation model with a nonconvex flux, which is proposed by Jin and Xin [*Comm. Pure Appl. Math.*, 48 (1995), pp. 555–563] to approximate an original hyperbolic system numerically under the subcharacteristic condition introduced by T. P. Liu [*Comm. Math. Phys.*, 108 (1987), pp. 153–175]. The travelling wave solutions with strong shock profile are shown to be asymptotically stable under small disturbances with integral zero using an elementary but technical energy method. Proofs involve detailed study of the error equation for disturbances using the same weight function introduced in [*Comm. Math. Phys.*, 165 (1994), pp. 83–96].

Key words. relaxation model, stability, travelling wave

AMS subject classifications. Primary, 39A11; Secondary, 35L65

PII. S003614109629903X

1. Introduction.

Relaxation occurs when the underlying material is in nonequilibrium and usually takes the form of hyperbolic conservation laws with source terms. Relaxation is often stiff when the relaxation rate is much shorter than the scales of other physical quantities.

The relaxation limit for nonlinear systems of the following form was first studied by Liu [4]:

$$\begin{cases} \partial_t u + \partial_x f(u,v) &= 0, \\ \partial_t v + \partial_x g(u,v) &= \frac{v_*(u)-v}{\tau(u)}, \end{cases} \tag{1.0}$$

provided that the travelling waves are weak and $f(u, v_*(u))$ is a convex function. And the subcharacteristic condition for stability was formulated in [4]. The dissipative entropy condition was formulated for general nonlinear relaxation systems later by Chen, Levermore, and Liu [1].

Recently, a class of relaxation models were proposed by Jin and Xin [10] to approximate the original conservation laws numerically. The special structure of these relaxation systems enables one to solve them numerically with underresolved stable discretization without using either Riemann solvers spatially or nonlinear systems of algebraic equations solvers temporally.

In this paper, we study the following relaxation model introduced in [10]:

$$\begin{cases} u_t + v_x = 0, \quad x \in R^1, \\ v_t + a u_x = -\frac{1}{\varepsilon}(v - f(u)), \end{cases} \tag{1.1}$$

*Received by the editors February 16, 1996; accepted for publication (in revised form) September 5, 1996. The research of the authors was supported in part by the Strategic Research grant 700416 of City University of Hong Kong. Research by the first two authors was also supported in part by the National Natural Science Foundation of China.

http://www.siam.org/journals/sima/29-1/29903.html

†Department of Mathematics, Henan Normal University, Xinxiang, 453002, P.R. China.

‡Institute of Systems Science, Academia Sinica, Beijing 100080, P.R. China (jwang@iss06.iss.ac.cn).

§Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (matyang@sobolev.cityu.hk).

with the initial data

$$(u,v)(x,0) = (u_0,v_0)(x) \to (u_\pm, v_\pm) \ \ as x \to \pm\infty, \quad v_\pm = f(u_\pm), \tag{1.2}$$

where a is a positive constant satisfying

$$-\sqrt{a} < f'(u) < \sqrt{a} \tag{1.3}$$

for all u under consideration. (1.3) is the *subcharacteristic condition* introduced by Liu [4]. We will show that the travelling wave solutions are stable as $\epsilon \to 0$.

In the relaxation limit, $\varepsilon \to 0^+$, the leading order of the relaxation system (1.1) is

$$\begin{aligned} v &= f(u), \\ u_t + f(u)_x &= 0. \end{aligned} \tag{1.4}$$

In fact, (1.1) was the prototype of the relaxation model introduced in [10] to solve (1.4) using a local relaxation approximation.

Using Chapman–Enskog expansion [4], the first-order approximation to (1.1) is

$$\begin{aligned} v &= f(u) - \varepsilon(a - f'(u)^2)u_x, \\ u_t + f(u)_x &= \varepsilon((a - f'(u)^2)u_x)_x. \end{aligned} \tag{1.5}$$

Since (1.5) is dissipative provided that condition (1.3) is satisfied, then similar to the diffusion, the relaxation term has smoothing and dissipative effects for the hyperbolic conservation laws. The stability of the viscous travelling waves with nonconvex flux was investigated by many authors, cf. [2], [5], [7], [8], etc. Using a weight function introduced in [7], we study the stability of strong travelling waves for the relaxation model (1.1) with a nonconvex flux. The behavior of solutions as $\varepsilon \to 0$ when subcharacteristic condition is violated was investigated by R. Leveque and J. Wang [3] under the assumption that the relaxation term is linear.

Under the scaling $(x,t) \to (\varepsilon x, \varepsilon t)$, equation (1.1) becomes

$$\begin{cases} u_t + v_x = 0, \quad x \in R^1, \\ v_t + a u_x = f(u) - v. \end{cases} \tag{1.6}$$

The behavior of the solution (u,v) of (1.1) and (1.2) at any fixed time t as $\varepsilon \to 0^+$ is equivalent to the long time behavior of (u,v) of (1.6) as $t \to \infty$.

In section 2, we will show that there exist travelling wave solutions with shock profile for (1.6), i.e.,

$$(u,v)(x,t) = (U,V)(x-st) \equiv (U,V)(z), \quad (U,V)(z) \to (u_\pm, v_\pm) \ \text{ as } \ z \to \pm\infty,$$

if the shock speed s lies between $-\sqrt{a}$ and $\sqrt{a}$ and (u_-, u_+) is an admissible shock of (1.4), that is, the constants $u_\pm$ and s (shock speed) satisfy the Rankine–Hugoniot condition

$$-s(u_+ - u_-) + f(u_+) - f(u_-) = 0 \tag{1.7}$$

and the entropy condition

$$Q(u) \equiv f(u) - f(u_\pm) - s(u - u_\pm) \begin{cases} < 0 & \text{for } \ u_+ < u < u_-, \\ > 0 & \text{for } \ u_- < u < u_+. \end{cases} \tag{1.8}$$

Note that the U component of a travelling wave solution of (1.6) is a travelling wave solution of the viscous conservation law

$$u_t + f(u)_x = \mu u_{xx} \tag{1.9}$$

with $\mu = a - s^2$. This also gives another justification of the dynamic subcharacteristic condition $s^2 < a$ [4].

The purpose of this paper is to show the stability of the strong travelling wave satisfying $s^2 < a$ for any nonconvex flux f which satisfies the entropy condition (1.7) and (1.8); our result also gives a justification of relaxation schemes introduced in [4] for the case of scalar nonconvex conservation laws.

Notation. Hereafter, C denotes a generic positive constant. L^2 denotes the space of square integrable functions on R with the norm

$$||f|| = \left(\int_R |f|^2 dx \right)^{1/2}.$$

Without any ambiguity, the integral region R will be omitted. $H^j (j > 0)$ denotes the usual jth-order Sobolev space with the norm

$$||f||_{H_j} = ||f||_j = \left(\sum_{k=0}^{j} ||\partial_x^k f||^2 \right)^{1/2}.$$

For a weight function $w > 0$, L^2_w denotes the space of measurable functions f satisfying $\sqrt{w} f \in L^2$ with the norm

$$|f|_w = \left(\int w(x) |f(x)|^2 dx \right)^{1/2}.$$

2. Preliminaries and theorem. We first state the existence of the travelling wave solution with shock profile for the system (1.6). Substituting

$$(u, v)(x, t) = (U, V)(z), \quad z = x - st,$$

into (1.6), we have

$$\begin{cases} -sU_z + V_z = 0, \\ -sV_z + aU_z = f(U) - V, \end{cases} \tag{2.1}$$

hence

$$(a - s^2) U_z = f(U) - V. \tag{2.2}$$

Integrating the first equation of (2.1) over $(\pm\infty, z)$ and using $(U, V)(\pm\infty) = (u_\pm, v_\pm)$ and $v_\pm = f(u_\pm)$ yields

$$-sU + V = -su_\pm + v_\pm = -su_\pm + f(u_\pm). \tag{2.3}$$

Combining (2.2) and (2.3), we obtain

$$U_z = \frac{Q(U)}{a - s^2}, \tag{2.4}$$

where $Q(U) \equiv f(U) - f(u_\pm) - s(U - u_\pm)$ and

$$s = \frac{v_+ - v_-}{u_+ - u_-} = \frac{f(u_+) - f(u_-)}{u_+ - u_-}.$$

Since (2.4) is a scalar ordinary differential equation of U, the trajectories satisfying boundary conditions $U(\pm\infty) = u_\pm$ necessarily connect adjacent equilibria u_- and u_+. It is easy to check that there is a trajectory from u_- to u_+ if and only if condition $(u_+ - u_-)\frac{Q(U)}{a-s^2} > 0$ that holds for u lies strictly between u_+ and u_-. By virtue of $s^2 < a$, this implies

$$Q(u)(u_+ - u_-) > 0$$

for u that lies strictly between u_+ and u_-, i.e., if and only if

$$u = \begin{cases} u_-, x - st < 0, \\ u_+, x - st > 0 \end{cases}$$

is an admissible shock for (1.4).

Without loss of generality, we study only the following case:

$$u_+ < u_- \quad \text{and} \quad U_z < 0. \tag{2.5}$$

Then the ordinary differential equation (2.4) with boundary condition $U(\pm\infty) = u_\pm$ has a unique smooth solution. Moreover, if $f'(u_+) < s < f'(u_-)$ or $Q'(u_\pm) \neq 0$, then $Q(U) \sim -|U-u_\pm|$ as $U \to u_\pm$. Hence $|(U-u_\pm, V-v_\pm)(z)| \sim \exp(-c_\pm|z|)$ as $z \to \pm\infty$ for some constants $c_\pm > 0$. While if $s = f'(u_+)$ or $Q'(u_+) = 0$, $|(U-u_+, V-v_+)(z)| \sim z^{-\frac{1}{k_+}}$ as $z \to +\infty$ provided $Q(U) \sim -|U-u_+|^{1+k_+}$ for $k_+ > 0$. Note $k_+ = n$ if $Q'(u_+) = \cdots = Q^{(n)}(u_+) = 0$ and $Q^{(n+1)}(u_+) \neq 0$.

Thus we have the existence of travelling wave solutions.

LEMMA 2.1. *Assume that $Q(U) < 0$ for $U \in (u_+, u_-)$, $s = \frac{f(u_+)-f(u_-)}{u_+-u_-}$, $v_\pm = f(u_\pm)$, and $|Q(U)| \sim |U-u_+|^{1+k_+}$ as $U \to u_+$ with $k_+ \geq 0$. Then there exists a travelling wave solution $(U,V)(x-st)$ of (1.1) with $(U,V)(\pm\infty) = (u_\pm, v_\pm)$, which is unique up to a shift and the speed satisfies*

$$s^2 < a. \tag{2.6}$$

Moreover, it holds as $z \to \pm\infty$

$$|(U-u_\pm, V-v_\pm)(z)| \sim \exp(-c_\pm|z|) \quad \text{if } f'(u_+) < s < f'(u_-);$$

$$|(U-u_+, V-v_+)(z)| \sim z^{-\frac{1}{k_+}} \quad \text{if } s = f'(u_+).$$

For the initial disturbance, without loss of generality, we assume

$$\int_{-\infty}^{+\infty} (u_0 - U)(x)dx = 0. \tag{2.7}$$

For a pair of travelling wave solutions given by Lemma 2.1, we let

$$(\phi_0, \psi_0)(x) = \left(\int_{-\infty}^{x} (u_0 - U)(y)dy, (v_0 - V)(x) \right). \tag{2.8}$$

Our goal is to show that the solution $(u,v)(x,t)$ of (1.6), (1.2) will approach the travelling wave solution $(U,V)(x-st)$ as $t \to \infty$; the main theorem is as follows.

THEOREM 2.2 (stability). *Suppose that (1.7)–(1.8) hold and $f'(u)^2 < a$, where $a > 0$ is a suitably large constant and $f(u)$ is a smooth function. Let $(U,V)(x-st)$ be a travelling wave solution determined by (2.7) with speed $s^2 < a$, and assume that*

$u_0 - U$ is integrable on R and $\phi_0 \in H^3, \psi_0 \in H^2$. Then there exists a constant $\varepsilon_0 > 0$ independent of $(u_\pm, v_\pm)$ such that if

$$N(0) \equiv ||u_0 - U, v_0 - V||_2 + ||\phi_0, \psi_0|| < \varepsilon_0,$$

the initial value problem (1.6), (1.2) *has a unique global solution $(u, v)(x, t)$ satisfying*

$$(u - U, v - V) \in C^0(0, \infty; H^2) \cap L^2(0, \infty; H^2).$$

Furthermore, the solution satisfies

$$\sup_{x \in R} |(u, v)(x, t) - (U, V)(x - st)| \to 0 \quad \text{as} \quad t \to +\infty. \tag{2.9}$$

3. Reformulation of the problem. The proof of Theorem 2.2 is based on L^2 energy estimates. We first rewrite the problem (1.6), (1.2) using the moving coordinate $z = x - st$. Under the assumption of (2.7), we will look for a solution of the following form:

$$(u, v)(x, t) = (U, V)(z) + (\phi_z, \psi)(z, t), \tag{3.1}$$

where (ϕ, ψ) is in some space of integrable functions which will be defined later.

We substitute (3.1) into (1.6), by virtue of (2.1), and integrate the first equation once with respect to z; the perturbation (ϕ, ψ) satisfies

$$\begin{cases} \phi_t - s\phi_z + \psi = 0, \\ \psi_t - s\psi_z + a\phi_{zz} = f(U + \phi_z) - f(U) - \psi. \end{cases} \tag{3.2}$$

The first equation of (3.2) gives

$$\psi = -(\phi_t - s\phi_z). \tag{3.3}$$

Substituting (3.3) into the second equation of (3.2), we get a closed equation for ϕ:

$$L(\phi) \equiv (\phi_t - s\phi_z)_t - s(\phi_t - s\phi_z)_z - a\phi_{zz} + \phi_t + \lambda\phi_z = -F(U, \phi_z), \tag{3.4}$$

where $F(U, \phi_z) = f(U + \phi_z) - f(U) - f'(U)\phi_z = O(1)(\phi_z^2)$ is a higher order term and $\lambda = Q'(U) = f'(U) - s$.

The corresponding initial data for (3.4) becomes

$$\phi(z, 0) = \phi_0(z), \;\; \phi_t(z, 0) = s\phi_0'(z) - \psi_0 = \phi_1(z). \tag{3.5}$$

The asymptotic stability of the profile (U, V) means that the perturbation (ϕ_z, ψ) decays to zero as $t \to \infty$. The left-hand side of (3.4) contains a first-order term with speed λ which plays the essential role of governing the large-time behavior of the solution.

Now, we introduce the solution space of the problem (3.4), (3.5) as follows:

$$X(0, T) = \{\phi(z, t) : \;\; \phi \in C^0([0, T); H^3) \cap C^1(0, T; H^2), \;\; (\phi_z, \phi_t) \in L^2(0, T; H^2)\},$$

with $0 < T \leq +\infty$. By virtue of (3.3), we have

$$\psi \in C^0([0, T); H^2) \cap L^2(0, T; H^2).$$

By the Sobolev embedding theorem, if we let

$$N(t) = \sup_{0 \leq \tau \leq t} \{||\phi(\tau)||_3 + ||\phi_t(\tau)||_2\},$$

then

$$\sup_{z\in R}\{|\phi|,|\phi_z|,|\phi_{zz},|\phi_t|,|\phi_{tz}|\}\le CN(t). \tag{3.6}$$

Thus Theorem 2.2 is a consequence of the following theorem.

THEOREM 3.1. *Under the conditions of Theorem* 2.2, *there exists a positive constant* δ_1 *such that if*

$$N(0)=||\phi_0||_3+||\phi_1||_2\le\delta_1, \tag{3.7}$$

then the problem (3.4), (3.5) *has a unique global solution* $\phi\in X(0,+\infty)$ *satisfying*

$$||\phi(t)||_3^2+||\phi_t||_2^2+\int_0^t||(\phi_t,\phi_z)(\tau)||_2^2d\tau\le CN(0)^2 \tag{3.8}$$

for $t\in[0,+\infty)$. *Furthermore,*

$$\sup_{z\in R}|(\phi_z,\phi_t)(z,t)|\to 0\quad as\ \ t\to\infty. \tag{3.9}$$

For the solution ϕ in the above theorem, we define (ϕ,ψ) by (3.3). Then it becomes a global solution of the problem (3.2) with $(\phi,\psi)(z,0)=(\phi_0,\psi_0)(z)$, and consequently we have the desired solution of the problem (1.6), (1.2) through the relation (3.1). On the other hand the solution of (1.6) is unique in the space $C^0(0,T;H^2)$. Therefore Theorem 2.2 follows from Theorem 3.1. Global existence for ϕ will be derived from the following local existence theorem for ϕ combined with an a priori estimate. (3.8) gives

$$||\phi_t,\phi_z||_1^2\to 0\quad \text{as}\ \ t\to\infty, \tag{3.10}$$

from which we have

$$\begin{aligned}\phi_t^2+\phi_z^2 &= \int_{-\infty}^{z}(2\phi_t\phi_{tz}+2\phi_z\phi_{zz})(y,t)dy\\ &\le \left(\int_{-\infty}^{+\infty}(\phi_t^2+\phi_z^2)dy\right)^{1/2}\left(\int_{-\infty}^{+\infty}(\phi_{tz}^2+\phi_{zz}^2)dy\right)^{1/2}\to 0,\ \ \text{as}\ \ t\to\infty.\end{aligned}$$

PROPOSITION 3.2 (local existence). *For any* $\delta_0>0$, *there exists a positive constant* T_0 *depending on* δ_0 *such that if* $\phi_0\in H^3$ *and* $\phi_1\in H^2$, *with* $N(0)<\delta_0/2$, *then the problem* (3.4), (3.5) *has a unique solution* $\phi\in X(0,T_0)$ *satisfying*

$$N(t)<2N(0) \tag{3.11}$$

for any $0\le t\le T_0$.

PROPOSITION 3.3 (a priori estimate). *Let* $\phi\in X(0,T)$ *be a solution for a positive constant* T; *then there exists a positive constant* δ_2 *independent of* T *such that if*

$$N(t)<\delta_2,\ \ t\in[0,T],$$

then ϕ *satisfies* (3.8) *for any* $0\le t\le T$.

Proposition 3.2 can be proved in the standard way, so we omit the proof; cf. [9]. To prove Proposition 3.3 is our main task in the following section.

Here we prove Theorem 3.1 by the continuation arguments based on Proposition 3.2 and Proposition 3.3.

Proof of Theorem 3.1. By the definition of $N(t)$, we have

$$N(t)^2 \le 2 \sup_{0\le\tau\le t} [||\phi(\tau)||_3^2 + ||\phi_t(\tau)||_2^2]. \tag{3.12}$$

Then the inequality (3.8) implies

$$N(t) < \sqrt{2C}N(0). \tag{3.13}$$

Choose δ_1 such that $\delta_1 = \min\{\frac{\delta_2}{2}, \frac{\delta_2}{2\sqrt{2C}}\}$; then the local solution of (3.4) can be continued globally in time, provided the smallness condition $N(0) \le \delta_1$ is satisfied. In fact we have $N(0) < \delta_1 \le \delta_2/2$. Therefore, by Proposition 3.2, there is a positive constant $T_0 = T_0(\delta_2)$ such that a solution exists on $[0, T_0]$ and satisfies $N(t) < 2N(0) \le \delta_2$ for $t \in [0, T_0]$.

Hence we can apply Proposition 3.3 with $T = T_0$ and get the estimate (3.8), that is, $N(t) \le \sqrt{2C}N(0) \le \frac{\delta_2}{2}$ for $t \in [0, T_0]$. Then we apply Proposition 3.2 by taking $t = T_0$ as the new initial time. We have a solution on $[T_0, 2T_0]$ with the estimate $N(t) \le 2N(T_0) \le \delta_2$ for $t \in [T_0, 2T_0]$. Therefore $N(t) \le \delta_2$ holds on $[0, 2T_0]$. Hence this again gives the estimate (3.8) for $t \in [0, 2T_0]$. In the same way we can extend the solution to the interval $[0, nT_0]$ successively, $n = 1, 2, \ldots$, and get a global solution ϕ. This completes the proof of Theorem 3.1. ∎

4. Energy estimates. In this section, we will complete the proof of our stability theorem. We establish the basic L^2 estimate as follows.

LEMMA 4.1. *There are positive constants C such that if*

$$-\sqrt{a} < f'(u) < \sqrt{a}, \quad u \in (u_+, u_-),$$

and a is sufficiently large, then

$$\begin{aligned} &||\phi(t)||_1^2 + ||\phi_t(t)||^2 + \int_0^t ||(\phi_t, \phi_z)(\tau)||^2 d\tau + \int_0^t \int_R |U_z|\phi^2 dz d\tau \\ &\le C\{||\phi_0||_1^2 + ||\phi_1||^2 + \int_0^t \int_R |F|(|\phi| + |(\phi_t, \phi_z)|) dz d\tau\} \end{aligned} \tag{4.1}$$

holds for $t \in [0, T]$.

Proof. When f is a nonconvex function, the standard energy method used in [6] does not work for our problem (3.4), (3.5). To overcome this difficulty, we use a weight function $w(U)$ introduced in [7] depending on the shock profile U.

First, by multiplying (3.4) by $2w(U)\phi$, we obtain

$$2w(U)\phi \cdot L(\phi) = -2Fw(U)\phi. \tag{4.2}$$

The left-hand side of (4.2) can be reduced to

(4.3)

$$\begin{aligned} &2[(\phi_t - s\phi_z)_t - s(\phi_t - s\phi_z)_z - a\phi_{zz}]w\phi + 2(\phi_t + \lambda\phi_z)w\phi \\ &= [2w\phi(\phi_t - s\phi_z)]_t - 2w\phi_t(\phi_t - s\phi_z) - 2s[w\phi(\phi_t - s\phi_z)]_z \\ &\quad + 2sw_z\phi(\phi_t - s\phi_z) + 2sw\phi_z(\phi_t - s\phi_z) - 2a(w\phi\phi_z)_z + 2aw\phi_z^2 \\ &\quad + (aw_z\phi^2)_z - aw_{zz}\phi^2 + (w\phi^2)_t + (\lambda w\phi^2)_z - \phi^2(\lambda w)_z \\ &= [w\phi^2 + 2w\phi(\phi_t - s\phi_z)]_t - 2w(\phi_t - s\phi_z)^2 + 2aw\phi_z^2 - aw_{zz}\phi^2 \\ &\quad - (\lambda w)_z\phi^2 + sw_z(\phi^2)_t - s^2\{w_z(\phi^2)\}_z + s^2w_{zz}\phi^2 \\ &\quad + \{-2sw\phi(\phi_t - s\phi_z) - 2aw\phi\phi_z + aw_z\phi^2 + \lambda w\phi^2\}_z \\ &= [w\phi^2 + 2w\phi(\phi_t - s\phi_z) + sw_z\phi^2]_t - 2w(\phi_t - s\phi_z)^2 + 2aw\phi_z^2 + A\phi^2 + \{\cdots\}_z; \end{aligned}$$

here $A = (s^2 - a)w_{zz} - (\lambda w)_z$, $\{\cdots\}_z$ denotes the terms which will disappear after integration with respect to $z \in R$.

Secondly, we calculate

$$2(\phi_t - s\phi_z)w \cdot L(\phi) = -2F(\phi_t - s\phi_z)w. \tag{4.4}$$

The left-hand side of (4.4) is

$$\begin{aligned}
&2[(\phi_t - s\phi_z)_t - s(\phi_t - s\phi_z)_z - a\phi_{zz}]w(\phi_t - s\phi_z)\\
&\quad +2w(\phi_t - s\phi_z)(\phi_t - s\phi_z + f'(U)\phi_z)\\
&= [w(\phi_t - s\phi_z)^2]_t - s[w(\phi_t - s\phi_z)^2]_z + sw_z(\phi_t - s\phi_z)^2\\
&\quad -2a[w\phi_z(\phi_t - s\phi_z)]_z + 2aw_z\phi_z(\phi_t - s\phi_z) + 2aw\phi_z(\phi_t - s\phi_z)_z\\
&\quad +2w(\phi_t - s\phi_z)^2 + 2wf'(U)\phi_z(\phi_t - s\phi_z)\\
&= [w(\phi_t - s\phi_z)^2]_t + (2w + sw_z)(\phi_t - s\phi_z)^2 + 2aw_z\phi_z(\phi_t - s\phi_z)\\
&\quad +2wf'(U)\phi_z(\phi_t - s\phi_z) + [aw\phi_z^2]_t - [asw\phi_z^2]_z + asw_z\phi_z^2\\
&\quad -[sw(\phi_t - s\phi_z)^2 + 2aw\phi_z(\phi_t - s\phi_z)]_z\\
&= [aw\phi_z^2 + w(\phi_t - s\phi_z)^2]_t + (2w + sw_z)(\phi_t - s\phi_z)^2\\
&\quad +saw_z\phi_z^2 + 2f'(U)w\phi_z(\phi_t - s\phi_z) + 2aw_z\phi_z(\phi_t - s\phi_z)\\
&\quad -[sw(\phi_t - s\phi_z)^2 + 2aw\phi_z(\phi_t - s\phi_z) + asw\phi_z^2]_z.
\end{aligned} \tag{4.5}$$

Hence, the combination (4.2) $\times\mu+$ (4.4) with a positive constant μ yields

$$\begin{aligned}
&\{E_1(\phi, (\phi_t - s\phi_z)) + E_3(\phi_z)\}_t + E_2(\phi_z, (\phi_t - s\phi_z)) + E_4(\phi) + \{\cdots\}_z\\
&\quad = -2Fw\{\mu\phi + (\phi_t - s\phi_z)\},
\end{aligned} \tag{4.6}$$

where

$$\begin{aligned}
E_1(\phi, (\phi_t - s\phi_z)) &= w(\phi_t - s\phi_z)^2 + 2\mu w\phi(\phi_t - s\phi_z) + \mu(w + sw_z)\phi^2,\\
E_3(\phi_z) &= aw\phi_z^2,\\
E_2(\phi_z, (\phi_t - s\phi_z)) &= (2w + sw_z - 2\mu w)(\phi_t - s\phi_z)^2\\
&\quad +2(f'(U)w + aw_z)\phi_z(\phi_t - s\phi_z) + a(2\mu w + sw_z)\phi_z^2,\\
E_4(\phi) &= \mu A\phi^2.
\end{aligned} \tag{4.7}$$

Due to $(a - s^2)U_z = Q(U)$ and $w = w(U)$, we have

$$\begin{aligned}
A &= -\{(a - s^2)w'(U)U_z + \lambda w\}_z\\
&= -\{w'(U)Q(U) + Q'(U)w\}_z\\
&= -\{wQ\}''U_z.
\end{aligned} \tag{4.8}$$

The monotonicity of the shock profile U implies $U_z < 0$; thus we need to choose $w \in C^2[u_+, u_-]$ such that

$$(wQ)'' \geq \nu > 0. \tag{4.9}$$

On the other hand, we need to choose a constant $\mu > 0$ and w such that the discriminants of E_i $(i = 1, 2)$ are negative; that is, the inequalities

$$\sup_j D_j < 0, \quad j = 1, 2, \tag{4.10}$$

hold uniformly in $(u_\pm, v_\pm)$, where D_j is the discrimant of the functions $E_j (j = 1, 2)$, respectively.

$$D_1 = 4\mu w[(\mu - 1)w - sw_z],$$

$$D_2 = 4\{(f'w + aw_z)^2 - a(2\mu w + sw_z)(2w + sw_z - 2\mu w)\},$$

and $2\mu w + sw_z > 0$. For this choice of μ and w, there exist positive constants c and C such that

$$\left\{ \begin{array}{lllll} c\{\phi^2 + (\phi_t - s\phi_z)^2\} & \le & E_1 & \le & C\{\phi^2 + (\phi_t - s\phi_z)^2\}, \\ c\{\phi_z^2 + (\phi_t - s\phi_z)^2\} & \le & E_2. & & \end{array} \right. \tag{4.11}$$

On the other hand, (4.8) and $a > 0$ gives

$$\left\{ \begin{array}{lllll} 0 & \le & E_3 & = & aw\phi_z^2, \\ & & E_4 & \ge & \mu\nu|U_z|\phi^2 \ge 0. \end{array} \right. \tag{4.12}$$

Thus the equality (4.6) together with the estimates (4.11)–(4.12) give the desired estimate (4.1) after integration with respect to t and z.

It remains to check conditions (4.8)–(4.10). First we choose the weight function $w(U)$ introduced in [7] for the scalar viscous conservation laws with nonconvex flux

$$w(U) = \frac{(U - u_+)(U - u_-)}{Q(U)}. \tag{4.13}$$

Then $w \in C^2[u_+, u_-]$ and (4.8) holds, i.e., $(wQ)'' = \nu = 2$. Furthermore, choosing $\mu = \frac{1}{2}$, the two inequalities in (4.10) are equivalent to

$$1 + 2s\frac{w_z}{w} > 0, \tag{4.14}$$

$$\left(f' + a\frac{w_z}{w}\right)^2 < a\left(1 + s\frac{w_z}{w}\right)^2, \tag{4.15}$$

since

$$\frac{w_z}{w} = \frac{w'}{w}\frac{Q}{a - s^2} = \frac{O(1)}{a - s^2},$$

which is small provided a is suitably large. This fact, together with $f'^2 < a$, gives us (4.14) and (4.15); thus conditions (4.8) and (4.10) are satisfied. This completes the proof of Lemma 4.1. □

Next we estimate the higher derivatives of ϕ, multiplying the derivative of (3.4) with respect to z by ϕ_z and $(\phi_t - s\phi_z)_z$, respectively; we have

$$2\partial_z L(\phi) \cdot \phi_z = -2F_z\phi_z,$$

$$2\partial_z L(\phi) \cdot (\phi_t - s\phi_z)_z = -2F_z(\phi_t - s\phi_z)_z.$$

Letting $\phi_z = \Phi$, then

$$\begin{array}{rcl} \partial_z L(\phi) & = & (\phi_{zt} - s\phi_{zz})_t - s(\phi_{zt} - s\phi_{zz})_z - a\phi_{zzz} + \phi_{zt} + \lambda\phi_{zz} + \lambda_z\phi_z \\ & = & L(\phi_z) + \lambda_z\phi_z = L(\Phi) + \lambda_z\Phi. \end{array} \tag{4.16}$$

By a similar argument to obtain (4.3) and (4.5) with $w = 1$, we have

$$\begin{aligned}[\Phi^2 + 2\Phi(\Phi_t - s\Phi_z)]_t + 2a\Phi_z^2 - 2(\Phi_t - s\Phi_z)^2 - \lambda_z\Phi^2 + 2\lambda_z\Phi^2 + \{\cdots\}_z \\ = -2F_z\Phi,\end{aligned} \tag{4.17}$$

and

$$\begin{aligned}[(\Phi_t - s\Phi_z)^2 + a\Phi_z^2]_t + 2(\Phi_t - s\Phi_z)^2 + 2f'(U)\Phi_z(\Phi_t - s\Phi_z) \\ +2\lambda_z\Phi(\Phi_t - s\Phi_z) + \{\cdots\}_z \\ = -2F_z(\Phi_t - s\Phi_z).\end{aligned} \tag{4.18}$$

The combination (4.17)$\times\frac{1}{2}$+(4.18) yields

$$\begin{aligned}\{E_1(\Phi, (\Phi_t - s\Phi_z)) + E_2(\Phi_z)\}_t + E_3(\Phi_z, (\Phi_t - s\Phi_z)) + G + \{\cdots\}_z \\ = -F_z\{\Phi + 2(\Phi_t - s\Phi_z)\},\end{aligned} \tag{4.19}$$

where

$$\begin{aligned} G &= \tfrac{\lambda_z}{2}\Phi^2 + 2\lambda_z\Phi(\Phi_t - s\Phi_z), \\ E_1(\Phi, (\Phi_t - s\Phi_z)) &= (\Phi_t - s\Phi_z)^2 + \Phi(\Phi_t - s\Phi_z) + \tfrac{1}{2}\Phi^2, \\ E_2(\Phi_z) &= a\Phi_z^2, \\ E_3(\Phi_z, (\Phi_t - s\Phi_z)) &= (\Phi_t - s\Phi_z)^2 + 2f'(U)\Phi_z(\Phi_t - s\Phi_z) + a\Phi_z^2. \end{aligned} \tag{4.20}$$

After integration with respect to t and z, (4.19) together with (4.20) gives the following estimate:

(4.21)

$$\begin{aligned}&||\Phi(t)||_1^2 + ||\Phi_t(t)||^2 + \int_0^t ||(\Phi_t, \Phi_z)(\tau)||^2 d\tau \\ &\quad \le C\left\{||\Phi_0||_1^2 + ||\Phi_1||^2 + \int_0^t \int |G| dz d\tau + \int_0^t \int_R |F_z|(|\Phi| + |(\Phi_t, \Phi_z)|) dz d\tau\right\};\end{aligned}$$

here $\Phi_0 = \phi_0'$ and $\Phi_1 = \phi_1'$.

Using the estimate (4.1), we obtain

(4.22)

$$\begin{aligned}\int_0^t \int |G| dz d\tau &\le \int_0^t \int \left[\frac{|\lambda_z|}{2}\Phi^2 + 2|\lambda_z|^2\Phi^2 + \frac{1}{2}\Phi_t^2 + 2s^2|\lambda_z|^2\Phi^2 + \frac{1}{2}\Phi_z^2\right] dz d\tau \\ &\le \frac{1}{2}\int_0^t ||(\Phi_t, \Phi_z)(\tau)||^2 d\tau + C\int_0^t \int \Phi^2 dz d\tau \\ &\le \frac{1}{2}\int_0^t ||(\Phi_t, \Phi_z)(\tau)||^2 d\tau \\ &\quad + C\left\{||\phi_0||_1^2 + ||\phi_1||^2 + \int_0^t \int |F|(|\phi| + |(\phi_t, \phi_z)|) dz d\tau\right\},\end{aligned}$$

where we have used Lemma 4.1 and the boundness of $|\lambda_z|$.

Substituting (4.22) into (4.21) and replacing Φ by $\partial_z\phi$, we have the following lemma.

LEMMA 4.2. *There are positive constants C such that if*

$$-\sqrt{a} < f'(u) < \sqrt{a} \ \text{ for } \ u \in (u_+, u_-),$$

then

$$
\begin{aligned}
&||\partial_z\phi(t)||_1^2+||\partial_z\phi_t||^2+\frac{1}{2}\int_0^t||(\partial_z\phi_t,\partial_z\phi_z)(\tau)||^2d\tau\\
&\quad\le C\left\{||\phi_0||_2^2+||\phi_1||_1^2+\int_0^t\int|F_z|(|\partial_z\phi|+|(\partial_z\phi_t,\partial_z\phi_z)|)dzd\tau\right.\\
&\quad\left.+\int_0^t\int|F|(|\phi|+|(\phi_t,\phi_z)|)dzd\tau\right\}
\end{aligned}
\tag{4.23}
$$

holds for $t\in[0,T]$.

Next we calculate the equality

$$\partial_z^2\phi\cdot\partial_z^2L(\phi)+2\partial_z^2(\phi_t-s\phi_z)\cdot\partial_z^2L(\phi)=-\partial_z^2F\{\partial_z^2\phi+2\partial_z^2(\phi_t-s\phi_z)\}$$

in the same way as for the proof of Lemma 4.2; it is easy to get the following equality for $\Psi=\partial_z^2\phi$:

(4.24)

$$
\begin{aligned}
&\left[(\Psi_t-s\Psi_z)^2+a\Psi_z^2+\Psi(\Psi_t-s\Psi_z)+\frac{1}{2}\Psi^2\right]_t+(\Psi_t-s\Psi_z)^2+2f'(U)\Psi_z(\Psi_t-s\Psi_z)\\
&\quad+a\Psi_z^2+4\lambda_z\Psi(\Psi_t-s\Psi_z)+\frac{3}{2}\lambda_z\Psi^2+\lambda_{zz}\Psi\phi_z+2\lambda_{zz}\phi_z(\Psi_t-s\Psi_z)+\{\cdots\}_z\\
&=-F_{zz}[\Psi+2(\Psi_t-s\Psi_z)].
\end{aligned}
$$

Thus, noting $\Psi=\phi_{zz}$, we have from (4.24) that

(4.25)

$$
\begin{aligned}
&||\partial_z^2\phi(t)||_1^2+||\partial_z^2\phi_t||^2+\frac{1}{3}\int_0^t||(\partial_z^2\phi_t,\partial_z^2\phi_z)(\tau)||^2d\tau-C\int_0^t\{||\partial_z^2\phi||^2+||\phi_z||^2\}d\tau\\
&\quad\le C\left\{||\phi_0||_3^2+||\phi_1||_2^2+\int_0^t\int|F_{zz}|(|\partial_z^2\phi|+|(\partial_z^2\phi_t,\partial_z^2\phi_z)|)dzd\tau\right\},
\end{aligned}
$$

where we have used the fact that $\lambda_z,\ \ \lambda_{zz}$ are smooth bounded functions and the Young inequality for the terms $4\lambda_z\Psi(\Psi_t-s\Psi_z)$ and $2\lambda_{zz}\phi_z(\Psi_t-s\Psi_z)$. Combining successively the estimates (4.1), (4.23), and (4.25), we have

(4.26)

$$
\begin{aligned}
&||\phi(t)||_3^2+||\phi_t(t)||_2^2+\int_0^t\int|\lambda_z|\phi^2dzd\tau+\int_0^t||(\phi_t,\phi_z)||_2^2d\tau\\
&\quad\le C\left\{||\phi_0||_3^2+||\phi_1||_2^2+\int_0^t\int\{|F|(|\phi|+|(\phi_t,\phi_z)|)+|F_z|(|\partial_z\phi|+|(\partial_z\phi_t,\partial_z\phi_z)|)\right.\\
&\qquad\left.+|F_{zz}|(|\partial_z^2\phi|+|(\partial_z^2\phi_t,\partial_z^2\phi_z)|)\}dzd\tau\right\}.
\end{aligned}
$$

Since $F=f(U+\phi_z)-f'(U)\phi_z-f(U)$, we have

$$|F|=O(1)(\phi_z^2),\quad |F_z|=O(1)(\phi_z^2+\phi_{zz}^2),$$

$$|F_{zz}|=O(1)(\phi_z^2+\phi_{zz}^2+|\phi_z\phi_{zzz}|).$$

By virtue of (3.6), the integral on the right-hand side of (4.26) is majored by

$$CN(t)\int_0^t||(\phi_t,\phi_z)||_2^2d\tau;$$

then we have

$$N^2(t) + \int_0^t ||(\phi_t, \phi_z)||_2^2 d\tau + \int_0^t \int |\lambda_z| \phi^2 dz d\tau \le N(0)^2 + CN(t) \int_0^t ||(\phi_t, \phi_z)||_2^2 d\tau.$$

Therefore, by assuming $N(T) \le \frac{1}{2C}$, we obtain the desired estimate

$$N^2(t) + \int_0^t ||(\phi_t, \phi_z)||_2^2 d\tau \le CN(0)^2 \quad \text{for} \quad t \in [0, T].$$

Thus the proof of Proposition 3.3 is completed. □

Remark. When $s = f'(u_+)$ or $s = f'(u_-)$, we need a weight of the order $\langle x \rangle = \sqrt{1 + x^2}$ as $x \to +\infty$ or $-\infty$ for a stability theorem. The stability analysis for ϕ in this case can be investigated similarly using the weighted function space

$$X_w(0, T) = \{\phi(z, t) : \quad \phi \in C^0([0, T); H^3 \cap L^2_{w(U)}) \cap C^1(0, T; H^2 \cap L^2_{w(U)}), \\ (\phi_z, \phi_t) \in L^2(0, T; H^2 \cap L^2_{w(U)})\},$$

where $w(U(z)) \sim \langle z \rangle$ as $z \to \pm\infty$ by virtue of Lemma 2.1 and the definition of $w(U)$ in (4.13).

Acknowledgment. The authors are grateful to Prof. M. Slemrod for stimulating discussions.

REFERENCES

[1] G.-Q. Chen, C. D. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1993), pp. 787–830.
[2] C. Jones, R. Gardner, and T. Kapitula, *Stability of travelling waves for non-convex scalar viscous conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 505–526.
[3] R. J. Leveque and J. Wang, *A linear hyperbolic system with stiff source terms*, in Proc. 4th Int'l. Conf. on Hyperbolic Problems, Taormina, Italy, 1992.
[4] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.
[5] H. L. Liu, *Asymptotic stability of shock profiles for non-convex convection-diffusion equation*, Appl. Math. Lett., 10 (1997), pp. 129–134.
[6] H. L. Liu and J. Wang, *Asymptotic stability of travelling wave solutions of a hyperbolic system with relaxation terms*, Beijing Math., 2 (1996), pp. 119–130.
[7] A. Matsumura and K. Nishihara, *Asymptotic stability of travelling waves of scalar viscous conservation laws with non-convex nonlinearity*, Comm. Math. Phys., 165 (1994), pp. 83–96.
[8] M. Mei, *Stability of shock profiles for non-convex scalar conservation laws*, Math. Models Method Appl. Sci., 5 (1995), pp. 279–296.
[9] T. N*ishida*, *Nonlinear Hyperbolic Equations and Related Topics in Fluid Dynamics*, Publ. Math. D'osay 78, 02 Dept. de Math., Paris-sud, France, 1978.
[10] S. Jin and Z. Xin, *The relaxing schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 555–563.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 30–47, January 1998

003

A CONDENSATION–EVAPORATION PROBLEM IN KINETIC THEORY*

L. ARKERYD† AND A. NOURI‡

Abstract. A linear Boltzmann model is used for studying a condensation–evaporation problem in a bounded domain. First the time asymptotic limit is derived, which solves the associated stationary problem. Then the Milne problem is discussed for the boundary layer. Finally a fluid approximation is obtained in the small mean free path limit with initial and boundary layers of zeroth order.

Key words. boundary layer, condensation, evaporation, hydrodynamic limit, initial layer, Milne problem, time asymptotics

AMS subject classification. 58G20

PII. S0036141096301737

Introduction. The kinetic description of a rarefied gas can be given through the Boltzmann equation for the density function $f(t,x,v)$ of particles with velocity v at position x and time t. A coarser theory consists of describing the gas as a continuous fluid with local density $\rho(t,x)$, velocity $u(t,x)$, and temperature $T(t,x)$ satisfying the Euler or Navier–Stokes equations. In the limit of small mean free path, the fluid dynamic equations may be derived from the Boltzmann equation through either a Hilbert or Chapman–Enskog expansion; see, e.g., [2, 8, 9, 12]. However, the fluid dynamic limits fail near shocks and for general indata near spatial or temporal boundaries.

Among the many studies of the boundary layer structure let us mention the following. In [3], the steady nonlinear Boltzmann equation for a gas with zero bulk velocity between two plates at two different temperatures is solved for a small mean free path, using a Chapman–Enskog expansion between the two plates. Here the fluid part of the solution contains Fourier's law for heat conduction which can be made to satisfy different temperature values at the two plates. This is why the boundary layer terms only need to be of first order with respect to the mean free path. An analogous study also including the initial layer is performed in [16] for the linear semiconductor case where further references in the field may also be found. For more results in the area see also [5, 10, 13, 19].

The present paper addresses the added presence of condensation–evaporation on the boundary. In this context a formal analysis and numerical computations are carried out in [17, 18] for a rarefied gas with varying temperatures and condensation–evaporation on the boundaries. On the basis of the linearized Boltzmann equation for hard sphere molecules, zeroth-order boundary layer terms are needed for solving the problem. Our paper considers the same problem for a rarefied solute in a solvent gas, and with varying temperatures on the boundary. The linear Boltzmann equation is used as a model for the solute. We prove that a fluid approximation in the interior together with initial and boundary layer structures are available to describe the solute

*Received by the editors April 10, 1996; accepted for publication (in revised form) November 12, 1996.

http://www.siam.org/journals/sima/29-1/30173.html

†Department of Mathematics, Chalmers University of Technology, S 41296 Gothenburg, Sweden.

‡Laboratoire J. A. Dieudonné, URA 168 du CNRS, UNSA, Parc Valrose, 06108 Nice Cédex 02, France (nouri@gaston.unice.fr).

gas. Here the fluid approximation is derived from the boundary layer analysis. Indeed, like [17, 18] this boundary layer structure requires zeroth-order terms with respect to the mean free path.

In the first section an existence and uniqueness result for the initial boundary value problem with given indata in a bounded region is recalled. We then determine the solution to the stationary boundary value problem from the time asymptotics of the initial boundary value solution. The approach is designed for prospective future use in the nonlinear case. For another approach to the nonlinear stationary problem see [1]. Section 2 is devoted to the solution of the Milne problem. For indepth discussions and bibliography see [4, 6]. Depending on the sign of the normal velocity of the solvent gas, two kinds of solutions are of interest for the following boundary layer analysis. In the last section we perform in the slab case a fluid approximation with respect to the mean free path by splitting the solution into a zeroth-order initial layer term together with a stationary boundary value contribution having a fluid part with zeroth-order boundary layer terms and a first order remainder term.

1. The initial boundary value problem and its time asymptotic behavior. The linear Boltzmann equation models the interaction between a solvent gas and a solute gas. The solute gas is rarefied enough so that collisions with itself are negligible in comparison with collisions with the solvent gas. Both gases are located in a bounded convex domain $\Omega \subset \mathbb{R}^3$. The distribution function $f(t,x,v)$ of the solute gas satisfies the linear Boltzmann equation

$$\partial_t f + v \cdot \nabla_x f = Q(f), \tag{1.1}$$

where

$$Q(f)(t,x,v) = \int B(\theta, w)(f'F'_* - fF_*)dv_* d\theta d\epsilon = Q^+(f) - \nu f.$$

Here

$$\begin{gathered} f' = f(t,x,v'), \quad F'_* = F(t,x,v'_*), \\ f = f(t,x,v), \quad F_* = F(t,x,v_*), \\ w = |v - v_*|, \quad v' = v - \frac{2}{1+\kappa}((v-v_*)\cdot e)e, \\ v'_* = v_* + \frac{2\kappa}{1+\kappa}((v-v_*)\cdot e)e, \quad e \in S^2. \end{gathered}$$

F is the solvent distribution function, assumed to be known, and κ is the ratio between the solute molecular mass m and the solvent molecular mass m_*.

Assuming that the collisions between the two gases are governed by a cut-off inverse power law interaction potential $U(\rho) = c\rho^{-k+1}, k > 2$ depending on the distance ρ of two colliding particles, the weight function B is $B(\theta, w) = w^\gamma b(\theta)$, $0 \le \theta < \frac{\pi}{2}$, $w > 0$ (cf. [7]), where $\gamma = \frac{k-5}{k-1}$ and b is a nonnegative L^1-function defined on $[0, \frac{\pi}{2}]$, with $\int_0^{\frac{\pi}{2}} b(\theta)d\theta > 0$. We assume hard interactions, i.e., $k > 5$ or $0 < \gamma < 1$. A principle of detailed balance only holds [14], when F is a Maxwellian,

$$F(v) = \left(\frac{2\pi T}{m_*}\right)^{-\frac{3}{2}} \exp\left(-m_* \frac{(v-U)^2}{2T}\right).$$

This is also assumed throughout the paper.

The collision frequency $\nu(v)$ is bounded from above and below by a positive multiple of $(1+|v|)^\gamma$. The choice of the bulk velocity $U=(u,0,0)\in\mathbb{R}^3$ in connection with the given boundary temperature follows from the boundary value problem for the solvent gas. The present study of the solute holds for any U and boundary temperature. The solute Maxwellian with the same bulk velocity U and temperature T is $M(v)=(\frac{2\pi T}{m})^{-\frac{3}{2}}\exp(-m\frac{(v-U)^2}{2T})$. It satisfies

$$F_*M=F'_*M'. \tag{1.2}$$

(1.1) is complemented with an initial condition

$$f(0,x,v)=f_i(x,v) \tag{1.3}$$

and given indata on the boundary

$$f(t,x,v)=f_b(x,v),\quad x\in\partial\Omega,\quad v\cdot n(x)>0. \tag{1.4}$$

Here $n(x)$ denotes the inward normal at x. Let $(\partial\Omega\times\mathbb{R}^3)^+$ and $(\partial\Omega\times\mathbb{R}^3)^-$ denote the sets of $(x,v)\in\partial\Omega\times\mathbb{R}^3$ such that $v\cdot n(x)>0$ and $v\cdot n(x)<0$, respectively.

For $\partial\Omega$ sufficiently smooth, say C^1, the existence and uniqueness approach of [15] can be used to prove the following theorem.

THEOREM 1.1. *If $(1+|v|)^\gamma f_i$ and $(1+|v|)^\gamma f_b$ belong to $L^1(\Omega\times\mathbb{R}^3)$ and $L^1_{v\cdot n(x)}((\partial\Omega\times\mathbb{R}^3)^+)$, respectively, then there exists a unique solution f of (1.1)–(1.3–1.4) with $f(t)(1+|v|)^\gamma\in L^1(\Omega\times\mathbb{R}^3)$ for $t>0$. Moreover, f is nonnegative whenever f_i and f_b are nonnegative.*

Let us next discuss the collisions and the collision operator in velocity space. The momentum and energy conservations imply

$$\begin{aligned} mv+m_*v_*&=mv'+m_*v'_*,\\ m|v|^2+m_*|v_*|^2&=m|v'|^2+m_*|v'_*|^2. \end{aligned}$$

A transformation to the equal mass situation $m=m_*$ is given by

$$\tilde v=v-\frac{\alpha}{2}(v-v_*),\quad \tilde v_*=v_*-\frac{\alpha}{2}(v-v_*),$$

where $\alpha=\frac{m_*-m}{m_*+m}$. Hence

$$\begin{aligned} \tilde v+\tilde v_*&=\tilde v'+\tilde v'_*,\\ |\tilde v|^2+|\tilde v_*|^2&=|\tilde v'|^2+|\tilde v'_*|^2. \end{aligned}$$

Denote by $\tilde f$, $\tilde Q^+(\tilde f)$, and $\tilde Q(\tilde f)$

$$\tilde f=f\sqrt{\frac{\nu}{M}},\quad \tilde Q^+(\tilde f)=\frac{1}{\sqrt{\nu M}}\,Q^+(f),\quad \tilde Q(\tilde f)=\tilde Q^+(\tilde f)-\tilde f.$$

By (1.2)

$$\tilde Q^+(\tilde f)=\int B\sqrt{\frac{F'_*F_*}{\nu'\nu}}\;\tilde f'dv_*d\theta d\epsilon.$$

Let $(\,,\,)$ denote the scalar product in $L^2(\mathbb{R}^3)$.

LEMMA 1.2. *Every $\tilde{f} \in L^2(\mathbb{R}^3)$ can uniquely be written*

$$\tilde{f} = c_f\sqrt{\nu M} + \tilde{w}_f, \tag{1.5}$$

with $(\sqrt{\nu M},\ \tilde{w}_f) = 0$. Moreover

$$|(\tilde{Q}^+\tilde{w}_f, \tilde{w}_f)| \le (1-\sigma)\|\tilde{w}_f\|^2_{L^2} \tag{1.6}$$

for some constant σ such that $0 < \sigma < 1$.

Proof. $\tilde{Q}^+$ satisfies Grad's conditions [12], so $\tilde{Q}^+$ is a compact operator in $L^q_s := L^q(\mathbb{R}^3, 1+|v|^s)$, $1 \le q < \infty$, $s \in \mathbb{R}$. Moreover, $\tilde{Q}^+$ is symmetric in L^2. Hence its eigenvector spaces span L^2 and are finite dimensional for nonzero eigenvalues. Then

$$|(\tilde{Q}^+\tilde{f}, \tilde{f})| = \int B\sqrt{\frac{F'_*}{\nu'}}\ \tilde{f}'\ \overline{\sqrt{\frac{F_*}{\nu}}\ \tilde{f}} \le \int B\frac{F_*}{\nu}|\tilde{f}|^2 = \int |\tilde{f}|^2,$$

so $-\tilde{Q}$ is positive in L^2 and $\|\tilde{Q}^+\| \le 1$. The $\tilde{Q}^+$-eigenvalue 1 is simple. Indeed, $\tilde{Q}^+\tilde{f} = \tilde{f}$ implies $(\tilde{Q}\tilde{f}, \tilde{f}) = 0$, which can be written

$$\int B\left(\sqrt{\frac{F'_*}{\nu'}}\ \tilde{f}' - \sqrt{\frac{F_*}{\nu}}\tilde{f}\right)^2 dvdv_*\ d\theta d\epsilon = 0.$$

Hence

$$\sqrt{\frac{F'_*}{\nu'}}\ \tilde{f}' = \sqrt{\frac{F_*}{\nu}}\ \tilde{f},$$

or $\frac{f'}{M'} = \frac{f}{M}$ by (1.2). It follows (see [14]) that $\tilde{f} = c\sqrt{\nu M}$, where c is a constant. Now -1 is not an eigenvalue of $\tilde{Q}^+$. Otherwise, $\tilde{Q}^+\tilde{f} = -\tilde{f}$ for some $\tilde{f}$ implies

$$\int B\left(\sqrt{\frac{F'}{\nu'}}\ \tilde{f}' + \sqrt{\frac{F_*}{\nu}}\ \tilde{f}\right)^2 dvdv_*\ d\theta d\epsilon = 0,$$

so $\frac{f'}{M'} = -\frac{f}{M}$. Varying v_* and the angular coordinate for v fixed gives that f has a constant sign. Hence $\tilde{f} = 0$. Since $\tilde{Q}^+$ is compact and symmetric, $\|\tilde{Q}^+\| \le 1$, -1 is not an eigenvalue, and the eigenspace of 1 is $c\sqrt{\nu M}$, it follows that every $\tilde{f} \in L^2$ can be uniquely written as

$$\tilde{f} = c_f\sqrt{\nu M} + \tilde{w}_f, \quad \text{with } (\sqrt{\nu M}, \tilde{w}_f) = 0$$

and

$$|(\tilde{Q}^+\tilde{w}_f, \tilde{w}_f)| \le (1-\sigma)\|\tilde{w}_f\|^2_{L^2}, \quad 0 < \sigma < 1. \qquad \square$$

Let us next describe the time asymptotics for the solution of the initial boundary value problem (1.1)–(1.3–1.4).

THEOREM 1.3. *Let f_i and f_b be functions belonging to $L^2_{\frac{1}{M}}(\Omega \times \mathbb{R}^3)$ and $L^2_{\frac{v\cdot n(x)}{M}}((\partial\Omega \times \mathbb{R}^3)^+)$. When t tends to infinity, the solution to the initial boundary value problem (1.1)–(1.3–1.4) converges in $L^1(\Omega \times \mathbb{R}^3)$ to the unique stationary solution g of the linear stationary Boltzmann equation*

$$v \cdot \nabla_x g = Q(g), \tag{1.7}$$

with $\tilde{g} \in L^2$, complemented with the boundary condition

$$g(x,v) = f_b(x,v), \quad (x,v) \in (\partial\Omega \times \mathbb{R}^3)^+. \tag{1.8}$$

Proof. Due to the linearity of (1.1), f can be split into the sum of the solution to (1.1) with initial condition f_i and zero boundary condition, and the solution to (1.1) with a zero initial condition and f_b boundary condition. Again by linearity it is enough to consider nonnegative initial and boundary values. Let us first prove that the first part tends to zero in $L^1_{x,v}$ when t tends to infinity. Let $d\alpha(x)$ denote the measure on the boundary $\partial\Omega$. The Green formula applied to (1.1), together with (1.6), implies

$$\begin{aligned}
&\int_{\Omega\times\mathbb{R}^3} \frac{|\tilde f(t,x,v)|^2}{\nu(v)}dxdv \\
&+\int_0^t\int_{(\partial\Omega\times\mathbb{R}^3)^-} \frac{|v\cdot n(x)|}{\nu(v)}|\tilde f(s,x,v)|^2 dsd\alpha(x)dv \\
&+\sigma\int_0^t\int_{\Omega\times\mathbb{R}^3}|\tilde w(s,x,v)|^2 dsdxdv \le \int_{\Omega\times\mathbb{R}^3} \frac{|\tilde f(t,x,v)|^2}{\nu(v)}dxdv \\
&+\int_0^t\int_{(\partial\Omega\times\mathbb{R}^3)^-} \frac{|v\cdot n(x)|}{\nu(v)}|\tilde f(s,x,v)|^2 dsd\alpha(x)dv \\
&-\int_0^t\int_\Omega (\tilde Q\tilde w,\tilde w)(s,x)dsdx = \int_{\Omega\times\mathbb{R}^3} \frac{|\tilde f_i(x,v)|^2}{\nu(v)}dxdv. \qquad (1.9)
\end{aligned}$$

It follows that $\int_{\Omega\times\mathbb{R}^3}\frac{1}{\nu(v)}|\tilde f(t,x,v)|^2dxdv$ decreases with time. Moreover, there is a sequence t_j tending to infinity and a function $\tilde f^\infty$ such that $\tilde f(t_j+t)$ tends weakly to $\tilde f^\infty$ in $L^2_{\frac{1}{\nu}}$, and $\int_0^t\int_{\Omega\times\mathbb{R}^3}|\tilde w(t_j+t,x,v)|^2\ dtdxdv$ and $\int_0^1\int_{(\partial\Omega\times\mathbb{R}^3)^-}\frac{|v\cdot n(x)|}{\nu(v)}|\tilde f(t_j+t,x,v)|^2\,dtd\alpha(x)dv$ tend to zero when j tends to infinity. The function f^∞ is a weak solution to the equation (1.1) with zero boundary condition and $\tilde w_{f^\infty}=0$. It follows that $f^\infty = c^\infty M$ for some constant c^∞. The null boundary conditions imply that $c^\infty = 0$. Hence $\tilde f(t)$ weakly converges to zero in $L^2_{\frac{1}{v}}$. Since

$$\|f(t)\|_{L^1_{x,v}} = \int_\Omega\int_{\mathbb{R}^3}\sqrt{\frac{M(v)}{\nu(v)}}\ \tilde f(t,x,v)dxdv,$$

$f(t)$ tends to zero strongly in $L^1_{x,v}$, when t tends to infinity.

Let us prove that the solution to the initial boundary value problem (1.1)–(1.3–1.4) with null initial condition and boundary condition f_b tends to a stationary solution g to (1.7–1.8). In view of possible future applications we prefer not to give a proof based on the existence of stationary solutions being known but instead to deduce their existence from the long time behavior. By translation invariance in time, the solution at time $t+s$ is the sum of the solution at time t and the contribution at time s carried forward with zero boundary values to $t+s$. So $f(t,x,v)$ is increasing with time and converges pointwise in x,v to a measurable function f^∞, when t tends to infinity. Let us prove that $\tilde f^\infty$ belongs to L^2. For any set $\Gamma\subset\Omega\times\mathbb{R}^3$, multiplying (1.1) by $\tilde f$ and using Green's formula leads to

$$\begin{aligned}
&\int_{\Gamma^c}\frac{1}{\nu(v)}(|\tilde f(t+s)|^2-|\tilde f(t)|^2)dxdv+\int_\Gamma\frac{1}{\nu(v)}|\tilde f(t+s)|^2dxdv \\
&+\sigma\int_t^{t+s}\int_{\Omega\times\mathbb{R}^3}|\tilde w(\tau,x,v)|^2d\tau dxdv
\end{aligned}$$

$$+\int_t^{t+s}\int_{(\partial\Omega\times\mathbb{R}^3)^-}\frac{|v\cdot n(x)|}{\nu(v)}|\tilde f(\tau,x,v)|^2 d\tau d\alpha(x)dv$$

$$\leq \int_\Gamma \frac{1}{\nu(v)}|\tilde f(t)|^2 dxdv + sc, \tag{1.10}$$

where

$$c := \int_{(\partial\Omega\times\mathbb{R}^3)^+}\frac{v\cdot n(x)}{\nu(v)}\,|\tilde f_b(x,v)|^2 d\alpha(x)dv.$$

Let $\Gamma_{s\epsilon}\subset\Omega\times\mathbb{R}^3$ be the set of (y,v) such that $|v|\leq\frac{1}{\epsilon}$ and the characteristic starting at (t,y,v), namely, $\{(t+\tau,y+\tau v,v);\,\tau\geq 0\}$, reaches $(\partial\Omega\times\mathbb{R}^3)^-$ at a time smaller than $t+s$. Then from the exponential form of the equation

$$\int_t^{t+s}\int_{(\partial\Omega\times\mathbb{R}^3)^-}\frac{|v\cdot n(x)|}{\nu(v)}\,|\tilde f(\tau,x,v)|^2 d\tau d\alpha(x)dv$$

$$\geq\ c(s,\epsilon)\int_{\Gamma_{s\epsilon}}\frac{1}{\nu(v)}\,|\tilde f(t,x,v)|^2 dxdv$$

for some $c(s,\epsilon)\in(0,1)$. Hence by (1.10)

$$\int_{\Gamma_{s\epsilon}}\frac{1}{\nu}|\tilde f(t+s)|^2 dxdv \leq (1-c(s,\epsilon))\int_{\Gamma_{s\epsilon}}\frac{1}{\nu}|\tilde f(t)|^2 dxdv + sc.$$

It follows that

$$\sup_{t>0}\int_{\Gamma_{s\epsilon}}\frac{1}{\nu(v)}\,|\tilde f(t,x,v)|^2 dxdv$$

is finite. Then by (1.10)

$$\sup_{t>0}\int_0^{\frac{3}{2}}\int_{\Omega\times\mathbb{R}^3}|\tilde w(t+s,x,v)|^2 dsdxdv$$

is finite. Hence, by the previous two lines,

$$\sup_{t>0}\int_0^{\frac{3}{2}}\int_{\Gamma_{s\epsilon}}|c_f(t+s,x)|^2 M(v)dsdxdv$$

is bounded. Since Ω is bounded and convex, it follows that (for ϵ small)

$$\inf_{x\in\Omega}\int_{(x,v)\in\Gamma_{\frac{1}{2}\epsilon}}\nu(v)M(v)dv > c\int_{\mathbb{R}^3}\nu(v)M(v)dv.$$

This implies that

$$\sup_{t>0}\int_{\frac{1}{2}}^{\frac{3}{2}}\int_{\Omega\times\mathbb{R}^3}|c_f(t+s,x)|^2\nu(v)M(v)dsdxdv<\infty.$$

And so

$$\sup_{t>0}\int_0^1\int_{\Omega\times\mathbb{R}^3}|c_f(t+s,x)|^2\nu(v)M(v)dsdxdv$$

is bounded. Finally

$$\sup_{t>0} \int_0^1 \int_{\Omega\times\mathbb{R}^3} |\tilde{f}(t+s,x,v)|^2 dsdxdv$$

and (since $\tilde{f}$ is an increasing function of time)

$$\sup_{t>0} \int_{\Omega\times\mathbb{R}^3} |\tilde{f}(t,x,v)|^2 dxdv$$

are bounded. Hence $\tilde{f}^\infty$ belongs to $L^2(\Omega\times\mathbb{R}^3)$. Moreover, f^∞ solves the stationary problem (1.7–1.8), and $\tilde{f}(t)$ tends to $\tilde{f}^\infty$ in $L^2(\Omega\times\mathbb{R}^3)$, when t tends to infinity. Finally the solution of the stationary problem is unique in the class of functions g such that $\tilde{g}\in L^2$. Indeed, let us prove that if a function g such that $\tilde{g}\in L^2$ satisfies

$$v\cdot\nabla_x g = Q(g), \tag{1.11}$$

$$g(x,v)=0, \quad (x,v)\in(\partial\Omega\times\mathbb{R}^3)^+, \tag{1.12}$$

then $g=0$. We notice that

$$\int Q^+(g)\mathrm{sign}(g)dv - \int \nu|g|dv \le \int Q^+(|g|)dv - \int \nu|g|dv = 0.$$

So, multiplying (1.11) by $\mathrm{sign}(g)$ and integrating implies that

$$\int_{(\partial\Omega\times\mathbb{R}^3)^-} v\cdot n|g| = 0.$$

Hence

$$g(x,v)=0, \quad (x,v)\in\partial\Omega\times\mathbb{R}^3. \tag{1.13}$$

$\tilde{g}$ belongs to L^2 and can be expressed by (1.5) as

$$\tilde{g} = c\sqrt{M\nu} + \tilde{w}. \tag{1.14}$$

It satisfies

$$\frac{1}{\nu(v)}\, v\cdot\nabla_x\tilde{g} = \tilde{Q}\tilde{w}. \tag{1.15}$$

Integrating (1.15) with respect to x and v using (1.6) implies by (1.13) that $\tilde{w}$ is equal to zero. Then $\tilde{g}=0$ follows from (1.11–1.12). □

2. The Milne problem. Write the velocity as $v=(\xi,v')$ with ξ the velocity component in the x-direction and v' the orthogonal velocity component. We consider the Milne problem

$$\frac{\xi}{\nu}\,\partial_x\tilde{f} = \tilde{Q}\tilde{f}, \quad x>0, \quad v\in\mathbb{R}^3, \tag{2.1}$$

$$\tilde{f}(0,v)=\tilde{\varphi}(v), \quad \xi>0. \tag{2.2}$$

THEOREM 2.1. *Let $\tilde{\varphi}\in L^2_{\frac{\xi}{\nu}}(\mathbb{R}_+\times\mathbb{R}^2)$. There is a solution to* (2.1–2.2) *in the set $\{\tilde{f};\exists c_\infty\in\mathbb{R},\ \tilde{f}-c_\infty\sqrt{\nu M}\in L^2(\mathbb{R}_+\times\mathbb{R}^3)\}$, which satisfies $\int\xi f(x,v)dv = c_\infty u$ for all $x\ge 0$. For $u<0$, this holds with $c_\infty=0$, i.e., $\int\xi f(x,v)dv=0$ for all $x\ge 0$.*

Proof. There is—by the approach of Theorem 1.3—a unique solution $\tilde{f}^a \in L^2([0,a]\times\mathbb{R}^3)$ of

$$\frac{\xi}{\nu}\,\partial_x \tilde{f}^a = \tilde{Q}\tilde{f}, \quad x\in[0,a], \quad v\in\mathbb{R}^3,$$

together with (2.2) and boundary conditions at $x=a$ suitable for our purpose. For $u\geq 0$, we take

$$\tilde{f}^a(a,\xi,v') = \tilde{f}^a(a,-\xi+2u,v'), \quad \xi<0, \tag{2.3}$$

whereas for $u<0$,

$$\tilde{f}^a(a,v) = \frac{\sqrt{M(v)\nu(v)}}{\int_{\xi<0}|\xi|M(v)dv}\int_{\xi>0}\xi f^a(a,v)dv, \quad \xi<0. \tag{2.4}$$

Remark. The boundary condition (2.3) can only be used for $u>0$. A desired nonnegativity (2.9) would not be obtained from the boundary condition (2.4) for $u>0$.

Clearly $\int \xi f^a(x,v)dv$ is constant in both cases, moreover equal to zero for $u\leq 0$. Denote by uc_a this constant and bound it for $u>0$ from above and below. First

$$uc_a = \int \xi f^a(0,v)dv \leq \int_{\xi>0}\xi\varphi(v)dv. \tag{2.5}$$

Let $\tilde{f}^a(x,v) = c^a(x)\sqrt{\nu(v)M(v)}+\tilde{w}^a(x,v)$ be the decomposition of $\tilde{f}^a$ from section 1. By orthogonality

$$-(\tilde{Q}\tilde{f}^a,\tilde{f}^a) = -(\tilde{Q}\tilde{w}^a,\tilde{w}^a) \geq \sigma(\tilde{w}^a,\tilde{w}^a). \tag{2.6}$$

Multiplying (2.1) by $\tilde{f}^a$ and integrating over $\mathbb{R}^3_v$ leads to

$$\partial_x\int\frac{\xi}{\nu}\,|\tilde{f}^a|^2(x,v)dv = 2(\tilde{Q}\tilde{f}^a,\tilde{f}^a) \leq -2\sigma\|\tilde{w}^a\|^2 \leq 0. \tag{2.7}$$

Hence for $u\geq 0$,

$$\begin{aligned}\int\frac{\xi}{\nu}\,|\tilde{f}^a|^2(x,v)dv &\geq \int\frac{\xi}{\nu}\,|\tilde{f}^a|^2(a,v)dv \\ &\geq \int_{\xi<0}\frac{\xi}{\nu}\,|\tilde{f}^a|^2(a,v)dv + \int_{\xi>2u}\frac{\xi}{\nu}\,|\tilde{f}^a|^2(a,v)dv \\ &= 2u\int_{\xi>2u}\frac{1}{\nu}\,|\tilde{f}^a|^2(a,v)dv \geq 0,\end{aligned} \tag{2.8}$$

whereas for $u<0$,

$$\begin{aligned}\int\frac{\xi}{\nu}\,|\tilde{f}^a|^2(x,v)dv &\geq \int\frac{\xi}{\nu}\,|\tilde{f}^a|^2(a,v)dv \\ &\geq \left(1-\frac{\int_{\xi>0}\xi M}{\int_{\xi<0}|\xi|M}\right)\int_{\xi>0}\frac{\xi}{\nu}\,|\tilde{f}^a|^2(a,v)dv \geq 0.\end{aligned} \tag{2.9}$$

Indeed, using (2.4)

$$\int_{\xi<0} \frac{|\xi|}{\nu(v)}\ |\tilde f^a|^2(a,v)dv = \frac{1}{\int_{\xi<0}|\xi|M}\ \left(\int_{\xi>0}\xi f^a(a,v)dv\right)^2$$

$$\leq\ \frac{\int_{\xi>0}\xi M}{\int_{\xi<0}|\xi|M}\ \int_{\xi>0}\frac{\xi}{\nu(v)}\ |\tilde f^a|^2(a,v)dv.$$

But $\frac{\int_{\xi>0}\xi M(v)dv}{\int_{\xi<0}|\xi|M(v)dv} < 1$ for $u<0$ and so (2.9) follows. Finally by (2.8)

$$uc_a \geq \int_{\xi<0}\xi f^a(0,v)dv$$

$$=\ -\int_{\xi<0}|\xi|\ \sqrt{\frac{M(v)}{\nu(v)}}\ \tilde f^a(0,v)dv$$

$$\geq\ -\left(\int_{\xi<0}|\xi|M(v)dv\int_{\xi<0}\ \frac{|\xi|}{\nu}\ |\tilde f^a|^2(0,v)dv\right)^{\frac12}$$

$$\geq\ -\left(\int_{\xi<0}|\xi|M(v)dv\int_{\xi>0}\ \frac{\xi}{\nu(v)}\ \tilde\varphi^2(v)dv\right)^{\frac12}. \tag{2.10}$$

In the case $u=0$ the theorem can from here be derived using, e.g., [2] or [16]. So let us only detail the case when $u\neq 0$. First $\tilde w^a$ is bounded in $L^2([0,a]\times\mathbb{R}^3)$ uniformly with respect to a. Indeed by (2.8), (2.9)

$$\sigma\int_0^a\int|\tilde w^a|^2(x,v)dxdv$$

$$\leq\ -\int_0^a(\tilde Q\tilde f^a,\tilde f^a) = -\int_0^a\int\ \frac{\xi}{\nu(v)}\ \tilde f^a\partial_x\tilde f^a dxdv$$

$$=\ \frac12\left(\int\frac{\xi}{\nu(v)}\ |\tilde f^a|^2(0,v)dv - \int\frac{\xi}{\nu}\ |\tilde f^a|^2(a,v)dv\right) \tag{2.11}$$

$$\leq\ \frac12\int_{\xi>0}\ \frac{\xi}{\nu(v)}\ \tilde\varphi^2(v)dv.$$

Since $f^a(x,v) = c^a(x)M(v) + w^a(x,v)$,

$$|c_a - c^a(x)| = \frac{1}{|u|}\ \left|\int\xi w^a(x,v)dv\right|$$

$$\leq\ \frac{1}{|u|}\left(\int\xi^2\ \frac{M(v)}{\nu(v)}\ dv\int|\tilde w^a|^2(x,v)dv\right)^{\frac12}$$

so that

$$\int_0^a\int\left|\tilde f^a(x,v) - c_a\sqrt{M(v)\nu(v)}\ \right|^2 dxdv \leq c\int_0^a\int|\tilde w^a|^2(x,v)dxdv. \tag{2.12}$$

By (2.5), (2.11), and (2.12), there exist a sequence (a_j) tending to infinity, a number c_∞, and a function $\tilde f$ such that c_{a_j} tends to c_∞ and $\tilde f^{a_j} - c_{a_j}\sqrt{\nu M}$ converges weakly

in L^2 to $\tilde{f} - c_\infty\sqrt{\nu M}$. One can then check that $\tilde{f}$ is a solution to the Milne problem (2.1–2.2) with the desired properties. ∎

For the boundary layer study in section 3, some decay of $\tilde{g} := \tilde{f} - c_\infty\sqrt{\nu M}$ is needed.

PROPOSITION 2.2. *Assume that*

$$\sup_{v\in\mathbb{R}^3_+} |\tilde{\varphi}(v)|(1+|v|)^s < \infty, \quad s \in \mathbb{R}_+. \tag{2.13}$$

Then for $s \in \mathbb{R}_+$, $\int \frac{|\xi|}{\nu(v)}|\tilde{g}(x,v)|^2 dv \le cx^{-s}, \quad x > 0.$

This result can essentially be found in [11]. For the convenience of the reader we give their proof with the differences introduced by the nonzero bulk velocity of the Maxwellian. The proof is based on the entropy method introduced by Bardos, Santos, and Sentis [2], and uses the following decay properties of $\tilde{g}$, pointwise in v and integral in x.

LEMMA 2.3. *Under* (2.13) *for* $s \in \mathbb{R}_+$,

$$\sup_{x>0, v\in\mathbb{R}^3} (1+|v|)^s|\tilde{g}(x,v)| \le c_1, \tag{2.14}$$

$$\int_0^\infty \int_{\mathbb{R}^3} x^s|\tilde{g}(x,v)|^2 dxdv \le c_2. \tag{2.15}$$

The constants c_1, c_2 *depend on* φ *and* s.

Proof of Proposition 2.2. Write

$$\int_{\mathbb{R}^3} \frac{|\xi|}{\nu}\,|\tilde{g}(x,v)|^2 dv \le \int_{|\xi|\le r} \frac{|\xi|}{\nu}\,|\tilde{g}(x,v)|^2 dv$$
$$+ \int_{|\xi|>r, |v|\le\rho} \frac{|\xi|}{\nu}\,|\tilde{g}(x,v)|^2 dv + \int_{|v|\ge\rho} \frac{|\xi|}{\nu}\,|\tilde{g}(x,v)|^2 dv$$
$$:= a + b + c.$$

By (2.14), $a(r)$ and $c(\rho)$ satisfy

$$a(r) \le cr \int_{|\xi|\le r} (1+|v|)^{-2s-\gamma} dv \le cr \quad \text{for } s > \frac{3}{2} - \frac{\gamma}{2}, \tag{2.16}$$

$$c(\rho) \le c \int_{|v|\ge\rho} (1+|v|)^{-2s+1-\gamma} dv \le c\rho^{-1} \quad \text{for } s > \frac{5}{2} - \frac{\gamma}{2}. \tag{2.17}$$

Evidently

$$b(r,\rho) \le cr^{-1}\rho^\gamma \int_{\mathbb{R}^3} \left|\frac{\xi\tilde{g}(x,v)}{\nu(v)}\right|^2 dv.$$

Now

$$\frac{\xi}{\nu}\,(\tilde{g}(y,v) - \tilde{g}(x,v)) = \int_x^y \tilde{Q}(\tilde{w}_g)(z,v)dz,$$

and so by (2.15)

$$\int_{|\xi|>r, |v|\le\rho} \left|\frac{\xi}{\nu}\,(\tilde{g}(y,v) - \tilde{g}(x,v))\right|^2 dv \le c\left(\int_x^y \left(\int_{\mathbb{R}^3} |\tilde{w}_g(z,v)|^2 dv\right)^{\frac{1}{2}} dz\right)^2$$
$$\le\ cx^{-s+1}\int_x^y \int_{\mathbb{R}^3} z^s|\tilde{g}(z,v)|^2 dvdz \le cx^{-s+1}.$$

Since $\tilde{g} \in L^2(\mathbb{R}_+ \times \mathbb{R}^3)$, a sequence $y_j \to \infty$ can be chosen so that $\lim_{j\to\infty} \tilde{g}(y_j, .) = 0$ in $L^2(\mathbb{R}^3)$. It follows that

$$\int_{|\xi|>r, |v|\le\rho} \left| \frac{\xi}{\nu} \tilde{g}(x,v) \right|^2 dv \le c x^{-s+1}$$

and so

$$b(r,\rho) \le c r^{-1} \rho^\gamma x^{-s+1}.$$

The choice $r = x^{-\frac{s}{3}}$, $\rho = x^{\frac{s}{3}}$ in (2.16–2.17) gives the desired result. □

Proof of (2.14). By [11, Prop. 4.3]

$$\sup_{x>0} \|\tilde{g}(x, .)\|_{L^2(\mathbb{R}^3)} \le c, \tag{2.18}$$

where c depends on $\tilde{\varphi}$ in the $L^2 \cap L^\infty$ sense. Also

$$\tilde{Q}^+(\tilde{g})(x,v) = \int_{\mathbb{R}^3} k(v, v_1) \tilde{g}(x, v_1) dv_1 \tag{2.19}$$

with

$$|k(v, v_1)| \le (1 + |v| + |v_1|)^{-1+\gamma} (1 + |v_1|)^{-\frac{\gamma}{2}} \phi(v, v_1)$$

and

$$\int_{\mathbb{R}^3} \phi^2(v, v_1) dv_1 \le c(1 + |v|)^{-1-\gamma}.$$

Hence

$$(1 + |v|)^{\frac{3}{2}-\frac{\gamma}{2}} |\tilde{Q}^+ \tilde{g}(x,v)| \le c \|\tilde{g}(x, .)\|_{L^2(\mathbb{R}^3)}.$$

The exponential form of (2.1–2.2) gives

$$\begin{aligned} (1 + |v|)^{\frac{3}{2}-\frac{\gamma}{2}} |\tilde{g}(x,v)| \le{}& |\tilde{\varphi}(v)| (1 + |v|)^{\frac{3}{2}-\frac{\gamma}{2}} \chi_{\xi>0} \\ &+ c \sup_{x>0} \|\tilde{g}(x, .)\|_{L^2(\mathbb{R}^3)}. \end{aligned} \tag{2.20}$$

Here $\chi_{\xi>0}$ is the characteristic function of the set $\{v \in \mathbb{R}^3;\ \xi > 0\}$. By (2.18) the right-hand side is finite. Also

$$\int_{\mathbb{R}^3} (1 + |v|)^{s+1} (1 + |v_1|)^{-s} k(v, v_1) dv_1 < \infty, \quad s \in \mathbb{R}_+.$$

Using this together with (2.20), a direct estimate in the exponential form of (2.1–2.2) gives (2.14). □

Proof of 2.15. By (2.11), which also holds for $\tilde{w}_g$, there is a sequence $y_j \to \infty$ such that

$$\int |\tilde{w}_g(y_j, v)|^2 dv \to 0. \tag{2.21}$$

It follows from Theorem 2.1 that $\int_{\mathbb{R}^3} \xi g(x,v)dv = 0$, $x \in \mathbb{R}_+$, and so the orthogonal decomposition $\tilde{g}(x,v) = c^\infty(x)\sqrt{\nu(v)M(v)} + \tilde{w}_g(x,v)$ gives

$$|uc^\infty(x)| = \left| \int \xi w_g(v) dv \right|$$

$$\leq \left(\frac{\xi^2 M(v)}{\nu(v)} dv \int |\tilde{w}_g(x,v)|^2 dv \right)^{\frac{1}{2}}. \tag{2.22}$$

In particular

$$\lim_{j\to\infty} c^\infty(y_j) = 0. \tag{2.23}$$

Now the proof is based on a study of the entropy flux

$$H(x) = \int \frac{\xi}{\nu} \, |\tilde{g}(x,v)|^2 dv.$$

Using the orthogonal decomposition of $\tilde{g}$ and splitting the domain of integration we get

$$\lim_{j\to\infty} H(y_j) \leq \lim_{j\to\infty} Cc(y_j)^2 + \lim_{j\to\infty} C \int_{\mathbb{R}^3} \tilde{w}_g(y_j,v)^2 dv$$
$$+ \lim_{j\to\infty} C \int_{|v|\geq\rho} \frac{\xi}{\nu(v)} \, \tilde{g}(y_j,v)^2 dv.$$

By (2.21) and (2.23) the first two of these limits are zero. By (2.14) the third one is bounded by

$$c \int_{|v|\geq\rho} \frac{|\xi|}{\nu(v)} \, (1+|v|)^{-5} dv \leq \frac{c}{\rho}.$$

It follows that $\lim_{j\to\infty} H(y_j) = 0$. A multiplication of (2.1) by $\tilde{g}$ and v-integration show that $H(x)$ is nonincreasing. And so

$$0 \leq H(x) \leq H(0) \leq \int_{\xi>0} \frac{\xi}{\nu} \, \tilde{\varphi}(v)^2 dv. \tag{2.24}$$

Since $\tilde{g} \in L^2(\mathbb{R}_+ \times \mathbb{R}^3)$, it is enough for (2.15) to consider

$$\int_1^\infty \int_{\mathbb{R}^3} x^s \tilde{g}(x,v)^2 dx dv.$$

A multiplication of (2.1) by $x^s\tilde{g}$ and integration gives

$$H(y)y^s + \int_1^y \left(x^s \int_{\mathbb{R}^3} \tilde{w}_g(x,v)^2 dv - s x^{s-1} H(x) \right) dx$$
$$\leq H(1) \leq \int_{\xi>0} \frac{\xi}{\nu} \, \tilde{\varphi}(v)^2 dv. \tag{2.25}$$

The positivity of $H(y)$ implies that

$$\int_1^y \left(x^s \int_{\mathbb{R}^3} \tilde{w}_g(x,v)^2 dv - s x^{s-1} H(x) \right) dx \leq H(1).$$

Now

$$\int_{|v|\le\rho} \frac{|\xi|}{\nu}\, \tilde g(x,v)^2 dv \le c\rho^{1-\gamma}\|\tilde g(x,.)\|^2_{L^2(\mathbb{R}^3)},$$

and by (2.14), for any $\lambda \in \mathbb{R}_+$,

$$\int_{|v|\ge\rho} \frac{|\xi|}{\nu}\, \tilde g(x,v)^2 dv \le c_\lambda \rho^{-\lambda}.$$

This together with (2.22) and (2.25) implies for some $\alpha > 0$ that

$$\int_1^y x^s \left(\int \tilde g(x,v)^2 dv\right)\left(\alpha - \frac{c\rho^{1-\gamma}}{x}\right) dx \le H(1) + c_\lambda \int_1^y \rho^{-\lambda} s x^{s-1} dx.$$

The choice $\rho(x) = (\frac{\alpha x}{2c})^{\frac{1}{1-\gamma}}, \lambda > s(1-\gamma)$ yields

$$\int_1^\infty x^s \int \tilde g(x,v)^2 dv dx \le c_s. \qquad \square$$

3. The fluid approximation with initial and boundary layers for nonzero bulk velocity. Introduce the mean free path $\epsilon > 0$ and take $u > 0$. This section considers the slab problem

$$\partial_t f_\epsilon + \frac{1}{\epsilon^2}\xi \partial_x f_\epsilon = \frac{1}{\epsilon^3} Q(f_\epsilon), \quad t > 0, \quad x \in (0,1), \quad v \in \mathbb{R}^3, \tag{3.1}$$

together with the initial condition

$$f_\epsilon(0,x,v) = f_i(x,v), \quad x \in (0,1), \quad v \in \mathbb{R}^3, \tag{3.2}$$

and the boundary conditions

$$f_\epsilon(t,0,v) = f_0(v), t > 0, \xi > 0; \qquad f_\epsilon(t,1,v) = f_1(v), \quad t > 0, \quad \xi < 0. \tag{3.3}$$

After an initial layer, the unique solution satisfies the stationary problem

$$\xi \partial_x g_\epsilon = \frac{1}{\epsilon} Q(g_\epsilon), \quad x \in (0,1), \quad v \in \mathbb{R}^3, \tag{3.4}$$

together with the boundary conditions

$$g_\epsilon(0,v) = f_0(v), \quad \xi > 0; \quad g_\epsilon(1,v) = f_1(v), \quad \xi < 0, \tag{3.5}$$

if one disregards the error term from the initial layer. Moreover, g_ϵ can be split into a fluid part cM in the interior of the domain together with boundary layers and with the error term tending to zero strongly in L^1, when ϵ tends to zero.

THEOREM 3.1. *Let f_i, f_0, f_1 be given with $\tilde f_i \in L^2_{\frac{1}{\nu}}((0,1)\times\mathbb{R}^3)$, $\tilde f_0 \in L^2_{\frac{\xi}{\nu}}(\mathbb{R}^\xi_+ \times \mathbb{R}^2)$, $\tilde f_1 \in L^2_{\frac{|\xi|}{\nu}}(\mathbb{R}^\xi_- \times \mathbb{R}^2)$. Denote by f_ϵ and g_ϵ the unique solutions of* (3.1–3.3)*, respectively,* (3.4–3.5) *with these given initial and boundary values. Then for $t > 0$*

$$\lim_{\epsilon\to 0} f_\epsilon(t,.) - g_\epsilon(.) = 0$$

strongly in $L^1((0,1)\times\mathbb{R}^3)$.

THEOREM 3.2. *Under the same hypotheses there are a constant c, boundary layer terms $l_\epsilon(x,v) = l^0(\frac{x}{\epsilon}, v)$, and $r_\epsilon(x,v) = r^0(\frac{x-1}{\epsilon}, v)$, with $\tilde{l}^0$ and $\tilde{r}^0$, respectively, belonging to $L^2(\mathbb{R}_+ \times \mathbb{R}^3)$ and $L^2(\mathbb{R}_- \times \mathbb{R}^3)$ such that*

$$g_\epsilon = cM + l_\epsilon + r_\epsilon + S_\epsilon. \tag{3.6}$$

Here the terms $\tilde{l}^0$ and $\tilde{r}^0$ have the decay properties of Proposition 2.2, and the remainder term S_ϵ tends to 0 in $L^1((0,1) \times \mathbb{R}^3_v)$, when ϵ tends to 0.

The proof of Theorem 3.1 is based on the following lemma.

LEMMA 3.3. *Let f_i be given with $0 \le \tilde{f}_i \in L^2_{\frac{1}{\nu}}((0,1) \times \mathbb{R}^3)$. Denote by $f_\epsilon(t,x,v)$ the solution of (3.1–3.3) with f_i as initial value and boundary values $f_0 = f_1 = 0$. For $s > 0$, $f_\epsilon(s,.,.)$ converges strongly in $L^1((0,1) \times \mathbb{R}^3)$ to zero, when ϵ tends to zero.*

Proof. After scaling $t \to \frac{t}{\epsilon^2}$, the solution (still denoted f_ϵ) satisfies

$$(\partial_t + \xi\partial_x) f_\epsilon = \frac{1}{\epsilon}\, Q(f_\epsilon), \quad t \in \mathbb{R}_+, \quad x \in (0,1), \quad v \in \mathbb{R}^3,$$
$$f_\epsilon(0,.) = f_i(.),$$
$$f_\epsilon(t,0,v) = 0, \quad t > 0, \quad \xi > 0, \quad f_\epsilon(t,1,v) = 0, \quad t > 0, \quad \xi < 0.$$

Green's formula implies that mass and entropy

$$\int_0^1 \int_{\mathbb{R}^3} f_\epsilon(t,x,v)\,dxdv, \quad \int_0^1 \int_{\mathbb{R}^3} \frac{\tilde{f}_\epsilon(t,x,v)^2}{\nu(v)}\,dxdv$$

are decreasing with time. Suppose that the lemma does not hold. Then for some $s > 0$, there is a sequence (ϵ_j) with $\lim_{j\to\infty} \epsilon_j = 0$ such that

$$\inf_j \int_0^1 \int_{\mathbb{R}^3} f_{\epsilon_j}(t_j,x,v)\,dxdv > 0.$$

Here $t_j = \frac{s}{\epsilon_j^2}$. The lemma follows if for a subsequence of (t_j) (still denoted (t_j)) there is a sequence (t'_j) with $0 \le t'_j \le t_j$ such that

$$\lim_{j\to\infty} \int_0^1 \int_{\mathbb{R}^3} f_{\epsilon_j}(t'_j,x,v)\,dxdv = 0.$$

With $\tilde{f}_{\epsilon_j} := \tilde{f}_j, \tilde{w}_{\epsilon_j} := \tilde{w}_j$, (1.9) gives

$$\int_0^1 \int_{\mathbb{R}^3} \frac{\tilde{f}_j(t_j,x,v)^2}{\nu(v)}\,dxdv + \frac{\sigma}{\epsilon_j} \int_0^{t_j} \int_0^1 \int_{\mathbb{R}^3} \tilde{w}_j(\tau,x,v)^2 d\tau dxdv \le \int_0^1 \int_{\mathbb{R}^3} \frac{\tilde{f}_i(x,v)^2}{\nu(v)}\,dxdv := \sigma c_1.$$

If each of the $s\epsilon_j^{-\frac{3}{2}}$ intervals $[l\epsilon_j^{-\frac{1}{2}}, (l+1)\epsilon_j^{-\frac{1}{2}}]$ of $[0,t_j]$ has

$$\frac{1}{\epsilon_j} \int_{l\epsilon_j^{-\frac{1}{2}}}^{(l+1)\epsilon_j^{-\frac{1}{2}}} \int_0^1 \int_{\mathbb{R}^3} \tilde{w}_j(\tau,x,v)^2 d\tau dxdv > \epsilon_j^{\frac{3}{2}}\, \frac{c_1}{s},$$

then

$$\frac{\sigma}{\epsilon_j}\int_0^{t_j}\int_0^1\int_{\mathbb{R}^3}\tilde{w}_j(\tau,x,v)^2 d\tau dxdv > c_1\sigma.$$

This contradiction implies that for some interval $I_j \subset [0,t_j]$ and of length $\epsilon_j^{-\frac{1}{2}}$

$$\frac{1}{\epsilon_j}\int_{I_j}\int_0^1\int_{\mathbb{R}^3}\tilde{w}_j(\tau,x,v)^2 d\tau dxdv \le e_j^{\frac{3}{2}}\ \frac{c_1}{s}.$$

In particular

$$\lim_{j\to\infty}\epsilon_j^{-2}\int_{I_j}\int_0^1\int_{\mathbb{R}^3}\tilde{w}_j(\tau,x,v)^2 d\tau dxdv = 0.$$

With $I_j = (t_j', t_j'')$ it follows that for $t \ge 0$ (and some subsequence of the j's)

$$\tilde{f}_j(t_j'+t,x,v) \rightharpoonup \tilde{f}_\infty(t,x,v)$$

weakly in $L^2_{\frac{1}{\nu}}((0,1)\times\mathbb{R}^3)$. Here

$$\int_0^1\int_{\mathbb{R}^3}\frac{\tilde{f}_\infty(t,x,v)^2}{\nu(v)}\ dxdv \le \sigma c_1.$$

By the equicontinuity in t, it is enough to prove the above weak L^2-convergence for rational t's. Using (1.9) we have for t fixed and j large enough that

$$\int_0^1\int_{\mathbb{R}^3}\frac{1}{\nu(v)}\ f_j(t_j'+t,x,v)^2 dxdv \le \sigma c_1.$$

So a subsequence of $\tilde{f}_j(t_j'+t)$ converges weakly when $j\to 0$. We conclude with a Cantor diagonalization argument.

Also for a.e. $t>0$,

$$\tilde{w}_j(t_j'+t,x,v)\to 0$$

strongly in $L^2((0,1)\times\mathbb{R}^3)$, and so

$$\tilde{f}_\infty(t,x,v) = c_\infty(t,x,)\sqrt{\nu(v)M(v)}.$$

But $\tilde{f}_\infty$ satisfies

$$(\partial_t+\xi\partial_x)\tilde{f}_\infty = 0, \quad t>0, \quad x\in(0,1), \quad v\in\mathbb{R}^3,$$
$$\tilde{f}_\infty(t,0,v) = 0, \quad t>0, \quad \xi>0; \quad \tilde{f}_\infty(t,1,v) = 0, \quad t>0, \quad \xi<0,$$

and so $\tilde{f}_\infty \equiv 0$. In particular $\lim_{j\to\infty}\tilde{f}_j(t_j',.,.) = 0$ weakly in $L^2_{\frac{1}{\nu}}((0,1)\times\mathbb{R}^3)$. It follows that

$$0 \le \lim_{j\to\infty}\int_0^1\int_{\mathbb{R}^3} f_j(t_j,x,v)dxdv$$
$$\le \lim_{j\to\infty}\int_0^1\int_{\mathbb{R}^3} f_j(t_j',x,v)dxdv = \lim_{j\to\infty}\int_0^1\int_{\mathbb{R}^3}\sqrt{\frac{M(v)}{\nu(v)}}\ \tilde{f}_j(t_j',x,v)dxdv$$
$$= 0.$$

This completes the proof of the lemma. □

Proof of Theorem 3.1. The function $f_\epsilon - g_\epsilon$ satisfies (3.1–3.3) with initial value $f_i - g_i$ and boundary value zero. By linearity it is enough to prove the theorem when the boundary values are zero and $f_i - g_i \geq 0$, and this case is contained in Lemma 3.3. □

Proof of Theorem 3.2. Essentially by section 1, there is a unique solution $g_\epsilon(x,v)$ with $\tilde{g}_\epsilon \in L^2$ to

$$\begin{aligned} \xi \partial_x g_\epsilon &= \frac{1}{\epsilon}\, Q(g_\epsilon), \\ g_\epsilon(0,v) &= f_0(v), \quad \xi > 0, \\ g_\epsilon(1,v) &= f_1(v), \quad \xi < 0. \end{aligned}$$

From the results on the Milne problem in Theorem 2.1, there is a constant c such that

$$\begin{aligned} \frac{\xi}{\nu(v)}\, \partial_x \tilde{q}(x,v) = \tilde{Q}\tilde{q}(x,v), \quad x > 0, \quad v \in \mathbb{R}^3, \\ \tilde{q}(0,v) = \tilde{f}_0(v), \quad \xi > 0 \end{aligned}$$

has a solution $\tilde{q} = c\sqrt{\nu M} + \tilde{l}$, with $\tilde{l} \in L^2(\mathbb{R}^+ \times \mathbb{R}^3_v)$. Define l^0 by

$$l^0(y,v) = \sqrt{\frac{M(v)}{\nu(v)}}\, \tilde{l}(y,v).$$

Also by Theorem 2.1 the Milne problem

$$\begin{aligned} \frac{\xi}{\nu(v)}\, \tilde{r}_x(x,v) = \tilde{Q}\tilde{r}(x,v), \quad x < 0, \quad v \in \mathbb{R}^3, \\ \tilde{r}(0,v) = \tilde{f}_1(v) - c\sqrt{\nu(v)M(v)}, \quad \xi < 0 \end{aligned}$$

has a solution $\tilde{r} \in L^2(\mathbb{R}_- \times \mathbb{R}^3_v)$. Indeed for $u > 0$, looking for a solution defined in $\mathbb{R}_-$ corresponds to considering $u < 0$ in the $\mathbb{R}_+$ situation. Define r^0 by

$$r^0(y,v) = \sqrt{\frac{M(v)}{\nu(v)}}\, \tilde{r}(y,v).$$

$S_\epsilon := g_\epsilon - cM - l_\epsilon - r_\epsilon$ satisfies

$$\begin{aligned} \xi \partial_x S_\epsilon = \frac{1}{\epsilon}\, Q(S_\epsilon), \quad x \in (0,1), \quad v \in \mathbb{R}^3, \\ S_\epsilon(0,v) = -r^0\left(-\frac{1}{\epsilon}, v\right), \quad \xi > 0, \\ S_\epsilon(1,v) = -l^0\left(\frac{1}{\epsilon}, v\right), \quad \xi < 0. \end{aligned} \tag{3.7}$$

Introduce as above the orthogonal decomposition

$$\tilde{S}_\epsilon(x,v) = c^\epsilon(x)\sqrt{\nu M} + \tilde{w}_\epsilon.$$

It follows from (3.4), Green's formula, and (1.6) that

$$\int_{\xi<0} \frac{|\xi|}{\nu}\ |\tilde{S}_\epsilon(0,v)|^2 dv + \int_{\xi>0} \frac{\xi}{\nu}\ |\tilde{S}_\epsilon(1,v)|^2 dv$$

$$+ \frac{\sigma}{\epsilon}\int_0^1 \int_{\mathbb{R}^3} |\tilde{w}_\epsilon(x,v)|^2 dx dv$$

$$\leq \int_{\xi>0} \frac{\xi}{\nu}\ \left|\tilde{r}^0\left(-\frac{1}{\epsilon},v\right)\right|^2 dv + \int_{\xi<0} \frac{|\xi|}{\nu}\ \left|\tilde{l}^0\left(\frac{1}{\epsilon},v\right)\right|^2 dv. \tag{3.8}$$

By Proposition 2.2 the right-hand side tends superalgebraically to zero, when ϵ tends to zero. By (3.4)

$$uc_\epsilon = \int \xi S_\epsilon(x,v) dv$$

is independent of x. Multiplying (3.7) with $\mathrm{sign} S_\epsilon$ and integrating we get

$$\int_{\xi<0} |\xi|\ |S_\epsilon(0,v)| dv + \int_{\xi>0} \xi |S_\epsilon(1,v)| dv$$

$$\leq \int_{\xi>0} \xi \left|r^0\left(-\frac{1}{\epsilon},v\right)\right| dv + \int_{\xi<0} \left|\xi l^0\left(\frac{1}{\epsilon},v\right)\right| dv.$$

Thus

$$|c_\epsilon| \leq \frac{1}{u}\int |\xi| S_\epsilon(0,v) dv$$

$$\leq \frac{c}{u}\left(\int_{\xi>0} \frac{\xi}{\nu}\ \left|\tilde{r}^0\left(-\frac{1}{\epsilon},v\right)\right|^2 dv + \int_{\xi<0} \frac{|\xi|}{\nu}\ \left|\tilde{l}^0\left(\frac{1}{\epsilon},v\right)\right|^2 dv\right), \tag{3.9}$$

which tends to zero superalgebraically, when ϵ tends to zero. As in (2.12)

$$\int_0^1 \int_{\mathbb{R}^3} \left|\tilde{S}_\epsilon(x,v) - c_\epsilon \sqrt{\nu(v) M(v)}\ \right|^2 dx dv$$

$$\leq c \int_0^1 \int_{\mathbb{R}^3} |\tilde{w}_\epsilon(x,v)|^2 dx dv. \tag{3.10}$$

By (3.5–3.7)

$$\int_0^1 \int_{\mathbb{R}^3} |\tilde{S}_\epsilon(x,v)|^2 dx dv$$

tends to zero superalgebraically, when ϵ tends to zero. □

Remark. The evaporation at $x = 0$ determines the (fluid dynamic) mass flux term cM through the boundary layer analysis. At the condensation boundary $x = 1$ this term is removed from the boundary layer correction.

Remark. It follows from this proof that the solution of the Milne problem in Theorem 2.1 is unique. It also follows that the convergence to zero of the error term in Theorem 3.2 is superalgebraic.

REFERENCES

[1] L. Arkeryd and A. Nouri, *A compactness result related to the stationary Boltzmann equation in a slab, with applications to the existence theory*, Indiana U. Math. J., 44 (1995), pp. 815–839.

[2] C. Bardos, R. Santos, and R. Sentis, *Diffusion approximation and computation of the critical size*, Trans. Amer. Math. Soc., 284 (1984), pp. 617–649.

[3] C. Bardos, R. Caflish, and B. Nicolaenko, *Thermal Layer Solutions of the Boltzmann Equation*, Prog. Phys. 10, Birkhäuser Boston, Cambridge, MA, 1985.

[4] C. Bardos, R. Caflish, and B. Nicolaenko, *Different aspects of the Milne problem*, Transport Theory Statist. Phys., 16 (1987), pp. 561–585.

[5] R. Caflisch, *The fluid dynamic limit of the nonlinear Boltzmann equation*, Comm. Pure Appl. Math., 33 (1980), pp. 651–666.

[6] C. Cercignani, *Half space problems in kinetic theory*, Lecture Notes in Phys., 249 (1987), pp. 35–51.

[7] C. Cercignani, *The Boltzmann Equation and its Applications*, Springer, Berlin, 1988.

[8] S. Chandrasekhar, *Radiative Transfer*, Dover, New York, 1950.

[9] A. Demasi, R. Esposito, and J. L. Lebowitz, *Incompressible Navier-Stokes and Euler limits of the Boltzmann equation*, Comm. Pure Appl. Math., 42 (1989), pp. 1189–1214.

[10] R. Ellis and M. Pinsky, *The first and second fluid approximation to the linearized Boltzmann equation*, J. Math. Pure Appl., 54 (1975), pp. 125–126.

[11] F. Golse and F. Poupaud, *Stationary solutions of the linearized Boltzmann equation in a half-space*, Math. Meth. Appl. Sci., 11 (1989), pp. 483–502.

[12] H. Grad, *Asymptotic theory of the Boltzmann equation*, Phys. Fluids, 6 (1963), pp. 147–181.

[13] T. Nishida, *Fluid dynamic limit of the nonlinear Boltzmann equation to the level of the compressible Euler equation*, Comm. Math. Phys., 61 (1978), pp. 119–148.

[14] R. Pettersson, *On weak and strong convergence to equilibrium for solutions to the linear Boltzmann equation*, J. Stat. Phys., 72 (1993), pp. 355–380.

[15] R. Pettersson, *On solutions and higher moments for the linear Boltzmann equation with infinite-range forces*, IMA J. Appl. Math., 38 (1987), pp. 151–166.

[16] F. Poupaud, *Diffusion approximation of the linear semiconductor Boltzmann equation: Analysis of boundary layers*, Asymptotic Anal., 4 (1991), pp. 293–317.

[17] Y. Sone, *Asymptotic theory of a steady flow of a rarefied gas past bodies for small Knudsen numbers*, Advances in Kinetic Theory and Continuum Mechanics, Springer, Berlin, 1991.

[18] Y. Sone, T. Ohwada, and K. Aoki, *Evaporation and condensation of a rarefied gas between its two parallel condensed phases with different temperatures and negative temperature gradient phenomenon*, Lecture Notes in Mathematics 1460, Springer, Berlin, 1991.

[19] S. Ukai and K. Asano, *The Euler limit and initial layer of the nonlinear Boltzmann equation*, Hokkaido Math. J., 12 (1983), pp. 303–324.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 48–68, January 1998

004

A VARIATIONAL PROBLEM RELATED TO THE GINZBURG–LANDAU MODEL OF SUPERCONDUCTIVITY WITH NORMAL IMPURITY INCLUSION*

SHIJIN DING†, ZUHAN LIU‡, AND WANGHUI YU†

Abstract. A variational problem related to the Ginzburg–Landau model of superconductivity with normal impurity inclusion is considered. A standing feature of this problem is the vortex-pinning effect (i.e., the zeros of solutions are attracted to the region occupied by the normal impurities) as some parameters are sufficiently small. Asymptotic behaviors of the solutions of this problem as these parameters tend to zero is studied, and the vortex-pinning effect is proved.

Key words. Ginzburg–Landau model, superconductivity, normal impurities, asymptotic behavior, vortex-pinning effect

AMS subject classifications. 35J55, 35Q40

PII. S0036141096303086

1. Introduction.

Let Ω_n and $\Omega(\Omega_n \subset\subset \Omega)$ be two bounded domains in $\mathbb{R}^2$, $g : \partial\Omega \to S^1$ be a smooth map with $\deg(g, \partial\Omega) > 0$. Set $\Omega_s \equiv \Omega\backslash\overline{\Omega_n}$. Consider the following variational problem: find a function $u \in H_g^1(\Omega)$ such that

$$E(u, \Omega) = \min_{v \in H_g^1(\Omega)} E(v, \Omega), \tag{1.1}$$

where

$$E(v, \Omega) \equiv \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{4\varepsilon^2}\int_{\Omega_s}(1 - |v|^2)^2 + \frac{1}{2\mu^2}\int_{\Omega_n}|v|^2, \tag{1.2}$$

$$H_g^1(\Omega) \equiv \{v : \Omega \to \mathcal{C} | v \in H^1(\Omega), \quad v = g \text{ on } \partial\Omega\}, \tag{1.3}$$

and ε, μ are small positive constants.

This problem is related to the Ginzburg–Landau model of superconductivity with normal impurity inclusion such as superconducting-normal junctions (cf. [1]). Ω_s and Ω_n represent the domains occupied by superconducting materials and normal conducting materials, respectively. The solution u is the Ginzburg–Landau complex order parameter. Zeros of u are known as Ginzburg–Landau vortices which are of significance in the theory of superconductivity. When the vortices move, resistance to the current is produced and causes loss of superconductivity. One way to prevent the movement of the vortices is to add some impurities such as normal conducting materials into the superconducting materials to provide pinning sites for the vortices. In this setting, a simplified Ginzburg–Landau's Gibbs free energy is given by (1.2) (cf. [1], [2]).

This paper is devoted to the study of asymptotic behaviors of the solution u as ε, μ, and diam Ω_n tend to zero as well as the vortex-pinning effect of (1.1)–(1.3). For simplicity, we shall let $\varepsilon = \mu$, $\Omega = B_1$, $\Omega_n = B_\rho$, and $\Omega_s = B_1\backslash\overline{B_\rho}$. Here $0 < \rho < \frac{1}{4}$

*Received by the editors May 6, 1996; accepted for publication (in revised form) November 19, 1996.

http://www.siam.org/journals/sima/29-1/30308.html

†Department of Mathematics, Suzhou University, Suzhou 215006, China.

‡Department of Mathematics, Yangzhou Normal College, Yangzhou 225002, China.

and B_r is the disc in $\mathbb{R}^2$ centered at the origin with radius r. Then our problem becomes

$$E(u, B_1) = \min_{v \in H_g^1(B_1)} E(v, B_1), \tag{1.4}$$

where

$$E(v, B_1) \equiv \frac{1}{2}\int_{B_1} |\nabla v|^2 + \frac{1}{4\varepsilon^2}\int_{B_1 \setminus B_\rho} (1 - |v|^2)^2 + \frac{1}{2\varepsilon^2}\int_{B_\rho} |v|^2 \tag{1.5}$$

and

$$H_g^1(B_1) \equiv \{v : B_1 \to \mathcal{C} | v \in H^1(B_1),\ v = g \text{ on } \partial B_1\}. \tag{1.6}$$

Our main result is the following theorem.

THEOREM. *Suppose that $g \in C^2(\partial B_1, S^1)$ and $deg(g, \partial B_1) = d > 0$. Given any sequence $\{\varepsilon_n, \rho_n\}_{n=1}^{+\infty}$ satisfying $\varepsilon_n \to 0$ and $\rho_n \to 0$ as $n \to +\infty$, let u_n be a solution of (1.4)–(1.6) with $\varepsilon = \varepsilon_n$ and $\rho = \rho_n$. Then, by passing to an appropriate subsequence, we have*

$$u_n \to u_* \text{ in } C^1_{\text{loc}}\left(\overline{B_1} \setminus \bigcup_{i=0}^{m} \{a_i\}\right) \quad \text{as } n \to +\infty,$$

where $a_0, a_1, \ldots, a_m$ are $m+1$ distinct points in B_1, $a_0 = 0$, $0 \le m \le d$ and u_* is a harmonic map in $\overline{B_1} \setminus (\cup_{j=0}^{m}\{a_j\})$ such that $deg(u_*, a_j) = 1 (j = 1, 2, \ldots, m)$, $deg(u_*, 0) = d - m$, and $u_* = g$ on ∂B_1.

Moreover, there is a constant $\mu \ge 1$ depending only on g such that

1° $\deg(u_*, 0) < k$ if $\lim\limits_{n \to +\infty} \dfrac{\varepsilon_n}{\rho_n^{2k-1}} \ge \mu,\ (k = 2, \ldots, d)$,

2° $\deg(u_*, 0) > k$ if $\overline{\lim\limits_{n \to +\infty}} \dfrac{\varepsilon_n}{\rho_n^{2k+1}} \le \dfrac{1}{\mu}, (k = 0, 1, \ldots, d-1)$.

In particular, we have the following vortex-pinning effect:

3° $\deg(u_*, 0) \ge 1$ if $\lim\limits_{n \to +\infty} \dfrac{\varepsilon_n}{\rho_n} = 0$,

4° $\deg(u_*, 0) = k$ if $\lim\limits_{n \to +\infty} \dfrac{\varepsilon_n}{\rho_n^{2k-1}} = 0$ and $\lim\limits_{n \to +\infty} \dfrac{\varepsilon_n}{\rho_n^{2k+1}} = +\infty$

$(k = 1, 2, \ldots, d-1)$,

5° $\deg(u_*, 0) = d$ if $\lim\limits_{n \to +\infty} \dfrac{\varepsilon_n}{\rho_n^{2d-1}} = 0$.

One can easily see from the arguments in the following sections that, if $\lim\limits_{n \to +\infty} \rho_n = \rho_0 > 0$, the theorem also holds with a_0 replaced by B_{ρ_0} and $\deg(u_*, 0)$ by $\deg(u_*, B_{\rho_0})$. Hence 5° implies $\deg(u_*, B_{\rho_0}) = d$ and $m = 0$. Note that our arguments can be applied for more general domains Ω and Ω_n.

We shall show some energy estimates in section 2 and prove the first part of the theorem above (i.e., the convergence of the solution u_n to u_* in $C^1_{\rm loc}(\overline{B_1}\setminus \cup_{j=0}^m \{a_j\})$ as $n \to +\infty$) in section 3. In the last section, i.e., section 4, we shall study the degrees of the singular points $\{a_0, a_1, \ldots, a_m\}$ and complete the proof of the theorem.

Throughout this paper, $x = (x_1, x_2) \equiv x_1 + ix_2$ denotes both a complex number and a point in $\mathbb{R}^2$, (r, θ) is the polar coordinate of $\mathbb{R}^2$, and the capital letter "C" denotes various constants which depend only on g. Moreover, for any domain $\Omega \subset \mathbb{R}^2$, we shall let

$$E(v, \Omega) \equiv \frac{1}{2}\int_{B_1\cap\Omega} |\nabla v|^2 + \frac{1}{4\varepsilon^2}\int_{(B_1\setminus B_\rho)\cap\Omega} (1-|v|^2)^2 + \frac{1}{2\varepsilon^2}\int_{B_\rho\cap\Omega} |v|^2. \tag{1.7}$$

2. Some energy estimates. By the standard theory of variational problems, (1.4)–(1.6) has at least one solution $u_{\varepsilon,\rho} \in C^1(\overline{B_1}) \cap C^2(\overline{B_1\setminus B_\rho}) \cap C^2(\overline{B_\rho})$ satisfying

$$-\Delta u = \frac{1}{\varepsilon^2} u(1-|u|^2) \quad \text{in } B_1\setminus B_\rho, \tag{2.1}$$

$$-\Delta u = -\frac{1}{\varepsilon^2} u \qquad \text{in } B_\rho, \tag{2.2}$$

$$u \text{ and } \nabla u \text{ are continuous across } \partial B_\rho, \tag{2.3}$$

and

$$u = g \qquad \text{on } \partial B_1. \tag{2.4}$$

It follows from (2.1)–(2.4) that

$$0 \le |u| \le 1 \quad \text{in } B_1 \tag{2.5}$$

and

$$|\nabla u| \le \frac{C}{\varepsilon} \qquad \text{in } B_1. \tag{2.6}$$

In fact, (2.5) can be proved by the maximum principle, and the proof of (2.6) can be found in [3] (see the Appendix in [3]).

In this section we shall show some energy estimates for the solution u which will be used in the next section.

LEMMA 2.1.

$$E(u, B_1) \le \pi j \log\frac{1}{\varepsilon} + \pi(d-j)^2 \log\frac{1}{\rho} + C\left[1 + \frac{\rho}{\varepsilon}\right] \tag{2.7}_j$$

for $j = 0, 1, \ldots, d$.

Proof. We can assume $\varepsilon < \frac{1}{4}$ by (2.5) and (2.6). For any integer $j (0 \le j \le d)$, choose j disjoint discs in $B_1\setminus B_{\frac{1}{2}}$: $\{B_{R_0}(x_k)\}_{k=0}^j$. Here R_0 is a small constant and $B_{R_0}(x_0) \equiv \phi$. Define a comparison function $v(x)$ in $\overline{B_1}$ by the following:

1° v is the harmonic map in $(B_1\setminus B_{\frac{1}{2}})\setminus(\bigcup_{k=0}^j B_{R_0}(x_k))$ satisfy

$$v\,|_{\partial B_1} = g, \quad v|_{\partial B_{\frac{1}{2}}} = e^{i(d-j)\theta}, \text{ and}$$

$$v|_{\partial B_{R_0}(x_k)} = \frac{x - x_k}{|x - x_k|}, \qquad k = 0, 1, \ldots, j,$$

2° $v = \dfrac{x - x_k}{|x - x_k|}$ in $B_{R_0}(x_k)\backslash B_{\varepsilon R_0}(x_k)$, $\qquad k = 0, 1, \ldots, j,$

3° v is a minimizer of the functional $E(v, B_{\varepsilon R_0}(x_k))$ in $H^1(B_{\varepsilon R_0}(x_k))$ with the boundary condition

$$v = \frac{x - x_k}{|x - x_k|} \text{ on } \partial B_{\varepsilon R_0}(x_k), \qquad k = 0, 1, \ldots, j,$$

where $E(v, B_{\varepsilon R_0}(x_k))$ is defined in (1.7).

4° $v = e^{i(d-j)\theta}$ in $B_{\frac{1}{2}}\backslash B_{\rho+\varepsilon}$,

5° $v = \dfrac{r - \rho}{\varepsilon} e^{i(d-j)\theta}$ in $B_{\rho+\varepsilon}\backslash B_\rho$,

6° $v \equiv 0$ in B_ρ.

Obviously, $v \in H^1_g(B_1)$. So $E(u, B_1) \le E(v, B_1)$.

It is easy to verify that

$$E(v, D) \le \pi j \log \frac{1}{\varepsilon} + C, \tag{2.8}$$

where $D \equiv (B_1\backslash B_{\frac{1}{2}})\backslash(\cup_{k=0}^{j} B_{\varepsilon R_0}(x_k))$,

$$E(v, B_{\frac{1}{2}}\backslash B_{\rho+\varepsilon}) \le \pi(d - j)^2 \log \frac{1}{\rho + \varepsilon} + C \tag{2.9}$$

and

$$E(v, B_{\rho+\varepsilon}) \le \pi(d - j)^2 \log \frac{\rho + \varepsilon}{\rho} + C\left[\frac{\rho}{\varepsilon} + 1\right]. \tag{2.10}$$

Set $x - x_k = \varepsilon y$ and $\widehat{v}(y) = v(x_k + \varepsilon y)$. Then

$$E(v, B_{\varepsilon R_0}(x_k)) = \int_{B_{R_0}} \left[\frac{1}{2}|\nabla \widehat{v}|^2 + (1 - |\widehat{v}|^2)^2\right] dy.$$

Hence, we have from the minimality of v (see the definition 3°) that

$$E(v, B_{\varepsilon R_0}(x_k)) \le C, \quad k = 0, 1, \ldots, j. \tag{2.11}$$

Thus (2.7) follows from (2.8)–(2.11). □

LEMMA 2.2.

$$E(u, B_{\gamma\rho}) \le \frac{C}{1 - \gamma}\left[1 + \frac{\varepsilon}{\rho} + \frac{\varepsilon}{\rho}\log\frac{1}{\rho}\right] \quad (0 < \gamma < 1). \tag{2.12}$$

Proof. It is easy to see from $(2.7)_0$ that there is a constant $\widehat{\gamma} \in (\gamma, 1)$ such that

$$\int_0^{2\pi} \left[|\nabla u(\widehat{\gamma}\rho, \theta)|^2 + \frac{1}{2\varepsilon^2}|u(\widehat{\gamma}\rho, \theta)|^2\right] d\theta \le \frac{C}{\rho^2(1 - \gamma^2)}\left[1 + \frac{\rho}{\varepsilon} + \log\frac{1}{\rho}\right]. \tag{2.13}$$

Multiplying (2.2) by u^* (the complex conjugate function of u) and then integrating over $B_{\widehat{\gamma}\rho}$, we find that

$$\int_{B_{\widehat{\gamma}\rho}} \left[|\nabla u|^2 + \frac{1}{\varepsilon^2}|u|^2\right] = \int_{\partial B_{\widehat{\gamma}\rho}} \frac{\partial u}{\partial n} u^*$$

$$\leq \left(\int_{\partial B_{\widehat{\gamma}\rho}} |\nabla u|^2\right)^{1/2} \left(\int_{\partial B_{\widehat{\gamma}\rho}} |u|^2\right)^{1/2}$$

$$\leq \frac{C\widehat{\gamma}\rho\varepsilon}{(1-\gamma^2)\rho^2}\left[\log\frac{1}{\rho} + \frac{\rho}{\varepsilon} + 1\right]$$

$$\leq \frac{C}{1-\gamma^2}\left[1 + \frac{\varepsilon}{\rho} + \frac{\varepsilon}{\rho}\log\frac{1}{\rho}\right].$$

Hence (2.12) follows. □

LEMMA 2.3.

$$E(u, B_1\backslash B_{\gamma\rho}) \leq \pi j \log\frac{1}{\varepsilon} + \pi(d-j)^2\log\frac{1}{\rho} + C_\gamma\left(1 + \log\frac{1}{\rho}\right), \tag{2.14}_j$$
$$\left(j = 0, 1, \ldots, d, \quad \gamma > 1, \quad \gamma\rho < \frac{1}{4}\right),$$

where $C_\gamma \equiv C(\frac{1}{\gamma^2-1} + \log\gamma + 1)$.

Proof. In the case $2\varepsilon \geq \rho$, (2.14) follows directly from (2.7). So we only need to prove (2.14) for $2\varepsilon < \rho < \frac{1}{4}$. Let

$$\bar{u}(r) = \left\{\frac{1}{2\pi}\int_0^{2\pi} |u(r,\theta)|^2 d\theta\right\}^{1/2} \quad \text{for } r \in [0,1]. \tag{2.15}$$

One can easily verify that

$$|\bar{u}(r)|^2 = \frac{1}{2\pi}\int_0^{2\pi} |u(r,\theta)|^2 d\theta \text{ for } r \in [0,1], \tag{2.16}$$

$$(1 - |\bar{u}(r)|^2)^2 \leq \frac{1}{2\pi}\int_0^{2\pi} (1 - |u(r,\theta)|^2)^2 d\theta, \qquad r \in [0,1], \tag{2.17}$$

and

$$\left|\frac{\partial\bar{u}(r)}{\partial r}\right|^2 \leq \frac{1}{2\pi}\int_0^{2\pi}\left|\frac{\partial u(r,\theta)}{\partial r}\right|^2 d\theta \text{ for a.e. } r \in [0,1]. \tag{2.18}$$

Since $(2.7)_0$ implies that there exists a constant $\widehat{\gamma} \in (1,\gamma)$ such that

$$\int_0^{2\pi}\left[\frac{1}{2}|\nabla u(\widehat{\gamma}\rho,\theta)|^2 + \frac{1}{4\varepsilon^2}(1 - |u(\widehat{\gamma}\rho,\theta)|^2)^2\right] d\theta$$

$$\leq \frac{C}{(\gamma^2-1)\rho^2}\left[\log\frac{1}{\rho} + \frac{\rho}{\varepsilon} + 1\right], \tag{2.19}$$

we have from (2.17) and (2.19) that

$$(1-|\bar{u}(\widehat{\gamma}\rho)|^2)^2 \le \frac{C}{\gamma^2-1}\left[\frac{\varepsilon}{\rho}+\frac{\varepsilon^2}{\rho^2}+\frac{\varepsilon^2}{\rho^2}\log\frac{1}{\rho}\right]. \tag{2.20}$$

Now define a function $\tilde{v}$ in $\overline{B_1}$ by the following:

1° $\tilde{v}=v\in\overline{B_1}\backslash B_{\frac{1}{2}}$,

where v is defined in the proof of Lemma 2.1.

2° $\tilde{v}=e^{i(d-j)\theta}$ in $\overline{B_{\frac{1}{2}}}\backslash B_\varepsilon$,

3° $\tilde{v}=\frac{r}{\varepsilon}e^{i(d-j)\theta}$ in B_ε.

It is obvious that

$$E(\tilde{v},B_{\frac{1}{2}}\backslash B_{\widehat{\gamma}\rho})\le \pi(d-j)^2\log\frac{1}{\rho}+C\log\gamma, \tag{2.21}$$

$$|\nabla\tilde{v}|\le\frac{C}{\rho},|\nabla^2\tilde{v}|\le\frac{C}{\rho^2}\text{ in }B_{\frac{1}{2}}\backslash B_{\frac{1}{2}\rho}, \tag{2.22}$$

and

$$|\tilde{v}|\le 1,|\nabla\tilde{v}|\le\frac{C}{\varepsilon}\text{ in }B_{\frac{1}{2}}. \tag{2.23}$$

Moreover, (2.8) and (2.11) give

$$E(\tilde{v},B_1\backslash B_{\frac{1}{2}})\le\pi j\log\frac{1}{\varepsilon}+C. \tag{2.24}$$

Set

$$w(x)=\begin{cases}\tilde{v} & \text{in } B_1\backslash B_{\widehat{\gamma}\rho+\varepsilon},\\ \left[\dfrac{1-\bar{u}(\widehat{\gamma}\rho)}{\varepsilon}(r-\widehat{\gamma}\rho)+\bar{u}(\widehat{\gamma}\rho)\right]\tilde{v} & \text{in } B_{\widehat{\gamma}\rho+\varepsilon}\backslash B_{\widehat{\gamma}\rho},\\ \bar{u}\tilde{v} & \text{in } B_{\widehat{\gamma}\rho}.\end{cases} \tag{2.25}$$

Then $w(x)\in H_g^1(B_1)$, and, consequently,

$$\begin{aligned}&E(u,B_1)\le E(w,B_1)\\ &=E(\tilde{v},B_1\backslash B_{\widehat{\gamma}\rho+\varepsilon})+E(w,B_{\widehat{\gamma}\rho+\varepsilon}\backslash B_{\widehat{\gamma}\rho})+E(w,B_{\widehat{\gamma}\rho}).\\ &E(w,B_{\widehat{\gamma}\rho+\varepsilon}\backslash B_{\widehat{\gamma}\rho})\le C\int_{B_{\widehat{\gamma}\rho+\varepsilon}\backslash B_{\widehat{\gamma}\rho}}\left[|\nabla\tilde{v}|^2+\frac{1}{\varepsilon^2}(1-\bar{u}(\widehat{\gamma}\rho))^2\right].\end{aligned} \tag{2.26}$$

By (2.20) and (2.22), we get after simple calculations that

$$E(w,B_{\widehat{\gamma}\rho+\varepsilon}\backslash B_{\widehat{\gamma}\rho})\le\frac{C}{\gamma^2-1}\left(1+\log\frac{1}{\rho}\right). \tag{2.27}$$

$$
\begin{aligned}
E(w, B_{\widehat{\gamma}\rho}) &= \int_{B_{\widehat{\gamma}\rho}} \frac{1}{2}|\tilde{v}\nabla\bar{u} + \bar{u}\nabla\tilde{v}|^2 + \frac{1}{4\varepsilon^2}\int_{B_{\widehat{\gamma}\rho}\setminus B_\rho}(1-|\bar{u}|^2)^2 \\
&\quad + \frac{1}{2\varepsilon^2}\int_{B_\rho}|\tilde{v}\bar{u}|^2 \\
&\le E(\bar{u}, B_{\widehat{\gamma}\rho}) + \frac{1}{2}\int_{B_{\widehat{\gamma}\rho}}|\bar{u}|^2|\nabla\tilde{v}|^2 \\
&\quad + \frac{1}{2}\int_{B_{\widehat{\gamma}\rho}}[\tilde{v}\bar{u}\nabla\bar{u}\cdot\nabla\tilde{v}^* + \tilde{v}^*\bar{u}\nabla\bar{u}\cdot\nabla\tilde{v}] \\
&\equiv E(\bar{u}, B_{\widehat{\gamma}\rho}) + I.
\end{aligned}
$$

Integrating by parts and using (2.22) and (2.23), we have

$$
\begin{aligned}
|I| &\le C\left\{\int_{B_{\widehat{\gamma}\rho}\setminus B_{\frac{1}{2}\rho}}|\Delta\tilde{v}| + \int_{\partial B_{\widehat{\gamma}\rho}\cup\partial B_{\frac{1}{2}\rho}}|\nabla\tilde{v}|\right. \\
&\quad \left. + \int_{B_{\widehat{\gamma}\rho}}|\bar{u}|^2|\nabla\tilde{v}|^2 + \int_{B_{\frac{1}{2}\rho}}|\nabla\bar{u}|^2\right\} \\
&\le C\left\{1 + \int_{B_{\widehat{\gamma}\rho}\setminus B_{\frac{1}{2}\rho}}|\nabla\tilde{v}|^2 + \int_{B_{\frac{1}{2}\rho}}\left[|\nabla\bar{u}|^2 + \frac{1}{\varepsilon^2}|\bar{u}|^2\right]\right\} \\
&\le C\left\{1 + \int_{B_{\frac{1}{2}\rho}}\left[|\nabla\bar{u}|^2 + \frac{1}{\varepsilon^2}|\bar{u}|^2\right]\right\}.
\end{aligned}
$$

By (2.15), (2.18), and Lemma 2.2, we get $|I| \le C$, therefore,

$$
E(w, B_{\widehat{\gamma}\rho}) \le E(\bar{u}, B_{\widehat{\gamma}\rho}) + C. \tag{2.28}
$$

Combining (2.26)–(2.28), (2.21), (2.24), and (2.16)–(2.18), we yield

$$
\begin{aligned}
E(u, B_1) &\le \pi j\log\frac{1}{\varepsilon} + \pi(d-j)^2\log\frac{1}{\rho} \\
&\quad + C\left(\frac{1}{\gamma^2-1} + \log\gamma + 1\right)\left(1 + \log\frac{1}{\rho}\right) + E(u, B_{\widehat{\gamma}\rho}).
\end{aligned}
$$

Hence, (2.14) follows. □

LEMMA 2.4.

$$
\frac{1}{\varepsilon^2}\int_{B_1\setminus B_{\gamma\rho\alpha}}(1-|u|^2) \le \frac{C}{1-\alpha}\left(\log\gamma + \frac{1}{\gamma^2-1} + 1\right), \tag{2.29}
$$

where $\gamma > 1$, $\gamma\rho < \frac{1}{4}$, $0 < \alpha < 1$.

Proof. Multiplying (2.1) by $x \cdot \nabla u^*$ and then integrating over $B_1 \backslash B_r (\rho < r < 1)$, after some elementary calculations we get

$$\begin{aligned} &\frac{1}{2}\int_{\partial B_1}\left|\frac{\partial u}{\partial \nu}\right|^2 + \frac{r}{2}\int_{\partial B_r}\left|\frac{\partial u}{\partial \tau}\right|^2 \\ &\quad + \frac{1}{4\varepsilon^2}\int_{B_1\backslash B_r}(1-|u|^2)^2 + \frac{r}{4\varepsilon^2}\int_{\partial B_r}(1-|u|^2)^2 \\ &= \frac{1}{2}\int_{\partial B_1}\left|\frac{\partial u}{\partial \tau}\right|^2 + \frac{r}{2}\int_{\partial B_r}\left|\frac{\partial u}{\partial \nu}\right|^2 + \frac{1}{4\varepsilon^2}\int_{\partial B_1}(1-|u|^2)^2 \\ &\qquad\qquad (\rho < r < 1), \end{aligned} \tag{2.30}$$

where ν is the unit normal vector and τ is the unit tangential vector (for details, see [2, Theorem III 2, p. 45]).

By Lemma 2.3 and the inequality

$$\frac{1}{2}\int_{\gamma\rho}^{\gamma\rho^\alpha}\frac{1}{r}\left(r\int_{\partial B_r}|\nabla u|^2\right)dr = \frac{1}{2}\int_{B_{\gamma\rho^\alpha}\backslash B_{\gamma\rho}}|\nabla u|^2 \le E(u, B_1\backslash B_{\gamma\rho}),$$

we can find a $\widehat{\gamma} \in (\gamma\rho, \gamma\rho^\alpha)$ such that

$$\widehat{\gamma}\int_{\partial B_{\widehat{\gamma}}}|\nabla u|^2 \le \frac{E(u, B_1\backslash B_{\gamma\rho})}{(1-\alpha)|\log\rho|} \le \frac{C}{1-\alpha}\left(\frac{1}{\gamma^2-1} + \log\gamma + 1\right). \tag{2.31}$$

Thus (2.29) follows from (2.30) and (2.31). □

Given any constant $\eta \in (0, \frac{1}{2})$, consider the following variational problem: find a function $\phi \in H_u^1(B_\eta\backslash B_\xi)$, subject to (s.t.)

$$E(\phi, B_\eta\backslash B_\xi) = \min_{v\in H_u^1(B_\eta\backslash B_\xi)} E(v, B_\eta\backslash B_\xi), \tag{2.32}$$

where $\xi = 2\max\{\rho, \varepsilon\}$, $\xi < \eta/2$, u is a solution of (1.4)–(1.6), $E(v, B_\eta\backslash B_\xi)$ is defined in (1.7), i.e.,

$$E(v, B_\eta\backslash B_\xi) \equiv \int_{B_\eta\backslash B_\xi}\left[\frac{1}{2}|\nabla v|^2 + \frac{1}{4\varepsilon^2}(1-|v|^2)^2\right], \tag{2.33}$$

and

$$H_u^1(B_\eta\backslash B_\xi) \equiv \{v : \overline{B_\eta\backslash B_\xi} \to \mathcal{C} | v \in H^1(B_\eta\backslash B_\xi), v = u \text{ on } \partial B_\eta\}. \tag{2.34}$$

By the standard theory, (2.32) has at least one solution $\phi \in C^2(\overline{B_\eta\backslash B_\xi})$ satisfying

$$-\Delta\phi = \frac{1}{\varepsilon^2}\phi(1-|\phi|^2) \quad \text{in } B_\eta\backslash B_\xi, \tag{2.35}$$

$$\phi = u \qquad \text{on } \partial B_\eta, \tag{2.36}$$

$$\frac{\partial\phi}{\partial\nu} = 0 \qquad \text{on } \partial B_\xi. \tag{2.37}$$

LEMMA 2.5. *Let ϕ be a solution of* (2.32); *then*

$$E(\phi, B_\eta \backslash B_\xi) \le C \log \frac{1}{\xi} \le C \log \frac{1}{\varepsilon}, \tag{2.38}$$

$$|\phi| \le 1 \qquad \text{in } B_\eta \backslash B_\xi, \tag{2.39}$$

$$|\nabla \phi| \le \frac{C}{\varepsilon} \qquad \text{in } B_\eta \backslash B_\xi, \tag{2.40}$$

and

$$\frac{1}{\varepsilon^2} \int_{B_{\xi^\alpha} \backslash B_\xi} (1 - |\phi|^2)^2 \le \frac{C}{1-\alpha} (0 < \alpha < 1,\ \xi^\alpha < \eta). \tag{2.41}$$

Moreover, if

$$\frac{1}{2} \int_{\partial B_\eta} |\nabla u|^2 + \frac{1}{4\varepsilon^2} \int_{\partial B_\eta} (1 - |u|^2)^2 \le \frac{C}{\eta} \log \frac{1}{\xi}, \tag{2.42}$$

then

$$\frac{1}{\varepsilon^2} \int_{B_{\varepsilon^\alpha}(x_0) \cap (B_\eta \backslash B_\xi)} (1 - |\phi|^2)^2 \le \frac{C}{\alpha(1-\alpha)} \tag{2.43}$$

for $0 < \alpha < 1$, $\quad x_0 \in \overline{B_\eta \backslash B_\xi}$, $\quad$ *and* $\quad \varepsilon^{\alpha/2} \le \eta$.

Proof. Since $\xi = 2 \max\{\varepsilon, \rho\}$, (2.38) follows from Lemmas 2.1 and 2.3 and the minimality of ϕ. (2.39) follows from (2.35)–(2.37), (2.5), and the maximum principle.

Set $\widehat{\phi}(y) = \phi(\varepsilon y)$. Then $\widehat{\phi}$ solves

$$\begin{cases} -\Delta \widehat{\phi} = \phi(1 - |\phi|^2) & \text{for } \dfrac{\xi}{\varepsilon} < |y| < \dfrac{\eta}{\varepsilon}, \\ \dfrac{\partial \widehat{\phi}}{\partial \nu} = 0 & \text{on } \ |y| = \dfrac{\xi}{\varepsilon}, \\ |\widehat{\phi}| \le 1 & \text{for } \dfrac{\xi}{\varepsilon} \le |y| \le \dfrac{\eta}{\varepsilon}. \end{cases}$$

Noting that $\frac{\xi}{\varepsilon} \ge 2$ and $\frac{\eta}{\varepsilon} - \frac{\xi}{\varepsilon} \ge \frac{\xi}{\varepsilon} \ge 2$, we can apply the local estimates of elliptic equations to $\widehat{\phi}$ in the set $\{\frac{\xi}{\varepsilon} \le |y| \le \frac{\xi}{\varepsilon} + 1\}$ and obtain

$$|\nabla \widehat{\phi}| \le C \ \text{ for } \ \frac{\xi}{\varepsilon} \le |y| \le \frac{\xi}{\varepsilon} + 1. \tag{2.44}$$

Similarly, set $\widetilde{\phi}(y) = \widehat{\phi}(y) - u(\varepsilon y)$. Then $\widetilde{\phi}$ satisfies

$$\begin{cases} -\Delta \widetilde{\phi} = -u(1 - |u|^2) + \phi(1 - |\phi|^2) \equiv f, & \dfrac{\xi}{\varepsilon} < |y| < \dfrac{\eta}{\varepsilon}, \\ \widetilde{\phi} = 0 & \text{for } |y| = \dfrac{\eta}{\varepsilon}, \\ |\widetilde{\phi}| \le 2, |f| \le 2 & \text{for } \dfrac{\xi}{\varepsilon} \le |y| \le \dfrac{\eta}{\varepsilon}. \end{cases}$$

By applying local estimates of elliptic equations to $\widetilde{\phi}$ in $\frac{\eta}{\varepsilon} - 1 \le |y| \le \frac{\eta}{\varepsilon}$, we get

$$|\nabla \widetilde{\phi}| \le C \qquad \text{for } \frac{\eta}{\varepsilon} - 1 \le |y| \le \frac{\eta}{\varepsilon},$$

which, together with (2.6), gives

$$|\nabla\widehat{\phi}| \le C \quad \text{for } \frac{\eta}{\varepsilon} - 1 \le |y| \le \frac{\eta}{\varepsilon}. \tag{2.45}$$

Now by (2.44), (2.45), and the interior estimates of elliptic equations, we yield

$$|\nabla\widehat{\phi}| \le C \quad \text{for } \frac{\xi}{\varepsilon} \le |y| \le \frac{\eta}{\varepsilon}. \tag{2.46}$$

Recalling $\widehat{\phi}(y) = \phi(\varepsilon y)$, we get (2.40).

(2.41) can be proved in terms of (2.37), (2.38), and an argument similar to that in the proof of Lemma 2.4.

If (2.42) holds, then, for any $x_0 \in \overline{B_\eta \backslash B_\xi}$,

$$(x - x_0)\cdot\nu \ge 0 \qquad \text{on } \partial B_\eta$$

and

$$\int_{\partial B_\eta \cap B_{\varepsilon^\alpha}(x_0)} |x - x_0| \cdot \left[\frac{1}{2}\left|\frac{\partial u}{\partial \tau}\right|^2 + \frac{1}{4\varepsilon^2}(1 - |u|^2)^2\right] dS$$

$$\le \frac{C\varepsilon^\alpha}{\eta}\log\frac{1}{\xi} \le C\varepsilon^{\frac{\alpha}{2}}\log\frac{1}{\varepsilon} \le \frac{C}{\alpha}.$$

Thus an argument similar to that in the proof of Lemma 2.4 gives (2.43) when $B_{\varepsilon^\alpha}(x_0) \cap \partial B_\eta \ne \emptyset$ (see [4]). Notice that (2.43) follows directly from (2.41) when $B_{\varepsilon^\alpha}(x_0) \cap \partial B_\xi \ne \emptyset$. Hence (2.43) is proved. □

LEMMA 2.6. *Assume* (2.42),

$$|u| \ge \frac{1}{2} \quad \textit{in } B_{2\eta}\backslash B_\eta, \tag{2.47}$$

and

$$|\deg(u, \partial B_\eta| \le C. \tag{2.48}$$

Then there are finite integers $\delta_0, \delta_1, \ldots, \delta_N$ *and a positive constant* $\varepsilon_0 = \varepsilon_0(\eta)$ *such that*

$$\sum_{k=0}^{n} \delta_k = \deg(u, \partial B_\eta), \tag{2.49}$$

$$0 \le N \le C, \quad |\delta_k| \le C \quad (k = 0, 1, \ldots, N), \tag{2.50}$$

and

$$E(u, B_{2\eta}\backslash B_\xi) \equiv \int_{B_{2\eta}\backslash B_\xi}\left[\frac{1}{2}|\nabla u|^2 + \frac{1}{4\varepsilon^2}(1 - |u|^2)^2\right]$$

$$\ge \pi\delta_0^2 \log\frac{\eta}{\xi} + \pi\left(\sum_{k=1}^{N}\delta_k^2\right)\log\frac{1}{\varepsilon} - C \qquad \textit{for } \varepsilon \le \varepsilon_0, \tag{2.51}$$

where u *is a solution of* (1.4)–(1.6), $\xi = 2\max\{\varepsilon, \rho\}$, *and* $\xi \le \eta/2$.

Proof. In terms of Lemma 2.5 and the arguments in [2, Chapter IV] and in [4], there are a positive constant $\lambda(\lambda \le C)$ and a finite collection of disjoint discs $\{B_{\lambda\varepsilon}(x_k)\}_{k=1}^N$ satisfying

$$x_k \in \overline{B_\eta \backslash B_\xi}, \quad k = 1, 2, \ldots, N, \quad N \le C, \tag{2.52}$$

$$|x_i - x_j| \ge 8\lambda\varepsilon, \quad 1 \le i < j \le N, \tag{2.53}$$

$$|\phi(x_k)| < \frac{1}{2}, \quad k = 1, 2, \ldots, N, \tag{2.54}$$

$$|\phi(x)| \ge \frac{1}{2} \qquad \text{in } (B_\eta \backslash B_\xi) \left\backslash \left(\bigcup_{k=1}^N B_{\lambda\varepsilon}(x_k) \right) \right., \tag{2.55}$$

and

$$|\kappa_k| \le C, \quad \kappa_k \equiv \deg(\phi, \partial B_{\lambda\varepsilon}(x_k)), \quad k = 1, \ldots, N. \tag{2.56}$$

Because of $N \le C$ and $\varepsilon \le \frac{\xi}{2}$, we can assume without loss of generality that

$$|x_k| \ge 8\xi \quad k = 1, 2, \ldots, N. \tag{2.57}$$

By (2.48) and (2.55)–(2.57), $\kappa_0 \equiv \deg(\phi, \partial B_\xi)$ is well defined and

$$\sum_{k=0}^N \kappa_k = \deg(u, \partial B_\eta), \quad |\kappa_k| \le C, \quad k = 0, 1, \ldots, N. \tag{2.58}$$

Extend ϕ to B_1 by $\phi = u$ in $B_1 \backslash B_\eta$ and let

$$\phi_0 \equiv \sum_{k=0}^N \left(\frac{x - x_k}{|x - x_k|} \right)^{\kappa_k}. \tag{2.59}$$

Combining the arguments in [2, Appendix IV, Theorem A.6] and [5], we can deduce from Lemma 2.5, (2.47), (2.48), and (2.52)–(2.57) that

$$\int_\Omega |\nabla\phi|^2 \ge \int_\Omega |\nabla\phi_0|^2 - C, \tag{2.60}$$

where $\Omega \equiv B_{2\eta} \backslash (\bigcup_{k=1}^N B_{\lambda\varepsilon}(x_k) \cup B_\xi)$. For the convenience of our readers, we shall prove (2.60) in the Appendix of this paper.

Without loss of generality, suppose $\deg(u, \partial B_\eta) \ge 0$ (otherwise, one can replace $u(r, \theta)$ by $u(r, -\theta)$). By repeating the proof of Theorem II.1 in [2], we have from Corollary II.1 in [2] that

$$\frac{1}{2} \int_\Omega |\nabla\phi_0|^2 \ge \pi \min_{(\delta_0, \delta_1, \ldots, \delta_N) \in P} \left\{ \delta_0^2 \log \frac{\eta}{\xi} + \left(\sum_{k=1}^N \delta_k^2 \right) \log \frac{1}{\varepsilon} \right\} - C, \tag{2.61}$$

where

$$P \equiv \left\{ (\delta_0, \delta_1, \ldots, \delta_N) \,|\, \delta_k \text{ is an integer}, \quad k = 0, 1, \ldots, N, \right.$$

$$\left. 0 \le \delta_k \le \frac{1}{2} \left(|\kappa_k| + \kappa_k \right), \quad k = 0, 1, \ldots, N, \sum_{k=0}^N \delta_k = \deg(u, \partial B_\eta) \right\}.$$

Now (2.51) follows from (2.60), (2.61), and the inequality $E(u, B_{2\eta}\backslash B_\xi) \geq E(\phi, B_{2\eta}\backslash B_\xi)$. □

3. Convergence of the solutions. Owing to (2.5), (2.6), Lemma 2.4, and the argument in [2, Chapter IV], there is a finite collection of disjoint discs $\{B_{\lambda\varepsilon}(x_k)\}_{k=0}^{\widehat{m}}$; $B_{\lambda\varepsilon}(x_0) \equiv \emptyset$; $x_k \in B_1\backslash B_{3\xi^{\frac{1}{2}}}$, $k = 1, 2, \ldots, \widehat{m}$, where $\xi = 2\max\{\varepsilon, \rho\}$; and λ is a positive constant depending only on g such that

$$|u(x_k)| < \frac{1}{2}, \quad k = 1, 2, \ldots, \widehat{m}, \tag{3.1}$$

$$|u(x)| \geq \frac{1}{2} \quad \text{in } (B_1\backslash B_{3\xi^{\frac{1}{2}}}) \left\backslash \left(\bigcup_{k=0}^{\widehat{m}} B_{\lambda\varepsilon}(x_k) \right) \right., \tag{3.2}$$

$$0 \leq \widehat{m} \leq C, \quad |\deg(u, \partial B_{\lambda\varepsilon}(x_k))| \leq C, \quad k = 1, \ldots, \widehat{m}. \tag{3.3}$$

For any sequence $\{\varepsilon_n, \rho_n\}_{n=1}^{+\infty}$, $\varepsilon_n \to 0$, $\rho_n \to 0$ as $n \to +\infty$, by passing to an appropriate subsequence, we can assume that the limit set of $\{x_k\}_{k=0}^{\widehat{m}}(x_k \equiv x_k(n)$, $\widehat{m} \equiv \widehat{m}(n))$ in $\overline{B_1}\backslash\{0\}$ as $n \to +\infty$ is given by $\{a_1, a_2, \ldots, a_m\} 0 \leq m \leq C$ (by $m = 0$ denote the limit set is empty).

Let $a_0 = 0$ and $0 < \eta \leq \frac{1}{4} \min\limits_{0 \leq i \leq j \leq m} |a_i - a_j|$. Then it follows from (3.2) that

$$|u_n| \geq \frac{1}{2} \text{ in } B_1 \left\backslash \left(\bigcup_{k=0}^{m} B_\eta(a_k) \right) \right. \tag{3.4}$$

for n large enough, where u_n is a solution of (1.4)–(1.6) with $\varepsilon = \varepsilon_n$ and $\rho = \rho_n$. Thus $d_k \equiv \deg(u_n, \partial B_\eta(a_k))(k = 0, 1, \ldots, m)$ are well defined and, by (3.3),

$$|d_k| \leq C, \quad k = 0, 1, \ldots, m, \sum_{k=0}^{m} d_k = d, \tag{3.5}$$

where $d \equiv \deg(g, \partial B_1) > 0$. Therefore, by passing to a further subsequence, we can assume d_k $(k = 0, 1, \ldots, m)$ are independent of n.

LEMMA 3.1.

$$E(u_n, B_{\gamma\rho_n}) \leq C(1 + \log\gamma)\left(\frac{\rho_n}{\varepsilon_n} + 1\right) (\gamma > 1, \quad \gamma\rho_n < 4\eta). \tag{3.6}$$

Proof. By Lemmas 2.1 and 2.3, we can find a constant $\widehat{\eta} \in [\frac{1}{2}\ \eta, \eta]$ such that

$$\frac{1}{2}\int_{\partial B_{\widehat{\eta}}} |\nabla u_n|^2 + \frac{1}{4\varepsilon^2}\int_{\partial B_{\widehat{\eta}}} (1 - |u_n|^2) \leq \frac{C}{\widehat{\eta}} \log\frac{1}{\xi}, \tag{3.7}$$

where $\xi = 2\max\{\varepsilon_n, \rho_n\}$. So it follows from Lemma 2.6 that

$$E(u_n, B_{2\eta}\backslash B_{\gamma\xi}) \geq \pi\delta_0^2 \log\frac{\eta}{\xi} + \pi\left(\sum_{k=1}^{N} \delta_k^2\right) \log\frac{\eta}{\varepsilon_n} - C(\log\gamma + 1) \tag{3.8}$$

for some integers $\delta_0, \delta_1, \ldots, \delta_N$ with $\sum_{k=0}^{N} \delta_k = d_0$ and $0 \leq \delta_k \leq C(k = 0, 1, \ldots, N)$.

By Lemma 2.4 and (3.4) we have

$$E(u_n, B_{2\eta}(a_k)) \geq \pi|d_k| \log\ \frac{\eta}{\varepsilon_n} - C, \quad k = 1, \ldots, m, \tag{3.9}$$

(see [2, Theorem V.2, p. 53]).

Now we yield from (3.8), (3.9), and Lemma 2.1 that

$$\begin{aligned} &E(u_n, B_{\gamma\xi}) \\ &\le \pi j \log\frac{1}{\varepsilon_n} + \pi(d-j)^2 \log\frac{1}{\rho_n} + C\left(1+\frac{\rho_n}{\varepsilon_n}\right) \\ &\quad -\pi\left(\sum_{k=1}^{m}|d_k|\right)\log\frac{\eta}{\varepsilon_n} - \pi\left(\sum_{k=1}^{N}\delta_k^2\right)\log\frac{\eta}{\varepsilon_n} \\ &\quad -\pi\delta_0^2\log\frac{\eta}{\xi} + C(1+\log\gamma). \end{aligned} \tag{3.10}$$

Suppose $\varepsilon_n \ge \rho_n$. Then $\xi = 2\varepsilon_n$. Letting $j = d$ in (3.10), we get

$$\begin{aligned} E(u_n, B_{\gamma\xi}) &\le \pi\left(d - \sum_{k=1}^{m}|d_k| - \sum_{k=0}^{N}\delta_k^2\right)\log\frac{1}{\varepsilon_n} \\ &\quad +\pi\left(\sum_{k=0}^{N}\delta_k^2 + \sum_{k=1}^{m}|d_k|\right)\log\frac{1}{\eta} + C(1+\log\gamma)\left(1+\frac{\rho_n}{\varepsilon_n}\right). \end{aligned} \tag{3.11}$$

Since

$$d - \left(\sum_{k=1}^{m}|d_k| + \sum_{k=0}^{N}\delta_k^2\right) \le d - \sum_{k=0}^{m} d_k = d - d = 0,$$

we yield (3.6) for $\varepsilon_n \ge \rho_n$ by taking $\eta = \eta_0 \equiv \frac{1}{4}\min_{0\le i\le j\le m}|a_i - a_j|$.

Suppose $\varepsilon_n \le \rho_n$. Then $\xi = 2\rho_n$. If $|\delta_0| > d$, let $j = 0$ in (3.10); if $|\delta_0| \le d$, let $j = d - |\delta_0|$ in (3.10). In both cases, we can get (3.6) for $\varepsilon_n \ge \rho_n$. □

LEMMA 3.2.

$$\begin{aligned} &E(u_n, B_1\backslash B_{\gamma\rho_n}) \le \pi j\log\frac{1}{\varepsilon_n} + \pi(d-j)^2\log\frac{1}{\rho_n} + \frac{C}{\gamma^2-1}(1+\log\gamma) \\ &\qquad (j = 0, 1, \dots, d, \gamma > 1). \end{aligned} \tag{3.12}$$

Proof. Obviously, we can suppose $\gamma\rho_n < 4\eta$. There is a constant $\widehat{\gamma} \in (1, \gamma)$ from (3.6) such that

$$\begin{aligned} &\int_0^{2\pi}\left[\frac{1}{2}|\nabla u_n(\widehat{\gamma}\rho,\theta)|^2 + \frac{1}{4\varepsilon_n^2}(1-|u_n(\widehat{\gamma}\rho,\theta)|^2)\right]d\theta \\ &\le \frac{C(1+\log\gamma)}{(\gamma^2-1)\rho^2}\left(1+\frac{\rho_n}{\varepsilon_n}\right). \end{aligned} \tag{3.13}$$

Replacing (2.19) by (3.13) and repeating the proof of Lemma 2.3, we can deduce (3.12). □

LEMMA 3.3.

$$E\left(u_n, B_1\backslash\bigcup_{k=0}^{m}B_{2\eta}(a_k)\right) \le \pi\left(\sum_{k=0}^{N}\delta_k^2 + \sum_{k=1}^{m}|d_k|\right)\log\frac{1}{\eta} + C, \tag{3.14}$$

where $0 < \eta \le \frac{1}{4}\min_{0\le i\le j\le m}|a_i - a_j|$.

Proof. (3.14) follows from (3.8), (3.9), and (3.12) in terms of the argument used in the proof of Lemma 3.1. □

From (3.14) and a result in [3] there exists a harmonic function u_* defined in $\overline{B_1}\backslash\bigcup_{k=0}^m\{a_k\}$ such that

$$u_n \to u_* \quad \text{in } C^1_{\text{loc}}\left(\overline{B_1}\backslash\bigcup_{k=0}^m\{a_k\}\right) \qquad \text{as } n\to+\infty. \tag{3.15}$$

4. The degrees of the singular points.

LEMMA 4.1.

$$d_0 \geq 0, \quad 0\leq m\leq d, \tag{4.1}$$

$$d_k = 1, \quad a_k\in B_1\backslash\{0\}, \qquad k=1,2,\dots,m \quad \text{if } m\geq 1. \tag{4.2}$$

Proof. Because $\sum_{k=0}^N \delta_k^2\leq C$, we can assume $\sum_{k=1}^N\delta_k^2$ and δ_0 are independent of n (by passing to a further subsequence).

Since $a_k(k=1,2,\dots,m)$ are the limit points of $\{x_k\}_{k=1}^{\widehat{m}}(x_k\equiv x_k(n),\ \widehat{m}=\widehat{m}(n))$ as $n\to+\infty$, (3.1) and (3.14) imply

$$d_k\neq 0, \quad k=1,\dots,m, \tag{4.3}$$

(see [2, Step 1 of the proof of Theorem VI. 2, p. 61]).

It follows from (3.14) and (3.15) that

$$d_k = \deg(u_*,\partial B_\eta), \quad k=0,1,\dots,m, \tag{4.4}$$

and

$$\frac{1}{2}\int_{B_1\backslash\bigcup_{k=0}^m B_\eta(a_k)}|\nabla u_*|^2\leq \pi\left(\sum_{k=0}^N\delta_k^2+\sum_{k=1}^m|d_k|\right)\log\frac{1}{\eta}+C, \tag{4.5}$$

where $0<\eta\leq\frac{1}{8}\min\limits_{0\leq i<j\leq m}|a_i-a_j|$.

Set $I\equiv\{k|0\leq k\leq m, a_k\in\partial B_1\}$ and $I'=\{0,1,\dots,m\}\backslash I$. Then we have from [2, Lemma VI.1, p. 63 and Corollary II.2, p. 33] that

$$\frac{1}{2}\int_{(B_1\cap B_{\eta_0}(a_k))\backslash B_\eta(a_k)}|\nabla u_*|^2\geq 2\pi|d_k|\log\frac{1}{\eta}-C \quad \text{for } k\in I \tag{4.6}$$

and

$$\frac{1}{2}\int_{B_{\eta_0}(a_k)\backslash B_\eta(a_k)}|\nabla u_*|^2\geq\pi|d_k|^2\log\frac{1}{\eta}-C \qquad \text{for } k\in I'; \tag{4.7}$$

here, $\eta_0\equiv\frac{1}{4}\min\limits_{0\leq i<j<m}|a_i-a_j|$.

Combining (4.5)–(4.7), we have

$$\left(2\sum_{k\in I}|d_k|+\sum_{k\in I'}|d_k|^2\right)\log\frac{1}{\eta}\leq\left(\sum_{k=0}^N\delta_k^2+\sum_{k=1}^m|d_k|\right)\log\frac{1}{\eta}+C. \tag{4.8}$$

Taking η so small that $C/\log\frac{1}{\eta} \le \frac{1}{2}$, we get from (4.8) that

$$2\sum_{k\in I}|d_k| + \sum_{k\in I'}|d_k|^2 \le \sum_{k=0}^{N}\delta_k^2 + \sum_{k=1}^{m}|d_k|. \tag{4.9}$$

It is obvious from (3.8), (3.9), and Lemma 3.2 that

$$\begin{aligned}&\left(\sum_{k=1}^{m}|d_k| + \sum_{k=1}^{N}\delta_k^2\right)\log\frac{1}{\varepsilon_n} + \delta_0^2\log\frac{1}{\xi}\\ &\le j\log\frac{1}{\varepsilon_n} + (d-j)^2\log\frac{1}{\rho_n} + C, \quad j = 0,1,\dots,d.\end{aligned} \tag{4.10}$$

Suppose $\xi \equiv 2\max\{\varepsilon_n,\rho_n\} = 2\varepsilon_n \ge 2\rho_n$. Then we get from (4.9) and (4.10) for $j = d$ that

$$\left(2\sum_{k\in I}|d_k| + \sum_{k\in I'}|d_k|^2 - \sum_{k=0}^{m}d_k\right) \le \frac{C}{|\log\varepsilon_n|};$$

therefore, for n large enough,

$$\sum_{k\in I}(2|d_k| - d_k) + \sum_{k\in I'}(|d_k|^2 - d_k) \le 0, \tag{4.11}$$

which, together with (4.3), gives $I = \emptyset$, $0 \le d_0 \le 1$, and $d_k = 1 (k = 1,2,\dots,m)$. Thus (4.1) and (4.2) are true if $\varepsilon_n \ge \rho_n$.

Suppose $\xi \equiv 2\max\{\varepsilon_n,\rho_n\} = 2\rho_n \ge 2\varepsilon_n$. Then we find by taking $j = d$ in (4.10) and letting $n \to +\infty$ that

$$\sum_{k=1}^{m}|d_k| \le d; \tag{4.12}$$

consequently,

$$d_0 = d - \sum_{k=1}^{m}d_k \ge d - \sum_{k=1}^{m}d_k \ge 0. \tag{4.13}$$

Let $j = \sum_{k=1}^{m}|d_k|$; then $0 \le j \le d$ and, by (4.10),

$$\sum_{k=0}^{N}\delta_k^2 \le \left(d - \sum_{k=1}^{m}|d_k|\right)^2 \le d_0^2.$$

In particular, we have

$$|\delta_0| \le d_0. \tag{4.14}$$

So we can let $j = d - |\delta_0|$ in (4.10), and by noting (4.9) we get

$$2\sum_{k\in I}|d_k| + \sum_{k\in I'}|d_k|^2 - \delta_0^2 \le d - |\delta_0|,$$

that is,

$$\begin{aligned}\sum_{k\in I}(2|d_k|-d_k)+\sum_{k\in I'\setminus\{0\}}(|d_k|^2-d_k)&\\ \le\ \delta_0^2-d_0^2+d_0-|\delta_0|&\\ =\ (d_0-|\delta_0|)(1-d_0-|\delta_0|)\le 0;&\end{aligned}$$

consequently, $I=\phi$ and $d_k=1$ $(k=1,2,\dots,m)$. Thus (4.1) and (4.2) are also true when $\varepsilon_n\le\rho_n$ for n large enough. □

LEMMA 4.2. *There is a constant $\mu\ge 1$, depending only on g, such that for n large enough*

$$d_0<k \quad \text{if } \varepsilon_n\ge\mu\rho_n^{2k-1} \quad (k=2,\dots,d) \tag{4.15}$$

and

$$d_0>k \quad \text{if } \varepsilon_n\le\frac{1}{\mu}\rho_n^{2k+1} \quad (k=0,1,\dots,d-1). \tag{4.16}$$

In particular, for n large enough,

$$d_0\ge 1 \quad \text{if } \varepsilon_n\le\frac{1}{\mu}\rho_n, \tag{4.17}$$

$$d_0=k \quad \text{if } \mu\rho_n^{2k+1}\le\varepsilon_n\le\frac{1}{\mu}\rho_n^{2k-1} \quad (k=1,2,\dots,d-1), \tag{4.18}$$

and

$$d_0=d \quad \text{if } \varepsilon_n\le\frac{1}{\mu}\rho_n^{2d-1}. \tag{4.19}$$

Proof. Suppose $\varepsilon_n\ge\rho_n$ for n large enough. Then $0\le d_0\le 1$ by (4.11) and, therefore, (4.15) holds. Note that the condition of (4.16) does not hold when $\varepsilon_n\ge\rho_n$ for n large enough.

Suppose $\varepsilon_n\le\rho_n$ for n large enough. Then by (4.9) and (4.10) we have

$$(d-d_0)\log\frac{1}{\varepsilon_n}+d_0^2\log\frac{1}{\rho_n}<j\log\frac{1}{\varepsilon_n}+(d-j)^2\log\frac{1}{\rho_n}+\log\mu,$$

where $\mu=e^C$. Hence

$$(d_0-d+j)(d_0+d-j)\log\frac{1}{\rho_n}<(d_0-d+j)\log\frac{1}{\varepsilon_n}+\log\mu. \tag{4.20}$$

Suppose $d_0\ge k$ for $k=2,\dots,d$. Letting $j=d-d_0+1$ in (4.20), we find

$$(2d_0-1)\log\frac{1}{\rho_n}<\log\frac{1}{\varepsilon_n}+\log\mu,$$

which is followed by

$$(2k-1)\log\frac{1}{\rho_n}<\log\frac{1}{\varepsilon_n}+\log\mu,$$

that is

$$\varepsilon_n < \mu\rho_n^{2k-1}.$$

Thus (4.15) is true.

Suppose $d_0 \le k$ for $k = 0, 1, \ldots, d-1$. Letting $j = d - d_0 - 1$ in (4.20), then

$$(2d_0+1)\log\frac{1}{\rho_n} > \log\frac{1}{\varepsilon_n} - \log\mu,$$

which is followed by

$$\varepsilon_n > \frac{1}{\mu}\rho_n^{2d_0+1} \ge \frac{1}{\mu}\rho_n^{2k+1}.$$

So, (4.16) is also true.

(4.17)–(4.19) are direct consequences of (4.15) and (4.16). □

Now the proof of the theorem stated in section 1 is completed.

Appendix. Let B_R be the disc of radius R centered at 0. Let $p_1, p_2, \ldots, p_m$ be points in B_R such that

$$4R_1 \le |p_j| \le R/2, \quad j = 1, 2, \ldots, m, \tag{A.1}$$

and

$$|p_j - p_k| \ge 4R_0, \quad 1 \le k < j \le m, \quad 0 < R_0 \le R_1 \le 1. \tag{A.2}$$

Set

$$\Omega \equiv B_R \backslash \left(\bigcup_{j=1}^{m} B_{R_0}(p_j) \cup B_{R_1} \right). \tag{A.3}$$

Let u be a smooth map from Ω to $\mathcal{C}$ and assume

$$0 < a \le |u| \le 1 \quad \text{in } \Omega, \tag{A.4}$$

$$\frac{1}{R_0^2}\int_\Omega (|u|^2 - 1)^2 \le K \log\frac{1}{R_1}, \tag{A.5}$$

$$\frac{1}{R_0^2}\int_{\Omega \cap B_{R_0^{\frac{1}{2}}}(p_j)} (|u|^2 - 1)^2 \le K, \tag{A.6}$$

and

$$\frac{1}{R_1^2}\int_{B_{R_1^{\frac{1}{2}}}} (|u|^2 - 1)^2 \le K. \tag{A.7}$$

Set

$$u_0(x) = \prod_{j=0}^{m} \left(\frac{x - p_j}{|x - p_j|} \right)^{d_j},$$

where $p_0 = 0$, $d_0 = \deg(u, \partial B_{R_1})$, and $d_j = \deg(u, \partial B_{R_0}(p_j))$, $j = 1, 2, \ldots, m$.

THEOREM A. *Assume* (1)–(7)*; then*

$$\int_\Omega |\nabla u|^2 \geq \int_\Omega |\nabla u_0|^2 - C, \tag{A.8}$$

where C depends only on a, m, K, and d_j $(j = 0, 1, \ldots, m)$.

Proof. Set $\rho = |u|$. There is a well-defined single-valued function $\psi : B_R \to \mathbb{R}$ such that

$$u = \rho u_0 e^{i\psi} \quad \text{in } \Omega \tag{A.9}$$

and

$$\int_{B_R} |\nabla \psi|^2 \leq C \int_\Omega |\nabla \psi|^2, \tag{A.10}$$

where C is some universal constant (see [2, Appendix IV, Lemma A.1]). Write

$$u_0 = e^{i\varphi_0}, \varphi_0 = \sum_{j=0}^{m} d_j \theta_j, \quad \theta_j = \operatorname{Arg}\left(\frac{x - p_j}{|x - p_j|}\right). \tag{A.11}$$

Let $r_j = |x - p_j|$. Then

$$\nabla \theta_j = \frac{V_j}{r_j}, \quad j = 0, 1, \ldots, m, \tag{A.12}$$

where $V_j(x)$ is the unit vector tangent to the circle of radius $|x - p_j|$, centered at p_j.

We have from (9) that

$$|\nabla u|^2 = |\nabla \rho|^2 + \rho^2 |\nabla \varphi_0 + \nabla \psi|^2 \tag{A.13}$$

and consequently

$$\int_\Omega |\nabla u|^2 \geq \int_\Omega |\nabla \rho|^2 + \int_\Omega |\nabla u_0|^2 + \int_\Omega a^2 |\nabla \psi|^2 - X, \tag{A.14}$$

where

$$\begin{aligned} X &= \int_\Omega (1 - \rho^2) |\nabla \varphi_0|^2 + 2 \int_\Omega (1 - \rho^2) \nabla \varphi_0 \cdot \nabla \psi - 2 \int_\Omega \nabla \varphi_0 \nabla \psi \\ &\equiv X_1 + X_2 + X_3. \end{aligned} \tag{A.15}$$

Estimate of X_1.

$$|X_1| \leq C \int_\Omega (1 - \rho^2) \left(\sum_{j=0}^{m} d_j^2 |\nabla \theta_j|^2 \right) = C \sum_{j=0}^{m} d_j^2 \int_\Omega (1 - \rho^2) |\nabla \theta_j|^2.$$

For $j = 1, 2, \ldots, m$, we compute

$$\begin{aligned} \int_\Omega (1 - \rho^2) |\nabla \theta_j| &\leq \int_\Omega (1 - \rho^2) \frac{1}{r_j^2} \\ &= \left(\int_{\Omega \setminus B_{R_0^{\frac{1}{2}}}(p_j)} + \int_{\Omega \cap B_{R_0^{\frac{1}{2}}}(p_j)} \right) \frac{(1 - \rho^2)}{r_j^2} \\ &\equiv I_1 + I_2, \end{aligned}$$

$$
\begin{aligned}
I_1 &\le \left\{ \int_{\Omega \setminus B_{R_0^{\frac{1}{2}}}(p_j)} (1-\rho^2)^2 \right\}^{1/2} \left\{ \int_{\Omega \setminus B_{R_0^{\frac{1}{2}}}(p_j)} \frac{1}{r_j^4} \right\}^{1/2} \\
&\le \left(K R_0^2 \log \frac{1}{R_1} \right)^{1/2} \left(\frac{2\pi}{3} \right)^{\frac{1}{2}} \left(\frac{1}{R_0} - \frac{1}{R^2} \right)^{1/2} \\
&\le C \left(R_0 \log \frac{1}{R_1} \right)^{1/2} \\
&\le C \left(R_1 \log \frac{1}{R_1} \right)^{1/2} \le C.
\end{aligned}
$$

Here we have used (5) in the second inequality.

Similarly, using (6), we have

$$
\begin{aligned}
I_2 &\le \left\{ \int_{\Omega \cap B_{R_0^{\frac{1}{2}}}(p_j)} (1-\rho^2)^2 \right\}^{1/2} \left\{ \int_{\Omega \cap B_{R_0^{\frac{1}{2}}}(p_j)} \frac{1}{r_j^4} \right\}^{1/2} \\
&\le (K R_0^2)^{1/2} \left(\frac{2\pi}{3} \right)^{1/2} \left(\frac{1}{R_0^2} - \frac{1}{R_0} \right)^{1/2} \le C.
\end{aligned}
$$

Hence

$$
\int_\Omega (1-\rho^2) |\nabla \theta_j|^2 \le C, \quad j = 1, 2, \ldots, m.
$$

We can also show by the same method as above that

$$
\int_\Omega (1-\rho^2) |\nabla \theta_0|^2 \le C.
$$

Hence, we yield

$$
|X_1| \le C. \tag{A.16}
$$

Estimate of X_2.

$$
|X_2| \le \mu \int_\Omega |\nabla \psi|^2 + \frac{C}{\mu} \int_\Omega (1-\rho^2)^2 |\nabla \varphi_0|^2 \le \mu \int_\Omega |\nabla \psi|^2 + \frac{C}{\mu} X_1.
$$

So (1.6) gives

$$
|X_2| \le \mu \int_\Omega |\nabla \psi|^2 + \frac{C}{\mu} \quad (\mu > 0). \tag{A.17}
$$

Estimate of X_3.

$$
X_3 \equiv -2 \int_\Omega \nabla \varphi_0 \nabla \psi = -2 \sum_{j=0}^{m} d_j \int_\Omega \nabla \theta_j \cdot \nabla \psi = -2 \sum_{j=0}^{m} d_j \int_\Omega \frac{V_j}{r_j} \nabla \psi.
$$

For $j = 1, 2, \ldots, m$,

$$\int_\Omega \frac{V_j}{r_j}\nabla\psi = \int_{\Omega\setminus B_{R_0}(p_j)} \frac{V_j}{r_j}\nabla\psi - \sum_{\substack{k\neq j\\ k\neq 0}} \int_{B_{R_0}(p_k)} \frac{V_j}{r_j}\nabla\psi - \int_{B_{R_1}} \frac{V_j}{r_j}\nabla\psi$$
$$\equiv J_1 + J_2 + J_3.$$

Let $\rho_j = R - |p_j|$ $(j = 1, 2, \ldots, m)$. Then $\frac{R}{2} \leq \rho_j \leq R$ by (1). Since

$$\int_{\partial B_r(p_j)} V_j \nabla\psi = \int_{\partial B_r(p_j)} \frac{\partial\psi}{\partial\tau} = 0 \quad \text{for } 0 < r < \rho_j, j = 0, 1, \ldots, m,$$

we have

$$\int_{B_{\rho_j(p_j)}\setminus B_{R_0(p_j)}} \frac{V_j\nabla\psi}{r_j} = 0 \quad j = 0, 1, \ldots, m.$$

So

$$|J_1| \leq \frac{1}{\rho_j}\int_{\Omega\setminus B_{\rho_j}(p_j)} |\nabla\psi|$$
$$\leq \frac{|\Omega|^{1/2}}{\rho_j}\left(\int_\Omega |\nabla\psi|^2\right)^{1/2}$$
$$\leq \mu\int_\Omega |\nabla\psi|^2 + \frac{C}{\mu} \quad (\mu > 0).$$

The assumption (2) implies $r_j = |x - p_j| \geq |p_k - p_j| - R_0 \geq 3R_0$ for $x \in B_{R_0}(p_k)(k \neq j)$. Therefore

$$|J_2| \leq \frac{1}{3R_0}\int_{B_{R_0}(p_k)} |\nabla\psi|$$
$$\leq C\left\{\int_{B_{R_0}(p_k)} |\nabla\psi|^2\right\}^{1/2}$$
$$\leq \mu\int_\Omega |\nabla\psi|^2 + \frac{C}{\mu}.$$

Here we have used (10).

Similarly,

$$|J_3| \leq \mu\int_\Omega |\nabla\psi|^2 + \frac{C}{\mu}.$$

Hence

$$\left|\int_\Omega \frac{V_j}{r_j}\nabla\psi\right| \leq \mu\int_\Omega |\nabla\psi|^2 + \frac{C}{\mu}, \quad j = 1, 2, \ldots, m.$$

We can also deduce by the same method that

$$\left|\int_\Omega \frac{V_0}{r_0}\nabla\psi\right| \leq \mu\int_\Omega |\nabla\psi|^2 + \frac{C}{\mu}.$$

Thus

$$|X_3| \le \mu \int_\Omega |\nabla\psi|^2 + \frac{C}{\mu}. \tag{A.18}$$

Now, by combining (14)–(18), we yield

$$\int_\Omega |\nabla u|^2 \ge \int_\Omega |\nabla\rho|^2 + \int_\Omega |\nabla u_0|^2 + a^2 \int_\Omega |\nabla\psi|^2 - 3\mu \int_\Omega |\nabla\psi|^2 - \frac{C}{\mu} - C;$$

therefore, Theorem A is proved if we take $\mu = \frac{a^2}{6}$. □

REFERENCES

[1] S. J. Chapman, Q. Du, and M. D. Gunzburger, *A Ginzburg-Landau type model of superconducting-normal junctions including Josephson junction*, European J. Appl. Math., 6 (1995), pp. 97–114.

[2] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices*, Birkhäuser Boston, Cambridge, MA, 1994.

[3] F. Bethuel, H. Brezis, and F. Hélein, *Asymptotics for the minimization of a Ginzburg-Landau functional*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 123–148.

[4] M. Struwe, *On the asymptotic behaviour of minimizers of the Ginzburg-Landau model in* 2 *dimensions*, Differential Integral Equations, 7 (1994), pp. 1613–1624.

[5] M. Struwe, *Erratum* "*on the asymptotic behavior of minimizers of the Ginzburg-Landau model in* 2 *dimensions*," Differential Integral Equations, 5 (1995), p. 224.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 69–84, January 1998

005

ASYMPTOTIC BEHAVIOR OF THE SOLUTIONS TO A LANDAU–GINZBURG SYSTEM WITH VISCOSITY FOR MARTENSITIC PHASE TRANSITIONS IN SHAPE MEMORY ALLOYS*

JÜRGEN SPREKELS†, SONGMU ZHENG‡, AND PEICHENG ZHU‡

Abstract. In this paper, we investigate the system of partial differential equations governing the dynamics of martensitic phase transitions in shape memory alloys under the presence of a (possibly small) viscous stress. The corresponding free energy is assumed in Landau–Ginzburg form and nonconvex as a function of the order parameter. Results concerning the asymptotic behavior of the solution as time tends to infinity are proved, and the compactness of the orbit is shown.

Key words. nonlinear thermoviscoelasticity, shape memory alloys, phase transitions, asymptotic behavior, compact orbits, Landau–Ginzburg theory

AMS subject classifications. 35Q72, 73B30, 35B40

PII. S0036141096297698

1. Introduction.

In the present paper, we study the asymptotic behavior of the solutions to a system that arises in the thermomechanical developments in a one-dimensional heat-conducting viscous solid of constant mass density ϱ (assumed to be normalized to unity, i.e., $\varrho = 1$). The solid is subjected to heating and loading. We think of metallic solids that not only respond to a change of the strain ε by a (possibly nonlinear) elastic stress $\sigma = \sigma(\varepsilon)$, but also to a change of the curvature of their metallic lattice by a couple stress $\mu = \mu(\varepsilon_x)$.

We assume that the Helmholtz free energy density F is a potential of Landau–Ginzburg form, i.e.,

$$F = F(\varepsilon, \varepsilon_x, \theta), \tag{1.1}$$

where θ denotes the absolute temperature. To cover systems modelling first-order, stress-induced, and temperature-induced solid–solid phase transitions accompanied by hysteresis phenomena, we do not assume that F is a convex function of the order parameter ε.

A particular class of materials, in which both stress-induced and temperature-induced first-order phase transitions leading to a rather spectacular hysteretic behavior occur, are the so-called *shape memory alloys*. In these materials the metallic lattice is deformed by shear, and the assumption of a constant density is justified. The shape memory effect itself is due to martensitic phase transitions between different configurations of the crystal lattice, namely, austenite and martensitic twins. For an account of the physical properties of shape memory alloys, we refer the reader to Chapter 5 in the monograph [4]. In a series of papers (cf., for instance, [7], [8]), Falk has proposed a Landau–Ginzburg theory that uses the shear strain ε as order

*Received by the editors January 22, 1996; accepted for publication (in revised form) October 22, 1996.

http://www.siam.org/journals/sima/29-1/29769.html

†Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany (sprekels@wias-berlin.de). This author was partially supported by Deutsche Forschungsgemeinschaft, DFG–SPP "Echtzeit-Optimierung großer Systeme."

‡Institute of Mathematics, Fudan University, 200433 Shanghai, P. R. China (szheng@fudan.ac.cn). The second author was supported by NSF of China grant 19331040.

parameter in order to explain the occurrence of the martensitic transitions in shape memory alloys. In this connection, we also refer to the works of Müller (cf., [1], [14]).

The simplest form for the free energy density F that accounts quite well for the experimentally observed behavior and that takes couple stresses into account is (see Falk [7], [8]) given by

$$F(\varepsilon, \varepsilon_x, \theta) = F_0(\theta) + F_1(\varepsilon)\theta + F_2(\varepsilon) + \frac{\delta}{2}\varepsilon_x^2\,, \tag{1.2}$$

where

$$F_1(\varepsilon) = \alpha_1\varepsilon^2, \quad F_2(\varepsilon) = \alpha_3\varepsilon^6 - \alpha_2\varepsilon^4 - \alpha_1\theta_1\varepsilon^2\,, \tag{1.3}$$

$$F_0(\theta) = -C_V\theta\log\left(\frac{\theta}{\theta_2}\right) + C_V\theta + \tilde{C}\,, \tag{1.4}$$

with positive physical constants $\theta_1, \delta, \alpha_1, \alpha_2, \alpha_3, \theta_2, C_V, \tilde{C}$. The constant C_V denotes the specific heat. Observe that in the interesting range of temperatures, for θ close to θ_1, F is not a convex function of the shear strain ε. In fact, $F(\cdot, \varepsilon_x, \theta)$ may have up to three minima that correspond to the austenitic and the two martensitic phases.

We want to forecast the dynamics of the phase transitions in the one-dimensional situation. To this end, let $\Omega = (0,1)$, and, for $t > 0$, $\Omega_t = \Omega \times (0,t)$. Then the balance laws of linear momentum and internal energy read

$$u_{tt} - \sigma_x + \mu_{xx} = 0, \quad \text{in} \quad \Omega_\infty\,, \tag{1.5}$$

$$U_t + q_x - \sigma\varepsilon_t - \mu\varepsilon_{xt} = 0, \quad \text{in} \quad \Omega_\infty\,. \tag{1.6}$$

The second law of thermodynamics is expressed by the Clausius–Duhem inequality

$$S_t + \left(\frac{q}{\theta}\right)_x \geq 0, \quad \text{in} \quad \Omega_\infty\,. \tag{1.7}$$

Here, u, σ, μ, U, q, ε, S, and θ denote displacement, shear stress, couple stress, internal energy density, heat flux, shear strain, entropy density, and absolute temperature, in that order.

For one-dimensional homogeneous thermoviscoelastic materials, we have the constitutive relations

$$\varepsilon = u_x, \quad \sigma = \frac{\partial F}{\partial \varepsilon} + \gamma\varepsilon_t, \quad \mu = \frac{\partial F}{\partial \varepsilon_x}, \quad S = -\frac{\partial F}{\partial \theta}, \quad U = F + \theta S\,, \tag{1.8}$$

where $\gamma > 0$ is the viscosity. For the heat flux q, we assume Fourier's law

$$q = -k\theta_x, \tag{1.9}$$

where $k > 0$ is the heat conductivity (assumed constant). Obviously, this assumption implies the validity of (1.7), so the second law of thermodynamics is automatically satisfied.

Inserting the constitutive relations in the balance laws (1.5)–(1.6), we obtain the system of partial differential equations

$$u_{tt} - (f_1\theta + f_2)_x - \gamma\varepsilon_{xt} + \delta u_{xxxx} = 0, \quad \text{in} \quad \Omega_\infty, \tag{1.10}$$

$$C_V\theta_t - k\theta_{xx} - f_1\theta\varepsilon_t - \gamma\varepsilon_t^2 = 0, \quad \text{in} \quad \Omega_\infty, \tag{1.11}$$

$$\varepsilon = u_x, \quad \text{in} \quad \Omega_\infty, \tag{1.12}$$

where

$$f_1 = f_1(\varepsilon) = F_1'(\varepsilon), \quad f_2 = f_2(\varepsilon) = F_2'(\varepsilon). \tag{1.13}$$

In addition, we prescribe the initial and boundary conditions

$$u|_{x=0} = \varepsilon_x|_{x=0} = 0, \quad \varepsilon|_{x=1} = (\gamma u_{xt} - \delta u_{xxx} + \sigma_1)|_{x=1} = 0, \tag{1.14}$$

with

$$\sigma_1 = f_1\theta + f_2, \tag{1.15}$$

as well as

$$\theta_x|_{x=0,\,1} = 0, \tag{1.16}$$

$$u(x,0) = u_0(x), \quad u_t(x,0) = u_1(x), \quad \theta(x,0) = \theta_0(x) > 0, \ x \in \ \overline{\Omega}. \tag{1.17}$$

The physical meaning of the boundary conditions is clear; for instance, the second condition at $x = 1$ describes the stress-free situation.

Next, we employ an idea of Andrews [2] and Pego [17] to simplify the problem by introducing the *velocity potential*

$$p(x,t) = \int_1^x u_t(y,t)\,dy\,. \tag{1.18}$$

Then,

$$\varepsilon_t = p_{xx}, \quad \text{in} \quad \Omega_\infty, \tag{1.19}$$

and (1.10)–(1.11) can be rewritten as

$$p_t - \gamma p_{xx} + \delta\varepsilon_{xx} - \sigma_1 = 0, \quad \text{in} \quad \Omega_\infty, \tag{1.20}$$

$$C_V\theta_t - k\theta_{xx} - f_1\theta p_{xx} - \gamma p_{xx}^2 = 0, \quad \text{in} \quad \Omega_\infty. \tag{1.21}$$

Accordingly, the initial and boundary conditions (1.14), (1.16), (1.17) become

$$p_x|_{x=0} = p_{xxx}|_{x=0} = \varepsilon_x|_{x=0} = 0, \tag{1.22}$$

$$p|_{x=1} = p_{xx}|_{x=1} = \varepsilon|_{x=1} = 0, \tag{1.23}$$

$$\varepsilon(x,0) = \varepsilon_0 = u_{0x}, \quad p(x,0) = p_0(x) = \int_1^x u_1(y)\,dy, \quad \theta(x,0) = \theta_0, \ \ x \in \overline{\Omega}\,. \tag{1.24}$$

It is easy to see that if (u, v, θ) is a smooth solution to (1.10)–(1.17), then (ε, p, θ) is a smooth solution to (1.19)–(1.24), and vice versa. Therefore, it suffices to consider the problem (1.19)–(1.24). In what follows, we assume without loss of generality that $C_V = 1$.

Before stating and proving our results, let us first recall some related results in the literature. In the case $\delta = 0$, Dafermos [5], Dafermos and Hsiao [6], Chen and Hoffmann [9], and Jiang [11] proved the global existence of a classical solution

to the system of (1.10)–(1.12) with various boundary conditions for a class of solid-like materials. However, an analysis of the asymptotic behavior as $t \to \infty$ was not performed in these papers. Recently, on the basis of Dafermos [5] and Dafermos and Hsiao [6], Luo [13] further investigated the asymptotic behavior of smooth solutions as time tends to infinity for a special class of solid-like materials in which $e = C_V\,\theta$, $F_2 = 0$, and $\delta = 0$. Racke and Zheng [18] obtained global existence, uniqueness, and the asymptotic behavior of weak solutions to (1.10)–(1.12) for $\delta = 0$ if both ends of the rod are insulated and if at least one end is stress-free.

In the case $\delta > 0$, we refer to Sprekels and Zheng [20] if $\delta > 0, \gamma = 0$, and to Hoffmann and Zochowski [10] if $\delta > 0, \gamma > 0$, for global existence and uniqueness results for Falk's Landau–Ginzburg model of shape memory alloys. However, the a priori estimates for the solution obtained in these papers depend on t, and hence the asymptotic behavior of the solution for $t \to \infty$ could not be treated there.

We also refer to the works of Andrews [2], Andrews and Ball [3], and Pego [17] for the isothermal and purely viscoelastic case.

The purpose of our contribution is to study the asymptotic behavior as $t \to \infty$ of the solutions to the system (1.19)–(1.24) and to prove the compactness of the orbit. Next, we state the main result of this paper.

THEOREM 1.1. *Suppose that $\varepsilon_0, p_0 \in H^3$, and $\theta_0 \in H^1$ are given functions that satisfy the compatibility conditions* $p_{0x}|_{x=0} = \varepsilon_{0x}|_{x=0} = 0$, $p_0|_{x=1} = p_{xx}|_{x=1} = \varepsilon_x|_{x=1} = 0$, *and suppose that $\theta_0 > 0$ in $[0,1]$. Then the following results hold.*

(i) *The problem admits a unique global solution (ε, p, θ) satisfying*

$$\varepsilon \in C(\mathbb{R}^+; H^3), \quad \varepsilon_t \in C(\mathbb{R}^+; H^1) \cap L^2(\mathbb{R}^+; H^2);$$

$$p \in C(\mathbb{R}^+; H^3) \cap L^2(\mathbb{R}^+; H^4), \quad p_t \in C(\mathbb{R}^+; H^1) \cap L^2(\mathbb{R}^+; H^2);$$

$$\theta \in C(\mathbb{R}^+; H^1), \quad \theta_x \in L^2(\mathbb{R}^+; H^1), \quad \theta_t \in L^2(\mathbb{R}^+; L^2),$$

$$\theta(x,t) > 0 \ \ \forall\, (x,t) \in \ [0,1] \times \mathbb{R}^+. \tag{1.25}$$

(ii) *As $t \to \infty$, it holds that*

$$\|p(\cdot,t)\|_{H^3} \to 0, \quad \|p_t(\cdot,t)\|_{H^1} \to 0, \tag{1.26}$$

$$\|\delta\varepsilon_{xx}(\cdot,t) - \sigma_1(\cdot,t)\|_{H^1} \to 0, \quad \|\varepsilon_t(\cdot,t)\|_{H^1} \to 0, \quad \|\theta_x(\cdot,t)\| \to 0. \tag{1.27}$$

(iii) *For all $\nu > 0$,*

$$\varepsilon \in C([\nu, +\infty); H^4), \quad p \in C([\nu, +\infty); H^4), \quad \theta \in C([\nu, +\infty); H^3); \tag{1.28}$$

i.e., the orbit is compact in $H^3 \times H^3 \times H^1$.

(iv)

$$(\varepsilon(\cdot,t), p(\cdot,t), \theta(\cdot,t)) \to (\overline{\varepsilon}, 0, \overline{\theta}), \quad \textit{as} \ \ t \to \infty, \ \ \textit{in} \ \ H^3 \times H^3 \times H^1\,, \tag{1.29}$$

where $(\overline{\varepsilon},\ \overline{\theta})$ is one of the equilibria for the corresponding stationary problem.

The main difficulties in proving Theorem 1.1 are due to the higher degree of nonlinearity inherent in the system (1.19)–(1.21) and to the higher order derivative arising for $\delta > 0$. The presence of this higher order derivative makes the problem in two ways significantly different from the problem with $\delta = 0, \gamma > 0$: it renders the

orbit compact (while discontinuities of strain will persist in the case $\delta = 0, \gamma > 0$, as shown in [18]), and the technique needed to obtain the asymptotic behavior differs considerably from that used in the case $\delta = 0, \gamma > 0$. One of the main ingredients of the proof in this paper is to bound the norms of ε, p, as well as their derivatives, in terms of expressions of the form

$$1 + \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^{\alpha} + \left(\int_0^t \|\theta_t(\tau)\|^2 \, d\tau\right)^{\beta}, \tag{1.30}$$

where $0 \le \alpha \le \frac{3}{2}$, $0 \le \beta \le \frac{1}{2}$. This makes it possible to reduce the degree of nonlinearity via interpolation techniques. To study the asymptotic behavior, we will make repeated use of a basic lemma in analysis proved in Shen and Zheng [19]. In section 2, we will prove the uniform a priori estimates and the compactness of the orbit. In section 3, the asymptotic behavior is investigated.

The notation in this paper will be as follows: L^p, $1 \le p \le \infty$, $W^{m,\infty}$, $m \in \mathbb{N}$, $H^1 \equiv W^{1,2}$, and $H_0^1 = W_0^{1,2}$, respectively, denote the usual Lebesgue and Sobolev spaces on (0,1). By $(\cdot,\cdot)$, we denote the inner product in L^2, and $\|\cdot\|_B$ denotes the norm in the space B. We use the abbreviation $\|\cdot\| := \|\cdot\|_{L_2}$, and $C^k(I,B)$, $k \in \mathbb{N}_0$, denotes the space of k-times continuously differentiable functions from $I \subset \mathbb{R}$ into a Banach space B. The spaces $L^p(I,B)$, $1 \le p \le \infty$, are defined analogously. Finally, ∂_t or $\frac{d}{dt}$ or a subscript t and, likewise, ∂_x or a subscript x, denote the partial derivatives with respect to t and x, respectively.

2. Uniform a priori estimates. The general framework to prove global existence and uniqueness of solution has been established in earlier papers, for instance, in Sprekels and Zheng [20] and Hoffmann and Zochowski [10]. The setting will become more apparent soon during the derivation of uniform a priori estimates. Therefore, we can focus our attention on the study of the asymptotic behavior and on the compactness of the orbit. In order to get the asymptotic behavior of the solution as $t \to \infty$, we shall prove uniform a priori estimates on ε, p, and θ with respect to t. From now on, we will always denote by C a universal positive constant that may depend on the initial data, but not on t.

LEMMA 2.1. *For any $t > 0$, the following estimates hold.*

$$\|\varepsilon(t)\| + \|\varepsilon(t)\|_{L^6} + \|p_x(t)\| + \|\varepsilon_x(t)\| + \|\theta(t)\|_{L^1} \le C, \tag{2.1}$$

$$\|p(t)\|_{L^\infty} + \|\varepsilon(t)\|_{L^\infty} \le C, \tag{2.2}$$

$$\theta(x,t) > 0 \quad \forall\, (x,t) \in [0,1] \times \mathbb{R}^+. \tag{2.3}$$

Proof. First, applying the maximum principle to (1.21), we find that

$$\theta(x,t) > 0 \quad \forall\, (x,t) \in [0,1] \times \mathbb{R}^+ . \tag{2.4}$$

Next, multiplying (1.20) by $-p_{xx}$, adding the result to (1.21), and integrating with respect to x over Ω, we arrive at

$$\frac{d}{dt}\int_0^1 \left(\theta + F_2(\varepsilon) + \frac{1}{2}p_x^2 + \frac{\delta}{2}\varepsilon_x^2\right)(t)\, dx = 0. \tag{2.5}$$

Thus,

$$\int_0^1 \left(\theta + F_2(\varepsilon) + \frac{1}{2}p_x^2 + \frac{\delta}{2}\varepsilon_x^2\right)(t)\, dx = E_1 , \tag{2.6}$$

where E_1 is a constant depending only on the initial data.

Using Young's inequality, we see that

$$F_2(\varepsilon) \geq C_1\varepsilon^6 - C_2\,, \tag{2.7}$$

and thus,

$$\|\varepsilon(t)\| + \|p_x(t)\| + \|\varepsilon_x(t)\| + \|\varepsilon(t)\|_{L^6} + \|\theta(t)\|_{L^1} \leq C. \tag{2.8}$$

By virtue of the boundary conditions and of Poincaré's inequality, we find

$$\|p(t)\|_{L^\infty} + \|\varepsilon(t)\|_{L^\infty} \leq C\,, \tag{2.9}$$

from which the assertion follows. □

LEMMA 2.2. *For any $t > 0$, the following estimates hold.*

$$\int_0^t \int_0^1 \left(\frac{\theta_x^2}{\theta^2} + \frac{p_{xx}^2}{\theta}\right) dx\, d\tau \leq C, \tag{2.10}$$

$$\int_0^t \|p_x(\tau)\|^2 d\tau \leq \int_0^t \|p_x(\tau)\|_{L^\infty}^2 d\tau \leq C, \quad \int_0^t \|p(\tau)\|_{L^\infty}^2 d\tau \leq C\,, \tag{2.11}$$

$$\int_0^t \|p_x(\tau)\|^{n+2} d\tau \leq C \quad \forall \quad n \geq 0. \tag{2.12}$$

Proof. Multiplication of (1.21) by θ^{-1} and integration with respect to x over Ω yield

$$\frac{d}{dt}\int_0^1 \big(\log\theta - F_1(\varepsilon)\big)(t)\, dx - \int_0^1 \left(\frac{k\theta_x^2}{\theta^2} + \frac{\gamma p_{xx}^2}{\theta}\right)(t)\, dx = 0. \tag{2.13}$$

Since $\log\theta \leq \theta - 1$ for all $\theta > 0$, we obtain

$$\int_0^t \int_0^1 \left(\frac{k\theta_x^2}{\theta^2} + \frac{\gamma p_{xx}^2}{\theta}\right) dx\, d\tau \leq C. \tag{2.14}$$

From $p_x|_{x=0} = 0$ it follows that

$$p_x(x,t) = p_x(0,t) + \int_0^x p_{xx}(y,t)\, dy = \int_0^x p_{xx}(y,t)\, dy. \tag{2.15}$$

Hence,

$$\begin{aligned}
&\int_0^t \|p_x(\tau)\|_{L^\infty}^2 d\tau \leq \int_0^t \Big(\int_0^1 |p_{xx}(x,\tau)|\, dx\Big)^2 d\tau \\
&\leq \int_0^t \left(\int_0^1 \sqrt{\theta}\, \frac{|p_{xx}|}{\sqrt{\theta}}\, dx\right)^2 d\tau \leq \int_0^t \Big(\int_0^1 \theta\, dx\Big)\Big(\int_0^1 \frac{p_{xx}^2}{\theta}\, dx\Big)\, d\tau \\
&\leq C\int_0^t \int_0^1 \frac{p_{xx}^2}{\theta}\, dx\, d\tau \leq C.
\end{aligned} \tag{2.16}$$

Thus,

$$\int_0^t \|p_x(\tau)\|^2 d\tau \leq \int_0^t \|p_x(\tau)\|_{L^\infty}^2 d\tau \leq C. \tag{2.17}$$

Combining (2.11) with (2.8), a simple induction yields that to any $n \in \mathbb{N}$ there is some $C = C(n)$ such that

$$\int_0^t \|p_x(\tau)\|^{n+2} d\tau \le C\,. \tag{2.18}$$

The proof of the assertion is complete. □

In what follows we will see that (2.18) is very useful for reducing the degree of nonlinearity. To get further estimates, we will now derive estimates for the derivatives of the norms of ε, p by expressions of the form (1.30).

LEMMA 2.3. *For any $t > 0$, the following estimates hold.*

$$\int_0^t (\|\varepsilon_t(\tau)\|^2 + \|p_{xx}(\tau)\|^2)\, d\tau \le C \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}, \tag{2.19}$$

$$\int_0^t \|\theta_x(\tau)\|^2 d\tau \le C \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^2. \tag{2.20}$$

Proof. Using Lemma 2.2, we obtain

$$\begin{aligned}\int_0^t \|p_{xx}(\tau)\|^2 d\tau &= \int_0^t \left\| \sqrt{\theta} \frac{p_{xx}}{\sqrt{\theta}}(\tau) \right\|^2 d\tau \\ &\le \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty} \int_0^t \left\| \frac{p_{xx}}{\sqrt{\theta}}(\tau) \right\|^2 d\tau \\ &\le C \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}. \end{aligned} \tag{2.21}$$

Similarly, we have

$$\int_0^t \|\theta_x(\tau)\|^2 d\tau \le C \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^2. \tag{2.22}$$

The proof is complete. □

We can now show further estimates.

LEMMA 2.4. *For any $t > 0$, the following estimates hold.*

$$\begin{aligned}&\|p_{xt}(t)\|^2 + \|p_{xxx}(t)\|^2 + \int_0^t (\|p_{xxt}(\tau)\|^2 + \|\varepsilon_{tt}(\tau)\|^2)\, d\tau \\ &\le C \left(1 + \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^3 + \int_0^t \|\theta_t(\tau)\|^2\, d\tau \right), \end{aligned} \tag{2.23}$$

$$\begin{aligned}&\|\varepsilon_{xt}(t)\|^2 + \int_0^t (\|p_{xxxx}(\tau)\|^2 + \|\varepsilon_{xxt}(\tau)\|^2)\, d\tau \\ &\le C \left(1 + \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^3 + \int_0^t \|\theta_t(\tau)\|^2\, d\tau \right). \end{aligned} \tag{2.24}$$

Proof. First, differentiating (1.20) with respect to t, multiplying the result by $-\varepsilon_{tt}$, and integrating with respect to x over Ω, we obtain

$$\begin{aligned}0 &= (p_{tt}(t), -p_{xxt}(t)) + \gamma \|\varepsilon_{tt}(t)\|^2 + (\delta\varepsilon_{xt}(t), \varepsilon_{xtt}(t)) + \int_0^1 \sigma_{1t}(t)\, \varepsilon_{tt}(t)\, dx \\ &= (p_{xtt}(t), p_{xt}(t)) + \gamma \|\varepsilon_{tt}(t)\|^2 + \delta(\varepsilon_{xt}(t), \varepsilon_{xtt}(t)) \\ &\quad + \int_0^1 \big(f_1'(\varepsilon)\, \varepsilon_t\, \theta + f_2'(\varepsilon)\, \varepsilon_t + f_1(\varepsilon)\, \theta_t \big)(t)\, \varepsilon_{tt}(t)\, dx. \end{aligned} \tag{2.25}$$

Combination with (2.9) yields

$$\frac{1}{2}\frac{d}{dt}(\|p_{xt}(t)\|^2+\delta\|\varepsilon_{xt}(t)\|^2)+\gamma\|\varepsilon_{tt}(t)\|^2 \le \frac{\gamma}{2}\|\varepsilon_{tt}(t)\|^2+C\int_0^1 \big(\theta^2\varepsilon_t^2+\varepsilon_t^2+\theta_t^2\big)(t)\,dx. \tag{2.26}$$

Integrating (2.26) with respect to t and applying Lemma 2.3, we arrive at

$$\begin{aligned}
&\|p_{xt}(t)\|^2 + \|\varepsilon_{xt}(t)\|^2 + \int_0^t \|\varepsilon_{tt}(\tau)\|^2 d\tau \\
&\le C + C\int_0^t (\|\theta(\tau)\,\varepsilon_t(\tau)\|^2 + \|\varepsilon_t(\tau)\|^2 + \|\theta_t(\tau)\|^2)\,d\tau \\
&\le C + C\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}^2 \int_0^t \|\varepsilon_t(\tau)\|^2\,d\tau + C\int_0^t (\|\varepsilon_t(\tau)\|^2 + \|\theta_t(\tau)\|^2)\,d\tau \\
&\le C\Big(1+\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}^3 + \int_0^t \|\theta_t(\tau)\|^2 d\tau\Big). \qquad (2.27)
\end{aligned}$$

Here, we have used Young's inequality in the form $a \le Ca^3 + C'$.

Next, we differentiate (1.20) with respect to t, then multiply by ε_{xxt}, and integrate the result with respect to x over Ω to obtain

$$\begin{aligned}
0 &= (p_{tt}(t),\varepsilon_{xxt}(t)) - \gamma(\varepsilon_{tt}(t),\varepsilon_{xxt}(t)) + \delta\|\varepsilon_{xxt}(t)\|^2 - \int_0^1 \varepsilon_{xxt}(t)\,\sigma_{1t}(t)\,dx \\
&= (p_{xxtt}(t),\varepsilon_t(t)) + \gamma(\varepsilon_{xtt}(t),\varepsilon_{xt}(t)) + \delta\|\varepsilon_{xxt}(t)\|^2 - \int_0^1 \varepsilon_{xxt}(t)\,\sigma_{1t}(t)\,dx \\
&= \frac{d}{dt}(p_{xxt}(t),\varepsilon_t(t)) - \|p_{xxt}(t)\|^2 + \frac{\gamma}{2}\frac{d}{dt}\|\varepsilon_{xt}(t)\|^2 + \delta\|\varepsilon_{xxt}(t)\|^2 \\
&\quad - \int_0^1 \varepsilon_{xxt}(t)\,\sigma_{1t}(t)\,dx. \qquad (2.28)
\end{aligned}$$

However, by integration by parts, we have

$$(p_{xxt}(t),\varepsilon_t(t)) = -(p_{xt}(t),\varepsilon_{xt}(t)). \tag{2.29}$$

Combining this with (2.28), and using (2.23) and Young's inequality, we find

$$\begin{aligned}
&\frac{\gamma}{4}\|\varepsilon_{xt}(t)\|^2 + \delta\int_0^t \|\varepsilon_{xxt}(\tau)\|^2 d\tau \\
&\le C + \frac{\delta}{2}\int_0^t \|\varepsilon_{xxt}(\tau)\|^2 d\tau + C\Big(\|p_{xt}(t)\|^2 + \int_0^t (\|\sigma_{1t}(\tau)\|^2 + \|p_{xxt}(\tau)\|^2)\,d\tau\Big) \\
&\le C\Big(1+\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}^3 + \int_0^t \|\theta_t(\tau)\|^2 d\tau\Big) + \frac{\delta}{2}\int_0^t \|\varepsilon_{xxt}(\tau)\|^2 d\tau\,. \qquad (2.30)
\end{aligned}$$

The proof of the lemma is complete. □

In what follows, we will find that the above lemma plays a crucial role in reducing the degree of nonlinearity.

LEMMA 2.5. *For any $t > 0$, the following estimates hold.*

$$\|\theta_x(t)\|^2 + \int_0^t \|\theta_t(\tau)\|^2 d\tau \le C. \tag{2.31}$$

$$\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty} \le C. \tag{2.32}$$

Proof. Multiplying (1.21) by θ_t and integrating with respect to x over Ω, we obtain

$$\begin{aligned}
&\frac{k}{2}\frac{d}{dt}\|\theta_x(t)\|^2 + \|\theta_t(t)\|^2 \;=\; \int_0^1 \big(f_1(\varepsilon)\,\theta\,\theta_t\,p_{xx} + \gamma\,\theta_t\,p_{xx}^2\big)(t)\,dx \\
&\le C\left(\|\theta(t)\,p_{xx}(t)\|\,\|\theta_t(t)\| + \Big(\int_0^1 p_{xx}^4(t)\,dx\Big)^{\frac12}\|\theta_t(t)\|\right) \\
&\le C\left(\|\theta(t)\|_{L^\infty}^{\frac12}\,\|p_{xx}(t)\|_{L^\infty}\Big(\int_0^1 \theta(t)\,dx\Big)^{\frac12}\|\theta_t(t)\| \;+\; \|p_{xx}(t)\|_{L^4}^2\,\|\theta_t(t)\|\right). \qquad (2.33)
\end{aligned}$$

Therefore, integration with respect to t yields

$$\begin{aligned}
&\|\theta_x(t)\|^2 + \int_0^t \|\theta_t(\tau)\|^2 d\tau \\
&\le C\left(\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}^{\frac12}\Big(\int_0^t \|p_{xx}(\tau)\|_{L^\infty}^2 d\tau\Big)^{\frac12}\Big(\int_0^t \|\theta_t(\tau)\|^2 d\tau\Big)^{\frac12}\right. \\
&\qquad \left. + \Big(\int_0^t \|p_{xx}(\tau)\|_{L^4}^4 d\tau\Big)^{\frac12}\Big(\int_0^t \|\theta_t(\tau)\|^2 d\tau\Big)^{\frac12} \;+\; 1\right) \\
&= C\,(I_1 + I_2 + 1)\,. \qquad (2.34)
\end{aligned}$$

We now estimate the terms I_1, I_2. By virtue of Nirenberg's inequality and the boundary conditions, we obtain

$$\|p_{xx}(t)\|_{L^\infty} \le C\,\|p_{xxxx}(t)\|^{\frac12}\;\|p_x(t)\|^{\frac12}, \qquad (2.35)$$

$$\|p_{xx}(t)\|_{L^4} \le C\,\|p_{xxxx}(t)\|^{\frac{5}{12}}\;\|p_x(t)\|^{\frac{7}{12}}. \qquad (2.36)$$

Hence,

$$\begin{aligned}
I_1 &= C\left(\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}\int_0^t \|\theta_t(\tau)\|^2 d\tau \int_0^t \|p_{xx}(\tau)\|_{L^\infty}^2 d\tau\right)^{\frac12} \\
&\le C\left(\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}\int_0^t \|\theta_t(\tau)\|^2 d\tau \int_0^t \|p_{xxxx}(\tau)\|\,\|p_x(\tau)\|\,d\tau\right)^{\frac12} \\
&\le C\left(\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}\int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac12}\left(\int_0^t \|p_{xxxx}(\tau)\|^2 d\tau \int_0^t \|p_x(\tau)\|^2 d\tau\right)^{\frac14}. \qquad (2.37)
\end{aligned}$$

Using Lemmas 2.2 and 2.4 and Young's inequality, we conclude that

$$\begin{aligned}
I_1 &\le C\left(\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}\int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac12}\left(\int_0^t \|p_{xxxx}(\tau)\|^2 d\tau\right)^{\frac14} \\
&\le C\left(\sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}\int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac12}\left(1 + \sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}^3 + \int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac14} \\
&\le \frac14\int_0^t \|\theta_t(\tau)\|^2 d\tau + C\left(1 + \sup_{0\le\tau\le t}\|\theta(\tau)\|_{L^\infty}^{\frac52}\right). \qquad (2.38)
\end{aligned}$$

Next, owing to Schwarz's inequality and (2.36), we have

$$\begin{aligned} I_2 &= C\left(\int_0^t \|\theta_t(\tau)\|^2 d\tau \int_0^t \|p_{xx}(\tau)\|_{L^4}^4 d\tau\right)^{\frac{1}{2}} \\ &\le C\left(\int_0^t \|\theta_t(\tau)\|^2 d\tau \int_0^t \|p_{xxxx}(\tau)\|^{\frac{5}{3}} \, \|p_x(\tau)\|^{\frac{7}{3}} d\tau\right)^{\frac{1}{2}} \\ &\le C\left(\int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac{1}{2}} \left(\int_0^t \|p_{xxxx}(\tau)\|^2 d\tau\right)^{\frac{5}{12}} \left(\int_0^t \|p_x(\tau)\|^{14} d\tau\right)^{\frac{1}{12}}. \end{aligned} \tag{2.39}$$

Applying (2.18) with $n = 12$ and Lemma 2.4, we get

$$\begin{aligned} I_2 &\le C\left(\int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac{1}{2}} \left(\int_0^t \|p_{xxxx}(\tau)\|^2 d\tau\right)^{\frac{5}{12}} \\ &\le C\left(\int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac{1}{2}} \left(1 + \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^3 + \int_0^t \|\theta_t(\tau)\|^2 d\tau\right)^{\frac{5}{12}} \\ &\le \frac{1}{4}\int_0^t \|\theta_t(\tau)\|^2 d\tau + C \sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^{\frac{5}{2}} + C. \end{aligned} \tag{2.40}$$

Owing to Nirenberg's inequality and (2.1), we have

$$\|\theta(t)\|_{L^\infty} \le C\,\|\theta_x(t)\|^{\frac{2}{3}} \, \|\theta(t)\|_{L^1}^{\frac{1}{3}} + C\,\|\theta(t)\|_{L^1} \le C\,\|\theta_x(t)\|^{\frac{2}{3}} + C. \tag{2.41}$$

Combining (2.38)–(2.40) with (2.34) and (2.41) and applying Young's inequality, we find

$$\begin{aligned} \|\theta_x(t)\|^2 + \int_0^t \|\theta_t(\tau)\|^2 d\tau &\le \frac{1}{2}\int_0^t \|\theta_t(\tau)\|^2 d\tau + C\Big(1 + \sup_{0\le\tau\le t} \|\theta_x(\tau)\|^{\frac{5}{3}}\Big) \\ &\le C + \frac{1}{2}\left(\int_0^t \|\theta_t(\tau)\|^2 d\tau + \sup_{0\le\tau\le t} \|\theta_x(\tau)\|^2\right). \end{aligned} \tag{2.42}$$

Taking the supremum with respect to t in (2.42) yields

$$\sup_{0\le\tau\le t} \|\theta_x(\tau)\|^2 + \int_0^t \|\theta_t(\tau)\|^2 d\tau \le \frac{1}{2}\Big(\int_0^t \|\theta_t(\tau)\|^2 d\tau + \sup_{0\le\tau\le t} \|\theta_x(\tau)\|^2\Big) + C. \tag{2.43}$$

Hence,

$$\sup_{0\le\tau\le t} \|\theta_x(\tau)\|^2 + \int_0^t \|\theta_t(\tau)\|^2 d\tau \le C. \tag{2.44}$$

Thus, using (2.41),

$$\sup_{0\le\tau\le t} \|\theta(\tau)\|_{L^\infty}^2 \le C\,, \tag{2.45}$$

which concludes the proof of the assertion. □

Combining Lemmas 2.3–2.5 and using the system equations, we immediately conclude with the following lemma.

LEMMA 2.6. *For any $t > 0$, the following estimates hold.*

$$\int_0^t (\|p_{xx}(\tau)\|^2 + \|\varepsilon_t(\tau)\|^2 + \|\theta_x(\tau)\|^2_{H^1})\, d\tau \le C, \tag{2.46}$$

$$\int_0^t (\|p_{xxt}(\tau)\|^2 + \|\varepsilon_{tt}(\tau)\|^2 + \|p_{xxxx}(\tau)\|^2 + \|\varepsilon_{xxt}(\tau)\|^2)\, d\tau \le C, \tag{2.47}$$

$$\|p_{xt}(t)\|^2 + \|p_{xxx}(t)\|^2 + \|\varepsilon_{xt}(t)\|^2 + \|\varepsilon_{xxx}(t)\|^2 \le C. \tag{2.48}$$

LEMMA 2.7. *For any $t > 0$, the following estimates hold.*

$$\int_0^t (\|p_t(\tau)\|^2 + \|p_t(\tau)\|^2_{L^\infty} + \|p_{xt}(\tau)\|^2 + \|p_{xt}(\tau)\|^2_{L^\infty})\, d\tau \le C, \tag{2.49}$$

$$\int_0^t (\|\delta\varepsilon_{xx}(\tau) - \sigma_1(\tau)\|^2 d\tau + \|(\delta\varepsilon_{xx} - \sigma_1)_t(\tau)\|^2)\, d\tau \le C, \tag{2.50}$$

$$\int_0^t (\|p_{xx}(\tau)\|^2_{L^\infty} + \|p_{xxx}(\tau)\|^2 + \|p_{xxx}(\tau)\|^2_{L^\infty} + \|p_{tt}(\tau)\|^2)\, d\tau \le C, \tag{2.51}$$

$$\|p_t(t)\|^2 + \|p_{xx}(t)\|^2 + \|p_t(t)\|^2_{L^\infty} + \|p_x(t)\|^2_{L^\infty} + \|p_{xx}(t)\|^2_{L^\infty} \le C. \tag{2.52}$$

Proof. These estimates can easily be derived from the system equations and from Lemmas 2.5 and 2.6. □

Now we proceed to investigate the compactness of the orbit of the solution for $t > 0$ in $H^3 \times H^3 \times H^1$. For the time being, we assume that the initial data are so smooth that the solution will have enough smoothness to carry out the following argument: if the initial data belonged just to $H^3 \times H^3 \times H^1$, we could approximate them by smooth functions and then pass to the limit.

Differentiating (1.20) twice with respect to t, we find that

$$p_{ttt} - \gamma p_{xxtt} + \delta\varepsilon_{xxtt} - \sigma_{1tt} = 0\,. \tag{2.53}$$

A straightforward calculation yields

$$\sigma_{1tt} = f_1'(\varepsilon)\,\varepsilon_{tt}\,\theta + 2\,f_1'(\varepsilon)\,\varepsilon_t\,\theta_t + f_1(\varepsilon)\,\theta_{tt} + f_2''(\varepsilon)\,\varepsilon_t^2 + f_2'\varepsilon)\,\varepsilon_{tt}. \tag{2.54}$$

Multiplying (2.53) by p_{tt} and integrating with respect to x over Ω, we find

$$\begin{aligned} 0 &= \frac{1}{2}\frac{d}{dt}\|p_{tt}(t)\|^2 - \gamma(p_{xxtt}(t), p_{tt}(t)) + \delta(\varepsilon_{xxtt}(t), p_{tt}(t)) - (\sigma_{1tt}(t), p_{tt}(t)) \\ &= \frac{1}{2}\frac{d}{dt}\|p_{tt}(t)\|^2 + \gamma\|p_{xtt}(t)\|^2 + \delta(\varepsilon_{tt}(t), p_{xxtt}(t)) - (\sigma_{1tt}(t), p_{tt}(t)) \\ &= \frac{1}{2}\frac{d}{dt}\Big(\|p_{tt}(t)\|^2 + \delta\|\varepsilon_{tt}(t)\|^2\Big) + \gamma\|p_{xtt}(t)\|^2 - (\sigma_{1tt}(t), p_{tt}(t))\,. \end{aligned} \tag{2.55}$$

Multiplying (2.55) by t^2 and using (2.32), as well as Lemmas 2.6 and 2.7, we obtain that

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\Big(t^2\|p_{tt}(t)\|^2 + t^2\delta\|\varepsilon_{tt}\|^2\Big) - t\,(\|p_{tt}(t)\|^2 + \delta\|\varepsilon_{tt}(t)\|^2) + \gamma t^2\|p_{xtt}(t)\|^2 \\ &\le t^2\|p_{tt}(t)\|^2 + Ct^2\|\sigma_{1tt}(t)\|^2 \\ &\le t^2\|p_{tt}(t)\|^2 + Ct^2(\|\varepsilon_{tt}(t)\|^2 + \|\theta_t(t)\|^2 + \|\theta_{tt}(t)\|^2 + \|\varepsilon_t(t)\|^2)\,. \end{aligned} \tag{2.56}$$

Hence, it follows from (2.31), (2.46), and (2.47) that

$$t^2(\|p_{tt}(t)\|^2+\delta\|\varepsilon_{tt}(t)\|^2)+\int_0^t \tau^2\|p_{xtt}(\tau)\|^2 d\tau \le C_1+Ct^2+C\int_0^t \tau^2\|\theta_{tt}(\tau)\|^2 d\tau\,, \tag{2.57}$$

where $C_1 = C(\|\varepsilon_0\|_{H^3}, \|p_0\|_{H^3}, \|\theta_0\|_{H^1})$.

On the other hand, differentiating (1.21) with respect to t, we get

$$\theta_{tt} - k\theta_{xxt} - (f_1(\varepsilon)\theta p_{xx} + \gamma p_{xx}^2)_t = 0. \tag{2.58}$$

Multiplying by θ_{tt} and integrating with respect to x, we arrive at

$$\begin{aligned} \frac{k}{2}\frac{d}{dt}\|\theta_{xt}(t)\|^2 + \|\theta_{tt}(t)\|^2 &\le \frac{1}{2}\|\theta_{tt}(t)\|^2 + \frac{1}{2}\|(f_1(\varepsilon)\theta p_{xx} + \gamma p_{xx}^2)_t(t)\|^2 \\ &\le \frac{1}{2}\|\theta_{tt}(t)\|^2 + C\left(\|p_{xx}(t)\|^2 + \|\theta_t(t)\|^2 + \|p_{xxt}(t)\|^2\right). \end{aligned} \tag{2.59}$$

Multiplication of (2.59) by t^2 yields

$$\frac{k}{2}\frac{d}{dt}(t^2\|\theta_{xt}(t)\|^2) - kt\|\theta_{xt}(t)\|^2 + \frac{t^2}{2}\|\theta_{tt}(t)\|^2 \le Ct^2(\|p_{xx}(t)\|^2+\|\theta_t(t)\|^2+\|p_{xxt}(t)\|^2). \tag{2.60}$$

In order to estimate $\int_0^t \tau\|\theta_{xt}(\tau)\|^2 d\tau$, we multiply (2.58) by θ_t and then integrate with respect to x over Ω to obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\theta_t(t)\|^2 + k\|\theta_{xt}(t)\|^2 &\le \frac{1}{2}\|\theta_t(t)\|^2 + \frac{1}{2}\|(f_1(\varepsilon)\theta p_{xx} + \gamma p_{xx}^2)_t(t)\|^2 \\ &\le \frac{1}{2}\|\theta_t(t)\|^2 + C\left(\|\varepsilon_t(t)\|^2 + \|\theta_t(t)\|^2 + \|p_{xxt}(t)\|^2\right). \end{aligned} \tag{2.61}$$

Multiplying (2.61) by t, we find

$$\frac{1}{2}\frac{d}{dt}(t\|\theta_t(t)\|^2) + kt\|\theta_{tx}(t)\|^2 \le C\left(\|\theta_t(t)\|^2 + t\|\theta_t(t)\|^2 + t\left(\|\varepsilon_t(t)\|^2+\|\theta_t(t)\|^2+\|p_{xxt}(t)\|^2\right)\right). \tag{2.62}$$

Therefore,

$$t\|\theta_t(t)\|^2 + \int_0^t \tau\|\theta_{xt}(\tau)\|^2 d\tau \le Ct + C_2\,, \tag{2.63}$$

where $C_2 = C(\|\varepsilon_0\|_{H^3}, \|p_0\|_{H^3}, \|\theta_0\|_{H^1})$.

Combination of (2.63) with (2.60) yields

$$\int_0^t \tau^2\|\theta_{tt}(\tau)\|^2 d\tau \le C_3 + Ct^2\,, \tag{2.64}$$

with $C_3 = C(\|\varepsilon_0\|_{H^3}, \|p_0\|_{H^3}, \|\theta_0\|_{H^1})$.

Thus, it follows from (2.57) that

$$\|p_{tt}(t)\|^2 + \|\varepsilon_{tt}(t)\|^2 \le C_4 t^{-2} + C. \tag{2.65}$$

Also, using (2.63) and (2.60),

$$\|\theta_t(t)\|^2 \le C + C_4 t^{-1}, \quad \|\theta_{xt}(t)\|^2 \le C_4 t^{-2} + C\,, \tag{2.66}$$

with C_4 depending only on $\|\varepsilon_0\|_{H^3}, \|p_0\|_{H^3}, \|\theta_0\|_{H^1}$.

Thus, it easily follows from equations (1.19) to (1.21) that for any initial data in $H^3 \times H^3 \times H^1$, the following holds:

$$(\varepsilon(\cdot,t), p(\cdot,t), \theta(\cdot,t)) \in H^4 \times H^4 \times H^3 \quad \forall\, t > 0. \tag{2.67}$$

Moreover, we can infer from Lemmas 2.5 to 2.7, and from (2.55), (2.59), and (2.61), that for any $\nu > 0$ the triple (ε, p, θ) is bounded in $C([\nu, +\infty); H^4 \times H^4 \times H^3)$. From this the compactness of the orbit in $H^3 \times H^3 \times H^1$ follows. □

3. Asymptotic behavior. In this section, we will prove the results on the asymptotic behavior of the solution given in Theorem 1.1. In what follows, a convergence symbol "$\to$" is always to be understood as $t \to \infty$. We will make use of the following basic lemma from Shen and Zheng [19].

LEMMA 3.1. *Suppose that y and h are nonnegative functions on $(0,\infty)$ such that y' is locally integrable and such that y, h satisfy*

$$\forall t \ge 0: \quad y'(t) \le A_1 y^2(t) + A_2 + h(t), \tag{3.1}$$

$$\forall T > 0: \quad \int_0^T y(\tau)d\tau \le A_3, \quad \int_0^T h(\tau)d\tau \le A_4, \tag{3.2}$$

where A_1, A_2, A_3, A_4 denote positive constants which are independent of t and T. Then, for any $r > 0$,

$$\forall t \ge 0: \quad y(t+r) \le \left(\frac{A_3}{r} + A_2 r + A_4\right) e^{A_1 A_2}. \tag{3.3}$$

Moreover,

$$\lim_{t\to\infty} y(t) = 0. \tag{3.4}$$

LEMMA 3.2. *It holds that*

$$\|p(t)\|_{H^3} \to 0, \quad \|p_t(t)\|_{H^1} \to 0, \tag{3.5}$$

$$\|\varepsilon_t(t)\|_{H^1} \to 0, \quad \|(\delta\varepsilon_{xx} - \sigma_1)(t)\|_{H^1} \to 0, \tag{3.6}$$

$$\|u_t(t)\|_{H^2} \to 0. \tag{3.7}$$

Proof. It follows from (2.26) and (2.32) that

$$\begin{aligned} &\frac{d}{dt}(\|p_{xt}(t)\|^2 + \delta\|\varepsilon_{xt}(t)\|^2) + \gamma\|\varepsilon_{tt}(t)\|^2 \\ &\le C\,(\|\theta(t)\,\varepsilon_t(t)\|^2 + \|\varepsilon_t(t)\|^2 + \|\theta_t(t)\|^2) \\ &\le C\,(\|\varepsilon_t(t)\|^2 + \|\theta_t(t)\|^2). \end{aligned} \tag{3.8}$$

Combining (3.8) with (2.51), (2.46), (2.49), (2.31) and applying Lemma 3.1, we arrive at

$$\|p_{xt}(t)\|^2 + \|\varepsilon_{xt}(t)\|^2 \to 0. \tag{3.9}$$

Hence, $\|p_{xxx}(t)\|^2 \to 0$, and thus $\|u_t\|_{H^2} \to 0$.

Next, we differentiate (1.20) with respect to t, then multiply by $\delta\varepsilon_{xx} - \sigma_1$, and integrate with respect to x over Ω. It follows that

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\delta\varepsilon_{xx}(t) - \sigma_1(t)\|^2 &= -(p_{tt}(t) - \gamma\varepsilon_{tt}(t), \delta\varepsilon_{xx}(t) - \sigma_1(t)) \\ &\le \frac{1}{2}\|\delta\varepsilon_{xx}(t) - \sigma_1(t)\|^2 + C(\|p_{tt}(t)\|^2 + \|\varepsilon_{tt}(t)\|^2). \end{aligned} \tag{3.10}$$

Combining (3.10) with (2.50), (2.51), (2.47) and applying Lemma 3.1, we conclude that

$$\|\delta\varepsilon_{xx}(t) - \sigma_1(t)\|^2 \to 0. \tag{3.11}$$

From (1.20) and (3.9), we also get

$$\|(\delta\varepsilon_{xx} - \sigma_1)_x(t)\|^2 \to 0. \tag{3.12}$$

The assertions of Lemma 3.2 now follow from the above estimates and from Poincaré's inequality. □

LEMMA 3.3. *It holds that*

$$\|\theta_x(t)\| \to 0\,. \tag{3.13}$$

Proof. We multiply (1.21) by θ_t and integrate with respect to x over Ω to get

$$\begin{aligned} \frac{k}{2}\frac{d}{dt}\|\theta_x(t)\|^2 + \|\theta_t(t)\|^2 &= \int_0^1 \Big(\gamma\, p_{xx}^2\, \theta_t \,+\, f_1(\varepsilon)\, \theta\, \theta_t\, p_{xx}\Big)(t)\, dx \\ &\le \frac{1}{2}\|\theta_t(t)\|^2 + \|\theta(t)\, p_{xx}(t)\|^2 + \|p_{xx}^2(t)\|^2. \end{aligned} \tag{3.14}$$

Combining (3.14) with (2.32) and (2.52), we see that

$$k\frac{d}{dt}\|\theta_x(t)\|^2 + \|\theta_t(t)\|^2 \le C\|p_{xx}(t)\|^2. \tag{3.15}$$

Hence, we can infer from (2.46) and Lemma 3.1 that

$$\|\theta_x(t)\|^2 \to 0\,,$$

which concludes the proof. □

Concerning the convergence of ε, u, θ, we have the following result.

LEMMA 3.4. *It holds that*

$$(\varepsilon(\cdot,t), p(\cdot,t), \theta(\cdot,t)) \to (\overline{\varepsilon}, 0, \overline{\theta}), \quad \textit{in} \;\; H^3 \times H^3 \times H^1\,, \tag{3.16}$$

$$u(\cdot,t) \to \overline{u}\,, \quad \textit{in} \quad H^4\,, \quad \textit{with} \;\; \overline{u}(x) = \int_0^x \overline{\varepsilon}(y)\, dy \quad \forall\, x \in [0,1]\,, \tag{3.17}$$

where $(\overline{\varepsilon}, \overline{\theta})$ *is one of the equilibria for the corresponding stationary problem*

$$\delta\varepsilon_{xx} - f_1(\varepsilon)\theta - f_2(\varepsilon) = 0, \tag{3.18}$$

$$\varepsilon_x|_{x=0} = 0, \quad \varepsilon|_{x=1} = 0, \tag{3.19}$$

$$\theta = \text{Const.}, \tag{3.20}$$

$$\int_0^1 \Big(\theta + F_2(\varepsilon) + \frac{\delta}{2}\varepsilon_x^2\Big)\, dx = E_1. \tag{3.21}$$

Proof. It is easy to see from (2.4) and (2.12) that, for any $0 < \nu < 1$,

$$\begin{aligned} &\frac{d}{dt}\int_0^1 \Big(\theta - \nu\,\log\theta + F_2(\varepsilon) + \nu F_1(\varepsilon) + \frac{1}{2}p_x^2 + \frac{\delta}{2}\varepsilon_x^2\Big)(t)\, dx \\ &\qquad +\nu\int_0^1 \left(\frac{k\theta_x^2}{\theta^2} + \frac{\gamma p_{xx}^2}{\theta}\right)(t)\, dx = 0. \end{aligned} \tag{3.22}$$

Thus the system (1.19)–(1.21) has a Lyapunov function of the form

$$\int_0^1 \Big(\theta - \nu\,\log\theta + F_2(\varepsilon) + \nu F_1(\varepsilon) + \frac{1}{2}p_x^2 + \frac{\delta}{2}\varepsilon_x^2\Big)(t)\, dx\,.$$

Since the orbit is compact, as proved in previous section, it follows from the standard theory of dynamical systems that the ω-limit set is connected, compact, and consists of equilibria. Since the corresponding stationary problem admits only a finite number of solutions (see Zhou [22], and also Luckhaus and Zheng [12], Novick-Cohen and Zheng [16], Zheng [21]), (3.16) follows. In view of the boundary condition $u|_{x=0} = 0$, we also get (3.17). Therefore, the proof is complete. □

REFERENCES

[1] M. Achenbach and I. Müller, *Creep and yield in martensitic transformations*, Ingenieur-Archiv, 53 (1983), pp. 73–83.

[2] G. Andrews, *On the existence of solutions to the equation* $u_{tt} = u_{xxt} + \sigma(u_x)_x$, J. Differential Equations, 35 (1980), pp. 200–231.

[3] G. Andrews and J. M. Ball, *Asymptotic behaviour and changes of phase in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 44 (1982), pp. 306–341.

[4] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Appl. Math. Sci. 121, Springer, New York, 1996.

[5] 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.

[6] C. M. Dafermos and L. Hsiao, *Global smooth thermomechanical processes in one-dimensional nonlinear thermoviscoelasticity*, Nonlinear Anal., 6 (1982), pp. 435–454.

[7] F. Falk, *Ginzburg-Landau theory of static domain walls in shape memory alloys*, Physica B, 51 (1983), pp. 177–185.

[8] F. Falk, *Ginzburg-Landau theory and solitary waves in shape memory alloys*, Physica B, 54 (1984), pp. 159–167.

[9] Z. Chen and K.-H. Hoffmann, *On a one-dimensional nonlinear thermoviscoelastic model for structural phase transitions in shape memory alloys*, J. Differential Equations, 112 (1994), pp. 325-350.

[10] K.-H. Hoffmann and A. Zochowski, *Existence of solutions to some non-linear thermoelastic systems with viscosity*, Math. Meth. Appl. Sci., 15 (1992), pp. 187–204.

[11] S. Jiang, *Global large solutions to initial boundary value problems in one-dimensional nonlinear thermoviscoelasticity*, Quart. Appl. Math., 51 (1993), pp. 731–744.

[12] S. Luckhaus and S. Zheng, *A nonlinear boundary value problem involving a nonlocal term*, Nonlinear Anal., 22 (1994), pp. 129–135.

[13] T. Luo, *Qualitative Behavior to Nonlinear Evolution Equations with Dissipation*, Ph.D. thesis, Institute of Mathematics, Academy of Sciences of China, Beijing, 1994.

[14] I. Müller and K. Wilmański, *A model for phase transitions in pseudoelastic bodies*, Nuovo Cimento B, 57 (1980), pp. 283–318.

[15] M. Niezgódka and J. Sprekels, *Existence of solutions for a mathematical model of structural phase transitions in shape memory alloys*, Math. Meth. Appl. Sci., 10 (1988), pp. 197–223.

[16] A. Novick-Cohen and S. Zheng, *The Penrose–Fife type equations: Counting the one-dimensional stationary solutions*, Proc. Royal Soc. Edinburgh Ser. A, 126 (1996), pp. 483–504.

[17] R. L. Pego, *Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability*, Arch. Rational Mech. Anal., 97 (1987), pp. 353–394.

[18] R. Racke and S. Zheng, *Global existence and asymptotic behavior in nonlinear thermoviscoelasticity*, J. Differential Equations, 134 (1997), pp. 46–67.

[19] W. Shen and S. Zheng, *On the coupled Cahn-Hilliard equations*, Comm. Partial Differential Equations, 18 (1993), pp. 701–727.

[20] J. Sprekels and S. Zheng, *Global solutions to the equations of a Ginzburg-Landau theory for structural phase transitions in shape memory alloys*, Physica D, 39 (1989), pp. 59–76.

[21] S. Zheng, *Nonlinear Parabolic Equations and Hyperbolic–Parabolic Coupled Systems*, Pitman Series Monographs and Surveys in Pure and Applied Mathematics 76, Longman Group Limited, London, 1995.

[22] P. Zhou, *Multiplicity of solutions to a nonlinear boundary value problem*, J. Math. Anal. Appl., submitted.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 85–105, January 1998

006

A NONLOCAL REGULARIZATION OF SOME OVER-DETERMINED BOUNDARY-VALUE PROBLEMS I*

D. E. EDMUNDS† AND N. I. POPIVANOV‡

Abstract. Some three-dimensional analogues of the plane Darboux problems for hyperbolic equations with degeneracy are investigated. In 1954, Protter initiated the study of such three-dimensional problems, and it is now well known that for an infinite number of smooth right-hand sides these problems have solutions with a strong power-type singularity on the characteristic cone. This effect appears even for small perturbations of certain C_0^∞ right-hand sides. Using Friedrichs' theory of symmetric positive operators, we find and investigate a nonlocal problem which is a regularizer, in some sense, of these ill-posed problems.

Key words. ill-posed problems, regularization methods, boundary-value problems, nonlocal problems, degenerate hyperbolic equations

AMS subject classifications. 35L20, 35L50, 35L80, 35R25

PII. S0036141096303232

1. Introduction.

To set the scene we denote points of $\mathbb{R}^3$ by $x = (x_1, x_2, t)$ and put $\rho = \sqrt{x_1^2 + x_2^2}$, $\varphi = \arctan(x_2/x_1)$. Let $K : [0, \infty) \to \mathbb{R}$ be of class C^1 and such that $K(0) = 0, K'(0) > 0$ with $K(t) > 0$, and $K'(t) > 0$ if $t > 0$. Let G be the domain

$$G = \left\{ x \in \mathbb{R}^3 : 0 < t < d,\ \int_0^t \sqrt{K(\tau)}d\tau < \rho < 1 - \int_0^t \sqrt{K(\tau)}d\tau \right\},$$

where d is the (unique) solution of the equation $2\int_0^d \sqrt{K(\tau)}d\tau = 1$. The boundary of G is $\partial G = S_0 \cup S_1 \cup S_2$, where S_0 is the disc $S_0 = \{x : t = 0, 0 \le \rho \le 1\}$ and

$$S_1 = \left\{ 0 \le t \le d, \rho = 1 - \int_0^t \sqrt{K(\tau)}d\tau \right\}, S_2 = \left\{ 0 \le t \le d, \rho = \int_0^t \sqrt{K(\tau)}d\tau \right\}.$$

We shall consider the equation

$$Lu := K(t)\left(u_{x_1x_1} + u_{x_2x_2}\right) - u_{tt} \equiv K(t)\left\{\rho^{-1}\left(\rho u_\rho\right)_\rho + \rho^{-2}u_{\varphi\varphi}\right\} - u_{tt} = f, \tag{1.1}$$

where f is a prescribed function; S_1 and S_2 are characteristics of (1.1).

Problem P. Is there a solution of (1.1) in G, which satisfies the condition

$$u = 0 \quad \text{on} \quad S_0 \cup S_1? \tag{1.2}$$

Problem P*. Is there a solution of (1.1) in G, which satisfies the condition

$$u = 0 \quad \text{on} \quad S_0 \cup S_2? \tag{1.3}$$

Protter [28, 29] formulated these adjoint problems as multidimensional analogues of the Darboux problem in the plane. He worked with the wave equation corresponding to $K(t) = 1$ and also investigated (1.1) in a domain which contained G in its

*Received by the editors May 8, 1996; accepted for publication October 28, 1996. This research was partially supported by the Bulgarian NSF under grant MM-512/95.
http://www.siam.org/journals/sima/29-1/30323.html
†Center for Mathematical Analysis and its Applications, University of Sussex, Brighton, BN1 9QH, England (mmfb7@sussex.ac.uk).
‡Department of Mathematics and Informatics, University of Sofia, 1164 Sofia, Bulgaria (nedyu@fmi.uni-sofia.bg).

hyperbolic part and contained a set G' in which (1.1) is elliptic. For equation (1.1), which is of changing type in $G \cup G' \cup S_0$, he formulated certain other problems, which are three-dimensional analogues of a plane problem examined by Morawetz [20] and Lax and Phillips [17].

When equation (1.1) is of changing type, Protter's problem given in Protter [29] was studied by Aziz and Schneider [6], Bitsadze [7], Didenko [10], Salzman and Schneider [31], Papadakis [21], and others. Problems P and P* for (1.1) in the domain G were considered by Didenko [10] in the case of the Tricomi equation (in which $K(t) \equiv t$). In the same case, after the paper of Kan Cher [8], Popivanov showed in 1986 that the homogeneous Problem P* has infinitely many classical solutions v_n $(n = 4, 5, ...)$, where

$$v_n(t, \rho, \varphi) = t\rho^{-n} \left(\rho^2 - 4t^3/9\right)^{n-4/3} \ (a_n \cos n\varphi + b_n \sin n\varphi) \tag{1.4}$$

and a_n, b_n are arbitrary constants. This corresponds to the result of Kwang-Chang [16] for the wave equation and implies that for classical solvability of Problem P an infinite number of conditions of the form $f \perp v_n \quad (n = 4, 5, ...)$ are necessary.

Some interesting results concerning Protter's problems for equation (1.1), both in changing-type domains and in G, are provided by Sorokina [32, 33]; we discuss these in section 7. For further work on the problem (1.1), (1.2), see Aldashev [1, 2, 3], Kan Cher [8], Popivanov and Schneider [26, 27], and the references cited in these works.

Popivanov and Schneider [26] proceeded in another way, being interested in the question: why does Problem P not have a classical solution when $f = v_n$ (v_n as in (1.4))? They introduced a new class of "generalized solutions" of Problem P and proved that some kind of "generalized solution" exists and is unique but that it is unstable and has a very strong singularity on the characteristic cone S_2. More precisely, they showed that given any $\ell \in \mathbb{N}$, there is a function $f_\ell \in C^\ell\left(\overline{G}\right)$ such that the corresponding "generalized solution" $u_\ell \in C\left(\overline{G} \setminus S_2\right)$ of Problem P exists, is unique, and satisfies the estimate

$$\int_{S_{2,\varepsilon}} |u_\ell| \, ds \geq \varepsilon^{-\ell}, \quad 0 < \varepsilon < 1, \tag{1.5}$$

where $S_{2,\varepsilon} = \{x \in G : \rho = \varepsilon + \int_0^t \sqrt{K(\tau)} d\tau\}$.

This situation can be interpreted in terms of improperly posed (or ill-posed) problems: we recall that these are problems which fail to have a unique global solution which depends continuously on the data. For investigations of such problems for partial differential equations we refer to the monographs of Payne [22], Tikhonov and Arsenin [34], and Lavrentiev, Romanov, and Shishatskii [19]; the book by Lattès and Lions [18] describes a regularization method for approximating solutions to ill-posed problems. We also refer to the papers by Ames, Levine, and Payne [5], Ames [4], as well as to the many references cited in these works; and to Tikhonov and Arsenin [34] for numerous regularization methods.

In Problem P, the position is the following:

(i) According to the results of Popivanov and Schneider cited above, there are infinitely many distinct right-hand sides f_n $(n \in \mathbb{N})$ of (1.1) for which there is a generalized solution with a strong singularity, of at least power-type (see (1.5)).

(ii) Let $u_0 \in C_0^\infty(G)$ be fixed and suppose that $K(t) \equiv t$. Then $f_0 := Lu_0 \in C_0^\infty(G)$ and for any right-hand side

$$f_{n,\delta} := f_0 + \delta f_n \qquad (n \in \mathbb{N}, \delta \neq 0) \tag{1.6}$$

there is a generalized solution with a strong singularity (see (1.5)) and there is no classical solution. This shows that the Problem P for the Tricomi equation is very unstable, even though $f_0 \in C_0^\infty(G)$, because a small perturbation (1.6) in an infinite number of directions has such a strong effect.

With this in mind, and having regard to the work on ill-posed problems which was mentioned earlier, it is appropriate to consider a new problem which regularizes Protter's problem. This new problem should be such that its solutions are free of the singularity, typified by (1.5), which appears on the characteristic cone S_2. This suggests connecting points from G with ones on the cone S_2, and so to investigate Problems P and P* we introduce a new, nonlocal problem. Let α be a small positive parameter. Given any t_0, ρ_0 and with $C = \{t_0 + \alpha\rho_0/(\alpha+1)\}\rho_0^\alpha$, let $(p_\alpha(t_0,\rho), q_\alpha(t_0,\rho_0))$ be the point of intersection of the curves

$$\rho = \int_0^t \sqrt{K(\tau)}d\tau, \quad t = C\rho^{-\alpha} - \alpha\rho/(\alpha+1).$$

Problem A. Is there a solution $\omega(t,\rho,\varphi)$ of the equation

$$(L\omega)(t,\rho,\varphi) - \rho^{-2}K(t)\omega_{\varphi\varphi}(p_\alpha(t,\rho), q_\alpha(t,\rho),\varphi) = f(t,\rho,\phi) \tag{1.7}$$

in G which satisfies the boundary conditions (1.2), i.e., $\omega = 0$ on $S_0 \cup S_1$?

Equation (1.7) is nonlocal because it involves points with coordinates (t,ρ,φ) and $(p_\alpha(t,\rho), q_\alpha(t,\rho),\varphi)$. We remark here merely that in our nonlocal Problem A, in the additional term

$$\rho^{-2}K(t)\omega_{\varphi\varphi}(p_\alpha(t,\rho), q_\alpha(t,\rho),\varphi)$$

of (1.7), the point $(p_\alpha, q_\alpha, \varphi)$ lies just on the characteristic cone S_2, where the big singularity shown by (1.5) appears in the "generalized solution" of the original Problem P. The derivative $\omega_{\varphi\varphi}$ is tangential to S_2 at that point.

Unlike [26], where the "generalized solution" belongs to a weighted space of smooth functions, we work here in a weighted Sobolev space. Following the work of Morawetz [20] and Lax and Phillips [17] in the two-dimensional case and Sorokina [32] in the multidimensional case, we introduce the weighted Sobolev space

$$\widetilde{W}_2^1(G) := \left\{\omega : \|\omega\|_{\widetilde{W}_2^1(G)} = \left(\int_G \left(\omega^2 + \omega_t^2 + r\left(\omega_{x_1}^2 + \omega_{x_2}^2\right)\right)dx\right)^{1/2} < \infty\right\}, \tag{1.8}$$

where $r = \sqrt{x_1^2 + x_2^2 + t^2}$. (The weight in Sorokina [33] is different.) In this space we establish existence, uniqueness, and an a priori estimate of a generalized solution of Problem A for every $f \in L_2(G)$. We also prove the infinite smoothness of the solution with respect to φ. We remark that analogous results were given, without proofs, for the Tricomi equation in G in Popivanov [25] and concerning the wave equation in Popivanov [24].

What is the connection between Problems P and A? We note that Garabedian [15] proved uniqueness of a classical solution of Problem P for the wave equation; an analogous result for the equation (1.1) follows from [26]. But in both cases we do not know whether or not the unique solution depends continuously on f. Following Didenko [11], we investigate another problem.

Problem P$_\varphi$. Is there a solution u of Problem P which satisfies the extra condition $\partial u/\partial\varphi = 0$ on S_2?

We prove that its solution u_f (when it exists) coincides with the solution ω_f of the Problem A, i.e., $u_f \equiv \omega_f$. Accordingly we can say that Problem A is a "nonlocal regularizer" of the strongly over-determined Problems P_φ and P. Using the results about Problem A we prove that the solution u_f of Problem P_φ depends continuously on f. For the adjoint Problem P^* (with $\dim KerP^* = \infty$) we find some additional conditions, under which we prove the uniqueness of solutions in $\widetilde{W}_2^1(G)$.

The plan of the paper is simple. Influenced by the works of Friedrichs [14], Morawetz [20], Lax and Phillips [17], and Sorokina [32], we investigate in section 2 a system of partial differential equations which is connected with Problem P and formulate corresponding boundary-value problems. In section 3 we examine the problem of coincidence of weak and strong solutions of these problems; sections 4, 5, and 6 are concerned with the proof of the existence, uniqueness, and smoothness (in φ) of the generalized solution of Problem A. The final section, section 7, deals with the connection between Problems P, P*, and the nonlocal Problem A. It is shown that, under appropriate conditions on f, the solutions of Problem P coincide with those of Problem A.

2. Investigation of a related system of equations. Given a vector-valued function $\hat{u} = (u_1, u_2, u_3)$ we introduce the formal notation

$$u_\rho := (x_1 u_1 + x_2 u_2)/\rho, \quad u_\varphi := x_1 u_2 - x_2 u_1, \tag{2.1}$$

and the derivatives $\frac{\partial}{\partial\rho} = \frac{x_1}{\rho}\frac{\partial}{\partial x_1} + \frac{x_2}{\rho}\frac{\partial}{\partial x_2}$, $\frac{\partial}{\partial\varphi} = x_1\frac{\partial}{\partial x_2} - x_2\frac{\partial}{\partial x_1}$. We consider the system

$$\begin{cases} \dfrac{K}{\rho}\dfrac{\partial}{\partial\rho}(\rho u_\rho) + \dfrac{K}{\rho^2}\dfrac{\partial u_\varphi}{\partial\varphi} - \dfrac{\partial u_3}{\partial t} = f, \\ K\left\{\dfrac{a}{\rho}\left(\dfrac{\partial u_\varphi}{\partial t} - \dfrac{\partial u_3}{\partial\varphi}\right) - \left(\dfrac{\partial u_\varphi}{\partial\rho} - \dfrac{\partial u_\rho}{\partial\varphi}\right)\right\} = 0, \\ \dfrac{\partial u_\rho}{\partial t} - \dfrac{\partial u_3}{\partial\rho} = 0, \end{cases} \tag{2.2}$$

where $a = a(t, \rho, \varphi)$ is a function to be chosen later. In matrix form this becomes

$$\hat{L}_0\hat{u} := \left(A^1\frac{\partial}{\partial x_1} + A^2\frac{\partial}{\partial x_2} + A^3\frac{\partial}{\partial t}\right)\hat{u} = \hat{f}_1, \tag{2.3}$$

where $\hat{f}_1 = (f, 0, 0)$. We shall see that the equation (1.1) and the system (2.3) are not equivalent. In addition, for (2.3) we have a new characteristic, S_0; thus all the boundary surfaces S_0, S_1, and S_2 are characteristics. To use the Friedrichs [14] theory of positive systems we reduce (2.3) to symmetric form by left multiplication by

$$\Lambda := \begin{pmatrix} -x_1 & x_2/\rho & -Kax_1/\rho \\ -x_2 & -x_1/\rho & -Kax_2/\rho \\ a & 0 & \rho \end{pmatrix}. \tag{2.4}$$

This gives the symmetric system

$$\begin{aligned} \hat{L}\hat{u} \;:=\; & \begin{pmatrix} -Kx_1 & -Kx_2 & Ka \\ -Kx_2 & Kx_1 & 0 \\ Ka & 0 & -x_1 \end{pmatrix}\frac{\partial\hat{u}}{\partial x_1} + \begin{pmatrix} Kx_2 & -Kx_1 & 0 \\ -Kx_1 & -Kx_2 & Ka \\ 0 & Ka & -x_2 \end{pmatrix}\frac{\partial\hat{u}}{\partial x_2} \\ & + \begin{pmatrix} -Ka & 0 & x_1 \\ 0 & -Ka & x_2 \\ x_1 & x_2 & -a \end{pmatrix}\frac{\partial\hat{u}}{\partial t} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} := \hat{f}, \end{aligned} \tag{2.5}$$

where $f_1 = -x_1 f$, $f_2 = -x_2 f$, $f_3 = af$. We observe that $\det \Lambda = \rho^2 - Ka^2$; this leads us to impose the following condition on the function $a : \rho - a\sqrt{K(t)} > 0$ in G. Following the notation of Friedrichs we have

$$\left(\hat{u}, \hat{L}\hat{u}\right)_{L_2(G)} = (\hat{u}, \varkappa\hat{u})_{L_2(G)} + \frac{1}{2}\int_{\partial G} \hat{u} \cdot \beta\hat{u} ds; \tag{2.6}$$

here the boundary matrix β is given by $\beta(x) = \sum_{j=1}^{3} n_j(x)\Lambda(x)A^j(x)$, where $n(x) = (n_1(x), n_2(x), n_3(x))$ is the unit exterior normal vector at $x \in \partial G$, and the matrix $\varkappa$ is defined by

$$\varkappa = \begin{pmatrix} (Ka)_t & 0 & -Ka_{x_1} \\ 0 & (Ka)_t & -Ka_{x_2} \\ -Ka_{x_1} & -Ka_{x_2} & 2 + a_t \end{pmatrix}. \tag{2.7}$$

In the paper Friedrichs [14], the matrix corresponding to our matrix $\varkappa$ is positive in $\overline{G}$. To fit in with this we work in a suitably weighted Sobolev space and choose

$$a(t, \rho, \varphi) = \alpha(\rho + t) \text{ in } G, \tag{2.8}$$

where $d > 0$ is a parameter, as in Sorokina [32]. With this choice we have

$$\hat{u} \cdot \varkappa\hat{u} \geq \alpha K'(t)(\rho + t)\left(u_1^2 + u_2^2\right) + (2 + \alpha - \alpha K)u_3^2. \tag{2.9}$$

This leads us to impose the following two conditions on α :

(E1) $\rho - \alpha(\rho + t)\sqrt{K(t)} > 0$ in $\overline{G}\setminus\{(0,0,0)\}$,

(E2) $\alpha K(d) < 2 + \alpha$.

Remark. It easy to see that near the point $(0,0,0)$ the condition (E1) is equivalent to $0 < \alpha < 2/3$. Note that (E1) and (E2) are satisfied for every sufficiently small $\alpha > 0$. For example, if $K(t) = t$, then they hold if $0 < \alpha < 2/\left(3 + \sqrt[3]{6}\right)$.

According to the Friedrichs theory, the boundary conditions

$$\begin{cases} x_1u_1 + x_2u_2 = 0 \text{ on } S_0, \qquad x_1u_2 - x_2u_1 = 0 \text{ on } S_2; \\ \sqrt{K}\,(x_1u_1 + x_2u_2) - \rho u_3 = 0 \text{ on } S_1 \end{cases} \tag{2.10}$$

are admissible for $\hat{L}$. The adjoint boundary conditions are

$$\begin{cases} x_1v_1 + x_2v_2 - av_3 = 0 \text{ on } S_0, \qquad x_2v_1 - x_1v_2 = 0 \text{ on } S_1; \\ \sqrt{K}\,(x_1v_1 + x_2v_2) + \rho v_3 = 0 \text{ on } S_2, \end{cases} \tag{2.11}$$

and these are admissible for the adjoint operator $\hat{L}^*$. Following the work of Morawetz [20] and Lax and Phillips [17] in the two-dimensional case and Sorokina [32] in the multidimensional case, we introduce the weighted Lebesgue spaces

$$\begin{aligned} H^*(G) &:= \left\{\hat{u} : \|\hat{u}\|^* := \left(\int_G \left[r^{-1}\left(u_1^2 + u_2^2\right) + u_3^2\right] dx\right)^{1/2} < \infty\right\}, \\ H_*(G) &:= \left\{\hat{u} : \|\hat{u}\|_* := \left(\int_G \left[r\left(u_1^2 + u_2^2\right) + u_3^2\right] dx\right)^{1/2} < \infty\right\}, \end{aligned} \tag{2.12}$$

where $r = \sqrt{\rho^2 + t^2}$. By $(\cdot, \cdot)$ we shall mean the inner product in $L_2(G)$, and $\|\cdot\|$ will stand for the corresponding norm.

DEFINITION 2.1. *A function* $\hat{u} \in H_*(G)$ *is said to be a weak solution of the problem* (2.5), (2.10) *if*

$$\left(\hat{u}, \hat{L}^*\hat{v}\right) = \left(\hat{f}, \hat{v}\right) \tag{2.13}$$

for every $\hat{v} \in C^1\left(\overline{G}\right)$ *which satisfies the adjoint boundary conditions* (2.11) *and vanishes in some neighborhood of the point* $(0,0,0)$.

DEFINITION 2.2. *A function* $\hat{u} \in H_*(G)$ *is called a strong solution of the problem* (2.5), (2.10) *if, and only if, there are functions* $\hat{u}_m \in C^1\left(\overline{G}\right)$ $(m \in \mathbb{N})$ *each of which satisfies the boundary conditions* (2.10) *and vanishes in some neighborhood of* $(0,0,0)$ *such that*

$$\|\hat{u}_m - \hat{u}\|_* \to 0 \text{ and } \left\|\hat{L}\hat{u}_m - \hat{f}\right\|^* \to 0 \text{ as } m \to \infty. \tag{2.14}$$

From the Friedrichs theory and (2.9), standard procedures now show that the following theorem holds.

THEOREM 2.3. *Let* $\alpha > 0$ *be so small that* (E1) *and* (E2) *hold. Then for any* $\hat{f} \in H^*(G)$, *there exists a weak solution* $\hat{u} \in H_*(G)$ *of the problem* (2.5), (2.10). *If there is a strong solution, it is unique and satisfies the a priori estimate*

$$\|\hat{u}\|_* \le C_\alpha \left\|\hat{f}\right\|^*,$$

where C_α *is a constant which does not depend on* $\hat{u}$.

Every strong solution of (2.5), (2.10) is a weak solution. However, we also have the following theorem.

THEOREM 2.4. *Every weak solution is a strong solution.*

From this result (to be proved in the next section) and Theorem 2.3 it follows that given any $\hat{f} \in H^*(G)$, there are a unique weak solution and a unique strong solution, which coincide.

3. Proof of Theorem 2.4.

Proof. By partition of the unity argument it is enough to show that if the support of a weak solution $\hat{u}$ is concentrated in a small neighborhood of an arbitrary point of $\overline{G}$, then it is a strong solution; that is, there are functions $\hat{u}_m \in C^1\left(\overline{G}\right)$ satisfying the conditions of Definition 2.2. Far from $(0,0,0)$, the L_2-norm and the norms on $H^*(G)$ and $H_*(G)$ are equivalent. The point $(0,0,0)$ requires separate treatment. For each other point in G we use the method of mollifiers developed by Friedrichs [13], Lax and Phillips [17], Peyser [23], Rauch [30], and others. After suitable change of variables we look for an integral operator R_ε, which depends on a parameter $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$, such that $R_\varepsilon\hat{u}$ satisfies the boundary conditions (2.10), and with an adjoint R_ε^* such that $R_\varepsilon^*\hat{v}$ satisfies the adjoint boundary conditions (2.11) for every pair of functions $u, v \in L_2(G)$ with support in a small neighborhood of the point considered. The most difficult problem usually is to prove $\|\hat{L}R_\varepsilon\hat{u} - \hat{f}\| \to 0$ when $\varepsilon \to 0$ in a special way. Since $(\hat{L}R_\varepsilon^*)^*\hat{u} = R_\varepsilon\hat{f}$, we must show that

$$\left\|\left(\hat{L}^*R_\varepsilon^*\right)^*\hat{u} - \hat{L}R_\varepsilon\hat{u}\right\| \to 0 \quad \text{as} \quad \varepsilon \to 0. \tag{3.1}$$

If the kernel k_ε of the integral operator R_ε is $k_\varepsilon(x,y) = k_\varepsilon(x-y)$, then we have the integral representation

$$\left(\hat{L}^* R_\varepsilon^*\right)^* \hat{u}(y) - \hat{L} R_\varepsilon \hat{u}(y) = \int_G \left\{ \sum_{j=1}^{3} \frac{\partial}{\partial z_j} \left[A^j(y) k_\varepsilon(y-z) - k_\varepsilon(y-z) A^j(z) \right] \right. \tag{3.2}$$
$$\left. - \left[B(y) k_\varepsilon(y-z) - k_\varepsilon(y-z) B(z) \right] \right\} \hat{u}(z)\, dz.$$

We prove here local coincidence only in the case when the support of the solution lies in a small neighborhood of some point of $S_0 \cap S_1$. For the other cases we refer to Lax and Phillips [17], Peyser [23], Rauch [30], and Popivanov [24] for indications of how to proceed.

Let $P_0 \in S_0 \cap S_1$ and suppose that $u \in L_2(G)$ is a weak solution x of the boundary value problem (2.5), (2.10) with $u = 0$ in some neighborhood of S_2. Because the surface S_2 is not C^2 near P_0, we work with the variables (t, ρ, φ). Note that the system (2.5) in variables

$$\widetilde{u} = \begin{pmatrix} u_\rho \\ -\rho^{-1} u_\varphi \\ u_3 \end{pmatrix} = F_1 \hat{u} := \begin{pmatrix} x_1/\rho & x_2/\rho & 0 \\ x_2/\rho & -x_1/\rho & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \tag{3.3}$$

becomes

$$\hat{L}_1 \widetilde{u} := F_1 \hat{L} \hat{u} := F_1 \hat{L} F_1 \widetilde{u} \quad := \left[\begin{pmatrix} -Ka & 0 & \rho \\ 0 & -Ka & 0 \\ \rho & 0 & -a \end{pmatrix} \frac{\partial}{\partial t} \right.$$
$$+ \begin{pmatrix} -K\rho & 0 & Ka \\ 0 & K\rho & 0 \\ Ka & 0 & -\rho \end{pmatrix} \left. \frac{\partial}{\partial \rho} + A_2^\varphi \frac{\partial}{\partial \varphi} + B_2 \right] \widetilde{u} = \widetilde{f} \tag{3.4}$$

and that the boundary conditions (2.10) become

$$\sqrt{K}\widetilde{u}_1 - \widetilde{u}_3 = 0 \text{ on } S_1, \quad \widetilde{u}_1 = 0 \text{ on } S_0. \tag{3.5}$$

To diagonalize the matrix A^t we multiply the system (3.4) by the matrix $F_2 \in C^1$ (supp u), where

$$F_2 = \frac{1}{\rho^2 - Ka^2} \begin{pmatrix} a & 0 & \rho \\ 0 & \dfrac{Ka^2 - \rho^2}{a} & 0 \\ \rho & 0 & Ka \end{pmatrix}. \tag{3.6}$$

This gives the system

$$\hat{L}_2 \widetilde{u} := F_2 \hat{L}_1 \widetilde{u} \quad := \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & K(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \frac{\partial}{\partial t} \right.$$
$$+ \begin{pmatrix} 0 & 0 & -1 \\ 0 & -\dfrac{K\rho}{a} & 0 \\ -K & 0 & 0 \end{pmatrix} \left. \frac{\partial}{\partial \rho} + A_2^\varphi \frac{\partial}{\partial \varphi} + B_2 \right] \widetilde{u} = \widetilde{f}_1; \tag{3.7}$$

the boundary conditions (3.5) remain the same. The new adjoint conditions are

$$\widetilde{v}_2 = 0 \text{ on } S_1, \quad \widetilde{v}_3 = 0 \text{ on } S_0. \tag{3.8}$$

We define approximating functions by

$$R_\varepsilon \widetilde{u}(y) = \int_D k_\varepsilon(y-z)\widetilde{u}(z)dz \quad \left(y=(t,\rho,\varphi),\ z=\left(\overline{t},\overline{\rho},\overline{\varphi}\right)\right) \tag{3.9}$$

in the domain $D = \{(t,\rho,\varphi) : \frac{2}{3} < \rho < 1 - \int_0^t \sqrt{K(\tau)}\, d\tau, c_3 < \varphi < c_4 t\}$, where $0 < c_4 - c_3 < 2\pi$; as usual $j \in C^\infty(\mathbb{R}), j(s) = 0$ for $\mid s \mid > 1$, $\int j(s)ds = 1$, and

$$\begin{aligned} k_\varepsilon(y) := \quad & \operatorname{diag}\{J_1, J_2, J_3\} = \frac{1}{\varepsilon_1\varepsilon_2\varepsilon_3} j\left(\frac{\varphi}{\varepsilon_3}\right) \operatorname{diag}\left\{ j\left(\frac{t}{\varepsilon_1}-2\right) j\left(\frac{\rho}{\varepsilon_2}+E\right), \right. \\ & \left. j\left(\frac{t}{\varepsilon_1}+2\right) j\left(\frac{\rho}{\varepsilon_2}-E\right), j\left(\frac{t}{\varepsilon_1}+2\right) j\left(\frac{\rho}{\varepsilon_2}+E\right)\right\}, \end{aligned} \tag{3.10}$$

where $0 < \varepsilon_1 \le \varepsilon_2 \le \varepsilon_3$ and E is an appropriate constant.

The functions (3.9) satisfy the boundary conditions (3.5) on $t = 0$, while the functions $R_\varepsilon^* v$ satisfy the adjoint boundary conditions (3.8) on $t = 0$. On the surface S_1 we have $\rho = 1 - \int_0^t \sqrt{K(\tau)}\, d\tau$ and if $E \ge 2 + 3\sqrt{K(d)}$, then $R_\varepsilon \widetilde{u}$ satisfies the boundary condition (3.5) on S_1 and $R_\varepsilon^* v$ the adjoint conditions (3.8) on S_1. Then $(\hat{L}^* R_\varepsilon^*)^* \widetilde{u} = R_\varepsilon \widetilde{f}_1$ and we have to prove only the convergence (3.1). Using the representation (3.2) we must deal with the problems of convergence of the terms involving (a) $\partial/\partial t$, (b) $\partial/\partial\rho$, (c) $\partial/\partial\varphi$, (d) B_2.

It is easy to handle problem (a) because after the transformation (3.6), A^t depends only on t, and not on ρ or φ; we have estimates such as

$$\left|K(t) - K\left(\overline{t}\right)\right| \le M\left|t - \overline{t}\right| \le 3M\varepsilon_1 \text{ on } \operatorname{supp} k_\varepsilon. \tag{3.11}$$

The most interesting is problem (b), that is, to establish the convergence in $L_2(D)$ as $\varepsilon \to 0$ of the expression

$$I_\varepsilon = \int_D \frac{\partial}{\partial\overline{\rho}} \left\{A^\rho(z)k_\varepsilon(y-z) - k_\varepsilon(y-z)A^\rho(y)\right\} \widetilde{u}(z)dz. \tag{3.12}$$

Note that $A^\rho(z)K_\varepsilon(y-z) - K_\varepsilon(y-z)A^\rho(y) =$

$$\begin{pmatrix} 0 & 0 & J_1 - J_3 \\ 0 & \left\{a_1\left(\overline{t},\overline{\rho}\right) - a_1(t,\rho)\right\} J_2 & 0 \\ K\left(\overline{t}\right) J_3 - K(t) J_1 & 0 & 0 \end{pmatrix},$$

where $a_1(t,\rho) = \rho K(t)/a(t,\rho)$. We remark that S_1 is a characteristic surface and that we have some components which are simultaneously "free" for both the boundary and the adjoint boundary conditions. This fact is crucial and explains why we can choose the kernel of the mollifier in such a way that J_1 and J_3 depend in the same way on ρ.

Note that for any constant c and any $w \in L_2(D)$ with $w = 0$ in $\mathbb{R}^3 \backslash D$,

$$\left\| \int_{\mathbb{R}} \frac{1}{\varepsilon_1} j\left(\frac{t-\overline{t}}{\varepsilon_1} + c\right) w\left(\overline{t},\rho,\varphi\right) d\overline{t} - w(t,\rho,\varphi) \right\|_{L_2(D)} \to 0 \tag{3.13}$$

as $\varepsilon_1 \to 0$. In our case, we have

$$I'_\varepsilon := \left\| \left(\varepsilon_1 \varepsilon_2^2 \varepsilon_3\right)^{-1} \int_D \left\{ j\left(\frac{t-\bar{t}}{\varepsilon_1}+2\right) - j\left(\frac{t-\bar{t}}{\varepsilon_1}-2\right)\right\} j'\left(\frac{\rho-\overline{\rho}}{\varepsilon_2}+E\right) j\left(\frac{\varphi-\overline{\varphi}}{\varepsilon_3}\right) \widetilde{u}_3(z)dz \right\|_{L_2(D)} \to 0 \tag{3.14}$$

as $\varepsilon_1 \to 0$, ε_2 and ε_3 being fixed. In a similar way we find, using (3.13), that

$$I''_\varepsilon := \left\| \left(\varepsilon_1 \varepsilon_2^2 \varepsilon_3\right)^{-1} \int_D \left\{ K\left(\bar{t}\right) j\left(\frac{t-\bar{t}}{\varepsilon_1}+2\right) - K(t) j\left(\frac{t-\bar{t}}{\varepsilon_1}-2\right)\right\} j'\left(\frac{\rho-\overline{\rho}}{\varepsilon_2}+E\right) j\left(\frac{\varphi-\overline{\varphi}}{\varepsilon_3}\right) \widetilde{u}_1(z)dz \right\|_{L_2(D)} \to 0 \tag{3.15}$$

as $\varepsilon_1 \to 0$, ε_2 and ε_3 being fixed. We also have that

$$I'''_\varepsilon := \left\| \left(\varepsilon_1 \varepsilon_2 \varepsilon_3\right)^{-1} \int_D \frac{\partial}{\partial \overline{\rho}} \left\{ \left[a_1(\bar{t},\bar{\rho}) - a_1(t,\rho)\right] j\left(\frac{\rho-\overline{\rho}}{\varepsilon_2}-E\right)\right\} j\left(\frac{t-\bar{t}}{\varepsilon_1}+2\right) j\left(\frac{\varphi-\overline{\varphi}}{\varepsilon_3}\right) \widetilde{u}_2(z)dz \right\|_{L_2(D)} \to 0 \tag{3.16}$$

as $\varepsilon_1 \to 0$, $\varepsilon_2 \to 0$ with $0 < \varepsilon_1 \le \varepsilon_2$, and with ε_3 being fixed, since

$$\left|a_1\left(\bar{t},\overline{\rho}\right) - a_1(t,\rho)\right| \varepsilon_2^{-1} \le M\left\{\left|t-\bar{t}\right| + |\rho-\overline{\rho}|\right\}\varepsilon_2^{-1} \le M_1(\varepsilon_1\varepsilon_2^{-1}+1). \tag{3.17}$$

This completes the discussion of problem (b). Problems (c) and (d), involving A_2^φ and B_2 (see (3.2), (3.7)), are handled in a standard way. The investigation in a neighborhood of a point from $S_0 \cap S_1$ is finished. To deal with the point $(0,0,0)$ we follow the Lax–Phillips scheme [17] for the two-dimensional problem, but with some changes. We use a cut-off function $\psi \in C^\infty(\mathbb{R})$, with $\psi(s) = 0$ for $s \le 1$ and $\psi(s) = 1$ for $s \ge 2$, and for each $m \in \mathbb{N}$ define $\psi_m \in C^\infty\left(\overline{G}\right)$ by $\psi_m(t,\rho) = \psi\left(m\left(\rho + t^{3/2}\right)\right)$. Note that the function $\hat{u}_m := \psi_m \hat{u}$ is a weak solution of the system

$$\hat{L}\hat{u} = \hat{f}_m := \psi_m \hat{f} + m\left(\frac{x_1}{\rho}A^1 + \frac{x_2}{\rho}A^2 + \frac{3}{2}\sqrt{t}A^3\right)\psi'\left(m\left(\rho+t^{3/2}\right)\right)\hat{u}.$$

To finish the proof of Theorem 2.4 we need the next lemma, which follows in a standard way, using the fact that on $\mathrm{supp}\psi'$, we have $\rho \le 2/m$ and $t \le 2/m^{2/3}$.

LEMMA 3.1. *Let $\hat{u} \in H_*(G)$ and $\hat{f} \in H^*(G)$. Then*

$$\left\|\hat{f}_m - \hat{f}\right\|^* \to 0 \text{ as } m \to \infty. \tag{3.18}$$

The proof of Theorem 2.4 is complete. □

THEOREM 3.2. *Let $\hat{f} \in H^*(G)$ and suppose that $\hat{u} \in H_*(G)$ is a weak solution of the boundary-value problem* (2.5), (2.10). *Then it is a strong solution, it is unique, and it satisfies the a priori estimate*

$$\|\hat{u}\|_* \le C_\alpha \left\|\hat{f}\right\|^*. \tag{3.19}$$

Proof. We wish to use Theorems 2.3 and 2.4. For a weak solution $\hat{u}$ from Theorem 2.3 and for each $m \in \mathbb{N}$ define functions $\hat{u}_m = \psi_m \hat{u}$ as above. We know that $\hat{u}_m = 0$ in $G \cap \left\{\rho + t^{3/2} \le m^{-1}\right\}$. From the results above it follows that far from $(0,0,0)$ every weak solution is a strong solution. But $\hat{u}_m$ satisfies this condition. In view of this and Lemma 3.1 the proof is complete. □

4. The nonlocal problem: Existence of a generalized solution. Here we deal with Problem A. We recall that we consider the curve $t = C\rho^{-\alpha} - \alpha\rho/(\alpha+1)$, α being a small positive parameter, and that this curve contains the point (t_0, ρ_0) provided that $C = \{t_0 + \alpha\rho_0/(\alpha+1)\}\rho_0^{\alpha}$; under this condition we denote by $(p_\alpha(t_0,\rho_0), q_\alpha(t_0,\rho_0))$ the common point of the curve and the curve $\rho = \int_0^t \sqrt{K(\tau)}d\tau$ (for the details see section 7).

DEFINITION 4.1. *A function $\omega(t,\rho,\varphi)$ is called a generalized solution of Problem* A *if* (a) $\omega, \frac{\partial\omega}{\partial t}, \sqrt{r}\frac{\partial\omega}{\partial\rho} \in L_2(G)$; (b) $\sqrt{r}\rho^{-1}\frac{\partial\omega^1}{\partial\varphi} \in L_2(G)$, *where*

$$\omega^1(t,\rho,\varphi) := \omega(t,\rho,\varphi) - \omega\left(p_\alpha(t,\rho),\ q_\alpha(t,\rho),\varphi\right); \tag{4.1}$$

(c) $\omega = 0$ *on* $S_0 \cup S_1$; (d) *for any function* $v \in C^* := \{v \in C^1(\overline{G}) : v = 0$ *on* $S_0 \cup S_2$, *and in some neighborhood of* $(0,0,0)\}$, *the equality*

$$\int_G \left(\frac{\partial\omega}{\partial t}\frac{\partial v}{\partial t} - K(t)\frac{\partial\omega}{\partial\rho}\frac{\partial v}{\partial\rho} - \frac{K(t)}{\rho^2}\frac{\partial\omega^1}{\partial\varphi}\frac{\partial v}{\partial\varphi} - fv\right)\rho\, d\rho\, d\varphi\, dt = 0 \tag{4.2}$$

holds.

Remark. The trace in (4.1) exists because $p_\alpha, q_\alpha \in C^1(G)$ (see section 7).

THEOREM 4.2. *Suppose that the parameter $\alpha > 0$ is so small that conditions* (E1) *and* (E2) *are satisfied. Then for any $f \in L_2(G)$, there is a generalized solution of Problem* A.

Proof. Let $f \in L_2(G)$; then $\hat{f} = (-x_1 f, -x_2 f, \alpha(\rho+t)f) \in H^*(G)$. By Theorem 2.3, there exists a weak solution $\hat{u} = (u_1, u_2, u_3) \in H_*(G)$ of the problem (2.5), (2.10); by Theorem 2.4, it is a strong solution. Hence there are functions $\hat{u}_m = (u_{m1}, u_{m2}, u_{m3}) \in C^1\left(\overline{G}\right)$ $(m \in \mathbb{N})$ such that

$$\begin{aligned} &x_1 u_{m1} + x_2 u_{m2} = 0 \text{ on } S_0, \sqrt{K}\left(x_1 u_{m1} + x_2 u_{m2}\right) - \rho u_{m3} = 0 \text{ on } S_1, \\ &x_1 u_{m2} - x_2 u_{m1} = 0 \text{ on } S_2, \hat{u}_m = 0 \text{ in a neighborhood of } (0,0,0) \end{aligned} \tag{4.3}$$

and

$$\|\hat{u}_m - \hat{u}\|_* \to 0, \left\|\hat{L}\hat{u}_m - \hat{f}\right\|^* \to 0 \text{ as } m \to \infty. \tag{4.4}$$

Recalling that $\hat{L} = \Lambda L_0$ we put $u_{m\rho} = (x_1 u_{m1} + x_2 u_{m2})/\rho$, $u_{m\varphi} = x_1 u_{m2} - x_2 u_{m1}$,

$$\begin{cases} w_{m1} = L_0^1 \hat{u}_m := K(t)\left(\dfrac{\partial u_{m1}}{\partial x_1} + \dfrac{\partial u_{m2}}{\partial x_2}\right) - \dfrac{\partial u_{m3}}{\partial t}, \\ w_{m2} = (1/K(t))\, L_0^2 \hat{u}_m := \dfrac{a}{\rho}\left(\dfrac{\partial u_{m\varphi}}{\partial t} - \dfrac{\partial u_{m3}}{\partial\varphi}\right) - \left(\dfrac{\partial u_{m\varphi}}{\partial\rho} - \dfrac{\partial u_{m\rho}}{\partial\varphi}\right), \\ w_{m3} = L_0^3 \hat{u}_m := \dfrac{\partial u_{m\rho}}{\partial t} - \dfrac{\partial u_{m3}}{\partial\rho}. \end{cases} \tag{4.5}$$

Since $\det \Lambda = \rho^2 - a^2 K \geq c\left(\rho^2 + t^3\right)$ in G, $c > 0$, the second part of (4.4) gives

$$\left\| \left(\rho^2 + t^3\right)^{1/3} (w_{m1} - f) \right\|_{L_2(G)} \to 0, \|t w_{m2}\|_{L_2(G)} \to 0, \tag{4.6}$$

$$\left\| \left(\rho^2 + t^3\right)^{1/2} w_{m3} \right\|_{L_2(G)} \to 0. \tag{4.7}$$

Now we define functions $\omega_m \in C^1\left(\overline{G}\right)$ by

$$\omega_m(t, \rho, \varphi) = \int_0^t u_{m3}(\tau, \rho, \varphi)\, d\tau \quad (m \in \mathbb{N}). \tag{4.8}$$

In view of (4.4), there is a function $\omega \in L_2(G)$ such that

$$\|\omega_m - \omega\| \to 0, \left\| \frac{\partial u_m}{\partial t} - u_3 \right\| \to 0 \text{ as } m \to \infty, \tag{4.9}$$

and $\omega_t = u_3 \in L_2(G)$. From the boundary conditions (2.10) and from (4.7), for every $\delta > 0$ we obtain

$$\left\| \frac{\partial \omega_m}{\partial \rho} - (x_1 u_1 + x_2 u_2)/\rho \right\|_{L_2(G_\delta)} \to 0 \text{ as } m \to \infty, \tag{4.10}$$

where G_δ is the set of all those points of G with $\rho > \delta$. Hence

$$\frac{\partial \omega}{\partial \rho} = (x_1 u_1 + x_2 u_2)/\rho,\ \sqrt{r}\frac{\partial \omega}{\partial \rho} \in L_2(G).$$

The function ω satisfies the required boundary conditions (1.2) on $S_0 \cup S_1$. Thus ω satisfies conditions (a) and (c) of Definition 4.1.

We can now turn to the derivative with respect to φ. Put

$$Q_m := \frac{\partial \omega_m}{\partial \varphi} - u_{m\varphi} = \frac{\partial \omega_m}{\partial \varphi} - (x_1 u_{m2} - x_2 u_{m1}). \tag{4.11}$$

Use of the equations in system (2.2) and notation (4.5) shows that Q_m satisfies

$$\frac{a}{\rho}\frac{\partial Q_m}{\partial t} - \frac{\partial Q_m}{\partial \rho} = -w_{m2} + \frac{a}{\rho}\frac{\partial}{\partial \varphi}\int_0^t w_{m3}(\tau, \rho, \varphi) d\tau := -G_m(t, \rho, \varphi). \tag{4.12}$$

To integrate (4.12) we observe that $a = \alpha(\rho + t)$, make the change of variables $\xi = (t + \alpha\rho/(\alpha + 1))\rho^\alpha$, $s = \rho$, and write

$$\overline{Q}_m(\xi, s, \varphi) := Q_m\left(\xi s^{-\alpha} - \frac{\alpha}{\alpha + 1} s, s, \varphi\right), \tag{4.13}$$

with similar definition for $\overline{G}_m$. From (4.12) we have

$$\overline{Q}_m(\xi, s, \varphi) = F_m(\xi, \varphi) + \int_{s(\xi)}^s \overline{G}_m(\xi, \sigma, \varphi) d\sigma. \tag{4.14}$$

We choose $s(\xi_0)$ to be the ρ-coordinate of the point of intersection of S_2 and the curve $\xi = \xi_0$. Since $\overline{u}_{m\varphi}(\xi, s(\xi), \varphi) = 0$, by (4.3), then

$$\overline{Q}_m(\xi, s, \varphi) - \frac{\partial \overline{\omega}_m}{\partial \varphi}(\xi, s(\xi), \varphi) = \int_{s(\xi)}^s \overline{G}_m(\xi, \sigma, \varphi) d\sigma \tag{4.15}$$

and we finally have

$$\begin{aligned} Q_m(t,\rho,\varphi) \quad &- \frac{\partial\overline{\omega}_m}{\partial\varphi}\left(t\rho^\alpha + \frac{\alpha}{\alpha+1}\rho^{\alpha+1},\ s\left(t\rho^\alpha + \frac{\alpha}{\alpha+1}\rho^{\alpha+1}\right),\varphi\right) \\ &= \int_{s\left(t\rho^\alpha+\frac{\alpha}{\alpha+1}\rho^{\alpha+1}\right)}^{s} \overline{G}_m d\sigma. \end{aligned} \tag{4.16}$$

But this means that

$$\frac{\partial\omega_m^1}{\partial\varphi} - u_{m\varphi} = \int_{q_\alpha}^{\rho} \overline{G}_m(\xi,s,\varphi)ds \tag{4.17}$$

according to the notation (4.1). The integral

$$\int_{q_\alpha(t,\rho)}^{\rho} \overline{G}_m(\xi,s,\varphi)ds = \int_{q_\alpha(t,\rho)}^{\rho} G_m\left(\left(\frac{\rho}{s}\right)^\alpha\left(t+\frac{\alpha}{\alpha+1}\rho\right) - \frac{\alpha}{\alpha+1}s, s, \varphi\right)ds$$

can be split into two parts, by (4.12). For these we have the following estimates: for every $\delta > 0$,

$$\left\|\int_{q_\alpha}^{\rho} w_{m2}ds\right\|_{L^2(G^\delta)} \to 0, \quad \left\|\int_{q_\alpha}^{\rho}\int_0^t w_{m3}d\tau ds\right\|_{L^2(G^\delta)} \to 0, \tag{4.18}$$

as $m \to \infty$; here G^δ is the set of all those points of G with $t > \delta$. To prove this we need the properties of w_{m2} and w_{m3} given by (4.6) and (4.7), together with the fact that if some characteristic of the first-order partial differential equation (4.12) starts from a point on S_2 far from $(0,0,0)$, then it will remain far from $(0,0,0)$ at all times, so we have some uniformity. It should also be observed that the weight t in the estimate (4.6) for w_{m2} causes no difficulty in our case, because starting from a point in G we integrate along a curve through this point which goes to S_2 but remains away from $t = 0$. These considerations enable us to prove that for all $\psi \in C_0^\infty(G)$,

$$-\left(\omega^1, \frac{\partial\psi}{\partial\varphi}\right) = (u_\varphi, \psi), \tag{4.19}$$

that is, $\frac{\partial\omega^1}{\partial\varphi} = (-x_2u_1 + x_1u_2)$, $\frac{\sqrt{r}}{\rho}\frac{\partial\omega^1}{\partial\varphi} \in L_2(G)$, and the proof that ω satisfies condition (b) of Definition 4.1 is complete.

As for condition (d), let $v \in C^1(\overline{G})$ be such that $\frac{\partial v}{\partial\varphi} \in C^1(\overline{G})$ and v vanishes on S_2 and in some neighborhood of S_0. Then from (4.5), (4.6), and (4.7) it follows that

$$\left(K\frac{\partial u_{m1}}{\partial x_1} + K\frac{\partial u_{m2}}{\partial x_2} - \frac{\partial u_{m3}}{\partial t}, v\right) \to (f,v) \text{ as } m \to \infty. \tag{4.20}$$

Integrating by parts in view of the boundary conditions (4.3) and with the aid of the representations of derivatives of ω_m, it now follows that (4.2) holds, when v is restricted as stated above. That condition (d) is fulfilled for the given class of functions v results from a density argument. The proof of Theorem 4.2 is complete. □

5. Uniqueness of the generalized solution.

THEOREM 5.1. *Let $f \in L_2(G)$ and suppose that the positive parameter α is so small that it satisfies conditions* (E1) *and* (E2). *Then Problem* A *has at most one*

generalized solution, and there is a constant C such that if ω is a generalized solution it satisfies the a priori estimate

$$\|\omega\| + \left\|\frac{\partial\omega}{\partial t}\right\| + \left\|\sqrt{r}\frac{\partial\omega}{\partial\rho}\right\| + \left\|\frac{\sqrt{r}}{\rho}\frac{\partial\omega^1}{\partial\varphi}\right\| \le C\|f\|, \tag{5.1}$$

where we recall that $\|\cdot\|$ stands for $\|\cdot\|_{L^2(G)}$.

Proof. Suppose that ω is a generalized solution of Problem A. Put

$$\begin{aligned} &u_\rho = \frac{\partial\omega}{\partial\rho}, \quad u_3 = \frac{\partial\omega}{\partial t}, \quad u_\varphi = \frac{\partial\omega^1}{\partial\varphi}, \quad u_1 = x_1u_\rho\rho^{-1} - x_2u_\varphi\rho^{-2}, \\ &u_2 = x_2u_\rho\rho^{-1} + x_1u_\varphi\rho^{-2}. \end{aligned} \tag{5.2}$$

Then u_3, $\sqrt{r}u_1$, $\sqrt{r}u_2 \in L_2(G)$. We claim that the following result holds.

LEMMA 5.2. *The function $\hat{u} = (u_1, u_2, u_3)$ is a weak solution of problem* (2.5), (2.10) *for the function $\hat{f} = (-x_1f, -x_2f, \alpha(\rho+t)f) \in H^*(G)$.*

Assuming this for the moment, it follows from Theorem 3.2 that $\hat{u}$ is a strong solution of (2.5), (2.10), that it is unique, and that it satisfies the inequality (2.5), which in the notation (5.2) means

$$\|\omega_t\| + \left\|\sqrt{r}\omega_\rho\right\| + \left\|\sqrt{r}\rho^{-1}\omega^1_\varphi\right\| \le C\|f\|.$$

Since $\omega = 0$ on S_0 and $\omega = \int_0^t \omega_t(\tau,\rho,\varphi)d\tau$, the a priori estimate (5.1) follows.

To conclude the proof of Theorem 5.1 we establish Lemma 5.2. From (4.2) it follows that if $w_1 \in C^*$, i.e., $w_1 \in C^1(\overline{G})$, w_1 is zero on $S_0 \cup S_2$ and in a neighborhood of $(0,0,0)$, then (with $(\cdot,\cdot)$ denoting the inner product in $L_2(G)$)

$$\left(u_3, \frac{\partial w_1}{\partial t}\right) - \left(Ku_1, \frac{\partial w_1}{\partial x_1}\right) - \left(Ku_2, \frac{\partial w_1}{\partial x_2}\right) = (f, w_1). \tag{5.3}$$

If $w_3 \in C^2\left(\overline{G}\right)$ is zero on S_2 and in a neighborhood of $(0,0,0)$, then plainly

$$\left(u_\rho, \frac{\partial w_3}{\partial t}\right) - \left(u_3, \frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho w_3)\right) = 0. \tag{5.4}$$

If $w_2 \in C^2\left(\overline{G}\right)$ is zero on S_1 and in a neighborhood of S_0, then we claim that

$$I(\omega) := \left(au_3 - \rho u_\rho, \frac{1}{\rho}\frac{\partial}{\partial\varphi}(Kw_2)\right) - \left(\frac{1}{\rho}u_\varphi, \frac{\partial}{\partial t}(aKw_2) - \frac{\partial}{\partial\rho}(\rho Kw_2)\right) = 0. \tag{5.5}$$

To verify this, observe that we may approximate the function $\omega(t,\rho,\varphi)$ by smooth functions $\omega_m(t,\rho,\varphi)$ in $W_2^1(\mathrm{supp}w_2)$. Using the notation ω_m^1 from (4.1) we see that $\omega_m^1 = 0$ on S_2, by definition of the functions p_α and q_α. Then $I(\omega_m) \to I(\omega)$ as $m \to \infty$; we see that $I(\omega_m) = 0$. This follows from the boundary conditions and because the functions $\omega_m(p_\alpha(t,\rho), q_\alpha(t,\rho), \varphi)$ depend actually only on φ and $\xi = t\rho^\alpha + \dfrac{\alpha}{\alpha+1}\rho^{\alpha+1}$. It is easy to check in that case that

$$\left\{\frac{\partial}{\partial\rho} - \frac{\alpha}{\rho}(t+\rho)\frac{\partial}{\partial t}\right\}\omega_m(p_\alpha(t,\rho), q_\alpha(t,\rho), \varphi) = 0. \tag{5.6}$$

From all this (5.5) follows. From (5.3), (5.4), (5.5), and the representation of the operator $\hat{L}_0$ we have

$$\left(\hat{u}, \hat{L}_0^*\hat{w}\right) = \left(\hat{f}_1, \hat{w}\right) = (f, w_1) \tag{5.7}$$

for every function $\hat{w} = (w_1, w_2, w_3)$ such that

$$\begin{cases} \hat{w} \in C^2\left(\overline{G}\right), \hat{w} = 0 \text{ in a neighborhood of } (0,0,0), \\ w_1 = 0 \text{ on } S_0 \cup S_2, w_2 = 0 \text{ on } S_1, w_3 = 0 \text{ on } S_2. \end{cases} \tag{5.8}$$

Recalling that $\hat{L} = \Lambda \hat{L}_0$, we are led to solve the equation $\hat{w} = \Lambda^* \hat{v}$. This gives

$$\begin{cases} w_1 = -\left(x_1 v_1 + x_2 v_2 - a v_3\right), \ w_2 = \left(x_2 v_1 - x_1 v_2\right)/\rho, \\ w_3 = \rho v_3 - Ka(x_1 v_1 + x_2 v_2)\rho^{-1}. \end{cases} \tag{5.9}$$

It is easy to see that for every $\hat{v} \in C^2\left(\overline{G}\right)$ which satisfies conditions (2.11) and the additional conditions

$$x_1 v_1 + x_2 v_2 = 0, \ v_3 = 0 \text{ on } S_2, \tag{5.10}$$

the corresponding function $\hat{w}$ given by (5.9) satisfies (5.8), and so the equality

$$\left(\hat{u}, \hat{L}^* \hat{v}\right) = \left(\hat{f}, \hat{v}\right), \ \hat{f} = (-x_1 f, -x_2 f, af)$$

holds for all such functions $\hat{v}$. Some density arguments remove the additional conditions (5.10) and thus show that $\hat{u}$ is a weak solution of problem (2.5), (2.10).

The proof of Lemma 5.2 is complete. □

6. Smoothness of the generalized solution. The matter investigated here is the smoothness with respect to φ of the generalized solution of the nonlocal Problem A. We begin with the following lemma.

LEMMA 6.1. *Suppose that the function ω satisfies all conditions from Definition* 4.1*, except that the equality* (4.2) *holds only for all functions $v \in C_0^\infty(G)$. Then ω is a generalized solution of Problem* A*.*

We omit this proof, because we shall not use this fact here.

THEOREM 6.2. *Suppose that the positive parameter α satisfies* (E1) *and* (E2)*, and let $f \in L_2(G)$ be such that $\frac{\partial^k f}{\partial \varphi^k} \in L_2(G)$ for $k = 1, \ldots, \ell$ for some $\ell \in \mathbb{N}$. Then there is a unique generalized solution ω of Problem* A *for which*

$$\frac{\partial^k \omega}{\partial \varphi^k} \in \widetilde{W}_2^1(G) \ (k = 0, \ldots, \ell - 1), \ \frac{\partial^{\ell+1} \omega}{\partial t \partial \varphi^\ell} \in L_2(G), \ \sqrt{r}\frac{\partial^{\ell+1} \omega}{\partial \rho \partial \varphi^\ell} \in L_2(G) \tag{6.1}$$

and

$$\sum_{k=0}^{\ell-1} \left\| \frac{\partial^k \omega}{\partial \varphi^k} \right\|_{\widetilde{W}_2^1(G)} + \left\| \frac{\partial^{\ell+1} \omega}{\partial t \partial \varphi^\ell} \right\|_{L_2(G)} + \left\| \sqrt{r}\frac{\partial^{\ell+1} \omega}{\partial \rho \partial \varphi^\ell} \right\|_{L_2(G)} \le C_\alpha \sum_{k=0}^{\ell} \left\| \frac{\partial^k f}{\partial \varphi^k} \right\|_{L_2(G)}. \tag{6.2}$$

Proof. First suppose that $g \in L_2(G)$ is such that $\partial g/\partial \varphi \in L_2(G)$. Let u be a generalized solution of Problem A, with $f = \partial g/\partial \varphi$; Theorem 4.2 ensures that such a solution exists. Then $u, \partial u/\partial t, \sqrt{r}\partial u/\partial \rho, \sqrt{r}\rho^{-1}\partial u^1/\partial \varphi \in L_2(G)$, $u = 0$ on $S_0 \cup S_1$, and

$$(u_t, v_t) - (Ku_\rho, v_\rho) - \left(K\rho^{-2}u_\varphi^1, v_\varphi\right) = (g_\varphi, v) = (g, v_\varphi) \tag{6.3}$$

for all $v \in C^*$. We have also in each set $G_\delta = G \cap \{\rho > \delta\}$ an approximation sequence $\{\omega_m\} \subset C^1(\bar{G}_\delta)$, as in Theorem 4.2.

Put $u_1(t,\rho,\varphi) = \int_0^{2\pi} u(t,\rho,\varphi)d\varphi$; then we have

$$\frac{\partial u_1}{\partial t} = \int_0^{2\pi} \frac{\partial u}{\partial t}(t,\rho,\varphi)d\varphi \in L_2(G), \frac{\partial u_1}{\partial \rho} = \int_0^{2\pi} \frac{\partial u}{\partial \rho}(t,\rho,\varphi)d\varphi, \frac{\partial u_1}{\partial \varphi} = 0. \tag{6.4}$$

Moreover, $u_1 = 0$ on $S_0 \cup S_1$. Let $v_1 \in C^*$ be independent of φ. Then from (6.3) we have

$$\int_\Delta \left(\frac{\partial u_1}{\partial t}\frac{\partial v_1}{\partial t} - K\frac{\partial u_1}{\partial \rho}\frac{\partial v_1}{\partial \rho} - \frac{K}{\rho^2}\frac{\partial u_1^1}{\partial \varphi}\frac{\partial v_1}{\partial \varphi} \right) \rho d\rho dt = 0, \tag{6.5}$$

where $\Delta = \{(t,\rho) : t > 0, \int_0^t \sqrt{K(\tau)}d\tau < \rho < 1 - \int_0^t \sqrt{K(\tau)}d\tau\}$.

Given $v \in C^*$, define $v_1 = \int_0^{2\pi} v d\varphi$; then (6.5) holds for this v_1 and we have

$$\int_G \left(\frac{\partial u_1}{\partial t}\frac{\partial v}{\partial t} - K\frac{\partial u_1}{\partial \rho}\frac{\partial v}{\partial \rho} - \frac{K}{\rho^2}\frac{\partial u_1^1}{\partial \varphi}\frac{\partial v}{\partial \varphi} \right) dx = 0 \tag{6.6}$$

for all $v \in C^*$. This shows that u_1 is a generalized solution of Problem A with $f = 0$. From the uniqueness part of Theorem 5.1 we have $u_1 = 0$ in G, that is, $\int_0^{2\pi} u(t,\rho,\varphi)d\varphi = 0$ for $(t,\rho) \in \Delta$.

Define

$$y(t,\rho,\varphi) = \int_0^\varphi u(t,\rho,\lambda)d\lambda. \tag{6.7}$$

Since $u_1 = 0$ in G, this function has a generalized φ-derivative $\frac{\partial y}{\partial \varphi} = u$, and by (6.3),

$$(y_t, v_{\varphi t}) - (Ky_\rho, v_{\varphi\rho}) - \left(K\rho^{-2}y_\varphi^1, v_{\varphi\varphi}\right) = (g, v_\varphi) \tag{6.8}$$

for all $v \in C^*$ such that $v_\varphi \in C^1\left(\overline{G}\right)$. Note that if $g \in L_2(G)$, there is a generalized solution ω of Problem A with $f = g$; for this solution (6.8) is also satisfied. With $z = \omega - y$ we then have

$$(z_t, v_{\varphi t}) - (Kz_\rho, v_{\varphi\rho}) - \left(K\rho^{-2}z_\varphi^1, v_{\varphi\varphi}\right) = 0 \tag{6.9}$$

for the same functions v. If $\psi \in C^2\left(\overline{G}\right) \cap C^*$, then

$$v_\varphi(t,\rho,\varphi) := \psi(t,\rho,\varphi) - \frac{1}{2\pi}\int_0^{2\pi} \psi(t,\rho,\varphi)d\varphi \tag{6.10}$$

can be substituted in (6.9), giving

$$\begin{aligned} \left(\left(z - \frac{1}{2\pi}\int_0^{2\pi} z d\varphi \right)_t, \psi_t \right) &- \left(K\left(z - \frac{1}{2\pi}\int_0^{2\pi} z d\varphi \right)_\rho, \psi_\rho \right) \\ &- \left(K\rho^{-2}\left(z - \frac{1}{2\pi}\int_0^{2\pi} z d\varphi \right)_\varphi^1, \psi_\varphi \right) = 0. \end{aligned} \tag{6.11}$$

It follows from Theorem 5.1 that

$$z(t,\rho,\varphi) - \frac{1}{2\pi}\int_0^{2\pi} z(t,\rho,\varphi)d\varphi = 0 \text{ in } G, \tag{6.12}$$

that is, $\omega_\varphi = u$ in G. The properties of the generalized solutions ω and u now lead to the conclusion of the theorem, once the estimate

$$\left\|\sqrt{r}\rho^{-1}\omega_\varphi\right\| \le C\,\|f_\varphi\| \tag{6.13}$$

has been established. To do this we use the fact that $\omega = 0$ in S_0 and the following lemma.

LEMMA 6.3. *For every* $w \in L_2(G)$,

$$\left\|\sqrt{r}\rho^{-1}\int_0^t w(\tau,\rho,\varphi)d\tau\right\| \le C\,\|w\|\,.$$

The proof of Theorem 6.2 is complete. □

7. Local and nonlocal problems, and their connection. We investigate here the connection between solutions of Problem P and those of Problem A.

DEFINITION 7.1. *A function* $u \in \tilde{W}_2^1(G)$ *is called a generalized solution of Problem* P *if, and only if,* $u = 0$ *on* $S_0 \cup S_1$ *and*

$$\int_G \left\{u_t v_t - K(t)u_\rho v_\rho - K(t)\rho^{-2}u_\varphi v_\varphi - fv\right\}dx = 0 \tag{7.1}$$

holds for all $v \in C^*$.

We shall also consider the following problem.

Problem P_φ. Is there a generalized solution of Problem P which satisfies the extra condition

$$u_\varphi = 0 \text{ on } S_2 \text{ in a weak sense,} \tag{7.2}$$

that is, $\int_{S_2} uv_\varphi\,ds = 0$ for all $v \in C_0^\infty(S_2)$?

For the Tricomi equation, Problem P_φ was formulated by Didenko [11].

Denote by H the set of all $f \in L_2(G)$ for which there exists a generalized solution u_f of Problem P; denote by H_φ the subspace of functions $f \in H$ for which the solution u_f satisfies (7.2). We know that in the case of the Tricomi equation, where $K(t) = t$,

$$\dim\left(L_2(G)/H\right) = \infty \tag{7.3}$$

since all the functions v_n given in (1.4) are orthogonal to H.

Remark. Sorokina [33] studied a variant of Problem P_φ: she called a function $u \in \tilde{W}_2^1(G)$ a "generalized solution" of this problem if, and only if, u is a solution of equation (1.1) in the sense of distributions (that is, (7.1) holds for every $v \in C_0^\infty(G)$), $u = 0$ on $S_0, u_\varphi = 0$ on S_2, and $u_\nu = 0$ on S_1 in a weak sense, where $u_\nu = K(n_1u_{x_1} + n_2u_{x_2}) - n_3u_{x_3}$ is the conormal derivative. Theorem 2 of Sorokina [33] states that given any $f \in L_2(G)$, there is a "generalized solution" of the problem, and any such solution is a strong solution. This result seems to us to be incorrect: the problem as formulated in Sorokina [33] appears to be strongly over determined (compare with (7.3) and Lemma 6.1).

We shall examine the uniqueness and continuous dependence on the data of solutions of Problem P_φ. The nonlocal Problem A will give some information about this.

THEOREM 7.2. *Let* $u \in \tilde{W}_2^1(G)$ *be a generalized solution of Problem* P_φ. *Then it is also a generalized solution of Problem* A *and so is unique and satisfies the a priori estimate*

$$\|u\|_{\tilde{W}_2^1(G)} \le C\,\|f\|_{L_2(G)}\,. \tag{7.4}$$

Proof. Let $v \in C^2(\overline{G})$ be zero in some neighborhoods of $S_1 \cap S_2$, $S_0 \cap S_1$, and $(0,0,0)$; suppose that $\alpha > 0$ is so small that conditions (E1) and (E2) hold. We investigate the nonlocal term in the integral equality (4.2):

$$\begin{aligned} I(v) := & \int_G K(t)\rho^{-2}u\left(p_\alpha(t,\rho), q_\alpha(t,\rho), \varphi\right) v_{\varphi\varphi}(t,\rho,\varphi)\rho d\rho dt d\varphi \\ = & \int_{S_2} u(s) w_\varphi(s) ds \end{aligned} \tag{7.5}$$

for some $w \in C_0^1(S_2)$, obtained by integration of v over the lines $t = C\rho^{-\alpha} - \alpha\rho/(\alpha+1)$. More precisely, let (t_0, ρ_0, φ) be a point in the domain G. This lies on the curve with equation $t = C\rho^{-\alpha} - \alpha\rho/(\alpha+1)$, where $C = (t_0 + \alpha\rho_0/(\alpha+1))\rho_0^\alpha$. This curve intersects $\rho = \int_0^t \sqrt{K(\tau)}d\tau$ in a point $(p_\alpha(t_0,\rho_0), q_\alpha(t_0,\rho_0), \varphi)$, where $t = p_\alpha(t_0,\rho_0)$ is a solution of the equation

$$F(t) := t\left[\int_0^t \sqrt{K(\tau)}d\tau\right]^\alpha + \frac{\alpha}{\alpha+1}\left[\int_0^t \sqrt{K(\tau)}d\tau\right]^{\alpha+1} = t_0\rho_0^\alpha + \alpha\rho_0^{\alpha+1}/(\alpha+1);$$

this exists because $F(t) > 0$ for $t > 0$ and $F(0) = 0$. Then

$$p_\alpha(t_0,\rho_0) = F^{-1}\left(t_0\rho_0^\alpha + \alpha\rho_0^{\alpha+1}/(\alpha+1)\right), \quad q_\alpha(t_0,\rho_0) = \int_0^{p_\alpha(t_0,\rho_0)} \sqrt{K(\tau)}d\tau.$$

We note here that $0 < p_\alpha(t_0,\rho_0) < d$; more precisely, the curve $t = C\rho^{-\alpha} - \alpha\rho/(\alpha+1)$ crosses every characteristic $\rho = \beta - \int_0^t \sqrt{K(\tau)}d\tau$ at least once, because at the common point (t_0,ρ_0) we have

$$t_1'(\rho_0) = -a(t_0,\rho_0)/\rho_0 > -1 \Big/ \sqrt{K(t_0)} = t_2'(\rho_0),$$

in view of condition (E1).

Let us denote by $(t,\psi_1(t))$ and $(t,\psi_2(t))$, $\psi_1(t) < \psi_2(t)$, the points of intersection of the curve $t = C\rho^{-\alpha} - \alpha\rho/(\alpha+1)$ and ∂G. We note here that $\psi_1 \in C^1(0,d)$, while $\psi_2 \in C^1((0,d)\backslash\{t'\})$, where $\psi_2(t') = 1$. Then the function $w(t,\varphi)$ in (7.5) will be given by

$$w(t,\varphi) = \frac{F'(t)}{\sqrt{1+K(t)}} \int_{\psi_1(t)}^{\psi_2(t)} (Kv_\varphi)\left(F(t)\rho^{-\alpha} - \alpha\rho/(\alpha+1), \rho, \varphi\right)\rho^{-1-\alpha}d\rho. \tag{7.6}$$

In view of condition (7.2) it follows that for every function v such that $w \in C_0^\infty(S_2)$, we have $I(v) = 0$. Thus $I(v) = 0$ for every $v \in C^*$. Comparison of (4.2) and (7.1) now shows that u is a generalized solution of Problem A. The rest of the proof now follows from Theorem 5.1. □

COROLLARY 7.3. *If $f \in H_\varphi$, then a generalized solution u_f of Problem P_φ exists and coincides with the generalized solution ω_f of Problem* A.

Theorems 6.2 and 7.2 also give information about the smoothness, with respect to φ, of generalized solutions of Problem P_φ.

COROLLARY 7.4. *If $f \in H_\varphi$ and $\partial^k f/\partial\varphi^k \in L_2(G)$ for $k = 0, ..., l$, then the conclusions of Theorem* 6.2 *hold for a generalized solution u_f of Problem P_φ.*

Of course, for $f \in H\backslash H_\varphi$ a generalized solution of Problem P need not coincide with the generalized solution ω_f of Problem A; and if $f \in L_2(G)\backslash H$, there is no

generalized solution of Problem P. Thus the nonlocal Problem A (for which Theorem 6.2 holds, ensuring uniqueness, existence, and differentiability with respect to φ) is, so to speak, a "regular continuation" of the strongly over determined Problem P_φ when $f \in L_2(G)\backslash H_\varphi$. In this sense, we may regard Problem A to be a "nonlocal regularization" of Problems P and P_φ. All this suggests the following procedure for tackling the ill-posed Problem P. For the given function $f \in L_2(G)$ we first try to solve the nonlocal Problem A. To do that, it is possible first to find the solution $\hat{u} = (u_1, u_2, u_3)$ of the local problem (2.5), (2.10) for the corresponding system of partial differential equations and then to find a solution ω_f of Problem A by integration of $u_3(t, \rho, \varphi)$. Then we check the value of the derivative $(\omega_f)_{\varphi\varphi}$ on the characteristic cone S_2 and if that value is very small, we might conclude that the solution u_f of the Problem P exists and is very close to the function ω_f already found.

Remark. Note that Eskin and Vishik solved the strongly overdetermined Cauchy problem for the Poisson equation $\Delta_n u = f$ (see Eskin [12]) by changing the equation to

$$\widehat{\Delta}_n u + G(v) = f,$$

where $G(v)$ is some potential with unknown density v which depends only on $(n-1)$ variables. They established existence and uniqueness results about the pair of functions (u, v); the addition of the potential $G(v)$ removed the overdeterminacy. In our approach we change equation (1.1) to (1.7), but our additional term depends only on the function ω, not on some new function.

We now return to Problem P^*, the homogeneous form of which has an infinite number of classical solutions. By introduction of additional conditions we seek to eliminate this nonuniqueness, and formulate the following nonlocal problem.

Problem $\tilde{\mathrm{P}}^*$. Is there a solution of the equation

$$K(t)\left(v_{x_1x_1} + v_{x_2x_2}\right) - v_{tt} = g \text{ in } G \tag{7.7}$$

which satisfies the boundary condition (1.3) and the additional nonlocal condition

$$\frac{\partial}{\partial\varphi}\int_{\psi_1(t)}^{\psi_2(t)} (Kv)(\tau(t,\rho),\rho,\varphi)\rho^{-1-\alpha}d\rho = 0 \quad (0 < \varphi < 2\pi, 0 < t < d)\,, \tag{7.8}$$

where $\tau(t,\rho) = F(t)\rho^{-\alpha} - \alpha\rho/(\alpha+1)$ (see (7.5) and (7.6) above)? The integration is over the intersection of G and the curve $(t + \frac{\alpha}{\alpha+1}\rho)\rho^\alpha = \text{constant}$, and it is assumed that the parameter α (> 0) satisfied conditions (E1) and (E2).

We must, of course, make precise the notion of a solution with which we shall be dealing. This leads to the following definition.

DEFINITION 7.5. *A function $v \in \tilde{W}_2^1(G)$ is called a generalized solution of Problem $\tilde{\mathrm{P}}^*$ if* (i) *$v = 0$ on $S_0 \cup S_2$;* (ii) *for every $u \in C_L := \{u \in C^1\left(\overline{G}\right) : u = 0$ on $S_0 \cup S_1$ and in some neighborhood of $(0,0,0)\}$, we have*

$$\int_G \left\{u_t v_t - K(t)u_\rho v_\rho - K(t)\rho^{-2}u_\varphi v_\varphi - ug\right\} dx = 0; \tag{7.9}$$

(iii) *for all $u \in C_0^\infty\left(S_2\right)$,*

$$\int_{S_2} u_\varphi(t,\varphi)\left\{\int_{\psi_1(t)}^{\psi_2(t)} (Kv)(\tau(t,\rho),\rho,\varphi)\rho^{-1-\alpha}d\rho\right\} dtd\varphi = 0. \tag{7.10}$$

For some functions, Problem $\tilde{\mathrm{P}}^*$ coincides with the problem adjoint to the nonlocal Problem A. This is the content of the following theorem.

THEOREM 7.6. *A function $v \in \tilde{W}_2^1(G)$ is a generalized solution of Problem $\tilde{\mathrm{P}}^*$ if, and only if, $v = 0$ on $S_0 \cup S_2$ and*

$$(7.11)\int_G \left\{u_t v_t - K(t)u_\rho v_\rho - K(t)\rho^{-2}[u_\varphi - u_\varphi\left(p_\alpha(t,\rho), q_\alpha(t,\rho), \varphi\right)]v_\varphi - ug\right\} dx = 0$$

for all $u \in C_L$.

Proof. Suppose that v satisfies (7.11), and let $u \in C_L$ be zero in some neighborhood of $S_1 \cap S_2$. Let $\psi \in C^\infty(\mathbb{R})$ be such that $\psi(s) = 0$ if $s \le 1$, $\psi(s) = 1$ if $s \ge 2$, and for each $m \in \mathbb{N}$ define

$$\psi_m(t,\rho) = \psi\left(m\left(\rho - \int_0^t \sqrt{K(\tau)}d\tau\right)\right);$$

$\psi_m u$ is zero in some neighborhood of S_2. Use of $\psi_m u$ in (7.11) gives

$$(7.12)\qquad \begin{aligned}(\psi_{mt}v_t - K\psi_{m\rho}v_\rho, u) \quad &+ (\psi_m u_t, v_t) - (\psi_m K u_\rho, v_\rho)\\ &- \left(\psi_m \rho^{-2} K u_\varphi, v_\varphi\right) = (\psi_m u, g)\end{aligned}$$

and by Hardy's inequality we see that the first term converges to zero. From this and (7.12) it follows that (7.9) holds, from which and (7.11) we obtain (7.10). Thus v is a generalized solution of Problem $\tilde{\mathrm{P}}^*$. The converse is immediate. ∎

THEOREM 7.7. *Problem $\tilde{\mathrm{P}}^*$ has at most one generalized solution.*

Proof. By Theorem 7.6, it is enough to prove that there is at most one generalized solution of the problem adjoint to Problem A. To do this, suppose that $v \in \tilde{W}_2^1(G)$, $v = 0$ on $S_0 \cup S_2$, and satisfies (7.11) with $g = 0$. Since $v \in L_2(G)$ and $\sqrt{r}\rho^{-1}v_\varphi \in L_2(G)$, by Theorem 6.2 there is a generalized solution ω of Problem A with $f = v$, such that $\omega \in \tilde{W}_2^1(G)$, $\frac{\partial^2\omega}{\partial t \partial\varphi} \in L_2(G)$, $\sqrt{r}\frac{\partial^2\omega}{\partial\rho\partial\varphi} \in L_2(G)$, and

$$(7.13)\qquad \int_G \left(\omega_t w_t - K\omega_\rho w_\rho - K\rho^{-2}\frac{\partial\omega^1}{\partial\varphi}w_\varphi - vw\right) dx = 0$$

for every $w \in C^*$.

We wish to put $w = v$ in (7.13). To justify this we use the function ψ employed in the proof of Theorem 7.6. Then (7.13) holds with $w = \psi(\rho m)v$, and so

$$\int_G \left(\omega_t v_t - K\omega_\rho v_\rho - K\rho^{-2}\frac{\partial\omega^1}{\partial\varphi}v_\varphi - v^2\right)\psi dx + m\int_G K\psi'(m\rho)\omega_\rho v dx = 0.$$

As $m \to \infty$, the second integral above converges to zero, since $|K(t)| \le Ct \le C_1\rho^{3/2}$ in G, and

$$m\left|\int_G K(t)\psi'(m\rho)\omega_\rho v dx\right| \le C\left\|\sqrt{r}\omega_\rho\right\|_{L_2(G_m)}\|v\|_{L_2(G_m)} \to 0,$$

where $G_m = G \cap \left\{(t,\rho,\varphi) : \rho \le 2m^{-1}\right\}$. Thus (7.11) with $g = 0$ and $u = w$ and (7.13) with $w = v$ show that $v = 0$ in G, and the proof is complete. ∎

We conjecture that for any $g \in \tilde{W}_2^1(G)$, there exists some kind of generalized solution of the Problem $\tilde{\mathrm{P}}^*$, possibly not belonging to $\tilde{W}_2^1(G)$. We feel that this should

follow from the uniqueness theorem relating to the general solution of the Problem A. Here we prove a weaker result.

THEOREM 7.8. *For every function $g \in L_2(G)$ there exists a weak solution of Problem* A**, that is, a function $v \in L_2(G)$ for which the equality*

$$\left(L\omega - K(t)\rho^{-2}\omega_{\varphi\varphi}\left(p_\alpha(t,\rho), q_\alpha(t,\rho), \varphi\right), v\right) = (\omega, g) \tag{7.14}$$

holds for every $\omega \in C_L \cap C^2\left(\overline{G}\right)$.

THEOREM 7.9. *This follows from the a priori estimate* (5.1)*, that is,*

$$\|\omega\|_{L_2(G)} \le C \left\|L\omega - K\rho^{-2}\omega_{\varphi\varphi}\left(p_\alpha, q_\alpha, \varphi\right)\right\|_{L_2(G)}.$$

Remark. Some other additional conditions relating to Problem P* for the wave equation (instead of equation (1.1)) were formulated by Kan Cher [9]. These conditions concern the boundedness of some integrals of Fourier-coefficients of solutions of Problem P*.

Acknowledgments. The essential part of this paper was finished while the second author was visiting the University of Sussex. He would like to thank the Royal Society for support and the University of Sussex for hospitality.

REFERENCES

[1] S. A. ALDASHEV, *Some boundary-value problems for linear multidimensional second-order hyperbolic equations*, Ukrainian Math. J., 43 (1991), pp. 379–384.
[2] S. A. ALDASHEV, *Multidimensional Darboux problems for the degenerate hyperbolic equation*, Differential Equations, 29 (1993), pp. 1829–1834.
[3] S. A. ALDASHEV, *On the well-posedness of multidimensional Darboux problems for the wave equation*, Ukrainian Math. J., 45 (1993), pp. 1456–1464.
[4] K. A. AMES, *Improperly posed problems for nonlinear partial differential equations*, in Nonlinear Equations in the Applied Sciences, W. F. Ames and C. Rogers, eds., Mathematics in Sciences and Engineering 185, Academic Press, New York, 1992, pp. 1–28.
[5] K. A. AMES, H. A. LEVINE, AND L. E. PAYNE, *Improved continuous dependence results for a class of evolutionary equations*, in Inverse and Ill-Posed Problems, H. W. Engl and C. W. Groetsch, eds., Mathematics in Sciences and Engineering 4, Academic Press, New York, 1987, pp. 433–450.
[6] A. K. AZIZ AND M. SCHNEIDER, *Frankl–Morawetz problems in R^3*, SIAM J. Math. Anal., 10 (1979), pp. 913–921.
[7] A. V. BITSADZE, *On some problems about mixed type equation in multidimensional regions*, Dokl. Akad. Nauk UzSSR, 110 (1956), pp. 1021–1024.
[8] K. KAN CHER, *Nonuniqueness of solutions of the Darboux problem*, Siberian Math. J., 26 (1985), pp. 286–288.
[9] K. KAN CHER, *An estimate of the solution of Darboux-Protter problems for the two-dimensional wave equation*, Soviet Math. Dokl., 43 (1991), pp. 887–891.
[10] V. P. DIDENKO, *On boundary-value problems for multidimensional hyperbolic equation with degeneration*, Soviet Math. Dokl., 13 (1972), pp. 998–1002.
[11] V. P. DIDENKO, *Boundary-value problems for three-dimensional hyperbolic equation of mixed type*, Differential Equations, 11 (1975), pp. 25–28.
[12] G. I. ESKIN, *Boundary-value problems for elliptic pseudodifferential equations*, Translation of Math. Monographs 52, AMS, Providence, RI, 1981.
[13] K. O. FRIEDRICHS, *The identity of weak and strong extensions of differential operators*, Trans. Amer. Math. Soc., 55 (1944), pp. 132–151.
[14] K. O. FRIEDRICHS, *Symmetric positive linear differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.
[15] P. R. GARABEDIAN, *Partial differential equations with more than two variables in the complex domain*, J. Math. Mech., 9 (1960), pp. 241–271.
[16] T. KWANG-CHANG, *On a boundary value problem for the wave equation*, Sci. Record, New Series, 1 (1957), pp. 1–3.

[17] P. D. Lax and R. S. Phillips, *Local boundary conditions for dissipative symmetric linear differential operators*, Comm. Pure Appl. Math., 13 (1960), pp. 427–455.
[18] R. Lattès and J. L. Lions, *The Method of Quasireversibility*, American Elsevier, New York, 1969.
[19] M. M. Lavrentiev, V. G. Romanov, and S. P. Shishatskii, *Ill-Posed Problems of Mathematical Physics and Analysis*, Translations of Mathematical Monographs 64, AMS, Providence, RI, 1986.
[20] C. S. Morawetz, *A weak solution for a system of equations of elliptic-hyperbolic type*, Comm. Pure Appl. Math., 11 (1958), pp. 315–331.
[21] J. S. Papadakis, *A boundary-value problem for equations of mixed type*, Bull. Soc. Math. Grèce, 14 (1973), pp. 157–171.
[22] L. E. Payne, *Improperly Posed Problems in Partial Differential Equations*, Regional Conference Series in Applied Mathematics 22, SIAM, Philadelphia, PA, 1975.
[23] G. Peyser, *On the identity of weak and strong solutions of differential equations with local boundary conditions*, Amer. J. Math., 87 (1965), pp. 267–277.
[24] N. I. Popivanov, *Overdetermined problems for the wave equation and their nonlocal regularization*, Differential Equations, 24 (1988), pp. 1301–1312.
[25] N. I. Popivanov, *Nonlocal regularization of some overdetermined boundary value problems*, in Proc. 1987 EQUADIFF Conference, C. Dafermos, G. Ladas, and G. Papanikolau, eds., Lecture Notes Pure Appl. Math. 118, 1989, pp. 581–586.
[26] N. I. Popivanov and M. Schneider, *The Darboux problem in R^3 for a class of degenerating hyperbolic equations*, J. Math. Anal. Appl., 175 (1993), pp. 537–579.
[27] N. I. Popivanov and M. Schneider, *On M. H. Protter problems for the wave equation in R^3*, J. Math. Anal. Appl., 194 (1995), pp. 50–77.
[28] M. H. Protter, *A boundary value problem for the wave equation and mean value problems*, Ann. Math. Studies, 33 (1954), pp. 247–257.
[29] M. H. Protter, *New boundary value problem for the wave equation and equations of mixed type*, J. Rat. Mech. Anal, 3 (1954), pp. 435–446.
[30] J. Rauch, *Symmetric positive systems with boundary characteristics of constant multiplicity*, Trans. Amer. Math. Soc., 291 (1985), pp. 167–187.
[31] H. Salzman and M. Schneider, *Schwache Lösungen des Frankl-Morawetz Problems in R^3*, Mh. Math., 84 (1977), pp. 237–246.
[32] N. G. Sorokina, *Multidimensional generation of a theorem of C. Morawetz for weak solvability of differential equations of mixed type*, Differential Equations, 13 (1977), pp. 128–134.
[33] N. G. Sorokina, *Multidimensional analogues of the Tricomi problem*, in Proc. Conference PDE, Moscow State University, 1978, pp. 452–454 (in Russian).
[34] A. N. Tikhonov and V. Y. Arsenin, *Solution on Ill-Posed Problems*, Wiley Press, New York, 1977.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 106–115, January 1998

007

ON A CHARACTERIZATION OF THE KERNEL OF THE DIRICHLET-TO-NEUMANN MAP FOR A PLANAR REGION*

DAVID INGERMAN† AND JAMES A. MORROW†

Abstract. We will show that the Dirichlet-to-Neumann map Λ for the electrical conductivity equation on a simply connected plane region has an *alternating property*, which may be considered as a generalized maximum principle. Using this property, we will prove that the kernel, K, of Λ satisfies a set of inequalities of the form $(-1)^{\frac{n(n+1)}{2}} \det K(x_i, y_j) > 0$. We will show that these inequalities imply Hopf's lemma for the conductivity equation. We will also show that these inequalities imply the alternating property of a kernel.

Key words. conductivity, Dirichlet-to-Neumann, kernel

AMS subject classifications. 35R30, 94C15

PII. S0036141096300483

1. Introduction. In this paper we will derive some properties of the Dirichlet-to-Neumann map for the electrical conductivity equation in $\mathbf{R}^2$. These properties are analogs of properties which characterize the Dirichlet-to-Neumann maps for electrical networks (see [1], [2], and [3]). We recall some definitions. Let Ω be a relatively compact, simply connected open set in $\mathbf{R}^2$ with C^2 boundary. Let $\gamma(p) > 0$ be a C^2 function on $\overline{\Omega}$. Let f be a function defined on $\partial\Omega$. Then there is a unique function u, defined on $\overline{\Omega}$, such that

$$\nabla(\gamma\nabla u) = 0 \tag{1.1}$$

and $u(p) = f(p)$ for $p \in \partial\Omega$. (Equation (1.1) is the electrical conductivity equation and a function, u, that satisfies (1.1) is called a γ-*harmonic function.*) Let $\frac{\partial u}{\partial n}(p)$ be the directional derivative of u in the direction of the outward pointing unit normal n at the point $p \in \partial\Omega$. Then the Dirichlet-to-Neumann map, Λ, is defined by the formula

$$\Lambda f(p) = \gamma(p)\frac{\partial u}{\partial n}(p). \tag{1.2}$$

The domain of Λ may be taken to be $H^{\frac{1}{2}}(\partial\Omega)$ and the image is in $H^{-\frac{1}{2}}(\partial\Omega)$. Λ is a pseudodifferential operator of order 1 and as such has a kernel, $K(x, y)$, defined as a distribution on $\partial\Omega \times \partial\Omega$. The kernel gives a representation of Λ by the formula

$$\Lambda f(x) = \int_{\partial\Omega} K(x, y)f(y)dy, \tag{1.3}$$

where x and y are arc length coordinates on $\partial\Omega$. For the pseudodifferential operator Λ, the kernel K is a symmetric function, $K(x, y) = K(y, x)$, and for a fixed $x \in \partial\Omega$, $\lim_{y\to x} |K(x, y)| = \infty$. More precisely,

$$K(x, y) = \frac{k(x, y)}{|x - y|^2} + D(x, y), \tag{1.4}$$

*Received by the editors March 11, 1996; accepted for publication (in revised form) September 30, 1996.

http://www.siam.org/journals/sima/29-1/30048.html

†Department of Mathematics, University of Washington, Box 354350, Seattle, WA 98195-4350 (ingerman@math.washington.edu, morrow@math.washington.edu).

where k is continuous on $\partial\Omega \times \partial\Omega$, $k(x,y) = k(y,x)$, $k(x,x) \neq 0$, and D is a distribution supported on $\Delta = \{(x,x) : x \in \partial\Omega\}$. (In this formula, $|x-y|$ is the separation in arc length of points with arc length coordinates x and y and the continuous term in this expansion has been incorporated into the term $\frac{k(x,y)}{|x-y|^2}$.) If $x \notin \text{supp}(f)$, then the integral is an ordinary integral and there are no convergence questions. Since we will be interested in the behavior of $K(x,y)$ for $x \neq y$ we will ignore D and will pretend that $K(x,y) = \frac{k(x,y)}{|x-y|^2}$. The expansion (1.4) follows from Lemma 3.7 of [6] or Theorem 0.1 in [7]. The boundary, $\partial\Omega$, is a Jordan curve and hence is homeomorphic to a circle. Pick an orientation on $\partial\Omega$. We say that $(x_1, \ldots, x_n; y_1, \ldots, y_n)$ is a *circular pair* if there are points $p, q \in \partial\Omega$ which divide $\partial\Omega$ into two connected components, A, B such that $\{x_1, \ldots, x_n\} \subset A$, $\{y_1, \ldots, y_n\} \subset B$, and $x_1, \ldots, x_n, y_1, \ldots, y_n$ are in circular order on $\partial\Omega$. (Note that this definition is modified from the definition in [2].) The main theorem of this paper is the following theorem, which we prove to be equivalent to the alternating property stated in section 2.

THEOREM 1.1. *Let $(x_1, \ldots, x_n; y_1, \ldots, y_n)$ be a circular pair on $\partial\Omega$. Let $L = (l_{ij})$ be the $n \times n$ matrix with entries defined by $l_{ij} = K(x_i, y_j)$. Then*

$$(-1)^{\frac{n(n+1)}{2}} \det(L) > 0. \tag{1.5}$$

We consider this to be a generalization of a result in [2]. We will see how it implies the classical Hopf lemma for the conductivity equation in dimension 2.

2. The alternating property. We first restate and prove a result of [2]. Suppose that $\partial\Omega = I \cup J$, where I and J are disjoint connected arcs. Then we have the following theorem.

THEOREM 2.1. *Let f be a smooth function on $\partial\Omega$ such that $f = 0$ on I. Suppose there is a sequence of points $\{p_1, \ldots, p_n\} \subset I$ in circular order such that*

$$(-1)^{i+1} \Lambda f(p_i) > 0. \tag{2.1}$$

Then there is a sequence of points $\{q_1, \ldots, q_n\} \subset J$ in circular order such that

$$(-1)^n \Lambda f(p_i) f(q_i) > 0. \tag{2.2}$$

Proof. Equation (2.2) is equivalent to

$$\Lambda f(p_i) f(q_{n+1-i}) < 0. \tag{2.3}$$

We first describe how to pick the point q_n. Let u be the solution of (1.1) such that $u = f$ on $\partial\Omega$. By (2.1) $\frac{\partial u}{\partial n}(p_1) > 0$. Hence there is a small open line segment, α, such that $\alpha \subset \Omega$, p_1 is one end of α, and $u < 0$ on α. Let W be the connected component of $\{z \in \Omega \ : \ u(z) < 0\}$ that contains α. Suppose that $\overline{W} \cap J = \emptyset$. Then $u = 0$ on ∂W. But this contradicts the maximum principle since $u < 0$ in W and $W \neq \emptyset$. Thus $\overline{W} \cap J \neq \emptyset$. Now $u = 0$ at every point of ∂W that is in Ω. Using the maximum principle again we see that there is a $q_n \in \overline{W} \cap J$ such that $f(q_n) < 0$ and there is an open line segment $\beta \subset W$ such that q_n is an end point of β. Now we can connect the ends of α and β that are inside W by a smooth curve in W. Hence there is a smooth curve C_1 such that C_1 is diffeomorphic to a line segment, has end points p_1 and q_n, and $C_1 - p_1 - q_n \subset W$. Then $u(z) < 0$ for all $z \in C_1 - p_1$. We can repeat this argument to produce curves C_j such that C_j joins p_j to a point $q_{n+1-j} \in J$, $C_j - p_j - q_{n+1-j} \subset \Omega$, and $(-1)^j u(z) < 0$ for all $z \in C_j - p_j$. These curves cannot

intersect and by the Jordan curve theorem the points $p_1, \ldots, p_n, q_1, \ldots, q_n$ must be in circular order on $\partial\Omega$. It is easy to see that these points satisfy (2.3). □

We have referred to this property as the alternating property. Elsewhere [5] a similar property has been called the variation diminishing property. See also section 6 of this paper.

3. The weak inequality. We first prove the weaker statement.

THEOREM 3.1. *Let $(x_1, \ldots, x_n; y_1, \ldots, y_n)$ be a circular pair on $\partial\Omega$. Let $L = (l_{ij})$ be the $n \times n$ matrix with entries defined by $l_{ij} = K(x_i, y_j)$. Then*

$$(-1)^{\frac{n(n+1)}{2}} \det(L) \geq 0. \tag{3.1}$$

Proof. The proof is by induction on n. We first consider $n = 1$. The proof goes by contradiction. Suppose that there are points $p, q \in \partial\Omega$ with $p \neq q$ and $K(p, q) > 0$. Then there is an $\epsilon > 0$ such that $p \notin D_\epsilon = \{y : |y-q| < \epsilon\}$ and $K(p, y) > 0$ for $y \in D_\epsilon$. Let $f(y)$ be a continuous function on $\partial\Omega$ such that $\text{supp}(f) \subset D_\epsilon = \{y : |y - q| < \epsilon\}$, $f(q) > 0$, and $f(s) \geq 0$ for all $s \in \partial\Omega$. Then

$$\gamma(p)\frac{\partial u}{\partial n}(p) = \Lambda f(p) = \int_{D_\epsilon} K(p, y) f(y) dy > 0,$$

where u satisfies (1.1) and $u(s) = f(s),\ s \in \partial\Omega$. But then there must be a point z near p in Ω such that $u(z) < 0$. This contradicts the maximum principle.

Next we assume that the result is true for all $(n-1) \times (n-1)$ matrices and prove that it is true for $n \times n$ matrices. If the result is not true, then we have a circular pair $(x_1, \ldots, x_n; y_1, \ldots, y_n)$ such that

$$(-1)^{\frac{n(n+1)}{2}} \det(L) < 0. \tag{3.2}$$

Consider the matrix L^{-1} with entries (h_{ij}). Then

$$h_{ij} = (-1)^{i+j} \frac{\det(L_{ij})}{\det(L)}, \tag{3.3}$$

where L_{ij} is the (i, j) minor of L. By induction, (3.2), and (3.3),

$$(-1)^{i+j+\frac{n(n-1)}{2}+\frac{n(n+1)}{2}+1} h_{ij} = (-1)^{i+j+n+1} h_{ij} \geq 0. \tag{3.4}$$

Since L is nonsingular, for fixed i there must be some j for which

$$(-1)^{i+j+n+1} h_{ij} > 0. \tag{3.5}$$

Now let $w = [1, -1, 1, \ldots, (-1)^{n+1}]^T$ be an n-vector with alternating signs. Let $z = L^{-1}w$. Then using (3.4) and (3.5) it is easy to verify that

$$(-1)^{i+n} z_i > 0. \tag{3.6}$$

To summarize, we have a vector z such that

$$(-1)^{i+1} = w_i = \sum_{j=1}^{n} K(x_i, y_j) z_j \tag{3.7}$$

and

$$(-1)^{n+1} z_i w_i > 0. \tag{3.8}$$

Now, choose small intervals D_j around the points y_j such that the D_j are disjoint and do not contain any of the points x_i. Choose the D_j so small that

$$|K(x_i, y) - K(x_i, y_j)| < \epsilon,\ y \in D_j,\ i = 1, \ldots, n. \tag{3.9}$$

Also choose functions f_j such that

$$\operatorname{supp}(f_j) \subset D_j,\ z_j f_j(y) \geq 0, \text{and} \int_{D_j} f_j = z_j. \tag{3.10}$$

Let $f = \sum f_j$. Then

$$\begin{aligned} |\Lambda f(x_i) - w_i| &= \left| \int_{\partial\Omega} K(x_i, y) f(y) dy - \sum_{j=1}^{n} K(x_i, y_j) z_j \right| \\ &= \left| \int_{\partial\Omega} (K(x_i, y) - K(x_i, y_j)) f(y) dy \right| \\ &\leq \epsilon \sum_{j=1}^{n} |z_i|. \end{aligned} \tag{3.11}$$

Thus we conclude that for ϵ small enough $\Lambda f(x_i)$ has the same sign as w_i. By the alternating property, there would have to be a set of n points t_i in circular order such that

$$(-1)^n w_i f(t_i) > 0. \tag{3.12}$$

For such a set of points we would have to have $t_i \in D_i$ and hence $f(t_i)$ would have the same signs as z_i. This contradicts (3.8). □

4. The strong inequality. We now prove Theorem (1.1). We consider the cases $n = 1$ and $n > 1$ separately. Let us assume the arc length of $\partial\Omega$ is S and that points on $\partial\Omega$ are parametrized by the numbers in the interval $[0, S)$. When $n = 1$, suppose there is a pair of points x_1, y_1 with $0 \leq x_1 < y_1$ and $K(x_1, y_1) = 0$. By (1.4) there is no sequence of points z_j such that $x_1 < z_j < y_1$, $\lim_{j\to\infty} z_j = x_1$, and $\lim_{j\to\infty} K(x_1, z_j) = 0$. Hence there is a point η_2 with $x_1 < \eta_2 < y_1$ such that

$$K(x_1, \eta_2) = 0 \text{ and } K(x_1, \eta) < 0 \text{ for } x_1 < \eta < \eta_2.$$

Let x be any number such that $x_1 < x < \eta_2$ and choose η_1 so that $x < \eta_1 < \eta_2$. Then $(x_1, x; \eta_1, \eta_2)$ is a circular pair and hence

$$\begin{vmatrix} K(x_1, \eta_1) & K(x_1, \eta_2) \\ K(x, \eta_1) & K(x, \eta_2) \end{vmatrix} \leq 0. \tag{4.1}$$

Since

$$K(x_1, \eta_2) = 0,\ K(x, \eta_2) \leq 0,\ \text{and } K(x_1, \eta_1) < 0,$$

it follows that

$$K(x, \eta_2) = 0. \tag{4.2}$$

This shows that for *all* x, with $x_1 < x < \eta_2$, $K(x, \eta_2) = 0$. Hence we get the contradiction that $\lim_{x \to \eta_2} K(x, \eta_2) = 0$.

The proof for $n > 1$ makes use of the following result in [2]. It was later pointed out to us that Charles Dodgson (Lewis Carroll) used a version of this identity in [4]. Let $(x_1, \dots, x_n; y_1, \dots, y_n)$ be a circular pair. We assume that the coordinates on $\partial\Omega$ are chosen so that $0 \le x_1 < \cdots < x_n < y_1 < \cdots < y_n < S$. Let L be the matrix with i, j entry equal to $K(x_i, y_j)$. We will use the notation

$$\kappa(x_1, \dots, x_n; y_1, \dots, y_n) = \det(L). \tag{4.3}$$

LEMMA 4.1. *Let* $(a_1, \dots, a_{n+1}; b_1, \dots, b_{n+1})$ *be a circular pair. Then*

$$\begin{aligned} &\kappa(a_1, \dots, a_{n+1}; b_1, \dots, b_{n+1})\kappa(a_1, \dots, a_{n-1}; b_3, \dots, b_{n+1}) \\ &= \kappa(a_1, \dots, a_n; b_1, b_3, \dots, b_{n+1})\kappa(a_1, \dots, a_{n-1}, a_{n+1}; b_2, \dots, b_{n+1}) \\ &-\kappa(a_1, \dots, a_n; b_2, \dots, b_{n+1})\kappa(a_1, \dots, a_{n-1}, a_{n+1}; b_1, b_3, \dots, b_{n+1}). \end{aligned} \tag{4.4}$$

Assume that

$$\kappa(x_1, \dots, x_n; y_1, \dots, y_n) = 0 \tag{4.5}$$

for some circular pair. First we claim that there is no sequence of points z_j such that $x_n < z_j < y_1$, $\lim_{j \to \infty} z_j = x_n$, and $\lim_{j \to \infty} \kappa(x_1, \dots, x_n; z_j, y_2, \dots, y_n) = 0$. For this would imply that there are constants c_k (independent of j) so that

$$K(x_n, z_j) = \sum_{k<n} c_k K(x_k, z_j), \tag{4.6}$$

and hence

$$\lim_{j \to \infty} K(x_n, z_j) = \sum_{k<n} c_k K(x_k, x_n), \tag{4.7}$$

contradicting (1.4). Thus there is a number η_1 with $x_n < \eta_1 < y_1$ such that

$$\kappa(x_1, \dots, x_n; \eta_1, y_2, \dots, y_n) = 0 \text{ and} \tag{4.8}$$

$$\kappa(x_1, \dots, x_n; \eta, y_2, \dots, y_n) \neq 0 \text{ for } x_n < \eta < \eta_1. \tag{4.9}$$

Let x be such that $x_n < x < \eta_1$. Then there is an η such that $x < \eta < \eta_1$ and hence $(x_1, \dots, x_n, x; \eta, \eta_1, y_2, \dots, y_n)$ is a circular pair. By (4.4), (4.5), and (3.1),

$$\begin{aligned} 0 &\ge \kappa(x_1, \dots, x_n, x; \eta, \eta_1, y_2, \dots, y_n)\kappa(x_1, \dots, x_{n-1}; y_2, \dots, y_n) \\ &= \kappa(x_1, \dots, x_n; \eta, y_2, \dots, y_n)\kappa(x_1, \dots, x_{n-1}, x; \eta_1, y_2, \dots, y_n) \\ &\quad -\kappa(x_1, \dots, x_n; \eta_1, y_2, \dots, y_n)\kappa(x_1, \dots, x_{n-1}, x; \eta, y_2, \dots, y_n) \\ &= \kappa(x_1, \dots, x_n; \eta, y_2, \dots, y_n)\kappa(x_1, \dots, x_{n-1}, x; \eta_1, y_2, \dots, y_n) \ge 0. \end{aligned} \tag{4.10}$$

Using this and (4.9) we see that

$$\kappa(x_1, \dots, x_{n-1}, x; \eta_1, y_2, \dots, y_n) = 0 \text{ for } x_n < x < \eta_1. \tag{4.11}$$

As above, this contradicts (1.4) and proves the theorem.

5. The Hopf lemma. We now show how the fact that $K(x,y) < 0$ for $x \neq y$ implies the Hopf lemma (reference) for the conductivity equation.

THEOREM 5.1. *Let u be a nonconstant solution of $\nabla(\gamma\nabla u) = 0$, and let $p \in \partial\Omega$ be a point where u assumes a minimum. Then*

$$\frac{\partial u}{\partial n}(p) < 0. \tag{5.1}$$

Proof. We may assume that $u(p) = 0$. Let $f = u|_{\partial\Omega}$. Since u is not constant, $\mathrm{supp}(f)$ is not empty. Thus there is an interval D around p in $\partial\Omega$ such that $\mathrm{supp}(f) - D$ is not empty. Let ψ be a smooth function on $\partial\Omega$ such that $\psi = 1$ on $\mathrm{supp}(f) - D$, $\psi = 0$ on an interval around p, and $0 \leq \psi \leq 1$. Let $g = \psi f$ and let v be the solution of $\nabla(\gamma\nabla v) = 0$ with $v|_{\partial\Omega} = g$. Since $f \geq g$ it follows that $u \geq v$. It is also true that $g \geq 0$. Since $p \notin \mathrm{supp}(g)$ and $K(p,y) < 0$,

$$0 > \int_{\partial\Omega} K(p,y)g(y)dy = \gamma(p)\frac{\partial v}{\partial n}(p) \geq \gamma(p)\frac{\partial u}{\partial n}(p), \tag{5.2}$$

which proves the theorem. □

6. The variation diminishing property. We will use the following notation. Let $M(x,y)$ be a continuous function on $[c,d] \times [a,b]$. Let $c \leq x_1 < x_2 < \cdots < x_n \leq d$, $a \leq y_1 < y_2 < \cdots < y_n \leq b$. Let T be the $n \times n$ matrix with i,j entry equal to $M(x_i, y_j)$. Let

$$\mu(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n) = \det(T).$$

The following lemma from [5] is sometimes paraphrased by saying that the kernel M has the *variation diminishing property*. It will be used to show that the inequalities (1.5) imply the alternating property.

LEMMA 6.1. *Let f be a continuous, not identically 0, function defined on the interval $[a,b]$ such that f changes its sign on this interval no more than $n-1$ times. Let $M(x,y)$, $x,y \in [c,d] \times [a,b]$, be a continuous kernel with the property that*

$$\mu(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n) > 0, \tag{6.1}$$

whenever $c \leq x_1 < x_2 < \cdots < x_n \leq d$, $a \leq y_1 < y_2 < \cdots < y_n \leq b$. Then the function

$$g(x) = \int_a^b M(x,y)f(y)dy$$

vanishes in $[c,d]$ no more than $n-1$ times.

By saying that function f changes its sign k times on the interval $[a,b]$ we mean that there are $k+1$ points $x_1 < x_2 < \cdots < x_{k+1}$ in $[a,b]$ such that for $i = 1, 2, \ldots, k$

$$f(x_i)f(x_{i+1}) < 0. \tag{6.2}$$

Proof. By hypothesis there are points $a = s_0 < s_1 < s_2 < \cdots < s_{n-1} < s_n = b$ such that in each interval (s_{i-1}, s_i), $i = 1, 2, \ldots, n$ function f does not change its sign and is not identically 0. For $i = 1, 2, \ldots, n$ let

$$g_i(x) = \int_{s_{i-1}}^{s_i} M(x,y)f(y)dy. \tag{6.3}$$

Then

$$g(x) = \sum_{i=1}^{n} g_i(x). \tag{6.4}$$

For any $c \le x_1 < x_2 < \cdots < x_n \le d$ the determinant

$$\det(\{g_i(x_j\}) = \int_{s_{n-1}}^{s_n} \cdots \int_{s_0}^{s_1} \mu(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n) f(y_1) \cdots f(y_n) dy_1 \cdots dy_n \tag{6.5}$$

is not 0 since the integrand is not identically zero and has constant sign. This shows that there is no nontrivial linear combination of g_i's vanishing at n points and hence that $g(x) = \sum_{i=1}^{n} g_i(x)$ cannot vanish at n points. □

We note that this proof only used the fact that $\mu(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n)$ has constant sign. We need one more lemma before coming to the proof of the alternating principal.

Let $K(x, y)$ be a kernel on $\partial\Omega \times \partial\Omega$. We assume that $K(x, y)$ is continuous when $x \neq y$, but we don't assume anything about K on the diagonal of $\partial\Omega \times \partial\Omega$. Let $\kappa(x_1, \ldots, x_n; y_1, \ldots, y_n)$ be defined as in section 4.

LEMMA 6.2. *Suppose that $\kappa(x_1, \ldots, x_n; y_1, \ldots, y_n)$ is never zero and has constant sign for all circular n-pairs $(x_1, \ldots, x_n; y_1, \ldots, y_n)$. Let $\partial\Omega = I \cup J$ where I and J are disjoint connected arcs. Let f be a continuous function on $\partial\Omega$ with* $\operatorname{supp}(f) \subset J$. *Let*

$$g(x) = \int_{\partial\Omega} K(x, y) f(y) dy. \tag{6.6}$$

Then if there is a sequence of $n+1$ points in I in circular order at which g alternates in sign, then there is a sequence of at least $n+1$ points in J in circular order at which f alternates in sign.

Proof. If there is no sequence of $n + 1$ points of J at which f alternates in sign, then f can change its sign no more than $n-1$ times in J. By Lemma 6.1, g can vanish no more than $n - 1$ times in I. But we are assuming that g has $n + 1$ alternations of sign in I and hence at least n zeros in I. This contradiction proves the lemma. □

We now state and prove the theorem.

THEOREM 6.3. *Using the notation of Lemma* 6.2, *suppose that*

$$(-1)^{\frac{n(n+1)}{2}} \kappa(x_1, \ldots, x_n; y_1, \ldots, y_n) > 0 \tag{6.7}$$

for all $n > 0$ and all circular n-pairs $(x_1, \ldots, x_n; y_1, \ldots, y_n)$. Let f be a continuous function on $\partial\Omega$ with $\operatorname{supp}(f) \subset J$. *Let*

$$g(x) = \int_{\partial\Omega} K(x, y) f(y) dy. \tag{6.8}$$

Suppose there is a sequence of points $\{p_1, \ldots, p_n\} \subset I$ in circular order such that

$$(-1)^{i+1} g(p_i) > 0. \tag{6.9}$$

Then there is a sequence of points $\{q_1, \ldots, q_n\} \subset J$ in circular order such that

$$(-1)^n g(p_i) f(q_i) > 0. \tag{6.10}$$

Proof. By Lemma 6.2 there is a sequence of points in J at which f alternates in sign. If there is no sequence with the desired alteration property then J is a disjoint union of subintervals J_i, in circular order, such that

1. f is not identically 0 on $J_i,\ i = 1, \ldots, n$,
2. f does not change its sign on $J_i,\ i = 1, \ldots, n$,
3. for some $z_i \in J_i$,

$$(-1)^{n+i} f(z_i) > 0. \tag{6.11}$$

We use the idea of Lemma 6.1. For $i = 1, 2, \ldots, n$ let

$$g_i(x) = \int_{J_i} K(x, y) f(y) dy. \tag{6.12}$$

Then

$$g(x) = \sum_{i=1}^{n} g_i(x). \tag{6.13}$$

Let

$$G = \begin{bmatrix} g_1(x_1) & g_2(x_1) & \ldots & g_n(x_1) \\ g_1(x_2) & \ldots & & g_n(x_2) \\ \vdots & & & \vdots \\ g_1(x_n) & \ldots & & g_n(x_n) \end{bmatrix}. \tag{6.14}$$

Let u be the n-vector with $u_i = 1,\ i = 1, \ldots, n$. Then

$$Gu = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_n) \end{bmatrix}. \tag{6.15}$$

Using (6.11) we will show that the signs of u are all negative. This contradiction will prove the theorem. We need to compute the signs of the entries of G^{-1}. Rather than get lost in a cloud of indices, we will give the proof in the case that $n = 3$ and leave the general proof to the reader. In this case the assumption (6.11) implies that $f(y) \geq 0$ in J_1, $f(y) \leq 0$ in J_2, and $f(y) \geq 0$ in J_3. As in section 3 we compute the signs of the cofactors of G. First we have

$$\det(G) = \int_{J_1} \int_{J_2} \int_{J_3} \kappa(x_1, x_2, x_3; y_1, y_2, y_3) f(y_1) f(y_2) f(y_3) dy_1 dy_2 dy_3 < 0. \tag{6.16}$$

We find that

$$\begin{vmatrix} g_2(x_2) & g_3(x_2) \\ g_2(x_3) & g_3(x_3) \end{vmatrix} = \int_{J_2} \int_{J_3} \kappa(x_2, x_3; y_2, y_3) f(y_2) f(y_3) dy_2 dy_3 > 0. \tag{6.17}$$

Hence $(G^{-1})_{11} < 0$. Next we compute that

$$(-1)^{1+2} \begin{vmatrix} g_1(x_2) & g_3(x_2) \\ g_1(x_3) & g_3(x_3) \end{vmatrix} = \int_{J_1} \int_{J_3} \kappa(x_2, x_3; y_1, y_3) f(y_1) f(y_3) dy_1 dy_3 > 0, \tag{6.18}$$

and thus $(G^{-1})_{21} < 0$. Continuing the calculation we find that the signs of G^{-1} are as follows:

$$G^{-1} = \begin{bmatrix} - & + & - \\ - & + & - \\ - & + & - \end{bmatrix}. \tag{6.19}$$

This yields the contradiction

$$\begin{bmatrix} - & + & - \\ - & + & - \\ - & + & - \end{bmatrix} \begin{bmatrix} + \\ - \\ + \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{6.20}$$

□

7. Remarks and conjectures. We have not tried to state the most general hypotheses under which our results are valid, but have stated them in such a way that the essential ideas of the proofs are clear. We can also prove determinant inequalities for certain "blocks" in Dirichlet-to-Neumann kernels for multiply-connected plane domains. We can differentiate our inequalities to get a set of inequalities involving determinants of derivatives of the Dirichlet-to-Neumann kernel. These inequalities are equivalent to the set of inequalities (1.5). In our arguments we seem to need to assume that γ is in $C^2(\Omega)$; however, it is possible that weaker assumptions would suffice.

We would like to single out the following conjecture on characterizing the kernel of a Dirichlet-to-Neumann map.

CONJECTURE 1. *Let Ω be a relatively compact, simply connected region in the plane with C^2 boundary. Let $K(x,y) = \frac{k(x,y)}{|x-y|^2}$, where $(x,y) \in \partial\Omega \times \partial\Omega - \Delta$, k is continuous on $\partial\Omega \times \partial\Omega$, $k(x,x) \neq 0$, and K satisfies* (1.5). *Then there is a distribution $D(x,y)$ on $\partial\Omega \times \partial\Omega$, supported on the diagonal, Δ, and a regularization of K as a distribution on $\partial\Omega \times \partial\Omega$, so that $L = K + D$ is the kernel of the Dirichlet-to-Neumann map for some conductivity, γ, on Ω. The distribution D is determined by the property that*

$$\int_{\partial\Omega} L(x,y)dy = 0. \tag{7.1}$$

Equation (7.1) is analogous to the fact that the Dirichlet-to-Neumann matrix for an electrical network has row sums equal to zero. This implies that the diagonal is determined by the off-diagonal terms. This is true as well in the continuous case.

Acknowledgments. We have discussed these results with many people and their suggestions and advice have been of great benefit. Among these people are Ed Curtis, John Sylvester, and Gunther Uhlmann. The importance of the alternating property was recognized some time ago by Ed Curtis.

REFERENCES

[1] Y. COLIN DE VERDIERE, I. GITLER, AND D. VERTIGAN, *Réseaux électriques planaire.* II, Comment. Math. Helv., 71 (1996), pp. 144–167.

[2] E. B. CURTIS, D. INGERMAN, AND J. MORROW, *Circular planar graphs and resistor networks*, Discrete Math., submitted.

[3] E. B. CURTIS, E. MOOERS, AND J. MORROW, *Finding the conductors in circular networks from boundary measurements*, Math. Modelling and Numer. Math. Modelling and Numer. Anal., 28 (1994), pp. 781–813.

[4] C. L. Dodgson, *Condensation of Determinants*, Proc. Roy. Soc. London, 15 (1866), pp. 150–155.

[5] F. R. Gantmakher and M. G. Krein, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, State Publishing House for Technical-Theoretical Literature, Moscow, 1950.

[6] J. Sylvester, *An anisotropic inverse boundary value problem*, Comm. Pure Appl. Math., 43 (1990), pp. 201–232.

[7] J. Sylvester and G. Uhlmann, *Inverse boundary value problems at the boundary – continuous dependence*, Comm. Pure Appl. Math., 41 (1988), pp. 197–219.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 116–133, January 1998

008

EXPONENTIALLY GROWING SOLUTIONS FOR NONSMOOTH FIRST-ORDER PERTURBATIONS OF THE LAPLACIAN*

CARLOS F. TOLMASKY†

Abstract. We construct exponentially growing solutions for first-order perturbations of the Laplacian which are not smooth. We apply this kind of solution to prove global uniqueness for an inverse boundary value problem for the Schrödinger equation in the presence of a magnetic field.

Key words. inverse boundary problems, Dirichlet–Neumann map, exponentially growing solutions

AMS subject classifications. 35J05, 35R30

PII. S0036141096301038

1. Introduction.

The idea of using exponentially growing solutions in the context of inverse boundary value problems goes back to Calderón [C]. Motivated by Calderón's idea, Sylvester and Uhlmann [S-U] constructed exponentially growing solutions in order to prove global uniqueness for the conductivity of a body knowing the so-called Dirichlet-to-Neumann map.

The method of constructing exponentially growing solutions has been applied to other inverse problems like the inverse scattering problem at a fixed energy [N] and inverse spectral problems. However, Sylvester and Uhlmann's methods cannot be applied when we are in the presence of the first-order perturbation. Examples of this situation are given by the following:

(1) the problem of determining both the electrical conductivity and permitivity of a body and its magnetic permeability by measuring the tangential components of the electric field and the magnetic field at the boundary [O-P-S];

(2) the problem involving measurements at the boundary of an elastic medium, in which one measures displacements at the boundary and the corresponding stress at the boundary [N-U];

(3) the problem of determining both the electrical potential and the magnetic potential from boundary observations [Su], [N-Su-U].

This last problem is modeled by the Schrödinger equation in the presence of a magnetic field and it will be considered in this paper as an example of our techniques.

Let Ω be a bounded domain with smooth boundary in $\mathbb{R}^n$, $n \geq 3$; then the Schrödinger equation in the presence of a magnetic field is given by

$$H_{\vec{C},q} = \sum_{j=1}^{n} \left(-i\frac{\partial}{\partial x_j} + C_j(x) \right)^2 + q(x), \tag{1.1}$$

where $\vec{C} = (C_1, ..., C_n)$ is the magnetic potential and q is the electric potential. Assume $\vec{C} \in C^1(\overline{\Omega})$, $q \in L^\infty(\Omega)$ and both are real valued. If we assume further that zero is not a Dirichlet eigenvalue of (1.1) in Ω then we have that for any $f \in H^{1/2}(\Omega)$

*Received by the editors March 25, 1996; accepted for publication (in revised form) October 29, 1996.

http://www.siam.org/journals/sima/29-1/30103.html

†University of Washington, Department of Mathematics, Box 354350, Seattle, WA 98195-4350 (tolmasky@math.washington.edu).

there exists a unique $u \in H^1(\Omega)$ which solves

$$\begin{cases} H_{\vec{C},q} = 0 & \text{in } \Omega, \\ u|_{\partial\Omega} = f. \end{cases} \tag{1.2}$$

We define the Dirichlet-to-Neumann map by

$$\Lambda_{\vec{C},q} : f \longrightarrow \frac{\partial u}{\partial \eta}_{|\partial\Omega} + i(\vec{C}\cdot\eta)f, \tag{1.3}$$

where $f \in H^{1/2}(\Omega), u$ solves (1.2), and η is the unit outer normal to Ω.

The problem we consider is the one of recovering information about $\vec{C}$ and q given knowledge of the Dirichlet-to-Neumann map. It is well known (see [Su], [N-Su-U]) that the Dirichlet-to-Neumann map is invariant under gauge transformations in the magnetic potential. Consider $g \in C^1_\Omega$, where

$$C^s_\Omega = \{g \in C^s(\mathbb{R}^n) : \operatorname{supp} g \subset \Omega\}, \tag{1.4}$$

and consider the magnetic potential $\vec{C} + \nabla g$. Then $\Lambda_{\vec{C},q} = \Lambda_{\vec{C}+\nabla g,q}$. Therefore, we cannot hope to recover $\vec{C}$ from the Dirichlet-to-Neumann map. However, we can see that $rot(\vec{C})$ is invariant under a gauge transformation. It is then natural to wonder whether full knowledge of the Dirichlet-to-Neumann map gives full knowledge of $rot(\vec{C})$ and of q. This question has an affirmative answer given by Z. Sun in the case that the magnetic potential belongs to C^2_Ω and $q \in L^\infty(\Omega)$ provided that $rot(\vec{C})$ is small in the L^∞ topology [Su], and by Nakamura, Sun, and Uhlmann in the case that $\vec{C} \in C^\infty_\Omega$ and $q \in L^\infty(\Omega)$ under no assumptions on $rot(\vec{C})$ [N-Su-U]. In this paper we prove the following theorem.

THEOREM 1.1. *Let $\vec{C}_j \in C^1_\Omega, q_j \in L^\infty(\overline{\Omega})$, $j = 1, 2$. Assume that zero is not a Dirichlet eigenvalue for $H_{\vec{C}_j,q_j}$, $j = 1, 2$. If*

$$\Lambda_{\vec{C}_1,q_1} = \Lambda_{\vec{C}_2,q_2}$$

then

$$rot(\vec{C}_1) = rot(\vec{C}_2) \quad \textit{and} \quad q_1 = q_2 \quad \textit{in} \quad \Omega.$$

In [N-U], Nakamura and Uhlmann proved a result that enables us to find exponentially growing solutions to any smooth first-order perturbation of the Laplacian. In the same paper they used the solutions to prove global uniqueness for the inverse problem related to the elasticity system mentioned above.

In this paper we give a general method to construct exponentially growing solutions in the case that the first-order perturbation is not smooth. We note that the problem of global uniqueness in the case of a conductivity having less than two derivatives has been addressed by R. Brown in [Br]. Our method works by splitting the first-order term into a smooth part, with which we deal by following [N-U], and a nonsmooth one, for which we need estimates. Namely, we consider

$$S_{\vec{C}}(u) = (\Delta + \vec{C}\cdot\nabla)u = f$$

with $\vec{C} \in C^{2/3+\epsilon}(\overline{\Omega})$, $\epsilon > 0$. After conjugating by $e^{x\cdot\rho}$ ($\rho \in \mathbb{C}^n$, $\rho\cdot\rho = 0$) we get

$$S_{\vec{C},\rho} = \Delta_\rho u + \vec{C}\cdot\nabla_\rho u = f. \tag{1.5}$$

We can now decompose $\vec{C} \cdot \nabla_\rho$ into a pseudodifferential operator in the Shubin class (see [S]) plus a pseudodifferential operator with nonregular symbol depending on the parameter ρ for which we prove estimates.

The paper is organized as follows. In section 1 we define the spaces of symbols with limited regularity and we will show how to "smooth out" such symbols. The following section contains the proof of the dependence on the parameter ρ of the norm of the pseudodifferential operator associated with a symbol with limited regularity between Sobolev spaces. In section 3 we give the construction of the exponentially growing solutions for a generic perturbation of the Laplacian and in section 4 we prove Theorem 1.1.

2. Symbol smoothing. In this section we are going to smooth out nonregular symbols depending on a parameter. The ideas are similar as if we were working with no parameter at all (see [T, section 1.3]).

We recall now the definitions of the spaces $C^s(\mathbb{R}^n)$ and $C^s_*(\mathbb{R}^n)$.

DEFINITION 2.0.1. *Given* $0 < s < 1$, $C^s(\mathbb{R}^n)$ *is defined as the set of functions* u *such that*

$$|u(x+y) - u(x)| \leq C\ |y|^s.$$

For $k = 0, 1, 2, ...$, we take $C^k(\mathbb{R}^n)$ as the set of bounded continuous functions u so that $D^\beta u$ is bounded and continuous for any β such that $|\beta| \leq k$.

Then, if $s = k + r$, $0 \leq r \leq 1$, we define $C^s(\mathbb{R}^n)$ as the set of functions $u \in C^k(\mathbb{R}^n)$ so that $D^\beta u$ belongs to $C^r(\mathbb{R}^n)$ for $|\beta| = k$.

Let us consider the partition of unity

$$1 = \sum_{j=0}^{\infty} \psi_j(\xi)\ ,$$

where ψ_j is supported on $(1+|\xi|^2)^{1/2} \sim 2^j$ (by this we mean that there are constants M_1 and M_2 so that $(1+|\xi|^2)^{1/2} \leq M_1 2^j$ and $(1+|\xi|^2)^{1/2} \geq M_2 2^j$).

DEFINITION 2.0.2. *If* $s > 0$ *we say that* $u \in C^s_*(\mathbb{R}^n)$ *if and only if*

$$\sup_k\ 2^{ks} \|\psi_k(D)u\|_{L^\infty} < \infty,$$

and we define $\|\ \|_{C^s_*}$ *as the supremum of those numbers.*

A family of spaces $\{X^s : s \in \Sigma\}$ will be called a scale. We will be working with the spaces C^s and C^s_*, so in our case $\Sigma = (0, \infty)$.

We will introduce classes of symbols with limited regularity.

DEFINITION 2.0.3. *Let* $\delta \in [0, 1]$*:*

(a) $p_\rho(x, \xi) \in C^s_* S^m_{1,\delta,\rho}(\mathbb{R}^n)$ *if and only if*

$$|D^\alpha_\xi p_\rho(x, \xi)| \leq C_\alpha((1 + |\xi|^2 + |\rho|^2)^{\frac{1}{2}})^{m-|\alpha|}$$

and

$$\|D^\alpha_\xi p_\rho(\cdot, \xi)\|_{C^s_*} \leq C_\alpha\ ((1 + |\xi|^2 + |\rho|^2)^{\frac{1}{2}})^{m-|\alpha|+s\delta}$$

for any $\alpha \in \mathbf{Z}^n_+$.

(b) $p_\rho(x, \xi) \in C^s S^m_{1,\delta,\rho}(\mathbb{R}^n)$ *if the conditions on* (a) *are satisfied and, additionally,*

$$\|D^\alpha_\xi p_\rho(\cdot, \xi)\|_{C^j} \leq C_\alpha\ ((1 + |\xi|^2 + |\rho|^2)^{\frac{1}{2}})^{m-|\alpha|+j\delta}$$

for any $\alpha \in \mathbf{Z}^n_+$ *and* $0 \leq j \leq s, j \in \mathbf{N}$.

Our goal is to write a symbol $p_\rho(x,\xi) \in X^s S^s_{1,0,\rho}(\mathbb{R}^n)$ (where $X^s = C^s$ *or* C^s_*) as a sum of a smooth symbol and a reminder of lower order. We will find that the smooth part does not belong to $S^m_{1,0,\rho}$, but rather to one of the classes $S^m_{1,\delta,\rho}$.

We will need a partition of unity ψ^j_ρ of $\mathbb{R}^n$ such that

$$1 = \sum_{j=0}^{\infty} \psi^j_\rho(\xi),$$

where ψ^j_ρ is supported on $(1+|\xi|^2)^{1/2} \sim 2^j(1+|\rho|^2)^{1/2}$.

To construct such a partition we take $\psi^0_\rho(\xi)$ to be positive such that

$$\psi^0_\rho(\xi) = \begin{cases} 1 & \text{if } |\xi| \le (1+|\rho|^2)^{1/2}, \\ 0 & \text{if } |\xi| \ge 2(1+|\rho|^2)^{1/2}. \end{cases}$$

Finally, we set $\Psi^j_\rho(\xi) = \psi^0_\rho(2^{-j}\xi)$ and $\psi^j_\rho(\xi) = \Psi^j_\rho(\xi) - \Psi^{j-1}_\rho(\xi)$.

Now, given $p(x,\xi) \in X^s S^s_{1,0,\rho}$, choose $\delta \in (0,1]$ and set

$$p^\sharp_\rho(x,\xi) = \sum_{j=0}^{\infty} J_{\epsilon^j_\rho}\ p(x,\xi)\psi^j_\rho(\xi),$$

where

$$J_\epsilon f(x) = \theta(\epsilon D) f(x) \tag{2.1}$$

with $\theta \in C^\infty_0(R^n),\ \theta(\xi) = 1$ for $|\xi| \le 1$. We take

$$\epsilon^j_\rho = (2^{-j}(1+|\rho|^2)^{-1/2})^\delta. \tag{2.2}$$

We then define $p^b_\rho(x,\xi)$ to be $p_\rho(x,\xi) - p^\sharp_\rho(x,\xi)$, so our decomposition is

$$p_\rho(x,\xi) = p^\sharp_\rho(x,\xi) + p^b_\rho(x,\xi). \tag{2.3}$$

DEFINITION 2.0.4. *A scale X^s is called microlocalizable if, for $m \in \mathbf{R}$, $s, s+m \in \Sigma$,*

$$OPS^m_{1,0} : X^{s+m} \longrightarrow X^s.$$

Examples.

(1) The Sobolev spaces $H^{s,p}(\mathbb{R}^n)$ are microlocalizables provided $p \in (1,\infty)$.

(2) The property fails for the spaces C^s if s is an integer, but it turns out to be true for the Zygmund spaces C^s_*.

The following lemma will be useful to analyze the two terms in the decomposition (2.3); for the proof we refer to [T, Lemma 1.3.A].

LEMMA 2.1. *Let $\{X^s : s \in \Sigma\}$ be a microlocalizable scale; then, for $\epsilon \in (0,1]$,*

$$\|D^\beta_x J_\epsilon f\|_{X^s} \ \le\ C_\beta\ \epsilon^{-|\beta|}\|f\|_{X^s} \tag{2.4}$$

and

$$\|f - J_\epsilon f\|_{X^{s-t}} \ \le\ C\ \epsilon^t\ \|f\|_{X^s} \quad \textit{for } s, s-t \in \Sigma, \quad t \ge 0. \tag{2.5}$$

Using this, we derive the following proposition.

PROPOSITION 2.2. *If X^s is a microlocalizable scale and $p_\rho(x,\xi) \in X^s S^m_{1,0,\rho}$, then we have*

$$p^\sharp_\rho(x,\xi) \in S^m_{1,\delta,\rho} \tag{2.6}$$

and

$$p^b_\rho(x,\xi) \in X^{s-t} S^{m-t\delta}_{1,0,\rho} \quad if\ s, s-t \in \Sigma. \tag{2.7}$$

Proof. Let j be a nonnegative integer and suppose that $(1+|\xi|^2)^{1/2} \sim 2^j(1+|\rho|^2)^{1/2}$, i.e.,

$$(1+|\xi|^2)^{1/2} \leq C_1\, 2^j(1+|\rho|^2)^{1/2} \quad \text{and } 2^j(1+|\rho|^2)^{1/2} \leq C_2\,(1+|\xi|^2)^{1/2} \tag{2.8}$$

for some constants $C_1 > 0$ and $C_2 > 0$. Then

$$\begin{aligned} (1+|\xi|^2+|\rho|^2)^{1/2} &\leq (C_1^2\ 2^{2j}\ (1+|\rho|^2) + |\rho|^2)^{1/2} \\ &\leq (C_1^2\ 2^{2j} + 1)^{1/2}\ (1+|\rho|^2)^{1/2} \\ &\leq (C_1^2 + 1)^{1/2}\ 2^j\ (1+|\rho|^2)^{1/2} \end{aligned} \tag{2.9}$$

and

$$\begin{aligned} 2^j\ (1+|\rho|^2)^{1/2} &\leq C_2(1+|\xi|^2)^{1/2} \\ &\leq C_2(1+|\xi|^2+|\rho|^2)^{1/2}. \end{aligned} \tag{2.10}$$

From (2.9) and (2.10) we obtain that

$$(1+|\xi|^2)^{1/2} \sim 2^j(1+|\rho|^2)^{1/2} \iff (1+|\xi|^2+|\rho|^2)^{1/2} \sim 2^j(1+|\rho|^2)^{1/2}.$$

Let us analyze first the smooth part $p^\sharp_\rho(x,\xi)$. We fix j a nonnegative integer and ξ so that $(1+|\xi|^2)^{1/2} \sim 2^j(1+|\rho|^2)^{1/2}$. By (2.2), (2.4), (2.10), and $p_\rho(x,\xi) \in X^s S^m_{1,0,\rho}$ we have

$$\begin{aligned} \|D^\beta_x\, J_{\epsilon^j_\rho} p_\rho(\,\cdot\,,\xi)\|_{X^s} &\leq C_\beta\,(\epsilon^j_\rho)^{-|\beta|}\,\|p_\rho(\,\cdot\,,\xi)\|_{X^s} \\ &\leq C_\beta\,(2^j(1+|\rho|^2)^{1/2})^{\frac{\delta|\beta|}{2}}((1+|\xi|^2+|\rho|^2)^{1/2})^m \\ &\leq C_\beta\,((1+|\xi|^2+|\rho|^2)^{1/2})^{m+\delta|\beta|}. \end{aligned} \tag{2.11}$$

To prove (2.7) we first notice that derivatives in the x variable commute with the operators defined in (2.1); then it is enough to prove the estimates for $p^b_\rho(\,\cdot\,,\xi)$. Then, by (2.2), (2.5), (2.10), and $p_\rho(x,\xi) \in X^s S^m_{1,0,\rho}$ we have

$$\begin{aligned} \|p^b_\rho(\,\cdot\,,\xi)\|_{X^{s-t}} &\leq C\,(\epsilon^j_\rho)^t\,\|p_\rho(\,\cdot\,,\xi)\|_{X^s} \\ &\leq C\,(2^j(1+|\rho|^2)^{1/2})^{-\frac{\delta t}{2}}((1+|\xi|^2+|\rho|^2)^{1/2})^m \\ &\leq C\,((1+|\xi|^2+|\rho|^2)^{1/2})^{m-t\delta}. \end{aligned} \tag{2.12}$$

3. Continuity in $H^s(\mathbb{R}^n)$ of symbols in $C^r_* S^m_{1,0,\rho}(\mathbb{R}^n)$. The crucial step in the construction of the exponentially growing solutions is to know the dependence on ρ of the norm of operators with nonsmooth symbols. More specifically we will use the following result.

THEOREM 3.1. *Let $r > 0$ and $p_\rho(x,\xi) \in C^r_* S^m_{1,0,\rho}(\mathbb{R}^n)$. Then*

$$p_\rho(x,D): H^{s+m,p}(\mathbb{R}^n) \longrightarrow H^{s,p}(\mathbb{R}^n)$$

with

$$\|p_\rho(x,D)\|_{s+m,s} \le C\ ((1+|\rho|^2)^{\frac{1}{2}})^{s+m}$$

where $0 < s < r$, $p \in (1,\infty)$, and $\|\ \|_{s+m,s}$ denotes the operator norm between Sobolev spaces.

We will use the following results from the Littlewood–Paley theory.

LEMMA 3.2. *Let $f_k \in S'(\mathbb{R}^n)$ be such that, for some $A > 0$,*

$$supp\ \hat{f}_k \subset \{\xi\ ; A\,2^{k-1} \le |\xi| \le A\,2^{k+1}\}, \qquad k \ge 1,$$

and $\hat{f}_0$ has compact support. Then, for $p \in (1,\infty)$, $s \in \mathbb{R}$, we have

$$\left\| \sum_{k=0}^{\infty} f_k \right\|_{H^{s,p}} \sim \left\| \left\{ \sum_{k=0}^{\infty} 4^{ks}\ |f_k|^2 \right\}^{\frac{1}{2}} \right\|_{L^p}.$$

LEMMA 3.3. *Let $f_k \in S'(\mathbb{R}^n)$ be such that*

$$supp\ \hat{f}_k \subset \{\xi\ ;\ |\xi| \le A(1+|\rho|^2)^{\frac{1}{2}}\ 2^{k+1}\}, \qquad k \ge 0.$$

Then, for $p \in (1,\infty)$, $s > 0$, we have

$$\left\| \sum_{k=0}^{\infty} f_k \right\|_{H^{s,p}} \le C \left\| \left\{ \sum_{k=0}^{\infty} 4^{ks}\ (1+|\rho|^2)^{\frac{s}{2}}\ |f_k|^2 \right\}^{\frac{1}{2}} \right\|_{L^p}.$$

Proof of Theorem 3.1. The idea of the proof was taken from [T, Theorem 2.1.A] and it follows pioneering work of Coifman and Meyer [C-M]. See also Bourdaud [B]. By following [C-M] we can restrict ourselves to considering elementary symbols in $C^r_* S^m_{1,0,\rho}(\mathbb{R}^n)$. An elementary symbol in $C^r_* S^m_{1,0,\rho}(\mathbb{R}^n)$ can be written as

$$q_\rho(x,\xi) = \sum_{k=0}^{\infty} Q_k(x)\ \varphi^k_\rho(\xi)$$

with $supp\ \varphi^k_\rho$ on $(1+|\xi|^2)^{1/2} \sim 2^k\ (1+|\rho|^2)^{1/2}$.

So, by the definition of the class $C^r_* S^m_{1,0,\rho}(\mathbb{R}^n)$ we have that

$$\|Q_k(\cdot)\ \varphi^k_\rho(\xi)\ \|_{C^r_*} \le A\ ((1+|\xi|^2+|\rho|^2)^{1/2})^m. \tag{3.1}$$

We now consider a partition of unity ψ_j with $supp\ \psi_j$ on $(1+|\xi|^2)^{\frac{1}{2}} \sim 2^j$. Let $Q_{kj}(x) = \psi_j(D)\ Q_k(x)$ and $f_k = \varphi^k_\rho(D) f$.

Then

$$q_\rho(x,\xi) = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k-4} Q_{kj}(x) + \sum_{j=k-3}^{k+3} Q_{kj}(x) + \sum_{j=k+4}^{\infty} Q_{kj}(x) \right) \varphi^k_\rho(\xi)$$

$$= q^1_\rho(x,\xi) + q^2_\rho(x,\xi) + q^3_\rho(x,\xi). \tag{3.2}$$

We will work with each $q^i_\rho, i = 1,2,3$ separately.

Let us start with $q^1_\rho(x,D)$.
We have

$$\operatorname{supp}\left(\sum_{j=0}^{k-4} Q_{kj}\ f_k\right)\widehat{\ }$$

$$\subset\ \{\xi; 2^{k-1}(1+|\rho|^2)^{1/2} \leq |\xi| \leq 2^{k+1}(1+|\rho|^2)^{1/2}\} + \{\xi; |\xi| \leq 2^{k-3}\}$$

$$\subset \{\xi; 2^{k-1}(1+|\rho|^2)^{1/2} \leq |\xi| \leq 2^{k+1}(1+|\rho|^2)^{1/2}\} + \{\xi; |\xi| \leq 2^{k-3}(1+|\rho|^2)^{1/2}\}.$$

$$\subset \{\xi; A\ 2^{k-1}(1+|\rho|^2)^{1/2} \leq |\xi| \leq A\ 2^{k+1}(1+|\rho|^2)^{1/2}\}.$$

By Lemma 3.3 we get

$$\|q^1_\rho(x,D)f\|_{H^{s,p}}\ \leq\ C\ (1+|\rho|^2)^{s/2}\ \left\|\left\{\sum_{k=4}^{\infty}\ 4^{ks}\Big|\sum_{j=0}^{k-4} Q_{kj}\ f_k\ |^2\right\}^{\frac{1}{2}}\right\|_{L^p}. \tag{3.3}$$

From (3.1) and the definition of the spaces C^r_* we get

$$\|Q_{kj}(x)\|_\infty\ \leq\ B\ 2^{-jr}\ 2^{km}\ ((1+|\rho|^2)^{\frac{1}{2}})^m, \tag{3.4}$$

from which we obtain

$$\begin{aligned}\|q^1_\rho(x,D)f\|_{H^{s,p}} &\leq\ C_1\ \left\|\left(\sum_{k=4}^{\infty} 4^{k(s+m)}((1+|\rho|^2)^{\frac{1}{2}})^{s+m}|f_k|^2\right)^{1/2}\right\|_{L^p}\\ &\leq\ C_1\,\|f\|_{H^{s+m,p}}.\end{aligned} \tag{3.5}$$

We now make use of Lemma 3.2 to get

$$\left\|\left(\sum_{k=1}^{\infty}\ 4^{k(s+m)}\ ((1+|\rho|^2)^{1/2})^{s+m}|f_k\ |^2\right)^{\frac{1}{2}}\right\|_{L^p} \leq C\,\|f\|_{H^{s+m,p}},$$

which, combined with (3.5), gives

$$\|q^1_\rho(x,D)f\|_{H^{s,p}}\ \leq\ C_1\ \|f\|_{H^{s+m,p}}. \tag{3.6}$$

So, the result holds for $q^1_\rho(x,D)$. Let us proceed with $q^2_\rho(x,D)$.
In this case,

$$\operatorname{supp}\left(\sum_{j=k-3}^{k+3} Q_{kj}\ f_k\right)\widehat{\ }$$

$$\subset\ \{\xi; 2^{k-1}(1+|\rho|^2)^{1/2} \leq |\xi| \leq 2^{k+1}(1+|\rho|^2)^{1/2}\} + \{\xi; 2^{k-4} \leq |\xi| \leq 2^{k+4}\}$$

$$\subset \{\xi; 2^{k-1}(1+|\rho|^2)^{1/2} \leq |\xi| \leq 2^{k+1}(1+|\rho|^2)^{1/2}\} + \{\xi; |\xi| \leq 2^{k+4}(1+|\rho|^2)^{1/2}\}$$

$$\subset \{|\xi| \leq A\ 2^k(1+|\rho|^2)^{1/2}\}.$$

By Lemma 3.3,

$$\|q_\rho^2(x,D)f\|_{H^{s,p}} \;\le\; C\;(1+|\rho|^2)^{s/2}\;\left\|\left\{\sum_{k=0}^{\infty}\;4^{ks}\left|\sum_{j=k-3}^{k+3}\;Q_{kj}\;f_k\;\right|^2\right\}^{\frac12}\;\right\|_{L^p}.$$

Therefore, by (3.4), $|\sum_{j=k-3}^{k+3}\;Q_{kj}|\;\le\;C\;2^{km}\;((1+|\rho|^2)^{\frac12})^m$, and then we obtain

$$\|q_\rho^2(x,D)f\|_{H^{s,p}} \;\le\; C\;((1+|\rho|^2)^{1/2})^{s+m}\left\|\left\{\sum_{k=0}^{\infty}\;4^{k(s+m)}\;|f_k\;|^2\right\}^{\frac12}\;\right\|_{L^p}$$

$$\le\; C\;((1+|\rho|^2)^{1/2})^{s+m}\left\|\;|f_0|+\left(\sum_{k=1}^{\infty}\;4^{k(s+m)}\;|f_k\;|^2\right)^{\frac12}\;\right\|_{L^p}$$

$$\le\; C\;\left(((1+|\rho|^2)^{1/2})^{s+m}\|f_0\|_{L^p}+\left\|\left(\sum_{k=1}^{\infty}\;4^{k(s+m)}\;((1+|\rho|^2)^{1/2})^{s+m}|f_k\;|^2\right)^{\frac12}\;\right\|_{L^p}\right).$$

The estimate for q_ρ^2 follows if we notice that $\|f_0\|_{L^p}\le C\,\|f\|_{L^p}$ and that

$$\left\|\left(\sum_{k=1}^{\infty}\;4^{k(s+m)}\;((1+|\rho|^2)^{1/2})^{s+m}|f_k\;|^2\right)^{\frac12}\;\right\|_{L^p}\le C\,\|f\|_{H^{s+m,p}}.$$

Finally, we prove the result for $q_\rho^3(x,D)$.

$$supp\left(\sum_{k=0}^{j-4}\;Q_{kj}\;f_k\right)\widehat{\ }\subset\;\bigcup_{k=0}^{j-4}\;(\{\xi;2^{j-1}\le|\xi|\le 2^{j+1}\}$$

$$+\{\xi;2^{k-1}(1+|\rho|^2)^{1/2}\le|\xi|\le 2^{k+1}(1+|\rho|^2)^{1/2}\})$$

$$\subset\bigcup_{k=0}^{j-4}\;(\{\xi;|\xi|\le 2^{j+1}(1+|\rho|^2)^{1/2}\}+\{\xi;|\xi|\le 2^{k+1}(1+|\rho|^2)^{1/2}\})$$

$$\subset\{|\xi|\le A\;2^j(1+|\rho|^2)^{1/2}\}.$$

Applying once more Lemma 3.3 we obtain

$$\|q_\rho^3(x,D)f\|_{H^{s,p}} \;\le\; C\;(1+|\rho|^2)^{s/2}\;\left\|\left\{\sum_{j=4}^{\infty}\;4^{js}\;\left|\sum_{k=0}^{j-4}\;Q_{kj}\;f_k\;\right|^2\right\}^{\frac12}\;\right\|_{L^p}$$

$$\le\; C\;((1+|\rho|^2)^{1/2})^{s+m}\left\|\left(\sum_{j=4}^{\infty}\left|\sum_{k=0}^{j-4}\;2^{js}\;2^{-jr}\;2^{km}\;f_k\right|^2\right)^{1/2}\right\|_{L^p}$$

$$= C\ ((1+|\rho|^2)^{1/2})^{s+m} \left\| \left(\sum_{j=4}^{\infty} \left| \sum_{k=0}^{j-4} 2^{(k-j)(r-s)}\ 2^{k(s+m)}\ f_k \right|^2 \right)^{1/2} \right\|_{L^p},$$

where we used (3.4) and the fact that $2^{-kr} \leq C$. We now set

$$F_k = 2^{k(s+m)}|f_k|\ , \quad G_j = \sum_{k=0}^{j-4} 2^{(k-j)(r-s)}\ F_k.$$

Recalling that $0 < s < r$ we get, by using Young's inequality, that

$$\|q^3_\rho(x,D)f\|_{H^{s,p}} \ \leq\ C\ ((1+|\rho|^2)^{1/2})^{s+m} \left\| \left\{ \sum_{k=0}^{\infty} 4^{k(s+m)}\ |f_k\ |^2 \right\}^{\frac{1}{2}} \right\|_{L^p}$$

and we can now follow the steps from the previous case.

So, the theorem is proved by putting together the estimates for each one of the three pieces.

4. Constructing the solutions. Let us consider the equation

$$S_{\vec{C}}(\omega) = (\Delta + \vec{C}\cdot\nabla)\omega = f$$

with $\vec{C} \in C^{r+\epsilon}(\overline{\Omega})$, $\epsilon > 0$, and $r > 0$ to be determined.

Let $S_{\vec{C},\rho} = e^{-x\cdot\rho} S_{\vec{C}} e^{x\cdot\rho}$. Then

$$S_{\vec{C},\rho} = \Delta_\rho\omega\ +\ \vec{C}\cdot\nabla_\rho\omega = f, \tag{4.1}$$

where $\Delta_\rho = e^{-x\cdot\rho}\Delta e^{x\cdot\rho}$, $\nabla_\rho = e^{-x\cdot\rho}\nabla e^{x\cdot\rho}$, and $\rho \in \mathbb{C}^n$ with $\rho\cdot\rho = 0$.

In this section we prove the following proposition.

PROPOSITION 4.1. *Let $r = \frac{2}{3}$; then $S_{\vec{C},\rho}$ admits a bounded inverse $S^{-1}_{\vec{C},\rho} : L^2(\Omega) \longrightarrow H^1(\Omega)$. Moreover, if $f \in L^2(\Omega)$ and $\omega = S^{-1}_{\vec{C},\rho}(f)$ we have*

$$\|\omega\|_{L^2(\Omega)} \leq \frac{C}{|\rho|}\|f\|_{L^2(\Omega)}, \tag{4.2}$$

$$\|\omega\|_{H^1(\Omega)} \leq C\|f\|_{L^2(\Omega)}, \tag{4.3}$$

with C independent of ρ.

By the regularity assumed on $\vec{C}$ we get that $\vec{C}\cdot\nabla_\rho \in OPC^{r+\epsilon}S^1_{1,0,\rho}(\mathbb{R}^n)$.

Let us decompose $\vec{C}\cdot\nabla_\rho$ in two parts:

$$\vec{C}\cdot\nabla_\rho = N^\sharp_\rho(x,D)\ +\ N^b_\rho(x,D), \tag{4.4}$$

where

$$N^\sharp_\rho(x,D) \in OPS^1_{1,\delta,\rho}(\mathbb{R}^n)$$

and

$$N^b_\rho(x,D) \in OPC^{r+\epsilon-t}S^{1-t\delta}_{1,0,\rho}(\mathbb{R}^n)$$

for any $\delta \in (0,1)$, and $r < t < r+\epsilon$.

Now take $t = r + \frac{\epsilon}{2}$ and $t\delta > r$. Then we get that

$$N_\rho^b(x,D) \in OPC^{\frac{\epsilon}{2}} S^\eta_{1,0,\rho}(\mathbb{R}^n) \tag{4.5}$$

with $\eta = 1 - t\delta$.

So, in order to find solutions of (4.1), it would be enough to find solutions of

$$(\Delta_\rho \ + \ N_\rho^\sharp(x,D))\omega = f \tag{4.6}$$

and to use Theorem 3.1.

To solve (4.6) we make use of the following result due to Nakamura and Uhlmann.

LEMMA 4.2. *Let $N \in \mathbb{N}$. There exist operators $A_\rho(x,D)$ and $B_\rho(x,D)$ properly supported and belonging to $OPS^0_{1,\delta,\rho}(\mathbb{R}^n)$ such that for any $\phi_1 \in C_0^\infty(\mathbb{R}^n)$ there exist $\phi_2, \phi_3 \in C_0^\infty(\mathbb{R}^n)$, and $M > 0$ so that*

$$\phi_1(\Delta_\rho \ + \ N_\rho^\sharp(x,D))A_\rho(x,D)v = \phi_1 B_\rho(x,D)\ (\Delta_\rho + \phi_2 R_\rho^{-N(1-\delta)}(x,D)\phi_3)v, \tag{4.7}$$

where

$$\phi_2 R_\rho^{-N(1-\delta)}(x,D)\phi_3 : H^\alpha(\mathbb{R}^n) \longrightarrow H^\alpha(\mathbb{R}^n)$$

is a bounded linear operator with

$$\|\phi_2 R_\rho^{-N(1-\delta)}(x,D)\phi_3\|_{H^\alpha(\mathbb{R}^n), H^\alpha(\mathbb{R}^n)} \ \leq \ C\ ((1+|\rho|^2)^{1/2})^{-N(1-\delta)}$$

for $\rho \in Z, |\rho| \geq M$, and any $\alpha \in \mathbb{R}$. The functions ϕ_1, ϕ_2, ϕ_3 are taken to satisfy

$$\phi_1\phi_2 = \phi_1, \qquad \phi_1\phi_3 = \phi_1,$$

and

$$\phi_1 B_\rho(x,D)\phi_2 = \phi_1 B_\rho(x,D).$$

It is clear that we can assume that ϕ_1, ϕ_2, ϕ_3 also satisfy

$$\phi_1 N_\rho^b(x,D) = \phi_1 N_\rho^b(x,D)\phi_2, \qquad \phi_2 A_\rho(x,D)\phi_3 = \phi_2 A_\rho(x,D).$$

In terms of (4.1) we see that we would get

$$\begin{aligned} &\phi_1(S_{\vec{C},\rho})A_\rho(x,D)v \\ &= \phi_1 B_\rho(x,D)\ (\Delta_\rho + \phi_2 R_\rho^{-N(1-\delta)}(x,D)\phi_3)v + \phi_1 N_\rho^b(x,D)\phi_2 A_\rho(x,D)\phi_3 v. \end{aligned} \tag{4.8}$$

So, for any $f \in L^2(\mathbb{R}^n)$ it is enough to solve

$$\phi_1(B_\rho(x,D)\ (\Delta_\rho + \phi_2 R_\rho^{-N(1-\delta)}(x,D)\phi_3) + \phi_1 N_\rho^b(x,D)\phi_2 A_\rho(x,D)\phi_3)v \ = \ f \tag{4.9}$$

or, which is the same,

$$\begin{aligned} (\Delta_\rho + \phi_2 R_\rho^{-N(1-\delta)}(x,D)\phi_3 + \phi_1'(B_\rho(x,D))^{-1}\phi_1 N_\rho^b(x,D)\phi_2 A_\rho(x,D)\phi_3)v \\ = \ \phi_1'(B_\rho(x,D))^{-1}\phi_1(f). \end{aligned} \tag{4.10}$$

If we call

$$T_\rho = \phi_2 R_\rho^{-N(1-\delta)}(x,D)\phi_3 + \phi_1'(B_\rho(x,D))^{-1}\phi_1 N_\rho^b(x,D)\phi_2 A_\rho(x,D)\phi_3$$

we see that we need to study mapping properties for T_ρ. By Theorem (3.1) we know that

$$N^b_\rho(x,D) : H^{\eta+\frac{\epsilon}{2}\gamma}(\mathbb{R}^n) \longrightarrow H^{\frac{\epsilon}{2}\gamma}(\mathbb{R}^n) \tag{4.11}$$

with

$$\|N^b_\rho(x,D)\|_{H^{\eta+\frac{\epsilon}{2}\gamma}(\mathbb{R}^n),H^{\frac{\epsilon}{2}\gamma}(\mathbb{R}^n)} \leq C\ ((1+|\rho|^2)^{1/2})^{\eta+\frac{\epsilon}{2}\gamma} \tag{4.12}$$

for any $\gamma \in (0,1)$.

By [S, Theorem (9.1)] we know that

$$\phi_2 A_\rho(x,D)\phi_3 : H^{\eta+\frac{\epsilon}{2}\gamma}(\mathbb{R}^n) \longrightarrow H^{\eta+\frac{\epsilon}{2}\gamma}(\mathbb{R}^n)$$

with

$$\|\phi_2 A_\rho(x,D)\phi_3\|_{H^{\eta+\frac{\epsilon}{2}\gamma}(\mathbb{R}^n),H^{\eta+\frac{\epsilon}{2}\gamma}(\mathbb{R}^n)} \ \leq\ C\ ((1+|\rho|^2)^{1/2})^{\eta+\frac{\epsilon}{2}\gamma}$$

and that

$$\phi_1'(B_\rho(x,D))^{-1}\phi_1 : H^{\frac{\epsilon}{2}\gamma}(\mathbb{R}^n) \longrightarrow L^2(\mathbb{R}^n)$$

with

$$\|\phi_1'(B_\rho(x,D))^{-1}\phi_1\|_{H^{\frac{\epsilon}{2}\gamma}(\mathbb{R}^n),L^2(\mathbb{R}^n)} \ \leq\ C.$$

Putting all this together we obtain the following lemma.

LEMMA. *$T_\rho|_K : H^{\eta+\frac{\epsilon}{2}\gamma}(K) \longrightarrow L^2(K)$ for any compact set $K \supset \Omega$ and we have*

$$\|T_\rho|_K\|_{H^{\eta+\frac{\epsilon}{2}\gamma}(K),L^2(K)} \ \leq\ C\ ((1+|\rho|^2)^{\frac{1}{2}})^{2\eta+\epsilon\gamma} \tag{4.13}$$

with $\gamma \in (0,1)$ and C depending only on Ω.

We now compose $T_\rho|_\Omega$ with Δ_ρ^{-1} and make use of the fact that $\Delta_\rho^{-1} : L^2(\Omega) \longrightarrow H^s(\Omega)$ with

$$\|\Delta_\rho^{-1}\|_{L^2(\Omega),H^s(\Omega)} \ \leq\ \frac{C}{((1+|\rho|^2)^{\frac{1}{2}})^{1-s}} \tag{4.14}$$

for $s \in [0,1]$ and C depending only on Ω. We therefore get $T_\rho|_\Omega \ \circ\ \Delta_\rho^{-1} : L^2(\Omega) \longrightarrow L^2(\Omega)$ with

$$\|T_\rho|_\Omega \ \circ\ \Delta_\rho^{-1}\|_{L^2(\Omega),L^2(\Omega)} \ \leq\ C\ \frac{1}{((1+|\rho|^2)^{\frac{1}{2}})^{1-3\eta-2\epsilon\gamma}}. \tag{4.15}$$

So we see that, in order for this last expression to decay for $|\rho|$ big enough we need $\eta\ <\ \frac{1}{3}$. If we recall that $\eta = 1 - t\delta$ and that $t\delta > r$ we arrive at the conclusion that r must be at least $\frac{2}{3}$.

Then we can solve (4.10) on Ω with the following estimates:

$$\|\psi_1 v\|_{L^2(\mathbb{R}^n)} \ \leq\ \frac{C}{|\rho|}\ \|\tilde{f}\|_{L^2(\mathbb{R}^n)}, \tag{4.16}$$

where $\psi_1 \in C_0^\infty(\mathbb{R}^n)$ is taken so that

$$\phi_1(S_{\vec{C},\rho})A_\rho(x,D)\psi_1 = \phi_1(S_{\vec{C},\rho})A_\rho(x,D)$$

and

$$\tilde{f} = \begin{cases} f & \text{in } \Omega, \\ 0 & \text{in } \mathbb{R}^n \backslash \Omega. \end{cases} \tag{4.17}$$

So by now taking $\psi_2 \in C_0^\infty(\mathbb{R}^n)$ so that $\psi_2 \equiv 1$ on Ω and calling $\omega = \psi_2 A_\rho(x, D)\psi_1\, v$ we obtain, from (4.10), (4.16), and [S, Theorem 9.1],

$$S_{\vec{C},\rho}\, \omega \;=\; f \qquad \text{in} \quad \Omega. \tag{4.18}$$

For the estimate involving derivatives we are going to use the following standard interior estimate which we state without proof.

LEMMA 4.3. *Let $\omega \in H^1(\Omega)$ a solution of*

$$\Delta\omega \;=\; -F \qquad \textit{in} \;\; \Omega';$$

then

$$\|\omega\|_{H^1(\Omega)} \;\leq\; C\, \left(\|F\|_{H^{-1}(\Omega')} \;+\; \|\omega\|_{L^2(\Omega')}\right), \tag{4.19}$$

provided $\Omega \subset\subset \Omega'$.

Proof of (4.3). Without loss of generality we can assume that (4.18) holds in a slightly bigger domain Ω' and assume further that $\vec{C}$ is defined in Ω'. From the lemma we obtain

$$\|\omega\|_{H^1(\Omega)} \;\leq\; C\, (\| - 2\rho \cdot\; \nabla\omega - \vec{C}\cdot(\nabla + \rho)\omega + f\|_{H^{-1}(\Omega')} + \|\omega\|_{L^2(\Omega')}). \tag{4.20}$$

So the inequality follows by the fact that $\|\nabla\omega\|_{H^{-1}(\Omega')} \leq C\, \|\omega\|_{L^2(\Omega')}$ and by using (4.2) with Ω replaced by Ω' and f replaced by $\tilde{f}|_{\Omega'}$ (notice that $\|f\|_{L^2(\Omega)} = \|\tilde{f}\|_{L^2(\Omega')}$).

5. Application to the Schrödinger equation in a magnetic field. The purpose of this section is to apply the solutions we constructed to an inverse boundary value problem for the Schrödinger equation in the presence of a magnetic field.

Let Ω be a bounded domain in $\mathbb{R}^n, n \geq 3$, with smooth boundary. The equation we are going to study is given by

$$H_{\vec{C},q} = \sum_{j=1}^{n} \left(-i\frac{\partial}{\partial x_j} + C_j(x)\right)^2 + q(x), \tag{5.1}$$

where $\vec{C} = (C_1, ..., C_n) \in C^1(\overline{\Omega})$ is the magnetic potential and the scalar function $q \in L^\infty(\Omega)$ is the electric potential. We assume both to be real valued.

We assume further that zero is not a Dirichlet eigenvalue of (5.1) on Ω. Then the boundary value problem

$$\begin{cases} H_{\vec{C},q} u = 0 & \text{in } \Omega, \\ u|_{\partial\Omega} = f \;\in H^{1/2}(\partial\Omega) \end{cases} \tag{5.2}$$

has a unique solution $u \in H^1(\Omega)$. The usual computations give that the Dirichlet-to-Neumann map in this case is given by

$$\Lambda_{\vec{C},q} : f \longrightarrow \frac{\partial u}{\partial \nu} + i(\vec{C}\cdot\nu)f\;, \qquad f \;\in H^{1/2}(\partial\Omega), \tag{5.3}$$

where u is the unique solution to (5.2) and ν is the outer normal to $\partial\Omega$.

The question here is under what assumptions we can recover $\vec{C}$ and q.

It is well known that two magnetic potentials $\vec{C}$ and $\vec{C}+\nabla g$, where

$$g \in C^1_\Omega = \{f \in C^1(\mathbb{R}^n) : \operatorname{supp} f \subset \Omega\}, \tag{5.4}$$

produce the same Dirichlet-to-Neumann map [Su]. As $rot(\vec{C}) = rot(\vec{C}+\nabla g)$ it is natural to ask whether $rot(\vec{C})$ and q can be uniquely determined by $\Lambda_{\vec{C},q}$. Z. Sun [Su] proved that this is actually true in the case that the magnetic potential $\vec{C}$ belongs to C^2_Ω and the electric potential q is a bounded function on Ω provided $rot(\vec{C})$ is small in $L^\infty(\Omega)$. He proved the following theorem.

THEOREM 5.1 (Su). *Let $\vec{C}_j \in C^2_\Omega$ (see (1.4)), $q_j \in L^\infty(\Omega)$, $j=1,2$. Assume that zero is not a Dirichlet eigenvalue for $H_{\vec{C}_j,q_j}$, $j=1,2$. Then there exists a constant $\epsilon = \epsilon(\Omega) > 0$ such that if $\|rot(\vec{C}_j)\|_{L^\infty(\Omega)} \leq \epsilon$, $j=1,2$, and*

$$\Lambda_{\vec{C}_1,q_1} = \Lambda_{\vec{C}_2,q_2},$$

then

$$rot(\vec{C}_1) = rot(\vec{C}_2) \quad and \quad q_1 = q_2 \quad in \quad \Omega.$$

The restriction on $rot(\vec{C})$ was removed in the case that $\vec{C}$ is in the C^∞ class by Nakamura, Sun, and Uhlmann [N-Su-U]. They proved the following theorem.

THEOREM 5.2 (see [N-Su-U]). *Let $\vec{C}_j \in C^\infty_\Omega$ (see (1.4)), $q_j \in L^\infty(\Omega)$, $j=1,2$. Assume that zero is not a Dirichlet eigenvalue for $H_{\vec{C}_j,q_j}$, $j=1,2$. If*

$$\Lambda_{\vec{C}_1,q_1} = \Lambda_{\vec{C}_2,q_2}$$

then

$$rot(\vec{C}_1) = rot(\vec{C}_2) \quad and \quad q_1 = q_2 \quad in \quad \Omega.$$

Using the solution we have constructed already we are able to prove Sun's theorem without the restriction on $rot(\vec{C})$.

As we said before, recovering $rot(\vec{C})$ is the most we can expect from the Dirichlet-to-Neumann map. Therefore, this result appears to be optimal.

Remark. We could follow [Su] up to the point in which we use the solutions constructed in the previous section. However, by doing so we would get a weaker result; namely, we would end up needing $\vec{C}_j \in C^2_\Omega, i=1,2$. We sketch how to remove the condition on $rot(\vec{C}_j), j=1,2$ following Sun's proof.

Let us look for solutions of $H_{\vec{C},q}$ of the form

$$u(x,\rho) = e^{x\cdot\rho+\phi(x,\rho)}(1+\omega(x,\rho)), \tag{5.5}$$

where $\rho \in \mathbb{C}^n$ is a complex vector satisfying $\rho\cdot\rho = 0$ and $\omega(x,\rho)$ has decay properties that we will state. By plugging into the equation $H_{\vec{C},q}u(x,\rho) = 0$ we get the following couple of equations:

$$\rho\cdot\nabla\phi = -i\rho\cdot\vec{C}, \tag{5.6}$$

$$\Delta\omega + 2(\rho+\nabla\phi+i\vec{C})\cdot\nabla\omega - g\omega = g, \tag{5.7}$$

where

$$g = \vec{C}^2 - i\nabla \cdot \vec{C} + q - 2i\vec{C} \cdot \nabla\phi - \nabla\phi \cdot \nabla\phi - \Delta\phi. \tag{5.8}$$

We use (2.6) in [Su] to solve (5.6). We get that $\phi \in C^2(\overline{\Omega})$ and

$$\|\phi\left(\cdot, \frac{\rho}{|\rho|}\right)\|_{C^2(\overline{\Omega})} \leq C \|\vec{C}\|_{C^2_\Omega}. \tag{5.9}$$

From (5.8) and (5.9) it is clear that $\vec{C}_j \in C^2_\Omega, i = 1, 2$ cannot be lowered. We will come to this later. We can rewrite (5.7) as

$$\Delta_\rho \omega + \vec{D} \cdot \nabla\omega - h\omega = h, \tag{5.10}$$

where

$$\Delta_\rho = \Delta + 2\rho \cdot \nabla, \quad \vec{D} = 2(\nabla\phi + i\vec{C})\psi, \quad h = g\psi \tag{5.11}$$

and $\psi \in C_0^\infty(\mathbb{R}^n)$ with $\psi|_\Omega \equiv 1$.

By (5.9) and the regularity assumed on $\vec{C}$ we see that $\nabla\phi \in C^1(\overline{\Omega})$ and then $\vec{D} \in C^1(\overline{\Omega})$. By the results from the previous section we know that for any $f \in L^2(\Omega)$ we can find $w \in H^1(\Omega)$ so that

$$(\Delta_\rho + \vec{D} \cdot \nabla)w = f \quad \text{in } \Omega.$$

Moreover,

$$\|w\|_{L^2(\Omega)} \leq \frac{C}{|\rho|} \|f\|_{L^2(\Omega)}, \tag{5.12}$$

$$\|w\|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}. \tag{5.13}$$

Therefore, it would not take much effort to prove Sun's theorem without the smallness condition. However, it pays off to take a slightly different path. By doing this we will be able to remove the smallness condition and also to sharpen the requirement on the regularity of the magnetic potential. We rewrite the Schrödinger equation as

$$H_{\vec{C},q} = -\Delta - 2\vec{C}(x) \cdot \nabla + G(x), \tag{5.14}$$

where $G = \vec{C}^2 - 2i\nabla \cdot \vec{C} + q$.

If we look for solutions of the form $e^{x\cdot\rho}(1 + \omega(x, \rho))$, we get

$$H_{\vec{C},q,\rho}\omega = (\Delta_\rho + 2\vec{C} \cdot \nabla_\rho + G(x))\omega = -G(x).$$

In the previous section we proved that $\Delta_\rho + 2\vec{C} \cdot \nabla_\rho$ is invertible as a map from $H^1(\Omega) \longrightarrow L^2(\Omega)$ and satisfies the usual estimates. We now prove that the same is true for $\Delta_\rho + 2\vec{C} \cdot \nabla_\rho + G(x)$.

LEMMA 5.3. *$H_{\vec{C},q,\rho}$ admits a bounded inverse $H^{-1}_{\vec{C},q,\rho} : L^2(\Omega) \longrightarrow H^1(\Omega)$. Moreover, if $g \in L^\infty(\Omega)$ we have*

$$\|H^{-1}_{\vec{C},q,\rho}(g)\|_{L^2(\Omega)} \leq \frac{C}{|\rho|}\|g\|_{L^2(\Omega)},$$

$$\|H^{-1}_{\vec{C},q,\rho}(g)\|_{H^1(\Omega)} \leq C\|g\|_{L^2(\Omega)},$$

with C independent of ρ.

Proof. Applying $(\Delta_\rho + 2\vec{C}\cdot\nabla_\rho)^{-1}$ to both sides, we get

$$(I+T)\omega = -(\Delta_\rho + 2\vec{C}\cdot\nabla_\rho)^{-1}(G), \tag{5.15}$$

where $T = (\Delta_\rho + 2\vec{C}\cdot\nabla_\rho)^{-1}\circ(G)$. By the regularity assumed on $\vec{C}$, $G \in L^\infty(\Omega)$ and then we obtain that $T : L^2(\Omega) \longrightarrow L^2(\Omega)$ with

$$\|T\|_{L^2(\Omega),L^2(\Omega)} \leq \frac{C}{|\rho|}\|G\|_{L^\infty}.$$

Therefore (5.15) has a unique solution provided $|\rho|$ is big enough. So the first estimate follows from $\|(\Delta_\rho+2\vec{C}\cdot\nabla_\rho)^{-1}(G)\|_{L^2(\Omega)} \leq \frac{C}{|\rho|}\|G\|_{L^2(\Omega)}$. We can now get the remaining estimate by writing

$$\omega = (\Delta_\rho + 2\vec{C}\cdot\nabla_\rho)^{-1}(G(1+\omega)).$$

We now turn to analyzing the behavior of the operator A_ρ when $|\rho| \longrightarrow \infty$.

LEMMA 5.4. *Let ϕ solve (5.6) and $\rho = s\rho_0$, with $s \in \mathbb{R}$ and $|\rho_0| = 1$. Then if $f \in C_0^\infty(\mathbb{R}^n)$ we have that*

$$\lim_{s\to\infty} A_\rho(x,D)f(x) = e^{\phi(x,\frac{\rho}{|\rho|})}f(x) \tag{5.16}$$

uniformly on compact sets.

Proof. Let us fix a compact set $K \subset \mathbb{R}^n$. By developing $a_\rho(x,\xi)$ (the full symbol of $A_\rho(x,D)$) in Taylor series in ξ we get:

$$a_\rho(x,\xi) = a_\rho(x,0) + \sum_{i=1}^n \xi_i \frac{\partial}{\partial\xi_i}a_\rho(x,0) + R_\rho(x,\xi,0), \tag{5.17}$$

where

$$R_\rho(x,\xi,0) = \sum_{i,j=1}^n \int_0^1 (1-t)\xi_i\xi_j \frac{\partial^2}{\partial\xi_i\partial\xi_j}a_\rho(x,t\xi)dt. \tag{5.18}$$

Now apply the operator given by (5.17) to $f \in C_0^\infty(\mathbb{R}^n)$. We need to prove that

$$\sum_{i=1}^n \frac{\partial}{\partial\xi_i}a_\rho(x,0)D_{x_i}(f) \quad\text{and}\quad \sum_{i,j=1}^n \int e^{ix\cdot\xi}\left(\int_0^1 (1-t)\frac{\partial^2}{\partial\xi_i\partial\xi_j}a_\rho(x,t\xi)dt\right)\xi_i\xi_j\hat{f}(\xi)d\xi$$

tend to 0 uniformly on K as $s\to\infty$. By the estimates on $a_\rho(x,\xi)$ it is easy to see this for the first term. For the second one we use

$$\left|\frac{\partial^2}{\partial\xi_i\partial\xi_j}a_\rho(x,t\xi)\right| \;\leq\; \frac{C_K}{1+|t\xi|^2+|\rho|^2} \;\leq\; \frac{C_K}{1+|\rho|^2},$$

which is valid for any $x \in K$. This proves that $|A_\rho(x,D)f(x) - a_\rho(x,0)f(x)| \longrightarrow 0$ as $s\to\infty$ uniformly on K.

Now, $a_\rho(x,0)$ solves the following equation:

$$\rho\cdot\nabla log(a_\rho(x,0)) = -iJ_{\epsilon_\rho^0}(\vec{C}(x)\cdot\rho).$$

Solving it we get

$$log(a_\rho(x,0)) = \frac{1}{(2\pi)^n}\int e^{-ix\eta}\frac{\rho\cdot\theta(\epsilon_\rho^0\eta)\vec{C}(\eta)}{\rho\cdot\eta}d\eta$$

or

$$log(a_\rho(x,0)) = \frac{1}{(2\pi)^n}\int e^{-ix\eta}\frac{\frac{\rho}{|\rho|}\cdot\theta(\epsilon_\rho^0\eta)\vec{C}(\eta)}{\frac{\rho}{|\rho|}\cdot\eta}d\eta.$$

As $s\to\infty, \epsilon_\rho^0\to 0$, so the last expression approaches

$$\frac{1}{(2\pi)^n}\int e^{-ix\eta}\frac{\frac{\rho}{|\rho|}\cdot\vec{C}(\eta)}{\frac{\rho}{|\rho|}\cdot\eta}d\eta$$

as $s\to\infty$ uniformly on compact sets which says that $\lim_{s\to\infty} log(a_\rho(x,0))$ satisfies (5.6).

We recall, without proof, the following identity proved in [S].

PROPOSITION 5.5. *Let $\vec{C}_j\in C^1_\Omega, q_j\in L^\infty(\overline{\Omega}),\ j=1,2$. Then*

$$i\int_\Omega(\vec{C}_1-\vec{C}_2)\cdot(u_1\nabla\overline{u_2}-\overline{u_2}\nabla u_1)dx+\int_\Omega(\vec{C}_1^2-\vec{C}_2^2+q_1-q_2)u_1\overline{u_2}dx$$

$$=-\int_{\partial\Omega}\overline{u_2}(\Lambda_{\vec{C}_1,q_1}-\Lambda_{\vec{C}_2,q_2})u_1ds$$

holds for the arbitrary u_j solution of $H_{\vec{C}_j,q_j}u_j=0, j=1,2$.

COROLLARY 5.6. *Assume the conditions from the previous proposition. If $\Lambda_{\vec{C}_1,q_1}=\Lambda_{\vec{C}_2,q_2}$ then*

$$i\int_\Omega(\vec{C}_1-\vec{C}_2)\cdot(u_1\nabla\overline{u_2}-\overline{u_2}\nabla u_1)dx+\int_\Omega(\vec{C}_1^2-\vec{C}_2^2+q_1-q_2)u_1\overline{u_2}dx=0 \tag{5.19}$$

holds for the arbitrary u_j solution of $H_{\vec{C}_j,q_j}u_j=0, j=1,2$.

Proof of Theorem 1.1. Let k,γ_1,γ_2 be three mutually orthogonal vectors $\in\mathbb{R}^n$ with $|\gamma_1|=|\gamma_2|=1$. Let $\zeta,\rho\in\mathbb{C}^n$ be given by

$$\zeta=\gamma_1+i\gamma_2,\qquad \rho=s\zeta+g(s,k)\gamma_1, \tag{5.20}$$

where s is a positive real parameter and

$$g(s,k)=2^{-1}|k|^2((|k|^2+4s^2)^{1/2}+4s)^{-1}. \tag{5.21}$$

Let $\rho_1,\rho_2\in\mathbb{C}^n$ be given by

$$\rho_1=i\frac{k}{2}+\rho,\qquad \overline{\rho_2}=i\frac{k}{2}-\rho. \tag{5.22}$$

We have

$$\rho_1\cdot\rho_1=\rho_2\cdot\rho_2=0,\qquad \rho_1+\overline{\rho_2}=ik,\qquad \rho_1-\overline{\rho_2}=2\rho, \tag{5.23}$$

$$\frac{\rho_1}{s} \to \zeta, \qquad \frac{\overline{\rho_2}}{s} \to -\zeta, \qquad \text{as } \ s \to \infty. \tag{5.24}$$

Let us construct solutions:

$$u_j(x,\rho_j) = e^{x\cdot\rho_j}\psi_2(x)A_{\rho_j}(x,D)\psi_1(x)(1+\omega_j(x,\rho_j)) \tag{5.25}$$

which is the solution of $H_{\vec{C}_j,q_j}u_j = 0, j = 1,2$ in Ω, $\psi_1,\psi_2 \in C_0^\infty(\mathbb{R}^n), \psi_j = 1$ in $\Omega, j = 1,2$, and $\omega_j, j = 1,2$ satisfying

$$\|\omega_j\|_{L^2(\Omega)} \leq \frac{C}{|\rho|} \quad \text{and} \qquad \|\nabla\omega_j\|_{H^1(\Omega)} \leq C \tag{5.26}$$

with C depending only on Ω, $\|\vec{C}_j\|_{C^1(\Omega)}$ and $\|q_j\|_{L^\infty(\Omega)}, j = 1,2$.

We now plug (5.25) into (5.19) to obtain

$$F + G + H + I + J + K + L = 0, \tag{5.27}$$

where F and G are functions of s, k, γ_1, γ_2, and they are defined by

$$F = -2i\int_\Omega e^{ix\cdot k}\rho\cdot(\vec{C}_1 - \vec{C}_2)A_{\rho_1}(x,D)(\psi_1)\overline{A_{\rho_2}(x,D)(\psi_1)}dx, \tag{5.28}$$

$$G = i\int_\Omega e^{ix\cdot k}A_{\rho_1}(x,D)(\psi_1)\overline{(\nabla(A_{\rho_2}(x,D)(\psi_1)) + \nabla(A_{\rho_2}(x,D)(\psi_1\omega_2)))}dx, \tag{5.29}$$

$$H = i\int_\Omega e^{ix\cdot k}A_{\rho_1}(x,D)(\psi_1\omega_1)\overline{(\nabla(A_{\rho_2}(x,D)(\psi_1)) + \nabla(A_{\rho_2}(x,D)(\psi_1\omega_2)))}dx, \tag{5.30}$$

$$I = -i\int_\Omega e^{ix\cdot k}\nabla(A_{\rho_1}(x,D)(\psi_1))\overline{(A_{\rho_2}(x,D)(\psi_1) + A_{\rho_2}(x,D)(\psi_1\omega_2))}dx, \tag{5.31}$$

$$J = -i\int_\Omega e^{ix\cdot k}\nabla(A_{\rho_1}(x,D)(\psi_1\omega_1))\overline{(A_{\rho_2}(x,D)(\psi_1) + A_{\rho_2}(x,D)(\psi_1\omega_2))}dx, \tag{5.32}$$

$$\begin{aligned} K = \int_\Omega e^{ix\cdot k}(&\vec{C_1}^{\,2} - \vec{C_2}^{\,2} \\ &+q_1 - q_2)A_{\rho_1}(x,D)(\psi_1)\overline{(A_{\rho_2}(x,D)(\psi_1) + A_{\rho_2}(x,D)(\psi_1\omega_2))}dx, \end{aligned} \tag{5.33}$$

$$\begin{aligned} L = \int_\Omega e^{ix\cdot k}(&\vec{C_1}^{\,2} - \vec{C_2}^{\,2} \\ &+q_1 - q_2)A_{\rho_1}(x,D)(\psi_1\omega_1)\overline{(A_{\rho_2}(x,D)(\psi_1) + A_{\rho_2}(x,D)(\psi_1\omega_2))}dx. \end{aligned} \tag{5.34}$$

We now apply (5.26) and Lemma (5.4) to get

$$\lim_{s\to\infty}\frac{G}{s} = \lim_{s\to\infty}\frac{H}{s} = \lim_{s\to\infty}\frac{I}{s} = \lim_{s\to\infty}\frac{J}{s} = 0 \lim_{s\to\infty}\frac{K}{s} = 0 \lim_{s\to\infty}\frac{L}{s} = 0 \tag{5.35}$$

and then

$$\lim_{s\to\infty}\frac{F}{s} = -2i\int_\Omega e^{ix\cdot k+\phi_1+\overline{\phi_2}}\zeta\cdot(\vec{C_1} - \vec{C_2})dx = 0, \tag{5.36}$$

where ϕ_j solves $\zeta\cdot\nabla\phi_1 = -i\zeta\cdot\vec{C}_1$ and $\zeta\cdot\nabla\overline{\phi_2} = i\zeta\cdot\vec{C}_2$.

(5.36) is all we need to end the proof of the theorem. The proof is finished following Sun's arguments. We refer to [Su] for details.

REFERENCES

[B] G. Bourdaud, *L^p-estimates for certain non-regular pseudo-differential operators*, Comm. Partial Differential Equations, 7 (1982), pp. 1023–1033.

[Br] R. Brown, *Global uniqueness in the impedance imaging problem for less regular conductivities*, SIAM J. Math Anal., 27 (1996), pp. 1049–1056.

[C] A. P. Calderón, *On an inverse boundary value problem*, Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasileira de Matemática, Rio de Janeiro, 1980, pp. 65–73.

[C-M] R. Coifman and Y. Meyer, *Au delà des opérateurs pseudodifférentiels*, Astérisque 57, Soc. Math de France, 1978.

[N] R. Novikov, *Multidimensional inverse spectral problems for the equation $-\Delta\psi+(v(x)-Eu(x))\psi=0$*, Funct. Anal. Appl., 22 (1988), pp. 263–272.

[N-U] G. Nakamura and G. Uhlmann, *Global uniqueness for an inverse boundary problem arising in elasticity*, Invent. Math., 118 (1994), pp. 457–474.

[N-Su-U] G. Nakamura, Z. Sun, and G. Uhlmann, *Global identifiability for an inverse problem for the Schrödinger equation in a magnetic field*, Math. Ann., 303 (1995), pp. 377–388.

[O-P-S] P. Ola, L. Päivärinta, and E. Somersalo, *An inverse boundary value problem in electrodynamics*, Duke Math. J., 70 (1993), pp. 617–653.

[S] M. A. Shubin, *Pseudodifferential operators and spectral theory*, Springer Series in Soviet Mathematics, Springer-Verlag, Berlin, New York, 1987.

[Su] Z. Sun, *An inverse boundary value problem for Schrödinger operators with vector potentials*, Trans. Amer. Math. Soc., 2 (1993), pp. 953–969.

[S-U] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math., 125 (1987), pp. 153–169.

[T] M. Taylor, *Pseudodifferential operators and nonlinear PDE*, Progress in Mathematics, Birkhäuser, Boston, MA, 1991.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 134–154, January 1998

009

LOCAL AVERAGE LIAPUNOV FUNCTIONS AND PERSISTENCE IN POPULATION DYNAMICS*

W. H. RUAN†

Abstract. This paper is concerned with the problem of uniform persistence in population dynamics. We consider systems of reaction–diffusion equations which model ecosystems in bounded habitats with diffusion. It is shown that a system is persistent if every chain-recurrent set in the boundary of the positive cone does not attract trajectories from the interior of the positive cone, and this property can be determined by using localized average Liapunov functions. Some results on constructing local average Liapunov functions are given. A system describing a food chain model is discussed as an example.

Key words. population dynamics, average Liapunov functions, persistence

AMS subject classifications. 35K57, 35B40, 92D40

PII. S0036141096297376

1. Introduction.

This paper is concerned with the uniform persistence of dynamical systems defined by systems of nonlinear partial differential equations of parabolic type. We consider the following initial-boundary value problem:

$$\begin{aligned} &\partial u_i/\partial t - L_i u_i = u_i f_i(u), && (x \in \Omega,\ t > 0), \\ &B_i u_i = 0, && (x \in \partial\Omega,\ t > 0), \quad (i = 1, \dots, N), \\ &u_i(x, 0) = u_i^0(x), && (x \in \Omega), \end{aligned} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^n$ is a bounded domain with a smooth boundary $\partial\Omega$; $u = (u_1, \dots, u_N)$ is a vector function with each component nonnegative; L_i is a uniformly strongly elliptic operator having the divergence form

$$L_i u = \sum_{i,j=1}^{n} (a_{ij}(x) u_{x_i})_{x_j} + \sum_{i=1}^{n} b_i(x) u_{x_i}; \tag{1.2}$$

B_i is a boundary operator such that

$$B_i u(x) = \delta_i \frac{\partial u(x)}{\partial \nu(x)} + \beta_i(x) u, \tag{1.3}$$

where δ_i is either 0 (Dirichlet condition) or 1 (Neumann or Robin condition); β_i is a nonnegative function which is the constant 1 if $\delta = 0$; $\nu(x) = (\nu_1, \dots, \nu_n)$ is the outward conormal vector at $x \in \partial\Omega$ with

$$\nu_i = \sum_{j=1}^{n} a_{ij}(x) \zeta_j, \qquad i = 1, \dots, n;$$

and $\zeta = (\zeta_1, \dots, \zeta_n)$ is the outward normal vector at $x \in \partial\Omega$. It is assumed that $a_{ij},\ b_i \in C^1(\overline{\Omega})$, $\beta_i \in C^{1+\alpha}(\partial\Omega)$, $u_i^0 \in C(\overline{\Omega})$, $f_i \in C^1(\mathbb{R}^+)$, and $\partial\Omega \in C^{1+\alpha}$, where

*Received by the editors January 17, 1996; accepted for publication (in revised form) October 1, 1996.

http://www.siam.org/journals/sima/29-1/29737.html

†Department of Mathematics, Computer Science and Statistics, Purdue University Calumet, Hammond, IN 46323 (ruan@nwi.calumet.purdue.edu).

$\alpha > 0$ is a constant. Such a system often arises in the study of population dynamics of ecological systems. Typically, Ω represents the habitat of several interacting species, u_i represents the density of the ith species (usually after scaling), f_i describes the effect of interaction among species on the growth rate of the ith species, and L_i and B_i account for diffusion and migration of the ith species in the interior and on the boundary of the habitat, respectively.

An important problem in ecology is to determine conditions under which all species survive in the long term. In mathematical language, it is the problem of persistence of the dynamical system. Let X_i be a Banach space such that $X_i = C_0(\Omega)$ if B_i is the Dirichlet condition and

$$X_i = \{u \in C^1(\overline{\Omega}) : \ B_i u = 0\}$$

if B_i is Neumann or Robin condition, and let X be the positive cone of the product space $\prod_{i=1}^N X_i$, i.e.,

$$X = \left\{ u \in \prod_{i=1}^N X_i : \ u_i \geq 0 \quad \text{for } i = 1, \ldots, N \right\}.$$

It is a consequence of the maximum principle that X is forwardly invariant for the semiflow π generated by system (1.1); that is, given any $\xi \in X$, the solution $u(\cdot, t)$ of (1.1) with $u^0 = \xi$ lies in X for all $t > 0$. Furthermore, by the uniqueness of solution, each face $S_i \equiv \{u \in X : u_i \equiv 0\}$ of the boundary of X is also forwardly invariant for the semiflow, and so is the total face $S \equiv \bigcup_{i=1}^N S_i$. The system (1.1), or equivalently the semiflow π, is called *persistent* if for all $\xi \in X \setminus S$, $\liminf_{t\to\infty} d(\xi\pi t, S) > 0$, and it is called *uniformly persistent* if there exists $\epsilon_0 > 0$ such that for all $\xi \in X \setminus S$, $\liminf_{t\to\infty} d(\xi\pi t, S) \geq \epsilon_0$ (cf. [4]). Here, $d(\cdot, \cdot)$ is the distance function in X defined by the norm, and $\xi\pi t$ is the solution $u(\cdot, t)$ of problem (1.1) with $u^0 = \xi$. In the case where the system is compact and point dissipative (the latter means there is a bounded nonempty set $B \subset X$ such that for any $\xi \in X$ there exists a $t_0 > 0$ such that $\xi\pi t \in B$ for all $t \geq t_0$), a result of Hale and Waltman [11] shows that persistence is equivalent to uniform persistent. This is true for many reaction–diffusion systems modeling ecosystems, and hence in this paper, we only discuss conditions for systems to be persistent.

The study of persistence of dynamical systems has attracted a considerable amount of attention in recent years (cf., e.g., [3, 4, 5, 6, 7, 9, 11, 13, 14, 17] and references therein). There are a number of important directions. One is to analyze the behavior of invariant sets on the boundary S and determine conditions for the maximal invariant set $I(S)$ in S to be repellent to trajectories in $X \setminus S$. Recall that a set in X is called invariant if it is a collection of full orbits, that is, orbits of solutions of (1.1) that are defined for all $t \in \mathbb{R}$. In view of Theorem 2.2 of [11], the maximal invariant set $I(S)$ in S exists and attracts all trajectories in S provided that the system is compact and point dissipative. It is found that if $I(S)$ has a finite covering by a family of isolated invariant sets such that there is no cycle in the family and each invariant set does not attract any trajectory from $X \setminus S$, then the system is persistent [3, 4, 9, 11]. Another direction is to use the so-called average Liapunov functions. An average Liapunov function is a continuous function $P : \ X \mapsto \mathbb{R}$ such that (i) $P(\xi) > 0$ for $\xi \in X \setminus S$, and (ii) for each $\xi \in I(S)$, there is $t > 0$ such that

$$\liminf_{\eta \in X\setminus S,\, \eta \to \xi} \frac{P(\eta\pi t)}{P(\eta)} > 1.$$

Whenever such a function exists, the system is persistent (cf. [13]). This method has been applied to several Lotka–Volterra-type ecological models in [5, 6, 7, 13]. Sufficient conditions for the systems to be persistent are obtained by constructing various such functions. It is the latter direction that we shall pursue mainly in this paper, although our approach is closely related to the former.

Generally, a difficulty in using average Liapunov functions arises from its "global" nature. To construct such a function, all the behavior of the semiflow on S must be taken into consideration simultaneously. Thus the larger the maximal invariant set $I(S)$ of S is, the more difficult it is to construct the function. When the semiflow has a complicated structure on S, the task of construction becomes overwhelming. Hence, it appears that the method would be much improved if average Liapunov functions can be used in a "local" sense, that is, not just one single global average Liapunov function but a number of local functions, one for each member of certain decomposition of $I(S)$, so that collectively, they ensure the persistence of the semiflow. The goal of this paper is to develop such a method. Let an average Liapunov function be defined only in a neighborhood of an isolated invariant set in S, i.e., an invariant set in S which is maximal in a neighborhood. We show that its existence ensures that the invariant set does not attract any trajectory from $X \setminus S$. Since the smaller the invariant set is, the easier the average Liapunov function can be constructed, the question becomes to determine the smallest invariant sets such that if each does not attract trajectories from $X \setminus S$, then neither does the entire face S. We show that the smallest of such invariant sets are the isolated connected chain-recurrent sets in S. Recall that an invariant set M is called chain recurrent if for each $x \in M$ and each $\varepsilon > 0$, there exist points $x = x_0, \ldots, x_n = x$ and times $t_1, \ldots, t_n \geq 1$ such that $d(x_i, x_{i-1}\pi t_i) < \varepsilon$ [8]. Our results show that if the system is not persistent, then any trajectory starting in $X \setminus S$ which is attracted to S is actually attracted to a chain-recurrent set in S. As a result, to show that the system is persistent, one needs to construct an average Liapunov function for each connected component of chain-recurrent sets in $I(S)$.

The paper is organized as follows. In section 2, we investigate the behavior of a trajectory of a general nonpersistent semiflow when it is attracted to S and show that the trajectory is actually attracted to a chain-recurrent set. In the case where the chain-recurrent set contains more than one equilibria, the trajectory will enter and exit any small neighborhood of each point of the set infinitely many times, with progressively slower pace. In section 3, we first show that for any isolated invariant set in S, the existence of an average Liapunov function in a neighborhood of the invariant set ensures that the invariant set does not attract any trajectory from $X \setminus S$. We then discuss the reaction–diffusion system (1.1) and show how the local average Liapunov functions can be constructed. As a result, we give sufficient conditions for isolated invariant sets in S not to attract trajectories from $X \setminus S$. The final section illustrates the techniques with a food chain model of Lotka–Volterra type.

2. Behavior of trajectories attracted to S. In this section, we consider nonpersistent systems and show that any trajectory attracted from $X \setminus S$ to S is necessarily attracted to a connected chain-recurrent set in S. As a result, we obtain sufficient conditions for the system to be persistent by requiring that each isolated connected chain-recurrent set in S repels trajectories in $X \setminus S$. We also give sufficient conditions for persistence in terms of Morse decompositions of the maximal invariant set $I(S)$ in S. Some other property of a trajectory when it is attracted to S in a nonpersistent system will also be given. The results in this section are valid for general dynamical systems.

Consider a semiflow π in the metric space (X, d). Let $S \subset X$ be a closed set such that both S and $X \setminus S$ are forwardly invariant for π. Let also $T(t) : X \mapsto X$ denote the semigroup defined by $T(t)x = x\pi t$ for $x \in X$ and $t \in \mathbb{R}^+$. We impose the following general assumptions:

(A) (i) The semiflow π is point dissipative in X [10].

(ii) There is a $t_0 > 0$ such that the semigroup $T(t)$ is compact in X for $t \geq t_0$.

In view of Theorem 2.2 of [11], the above condition (A) ensures that the semiflow π restricted to S has a compact nonempty global attractor $I(S)$ in S. We further assume the following:

(B) The maximal invariant set $I(S)$ in S is an isolated invariant set for π in X.

Let $M \subset S$ be an invariant set for π. We consider the situation when there is $x \in X \setminus M$ attracted to M. By x being attracted to (resp., repelled by) M we mean the ω-limit set $\omega(x)$ (resp., the α-limit set $\alpha(x)$) is nonempty and is enclosed in M (resp., $\alpha(x) \subset M$). Recall that $\omega(x)$ (resp., $\alpha(x)$) is the set of all $y \in X$ for which there is a sequence $t_n \to \infty$ (resp., $t \to -\infty$) as $n \to \infty$ such that $x\pi t_n \to y$. The set of all x attracted to (resp., repelled by) M is called the *stable* (resp., *unstable*) *set* of M and is denoted as $W^s(M)$ (resp., $W^u(M)$) (cf. [10]). For each $x \in X$, we let $\gamma^+(x) \equiv x\pi[0, T_{\max})$ and $\gamma^-(x) \equiv x\pi(T_{\min}, 0]$ denote the forward and backward semi-orbits, respectively. Here $T_{\max}$ (resp., $T_{\min}$) is the supreme of $t > 0$ (resp., infimum of $t \leq 0$) such that $x\pi t$ exists. One notes that for semiflows generated by the PDE (1.1), $\gamma^-(x)$ may not exist or may not be unique. Finally, let N be an open set of X; we say π does not explode in N if, whenever $x \in X$ and $x\pi[0, T_{\max}) \subset N$, $T_{\max} = \infty$. The next result from Rybakowski [15] plays an important role in our discussion.

PROPOSITION 2.1. *Let M be an isolated invariant set and N be a closed isolating neighborhood of M such that π does not explode in N. Let $x \in X$ be such that $\gamma^+(x)$ is bounded, $\omega(x) \cap M \neq \emptyset$, and $\omega(x) \setminus M \neq \emptyset$. Then there exist points x^s, $x^u \in \partial N \cap \omega(x)$ such that $\gamma^+(x^s) \subset N$ and all backward orbits $\gamma^-(x^u)$ through x^u are contained in N.*

Remark. By the invariance of ω-limit sets, through any point $y \in \omega(x)$ a full orbit $\gamma(y) \equiv y\pi\mathbb{R}$ exists. Hence by the compactness of $\omega(x)$, both $\omega(y)$ and $\alpha(y)$ are nonempty. Since, by definition, M is the maximal invariant set in N, it is clear that in the above proposition, $\omega(y^s) \subset M$ and $\alpha(y^u) \subset M$. That is, $y^s \in W^s(M)$ and $y^u \in W^u(M)$.

Using this proposition, we can show that if $x \in X \setminus S$ has an ω-limit point in S, then either x itself or some $y \in \omega(x) \setminus S$ is attracted to S.

COROLLARY 2.2. *Suppose conditions* (A) *and* (B) *hold. If $x \in X \setminus S$ has a bounded forward semi-orbit $\gamma^+(x)$, and $\omega(x) \cap S \neq \emptyset$, then either $\omega(x) \subset S$, or there is $y \in \omega(x) \setminus S$ such that $\omega(y) \subset S$.*

Proof. Since, by assumption, $\gamma^+(x)$ is bounded and $T(t)$ is compact, $\omega(x)$ is nonempty and contains only full orbits of the semiflow π. Hence $\omega(x) \cap S \subset I(S)$. Suppose $\omega(x) \not\subset S$. Then by assumption (B) and Proposition 2.1, for each isolating neighborhood N of $I(S)$, there is $y \in \omega(x) \cap \partial N$ such that the forward semi-orbit $\gamma^+(y) \subset N$. This implies that $\omega(y) \subset I(S)$ and there is a full orbit passing through y. Hence we must have $y \notin S$, because otherwise, by the invariance of S, the full orbit $\gamma(y)$ must be contained in S, and hence $y \in I(S)$, contradicting $y \in \partial N$. This proves the existence of $y \in \omega(x) \setminus S$ such that $\omega(y) \subset S$. □

The above result can be improved if we replace S by any isolated invariant set $M \subset S$, which is a repeller in S. Recall that an invariant set $M \subset S$ is called an

attractor if it attracts all elements in a neighborhood of itself, and an invariant set $M' \subset S$ is called a repeller dual to M if

$$M' = \{x \in I(S) : \omega(x) \cap M = \emptyset\}.$$

For an isolated invariant set M, an ordered finite collection of subsets $D = \{M_1, \ldots, M_n\}$ in M is called a Morse decomposition of M if each member M_i is invariant and if each $x \in M$ either lies in a set M_k or has its $\omega(x)$ and $\alpha(x)$ enclosed in two distinct sets M_i and M_j, respectively, with $i < j$. The members of D are called Morse sets (cf. [8]). The next theorem gives a result more general than Corollary 2.2. It shows that if a repeller M has the Morse decomposition $\{M_1, \ldots, M_n\}$, and if $x \in X \setminus S$ is such that $\omega(x) \cap M \neq \emptyset$, then either x or some $y \in \omega(x) \setminus S$ must be attracted to one of the Morse sets.

THEOREM 2.3. *Suppose conditions* (A) *and* (B) *hold. Let $M \subset S$ be an isolated invariant set of π which is a repeller in S (dual to an attractor). Suppose $\{M_1, \ldots, M_n\}$ is a Morse decomposition of M in S, and suppose that there is an $x \in X \setminus S$ having a bounded forward semi-orbit $\gamma^+(x)$ and $\omega(x) \cap M \neq \emptyset$. Then there is a Morse set M_i such that either x or some $y \in \omega(x) \setminus S$ is contained in $W^s(M_i)$.*

Proof. Assume by contradiction that such a Morse set does not exist. Let $z \in \omega(x) \cap M$. Then there is a Morse set M_k such that $\omega(z) \subset M_k$, and by the invariance of ω-limit sets, $\omega(x) \cap M_k \neq \emptyset$. Let j be the maximum of the subscripts k such that $\omega(x) \cap M_k \neq \emptyset$. Since, by assumption, $x \notin W^s(M_j)$, Proposition 2.1 shows that there is $y \in \omega(x)$ such that $y \notin M_j$ and $y \in W^s(M_j)$. Again by assumption, $y \notin X \setminus S$. Since $y \in \omega(x)$ implies that there is a full orbit passing through y, it follows that this orbit lies in S. Hence either $\alpha(y) \not\subset M$ or there is $i > j$ such that $\alpha(y) \subset M_i$. The former contradicts the fact that M is a repeller. The latter implies that $\omega(x) \cap M_i \supset \alpha(y) \cap M_i \neq \emptyset$, which contradicts the maximality of j. This completes the proof. □

Remark. Using a similar argument, one can show that the result of the theorem holds if $\omega(x) \subset M$ while M is not necessarily a repeller. The modification of the proof is straightforward.

It is known that given two Morse decompositions D_1, D_2 of an isolated invariant set M one can construct a third Morse decomposition D_3 finer than both D_1 and D_2. Let $M(D_i)$ denote the union of all the Morse sets in the decomposition D_i. Then necessarily, $M(D_3) \subset M(D_1) \cap M(D_2)$. It has been shown by Conley [8] that the intersection of the unions of Morse sets over all Morse decompositions is a chain recurrent set. This suggests that if there is $x \in X \setminus S$ which is attracted to S, then x is actually attracted to a chain recurrent set in S. We show this result in the next theorem.

THEOREM 2.4. *Suppose* (A) *and* (B) *hold and suppose $x \in X \setminus S$ has a bounded forward semi-orbit $\gamma^+(x)$ and satisfies $\omega(x) \cap S \neq \emptyset$. Then there is an invariant set $C \subset S$ which is connected and chain recurrent for π in S such that either x or some $y \in \omega(x) \cap (X \setminus S)$ is contained in $W^s(C)$.*

Proof. In view of Corollary 2.2, there exists $x^s \in X \setminus S$ such that $\omega(x^s) \subset S$, and x^s is either x itself or lies in $\omega(x)$. Let $R(S)$ be the maximal chain recurrent set of the semiflow π restricted in $I(S)$. According to [8], in every neighborhood of $R(S)$, there is a Morse decomposition D in S such that the union $M(D)$ of the Morse sets is in that neighborhood. Choose a decreasing sequence $\epsilon_n \to \infty$ and a corresponding sequence of Morse decompositions $\{D^n\}$ such that $M(D^n)$ lies in the ϵ_n-neighborhood of $R(S)$. By using intersections of Morse sets, we may assume without loss of generality that D^n becomes finer as n increases.

We select a decreasing sequence of Morse sets $M^n \in D^n$ and a precompact sequence of points $x_n \in X \setminus S$ such that $\omega(x_n) \subset M^n$ as follows. Let N be a fixed isolating neighborhood of $I(S)$ in X. By either reducing N or following the trajectory from x^s if necessary, we may assume $x^s \in \partial N$. Without loss of generality, we may also assume that $M(D^1) \subset N$. Applying Theorem 2.3 to x^s and D^1, we see that there is an $x_1 \in X \setminus S$ such that $\omega(x_1) \in M^1$ for some $M^1 \in D^1$, and x_1 is either x^s or lies in $\omega(x^s)$. We show that x_1 can be chosen on ∂N. Indeed, if $x_1 = x^s$, we already have $x_1 \in \partial N$. If $x_1 \in \omega(x^s)$, then by the invariance of ω-limit sets, there is a full orbit $\gamma(x_1)$ passing through x_1. Since $x_1 \notin S$ and since N is an isolating neighborhood of $I(S)$, it follows that $\gamma(x_1) \not\subset N$. Using the fact that $\omega(x_1) \subset M^1 \subset I(S)$, one sees that there exists $x_1' \in \gamma(x_1)$ such that $x_1' \in \partial N$ and $\omega(x_1') \in M^1$. We can then choose x_1' to replace x_1.

Next, assume that there exist Morse sets $M^k \in D^k$ and points $x_k \in \partial N \setminus S$, $k = 1, \ldots, n$, such that

$$M^1 \supset M^2 \supset \cdots \supset M^n$$

and $\omega(x_k) \subset M^k$. Assume also that each x_k is either x^s or lies in $\omega(x^s)$. It is easy to see that the intersections of M^n with the Morse sets of D^{n+1} constitute a Morse decomposition of M^n. Hence, in view of the remark following the proof of Theorem 2.3, there exist a Morse set $M^{n+1} \in D^{n+1}$ and a point $x_{n+1} \in X \setminus S$ such that $M^{n+1} \subset M^n$ and $\omega(x_{n+1}) \subset M^{n+1}$. Furthermore, x_{n+1} is either the same as x_n or lies in $\omega(x_n)$. By the transitivity of ω-limit sets, we see that x_{n+1} is either the same as x^s or lies in $\omega(x^s)$. Finally, using a similar argument in the preceding paragraph, x_{n+1} can be chosen to lie on ∂N. This completes the induction, and also the construction of sequences $\{M^n\}$ and $\{x_n\}$.

Let $C \equiv \bigcap_{n=1}^{\infty} M^n$. Then clearly, C is invariant and chain recurrent for π in S. By the compactness of $\omega(x^s)$, we see that there exists a convergent subsequence $\{x_{n'}\}$ of $\{x_n\}$ such that

$$\lim_{n' \to \infty} x_{n'} = y \in \partial N.$$

We show that $\omega(y) \subset C$ and y is either x itself or lies in $\omega(x) \setminus S$. First observe that by the construction of $\{x_n\}$, for each fixed n' and each $k \geq n'$, x_k is either $x_{n'}$ itself or lies in $\omega(x_{n'})$. Hence by the closedness of ω-limit sets, the same is true for y. This implies that $\omega(y) \subset M^{n'}$ for all n', which leads to $\omega(y) \subset C$. Next, since each x_n is either x^s itself or lies in $\omega(x^s)$, and x^s is either x itself or lies in $\omega(x)$, by the transitivity, the same is true for y. In the case where $y \in \omega(x)$, there is a full orbit passing through y. Since $y \in \partial N$ implies that $y \notin I(S)$, it follows that $y \notin S$. Finally, since $\omega(x)$ is connected, in case C is not, we may simply replace C by a component containing $\omega(x)$. This completes the proof. □

The above Theorems 2.3 and 2.4 lead to the following results for the persistence of the semiflow.

COROLLARY 2.5. *Let conditions* (A) *and* (B) *hold. Then the semiflow π is persistent if either* (i) *$I(S)$ has a Morse decomposition $\{M_1, \ldots, M_n\}$ such that $W^s(M_i) \cap (X \setminus S) = \emptyset$ for $i = 1, \ldots, n$, or* (ii) *each component $C \subset I(S)$ of the chain-recurrent set of π in S satisfies $W^s(C) \cap (X \setminus S) = \emptyset$.*

Proof. Suppose by contradiction that π is not persistent. Then there is $x \in X \setminus S$ such that $\gamma^+(x)$ is bounded and $\omega(x) \cap S \neq \emptyset$. In view of Corollary 2.2, we may assume $\omega(x) \subset S$ without loss of generality. Suppose $I(S)$ has a Morse decomposition $\{M_1, \ldots, M_n\}$. Then since $I(S)$ is a repeller in S (dual to $\emptyset$), Theorem 2.3 with

$M = I(S)$ asserts that there is a Morse set M_i attracting trajectories from $X \setminus S$. Hence condition (i) does not hold. Also, since by Theorem 2.4 there is at least one connected chain-recurrent set of π in S attracting trajectories from $X \setminus S$, we see that condition (ii) does not hold either. This completes the proof. □

By Theorem 2.4, if any $x \in X \setminus S$ is attracted to S it is actually attracted to a chain-recurrent set $C \subset S$ of the semiflow. In the case when $\omega(x)$ contains more than one equilibrium point, we can show that the trajectory through x will visit any neighborhood of each point of $\omega(x)$ infinitely many times, with increasingly slower pace, provided that each equilibrium itself does not attract any trajectory from $X \setminus S$. More generally, we have the following theorem.

THEOREM 2.6. *Suppose* (A) *and* (B) *hold and suppose $C \subset S$ is a connected chain-recurrent set of π in S such that for each equilibrium $e \in C$, $W^s(e) \cap (X \setminus S) = \emptyset$. Let $x \in W^s(C) \cap (X \setminus S)$, and let $y \in \omega(x)$. Suppose there exists an equilibrium $y' \neq y$ such that $y' \in \omega(x)$. Then for each $T > 0$ there exist constants $\varepsilon > 0$ and $T_3 > T_2 > T_1 > T_0 > 0$ such that $x\pi T_0 \in N_\varepsilon(y)$, $x\pi T_3 \in N_\varepsilon(y)$, $x\pi t \notin N_\varepsilon(y)$ for $t \in [T_1, T_2]$, and $T_2 - T_1 \geq T$.*

Proof. Let $\{t_n\}$ and $\{t'_n\}$ be sequences tending to ∞ such that $x\pi t_n \to y$, $x\pi t'_n \to y'$, where y' is an equilibrium. Choose ε sufficiently small so that the ε-neighborhoods $N_\varepsilon(y)$ and $N_\varepsilon(y')$ of y and y' are disjoint. Then for large n, $x\pi t_n \in N_\varepsilon(y)$ and $x\pi t'_n \in N_\varepsilon(y')$. By taking subsequences and relabeling them, we may assume that for all n, $t_n < t'_n < t_{n+1}$. Define

$$s_n = \sup\{t : \ t'_n < t < t_{n+1}, \ x\pi[t'_n, t) \subset N_\epsilon(y')\}.$$

It follows that $t_n < t'_n < s_n < t_{n+1}$, $x\pi[t'_n, s_n] \subset \overline{N_\varepsilon(y')}$ for all n, and $x\pi s_n \in \partial N_\varepsilon(y')$. In particular, $x\pi t \notin N_\varepsilon(y)$ for $t \in [t'_n, s_n]$. We assert that $s_n - t'_n \to \infty$ as $n \to \infty$. For if not, by choosing subsequences, we may assume without loss of generality that $s_n - t'_n \to c \in \mathbb{R}^+$ and $x\pi s_n \to y'' \in \partial N_\varepsilon(y')$ as $n \to \infty$. However, since y' is an equilibrium, it follows that $x\pi s_n = x\pi t'_n \pi(s_n - t'_n) \to y'\pi c = y'$. This contradicts $y'' \in \partial N_\varepsilon(y')$.

Now, choose n large so that $s_n - t'_n > T$. We see that for $T_0 = t_n$, $T_1 = t'_n$, $T_2 = s_n$ and $T_3 = t_{n+1}$, the conclusion of the theorem holds. □

3. Local average Liapunov functions. In view of Theorems 2.4 and 2.6, to determine the persistence of a semiflow π, we need to examine a family of subsets of $I(S)$ and show that each member of the family does not attract trajectories from $X \setminus S$. The family can be either a Morse decomposition of $I(S)$ or the collection of connected chain-recurrent sets of π in S. In this section, we show how the existence of an average Liapunov function defined in a neighborhood of an isolated invariant set in S ensures that the invariant set does not attract trajectories from $X \setminus S$. Our approach in this section follows that in Hutson [13].

Let $M \subset S$ be an isolated invariant set for π and let N be an isolating neighborhood of M in X. We say that a continuous function $P : \ N \mapsto \mathbb{R}$ is a (local) average Liapunov function in N with respect to S if

(i) $P(x) > 0$ for $x \in N \setminus S$,

(ii) for each $x \in M$ there is $t > 0$ such that $\liminf_{y \in N \setminus S; \ y \to x} P(y\pi t)/P(y) > 0$.

Define a function $Q(x, t)$ for $x \in N$ and $t \geq 0$ by

$$Q(x,t) = \begin{cases} P(x\pi t)/P(x) & x \in N \setminus S, \\ \liminf\limits_{y \in N \setminus S; \ y \to x} P(y\pi t)/P(y) & x \in N \cap S. \end{cases} \tag{3.1}$$

Then part (ii) of the definition of P is equivalent to saying that for each $x \in M$ there is $t > 0$ such that $Q(x,t) > 0$. It is easy to see that $Q(\cdot,t)$ is lower semicontinuous. This is obvious if $x \in N \setminus S$. Suppose $x \in N \cap S$. Let $\varepsilon > 0$. Then there is $\delta > 0$ such that

$$Q(y,t) \equiv P(y\pi t)/P(y) > Q(x,t) - \varepsilon$$

for all $y \in N \setminus S$, $d(x,y) < \delta$. Hence for any $z \in N \cap S$ with $d(z,x) < \delta$, we have $Q(z,t) \geq Q(x,t) - \varepsilon$. This proves the semicontinuity of $Q(\cdot,t)$ at x.

In the next theorem, we use functions P and Q to give a sufficient condition for M not to attract any trajectory from $X \setminus S$.

THEOREM 3.1. *Let the condition* (A)(ii) *hold. Suppose $M \subset S$ is an isolated invariant set of π in X, and suppose that there exists an average Liapunov function P defined in a neighborhood $N \supset M$ with respect to S. If for each $x \in M$ there is $t > 0$ such that $Q(x,t) > 1$, then $W^s(M) \cap (X \setminus S) = \emptyset$.*

Proof. Since the function $Q(x,t)$ is lower semicontinuous, it follows that for each $h > 0$ and $t > 0$, the set

$$U(h,t) = \{x \in N^\circ : \ Q(x,t) > h\} \tag{3.2}$$

is open. (Here N° is the interior of N in X.) Clearly $U(h,t)$ is monotone for any fixed t, i.e., $U(h_1,t) \supset U(h_2,t)$ if $h_1 < h_2$. Using the assumption that for each $x \in M$ there is $t > 0$ such that $Q(x,t) > 1$, we see that M is covered by the family $\{U(h,t) : \ h > 1, \ t > 0\}$. Since M is compact by the assumption (A), there exist $h_0 > 1$ and $t_1, \ldots, t_n > 0$ such that

$$M \subset U \equiv \bigcup_{i=1}^{n} U(h_0, t_i).$$

It is clear that $U \subset N$ and is open in X.

Suppose now by contradiction that there is $y \in X \setminus S$ such that $\omega(y) \subset M$. Then there exists $T > 0$ such that $y\pi t \in U$ for all $t \geq T$. Use a translation in t if necessary; we may assume that $y \in U$. Let $\bar{t} = \max(t_1, \ldots, t_n)$ and $\underline{t} = \min(t_1, \ldots, t_n)$. Then given any $t \geq 0$, there is $\underline{t} \leq \tau \leq \bar{t}$ such that

$$P(y\pi(t+\tau)) \geq h_0 P(y\pi t).$$

Repeated use of this inequality and the fact that $P(y) > 0$ shows that there is a sequence $\{\tau_n\} \to \infty$ such that $P(y\pi\tau_n) \to \infty$. However, since $\omega(y) \subset M$, by choosing a subsequence if necessary, we may assume that $\{y\pi\tau_n\}$ converges to a point in M. Hence by the continuity of P, the set $\{P(y\pi\tau_n)\}$ is bounded. This contradiction shows that such y does not exist. □

The conditions in Theorem 3.1 can be relaxed as follows. Let M_ω be the closure of the set $\bigcup_{x \in M} \omega(x)$. Then instead of requiring $Q(x,t) > 1$ for each $x \in M$ and a corresponding $t > 0$, it suffices to require that the inequality be satisfied in M_ω only. This is shown in the next theorem.

THEOREM 3.2. *The conclusion of Theorem* 3.1 *remains true if there exists an average Liapunov function P defined in a neighborhood $N \supset M$ with respect to S such that for each $x \in M_\omega$, there is $t > 0$ for which $Q(x,t) > 1$.*

Proof. We first show that for any $x \in M$ and any $t_0 > 0$ there is a $\tau > t_0$ such that $Q(x,\tau) > 0$. Let $U(h,t)$ be the open set defined in (3.2). Since P is an average

Liapunov function, it follows from part (ii) of the definition that for each $x \in M$ there is $t > 0$ such that $Q(x,t) > 0$. Hence M is covered by the family of open sets $\{U(h,t) : \ h > 0, t > 0\}$, and by the compactness of M, there exist $h_0 > 0$ and $t_1, \dots, t_n > 0$ such that

$$M \subset U \equiv \bigcup_{i=1}^{n} U(h_0, t_i) \subset N.$$

Clearly U is a neighborhood of M in X. Let $\bar{t} = \max(t_1, \dots, t_n)$ and $\underline{t} = \min(t_1, \dots, t_n)$. We choose an integer k such that $k\underline{t} \geq t_0$ and a neighborhood $N(x)$ such that $N(x)\pi[0, k\bar{t}] \subset U$. Then for any $y \in N(x) \setminus S$, there exist $0 = \tau_0 < \tau_1 < \cdots < \tau_k \leq k\bar{t}$ such that $\underline{t} \leq \tau_i \leq \bar{t}$ and $P(y\pi\tau_i) > h_0 P(y\pi\tau_{i-1})$ for $i = 1, \dots, k$. This implies that $P(y\pi\tau_k) > h_0^k P(y)$ for all $y \in N(x) \setminus S$. Hence $Q(x, \tau_k) \geq h_0^k > 0$. Since $\tau_k \geq k\underline{t} \geq t_0$, the assertion is proven with $\tau = \tau_k$.

We now show that the condition of Theorem 3.2 implies the condition of Theorem 3.1. That is, we show that for each $x \in M$ there is a $t > 0$ such that $Q(x,t) > 1$. It is clear that M_ω is a closed subset of the compact set M. Since by the assumption of the theorem M_ω is covered by $\bigcup_{h>1,t>0} U(h,t)$, it follows that there is a finite covering

$$M_\omega \subset U_1 \equiv \bigcup_{i=1}^{m} U(h_1, t_i),$$

where $h_1 > 1$ and $t_1, \dots, t_m$ are positive constants. Thus U_1 is an open neighborhood of M_ω in X. Let $t' = \max(t_1, \dots, t_m)$. Suppose $x \in M$. Then there exist t_0 and τ such that $t_0 \leq \tau$, $x\pi t \in U_1$ for all $t \geq t_0$, and $Q(x,\tau) = \delta > 0$. Choose an integer ν such that $h_1^\nu \delta/2 > 1$, and choose a neighborhood $N(x)$ of x such that $N(x)\pi[\tau, \tau + \nu t'] \subset U_1$ and

$$P(y\pi\tau)/P(y) > \delta/2 \qquad \text{for } y \in N(x). \tag{3.3}$$

Then for each $y \in N(x)$, we have $z \equiv y\pi\tau \in U_1$ and $z\pi[0, \nu t'] \subset U_1$. This implies that there exist $0 = \tau_0 < \tau_1 < \cdots < \tau_\nu \leq \nu t'$ such that $\tau_i \leq t'$ and $P(z\pi\tau_i) > h_1 P(z\pi\tau_{i-1})$ for $i = 1, \dots, \nu$. Therefore, by (3.3),

$$P(y\pi(\tau + \tau_\nu)) = P(z\pi\tau_\nu) > h_1^\nu P(z) = h_1^\nu P(y\pi\tau) > h_1^\nu (\delta/2) P(y).$$

This leads to $Q(x,t) \geq h_1^\nu \delta/2 > 1$ for $t = \tau + \tau_\nu$. The assertion is proven.

The conclusion of the theorem follows now from Theorem 3.1. □

In the remainder of this section, we consider the system (1.1) of parabolic partial differential equations. We present certain techniques of constructing local average Liapunov functions and use them to obtain conditions for an invariant set not to attract trajectories from $X \setminus S$. In the following, X is the positive cone of the product space $\prod_{i=1}^{N} X_i$ where $X_i = C_0(\Omega)$ if B_i is the Dirichlet condition and

$$X_i = \{u \in C^1(\overline{\Omega}) : \ B_i u = 0\}$$

if B_i is Neumann or Robin condition, and $S \equiv \bigcup_{i=1}^{N} S_i$ is the union of the faces $S_i \equiv \{u \in X : \ u_i \equiv 0\}$. It is clear from the uniqueness of the solution of (1.1) that each face S_i is forwardly invariant for the semiflow π generated by (1.1). Hence S is also forwardly invariant.

We first use eigenfunctions of adjoint operators to construct average Liapunov functions. For $i = 1, \dots, N$, let L_i^* and B_i^* be the adjoint operators corresponding to L_i and B_i given by (1.2) and (1.3), i.e.,

$$L_i^* u = \sum_{i,j=1}^{n} (a_{ij} u_{x_j})_{x_i} - \sum_{i=1}^{n} b_i u_{x_i} - u \sum_{i=1}^{n} (b_i)_{x_i},$$
$$B_i^* u = \delta_i \frac{\partial u(x)}{\partial \nu^*(x)} + \left(\delta_i \sum_{i=1}^{n} b_i \zeta_i + \beta \right) u,$$

where $\nu^* = (\nu_1^*, \dots, \nu_n^*)$ is the adjoint conormal vector at $x \in \partial\Omega$ with $\nu_j^* = \sum_{i=1}^{n} a_{ij} \zeta_i$. A simple computation shows that

$$\int_\Omega v L_i u \, dx = \int_\Omega u L_i^* v \, dx \quad \text{if } Bu = B^* v = 0. \tag{3.4}$$

Let ϕ be a positive function that satisfies the relation

$$L_i^* \phi = g(x)\phi \quad (x \in \Omega), \qquad B_i^* \phi = 0 \quad (x \in \partial\Omega) \tag{3.5}$$

for some function g. Define the function $P : X \mapsto \mathbb{R}$ by

$$P(\xi) = \int_\Omega \xi_i \phi \, dx. \tag{3.6}$$

Then $P(\xi) > 0$ for all $\xi \in X \setminus S_i$ and

$$\begin{aligned} \frac{P(\xi \pi t)}{P(\xi)} &= \exp\left(\ln \int_\Omega \phi u_i(x,t) \, dx - \ln \int_\Omega \phi u_i(x,0) \, dx \right) \\ &= \exp\left(\int_0^t \frac{\int_\Omega \phi u_{i,t}(x,\tau) \, dx}{\int_\Omega \phi u_i(x,\tau) \, dx} \, d\tau \right), \end{aligned} \tag{3.7}$$

where $u = (u_1, \dots, u_n)$ is the solution of (1.1) with $u(x,0) = \xi$ and $u_{i,t}$ is the partial derivative of u_i with respect to t. Since by (1.1) and (3.4),

$$\begin{aligned} \int_\Omega \phi u_{i,t} \, dx &= \int_\Omega \phi (L_i u_i + u_i f_i(u)) \, dx \\ &= \int_\Omega \phi u_i (g(x) + f_i(u(x,t))) \, dx \\ &\geq c_0 \int_\Omega \phi u_i \, dx, \end{aligned} \tag{3.8}$$

where $c_0 = \inf_{x \in \Omega} \{ g(x) + f_i(u(x,t)) \}$, it follows from the definition of $Q(x,t)$ in (3.1) that $Q(\xi, t) \geq \exp(c_0) > 0$ for all $\xi \in X$ and $t > 0$. This shows that P is a local average Liapunov function for π with respect to S_i. It is clear that for any constant $\alpha > 0$, the function P^α is also a local average Liapunov function with respect to S_i. Using this construction and Theorem 3.1 we obtain the following result.

THEOREM 3.3. *Let $M \subset S_i$ be an isolated invariant set of π. Suppose the semiflow defined by* (1.1) *satisfies the condition* (A)(ii), *and suppose that there exists a function $\phi > 0$ that satisfies the relation in* (3.5) *for some function g such that $g(x) > -f_i(\xi)$ for all $\xi \in M$ and $x \in \overline{\Omega}$. Then $W^s(M) \cap (X \setminus S_i) = \emptyset$.*

Proof. Let P be defined by (3.6). As it is shown above, P is a local average Liapunov function of π with respect to S_i. Let Q be the corresponding semicontinuous function defined by (3.1). Then by (3.5) and (3.7),

$$Q(\xi,t) = \liminf_{\eta\in X\setminus S_i;\,\eta\to\xi} P(\eta\pi t)/P(\eta) \geq e^{c_0 t} \qquad \text{for } \xi\in M,\ t>0,$$

where $c_0 = \inf_{x\in\Omega,\,\xi\in M}\{g(x)+f_i(\xi)\}$. Since by the assumption of the theorem $c_0>0$, it follows that $Q(\xi,t)>1$. The conclusion of the theorem now follows from Theorem 3.1. □

Remark. Since by the Fredholm alternative and the Krein–Rutman theorem problem (3.5) has a positive solution if and only if the problem

$$L_i\psi = g(x)\psi \quad (x\in\Omega), \qquad B_i\psi = 0 \quad (x\in\partial\Omega) \tag{3.9}$$

has a positive solution, the conclusion of Theorem 3.3 is valid if equation (3.5) is replaced by (3.9).

In the case when $M\equiv\{u^0\}\subset S_i$ is a singleton, the attractivity of u^0 to $X\setminus S$ is described by the eigenvalue problem:

$$L_i\psi + f_i(u^0)\psi = \lambda\psi \quad (x\in\Omega), \qquad B_i\psi = 0 \quad (x\in\partial\Omega). \tag{3.10}$$

As a special case of Theorem 3.3, we have the following result.

COROLLARY 3.4. *Let $u^0\in S$ be an isolated equilibrium of π which also constitutes an isolated invariant set of π in X. Then $W^s(u^0)\cap(X\setminus S)=\emptyset$ if problem* (3.10) *has a positive eigenvalue for some $i\in\{1,\dots,N\}$ such that $u^0\in S_i$.*

Proof. Suppose problem (3.10) has a positive solution ψ for some $\lambda>0$. Then the function $g(x)$ in (3.9) satisfies

$$g(x) = -f(u^0)+\lambda > -f(u^0).$$

Hence by Theorem 3.3, the conclusion follows. □

More complicated average Liapunov functions can be constructed as follows. Suppose P_1 and P_2 are two functions defined by

$$P_1(\xi) = \int_\Omega \xi_{i_1}\phi_1\,dx, \quad P_2(\xi) = \int_\Omega \xi_{i_2}\phi_2\,dx, \tag{3.11}$$

where $i_1,\ i_2\in\{1,\dots,n\}$ and $\phi_1,\ \phi_2$ satisfy the relation

$$\begin{aligned} L_{i_1}^*\phi_1 &= g_1(x)\phi_1 \quad (x\in\Omega), \qquad B_{i_1}^*\phi_1 = 0 \quad (x\in\partial\Omega),\\ L_{i_2}^*\phi_2 &= g_2(x)\phi_2 \quad (x\in\Omega), \qquad B_{i_2}^*\phi_2 = 0 \quad (x\in\partial\Omega), \end{aligned}$$

for some functions g_1 and g_2.

THEOREM 3.5. *Let P_1 and P_2 be defined by* (3.11) *and let α_1, α_2, δ_1, δ_2 be positive constants. Then* (i) *the function $P_1^{\alpha_1}P_2^{\alpha_2}$ is a local average Liapunov function of π with respect to $S_{i_1}\cup S_{i_2}$, and* (ii) *the function $\delta_1P_1^{\alpha_1}+\delta_2P_2^{\alpha_2}$ is a local average Liapunov function of π with respect to $S_{i_1}\cap S_{i_2}$.*

Proof. (i) Let $P=P_1^{\alpha_1}P_2^{\alpha_2}$. Then clearly, $P(\xi)>0$ for $\xi\notin S_{i_1}\cup S_{i_2}$. Let Q be the semicontinuous functions defined by (3.1) and let Q_1 and Q_2 be defined similarly with P replaced by P_1 and P_2, respectively. Then by taking limit-infimum, $Q\geq Q_1^{\alpha_1}Q_2^{\alpha_2}$. Since $Q_1(\xi,t)>0$ and $Q_2(\xi,t)>0$ for all $\xi\in K$ and $t>0$, the same is true for Q. This proves that P is a local average Liapunov function with respect to $S_{i_1}\cup S_{i_2}$.

(ii) Let $P = \delta_1 P_1^{\alpha_1} + \delta_2 P_2^{\alpha_2}$ and let $\xi \in K \setminus (S_{i_1} \cap S_{i_2})$. Suppose $u(t) = \xi\pi t$ for $t \geq 0$. Then, by computation,

$$P(\xi\pi t)/P(\xi) = \exp\left(\ln(P(\xi\pi t)) - \ln(P(\xi))\right) = \exp\left(\ln(P(u(t)) - \ln(P(u(0)))\right)$$
$$= \exp\left(\int_0^t \dot{P}(u(\tau))/P(u(\tau))\, d\tau\right),$$

where $\dot{P}$ is the derivative of $P(u(t))$ with respect to t along the trajectory. Using the definition of P_1 and P_2 in (3.11), we compute

$$\dot{P}(u(t)) = \delta_1\alpha_1 \int_\Omega \phi_1 u_{i_1,t}\, dx \left(\int_\Omega \phi_1 u_{i_1}\, dx\right)^{\alpha_1 - 1}$$
$$+ \delta_2\alpha_2 \int_\Omega \phi_2 u_{i_2,t}\, dx \left(\int_\Omega \phi_2 u_{i_2}\, dx\right)^{\alpha_2 - 1},$$

where $u_{i_1,t}$ and $u_{i_2,t}$ are the respective partial derivatives of u_{i_1} and u_{i_2} with respect to t. Following the derivation in (3.8), we have

$$\int_\Omega \phi_k u_{i_k,t}\, dx \geq c_k \int_\Omega \phi_k u_{i_k}\, dx$$

with the constants

$$c_k = \inf_{x\in\Omega}\{g_k(x) + f_{i_k}(u(x,t))\}, \qquad k = 1, 2.$$

Thus, by letting $c_0 = \min\{\alpha_1 c_1, \alpha_2 c_2\}$, it follows that

$$\dot{P}(u(t)) \geq \delta_1\alpha_1 c_1 \left(\int_\Omega \phi_1 u_{i_1}\, dx\right)^{\alpha_1} + \delta_2\alpha_2 c_2 \left(\int_\Omega \phi_2 u_{i_2}\, dx\right)^{\alpha_2} \geq c_0 P(u(t)).$$

Hence by dividing $P(u(t))$ and taking limit-infimum, we arrive at

$$Q(\xi, t) \geq e^{c_0 t} > 0 \qquad \text{for all } t > 0.$$

This proves the part (ii) of the theorem. □

In the case where $I_\omega(S)$, the closure of the set $\cup_{x\in S}\omega(x)$ is a finite set, we give a condition for the system to be persistent. It is clear that in this case, each member u^j of $I_\omega(S)$ is an equilibrium of the system, and there is $i \in \{1, \ldots, N\}$ such that $u^j \in S_i$. An equilibrium u^i is said to be chained to another equilibrium u^j if there exists a connecting orbit from u^i to u^j. A finite sequence $u^{i_1}, \ldots, u^{i_k}$ is called a cycle if $u^{i_1} = u^{i_k}$ and u^{i_j} is chained to $u^{i_{j+1}}$ for $j = 1, \ldots, k$. We show that the system is uniformly persistent if the following conditions are satisfied:

(C) (i) The set $I_\omega(S) = \{u^j\}$ contains no cycle.
(ii) At each equilibrium u^j, there is $i \in \{1, \ldots, N\}$ such that $u^j \in S_i$ and problem (3.10) has a positive eigenvalue.

THEOREM 3.6. *Suppose the system* (1.1) *satisfies conditions* (A) *and* (B) *and suppose that* $I_\omega(S) = \{u^j\}$ *is a finite set which satisfies condition* (C). *Then the system defined by* (1.1) *is uniformly persistent.*

Proof. Since $I_\omega(S)$ is a finite set and contains no cycle, its members constitute a Morse decomposition D for the invariant set $I(S)$. Rearranging the subscripts if

necessary, we may assume that $D = \{M_1, \dots, M_n\}$ with $M_j = \{u^j\}$ for $j = 1, \dots, n$. In view of Theorem 2.3, the semiflow π is uniformly persistent with respect to S if $W^s(u^j) \cap (X \setminus S) = \emptyset$ for each $j \in \{1, \dots, n\}$. This condition is ensured by the assumption (C)(ii) and Corollary 3.4. □

Finally, we comment that for the semiflow generated by the parabolic system (1.1), the condition (A) is the consequence of the following conditions:

(D) (i) *Uniform boundedness.* Given $\alpha > 0$, there exists a constant $B(\alpha)$ such that $||u^0||_\infty \le \alpha$ implies that the solution $u(x,t)$ of (1.1) satisfies $||u(\cdot,t)||_\infty < B(\alpha)$ for all $t > 0$.

(ii) *Dissipativity in L^∞.* There exists a positive constant γ such that for each $u^0 \in X$, there is a $t_0 > 0$ such that

$$||u(\cdot,t)||_\infty \le \gamma \qquad \text{for all } t \ge t_0.$$

Here $||\cdot||_\infty$ is the L^∞ norm. This result is essentially Theorem 3.3 in Cantrell, Cosner, and Hutson, [6], although in [6] the differential operators L_i are restricted to the type $\mu_i \Delta$, where μ_i are positive constants and Δ is the Laplacian, and the boundary conditions are either the homogeneous Dirichlet type or the homogeneous Neumann type. The proof is valid for the more general (1.1). Furthermore, condition (D) can often be deduced by the comparison principle and the method of upper and lower solutions, as the example in the next section shows.

4. A food chain model. In this final section, we discuss a Lotka–Volterra food chain model of three species as an illustration of the results obtained in the previous sections. For the sake of simplicity, we assume that the diffusion coefficients for all three species are identical and the interacting mechanisms are also identical. The analysis for the general case is similar, but lengthy. Our model has the form

$$\begin{aligned} &\partial u_1/\partial t - d\Delta u_1 = u_1(a - u_1 - bu_2 + cu_3) \\ &\partial u_2/\partial t - d\Delta u_2 = u_2(a + cu_1 - u_2 - bu_3) \qquad (x \in \Omega), \\ &\partial u_3/\partial t - d\Delta u_3 = u_3(a - bu_1 + cu_2 - u_3) \\ &\qquad Bu_1 = Bu_2 = Bu_3 = 0 \qquad (x \in \partial\Omega), \end{aligned} \tag{4.1}$$

where a, b, c and d are positive constants, $\Omega \subset \mathbb{R}^n$ is a bounded domain with a smooth boundary $\partial\Omega \in C^{1+\alpha}$ for some $\alpha > 0$, and

$$B = \delta \frac{\partial}{\partial \nu} + \beta(x),$$

where δ is either 0 or 1, $\beta \in C^{1+\delta}(\partial\Omega)$ is a nonnegative function such that $\beta \equiv 1$ if $\delta = 0$, and ν is the outward normal vector on $\partial\Omega$. We consider two cases. In the first case, the system has semitrivial solutions with two positive components. Hence, when one species is absent, the other two may coexist. In the second case, the system has no semitrivial solution with two positive components. Thus if one species is absent, one of the others cannot survive. We will see that in this case, the three semitrivial solutions (each has only one positive component) form a cycle. For each case, we use local average Liapunov functions to obtain conditions for the system to be persistent.

Our basic assumption for this section is

$$a > \lambda_0, \quad 0 < c < 1, \tag{4.2}$$

where λ_0 is the principal eigenvalue of the problem

$$-d\Delta\psi = \lambda\psi \quad (x \in \Omega), \qquad B\psi = 0 \quad (x \in \partial\Omega).$$

The next lemma gives the compactness and dissipativity of the semiflow.

LEMMA 4.1. *Suppose a, c, and d are positive constants, b is a nonnegative constant, and $c < 1$. Then the semiflow π generated by the system* (4.1) *satisfies condition* (A).

Proof. In view of the remark at the end of the preceding section, it suffices to verify condition (D). Let α be a positive constant, and let $u(x,t)$ be a solution of (4.1) with $||u^0||_\infty \le \alpha$. We choose the constant $M = \max\{\alpha, a/(1-c)\}$ and let $\mathbf{y} \equiv (y_1, y_2, y_3)$ be the solution of the system

$$\begin{aligned} y_1' &= y_1(a - y_1 + cy_3), \\ y_2' &= y_2(a + cy_1 - y_2), \\ y_3' &= y_3(a + cy_2 - y_3), \\ y_1(0) &= y_2(0) = y_3(0) = M. \end{aligned} \tag{4.3}$$

It is easy to see that the functions $\mathbf{y}$ and $\mathbf{0} \equiv (0,0,0)$ form a pair of ordered upper and lower solutions for the problem (4.1) (cf. [16, Chapter 8]). Hence

$$0 \le u_i(x,t) \le y_i(t) \qquad \text{for } (x,t) \in \Omega \times (0,\infty),\ i = 1,2,3. \tag{4.4}$$

On the other hand, due to the symmetry of (4.3), y_1, y_2, y_3 are identical and they solve the scalar problem

$$y' = y(a - (1-c)y), \quad y(0) = M.$$

Since by assumption $c < 1$ and $M \ge a/(1-c)$, it follows that y_i's are monotone nonincreasing in t and have the limit $a/(1-c)$. Thus by (4.4), $||u(\cdot,t)||_\infty \le M$ for all $t > 0$, and

$$\limsup_{t\to\infty} ||u(\cdot,t)||_\infty \le a/(1-c).$$

This proves the fulfillment of condition (D), and hence (A). □

We next identify the equilibria on the boundary of the positive cone. The system clearly has a trivial solution $\mathbf{0}$. A spectral analysis shows that it is unstable. Since by assumption $a > \lambda_0$, it is well known (cf. [2]) that the problem

$$-d\Delta\phi = \phi(a - \phi) \quad (x \in \Omega), \qquad B\phi = 0 \quad (x \in \partial\Omega) \tag{4.5}$$

has a unique positive solution ϕ. Hence problem (4.1) has three semitrivial solutions $\Phi_1 \equiv (\phi, 0, 0)$, $\Phi_2 \equiv (0, \phi, 0)$, and $\Phi_3 \equiv (0, 0, \phi)$. A spectral analysis shows that all are unstable. To determine the existence of a semitrivial solution with two positive components, we consider the two-component system

$$\begin{aligned} \partial u/\partial t - d\Delta u &= u(a - u - bv) \\ \partial v/\partial t - d\Delta v &= v(a + cu - v) \\ Bu = Bv &= 0 \end{aligned} \qquad \begin{aligned} &(x \in \Omega), \\ &\\ &(x \in \partial\Omega). \end{aligned} \tag{4.6}$$

Let $\lambda_0(p)$ denote the principal eigenvalue of the problem

$$-d\Delta\psi + p\psi = \lambda\psi \quad (x \in \Omega), \qquad B\psi = 0 \quad (x \in \partial\Omega), \tag{4.7}$$

where p is a continuous function. It is easy to see that whenever problem (4.6) has a positive steady-state solution (u_s, v_s), then $a > \lambda_0(bv_s)$ (see [2]). Since the principal eigenvalue $\lambda_0(p)$ is monotone increasing in p (which can be easily seen from the variational formulation of the principal eigenvalue), it follows that a necessary condition for the existence of a positive solution is $a > \lambda_0(0) \equiv \lambda_0$. The next lemma shows that another necessary condition is $b \leq 1$.

LEMMA 4.2. *Suppose $a > \lambda_0$ and $b > 1$. Then*

(i) *every solution (u, v) with $u(x, 0) \geq 0$, $u(x, 0) \not\equiv 0$, and $v(x, 0) \equiv 0$ converges to $(\phi, 0)$ as $t \to \infty$;*

(ii) *every solution (u, v) with $u(x, 0) \geq 0$, $v(x, 0) \geq 0$, and $v(x, 0) \not\equiv 0$ converges to $(0, \phi)$ as $t \to \infty$.*

In each case, the convergence is uniform for $x \in \overline{\Omega}$.

Proof. Consider the scalar problem

$$\partial z/\partial t - d\Delta z = z(a - p - z) \quad (x \in \Omega), \qquad Bz = 0 \quad (x \in \partial\Omega), \tag{4.8}$$

where p is a smooth function. It is well known that if $a \leq \lambda_0(p)$, then all nonnegative solutions converge uniformly to 0, and if $a > \lambda_0(p)$, then each solution $z(x, t)$ with $z(x, 0) \geq 0$, $z(x, 0) \not\equiv 0$ converges uniformly to the unique positive steady-state solution z_s (cf. e.g., [16, Chapter 8]). Part (i) of the lemma is a consequence of this result with $p = 0$ and $z_s = \phi$. We prove part (ii) of the lemma as follows.

We first show that the steady-state solution $(0, \phi)$ of (4.6) is asymptotically stable. The eigenvalue problem with respect to $(0, \phi)$ is

$$\begin{aligned} d\Delta\xi + (a - b\phi)\xi &= \lambda\xi \\ d\Delta\eta + c\phi\xi + (a - 2\phi)\eta &= \lambda\eta \\ B\xi = B\eta &= 0 \end{aligned} \qquad \begin{aligned} &(x \in \Omega), \\ &\\ &(x \in \partial\Omega). \end{aligned} \tag{4.9}$$

Since ϕ is a positive solution of (4.7) with $\lambda = a$ and $p = \phi$, it follows that $a = \lambda_0(\phi)$. Thus by the monotonicity of $\lambda_0(\cdot)$ and the assumption $b > 1$, we have $a < \lambda_0(b\phi)$ and $a < \lambda_0(2\phi)$. This implies that the first equation has a nontrivial solution only for some $\lambda < 0$, and so does the second equation with $\xi = 0$ to have a nontrivial solution. Hence all the eigenvalues of (4.9) are negative, which implies that $(0, \phi)$ is asymptotically stable.

Let ω_0 be the ω-limit set of the trajectory starting with $(u(\cdot, 0), v(\cdot, 0))$. It is well known for the Lotka–Volterra system (4.6) that any trajectory with a nonnegative initial condition has a connected, compact ω-limit set in the positive cone of the Banach space X^2, where $X = C_0(\Omega)$ if B is of Dirichlet type and $X = C^1(\overline{\Omega})$ if B is of Neumann or Robin type. To show that the solution $(u(\cdot, t), v(\cdot, t))$ converges to $(0, \phi)$, it suffices to show that $(0, \phi) \in \omega_0$. Because this would ensure that the trajectory will enter any small neighborhood of $(0, \phi)$ in finite time. Thus by the asymptotic stability, the trajectory must be attracted to $(0, \phi)$.

We first show that for each $(u^*, v^*) \in \omega_0$ the inequalities

$$u^*(x) \leq \phi(x), \quad v^*(x) \geq \phi(x) \qquad (x \in \Omega) \tag{4.10}$$

hold. To see this, let $\tilde{u}(x, t)$ and $\hat{v}(x, t)$ be solutions of (4.8) with $p = 0$ such that $\tilde{u}(x, 0) = u(x, 0)$ and $\hat{v}(x, 0) = v(x, 0)$. Then since u and v are nonnegative, it follows from the comparison principle (cf. e.g., [16, Theorem 2.3 of Chapter 4]) that $u(x, t) \leq \tilde{u}(x, t)$ and $v(x, t) \geq \hat{v}(x, t)$ for all $x \in \Omega$, $t > 0$. Since by the assumption

of the lemma $\lambda_0(0) < a < \lambda_0(b\phi)$ and $v(x,0) \not\equiv 0$, we have $\lim_{t\to\infty} \tilde{u}(\cdot,t) \le \phi$, and $\lim_{t\to\infty} \hat{v}(\cdot,t) = \phi$ uniformly in $\overline{\Omega}$. This proves (4.10).

Next, we let $(u^*(x,t), v^*(x,t))$ be a solution of (4.6) with $(u^*(\cdot,0), v^*(\cdot,0)) \in \omega_0$ and show that $(u^*(\cdot,t), v^*(\cdot,t)) \to (0,\phi)$ as $t \to \infty$. Once this is shown, then by the invariance of ω-limit sets, it is necessary that $(0,\phi) \in \omega_0$, which completes the proof of part (ii) of the lemma. Let $\tilde{u}(x,t)$ be a solution of (4.8) with $p = b\phi$ and $\tilde{u}(x,0) = u^*(x,0)$. Since by invariance $(u^*(\cdot,t), v^*(\cdot,t)) \in \omega_0$ for all $t > 0$, it follows from (4.10) that

$$u^*(x,t) \le \phi(x), \quad v^*(x,t) \ge \phi(x) \qquad \text{for all } x \in \Omega,\ t \ge 0. \tag{4.11}$$

Hence, by the comparison principle, $\tilde{u}(x,t) \ge u^*(x,t)$ for all $x \in \Omega$ and $t > 0$. Using the relation $a < \lambda_0(b\phi)$ we see that $\tilde{u}(\cdot,t) \to 0$ as $t \to \infty$, uniformly in $\overline{\Omega}$. Hence $u^*(\cdot,t) \to 0$. To find the limit of $v^*(\cdot,t)$, we observe that for any $\epsilon > 0$, there is $T > 0$ such that $cu^*(x,t) \le \varepsilon$ for all $x \in \Omega$ and $t \ge T$. Let $\hat{v}(x,t)$ be a solution of (4.8) with $p = -\varepsilon$ and $\tilde{v}(x,0) = v^*(x,T)$. Then by the comparison principle, $\tilde{v}(\cdot,t-T) \ge v^*(\cdot,t)$ for $t \ge T$. Since by the monotonicity of the principal eigenvalue $\lambda_0(\cdot)$,

$$a > \lambda_0(0) > \lambda_0(-\varepsilon),$$

it follows that $\tilde{v}(\cdot,t) \to \phi_\varepsilon$ as $t \to \infty$, where ϕ_ε is the positive solution of problem (4.5) with a replaced by $a + \varepsilon$. This leads to $v^*(x,t) \le \phi_\varepsilon(x)$ for t sufficiently large. Moreover, by the continuity of the solution of (4.5) with respect to a, $\phi_\varepsilon \to \phi$ as $\varepsilon \to 0^+$. Hence, we have

$$\limsup_{t\to\infty} v^*(x,t) \le \phi(x) \qquad \text{for all } x \in \Omega.$$

This, together with (4.11), shows that $v^*(\cdot,t) \to \phi$. The proof is complete. □

Remark. It can be shown using the degree theory that a sufficient condition for problem (4.6) to have a positive solution is $a > \lambda_0$ and $b < 1$.

Case 1*: when there exist semitrivial solutions with two positive components.* We show below that if $a > \lambda_0$, $c < 1$, and $b(1+c) < 1$ then the system is persistent. To prove this, we first find a priori bounds for the invariant sets of the system in S. This is done by using upper and lower solutions. Let $(\tilde{u}, \tilde{v})$ and $(\hat{u}, \hat{v})$ be smooth functions satisfying the inequalities

$$\tilde{u} \ge \hat{u} \ge 0, \qquad \tilde{v} \ge \hat{v} \ge 0,$$

and

$$\begin{aligned} -d\Delta\tilde{u} &\ge \tilde{u}(a - \tilde{u} - b\hat{v}), \quad -d\Delta\tilde{v} \ge \tilde{v}(a + c\tilde{u} - \tilde{v}), \\ -d\Delta\hat{u} &\le \hat{u}(a - \hat{u} - b\tilde{v}), \quad -d\Delta\hat{v} \le \hat{v}(a + c\hat{u} - \hat{v}), \qquad (x \in \Omega) \\ B\tilde{u} &\ge 0 \ge B\hat{u}, \quad B\tilde{v} \ge 0 \ge B\hat{v} \qquad (x \in \partial\Omega). \end{aligned}$$

It is well known that any solution (u,v) of (4.6) with the initial condition

$$\hat{u}(x) \le u(x,0) \le \tilde{u}(x), \quad \hat{v}(x) \le v(x,0) \le \tilde{v}(x) \qquad (x \in \Omega)$$

satisfies

$$\hat{u}(x) \le u(x,t) \le \tilde{u}(x), \quad \hat{v}(x) \le v(x,t) \le \tilde{v}(x) \qquad (x \in \Omega,\ t > 0).$$

Furthermore, there exist functions $(\overline{u}, \overline{v})$ and $(\underline{u}, \underline{v})$ satisfying the relations

$$\hat{u} \le \underline{u} \le \overline{u} \le \tilde{u}, \quad \hat{v} \le \underline{v} \le \overline{v} \le \tilde{v} \qquad \text{in } \Omega$$

such that

$$\begin{aligned} &\underline{u}(x) \le \liminf_{t\to\infty} u(x,t) \le \limsup_{t\to\infty} u(x,t) \le \overline{u}(x), \\ &\underline{v}(x) \le \liminf_{t\to\infty} v(x,t) \le \limsup_{t\to\infty} v(x,t) \le \overline{v}(x) \end{aligned} \qquad (x \in \Omega).$$

In fact, $(\overline{u}, \overline{v})$ and $(\underline{u}, \underline{v})$ form a "quasi-solution" of problem (4.6) in the sense that they satisfy

$$\begin{aligned} -d\Delta\overline{u} &= \overline{u}(a - \overline{u} - b\underline{v}), \quad -d\Delta\overline{v} = \overline{v}(a + c\overline{u} - \overline{v}), && (x \in \Omega) \\ -d\Delta\underline{u} &= \underline{u}(a - \underline{u} - b\overline{v}), \quad -d\Delta\underline{v} = \underline{v}(a + c\underline{u} - \underline{v}), && \\ B\overline{u} &= B\underline{u} = 0, \quad B\overline{v} = B\underline{v} = 0 && (x \in \partial\Omega) \end{aligned} \tag{4.12}$$

(see Theorem 4.2 of Chapter 8 in [16]). The next lemma gives upper and lower bounds of these functions.

LEMMA 4.3. *Suppose $a > \lambda_0$, $b(1+c) < 1$, $\hat{u} \not\equiv 0$, and $\hat{v} \not\equiv 0$. Then the inequalities*

$$(1 - b - bc)\phi \le \underline{u} \le \overline{u} \le \phi, \quad \phi \le \underline{v} \le \overline{v} \le (1 + c)\phi$$

hold for $x \in \Omega$.

Proof. Since by assumption $\hat{u}$ and $\hat{v}$ are nonnegative functions and none of them is identically zero, it follows from the maximum principle that $\underline{u}$ and $\underline{v}$ are both positive functions in Ω. Using the comparison principle to equations (4.5) and (4.12), we see that $\underline{v} \ge \phi$ and $\overline{u} \le \phi$. On the other hand, since the function $\xi = (1 + \mu)\phi$ satisfies

$$\begin{aligned} -d\Delta\xi &= \xi(a + \mu\phi - \xi) && (x \in \Omega), \\ B\xi &= 0 && (x \in \partial\Omega) \end{aligned}$$

for any $\mu \in \mathbb{R}$, it follows again from the comparison principle that $\overline{v} \le (1 + c)\phi$. Finally, since $\underline{u} > 0$ in Ω, by comparing the above equation with (4.12), we have $\underline{u} \ge (1 - b(1 + c))\phi$. This proves the lemma. □

Lemma 4.3 and Theorems 2.4 and 3.3 lead to the following persistence result.

THEOREM 4.4. *Suppose $a > \lambda_0$, $c < 1$, and $b(1 + c) < 1$. Then system (4.1) is uniformly persistent.*

Proof. Let M_1 be the maximal invariant set of the semiflow π in the region

$$\{u_1 = 0,\ (1 - b - bc)\phi \le u_2 \le \phi,\ \phi \le u_3 \le (1 + c)\phi\} \subset S_1.$$

We show that $W^s(M_1) \cap (X \setminus S) = \emptyset$. First observe that by assumption, $1 - b - bc > 0$; we have $M_1 \cap (S_2 \cup S_3) = \emptyset$. Thus the relation is equivalent to $W^s(M_1) \cap (X \setminus S_1) = \emptyset$. In view of Theorem 3.3, we need only to find a function $g(x)$ with the property that (1) $g(x) > -a + bu_2 - cu_3$ in Ω for all u_2 and u_3 such that $(0, u_2, u_3) \in M_1$, and (2) the problem

$$d\Delta\psi = g(x)\psi \quad (x \in \Omega), \qquad B\psi = 0 \quad (x \in \partial\Omega)$$

has a positive solution ψ. We choose g as follows. Since by assumption $b(1 + c) < 1$, it follows that $b < 1 < 1 + c$. Hence by the monotonicity of $\lambda_0(\cdot)$,

$$a = \lambda_0(\phi) > \lambda_0((b - c)\phi).$$

This ensures that for some sufficiently small ϵ, there is a positive solution ψ to the scalar problem

$$-d\Delta\psi = \psi(a - \varepsilon - (b-c)\phi - \psi) \quad (x \in \Omega), \qquad B\psi = 0 \quad (x \in \partial\Omega).$$

Choose $g = -a + \epsilon + (b-c)\phi + \psi$. Then

$$g(x) > -a + bu_2 - cu_3 \quad \text{if } u_2 \le \phi \text{ and } u_3 \ge \phi.$$

In particular, this holds if $(0, u_2, u_3) \in M_1$. The assertion $W^s(M_1) \cap (X \setminus S) = \emptyset$ thus follows from Theorem 3.3.

Define $M_2 \subset S_2$ and $M_3 \subset S_3$ analogously as the maximal invariant sets in the subregions

$$\{\phi \le u_1 \le (1+c)\phi,\ u_2 = 0,\ (1-b-bc)\phi \le u_3 \le \phi\}$$

and

$$\{(1-b-bc)\phi \le u_1 \le \phi,\ \phi \le u_2 \le (1+c)\phi,\ u_3 = 0\},$$

respectively. Then a similar proof shows that $W^s(M_i) \cap (X \setminus S) = \emptyset$ for $i = 1, 2$.

To show that the system (4.1) is persistent, we observe that the invariant set $I(S)$ has the Morse decomposition $\{M_1, M_2, M_3, \Phi_1, \Phi_2, \Phi_3, \mathbf{0}\}$. In view of Theorem 2.4, we need only show that each Morse set does not attract trajectories from $X \setminus S$. This has been done for M_i $(i = 1, 2, 3)$ in the preceding paragraphs. To see that the same is true for Φ_i and $\mathbf{0}$, we use Corollary 3.4. Consider Φ_1. It is clear that $\Phi_1 \in S_3$ and the eigenvalue problem (3.10) with $i = 3$ has the form

$$d\Delta\psi + (a - b\phi)\psi = \lambda\psi \quad (x \in \Omega), \qquad B\psi = 0 \quad (x \in \partial\Omega).$$

This problem has a positive solution if and only if $a - \lambda = \lambda_0(b\phi)$. Since $b < 1$, it follows that $a = \lambda_0(\phi) > \lambda_0(b\phi)$. Hence $\lambda > 0$, and by Corollary 3.4, $W^s(\Phi_1) \cap (X \setminus S) = \emptyset$. A similar argument gives the property for Φ_2 and Φ_3.

Finally, since $\mathbf{0} \in S_1$, its corresponding eigenvalue problem has the form

$$d\Delta\psi + a\psi = \lambda\psi \quad (x \in \Omega), \qquad B\psi = 0 \quad (x \in \partial\Omega).$$

Since by assumption $a > \lambda_0$, it certainly has a positive eigenvalue. Hence $W^s(\mathbf{0}) \cap (X \setminus S) = \emptyset$. This completes the proof. □

Case 2: when there exists no semitrivial solution with two positive components. Suppose $a > \lambda_0$ and $b > 1$. In view of Lemma 4.2, the system has no semitrivial solution with two positive components. The phase portrait of the semiflow in S is described in Figure 1. As can be seen, the maximal invariant set $I(S)$ has the Morse decomposition $\{M_1, M_2\}$, where M_1 consists of semitrivial solutions Φ_1, Φ_2, Φ_3, and the connecting orbits, and M_2 consists of the trivial solution $\mathbf{0}$. (Since each Φ_i has a one-dimensional unstable manifold as can be verified by examining the corresponding eigenvalue problem, the connecting orbit between each pair of semitrivial solutions is a one-dimensional manifold.) It is clear that M_1 is a chain-recurrent set in S. We consider the conditions for $W^s(M_i) \cap (X \setminus S) = \emptyset$ for $i = 1, 2$.

First consider $M_2 = \{\mathbf{0}\}$, which is a singleton. By Corollary 3.4, $W^s(M_2) \cap (X \setminus S) = \emptyset$ if for some $i \in \{1, 2, 3\}$, the problem

$$d\Delta\psi + a\psi = \lambda\psi \quad (x \in \Omega), \qquad B\psi = 0 \quad (x \in \partial\Omega)$$

has a positive eigenvalue. This is ensured by the assumption $a > \lambda_0$.

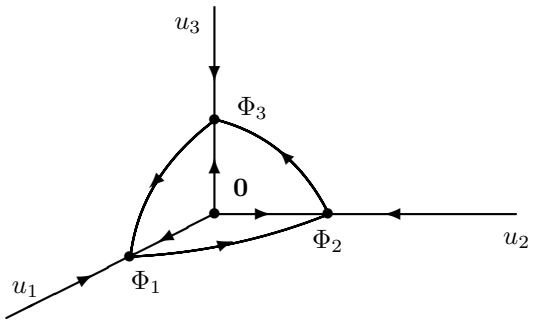

FIG. 1. *The boundary flow with a cycle.*

To obtain conditions for $W^s(M_1) \cap (X \setminus S) = \emptyset$, we construct the function P as follows. Let λ^* be the principal eigenvalue of the problem

$$d\Delta\varphi + (a + c\phi)\varphi = \lambda\varphi \quad (x \in \Omega), \qquad B\varphi = 0 \quad (x \in \partial\Omega). \tag{4.13}$$

Since by assumption and the monotonicity of principal eigenvalue,

$$a > \lambda_0(0) > \lambda_0(-c\phi),$$

it follows that $\lambda^* > 0$. Choose a positive eigenfunction φ and define

$$P(u) \equiv \prod_{i=1}^{3} \int_\Omega \varphi u_i \, dx.$$

In view of Theorem 3.5, P is a local average Liapunov function with respect to S. A direct computation shows that

$$\frac{P(\eta\pi t)}{P(\eta)} = \exp\left[\int_0^t \sum_{i=1}^{3} \frac{\int_\Omega \varphi u_{i,t}(x,\tau)\,dx}{\int_\Omega \varphi u_i(x,\tau)\,dx}\, d\tau\right] \quad \text{for } \eta \in X \setminus S, \tag{4.14}$$

where $u(\cdot,t) \equiv (u_1(\cdot,t), u_2(\cdot,t), u_3(\cdot,t)) = \eta\pi t$ and $u_{i,t} = \partial u_i/\partial t$. Since M_1 has the ω-limit set $M_{1,\omega} = \{\Phi_1, \Phi_2, \Phi_3\}$, it follows from Theorem 3.2 that $W^s(M_1) \cap (X \setminus S) = \emptyset$ provided that for each $j = 1, 2, 3$ there is a $t = t(j) > 0$ such that

$$Q(\Phi_j, t) = \liminf_{\eta \in X\setminus S,\, \eta \to \Phi_j} \frac{P(\eta\pi t)}{P(\eta)} > 1. \tag{4.15}$$

Once this holds, Theorem 2.4 ensures that the system is uniformly persistent. Such condition is given in the following theorem.

THEOREM 4.5. *Suppose* $a > \lambda_0$, $c < 1$, $b > 1$, *and*

$$2\lambda^* > (b + c)||\phi||_{C(\overline{\Omega})}. \tag{4.16}$$

Then the system (4.1) *is uniformly persistent.*

Proof. It suffices to show that for each $j = 1, 2, 3$, there is $t = t(j) > 0$ such that (4.15) holds. We only prove the assertion for $j = 1$. The proof for other values of j is similar. From equations (4.1) and (4.13) we see with the Green's identity that

$$\int_\Omega \varphi u_{1,t}(x,\tau)\,dx = \int_\Omega \varphi u_1(\lambda^* - u_1 - bu_2 - c(\phi - u_3))\,dx. \tag{4.17}$$

Since by the continuity of the solution of (4.1) with respect to the initial value $\lim_{\eta\to\Phi_1} u = \Phi_1$, it follows that the right side of (4.17) tends to $\int_\Omega \varphi\phi(\lambda^* - (c+1)\phi)\,dx$. Using (4.13) and the Green's identity, we have

$$\begin{aligned}\int_\Omega \varphi\phi(\lambda^* - (c+1)\phi)\,dx &= \int_\Omega (d\phi\Delta\varphi + (a+c\phi)\phi\varphi - (c+1)\phi^2\varphi)\,dx \\ &= \int_\Omega \varphi(d\Delta\phi + a\phi - \phi^2)\,dx,\end{aligned}$$

which is 0 by (4.5). Also using (4.13) and the Green's identity, we derive

$$\begin{aligned}\int_\Omega \varphi u_{2,t}(x,\tau)\,dx &= \int_\Omega \varphi u_2(\lambda^* - c(\phi - u_1) - u_2 - bu_3)\,dx, \\ \int_\Omega \varphi u_{3,t}(x,\tau)\,dx &= \int_\Omega \varphi u_3(\lambda^* - bu_1 - c(\phi - u_2) - u_3)\,dx.\end{aligned}$$

By the continuity of the solution u with respect to the initial value η, it follows that for any $\varepsilon > 0$ there is a constant $T > 0$ and a neighborhood $N(\Phi_1)$ such that

$$||c(\phi - u_1) + u_2 + bu_3||_{C(\overline{\Omega})} < \varepsilon, \quad ||bu_1 + c(\phi - u_2) + u_3||_{C(\overline{\Omega})} < ||(b+c)\phi||_{C(\overline{\Omega})} + \varepsilon$$

for $0 \le t \le T$. Hence in this neighborhood $N(\Phi_1)$,

$$\begin{aligned}\int_\Omega \varphi u_{2,t}(x,\tau)\,dx \Big/ \int_\Omega \varphi u_2(x,\tau)\,dx &\ge \lambda^* - \varepsilon, \\ \int_\Omega \varphi u_{3,t}(x,\tau)\,dx \Big/ \int_\Omega \varphi u_3(x,\tau)\,dx &\ge \lambda^* - (b+c)||\phi||_{C(\overline{\Omega})} - \varepsilon.\end{aligned}$$

Thus, using the assumption (4.16), we may choose $N(\Phi_1)$ sufficiently small such that

$$2\lambda^* - (b+c)||\phi||_{C(\overline{\Omega})} > 2\varepsilon.$$

This ensures that

$$\sum_{i=1}^{3} \frac{\int_\Omega \varphi u_{i,t}(x,\tau)\,dx}{\int_\Omega \varphi u_i(x,\tau)\,dx} > 0 \qquad \text{for } \eta \in N(\Phi_1) \text{ and } \tau \in (0,T].$$

Therefore, by (4.14),

$$\liminf_{\eta\in X\setminus S,\, \eta\to\Phi_1} \frac{P(\eta\pi t)}{P(\eta)} > 1 \qquad \text{for } t \in (0,T].$$

The proof is complete. □

Remark. In the case when the boundary operator B is of the Neumann type, we have $\phi = a$ and $\lambda^* = a(1+c)$. Hence condition (4.16) is reduced to $b < c + 2$.

REFERENCES

[1] H. Amann, *Existence and multiplicity theorems for semi-linear elliptic boundary value problems*, Math Z., 150 (1976), pp. 281–295.

[2] H. Berestycki and P. L. Lions, *Some applications of the method of super and subsolutions*, in Bifurcation and Nonlinear Eigenvalue Problems, Lecture Notes in Mathematics 782, Springer-Verlag, New York, 1980, pp. 16–41.

[3] G. Butler, H. I. Freedman, and P. Waltman, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425–430.
[4] G. Butler and P. Waltman, *Persistence in dynamical systems*, J. Differential Equations, 63 (1986), pp. 255–263.
[5] R. S. Cantrell, C. Cosner, and V. Hutson, *Permanence in some diffusive Lotka-Volterra models for three interacting species*, Dynamic System Appl., 2 (1993), pp. 505–530.
[6] R. S. Cantrell, C. Cosner, and V. Hutson, *Permanence in ecological systems with spatial heterogeneity*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 533–559.
[7] R. S. Cantrell, C. Cosner, and V. Hutson, *Ecological models, permanence and spatial heterogeneity*, Rocky Mountain J. Math., 26 (1996), pp. 1–35.
[8] C. Conley, *Isolated Invariant Sets and the Morse Index*, Conf. Board Math. Sci. 38, Amer. Math. Soc., Providence, RI, 1978.
[9] S. R. Dunbar, K. P. Rybakowski, and K. Schmitt, *Persistence in models of predator-prey populations with diffusion*, J. Differential Equations, 65 (1986), pp. 117–138.
[10] J. K. Hale, *Asymptotic Behavior of Dissipative Systems*, Mathematics Surveys and Monographs, Amer. Math. Soc., Providence, RI, 1988.
[11] J. K. Hale and P. Waltman, *Persistence in infinite-dimensional systems*, SIAM J. Math. Anal., 20 (1989), pp. 388–395.
[12] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer, Berlin, 1981.
[13] V. Hutson, *A theorem on average Liapunov functions*, Mh. Math., 98 (1984), pp. 267–275.
[14] V. Hutson and R. Law, *Permanent coexistence in general models of three interacting species*, J. Math. Biol., 21 (1985), pp. 285–298.
[15] K. P. Rybakowski, *The Homotopy Index and Partial Differential Equations*, Springer-Verlag, Berlin, 1987.
[16] C. V. Pao, *Nonlinear Parabolic and Elliptic Equations*, Plenum, New York, 1992.
[17] P. Waltman, *A brief survey of persistence in dynamical systems*, in Delay-Differential Equations and Dynamical Systems, Lecture Notes in Mathematics 1475, S. Burenber and Martelli, eds., Springer, Berlin, 1991, pp. 31–40.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 155–182, January 1998

010

EXPONENTIAL STABILITY OF A THERMOELASTIC SYSTEM WITH FREE BOUNDARY CONDITIONS WITHOUT MECHANICAL DISSIPATION*

GEORGE AVALOS† AND IRENA LASIECKA‡

Abstract. We show herein the uniform stability of a thermoelastic plate model with no added dissipative mechanism on the boundary (uniform stability of a thermoelastic plate with added boundary dissipation was shown in [J. LAGNESE, *Boundary Stabilization of Twin Plates*, SIAM Stud. Appl. Math. 10, SIAM, Philadelphia, PA, 1989], as was that of the analytic case—where rotational forces are neglected—in [Z. LIU and S. ZHENG, *Quarterly Appl. Math.*, 55 (1997), pp. 551–564]). The proof is constructive in the sense that we make use of a multiplier with respect to the coupled system involved so as to generate a fortiori the desired estimates; this multiplier is of an operator theoretic nature, as opposed to the more standard differential quantities used for related work. Moreover, the particular choice of our multiplier becomes clear only after recasting the PDE model into an associated abstract evolution equation.

Key words. thermoelastic plates, uniform stability, free boundary conditions

AMS subject classification. 35

PII. S0036141096300823

1. Introduction.

1.1. Statement of the problem. Let Ω be a bounded open subset of $\mathbb{R}^2$ with sufficiently smooth boundary $\Gamma = \Gamma_0 \cup \Gamma_1$, Γ_0 and Γ_1 both nonempty, and $\overline{\Gamma_0} \cap \overline{\Gamma_1} = \varnothing$. We consider here the following thermoelastic system taken from J. Lagnese's monograph [12]:

$$
\begin{cases}
\begin{cases} \omega_{tt} - \gamma\Delta\omega_{tt} + \Delta^2\omega + \alpha\Delta\theta = 0 \\ \beta\theta_t - \eta\Delta\theta + \sigma\theta - \alpha\Delta\omega_t = 0 \end{cases} & \text{on } (0,\infty)\times\Omega; \\
\omega = \dfrac{\partial\omega}{\partial\nu} = 0 \ \text{ on } (0,\infty)\times\Gamma_0; \\
\begin{cases} \Delta\omega + (1-\mu)B_1\omega + \alpha\theta = 0 \\ \dfrac{\partial\Delta\omega}{\partial\nu} + (1-\mu)\dfrac{\partial B_2\omega}{\partial\tau} - \gamma\dfrac{\partial\omega_{tt}}{\partial\nu} + \alpha\dfrac{\partial\theta}{\partial\nu} = 0 \end{cases} & \text{on } (0,\infty)\times\Gamma_1; \\
\dfrac{\partial\theta}{\partial\nu} + \lambda\theta = 0 \qquad \text{on } (0,\infty)\times\Gamma,\ \lambda \geq 0; \\
\omega(t=0) = \omega^0,\ \omega_t(t=0) = \omega^1,\ \theta(t=0) = \theta^0 \text{ on } \Omega.
\end{cases} \tag{1.1}
$$

*Received by the editors March 22, 1996; accepted for publication September 30, 1996. The research of the second author was partially supported by the NSF grant DMS-9504822 and ARO grant DAAH04-96-1-0059.

http://www.siam.org/journals/sima/29-1/30082.html

†Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55455-0436 (avalos@mercer.me.ttu.edu).

‡Department of Applied Mathematics, Thornton Hall, University of Virginia, Charlottesville, VA 22903 (il2v@virginia.edu).

Here, α, β, and η are strictly positive constants; positive constant γ is proportional to the thickness of the plate and assumed to be small with $0 < \gamma \leq M$; the constant $\sigma \geq 0$ and the boundary operators B_i are given by

$$B_1\omega \equiv 2\nu_1\nu_2\frac{\partial^2\omega}{\partial x\partial y} - \nu_1^2\frac{\partial^2\omega}{\partial y^2} - \nu_2^2\frac{\partial^2\omega}{\partial x^2};$$

(1.2)

$$B_2\omega \equiv (\nu_1^2 - \nu_2^2)\frac{\partial^2\omega}{\partial x\partial y} + \nu_1\nu_2\left(\frac{\partial^2\omega}{\partial y^2} - \frac{\partial^2\omega}{\partial x^2}\right);$$

the constant μ is the familiar Poisson's ratio $\in (0, \frac{1}{2})$, and $[\nu_1, \nu_2]$ denotes the outward unit normal to the boundary. The given model mathematically describes a Kirchoff plate, the displacement of which is represented by the function ω subjected to a thermal damping as quantified by θ. We are concerned here with the uniform stability of solutions $[\omega, \theta]$ to (1.1).

1.2. Preliminaries and abstract formulation. As a departure point for obtaining the proofs of well posedness and of exponential stability, we will consider the system (1.1) as an abstract evolution equation in a certain Hilbert space, for which we introduce the following definitions and notation:

- With $H^k_{\Gamma_0}(\Omega) \equiv \{\omega \in H^k(\Omega) : \frac{\partial^j\omega}{\partial\nu^j}|_{\Gamma_0} = 0 \text{ for } j = 0, ..., k-1\}$, we define $\mathring{\mathbf{A}}$: $L^2(\Omega) \supset D\left(\mathring{\mathbf{A}}\right) \to L^2(\Omega)$ to be $\mathring{\mathbf{A}} = \Delta^2$, with domain

$$D(\mathring{\mathbf{A}}) = \Bigg\{\omega \in H^4(\Omega) \cap H^2_{\Gamma_0}(\Omega) : \Delta\omega + (1-\mu)B_1\omega = 0 \text{ on } \Gamma_1 \text{ and}$$

(1.3)

$$\frac{\partial\Delta\omega}{\partial\upsilon} + (1-\mu)\frac{\partial B_2\omega}{\partial\tau} = 0 \text{ on } \Gamma_1\Bigg\}.$$

- $\mathring{\mathbf{A}}$ is then positive definite, self-adjoint, and consequently from [8] we have the characterizations

$$D(\mathring{\mathbf{A}}^{\frac{1}{4}}) = H^1_{\Gamma_0}(\Omega);$$

(1.4)

$$D(\mathring{\mathbf{A}}^{\frac{1}{2}}) = H^2_{\Gamma_0}(\Omega);$$

$$D(\mathring{\mathbf{A}}^{\frac{3}{4}}) = \left\{\omega \in H^3(\Omega) \cap H^2_{\Gamma_0}(\Omega) : \Delta\omega + (1-\mu)B_1\omega = 0 \text{ on } \Gamma_1\right\}.$$

Moreover, using the Green's formula in [12], we have that for ω, $\widehat{\omega}$ "smooth enough,"

$$\begin{aligned}\int_\Omega (\Delta^2\omega)\widehat{\omega}d\Omega &= a(\omega, \widehat{\omega}) \\ &+ \int_\Gamma \left[\frac{\partial\Delta\omega}{\partial\nu} + (1-\mu)\frac{\partial B_2\omega}{\partial\tau}\right]\widehat{\omega}d\Gamma \\ &- \int_\Gamma \left[\Delta\omega + (1-\mu)B_1\omega\right]\frac{\partial\widehat{\omega}}{\partial\nu}d\Gamma,\end{aligned} \tag{1.5}$$

where $a(\cdot,\cdot)$ is defined by

(1.6)

$$a(\omega, \widehat{\omega}) \equiv \int_\Omega \left[\omega_{xx}\widehat{\omega}_{xx} + \omega_{yy}\widehat{\omega}_{yy} + \mu\left(\omega_{xx}\widehat{\omega}_{yy} + \omega_{yy}\widehat{\omega}_{xx}\right) + 2(1-\mu)\omega_{xy}\widehat{\omega}_{xy}\right]d\Omega.$$

In particular, this formula and the second characterization in (1.4) give that for all ω, $\widehat{\omega} \in D(\mathring{\mathbf{A}}^{\frac{1}{2}})$,

$$\left\langle \mathring{\mathbf{A}}\omega, \widehat{\omega} \right\rangle_{\left[D(\mathring{\mathbf{A}}^{\frac{1}{2}})\right]' \times D(\mathring{\mathbf{A}}^{\frac{1}{2}})} = \left(\mathring{\mathbf{A}}^{\frac{1}{2}}\omega, \mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\omega}\right)_{L^2(\Omega)} = a\left(\omega, \widehat{\omega}\right)_{L^2(\Omega)}, \tag{1.7}$$

and in addition,

$$\|\omega\|^2_{D(\mathring{\mathbf{A}}^{\frac{1}{2}})} = \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} = a\left(\omega, \omega\right). \tag{1.8}$$

- We define $A_D : L^2(\Omega) \supset D\left(A_D\right) \to L^2(\Omega)$ to be $A_D = -\Delta$, with Dirichlet boundary conditions, viz.

$$D(A_D) = H^2(\Omega) \cap H^1_0(\Omega). \tag{1.9}$$

A_D is also positive definite, self-adjoint, and, by [8],

$$D(A_D^{\frac{1}{2}}) = H^1_0(\Omega). \tag{1.10}$$

- The space $L^2_{\sigma+\lambda}(\Omega)$ will be defined as

$$L^2_{\sigma+\lambda}(\Omega) \equiv \begin{cases} L^2(\Omega) & \text{if } \sigma + \lambda > 0, \\ \\ L^2_0(\Omega) & \text{if } \sigma + \lambda = 0, \end{cases} \tag{1.11}$$

where $L^2_0(\Omega) = \left\{\theta \in L^2(\Omega) \ni \int_\Omega \theta = 0\right\}$.

- We designate as $A_R : L^2(\Omega) \supset D\left(A_R\right) \to L^2(\Omega)$ the following second-order elliptic operator:

$$\begin{aligned} A_R &= -\Delta + \frac{\sigma}{\eta}\mathbf{I}, \\ D(A_R) &= \left\{\theta \in H^2(\Omega) : \frac{\partial \theta}{\partial \nu} + \lambda\theta = 0\right\}; \end{aligned} \tag{1.12}$$

A_R is self-adjoint, positive semidefinite on $L^2(\Omega)$, and, once more by [8],

$$D(A_R^{\frac{1}{2}}) = H^1(\Omega). \tag{1.13}$$

When $\lambda = \sigma = 0$, we shall denote the corresponding operator as A_N (instead of as A_R).

Furthermore, as the bilinear form $(\nabla\theta, \nabla\widetilde{\theta})_{L^2(\Omega)}$ is $H^1(\Omega)$-elliptic on $H^1(\Omega) \cap L^2_0(\Omega)$, we can define the norm-inducing inner product on $H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)$ as

$$\left(\theta, \widetilde{\theta}\right)_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} \equiv \left(\nabla\theta, \nabla\widetilde{\theta}\right)_{L^2(\Omega)} + \lambda\left(\theta, \widetilde{\theta}\right)_{L^2(\Gamma)} + \frac{\sigma}{\eta}\left(\theta, \widetilde{\theta}\right)_{L^2(\Omega)}. \tag{1.14}$$

- (γ_0, γ_1) will denote the Sobolev trace maps, which yield for $f \in C^\infty(\overline{\Omega})$

$$\gamma_0 f = f|_\Gamma\,; \gamma_1 f = \left.\frac{\partial f}{\partial \nu}\right|_\Gamma. \tag{1.15}$$

- We define the elliptic operators G_1, G_2, and D as thus:

$$G_1 h = v \Longleftrightarrow \begin{cases} \Delta^2 v = 0 \text{ in } (0,\infty)\times\Omega; \\ v = \dfrac{\partial v}{\partial \nu} = 0 \text{ on } (0,\infty)\times\Gamma_0; \\ \begin{cases} \Delta v + (1-\mu) B_1 v = h \\ \dfrac{\partial \Delta v}{\partial \nu} + (1-\mu)\dfrac{\partial B_2 v}{\partial \tau} = 0 \end{cases} \text{on } (0,\infty)\times\Gamma_1; \end{cases} \tag{1.16}$$

$$G_2 h = v \Longleftrightarrow \begin{cases} \Delta^2 v = 0 \text{ in } (0,\infty)\times\Omega; \\ v = \dfrac{\partial v}{\partial \nu} = 0 \text{ on } (0,\infty)\times\Gamma_0; \\ \begin{cases} \Delta v + (1-\mu) B_1 v = 0 \\ \dfrac{\partial \Delta v}{\partial \nu} + (1-\mu)\dfrac{\partial B_2 v}{\partial \tau} = h \end{cases} \text{on } (0,\infty)\times\Gamma_1; \end{cases} \tag{1.17}$$

$$Dh = v \Longleftrightarrow \begin{cases} \Delta v = 0 \text{ on } (0,\infty)\times\Omega; \\ v|_{\Gamma} = h \text{ on } (0,\infty)\times\Gamma. \end{cases} \tag{1.18}$$

The classic regularity results of [19, p. 152] then provide that for $s \in \mathbb{R}$,

$$\begin{cases} D \in \mathcal{L}\left(H^s(\Gamma), H^{s+\frac{1}{2}}(\Omega)\right); \\ G_1 \in \mathcal{L}\left(H^s(\Gamma_1), H^{s+\frac{5}{2}}(\Omega)\right); \\ G_2 \in \mathcal{L}\left(H^s(\Gamma_1), H^{s+\frac{7}{2}}(\Omega)\right). \end{cases} \tag{1.19}$$

With the operators $\mathring{\mathbf{A}}$ and G_i as defined above, one can readily show with the use of the Green's formula (1.5) that $\forall\ \omega \in D(\mathring{\mathbf{A}}^{\frac{1}{2}})$ the adjoints $G_i^* \mathring{\mathbf{A}} \in \mathcal{L}(D(\mathring{\mathbf{A}}^{\frac{1}{2}}), L^2(\Gamma))$ satisfy, respectively,

$$G_1^* \mathring{\mathbf{A}} \omega = \begin{cases} \left.\dfrac{\partial \omega}{\partial \nu}\right|_{\Gamma_1} & \text{on } (0,\infty)\times\Gamma_1; \\ 0 & \text{on } (0,\infty)\times\Gamma_0; \end{cases}$$

$$G_2^* \mathring{\mathbf{A}} \omega = \begin{cases} -\omega|_{\Gamma_1} & \text{on } (0,\infty)\times\Gamma_1; \\ 0 & \text{on } (0,\infty)\times\Gamma_0. \end{cases} \tag{1.20}$$

- We define the operator P_γ by

$$P_\gamma \equiv \mathbf{I} + \gamma A_N, \tag{1.21}$$

and make the following points:
(i) With the parameter $\gamma > 0$, we define a space $H^1_{\Gamma_0,\gamma}(\Omega)$ equivalent to $H^1_{\Gamma_0}(\Omega)$ with its inner product being

$$(\omega_1, \omega_2)_{H^1_{\Gamma_0,\gamma}(\Omega)} \equiv (\omega_1, \omega_2)_{L^2(\Omega)} + \gamma\,(\nabla\omega_1, \nabla\omega_2)_{L^2(\Omega)} \quad \forall\, \omega_1, \omega_2 \in H^1_{\Gamma_0}(\Omega), \tag{1.22}$$

and with its dual (pivotal with respect to L_2 inner product) denoted as $H^{-1}_{\Gamma_0,\gamma}(\Omega)$. After recalling that $H^1(\Omega) = D(A_N^{1/2})$, two extensions by continuity will then yield that

$$P_\gamma \in \mathcal{L}\left(H^1_{\Gamma_0,\gamma}(\Omega), H^{-1}_{\Gamma_0,\gamma}(\Omega)\right), \text{ with} \tag{1.23}$$

$$\langle P_\gamma \omega_1, \omega_2\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega)\times H^1_{\Gamma_0,\gamma}(\Omega)} = (\omega_1,\omega_2)_{H^1_{\Gamma_0,\gamma}(\Omega)}. \tag{1.24}$$

Furthermore, the obvious $H^1_{\Gamma_0,\gamma}(\Omega)$-ellipticity of P_γ and Lax–Milgram give us that P_γ is boundedly invertible, i.e.,

$$P_\gamma^{-1} \in \mathcal{L}\left(H^{-1}_{\Gamma_0,\gamma}(\Omega), H^1_{\Gamma_0,\gamma}(\Omega)\right); \tag{1.25}$$

and moreover, P_γ being positive definite and self-adjoint as an operator $P_\gamma : L^2(\Omega) \supset D(P_\gamma) \to L^2(\Omega)$, the square root $P_\gamma^{1/2}$ is consequently well defined with $D(P_\gamma^{1/2}) = H^1_{\Gamma_0,\gamma}(\Omega)$ (using the interpolation theorem in [19, p. 10]; it then follows from (1.22) and (1.24) that for ω and $\widehat{\omega} \in H^1_{0,\gamma}(\Omega)$,

$$\left\|P_\gamma^{\frac12}\omega\right\|^2_{L^2(\Omega)} = \|\omega\|^2_{L^2(\Omega)} + \gamma\,\|\nabla\omega\|^2_{L^2(\Omega)} = \|\omega\|^2_{H^1_{\Gamma_0,\gamma}(\Omega)}; \tag{1.26}$$

$$\left(P_\gamma^{\frac12}\omega, P_\gamma^{\frac12}\widehat{\omega}\right)_{L^2(\Omega)} = (\omega,\widehat{\omega})_{H^1_{\Gamma_0,\gamma}(\Omega)}. \tag{1.27}$$

(ii) Finally, inasmuch as Green's formula yields for ω, $\widehat{\omega} \in D(\mathring{\mathbf{A}}^{\frac12})$,

$$\begin{aligned}&\gamma\left\langle\left(\Delta + \mathring{\mathbf{A}}G_2\gamma_1\right)\omega, \widehat{\omega}\right\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega)\times H^1_{\Gamma_0,\gamma}(\Omega)}\\ &= -\gamma\,(\nabla\omega, \nabla\widehat{\omega})_{L^2(\Omega)} + \gamma\left(\frac{\partial\omega}{\partial\nu}, \widehat{\omega}\right)_{L^2(\Gamma_1)} + \gamma\left(\gamma_1\omega, G_2^*\mathring{\mathbf{A}}\widehat{\omega}\right)_{L^2(\Gamma_1)}\\ &= -\gamma\,(\nabla\omega, \nabla\widehat{\omega})_{L^2(\Omega)} = -\gamma\left\langle A_N\omega, \widehat{\omega}\right\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega)\times H^1_{\Gamma_0,\gamma}(\Omega)},\end{aligned} \tag{1.28}$$

after using (1.20). We thus obtain after two extensions by continuity to $H^1_{\Gamma_0,\gamma}(\Omega)$ that

$$P_\gamma = \mathbf{I} - \gamma\left(\Delta + \mathring{\mathbf{A}}G_2\gamma_1\right) \text{ as elements of } \mathcal{L}\left(H^1_{\Gamma_0,\gamma}(\Omega), H^{-1}_{\Gamma_0,\gamma}(\Omega)\right). \tag{1.29}$$

In obtaining the equality above, we have used implicitly the fact that for every $\varpi^* \in H^{-1}_{\Gamma_0,\gamma}(\Omega)$ and $\varpi \in D(\mathring{\mathbf{A}}^{1/2})$,

$$\langle\varpi^*, \varpi\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega)\times H^1_{\Gamma_0,\gamma}(\Omega)} = \langle\varpi^*, \varpi\rangle_{[D(\mathring{\mathbf{A}}^{\frac12})]'\times D(\mathring{\mathbf{A}}^{\frac12})}. \tag{1.30}$$

- We denote the Hilbert space $\mathbf{H}_\gamma$ to be

$$\mathbf{H}_\gamma \equiv D(\mathring{\mathbf{A}}^{\frac12}) \times H^1_{\Gamma_0,\gamma}(\Omega) \times L^2_{\sigma+\lambda}(\Omega), \tag{1.31}$$

with the inner product

$$\begin{aligned}&\left(\begin{bmatrix}\omega_1\\ \omega_2\\ \theta\end{bmatrix}, \begin{bmatrix}\widehat{\omega}_1\\ \widehat{\omega}_2\\ \widehat{\theta}\end{bmatrix}\right)_{\mathbf{H}_\gamma}\\ &= \left(\mathring{\mathbf{A}}^{\frac12}\omega_1, \mathring{\mathbf{A}}^{\frac12}\widehat{\omega}_1\right)_{L^2(\Omega)} + \left(P_\gamma^{\frac12}\omega_2, P_\gamma^{\frac12}\widehat{\omega}_2\right)_{L^2(\Omega)} + \beta\left(\theta,\widehat{\theta}\right)_{L^2(\Omega)}.\end{aligned} \tag{1.32}$$

- With the above definitions, we then set $\mathcal{A}_\gamma : \mathbf{H}_\gamma \supset D(\mathcal{A}_\gamma) \rightarrow \mathbf{H}_\gamma$ to be

(1.33)
$$\mathcal{A}_\gamma \equiv \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & P_\gamma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} 0 & \mathbf{I} & 0 \\ -\mathring{\mathbf{A}} & 0 & (\clubsuit) \\ 0 & -\frac{\alpha}{\beta} A_D(\mathbf{I} - D\gamma_0) & -\frac{\eta}{\beta} A_R \end{pmatrix},$$

$$\text{where } (\clubsuit) \equiv \alpha \left(A_R - \frac{\sigma}{\eta} - \mathring{\mathbf{A}} G_1 \gamma_0 + \lambda \mathring{\mathbf{A}} G_2 \gamma_0 \right),$$

$$\text{with } D(\mathcal{A}_\gamma) = \Big\{ [\omega_1, \omega_2, \theta] \in D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times D(A_R) \cap L^2_{\sigma+\lambda}(\Omega)$$
$$\text{such that } \mathring{\mathbf{A}}\omega_1 + \alpha \mathring{\mathbf{A}} G_1 \gamma_0 \theta - \alpha\lambda \mathring{\mathbf{A}} G_2 \gamma_0 \theta \in H^{-1}_{\Gamma_0,\gamma}(\Omega)$$
$$\text{and } \alpha\, \Delta\omega_2 + \eta \Delta\theta \in L^2_{\sigma+\lambda}(\Omega) \Big\}.$$

If we take the initial data $\left[\omega^0, \omega^1, \theta^0\right]$ to be in $\mathbf{H}_\gamma$, then the coupled system (1.1) becomes the operator theoretic model

(1.34)
$$\frac{d}{dt} \begin{bmatrix} \omega \\ \omega_t \\ \theta \end{bmatrix} = \mathcal{A}_\gamma \begin{bmatrix} \omega \\ \omega_t \\ \theta \end{bmatrix},$$
$$\begin{bmatrix} \omega(0) \\ \omega_t(0) \\ \theta(0) \end{bmatrix} = \begin{bmatrix} \omega^0 \\ \omega^1 \\ \theta^0 \end{bmatrix}.$$

REMARK 1. *For initial data* $\left[\omega^0, \omega^1, \theta^0\right]$ *in* $D(\mathcal{A}_\gamma)$*, the two equations of* (1.1) *may be written pointwise as*

$$P_\gamma \omega_{tt} = -\mathring{\boldsymbol{A}}\omega - \alpha \mathring{\boldsymbol{A}} G_1 \gamma_0 \theta + \alpha\lambda \mathring{\boldsymbol{A}} G_2 \gamma_0 \theta - \alpha \Delta\theta \ \text{ in } H^{-1}_{\Gamma_0,\gamma}(\Omega); \tag{1.35}$$

$$\beta\theta_t = \eta\Delta\theta - \sigma\theta + \alpha\Delta\omega_t \ \text{ in } L^2_{\sigma+\lambda}(\Omega). \tag{1.36}$$

1.3. Previous literature. In recent years, questions related to the controllability and stabilization of thermoelastic plates have drawn considerable attention in the recent past (see [10], [12], [9], [21], [22], and [24]); we shall concentrate here on detailing results of strong and uniform stability related to the present model, that of the two-dimensional Kirchoff plate coupled with the heat equation. This particular model, associated with free boundary conditions, was introduced by J. Lagnese in [12]. In that work, he established the well posedness and exponential stability of (1.1) with γ strictly positive, and with the appropriately chosen feedback mechanisms $[\mathcal{F}_1(\omega_t), \mathcal{F}_2(\omega_t)]$ inserted into the natural boundary conditions of the Kirchoff plate component of the system, viz.

(1.37)
$$\begin{cases} \omega = \dfrac{\partial\omega}{\partial\nu} = 0 \text{ on } (0,\infty) \times \Gamma_0, \\ \\ \Delta\omega + (1-\mu) B_1 \omega + \alpha\theta = \mathcal{F}_1(\omega_t) \text{ on } (0,\infty) \times \Gamma_1, \\ \\ \dfrac{\partial \Delta\omega}{\partial\nu} + (1-\mu)\dfrac{\partial B_2 \omega}{\partial\tau} - \gamma \dfrac{\partial \omega_{tt}}{\partial\nu} + \alpha \dfrac{\partial\theta}{\partial\nu} = \mathcal{F}_2(\omega_t) \text{ on } (0,\infty) \times \Gamma_1; \end{cases}$$

the proof of Lagnese is based on the use of differential multipliers, and it exploits the fact that $\gamma > 0$. Since, from a physical point of view, the thermal effects present should induce some measure of energy dissipation (in fact, one can show the homogeneous system's strong stability by routine methods; see [12, Chap. 7], including the remark at the end of sect. 2.3 on p. 161), a natural question arising in this context is whether the system is actually (uniformly) stable without the boundary feedbacks $\mathcal{F}_1(\omega_t)$, $\mathcal{F}_2(\omega_t)$ in place, i.e., when there are no added mechanical forces. Indeed, in the case $\gamma = 0$ and with different boundary conditions than those in (1.1) imposed upon the system, the answer to the question is in the affirmative and has been provided by several authors. With $\gamma = 0$, J. Kim in [10] showed the uniform stability of (1.1) with the clamped boundary conditions $\omega = \frac{\partial \omega}{\partial \nu} = \theta = 0$ on Γ, as did J. Rivera and R. Racke in [25], who studied the coupled equation with the hinged boundary conditions $\omega = \Delta\omega = \theta = 0$. Also with $\gamma = 0$, Z. Liu and S. Zheng in [21] proved the exponential stability of (1.1) with the boundary conditions

$$\begin{cases} \omega = \dfrac{\partial \omega}{\partial \nu} = 0 \text{ on } (0,\infty) \times \Gamma_0, \\ \\ \omega = \Delta\omega + (1-\mu) B_1 \omega + \alpha\theta = 0 \text{ on } (0,\infty) \times \Gamma_1, \end{cases} \tag{1.38}$$

leaving the case of free boundary conditions as an open question, even in the case $\gamma = 0$. The proof of Liu and Zheng is indirect in the sense that it is based on a contradiction argument applied to the exponential decay stability criterion (due to L. Monauni, R. Nagel, and F. Huang), a criterion essentially dictating the uniform estimate for that part of the resolvent which lies on the imaginary axis. On the other hand, it is now known that the case $\gamma = 0$ is rather special as the corresponding system (at least for certain boundary conditions) generates an *analytic* semigroup (see [20]), a consequence of which will be the exponential stability of the system (recall that the system is strongly stable). Given these results, the question of interest now is whether the given thermoelastic system (without any additional boundary dissipation) is *uniformly* stable in the *nonanalytic* case, viz. $\gamma > 0$, with consequently the elastic part of the system being of hyperbolic character.

A partial answer to the question above was given by the present authors in [2], [3]: with $\gamma > 0$ in (1.1) and the boundary conditions

$$\begin{cases} \omega = (1-\chi) \dfrac{\partial \omega}{\partial \nu} = 0 \\ \\ \chi \left(\Delta\omega + (1-\mu)\, B_1 \omega \right) + \alpha\theta = 0 \end{cases} \qquad \text{on } (0,\infty) \times \Gamma \tag{1.39}$$

replacing the higher order ones for ω which are being considered in this work, where the parameter χ above is either 0 or 1, it is shown that the partial differential equation is uniformly stable with decay estimates which are "robust" with respect to the parameter γ. The proof of this stability result is through an implementation of the multiplier method (see [11] a for treatise of this technique), with an operator theoretic quantity taken as the particular multiplier of choice.

The main goal of the present paper is to provide an affirmative answer to the question of uniform stability of (1.1) with the free boundary conditions in place, again with $\gamma > 0$. The fact that the presence of these higher order boundary conditions greatly complicates the analysis was duly noted in [21], and the arguments employed in that work do not carry over for plates with free boundary conditions, even when $\gamma = 0$.

Like our earlier work in [2], [3], the proof of uniform decay here is "direct," based on pseudodifferential (or operator theoretic) multipliers, in contrast to the contradiction argument supplied in [21] for the case $\gamma = 0$ and clamped boundary conditions. In addition, our direct proof, making use as it does of the multiplier method, carries the advantage of providing explicit estimates of the decay rates. However, an application of the multiplier method alone is *not* enough to obtain the desired inequalities for the equation (1.1) in the case when free boundary conditions are present. Indeed, in proving the stability result (Theorem 1.3 below) we must couple the use of an operator theoretic multiplier with a decomposition of the solution ω into three separate components, and a subsequent and crucial invocation of recently derived trace regularity results to handle each of these in distinct fashion; in particular, we exploit the observation that the time derivative of one of these components (modulo a change of variable) solves a certain wave equation. This scrutiny of boundary traces for the hyperbolic component ω of the dynamics is a sine qua non for obtaining the necessary estimates for uniform decay. Finally, we must emphasize that the acute difficulty of the problem which necessitates the use of microlocally derived trace estimates is owing solely to the specific boundary conditions being considered here and does not appear for other combinations of lower order boundary conditions.

1.4. Statement of the results. We shall begin by giving preliminary results regarding the well posedness of the system (1.1) and the regularity of its solutions.

THEOREM 1.1 (well posedness). *Again with the parameter $\gamma > 0$, $\mathcal{A}_\gamma$, given by* (1.33)*, generates a C_0-semigroup of contractions $\left\{e^{\mathcal{A}_\gamma t}\right\}_{t\geq 0}$ on the energy space $\boldsymbol{H}_\gamma$; therefore for initial data $\left[\omega^0, \omega^1, \theta^0\right]$ in $\boldsymbol{H}_\gamma$, the solution $[\omega, \omega_t, \theta]$ to* (1.34)*, and consequently to* (1.1)*, is given by*

$$\left[\begin{array}{c} \omega \\ \omega_t \\ \theta \end{array}\right] = e^{\mathcal{A}_\gamma(\cdot)} \left[\begin{array}{c} \omega^0 \\ \omega^1 \\ \theta^0 \end{array}\right] \in C([0,T], \boldsymbol{H}_\gamma). \tag{1.40}$$

The following regularity result is needed to justify the computations performed below.

THEOREM 1.2. *For initial data $\left[\omega^0, \omega^1, \theta^0\right] \in D\left(\mathcal{A}_\gamma^2\right)$, we have the following:*

(i) *the solution $[\omega, \omega_t, \theta]$ to* (1.1) *is an element of $C([0,T]; H^4(\Omega) \times H^3(\Omega) \times H^2(\Omega))$.*

(ii) $\omega - \gamma G_2 \gamma_1 \omega_{tt} + \alpha G_1 \gamma_0 \theta - \alpha\lambda G_2 \gamma_0 \theta \in C([0,T];\, D(\boldsymbol{\mathring{A}}))$.

Our main result is as follows.

THEOREM 1.3 (uniform stability). *With $\gamma > 0$, the solution $[\omega, \omega_t, \theta]$ of* (1.1) *decays exponentially; that is to say, there exist constants $\delta > 0$ and $M_\delta \geq 1$ such that for all $t > 0$,*

$$\left\| \left[\begin{array}{c} \omega(t) \\ \omega_t(t) \\ \theta(t) \end{array}\right] \right\|_{\boldsymbol{H}_\gamma} \leq M_\delta e^{-\delta t} \left\| \left[\begin{array}{c} \omega^0 \\ \omega^1 \\ \theta^1 \end{array}\right] \right\|_{\boldsymbol{H}_\gamma}. \tag{1.41}$$

REMARK 2. *The estimates obtained in Theorem* 1.3 *are not uniform with respect to the parameter $\gamma > 0$. Indeed, the arguments used in the proof break down when $\gamma = 0$, and consequently the estimates leading to the statement in Theorem* 1.3 *blow up when $\gamma \to 0$. This is due to technicalities of the proof which rely on the strict hyperbolicity of the model (a property which is lost in the limit case $\gamma = 0$). On the other hand, in the case $\gamma = 0$, it has been recently shown in* [27] *that the thermoelastic*

system with free boundary conditions generates an analytic semigroup. Therefore, a posteriori (recalling the strong stability of the system), we conclude that uniform stability holds true also for the case $\gamma = 0$. However, these estimates cannot be reconstructed as a limiting case of the present problem when $\gamma > 0$. This is unlike the case of other boundary conditions associated with this model (see [3]*).*

2. Proofs. The proofs of well posedness and of regularity (Theorems 1.1 and 1.2) are by now fairly routine (see [12, Chap. 7] for related well posedness/regularity results). However, since these preliminaries are critical for our ultimate end of uniform stability, we provide their concise proofs for the sake of completeness.

2.1. Proof of Theorem 1.1. In establishing the semigroup generation of $\mathcal{A}_\gamma$, we will show that the conditions of the Lumer–Phillips theorem are satisfied; namely, we demonstrate here that $\mathcal{A}_\gamma$ is maximal dissipative.

To show the dissipativity of $\mathcal{A}_\gamma$: for $[\omega_1, \omega_2, \theta] \in D(\mathcal{A}_\gamma)$ we have

(2.1)
$$\left(\mathcal{A}_\gamma \begin{bmatrix} \omega_1 \\ \omega_2 \\ \theta \end{bmatrix}, \begin{bmatrix} \omega_1 \\ \omega_2 \\ \theta \end{bmatrix} \right)_{\mathbf{H}_\gamma}$$

$$= \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_2, \mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1 \right)_{L^2(\Omega)}$$

$$+ \left(P_\gamma^{\frac{1}{2}} P_\gamma^{-1} \left(-\mathring{\mathbf{A}} \omega_1 + \alpha A_R \theta - \frac{\alpha\sigma}{\eta} \theta - \alpha \mathring{\mathbf{A}} G_1 \gamma_0 \theta + \alpha\lambda \mathring{\mathbf{A}} G_2 \gamma_0 \theta \right), P_\gamma^{\frac{1}{2}} \omega_2 \right)_{L^2(\Omega)}$$

$$-\alpha \left(A_D (\mathbf{I} - D\gamma_0) \omega_2, \theta \right)_{L^2(\Omega)} - \left(\eta A_R \theta, \theta \right)_{L^2(\Omega)};$$

using the characterizations in (1.4) and (1.20), along with the equality posted in (2.23), we have upon the taking of adjoints that

$$(2.1) = \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_2, \mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1 \right)_{L^2(\Omega)} - \left\langle \mathring{\mathbf{A}} \omega_1, \omega_2 \right\rangle_{\left[D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \right]' \times D(\mathring{\mathbf{A}}^{\frac{1}{2}})}$$

$$+\alpha \left(A_R \theta - \frac{\sigma}{\eta} \theta, \omega_2 \right)_{L^2(\Omega)} - \alpha \left(\theta, \frac{\partial \omega_2}{\partial \nu} \right)_{L^2(\Gamma_1)} - \alpha\lambda \left(\theta, \omega_2 \right)_{L^2(\Gamma_1)}$$

$$-\alpha \left(A_D (\mathbf{I} - D\gamma_0) \omega_2, \theta \right)_{L^2(\Omega)} - \left(\eta A_R \theta, \theta \right)_{L^2(\Omega)}$$

$$= \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_2, \mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1 \right)_{L^2(\Omega)} - \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1, \mathring{\mathbf{A}}^{\frac{1}{2}} \omega_2 \right)_{L^2(\Omega)} - \alpha \left(\Delta\theta, \omega_2 \right)_{L^2(\Omega)}$$

$$-\alpha \left(\theta, \frac{\partial \omega_2}{\partial \nu} \right)_{L^2(\Gamma_1)} - \alpha\lambda \left(\theta, \omega_2 \right)_{L^2(\Gamma_1)} + \alpha \left(\Delta\omega_2, \theta \right)_{L^2(\Omega)} + \left(\eta\Delta\theta - \sigma\theta, \theta \right)_{L^2(\Omega)}$$

$$= \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_2, \mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1 \right)_{L^2(\Omega)} - \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1, \mathring{\mathbf{A}}^{\frac{1}{2}} \omega_2 \right)_{L^2(\Omega)} + \alpha \left(\nabla\theta, \nabla\omega_2 \right)_{L^2(\Omega)}$$

$$-\alpha \left(\nabla\omega_2, \nabla\theta \right)_{L^2(\Omega)} - \eta \left\| \nabla\theta \right\|^2_{L^2(\Omega)} - \lambda\eta \left\| \theta \right\|^2_{L^2(\Gamma)} - \sigma \left\| \theta \right\|^2_{L^2(\Omega)}$$

$$(2.2) \leq 0$$

(here, we are using the fact that $\frac{\partial\theta}{\partial\nu} = -\lambda\theta$); i.e., $\mathcal{A}_\gamma$ is dissipative.

To show the maximality of $\mathcal{A}_\gamma$: if for some $\xi > 0$ and arbitrary $[f_1, f_2, f_3] \in \mathbf{H}_\gamma$, $[\omega_1, \omega_2, \theta] \in D(\mathcal{A}_\gamma)$ solves the equation

$$(\xi \mathbf{I} - \mathcal{A}_\gamma) \begin{bmatrix} \omega_1 \\ \omega_2 \\ \theta \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}, \tag{2.3}$$

then this relation holds if and only if

(2.4)

$$\begin{cases} \xi\omega_1 - \omega_2 = f_1 \ \text{ in } D(\mathring{\mathbf{A}}^{\frac{1}{2}}), \\ \\ \xi\omega_2 + P_\gamma^{-1}\left(\mathring{\mathbf{A}}\omega_1 + \alpha\mathring{\mathbf{A}}G_1\gamma_0\theta - \alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta - \alpha A_R\theta + \dfrac{\alpha\sigma}{\eta}\theta\right) = f_2 \ \text{ in } H^1_{\Gamma_0,\gamma}(\Omega), \\ \\ \xi\theta + \dfrac{\alpha}{\beta}A_D(\mathbf{I} - D\gamma_0)\omega_2 + \dfrac{\eta}{\beta}A_R\theta = f_3 \ \text{ in } L^2_{\sigma+\lambda}(\Omega) \end{cases}$$

$$\Longleftrightarrow$$

(2.5)

$$\begin{cases} \xi^3 P_\gamma\omega_1 + \xi\mathring{\mathbf{A}}\omega_1 + \alpha\xi\mathring{\mathbf{A}}G_1\gamma_0\theta - \alpha\lambda\xi\mathring{\mathbf{A}}G_2\gamma_0\theta - \alpha\xi A_R\theta + \dfrac{\alpha\xi\sigma}{\eta}\theta \\ \qquad = \xi P_\gamma f_2 + \xi^2 P_\gamma f_1 \ \text{ in } H^{-1}_{\Gamma_0,\gamma}(\Omega), \\ \\ \alpha\xi A_D(\mathbf{I} - D\gamma_0)\omega_1 + \beta\xi\theta + \eta A_R\theta = \beta f_3 + \alpha A_D(\mathbf{I} - D\gamma_0)f_1 \text{ in } L^2(\Omega). \end{cases}$$

At this point we bring forth the following proposition.

PROPOSITION 2.1. *The operator* $\boldsymbol{F}$ *defined by*

$$\boldsymbol{F} \equiv \begin{bmatrix} \xi^3 P_\gamma + \xi\boldsymbol{\mathring{A}} & \alpha\xi\boldsymbol{\mathring{A}}G_1\gamma_0 - \alpha\lambda\xi\boldsymbol{\mathring{A}}G_2\gamma_0 - \alpha\xi A_R + \dfrac{\alpha\xi\sigma}{\eta}\boldsymbol{I} \\ \alpha\xi A_D(\boldsymbol{I} - D\gamma_0) & \beta\xi\boldsymbol{I} + \eta A_R \end{bmatrix} \tag{2.6}$$

is an element of $\mathcal{L}(D(\boldsymbol{\mathring{A}}^{1/2}) \times H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega), [D(\boldsymbol{\mathring{A}}^{1/2})]' \times [H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)]')$ *and is boundedly invertible.*

Proof of Proposition 2.1. Easily, from the definitions of the operators which make up the components of $\mathbf{F}$, all of which are given in section 1.2, we deduce that $\mathbf{F}$ is bounded in the asserted topology. Moreover, we note by Green's theorem that for arbitrary $\theta \in D(A_R)$ and $\omega \in D(\mathring{\mathbf{A}}^{1/2})$,

$$\left\langle A_R\theta + \alpha\lambda\xi\mathring{\mathbf{A}}G_2\gamma_0, \omega \right\rangle_{[D(\mathring{\mathbf{A}}^{\frac{1}{2}})]' \times D(\mathring{\mathbf{A}}^{\frac{1}{2}})} = (\nabla\theta, \nabla\omega)_{L^2(\Omega)} + \frac{\sigma}{\eta}(\theta, \omega)_{L^2(\Gamma)}; \tag{2.7}$$

the characterization (1.13) and an extension by continuity will then have that (2.7) holds for all θ in $H^1(\Omega)$, and so for θ in $H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)$. (2.7) in turn, when coupled with (2.23), (1.24), (1.14), (1.20), and Green's formula will provide the following coercivity inequality for all $[\omega, \theta] \in D(\mathring{\mathbf{A}}^{1/2}) \times H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)$:

$$\begin{aligned} \left\langle \mathbf{F}\begin{bmatrix} \omega \\ \theta \end{bmatrix}, \begin{bmatrix} \omega \\ \theta \end{bmatrix} \right\rangle &= \xi^3 \|\omega\|^2_{L^2(\Omega)} + \xi^3\gamma \|\nabla\omega\|^2_{L^2(\Omega)} + \xi \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} \\ &\quad -\alpha\xi(\nabla\theta, \nabla\omega)_{L^2(\Omega)} + \alpha\xi(\nabla\theta, \nabla\omega)_{L^2(\Omega)} \\ &\quad +\eta\|\nabla\theta\|^2_{L^2(\Omega)} + \lambda\eta\|\theta\|^2_{L^2(\Gamma)} + (\sigma + \beta\xi)\|\theta\|^2_{L^2(\Omega)} \\ &\quad \text{(after noting the cancellation of boundary terms)} \\ &\geq C\left[\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} + \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\right] \end{aligned} \tag{2.8}$$

(where $\langle\cdot,\cdot\rangle$ in (2.8) denotes the pairing between $D(\mathring{\mathbf{A}}^{1/2})\times H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)$ and its dual, and where the constant $C>0$). Thus, by Lax–Milgram, $\mathbf{F}^{-1}$ exists as an element of

$$\mathcal{L}\left(\left[D(\mathring{\mathbf{A}}^{\frac{1}{2}})\right]'\times\left[H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)\right]', D(\mathring{\mathbf{A}}^{\frac{1}{2}})\times H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)\right),$$

and the Proposition is proved.

To complete the proof of the maximality of $\mathcal{A}_\gamma$, we apply the inverse assured by Proposition 2.1 to both sides of (2.5) to obtain

$$\begin{cases}\begin{bmatrix}\omega_1\\ \theta\end{bmatrix}\equiv\mathbf{F}^{-1}\begin{bmatrix}\xi P_\gamma f_2+\xi^2 P_\gamma f_1\\ \beta f_3+\alpha A_D(\mathbf{I}-D\gamma_0)f_1\end{bmatrix},\\ \omega_2\equiv\xi\omega_1-f_1,\end{cases}\tag{2.9}$$

and a fortiori, one has, by using the second equation in (2.5), that

$$A_R\theta=-\frac{\beta\xi}{\eta}\theta-\frac{\alpha\xi}{\eta}A_D(\mathbf{I}-D\gamma_0)\omega_1+\frac{\beta}{\eta}f_3+\frac{\alpha}{\eta}A_D(\mathbf{I}-D\gamma_0)f_1\ \in L^2(\Omega),$$

viz. $\theta\in D(A_R)\cap L^2_{\sigma+\lambda}(\Omega)$. This additional regularity of θ, in conjunction with that implied in the first equation of (2.5) (namely, $\mathring{\mathbf{A}}\omega_1+\alpha\mathring{\mathbf{A}}G_1\gamma_0\theta-\alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta\in H^{-1}_{\Gamma_0,\gamma}(\Omega)$) and along with the third equation of (2.4), gives that our constructively acquired solution $[\omega_1,\omega_2,\theta]$ to (2.3) is in $D(\mathcal{A}_\gamma)$ as defined in (1.33). Hence, $\mathcal{A}_\gamma$ is maximal dissipative and the proof of Theorem 1.1 is complete.

2.2. Proof of Theorem 1.2. By definition, if $\left[\omega^0,\omega^1,\theta^0\right]\in D\left(\mathcal{A}_\gamma\right)$, then $\omega^1\in D(\mathring{\mathbf{A}}^{1/2})$ and $\theta^0\in D(A_R)$, and

$$\mathring{\mathbf{A}}\omega^0+\alpha\mathring{\mathbf{A}}G_1\gamma_0\theta^0-\alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta^0=g\in H^{-1}_{\Gamma_0,\gamma}(\Omega)=\left[D(\mathring{\mathbf{A}}^{\frac{1}{4}})\right]';\tag{2.10}$$

as $\mathring{\mathbf{A}}^{-1}:[D(\mathring{\mathbf{A}}^{1/4})]'\to D(\mathring{\mathbf{A}}^{3/4})\subset H^3(\Omega)$ (this containment deduced by the last characterization in (1.4)), we have after applying $\mathring{\mathbf{A}}^{-1}$ to (2.10), the use of trace theory and the regularity posted in (1.19) that

$$\omega^0=\mathring{\mathbf{A}}^{-1}g-\alpha G_1\gamma_0\theta^0+\alpha\lambda G_2\gamma_0\theta^0\in H^3(\Omega).\tag{2.11}$$

Thus for $\left[\omega^0,\omega^1,\theta^0\right]\in D\left(\mathcal{A}^2_\gamma\right)$,

(2.12)

$$\mathcal{A}_\gamma\begin{bmatrix}\omega^0\\ \omega^1\\ \theta^0\end{bmatrix}=\begin{bmatrix}\omega^1\\ P_\gamma^{-1}\left[-\mathring{\mathbf{A}}\omega^0-\alpha\mathring{\mathbf{A}}G_1\gamma_0\theta^0+\alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta^0+\alpha\left(A_R\theta^0-\frac{\sigma}{\eta}\theta^0\right)\right]\\ -\frac{\eta}{\beta}A_R\theta^0-\frac{\alpha}{\beta}A_D(\mathbf{I}-D\gamma_0)\omega^1\end{bmatrix}\in D\left(\mathcal{A}_\gamma\right),$$

and (2.12) coupled with (2.11) implies that

$$\omega^1\in H^3(\Omega).\tag{2.13}$$

Moreover, (2.12) also has that

$$P_\gamma^{-1}\left[\mathring{\mathbf{A}}\omega^0+\alpha\mathring{\mathbf{A}}G_1\gamma_0\theta^0-\alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta^0-\alpha\left(A_R\theta^0-\frac{\sigma}{\eta}\theta^0\right)\right]=g, \tag{2.14}$$

where $g\in D(\mathring{\mathbf{A}}^{1/2})$, or equivalently

$$\mathring{\mathbf{A}}\omega^0+\gamma\mathring{\mathbf{A}}G_2\gamma_1 g+\alpha\mathring{\mathbf{A}}G_1\gamma_0\theta^0-\alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta^0=g-\gamma\Delta g-\alpha\Delta\theta^0\in L^2(\Omega), \tag{2.15}$$

after using (1.29). A fortiori then, $\omega^0+\gamma G_2\gamma_1 g+\alpha G_1\gamma_0\theta^0-\alpha\lambda G_2\gamma_0\theta^0\in D(\mathring{\mathbf{A}})\subset H^4(\Omega)$. But trace theory and the smoothing specified in (1.19) give that $G_2\gamma_1 g$, $G_1\gamma_0\theta^0$ and $G_2\gamma_0\theta^0\in H^4(\Omega)$, and thus $D(\mathcal{A}_\gamma^2)\subset H^4(\Omega)\times H^3(\Omega)\times H^2(\Omega)$ with the inclusion being continuous. The solution $[\omega,\omega_t,\theta]$ will consequently have the asserted regularity upon consideration of the fundamental property that for $\xi\geq 0$, $\left[\omega^0,\omega^1,\theta^0\right]\in D\left(\mathcal{A}_\gamma^\xi\right)\Rightarrow$

$$\begin{bmatrix}\omega\\ \omega_t\\ \theta\end{bmatrix}=e^{\mathcal{A}_\gamma(\cdot)}\begin{bmatrix}\omega^0\\ \omega^1\\ \theta^0\end{bmatrix}\in C\left([0,T];D\left(\mathcal{A}_\gamma^\xi\right)\right)\ . \tag{2.16}$$

To prove (ii), we note that with $\left[\omega^0,\omega^1,\theta^0\right]\in D\left(\mathcal{A}_\gamma^2\right)$, $\omega_{tt}\in C([0,T];D(\mathring{\mathbf{A}}^{1/2}))$, so the solution $[\omega,\omega_t,\theta]$ to (1.1) satisfies

$$-\mathring{\mathbf{A}}\omega+\gamma\mathring{\mathbf{A}}G_2\gamma_1\omega_{tt}-\alpha\mathring{\mathbf{A}}G_1\gamma_0\theta+\alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta=\omega_{tt}-\gamma\Delta\omega_{tt}+\alpha\Delta\theta \tag{2.17}$$

in $C([0,T];\,L^2(\Omega))$, which establishes the result. □

REMARK 3. *Because of the regularity result posted in Theorem* 1.2 (ii)*, we have for sufficiently smooth initial data the valid pointwise representation*

$$\omega_{tt}+\Delta^2\omega-\gamma\Delta\omega_{tt}+\alpha\Delta\theta=0. \tag{2.18}$$

REMARK 4. *If either λ or $\sigma>0$, then for initial data $\left[\omega^0,\omega^1,\theta^0\right]\in D\left(\mathcal{A}_\gamma^2\right)$, we will also have that the solution component θ of* 1.1 *is in $C([0,T];H^3(\Omega))$. In fact, the last component on the right-hand side of* (2.12)*, the definition of $D\left(\mathcal{A}_\gamma\right)$, and* (2.13) *give that*

$$A_R\theta^0=h+\frac{\alpha}{\eta}\Delta\omega^1\in H^1(\Omega), \tag{2.19}$$

where $h\in H^2(\Omega)$. Applying ${A_R}^{-1}$ (which exists for λ or $\sigma>0$) to both sides of (2.19) *thus yields*

$$\theta^0\in H^3(\Omega), \tag{2.20}$$

and the result will follow from the semigroup property posted in (2.16).

2.3. Proof of Theorem 1.3. In proving Theorem 1.3, we begin with a preliminary energy identity.

LEMMA 2.2. *Again, with initial data $\left[\omega^0,\omega^1,\theta^0\right]\in\boldsymbol{H}_\gamma$, we have that the component θ of the solution of* (1.1) *is an element of $L^2\left(0,\infty;H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)\right)$; indeed, we have the following relation $\forall\ T>0$:*

$$-2\eta\int_0^T\|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\,dt=E_\gamma(T)-E_\gamma(0), \tag{2.21}$$

where the "energy" $E_\gamma(t)$ is defined by

$$E_\gamma(t) \equiv \left\| \mathring{\boldsymbol{A}}^{\frac{1}{2}} \omega(t) \right\|^2_{L^2(\Omega)} + \left\| P_\gamma^{\frac{1}{2}} \omega_t(t) \right\|^2_{L^2(\Omega)} + \|\theta\|^2_{L^2_{\sigma+\lambda}(\Omega)}, \tag{2.22}$$

and where the norm of $H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)$ is as defined in (1.14).

Proof. Starting with initial data in $D(\mathcal{A}_\gamma)$ which will provide $\forall\ T > 0$ that the solution $[\omega, \omega_t, \theta] \in C([0,T]; D(\mathcal{A}_\gamma))$ and $[\omega_t, \omega_{tt}, \theta_t] \in C([0,T]; \mathbf{H}_\gamma)$, we have pointwise on $(0,T)$

$$\frac{d}{dt} \left\| \begin{bmatrix} \omega(t) \\ \omega_t(t) \\ \theta(t) \end{bmatrix} \right\|^2_{\mathbf{H}_\gamma} = 2 \left(\mathcal{A}_\gamma \begin{bmatrix} \omega(t) \\ \omega_t(t) \\ \theta(t) \end{bmatrix}, \begin{bmatrix} \omega(t) \\ \omega_t(t) \\ \theta(t) \end{bmatrix} \right)_{\mathbf{H}_\gamma},$$

and for this special choice of initial data we will have the desired equality (2.21) upon integration and using the fact from (1.12) that

$$\begin{aligned} (A_R\theta, \theta)_{L^2(\Omega)} &= \left(-\Delta\theta + \frac{\sigma}{\eta}\theta, \theta \right)_{L^2(\Omega)} \\ &= \|\nabla\theta\|^2_{L^2(\Omega)} + \frac{\sigma}{\eta} \|\theta\|^2_{L^2(\Omega)} + \lambda \|\theta\|^2_{L^2(\Gamma)} \quad \text{for } \theta \in D(A_R). \end{aligned} \tag{2.23}$$

The asserted L^2-regularity follows immediately from (2.21), using the norm definition (1.14) for $H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)$, and the fact that $\left\{ e^{\mathcal{A}_\gamma t} \right\}_{t \geq 0}$ is a contraction semigroup. A density argument concludes the proof. □

REMARK 5. *J. Lagnese in* [12] *first showed the dissipativity property* (2.21) *through a formal integration and a subsequent justification through variational arguments, and the alternate proof is included here as a simple consequence of contractive semigroups.*

We next derive a trace regularity result for the model under consideration here, a regularity which does not follow from the standard Sobolev trace theory, and which is critical in our estimates of uniform decay. We note that related trace regularity results for Euler–Bernoulli plates were proved in [18] and for Kirchoff plates in [14].

LEMMA 2.3. *One has the component ω of the solution $[\omega, \omega_t, \theta]$ of* (1.1) *satisfies $\Delta\omega|_{\Gamma_0} \in L^2(0,T; L^2(\Gamma_0))$ with the estimate*

$$\begin{aligned} \int_0^T \|\Delta\omega\|^2_{L^2(\Gamma_0)}\, dt \leq C \Bigg(\int_0^T \left[\left\| \mathring{\boldsymbol{A}}^{\frac{1}{2}} \omega \right\|^2_{L^2(\Omega)} + \left\| P_\gamma^{\frac{1}{2}} \omega_t \right\|^2_{L^2(\Omega)} + \|\theta\|^2_{H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)} \right] dt \\ + E_\gamma(T) + E_\gamma(0) \Bigg), \end{aligned} \tag{2.24}$$

where C does not depend on the parameter γ.

Proof. If we take initial data $\left[\omega^0, \omega^1, \theta^0\right]$ in $D(\mathcal{A}_\gamma^2)$, then Theorem 1.2 provides that $[\omega, \omega_t, \theta]$ is a classical pointwise solution of (1.1). We will work to extract the desired estimate (2.24) in this special case—and consequently for all initial data after an extension by continuity—by multiplying the first equation of (1.1) by the quantity $h \cdot \nabla\omega$, where $h(x,y) \equiv [h_1(x,y), h_2(x,y)]$ is a $\left[C^2(\overline{\Omega})\right]^2$ vector field[1] which satisfies

$$h|_\Gamma = \begin{cases} [\nu_1, \nu_2] & \text{on } \Gamma_0, \\ 0 & \text{on } \Gamma_1, \end{cases} \tag{2.25}$$

[1]Here is where we use the fact that Γ_0 and Γ_1 are separated.

followed by an integration from 0 to T; i.e., we will work with the equation

$$\int_0^T \left(\omega_{tt} - \gamma\Delta\omega_{tt} + \Delta^2\omega + \alpha\Delta\theta, h\cdot\nabla\omega\right)_{L^2(\Omega)} dt = 0. \tag{2.26}$$

(i) First,

$$\begin{aligned}
&\int_0^T (\omega_{tt}, h\cdot\nabla\omega)_{L^2(\Omega)}\, dt = (\omega_t, h\cdot\nabla\omega)_{L^2(\Omega)}\Big|_0^T - \int_0^T (\omega_t, h\cdot\nabla\omega_t)_{L^2(\Omega)}\, dt \\
&= (\omega_t, h\cdot\nabla\omega)_{L^2(\Omega)}\Big|_0^T - \frac{1}{2}\int_0^T\int_\Omega \operatorname{div}\left(\omega_t^2 h\right) dtd\Omega \\
&\quad + \frac{1}{2}\int_0^T\int_\Omega \omega_t^2\left[h_{1x} + h_{2y}\right] dtd\Omega \\
&= (\omega_t, h\cdot\nabla\omega)_{L^2(\Omega)}\Big|_0^T + \frac{1}{2}\int_0^T\int_\Omega \omega_t^2\left[h_{1x} + h_{2y}\right] dtd\Omega,
\end{aligned} \tag{2.27}$$

after making use of the divergence theorem and the fact that $\omega_t = 0$ on Γ_0.

(ii) Next,

$$\begin{aligned}
&\int_0^T (-\Delta\omega_{tt}, h\cdot\nabla\omega)_{L^2(\Omega)}\, dt \\
&= (\nabla\omega_t, \nabla(h\cdot\nabla\omega))_{L^2(\Omega)}\Big|_0^T - \int_0^T (\nabla\omega_t, \nabla(h\cdot\nabla\omega_t))_{L^2(\Omega)}\, dt \\
&= (\nabla\omega_t, \nabla(h\cdot\nabla\omega))_{L^2(\Omega)}\Big|_0^T - \frac{1}{2}\int_0^T\int_\Omega \operatorname{div}\left(|\nabla\omega_t|^2 h\right) dtd\Omega \\
&\quad - \int_0^T\int_\Omega \left[\frac{\omega_{tx}^2 h_{1x}}{2} + \frac{\omega_{ty}^2 h_{2y}}{2}\right] dtd\Omega - \int_0^T\int_\Omega \left[\omega_{tx}\omega_{ty}h_{2x} + \omega_{tx}\omega_{ty}h_{1y}\right] dtd\Omega \\
&\quad + \int_0^T\int_\Omega \left[\frac{\omega_{tx}^2 h_{2y}}{2} + \frac{\omega_{ty}^2 h_{1x}}{2}\right] dtd\Omega \\
&= (\nabla\omega_t, h\cdot\nabla\omega)_{L^2(\Omega)}\Big|_0^T \\
&\quad + \int_0^T\int_\Omega \left[\frac{\omega_{tx}^2 h_{2y}}{2} + \frac{\omega_{ty}^2 h_{1x}}{2} - \frac{\omega_{tx}^2 h_{1x}}{2} - \frac{\omega_{ty}^2 h_{2y}}{2}\right] dtd\Omega \\
&\quad - \int_0^T\int_\Omega \left[\omega_{tx}\omega_{ty}h_{2x} + \omega_{tx}\omega_{ty}h_{1y}\right] dtd\Omega,
\end{aligned} \tag{2.28}$$

after again using the divergence theorem and the fact that

$$\int_\Omega \operatorname{div}\left(|\nabla\omega_t|^2 h\right) d\Omega = \int_{\Gamma_0} |\nabla\omega_t|^2\, d\Gamma_0 = 0 (\text{as } \omega_t(t) \in H^2_{\Gamma_0}(\Omega)).$$

(iii) To handle the fourth-order term, we use Green's theorem (1.5), the given boundary conditions of (1.1), (2.25), and the fact that $\omega \in H^2_{\Gamma_0}(\Omega)$ to obtain

(2.29)
$$\int_0^T \left(\Delta^2\omega, h\cdot\nabla\omega\right)_{L^2(\Omega)} dt = \int_0^T a\left(\omega, h\cdot\nabla\omega\right) dt$$
$$+\alpha\int_0^T\int_{\Gamma_1} \theta\cdot\frac{\partial h\cdot\nabla\omega}{\partial\nu}\, d\Gamma_1 dt \;-\int_0^T\int_{\Gamma_0}\left(\Delta\omega + (1-\mu)B_1\omega\right)\frac{\partial^2\omega}{\partial\nu^2}\, d\Gamma_0 dt.$$

We note at this point that we can rewrite the first term on the right-hand side of (2.29) as

(2.30)
$$\int_0^T a\left(\omega, h\cdot\nabla\omega\right) dt = \frac{1}{2}\int_0^T\int_\Omega h\cdot\nabla\left[\omega_{xx}^2+\omega_{yy}^2+2\mu\omega_{xx}\omega_{yy}+2(1-\mu)\omega_{xy}^2\right] dtd\Omega$$
$$+\mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt\right),$$

where $\mathcal{O}(\int_0^T \|\mathring{\mathbf{A}}^{1/2}\omega\|^2_{L^2(\Omega)}dt)$ denotes a series of terms which can be majorized by the $L^2(0,T;D(\mathring{\mathbf{A}}^{1/2}))$-norm of ω; we consequently have by the divergence theorem that

$$\int_0^T a\left(\omega, h\cdot\nabla\omega\right) dt =$$
$$\frac{1}{2}\int_0^T\int_\Omega h\cdot\nabla\left[\omega_{xx}^2+\omega_{yy}^2+2\mu\omega_{xx}\omega_{yy}+2(1-\mu)\omega_{xy}^2\right] dtd\Omega$$
$$+\mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt\right)$$
$$= \frac{1}{2}\int_0^T\int_\Omega \text{ div }\left\{h\left[\omega_{xx}^2+\omega_{yy}^2+2\mu\omega_{xx}\omega_{yy}+2(1-\mu)\omega_{xy}^2\right]\right\}$$
(2.31)
$$+\,\mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt\right)$$
$$= \frac{1}{2}\int_0^T\int_{\Gamma_0}\left[\omega_{xx}^2+\omega_{yy}^2+2\mu\omega_{xx}\omega_{yy}+2(1-\mu)\omega_{xy}^2\right] dtd\Gamma_0$$
$$+\mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt\right)$$
$$= \frac{1}{2}\int_0^T\int_{\Gamma_0}(\Delta\omega)^2\, dt + \mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt\right),$$

where in the last step above, we have used the fact (as reasoned in [12, Ch. 4] that $\omega|_{\Gamma_0} = \frac{\partial\omega}{\partial\nu}|_{\Gamma_0} = 0$ implies that $\omega_{xx}^2+\omega_{yy}^2+2\mu\omega_{xx}\omega_{yy}+2(1-\mu)\omega_{xy}^2 = (\Delta\omega)^2$ on Γ_0.

To handle the last term on the right-hand side of (2.29), we note that $B_1\omega = 0$ on Γ_0, which implies that

$$\Delta\omega = \Delta\omega + (1-\mu)B_1\omega = \frac{\partial^2\omega}{\partial\nu^2} \text{ on } \Gamma_0 \ ; \tag{2.32}$$

we consequently have upon the insertion of (2.31) into (2.29), as well as by the consideration of (2.32), that

$$\begin{aligned}\int_0^T \left(\Delta^2\omega, h\cdot\nabla\omega\right)_{L^2(\Omega)} dt &= -\frac{1}{2}\int_0^T \|\Delta\omega\|^2_{L^2(\Gamma_0)}\, dt \\ +\alpha\int_0^T\int_{\Gamma_1} \theta\cdot\frac{\partial h\cdot\nabla\omega}{\partial\nu}\, d\Gamma_1 dt &+ \mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt\right).\end{aligned} \tag{2.33}$$

(iv) To handle the last term on the left-hand side of equation (2.26), we again use Green's theorem and the boundary conditions posted in (1.1) to obtain

$$\int_0^T (\Delta\theta, h\cdot\nabla\omega)_{L^2(\Omega)}\, dt = -\int_0^T (\nabla\theta, \nabla\,(h\cdot\nabla\omega))_{L^2(\Omega)}\, dt. \tag{2.34}$$

To finish the proof, we rewrite (2.26) by collecting the relations given above in (2.27), (2.28), (2.33), and (2.34) to attain the desired inequality (2.24), upon the taking of norms and a subsequent majorization. □

In showing the exponential decay of the semigroup $\left\{e^{\mathcal{A}_\gamma t}\right\}_{t\geq 0}$ (Theorem 1.3) it will suffice as usual, to prove that there exists a time $0 < T < \infty$ which satisfies for all initial data in $\mathbf{H}_\gamma$,

$$E_\gamma(T) \leq \xi E_\gamma(0) \text{ with } \xi < 1. \tag{2.35}$$

By a density argument, it will then be enough by Lemma 2.2 *to show the existence of a time* T, $0 < T < \infty$, *and a positive constant* C_T *(independent of* γ*) for initial data in* $\left[\omega^0, \omega^1, \theta^0\right] \in D\left(\mathcal{A}_\gamma^2\right)$ *such that*

$$E_\gamma(T) \leq C_T \int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\, dt, \tag{2.36}$$

to which end we will proceed to work.

2.4. Proof of inequality (2.36). Because of Theorem 1.2, we have for initial data $\left[\omega^0, \omega^1, \theta^0\right] \in D\left(\mathcal{A}_\gamma^2\right)$ a classical pointwise solution $[\omega, \omega_t, \theta]$ of (1.1); we can thus multiply the first equation in (1.1) by $A_D^{-1}\theta$ and integrate in time and space to obtain

$$\int_0^T \left(\omega_{tt} - \gamma\Delta\omega_{tt} + \Delta^2\omega + \alpha\Delta\theta, A_D^{-1}\theta\right)_{L^2(\Omega)} dt = 0; \tag{2.37}$$

the bulk of the work from here on out will be the scrutiny of the left-hand side of this equation.

(A.1) *Dealing with* $\int_0^T \left(\omega_{tt} - \gamma\Delta\omega_{tt}, A_D^{-1}\theta\right)_{L^2(\Omega)} dt$. Using an integration by parts, the second differential equation of (1.1) and the fact that $A_R\theta = -\Delta\theta + \frac{\sigma}{\eta}\theta = -\Delta\theta +$

$\Delta D\gamma_0\theta + \frac{\sigma}{\eta}\theta = A_D(\mathbf{I} - D\gamma_0)\theta + \frac{\sigma}{\eta}\theta$ produce

$$
\begin{aligned}
&\int_0^T \left(\omega_{tt} - \gamma\Delta\omega_{tt}, A_D^{-1}\theta\right)_{L^2(\Omega)} dt \\
&= \left(\omega_t, A_D^{-1}\theta\right)_{L^2(\Omega)}\Big|_0^T + \gamma\left(\nabla\omega_t, \nabla A_D^{-1}\theta\right)_{L^2(\Omega)}\Big|_0^T \\
&\quad - \int_0^T \left[\left(\omega_t, A_D^{-1}\theta_t\right)_{L^2(\Omega)} + \gamma\left(\nabla\omega_t, \nabla A_D^{-1}\theta_t\right)_{L^2(\Omega)}\right] dt \\
&= \alpha\beta^{-1}\int_0^T \left[\|\omega_t\|^2_{L^2(\Omega)} + \gamma\|\nabla\omega_t\|^2_{L^2(\Omega)}\right] dt \\
&\quad -\alpha\beta^{-1}\int_0^T \left[\left(\omega_t, D\gamma_0\omega_t\right)_{L^2(\Omega)} + \gamma\left(\nabla\omega_t, \nabla D\gamma_0\omega_t\right)_{L^2(\Omega)}\right] dt \\
&\quad +\eta\beta^{-1}\int_0^T \left[\left(\omega_t, (\mathbf{I} - D\gamma_0)\theta\right)_{L^2(\Omega)} + \gamma\left(\nabla\omega_t, \nabla(\mathbf{I} - D\gamma_0)\theta\right)_{L^2(\Omega)}\right] dt \\
&\quad +\sigma\beta^{-1}\int_0^T \left[\left(\omega_t, A_D^{-1}\theta\right)_{L^2(\Omega)} + \gamma\left(\nabla\omega_t, \nabla A_D^{-1}\theta\right)_{L^2(\Omega)}\right] dt \\
&\quad + \left(\omega_t, A_D^{-1}\theta\right)_{L^2(\Omega)}\Big|_0^T + \gamma\left(\nabla\omega_t, \nabla A_D^{-1}\theta\right)_{L^2(\Omega)}\Big|_0^T .
\end{aligned}
\tag{2.38}
$$

A further integration by parts, an application of Green's theorem (1.5) to the term $\int_0^T (\nabla\omega_t, \nabla D\gamma_0\omega_t)_{L^2(\Omega)}\, dt$, and a consideration of the boundary conditions posted in (1.1) yield

$$
\begin{aligned}
&-\gamma\int_0^T (\nabla\omega_t, \nabla D\gamma_0\omega_t)_{L^2(\Omega)}\, dt \\
&= -\gamma\left(\nabla\omega_t, \nabla D\gamma_0\omega\right)_{L^2(\Omega)}\Big|_0^T + \gamma\int_0^T \left(\nabla\omega_{tt}, \nabla D\gamma_0\omega\right)_{L^2(\Omega)} dt \\
&= -\gamma\left(\nabla\omega_t, \nabla D\gamma_0\omega\right)_{L^2(\Omega)}\Big|_0^T - \gamma\int_0^T \left(\Delta\omega_{tt}, D\gamma_0\omega\right)_{L^2(\Omega)} dt \\
&\quad +\gamma\int_0^T \left(\frac{\partial\omega_{tt}}{\partial\nu}, \gamma_0\omega\right)_{L^2(\Gamma_1)} dt \\
&= -\gamma\left(\nabla\omega_t, \nabla D\gamma_0\omega\right)_{L^2(\Omega)}\Big|_0^T - \int_0^T \left(\omega_{tt} + \Delta^2\omega + \alpha\Delta\theta, D\gamma_0\omega\right)_{L^2(\Omega)} dt \\
&\quad +\gamma\int_0^T \left(\frac{\partial\omega_{tt}}{\partial\nu}, \gamma_0\omega\right)_{L^2(\Gamma_1)} dt \\
&= -\gamma\left(\nabla\omega_t, \nabla D\gamma_0\omega\right)_{L^2(\Omega)}\Big|_0^T - \left(\omega_t, D\gamma_0\omega\right)_{L^2(\Omega)}\Big|_0^T + \int_0^T \left(\omega_t, D\gamma_0\omega_t\right)_{L^2(\Omega)} dt \\
&\quad - \int_0^T a\left(D\gamma_0\omega, \omega\right) dt - \int_0^T \left(\alpha\theta, \frac{\partial D\gamma_0\omega}{\partial\nu}\right)_{L^2(\Gamma_1)} dt - \int_0^T \left(\Delta\omega, \frac{\partial D\gamma_0\omega}{\partial\nu}\right)_{L^2(\Gamma_0)} dt \\
&\quad +\alpha\int_0^T \left(\nabla\theta, \nabla D\gamma_0\omega\right)_{L^2(\Omega)} dt.
\end{aligned}
\tag{2.39}
$$

Given that $D\gamma_0 \in \mathcal{L}(H^s(\Omega))$ for all real s and further using the fact that A_D^{-1} is "smoothing," viz. $\left\|A_D^{-1}\theta\right\|_{H^2(\Omega)} \leq C\left\|\theta\right\|_{L^2(\Omega)}$, we have the following estimates for the solution $[\omega, \omega_t, \theta]$ of (1.1) corresponding to arbitrary initial data in $\mathbf{H}_\gamma$:

$$\left\|(I - D\gamma_0)\theta\right\|_{L^2(\Omega)} + \left\|A_D^{-1}\theta\right\|_{L^2(\Omega)} \leq C\left\|\theta\right\|_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}; \tag{2.40}$$

$$\left\|\nabla(I - D\gamma_0)\theta\right\|_{L^2(\Omega)} + \left\|\nabla A_D^{-1}\theta\right\|_{L^2(\Omega)} \leq C\left\|\theta\right\|_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}; \tag{2.41}$$

$$\left\|\nabla D\gamma_0\omega\right\|_{L^2(\Omega)} \leq C\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|_{L^2(\Omega)}; \tag{2.42}$$

$$\left\|\frac{\partial D\gamma_0\omega}{\partial\nu}\right\|_{H^{\frac{1}{2}}(\Gamma)} \leq C\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|_{L^2(\Omega)}. \tag{2.43}$$

Thus a substitution of (2.39) into (2.38) and a subsequent majorization which makes use of the inequalities (2.40)–(2.43) will give the estimate

(2.44)

$$\begin{aligned}
&\left|\int_0^T \left(\omega_{tt} - \gamma\Delta\omega_{tt}, A_D^{-1}\theta\right)_{L^2(\Omega)} dt - \alpha\beta^{-1}\int_0^T \left[\left\|\omega_t\right\|^2_{L^2(\Omega)} + \gamma\left\|\nabla\omega_t\right\|^2_{L^2(\Omega)}\right] dt\right| \\
&\leq C\int_0^T \left[\left\|\omega_t\right\|_{L^2(\Omega)}\left\|\theta\right\|_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} + \gamma\left\|\nabla\omega_t\right\|_{L^2(\Omega)}\left\|\theta\right\|_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\right] dt \\
&\quad + C\left[E_\gamma(0) + E_\gamma(T)\right] + \left|\int_0^T a\left(D\gamma_0\omega, \omega\right) dt\right| \\
&\quad + \left|\int_0^T \left(\Delta\omega, \frac{\partial D\gamma_0\omega}{\partial\nu}\right)_{L^2(\Gamma_0)} dt\right| \\
&\leq \epsilon\int_0^T \left[\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} + \left\|\omega_t\right\|^2_{L^2(\Omega)} + \gamma\left\|\nabla\omega_t\right\|^2_{L^2(\Omega)}\right] dt + C_\epsilon\int_0^T \left\|\theta\right\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} dt \\
&\quad + C\left[E_\gamma(0) + E_\gamma(T)\right] + \frac{\alpha}{\beta}\left|\int_0^T a\left(D\gamma_0\omega, \omega\right) dt\right| \\
&\quad + \frac{\alpha}{\beta}\left|\int_0^T \left(\Delta\omega, \frac{\partial D\gamma_0\omega}{\partial\nu}\right)_{L^2(\Gamma_0)} dt\right|,
\end{aligned}$$

where the constants C and C_ϵ do not depend on γ, $0 < \gamma \leq M$.

(A.2) *Dealing with* $\int_0^T \left(\Delta^2\omega, A_D^{-1}\theta\right) dt$. Yet another application of Green's theorem in (1.5) and the use of the enforced boundary conditions in (1.1) give

$$\begin{aligned}
\int_0^T \left(\Delta^2\omega, A_D^{-1}\theta\right) dt &= \int_0^T a\left(\omega, A_D^{-1}\theta\right) dt - \int_0^T \left(\Delta\omega, \frac{\partial A_D^{-1}\theta}{\partial\nu}\right)_{L^2(\Gamma_0)} dt \\
&\quad + \alpha\int_0^T \left(\theta, \frac{\partial A_D^{-1}\theta}{\partial\nu}\right)_{L^2(\Gamma_1)} dt.
\end{aligned} \tag{2.45}$$

Estimating the right-hand side of (2.45) yields, after the use of trace theory, elliptic regularity and the mean inequality,

$$\begin{aligned}
&\left|\int_0^T \left(\Delta^2\omega, A_D^{-1}\theta\right) dt\right| \\
&\leq C_0 \int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|_{L^2(\Omega)} \|\theta\|_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} dt \\
&\quad + \frac{\epsilon}{2C}\int_0^T \|\Delta\omega\|^2_{L^2(\Gamma_0)} dt + C_\epsilon \int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} dt \\
&\quad \text{(where the inverted } C \text{ is the same constant present in (2.24))} \\
&\leq\ C_0 \int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|_{L^2(\Omega)} \|\theta\|_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} dt \\
&\quad + \frac{\epsilon}{2}\left[\int_0^T \left(\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} + \left\|P_\gamma^{\frac{1}{2}}\omega_t\right\|^2_{L^2(\Omega)}\right) dt + \right. \\
&\quad \left. + E_\gamma(0) + E_\gamma(T)\right] + C_\epsilon \int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} dt \\
&\quad \text{(by Lemma 2.3)} \\
&\leq \epsilon \int_0^T \left[\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} + \left\|P_\gamma^{\frac{1}{2}}\omega_t\right\|^2_{L^2(\Omega)}\right] dt \\
&\quad + C\left[E_\gamma(0) + E_\gamma(T)\right] + C_\epsilon \int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} dt, \qquad (2.46)
\end{aligned}$$

after the use of the mean inequality.

(A.3) *Dealing with* $\int_0^T \left(\alpha\Delta\theta, A_D^{-1}\theta\right)_{L^2(\Omega)} dt$. Finally, for the last term of (2.37), again using the fact that $A_R\theta = A_D(\mathbf{I} - D\gamma_0)\theta + \frac{\sigma}{\eta}\theta$, we have easily

$$\begin{aligned}
&\alpha\int_0^T \left(A_D(\mathbf{I} - D\gamma_0)\theta + \frac{\sigma}{\eta}\theta, A_D^{-1}\theta\right)_{L^2(\Omega)} dt \\
&\quad = \alpha\int_0^T \left[\|\theta\|^2_{L^2(\Omega)} - (D\gamma_0\theta, \theta)_{L^2(\Omega)} + \left(\frac{\alpha\sigma}{\eta}A_D^{-1}\theta, \theta\right)_{L^2(\Omega)}\right] dt \qquad (2.47) \\
&\quad \leq C\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\,.
\end{aligned}$$

(A.4) *Combining* (2.37), (2.44), (2.46)*, and* (2.47) *thus results in the following.* For $\epsilon > 0$ small enough there exists a constant $C > 0$ (independent of γ) such that the solution $[\omega, \omega_t, \theta]$ of (1.1) satisfies

$$\begin{aligned}
&\left(\frac{\alpha}{\beta} - 2\epsilon\right)\int_0^T \left[\|\omega_t\|^2_{L^2(\Omega)} + \gamma\|\nabla\omega_t\|^2_{L^2(\Omega)}\right] dt \\
&\leq C\left[\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)} dt + E_\gamma(T) + E_\gamma(0)\right] \\
&\quad + 2\epsilon\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt + \frac{\alpha}{\beta}\left|\int_0^T a\left(D\gamma_0\omega, \omega\right) dt\right| \\
&\quad + \frac{\alpha}{\beta}\left|\int_0^T \left(\Delta\omega, \frac{\partial D\gamma_0\omega}{\partial\nu}\right)_{L^2(\Gamma_0)} dt\right|, \qquad (2.48)
\end{aligned}$$

where the noncrucial dependence of C upon ϵ has not been noted.

(A.5) *Estimating the residual terms* $|\int_0^T a(D\gamma_0\omega,\omega)dt|$ and $|\int_0^T(\Delta\omega,\frac{\partial D\gamma_0\omega}{\partial\nu})_{L^2(\Gamma_0)}dt|$.[2] At this point we will find it advantageous to consider a decomposition of the solution component $[\omega,\omega_t]$ into $\omega=\omega^{(1)}+\omega^{(2)}+\omega^{(3)}$ (again with the corresponding initial data $[\omega_0,\omega_1]\in D(\mathcal{A}_\gamma^2)$), where the $\omega^{(i)}$ solve, respectively,

$$\begin{cases} -\gamma\Delta\omega_{tt}^{(1)}+\Delta^2\omega^{(1)}=-\alpha\Delta\theta \ \text{ on } (0,\infty)\times\Omega, \\ \omega^{(1)}=\dfrac{\partial\omega^{(1)}}{\partial\nu}=0 \ \text{ on } (0,\infty)\times\Gamma_0, \\ \begin{cases} \Delta\omega^{(1)}+(1-\mu)B_1\omega^{(1)}+\alpha\theta=0 \\ \dfrac{\partial\Delta\omega^{(1)}}{\partial\nu}+(1-\mu)\dfrac{\partial B_2\omega^{(1)}}{\partial\tau}-\gamma\dfrac{\partial\omega_{tt}^{(1)}}{\partial\nu}=0 \end{cases} \text{on } (0,\infty)\times\Gamma_1, \\ \omega^{(1)}(t=0)=\omega_t^{(1)}(t=0)=0; \end{cases} \tag{2.49}$$

$$\begin{cases} -\gamma\Delta\omega_{tt}^{(2)}+\Delta^2\omega^{(2)}=-\omega_{tt} \ \text{ on } (0,\infty)\times\Omega; \\ \omega^{(2)}=\dfrac{\partial\omega^{(2)}}{\partial\nu}=0 \ \text{ on } (0,\infty)\times\Gamma_0; \\ \begin{cases} \Delta\omega^{(2)}+(1-\mu)B_1\omega^{(2)}=0 \\ \dfrac{\partial\Delta\omega^{(2)}}{\partial\nu}+(1-\mu)\dfrac{\partial B_2\omega^{(2)}}{\partial\tau}-\gamma\dfrac{\partial\omega_{tt}^{(2)}}{\partial\nu}+\alpha\dfrac{\partial\theta}{\partial\nu}=0 \end{cases} \text{on } (0,\infty)\times\Gamma_1; \\ \omega^{(2)}(t=0)=\omega_t^{(2)}(t=0)=0. \end{cases} \tag{2.50}$$

$$\begin{cases} -\gamma\Delta\omega_{tt}^{(3)}+\Delta^2\omega^{(3)}=0 \ \text{ on } (0,\infty)\times\Omega; \\ \omega^{(3)}=\dfrac{\partial\omega^{(3)}}{\partial\nu}=0 \ \text{ on } (0,\infty)\times\Gamma_0; \\ \begin{cases} \Delta\omega^{(3)}+(1-\mu)B_1\omega^{(3)}=0 \\ \dfrac{\partial\Delta\omega^{(3)}}{\partial\nu}+(1-\mu)\dfrac{\partial B_2\omega^{(3)}}{\partial\tau}-\gamma\dfrac{\partial\omega_{tt}^{(3)}}{\partial\nu}=0 \end{cases} \text{on } (0,\infty)\times\Gamma_1; \\ \omega^{(3)}(0)=\omega^0; \ \ \omega_t^{(3)}(0)=\omega^1. \end{cases} \tag{2.51}$$

Through a semigroup formulation, the well posedness of (2.50) and (2.51) can be handled just as easily as the entire system (1.1); to wit, defining on the state space

[2]Notice that at this point, one might be tempted to straightaway majorize $\int_0^T a\,(D\gamma_0\omega,\omega)\,dt$ so as to obtain something like $|\int_0^T a(D\gamma_0\omega,\omega)|dt\le C\int_0^T\|\mathbf{\mathring{A}}^{1/2}\omega(t)\|^2_{L^2(\Omega)}dt$. However, this will not suffice as we do not have control over the constant C (C may not be small $<<1$). Therefore, we need a different, more complex argument which will culminate in the estimate (2.72) below; likewise for the term $|\int_0^T(\Delta\omega,\frac{\partial D\gamma_0\omega}{\partial\nu})_{L^2(\Gamma_0)}dt|$.

$D(\mathring{\mathbf{A}}^{1/2}) \times H^1_{\Gamma_0,\gamma}(\Omega)$ the operator $\widetilde{\mathcal{A}}_\gamma$ as

$$\widetilde{\mathcal{A}}_\gamma \equiv \begin{pmatrix} 0 & \mathbf{I} \\ -\widetilde{P}_\gamma^{-1}\mathring{\mathbf{A}} & 0 \end{pmatrix} \tag{2.52}$$

$$(\text{where } \widetilde{P}_\gamma \equiv \gamma A_N \in \mathcal{L}(H^1_{\Gamma_0,\gamma}(\Omega), H^{-1}_{\Gamma_0,\gamma}(\Omega))) \tag{2.53}$$

$$\text{with domain } D(\widetilde{\mathcal{A}}_\gamma) = \left\{ [\omega_1, \omega_2] \in D(\mathring{\mathbf{A}}^{\frac{3}{4}}) \times D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \right\}; \tag{2.54}$$

then with the same degree of effort as in the proof of Theorem 1.1, we can show that $\widetilde{\mathcal{A}}_\gamma$ generates a unitary C_0-group $\{e^{\widetilde{\mathcal{A}}_\gamma t}\}_{t\geq 0}$ on $D(\mathring{\mathbf{A}}^{1/2}) \times H^1_{\Gamma_0,\gamma}(\Omega)$ (note we are using the knowledge that $\widetilde{P}_\gamma^{-1}$ exists, inasmuch as A_N is elliptic on $H^1_{\Gamma_0,\gamma}(\Omega)$, and that $\widetilde{P}_\gamma = \gamma\left(\Delta + \mathring{\mathbf{A}}G_2\gamma_1\right)$ from (1.29)). Consequently we have that $\omega^{(2)} \in C([0,T]\,; D(\mathring{\mathbf{A}}^{1/2}) \times H^1_{\Gamma_0,\gamma}(\Omega))$, with this unique solution of (2.50) written explicitly as

$$\begin{bmatrix} \omega^{(2)}(t) \\ \omega_t^{(2)}(t) \end{bmatrix} = \int_0^t e^{\widetilde{\mathcal{A}}_\gamma(t-s)} \begin{bmatrix} 0 \\ \widetilde{P}_\gamma^{-1}\left(-\omega_{tt}(s) + \alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\theta(s)\right) \end{bmatrix} ds, \tag{2.55}$$

where again ω_{tt} is the second time derivative of the solution component ω. Recall that we are taking the initial data $\left[\omega^0, \omega^1, \theta^0\right]$ to be in $D(\mathcal{A}_\gamma^2)$, and so $\omega_{tt} \in C\left([0,T]\,; H^1_{\Gamma_0,\gamma}(\Omega)\right)$. Moreover, for arbitrary initial data, $\theta \in L^2(0,T; H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega))$, by Lemma 2.2, and this regularity, coupled with the facts contained in (1.17), (1.19), and (1.4), provide that $\mathring{\mathbf{A}}G_2\gamma_0\theta(t) \in L^2(0,T; H^{-1}_{\Gamma_0,\gamma}(\Omega))$. Hence the formula (2.55) is well defined. Likewise, $\omega^{(3)} \in C([0,T]\,; D(\mathring{\mathbf{A}}^{1/2}) \times H^1_{\Gamma_0,\gamma}(\Omega))$ with

$$\omega^{(3)}(t) = e^{\widetilde{\mathcal{A}}_\gamma t} \begin{bmatrix} \omega^0 \\ \omega^1 \end{bmatrix}. \tag{2.56}$$

Regarding the well posedness of the system (2.49), we have the following result from [14] and [13].

REGULARITY THEOREM. *For arbitrary initial data* $[\omega_0, \omega_1] \in D(\mathring{\mathbf{A}}^{1/2}) \times H^1_{\Gamma_0,\gamma}(\Omega)$, *parameter* $\xi \geq 0$, $f \in L^2(0,T; H^{-1}_{\Gamma_0,\gamma}(\Omega))$, *and* $g \in L^2(0,T; H^{1/2}(\Gamma_1))$, *the following system is well posed*:

$$\begin{cases} \xi\omega_{tt} - \gamma\Delta\omega_{tt} + \Delta^2\omega = f \ \text{ on } (0,\infty)\times\Omega; \\ \omega = \dfrac{\partial\omega}{\partial\nu} = 0 \ \text{ on } (0,\infty)\times\Gamma_0; \\ \begin{cases} \Delta\omega + (1-\mu)B_1\omega = g \\ \dfrac{\partial\Delta\omega}{\partial\nu} + (1-\mu)\dfrac{\partial B_2\omega}{\partial\tau} - \gamma\dfrac{\partial\omega_{tt}}{\partial\nu} = 0 \end{cases} \text{on } (0,\infty)\times\Gamma_1; \\ \omega(0) = \omega_0, \ \ \omega_t(0) = \omega_1, \end{cases} \tag{2.57}$$

with the solution $[\omega, \omega_t] \in C([0,T]\,; D(\mathring{\mathbf{A}}^{1/2}) \times H^1_{\Gamma_0,\gamma}(\Omega))$.

To make use of the above theorem for the resolution of (2.49) with arbitrary θ in $H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)$ subject to Robin boundary conditions, we note that $-\Delta =$

$A_R - \frac{\sigma}{\eta} \in \mathcal{L}(H^1(\Omega), \left[H^1(\Omega)\right]')$ and consequently $\Delta\theta \in L^2(0,T; H^{-1}_{\Gamma_0,\gamma}(\Omega))$; moreover, $\theta|_\Gamma \in L^2(0,T; H^{1/2}(\Gamma))$ by the trace theorem, and so the regularity theorem will give us that

$$\omega^{(1)} \in C([0,T]\,; D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)), \tag{2.58}$$

with the pointwise estimate

$$\begin{aligned} \left\| \begin{array}{c} \omega^{(1)}(t) \\ \omega^{(1)}_t(t) \end{array} \right\|^2_{D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)} &\leq C \left[\int_0^T \|\Delta\theta(t)\|^2_{H^{-1}_{\Gamma_0,\gamma}(\Omega)}\, dt + \alpha \int_0^T \|\theta(t)\|^2_{H^{\frac{1}{2}}(\Gamma_1)}\, dt \right] \\ &\leq C \int_0^T \|\theta(t)\|^2_{H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega)}\, dt. \end{aligned} \tag{2.59}$$

A simple uniqueness argument which makes use of the regularity theorem verifies that indeed the solution component $\omega \equiv \omega^{(1)} + \omega^{(2)} + \omega^{(3)}$. Moreover, concerning the explicit representation (2.55), an integration by parts has that

$$\begin{aligned} &\int_0^t e^{\widetilde{\mathcal{A}}_\gamma (t-s)} \left[\begin{array}{c} 0 \\ \widetilde{P}^{-1}_\gamma \omega_{tt}(s) \end{array} \right] ds = e^{\widetilde{\mathcal{A}}_\gamma (t-s)} \left[\begin{array}{c} 0 \\ \widetilde{P}^{-1}_\gamma \omega_t(s) \end{array} \right] \Big|_0^t \\ &+ \int_0^t e^{\widetilde{\mathcal{A}}_\gamma (t-s)} \widetilde{\mathcal{A}}_\gamma \left[\begin{array}{c} 0 \\ \widetilde{P}^{-1}_\gamma \omega_t(s) \end{array} \right] ds \\ &= e^{\widetilde{\mathcal{A}}_\gamma (t-s)} \left[\begin{array}{c} 0 \\ \widetilde{P}^{-1}_\gamma \omega_t(s) \end{array} \right] \Big|_0^t - \int_0^t e^{\widetilde{\mathcal{A}}_\gamma (t-s)} \left[\begin{array}{c} \widetilde{P}^{-1}_\gamma \omega_t(s) \\ 0 \end{array} \right] ds, \end{aligned} \tag{2.60}$$

where the last equality above makes sense pointwise in $[D(\widetilde{\mathcal{A}_\gamma}^*)]' = [D(\mathring{\mathbf{A}}^{\frac{3}{4}})]' \times [D(\mathring{\mathbf{A}}^{1/2})]'$; hence upon majorizing (2.55) with the expression (2.60) in mind (and using the contraction of the semigroup $\{e^{\widetilde{\mathcal{A}}_\gamma (t)}\}_{t \geq 0}$), we have

$$\left\| \begin{array}{c} \omega^{(2)}(t) \\ \omega^{(2)}_t(t) \end{array} \right\|^2_{D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)} \leq C_T \left[\|\theta\|^2_{L^2(0,T; H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega))} + \|\omega_t\|^2_{C([0,T]; L^2(\Omega))} \right]. \tag{2.61}$$

Thus, using (2.59), (2.61), and the explicit representation (2.56), we have

$$\left\| \left[\begin{array}{c} \omega^{(1)}(t) + \omega^{(2)}(t) \\ \omega^{(1)}(t) + \omega^{(2)}_t(t) \end{array} \right] \right\|^2_{D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)}$$

$$\leq C_T \left[\|\theta\|^2_{L^2(0,T; H^1(\Omega) \cap L^2_{\sigma+\lambda}(\Omega))} + \|\omega_t\|^2_{C([0,T]; L^2(\Omega))} \right]; \tag{2.62}$$

$$\left\| \left[\begin{array}{c} \omega^{(3)}(t) \\ \omega^{(3)}_t(t) \end{array} \right] \right\|^2_{D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)} \leq E_\gamma(0). \tag{2.63}$$

Further analyzing $\omega^{(3)}$, if we make the substitution $z \equiv \Delta\omega^{(3)}$, we then note that z solves the wave equation

$$\gamma z_{tt} = \Delta z, \tag{2.64}$$

with $[z, z_t] \in C([0,T]; L^2(\Omega) \times H^{-1}(\Omega))$. Consequently, the recent regularity result of [26] (specifically, apply Theorem 3 therein together with Remark 2.3 and the remark after Theorem 9 in [26]) reveals that z has a "trace" on Γ with a positive constant $C(T,\gamma)$ and a $\rho > 0$ such that the following estimate holds:[3]

$$\| z|_\Gamma \|_{L^2(0,T;H^{-\frac{1}{2}+\rho}(\Gamma))} \le C(T,\gamma) \, \| [z, z_t] \|_{C([0,T];L^2(\Omega)\times H^{-1}(\Omega))}; \tag{2.65}$$

and as pointwise we have

$$\|z(t)\|^2_{L^2(\Omega)} + \|z_t(t)\|^2_{H^{-1}(\Omega)} \le C E_\gamma(0) \tag{2.66}$$

(from the estimate (2.63)), we end up with

$$\left\| \Delta\omega^{(3)}\Big|_\Gamma \right\|^2_{L^2(0,T;H^{-\frac{1}{2}+\rho}(\Gamma))} \le C(T,\gamma) E_\gamma(0). \tag{2.67}$$

Recall that $\omega^{(3)}$, as the solution of (2.51), satisfies

$$\Delta\omega^{(3)} - (1-\mu)\frac{\partial^2 \omega^{(3)}}{\partial \tau^2} = (1-\mu)\kappa \frac{\partial \omega^{(3)}}{\partial \nu} \quad \text{on } (0,T)\times\Gamma_1, \tag{2.68}$$

where κ denotes the curvature, and so (2.68), coupled with the estimates (2.67) and

$$\left\| \frac{\partial \omega^{(3)}}{\partial \nu} \right\|_{C([0,T];H^{\frac{1}{2}}(\Gamma_1))} \le C \left\| \omega^{(3)} \right\|_{C([0,T];H^2(\Omega))} \le C(T) E_\gamma(0),$$

gives that $\dfrac{\partial^2 \omega^{(3)}}{\partial \tau^2} \in L^2(0,T; H^{-\frac{1}{2}+\rho}(\Gamma_1))$ with

$$\left\| \frac{\partial^2 \omega^{(3)}}{\partial \tau^2} \right\|_{L^2(0,T;H^{-\frac{1}{2}+\rho}(\Gamma_1))} \le C(T,\gamma) E_\gamma(0), \tag{2.69}$$

and (2.69) is in turn equivalent to

$$\left\| \gamma_0 \omega^{(3)} \right\|_{L^2(0,T;H^{\frac{3}{2}+\rho}(\Gamma_1))} \le C(T,\gamma) E_\gamma(0). \tag{2.70}$$

REMARK 6. *The estimate in* (2.70) *can also be derived independently of Tataru's result in* [26] *by decomposing problem* (2.51) *microlocally into respective elliptic and hyperbolic parts. In the elliptic sector, we can use standard elliptic regularity and the boundary conditions on* Γ_1 *to deduce the regularity of the trace* $\gamma_0\omega^{(3)}$ *in* $H^2(0, T\times\Gamma_1)$. *In the hyperbolic sector, we apply the transformation* $z \equiv \Delta\omega^{(3)}$, *and we are subsequently led to the study of the wave equation with its forcing term in* $L^2(0,T;H^{-1}(\Omega))$ *(due to microlocalization). The arguments presented in* [16] *and* (*see also* [17]) *apply to the hyperbolic sector specifically and provide the estimate* (2.70) *valid in that sector. Combining elliptic and hyperbolic estimates yields* (2.70) *with the value of* ρ *being at least* $\frac{1}{10}$. *Instead, the estimate obtained by using Tataru's result* [26] *leads to the optimal value of* $\rho = \frac{1}{6}$.

[3] We note that the value of ρ depends on the geometry; however, we *always* have $\rho > 0$.

Given this extra regularity for the trace of $\omega^{(3)}\big|_{\Gamma_1}$, we can hence invoke a classical PDE interpolation inequality to finally obtain

(2.71)

$$\left\|\gamma_0\omega^{(3)}\right\|^2_{L^2(0,T;H^{\frac{3}{2}}(\Gamma_1))} \leq C(T,\gamma)^{-1}\left\|\gamma_0\omega^{(3)}\right\|^2_{L^2(0,T;H^{\frac{3}{2}+\rho}(\Gamma_1))}$$

$$+C_{T,\gamma}\left\|\gamma_0\omega^{(3)}\right\|^2_{L^2(0,T;H^{\frac{1}{2}}(\Gamma_1))}$$

(where $C(T,\gamma)$ is as in (2.70), and $C_{T,\gamma}$ denotes another positive constant depending on T and γ)

$$\leq E_\gamma(0) + C_{T,\gamma}\left\|\omega^{(3)}\right\|^2_{L^2(0,T;H^1(\Omega))}$$

(after using the estimate (2.70) and trace theory)

$$\leq E_\gamma(0) + C_{T,\gamma}\left\|\omega\right\|^2_{L^2(0,T;H^1(\Omega))} + C_{T,\gamma}\left\|\omega^{(1)}+\omega^{(2)}\right\|^2_{L^2(0,T;H^1(\Omega))}$$

(after using the decomposition $\omega = \omega^{(1)}+\omega^{(2)}+\omega^{(3)}$)

$$\leq E_\gamma(0) + C_{T,\gamma}\left[\|\theta\|^2_{L^2(0,T;H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega))} + \|\omega\|^2_{L^2(0,T;H^1(\Omega))} + \|\omega_t\|^2_{C([0,T];L^2(\Omega))}\right],$$

after using the inequality (2.62).

With the decomposition of ω in hand, along with its accompanying norm estimates, particularly that of the trace $\gamma_0\omega^{(3)}$ in (2.71), we can now deal with the recalcitrant terms $|\int_0^T a\left(D\gamma_0\omega,\omega\right)dt|$ and $|\int_0^T(\Delta\omega,\frac{\partial D\gamma_0\omega}{\partial\nu})_{L^2(\Gamma_0)}dt|$:

(A5.i) *Dealing with* $\left|\int_0^T a\left(D\gamma_0\omega,\omega\right)dt\right|$:

$$\left|\int_0^T a\left(D\gamma_0\omega,\omega\right)dt\right|$$

$$= \left|\int_0^T a\left(D\gamma_0\left(\omega^{(1)}+\omega^{(2)}+\omega^{(3)}\right),\omega\right)dt\right|$$

$$\leq \int_0^T C\left\|D\gamma_0\left(\omega^{(1)}+\omega^{(2)}+\omega^{(3)}\right)\right\|_{H^2(\Omega)}\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|_{L^2(\Omega)}dt$$

(after using the fact that $D\gamma_0\in\mathcal{L}(H^2(\Omega))$)

$$\leq \epsilon\int_0^T\left\|\mathring{\mathbf{A}}\omega\right\|^2_{L^2(\Omega)}dt + C_{T,\gamma}\left[\int_0^T\|\theta\|^2_{H^1(\Omega)\cap L^2(\Omega)}\,dt + \|\omega\|^2_{L^2(0,T;H^1(\Omega))}\right.$$

$$\left.+\|\omega_t\|^2_{C([0,T];L^2(\Omega))}\right]$$

$$+C\left[E_\gamma(T)+E_\gamma(0)\right], \tag{2.72}$$

after using the boundedness of the Dirichlet map D followed by the standard mean inequality as well as the crucial estimates (2.71) and (2.62) (here we have not noted the noncrucial dependence of ϵ in the constant $C_{T,\gamma}$).

(A.5ii) *Dealing with* $|\int_0^T(\Delta\omega,\frac{\partial D\gamma_0\omega}{\partial\nu})_{L^2(\Gamma_0)}dt|$. By Lemma 2, $\Delta\omega|_{\Gamma_0}\in L^2(0,T;L^2(\Gamma_0))$, and so with this bit of information we have

$$\left|\int_0^T\left(\Delta\omega,\frac{\partial D\gamma_0\omega}{\partial\nu}\right)_{L^2(\Gamma_0)}dt\right|$$

$$\leq C\int_0^T\|\Delta\omega\|_{L^2(\Gamma_0)}\|D\gamma_0\omega\|_{H^2(\Omega)}\,dt$$

(by the trace theorem)

$$
\begin{aligned}
&= C\int_0^T \|\Delta\omega\|_{L^2(\Gamma_0)} \left\| D\gamma_0\left(\omega^{(1)}+\omega^{(2)}+\omega^{(3)}\right)\right\|_{H^2(\Omega)} dt \\
&\le \frac{\epsilon}{C}\int_0^T \|\Delta\omega\|^2_{L^2(\Gamma_0)}\, dt + C_{T,\gamma}\left[\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2(\Omega)}\, dt + \|\omega\|^2_{L^2(0,T;H^1(\Omega))}\right. \\
&\quad \left. + \|\omega_t\|^2_{C([0,T];L^2(\Omega))}\right] \\
&\quad +C\left[E_\gamma(T)+E_\gamma(0)\right] \\
&\quad \text{(again using the mean inequality followed by (2.71) and (2.62),} \\
&\quad \text{and where the inverted positive constant } C \text{ is that in (2.24))} \\
&\le \epsilon\int_0^T \left[\left\|\mathring{\mathbf{A}}^{\frac12}\omega\right\|^2_{L^2(\Omega)} + \int_0^T \left\|P_\gamma^{\frac12}\omega\right\|^2_{L^2(\Omega)}\right] dt + C_{T,\gamma}\left[\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2(\Omega)}\, dt\right. \\
&\quad \left. + \|\omega\|^2_{L^2(0,T;H^1(\Omega))} + \|\omega_t\|^2_{C([0,T];L^2(\Omega))}\right] + C\left[E_\gamma(0)+E_\gamma(T)\right]. \qquad (2.73)
\end{aligned}
$$

Combining (2.48), (2.72), and (2.73), we finally have

$$
\begin{aligned}
&\left(\frac{\alpha}{\beta}-3\epsilon\right)\int_0^T \left[\|\omega_t\|^2_{L^2(\Omega)} + \gamma\|\nabla\omega_t\|^2_{L^2(\Omega)}\right] dt \\
&\le 3\epsilon\int_0^T \left\|\mathring{\mathbf{A}}^{\frac12}\omega\right\|^2_{L^2(\Omega)} dt + C_{T,\gamma}\left[\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\, dt\right. \\
&\quad \left. + \|\omega\|^2_{L^2(0,T;H^1(\Omega))} + \|\omega_t\|^2_{C([0,T];L^2(\Omega))}\right] \\
&\quad +C\left[E_\gamma(0)+E_\gamma(T)\right]. \qquad (2.74)
\end{aligned}
$$

(B) *Conclusion of the Proof of Theorem* 3. To majorize the norm of the component ω, we multiply (1.35) by ω, integrate from 0 to T and employ Green's theorem to obtain (after accounting for the boundary conditions and using (1.20))

$$
\begin{aligned}
&\left(P_\gamma^{\frac12}\omega_t, P_\gamma^{\frac12}\omega\right)_{L^2(\Omega)}\Big|_0^T - \int_0^T \left\|P_\gamma^{\frac12}\omega_t\right\|^2_{L^2(\Omega)} dt \\
&= -\int_0^T \left\|\mathring{\mathbf{A}}^{\frac12}\omega\right\|^2_{L^2(\Omega)} dt - \alpha\int_0^T \left(\theta, \frac{\partial\omega}{\partial\nu}\right)_{L^2(\Gamma_1)} dt \\
&\quad +\alpha\int_0^T (\nabla\theta, \nabla\omega)_{L^2(\Omega)}\, dt; \qquad (2.75)
\end{aligned}
$$

since by the trace theorem we have pointwise

$$
\begin{aligned}
&\left|\left(\theta, \frac{\partial\omega}{\partial\nu}\right)_{L^2(\Gamma_1)}\right| + \left|(\nabla\theta, \nabla\omega)_{L^2(\Omega)}\right| \\
&\le C\left[\|\theta\|_{H^{\frac12}(\Gamma)}\left\|\frac{\partial\omega}{\partial\nu}\right\|_{H^{\frac12}(\Gamma_1)} + \|\theta\|_{H^1(\Omega)}\|\omega\|_{H^1(\Omega)}\right] \\
&\le C\|\theta\|_{H^1(\Omega)}\|\omega\|_{H^2(\Omega)} \le \epsilon\left\|\mathring{\mathbf{A}}^{\frac12}\omega\right\|^2_{L^2(\Omega)} + C_\epsilon\|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}, \qquad (2.76)
\end{aligned}
$$

we thus arrive at the following.

There exists a constant $C > 0$ such that for $\epsilon > 0$ small enough, the solution $[\omega, \omega_t, \theta]$ of (1.1) satisfies

$$(1-\epsilon)\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} dt \leq C\int_0^T \left[\|\omega_t\|^2_{L^2(\Omega)} + \gamma\|\nabla\omega_t\|^2_{L^2(\Omega)}\right] dt$$

$$+C\left(\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\, dt + E_\gamma(T) + E_\gamma(0)\right), \tag{2.77}$$

where the noncrucial dependence of C upon ϵ has not been noted.

Thus, if ϵ is small enough, we then have, upon combining (2.74) and (2.77), the existence of constants C and $C_{T,\gamma}$ such that

$$\int_0^T \left[\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} + \|\omega_t\|^2_{L^2(\Omega)} + \gamma\|\nabla\omega_t\|^2_{L^2(\Omega)} + \|\theta\|^2_{L^2(\Omega)}\right] dt$$

$$\leq C_{T,\gamma}\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\, dt + C\left[E_\gamma(T) + E_\gamma(0)\right]$$

$$+C_{T,\gamma}\left[\|\omega\|^2_{L^2(0,T;H^1(\Omega))} + \|\omega_t\|^2_{C([0,T];L^2(\Omega))}\right]. \tag{2.78}$$

From here, we apply the relation (2.21) and its inherent dissipativity property (that is, $E_\gamma(T) \leq E_\gamma(t)\ \forall\ 0 \leq t \leq T$) to (2.78) to finally attain the preliminary inequality; namely, for $T > 2C$ (with C as in (2.78) independent of T),

$$E_\gamma(T) \leq \frac{C_{T,\gamma} + 2C\eta}{T - 2C}\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\, dt$$

$$+C_{T,\gamma}\left[\|\omega\|^2_{L^2(0,T;H^1(\Omega))} + \|\omega_t\|^2_{C([0,T];L^2(\Omega))}\right]. \tag{2.79}$$

A straightforward compactness–uniqueness argument similar to that employed in [15] and [1] will subsequently eliminate the lower order terms in (2.79), viz. we have the following proposition.

PROPOSITION 2.4. *The presence of the inequality* (2.79) *implies that there exists a constant* C_T *which satisfies*

$$\|\omega\|^2_{L^2(0,T;H^1(\Omega))} + \|\omega_t\|^2_{C([0,T];H^2(\Omega))}\, dt \leq C_T\int_0^T \|\theta\|^2_{H^1(\Omega)\cap L^2_{\sigma+\lambda}(\Omega)}\, dt. \tag{2.80}$$

Hence, the inequalities (2.79) and (2.80) give the desired estimate (2.36) (and consequently (2.35)), and so the proof of Theorem 1.3 is now complete.

Note added in proof. As one reads through the arguments in the present paper, he or she gathers the understanding that the key ingredient in our stability proof is the selection of the "right" multiplier $A_D^{-1}\theta$ (which is novel when compared to the standard differential multipliers used in plate theory). This multiplier was first devised in our paper [3] (which initially considered the easier case of the thermoelastic plate with lower order "clamped" or "hinged" boundary conditions), and we have since invoked it in later problems (see [4], [13], [6], [5]. In particular, [4] is a preliminary version

of our present paper). In our present paper, it is this particular choice of multiplier which allows us to obtain sharp results on the uniform stabilization of thermoelastic plates with the higher order "free" boundary conditions in place, results which include the attainment of explicit decay rates.

Related work on this problem includes that of E. Bisognin, V. Bisognin, P. Menzala, and E. Zuazua in [7], who employed an alternative and indirect argument for the stabilization of the nonlinear thermoelastic plate in the case of *clamped/hinged* boundary conditions *only*. This method, even in the case of linear models, yielded weaker results than those posted in [2], [3]. (We assume that at the time of their work the four authors were unaware of [3].) Indeed, the indirect (proof by contradiction) method in [7] has the following shortcomings:

(i) The method requires two different treatments of the problem, corresponding to the respective cases $\gamma > 0$ and $\gamma = 0$. This dichotomy is necessitated by the fact that the accompanying decay rates they obtain blow up as $\gamma \downarrow 0$.

(ii) The decay rates they obtain are not explicit.

(iii) In the specific case $\gamma = 0$, the analyticity of the underlying semigroup is used in an essential way, which precludes the possibility that their indirect method can be adjusted so as to give a unified treatment of the problem for *all* cases $\gamma \geq 0$ (recall that $\gamma > 0$ corresponds to hyperbolic-like dynamics).

In contrast, the paper [3] (which is critical and constitutes a basis for the present paper) obtains decay estimates which are *uniform* in the parameter $\gamma \geq 0$, this being accomplished via the use of the multiplier $A_D^{-1}\theta$. As the authors of [7] were apparently informed much before the date of submission of [23] of this comparison between their work and that in [3] (this is a documented fact), one may then view as perplexing the subsequent appearance of the paper [23], which now claims for itself the right (and much improved with respect to [7]) result using the *very same* techniques and ideas as in [3] (which again are radically different from those in [7]). In particular, [23] uses the same multiplier and the same trace result, the latter being proclaimed therein as "hidden regularity." Perhaps adding to the perplexity is the fact that the two authors in [23], while freely addressing the aforementioned shortcomings of [7], make neither acknowledgment nor reference to [3]. Our main point here is to stress the fact that the critical multiplier and the resulting technique for proving uniform decay rates for thermoelastic plates takes its origin in [2], [3], and not in [23].

REFERENCES

[1] G. Avalos, *The exponential stability of a coupled hyperbolic/parabolic system arising in structural acoustics*, Abstract Appl. Anal., 1 (1996), pp. 203–219.

[2] G. Avalos and I. Lasiecka, *Exponential stability of a thermoelastic system without mechanical dissipation*, Rendiconti Di Instituto Di Matematica, an invited paper in a special issue honoring Pierre Grisvard; Rendiconti Di Instituto Di Matematica Dell'Universitá di Trieste, E. Mitidieri and Philippe Clement, eds., Suppl. Vol. XXVIII, 1997, p. 1–27.

[3] G. Avalos and I. Lasiecka, *Exponential Stability of a Thermoelastic System without Mechanical Dissipation*, IMA Preprint Series #1357, November 1995.

[4] G. Avalos and I. Lasiecka, *Exponential Stability of a Simply Supported Thermoelastic System without Mechanical Dissipation* II*: The Case of Simply Supported Boundary Conditions*, IMA Preprint # 1397, March, 1996.

[5] G. Avalos and I. Lasiecka, *Uniform decay rates for a nonlinear thermoelastic system*, in Proc. European Control Conference, Brussels, Belgium, 1997.

[6] G. Avalos and I. Lasiecka, *Uniform decay rates in nonlinear thermoelastic systems without mechanical dissipation*, in Proc. IFIP WG 7.2 Conference on Optimal Control: Theory, Algorithms, and Applications, Gainesville, FL, Kluwer Academic Publishers, Norwell, MA, February 1997.

[7] E. Bisognin, V. Bisognin, G. Perla Menzala, and E. Zuazua, *On the Exponential Stability for von Karman Equations in the Presence of Thermal Effects*, Preprint, 1996.
[8] P. Grisvard, *Caracterization de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40–63.
[9] S. Hansen, *Boundary control of a one-dimensional linear thermoelastic rod*, SIAM J. Control Optim., 32 (1994), pp. 1052–1074.
[10] J. U. Kim, *On the energy decay of a linear thermoelastic bar and plate*, SIAM J. Math. Anal., 23 (1992), pp. 889–899.
[11] V. Komornik, *Exact controllability and stabilization: The multiplier method*, Research in Applied Mathematics, John Wiley & Sons, New York, 1994.
[12] J. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Stud. Appl. Math. 10, SIAM, Philadelphia, PA, 1989.
[13] I. Lasiecka, *Control and stabilization of interactive structures*, in Systems and Control in the Twenty-First Century, Birkhäuser Boston, Boston, MA, 1997, pp. 245–263.
[14] I. Lasiecka and R. Triggiani, *Exact controllability and uniform stabilization of Kirchoff plates with boundary control only on* $\Delta\omega|_{\Sigma}$ *and homogeneous boundary displacement*, J. Differential Equations, 88 (1991), pp. 62–101.
[15] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.
[16] I. Lasiecka and R. Triggiani, *Sharp regularity theory for second-order hyperbolic equations of Neumann type*, Annali Mat. Pura & Appl. IV, 207 (1990), pp. 285–367.
[17] I. Lasiecka and R. Triggiani, *Recent advances in regularity of second-order hyperbolic mixed problems and applications*, in Dynamics Reported, Expositions in Dynamical Systems, C. K. R. T. Jones, U. Kirchgraber, and H. O. Wather, eds., Springer-Verlag, New York, 1994, pp. 104–162.
[18] J.L. Lions, *Contrôlabilité exacte, perturbations et stabilization de systèmes distribués*, Vol. 1, Masson, Paris, 1989.
[19] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.
[20] Z. Liu and M. Renardy, *A note on the equations of a thermoelastic plate*, Appl. Math. Lett., 8 (1995), pp. 1–6.
[21] Z. Liu and S. Zheng, *Exponential stability of the Kirchoff plate with thermal or viscoelastic damping*, Quarterly Appl. Math., 55 (1997), pp. 551–564.
[22] Z. Liu and S. Zheng, *Exponential stability of semigroup associated with a thermal elastic system*, Quarterly Appl. Math., 52 (1993), pp. 535–545.
[23] G. Perla Menzala and E. Zuazua, *Explicit exponential decay rates for solutions of von Karman's system of thermoelastic plates*, C.R. Acad. Sci. Paris, 324 (1997), pp. 49–54.
[24] K. Narukawa, *Boundary value control of thermoelastic systems*, Hiroshima Math. J., 13 (1983), pp. 227–272.
[25] J. E. M. Rivera, *Energy decay rate in linear thermoelasticity*, Funkcial. Ekvac., 35 (1992), pp. 19–30.
[26] D. Tataru, *On the Regularity of Boundary Traces for the Wave Equation*, Ann. Scuóla Normale, to appear.
[27] I. Lasiecka and R. Triggiani, *Analyticity and lack thereof, of thermo-elastic semigroups*, in Proc. Conference on Control of PDEs, CIRM, Luminy, France, June 1997.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 183–207, January 1998

011

STABILITY OF N-FRONTS BIFURCATING FROM A TWISTED HETEROCLINIC LOOP AND AN APPLICATION TO THE FITZHUGH–NAGUMO EQUATION*

BJÖRN SANDSTEDE†

Abstract. In this article, existence and stability of N-front travelling-wave solutions of partial differential equations on the real line is investigated. The N-fronts considered here arise as heteroclinic orbits bifurcating from a twisted heteroclinic loop in the underlying ordinary differential equation describing travelling-wave solutions. It is proved that the N-front solutions are linearly stable provided the fronts building the twisted heteroclinic loop are linearly stable. The result is applied to travelling waves arising in the FitzHugh–Nagumo equation.

Key words. heteroclinic orbits, stability, FitzHugh–Nagumo equation

AMS subject classifications. 34C37, 35B35, 58F14

PII. S0036141096297388

1. Introduction.

In this article, existence and stability of N-front solutions of parabolic equations

$$U_t = A\,U + F(U,\epsilon), \qquad x \in \mathbb{R} \tag{1.1}$$

on the real line is investigated. Here, the differential operator A generates a C^0-semiflow on $BU(\mathbb{R},\mathbb{R}^m)$—the space of bounded, uniformly continuous functions from $\mathbb{R}$ to $\mathbb{R}^m$—and F is a superposition operator—that is, $F(U,\epsilon)(x)$ depends only on the values of U and possibly derivatives of U at the point x—defined on the same space. Fronts and backs are travelling-wave solutions $U(\xi) = U(x + c\,t)$ which are asymptotically constant for $\xi \to \pm\infty$. Transforming (1.1) into a moving coordinate frame $(x,t) \mapsto (x + c\,t, t) = (\xi, t)$ yields

$$U_t = A\,U - c\,U_\xi + F(U,\epsilon), \qquad \xi \in \mathbb{R}. \tag{1.2}$$

Then fronts and backs of (1.1) with wave speed c correspond to equilibria of (1.2) solving

$$\begin{aligned} A\,U - c\,U_\xi + F(U,\epsilon) &= 0, \\ \lim_{\xi\to\pm\infty} U(\xi) &= U_\pm. \end{aligned} \tag{1.3}$$

Stability of a front U is often determined by the spectrum of the linearized operator

$$L(U)\,V = A\,V - c\,V_\xi + D_U F(U,\epsilon)\,V. \tag{1.4}$$

A front or back is called linearly stable if the spectrum of L is contained in the left half-plane with the exception of a simple eigenvalue at zero which is inevitable due

*Received by the editors January 16, 1996; accepted for publication (in revised form) August 29, 1996. This author was partially supported by a Feodor-Lynen Fellowship of the Alexander von Humboldt Foundation.

http://www.siam.org/journals/sima/29-1/29738/html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (sandsted@cfm.brown.edu). Permanent address: WIAS, Mohrenstraße 39, 10117 Berlin, Germany.

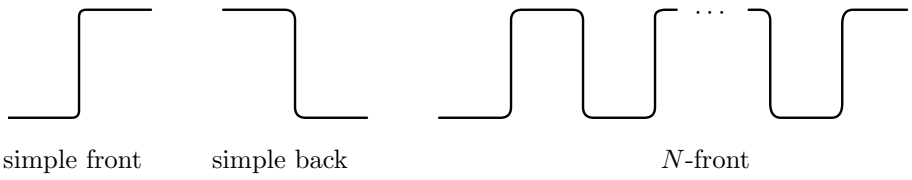

FIG. 1. *N-front solutions consist of* $2N+1$ *concatenated copies of a simple front and back.*

to translational invariance. Under rather general assumptions on A, linear stability implies nonlinear stability; see [Hen81] or [BJ89].

Suppose now that for $(c,\epsilon) = (c_0,\epsilon_0)$ linearly stable front and back waves do exist simultaneously. Then, upon varying $\mu := (c,\epsilon)$, other front solutions may arise. In particular, so-called N-fronts which are formed by alternately concatenating $2N+1$ copies of the simple front and back may bifurcate; see Figure 1. A natural and interesting question is whether the bifurcating N-fronts U_N inherit the linear stability from the simple front and back. For a fairly general class of operators A, it follows from [AGJ90] that the spectrum of $L(U_N)$ is bounded to the left of the imaginary axis except for $2N+1$ eigenvalues near zero. It therefore suffices to calculate these critical eigenvalues, that is, solutions (λ, V) of

$$A\,V - c_N\,V_\xi + D_U F(U_N,\epsilon_N)\,V = \lambda V \tag{1.5}$$

for λ close to zero, where U_N is the N-front existing for $(c,\epsilon) = (c_N,\epsilon_N)$.

Notice that the steady-state equation (1.3) and the eigenvalue problem (1.5) are ordinary differential equations in the time variable ξ. As such they can be written as first-order systems

$$\dot u = f(u,\mu), \qquad \mu = (c,\epsilon), \tag{1.6}$$

$$\dot v = (D_u f(u,\mu) + \lambda B)\,v, \tag{1.7}$$

respectively. Simple fronts and backs of (1.3) correspond to heteroclinic solutions $q_1(\xi)$ and $q_2(\xi)$ of (1.6) connecting two equilibria p_1 and p_2.

In this article, we investigate the existence and stability of N-fronts (and N-backs) under the assumption that the simple heteroclinic orbits q_1 and q_2 form a twisted heteroclinic loop; see Figure 2. Under certain generic assumptions, we prove existence of N-fronts of (1.6) for any $N \geq 1$ and determine all eigenvalues λ of (1.7) with $|\lambda|$ small. The N-fronts are either all stable or all unstable depending only on conditions on the simple front and back solution. The proof relies on a geometric reduction of the flow onto a two-dimensional invariant manifold containing the heteroclinic loop; see [Hom96], [San93], and [San95]. The reduction allows for a smooth linearization of the vector field near both equilibria. The existence of N-fronts is then proved using the Lyapunov–Schmidt reduction for the resulting vector field in $\mathbb{R}^2$ in the spirit of [Lin90] and [San93]. Finally, the critical eigenvalues of the operator (1.5) are calculated using [San96].

Deng [Den91a] proved the existence of N-fronts bifurcating from a twisted heteroclinic loop under the additional assumption that the stable manifolds of the relatively contractive equilibria p_1 and p_2 are one-dimensional using topological methods; see [Den91a, section 7(a)]. Shashkov [Sha92] asserts the existence of N-fronts for two-dimensional vector fields of class C^3, however, without giving a proof.

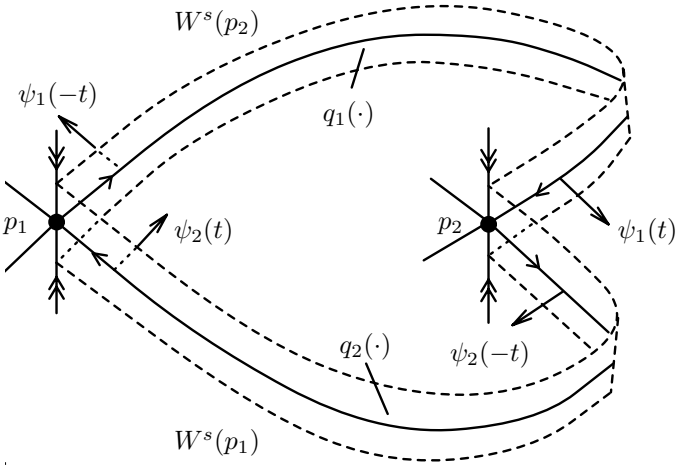

FIG. 2. *A twisted heteroclinic loop. The dashed curves indicate the stable manifolds of the equilibria continued backward in time.*

Finally, we apply the stability result to the FitzHugh–Nagumo equation

$$\begin{aligned} u_t &= u_{xx} + f(u) - w, \\ w_t &= \epsilon(u - \gamma w). \end{aligned}$$

Deng [Den91b] showed that the hypotheses of his existence result [Den91a] are satisfied, while Yanagida [Yan89] proved that the simple front and back are both linearly stable. Nii [Nii95b] proved linear stability of the 1-front provided f is linear near both equilibria. We show that in fact all N-fronts are linearly stable. Recently, Nii (personal communication) announced an extension of his result to N-fronts under the same restrictive hypothesis on f using topological methods.

The paper is organized as follows. In section 2, we state the basic assumptions and the main results about existence and stability of N-front solutions. The existence theorem is proved in section 3, the stability result in section 4. Finally, in section 5, the application to the FitzHugh–Nagumo system is given.

2. Main results. Consider the equation

$$\dot{u} = f(u, \mu), \qquad (u, \mu) \in \mathbb{R}^n \times \mathbb{R}^2, \tag{2.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^2 \to \mathbb{R}^n$ is C^2. We assume that equation (2.1) possesses two hyperbolic equilibria $p_1(\mu)$ and $p_2(\mu)$ for all μ. Moreover, the spectrum $\sigma(D_u f(p_k(\mu), \mu))$ for $k = 1, 2$ of the linearized vector field at these equilibria decomposes as follows.

(H1) *We assume that* $\dim W^s(p_1(0), 0) = \dim W^s(p_2(0), 0)$ *and*

$$\sigma(D_u f(p_k(\mu), \mu)) = \sigma_k^{ss} \cup \{-\alpha_k^s(\mu), \alpha_k^u(\mu)\} \cup \sigma_k^{uu}, \qquad 0 < \alpha_k^s(\mu) < \alpha_k^u(\mu)$$

hold with $\operatorname{Re}\sigma_k^{ss} < -\alpha_k^s(\mu)$, $\operatorname{Re}\sigma_k^{uu} > \alpha_k^u(\mu)$ *for* $k = 1, 2$ *and all* μ. *Moreover,* $-\alpha_k^s(\mu)$ *and* $\alpha_k^u(\mu)$ *are simple eigenvalues for* $k = 1, 2$. *We define* $\alpha_k(\mu) = \alpha_k^u(\mu)/\alpha_k^s(\mu) > 1$. *Also, let* $\alpha_k := \alpha_k(0)$ *and* $\alpha_k^i := \alpha_k^i(0)$ *for* $i = s, u$ *and* $k = 1, 2$.

We choose coordinates such that the equilibria do not depend on μ. Suppose that for $\mu = 0$ there exist two heteroclinic orbits $q_1(t)$ and $q_2(t)$ connecting p_1 to p_2 and vice versa; see (H2).

(H2) *The solution* $q_1(t)$ *satisfies* $\lim_{t\to-\infty} q_1(t) = p_1$ *and* $\lim_{t\to\infty} q_1(t) = p_2$ *while* $q_2(t)$ *satisfies* $\lim_{t\to-\infty} q_2(t) = p_2$ *and* $\lim_{t\to\infty} q_2(t) = p_1$.

Due to hypothesis (H1), the next assumption is met for generic vector fields.

(H3) *The heteroclinic solutions* $q_1(t)$ *and* $q_2(t)$ *are nondegenerate, that is,*

$$\begin{aligned} T_{q_1(0)}W^u(p_1,0) \cap T_{q_1(0)}W^s(p_2,0) &= \mathbb{R}\dot{q_1}(0), \\ T_{q_2(0)}W^u(p_2,0) \cap T_{q_2(0)}W^s(p_1,0) &= \mathbb{R}\dot{q_2}(0) \end{aligned}$$

hold.

Due to (H3), there exist two unique (up to constant multiples) bounded solutions $\psi_k(t)$ of the adjoint variational equation

$$\dot{w} = -D_u f(q_k(t),0)^*\, w$$

evaluated at $q_k(t)$ for $k = 1, 2$, respectively. As a matter of fact, they satisfy

$$\psi_k(t) \perp \big(T_{q_k(t)}W^u(p_k,0) + T_{q_k(t)}W^s(p_{k+1},0)\big). \tag{2.2}$$

Here, the index k is taken modulo two. Upon changing the parameter μ, the heteroclinic solutions $q_k(t)$ should break up. This is made precise in the next hypothesis.

(H4) *The Melnikov integrals*

$$N_k := \int_{-\infty}^{\infty} \langle \psi_k(t), D_\mu f(q_k(t),0)\rangle\, dt \in \mathbb{R}^2, \qquad k = 1, 2$$

are linearly independent (and in particular nonzero).

We need to assume that $q_k(t)$ and $\psi_k(t)$ converge along the leading directions to the equilibria and zero, respectively.

(H5) *Assume that the limits*

$$\begin{aligned} \lim_{t\to-\infty} e^{-\alpha_k^u t}\,\dot{q}_k(t) &=: v_k^-, & \lim_{t\to\infty} e^{\alpha_{k+1}^s t}\,\dot{q}_k(t) &=: v_{k+1}^+, \\ \lim_{t\to-\infty} e^{-\alpha_k^s t}\,\psi_k(t) &=: w_k^+, & \lim_{t\to\infty} e^{\alpha_{k+1}^u t}\,\psi_k(t) &=: w_{k+1}^- \end{aligned}$$

are nonzero for $k = 1, 2$*; see Figure* 2*. Again, the index* k *is taken modulo two.*

Then $v_k^\pm$ and $w_k^\pm$ are right and left eigenvectors of $D_u f(p_k, 0)$ for the eigenvalues $\alpha_k^{s,u}$. Due to (2.2), hypothesis (H5) is equivalent to the strong inclination property. Finally, we suppose that both heteroclinic orbits are twisted.

(H6) *Suppose that the scalar products* $\langle w_k^-, v_k^-\rangle > 0$ *and* $\langle w_k^+, v_k^+\rangle > 0$ *are positive for* $k = 1, 2$*; see Figure* 2*. Note that the scalar products do not vanish according to hypotheses* (H1) *and* (H5).

Choose two sections Σ_k transverse to the vector field and placed at $q_k(0)$ for $k = 1, 2$. We call the heteroclinic solutions $q_1(t)$ and $q_2(t)$ simple fronts and backs, respectively. An N-front solution is a heteroclinic orbit connecting p_1 to p_2 and intersecting Σ_2 N-times; see Figure 3. In other words, it follows the heteroclinic loop $N+\frac{1}{2}$ times and hits the set $\Sigma_1 \cup \Sigma_2$ $2N+1$ times. Similarly, N-backs are defined connecting p_2 to p_1.

Associated with each N-front are $2N$ return times T_j for $j = 0, \ldots, 2N-1$. With $l = 0, \ldots, N-1$, the numbers T_{2l} are the times consecutively spent between Σ_1 and Σ_2, that is, near the equilibrium p_2, while T_{2l+1} are the times spent between Σ_2 and Σ_1, that is, near the equilibrium p_1.

We remark that, on account of hypothesis (H4), there is a change of parameters of class C^2 such that the Melnikov integrals coincide with the coordinate axes

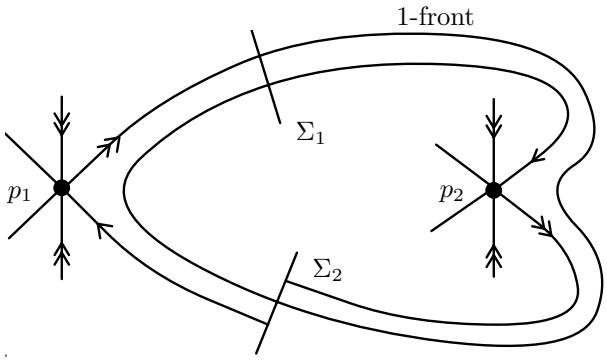

FIG. 3. *An N-front solution for $N = 1$. There are two return times T_0 and T_1 associated with the 1-front. T_0 is the time spent between Σ_1 and Σ_2, while T_1 is the time spent between Σ_2 and Σ_1.*

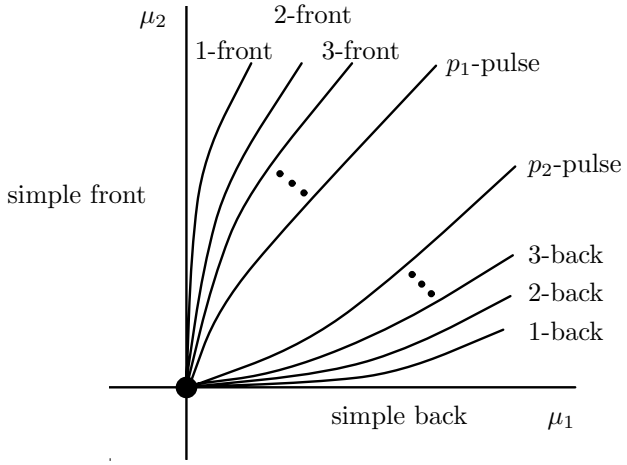

FIG. 4. *The bifurcation diagram. The parameters have been transformed such that a simple front (back) exists precisely when $\mu_1 = 0$ ($\mu_2 = 0$) and such that the Melnikov integrals are given by $N_1 = (1, 0)$ and $N_2 = (0, 1)$. The twisted heteroclinic loop exists for $\mu = 0$.*

in $\mathbb{R}^2$, that is, $N_1 = (1, 0)$ and $N_2 = (0, 1)$, and that simple fronts (backs) exist precisely when $\mu_1 = 0$ ($\mu_2 = 0$); see Figure 4. We refer to section 3.2 for the proof.

The first result is an extension of the existence theorem proved by Deng [Den91a], who assumed that the unstable manifolds are one-dimensional.

THEOREM 2.1. *Assume that* (H1)–(H6) *are satisfied. Then, for each $N \geq 1$, there exists a unique curve $\bar{\mu}_N(r)$ in parameter space defined for $r \in [0, r_0)$ with $\bar{\mu}_N(0) = 0$ such that* (2.1) *has an N-front solution (u, μ) if and only if there exists an r such that $\mu = \bar{\mu}_N(r)$. The N-fronts are unique and the curves $\bar{\mu}_N$ are of class C^1. The return times (as defined right before this theorem) of the N-fronts are given by*

$$
\begin{aligned}
T_{2l} &= -\tfrac{1}{\alpha_2^s}\,(1 + \mathrm{o}(1))\,\ln r, && \text{time spent near } p_2,\\
T_{2l+1} &= -\tfrac{\alpha_2+\theta_{N-l}}{\alpha_1^s}\,(1 + \mathrm{o}(1))\,\ln r, && \text{time spent near } p_1
\end{aligned}
\tag{2.3}
$$

for $l = 0, \dots, N-1$ as $r \to 0$. Here, the sequence θ_l is defined recursively by $\theta_1 = 0$, $\theta_2 = \alpha_1\alpha_2 - 1 > 0$, and $\theta_{l+1} := \alpha_1\theta_l + \theta_2 > \theta_l$. Different choices of the sections Σ_k do not change the leading order term in (2.3)*. Analogous results hold for N-backs.*

Now, assume in addition that parameters have been transformed according to the remark stated right before this theorem. Then the curves $\bar\mu_N$ satisfy

$$\bar\mu_N(r) = (r^{\alpha_2}(1 + \mathrm{o}(1)), r)$$

as $r \to 0$, and the bifurcation diagram is given in Figure 4*.*

Next we describe the bounded solutions $v \in C^1(\mathbb{R}, \mathbb{R}^n)$ of the equation

$$\dot v = \big(D_u f(q_N(r)(t), \bar\mu_N(r)) + \lambda B(t)\big)v \tag{2.4}$$

for $\lambda \in \mathbb{C}$ with $|\lambda|$ small, where $q_N(r)$ denotes the N-front existing for $\mu = \bar\mu_N(r)$. Here, B is a bounded, continuous, and matrix-valued function. Equation (2.4) is a generalized eigenvalue problem of the form

$$Lv = \lambda Bv.$$

Generalized eigenfunctions of (2.4) corresponding to an eigenvalue λ are functions v_i satisfying

$$Lv_i = \lambda Bv_i + Bv_{i-1}$$

with $v_0 = 0$. The algebraic multiplicity of eigenvalues can be defined in the usual way. We assume a nondegeneracy assumption with respect to λ.

(H7) *Suppose that the Melnikov integrals*

$$M_k := \int_{-\infty}^{\infty} \langle \psi_k(t), B(t)\,\dot q_k(t)\rangle\, dt \neq 0$$

are nonzero for $k = 1, 2$, where ψ_k is chosen according to hypothesis (H6)*.*

The next theorem—which is the main result of the present paper—describes the set of $\lambda \in \mathbb{C}$ with $|\lambda|$ small for which (2.4) possesses a bounded solution v.

THEOREM 2.2. *Suppose that the assumptions* (H1)–(H7) *are satisfied. Then there exists a $\delta > 0$ independent of N such that the following holds. For any $N \geq 1$ and $r_0 = r_0(N) > 0$ sufficiently small there exist precisely $2N+1$ solutions $(\lambda_j, v_j) \in \mathbb{C} \times C^1(\mathbb{R}, \mathbb{R}^n)$ of* (2.4) *with $|\lambda| < \delta$. The eigenvalues are counted with multiplicity and are given by*

$$\begin{aligned} \lambda_{2l-1} &= (c_{2l-1} + \mathrm{o}(1))\, r, \\ \lambda_{2l} &= (c_{2l} + \mathrm{o}(1))\, r^{\alpha_2 + \theta_{N+1-l}}, \\ \lambda_{2N+1} &= 0 \end{aligned}$$

for $l = 1, \dots, N$ as $r \to 0$, where the exponents θ_{N+1-l} have been defined in Theorem 2.1*.*

The constants c_j are nonzero and satisfy $\operatorname{sign} c_{2l} = \operatorname{sign} M_1$ *and* $\operatorname{sign} c_{2l-1} = \operatorname{sign} M_2$*. In particular, the eigenvalues λ_j are contained in the left half-plane for $j = 1, \dots, 2N$ provided $M_1, M_2 < 0$ are negative. Analogous results hold for N-backs.*

The second theorem establishes stability of the N-front solutions with respect to the underlying partial differential equation; see section 5 for an example.

Notice that there exist precisely two pulses converging to p_1 and p_2, respectively; see Figure 4. The existence proof is implicitly contained in section 3.3. As far as their stability is concerned, the same statement as for the N-fronts holds. This follows from [Nii95a] or section 4 of the present article.

3. Existence. In order to prove existence of N-fronts, a geometric reduction onto a two-dimensional invariant manifold in phase space is employed. The manifold is diffeomorphic to an annulus. Next, a system of $2N+1$ equations is derived using the Lyapunov–Schmidt reduction applied to the flow on the invariant manifold. In the final section, this system is being solved for using an implicit function theorem.

Throughout we assume that hypotheses (H1)–(H6) are met.

3.1. Center-manifold reduction. We have the following lemma.

LEMMA 3.1. *There exists a two-dimensional, locally invariant, and normally hyperbolic manifold $W^c_{hom} \subset \mathbb{R}^n$ of class $C^{1,\rho}$ jointly in (u,μ) for some $\rho > 0$. All solutions staying near the heteroclinic loop for all times and for parameter values close to zero are contained in W^c_{hom}. The manifold is homeomorphic to an annulus.*

Moreover, the flow restricted to W^c_{hom} is C^1-conjugated to the flow of an appropriate vector field $g(u,\mu)$ of class C^1 defined on $\mathbb{R}^2$. The hypotheses (H1)–(H6) *are still satisfied for g and, in addition, g is linear locally near both equilibria.*

Proof. The existence of W^c_{hom} is an application of [San95, Theorem 1]. We shall verify the assumptions of that theorem using the decomposition

$$\sigma(D_u f(p_k,0)) = \sigma_k^{ss} \cup \sigma_k^c \cup \sigma_k^{uu}, \qquad \sigma_k^c = \{-\alpha_k^s, \alpha_k^u\}.$$

Then [San95, (H1), ($\widetilde{\text{H3}}$)] are satisfied due to (H1) and (H5), while [San95, (H4)] is void. It remains to verify [San95, ($\widetilde{\text{H2}}$)] which reads

$$\begin{aligned} T_{q_1(0)}W^{uu}(p_1) \oplus T_{q_1(0)}W^{u,s,ss}(p_2) &= \mathbb{R}^n, \\ T_{q_1(0)}W^{s,u,uu}(p_1) \oplus T_{q_1(0)}W^{ss}(p_2) &= \mathbb{R}^n, \end{aligned}$$

and the analogous condition for $q_2(t)$. Here, $W^{u,s,ss}(p_2)$ denotes an invariant manifold tangent to the generalized eigenspace $E^{u,s,ss}$ associated with $\sigma_2^{ss} \cup \sigma_2^c$ at p_2 and similarly for $W^{s,u,uu}(p_1)$. On account of (H1), it suffices to prove that

$$\begin{aligned} T_{q_1(0)}W^{uu}(p_1) \cap T_{q_1(0)}W^{u,s,ss}(p_2) &= \{0\}, \\ T_{q_1(0)}W^{s,u,uu}(p_1) \cap T_{q_1(0)}W^{ss}(p_2) &= \{0\}. \end{aligned} \tag{3.1}$$

We have

$$T_{q_1(0)}W^{u,s,ss}(p_2) = T_{q_1(0)}W^{s}(p_2) \oplus \mathbb{R}v^u$$

for some nonzero v^u. On account of (H3) and (H5), the intersection

$$T_{q_1(0)}W^{uu}(p_1) \cap T_{q_1(0)}W^{s}(p_2) = \{0\}$$

is trivial. Therefore, if the first equation of (3.1) does not hold, there exists a vector $w \in T_{q_1(0)}W^s(p_2)$ such that

$$v^u + w \in T_{q_1(0)}W^{uu}(p_1) \cap T_{q_1(0)}W^{u,s,ss}(p_2).$$

Let $v^u(t)$ and $w(t)$ be the solutions of the variational equation along $q_1(t)$ satisfying $v^u(0) = v^u$ and $w(0) = w$. Choose $q_1(0)$ close to p_2, so that $T_{q_1(0)}W^{u,s,ss}(p_2)$ is close to $E^{u,s,ss}$. Then, due to (H5), $\langle \psi_1(0), v^u \rangle \neq 0$. However, the solution $v^u(t) + w(t) \in T_{q_1(t)}W^{uu}(p_1)$ decays exponentially to zero for $t \to -\infty$, while

$$\langle \psi_1(t), v^u(t) + w(t) \rangle \overset{(2.2)}{=} \langle \psi_1(t), v^u(t) \rangle = \langle \psi_1(0), v^u(0) \rangle \neq 0$$

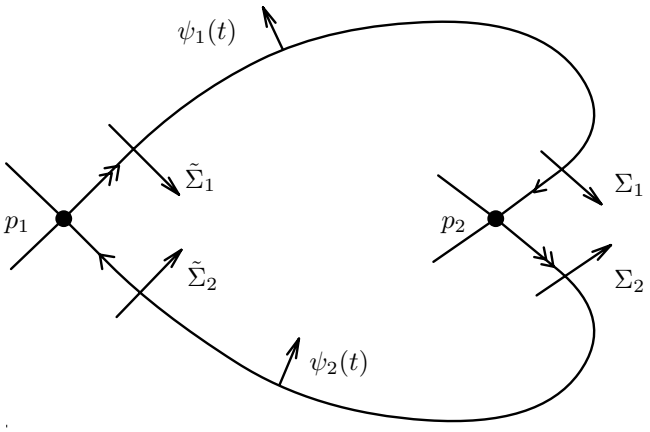

FIG. 5. *The choice of the sections in* $\mathbb{R}^2$. *The arrows denote the positive direction once sections are identified with intervals in* $\mathbb{R}$.

is independent of t as $\psi_1(t)$ solves the adjoint equation. This is a contradiction to $\psi_1(t)$ being bounded; thus

$$T_{q_1(0)}W^{uu}(p_1) \cap T_{q_1(0)}W^{u,s,ss}(p_2) = \{0\}.$$

The argument for the second equation of (3.1) is similar. Thus we can apply [San95, Theorem 1] to conclude the existence of an invariant manifold W^c_{hom}. Moreover, by the construction in [San95], W^c_{hom} is homeomorphic to an annulus owing to (H6). That the flow on W^c_{hom} is C^1-conjugated to the flow of a C^1-vector field in $\mathbb{R}^2$ follows from [San95, section 3.5]. The statement about the smooth linearization is proved in [Hom96, Proposition A.1.1]. □

Hence we can restrict the analysis to a C^1-vector field g in $\mathbb{R}^2$ satisfying ($\widetilde{\text{H1}}$), (H2)–(H6), and being linear locally near both equilibria, where hypothesis ($\widetilde{\text{H1}}$) is given by the following.

($\widetilde{\text{H1}}$) *We assume that* $\dim W^s(p_1, 0) = \dim W^s(p_2, 0) = 1$ *and*

$$\sigma(D_u f(p_k(\mu), \mu)) = \{-\alpha^s_k(\mu), \alpha^u_k(\mu)\}, \qquad 0 < \alpha^s_k(\mu) < \alpha^u_k(\mu)$$

hold for $k = 1, 2$. *We define* $\alpha_k(\mu) = \alpha^u_k(\mu)/\alpha^s_k(\mu) > 1$.

3.2. Lin's method in $\mathbb{R}^2$. According to the last section, it suffices to consider a vector field

$$\dot{u} = g(u, \mu), \qquad (u, \mu) \in \mathbb{R}^2 \times \mathbb{R}^2, \tag{3.2}$$

with $g \in C^1$ such that ($\widetilde{\text{H1}}$) and (H2) up to (H6) are satisfied and the flow near the equilibria p_k for $k = 1, 2$ is linear. Choose Poincaré sections Σ_k and $\tilde{\Sigma}_k$ for $k = 1, 2$ as in Figure 5. All sections are chosen inside the regions near the equilibria p_k where the flow is linear. Moreover, we shall identify the one-dimensional sections with intervals in $\mathbb{R}$ as shown in Figure 5. Next, we compute various Poincaré maps. Denoting the time spent between Σ_1 and Σ_2 by T, the map from Σ_1 to Σ_2 is given by

$$\begin{array}{ccc} \Sigma_1 & \to & \Sigma_2, \\ e^{-\alpha^u_2(\mu)T} & \mapsto & e^{-\alpha^s_2(\mu)T}, \end{array} \tag{3.3}$$

using the fact that the vector field is linear. Similarly, the map from $\tilde{\Sigma}_2$ to $\tilde{\Sigma}_1$ equals

$$\begin{array}{ccc} \tilde{\Sigma}_2 & \to & \tilde{\Sigma}_1, \\ e^{-\alpha_1^u(\mu)\tau} & \mapsto & e^{-\alpha_1^s(\mu)\tau}. \end{array} \tag{3.4}$$

The maps

$$\begin{array}{rccl} \tilde{\Pi}_k(u,\mu) & : & \tilde{\Sigma}_k & \to & \Sigma_k, \\ & & u & \mapsto & -\Pi_k(u,\mu) + d_k(\mu) \end{array} \tag{3.5}$$

are diffeomorphisms with $\Pi_k(u,\mu) \in C^1$, $\Pi_k(0,\mu) = 0$, and $D_u\Pi_k(0,\mu) > 0$ for $k = 1, 2$. The minus sign before the term $\Pi_k(u,\mu)$ appearing in (3.5) is a consequence of hypothesis (H6); see Figure 5. Due to hypothesis (H4), we may assume that $d_k(\mu) = \mu_k$ by a C^1-transformation of parameters. Indeed, with $u_k^u(\mu) \in W^u(p_k,\mu)$ and $u_k^s(\mu) \in W^s(p_k,\mu)$ chosen properly, and after possibly shifting time such that $q_k(0) \in \Sigma_k$, the separation function $d_k(\mu) = \langle \psi_k(0), u_k^u(\mu) - u_{k+1}^s(\mu)\rangle$ measures the signed distance of the one-dimensional stable and unstable manifolds of the equilibria at the section Σ_k; see, for instance, [Kok88], [Lin90], or [Den91b]. The integrals N_k appearing in (H4) are in fact the derivatives of $d_k(\mu)$ at $\mu = 0$. For consistency with the remark stated before Theorem 2.1, we note that parameters can be changed in the original system before applying the center-manifold reduction resulting in a C^2-transformation.

Summarizing the above, we obtain a map

$$\begin{array}{ccc} \Sigma_2 & \to & \Sigma_1, \\ -\Pi_2(e^{-\alpha_1^u(\mu)\tau},\mu) + \mu_2 & \mapsto & -\Pi_1(e^{-\alpha_1^s(\mu)\tau},\mu) + \mu_1. \end{array} \tag{3.6}$$

All solutions being mapped from Σ_2 to Σ_1 are captured by the above parametrization. The next step consists in formulating the Poincaré map by means of the return time with respect to the sections Σ_k instead of the one for $\tilde{\Sigma}_k$.

The times needed for initial points $u \in \tilde{\Sigma}_k$ to reach the sections Σ_k are given by functions $\Omega_k(u,\mu)$. Both functions $\Omega_k(u,\mu)$ are in C^1 and bounded uniformly in u. Thus the time T needed for the initial point

$$-\Pi_2(e^{-\alpha_1^u(\mu)\tau},\mu) + \mu_2 \in \Sigma_2$$

to reach

$$-\Pi_1(e^{-\alpha_1^s(\mu)\tau},\mu) + \mu_1 \in \Sigma_1$$

is given by

$$T = \tau + \Omega_1(e^{-\alpha_1^s(\mu)\tau},\mu) + \Omega_2(e^{-\alpha_1^u(\mu)\tau},\mu).$$

By the implicit function theorem, we can solve this equation with respect to τ yielding a C^1-function $\tau(T,\mu)$, with

$$\tau(T,\mu) = T - \Omega_1(e^{-\alpha_1^s(\mu)\,\tau(T,\mu)},\mu) - \Omega_2(e^{-\alpha_1^u(\mu)\,\tau(T,\mu)},\mu). \tag{3.7}$$

Therefore, we obtain the following lemma.

LEMMA 3.2. *The Poincaré maps from Σ_1 to Σ_2 and vice versa are given by*

$$\begin{array}{ccc} \Sigma_1 & \to & \Sigma_2, \\ e^{-\alpha_2^u(\mu)T} & \mapsto & e^{-\alpha_2^s(\mu)T} \end{array} \tag{3.8}$$

and

$$\begin{array}{ccc} \Sigma_2 & \to & \Sigma_1, \\ -\Pi_2(e^{-\alpha_1^u(\mu)\,\tau(T,\mu)},\mu)+\mu_2 & \mapsto & -\Pi_1(e^{-\alpha_1^s(\mu)\,\tau(T,\mu)},\mu)+\mu_1, \end{array} \tag{3.9}$$

respectively. The C^1-function $\tau(T,\mu)$ defined in (3.7) *satisfies*

$$\Big|\frac{d}{dT}\tau(T,\mu)-1\Big| \ll 1,$$

and the C^1-maps $\Omega_k(u,\mu)$ are bounded uniformly in u. Moreover, $\Pi_k(u,\mu)\in C^1$, $\Pi_k(0,\mu)=0$, and $D_u\Pi_k(0,\mu)>0$ for $k=1,2$. Up to this point, the construction looks pretty much like using Shilnikov variables. However, in order to describe solutions following the original heteroclinic loop several times, we shall adopt a boundary-value-point-of-view. That is, we are not going to iterate the Poincaré maps given in the previous lemma, but shall derive matching conditions in the sections.

Using Lemma 3.2, the existence of N-front solutions is equivalent to the existence of return times $T_j<\infty$ for $j=0,\dots,2N-1$ and parameter values μ such that

$$\begin{array}{rcll} e^{-\alpha_2^u(\mu)T_0} &=& \mu_1, & \\ e^{-\alpha_2^s(\mu)T_{2j}} &=& -\Pi_2(e^{-\alpha_1^u(\mu)\,\tau(T_{2j+1},\mu)},\mu)+\mu_2, & j=0,\dots,N-1, \\ e^{-\alpha_2^u(\mu)T_{2j}} &=& -\Pi_1(e^{-\alpha_1^s(\mu)\,\tau(T_{2j-1},\mu)},\mu)+\mu_1, & j=1,\dots,N-1, \\ 0 &=& -\Pi_1(e^{-\alpha_1^s(\mu)\,\tau(T_{2N-1},\mu)},\mu)+\mu_1 & \end{array} \tag{3.10}$$

holds. Indeed, then the various pieces of solutions defined in between the sections will fit together. Moreover, the first and last equation assert that the solution is contained in the unstable and stable manifolds of the equilibria p_1 and p_2, respectively. In fact, T_{2j+1} and T_{2j} are the times spent near the equilibria p_1 and p_2, respectively. Define

$$\begin{array}{rclcrcl} a_{2j+1}\,s &=& e^{-\alpha_1^s(\mu)\,\tau(T_{2j+1},\mu)}, & & s &=& e^{-\alpha_1^s(\mu)\,\tau(T_{2N-1},\mu)}, \\ a_{2j}\,r &=& e^{-\alpha_2^s(\mu)T_{2j}}, & & r &=& e^{-\alpha_2^s(\mu)T_0} \end{array} \tag{3.11}$$

for $j=0,\dots,N-1$ such that $a_0=a_{2N-1}=1$ and $a_1,\dots,a_{2N-2}$ are bounded. In the new variables a_j, r, and s, equation (3.10) reads

$$\begin{array}{rcll} r^{\alpha_2(\mu)}-\mu_1 &=& 0, & \\ r+\Pi_2((a_1 s)^{\alpha_1(\mu)},\mu)-\mu_2 &=& 0, & \\ (a_{2j}r)^{\alpha_2(\mu)}+\Pi_1(a_{2j-1}s,\mu)-\mu_1 &=& 0, & j=1,\dots,N-1, \\ a_{2j}r+\Pi_2((a_{2j+1}s)^{\alpha_1(\mu)},\mu)-\mu_2 &=& 0, & j=1,\dots,N-1, \\ \Pi_1(s,\mu)-\mu_1 &=& 0 & \end{array} \tag{3.12}$$

with $\alpha_k(\mu)=\alpha_k^u(\mu)/\alpha_k^s(\mu)>1$. Whenever (a_j,r,s) solve (3.12) such that $a_j>0$ and $r,s>0$, we obtain associated return times $T_j<\infty$ which solve (3.10) by using (3.11). Indeed, we have

$$\begin{array}{rcl} \tau(T_{2j+1},\mu) &=& -\frac{1}{\alpha_1^s(\mu)}\ln(a_{2j+1}s), \\ T_{2j} &=& -\frac{1}{\alpha_2^s(\mu)}\ln(a_{2j}r), \end{array} \tag{3.13}$$

and Lemma 3.2 implies that $\tau(T,\mu)$ is invertible with respect to T. Hence, it suffices to consider (3.12) keeping in mind that only positive solutions of this system correspond to solutions of the original problem.

3.3. Existence of N-fronts bifurcating from a twisted heteroclinic cycle. We shall solve (3.12). Note that the functions Π_1 and Π_2 are in C^1. By convention, for $\alpha > 1$, define x^α to be zero for negative values of x yielding a C^1-function, too. Then (3.12) is defined for all a_j bounded and r, s small including negative values. Throughout this section, the range of the index j is $j = 1, \dots, N-1$.

First, solve

$$\begin{aligned} \mu_1 &= r^{\alpha_2(\mu)}, \\ \Pi_1(s,\mu) &= r^{\alpha_2(\mu)} \end{aligned} \tag{3.14}$$

with respect to (μ_1, s) near $(r, s, \mu) = 0$ by the implicit function theorem using Lemma 3.2. Denote the solutions by $\mu_1(\mu_2, r)$ and $s(\mu_2, r)$, both of which are of class C^1. Observe that, owing to $\Pi_1(0, \mu) = 0$, the estimates

$$|s(\mu_2, r)|, |D_{\mu_2} s(\mu_2, r)| \le C_\delta\, r^{\alpha_2 - \delta} \tag{3.15}$$

hold for arbitrary small positive δ. Using the ansatz $\mu_2 = \epsilon r$, the second equation in (3.12) reads

$$r + \Pi_2((a_1 s)^{\alpha_1(\mu)}, \mu) - \mu_2 = r + \Pi_2\big((a_1 s(\epsilon r, r))^{\alpha_1(\epsilon r, r)}, \mu_1(\epsilon r, r), \epsilon r\big) - \epsilon r = 0. \tag{3.16}$$

Here and in the following, we will be a bit sloppy concerning the dependence of $\alpha_k(\mu)$ and Π_k on ϵ and r to avoid unnecessary complicated notation. Dividing (3.16) by r yields

$$1 + r^{-1}\Pi_2\big((a_1 s(\epsilon r, r))^{\alpha_1(\epsilon r, r)}, \mu_1(\epsilon r, r), \epsilon r\big) - \epsilon = 0, \tag{3.17}$$

which is C^1 in (ϵ, a_1) for $r \ge 0$ owing to (3.15) and since the dependence on ϵ is due to $\mu_2 = \epsilon r$. Using (3.15), we can solve (3.17) with respect to ϵ near $\epsilon = 1$, $r = 0$, and arbitrary bounded a_1 yielding a C^1-function

$$\epsilon = \epsilon(a_1, r) = 1 + r^{-1}\Pi_2\big((a_1 \tilde{s}(a_1, r))^{\tilde{\alpha}_1(a_1, r)}, \tilde{\mu}_1(a_1, r), \epsilon(a_1, r) r\big), \tag{3.18}$$

where

$$\begin{aligned} \tilde{s}(a_1, r) &= s(\epsilon(a_1, r) r, r), \\ \tilde{\alpha}_k(a_1, r) &= \alpha_k(\epsilon(a_1, r) r, r), \\ \tilde{\mu}_1(a_1, r) &= \mu_1(\epsilon(a_1, r) r, r). \end{aligned}$$

Notice that the dependence of all these functions on a_1 is due to terms of the form $\epsilon(a_1, r) r$. It remains to solve the system

$$\begin{aligned} \Pi_1\big(a_{2j-1}\tilde{s}(a_1, r), \tilde{\mu}(a_1, r)\big) + (a_{2j} r)^{\tilde{\alpha}_2(a_1, r)} - \tilde{\mu}_1(a_1, r) &= 0, \\ a_{2j} r + \Pi_2\big((a_{2j+1}\tilde{s}(a_1, r))^{\tilde{\alpha}_1(a_1, r)}, \tilde{\mu}(a_1, r)\big) - \epsilon(a_1, r) r &= 0 \end{aligned}$$

for $j = 1, \dots, N-1$. Dividing by $r^{\tilde{\alpha}_2(a_1, r)}$ and r, respectively, yields

$$\begin{aligned} r^{-\tilde{\alpha}_2(a_1, r)}\, \Pi_1\big(a_{2j-1}\tilde{s}(a_1, r), \tilde{\mu}(a_1, r)\big) + a_{2j}^{\tilde{\alpha}_2(a_1, r)} - 1 &= 0, \\ a_{2j} + r^{-1}\Pi_2\big((a_{2j+1}\tilde{s}(a_1, r))^{\tilde{\alpha}_1(a_1, r)}, \tilde{\mu}(a_1, r)\big) - \epsilon(a_1, r) &= 0. \end{aligned} \tag{3.19}$$

The functions

$$\begin{aligned} &r^{-\tilde{\alpha}_2(a_1, r)}\, \Pi_1\big(a_{2j-1}\tilde{s}(a_1, r), \tilde{\mu}(a_1, r)\big), \\ &r^{-1}\Pi_2\big((a_{2j-1}\tilde{s}(a_1, r))^{\tilde{\alpha}_1(a_1, r)}, \tilde{\mu}(a_1, r)\big) \end{aligned}$$

are C^1 in (a_{2j-1}, a_1) up to $r = 0$ owing to (3.14) and the above comment about the dependence on a_1. Moreover, the derivative with respect to a_{2j-1} at $r = 0$ equals one for the first and zero for the second function. Therefore, $a_{2j} = 1$ and $a_{2j-1} = 0$ for $j = 1, \ldots, N-1$ solve (3.19) with $r = 0$, and we can use the implicit function theorem to obtain solutions $a_{2j}(r)$ and $a_{2j-1}(r)$ for positive r.

It remains to show that $a_{2j-1}(r) > 0$ is positive for $r > 0$. Define constants γ_j recursively by

$$\begin{aligned} \gamma_N &:= 0, \\ \gamma_{N-1} &:= \alpha_1\alpha_2 - 1 > 0, \\ \gamma_{j-1} &:= \alpha_1\gamma_j + \gamma_{N-1} > \gamma_j \end{aligned} \tag{3.20}$$

for $j = 1, \ldots, N-1$. These constants are related to the numbers θ_j via

$$\gamma_j = \theta_{N+1-j}. \tag{3.21}$$

We will, however, work with the γ_j in order to keep the notation simpler. Let

$$\begin{aligned} a_{2j-1} &= b_{2j-1}\, r^{\gamma_j}, \\ a_{2j} &= 1 - b_{2j}\, r^{\gamma_j} \end{aligned} \tag{3.22}$$

for $j = 1, \ldots, N-1$, and set $b_{2N-1} = 1$. Substituting these expressions together with (3.18) into equation (3.19) yields

$$\begin{aligned} 0 &= r^{-\hat\alpha_2(b_1,r)}\,\Pi_1\big(b_{2j-1}r^{\gamma_j}\,\hat s(b_1,r), \hat\mu(b_1,r)\big) - 1 + (1 - b_{2j}r^{\gamma_j})^{\hat\alpha_2(b_1,r)}, \\ 0 &= b_{2j}r^{\gamma_j} + r^{-1}\big(\Pi_2\big((b_1r^{\gamma_1}\hat s(b_1,r))^{\hat\alpha_1(b_1,r)}, \hat\mu(b_1,r)\big) \\ &\quad -\Pi_2\big((b_{2j+1}r^{\gamma_{j+1}}\hat s(b_1,r))^{\hat\alpha_1(b_1,r)}, \hat\mu(b_1,r)\big)\big), \end{aligned}$$

where

$$\begin{aligned} \hat s(b_1,r) &= s(\epsilon(b_1r^{\gamma_1},r)r,r), \\ \hat\alpha_k(b_1,r) &= \alpha_k(\epsilon(b_1r^{\gamma_1},r)r,r), \\ \hat\mu_1(b_1,r) &= \mu_1(\epsilon(b_1r^{\gamma_1},r)r,r), \\ \hat\mu_2(b_1,r) &= \epsilon(b_1r^{\gamma_1},r)r. \end{aligned} \tag{3.23}$$

Dividing these equations by r^{γ_j} reads

$$\begin{aligned} 0 &= r^{-(\hat\alpha_2(b_1,r)+\gamma_j)}\,\Pi_1\big(b_{2j-1}r^{\gamma_j}\,\hat s(b_1,r), \hat\mu(b_1,r)\big) \\ &\quad + r^{-\gamma_j}\big((1 - b_{2j}r^{\gamma_j})^{\hat\alpha_2(b_1,r)} - 1\big), \\ 0 &= b_{2j} + r^{-(1+\gamma_j)}\big(\Pi_2\big((b_1r^{\gamma_1}\hat s(b_1,r))^{\hat\alpha_1(b_1,r)}, \hat\mu(b_1,r)\big) \\ &\quad -\Pi_2\big((b_{2j+1}r^{\gamma_{j+1}}\hat s(b_1,r))^{\hat\alpha_1(b_1,r)}, \hat\mu(b_1,r)\big)\big). \end{aligned} \tag{3.24}$$

As before, using the recursive relations (3.20), it is tedious but straightforward to see that the functions appearing in (3.24) are C^1 up to $r = 0$. Moreover, for $r = 0$, (3.24) boils down to

$$\begin{aligned} b_{2i-1} - \alpha_2\, b_{2i} &= 0, && i = 1, \ldots, N-1, \\ b_{2i} - D_u\Pi_2(0,0)\, D_u\Pi_1(0,0)^{-\alpha_1}\, b_{2i+1}^{\alpha_1} &= 0, && i = 1, \ldots, N-2, \\ b_{2N-2} - D_u\Pi_2(0,0)\, D_u\Pi_1(0,0)^{-\alpha_1} &= 0 \end{aligned} \tag{3.25}$$

owing to (3.14). It is straightforward to check that the Jacobian of (3.25) with respect to (b_j) is upper triangular with nonzero diagonal elements. Equation (3.24) can therefore be solved near

$$\begin{aligned} b_{2N-2} &= D_u\Pi_2(0,0)\, D_u\Pi_1(0,0)^{-\alpha_1}, \\ b_{2i-1} &= \alpha_2\, b_{2i}, && i = 1,\dots,N-1, \\ b_{2i-2} &= b_{2N-2}\, b_{2i-1}^{\alpha_1}, && i = 2,\dots,N-1 \end{aligned} \tag{3.26}$$

by invoking an implicit function theorem. This proves that

$$\begin{aligned} a_{2j-1} &= (b_{2j-1} + \mathrm{o}(1))\, r^{\gamma_j}, \\ a_{2j} &= 1 - (b_{2j} + \mathrm{o}(1))\, r^{\gamma_j} \end{aligned} \tag{3.27}$$

holds for $j = 1,\dots,N-1$. In particular, $a_{2j-1}(r) > 0$ is positive for $r > 0$ thanks to (3.26) and Lemma 3.2.

The expansion (2.3) of the return times is now an easy consequence of (3.13) and (3.27) using the relation (3.21) of γ_j and θ_j.

Finally, we prove the claim about the ordering of the bifurcation curves in Figure 4. Summarizing the results obtained thus far and using the exponents θ_j instead of γ_j, the bifurcation curve $\bar\mu_N(r)$ for N-fronts is given by

$$\begin{aligned} \mu_1(r) &= r^{\alpha_2(\mu_1(r),\mu_2(r))}, \\ \mu_2(r) &= r + (C_N + \mathrm{o}(1))\, r^{\alpha_1(\alpha_2+\theta_N)} =: r + \rho_N(r) \end{aligned} \tag{3.28}$$

for some positive constant C_N. Indeed, the first equation is (3.14), while the second is obtained by substituting (3.27) for $j = 1$ and the solution $s(\mu_2, r)$ of (3.14) into (3.16). Define the function $\hat\mu_1(r)$ by solving

$$\hat\mu_1 = r^{\alpha_2(\hat\mu_1, r)}$$

with respect to $\hat\mu_1$. This definition allows us to separate the parts of $\bar\mu_N(r)$ which are independent of N from those which are not. Write

$$(\mu_1, \mu_2)(r) = (\hat\mu_1(r) + \sigma_N, r + \rho_N(r)).$$

Then, using (3.28), a straightforward calculation shows that

$$\sigma_N(r) = \mathrm{O}(r^{\alpha_2(\hat\mu_1(r),r)}\, r^{\alpha_1(\alpha_2+\theta_N)-\delta}),$$

where $\delta > 0$ can be chosen as small as we wish. Thus, the bifurcation curve $\bar\mu_N(r)$ for N-fronts is given by

$$\begin{aligned} \mu_1(r) &= r^{\alpha_2(\hat\mu_1(r),r)}(1 + \mathrm{O}(r^{\alpha_1(\alpha_2+\theta_N)-\delta})), \\ \mu_2(r) &= r + (C_N + \mathrm{o}(1))\, r^{\alpha_1(\alpha_2+\theta_N)}. \end{aligned}$$

As the exponent $\alpha_1(\alpha_2 + \theta_N)$ is larger than one, it is possible to write r as a function of μ_2:

$$r(\mu_2) = \mu_2 - (C_N + \mathrm{o}(1))\, \mu_2^{\alpha_1(\alpha_2+\theta_N)}$$

for $\mu_2 \ge 0$. Therefore, using that $\delta > 0$ can be chosen smaller than one, we obtain

$$\begin{aligned} \mu_1 &= \mu_2^{\hat\alpha_2(\mu_2)} \Big(1 - (C_N + \mathrm{o}(1))\, \mu_2^{\alpha_1(\alpha_2+\theta_N)-1}\Big)^{\hat\alpha_2(\mu_2)} \Big(1 + \mathrm{O}(\mu_2^{\alpha_1(\alpha_2+\theta_N)-\delta})\Big) \\ &= \mu_2^{\hat\alpha_2(\mu_2)} \Big(1 - \hat\alpha_2(\mu_2)\, C_N\, \mu_2^{\alpha_1(\alpha_2+\theta_N)-1} + \mathrm{O}(\mu_2^{\alpha_1(\alpha_2+\theta_N)-\delta})\Big) \end{aligned}$$

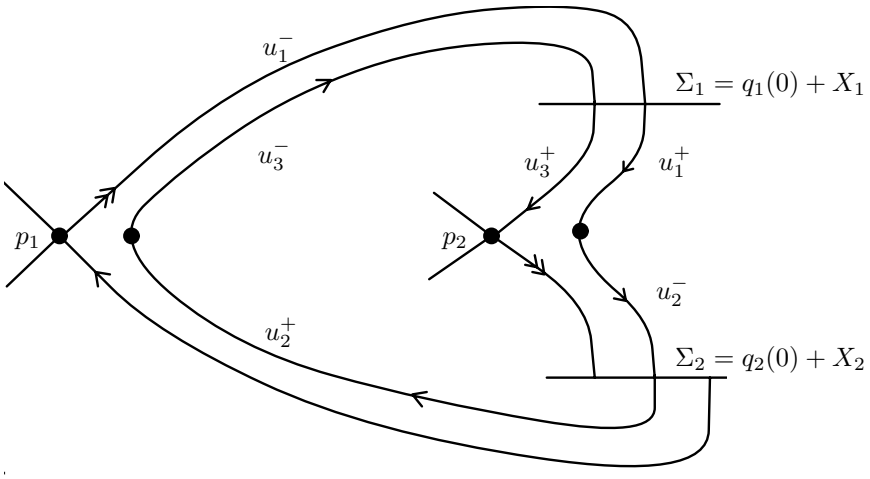

FIG. 6. *Description of N-front solutions for $N = 1$.*

with $\hat{\alpha}_2(\mu_2) := \alpha_2(\hat{\mu}_1(r(\mu_2)), r(\mu_2))$. Now, the claim about the ordering of the curves $\bar{\mu}_N(r)$ is an easy consequence of the fact that θ_N is strictly increasing in N.

Hence the proof of Theorem 2.1 is complete. ∎

4. Stability. This section is devoted to the proof of Theorem 2.2. The basic technique used is Lin's method applied to the eigenvalue problem (2.4). We shall use the abstract results from [San96] together with certain modifications needed in the present situation. As for the concrete bifurcation investigated here, we are again going to exploit the reduction to a two-dimensional invariant manifold. Finally, the eigenvalues of the resulting tridiagonal matrix are calculated.

Throughout we suppose that hypotheses (H1)–(H7) are met.

Convention. Throughout this section, we use the convention that the ranges of the indices i and j are $i = 1, \dots, 2N+1$ and $j = 1, \dots, 2N$ unless stated otherwise. Moreover, we define $i \bmod 2 \in \{1, 2\}$ by convention. The Landau symbol $o(1)$ is taken with respect to $r \to 0$.

4.1. Abstract reduction of the eigenvalue problem. We consider equation (2.1) and (2.4) in $\mathbb{R}^n$ keeping in mind that the N-fronts are actually contained in the invariant C^1-manifold W^c_{hom}. We also extend the sections Σ_k for $k = 1, 2$ to sections in $\mathbb{R}^n$ without changing notation.

Any solution with initial point in Σ_k and end point in Σ_{k+1} is uniquely described by the associated return time T. In particular, any N-front $q_N(t)$ is determined by $2N$ return times T_j for $j = 0, \dots, 2N-1$; see Theorem 2.1 and the proof in the last section. Define $u_i^{\pm}(t)$ by

$$q_N\Big(t + \sum_{j=0}^{i-2} T_j\Big) = \begin{cases} u_i^-(t) & \text{for } t \in [-\frac{1}{2}T_{i-2}, 0], \\ u_i^+(t) & \text{for } t \in [0, \frac{1}{2}T_{i-1}] \end{cases} \tag{4.1}$$

for $i = 1, \dots, 2N+1$ and with $T_{-1} = T_{2N} = \infty$; see Figure 6. As $q_N(t)$ is a solution of (2.1), the functions $u_i^{\pm}$ satisfy

$$\begin{aligned} u_i^+(0) &= u_i^-(0), && i = 1, \dots, 2N+1, \\ u_j^+(\tfrac{1}{2}T_{j-1}) &= u_{j+1}^-(-\tfrac{1}{2}T_{j-1}), && j = 1, \dots, 2N. \end{aligned} \tag{4.2}$$

The eigenvalue problem (2.4)

$$\dot v = \big(D_u f(q_N(t), \bar\mu_N) + \lambda B(t)\big) v, \qquad\qquad t \in \mathbb{R}$$

can be written as

$$\begin{array}{rcll} \dot v_i^- & = & (D_u f(u_i^-(t), \bar\mu_N) + \lambda B(t))\, v_i^- & \text{for } t \in (-\frac12 T_{i-2}, 0), \\ \dot v_i^+ & = & (D_u f(u_i^+(t), \bar\mu_N) + \lambda B(t))\, v_i^+ & \text{for } t \in (0, \frac12 T_{i-1}), \\ v_i^+(0) & = & v_i^-(0), & \\ v_j^+(\frac12 T_{j-1}) & = & v_{j+1}^-(-\frac12 T_{j-1}), & \end{array} \tag{4.3}$$

considered as equations over the complex field. Exploiting the fact that $\dot q_N(t)$ solves (2.4) for $\lambda = 0$ and using (4.1), we take the ansatz

$$v_i^\pm(t) = \dot u_i^\pm(t)\, d_i + w_i^\pm(t),$$

with $d_i \in \mathbb{R}$. On account of [San96, section 3.1] and (4.2), equation (4.3) is then equivalent to

$$\begin{array}{rcl} \dot w_i^\pm & = & (D_u f(u_i^\pm(t), \bar\mu_N) + \lambda B(t))\, w_i^\pm + \lambda\, B(t)\, \dot u_i^\pm(t)\, d_i \\ & & \text{for } t \in (-\frac12 T_{i-2}, 0) \text{ and } t \in (0, \frac12 T_{i-1}), \text{ respectively,} \\ w_i^+(0) & = & w_i^-(0), \\ w_i^\pm(0) & \in & X_{i \bmod 2}, \\ w_j^+(\frac12 T_{j-1}) & = & w_{j+1}^-(-\frac12 T_{j-1}) + \dot u_{j+1}^-(-\frac12 T_{j-1})(d_{j+1} - d_j), \end{array} \tag{4.4}$$

where the (complexified) subspaces X_k are defined by $\Sigma_k = q_k(0) + X_k$ for $k = 1, 2$. Following [San96], we shall investigate the system

$$\begin{array}{rcl} \dot w_i^\pm & = & (D_u f(u_i^\pm(t), \bar\mu_N) + \lambda B(t))\, w_i^\pm + \lambda\, B(t)\, \dot u_i^\pm(t)\, d_i \\ & & \text{for } t \in (-\frac12 T_{i-2}, 0) \text{ and } t \in (0, \frac12 T_{i-1}), \text{ respectively,} \\ w_i^+(0) - w_i^-(0) & \in & \mathbb{C} T_{u_i^+(0)} W^c_{hom}(\bar\mu_N) \cap X_{i \bmod 2} \;\cong\; \mathbb{C}, \\ w_i^\pm(0) & \in & X_{i \bmod 2}, \\ w_j^+(\frac12 T_{j-1}) & = & w_{j+1}^-(-\frac12 T_{j-1}) + \dot u_{j+1}^-(-\frac12 T_{j-1})(d_{j+1} - d_j). \end{array} \tag{4.5}$$

Define the signed distances

$$\xi_i := \langle \psi_{i \bmod 2}(0), w_i^+(0) - w_i^-(0) \rangle \in \mathbb{C}; \tag{4.6}$$

see Figure 5. Then we have the following lemma.

LEMMA 4.1. *Equation* (4.5) *possesses a unique bounded solution* $w = W(\lambda)d$ *linear in* d *and analytic in* λ. *Moreover,* w *solves* (4.4) *if and only if*

$$\xi = S(\lambda)\, d = \big(A(r) - \lambda\,(M + \mathrm{o}(1)) + \mathrm{O}(|\lambda|^2)\big)\, d = 0 \tag{4.7}$$

for some analytic, matrix-valued function $S(\lambda)$ *and*

$$M = \mathrm{diag}(M_1 K_1, M_2 K_2, \ldots, M_1 K_1)$$

with $K_1, K_2 > 0$ *positive. The matrix* $A(r)$ *is determined by* (4.5) *with* $\lambda = 0$. *Any solution of* (2.4) *with* $|\lambda|$ *small is given by the above function* $W(\lambda)$. *In particular,* $d = (1, \ldots, 1)$ *solves* $S(0)\, d = 0$.

With the equivalence of (2.4) and (4.1) as well as Lemma 4.1 at hand, it therefore remains to solve the reduced equation

$$\det S(\lambda) = 0. \tag{4.8}$$

Proof. The proof of the lemma is essentially contained in [San96], where the analysis was done for N-pulses. We will briefly mention the changes needed here.

The hypotheses (H1) and (H3) ensure that the technique developed in [San96] works in the present context. The only difference is that the linearized flows for the heteroclinic solutions are used instead of linearizing along a single homoclinic orbit. The major change made here in comparison with [San96] is that we allow for jumps in

$$w_i^+(0) - w_i^-(0) \in \mathbb{C}T_{u_i^+(0)}W^c_{hom}(\bar\mu_N) \cap X_{i \bmod 2} \cong \mathbb{C}$$

compared with jumps in $\mathbb{C}\psi(0)$

$$w_i^+(0) - w_i^-(0) \in \mathbb{C}\psi_{i \bmod 2}(0),$$

where $\psi_k(t)$ are the unique bounded solutions of the adjoint equation; see section 2. However, the only property of $\mathbb{C}\psi_k(0)$ used in [San96] is the transversality condition

$$\mathbb{R}\psi_k(0) \oplus \mathbb{R}\dot q_k(0) \oplus T_{q_k(0)}W^{uu}(p_k) \oplus T_{q_k(0)}W^{ss}(p_{k+1}) = \mathbb{R}^n$$

for $k = 1, 2$; see [San96, Lemma 3.5]. The corresponding relations

$$\big(T_{u_i^+(0)}W^c_{hom}(\bar\mu_N) \cap X_k\big) \oplus \mathbb{R}\dot q_k(0) \oplus T_{q_k(0)}W^{uu}(p_k) \oplus T_{q_k(0)}W^{ss}(p_{k+1}) = \mathbb{R}^n$$

are satisfied where $k = i \bmod 2$. Indeed, this is a consequence of (2.2) and the proof of Lemma 3.1. The statement about the matrix M follows from [San96, Lemma 3.6] and the above discussion. Indeed, taking the limit $r \to 0$ is equivalent to computing the matrix M by investigating the eigenvalue problem (2.4) for the primary heteroclinic orbits $q_k(t)$ for $k = 1, 2$ as $u_i \to q_{i \bmod 2}$ for $r \to 0$ in the sup-norm. The positive factors K_1 and K_2 stem from the projection of $\psi_k(0)$ onto the tangent spaces$T_{q_k(0)}W^c_{hom}$ for $k = 1, 2$. ∎

4.2. Determining the reduced problem using center-manifolds. In order to solve (4.8)

$$\det S(\lambda) = \det\big(A(r) - \lambda\,(M + \mathrm{o}(1)) + \mathrm{O}(|\lambda|^2)\big) = 0,$$

we have to determine the matrix $A(r)$. By definition, with $\lambda = 0$,

$$\xi = (\langle \psi_{i \bmod 2}(0), w_i^+(0) - w_i^-(0)\rangle)_{i=1,\dots,2N+1} = A(r)\,d,$$

where $w = W(0)\,d$ solves (4.5) with $\lambda = 0$; that is,

$$\begin{aligned}
&\text{(i)} && \dot w_i^\pm && = && D_u f(u_i^\pm, \bar\mu_N)\, w_i^\pm, \\
& && && && \text{for } t \in (-\tfrac12 T_{i-2}, 0) \text{ and } t \in (0, \tfrac12 T_{i-1}), \text{ respectively,} \\
&\text{(ii)} && w_i^+(0) - w_i^-(0) && \in && \mathbb{C}T_{u_i^+(0)}W^c_{hom}(\bar\mu_N) \cap X_{i \bmod 2}, \\
&\text{(iii)} && w_i^\pm(0) && \in && X_{i \bmod 2}, \\
&\text{(iv)} && w_j^+(\tfrac12 T_{j-1}) && = && w_{j+1}^-(-\tfrac12 T_{j-1}) + \dot u_{j+1}^-(-\tfrac12 T_{j-1})(d_{j+1} - d_j).
\end{aligned} \tag{4.9}$$

Therefore, the solutions w_i have to solve the variational equation along the N-front. Since W^c_{hom} is locally invariant and C^1, its continuous tangent bundle is invariant under the linearized flow. Since $\dot u_i \in T_{q_N} W^c_{hom}$ and the jumps of w_i are required to be in $T_{q_N} W^c_{hom}$, too, we expect that the solutions $w_i \in T_{q_N} W^c_{hom}$ are contained in the tangent bundle as well. By uniqueness of w as stated in Lemma 4.1, it is therefore sufficient to prove that we can solve (4.9) with $w_i \in T_{u_i} W^c_{hom}$. Since the linearized flow is still C^0-conjugated to the linearized flow in $\mathbb{R}^2$, see Lemma 3.1, it suffices to consider (4.9) for the vector field in $\mathbb{R}^2$ investigated in section 3—note that we do not need any differentiability further on.

Hence consider $w \in \mathbb{R}^2$ from now on. Denote the evolution of

$$\dot w = D_u f(u_i^\pm(t), \bar\mu_N)\, w$$

by $\Phi_i^\pm(t,s)$, then $w_i^\pm(t) = \Phi_i^\pm(t,0)\, w_i^\pm(0)$ solves (4.9)(i) and (iii) for arbitrary $w_i^\pm(0) \in X_k$. Note that (4.9)(ii) is then satisfied, too, as the subspaces $X_k \subset \mathbb{R}^2$ are one-dimensional. We shall solve (4.9)(iv)

$$w_j^+(\tfrac12 T_{j-1}) = w_{j+1}^-(-\tfrac12 T_{j-1}) + \dot u_{j+1}^-(-\tfrac12 T_{j-1})(d_{j+1} - d_j) \tag{4.10}$$

for given $d = (d_i)_{i=1,\dots,2N+1}$ and $j = 1, \dots, 2N$. Observe that these equations decouple as we can choose $w_i^\pm(0) \in X_k$ arbitrarily.

First, consider (4.10) for odd $j = 2l+1$ for $l = 0, \dots, N-1$. Then

$$\Phi^+_{2l+1}(t,0) = \Phi^-_{2l+2}(t,0) = \begin{pmatrix} e^{-\alpha_2^s(\mu)t} & 0 \\ 0 & e^{\alpha_2^u(\mu)t} \end{pmatrix}$$

as the flow is linear. Also,

$$\dot u^-_{2l+2}(-\tfrac12 T_{2l}) = (-\alpha_2^s(\mu)\, e^{-\frac12 \alpha_2^s(\mu) T_{2l}}, \alpha_2^u(\mu)\, e^{-\frac12 \alpha_2^u(\mu) T_{2l}})$$

and

$$\begin{aligned} w^+_{2l+1}(\tfrac12 T_{2l}) &= (0, e^{\frac12 \alpha_2^u(\mu) T_{2l}}\, w^+_{2l+1}(0)), \\ w^-_{2l+2}(-\tfrac12 T_{2l}) &= (e^{\frac12 \alpha_2^s(\mu) T_{2l}}\, w^-_{2l+2}(0), 0), \end{aligned}$$

identifying the subspaces X_k with $\mathbb{R}$ as in Figure 5. Thus, we conclude that

$$\begin{aligned} w^+_{2l+1}(0) &= \alpha_2^u(\mu)\, e^{-\alpha_2^u(\mu) T_{2l}}\, (d_{2l+2} - d_{2l+1}) = \mathrm{o}(r)\, (d_{2l+2} - d_{2l+1}), \\ w^-_{2l+2}(0) &= \alpha_2^s(\mu)\, e^{-\alpha_2^s(\mu) T_{2l}}\, (d_{2l+2} - d_{2l+1}) \\ &= \alpha_2^s\, (1 + \mathrm{o}(1))\, r\, (d_{2l+2} - d_{2l+1}), \end{aligned} \tag{4.11}$$

using (3.7) and (3.27).

Next, consider (4.10) for even $j = 2l$ for $l = 1, \dots, N$. Then

$$\begin{aligned} \Phi^+_{2l}(t,0) &= \begin{pmatrix} e^{-\alpha_1^s(\mu)(t-\Omega_2)} & 0 \\ 0 & e^{\alpha_1^u(\mu)(t-\Omega_2)} \end{pmatrix} \Phi^+_{2l}(\Omega_2, 0), \\ \Phi^-_{2l+1}(-t,0) &= \begin{pmatrix} e^{-\alpha_1^s(\mu)(-t+\Omega_1)} & 0 \\ 0 & e^{\alpha_1^u(\mu)(-t+\Omega_1)} \end{pmatrix} \Phi^+_{2l+1}(-\Omega_1, 0) \end{aligned}$$

for $t > 0$ large and with

$$\begin{aligned} \Omega_1 &= \Omega_1(e^{-\alpha_1^s(\mu)\, \tau(T_{2l-1}, \mu)}, \mu), \\ \Omega_2 &= \Omega_2(e^{-\alpha_1^u(\mu)\, \tau(T_{2l-1}, \mu)}, \mu); \end{aligned}$$

see section 3.2. Therefore, we obtain

$$\begin{aligned} w^+_{2l}(\tfrac{1}{2}T_{2l-1}) &= (e^{\alpha_1^s(\mu)(-\frac{1}{2}T_{2l-1}+\Omega_2)}\,\pi^s_{2l}, e^{\alpha_1^u(\mu)(\frac{1}{2}T_{2l-1}-\Omega_2)}\,\pi^u_{2l})\,w^+_{2l}(0),\\ w^-_{2l+1}(-\tfrac{1}{2}T_{2l-1}) &= (e^{\alpha_1^s(\mu)(\frac{1}{2}T_{2l-1}-\Omega_1)}\,\pi^s_{2l+1}, e^{\alpha_1^u(\mu)(-\frac{1}{2}T_{2l-1}+\Omega_1)}\,\pi^u_{2l+1})\,w^-_{2l+1}(0) \end{aligned}$$

for some constants π^k_{2l}, π^k_{2l+1} uniformly bounded in T_{2l-1} for $k = s, u$ such that

$$\pi^u_{2l}, \pi^s_{2l+1} < -\delta < 0 \tag{4.12}$$

for some δ owing to the sign convention for the sections—we identify the subspaces X_k with $\mathbb{R}$ in the same way as we did for Σ_k; see Figure 5. The time derivative is given by

$$\dot{u}^-_{2l+1}(-\tfrac{1}{2}T_{2l-1}) = (-\alpha_1^s(\mu)\,e^{-\alpha_1^s(\mu)(\frac{1}{2}T_{2l-1}-\Omega_2)}, \alpha_1^u(\mu)\,e^{-\alpha_1^u(\mu)(\frac{1}{2}T_{2l-1}-\Omega_1)}).$$

Thus, (4.10) reads

$$\begin{pmatrix} -e^{\alpha_1^s(\mu)(\frac{1}{2}T_{2l-1}-\Omega_1)}\,\pi^s_{2l+1} & e^{\alpha_1^s(\mu)(-\frac{1}{2}T_{2l-1}+\Omega_2)}\,\pi^s_{2l} \\ -e^{\alpha_1^u(\mu)(-\frac{1}{2}T_{2l-1}+\Omega_1)}\,\pi^u_{2l+1} & e^{\alpha_1^u(\mu)(\frac{1}{2}T_{2l-1}-\Omega_2)}\,\pi^u_{2l} \end{pmatrix} \begin{pmatrix} w^-_{2l}(0) \\ w^+_{2l+1}(0) \end{pmatrix}$$
$$= \begin{pmatrix} -\alpha_1^s(\mu)\,e^{-\alpha_1^s(\mu)(\frac{1}{2}T_{2l-1}-\Omega_2)} \\ \alpha_1^u(\mu)\,e^{-\alpha_1^u(\mu)(\frac{1}{2}T_{2l-1}-\Omega_1)} \end{pmatrix} (d_{2l+1}-d_{2l}),$$

and it is straightforward to calculate that for some $\delta > 0$

$$\begin{aligned} w^+_{2l}(0) &= \alpha_1^u(\mu)\,e^{-\alpha_1^u(\mu)(T_{2l-1}-\Omega_1-\Omega_2)}\,\pi^u_{2l}\,(1+\mathrm{O}(e^{-\delta T_{2l-1}}))\,(d_{2l+1}-d_{2l})\\ &= \alpha_1^u(\mu)\,e^{-\alpha_1^u(\mu)\,\tau(T_{2l-1})}\,\pi^u_{2l}\,(1+\mathrm{O}(e^{-\delta\tau(T_{2l-1})}))\,(d_{2l+1}-d_{2l})\\ &= \mathrm{o}(r^{\alpha_2+\gamma_l})\,(d_{2l+1}-d_{2l}),\\ w^-_{2l+1}(0) &= \alpha_1^s(\mu)\,e^{-\alpha_1^s(\mu)(T_{2l-1}-\Omega_1-\Omega_2)}\,\pi^s_{2l+1}\,(1+\mathrm{O}(e^{-\delta T_{2l-1}}))\,(d_{2l+1}-d_{2l})\\ &= \alpha_1^s(\mu)\,e^{-\alpha_1^s(\mu)\,\tau(T_{2l-1})}\,\pi^s_{2l+1}\,(1+\mathrm{O}(e^{-\delta\tau(T_{2l-1})}))\,(d_{2l+1}-d_{2l})\\ &= \alpha_1^s\,(b_{2l-1}+\mathrm{o}(1))\,\pi^s_{2l+1}\,r^{\alpha_2+\gamma_l}\,(d_{2l+1}-d_{2l}); \end{aligned} \tag{4.13}$$

see again (3.7) and (3.27). It is convenient to check the signs appearing in (4.11) and (4.13) by inspecting Figures 5 and 6.

Thus, the differences of $w^\pm_i(0)$ for $i = 1, \ldots, 2N+1$ with $\lambda = 0$ are given by

$$\begin{aligned} w^+_{2l}(0) - w^-_{2l}(0) &= \mathrm{o}(r^{\alpha_2+\gamma_l})\,(d_{2l+1}-d_{2l}) - \alpha_2^s\,(1+\mathrm{o}(1))\,r\,(d_{2l}-d_{2l-1}),\\ w^+_{2l+1}(0) - w^-_{2l+1}(0) &= \mathrm{o}(r)\,(d_{2l+2}-d_{2l+1})\\ &\quad -\alpha_1^s\,(b_{2l-1}+\mathrm{o}(1))\,\pi^s_{2l+1}\,r^{\alpha_2+\gamma_l}\,(d_{2l+1}-d_{2l}), \end{aligned}$$

and the jumps ξ_i read

$$\begin{aligned} \xi_{2l} &= \langle \psi_2(0), w^+_{2l}(0) - w^-_{2l}(0)\rangle\\ &= r\big(\mathrm{o}(r^{\alpha_2+\gamma_l-1})\,(d_{2l+1}-d_{2l}) + \alpha_2^s\,(1+\mathrm{o}(1))\,(d_{2l}-d_{2l-1})\big),\\ \xi_{2l+1} &= \langle \psi_1(0), w^+_{2l+1}(0) - w^-_{2l+1}(0)\rangle\\ &= r\big(\mathrm{o}(1)\,(d_{2l+2}-d_{2l+1})\\ &\quad -\alpha_1^s\,(b_{2l-1}+\mathrm{o}(1))\,\pi^s_{2l+1}\,r^{\alpha_2+\gamma_l-1}\,(d_{2l+1}-d_{2l})\big). \end{aligned} \tag{4.14}$$

Notice that the sign changes in the first equation since $\psi_2(0)$ points in the negative direction of X_2; see Figure 5. We rewrite (4.14) according to

$$\begin{aligned}\xi_{2l} &= r\,(-\kappa_{2l-1}\,d_{2l-1} + (\kappa_{2l-1} - \tilde\kappa_{2l})\,d_{2l} + \tilde\kappa_{2l}\,d_{2l+1}),\\ \xi_{2l+1} &= r\,(-\kappa_{2l}\,d_{2l} + (\kappa_{2l} - \tilde\kappa_{2l+1})\,d_{2l+1} + \tilde\kappa_{2l+1}\,d_{2l+2}),\end{aligned}$$

using the definitions

$$\begin{array}{llll}\kappa_{2l-1} &:= c_{2l-1} + \mathrm{o}(1) &:= \alpha_2^s\,(1+\mathrm{o}(1)),\\ \tilde\kappa_{2l-1} &:= \mathrm{o}(1),\\ \kappa_{2l} &:= (c_{2l} + \mathrm{o}(1))\,r^{\beta_l} &:= -\alpha_1^s\,(b_{2l-1} + \mathrm{o}(1))\,\pi_{2l+1}^s\,r^{\alpha_2+\gamma_l-1},\\ \tilde\kappa_{2l} &:= \mathrm{o}(r^{\beta_l}) &:= \mathrm{o}(r^{\alpha_2+\gamma_l-1})\end{array} \tag{4.15}$$

for $l = 1, \dots, N$ and

$$\kappa_0 = \tilde\kappa_0 = \kappa_{2N+1} = \tilde\kappa_{2N+1} = 0.$$

The exponents β_l and the constants c_j satisfy

$$\begin{array}{ll}\beta_l := \alpha_2 + \gamma_l - 1, & l = 1, \dots, N,\\ 0 < \alpha_2 - 1 = \beta_N < \beta_l < \beta_{l-1}, & l = 2, \dots, N-1,\\ c_j > 0, & j = 1, \dots, 2N,\end{array} \tag{4.16}$$

due to (3.20), (3.26), and (4.12).

Therefore, we end up with computing solutions of

$$\det\left(r\tilde A(r) - M\lambda + \mathrm{O}(|\lambda|(|\lambda| + \mathrm{o}(1)))\right) = 0, \tag{4.17}$$

where

$$M = \mathrm{diag}(M_1K_1, M_2K_2, \dots, M_1K_1)$$

for some positive constants $K_1, K_2 > 0$ and

$$\tilde A(r) = \begin{pmatrix} -\tilde\kappa_1 & \tilde\kappa_1 & & & & \\ -\kappa_1 & \kappa_1 - \tilde\kappa_2 & \tilde\kappa_2 & & & \\ & -\kappa_2 & \kappa_2 - \tilde\kappa_3 & \tilde\kappa_3 & & \\ & & \ddots & \ddots & & \\ & & & & -\kappa_{2N} & \kappa_{2N} \end{pmatrix}. \tag{4.18}$$

As we are mainly interested in stable N-front solutions, we assume

$$\operatorname{sign} M_1 = \operatorname{sign} M_2 = -1$$

from now on, and, by rescaling the solutions $\psi_k(t)$, we obtain

$$M = -\operatorname{id}.$$

The other cases can be handled similarly.

4.3. Solving the reduced eigenvalue problem. Thus we shall solve (4.17). By Rouché's theorem, there exist precisely $2N+1$ solutions of (4.17), since $S(\lambda)$ is analytic in λ and

$$\det S(\lambda) = \lambda^{2N+1} + \mathrm{o}(1)$$

near $\lambda = 0$.

One of these solutions is equal to zero,

$$\lambda_{2N+1} = 0, \tag{4.19}$$

due to translational invariance. By construction, the associated eigenvector is given by $v = (1, \dots, 1)$; see Lemma 4.1.

Substituting $\lambda = \nu r$ and $M = -\mathrm{id}$ into (4.17) and dividing by r^{2N+1} yields

$$\det\left(\tilde{A}(r) + \nu\,(\mathrm{id} + \mathrm{o}(1))\right) = 0. \tag{4.20}$$

There are another N eigenvalues which can be computed easily. Indeed, setting $r = 0$ in (4.20), we obtain

$$\det(\tilde{A}(0) + \nu\,\mathrm{id}) = \nu^{N+1} \prod_{l=1}^{N} (c_{2l-1} + \nu).$$

Hence, again by Rouché's theorem, there exist precisely N solutions $\nu_{2l-1}(r)$ of (4.20) counted with multiplicity and continuous in r such that

$$\nu_{2l-1}(0) = -c_{2l-1} < 0.$$

They correspond to N eigenvalues $\lambda_{2l-1}(r)$ of (4.17) given by

$$\lambda_{2l-1}(r) = \nu_{2l-1}(r)\, r = -(c_{2l-1} + \mathrm{o}(1))\, r < 0, \qquad l = 1, \dots, N. \tag{4.21}$$

It remains to calculate the remaining N eigenvalues of (4.20). The columns of the matrix $S(\nu, r) = \tilde{A}(r) + \nu\,(\mathrm{id} + \mathrm{o}(1))$ are given by

$$\begin{aligned}
C_1 &= (-\tilde{\kappa}_1 + \nu, -\kappa_1, 0, \dots, 0) + \mathrm{o}(1)\nu, \\
C_j &= \Big(0, \dots, 0, \tilde{\kappa}_{j-1}, \underbrace{\kappa_{j-1} - \tilde{\kappa}_j + \nu}_{j\text{th}}, -\kappa_j, 0, \dots, 0\Big) + \mathrm{o}(1)\nu, \qquad j = 2, \dots, 2N, \\
C_{2N+1} &= (0, \dots, 0, \tilde{\kappa}_{2N}, \kappa_{2N} + \nu) + \mathrm{o}(1)\nu;
\end{aligned}$$

see (4.18). Adding successively the jth column C_j to C_{j-1} for $j = 2N+1, \dots, 2$ yields a matrix with columns

$$\begin{aligned}
C_1 &= (\nu, \dots, \nu) + \mathrm{o}(1)\nu, \\
C_j &= \Big(0, \dots, 0, \tilde{\kappa}_{j-1}, \underbrace{\kappa_{j-1} + \nu}_{j\text{th}}, \nu, \dots, \nu\Big) + \mathrm{o}(1)\nu, \qquad j = 2, \dots, 2N, \\
C_{2N+1} &= (0, \dots, 0, \tilde{\kappa}_{2N}, \kappa_{2N} + \nu) + \mathrm{o}(1)\nu.
\end{aligned}$$

Note that this transformation does not change the determinant. Moreover, recall from (4.15) that

$$\begin{aligned}
\kappa_{2l-1} &= c_{2l-1} + \mathrm{o}(1), & \tilde{\kappa}_{2l-1} &= \mathrm{o}(1) = \mathrm{o}(\kappa_{2l-1}), \\
\kappa_{2l} &= (c_{2l} + \mathrm{o}(1))\, r^{\beta_l}, & \tilde{\kappa}_{2l} &= \mathrm{o}(r^{\beta_l}) = \mathrm{o}(\kappa_{2l})
\end{aligned}$$

for positive constants $c_j > 0$ and exponents $\beta_l > 0$ strictly decreasing in l; see (4.16). For fixed k satisfying $1 \le k \le N$, we make the ansatz

$$\nu = r^{\beta_k}\,\eta.$$

Substituting it into the matrix yields

$$\begin{aligned}
C_1 &= \big[(\eta,\dots,\eta) + \mathrm{o}(1)\big]\, r^{\beta_k},\\
C_{2l} &= \left[\left(0,\dots,0,\underbrace{c_{2l-1}}_{(2l)\mathrm{th}},0\dots,0\right) + \mathrm{o}(1)\right],\\
C_{2l+1} &= \begin{cases}
\left[\left(0,\dots,0,\underbrace{\eta}_{(2l+1)\mathrm{th}},\eta,\dots,\eta\right) + \mathrm{o}(1)\right] r^{\beta_k}, & l < k,\\
\left[\left(0,\dots,0,\underbrace{c_{2k}+\eta}_{(2k+1)\mathrm{th}},\eta,\dots,\eta\right) + \mathrm{o}(1)\right] r^{\beta_k}, & l = k,\\
\left[\left(0,\dots,0,\underbrace{c_{2l}}_{(2l+1)\mathrm{th}},0,\dots,0\right) + \mathrm{o}(1)\right] r^{\beta_l}, & l > k
\end{cases}
\end{aligned}$$

for $l = 1,\dots,N$. Thus, factorizing the powers of r multiplying each column, the determinant of the matrix $S(r^{\beta_k}\eta, r)$ equals

$$\det\, S(r^{\beta_k}\,\eta, r) = \big(\det\, \tilde{S}(\eta, r)\big)\, r^{(k+1)\beta_k} \prod_{l=k+1}^{N} r^{\beta_l},$$

where the columns of $\tilde{S}(\eta, r)$ are given by

$$\begin{aligned}
C_1 &= \big[(\eta,\dots,\eta) + \mathrm{o}(1)\big],\\
C_{2l} &= \left[\left(0,\dots,0,\underbrace{c_{2l-1}}_{(2l)\mathrm{th}},0\dots,0\right) + \mathrm{o}(1)\right],\\
C_{2l+1} &= \begin{cases}
\left[\left(0,\dots,0,\underbrace{\eta}_{(2l+1)\mathrm{th}},\eta,\dots,\eta\right) + \mathrm{o}(1)\right], & l < k,\\
\left[\left(0,\dots,0,\underbrace{c_{2k}+\eta}_{(2k+1)\mathrm{th}},\eta,\dots,\eta\right) + \mathrm{o}(1)\right], & l = k,\\
\left[\left(0,\dots,0,\underbrace{c_{2l}}_{(2l+1)\mathrm{th}},0,\dots,0\right) + \mathrm{o}(1)\right], & l > k.
\end{cases}
\end{aligned}$$

As we are interested in zeroes for $r > 0$, it suffices to solve

$$\det\, \tilde{S}(\eta, r) = 0. \tag{4.22}$$

This matrix, however, is upper triangular up to terms of order o(1). Its determinant is therefore given by

$$\begin{aligned}\det \tilde{S}(\eta, r) &= \det \tilde{S}(\eta, 0) + \mathrm{o}(1) \\ &= \eta^k \big(\textstyle\prod_{l=k+1}^{N} c_{2l} \big)(\eta + c_{2k}) \big(\textstyle\prod_{l=1}^{N} c_{2l-1} \big) + \mathrm{o}(1).\end{aligned}$$

Again by Rouché's theorem, there is a unique solution $\eta_{2l}(r)$ of (4.22) satisfying

$$\eta_{2l}(0) = -c_{2l}$$

for $l = 1, \dots, N$. The corresponding solution $\lambda_{2l}(r)$ of (4.17) is given by

$$\begin{aligned}\lambda_{2l}(r) = \nu_{2l}(r)\, r = \eta_{2l}(r)\, r^{1+\beta_l} &= -(c_{2l} + \mathrm{o}(1))\, r^{1+\beta_l} \\ &= -(c_{2l} + \mathrm{o}(1))\, r^{\alpha_2 + \gamma_{N+1-l}}\end{aligned} \tag{4.23}$$

for $l = 1, \dots, N$; see (4.16) for the last identity. Note that these solutions are not the same for different values of l owing to (4.16). Moreover, they converge faster to zero than the eigenvalues λ_{2l-1} obtained in (4.21).

Summarizing the facts obtained above, we have calculated $2N+1$ solutions λ_j of (4.17) appearing in (4.19), (4.21), and (4.23). According to the remark above, they are pairwise distinct, from which we have found all solutions. This proves Theorem 2.2. □

5. Application to the FitzHugh–Nagumo equation. Consider the FitzHugh–Nagumo equation

$$\begin{aligned}u_t &= u_{xx} + f(u) - w, \\ w_t &= \epsilon(u - \gamma w)\end{aligned} \tag{5.1}$$

for $x \in \mathbb{R}$ with $f(u) = u(1-u)(u-a)$ and $a \in (0, \frac{1}{2})$ fixed. This equation is a simplification of the Hodgkin–Huxley equation modeling the propagation of impulses in nerve axons. Being interested in travelling waves $(u, w)(x, t) = (u, w)(x + ct)$, we introduce new variables $(\xi, t) = (x + ct, t)$ in which (5.1) takes the form

$$\begin{aligned}u_t &= u_{\xi\xi} - c u_\xi + f(u) - w, \\ w_t &= -c w_\xi + \epsilon(u - \gamma w).\end{aligned} \tag{5.2}$$

The existence of fronts travelling with wave speed c boils down to investigating heteroclinic orbits of the ordinary differential equation

$$\begin{aligned}\dot{u} &= v, \\ \dot{v} &= cv - f(u) + w, \\ \dot{w} &= \tfrac{\epsilon}{c}(u - \gamma w),\end{aligned} \tag{5.3}$$

which is the steady-state equation corresponding to (5.2). Here $\dot{} = d/d\xi$. Linearized stability of equilibria (u, w) of (5.2) is determined by the spectrum of the linear operator

$$L(U, W) = \begin{pmatrix} U_{\xi\xi} - c\, U_\xi + D_u f(u) U - W \\ -c\, W_\xi + \epsilon(U - \gamma W) \end{pmatrix}. \tag{5.4}$$

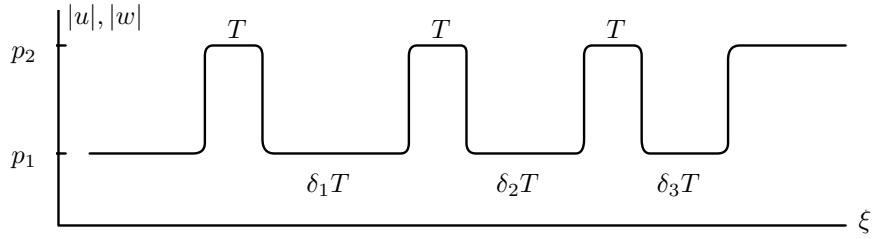

FIG. 7. *The N-front wave solution for $N = 3$. The distances of the layers are given by T and $\delta_j T = \frac{\alpha_2^u + \alpha_2^s \theta_{N+1-j}}{\alpha_1^s} T$ with $\theta_{N+1-j} > 0$ strictly decreasing in j; see Theorem* 2.1.

In particular, eigenvalues λ with corresponding eigenfunction (U, W) of L are given by bounded solutions of

$$\begin{aligned} \dot{U} &= V, \\ \dot{V} &= cV - D_u f(u)U + W + \lambda U, \\ \dot{W} &= \tfrac{\epsilon}{c}(U - \gamma W) - \tfrac{\lambda}{c} W. \end{aligned} \tag{5.5}$$

Deng proved in [Den91b] that there is a curve $(\gamma(\epsilon), c(\epsilon))$ for all $\epsilon > 0$ sufficiently small such that the FitzHugh–Nagumo equation (5.3) possesses a twisted heteroclinic loop for these values of parameters. In particular, he concluded the existence of N-fronts for any $N \geq 1$ using his result [Den91a]. Theorem 2.1 of the present article provides the distance of the layers; see Figure 7. Yanagida proved in [Yan89] that the simple fronts $q_1(t)$ and $q_2(t)$ building the heteroclinic loop are linearly stable with respect to the partial differential equation; that is, the spectrum of the linearized operator (5.4) is contained in the left half-plane except for a simple eigenvalue at zero. Finally, Nii [Nii95b] proved that the 1-fronts are linearly stable, too, using topological methods—however, he had to assume that the flow of (5.3) is linear near both equilibria. The next result asserts that in fact all N-fronts are linearly stable and provides asymptotic expansions of the critical eigenvalues.

THEOREM 5.1. *The N-fronts (and N-backs) of* (5.1) *proved to exist by Deng [Den*91*b] are linearly stable for all N. The $2N+1$ critical eigenvalues near zero are given by Theorem* 2.2. Note that linear stability implies nonlinear stability by [BJ89].

Proof. We shall use Theorem 2.2 to conclude linear stability of the N-fronts. First note that the hypotheses (H1)–(H6) needed in that theorem are met by [Den91b]. Moreover, by the results in [AGJ90] and the stability of the simple fronts proved in [Yan89], it is sufficient to calculate eigenvalues of the linearized operator (5.4) near zero; see for example [Nii95b] for a discussion. Indeed, the spectrum of (5.4) does not contain eigenvalues with nonnegative real part and large modulus; see [Eva75]. Comparing the eigenvalue problem (5.5) and the travelling wave equation (5.3) with equations (2.1) and (2.4), we see that they are of the same form by taking B according to

$$B = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{c} \end{pmatrix}.$$

Hence it suffices to prove that the Melnikov integrals

$$\int_{-\infty}^{\infty} \langle \psi_k(t), B\dot{q}_k(t) \rangle \, dt < 0 \tag{5.6}$$

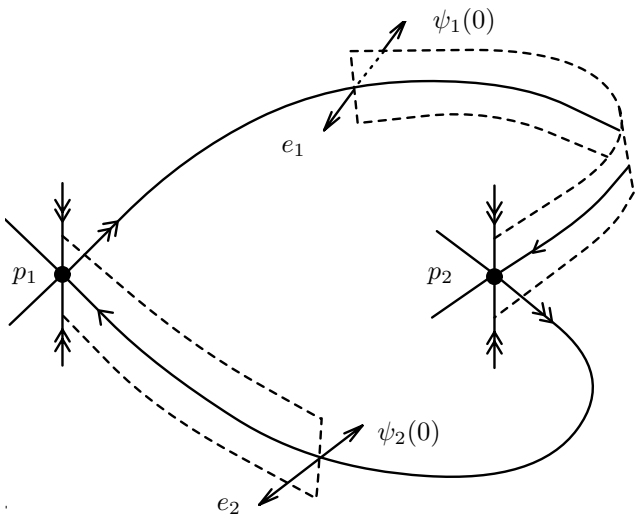

FIG. 8. *Conventions used by Deng and the present article.*

are negative for $k = 1, 2$, where $\psi_k(t)$ are chosen according to hypothesis (H6); see Figures 2 or 8. Indeed, then the statement of the theorem follows immediately from Theorem 2.2.

In order to do so, notice that for any solution (u, v, w) of (5.3)

$$B \begin{pmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{pmatrix} = \begin{pmatrix} 0 \\ \dot{u} \\ -\frac{1}{c}\dot{w} \end{pmatrix} = \begin{pmatrix} 0 \\ v \\ -\frac{\epsilon}{c^2}(u - \gamma w) \end{pmatrix} = D_c F(u, v, w, c)$$

holds, where F denotes the right-hand side of (5.3). In particular, we obtain

$$\int_{-\infty}^{\infty} \langle \psi_k(t), B\dot{q}_k(t) \rangle \, dt = \int_{-\infty}^{\infty} \langle \psi_k(t), D_c F(q_k(t), c) \rangle \, dt. \tag{5.7}$$

The second integral in the above formula is the derivative with respect to c of the signed distance of unstable and unstable manifolds measured in the direction $\psi_k(0)$, that is,

$$\int_{-\infty}^{\infty} \langle \psi_k(t), D_c F(q_k(t), c) \rangle \, dt = \frac{d}{dc} \langle \psi_k(0), p_k^u(c) - p_{k+1}^s(c) \rangle, \tag{5.8}$$

where $p_k^u(c) \in W^u(p_k, c)$ and $p_k^s(c) \in W^s(p_k, c)$; see, for instance, [Kok88], [Lin90], or [Den91b]. Here, and in the following, the index k is taken modulo two. The last quantity appearing in (5.8) has been computed in [Den91b]. What is actually computed therein is

$$\frac{d}{dc} Q_k = \frac{d}{dc} \langle e_k, p_{k+1}^s(c) - p_k^u(c) \rangle < 0; \tag{5.9}$$

see [Den91b, eq. (3.1)] for the definition and [Den91b, eqs. (5.3a), (5.4a)] for the actual computation. Moreover, the vectors e_k appearing in (5.9) above are chosen in [Den91b, pp. 1641 and 1644] such that

$$e_k = -\psi_k(0); \tag{5.10}$$

see Figure 8. Summarizing, we obtain from (5.7) and (5.8) that the Melnikov integrals

$$\int_{-\infty}^{\infty} \langle \psi_k(t), B\dot{q}_k(t) \rangle \, dt \overset{(5.7),(5.8)}{=} \frac{d}{dc} \langle \psi_k(0), p_k^u(c) - p_{k+1}^s(c) \rangle$$

$$\overset{(5.10)}{=} \frac{d}{dc} \langle -e_k(0), p_k^u(c) - p_{k+1}^s(c) \rangle$$

$$\overset{(5.9)}{=} \frac{d}{dc} Q_k \; < \; 0$$

are indeed negative. Thus the theorem is proved. □

Acknowledgments. I am very grateful to Sanjeeva Balasuriya, Christopher K.R.T. Jones, and Daniela Peterhof for helpful discussions. Moreover, I wish to thank Hiroshi Kokubu and Shunsaku Nii for informing me about their related work.

REFERENCES

[AGJ90] J. C. ALEXANDER, R. A. GARDNER, AND C. K. R. T. JONES, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[BJ89] P. W. BATES AND C. K. R. T. JONES, *Invariant manifolds for semilinear partial differential equations*, in Dynamics Reported, U. Kirchgraber and H.-O. Walther, eds., Vol. 2, John Wiley & Sons and Teubner, 1989, pp. 1–38.

[Den91a] B. DENG, *The bifurcations of countable connections from a twisted heteroclinic loop*, SIAM J. Math. Anal., 22 (1991), pp. 653–679.

[Den91b] B. DENG, *The existence of infinitely many travelling front and back waves in the Fitzhugh–Nagumo equations*, SIAM J. Math. Anal., 22 (1991), pp. 1631–1650.

[Eva75] J. W. EVANS, *Nerve axon equations,* IV*: The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[Hen81] D. HENRY, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Mathematics 804, Springer, Berlin, Heidelberg, New York, 1981, pp. 136–140.

[Hom96] A. J. HOMBURG, *Some Global Aspects of Homoclinic Bifurcations of Vector Fields*, Memoirs of the AMS 578, American Mathematical Society, Providence, RI, 1996.

[Kok88] H. KOKUBU, *Homoclinic and heteroclinic bifurcations in vector fields*, Japan J. Appl. Math., 5 (1988), pp. 455–501.

[Lin90] X.-B. LIN, *Using Melnikov's method to solve Silnikov's problems*, Proc. Roy. Soc. Edinburgh, 116A (1990), pp. 295–325.

[Nii95a] S. NII, *An extension of the stability index for traveling-wave solutions and its application for bifurcations*, SIAM J. Math. Anal., 28 (1997), pp. 402–433.

[Nii95b] S. NII, *Stability of the travelling multiple-front (multiple-back) wave solutions of the FitzHugh–Nagumo equations*, SIAM J. Math. Anal., 28 (1997), pp. 1094–1112.

[San93] B. SANDSTEDE, *Verzweigungstheorie homokliner Verdopplungen*, Doctoral thesis, University of Stuttgart, Germany, 1993.

[San95] B. SANDSTEDE, *Center manifolds for homoclinic solutions*, WIAS Preprint 186, 1995.

[San96] B. SANDSTEDE, *Stability of multiple-pulse solutions*, Trans. Amer. Math. Soc., to appear.

[Sha92] M. V. SHASHKOV, *On the bifurcations of separatrix contours on two-dimensional surfaces*, Selecta Math. Soviet., 11 (1992), pp. 341–353.

[Yan89] E. YANAGIDA, *Stability of travelling front solutions of the Fitzhugh-Nagumo equations*, Math. Comput. Modelling, 12 (1989), pp. 289–301.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 208–234, January 1998

012

MONOTONICITY OF PHASELOCKED SOLUTIONS IN CHAINS AND ARRAYS OF NEAREST-NEIGHBOR COUPLED OSCILLATORS*

LIWEI REN† AND G. BARD ERMENTROUT†

Abstract. The existence of phaselocked solutions in chains of weakly coupled oscillators is proven rigorously. The solutions show interesting monotonicity which plays an important role for the existence proofs. Under some conditions, we show that two-dimensional arrays can be decomposed into two one-dimensional problems. With this theory of decomposition, target patterns can be explained. Numerical results are provided to illustrate the theorems on the chain problem and to show traveling waves in the chains and arrays.

Key words. coupled oscillators, target patterns, phaselocking, neurons

AMS subject classifications. 34C29, 34C15, 58F22, 92C20

PII. S0036141096298837

1. Introduction.

Coupled oscillators play an increasingly important role in our understanding of various types of repetitive activity in the nervous system. There have been numerous analytic and numerical studies of the behavior of systems of coupled oscillators. These range from models of cognitive processing and binding [1] to attempts to model locomotor patterns [2]. Several connection topologies have been explored primarily due to their mathematical tractability. The simplest topology is a one-dimensional chain of oscillators. Mathematically, the case in which the two ends are connected is the easiest to analyze, but in realistic applications, this rarely arises. However, the chain topology is quite natural for models of systems such as the lamprey swim central pattern generator [2] or the central pattern generator of the leech [3]. The behavior of weakly coupled oscillators in a chain has been the object of extensive work by several authors [4, 5, 6, 7, 8].

Two-dimensional arrays of oscillators have been subject to far less mathematical analysis; most work deals exclusively with numerical simulations. They arise more naturally than chains in attempts to understand oscillatory neural behavior in neural tissue which is typically arranged in distinct two-dimensional sheets. Furthermore, there are many phenomena that can occur in two- and three-dimensional systems of oscillators that are not possible in one dimension.

It was shown in [5] that the phaselocked behavior of a sufficiently long chain of weakly coupled oscillators can be described by the solutions of a singularly perturbed two-point boundary value problem. The point of this reduction is that the analysis of phaselocking and the behavior of the chain in the presence of inhomogeneities and anisotropic coupling is much easier for the continuum model than for its discrete analogue. In this paper, we will use another approach to investigate the phaselocked behavior with *any number of oscillators*. That is, we do not require the length of the chain to tend to infinity.

Coupled oscillators present an almost impossible problem to analyze in any generality. Thus, we will restrict our attention to a class of so-called phase models that

*Received by the editors February 14, 1996; accepted for publication (in revised form) October 17, 1996. This work was supported by NIH grant NIMH-47150 and NSF grant DMS 96-26728.
http://www.siam.org/journals/sima/29-1/29883.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (bard@math.pitt.edu).

arise when oscillatory elements are weakly coupled. As was the case in [5], we will restrict our attention in this paper to nearest-neighbor coupling. In a later paper, we investigate coupling with greater spread. We first consider one-dimensional chains of oscillators. Then, we turn our attention to two-dimensional arrays. Under the condition that the distribution of intrinsic frequencies is a sum of two stripe distributions: one with constant frequencies along each row and another with constant frequencies along each column, we are able to decompose the two-dimensional problem into a set of one-dimensional problems and from this gain insight into the global phaselocked behavior. The techniques for two-dimensional arrays can be generalized in an obvious fashion to three- and higher dimensional arrays.

The equations to be considered have the form

$$\theta_i' = \omega_i + H^+(\theta_{i+1} - \theta_i) + H^-(\theta_{i-1} - \theta i), \tag{1.1}$$

where $i = 1, \ldots, n+1$, both H^+ and H^- are smooth 2π-periodic functions of their arguments, and ω_i is the frequency for each oscillator. Note that (1.1) is a nearest-neighbor coupled system. The term H^- (respectively, H^+) will be ignored for $i = 1$ (respectively, $i = n+1$). Equation (1.1) arises naturally in systems of weakly coupled oscillators. We assume that without coupling, each component of the chain has an asymptotically stable limit cycle. Thus, without coupling, each oscillator is described by a single coordinate, the phase, θ_i. The phase space of the $n+1$ oscillators then lies in an $n+1$ torus. If the oscillators interact weakly, then this invariant torus persists, and it follows from averaging theory that the equations for the phases of the $n+1$ oscillators is exactly equation (1.1). (For details on the derivation of these equations, see, e.g., [4].) The interaction functions $H^{\pm}$ are easily computed once the uncoupled oscillation is known and a formula is given for the interaction between the oscillators.

We point out that if two oscillators are coupled by diffusion, then the interaction functions $H^{\pm}$ vanish at 0. Thus, if there are no local differences in the oscillators (ω_i is independent of i) then the synchronous state $\theta_i(t) = \omega t$ is one possible solution. However, if the coupling between oscillators is based on chemical transmission then one does not expect that $H^{\pm}(0)$ will vanish. Because oscillators on the boundary (at the ends in one dimension, on the edges in two dimensions, etc.) receive less synaptic input than oscillators in the interior, this sets up a natural frequency difference between the oscillators. This makes it possible to induce a pattern of relative phases such as a traveling wave in one dimension and target patterns in two dimensions. In [5] we analyzed chains of oscillators in which there is an intrinsic anisotropy in the coupling so that H^+ and H^- are not necessarily the same. This was exploited in order to suggest a mechanism for the uniform traveling wave of electrical activity in the lamprey spinal cord. In this paper, we are mainly concerned with couplings for which $H^{\pm}$ are identical. In [7] the behavior of the chain is understood by letting n get very large and converting to a continuum equation. Here we do not restrict the size of n; the results hold for both small and large n. The main reason that we first analyze the one-dimensional chain is that we can then use these results to analyze a class of solutions in two and higher dimensions.

In section 2, we shall take the technique used in [9] to prove the existence of phaselocked solutions for several general cases. The monotonicity of the phaselocked solutions is also obtained. The monotonicity does not have any specific implication for traveling wave, but it does play a critical role in the existence proof of the phaselocked solutions.

In section 3, we shall investigate the two-dimensional arrays of weakly coupled oscillators based on the existence results of section 2. As in the one-dimensional case,

we restrict our attention to nearest-neighbor coupling, but the coupling in each of the four directions need not be the same. Under some conditions on the frequencies ω_{ij}, we can reduce this problem to two independent chain problems such that we can apply the results obtained from section 2 to describe the behavior of two-dimensional arrays of weakly coupled oscillators. One of the main results is that with isotropic "synaptic coupling," target patterns spontaneously form and synchrony cannot occur. This is due to the effects of boundaries in synaptically coupled cells.

Finally, we discuss some other two-dimensional solutions as well as how small chains can qualitatively differ from very long chains.

2. Chains of oscillators. For convenience, the equations (1.1) are written in the form

$$\begin{aligned} \theta_1' &= \omega_1 + H^+(\theta_2 - \theta_1), \\ \theta_i' &= \omega_i + H^-(\theta_{i-1} - \theta_i) + H^+(\theta_{i+1} - \theta_i), \\ \theta_{n+1}' &= \omega_{n+1} + H^-(\theta_n - \theta_{n+1}). \end{aligned} \tag{2.1}$$

We take $\phi_i = \theta_{i+1} - \theta_i, \beta_i = \omega_{i+1} - \omega_i, i = 1, \ldots, n$. Also, we define two functions f and g related to H^+ and H^- as $f(\phi) + g(\phi) = H^+(\phi)$ and $f(\phi) - g(\phi) = H^-(-\phi)$. In (2.1), if the ith equation is subtracted from the $(i+1)$th one, we have

$$\begin{aligned} \phi_1' &= \beta_1 + f(\phi_2) + g(\phi_2) - 2g(\phi_1), \\ \phi_i' &= \beta_i + f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}), \\ & \qquad i = 2, \ldots, n-1, \\ \phi_n' &= \beta_n - f(\phi_{n-1}) - 2g(\phi_n) + g(\phi_{n-1}). \end{aligned} \tag{2.2}$$

Two numbers ϕ_L and ϕ_R need to be considered. They are defined as $f(\phi_L) = g(\phi_L)$, i.e., $H^-(-\phi_L) = 0$, and $f(\phi_R) = -g(\phi_R)$, i.e., $H^+(\phi_R) = 0$.

We assume some hypotheses on f and g in a sufficiently large interval J around $\phi = 0$:

(H1) $g'(\phi) > |f'(\phi)|$ *for* $\phi \in J$;

(H2) *There exists a unique solution* ϕ_L *(respectively,* ϕ_R*) to* $f = g$ *(respectively,* $f = -g$*) for* $\phi \in J$.

These conditions are proposed in [5] with other conditions. Note that $\phi_R < 0 < \phi_L$ if $f(0) > |g(0)|$ and $\phi_L < 0 < \phi_R$ if $f(0) < -|g(0)|$.

2.1. Isotropic case with $\beta_i = 0, i = 1, \ldots, n$. We investigate the case with $H^+ = H^-$ and $\beta_i = 0, i = 0, \ldots, n$. In this case, f is an even function and g an odd one. And we have $\phi_L = -\phi_R$. Then (2.2) can be rewritten as

$$\begin{aligned} \phi_1' &= f(\phi_2) + g(\phi_2) - 2g(\phi_1), \\ \phi_i' &= f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}), \\ & \qquad i = 2, \ldots, n-1, \\ \phi_n' &= -f(\phi_{n-1}) - 2g(\phi_n) + g(\phi_{n-1}). \end{aligned} \tag{2.3}$$

First of all, let's look at the initial value problem (IVP) (2.1) with $\theta_i(0) = c$ where c is any real number. Then by the facts that $H^+ = H^-$ and $\omega_i \equiv \omega$ (since $\beta_i = 0$, $i = 1, \ldots, n$), we have $\theta_i(t) = \theta_{n+2-i}(t)$ for $t \geq 0, i = 1, \ldots, n+1$. Then the IVP (2.3) with $\phi_i(0) = 0$ shall yield $\phi_i(t) \equiv -\phi_{n+1-i}(t)$. That inspires us to study the system including only half the number of equations of (2.3).

LEMMA 2.1. *Let $n = 2m - 1$. Assume that f and g satisfy the conditions* (H1), (H2), *and $f(0) > 0$; then the IVP* (2.3) *with $\phi_i(0) = 0, i = 1, \dots, n$, has the following monotonicity along the trajectory:*

$$\phi_L > \phi_1(t) > \phi_2(t) > \cdots > \phi_{m-1}(t) > \phi_m(t) \equiv 0 \tag{2.4}$$

and

$$\phi_i'(t) > 0, i = 1, \dots, m-1 \tag{2.5}$$

for $0 < t < \hat{t}$, where $\hat{t}$ is such that $\phi_i'(\hat{t}) = 0$, $i = 1, \dots, m$, or $\hat{t} = +\infty$.

Remark. The fixed point always happens at $t = +\infty$ for an autonomous system. So we should have $\hat{t} = +\infty$ here. But a finite positive $\hat{t}$ does not affect our results. Hence we define $\hat{t}$ in the above way for the convenience of proof.

Proof. As we mentioned, $\phi_m(t) = -\phi_{n+1-m}(t) = -\phi_m(t)$ for $t \geq 0$. Then $\phi_m(t) \equiv 0$ is obvious. Since we only use half the number of equations (2.3), we restate them as

$$\begin{aligned} \phi_1' &= f(\phi_2) + g(\phi_2) - 2g(\phi_1), \\ \phi_i' &= f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}), \\ &\quad i = 2, \dots, m-2, \\ \phi_{m-1}' &= f(0) - f(\phi_{m-2}) - 2g(\phi_{m-1}) + g(\phi_{m-2}). \end{aligned} \tag{2.6}$$

Therefore $\phi_1'(0) = f(0) > 0, \phi_i'(0) = 0, i = 2, \dots, m-1$ (where we use the fact that $g(0) = 0$ since g is odd).

Furthermore, one can show by induction that

$$\begin{aligned} \phi_1'(0) &> 0, \\ \phi_i'(0) = \cdots = \phi_i^{(i-1)}(0) &= 0, \\ \phi_i^{(i)}(0) = g'(0)\phi_{i-1}^{(i-1)}(0) &> 0, \quad i = 2, \dots, m-1. \end{aligned} \tag{2.7}$$

Remark. By (H1), $g'(0) > 0$ such that $\phi_i^{(i)}(0) = [g'(0)]^{i-1}\phi_1'(0) > 0$.

So there exists small $\delta > 0$ such that (2.4) and (2.5) hold for $0 < t < \delta$ if one applies the Taylor's expansion for $\phi_i(t)$ and $\phi_i'(t)$ around $t = 0$. Starting with this result, we need to show that (2.4) and (2.5) are always true for $t > 0$.

By contradiction, suppose that there is a first place t_0 where (2.4) and (2.5) break down. Then we need to study the following cases.

CASE 1. $\phi_L = \phi_1(t_0) \geq \phi_2(t_0) \geq \cdots \geq \phi_{m-1}(t_0) > \phi_m(t_0) \equiv 0$ *and* $\phi_i'(t_0) \geq 0$, $i = 1, \dots, m-1$.

Then

$$\begin{aligned} 0 \leq \phi_1'(t_0) &= f(\phi_2(t_0)) + g(\phi_2(t_0)) - 2g(\phi_1(t_0)) \\ &= f(\phi_2(t_0)) + g(\phi_2(t_0)) - 2g(\phi_L) \\ &= f(\phi_2(t_0)) + g(\phi_2(t_0)) - f(\phi_L) - g(\phi_L) \\ &= [f'(\xi) + g'(\xi)](\phi_2(t_0) - \phi_L) \\ &\leq 0, \end{aligned}$$

where $\xi \in (\phi_2(t_0), \phi_L)$ by the mean value theorem and $(f' + g')(\xi) > 0$ by (H1). This leads to $\phi_2(t_0) = \phi_L$.

By induction on i, we shall gain $\phi_i(t_0) = \phi_L$, $i = 2, \ldots, m-1$.

Then we have

$$\begin{aligned} 0 &\le f(0) - f(\phi_L) - g(\phi_L) \\ &= f(0) + g(0) - f(\phi_L) - g(\phi_L) \\ &= (f' + g')(\xi)(0 - \phi_L), \end{aligned}$$

which implies $\phi_L \le 0$. This leads to contradiction since $\phi_L > 0$. Therefore *Case* 1 *is impossible.*

CASE 2. $\phi_L > \phi_1(t_0) > \cdots > \phi_j(t_0) = \phi_{j+1} \ge \cdots \ge \phi_{m-1}(t_0) > \phi_m(t_0) = 0$ *for some* $j \in \{1, 2, \ldots, m-2\}$ *and* $\phi_i'(t_0) \ge 0$ $\forall i \in \{1, \ldots, m-1\}$.

Then

$$\begin{aligned} 0 \le \phi_{j+1}'(t_0) &= f(\phi_{j+2}(t_0)) - f(\phi_j(t_0)) + g(\phi_{j+2}(t_0)) - 2g(\phi_{j+1}(t_0)) + g(\phi_j(t_0)) \\ &= f(\phi_{j+2}(t_0)) - f(\phi_{j+1}(t_0)) + g(\phi_{j+2}(t_0)) - g(\phi_{j+1}(t_0)) \\ &= [f' + g'](\xi)(\phi_{j+2}(t_0) - \phi_{j+1}(t_0)) \\ &\le 0, \end{aligned}$$

which implies $\phi_{j+2}(t_0) = \phi_{j+1}(t_0)$ (since $f' + g' > 0$ in J and $\phi_{j+1}(t_0) \ge \phi_{j+2}(t_0)$).

By induction, we have $\phi_L > \phi_1(t_0) > \cdots > \phi_j(t_0) = \phi_{j+1}(t_0) = \cdots = \phi_{m-1}(t_0) > \phi_m(t_0) = 0$.

Then

$$\begin{aligned} 0 \le \phi_{m-1}'(t_0) &= f(0) - f(\phi_{m-1}(t_0)) - g(\phi_{m-1}(t_0)) \\ &= f(0) + g(0) - f(\phi_{m-1}(t_0)) - g(\phi_{m-1}(t_0)) \\ &= [f' + g'](\xi)(0 - \phi_{m-1}(t_0)), \end{aligned}$$

which implies $\phi_{m-1}(t_0) \le 0$: a contradiction!

Therefore we *eliminate the possibility of Case* 2.

CASE 3. $\phi_L > \phi_1(t_0) > \cdots > \phi_{m-1}(t_0) > \phi_m(t_0) = 0$ *and* $\phi_i'(t_0) \ge 0$ $\forall i$ *and* $\phi_j'(t_0) = 0$ *for some* $j \in \{1, \ldots, m-1\}$.

First of all, if $j = 1$, i.e., $\phi_1'(t_0) = 0$, then we must have $\phi_2'(t_0) = 0$. Otherwise $\phi_2'(t_0) > 0$; then for $\varepsilon > 0$ small enough, we have

$$\begin{aligned} \phi_1'(t_0 - \varepsilon) &= \phi_1'(t_0) - \phi_1''(t_0)\varepsilon + o(\varepsilon^2) \\ &= \phi_1'(t_0) - [f'(\phi_2(t_0)) + g'(\phi_2(t_0))]\phi_2'(t_0)\varepsilon + 2g'(\phi_1(t_0))\phi_1'(t_0)\varepsilon + o(\varepsilon^2) \\ &= -[f'(\phi_2(t_0)) + g'(\phi_2(t_0))]\phi_2'(t_0)\varepsilon + o(\varepsilon^2) \\ &< 0. \end{aligned}$$

This is a contradiction since t_0 is the first place where (2.4) and (2.5) break down.

Furthermore, we can get $\phi_i'(t_0) = 0, i = 2, \ldots, m-1$ by using the techniques of induction and contradiction. Taking $\hat{t} = t_0$, we are done with the proof.

Secondly, assume that $\phi_i'(t_0) > 0, i = 1, \ldots, j-1$, and $\phi_j'(t_0) = 0$ for some $j \in \{2, \ldots, m-1\}$. Then by applying the same technique above and noting that $g' - f' > 0$ in J, we will obtain $\phi_{j-1}'(t_0) = 0$, which is a contradiction. Hence *we eliminate Case* 3.

Now by getting rid of Cases 1–3, we can conclude that either there exists a $\hat{t} > 0$ such that (2.4) and (2.5) hold for $0 < t < \hat{t}$ and $\phi_i'(\hat{t}) = 0, i = 1, \ldots, m$, or the first place t_0 where (2.4) and (2.5) break down does not exist. The proof is completed. □

Remark 1. In the proof of Lemma 2.1, the monotonicity of solution along the trajectory plays an important role. In order to get monotonicity at the start of the trajectory, we need the initial vector $\phi_i(0) = 0, i = 1, \ldots, n$. For other initial vectors, monotonicity fails.

Remark 2. Throughout this paper, we always start from $\phi_i(0) = 0$. As we can see in the following sections, if the monotonicity fails on the trajectory, we cannot continue the proof theoretically. But numerical experiments show that the solution trajectory of (2.2) always converges to the same equilibrium for any initial vector. It seems that the basin of attraction is infinitely large.

LEMMA 2.2. *Let $n = 2m$. Assume that f and g satisfy the conditions* (H1), (H2), *and $f(0) > 0$; then the IVP* (2.3) *with $\phi_i(0) = 0, i = 1, \ldots, n$, has the following monotonicity along the trajectory:*

$$\phi_L > \phi_1(t) > \phi_2(t) > \cdots > \phi_{m-1}(t) > \phi_m(t) > 0 \tag{2.8}$$

and

$$\phi_i'(t) > 0, \quad 0 < t < \hat{t}, \quad i = 1, \ldots, m, \tag{2.9}$$

where $\hat{t}$ is such that $\phi_i'(\hat{t}) = 0$, $i = 1, \ldots, m$, or $\hat{t} = +\infty$.

Proof. The proof is similar to the proof of Lemma 2.1. The difference is that we should restate the equations of (2.3) as

$$\begin{aligned} \phi_1' &= f(\phi_2) + g(\phi_2) - 2g(\phi_1), \\ \phi_i' &= f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}), \\ &\qquad i = 2, \ldots, m-1, \\ \phi_m' &= f(\phi_m) - f(\phi_{m-1}) - 3g(\phi_m) + g(\phi_{m-1}). \end{aligned} \tag{2.10}$$

All the techniques from Lemma 2.1 can be applied here so we ignore the details. ∎

THEOREM 2.3. *Assume f and g satisfy the same conditions as in Lemmas* 2.1 *and* 2.2; *then the IVP* (2.3) *with $\phi_i(0) = 0$, $i = 1, \ldots, n$ has the following properties:*

(i) *For each $i \in \{1, \ldots, n\}$, there exists $\bar{\phi}_i$ such that $\lim_{t \to \hat{t}} \phi_i(t) = \bar{\phi}_i$;*

(ii) *$(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ is the fixed point of the system* (2.3);

(iii) *$\phi_L > \bar{\phi}_1 > \bar{\phi}_2 > \cdots > \bar{\phi}_{n-1} > \bar{\phi}_n > \phi_R$;*

(iv) *$\bar{\phi}_i = -\bar{\phi}_{n+1-i}$, $i = 1, \ldots, n$.*

Proof. By the results of Lemmas 2.1 and 2.2, (i), (ii), and (iv) are easy to check. Also we have $\phi_L \geq \bar{\phi}_1 \geq \bar{\phi}_2 \geq \cdots \geq \bar{\phi}_{n-1} \geq \bar{\phi}_n \geq \phi_R$. We need to show that all the inequalities are strict. By contradiction, suppose $\phi_L = \bar{\phi}_1$. Then we have

$$\begin{aligned} 0 &= f(\bar{\phi}_2) + g(\bar{\phi}_2) - 2g(\bar{\phi}_1) \\ &= f(\bar{\phi}_2) + g(\bar{\phi}_2) - 2g(\phi_L) \\ &= f(\bar{\phi}_2) + g(\bar{\phi}_2) - [f(\phi_L) + g(\phi_L)] \\ &= [f' + g'](\xi)(\bar{\phi}_2 - \phi_L), \end{aligned}$$

which implies $\bar{\phi}_2 = \phi_L$.

Then we would have $\bar{\phi}_i = \phi_L$, $i = 1, \ldots, n$ by induction on i.

And $0 = -f(\phi_L) - 2g(\phi_L) + g(\phi_L) = -f(\phi_L) - g(\phi_L)$ by the last equation of (2.3) such that $f(\phi_L) = -g(\phi_L)$ which leads to $\phi_L = \phi_R$. This contradicts $\phi_L = -\phi_R$ since $\phi_L > 0$. Hence $\phi_L > \bar{\phi}_1$ must hold.

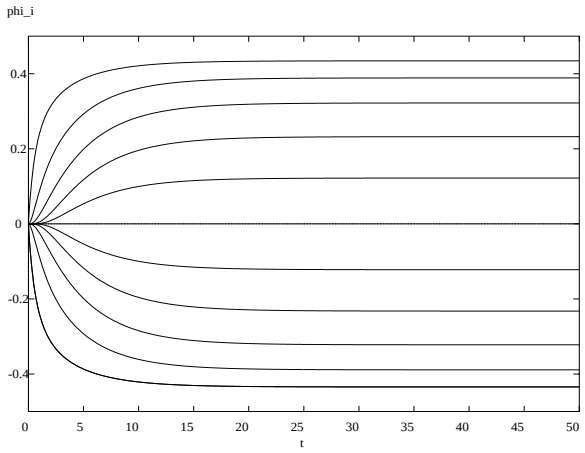

FIG. 2.1. *The isotropic case with* $H^+(\phi) = H^-(\phi) = H(\phi) = .5\cos\phi + \sin\phi$, $n = 11$, *and* $\beta_i = 0.0$.

Suppose $\bar{\phi}_1 = \bar{\phi}_2$; then $0 = f(\bar{\phi}_2) + g(\bar{\phi}_2) - 2g(\bar{\phi}_1)$ by (2.3). This is $0 = f(\bar{\phi}_1) - g(\bar{\phi}_1)$, which implies $\bar{\phi}_1 = \phi_L$. So we must have $\phi_L > \bar{\phi}_1 > \bar{\phi}_2$. By the symmetry, we have $\phi_R < \bar{\phi}_n < \bar{\phi}_{n-1}$.

Suppose i is the first index such that $\bar{\phi}_i = \bar{\phi}_{i+1}$; then

$$\begin{aligned} 0 &= f(\bar{\phi}_{i+1}) - f(\bar{\phi}_{i-1}) + g(\bar{\phi}_{i+1}) - 2g(\bar{\phi}_i) + g(\bar{\phi}_{i-1}) \\ &= f(\bar{\phi}_i) - f(\bar{\phi}_{i-1}) - g(\bar{\phi}_i) + g(\bar{\phi}_{i-1}) \\ &= f(\bar{\phi}_i) - g(\bar{\phi}_{i-1}) - g(\bar{\phi}_i) + g(\bar{\phi}_{i-1}) \\ &= (g' - f')(\xi)(\bar{\phi}_{i-1} - \bar{\phi}_i), \end{aligned}$$

which implies $\bar{\phi}_{i-1} = \bar{\phi}_i$, a contradiction.

Hence $\phi_L > \bar{\phi}_1 > \bar{\phi}_2 > \cdots > \bar{\phi}_{n-1} > \bar{\phi}_n > \phi_R$. ∎

In Figure 2.1 we illustrate the theory of Lemma 2.1 and Theorem 2.3 with a numerical example. Here we take $H^+(\phi) = H^-(\phi) = 0.5\cos\phi + \sin\phi$. Then $\phi_L = -\phi_R = \arctan(0.5) \approx 0.464$ and $J = (-\arctan 2, \arctan 2) \approx (-1.107, 1.107)$ for the conditions (H1) and (H2). We implemented the numerical computation by using the interactive package **XPPAUT** which was developed by B. Ermentrout. From the top to the bottom, the curves are $\phi_1(t), \ldots, \phi_{11}(t)$ ($n = 11$), respectively. Note $\phi_6(t) \equiv 0$ is on the x-axis. The figure shows the monotonicity and symmetry of solution $(\phi_1(t), \ldots, \phi_n(t))$ along the trajectory.

In Lemmas 2.1 and 2.2 and Theorem 2.3, we have the condition $f(0) > 0$. For $f(0) < 0$, the results and the proofs are very similar. We just state Theorem 2.4 without proof.

THEOREM 2.4. *Assume* f *and* g *satisfy* (H1), (H2), *and* $f(0) < 0$; *then the IVP* (2.3) *with* $\phi_i(0) = 0$, $i = 1, \ldots, n$ *has the following properties:*

(i) *For each* $i \in \{1, \ldots, n\}$, *there exists* $\bar{\phi}_i$ *such that* $\lim_{t\to\hat{t}} \phi_i(t) = \bar{\phi}_i$;

(ii) $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ *is the fixed point of the system* (2.3);

(iii) $\phi_L < \bar{\phi}_1 < \bar{\phi}_2 < \cdots < \bar{\phi}_{n-1} < \bar{\phi}_n < \phi_R$;

(iv) $\bar{\phi}_i = -\bar{\phi}_{n+1-i}$, $i = 1, \ldots, n$.

We turn our attention back to the system (2.1). Notice that we have

$$\omega + H^+(\bar{\phi}_1) = \omega + H^-(-\bar{\phi}_{i-1}) + H^+(\bar{\phi}_i) = \omega + H^-(-\bar{\phi}_n), i = 2, \ldots, n-1.$$

We take $\Omega = \omega + H^-(-\bar{\phi}_{i-1}) + H^+(\bar{\phi}_i)$; then $\theta_1 = \Omega t$, $\theta_i = \Omega t + \sum_{k=1}^{i-1} \bar{\phi}_k$, $i = 2, \ldots, n+1$, is the phaselocked solution of (2.1). Before showing that this phaselocked solution is stable, we state a general stability result due to Ermentrout [10].

THEOREM 2.5 (Ermentrout, 1992). *Consider the equations*

$$d\theta_k/dt = H_k(\theta_1 - \theta_k, \ldots, \theta_M - \theta_k), \quad k = 1, \ldots, M. \tag{2.11}$$

Let $\theta_k = \Omega t + \bar{\psi}_k$ be a phaselocked solution and let

$$a_{jk} = \partial H_k(z_1, \ldots, z_M)/\partial z_j \tag{2.12}$$

evaluated at $z_j = \bar{\psi}_j - \bar{\psi}_k$. Suppose that $a_{jk} \geq 0$ and the graph of the matrix (a_{jk}) is complete. Then the phaselocked solution is orbitally asymptotically stable in the sense that there is a simple zero eigenvalue corresponding to translation in time and all other eigenvalues have negative real parts.

Due to (H1), we have $g' \pm f' > 0$ in J. Then the phaselocked solution $\theta_1 = \Omega t$, $\theta_i = \Omega t + \sum_{k=1}^{i-1} \bar{\phi}_k$, $i = 2, \ldots, n+1$, satisfies the nonnegativity assumption in Theorem 2.5. The graph of (a_{jk}) is complete since $a_{i,i+1} > 0$ and $a_{i+1,i} > 0$ for $i = 1, \ldots, n$. So we have shown that the phaselocked solution is asymptotically stable. This result is summarized in the following theorem.

THEOREM 2.6. *Under the conditions of Theorem* 2.3 *or* 2.4, $\theta_1 = \Omega t$, $\theta_i = \Omega t + \sum_{k=1}^{i-1} \bar{\phi}_k$, $i = 2, \ldots, n+1$, *is the phaselocked solution of* (2.1), *orbitally asymptotically stable in the sense that there is a simple zero eigenvalue corresponding to translation in time and other eigenvalues have negative real parts.*

As a matter of fact, in Theorem 2.6, all the n nonzero eigenvalues with negative real parts are actually the eigenvalues of the system (2.3) linearized around the equilibrium $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$.

COROLLARY 2.7. *Under the conditions of Theorems* 2.3 *or* 2.4, $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ *is an asymptotically stable steady state of* (2.3) *and all the eigenvalues of the system* (2.3) *linearized around it have negative real parts.*

2.2. Isotropic case with $\beta_i = \beta \neq 0, i = 1, \ldots, n$. Throughout this section, without loss of generality, we assume $\beta < 0$. If $\beta > 0$, you can subtract the consecutive equations of (2.1) in another direction such that the frequency difference is less than zero. In this case, we still have $H^+ = H^-$, which implies that f is even and g odd such that $\phi_L = -\phi_R$. We restate (2.2) in the form

$$\begin{aligned} \phi_1' &= \beta + f(\phi_2) + g(\phi_2) - 2g(\phi_1), \\ \phi_i' &= \beta + f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}), \\ &\qquad i = 2, \ldots, n-1, \\ \phi_n' &= \beta - f(\phi_{n-1}) - 2g(\phi_n) + g(\phi_{n-1}). \end{aligned} \tag{2.13}$$

For $\beta = 0$, we have that $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ is the asymptotically stable steady state of (2.13) following Corollary 2.7. Then if $|\beta|$ is small enough, we should get an asymptotically stable steady state $\bar{\phi}_i(\beta)$, $i = 1, \ldots, n$ near $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ by the implicit function theorem. We denote the trajectory by $\phi_i(t, \beta)$, $i = 1, \ldots, n$ for the IVP (2.13) with $\phi_i(0) = 0$. By continuity, $\phi_i(t, \beta)$ should have the monotonicity and boundedness as in Lemmas 2.1 and 2.2, and $\phi_i(t, \beta) \to \bar{\phi}_i(\beta)$ as $t \to +\infty$ if $|\beta|$ is small enough. We summarize this fact in Theorem 2.8

THEOREM 2.8. *Assume that f and g satisfy* (H1) *and* (H2). *Let $|\beta|$ be small enough; then the IVP* (2.13) *with $\phi_i(0) = 0$ satisfies that*

(i) *if* $f(0) > 0$, *then*

$$\phi_L > \phi_1(t,\beta) > \cdots > \phi_n(t,\beta) > \phi_R; \tag{2.14}$$

(ii) *if* $f(0) < 0$, *then*

$$\phi_L < \phi_1(t,\beta) < \cdots < \phi_n(t,\beta) < \phi_R. \tag{2.15}$$

Also $\phi_i(t,\beta) \to \bar{\phi}_i(\beta)$ *as* $t \to +\infty$ *for* $i = 1,\ldots,n$, *where* $(\bar{\phi}_1(\beta),\ldots,\bar{\phi}_n(\beta))$ *is the asymptotically stable steady state of* (2.13) *near* $(\bar{\phi}_1,\ldots,\bar{\phi}_n)$.

Theorem 2.8 is not a particularly strong result. To keep the monotonicity and boundedness, $|\beta|$ has to be assumed very small. We would like to know when the monotonicity breaks down. This leads to the following theorem.

THEOREM 2.9. *Assume that* f *and* g *satisfy* (H1), (H2), *and* $f(0) > 0$. *Let* $|\beta|$ *be small enough that* $\phi_{n-1}(t) \geq \phi_R$, $t > 0$ *for the IVP* (2.13) *with* $\phi_i(0) = 0$, $i = 1,\ldots,n$. *Then we have the following properties along the trajectory:*

(i) *there is a sequence* $\{t_k\}_{k=1}^{\infty}$ *(it could be a finite sequence) such that* $0 = t_1 < t_2 < \cdots < t_k < \cdots < \hat{t}$, *and for each* k, *there is* $l_k \in \{1,\ldots,n\}$ *so that*

$$\phi_i'(t) > 0, \quad i = 1,\ldots,l_k, \quad t_k < t < t_{k+1},$$

$$\phi_j'(t) < 0, \quad j = l_k+1,\ldots,n, \quad t_k < t < t_{k+1}, \tag{2.16}$$

$$l_{k+1} \in \{0, l_k - 1, l_k, l_k + 1, n\}, \tag{2.17}$$

$$\text{either } \phi_{l_k}'(t_{k+1}) = 0 \text{ or } \phi_{l_k+1}'(t_{k+1}) = 0 \text{ (not both)}, \tag{2.18}$$

$$\phi_L > \phi_1(t) > \cdots > \phi_{n-1}(t) > \phi_n(t) > \phi_\beta, \quad t_k < t \leq t_{k+1} \tag{2.19}$$

where $\hat{t}$ *is such that* $\phi_i'(\hat{t}) = 0$, $i = 1,\ldots,n$ *or* $\hat{t} = +\infty$, *and* $\phi_\beta \in J$ *is such that* $f(\phi_\beta) + g(\phi_\beta) = \beta$ *(note that* $\phi_\beta < \phi_R$*).*

(ii) *for each* $i \in \{1,\ldots,n\}$, *there exists* $\bar{\phi}_i$ *such that*

$$\lim_{t \to \hat{t}} \phi_i(t) = \bar{\phi}_i, \tag{2.20}$$

$$\phi_L > \bar{\phi}_1 > \cdots > \bar{\phi}_n > \phi_\beta, \tag{2.21}$$

and $(\bar{\phi}_1,\ldots,\bar{\phi}_n)$ *is a fixed point of* (2.13).

Remark. The condition $\phi_{n-1}(t) \geq \phi_R, t > 0$ means that ϕ_{n-1} cannot cross ϕ_R along the trajectory. It is weaker than the condition in Theorem 2.8 since it allows ϕ_n to cross ϕ_R. It holds when $|\beta|$ is small enough (but not as small as in Theorem 2.8) according to the results of section 2.1.

The proof of the theorem is very long. We put it in the Appendix for interested readers.

Figure 2.2 is a numerical solution illustrating Theorem 2.9. Here $H^+(\phi) = H^-(\phi) = 0.5\cos\phi + \sin\phi$ and $\beta = -0.005$. The figure shows monotonicity of solution along the trajectory.

For the case $f(0) < 0$, we have results parallel to Theorem 2.9.

THEOREM 2.10. *Assume that* f *and* g *satisfy* (H1), (H2), *and* $f(0) < 0$. *Let* $|\beta|$ *be small enough that* $\phi_2(t) \geq \phi_L$, $t > 0$ *for the IVP* (2.13) *with* $\phi_i(0) = 0$, $i = 1,\ldots,n$. *Then we have the following properties along the trajectory:*

(i) *there is a sequence* $\{t_k\}_{k=1}^{\infty}$ *(it could be a finite sequence) such that* $0 = t_1 < t_2 < \cdots < t_k < \cdots < \hat{t}$ *and for each* k, *there is* $l_k \in \{1,\ldots,n\}$ *so that*

$$\phi_i'(t) < 0, \quad i = 1,\ldots,l_k, \quad t_k < t < t_{k+1},$$

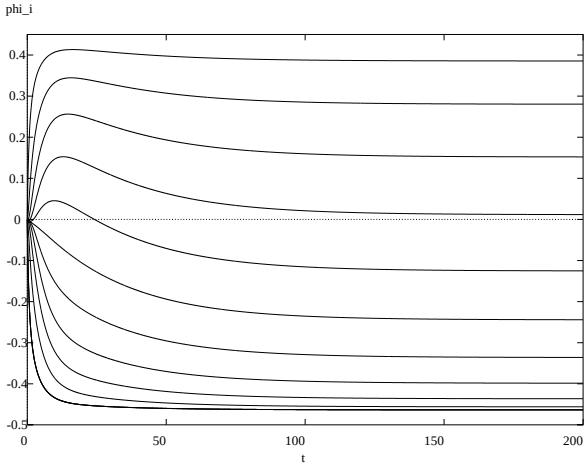

FIG. 2.2. *The isotropic case with* $H^+(\phi) = H^-(\phi) = H(\phi) = .5\cos\phi + \sin\phi$, $n = 11$, *and* $\beta_i = \beta = -0.005$.

$$\phi_j'(t) > 0, \quad j = l_k + 1, \dots, n, \quad t_k < t < t_{k+1}, \tag{2.22}$$

$$l_{k+1} \in \{0, l_k - 1, l_k, l_k + 1, n\}, \tag{2.23}$$

$$\text{either } \phi_{l_k}'(t_{k+1}) = 0 \text{ or } \phi_{l_k+1}'(t_{k+1}) = 0 \text{ (not both)}, \tag{2.24}$$

$$\phi_\beta < \phi_1(t) < \cdots < \phi_{n-1}(t) < \phi_n(t) < \phi_R, \quad t_k < t \le t_{k+1}, \tag{2.25}$$

where $\hat{t}$ *is such that* $\phi_i'(\hat{t}) = 0$, $i = 1, \dots, n$ *or* $\hat{t} = +\infty$, *and* $\phi_\beta \in J$ *is such that* $-f(\phi_\beta) + g(\phi_\beta) = \beta$ *(note that* $\phi_\beta < \phi_L$*).*

(ii) *for each* $i \in \{1, \dots, n\}$, *there exists* $\bar{\phi}_i$ *such that*

$$\lim_{t \to \hat{t}} \phi_i(t) = \bar{\phi}_i, \tag{2.26}$$

$$\phi_\beta < \bar{\phi}_1 < \cdots < \bar{\phi}_n < \phi_R, \tag{2.27}$$

and $(\bar{\phi}_1, \dots, \bar{\phi}_n)$ *is a fixed point of* (2.13).

As we did in section 2.1, if we let $\theta_1(t) = \Omega t$, $\theta_i(t) = \Omega t + \sum_{k=1}^{i-1} \bar{\phi}_k$, $i = 1, \dots, n+1$, where $(\bar{\phi}_1, \dots, \bar{\phi}_n)$ is the fixed point of (2.13) which we obtained in Theorems 2.9 and 2.10, then Theorem 2.5 assures us that $(\theta_1(t), \dots, \theta_{n+1}(t))$ is an orbitally asymptotically stable phaselocked solution of (2.1).

2.3. Nonisotropic case with $\beta_i = 0, i = 1, \dots, n$. In this case, we have $H^+ \neq H^-$ which implies that f is not even and g is not odd anymore. So $\phi_L \neq -\phi_R$ in general. And we would like to restate (2.2) in the form

$$\begin{aligned} \phi_1' &= f(\phi_2) + g(\phi_2) - 2g(\phi_1), \\ \phi_i' &= f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}), \\ &\qquad i = 2, \dots, n-1, \\ \phi_n' &= -f(\phi_{n-1}) - 2g(\phi_n) + g(\phi_{n-1}). \end{aligned} \tag{2.28}$$

THEOREM 2.11. *Assume that* f *and* g *satisfy* (H1), (H2), *and* $f(0) > |g(0)|$. *Then the IVP* (2.28) *with* $\phi_i(0) = 0$, $i = 1, \dots, n$ *has the following properties along the trajectory:*

(i) *There is a sequence* $\{t_k\}_{k=1}^{\infty}$ *(it could be a finite sequence) such that* $0 = t_1 < t_2 < \cdots < t_k < \cdots < \hat{t}$ *and for each* k*, there is* $l_k \in \{1, \ldots, n\}$ *so that*

$$\phi_i'(t) > 0, \quad i = 1, \ldots, l_k, \quad t_k < t < t_{k+1},$$

$$\phi_j'(t) < 0, \quad j = l_k + 1, \ldots, n, \quad t_k < t < t_{k+1}, \tag{2.29}$$

$$l_{k+1} \in \{0, l_k - 1, l_k, l_k + 1, n\}, \tag{2.30}$$

$$\text{either } \phi_{l_k}'(t_{k+1}) = 0 \text{ or } \phi_{l_k+1}'(t_{k+1}) = 0 \text{ (not both)}, \tag{2.31}$$

$$\phi_L > \phi_1(t) > \cdots > \phi_{n-1}(t) > \phi_n(t) > \phi_R, \quad t_k < t \le t_{k+1}, \tag{2.32}$$

where $\hat{t}$ *is such that* $\phi_i'(\hat{t}) = 0$*,* $i = 1, \ldots, n$*, or* $\hat{t} = +\infty$*.*

(ii) *For each* $i \in \{1, \ldots, n\}$*, there exists* $\bar{\phi}_i$ *such that*

$$\lim_{t \to \hat{t}} \phi_i(t) = \bar{\phi}_i, \tag{2.33}$$

$$\phi_L > \bar{\phi}_1 > \cdots > \bar{\phi}_n > \phi_R, \tag{2.34}$$

and $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ *is a fixed point of* (2.28)*.*

Proof. Note that $f(0) > |g(0)|$; then

$$\begin{aligned} \phi_1'(0) &= f(0) - g(0) > 0, \\ \phi_i'(0) &= 0, \quad i = 2, \ldots, n-1, \\ \phi_n'(0) &= -f(0) - g(0) < 0. \end{aligned} \tag{2.35}$$

Then by (2.28) and (2.35), we have

$$\begin{aligned} \phi_2''(0) &= [g'(0) - f'(0)][f(0) - g(0)] > 0, \\ \phi_{n-1}''(0) &= [g'(0) + f'(0)][-f(0) - g(0)] < 0. \end{aligned} \tag{2.36}$$

By induction on i, we can get that for $i = 3, \ldots, m-1$

$$\begin{aligned} \phi_i^{(k)}(0) &= 0, \quad k = 1, \ldots, i-1, \\ \phi_i^{(i)}(0) &= [g'(0) - f'(0)]^{i-1}[f(0) - g(0)] > 0, \\ \phi_{n-i+1}^{(k)}(0) &= 0, \quad k = 1, \ldots, i-1, \\ \phi_{n-i+1}^{(i)}(0) &= [g'(0) - f'(0)]^{i-1}[-f(0) + g(0)] < 0 \end{aligned} \tag{2.37}$$

whenever $n = 2m-1$ or $2m-2$. And when $n = 2m-1$, we have extra terms

$$\phi_m^{(k)}(0) = 0, \quad k = 1, \ldots, m. \tag{2.38}$$

Assume $\phi_m^{(m+1)}(0) \neq 0$ (otherwise we can figure out $\phi_m^{(M)}(0) \neq 0$ and $\phi_m^{(k)}(0) = 0, k = 1, \ldots, M-1$).

Without loss of generality, we assume $\phi_m^{m+1}(0) > 0$ when $n = 2m-1$.

Then we have $t_1 = 0, l_1 = m-1$ when $n = 2m-2$, and $t_1 = 0, l_1 = m$ when $n = 2m-1$. And the rest of the proof just mimics all the steps of proving Theorem 2.9 ∎

Again in Figure 2.3 we show the results of Theorem 2.11. Here $H^+(\phi) = H(\phi)$ and $H^-(\phi) = 0.2H(\phi)$ where $H(\phi) = 0.5\cos\phi + \sin\phi$. And $\phi_L = -\phi_R = \arctan(0.5)$ and $J = (-\arctan 2, \arctan 2)$. The monotonicity of the solution along the trajectory can be seen from the figure. Also we see that the solution converges to a fixed point.

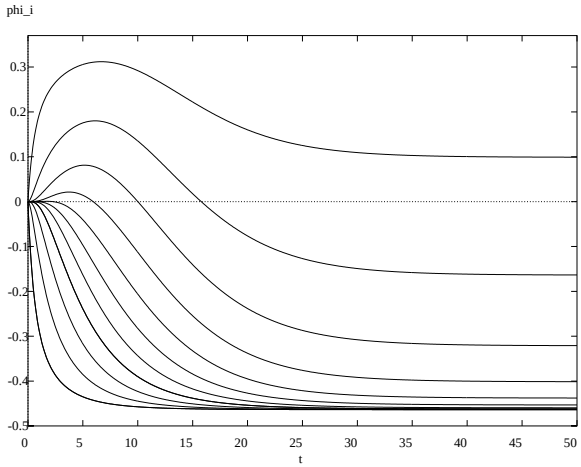

FIG. 2.3. *The nonisotropic case with* $H^+(\phi) = H(\phi), H^-(\phi) = 0.2H(\phi), H(\phi) = .5\cos\phi + \sin\phi$, $n = 11$, *and* $\beta_i = \beta = 0$.

THEOREM 2.12. *Assume that* f *and* g *satisfy* (H1), (H2), *and* $f(0) < -|g(0)|$. *Then the IVP* (2.28) *with* $\phi_i(0) = 0$, $i = 1, \ldots, n$ *has the following properties along the trajectory:*

(i) *there is a sequence* $\{t_k\}_{k=1}^{\infty}$ *(it could be a finite sequence) such that* $0 = t_1 < t_2 < \cdots < t_k < \cdots < \hat{t}$ *and for each* k, *there is* $l_k \in \{1, \ldots, n\}$ *so that*

$$\phi_i'(t) < 0, \quad i = 1, \ldots, l_k, \quad t_k < t < t_{k+1},$$

$$\phi_j'(t) > 0, \quad j = l_k + 1, \ldots, n, \quad t_k < t < t_{k+1}, \tag{2.39}$$

$$l_{k+1} \in \{0, l_k - 1, l_k, l_k + 1, n\}, \tag{2.40}$$

$$\text{either } \phi_{l_k}'(t_{k+1}) = 0 \text{ or } \phi_{l_k+1}'(t_{k+1}) = 0 \text{ (not both)}, \tag{2.41}$$

$$\phi_L < \phi_1(t) < \cdots < \phi_{n-1}(t) < \phi_n(t) < \phi_R, t_k < t \le t_{k+1}, \tag{2.42}$$

where $\hat{t}$ *is such that* $\phi_i'(\hat{t}) = 0$, $i = 1, \ldots, n$ *or* $\hat{t} = +\infty$.

(ii) *for each* $i \in \{1, \ldots, n\}$, *there exists* $\bar{\phi}_i$ *such that*

$$\lim_{t \to \hat{t}} \phi_i(t) = \bar{\phi}_i, \tag{2.43}$$

$$\phi_L < \bar{\phi}_1 < \cdots < \bar{\phi}_n < \phi_R, \tag{2.44}$$

and $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ *is a fixed point of* (2.28).

2.4. Nonisotropic case with $\beta_i = \beta \neq 0, i = 1, \ldots, n$. Throughout this section, without loss of generality, we assume $\beta < 0$. If $\beta > 0$, you can subtract the consecutive equations of (2.1) in another direction such that the frequency difference is less than zero. In this case, like in section 2.2, we have $H^+ \neq H^-$ which implies f is not even and g not odd such that $\phi_L \neq -\phi_R$. And we would like to restate (2.2) in the form

$$\begin{aligned} \phi_1' &= \beta + f(\phi_2) + g(\phi_2) - 2g(\phi_1), \\ \phi_i' &= \beta + f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}), \\ &\quad i = 2, \ldots, n-1, \\ \phi_n' &= \beta - f(\phi_{n-1}) - 2g(\phi_n) + g(\phi_{n-1}). \end{aligned} \tag{2.45}$$

THEOREM 2.13. *Assume that f and g satisfy* (H1), (H2), *and $f(0) > |g(0)|$. Let $|\beta|$ be small enough that $\phi_{n-1}(t) \geq \phi_R$, $t > 0$ for the IVP* (2.45) *with $\phi_i(0) = 0$, $i = 1, \ldots, n$. Then we have the following properties along the trajectory:*

(i) *there is a sequence $\{t_k\}_{k=1}^{\infty}$ (it could be a finite sequence) such that $0 = t_1 < t_2 < \cdots < t_k < \cdots < \hat{t}$ and for each k, there is $l_k \in \{1, \ldots, n\}$ so that*

$$\phi_i'(t) > 0, \quad i = 1, \ldots, l_k, \quad t_k < t < t_{k+1},$$

$$\phi_j'(t) < 0, \quad j = l_k + 1, \ldots, n, \quad t_k < t < t_{k+1}, \tag{2.46}$$

$$l_{k+1} \in \{0, l_k - 1, l_k, l_k + 1, n\}, \tag{2.47}$$

$$\text{either } \phi_{l_k}'(t_{k+1}) = 0 \text{ or } \phi_{l_k+1}'(t_{k+1}) = 0 \text{ (not both)}, \tag{2.48}$$

$$\phi_L > \phi_1(t) > \cdots > \phi_{n-1}(t) > \phi_n(t) > \phi_\beta, \quad t_k < t \leq t_{k+1}, \tag{2.49}$$

where $\hat{t}$ is such that $\phi_i'(\hat{t}) = 0$, $i = 1, \ldots, n$ or $\hat{t} = +\infty$, and $\phi_\beta \in J$ is such that $f(\phi_\beta) + g(\phi_\beta) = \beta$ (note that $\phi_\beta < \phi_R$).

(ii) *for each $i \in \{1, \ldots, n\}$, there exists $\bar{\phi}_i$ such that*

$$\lim_{t \to \hat{t}} \phi_i(t) = \bar{\phi}_i, \tag{2.50}$$

$$\phi_L > \bar{\phi}_1 > \cdots > \bar{\phi}_n > \phi_\beta, \tag{2.51}$$

and $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ is a fixed point of (2.13).

Figures 2.4(a) and 2.4(b) are numerical illustrations of Theorem 2.13. Here $H^+(\phi) = H(\phi)$ and $H^-(\phi) = 0.2H(\phi)$, where $H(\phi) = 0.5\cos\phi + \sin\phi$. And $\phi_L = -\phi_R = \arctan(0.5)$ and $J = (-\arctan 2, \arctan 2)$. Note that in Fig. 2.4(b), $\phi_{n-1}(t)$ crosses ϕ_R somewhere so that the monotonicity is destroyed. However, the trajectory still converges to a fixed point. Hence the monotonicity is not necessary for the convergence of the solution. In Fig. 2.4(a) the monotonicity is preserved since the $|\beta|$ is so small that $\phi_{n-1}(t)$ does not cross ϕ_R.

THEOREM 2.14. *Assume that f and g satisfy* (H1), (H2), *and $f(0) < -|g(0)|$. Let $|\beta|$ be small enough that $\phi_2(t) \geq \phi_L$, $t > 0$ for the IVP* (2.45) *with $\phi_i(0) = 0$, $i = 1, \ldots, n$. Then we have the following properties along the trajectory:*

(i) *there is a sequence $\{t_k\}_{k=1}^{\infty}$ (it could be a finite sequence) such that $0 = t_1 < t_2 < \cdots < t_k < \cdots < \hat{t}$ and for each k, there is $l_k \in \{1, \ldots, n\}$ so that*

$$\phi_i'(t) < 0, \quad i = 1, \ldots, l_k, \quad t_k < t < t_{k+1},$$

$$\phi_j'(t) > 0, \quad j = l_k + 1, \ldots, n, \quad t_k < t < t_{k+1}, \tag{2.52}$$

$$l_{k+1} \in \{0, l_k - 1, l_k, l_k + 1, n\}, \tag{2.53}$$

$$\text{either } \phi_{l_k}'(t_{k+1}) = 0 \text{ or } \phi_{l_k+1}'(t_{k+1}) = 0 \text{ (not both)}, \tag{2.54}$$

$$\phi_\beta < \phi_1(t) < \cdots < \phi_{n-1}(t) < \phi_n(t) < \phi_R, \quad t_k < t \leq t_{k+1}, \tag{2.55}$$

where $\hat{t}$ is such that $\phi_i'(\hat{t}) = 0$, $i = 1, \ldots, n$ or $\hat{t} = +\infty$, and $\phi_\beta \in J$ is such that $-f(\phi_\beta) + g(\phi_\beta) = \beta$ (note that $\phi_\beta < \phi_L$).

(ii) *for each $i \in \{1, \ldots, n\}$, there exists $\bar{\phi}_i$ such that*

$$\lim_{t \to \hat{t}} \phi_i(t) = \bar{\phi}_i, \tag{2.56}$$

$$\phi_\beta < \bar{\phi}_1 < \cdots < \bar{\phi}_n < \phi_R, \tag{2.57}$$

and $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ is a fixed point of (2.45).

By Theorem 2.5, the fixed points $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ from the two theorems above are asymptotically stable steady state of (2.45).

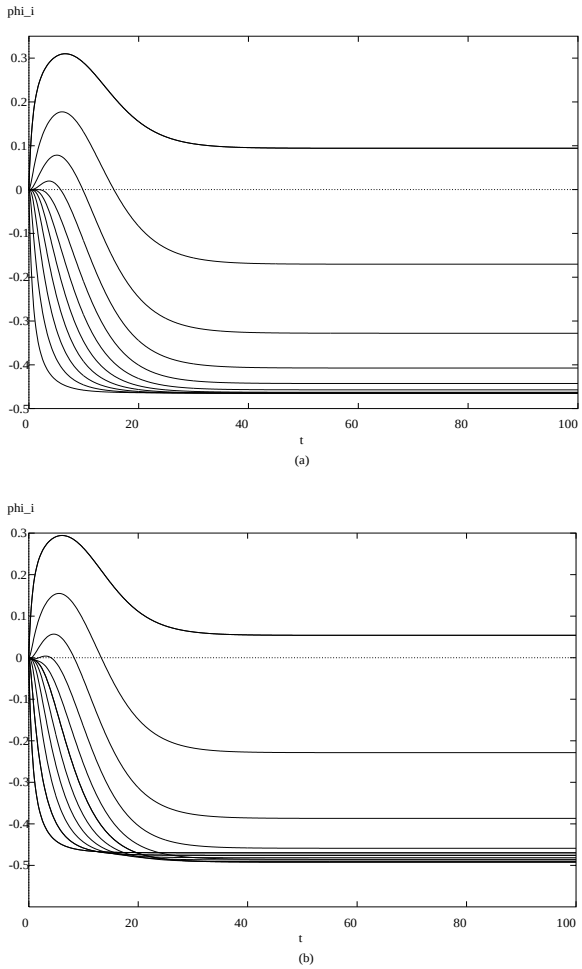

FIG. 2.4. *The nonisotropic case with* $H^+(\phi) = H(\phi), H^-(\phi) = 0.2H(\phi), H(\phi) = .5\cos\phi + \sin\phi$, $n = 11$, (a) $\beta_i = \beta = -0.0005$, (b) $\beta_i = \beta = -0.005$.

3. Arrays of oscillators. In this section, we consider a two-dimensional array of coupled oscillators. The equations to be considered have the form

$$\begin{aligned}\theta'_{ij} &= \omega_{ij} + H^{+X}(\theta_{i+1,j} - \theta_{ij}) + H^{-X}(\theta_{i-1,j} - \theta_{ij}) \\ &\quad + H^{+Y}(\theta_{i,j+1} - \theta_{ij}) + H^{-Y}(\theta_{i,j-1} - \theta_{ij}), \\ &\quad i, j = 1, \ldots, n+1, \end{aligned} \tag{3.1}$$

where H^{+X}, H^{+Y}, H^{-X}, and H^{-Y} are smooth 2π-periodic functions of the arguments and ω_{ij} is the frequency for each oscillator.

Note that in (3.1), each oscillator is coupled with its four nearest neighbors. The term H^{-X} (respectively, H^{+X} or H^{-Y} or H^{+Y}) is ignored for $i = 1$ (respectively, $i = n+1$ or $j = 1$ or $j = n+1$). We take

$$\begin{aligned}\phi_{ij} &= \theta_{i+1,j} - \theta_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n+1, \\ \psi_{ij} &= \theta_{i,j+1} - \theta_{ij}, \quad i = 1, \ldots, n+1, \quad j = 1, \ldots, n,\end{aligned}$$

$$\alpha_{ij} = \omega_{i+1,j} - \omega_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, n+1,$$
$$\beta_{ij} = \omega_{i,j+1} - \omega_{ij}, \quad i = 1, \dots, n+1, \quad j = 1, \dots, n$$

and define the functions f, g, p, and q as

$$\begin{aligned} f(\phi) + g(\phi) &= H^{+X}(\phi), \\ f(\phi) - g(\phi) &= H^{-X}(-\phi), \\ p(\psi) + q(\psi) &= H^{+Y}(\psi), \\ p(\psi) - q(\psi) &= H^{-Y}(-\psi). \end{aligned} \tag{3.2}$$

Then in (3.1), if we subtract the (i,j)th equation from the $(i+1,j)$th one and the $(i,j+1)$th one, respectively, we have

$$\begin{aligned} \phi'_{ij} = {} & \alpha_{ij} + f(\phi_{i+1,j}) - f(\phi_{i-1,j}) + g(\phi_{i+1,j}) - 2g(\phi_{ij}) + g(\phi_{i-1,j}) \\ & + p(\psi_{i+1,j}) + p(\psi_{i+1,j-1}) - p(\psi_{ij}) - p(\psi_{i,j-1}) \\ & + q(\psi_{i+1,j}) - q(\psi_{i+1,j-1}) - q(\psi_{ij}) + q(\psi_{i,j-1}), \\ & i = 1, \dots, n, j = 1, \dots, n+1, \end{aligned}$$

$$\begin{aligned} \psi'_{ij} = {} & \beta_{ij} + p(\psi_{i,j+1}) - p(\psi_{i,j-1}) + q(\psi_{i,j+1}) - 2q(\psi_{ij}) + q(\psi_{i,j-1}) \\ & + f(\phi_{i,j+1}) + f(\phi_{i-1,j+1}) - f(\phi_{ij}) - f(\phi_{i-1,j}) \\ & + g(\phi_{i,j+1}) - g(\phi_{i-1,j+1}) - g(\phi_{ij}) + g(\phi_{i-1,j}) \\ & i = 1, \dots, n+1, j = 1, \dots, n. \end{aligned} \tag{3.3}$$

Note that the index (i,j) for ϕ_{ij} should satisfy $1 \le i \le n$ and $1 \le j \le n+1$ and the index (i,j) for ψ_{ij} should satisfy $1 \le i \le n+1$, $1 \le j \le n$. Hence if (i,j) is out of range for ϕ_{ij} or ψ_{ij}, the corresponding terms on the right-hand sides of (3.3) are ignored.

Again we define several constants related to f, g, p, and q. We define

- ϕ_L as $f(\phi_L) = g(\phi_L)$, i.e., $H^{-X}(-\phi_L) = 0$;
- ϕ_R as $f(\phi_R) = -g(\phi_R)$, i.e., $H^{+X}(\phi_R) = 0$;
- ψ_L as $p(\psi_L) = q(\psi_L)$, i.e., $H^{-Y}(-\psi_L) = 0$;
- ψ_R as $p(\psi_R) = -q(\psi_R)$, i.e., $H^{+Y}(\psi_R) = 0$.

We assume some hypotheses on f, g, p, and q in sufficiently large intervals J_X and J_Y around $\phi = 0$ and $\psi = 0$, respectively:

(HX1) *$g'(\phi) > |f'(\phi)|$ for $\phi \in J_X$;*

(HX2) *there exists a unique ϕ_L (respectively, ϕ_R) to $f = g$ (respectively, $f = -g$) for $\phi \in J_X$;*

(HY1) *$q'(\psi) > |p'(\psi)|$ for $\psi \in J_Y$;*

(HY2) *there exists a unique ψ_L (respectively, ψ_R) to $p = q$ (respectively, $p = -q$) for $\psi \in J_Y$.*

Note that (HX1), (HX2), (HY1), and (HY2) are the assumptions extended from the chain model.

Our goal is to apply the results obtained from the chain model to this array model. In order to achieve this task, let us first consider a very special system of equations:

$$\begin{aligned} \phi'_{ij} &= F_{ij}(\Phi) + G_{ij}(\Psi), \quad i = 1, \dots, n, \quad j = 1, \dots, n+1, \\ \psi'_{ij} &= P_{ij}(\Psi) + Q_{ij}(\Phi), \quad i = 1, \dots, n+1, \quad j = 1, \dots, n, \end{aligned} \tag{3.4}$$

where

$$\Phi = (\phi_{ij})_{n\times(n+1)} = [\Phi_1, \ldots, \Phi_{n+1}]$$

with

$$\Phi_j = \begin{bmatrix} \phi_{1j} \\ \vdots \\ \phi_{nj} \end{bmatrix}, \quad j = 1, \ldots, n+1$$

and

$$\Psi = (\psi_{ij})_{(n+1)\times n} = \begin{bmatrix} \Psi_1 \\ \vdots \\ \Psi_{n+1} \end{bmatrix}$$

with

$$\Psi_i = [\psi_{i1}, \ldots, \psi_{in}], \quad i = 1, \ldots, n+1$$

and F_{ij}, P_{ij}, G_{ij}, and Q_{ij} satisfy the following assumptions:

(i) *if* $\Phi_1 = \Phi_2 = \cdots = \Phi_{n+1}$ *(i.e.,* ϕ_{ij} *is independent of the index* j*), then*

$$F_{i1}(\Phi) = F_{i2}(\Phi) = \cdots = F_{i,n+1}(\Phi), \quad i = 1, \ldots, n, \tag{3.5}$$

$$Q_{ij}(\Phi) = 0, i = 1, \ldots, n+1, \quad j = 1, \ldots, n; \tag{3.6}$$

(ii) *if* $\Psi_1 = \Psi_2 = \cdots = \Psi_{n+1}$ *(i.e.,* ψ_{ij} *is independent of the index* i*), then*

$$P_{1j}(\Psi) = P_{2j}(\Psi) = \cdots = P_{n+1,j}(\Psi), \quad j = 1, \ldots, n, \tag{3.7}$$

$$G_{ij}(\Psi) = 0, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n+1. \tag{3.8}$$

Remark. The special form of (3.4) is a generalization of the system (3.3). We will see this later. The conditions on F_{ij}, G_{ij}, P_{ij}, and Q_{ij} reflect a homogeneity requirement for the two-dimensional domain. That is, the phase lags between left and right neighbors are the same for each row. Similarly, the lags between top and bottom neighbors are the same for each column.

LEMMA 3.1. *The set* $S = \{(\Phi, \Psi) | \Phi_1 = \Phi_2 = \cdots = \Phi_{n+1}$ *and* $\Psi_1 = \Psi_2 = \cdots = \Psi_{n+1}\}$ *is an invariant set for the system* (3.4).

Proof. We only need to show that if $(\Phi(0), \Psi(0)) \in S$, then $\Phi_1'(0) = \cdots = \Phi_{n+1}'(0)$ and $\Psi_1'(0) = \cdots = \Psi_{n+1}'(0)$, i.e.,

$$\phi_{i1}'(0) = \phi_{i2}'(0) = \cdots = \phi_{i,n+1}'(0) \quad \text{for} \quad i = 1, \ldots, n, \tag{3.9}$$

$$\psi_{1j}'(0) = \psi_{2j}'(0) = \cdots = \psi_{n+1,j}'(0) \quad \text{for} \quad j = 1, \ldots, n. \tag{3.10}$$

By (3.4), (3.5), and (3.8), we have that for each $i \in \{1, \ldots, n\}$,

$$\begin{aligned} \phi_{ij}'(0) &= F_{ij}(\Phi(0)) + G_{ij}(\Psi(0)) \\ &= F_{ik}(\Phi(0)) + G_{ik}(\Psi(0)) \\ &= \phi_{ik}'(0). \end{aligned}$$

Hence (3.9) is proven. Also, we can prove (3.10) in the same way. ∎

LEMMA 3.2. *In the system* (3.3), *if we assume*

$$\alpha_{ij} = \alpha_i \text{ and } \beta_{ij} = \beta_j \tag{3.11}$$

then (3.3) *is a system of the type* (3.4).

Proof. (3.3) is a special case of (3.4) where

$$\begin{aligned}
F_{ij}(\Phi) &= \alpha_{ij} + f(\phi_{i+1,j}) - f(\phi_{i-1,j}) + g(\phi_{i+1,j}) - 2g(\phi_{ij}) + g(\phi_{i-1,j}),\\
G_{ij}(\Psi) &= p(\psi_{i+1,j}) + p(\psi_{i+1,j-1}) - p(\psi_{ij}) - p(\psi_{i,j-1})\\
&\quad + q(\psi_{i+1,j}) - q(\psi_{i+1,j-1}) - q(\psi_{ij}) + q(\psi_{i,j-1}),\\
P_{ij}(\Psi) &= \beta_{ij} + p(\psi_{i,j+1}) - p(\psi_{i,j-1}) + q(\psi_{i,j+1}) - 2q(\psi_{ij}) + q(\psi_{i,j-1}),\\
Q_{ij}(\Phi) &= f(\phi_{i,j+1}) + f(\phi_{i-1,j+1}) - f(\phi_{ij}) - f(\phi_{i-1,j})\\
&\quad + g(\phi_{i,j+1}) - g(\phi_{i-1,j+1}) - g(\phi_{ij}) + g(\phi_{i-1,j}).
\end{aligned}$$

Since we have (3.11), α_{ij} is independent of j and β_{ij} is independent of i. Then if ϕ_{ij} is independent of j and ψ_{ij} is independent of i, (3.5)–(3.8) are satisfied. The proof is completed. ∎

Remark. (3.11) means that the distribution of intrinsic frequencies is a sum of two stripe distributions: one with constant frequencies along each row, and another with constant frequencies along each column. Hence ω_{ij} is in the form of $\omega_{ij} = \omega_i^X + \omega_j^Y$.

LEMMA 3.3. *If the system* (3.3) *satisfies* (3.11), *then the IVP* (3.3) *with*

$$\phi_{ij}(0) = 0, \quad i = 1, \dots, n, \quad j = 1, \dots, n+1, \tag{3.12}$$

$$\psi_{ij}(0) = 0, \quad i = 1, \dots, n+1, \quad j = 1, \dots, n \tag{3.13}$$

has the following identity property:

$$\phi_{i1}(t) = \phi_{i2}(t) = \cdots = \phi_{i,n+1}(t), \quad i = 1, \dots, n, \tag{3.14}$$

$$\psi_{1j}(t) = \psi_{2j}(t) = \cdots = \psi_{n+1,j}(t), \quad j = 1, \dots, n \tag{3.15}$$

for $t \geq 0$.

Proof. This is an immediate result of Lemmas 3.1 and 3.2. ∎

Hence the IVP (3.3), (3.12), and (3.13) satisfying (3.11) is reduced to two independent systems of chain model, i.e.,

$$\begin{aligned}
\phi_1' &= \alpha_1 + f(\phi_2) + g(\phi_2) - 2g(\phi_1),\\
\phi_i' &= \alpha_i + f(\phi_{i+1}) - f(\phi_{i-1}) + g(\phi_{i+1}) - 2g(\phi_i) + g(\phi_{i-1}),\\
&\quad i = 2, \dots, n-1,\\
\phi_n' &= \alpha_n - f(\phi_{n-1}) - 2g(\phi_n) + g(\phi_{n-1})
\end{aligned} \tag{3.16}$$

and

$$\begin{aligned}
\psi_1' &= \beta_1 + p(\psi_2) + q(\psi_2) - 2q(\psi_1),\\
\psi_j' &= \beta_j + p(\psi_{j+1}) - p(\psi_{j-1}) + q(\psi_{j+1}) - 2q(\psi_j) + q(\psi_{j-1}),\\
&\quad j = 2, \dots, n-1,\\
\psi_n' &= \beta_n - p(\psi_{n-1}) - 2q(\psi_n) + q(\phi_{n-1}),
\end{aligned} \tag{3.17}$$

where $\phi_i = \phi_{i1} = \cdots = \phi_{i,n+1}$ and $\psi_j = \psi_{1j} = \cdots = \psi_{n+1,j}$.

Note that both (3.16) and (3.17) are exactly in the form of (2.2).

THEOREM 3.4. *If the trajectories of the IVP* (3.16) *with* $\phi_i(0) = 0$ *and the IVP with* (3.17) *with* $\psi_j(0) = 0$ *converge to the fixed point* $(\bar{\phi}_1, \ldots, \bar{\phi}_n)$ *of* (3.16) *and the fixed point* $(\bar{\psi}_1, \ldots, \bar{\psi}_n)$ *of* (3.17) *respectively, then the trajectory of the IVP* (3.3) *with* (3.12) *and* (3.13) *goes to* $((\bar{\phi}_{ij})_{n\times(n+1)}, (\bar{\psi}_{ij})_{(n+1)\times n})$ *which is the fixed point of* (3.3), *where*

$$\bar{\phi}_{ij} = \bar{\phi}_i, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n+1$$

and

$$\bar{\psi}_{ij} = \bar{\psi}_j, \quad i = 1, \ldots, n+1, \quad j = 1, \ldots, n.$$

Also, $\Omega \equiv \omega_{ij} + H^{+X}(\bar{\phi}_{ij}) + H^{-X}(-\bar{\phi}_{i-1,j}) + H^{+Y}(\bar{\psi}_{ij}) + H^{-Y}(-\bar{\psi}_{i,j-1})$ *(*$i, j = 1, \ldots, n+1$*) is the locked frequency of* (3.1).

Proof. This is a straightforward result of Lemma 3.3. □

Now if we let $\theta_{1,1}(t) = \Omega t$, $\theta_{ij}(t) = \Omega t + \sum_{k=1}^{i-1} \bar{\phi}_k + \sum_{k=1}^{j-1} \bar{\psi}_k$, $\{\theta_{ij}(t)\}$ is the phaselocked solution of (3.1). And it is orbitally asymptotically stable by Theorem 2.5.

Therefore, all the results which we obtained in section 2 can be extended to this system.

Remark 3. If the condition (3.11) is not satisfied, we will not achieve the reduction. But if $\omega_{ij} = \omega_i^X + \omega_j^Y + o(\varepsilon)$ for small ε, we still get a stable phaselocked solution by the implicit function theorem.

Remark 4. The reduction technique could be applied to three-dimensional arrays of oscillators as long as ω_{ijk} is in the form of $\omega_{ijk} = \omega_i^X + \omega_j^Y + \omega_k^Z$. And the array models could be reduced to three independent chain models.

The following are some numerical results for the two-dimensional arrays of oscillators (3.1) and the reduced chains (3.16) and (3.17). For all cases, $\omega_{ij} = \omega_i^X + \omega_j^Y$ is assumed. A basic function $H(\phi) = 0.5\cos\phi + \sin\phi$ is assumed.

Example 1. Let $H^{+X} = H^{-X} = H^{+Y} = H^{-Y} = H$ and $\omega_{ij} \equiv \omega > 0$. Then (HX1), (HX2), (HY1), and (HY2) are satisfied with $J_X = J_Y = (-\arctan 2, \arctan 2)$ and $\phi_L = -\phi_R = \psi_L = -\psi_R = \arctan(0.5)$. Also, $f = p$ and $g = q$. Since $\omega_{ij} \equiv \omega$, the condition (3.11) holds so that the array system (3.3) can be reduced to the two chain systems (3.16) and (3.17) by Lemma 3.3. And (3.16) and (3.17) have asymptotically stable equilibria following the results in section 2.1. Then (3.3) has an asymptotically stable equilibrium. Noting that $f = p$, $g = q$, and $\alpha_i = \beta_j = 0$, the solutions of (3.16) and (3.17) are the identical. So we only study the solution $\bar{\phi}_i$ of (3.16). Figure 3.1(a) is the plot for $\bar{\phi}_i$ where $(i/(n+1), \bar{\phi}_i)$ are the coordinates. We can see that there is a wave traveling outward in both directions from the midpoint of the chain [5, 11]. The wave speed is almost constant except near the middle. By Theorem 3.4, $\bar{\phi}_{ij} = \bar{\phi}_i$ and $\bar{\psi}_{ij} = \bar{\psi}_j$. Then for the array, we have a wave traveling outward from the midpoint of the array. Figure 3.1b shows this observation by plotting the relative phases. As we mentioned in the introduction, with isotropic "synaptic coupling" target patterns are the generic phaselocked behavior. (See the remarks at the end of this section for a discussion about other stable patterns.)

Example 2. Let $H^{+X} = H^{+Y} = 1.5H$, $H^{-X} = H^{-Y} = 0.5H$, and $\omega_{ij} \equiv \omega > 0$. Then (HX1), (HX2), (HY1), and (HY2) hold with $J_X = J_Y = (-\arctan 2, \arctan 2)$ and $\phi_L = -\phi_R = \psi_L = -\psi_R = \arctan(0.5)$. Also, $f = p$ and $g = q$. The reduction from an array to two chains is then obtained. These two chains are identical according to our choice of coupling functions. Figure 3.2 shows the results for the reduced chains and the array. There is a wave traveling from the left of chain to the right. Thus there is a wave traveling from the southwest corner to the northeast corner of the array.

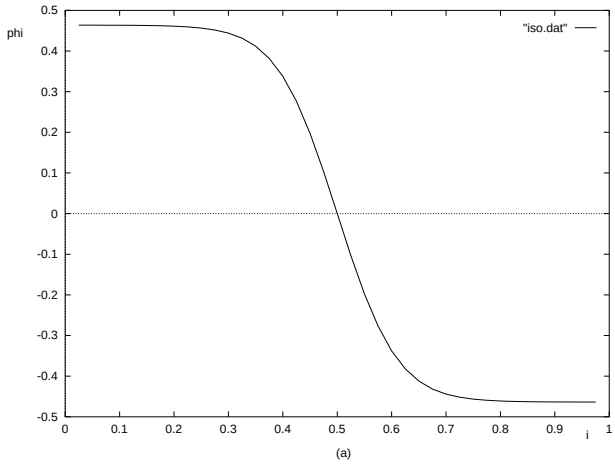

FIG. 3.1. $n+1=40$. (a) *Phase lags of the reduced chains. There is a wave traveling outward in both directions from the midpoint of the chain. The wave speed is almost constant except near the middle.* (b) *Relative phases of the array. There is a wave traveling outward from the midpoint of the array.*

Example 3. The coupling functions are the same as in Example 2. We choose $\omega_{ij} = 2\omega + 0.1[1 - i/(n+1)] + 0.1[1 - j/(n+1)]$ which is in the form $\omega_{ij} = \omega_i^X + \omega_j^Y$, where $\omega_i^X = \omega + 0.1[1 - i/(n+1)]$ and $\omega_j^X = \omega + 0.1[1 - j/(n+1)]$. Then the solutions of the two chain systems (3.16) and (3.17) are the same. Figure 3.3 shows the numerical solutions for the reduced chains and the array.

Example 4. In this example, we show how the size of the chain can apparently affect the qualitative features of the phases in one- and two-dimensional arrays. In Fig. 3.4(a), we show the results of a simulation with a 50×50 array of oscillators with no frequency gradient and all of the interactions functions identical and given by $H(\phi) = \sin\phi + 0.05\cos\phi + 0.8$. The phases give the appearance of a circularly symmetric target pattern, quite different from the rectangular-looking patterns of Figure 3.1. This effect can be understood by looking at the behavior of the chain. In Figure 3.4b, the phase-shifts between successive oscillators are shown for a chain with $n = 50$ and $n = 500$ oscillators. In the case of $n = 50$ the phase-difference

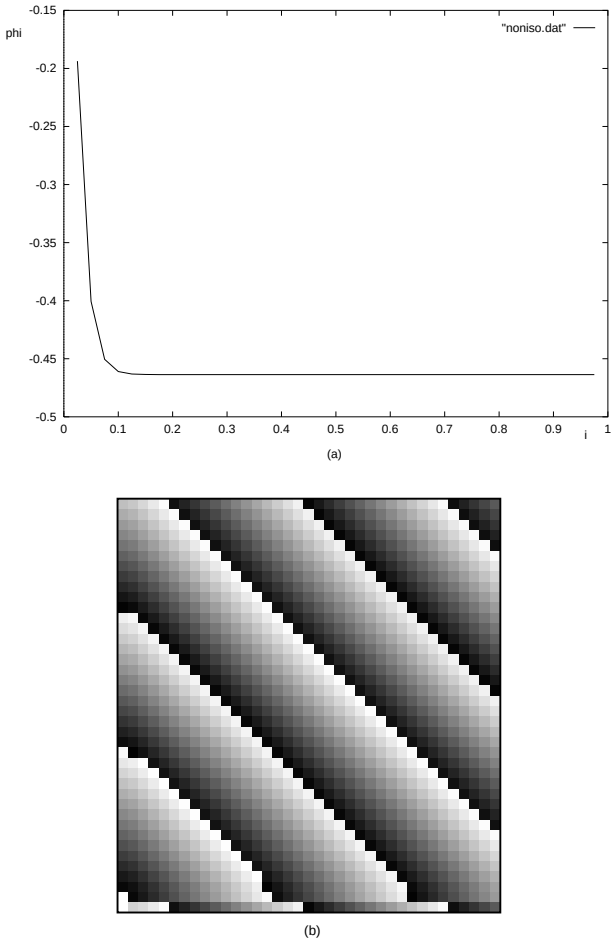

FIG. 3.2. $n + 1 = 40$. (a) *Phase lags of the reduced chains. There is a wave traveling from the left of chain to the right.* (b) *Relative phases of the array. There is a wave traveling from the southwest corner to the northeast corner.*

is nearly a straight line so that the relative phases (which are the "integral" of the phase differences) are quadratic. Since the results of this section show that the array behaves like two orthogonal chains, it is now clear why the relative phases in the square array have apparently circular contours; the relative phase along any axes of the array are nearly quadratic. This is actually an artifact of the chain size. For, as n increases, Figure 3.4(b) shows that the phase differences become piecewise constant and so the relative phases will be linear and, in the array, will look like Figure 3.1. This is also what is predicted by the continuum theory in [5]. However, due to the small size of the cosine coefficient, n must be very large before there is qualitative similarity to the continuum approximation.

3.1. Some remarks on the stability of the patterns. In one-dimensional chains with "synaptic coupling" the traveling wave solutions described in section 2 appear to be the only stable solutions. That is, no matter what the initial conditions, solutions converge to the monotone solutions described in section 2. On the other hand, if the one-dimensional chain has a ring geometry so that the two ends are

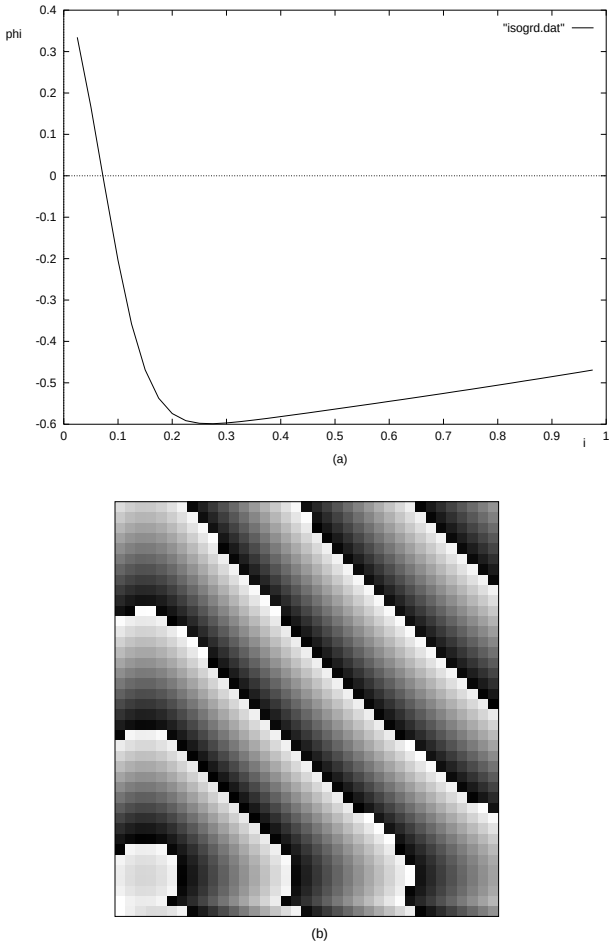

FIG. 3.3. $n+1=40$. (a) *Phase lags of the reduced chains.* (b) *Relative phases of the array.*

identified, then, there are several stable solutions that correspond to synchrony and traveling waves. Thus, the domain of attraction of any given solution varies and does not constitute the entire phase space. In particular, the larger the chain, the more different types of stable solutions are possible.

In two-dimensional systems, everything gets worse; there are many stable phaselocked patterns possible and a characterization of all of them remains a topic of current research. Finding domains of stability is even harder. Consider an $N \times N$ array where the coupling functions $H^{\pm X}, H^{\pm Y}$ are of the form

$$H(\phi) = \lambda \cos\phi + \sin\phi.$$

When $\lambda = 0$ one stable phaselocked solution is synchrony. As λ increases away from 0, the resulting phaselocked solution perturbs to the target-like patterns that we have discussed here. For $\lambda = 0$ Paullet and Ermentrout [9] have proven that there are also stable solutions analogous to spiral waves. Since these are stable, they persist for small λ and thus represent another phaselocked solution distinct from the target patterns described in this paper. Random initial data (rather than initial data identically 0)

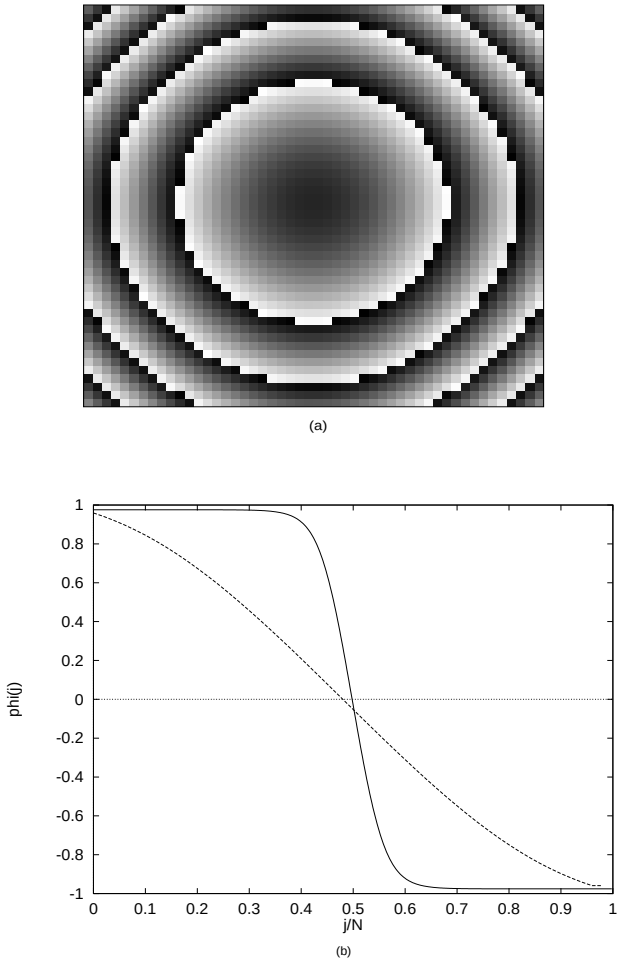

FIG. 3.4. *Relative phases in an array with almost circular symmetry and their analogue in a chain.* (a) *Relative phase for a* 50×50 *array.* (b) *A chain of length* 50 *and* 500 *showing how the almost quadratic behavior of the phase shifts for* $n = 50$ *becomes the piecewise linear phases for* $n = 500$ *as is predicted by the continuum equations.*

converge on phaselocked solutions, but sometimes they are not targets but rather are related to the spiral patterns. For small arrays, random initial data converge mainly to the target patterns but on larger arrays (e.g., 40×40) the tendency is to converge to series of broken spiral-like patterns. Thus, target patterns are "homotopes" of synchrony and have essentially the same global stability behavior. They are not unique phaselocked patterns, unlike their analogue in one dimension.

Appendix. Proof of Theorem 2.9.

(i) We prove it by applying induction on $k \in N$. Let $k = 1$. Note that $|\beta|$ is small; we have $f(0) + \beta > 0$. Then

$$\begin{aligned} \phi_1'(0) &= f(0) + \beta > 0, \\ \phi_i'(0) &= \beta < 0, \quad i = 2, \ldots, n-1, \\ \phi_n'(0) &= -f(0) + \beta < 0. \end{aligned} \tag{A.1}$$

Then by (2.13) and (A.1), we have

$$\begin{aligned} \phi_2''(0) &= g'(0)f(0), \\ \phi_{n-1}''(0) &= -g'(0)f(0). \end{aligned} \tag{A.2}$$

By induction on i, we can get that for $i = 3, \dots, m-1$

$$\begin{aligned} \phi_i'(0) = \beta, \phi_i^{(k)}(0) = 0, \quad k = 2, \dots, i-1, \\ \phi_i^{(i)}(0) = [g'(0)]^{i-1} f(0) > 0, \\ \phi_{n-i+1}'(0) = \beta, \phi_{n-i+1}^{(k)}(0) = 0, \quad k = 2, \dots, i-1, \\ \phi_{n-i+1}^{(i)}(0) = -[g'(0)]^{i-1} f(0) < 0, \end{aligned} \tag{A.3}$$

where $n = 2m-1$ or $n = 2m-2$ and

$$\phi_m'(0) = \beta, \phi_m''(0) = \cdots = \phi_m^{(m)}(0) = 0, \tag{A.4}$$

where $n = 2m-1$.

Hence by (A.1)–(A.4) and the fact that $\phi_i(0) = 0, i = 1, \dots, n$, one can apply the Taylor's formula to $\phi_i(t)$ and $\phi_i'(t)$. Then we have

$$\phi_1'(t) > 0, \phi_i'(t) < 0, \quad i = 2, \dots, n \tag{A.5}$$

and

$$\phi_L > \phi_1(t) > \cdots > \phi_n(t) > \phi_\beta \tag{A.6}$$

in $(0, \delta)$ for $\delta > 0$ small enough. Therefore $t_1 = 0$ and $l_1 = 1$.

CLAIM 1. *From $t = 0$, as long as* (A.5) *holds, we always have* (A.6).

Suppose that there is some first place t^* such that $\phi_L = \phi_1(t^*) \geq \phi_2(t^*) \geq \cdots \geq \phi_n(t^*) \geq \phi_\beta$. Then

$$\begin{aligned} \phi_1'(t^*) &= \beta + f(\phi_2(t^*)) + g(\phi_2(t^*)) - 2g(\phi_L), \\ &= \beta + f(\phi_2(t^*)) + g(\phi_2(t^*)) - f(\phi_L) - g(\phi_L), \\ &= \beta + [f'(\xi) + g'(\xi)](\phi_2(t^*) - \phi_L), \\ &\leq \beta. \end{aligned}$$

This is a contradiction since $\phi_1'(t^*) > 0$.

Now suppose that there is a first place t^* such that for some $i \in \{1, \dots, n-2\}$

$$\phi_L > \phi_1 > \phi_2 > \cdots > \phi_i = \phi_{i+1} \geq \cdots \geq \phi_{n-1} \geq \phi_n \geq \phi_\beta$$

at t^*. Then at this point t^*,

$$\begin{aligned} \phi_i' &= \beta + f(\phi_i) - f(\phi_{i-1}) + g(\phi_i) - 2g(\phi_i) + g(\phi_{i-1}) \\ &= \beta + f(\phi_i) - g(\phi_i) - f(\phi_{i-1}) + g(\phi_{i-1}) \\ &= \beta + [g' - f'](\xi_1)(\phi_{i-1} - \phi_i) \\ &> \beta \end{aligned}$$

and

$$\begin{aligned} \phi_{i+1}' &= \beta + f(\phi_{i+2}) - f(\phi_i) + g(\phi_{i+2}) - 2g(\phi_i) + g(\phi_i) \\ &= \beta + f(\phi_{i+2}) - f(\phi_i) + g(\phi_{i+2}) - g(\phi_i) \\ &= \beta + [g' - f'](\xi_2)(\phi_{i+2} - \phi_i) \\ &\leq \beta, \end{aligned}$$

so $\phi_i'(t^*) > \phi_{i+1}'(t^*)$. Therefore in a small neighborhood $(t^* - \delta, t^*)$ of t^* $(t^* > 0)$, we have $\phi_{i+1}(t) > \phi_i(t)$ since $\phi_i(t^*) = \phi_{i+1}(t^*)$. This leads to a contradiction.

Hence we can conclude that $\phi_L > \phi_1(t) > \phi_2(t) > \cdots > \phi_{n-1}(t) \geq \phi_n(t) \geq \phi_\beta$.

Suppose $\phi_L > \phi_1(t) > \phi_2(t) > \cdots > \phi_{n-1}(t) \geq \phi_n(t) = \phi_\beta$ at a first place t^*; then

$$\begin{aligned}\phi_n' &= \beta - f(\phi_{n-1}) + g(\phi_{n-1}) - 2g(\phi_\beta)\\ &= f(\phi_\beta) - g(\phi_\beta) - f(\phi_{n-1}) + g(\phi_{n-1})\\ &= [g' - f'](\xi)(\phi_{n-1} - \phi_\beta)\\ &\geq 0.\end{aligned}$$

This is a contradiction since we have $\phi_n'(t) < 0$ so far.

Hence $\phi_L > \phi_1(t) > \phi_2(t) > \cdots > \phi_{n-1}(t) \geq \phi_n(t) > \phi_\beta$.

Now suppose $\phi_L > \phi_1(t) > \phi_2(t) > \cdots > \phi_{n-1}(t) = \phi_n(t) > \phi_\beta$ at a first place t^*; then at t^*

$$\begin{aligned}\phi_n' &= \beta - f(\phi_n) + g(\phi_{n-1}) - 2g(\phi_n)\\ &= \beta - [f(\phi_{n-1}) + g(\phi_{n-1})]\end{aligned}$$

and

$$\begin{aligned}\phi_{n-1}' &= \beta + f(\phi_n) - f(\phi_{n-2}) + g(\phi_n) - 2g(\phi_{n-1}) + g(\phi_{n-2})\\ &= \beta + f(\phi_n) - f(\phi_{n-2}) - g(\phi_n) + g(\phi_{n-2})\\ &= \beta + [g' - f'](\xi_1)(\phi_{n-2} - \phi_n).\\ &> \beta.\end{aligned}$$

Since $\phi_{n-1}(t) \geq \phi_R$ for $t > 0$ by the assumption of the theorem,

$$\begin{aligned}f(\phi_{n-1}) + g(\phi_{n-1}) &= f(\phi_{n-1}) + g(\phi_{n-1}) - [f(\phi_R) + g(\phi_R)]\\ &= [f' + g'](\xi_2)(\phi_{n-1} - \phi_R)\\ &\geq 0\end{aligned}$$

at t^*. So $\phi_n'(t^*) < \phi_{n-1}'(t^*)$. Hence in a small neighborhood $(t^* - \delta, t^*)$ of t^*, we have $\phi_n(t) > \phi_{n-1}(t)$ which is a contradiction. Therefore Claim 1 is proven.

Suppose (A.5) breaks down at some first place $t_2 > 0$ (otherwise the proof is finished with $\hat{t} = +\infty$) and $\phi_i'(t_2) \neq 0$ for some $i \in \{1, \ldots, n\}$ (otherwise the proof is finished with $\hat{t} = t_2$). Then we have six cases to consider.

CASE 1. *There is some $l > 2$ such that $\phi_1'(t_2) \geq 0$, $\phi_{l-1}'(t_2) < 0$, $\phi_l'(t_2) = 0$, and $\phi_i'(t_2) \leq 0$ for $i \in \{2, \ldots, n\} - \{l-1, l\}$.*

CASE 2. *There is some $l \in \{3, \ldots, n-1\}$ such that $\phi_1'(t_2) \geq 0$, $\phi_i'(t_2) = 0$ for $i = 2, \ldots, l$ and $\phi_{l+1}'(t_2) < 0$, $l \in \{3, \ldots, n-1\}$.*

CASE 3. *$\phi_1'(t_2) = \phi_2'(t_2) = 0$ and $\phi_3'(t_2) < 0$, $\phi_i'(t_2) \leq 0$, $i = 4, \ldots, n$.*

CASE 4. *$\phi_1'(t_2) = 0$ and $\phi_i'(t_2) < 0$, $i = 2, \ldots, n$.*

CASE 5. *$\phi_1'(t_2) > 0$, $\phi_2'(t_2) = 0$ and $\phi_i'(t_2) < 0$, $i = 3, \ldots, n$.*

CASE 6. *$\phi_1'(t_2) > 0$ and $\phi_i'(t_2) = 0$, $i = 2, \ldots, n$.*

Assume Case 1 is true. Then we have

$$\begin{aligned}\phi_l'(t_2 - \varepsilon) = \beta &+ f(\phi_{l+1}(t_2 - \varepsilon)) - f(\phi_{l-1}(t_2 - \varepsilon))\\ &+ g(\phi_{l+1}(t_2 - \varepsilon)) - 2g(\phi_l(t_2 - \varepsilon)) + g(\phi_{l-1}(t_2 - \varepsilon))\end{aligned}$$

$$\begin{aligned}
&= \phi_l'(t_2) - f'(\phi_{l+1}(t_2))\phi_{l+1}'(t_2)\varepsilon + f'(\phi_{l-1}(t_2))\phi_{l-1}'(t_2)\varepsilon \\
&\quad - g'(\phi_{l+1}(t_2))\phi_{l+1}'(t_2)\varepsilon + 2g'(\phi_l(t_2))\phi_l'(t_2)\varepsilon \\
&\quad - g'(\phi_{l-1}(t_2))\phi_{l-1}'(t_2)\varepsilon + o(\varepsilon^2) \\
&= -\phi_{l+1}'(t_2)[g' + f'](\phi_{l+1}(t_2))\varepsilon \\
&\quad - \phi_{l-1}'(t_2)[g' - f'](\phi_{l-1}(t_2))\varepsilon + o(\varepsilon^2) \\
&> 0
\end{aligned}$$

for $\varepsilon > 0$ small enough (since $g' \pm f' > 0$ in J). This is a contradiction! So *Case* 1 *is eliminated in our concern.*

Assume Case 2 is true. Then

$$\begin{aligned}
\phi_l'(t_2 - \varepsilon) &= -\phi_{l+1}'(t_2)[g' + f'](\phi_{l+1}(t_2))\varepsilon + o(\varepsilon^2) \\
&> 0
\end{aligned}$$

for $\varepsilon > 0$ small enough. This is a contradiction! So *Case* 2 *is also eliminated.*

Case 3 *can be eliminated in the same way as Case* 2.

Hence we have Cases 4–6 left.

If case 4 is true, then

$$\begin{aligned}
\phi_1'(t_2 + \varepsilon) &= \phi_2'(t_2)[g' + f'](\phi_2(t_2))\varepsilon + o(\varepsilon^2) \\
&< 0
\end{aligned}$$

for small $\varepsilon > 0$. Then $l_2 = 0$ such that $l_2 = l_1 - 1$.

If Case 5 is true, then we have that for small $\varepsilon > 0$, either $\phi_2'(t) > 0$ in $(t_2, t_2 + \varepsilon)$ or $\phi_2'(t) < 0$ in $(t_2, t_2 + \varepsilon)$ (note that $\phi_2'(t) \equiv 0$ in $(t_2, t_2 + \varepsilon)$ cannot be true). Hence $l_2 = 2$, i.e., $l_2 = l_1 + 1$ or $l_2 = 1$, i.e., $l_2 = l_1$.

If Case 6 is true, then we can show that

$$\begin{aligned}
&\phi_2''(t_2) > 0, \\
&\phi_i^{(j)}(t_2) \;=\; 0, \quad j = 2, \ldots, i-1, \\
&\phi_i^{(i)}(t_2) \;>\; 0, \quad i = 3, \ldots, n
\end{aligned}$$

such that $\phi_i'(t) > 0$ $(i = 2, \ldots, n)$ in $(t_2, t_2 + \varepsilon)$ for ε small enough. Then $l_2 = n$.

And for Cases 4–6, we can prove by using the same techniques as above that

$$\phi_L > \phi_1(t_2) > \cdots > \phi_n(t_2) > \phi_\beta.$$

So we are done with $k = 1$.

Now suppose (2.16)–(2.19) hold for $1, 2, \ldots, k-1$ with $l_1, \ldots, l_k$, and $t_1 < t_2 < \ldots < t_k$.

Then for $t \in (t_k, t_k + \delta)$ ($\delta > 0$ is small)

$$\begin{aligned}
&\phi_i'(t) > 0, \quad i = 1, \ldots, l_k, \\
&\phi_j'(t) < 0, \quad j = l_k + 1, \ldots, n.
\end{aligned}$$

Claim 2. *From t_k, as long as*

$$\phi_i'(t) > 0, \quad i = 1, \ldots, l_k,$$

(A.7)

$$\phi_j'(t) < 0, \quad j = l_k + 1, \ldots, n,$$

we always have (A.6).

The proof is similar to Claim 1; we just ignore it here.

Suppose (A.7) breaks down at a first place $t_{k+1} > t_k$ and $\phi_i'(t_{k+1}) \neq 0$ for some $i \in \{1, \ldots, n\}$; then several cases should be considered carefully.

CASE 1. *There is $l < l_k$ such that $\phi_l'(t_{k+1}) \leq 0$, $\phi_{l+1}'(t_{k+1}) > 0$, $\phi_i'(t_{k+1}) \geq 0$ for $i \in \{1, \ldots, l_k\} - \{l, l+1\}$, and $\phi_j'(t_{k+1}) = 0$ for $j = l_k + 1, \ldots, n$.*

CASE 2. *There is $l > l_k + 1$ such that $\phi_i'(t_{k+1}) \geq 0$, $i = 1, \ldots, l_k$, $\phi_{l-1}'(t_{k+1}) < 0$, $\phi_l'(t_{k+1}) = 0$, and $\phi_j'(t_{k+1}) \leq 0$ for $j \in \{l_k + 1, \ldots, n\} - \{l-1, l\}$.*

CASE 3. *There is some $l \in \{2, \ldots, l_k - 1\}$ such that $\phi_i'(t_{k+1}) \geq 0$ for $i \in \{1, \ldots, l-2\}$, $\phi_{l-1}'(t_{k+1}) > 0, \phi_j'(t_{k+1}) = 0$ for $j \in \{l, \ldots, l_k\}$, and $\phi_j'(t_{k+1}) \leq 0$ for $j \in \{l_k + 1, \ldots, n\}$.*

CASE 4. *There is $l \in \{l_k + 2, \ldots, n-1\}$ such that $\phi_j'(t_{k+1}) \geq 0$ for $i \in \{1, \ldots, l_k\}$, $\phi_j'(t_{k+1}) = 0$ for $j \in \{l_k + 1, \ldots, l\}$, $\phi_{l+1}'(t_{k+1}) < 0$, and $\phi_j'(t_{k+1}) \leq 0$ for $j \in \{l + 2, \ldots, n\}$.*

CASE 5. *$\phi_i'(t_{k+1}) \geq 0$ for $i \in \{1, \ldots, l_k - 2\}$, $\phi_{l_k-1}'(t_{k+1}) > 0, \phi_{l_k}'(t_{k+1}) = \phi_{l_k+1}'(t_{k+1}) = 0$, and $\phi_i'(t_{k+1}) \leq 0$ for $i \in \{l_k + 2, \ldots, n\}$.*

CASE 6. *$\phi_i'(t_{k+1}) \geq 0$, $i \in \{1, \ldots, l_k - 1\}$, $\phi_{l_k}'(t_{k+1}) = \phi_{l_k+1}'(t_{k+1}) = 0$, $\phi_{l_k+2}'(t_{k+1}) < 0$, $\phi_j'(t_{k+1}) \leq 0$ for $j \in \{l_k + 3, \ldots, n\}$.*

CASE 7. *$\phi_i'(t_{k+1}) > 0$ for $i \in \{1, \ldots, l_k - 1\}$, $\phi_{l_k}'(t_{k+1}) = 0$, $\phi_j'(t_{k+1}) < 0$ for $j \in \{l_k + 1, \ldots, n\}$.*

CASE 8. *$\phi_i'(t_{k+1}) > 0$ for $i \in \{1, \ldots, l_k\}$, $\phi_{l_k+1}'(t_{k+1}) = 0$, and $\phi_j'(t_{k+1}) < 0$ for $j \in \{l_k + 2, \ldots, n\}$.*

CASE 9. *$\phi_i'(t_{k+1}) = 0$ for $i \in \{1, \ldots, l_k\}$, and $\phi_j'(t_{k+1}) < 0$ for $j \in \{l_k + 1, \ldots, n\}$.*

CASE 10. *$\phi_i'(t_{k+1}) > 0$ for $i \in \{1, \ldots, l_k\}$, and $\phi_j'(t_{k+1}) = 0$ for $j \in \{l_k + 1, \ldots, n\}$.*

By the techniques which we used in the case of $k = 1$, Cases 1–6 can be eliminated. Hence only Cases 7–10 are possible.

If Case 7 is true, then we have that for $\varepsilon > 0$ small enough, either $\phi_{l_k}'(t) > 0$ in $(t_{k+1}, t_{k+1} + \varepsilon)$ or $\phi_{l_k}' < 0$ in $(t_{k+1}, t_{k+1} + \varepsilon)$. Then $l_{k+1} = l_k$ or $l_{k+1} = l_k - 1$.

If Case 8 is true, then for $\varepsilon > 0$ small enough, either $\phi_{l_k+1}'(t) > 0$ in $(t_{k+1}, t_{k+1} + \varepsilon)$ or $\phi_{l_k+1}' < 0$ in $(t_{k+1}, t_{k+1} + \varepsilon)$. Then $l_{k+1} = l_k + 1$ or $l_{k+1} = l_k$.

If Case 9 is true, then for $\varepsilon > 0$ small enough, we can prove that $\phi_i'(t) < 0$ in $(t_{k+1}, t_{k+1} + \varepsilon)$ for $i = 1, \ldots, n$. Then $l_{k+1} = 0$.

If Case 10 is true, then for $\varepsilon > 0$ small enough, we can prove that $\phi_i'(t) > 0, i = 1, \ldots, n$; then $l_{k+1} = n$.

And for Cases 7–10, we can show that $\phi_L > \phi_1(t_{k+1}) > \cdots > \phi_n(t_{k+1}) > \phi_\beta$ always holds. Hence the proof is completed for this part.

(ii).

CLAIM 3. *Both $\phi_1'(t)$ and $\phi_n'(t)$ can change sign at most once. And if $\phi_1'(t)$ changes sign once, $\phi_n'(t)$ never changes sign. If $\phi_n'(t)$ changes sign once, $\phi_1'(t)$ never changes sign. That is,*

(a) *if $\phi_i'(t) < 0$, $i = 1, \ldots, n$ for $t \in (t_k, t_k + \varepsilon)$, then $\phi_i'(t) < 0$, $i = 1, \ldots, n$ for $t \in (t_k, \hat{t})$;*

(b) *if $\phi_i'(t) > 0$, $i = 1, \ldots, n$ for $t \in (t_k, t_k + \varepsilon)$, then $\phi_i'(t) > 0$, $i = 1, \ldots, n$ for $t \in (t_k, \hat{t})$.*

Claim 3 can be shown by contradiction. We ignore the proof here.

Hence by Claim 3, without loss of generality, we assume $\phi_1'(t)$ never changes sign; then we always have that $\phi_1'(t) > 0$ for $0 < t < \hat{t}$. So $\phi_1(t)$ increases as t increases. Since $\phi_L > \phi_1(t) > \cdots > \phi_n(t) > \phi_\beta$, $\phi_1(t)$ is bounded above by ϕ_L such that

$\lim_{t\to\hat{t}}\phi_1(t)=\bar{\phi}_1$ for some $\bar{\phi}_1\in[\phi_\beta,\phi_L]$. Also we have $\lim_{t\to\hat{t}}\phi_1'(t)=0$ such that

$$0=\beta+\lim_{t\to\hat{t}}[f+g](\phi_2(t))-2g(\bar{\phi}_1).$$

Then $\lim_{t\to\hat{t}}\phi_2(t)=\lim_{t\to\hat{t}}[f+g]^{-1}[f+g](\phi_2(t))$ exists. Let $\bar{\phi}_2=\lim_{t\to\hat{t}}\phi_2(t)$. By the boundedness, $\phi_L\geq\bar{\phi}_2\geq\phi_\beta$.

By induction we can show $\lim_{t\to\hat{t}}\phi_i(t)=\bar{\phi}_i$, where $\bar{\phi}_i\in[\phi_\beta,\phi_L]$. Since $\phi_L>\phi_1(t)>\cdots>\phi_n(t)>\phi_\beta$ for $t>0$, then $\phi_L\geq\bar{\phi}_1\geq\cdots\geq\bar{\phi}_n\geq\phi_\beta$.

If we apply the argument in the proof of Theorem 2.3, we also can show $\phi_L>\bar{\phi}_1>\cdots>\bar{\phi}_n>\phi_\beta$.

Remark. If we recall the proof of Claims 1 and 2, we need to assume $\phi_{n-1}\geq\phi_R$ along the trajectory. If this condition breaks down somewhere, the monotonicity may be destroyed.

REFERENCES

[1] M. Stemmler, M. Usher, and E. Niebur, *Lateral interactions in primary visual cortex - a model bridging physiology and psychophysics*, Science, 269 (1995), pp. 1877–1880.

[2] S. Grillner, *Neurobiological bases of rhythmic motor acts in vertebrates*, Science, 228 (1985), pp. 143–149.

[3] W. O. Friesen, M. Poon, and G. Stent, *Neuronal control of swimming in the medicinal leech* IV. *Identification of a network of oscillatory interneurons*, J. Exp. Biol., 75 (1978), pp. 25–43.

[4] G. B. Ermentrout and N. Kopell, *Frequency plateaus in a chain of weakly coupled oscillators*, SIAM J. Math. Anal., 15 (1984), pp. 215–237.

[5] G. B. Ermentrout and N. Kopell, *Symmetry and phase-locking in chains of weakly coupled oscillators*, Comm. Pure Appl. Math., 49 (1986), pp. 623–660.

[6] G. B. Ermentrout and N. Kopell, *Oscillator death in systems of coupled neural oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 125–146.

[7] G. B. Ermentrout and N. Kopell, *Phase transitions and other phenomena in chains of coupled oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 1014–1052.

[8] N. Kopell, W. Zhang, and G. B. Ermentrout, *Multiple coupling in chains of oscillators*, SIAM J. Math. Anal., 21 (1990), pp. 935–953.

[9] J. E. Paullet and G. B. Ermentrout, *Stable rotating waves in two-dimensional discrete active media*, SIAM J. Appl. Math., 54 (1994), pp. 1720–1744.

[10] G. B. Ermentrout, *Stable periodic solutions to discrete and continuum arrays of weakly coupled nonlinear oscillators*, SIAM J. Appl. Math., 52 (1992), pp. 1665–1687.

[11] N. Kopell, *Toward a theory of modelling central pattern generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. H. Cohen, S. Rossignol, and S. Grillner, eds., John Wiley, NY, 1988.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 235–250, January 1998

013

REFINABLE FUNCTION VECTORS*

ZUOWEI SHEN†

Abstract. Refinable function vectors are usually given in the form of an infinite product of their refinement (matrix) masks in the frequency domain and approximated by a cascade algorithm in both time and frequency domains. We provide necessary and sufficient conditions for the convergence of the cascade algorithm. We also give necessary and sufficient conditions for the stability and orthonormality of refinable function vectors in terms of their refinement matrix masks. Regularity of function vectors gives smoothness orders in the time domain and decay rates at infinity in the frequency domain. Regularity criteria are established in terms of the vanishing moment order of the matrix mask.

Key words. refinable function vectors, stable basis, cascade algorithm, regularity

AMS subject classifications. Primary, 41A15; 42A05; 42A15; 41A30; Secondary, 39B62; 42A38

PII. S0036141096302688

1. Introduction.

This paper presents a complete characterization of the convergence of the cascade algorithm and the stability and orthonormality of compactly supported refinable function vectors in terms of their refinement matrix masks. Regularity criteria for refinable function vectors are also established in terms of the vanishing moment order of the matrix mask.

We start with a finite set of compactly supported functions $\Phi \subset L_2(\mathbb{R}^s)$. The *FSI space* (finitely generated shift invariant; see [2]) $S(\Phi)$ generated by Φ is the smallest (closed) shift invariant subspace of $L_2(\mathbb{R}^s)$ containing Φ. Here we recall that a space is *shift invariant* if it is invariant under all *shifts*, i.e., invariant under all integer translations.

It is very convenient to discuss the shift invariant space in the frequency domain by using Gramian analysis. For a given set of functions Φ, the *pre-Gramian* matrix at $\omega \in \mathbb{T}^s$ is defined as a $\mathbb{Z}^s \times \Phi$ matrix by

$$J(\omega) := J_\Phi(\omega) := (\widehat{\varphi}(\omega + 2\pi\alpha))_{\alpha,\varphi},$$

where $\widehat{\varphi}$ is the Fourier transform of the function φ. Its adjoint matrix

$$J^*(\omega) := J_\Phi^*(\omega) := (\overline{\widehat{\varphi}(\omega + 2\pi\alpha)})_{\varphi,\alpha}$$

is a $\Phi \times \mathbb{Z}^s$ matrix. The *Gramian matrix* of functions Φ is a $\Phi \times \Phi$ matrix defined as the product of J^* and J, i.e., $J_\Phi^*(\omega) J_\Phi(\omega)$. The pre-Gramian matrix was first introduced in [22]; the basic properties of the pre-Gramian and its roles in the Gramian analysis for shift invariant spaces (not necessarily an FSI space) can be found in [22]. In this paper, we will often use the matrix $\overline{J^*J} =: G_\Phi =: G$ instead of J^*J. Since the properties of the Gramian matrix J^*J in which we are interested do not change when the conjugation is taken, we also call G_Φ the Gramian matrix of Φ.

This paper uses functions that are defined on $\mathbb{T}^s$, the s-dimensional torus. These can be viewed as 2π-periodic functions, via the standard transformation $\mathbb{R}^s \ni \omega \mapsto$

*Received by the editors April 29, 1996; accepted for publication (in revised form) October 15, 1996.

http://www.siam.org/journals/sima/29-1/30268.html

†Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260 (matzuows@leonis.nus.sg).

$e^{i\omega} := (e^{i\omega_1}, ..., e^{i\omega_s}) \in \mathbb{T}^s$. Though we may refer to such functions as defined on $\mathbb{T}^s$, we always treat their arguments as *real*. Thus, "multiplying a function defined on $\mathbb{T}^s$ by a function defined on $\mathbb{R}^s$" simply means "multiplying a 2π-periodic function by ...." Following this slight abuse of language, we write "$\Omega \subset \mathbb{T}^s$" to mean "$\Omega \subset [-\pi, \pi]^s$."

The functions Φ used in this paper are solutions to functional equations of the type

$$\Phi = \sum_{\alpha \in \mathbb{Z}^s} P_\alpha \Phi(2 \cdot -\alpha), \tag{1.1}$$

where the "coefficients" P_α are $\Phi \times \Phi$ matrices and Φ is a $\#\Phi$-dimensional column refinable function vector. We assume throughout that the refinement matrix masks are supported in $[0, N]^s$. Here we use Φ to denote both the set of functions Φ and the column function vector Φ.

Define

$$\mathbf{P} := 2^{-s} \sum_{\alpha \in \mathbb{Z}^s} P_\alpha \exp(-i\alpha \cdot).$$

Then, $\mathbf{P}$ is a $\Phi \times \Phi$ matrix, so each entry is a trigonometric polynomial such that their Fourier coefficients are supported in $[0, N]^s$. The functional equations (1.1) can be written as

$$\widehat{\Phi} := \mathbf{P}(\cdot/2)\widehat{\Phi}(\cdot/2). \tag{1.2}$$

Equations of the type (1.2) are called *vector refinement equations*; the matrix $\mathbf{P}$ is called the *refinement (matrix) mask*, and Φ is a ($\mathbf{P}$-) *refinable function vector*.

Since each entry of $\mathbf{P}$ is a trigonometric polynomial, the function matrix $\mathbf{P}$ satisfies

$$\|\mathbf{P}(\cdot) - \mathbf{P}(0)\| \le \text{const}\|\cdot\|,$$

where for any $d \times d$ matrix M, $\|M\| := \max_{\|\mathbf{v}\|=1} \|M\mathbf{v}\|/\|\mathbf{v}\|$, with $\|\mathbf{v}\|$ the Euclidean norm of the column vector $\mathbf{v} \in \mathbb{R}^d$.

If $\lim_{n\to\infty} \mathbf{P}(0)^n$ exists and is nontrivial, then the infinite product

$$\mathbf{P}^\infty := \prod_{k=1}^{\infty} \mathbf{P}(2^{-k}\cdot)$$

converges uniformly on compact sets. Further, $\widehat{\Phi} = \mathbf{P}^\infty \mathbf{a}$ is a solution of (1.2), where $\mathbf{a}$ is a right eigenvector of $\mathbf{P}(0)$ (see [12], [11] for the univariate case and [15] for the multivariate one). The functions Φ are compactly supported distributions with $\text{supp}(\Phi) \subset [0, N]^s$. We further remark that the existence of a solution $\widehat{\Phi}$ of (1.2) only requires the convergence of $\prod_{j=1}^n \mathbf{P}(2^{-j}\cdot)\mathbf{a}$, where $\mathbf{a}$ is a right eigenvector of $\mathbf{P}(0)$ corresponding to the eigenvalue 1 (see [12] and [4]). It has been shown in [4] that $\prod_{j=1}^n \mathbf{P}(2^{-j}\cdot)\mathbf{a}$ converges if $\rho(\mathbf{P}(0)) < 2$ (see also [12]).

We say that a matrix M (or linear operator) satisfies the condition on eigenvalues, or *Condition E* for short, if the spectral radius $\rho(M) \le 1$, 1 is required to be the only eigenvalue on the unit circle and must be a simple eigenvalue. Condition E is a useful concept in the wavelet theory and applications (see [3], [23], [24], and [18]).

Assume that $\mathbf{P}(0)$ satisfies Condition E. Then, there is a nonsingular matrix U so that $U\mathbf{P}(0)U^{-1}$ has the form

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \Lambda \end{pmatrix}, \tag{1.3}$$

where

$$\Lambda := \begin{pmatrix} \lambda_2 & \mu_2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_3 & \mu_3 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & \lambda_{\#\Phi} \end{pmatrix}$$

with $|\lambda_i| < 1$, and $\mu_i = 1$, or 0, $i = 2, \dots, \#\Phi$. Define $\mathbf{P}_1 = U\mathbf{P}U^{-1}$; then $\Phi^1 := U\Phi$ satisfies the refinement equations

$$\widehat{\Phi}^1 = \mathbf{P}_1(\cdot/2)\widehat{\Phi}^1(\cdot/2), \tag{1.4}$$

where Φ is a solution of (1.2).

The stability, regularity, and convergence of the cascade algorithm discussed in the paper do not change, even if we consider the refinement equation (1.4) instead of (1.1). Furthermore, as we will see in section 3, the problem of checking the orthonormality of Φ can be reduced to that of checking the stability of Φ^1. Therefore, we can always assume that $\mathbf{P}(0)$ has the form given in (1.3), without losing anything.

In this case the vector $\mathbf{i}_1^T := (1, 0, \dots, 0)$ is a left eigenvector and $\mathbf{i}_1$ is a right eigenvector of $\mathbf{P}(0)$ corresponding to eigenvalue 1. We further require that $\mathbf{P}$ have vanishing moments of at least order 1, that is, equivalent to the fact that

$$\mathbf{i}_1^T\mathbf{P}(\pi\nu) = \delta_\nu \mathbf{i}_1^T, \quad \nu \in \mathbb{Z}^s/2\mathbb{Z}^s.$$

This implies that

$$\mathbf{i}_1^T\widehat{\Phi}(2\pi\alpha) = \delta_\alpha,\ \alpha \in \mathbb{Z}^s.$$

Altogether, we assume, throughout this paper, that the mask $\mathbf{P}$ satisfies the following conditions.

BASIC CONDITIONS 1.5. *We say that* $\mathbf{P}$ *satisfies the basic conditions if*

(i) $\mathbf{P}(0)$ *has the form of* (1.3)*, and*

(ii) $\mathbf{i}_1^T\mathbf{P}(\pi\nu) = \delta_\nu \mathbf{i}_1^T, \quad \nu \in \mathbb{Z}^s/2\mathbb{Z}^s$.

It has further been shown in [15] (see also [12]) that if the basic conditions 1.5 hold for $\mathbf{P}$, then

$$\mathbf{P}^\infty = \prod_{j=1}^{\infty} \mathbf{P}(\cdot/2^j) = \begin{pmatrix} \widehat{\Phi} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}.$$

In particular, if $\widehat{\Phi}(0) \neq \mathbf{0}$, the solution Φ is determined uniquely up to a constant factor. In fact, $\widehat{\Phi} = c\mathbf{P}^\infty\mathbf{b}$, where $\mathbf{b}$ is an arbitrary vector satisfying $\mathbf{i}_1^T\mathbf{b} = 1$.

The functions Φ can be approximated by the following *cascade algorithm*: starting with a function vector Φ_0 which satisfies

$$\sum_{\alpha\in\mathbb{Z}^s} \mathbf{i}_1\Phi_0(\cdot - \alpha) = 1,$$

the function vector Φ_n is defined inductively by

$$\Phi_n := \sum_{\alpha\in\mathbb{Z}^s} P_\alpha \Phi_{n-1}(2\cdot - \alpha). \tag{1.5}$$

The cascade algorithm can be iterated in the frequency domain by taking the Fourier transform of (1.5):

$$\widehat{\Phi}_n = \mathbf{P}(\cdot/2)\widehat{\Phi}_{n-1}(\cdot/2). \tag{1.6}$$

It is clear that the sequence $(\Phi_n)_n$ converges in the L_2-norm if and only if the sequence $(\widehat{\Phi}_n)_n$ does. (We say $(\Phi_n)_n$ converges to Φ in the L_2-norm if each component of $(\Phi_n)_n$ converges to the corresponding component of Φ in the L_2-norm.) A sufficient condition for the convergence of the cascade algorithm is given in [4], under the assumption that Φ and its shifts are linearly independent; $s = 1$ and $\widehat{\Phi}_0 = \chi_{[-\pi,\pi]}\mathbf{a}$, where $\mathbf{a}$ is a right eigenvector of $\mathbf{P}(0)$ corresponding to eigenvalue 1. If the sequence $(\Phi_n)_n$ defined by (1.5) converges, then $\Phi \subset L_2(\mathbb{R}^s)$.

Define

$$S^k := S^k(\Phi) := \{f(2^k\cdot) : f \in S(\Phi)\}. \tag{1.7}$$

Then,

$$S^k \subset S^{k+1}. \tag{1.8}$$

It is proven in [14] that if

$$\cup_{k\in\mathbb{Z}} \cup_{\varphi\in\Phi} \operatorname{supp} \widehat{\varphi}(2^k\cdot) = \mathbb{R}^s \tag{1.9}$$

holds up to a null set and (1.8) holds, then $\cup_{k\in\mathbb{Z}} S^k$ is dense in $L_2(\mathbb{R}^s)$. If the refinable function vector Φ is compactly supported, (1.9) is always true. It has further been shown in [14] that if $\Phi \subset L_2(\mathbb{R}^s)$, then $\cap_{k\in\mathbb{Z}} S^k = \{0\}$. Altogether, we have the following result.

RESULT 1.1. *Let Φ be the compactly supported $\mathbf{P}$-refinable function vector. If $\Phi \subset L_2$, then*

$$\overline{\cup_{k\in\mathbb{Z}} S^k} = L_2(\mathbb{R}^s) \quad \textit{and} \quad \cap_{k\in\mathbb{Z}}\, S^k = \{0\}. \tag{1.10}$$

We say that a set of functions Φ is *stable* if Φ and their shifts form a Riesz basis of $S(\Phi)$, and a set of functions Φ is *orthonormal* if Φ and their shifts form an orthonormal basis of $S(\Phi)$.

A set of functions $\Phi \subset L_2(\mathbb{R}^s)$ is stable if and only if

$$0 < C_1 \le \|\lambda\|_\infty \le \|\Lambda\|_\infty \le C_2 < \infty, \quad \text{a.e.} \quad \omega \in \mathbb{T}^s,$$

where $\lambda(\omega)$ and $\Lambda(\omega)$ are the smallest and largest eigenvalues of the Gramian matrix $G_\Phi(\omega)$. If the set of functions Φ is compactly supported, then Φ is stable if and only if $\det G(\omega) \neq 0$ for all $\omega \in \mathbb{T}^s$. The set of functions Φ is orthonormal if and only if $G(\omega) = I$, a.e. $\omega \in \mathbb{T}^s$, where I is the identity matrix. The proofs of these results can be found in many articles (see, e.g., [13], [2], [9], [7], [22], [5], and [15]).

Once the set of functions $\Phi \subset L_2(\mathbb{R}^s)$ is stable (or orthonormal), it would be advantageous to know the regularity of Φ in order to make better use of Φ. An estimation of the regularity of Φ ($s = 1$) in terms of $\mathbf{P}$ has been given in [4], under the assumption that the refinable function vector Φ and its shifts are linearly independent.

By the above discussion, if the refinable function vector $\Phi \subset L_2(\mathbb{R}^s)$ is stable, or orthonormal, the sequence of subspaces $(S^k)_k$, $k \in \mathbb{Z}$ of $L_2(\mathbb{R}^s)$ forms a multiresolution; recall that a sequence $(S^k)_k$ forms a *multiresolution* if the sequence $(S^k)_k$ satisfies (1.10) and is refinable ($S^k \subset S^{k+1}$, $k \in \mathbb{Z}$) and if the refinable function vector Φ is stable or orthonormal.

The multiresolution generated by several functions was first introduced by [10], [9] (see also [1] and [6]). Result 1.1 is due to [14]. The first set of examples of orthonormal refinable function vectors Φ were given in [8], [7], and [6]. Examples of stable refinable function vectors Φ were given in [9]. Compactly supported wavelets and prewavelets from these examples were constructed in [7], [8], [25], and [16] (see also [5]).

It is of particular interest to construct compactly supported wavelets and prewavelets from compactly supported refinable function vectors and the refinement matrix masks. An algorithmic method in the construction of compactly supported wavelets and prewavelets from an arbitrarily given $\mathbf{P}$ refinable function vector Φ was obtained in [16], where $s = 1$. The problem of wavelet constructions is much more challenging in higher dimensions even when $\#\Phi = 1$ (see [13] and [14]). However, in dimensions no greater than 3, a method for the case $\#\Phi = 1$ has been provided in [19] and [20], under a mild condition on refinement masks.

Since the solutions Φ of (1.1) are defined via their Fourier transform by the refinement matrix mask $\mathbf{P}$, and since in practice only the refinement matrix mask is available for checking, it is useful to transfer the characterization of the stability and orthonormality of Φ by the Gramian matrix of Φ to characterization in terms of the mask. Similarly, it is necessary to characterize the convergence of the cascade algorithm defined by (1.6) and set criteria for the regularity of refinable function vectors in terms of the refinable matrix mask $\mathbf{P}$.

For this, we introduce the transition operator defined on $\mathbb{H}$, the space of all $\Phi \times \Phi$ matrices whose entries are trigonometric polynomials such that their Fourier coefficients are supported in $[-N, N]^s$. Here, we recall that the refinement mask $(P_\alpha)_\alpha$ is supported in $[0, N]^s$. The *transition operator* $\mathbf{T}_\Phi := \mathbf{T}$ is defined by

$$\mathbf{T}H := \sum_{\nu \in \mathbb{Z}^s/2\mathbb{Z}^s} \mathbf{P}(\cdot/2 + \pi\nu)H(\cdot/2 + \pi\nu)\mathbf{P}^*(\cdot/2 + \pi\nu), \quad H \in \mathbb{H}.$$

Then, $\mathbf{T}$ is a linear operator on $\mathbb{H}$.

Denote by $\mathbb{H}_M$ the space of all $\Phi \times \Phi$ matrices whose entries are trigonometric polynomials such that their Fourier coefficients are supported in $[-M, M]^s$. Then, if $M \geq N$, the transition operator $\mathbf{T}$ can be defined as an operator on $\mathbb{H}_M$. Further, since if $M > N$, $\mathbf{T}$ is a Fourier coefficient support reduced operator on $\mathbb{H}_M$, any eigenmatrix of nonzero eigenvalues of $\mathbf{T}_{|\mathbb{H}_M}$ is in $\mathbb{H}$. Therefore, all results of this paper can be stated in terms of the transition operator $\mathbf{T}$ on $\mathbb{H}_M$ for $M > N$, although they are stated in terms of the transition operator $\mathbf{T}$ on $\mathbb{H}$.

If the functions Φ are the solutions of refinable equations (1.1), and if one writes $J(\omega)$ as a column block matrix by

$$J(\omega) = (\widehat{\varphi}(\omega + 2\pi\nu + 4\pi\alpha))_{(\nu,\alpha)\times\varphi \in (\mathbb{Z}^s/2\mathbb{Z}^s \times \mathbb{Z}^s)\times\Phi},$$

then

$$\overline{J^*}(\omega) = (\mathbf{P}(\omega/2 + \pi\nu)\overline{J^*}(\omega/2 + \pi\nu))_{\nu \in \mathbb{Z}^s/2\mathbb{Z}^s}.$$

Hence

$$G(\omega) = \overline{J^*}(\omega)\overline{J}(\omega) = \sum_{\nu \in \mathbb{Z}^s/2\mathbb{Z}^s} \mathbf{P}(\omega/2 + \pi\nu)G(\omega/2 + \pi\nu)\mathbf{P}^*(\omega/2 + \pi\nu). \tag{1.11}$$

Therefore, the Gramian matrix $G_\Phi \in \mathbb{H}$ is an eigenmatrix of eigenvalue 1 of the transition operator $\mathbf{T}$.

Equation (1.11) also leads to the following result, which was proven by [10], [9], [8], [5], and [15].

RESULT 1.2. *If the compactly supported refinable function vector Φ is orthonormal, then*

$$I = \sum_{\nu\in\mathbb{Z}^s/2\mathbb{Z}^s} \mathbf{P}(\omega/2+\pi\nu)\mathbf{P}^*(\omega/2+\pi\nu), \quad \omega\in\mathbb{T}^s. \tag{1.12}$$

If the refinable function vector Φ is stable, then the matrix

$$\sum_{\nu\in\mathbb{Z}^s/2\mathbb{Z}^s} \mathbf{P}(\omega/2+\pi\nu)\mathbf{P}^*(\omega/2+\pi\nu)$$

is not singular for all $\omega\in\mathbb{T}^s$.

A mask $\mathbf{P}$ satisfying (1.12) is called a *conjugate quadrature filter*, or CQF.

Since $\mathbb{H}$ is a finite dimensional space, the operator $\mathbf{T}$ can be represented by a finite order matrix with respect to some fixed basis of $\mathbb{H}$. The matrix is also denoted by $\mathbf{T}$, and we will identify the operator $\mathbf{T}$ with the matrix $\mathbf{T}$.

We say that the cascade algorithm defined in (1.5) *converges*, if Φ_n defined by (1.5) converges to Φ with $\widehat{\Phi}(0)=\mathbf{i}_1$ for all Φ_0 which satisfy

$$\sum_{\alpha\in\mathbb{Z}^s} \mathbf{i}_1^T\Phi_0(\cdot-\alpha)=1, \quad \text{and} \quad G_{\Phi_0}\in\mathbb{H}. \tag{1.13}$$

We note that if $\mathbf{P}$ satisfies the basic conditions (1.5) and Φ_0 satisfies (1.13), then the Φ_n defined by the cascade algorithm (1.5) always converges to Φ with $\widehat{\Phi}(0)=\mathbf{i}_1$ in the distribution sense.

The rest of the paper is organized as follows. In section 2, we will prove that the cascade algorithm converges if and only if the transition operator $\mathbf{T}$ satisfies Condition E. In section 3, we will prove that Φ is stable if and only if the transition operator $\mathbf{T}$ satisfies Condition E and the corresponding eigenmatrix is nonsingular on $\mathbb{T}^s$. Consequently, we show that if Φ is stable, then the cascade algorithm converges; if $\mathbf{P}$ is a CQF mask, then Φ is orthonormal if and only if it is stable. Regularity criteria in terms of mask are established in section 4. We also remark that most of the results in this paper can be generalized to a general dilation matrix easily.

Finally, we remark that the corresponding results for the case $\#\Phi=1$ were obtained in [18] (convergence of the cascade algorithm), [17] (stability and orthonormality), and [21] (regularity).

2. Convergence of cascade algorithms. In this section we present a complete characterization of the convergence of the cascade algorithm defined by (1.5).

In what follows, we will identify the matrix $H\in\mathbb{H}$ with the corresponding unique sequence $(h_B^H)_{B\in\mathbb{B}}$ for a fixed basis $\mathbb{B}$, where

$$H=\sum_{B\in\mathbb{B}} h_B^H B.$$

We use the standard basis

$$\begin{aligned}\mathbb{B}_{st} :=&\{B_{i,j}^\alpha=(b_{l,l'}^\alpha)_{1\le l,l'\le\#\Phi}\in\mathbb{H}:\\ &b_{i,j}^\alpha=\exp(-i\alpha\cdot); \quad b_{l,l'}^\alpha=0, \quad (l,l')\ne(i,j); \quad \alpha\in[-N,N]^s\}\end{aligned}$$

in the proof of the sufficiency part of the next theorem.

We also note that a sequence of matrices $(\mathbf{T}^n)$ generated by a finite order matrix $\mathbf{T}$ converges to a nontrivial matrix if and only if the spectral radius $\rho(\mathbf{T}) \le 1$, and 1 is the only eigenvalue on the unit circle and is nondegenerate. Furthermore the sequence $(\mathbf{T}^n)$ converges if and only if for all $H \in \mathbb{H}$, the sequence $(\mathbf{T}^n H)$ converges. Here the convergence of the sequence $(\mathbf{T}^n H)$ is equivalent to the convergence of the sequence $(h_B^{\mathbf{T}^n H})_n$ for a fixed basis $\mathbb{B}$. Since $\mathbf{T}(\lim_n \mathbf{T}^{n-1} H) = \lim_n \mathbf{T}^n H$, the matrix $\lim_n \mathbf{T}^n H$ is an eigenmatrix of $\mathbf{T}$ corresponding to the eigenvalue 1. In particular, if $\mathbf{T}$ satisfies Condition E, then, for arbitrary $H \in \mathbb{H}$, $\lim_n \mathbf{T}^n H = \text{const} G_\Phi$.

A special basis of $\mathbb{H}$ is needed in the proof of the necessity part of the next theorem. The basis $\mathbb{B}_{sp}$ chosen is the one such that for each $B \in \mathbb{B}_{sp}$, there are Φ_0 and Ψ_0 satisfying (1.13) and $B = \overline{J^*_{\Phi_0} J_{\Psi_0}}$.

Define $\mathbb{D} = \mathbb{D}_1 \cup \mathbb{D}_2$, where

$$\begin{aligned}\mathbb{D}_1 := \{&D^\alpha_{1,1} = (d^\alpha_{l,l'})_{1\le l,l'\le \#\Phi} \in \mathbb{H}: \\ &d^\alpha_{1,1} = \exp(-i\alpha\cdot), \quad d^\alpha_{l,l'} = 0, \quad (l,l') \ne (1,1); \quad \alpha \in [-N,N]^s\};\end{aligned}$$

and

$$\begin{aligned}\mathbb{D}_2 :=\{&D^\alpha_{i,i} = (d^\alpha_{l,l'})_{1\le l,l'\le \#\Phi} \in \mathbb{H}: \quad d^\alpha_{1,1} = 1, \quad d^\alpha_{i,i} = \exp(-i\alpha\cdot), \quad 1 < i, \\ &d^\alpha_{1,i} = \exp(-i\alpha\cdot), \quad d_{i,1} = 1, \\ &d^\alpha_{l,l'} = 0, \quad \text{if } (l,l') \ne (i,i), (1,1),\ (i,1), \text{ and } (1,i);\ \alpha \in [-N,N]^s\}.\end{aligned}$$

Then define $\mathbb{E} = \mathbb{E}_1 \cup \mathbb{E}_2 \cup \mathbb{E}_3$, where

$$\begin{aligned}\mathbb{E}_1 :=\{&E^\alpha_{i,j} = (e^\alpha_{l,l'})_{1\le l,l'\le \#\Phi} \in \mathbb{H}: \\ &e^\alpha_{1,1} = 1, \quad e^\alpha_{i,j} = \exp(-i\alpha\cdot),\ 1 < i,j\ i \ne j, \\ &e^\alpha_{i,1} = 1, \quad e^\alpha_{1,j} = \exp(-i\alpha\cdot), \\ &e^\alpha_{l,l'} = 0, \text{ if } (l,l') \ne (1,1), (i,1), (1,j), \text{and } (i,j); \quad \alpha \in [-N,N]^s\};\end{aligned}$$

$$\begin{aligned}\mathbb{E}_2 :=\{&E^\alpha_{1,j} = (e^\alpha_{l,l'})_{1\le l,l'\le \#\Phi} \in \mathbb{H}: \quad e^\alpha_{1,1} = 1, \quad e^\alpha_{1,j} = \exp(-i\alpha\cdot), 1 < j, \\ &e^\alpha_{l,l'} = 0, \text{ otherwise}; \quad \alpha \in [-N,N]^s\};\end{aligned}$$

and

$$\mathbb{E}_3 := \{B^T : B \in \mathbb{E}_2\}.$$

Then, the set $\mathbb{B}_{sp} := \mathbb{D} \cup \mathbb{E}$ is a basis of $\mathbb{H}$.

For function vectors Φ_0 and Ψ_0 satisfying (1.13), define function vectors

$$\Phi_n = (\varphi^i_n)^T_{1\le i\le \#\Phi} \quad \text{and} \quad \Psi_n = (\psi^i_n)_{1\le i\le \#\Phi}$$

via their Fourier transform as

$$\widehat{\Phi}_n := \mathbf{P}(\cdot/2)\widehat{\Phi}_{n-1}(\cdot/2), \quad \text{and} \quad \widehat{\Psi}_n := \mathbf{P}(\cdot/2)\widehat{\Psi}_{n-1}(\cdot/2).$$

Let $G_n = \overline{J^*_{\Phi_n} J_{\Psi_n}}$. Then

$$\mathbf{T}G_{n-1} = \sum_{\nu\in\mathbb{Z}^s/2\mathbb{Z}^s} \mathbf{P}(\cdot/2+\pi\nu)\overline{J^*_{\Phi_{n-1}}}(\cdot/2+\pi\nu)\overline{J_{\Psi_{n-1}}}(\cdot/2+\pi\nu)\mathbf{P}^*(\cdot/2+\pi\nu) = G_n.$$

If the cascade algorithm converges in the L_2-norm, then each entry of G_n converges to the corresponding entry of G_Φ in the L_1-norm. This implies G_n converges to G_Φ in the $\|\cdot\|_1$-norm on $\mathbb{H}$, where for $H(\omega) := (h_{i,j}(\omega))_{1\le i,j\le \#\Phi}$,

$$\|H\|_1 := \sum_{1\le i,j\le \#\Phi} \|h_{i,j}(\cdot)\|_1.$$

For a fixed $\mathbb{B}$, let

$$G_n = \sum_{B\in\mathbb{B}} a_B^n B \quad \text{and} \quad G_\Phi = \sum_{B\in\mathbb{B}} a_B B.$$

Then, the convergence of the cascade algorithm implies that the sequence of the sequences $(a^n)_n$ converges to the sequence a.

Finally, we note that for any $\Phi := (\varphi^l)^T$ and $\Psi := (\psi^{l'})^T$, the $(\varphi^l, \psi^{l'})$th entry of $\overline{J_\Phi^* J_\Psi}$ can be written as

$$\sum_{\beta\in\mathbb{Z}^s} \widehat{\varphi^l}(\cdot + 2\pi\beta)\overline{\widehat{\psi^{l'}}(\cdot + 2\pi\beta)} = \sum_{\alpha\in\mathbb{Z}^s} (\varphi^l * \overline{\psi^{l'}(-\cdot)})(\alpha)\exp(-i\alpha\cdot). \tag{2.1}$$

We are ready to prove the following theorem.

THEOREM 2.1. *If the basic conditions* 1.5 *hold for* $\mathbf{P}$*, then the cascade algorithm converges if and only if the transition operator* $\mathbf{T}$ *satisfies Condition* E.

Proof. "$\Longrightarrow$" Since $\mathbf{T}$ satisfies Condition E, the sequence of linear operators $\mathbf{T}^n$ converges. Let Φ_n be a sequence of function vectors generated by the cascade algorithm (1.5) with Φ_0 satisfying (1.13). Then $G_{\Phi_n} = \mathbf{T}^n G_{\Phi_0}$ converges. Since for each fixed l, $\|\varphi_n^l\|$ is the coefficient of $B_{l,l}^0 \in \mathbb{B}_{st}$, if we express G_{Φ_n} by $\mathbb{B}_{st}$, and since $\mathbb{H}$ is a finite dimensional space, Φ_n is bounded in the L_2-norm. Hence any subsequence of Φ_n contains a weakly convergent subsubsequence of Φ_n. Since Φ_n converges to Φ in the distribution sense, and since weak convergence is stronger than convergence in the distribution sense, Φ_n converges to Φ weakly. Therefore, to show that Φ_n converges strongly to Φ, it remains to show that the L_2-norm of Φ_n converges to that of Φ. Since G_Φ is an eigenmatrix of $\mathbf{T}$ corresponding to the eigenvalue 1 and since $\mathbf{T}$ satisfies Condition E, G_Φ is the unique eigenmatrix of $\mathbf{T}$. Therefore,

$$\lim_{n\to\infty} \mathbf{T}^n G_{\Phi_0} = \lim_{n\to\infty} G_{\Phi_n} = \text{const} G_\Phi.$$

Since $\mathbf{i}_1^T G_{\Phi_0}(0)\mathbf{i}_1 = \mathbf{i}_1^T G_{\Phi_n}(0)\mathbf{i}_1 \neq 0$ for all n, we have

$$0 \neq \mathbf{i}_1^T G_{\Phi_0}(0)\mathbf{i}_1 = \lim_{n\to\infty} \mathbf{i}_1^T \mathbf{T}^n G_{\Phi_0}(0)\mathbf{i}_1$$

$$= \lim_{n\to\infty} \mathbf{i}_1^T G_{\Phi_n}(0)\mathbf{i}_1 = \mathbf{i}_1^T G_\Phi(0)\mathbf{i}_1 = \text{const}\mathbf{i}_1^T G_\Phi(0)\mathbf{i}_1.$$

Hence const $= 1$, and

$$\lim_{n\to\infty} \mathbf{T}^n G_{\Phi_0} = \lim_{n\to\infty} G_{\Phi_n} = G_\Phi.$$

Since for each fixed l, $\|\varphi_n^l\|$ is the coefficient of $B_{l,l}^0 \in \mathbb{B}_{st}$ and since $\mathbb{H}$ is a finite dimensional space, we must have $\lim_n \|\varphi_n^l\| = \|\varphi^l\|$. Hence, Φ_n converges to Φ strongly in the L_2-norm.

"$\Longleftarrow$" We first prove that for any element $B \in \mathbb{B}_{sp}$, there is a proper choice of $\Phi_0 := (\varphi^{l'})^T$ and $\Psi_0 := (\psi^l)^T$ satisfying (1.13) so that $\overline{J^*_{\Phi_0} J_{\Psi_0}} = B$.

For this, let $f := \chi_{[-1/2,1/2]^s}$.

First, if $B = D^\alpha_{1,1} \in \mathbb{D}_1$, then define $\Phi_0 = (\varphi^l_0)^T$ and $\Psi_0 = (\psi^{l'}_0)^T$ so that $\varphi^1_0 = f$, $\psi^1_0 = f(\cdot - \alpha)$, $\varphi^l_0 = 0$, and $\psi^{l'}_0 = 0$ if $l, l' \neq 1$. Then Φ_0 and Ψ_0 satisfy (1.13) and $\overline{J^*_{\Phi_0} J_{\Psi_0}} = D^\alpha_{1,1} \in \mathbb{D}_1$ by (2.1). For $E^\alpha_{i,j} \in \mathbb{E}_1$ (or $D^\alpha_{i,i} \in \mathbb{D}_2$), define $\Phi_0 = (\varphi^l_0)^T$ and $\Psi_0 = (\psi^{l'}_0)^T$ to be the function vectors such that $\varphi^1_0 = \psi^1_0 = f$ and $\varphi^i_0 = f$ and $\psi^j_0 = f(\cdot - \alpha)$, $\varphi^l_0 = 0$, $l \neq 1, i$, and $\psi^{l'}_0 = 0$ $l' \neq 1, j$. Then the function vectors Φ_0 and Ψ_0 satisfy (1.13). Further, the matrix $\overline{J^*_{\Phi_0} J_{\Psi_0}} = E^\alpha_{i,j} \in \mathbb{E}_1$ if $i \neq j$ (and $D_{i,i} \in \mathbb{D}_2$ if $i = j$) by (2.1). For $E^\alpha_{i,j} \in \mathbb{E}_2$, define $\Phi_0 = (\varphi^l_0)^T$ and $\Psi_0 = (\psi^{l'}_0)^T$ so that $\varphi^1_0 = \psi^1_0 = f$, and $\varphi^j_0 = 0$ and $\psi^j_0 = f(\cdot - \alpha)$ and $\varphi^l_0 = \psi^{l'}_0 = 0$ if $l, l' \neq 1, j$. Then Φ_0 and Ψ_0 satisfy (1.13) and $\overline{J^*_{\Phi_0} J_{\Psi_0}} = E^\alpha_{i,j} \in \mathbb{E}_2$ by (2.1).

Since the cascade algorithm converges, $\mathbf{T}^n B$ converges to G_Φ for all $B \in \mathbb{B}_{sp}$. Thus for any $H \in \mathbb{H}$, $\lim_{n\to\infty} \mathbf{T}^n H = \text{const} G_\Phi$; consequently, the sequence of matrices $(\mathbf{T}^n)$ converges. Therefore, the spectral radius $\rho(\mathbf{T}) \leq 1$ and 1 is the only eigenvalue of $\mathbf{T}$ on the unit circle. Further, 1 is a nondegenerate eigenvalue of $\mathbf{T}$. To prove that $\mathbf{T}$ satisfies Condition E, it remains to show that G_Φ is the only eigenmatrix of eigenvalue 1. Let $E \in \mathbb{H}$ so that $\mathbf{T}E = E$; then

$$\lim_{n\to\infty} \mathbf{T}^n E = E = \text{const} G_\Phi.$$

Hence, G_Φ is the only eigenmatrix (up to a constant multiple) of $\mathbf{T}$ corresponding to eigenvalue 1. □

3. Stability, orthonormality, and biorthonormality. In this section we will discuss the stability of refinable function vectors. We first give here a sufficient condition in terms of the eigenvalue of the transition operator $\mathbf{T}$ under which the function vector Φ is stable. Then, we will show that this condition is also necessary.

PROPOSITION 3.1. *Suppose that the basic conditions* (1.5) *hold for* $\mathbf{P}$ *and* $\Phi \subset L_2(\mathbb{R}^s)$. *If* 1 *is a simple eigenvalue of the transition operator* $\mathbf{T}$ *on* $\mathbb{H}$ *and the corresponding eigenmatrix nonsingular on* $\mathbb{T}^s$, *then the* $\mathbf{P}$*-refinable function vector* Φ *is stable.*

Proof. Since G_Φ is an eigenmatrix of $\mathbf{T}$ of the eigenvalue 1, the hypothesis of the theorem implies that the matrix $G_\Phi(\omega)$ is nonsingular on $\mathbb{T}^s$. Hence, the function vector Φ is stable. □

We note that if Φ is stable, the simplicity of the eigenvalue 1 of $\mathbf{T}$ implies that the corresponding eigenmatrix of the eigenvalue 1 of $\mathbf{T}$ is nonsingular on $\mathbb{T}^s$, since in this case G_Φ is the only eigenmatrix of $\mathbf{T}$. Therefore, to show that the condition in the above theorem is necessary, one only requires to show that if Φ is stable, then 1 is a simple eigenvalue of $\mathbf{T}$.

Define

$$V_1 := \{H \in \mathbb{H} : (\mathbf{i}_1^T H \mathbf{i}_1)(0) = 0\}.$$

Since for any $H \in \mathbb{H}$, $H = \sum_{B\in\mathbb{B}_{sp}} h_B B$, where the set $\mathbb{B}_{sp}$ is the basis defined in the previous section, a matrix $H \in V_1$ if and only if $\sum_{B\in\mathbb{B}_{sp}} h_B = 0$ by the structure of the element of $\mathbb{B}_{sp}$. Hence, the space V_1 has codimension 1. Since $(\mathbf{i}_1^T G_\Phi \mathbf{i}_1)(0) \neq 0$, $G_\Phi \notin V_1$. Since $\mathbf{P}$ satisfies the basic condition (1.5), for any $H \in V_1$,

$$(\mathbf{i}_1^T (\mathbf{T}H) \mathbf{i}_1)(0) = \sum_{\nu\in\mathbb{Z}^s/2\mathbb{Z}^s} \mathbf{i}_1^T \mathbf{P}(\pi\nu) H(\pi\nu) \mathbf{P}^*(\pi\nu) \mathbf{i}_1 = 0.$$

Hence, V_1 is a $\mathbf{T}$-invariant subspace of $\mathbb{H}$.

The proofs of the following two propositions (Propositions 3.2 and 3.3) were originally in our earlier drafts. Before completing the paper, we received a preprint of [15], which contains the same results (Proposition 3.3 and Theorems 5.1 and 5.2 in there) with similar proofs. Thus, we will only provide an outline of the proofs here.

PROPOSITION 3.2. *Let $H_1(\omega)$ and $H_2(\omega)$ be matrices so that each entry is a continuous function on $\mathbb{T}^s$. Then,*

$$\int_{\mathbb{T}^s} H_1(\omega)(\mathbf{T}^n H_2)(\omega)d\omega = \int_{\mathbb{R}^s} H_1(\omega)\Pi_n(\omega)H_2(2^{-n}w)\Pi_n^*(\omega)d\omega, \tag{3.1}$$

where

$$\Pi_n(\omega) := \chi_{2^n\mathbb{T}^s}(\omega)\prod_{j=1}^{n}\mathbf{P}(\omega/2^j);\ n = 1, 2, \ldots. \tag{3.2}$$

Here we define the transition operator as an operator on the space of the all $\Phi \times \Phi$ matrices whose entries are continuous functions on $\mathbb{T}^s$.

Proof. One can easily show that for any such H

$$\mathbf{T}^n H = \sum_{\alpha\in\mathbb{Z}^s} \Pi_n(\cdot + 2\pi\alpha)H(2^{-n}(\cdot + 2\pi\alpha))\Pi_n^*(\cdot + 2\pi\alpha),$$

by induction. Replacing H by H_2, multiplying by H_1, and integrating both sides of the above identity lead to the fact that for any H_1 and H_2, (3.1) holds. □

PROPOSITION 3.3. *Assume that the $\mathbf{P}$-refinable function vector Φ is stable and its mask $\mathbf{P}$ satisfies basic conditions* (1.5). *Then,*

(i) *for any $H_1 \in \mathbb{H}$ and $H_2 \in V_1$,*

$$\lim_{n\to\infty}\int_{\mathbb{R}^s} H_1(\omega)\Pi_n(\omega)H_2(2^{-n}\omega)\Pi_n^*(\omega)d\omega$$

$$= \int_{\mathbb{R}^s}\lim_{n\to\infty} H_1(\omega)\Pi_n(\omega)H_2(2^{-n}\omega)\Pi_n^*(\omega)d\omega = \mathbf{0};$$

(ii) *the transition operator $\mathbf{T}$ restricted to V_1 has spectral radius < 1.*

Proof. Since Φ is stable, $G_\Phi \geq cI$, with $c > 0$. This leads to the fact that the sequence $(\Pi_n\Pi_n^*)$ is uniformly integrable (details in the proof of Theorem 5.2 of [15]). Recall that the sequence

$$(\Pi_n(\omega)\Pi_n^*(\omega)), \quad n = 0, 1, \ldots,$$

is uniformly integrable if for an arbitrary $\varepsilon > 0$ there exist a finite measure set F and $\delta > 0$ so that

$$\int_{\mathbb{R}^s\setminus F} \Pi_n(\omega)\Pi_n^*(\omega)d\omega \leq \varepsilon$$

and

$$\int_D \Pi_n(\omega)\Pi_n^*(\omega)d\omega \leq \varepsilon$$

hold for all n for any measurable set D with the measure of $D \leq \delta$.

That the sequence $(\Pi_n(\omega)\Pi_n^*(\omega))_n$ is uniformly integrable implies that the sequence $(H_1(\omega)\Pi_n(\omega)H_2(2^{-n}\omega)\Pi_n^*(\omega))$ is uniformly integrable for any $H_1 \in \mathbb{H}$ and $H_2 \in V_1$. This implies that

$$\lim_{n\to\infty}\int_{\mathbb{R}^s} H_1(\omega)\Pi_n(\omega)H_2(2^{-n}\omega)\Pi_n^*(\omega)d\omega = \int_{\mathbb{R}^s}\lim_{n\to\infty} H_1(\omega)\Pi_n(\omega)H_2(2^{-n}\omega)\Pi_n^*(\omega)d\omega.$$

Since $(\mathbf{i}_1^T H_2\mathbf{i}_1)(0) = 0$ and since $\mathbf{P}^\infty = \begin{pmatrix}\widehat{\Phi} & \mathbf{0} & \cdots & \mathbf{0}\end{pmatrix}$,

$$\lim_{n\to\infty} H_1(\omega)\Pi_n(\omega)H_2(2^{-n}\omega)\Pi_n^*(\omega) = H_1(\omega)\mathbf{P}^\infty(\omega)H_2(0)\mathbf{P}^{*\infty}(\omega) = \mathbf{0}.$$

Hence, the first statement holds.

For the second statement, assume that λ is an eigenvalue of $\mathbf{T}$ restricted to V_1 and $H \in V_1$ is the corresponding nontrivial eigenmatrix. Then

$$\lambda^n\int_{\mathbb{T}^s} H^*(\omega)H(\omega)d\omega = \int_{\mathbb{R}^s} H^*(\omega)\Pi_n(\omega)H(2^{-n}\omega)\Pi_n^*(\omega)d\omega.$$

Hence, (i) implies that

$$\lim_{n\to\infty}\int_{\mathbb{R}^s} H^*(\omega)\Pi_n(\omega)H(2^{-n}\omega)\Pi_n^*(\omega)d\omega = \int_{\mathbb{R}^s}\lim_{n\to\infty} H^*(\omega)\Pi_n(\omega)H(2^{-n}\omega)\Pi_n^*(\omega)d\omega = \mathbf{0}.$$

This gives

$$\lim_{n\to\infty}\lambda^n\int_{\mathbb{T}^s} H^*(\omega)H(\omega)d\omega = 0.$$

Therefore $|\lambda| < 1$ by the fact

$$\int_{\mathbb{T}^s} H^*(\omega)H(\omega)d\omega \neq 0. \qquad \square$$

From the two propositions above, we obtain the following result.

LEMMA 3.4. *Assume that* $\mathbf{P}$ *satisfies basic conditions* (1.5) *and the corresponding refinable function vector* Φ *is stable; then the transition operator* $\mathbf{T}$ *satisfies Condition* E. *In particular,* 1 *is a simple eigenvalue of the transition operator* $\mathbf{T}$.

Proof. Let $\mathbb{B}_0$ be a basis of V_1. Since G_Φ is not in V_1 and V_1 has codimension 1, $G_\Phi \cup \mathbb{B}_0$ is a basis of $\mathbb{H}$. Therefore, an arbitrary $H \in \mathbb{H}$ can be written uniquely as

$$H = aG_\Phi + H_0, \quad H_0 \in V_1.$$

Let H be the eigenmatrix of the eigenvalue λ of $\mathbf{T}$,

$$\lambda aG_\Phi + \lambda H_0 = \mathbf{T}H = a\mathbf{T}G_\Phi + \mathbf{T}H_0 = aG_\Phi + \mathbf{T}H_0.$$

If $\lambda \neq 1$, then $a = 0$. Thus, $H = H_0 \in V_1$ is an eigenmatrix of λ. This implies $|\lambda| < 1$ by Proposition 3.3. If $\lambda = 1$, then $H_0 \in V_1$ is also the eigenmatrix of $\mathbf{T}$ corresponding to the eigenvalue 1; thus $H_0 = 0$ again by Proposition 3.3. Hence, $\rho(\mathbf{T}) \leq 1$ and 1 is the unique eigenvalue on the unit circle. Further, G_Φ is the only eigenmatrix of eigenvalue 1 up to a constant multiple.

Finally we need to show that 1 is a simple eigenvalue of $\mathbf{T}$. If not, it must be a degenerate eigenvalue with only one eigenmatrix. In this case, there exists a matrix $H \in \mathbb{H}$ such that $\mathbf{T}H = G_\Phi + H$. Let $H_1 = cG_\Phi + H$ so that $H_1 \in V_1$. Then

$$\int_{\mathbb{T}^s}(\mathbf{T}^n H_1)(\omega)d\omega = \int_{\mathbb{R}^s}\Pi_n(\omega)H_1(2^{-n}\omega)\Pi_n^*(\omega)d\omega = \int_{\mathbb{T}^s}((c+n)G_\Phi(\omega) + H(\omega))d\omega.$$

The left-hand side tends to 0 by the stability of the vector function Φ and (i) of Proposition 3.3, while the right-hand side tends to ∞, which is a contradiction. □

An immediate consequence of this lemma is the following corollary.

COROLLARY 3.5. *Assume that* $\mathbf{P}$ *satisfies the conditions* (1.5). *If the refinable vector function* Φ *is stable, then the cascade algorithm converges.*

The next theorem, the main result of this section, follows directly from Proposition 3.1 and Lemma 3.4.

THEOREM 3.6. *Assume that* $\mathbf{P}$ *satisfies the basic conditions* (1.5). *The* $\mathbf{P}$*-refinable function vector* Φ *is stable if and only if the corresponding transition operator* $\mathbf{T}$ *satisfies Condition* E *and the eigenmatrix of eigenvalue* 1 *is nonsingular on* $\mathbb{T}^s$.

Proof. If the transition operator satisfies Condition E, then the cascade algorithm converges and $\Phi \subset L_2(\mathbb{R}^s)$. Therefore, Φ is stable by Proposition 3.1. If Φ is stable, then Lemma 3.4 implies that the transition operator satisfies Condition E. G_Φ is the eigenmatrix of a simple eigenvalue 1 of $\mathbf{T}$ which is nonsingular on $\mathbb{T}^s$. □

REMARK 3.7. *If the transition operator* $\mathbf{T}$ *satisfies Condition* E *and if eigenvalue* 1 *has an eigenmatrix which is nonsingular on* $\mathbb{T}^s$, *then the compactly supported* $\mathbf{P}$*-refinable functions* $\Phi \subset L_2(\mathbb{R}^s)$ *by the fact that the corresponding cascade algorithm converges. Hence the sequence* $(S^k(\Phi))$ *forms a multiresolution of* $L_2(\mathbb{R}^s)$ *with the functions* Φ *and their shifts forming a Riesz basis of* $S(\Phi)$ *by Result* 1.1 *and Theorem* 3.6.

If $\mathbf{P}$ is CQF, the identity matrix I is an eigenmatrix of the transition operator $\mathbf{T}$ corresponding to eigenvalue 1. A consequence of Theorem 3.6 is as follows.

THEOREM 3.8. *Suppose that* $\mathbf{P}$ *is a CQF matrix mask which satisfies the basic conditions* (1.5)*; then the following statements are equivalent:*

(i) *the refinable function vector* Φ *is orthonormal,*

(ii) *the transition operator* $\mathbf{T}$ *satisfies Condition* E*,*

(iii) *the refinable function vector* Φ *is stable, and*

(iv) *the corresponding cascade algorithm converges.*

Remark 3.7 gives the following corollary.

COROLLARY 3.9. *Suppose that* $\mathbf{P}$ *is a CQF matrix mask which satisfies the basic conditions* (1.5). *If the corresponding transition operator* $\mathbf{T}$ *satisfies Condition* E*, then the sequence of spaces* $(S^k(\Phi))_k$ *forms a multiresolution of* $L_2(\mathbb{R}^s)$ *with the functions* Φ *and their shifts forming an orthonormal basis of* $S(\Phi)$.

In the rest of this section, we discuss the biorthonormality of two refinable function vectors Φ and Ψ. Let $\mathbf{P}_\Phi$ and $\mathbf{P}_\Psi$ be the refinement masks of functions Φ and Ψ satisfying the basic conditions (1.5) and the condition

$$\sum_{\nu \in \mathbb{Z}^s/2\mathbb{Z}^s} \mathbf{P}_\Phi(\cdot/2 + \pi\nu)\mathbf{P}_\Psi^*(\cdot/2 + \pi\nu) = I. \tag{3.3}$$

We say that Φ and Ψ are *biorthonormal* if both function vectors Φ and Ψ are stable and

$$\overline{J_\Phi^*}(\omega)\overline{J_\Psi}(\omega) = I, \quad \in \mathbb{T}^s.$$

Here again we are interested in characterizing the biorthonormality in terms of the matrix masks $\mathbf{P}_\Phi$ and $\mathbf{P}_\Psi$. The following result was shown in [15, Theorem 5.3], which is the main result of [15].

RESULT 3.10. *Let* $\mathbf{P}_\Phi$ *and* $\mathbf{P}_\Psi$ *be the refinement masks of refinable function vectors* Φ *and* Ψ *satisfying the basic conditions* (1.5) *and* (3.3). *Assume that* $G_\Phi(0) \geq \text{const}I$

and $G_\Psi(0) \geq \text{const} I$. *Then* Φ *and* Ψ *are biorthonormal if both* $\mathbf{T}_\Phi$ *and* $\mathbf{T}_\Psi$ *have the spectrum radius* < 1 *on* V_1.

We note that if the Φ and Ψ are stable, then by Proposition 3.3 (ii) the conditions in the above result are satisfied.

THEOREM 3.11. *Let* $\mathbf{P}_\Phi$ *and* $\mathbf{P}_\Psi$ *be the refinement matrix masks of* Φ *and* Ψ *which satisfy the basic conditions* (1.5) *and condition* (3.3). *Then the following statements are equivalent:*

(i) *the refinable function vectors* Φ *and* Ψ *are biorthonormal;*

(ii) *both* Φ *and* Ψ *are stable; and*

(iii) *the transition operators* $\mathbf{T}_\Phi$ *and* $\mathbf{T}_\Psi$ *satisfy Condition* E, *and the corresponding eigenmatrices of eigenvalue* 1 *of* $\mathbf{T}_\Phi$ *and* $\mathbf{T}_\Psi$ *are nonsingular on* $\mathbb{T}^s$.

Proof. The equivalence of (ii) and (iii) follows from Theorem 3.6. Since if Φ and Ψ are stable, the conditions in Result 3.10 are satisfied; (ii) implies (i) by that result. Finally, (i) implies (ii) by the definition of the biorthonormality of Φ and Ψ. □

4. Regularity of refinable function vectors. In this section, we establish some criteria for the regularity of refinable function vectors.

We say that the mask $\mathbf{P}$ satisfying the basic conditions (1.5) has *vanishing moment order* r if conditions

$$D^\beta(A^*(2\omega)\mathbf{P}(\omega))_{|\omega=\pi\nu} = i^{-|\beta|}(D^\beta A^*)(0)\delta_\nu, \quad \nu \in \mathbb{Z}^s/2\mathbb{Z}^s, \quad |\beta| \leq r-1, \tag{4.1}$$

hold for some

$$A = \sum_{|\beta|\leq r-1} \mathbf{a}_\beta^T \exp(-i\beta\cdot),$$

where $\mathbf{a}_\beta \in \mathbb{R}^{\#\Phi}$.

As we did in [21] for the case $\#\Phi = 1$, we will connect this vanishing moment order to the regularity of Φ.

We say that $\Phi := (\varphi^l)^T \in C^\gamma$ if each component $\varphi^l \in C^\gamma$. Recall that a function $\varphi \in C^\gamma$ for $n \leq \gamma < n+1$ provided that $\varphi \in C^n$ and

$$|D^\beta\varphi(x+t) - D^\beta\varphi(x)| \leq \text{const}|t|^{\gamma-n} \text{ for all } |\beta| = n \text{ and } |t| \leq 1$$

for some constant independent of x. This number is related to

$$\kappa_2 := \sup\left\{\kappa : \int_{\mathbb{R}^s} (1+|w|^2)^\kappa |\widehat{\varphi}(\omega)|^2 d\omega < \infty\right\}$$

by the inequality $\gamma \geq \kappa_2 - s/2$.

Define

$$V_r := \{H \in \mathbb{H} : (D^\beta(A^*(\omega)H(\omega))_{|\omega=0})^T = D^\beta(H(\omega)A(\omega))_{|\omega=0}, = \mathbf{0}, \quad |\beta| \leq r-1\}.$$

In the case that $r-1 > N$, we replace $\mathbb{H}$ by $\mathbb{H}_{r-1}$ in the above definition of V_r.

Since

$$D^\beta(A^*(2\omega)\mathbf{P}(\omega))_{|_{\omega=\pi\nu}} = i^{-|\beta|}(D^\beta A^*)(0)\delta_\nu, \quad \nu \in \mathbb{Z}^s/2\mathbb{Z}^s,\ |\beta| \le r-1, \tag{4.2}$$

we have that

$$(D^\beta(A^*(\omega)\mathbf{T}H(\omega))_{|_{\omega=0}})^T = D^\beta(\mathbf{T}H(\omega)A(\omega))_{|_{\omega=0}} = \mathbf{0}, \text{ for all } H \in V_r, |\beta| \le r-1.$$

Hence, the space V_r is an invariant subspace of the transition operator $\mathbf{T}$.

In the case $r = 1$, the space V_1 defined here is an invariant subspace of V_1 defined in section 3, since $A^*(0)$ is the left eigenvector of $\mathbf{P}$.

For each $H(\omega) := (h_{i,j}(\omega))_{1\le i,j\le \#\Phi} \in V_r$, define

$$\|H\|_F := \sum_{1\le i,j\le \#\Phi} \|h_{i,j}(\cdot)\|_\infty.$$

If H is a constant matrix, this norm is the sum of the modulus of all entries.

Then the operator norm $\|\mathbf{T}_{|_{V_r}}\| := \sup_{H\in V_r\setminus\{0\}} \frac{\|\mathbf{T}H\|_F}{\|H\|_F}$ on V_r satisfies

$$\lim_{n\to\infty} \|\mathbf{T}^n_{|_{V_r}}\|^{1/n} = \rho,$$

where ρ is the spectral radius of $\mathbf{T}_{|_{V_r}}$. Hence, there exists $N_{\mathbf{T}}$ such that for any $H \in V_r$ and for all $n > N_{\mathbf{T}}$,

$$\|\mathbf{T}^n H\|_F \le \|\mathbf{T}^n\|\|H\|_F \le \text{const}(\rho+\varepsilon)^n\|H\|_F, \tag{4.3}$$

where ε is arbitrarily small.

The proof of the following proposition is carried out by modifying the proofs of Proposition 3.6 of [21] and Theorem 5.2 of [15].

PROPOSITION 4.1. *Suppose* $\mathbf{P}$ *satisfies the basic conditions* 1.5 *and conditions* (4.1). *Then for the* $\mathbf{P}$*-refinable function* $\Phi := (\varphi^l)^T$, *there exists a constant* C *such that*

$$\int_{\mathbb{F}_n} \left|\widehat{\varphi}^l(\omega)\right|^2 d\omega \le C(\rho+\varepsilon)^{n+1},$$

where $\mathbb{F}_n := 2^n\mathbb{T}^s\backslash 2^{n-1}\mathbb{T}^s$ *for all* $n > N_{\mathbf{T}}$ *and* ε *is arbitrarily small.*

Proof. It follows from (4.3) for any $H \in V_r$,

$$\left\|\int_{\mathbb{F}_n} \mathbf{T}^n H(\omega)d\omega\right\|_F \le \text{const}(\rho+\varepsilon)^n\|H\|_F.$$

Since none of the choices of the constants in this proof depend on n, for simplicity we denote all constants by "const" even though the value of this may change with each occurrence.

Let $H(\omega) := (\sum_{\ell=1}^s (1-\cos w(\ell))^{r-1})I$. Since

$$(D^\beta(A^*(\omega)H(\omega))_{|_{\omega=0}})^T = D^\beta(H(\omega)A(\omega))_{|_{\omega=0}} = \mathbf{0}, \quad |\beta| \le r-1,$$

we have $H \in V_r$ and $H \ge I$ for all $\omega \in \mathbb{T}^s\backslash(1/2\mathbb{T}^s)$. Since $\|\mathbf{P}(\omega)-\mathbf{P}(0)\| \le \text{const}\|\omega\|$, the function $\widehat{\Phi}$ is bounded on $\mathbb{T}^s$.

We also note that

$$\widehat{\Phi}(\omega) = \Pi_n(\omega)\widehat{\Phi}(2^{-(n+1)}\omega).$$

Hence we have

$$\begin{aligned}\int_{\mathbb{F}_n} |\widehat{\varphi}^l(\omega)|^2 d\omega &= \int_{\mathbb{F}_n} \mathbf{i}_l^T \widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega)\mathbf{i}_l d\omega \\ &= \int_{\mathbb{F}_n} \mathbf{i}_l^T \Pi_n(\omega)\widehat{\Phi}(2^{-(n+1)}\omega)\widehat{\Phi}^*(2^{-(n+1)}\omega)\Pi_n^*(\omega)\mathbf{i}_l d\omega \\ &\le \text{const} \int_{\mathbb{F}_n} \mathbf{i}_l^T \Pi_n(\omega) H(2^{-(n+1)}\omega)\Pi_n^*(\omega)\mathbf{i}_l d\omega \\ &\le \text{const} \int_{2^n\mathbb{T}^s} \mathbf{i}_l^T \Pi_n(\omega) H(2^{-(n+1)}\omega)\Pi_n^*(\omega)\mathbf{i}_l d\omega \\ &\le \text{const} \| \int_{2^n\mathbb{T}^s} \Pi_n(\omega) H(2^{-(n+1)}\omega)\Pi_n^*(\omega) d\omega \|_F \\ &= \text{const} \| \int_{\mathbb{T}^s} (\mathbf{T}^{(n+1)} H)(\omega) d\omega \|_F \\ &\le \text{const}(\rho+\varepsilon)^{n+1},\end{aligned}$$

where $\mathbf{i}_l$ is the $\Phi \times 1$ column vector whose lth entry is 1 and all others are 0. ∎

This proposition together with the usual Littlewood–Paley technique leads to the following estimate of the regularity of the refinable function vector Φ.

THEOREM 4.2. *Suppose* $\mathbf{P}$ *satisfies the basic conditions* 1.5 *and the conditions* (4.1)*, and let* ρ *be the spectral radius of* $\mathbf{T}_{|V_r}$*. Then the function* $\Phi = (\varphi^l)^T$ *is in* $C^{\gamma-\varepsilon}$ *for any* $\varepsilon > 0$ *and* $\gamma = -\log\rho/(2\log 2) - s/2$.

Proof. Since when $n > N_{\mathbf{T}}$,

$$\int_{\mathbb{F}_n} |\widehat{\varphi}^l(\omega)|^2 d\omega \le \text{const}\rho^{n+1},$$

and since the function $\widehat{\varphi}^l$ is bounded on $2^{N_T}\mathbb{T}^s$,

$$\int_{\mathbb{R}^s} (1+|w|^2)^\kappa |\widehat{\varphi}^l(\omega)|^2 d\omega \le \text{const}\left(1 + \sum_{n=1}^{\infty} 2^{2n\kappa}\rho^{n+1}\right).$$

Hence $\varphi^l \in C^{\gamma-\varepsilon}$ where $\gamma = -\log\rho/(2\log 2) - s/2$. That is, $\Phi \in C^{\gamma-\varepsilon}$. ∎

REFERENCES

[1] B. K. Alpert and V. Rokhlin, *A fast algorithm for the evaluation of Legendre expansion*, SIAM J. Sci. Comput., 12 (1991), pp. 246–262.

[2] C. de Boor, R. DeVore, and A. Ron, *The structure of finitely generated shift-invariant spaces in* $L_2(\mathbb{R}^d)$, J. Funct. Anal., 119 (1994), pp. 37–78.

[3] A. Cohen and I. Daubechies, *A stability criterion for biorthogonal wavelet bases and their related subband coding scheme*, Duke Math. J., 68 (1992), pp. 313–335.

[4] A. Cohen I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., to appear.

[5] K. C. Chui and J. Lian, *A study on orthonormal multi-wavelets*, J. Appl. Numer. Math., 20 (1996), pp. 273–298.

[6] G. Donovan, J. S. Geronimo, and D. P. Hardin, *Intertwining Multiresolution Analysis and the Construction of Piecewise Polynomial Wavelets*, preprint.

[7] G. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1994), pp. 1158–1192.

[8] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Fractal functions and wavelet expansions based on several scaling functions*, J. Approx. Theory, 78 (1994), pp. 373–401.

[9] T. N. T. Goodman and S. L. Lee, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.
[10] T. N. T. Goodman, S. L. Lee, and W. S. Tang, *Wavelets in wandering subspaces*, Trans. Amer. Math. Soc., 338 (1993), pp. 639–654.
[11] P. Hervé, *Multi-resolution analysis of multiplicity d: Application to dynadic interpolation*, Appl. Comput. Harmonic. Anal., 1 (1994), pp. 299–315.
[12] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.
[13] R.-Q. Jia and C. A. Micchelli, *Using the refinement equation for the construction of pre-wavelets* II: *Powers of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.
[14] R.-Q. Jia and Z. Shen, *Multiresolution and wavelets*, Proc. Edinburgh Math. Soc., 37 (1994), pp. 271–300.
[15] R. Long, W. Chen, and S. Yuan, *Wavelets Generated by Vector Multiresolution Analysis*, preprint, 1995.
[16] W. Lawton, S. L. Lee, and Z. Shen, *An algorithm for matrix extension and wavelet construction*, Math. Comp., 65 (1996), pp. 723–737.
[17] W. Lawton, S. L. Lee, and Z. Shen, *Stability and orthonormality of multivariate refinable functions*, SIAM J. Math. Anal., 28 (1997), pp. 999–1014.
[18] W. Lawton, S. L. Lee, and Z. Shen, *Convergence of multidimensional cascade algorithm*, Numer. Math., to appear.
[19] S. D. Riemenschneider and Z. Shen, *Box splines, cardinal series and wavelets*, in Approximation Theory and Functional Analysis, C. K. Chui, ed., Academic Press, New York, 1991, pp. 133–149.
[20] S. D. Riemenschneider and Z. Shen, *Wavelets and pre-wavelets in low dimensions*, J. Approx. Theory, 71 (1992), pp. 18–38.
[21] S. D. Riemenschneider and Z. Shen, *Multidimensional interpolatory subdivision schemes*, SIAM J. Numer. Anal., 34 (1997), pp. 2357–2381.
[22] A. Ron and Z. Shen, *Frames and stable basis for shift invariant subspaces of* $L_2(\mathbb{R}^s)$, Canad. J. Math., 47 (1995), pp. 1951–1094.
[23] G. Strang, *Eigenvalues of* $(\downarrow 2)H$ *and convergence of cascade algorithm*, IEEE Trans. Sig. Proc., to appear.
[24] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley–Cambridge Press, Wellesley, MA, 1996.
[25] G. Strang and V. Strela, *Short wavelets and matrix dilation equations*, IEEE Trans. Signal Proc., 43 (1995), pp. 108–115.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 251–265, January 1998

014

ORTHOGONALITY OF SIEVED RANDOM WALK POLYNOMIALS FROM A NONSIEVED ANALOGUE*

BLAISE DESESA†

Abstract. The continuous component of the orthogonality relation for sieved random walk polynomials is derived in general from the orthogonality relation of another nonsieved random walk sequence called the companion polynomials. Conditions are stated on the three-term recurrence relation between the two polynomial sequences for a linear difference equation to hold. Results from Ismail are used to find the Stieltjes transform of the orthogonality measure of the sieved polynomials. The linear difference equation allows for the Stieltjes transform to be inverted. The theory is applied to the sieved associated ultraspherical polynomials, and in general to random walk polynomials with linear birth and death rates of $\beta_n = c\,n + a$ and $\delta_n = c\,n + b$, respectively.

Key words. associated ultraspherical polynomials, birth and death process, companion polynomials, orthogonal polynomials, random walk polynomials, sieved polynomials, Stieltjes transform

AMS subject classification. 33

PII. S0036141096298436

1. Introduction.

In this paper we will study sieved orthogonal polynomials and derive their orthogonality from a nonsieved random walk analogue. Random walk polynomials arise from a special stationary Markov process known as a birth and death process (BDP). Here the state space $\{X(t) : t \geq 0\}$ is the set of nonnegative integers. The transition probabilities $P_{m,n}(t)$ are the probabilities that the system goes from state m to state n in time t. The transition probabilities satisfy

$$\begin{aligned} &P_{n,n+1}(t) = \beta_n t + o(t), \\ &P_{n,n-1}(t) = \delta_n t + o(t), \\ &P_{n,n}(t) = 1 - \delta_n t - \beta_n t + o(t), \\ &o(t) \text{ otherwise, as } t \to 0^+. \end{aligned} \tag{1}$$

The β_n and δ_n are called the birth and death rates of the process at state n, and they satisfy the requirements $\beta_n > 0$, $\delta_{n+1} > 0$ for $n \geq 0$, and $\delta_0 > 0$.

The random walk polynomials of the BDP are defined as

$$\begin{aligned} &R_{-1}(x) = 0, \qquad R_0(x) = 1, \\ &xR_n(x) = B_n R_{n+1}(x) + D_n R_{n-1}(x), \quad n \geq 0, \end{aligned} \tag{2}$$

where

$$B_n = \frac{\beta_n}{\beta_n + \delta_n}, \qquad D_n = \frac{\delta_n}{\beta_n + \delta_n}. \tag{3}$$

Another set of orthogonal polynomials associated with a BDP is the birth and death polynomials $\{Q_n(x)\}$, generated by

$$\begin{aligned} &Q_0(x) = 1, \qquad Q_1(x) = \frac{\lambda_0 + \mu_0 - x}{\lambda_0}, \\ &-xQ_n(x) = \lambda_n Q_{n+1}(x) + \mu_n Q_{n-1}(x) - (\lambda_n + \mu_n)Q_n(x), \quad n > 0. \end{aligned}$$

*Received by the editors January 29, 1996; accepted for publication (in revised form) September 30, 1996.

http://www.siam.org/journals/sima/29-1/29843.html

†Mathematics Department, Allentown College of Saint Francis de Sales, 2755 Station Avenue, Center Valley, PA 18034-9568.

Most of the BDP and their random walk polynomials that have been studied have birth and death rates as linear functions of n. If $\beta_n = an + b$ and $\delta_n = cn + d$, then interpreting the states of the process as populations, "an" and "cn" represent the growth and decline of the population due to its current size, while "b" and "d" represent external forces.

Given a random walk sequence of polynomials, we may define the sieved random walk polynomials. Charris and Ismail in [6, 7] outlined this process and defined the sieved random walk polynomials of the first and second kinds in general. An original random walk polynomial sequence $\{R_n(x)\}$ is given, defined by the recursion (2). Next, another random walk sequence $\{S_n(x)\}$, called the dual sequence, is defined by the recursion

$$\begin{aligned} &S_{-1}(x) = 0, \qquad S_0(x) = 1, \\ &xS_n(x) = D_nS_{n+1}(x) + B_nS_{n-1}(x), \quad n \geq 0. \end{aligned} \tag{4}$$

The sieved random walk polynomials of the first kind, written $r_n(x;k)$, are next defined by

$$\begin{aligned} &r_0(x;k) = 1, \qquad r_1(x;k) = x, \\ &xr_n(x;k) = d_{n-1}r_{n+1}(x;k) + b_{n-1}r_{n-1}(x;k), \quad n > 0, \end{aligned} \tag{5}$$

where

$$\begin{aligned} &b_n = d_n = \frac{1}{2} \text{ if } n+1 \neq mk, \\ &b_{mk-1} = B_{m-1}, d_{mk-1} = D_{m-1}, m = 1, 2, \ldots. \end{aligned} \tag{6}$$

The sieved random walks of the second kind, $s_n(x;k)$, are defined similarly by

$$\begin{aligned} &s_0(x;k) = 1, \qquad s_1(x;k) = 2x, \\ &xs_n(x;k) = b_ns_{n+1}(x;k) + d_ns_{n-1}(x;k), \quad n > 0. \end{aligned} \tag{7}$$

The first sieved polynomials studied were the sieved analogues of the ultraspherical polynomials $C_n^{\lambda}(x)$ by Al-Salam, Allaway, and Askey [1, 2]. These were studied as special limiting cases of the q-continuous ultraspherical polynomials $C_n(x;\beta|q)$.

Explicit representations of the sieved polynomials of the first and second kind were established by Charris and Ismail [6] in terms of the random walk polynomials $R_n(x)$ and the Tchebyshev polynomials of the first and second kinds, $T_n(x)$ and $U_n(x)$, respectively. These formulas allowed Charris and Ismail to find the Stieltjes transform of the measure of orthogonality of the sieved polynomials of the first kind. The absolutely continuous component of the measure of orthogonality of the sieved polynomials of the first and second kind have been related to each other in Ismail [11].

2. Mathematical preliminaries. Given a distribution function $\mu(x)$, there is a sequence of polynomials $\{P_n(x)\}$, where $P_n(x)$ is of exact degree n, such that

$$\int_{-\infty}^{\infty} P_n(x)P_m(x)d\mu(x) = \lambda_n\delta_{nm}, \quad \lambda_n > 0 \tag{8}$$

(see Chihara [8, p. 14]). The sequence $\{P_n(x)\}$ is called an orthogonal polynomial sequence (OPS).

Given the distribution function $\mu(x)$, then the OPS $\{P_n(x)\}$ always satisfies a three-term recurrence relation

$$P_{n+1}(x) = (A_n x + B_n)P_n(x) - C_n P_{n-1}(x), \quad n = 1, 2, \ldots, \tag{9}$$

where the coefficients A_n, B_n, C_n are real and such that $A_{n-1}A_nC_n > 0$, $n = 1, 2, \ldots$, which is the positivity condition to ensure that the polynomials are orthogonal with respect to a positive measure [8, p. 19].

From the orthogonality condition (8) and the three-term recurrence (9), we obtain

$$\lambda_n = \frac{A_0}{A_n} \prod_{k=1}^{n} C_k. \tag{10}$$

The vth associated polynomials of $P_n(x)$, written $P_n^{(v)}(x)$, are defined by the recursion

$$\begin{aligned} &P_{-1}^{(v)}(x) = 0, \qquad P_0^{(v)}(x) = 1, \\ &P_{n+1}^{(v)}(x) = (A_{n+v}x + B_{n+v})P_n^{(v)}(x) - C_{n+v}P_{n-1}^{(v)}(x). \end{aligned} \tag{11}$$

These polynomials are orthogonal by Favard's theorem [8]. The numerator polynomials of $P_n(x)$ are defined as

$$P_n^*(x) = A_0 P_{n-1}^{(1)}(x). \tag{12}$$

The Stieltjes transform of a distribution function $\mu(t)$ is defined as the function

$$\chi(z) = \int_{-\infty}^{\infty} \frac{d\mu(t)}{z - t} \tag{13}$$

for wherever the integral is convergent. Markov's theorem asserts that when the support of $\mu(x)$ is compact, with $[\xi_1, \eta_1]$ being the smallest closed interval containing the support of $\mu(x)$, then $\chi(z)$ is an analytic function for $z \notin [\xi_1, \eta_1]$, and

$$\chi(z) = \lim_{n\to\infty} \frac{P_n^*(z)}{P_n(z)} = \lim_{n\to\infty} \frac{A_0 P_{n-1}^{(1)}(z)}{P_n(z)}. \tag{14}$$

(See Askey and Ismail [3] and Wall [19].) For random walk polynomials, the interval $[\xi_1, \eta_1]$ is always $[-1, 1]$ (see Karlin and McGregor [13, 14]).

Once the Stieltjes transform of a distribution function is known from Markov's theorem, we will want to recover $d\mu_1(x)$, the absolutely continuous component of the measure induced by the distribution function $\mu(x)$. To this end we employ a corollary to the Stieltjes inversion formula. If $\mu(x)$ is of bounded variation on $(-\infty, \infty)$, then

$$d\mu_1(t) = \frac{1}{2\pi i} \lim_{\varepsilon\to 0^+} [\chi(t - i\varepsilon) - \chi(t + i\varepsilon)], \quad t \in \text{supp } (\mu(t)). \tag{15}$$

(See Widder [20].)

Another method of finding the measure $d\mu(x)$ for an OPS $\{P_n(x)\}$ is given by the theorem of Nevai: let $P_n(x)$ be defined by the recursion (9), and let $\tilde{P}_n(x) = P_n(x)/\sqrt{\lambda_n}$ be the corresponding orthonormal series. If in (9) we have

$$\sum_{n=0}^{\infty} \left\{ |B_n A_n^{-1}| + \left| C_{n+1}^{1/2} A_n^{-1/2} A_{n+1}^{-1/2} - \frac{\gamma}{2} \right| \right\} < \infty \tag{16}$$

for some γ, then

$$d\mu(x) = d\mu_1(x) + d\mu_2(x), \tag{17}$$

where $d\mu_1(x)$ is continuous and a positive measure in $(-\gamma, \gamma)$, $\text{supp}(d\mu) = [-\gamma, \gamma]$ is the smallest closed interval containing the support of $\mu(x)$, and $\mu_2(x)$ is a step function that is constant in $(-\gamma, \gamma)$. Further,

$$\limsup_{n\to\infty} \left\{ d\mu_1(x)\sqrt{\gamma^2 - x^2}\ \tilde{P}_n^2(x) \right\} = \frac{2}{\pi} \tag{18}$$

holds almost everywhere (a.e.) in $\text{supp}(d\mu)$. (See Nevai [15, Corollary 36, p. 141 and Theorem 40, p. 143].)

A theorem useful for finding the isolated jumps in the distribution function $\mu(x)$ is due to Shohat and Tamarkin [18, Corollary 26, pp. 45–46]. Let $\{\tilde{P}_n(x)\}$ be the orthonormal polynomials with respect to the distribution function $\mu(x)$. Let

$$[\rho(x)]^{-1} = \sum_{n=0}^{\infty} \tilde{P}_n^2(x). \tag{19}$$

When the corresponding Hamburger moment problem is determined, then $\rho(x) = 0$ at all points of continuity of $\mu(x)$ and equals the jump of $\mu(x)$ at a point of discontinuity.

Let $\sigma_1(x)$ be the distribution function for the polynomials $r_n(x;k)$ and $\chi_1(x;k) = \int_{-1}^{1} \frac{d\sigma_1(t)}{x-t}$ be the Stieltjes transform for $r_n(x;k)$. Charris and Ismail [6] proved the result

$$\chi_1(x;k) = \lim_{m\to\infty} \frac{a_{mk} U_{k-1}(x) S_{m-1}(T_k(x))}{R_m(T_k(x)) - R_{m-2}(T_k(x))}, \quad x \notin [-1, 1], \tag{20}$$

where

$$a_{mk} = \{D_0 D_1 \cdots D_{m-2}\}/\{B_0 B_1 \cdots B_{m-1}\}, \quad m > 1. \tag{21}$$

Further, if $\sigma_2(x)$ is the distribution function for $s_n(x;k)$, then

$$d\sigma_2(x) = c(1 - x^2) d\sigma_1(x), \tag{22}$$

c is a constant. (See Ismail [11].)

In the first part of this paper conditions will be developed that allow for the difference $R_n(x) - R_{n-2}(x)$ that occur in the denominator of (20) to be expressed as a linear combination of another random walk polynomial $P_n(x)$, called the companion polynomials to $R_n(x)$. It will then be shown that if the orthogonality relationship for the companion polynomials is known, then by using (20), it will be possible to invert the Stieltjes transform for the sieved polynomials of the first kind of $R_n(x)$ and obtain its absolutely continuous component of the measure of orthogonality.

In the second section, this general theory will be applied to the random walk polynomials $R_n(x; a, b, c)$ having linear birth and death rates given by $\beta_n = cn + a$, $\delta_n = cn + b$. This will lead to a mixed recursion relation with interesting special cases. The orthogonality for the sieved polynomials of the first kind of $R_n(x; a, b, c)$ will be discussed after this.

We will use the usual notation for the Gaussian hypergeometric series, writing

$${}_2F_1(a, b; c; z) = {}_2F_1\left(\begin{matrix} a, b \\ c \end{matrix} ; z \right) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!}, \tag{23}$$

where

$$(a)_k = a(a+1)\cdots(a+k-1) = \Gamma(a+k)/\Gamma(a), \quad k = 1, 2, \ldots, \tag{24}$$
$$(a)_0 = 1.$$

3. A linear difference relation for sieved polynomials. In this section we seek a linear representation of the difference $R_n(x) - R_{n-2}(x)$ of a random walk polynomial in terms of another random walk polynomial. To this end we formulate the problem as follows. Let $R_n(x)$ and its dual $S_n(x)$ be given as defined by (2) and (4). These are the random walk polynomials whose sieved analogues we are trying to determine an orthogonality relationship for. We want to know under what conditions are there constants θ_n, ϕ_n, and ω_n, and another random walk sequence $\{P_n(x)\}$, such that

$$S_n(x) = \theta_n P_n^{(1)}(x), \quad n \geq 0, \tag{25}$$

$$R_n(x) - R_{n-2}(x) = \phi_n P_n(x) + x\omega_n P_{n-1}^{(1)}(x), \quad n \geq 1. \tag{26}$$

The polynomials $P_n(x)$ will be called the companion polynomials to $R_n(x)$, and $P_n^{(1)}(x)$ are the associated polynomials $P_n^{(v)}(x)$ with parameter $v = 1$. The companion polynomials are defined recursively by

$$P_{-1}(x) = 0, \qquad P_0(x) = 1, \tag{27}$$
$$xP_n(x) = \overline{B}_n P_{n+1}(x) + \overline{D}_n P_{n-1}(x), \quad n \geq 0,$$
$$\overline{B}_n + \overline{D}_n = 1, \quad n \geq 1.$$

The Stieltjes transform $\chi_P(x)$ for $P_n(x)$ will be assumed known, as well as the absolutely continuous component of the measure of orthogonality, $d\mu_P(x)$.

We state the following theorem.

THEOREM 1. *If*

$$B_n(1 - B_{n-1}) = \overline{B}_n(1 - \overline{B}_{n+1}), \quad n \geq 1, \tag{28}$$

then (25) *and* (26) *hold, with the constants* θ_n, ϕ_n, *and* ω_n *given by*

$$\theta_0 = 1, \qquad \theta_n = \frac{\overline{B}_1\overline{B}_2\cdots\overline{B}_n}{D_0 D_1 \cdots D_{n-1}}, \quad n \geq 1,$$
$$\phi_1 = \frac{1}{\overline{D}_1}, \qquad \phi_n = \frac{\overline{B}_1\overline{B}_2\cdots\overline{B}_{n-1}}{\overline{D}_1 B_1 B_2 \cdots B_{n-1}}, \quad n \geq 2, \tag{29}$$
$$\omega_n = \overline{D}_1\left(\frac{1}{B_0} - \frac{1}{\overline{B}_1\overline{D}_1}\right)\phi_n, \quad n \geq 1.$$

Proof. We shall proceed by induction on n. First, note (28) is equivalent to $\overline{B}_n\overline{D}_{n+1} = B_n D_{n-1}$, $n \geq 1$. The recurrence relation for $P_n^{(1)}(x)$ is

$$P_{-1}^{(1)}(x) = 0, P_0^{(1)}(x) = 1,$$
$$xP_n^{(1)}(x) = \overline{B}_{n+1}P_{n+1}^{(1)}(x) + \overline{D}_{n+1}P_{n-1}^{(1)}(x), \quad n \geq 0. \tag{30}$$

Thus $P_1^{(1)}(x) = \frac{1}{B_1}x$. Let $Q_n(x) = \theta_n P_n^{(1)}(x)$, $n \geq 0$, $Q_{-1}(x) = 0$. It is clear that $Q_n(x) = S_n(x)$ for $n = 0, 1$. Multiplying (30) throughout by θ_n, we obtain

$$xQ_n(x) = D_nQ_{n+1}(x) + \frac{\overline{D}_{n+1}\overline{B}_n}{D_{n-1}}Q_{n-1}(x) = D_nQ_{n+1}(x) + B_nQ_{n-1}(x), \quad n \geq 1.$$

Thus, $Q_n(x)$ has the same recursion and initial conditions as the dual polynomials $S_n(x)$ do. This establishes (25).

When $n = 1$, (26) becomes $R_1(x) = \phi_1 P_1(x) + x\omega_1$, which is easily verified as true. Likewise, using the recursion for $R_2(x)$, $P_2(x)$, and the definitions for ϕ_2 and ω_2, along with the fact that $B_1 + D_1 = 1$, it is routine to verify that (26) is true for $n = 2$.

Now assume (26) is true for $n \geq 2$ in general. Multiply both sides of (26) by x, then use the recursion relations for $R_n(x)$, $P_n(x)$, and $P_n^{(1)}(x)$. This yields

$$\begin{aligned} &B_nR_{n+1}(x) + (D_n - B_{n-2})R_{n-1}(x) - D_{n-2}R_{n-3}(x) \\ &= \phi_n[\overline{B}_nP_{n+1}(x) + \overline{D}_nP_{n-1}(x)] + x\omega_n[\overline{B}_nP_n^{(1)}(x) + \overline{D}_nP_{n-2}^{(1)}(x)]. \end{aligned} \tag{31}$$

Since $B_n + D_n = 1$ for $n \geq 0$, then $D_n - B_{n-2} = D_{n-2} - B_n$ for $n \geq 2$. Thus the left side of (31) becomes $B_n[R_{n+1}(x) - R_{n-1}(x)] + D_{n-2}[R_{n-1}(x) - R_{n-3}(x)]$. Applying the induction hypothesis to $R_{n-1}(x) - R_{n-3}(x)$ with $n \to n - 1$ and simplifying, we obtain

$$\begin{aligned} B_n[R_{n+1}(x) - R_{n-1}(x)] &= \phi_n\overline{B}_nP_{n+1}(x) + (\phi_n\overline{D}_n - \phi_{n-1}D_{n-2})P_{n-1}(x) \\ &+ x[\omega_n\overline{B}_nP_n^{(1)}(x) + (\omega_n\overline{D}_n - \omega_{n-1}D_{n-2})P_{n-2}^{(1)}(x)]. \end{aligned} \tag{32}$$

To simplify the above, first note that $\phi_{n+1} = \frac{\overline{B}_n}{B_n}\phi_n$, $\omega_{n+1} = \frac{\overline{B}_n}{B_n}\omega_n$, $n \geq 1$. Therefore $\phi_n\overline{D}_n = \frac{\overline{B}_{n-1}\overline{D}_n}{B_{n-1}}\phi_{n-1}$, $n \geq 2$. But $\overline{D}_n = \frac{B_{n-1}D_{n-2}}{\overline{B}_{n-1}}$, $n \geq 2$, so $\phi_n\overline{D}_n - \phi_{n-1}D_{n-2} = 0$, $n \geq 2$. Similarly, $\omega_n\overline{D}_n - \omega_{n-1}D_{n-2} = 0$, $n \geq 2$. Therefore, (32) simplifies to

$$R_{n+1}(x) - R_{n-1}(x) = \phi_{n+1}P_{n+1}(x) + x\omega_{n+1}P_n^{(1)}(x).$$

This is (26) with $n \to n + 1$, so the theorem is proved by induction. □

Once a suitable set of companion polynomials are selected, we may proceed to find $d\sigma_1(x; k)$, the absolutely continuous component of orthogonality for $r_n(x; k)$, the sieved polynomials of the first kind of $R_n(x)$, and the Stieltjes transform $\chi_1(x; k)$ of this measure.

By Theorem 1, $\frac{S_{n-1}(x)}{R_n(x)-R_{n-2}(x)} = \frac{\theta_{n-1}P_{n-1}^{(1)}(x)/P_n(x)}{\phi_n+x\omega_nP_{n-1}^{(1)}(x)/P_n(x)}$. Substituting the values in Theorem 1 for θ_n, ϕ_n, ω_n, and the values in (21) for a_{nk}, we have, after simplification and the use of Markov's theorem,

$$\lim_{n\to\infty} \frac{a_{n\,k}S_{n-1}(x)}{R_n(x) - R_{n-2}(x)} = \frac{\overline{B}_0}{B_0} \frac{\chi_P(x)}{\frac{1}{\overline{D}_1} + \overline{B}_0\left(\frac{1}{B_0} - \frac{1}{\overline{B}_0\overline{D}_1}\right)x\chi_P(x)}, \tag{33}$$

$x \notin [-1, 1]$. Since $x \in [-1, 1] \Leftrightarrow T_k(x) \in [-1, 1]$, replacing x by $T_k(x)$ in the above, and using (20), we have the following theorem.

THEOREM 2.

$$\chi_1(x;k)=\frac{\overline{B}_0}{B_0}\ \frac{U_{k-1}(x)\chi_P(T_k(x))}{\dfrac{1}{\overline{D}_1}+\overline{B}_0\left(\dfrac{1}{B_0}-\dfrac{1}{\overline{B}_0\overline{D}_1}\right)T_k(x)\chi_P(T_k(x))},\ x\notin[-1,1]. \tag{34}$$

To use the Stieltjes inversion formula on $\chi_1(x;k)$ to recover $d\mu_1(x;k)$, it is first necessary to mention a special case of a theorem due to Geronimo and Van Assche. If μ_0 is a probability measure with Stieltjes transform $\chi(z,\mu_0)$ and μ is another measure whose Stieltjes transform is given by $\chi(z,\mu)=U_{k-1}(z)\chi(T_k(z),\,\mu_0)$, where $T_k(x)$ and $U_k(x)$ are the Tchebyshev polynomials of the first and second kind, then the absolutely continuous components of the measures μ and μ_0 satisfy the relation $d\mu(x)=|U_{k-1}(x)|d\mu_0(T_k(x))$. (See Geronimo and Van Assche [10].)

If we let

$$\chi(x,\rho)=\lim_{n\to\infty}\frac{a_{nk}S_{n-1}(x)}{R_n(x)-R_{n-2}(x)},\quad x\notin[-1,1],$$

be the Stieltjes transform for some measure ρ, then $\chi_1(x;k)=U_{k-1}(x)\chi(T_k(x),\rho)$. Using (15), the corollary to the Stieltjes inversion formula, then (33) becomes

$$d\rho(x)=\frac{\overline{B}_0}{B_0\overline{D}_1}\ \frac{d\mu_P(x)}{\left|\dfrac{1}{\overline{D}_1}+\overline{B}_0\left(\dfrac{1}{B_0}-\dfrac{1}{\overline{B}_0\overline{D}_1}\right)x\chi_P(x)\right|^2},\quad x\in(-1,1). \tag{35}$$

Here the Stieltjes transform for the companion polynomials,

$$\chi_P(x)=\text{p.v.}\int_{-1}^{1}\frac{d\mu_P(t)}{x-t},\quad x\in(-1,1),$$

is understood to be a Cauchy principal value integral. Now applying the theorem of Geronimo and Van Assche to (35), we obtain the following theorem.

THEOREM 3.

$$d\sigma_1(x;k)=\frac{\overline{B}_0}{B_0\overline{D}_1}\ \frac{|U_{k-1}(x)|d\mu_P(T_k(x))}{\left|\dfrac{1}{\overline{D}_1}+\overline{B}_0\left(\dfrac{1}{B_0}-\dfrac{1}{\overline{B}_0\overline{D}_1}\right)T_k(x)\chi_P(T_k(x))\right|^2},\quad x\in(-1,1). \tag{36}$$

It is always possible to find companion polynomials $P_n(x)$ such that the formula $B_nD_{n-1}=\overline{B}_n\overline{D}_{n+1}$, $n\geq 1$, is satisfied. Choose $\overline{B}_n=D_{n-1}$ and $\overline{D}_n=B_{n-1}$. The requisite equation (28) is satisfied, and so is the random walk condition $\overline{B}_n+\overline{D}_n=1$, $n\geq 1$ for $P_n(x)$. These are called the natural companion polynomials and are defined as

$$\begin{aligned}&xP_n(x)=D_{n-1}P_{n+1}(x)+B_{n-1}P_{n-1}(x),\quad n\geq 1,\\ &P_{-1}(x)=0,\qquad P_0(x)=1,\qquad P_1(x)=\frac{x}{\overline{B}_0}.\end{aligned} \tag{37}$$

Here $\overline{B}_0$ and $\overline{D}_0$ need not be specified, since $\overline{D}_0$ plays no role as $P_{-1}(x)=0$. With this selection for $P_n(x)$, we see $P_n^{(1)}(x)=S_n(x)$. The constants θ_n, ϕ_n, and ω_n in

Theorem 1 become

$$\begin{aligned} &\theta_n = 1, \quad n \geq 0, \\ &\phi_1 = \frac{1}{B_0}, \qquad \phi_n = a_{nk}, \quad n \geq 2, \\ &\omega_1 = \frac{1}{B_0}\left(1 - \frac{1}{\overline{B}_0}\right), \qquad \omega_n = a_{nk}\left(1 - \frac{1}{\overline{B}_0}\right), \quad n \geq 2. \end{aligned} \tag{38}$$

Theorem 1 becomes

$$R_n(x) - R_{n-2}(x) = a_{nk}\left[P_n(x) + \left(1 - \frac{1}{\overline{B}_0}\right) x P^{(1)}_{n-1}(x)\right], \quad n \geq 2. \tag{39}$$

The formulas for the Stieltjes transform and measure of orthogonality for $r_n(x;k)$ become

$$\chi_1(x;k) = \frac{\overline{B}_0 U_{k-1}(x)\chi_P(T_k(x))}{1 + (\overline{B}_0 - 1)T_k(x)\chi_P(T_k(x))}, \quad x \notin [-1,1], \tag{40}$$

$$d\sigma_1(x;k) = \frac{\overline{B}_0 |U_{k-1}(x)| d\mu_P(T_k(x))}{|1 + (\overline{B}_0 - 1)T_k(x)\chi_P(T_k(x))|^2}, \quad x \in (-1,1). \tag{41}$$

4. Orthogonality of the associated ultraspherical polynomials. In this section we will study the random walk polynomials $R_n(x;a,b,c)$, or simply $R_n(x)$ when the dependence on the parameters is not necessary to state, with linear birth and death rates given by $\beta_n = cn + a$, $\delta_n = cn + b$. This will give us the background to then select a companion random walk polynomial sequence to $R_n(x)$ and to study the sieved analogues in the next section.

These polynomials satisfy the recursion

$$\begin{aligned} &R_{-1}(x) = 0, \qquad R_0(x) = 1, \\ &(cn + a)R_{n+1}(x) = (2cn + a + b)xR_n(x) - (cn + b)R_{n-1}(x), \quad n \geq 0. \end{aligned} \tag{42}$$

The case of $c = 0$ is seen to be the Tchebyshev polynomials $R_n(x) = (\frac{b}{a})^{n/2} U_n(\frac{a+b}{2\sqrt{ab}}x)$. Dispensing with this case, we may divide the recursion (42) by c, replace a by $\frac{a}{c}$ and b by $\frac{b}{c}$, thereby scaling c to 1. This leaves the recursion to the simplified form

$$(n + a)R_{n+1}(x;a,b) = (2n + a + b)xR_n(x;a,b) - (n + b)R_{n-1}(x;a,b), \quad n \geq 0. \tag{43}$$

These are seen to be the associated ultraspherical polynomials $C_n^{(\gamma)}(x;\beta) = R_n(x;\gamma+1, 2\beta+\gamma-1)$. The ultraspherical polynomials are $C_n^{\lambda}(x) = R_n(x;1,2\lambda-1)$.

The corresponding birth and death polynomials when $\beta_n = n+\alpha+c+1$, $\delta_n = n+c$, $n \geq 0$, are the associated Laguerre polynomials $L_n^{\alpha}(x;c)$. These have been studied in Askey and Wimp [4] and Ismail, Letessier, and Valent [12].

The positivity condition on the measure for $R_n(x)$ is found from (43) to be

$$\frac{(2n + a + b)(2n + a + b - 2)(n + b)}{(n + a)^2(n + a - 2)} > 0, \quad n = 1, 2, \ldots.$$

The intersection of these inequalities for $n = 1, 2, \ldots$, is found to be $a > 1, b > -1$.

Let $G(x,t)=\sum_{n=0}^{\infty}R_n(x;a,b)t^n$ be the generating function for the polynomials $R_n(x;a,b)$. Note that $G(x,0)=1$. Multiplying (43) by t^{n+1} and summing from $n=0$ to ∞, we obtain the differential equation

$$t(t-\alpha)(t-\beta)\frac{\partial G(x,t)}{\partial t}=-(1+b)(t-p)(t-q)G(x,t)+(a-1).$$

Here α,β are the roots of the equation $t^2-2tx+1=0$, with $|\alpha|\le|\beta|$. We note that $\alpha+\beta=2x$, $\alpha\beta=1$, and $|\alpha|=|\beta|\Leftrightarrow|x|\le 1$. Also, p,q are the roots of $-(1+b)t^2+(a+b)xt+(1-a)=0$. We have $p+q=\frac{a+b}{1+b}x$, $pq=\frac{a-1}{b+1}$. Let $A=1-a$, $B=\frac{a-b}{2}-1$. Solving the above differential equation, we obtain

$$G(x,t)=(a-1)t^A(1-2xt+t^2)^B\quad\int_0^t u^{-A-1}(1-2xu+u^2)^{-B-1}du.$$

For the integral to converge, we must have $A<0$, or $a>1$, which is implied from the positivity condition.

We may use the method of Darboux (see Olver [16, section 8.9]) to obtain asymptotic estimates for $R_n(x)$. Analyzing $x\notin[-1,1]$ first, we find a comparison function to be

$$\lim_{t\to\alpha}\left(1-\frac{t}{\alpha}\right)^{-B}G(x,t)=(a-1)\alpha^A\left(1-\frac{\alpha}{\beta}\right)^B\int_0^{\alpha}u^{-A-1}\left(1-\frac{u}{\alpha}\right)^{-B-1}\left(1-\frac{u}{\beta}\right)^{-B-1}du,$$

$B\neq 0,1,\ldots$. The above integral is a Hadamard integral (see Askey and Ismail [3, Chapter 5]). Using the binomial theorem for $(1-\frac{t}{\alpha})^B=\sum_{n=0}\frac{(-B)_n}{n!}\frac{t^n}{\alpha^n}$, and Stirling's asymptotic estimate $\frac{(-B)_n}{n!}\alpha^{-n}\sim\frac{\alpha^{-n}}{\Gamma(-B)}n^{-B-1}$, we obtain the asymptotic estimate

$$R_n(x)\sim(a-1)\alpha^A\left(1-\frac{\alpha}{\beta}\right)^B\frac{\alpha^{-n}}{\Gamma(-B)}n^{-B-1}\int_0^{\alpha}u^{-A-1}\left(1-\frac{u}{\alpha}\right)^{-B-1}\left(1-\frac{u}{\beta}\right)^{-B-1}du,$$

$x\in[-1,1]$.

To evaluate the above integral, we may use the binomial theorem on the factor $(1-\frac{u}{\beta})^{-B-1}$ inside the integral, reverse order of summation and integration, and use the definition of beta integrals to obtain after simplification the result

$$\begin{aligned}R_n(x;a,b)\sim{}&\frac{\Gamma(a)}{\Gamma\left(\frac{a+b}{2}\right)}\left(1-\frac{\alpha}{\beta}\right)^{\frac{a-b}{2}-1}\\&\times{}_2F_1\left(a-1,\frac{a-b}{2};\ \frac{a+b}{2};\ \frac{\alpha}{\beta}\right)\alpha^{-n}n^{(b-a)/2},\\&n\to\infty,x\notin[-1,1].\end{aligned}\tag{44}$$

There is no need to separately consider the case where $B=0,1,2,\ldots$, since the factor $\Gamma(-B)$ has cancelled.

For $x\in(-1,1)$, $|\alpha|=|\beta|$. Due to the two singularities of the generating function $G(x,t)$ that are equidistant from the origin, a comparison function to use in the method of Darboux in this case is

$$(a-1)\alpha^A\left(1-\frac{\alpha}{\beta}\right)^B\int_0^{\alpha}u^{-A-1}\left(1-\frac{u}{\alpha}\right)^{-B-1}\left(1-\frac{u}{\beta}\right)^{-B-1}du$$

$+$ complex conjugate.

Letting $x = \cos\theta$, $\alpha = e^{i\theta}$, $\beta = e^{-i\theta}$, we obtain the asymptotic estimate

$$\begin{aligned} R_n(x;a,b) &\sim 2\frac{\Gamma(a)}{\Gamma\left(\dfrac{a+b}{2}\right)}\ |1-e^{2i\theta}|^{(\frac{a-b}{2})-1} \\ &\times \left|{}_2F_1\left(\frac{a-b}{2}, a-1; \frac{a+b}{2}; e^{2i\theta}\right)\right| n^{(\frac{b-a}{2})}\cos(-n\theta+\phi), \\ &n\to\infty, \quad x\in(-1,1), \end{aligned} \tag{45}$$

where $\phi = \arg[(1-e^{2i\theta})^{(\frac{a-b}{2})-1}\ {}_2F_1\ (\frac{a-b}{2}, a-1; \frac{a+b}{2}; e^{2i\theta})]$.

To obtain the asymptotics at $x = 1$, we find the differential equation of the generating function $G(x,t)$ of $R_n(x)$ at $x = 1$ to be

$$t(t-1)^2\frac{\partial G(1,t)}{\partial t} = [-(b+1)t^2 + (a+b)t + 1 - a]G(1,t) + (a-1).$$

Solving this we obtain

$$G(1,t) = (a-1)t^{1-a}(1-t)^{a-b-2}\int_0^t u^{a-2}(1-u)^{b-a}du, \quad a>1.$$

A comparison function in the method of Darboux is

$$(a-1)(1-t)^{a-b-2}\int_0^1 u^{a-2}(1-u)^{b-a}du, \quad b-a\neq -1,-2,\ldots.$$

Evaluating the above Hadamard integral and using the binomial theorem and Stirling's asymptotic estimate, we obtain after simplification

$$\begin{aligned} R_n(1) &\sim \frac{1}{\Gamma(-a+b+2)}\ {}_2F_1(a-b, a-1; a; 1)n^{-a+b+1}, \\ &n\to\infty, \qquad b-a\neq -1,-2,\ldots. \end{aligned} \tag{46}$$

Since the recursion for $R_n(x)$ is symmetric, then $R_n(-1) = (-1)^n R_n(1)$, and therefore this case need not be considered separately (Chihara [8, p. 21]).

The numerator polynomials are defined in terms of the associated polynomials with parameter $v = 1$. We see that $R_n^{(1)}(x;a,b) = R_n(x;a+1,b+1)$. Therefore, by (44), we obtain the asymptotic estimate

$$\begin{aligned} R_n^{(1)}(x;a,b) &\sim \frac{\Gamma(a+1)}{\Gamma\left(\dfrac{a+b}{2}+1\right)}\ \left(1-\frac{\alpha}{\beta}\right)^{\frac{a-b}{2}-1} \\ &\times {}_2F_1\left(a, \frac{a-b}{2}; \frac{a+b}{2}+1; \frac{\alpha}{\beta}\right)\alpha^{-n}n^{(b-a)/2}, \\ &n\to\infty, \quad x\notin[-1,1]. \end{aligned} \tag{47}$$

The Stieltjes transform of the distribution function for the polynomials $R_n(x;a,b)$ can be obtained next. Let the orthogonality be written as

$$\int_{-1}^1 R_n(x;a,b)R_m(x;a,b)d\mu(x) = \lambda_n\delta_{mn}.$$

Let $\chi(x) = \int_{-1}^{1} \frac{d\mu(t)}{x-t}, x \notin [-1,1]$, be the Stieltjes transform. Applying Markov's theorem to asymptotic estimates (44) and (47), we obtain after simplification

$$\chi(x) = 2\alpha \frac{{}_2F_1\left(a, \frac{a-b}{2}; \frac{a+b}{2}+1; \frac{\alpha}{\beta}\right)}{{}_2F_1\left(a-1, \frac{a-b}{2}, \frac{a+b}{2}; \frac{\alpha}{\beta}\right)}, \quad x \notin [-1,1]. \tag{48}$$

Now that the asymptotic estimates for $R_n(x;a,b)$ are known for $x \in (-1,1)$, we may use the theorem of Nevai to obtain $d\mu(x)$, the continuous component of the measure of orthogonality. The hypotheses of the theorem are seen to be satisfied by a routine order estimation of the recursion coefficients. Using formula (10) and Stirling's formula, we have

$$\lambda_n = \frac{a+b}{a} \frac{(n+a)}{(2n+a+b)} \frac{(b+1)_n}{(a+1)_n} \sim \frac{a+b}{2b} \frac{\Gamma(a)}{\Gamma(b)} n^{b-a}. \tag{49}$$

Letting $x = \cos\theta$, we then have $\sqrt{1-x^2} = |\sin\theta|$, and $|1-e^{2i\theta}| = 2|\sin\theta|$. Therefore, using (45) and (49), we can write

$$\begin{aligned} d\mu(x) &= \limsup_{n\to\infty} \frac{2}{\pi} \frac{1}{\sqrt{1-x^2}} \frac{\lambda_n}{R_n^2(x,a,b)} \\ &= \frac{1}{\pi} 2^{b-a} \frac{a+b}{b} \frac{\Gamma^2\left(\frac{a+b}{2}\right)}{\Gamma(a)\Gamma(b)} |\sin\theta|^{b-a+1} \left| {}_2F_1\left(a-1, \frac{a-b}{2}; \frac{a+b}{2}; e^{2i\theta}\right)\right|^{-2} dx, \end{aligned} \tag{50}$$

$x \in (-1,1)$.

A partial analysis of mass points in the distribution function $\mu(x)$ may be obtained using (19). From asymptotic estimates (46) and (49), we have $\widetilde{R}_n^2(1) = \frac{R_n^2(1)}{\lambda_n} = O n^{-a+b+2}$. Therefore, if $a-b>3$, then $\sum_{n=0}^{\infty} \widetilde{R}_n^2(1)$ is divergent, and $\mu(x)$ does not have a mass point at $x=1$. The same analysis holds at $x=-1$.

The dual polynomials are defined by the recursion

$$(2n+a+b)xS_n(x;a,b) = (n+b)S_{n+1}(x;a,b) + (n+a)S_{n-1}(x;a,b), \tag{51}$$

with the usual initial conditions. By showing that the same recurrence and initial conditions are satisfied, we can make the following identifications:

$$S_n(x\colon a,b) = R_n(x\colon b,a) = R_n^{(1)}(x\colon b-1, a-1). \tag{52}$$

From asymptotic formula (44), we therefore have

$$S_n(x;a,b) \sim \left(1-\frac{\alpha}{\beta}\right)^{\frac{b-a}{2}-1} \frac{\Gamma(b)}{\Gamma\left(\frac{a+b}{2}\right)} {}_2F_1\left(b-1, \frac{b-a}{2}; \frac{a+b}{2}; \frac{\alpha}{\beta}\right) \alpha^{-n} n^{(a-b)/2}, \tag{53}$$

$n \to \infty, \quad x \notin [-1,1]$.

5. The sieved associated ultraspherical polynomials. We now have the background material to study the companion polynomials and the various sieved polynomials of $R_n(x)$. Let $r_n(x;k)$ be the sieved polynomials of the first kind of $R_n(x)$. From the birth and death rates $\beta_n = n + a$, $\delta_n = n + b$ of $R_n(x)$, we identify the coefficients

$$B_n = \frac{n+a}{2n+a+b}, \quad D_n = \frac{n+b}{2n+a+b}. \tag{54}$$

Therefore, the coefficients a_{mk} in (21) are given by

$$a_{mk} = (a+b+2m-2)\frac{(b)_{m-1}}{(a)_n} \sim \frac{2\Gamma(a)}{\Gamma(b)}\, m^{b-a}. \tag{55}$$

Let $\sigma_1(x;k)$ be the distribution function that the polynomials $r_n(x;k)$ are orthogonal with respect to, and let $\chi_{\ 1}(x;k)$ be the Stieltjes transform of $\sigma_1(x;k)$. Asymptotic estimates (44), (53), and (55) allow us to write

$$\lim_{m\to\infty} \frac{a_{mk} S_{m-1}(x;a,b)}{R_m(x;a,b) - R_{m-2}(x;a,b)} = \frac{2}{\beta-\alpha}\, \left(1-\frac{\alpha}{\beta}\right)^{(b-a)}\, \frac{{}_2F_1\left(b-1, \frac{b-a}{2}; \frac{a+b}{2}; \frac{\alpha}{\beta}\right)}{{}_2F_1\left(a-1, \frac{a-b}{2}; \frac{a+b}{2}; \frac{\alpha}{\beta}\right)},$$

$x \notin [-1,1]$.

Now let x be replaced everywhere by $T_k(x)$, the Tchebyshev polynomial of the first kind. Then α, β are replaced by

$$\alpha_k, \beta_k = T_k(x) \pm \sqrt{T_k^2(x) - 1}. \tag{56}$$

The statement $x \notin [-1,1]$ becomes $T_k(x) \notin [-1,1]$, which is identical to $x \notin [-1,1]$. By (20) we have

$$\chi_{\ 1}(x;k) = \frac{2}{\beta_k - \alpha_k}\, U_{k-1}(x)\, \left(1-\frac{\alpha_k}{\beta_k}\right)^{(b-a)}\, \frac{{}_2F_1\left(b-1, \frac{b-a}{2}; \frac{a+b}{2}; \frac{\alpha_k}{\beta_k}\right)}{{}_2F_1\left(a-1, \frac{a-b}{2}; \frac{a+b}{2}; \frac{\alpha_k}{\beta_k}\right)}, \tag{57}$$

$x \notin [-1,1]$.

5.1. The companion polynomials. We next study the companion polynomials for the $R_n(x;a,b)$ defined by the recursion $xP_n(x;a,b) = \overline{B}_n P_{n+1}(x;a,b) + \overline{D}_n P_{n-1}(x;a,b)$, with the usual initial conditions, and where

$$\overline{B}_n = \frac{n+a}{2n+a+b-2}, \qquad \overline{D}_n = \frac{n+b-2}{2n+a+b-2}, \quad n \geq 0. \tag{58}$$

We see the relation $\overline{B}_n \overline{D}_{n+1} = B_n D_{n-1}$ holds for $n \geq 1$, so the $P_n(x)$ are companion polynomials to $R_n(x)$, but these are not the natural companion polynomials of $R_n(x)$. The above companion polynomials were selected since they will furnish some interesting results. We first note the identities

$$\begin{aligned} P_n(x;a,b) &= R_n(x,a,b-2), \\ P_n^{(1)}(x;a,b) &= R_n(x;a+1,b-1). \end{aligned} \tag{59}$$

From (29), we identify the coefficients θ_n, ϕ_n, ω_n as

(60)
$$\theta_n = \frac{(a+1)_n}{(b)_n}, \quad \phi_n = \frac{a+b+2(n-1)}{b-1}, \quad \omega_n = \frac{1-a}{a}\ \frac{a+b+2(n-1)}{b-1}, \quad n \geq 1.$$

Using identities (59) and (60), formulas (25) and (26) become

$$S_n(x;a,b) = \frac{(a+1)_n}{(b)_n}\ R_n(x;a+1,b-1), \quad n \geq 0, \tag{61}$$

$$\begin{aligned} &R_n(x;a,b) - R_{n-2}(x;a,b) = \frac{a+b+2(n-1)}{b-1}, \\ &\times\ \left[R_n(x;a,b-2) + \frac{1-a}{a}\ x\ R_{n-1}(x;a+1,b-1)\right], \quad n \geq 1. \end{aligned} \tag{62}$$

In terms of the associated ultraspherical polynomials, formula (62) becomes the mixed recursion relation

$$\begin{aligned} &C_n^{(\gamma)}(x;\beta) - C_{n-2}^{(\gamma)}(x;\beta) = \\ &\frac{2(n+\beta+\gamma-1)}{2\beta+\gamma-2}\ \left[C_n^{(\gamma)}(x;\beta-1) - \frac{\gamma x}{\gamma+1}\ C_{n-1}^{(\gamma+1)}(x;\beta-1)\right]. \end{aligned} \tag{63}$$

The case of $\gamma = 0$ are the ultraspherical polynomials, written $C_n^{\lambda}(x)$, where λ is replacing β. Formula (63) reduces to

$$C_n^{\lambda}(x) - C_{n-2}^{\lambda}(x) = \frac{\lambda+n-1}{\lambda-1}\ C_n^{\lambda-1}(x),$$

a known result. (See Rainville [17, p. 283, eq. (39)].) Charris and Ismail [7] use this relation, along with the known weight function for the ultraspherical polynomials, to invert the Stieltjes transform for the sieved ultraspherical polynomials of the first kind, and thereby obtain the orthogonality relation for these polynomials.

The derivation of formula (63) was done originally in DeSesa [9, pp. 139–142]. First an expression for $C_n^{(\gamma)}(x;\beta)$ in terms of the Legendre functions of the first and second kinds was used. (See Bustoz and Ismail [5].) Then properties of the Legendre functions established (63). This was used to invert the Stieltjes transform for the sieved associated ultraspherical polynomials of the first kind. (See DeSesa [9, Chapter 4].)

The Stieltjes transform $\chi_P(z)$ for the companion polynomials is easily obtained by making the substitution $b \to b-2$ in formula (48), as justified by identity (59). Similarly, $d\mu_p(x)$, the absolutely continuous component for the measure of orthogonality of the companion polynomials, is found by using the same substitution in (50).

5.2. The sieved polynomials of the first kind. For $r_n(x;k)$, the sieved polynomials of the first kind, where B_n and D_n are given by (54), we note first

$$B_0 = \frac{a}{a+b}, \qquad \overline{B}_0 = \frac{a}{a+b-2}, \qquad \overline{D}_1 = \frac{b-1}{a+b}. \tag{64}$$

Using (64) and the substitution $b \to b-2$ in (48) for $\chi_P(x)$ in Theorem 2, formula (34), we obtain a second expression for $\chi_1(x;k)$ given by

$$\chi_1(x;k) = \frac{2(b-1)U_{k-1}(x)\alpha_k F_1}{(a+b-2)F_2 + 2(1-a)T_k(x)\alpha_k F_1}, \quad x \notin [-1,1], \tag{65}$$

where

$$F_1 = {}_2F_1\left(\begin{matrix} a, \dfrac{a-b}{2}+1 \\ \dfrac{a+b}{2} \end{matrix}; \frac{\alpha_k}{\beta_k}\right), \qquad F_2 = {}_2F_1\left(\begin{matrix} a-1, \dfrac{a-b}{2}+1 \\ \dfrac{a+b}{2}-1 \end{matrix}; \frac{\alpha_k}{\beta_k}\right),$$

and α_k, β_k are given by (56). The equivalence of (65) and (57) may be established more directly from a sequence of Kummer transformations and contiguous parameter identities of the Gaussian hypergeometric function.

To obtain the formula for $d\sigma_1(x;k)$, substitute the expressions found for $d\mu_P(x)$ and $\chi_P(x)$ into Theorem 3, formula (36). Note that when x is replaced by $T_k(x)$ for $x \in (-1,1)$, $\theta = \cos^{-1}(x)$ is replaced by $k\theta$, $\alpha_k = e^{ik\theta}$, and $\beta_k = e^{-ik\theta}$. This yields

$$d\sigma_1(x;k) = \frac{\dfrac{2^{b-a}}{\pi}(a+b)(b-1)\ \dfrac{\Gamma^2\left(\dfrac{a+b}{2}\right)}{\Gamma(a)\Gamma(b)}\ |\sin k\theta|^{b-a-1}|U_{k-1}(x)|}{|(a+b-2)^2F_3 + 2[(a+b)(b-2)-(a+b-2)^2]T_k(x)e^{ik\theta}F_4|^2}\ dx, \tag{66}$$

$x = \cos\theta,$

where

$$F_3 = {}_2F_1\left(\begin{matrix} a-1, \dfrac{a-b}{2}+1 \\ \dfrac{a+b}{2}-1 \end{matrix}; e^{2ik\theta}\right), \ F_4 = {}_2F_1\left(\begin{matrix} a, \dfrac{a-b}{2}+1 \\ \dfrac{a+b}{2} \end{matrix}; e^{2ik\theta}\right).$$

REFERENCES

[1] W. Al-Salam, W. Allaway, and R. Askey, *A characterization of the continuous q-ultraspherical polynomials*, Canad. Math. Bull., 27 (1984), pp. 329–336.

[2] W. Al-Salam, W. Allaway, and R. Askey, *Sieved ultraspherical polynomials*, Trans. Amer. Math. Soc., 284 (1984), pp. 39–55.

[3] R. Askey and M. Ismail, *Recurrence Relations, Continued Fractions and Orthogonal Polynomials*, Mem. Amer. Math. Soc., AMS, Providence, RI, 1984.

[4] R. Askey and J. Wimp, *Associated Laguerre and Hermite polynomials*, Proc. Royal Soc. Edinburgh Sect. A, 96 (1984), pp. 15–37.

[5] J. Bustoz and M. Ismail, *The associated ultraspherical polynomials and their q-analogues*, Canad. J. Math., 34 (1982), pp. 718–736.

[6] J. Charris and M. Imail, *On sieved orthogonal polynomials* II: *Random walk polynomials*, Canad. J. Math., 38 (1986), pp. 397–415.

[7] J. Charris and M. Ismail, *On sieved orthogonal polynomials* V*: Sieved Pollaczek polynomials*, SIAM J. Math. Anal., 18 (1987), pp. 1177–1218.

[8] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[9] B. DeSesa, *Sieved Orthogonal Polynomials*, Ph.D. thesis, Drexel University, Philadelphia, PA, 1992.

[10] J. Geronimo and W. Van Assche, *Orthogonal polynomials on several intervals via a polynomial mapping*, Trans. Amer. Math. Soc., 308 (1988), pp. 559–581.

[11] M. Ismail, *On sieved orthogonal polynomials* III*: Orthogonality on several intervals*, Trans. Amer. Math. Soc., 294 (1986), pp. 89–111.
[12] M. Ismail, J. Letessier, and G. Valent, *Linear Birth and Death Models and Associated Laguerre and Meixner Polynomials*, J. Approx. Theory, 55 (1988), pp. 337–348.
[13] S. Karlin and J. McGregor, *The classification of birth and death processes*, Trans. Amer. Math. Soc., 86 (1957), pp. 366–400.
[14] S. Karlin and J. McGregor, *Random walks*, Illinois J. Math., 3 (1959), pp. 66–81.
[15] P. Nevai, *Orthogonal Polynomials*, Mem. Amer. Math. Soc. 213, AMS, Providence, RI, 1979.
[16] F. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[17] E. Rainville, *Special Functions*, Macmillan, New York, 1960.
[18] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, Math. Surveys 1, AMS, Providence, RI, 1963.
[19] H. S. Wall, *Analytic Theory of Continued Fractions*, D. Van Nostrand, New York, 1948.
[20] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 266–278, January 1998

015

FRAMES CONTAINING A RIESZ BASIS AND PRESERVATION OF THIS PROPERTY UNDER PERTURBATIONS*

PETER G. CASAZZA† AND OLE CHRISTENSEN‡

Abstract. Aldroubi has shown how one can construct any frame $\{g_i\}_{i=1}^{\infty}$ starting with one frame $\{f_i\}_{i=1}^{\infty}$,using a bounded operator U on the space of square summable sequences $\ell^2(N)$. We study the overcompleteness of the frames in terms of properties of U. We also discuss perturbation of frames in the sense that two frames are "close" if a certain operator is compact. In this way we obtain an equivalence relation with the property that frames in the same equivalence class have the same overcompleteness. On the other hand we show that perturbation in the Paley–Wiener sense does not have this property.

Finally we construct a frame which is norm bounded below but which does not contain a Riesz basis. The construction is based on the delicate difference between the unconditional convergence of the frame representation and the fact that a convergent series in the frame elements need not converge unconditionally.

Key words. frames, Riesz bases, perturbations, overcompleteness

AMS subject classifications. 42C99, 46C99

PII. S0036141095294250

1. Introduction. The introduction of frames for a Hilbert space $\mathcal{H}$ goes back to the paper [9] from 1952, where they are used in nonharmonic Fourier analysis. A frame is a family $\{f_i\}_{i\in I}$ of elements in $\mathcal{H}$ which can be considered as an "overcomplete basis": every element in $\mathcal{H}$ can be written as a linear combination of the frame elements f_i, with square integrable coefficients, which do not need to be unique. A natural theoretical question (which is also important for applications, e.g., representation of an operator using a basis) is how far frames are away from bases, i.e., one may ask questions like (1) does a frame contain a Riesz basis? (2) which conditions imply that a frame consists of a Riesz basis plus finitely many elements? (3) what happens with the overcompleteness if the frame elements are perturbed? The reason for the interest in Riesz bases and not just bases is that frames and Riesz bases are closely related: a Riesz bases is just a frame $\{f_i\}_{i=1}^{\infty}$, where the elements are ω-*independent*, i.e.,

$$\sum_{i\in I} c_i f_i = 0, \quad \{c_i\}_{i=1}^{\infty} \in \ell^2(I) \Rightarrow c_i = 0 \quad \forall i \in I.$$

Some answers have been found by Holub [10], who concentrates on the second question. Here we go one step further, in that we are mainly interested in frames which just contain a Riesz basis. For such frames one defines the *excess* as the number of elements one should take away to obtain a Riesz basis.

In the first part of the paper we apply a result of Aldroubi [1], explaining how one can map a frame onto another using a bounded operator U on ℓ^2. Our results

*Received by the editors November 1, 1995; accepted for publication (in revised form) November 19, 1996. This research was carried out during the first author's visit to Odense University in the spring of 1995. It was partially supported by the Danish Natural Science Foundation Council grant 9401598 and by NSF grant DMS-9201357.

http://www.siam.org/journals/sima/29-1/29425.html

†Department of Mathematics, University of Missouri, Columbia, MO 65211 (pete@casazza.math.missouri.edu).

‡Mathematical Institute, Building 303, Technical University of Denmark, 2800 Lyngby, Denmark (olechr@mat.dtu.dk).

concern the relation between the frames involved and properties of U. Independent of that we construct a norm-bounded frame not containing a Riesz basis.

In section 3 we concentrate on the third question. We introduce the concept "compact perturbation." This leads to an equivalence relation on the set of frames, with the property that frames in the same equivalence class have the same overcompleteness properties; this means that if a frame contains a Riesz basis then all members in the class contain a Riesz basis, and all those frames have the same excess.

Finally we show that perturbation in the Paley–Wiener sense [7] does not have this pleasant property.

2. Frames containing a Riesz basis. Let $\mathcal{H}$ be a separable Hilbert space. A family $\{f_i\}_{i\in I}$ is called a *frame* for $\mathcal{H}$ if

$$\exists A, B > 0 : \ A||f||^2 \leq \sum_{i\in I} |\langle f, f_i\rangle|^2 \leq B||f||^2 \quad \forall f \in \mathcal{H}.$$

A and B are called *frame bounds*. The frame is *tight* if we can choose $A = B$. A *Riesz basis* is a family of elements which is the image of an orthonormal basis by a bounded invertible operator. Frequently we will use an equivalent characterization [16]: $\{f_i\}_{i\in I}$ is a Riesz basis if there exist numbers $A, B > 0$ such that

$$A\sum |c_i|^2 \leq \left\|\sum c_i f_i\right\|^2 \leq B\sum |c_i|^2 \tag{1}$$

for all finite sequences $\{c_i\}$.

Also, a basis $\{f_i\}_{i\in I}$ is a Riesz basis if and only if it is unconditional (meaning that if $\sum c_i f_i$ converges for some coefficients $\{c_i\}$, then it actually converges unconditionally) and $0 < \inf_i ||f_i|| \leq \sup_i ||f_i|| < \infty$.

There is a close connection between frames and Riesz bases:

$$\{f_i\}_{i\in I} \text{ is a Riesz basis}$$

$$\Updownarrow$$

$$\left[\{f_i\}_{i\in I} \text{ is a frame and } \sum_{i\in I} c_i f_i = 0, \{c_i\}_{i\in I} \in \ell^2(I) \Rightarrow c_i = 0 \quad \forall i\right].$$

In words: a Riesz basis is a frame, where the elements are ω-independent. If $\{f_i\}_{i\in I}$ is a Riesz basis, then the numbers A, B appearing in (1) are actually frame bounds. If $\{f_i\}_{i\in I}$ is a frame (or if only the upper frame condition is satisfied) then we define the *preframe operator* as an operator from the space of square summable sequences with index set I into $\mathcal{H}$:

$$T : \ell^2(I) \to \mathcal{H}, \quad T\{c_i\} := \sum_{i\in I} c_i f_i.$$

The operator T is bounded. Composing T with its adjoint

$$T^* : \mathcal{H} \to \ell^2(I), \quad T^* f = \{\langle f, f_i\rangle\}_{i\in I},$$

we get the *frame operator*

$$S = TT^* : \mathcal{H} \to \mathcal{H}, \quad Sf := \sum_{i\in I} \langle f, f_i\rangle f_i,$$

which is a bounded and invertible operator. This immediately leads to the *frame decomposition*; every $f \in \mathcal{H}$ can be written as

$$f = \sum_{i \in I} \langle f, S^{-1} f_i \rangle f_i,$$

where the series converges unconditionally. So a frame has a property similar to a basis: every element in $\mathcal{H}$ can be written as a linear combination of the frame elements. For more information about basic properties of frames we refer to the original paper [9] and the research tutorial [12]. The main difference between a frame $\{f_i\}_{i \in I}$ and a basis is that a frame can be overcomplete, so it might happen that $f \in \mathcal{H}$ has a representation $f = \sum_{i \in I} c_i f_i$ for some coefficients c_i which are different from the *frame coefficients* $\langle f, S^{-1} f_i \rangle$. In applications one might wish not to have "too much redundancy." In that spirit Holub [10] discusses *near-Riesz bases*, i.e., frames $\{f_i\}_{i \in I}$ consisting of a Riesz basis $\{f_i\}_{i \in I - \sigma}$ plus finitely many elements $\{f_i\}_{i \in \sigma}$. The number of elements in σ is called the *excess*. Let us denote the kernel of the operator T by N_T. If $\{f_i\}_{i \in I}$ is a frame, then

$$\{f_i\}_{i \in I} \text{ is a near-Riesz basis,}$$

$$\Updownarrow$$

$$N_T \text{ has finite dimension,}$$

$$\Updownarrow$$

$$\{f_i\}_{i \in I} \text{ is unconditional.}$$

The first of the above bi-implications is due to Holub [10], who also proves the second under the assumption that the frame is norm bounded below. The generalization above is proved by the authors in [4]. If the conditions above are satisfied, then the excess is equal to $\dim(N_T)$.

If $\dim(N_T) = \infty$, two things can happen: $\{f_i\}_{i \in I}$ consists of a Riesz basis plus infinitely many elements (in which case we will say that $\{f_i\}_{i \in I}$ has infinite excess) or $\{f_i\}_{i \in I}$ does not contains a Riesz basis at all. In the present paper we concentrate on frames which contain a Riesz basis. Every frame can be mapped onto such a frame (in fact, onto an arbitrary frame) using a construction of Aldroubi [1], which we now shortly describe.

For convenience, we will index our frames by the natural numbers. Let $\{f_i\}_{i=1}^{\infty}$ be a frame and $U : \ell^2(\mathbf{N}) \to \ell^2(\mathbf{N})$ a bounded operator. Let $\{u_{i,j}\}_{i,j \in N}$ be the matrix for U with respect to some basis. Define the family $\{g_i\}_{i=1}^{\infty} \in \mathcal{H}$ by

$$g_i = \sum_{j=1}^{\infty} u_{i,j} f_j. \tag{2}$$

By an abuse of notation we will sometimes write $\{g_i\}_{i=1}^{\infty} = U\{f_i\}_{i=1}^{\infty}$. A result of Aldroubi (differently formulated) states that

$$\{g_i\}_{i=1}^{\infty} \text{ is a frame} \Leftrightarrow \exists \gamma > 0 : ||UT^* f|| \geq \gamma \cdot ||T^* f|| \ \ \forall f \in \mathcal{H}.$$

It is important that *every* frame $\{g_i\}_{i=1}^{\infty}$ can be generated in this way; i.e., given the frame $\{g_i\}_{i=1}^{\infty}$ we just have to find the operator U mapping $\{f_i\}_{i=1}^{\infty}$ to $\{g_i\}_{i=1}^{\infty}$.

In connection with Aldroubi's construction there are (at least) two natural questions related to Holub's work: how is the excess of $\{g_i\}_{i=1}^\infty$ related to that of $\{f_i\}_{i=1}^\infty$, and which conditions imply that $\{g_i\}_{i=1}^\infty$ actually is a Riesz basis? We shall give answers to both questions in this section.

The definition of $\{g_i\}_{i=1}^\infty$ immediately shows that

$$\{\langle g_i, f\rangle\} = U\{\langle f_i, f\rangle\} \quad \forall f \in \mathcal{H};$$

this is true whether or not $\{g_i\}_{i=1}^\infty$ builds a frame. The formula leads to an expression for the preframe operator associated with $\{g_i\}_{i=1}^\infty$. We let U^T denote the transpose of U and $\overline{U}$ be the operator corresponding to the matrix where all entries in the matrix of U are complex conjugated. It is easy to prove that

$$\sum_{i=1}^\infty c_i g_i = TU^T\{c_i\}_{i=1}^\infty \quad \forall \{c_i\}_{i=1}^\infty \in \ell^2(\mathbf{N}).$$

So if $\{g_i\}_{i=1}^\infty$ contains a Riesz basis, then its excess is equal to $\dim(N_{TU^T})$. For the calculation of this number we need a lemma, the proof of which we leave to the reader. Corresponding to an operator V we denote its range by R_V.

LEMMA 2.1. *Let X, Y be vector spaces and $V : X \to Y$ a linear mapping. Given a subspace $Z \subseteq Y$, define $V^{-1}(Z) := \{x \in X \mid Vx \in Z\}$. Then*

$$\dim(V^{-1}(Z)) = \dim(Z \cap R_V) + \dim(N_V).$$

THEOREM 2.2. $\dim(N_{TU^T}) = \dim(R_{U^T} \cap N_T) + \dim(R_U^\perp)$.

Proof. Theorem 2.2 is an easy consequence of Lemma 2.1 and the calculation

$$\{\{c_i\}_{i=1}^\infty \mid TU^T\{c_i\}_{i=1}^\infty = 0\} = \{\{c_i\}_{i=1}^\infty \mid U^T\{c_i\}_{i=1}^\infty \in N_T\} = (U^T)^{-1}(N_T). \qquad \square$$

So if $\{g_i\}_{i=1}^\infty$ actually is a frame containing a Riesz basis, then Theorem 2.2 gives a recipe for calculation of the excess. In particular, if $\{f_i\}_{i=1}^\infty$ is a near-Riesz basis and R_U has finite codimension, then $\{g_i\}_{i=1}^\infty$ is also a near-Riesz basis. Observe that in the special case where $\{f_i\}_{i=1}^\infty$ is a Riesz basis, the excess of $\{g_i\}_{i=1}^\infty$ is equal to $\dim({R_U}^\perp) = \dim(N_{U^*})$.

Example 1. Let $\{f_i\}_{i=1}^\infty = \{e_i\}_{i=1}^\infty$ be an orthonormal basis for $\mathcal{H}$ and define

$$g_1 := e_1, \qquad g_i = e_{i-1} + \frac{1}{i}e_i, \quad i \geq 2.$$

According to (2), we have

$$U = \begin{pmatrix} 1 & 0 & 0 & . & . & . \\ 1 & 1/2 & 0 & 0 & . & . \\ 0 & 1 & 1/3 & 0 & 0 & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{pmatrix},$$

which certainly defines a bounded operator from $\ell^2(N)$ into $\ell^2(N)$. Since $\{f_i\}_{i=1}^\infty$ is an orthonormal basis, $R_{T^*} = \ell^2(N)$, so $\{g_i\}_{i=1}^\infty$ is a frame if and only if

$$\exists \gamma > 0 : \ ||U\{c_i\}_{i=1}^\infty|| \geq \gamma \cdot ||\{c_i\}_{i=1}^\infty|| \quad \forall \{c_i\}_{i=1}^\infty \in \ell^2(N).$$

But

$$||U\{c_i\}_{i=1}^{\infty}|| = \left\| \left(c_1, c_1 + \frac{1}{2}c_2, c_2 + \frac{1}{3}c_3, ... \right) \right\|$$

$$\geq ||(c_1, c_1, c_2, c_3, ...)|| - \left\| \left(0, \frac{c_2}{2}, \frac{c_3}{3}, ... \right) \right\| \geq \frac{1}{2}||\{c_i\}_{i=1}^{\infty}|| \quad \forall \{c_i\}_{i=1}^{\infty} \in \ell^2(N).$$

Now, since

$$U^* = \begin{pmatrix} 1 & 1 & 0 & . & . & . \\ 0 & 1/2 & 1 & 0 & . & . \\ 0 & 0 & 1/3 & 1 & 0 & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{pmatrix},$$

$N_{U^*} = \text{span}\{(1, -1, \frac{1}{2!}, \frac{-1}{3!}, ...)\}$. We conclude that $\{g_i\}_{i=1}^{\infty}$ has excess one.

Concerning Riesz bases we have another result, which can be proved by the interested reader.

PROPOSITION 2.3. *$\{g_i\}_{i=1}^{\infty}$ is a Riesz basis $\Leftrightarrow$ $\overline{U} : R_{T^*} \to \ell^2(N)$ is surjective.*

More generally one may wish that the frame at least contain a Riesz basis. As shown in [6], this is the case for a *Riesz frame*, which is a frame with the property that every subfamily is a frame for its closed linear span, with a common lower bound (see [4] for an extension.)

It is easy to construct a frame which does not contain a Riesz basis if one allows a subsequence of the frame elements to converge to 0 in norm. We now present an example showing that the same can be the case for a frame which is norm bounded below. Our approach is complementary to a work by Seip [14], who proves that there exist frames of complex exponentials for $L^2(-\pi, \pi)$ which do not contain a Riesz basis. While Seip relies on the theory for sampling and interpolation our approach is more elementary, just using functional analysis. Furthermore our construction puts focus on a different point, namely the difference between convergence and unconditional convergence of an expansion in the frame elements.

PROPOSITION 2.4. *There exists a tight frame for $\mathcal{H}$ which is norm bounded below but which does not contain a Riesz basis.*

The proof needs several lemmas, so let us shortly explain the basic idea. As we have seen, $\sum_{i \in I} c_i f_i$ converges unconditionally for every set of frame coefficients $\{c_i\}_{i \in I}$. But nothing guarantees that convergence of $\sum_{i \in I} c_i f_i$ implies unconditional convergence for general coefficients $\{c_i\}_{i \in I}$. Our proof consists of a construction of a frame where no total subset is unconditional, and hence not a Riesz basis. Technically the first step is to decompose $\mathcal{H}$ into a direct sum of finite dimensional subspaces of increasing dimension. The idea behind the proof might be useful in other situations as well.

LEMMA 2.5. *Let $\{e_i\}_{i=1}^{n}$ be an orthonormal basis for a finite dimensional space $\mathcal{H}_n$. Define*

$$f_j = e_j - \frac{1}{n} \sum_{i=1}^{n} e_i \quad \text{for } j = 1, \ldots, n,$$

$$f_{n+1} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i.$$

Then

$$\sum_{j=1}^{n+1} |\langle f, f_j\rangle|^2 = ||f||^2 \quad \forall f \in \mathcal{H}_n.$$

Proof. Given $f \in \mathcal{H}_n$, write $f = \sum_{i=1}^n a_i e_i,\ a_i = \langle f, e_i\rangle$. If we let P denote the orthogonal projection onto the unit vector $\frac{1}{\sqrt{n}} \sum_{i=1}^n e_i$, then

$$Pf = \frac{1}{n} \left\langle f, \sum_{i=1}^n e_i \right\rangle \sum_{i=1}^n e_i = \frac{\sum_{i=1}^n a_i}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i.$$

Therefore

$$||Pf||^2 = \frac{|\sum_{i=1}^n a_i|^2}{n} = |\langle f, f_{n+1}\rangle|^2.$$

Also

$$||(I-P)f||^2 = ||f - Pf||^2 = \left\| \sum_{i=1}^n a_i e_i - \frac{\sum_{j=1}^n a_j}{n} \sum_{i=1}^n e_i \right\|^2$$
$$= \left\| \sum_{i=1}^n \left(a_i - \frac{\sum_{j=1}^n a_j}{n} \right) e_i \right\|^2 = \sum_{i=1}^n \left| a_i - \frac{\sum_{j=1}^n a_j}{n} \right|^2 = \sum_{i=1}^n |\langle f, f_i\rangle|^2.$$

Putting the two results together we obtain

$$||f||^2 = ||Pf||^2 + ||(I-P)f||^2 = \sum_{i=1}^{n+1} |\langle f, f_i\rangle|^2,$$

and the proof is complete. □

Given a sequence $\{g_i\}_{i\in I} \subseteq \mathcal{H}$, its *unconditional basis constant* is defined as the number

$$\sup \left\{ \left\| \sum_{i\in I} \sigma_i c_i g_i \right\| \Big| \left\| \sum_{i\in I} c_i g_i \right\| = 1 \text{ and } \sigma_i = \pm 1 \quad \forall i \right\}.$$

As shown in [15], a total family $\{g_i\}_{i\in I}$ consisting of nonzero elements is an unconditional basis for $\mathcal{H}$ if and only if it has finite unconditional basis constant.

LEMMA 2.6. *Define* $\{f_1, ..., f_{n+1}\}$ *as in Lemma* 2.5. *Any subset of* $\{f_1, f_2, ..., f_{n+1}\}$ *which spans* $\mathcal{H}_n$ *has unconditional basis constant greater than or equal to* $\sqrt{n-1} - 1$.

Proof. Since $\sum_{i=1}^n f_i = 0$, any subset of $\{f_1, ..., f_{n+1}\}$ which spans $\mathcal{H}_n$ must contain $n-1$ elements from $\{f_1, ..., f_n\}$ plus f_{n+1}. By the symmetric construction it is enough to consider the family $\{f_1, ..., f_{n-1}, f_{n+1}\}$. We have

$$\left\| \sum_{i=1}^{n-1} f_i \right\| = \left\| \sum_{i=1}^{n-1} e_i - \frac{n-1}{n} \sum_{i=1}^n e_i \right\|$$
$$= \left\| (1 - \frac{n-1}{n}) \sum_{i=1}^{n-1} e_i - \frac{n-1}{n} e_n \right\|$$
$$= \left\| \frac{1}{n} \sum_{i=1}^{n-1} e_i - \frac{n-1}{n} e_n \right\|$$
$$= \sqrt{\frac{n-1}{n^2} + \frac{(n-1)^2}{n^2}}$$
$$= \frac{1}{n} \sqrt{n(n-1)} \leq 1.$$

Now consider $||\sum_{i=1}^{n-1}(-1)^n f_i||$; if n is odd this number is equal to $||\sum_{i=1}^{n-1}(-1)^n e_i|| = \sqrt{n-1}$, and if n is even it is equal to

$$\left\|\sum_{i=1}^{n-1}(-1)^i e_i - \frac{1}{n}\sum_{i=1}^{n} e_i\right\| \geq \left\|\sum_{i=1}^{n-1}(-1)^i e_i\right\| - \left\|\frac{1}{n}\sum_{i=1}^{n} e_i\right\| \geq \sqrt{n-1} - \frac{\sqrt{n}}{n} \geq \sqrt{n-1} - 1.$$

That is, in all cases,

$$\left\|\sum_{i=1}^{n-1}(-1)^n f_i\right\| \geq \sqrt{n-1} - 1.$$

Combining this with the norm estimate $||\sum_{i=1}^{n-1} f_i|| \leq 1$, it follows that the unconditional basis constant of $\{f_1, ..., f_{n-1}\}$ is greater than or equal to $\sqrt{n-1}-1$, so clearly the same is true for $\{f_1, ..., f_{n-1}, f_{n+1}\}$. □

Now we are ready to do the construction for Proposition 2.4. Let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis for $\mathcal{H}$ and define

$$\mathcal{H}_n := \text{span}\{e_{\frac{(n-1)n}{2}+1}, e_{\frac{(n-1)n}{2}+2}, ..., e_{\frac{(n-1)n}{2}+n}\}.$$

So $\mathcal{H}_1 = \text{span}\{e_1\}$, $\mathcal{H}_2 = \text{span}\{e_2, e_3\}, \mathcal{H}_3 = \text{span}\{e_4, e_5, e_6\}, ...$ By construction,

$$\mathcal{H} = \left(\sum_{n=1}^{\infty}\bigoplus \mathcal{H}_n\right)_{\mathcal{H}}.$$

That is, $g \in \mathcal{H} \Leftrightarrow g = \sum_{n=1}^{\infty} g_n, g_n \in \mathcal{H}_n$, and $||g||^2 = \sum_{n=1}^{\infty} ||g_n||^2$. We refer to [13] for details about such decompositions. For each space $\mathcal{H}_n$ we construct the sequence $\{f_i^n\}_{i=1}^{n+1}$ as in Lemma 2.5, starting with the orthonormal basis $\{e_{\frac{(n-1)n}{2}+1}, ..., e_{\frac{(n-1)n}{2}+n}\}$. Specifically, given $n \in N$,

$$f_i^n = e_{\frac{(n-1)n}{2}+i} - \frac{1}{n}\sum_{j=1}^{n} e_{\frac{(n-1)n}{2}+j}\ , \quad 1 \leq i \leq n,$$

$$f_{n+1}^n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} e_{\frac{(n-1)n}{2}+j}.$$

LEMMA 2.7. *$\{f_i^n\}_{i=1,n=1}^{n+1,\infty}$ is a frame for $\mathcal{H}$, with bounds $A = B = 1$.*

Proof. Write $g \in \mathcal{H}$ as $g = \sum_{n=1}^{\infty} g_n$, $g_n \in \mathcal{H}_n$. Given $n \in N$ it is clear that

$$\langle g, f_i^n\rangle = \langle g_n, f_i^n\rangle \text{ for } i = 1, ..., n+1.$$

From this calculation it follows that

$$\sum_{n=1}^{\infty}\sum_{i=1}^{n+1} |\langle g, f_i^n\rangle|^2 = \sum_{n=1}^{\infty}\sum_{i=1}^{n+1} |\langle g_n, f_i^n\rangle|^2 = \sum_{n=1}^{\infty} ||g_n||^2 = ||g||^2,$$

where we have used Lemma 2.5. □

LEMMA 2.8. *No subsequence of $\{f_i^n\}_{i=1,n=1}^{n+1,\infty}$ is a Riesz basis for $\mathcal{H}$.*

Proof. Any subsequence of $\{f_i^n\}_{i=1,n=1}^{n+1,\infty}$ which spans $\mathcal{H}$ must contain n elements from $\{f_i^n\}_{i=1}^{n+1}$ and so by Lemma 2.6, its unconditional basis constant is greater than or equal to $\sqrt{n-1}-1$ for every n. That is, the unconditional basis constant is infinite, hence the subsequence cannot be an unconditional basis for $\mathcal{H}$. □

Lemmas 2.7 and 2.8 prove Proposition 2.4. It would be interesting to determine whether Proposition 2.4 still holds if one only considers classes of frames with a special structure, for example Weyl–Heisenberg frames, wavelet frames, or frames consisting of translates of a single function.

Remark. Corresponding to a subfamily $\{f_i\}_{i=1}^n$ of a frame $\{f_i\}_{i=1}^\infty$, we define the frame operator by

$$S_n : \operatorname{span}\{f_i\}_{i=1}^n \to \operatorname{span}\{f_i\}_{i=1}^n, \quad S_n f = \sum_{i=1}^n \langle f, f_i\rangle f_i.$$

The orthogonal projection of $\mathcal{H}$ onto $\operatorname{span}\{f_i\}_{i=1}^n$ is given by

$$P_n f = \sum_{i=1}^n \langle f, {S_n}^{-1} f_i\rangle f_i.$$

According to [5, 6], we say that the *projection method* works if

$$\langle f, {S_n}^{-1} f_i\rangle \to \langle f, S^{-1} f_i\rangle \text{ for } n \to \infty \ \forall f \in \mathcal{H}, \ \forall i \in N.$$

The "block structure" of the frame $\{f_i^n\}_{i=1,n=1}^{n+1,\infty}$ constructed here shows that the projection method can be used. As shown in [6] the method can also be used for every Riesz basis. The more general questions whether a frame contains a Riesz basis and whether the projection method works do not seem to be strongly related.

3. Excess preserving perturbation. At several places in the following we need results for perturbation of frames and Riesz bases. We denote the frames by $\{f_i\}_{i=1}^\infty, \{g_i\}_{i=1}^\infty$, usually with the convention that $\{f_i\}_{i=1}^\infty$ is the frame we begin with, and $\{g_i\}_{i=1}^\infty$ is the perturbed family. Common to all these results is that they can be formulated using the *perturbation operator* K mapping a sequence $\{c_i\}$ of numbers to $\sum c_i(f_i - g_i)$.

THEOREM 3.1. *Let $\{f_i\}_{i=1}^\infty, \{g_i\}_{i=1}^\infty \subseteq \mathcal{H}$.*

(a) *If $\{f_i\}_{i=1}^\infty$ is a frame for $\mathcal{H}$ and K is compact as an operator from $\ell^2(N)$ into $\mathcal{H}$, then $\{g_i\}_{i=1}^\infty$ is a frame for its closed linear span.*

(b) *Suppose $\{f_i\}_{i=1}^\infty$ is a frame for $\mathcal{H}$ with bounds A, B. If there exist numbers $\lambda, \mu \geq 0$ such that $\lambda + \frac{\mu}{\sqrt{A}} < 1$ and*

$$\left\| \sum c_i(f_i - g_i) \right\| \leq \lambda \cdot \left\| \sum c_i f_i \right\| + \mu \sqrt{\sum |c_i|^2}$$

for all finite sequences $\{c_i\}$, then $\{g_i\}_{i=1}^\infty$ is a frame for $\mathcal{H}$ with bounds $A(1 - (\lambda + \frac{\mu}{\sqrt{A}}))^2, B(1 + \lambda + \frac{\mu}{\sqrt{B}})^2$.

(c) *If $\{f_i\}_{i=1}^\infty$ is a Riesz basis for $\overline{\operatorname{span}}\{f_i\}_{i=1}^\infty$ and the perturbation condition in* (b) *is satisfied, then $\{g_i\}_{i=1}^\infty$ is a Riesz basis for $\overline{\operatorname{span}}\{g_i\}_{i=1}^\infty$.*

For the proofs we refer to [6, 7, 8]. As an easy consequence of (a) we have the following.

COROLLARY 3.2. *If $\{f_i\}_{i=1}^\infty$ is a frame and $\sigma \subseteq N$ is finite, then $\{f_i\}_{i \in N-\sigma}$ is a frame for $\overline{\operatorname{span}}\{f_i\}_{i \in N-\sigma}$.*

Our next result connects Theorem 3.1 with the question about overcompleteness of the involved frames.

THEOREM 3.3. *Suppose that $\{f_i\}_{i=1}^\infty$ is a frame containing a Riesz basis, that $\{g_i\}_{i=1}^\infty$ is total, and that K is compact as a mapping from $l^2(N)$ into $\mathcal{H}$. Then $\{g_i\}_{i=1}^\infty$ is a frame for $\mathcal{H}$ containing a Riesz basis, and the frames $\{f_i\}_{i=1}^\infty$ and $\{g_i\}_{i=1}^\infty$ have the same excess.*

Proof. First assume that $\{f_i\}_{i=1}^{\infty}$ has finite excess equal to n. By changing the index set we may write $\{f_i\}_{i=1}^{\infty} = \{f_i\}_{i=1}^{n} \cup \{f_i\}_{i=n+1}^{\infty}$, where $\{f_i\}_{i=n+1}^{\infty}$ is a Riesz basis for $\mathcal{H}$. Let A be a lower frame bound for $\{f_i\}_{i=n+1}^{\infty}$ and choose $\mu < \sqrt{A}$. By compactness there exists a number $m > n$ such that

$$\left\| \sum_{i=m+1}^{\infty} c_i(f_i - g_i) \right\| \leq \mu \sqrt{\sum_{i=m+1}^{\infty} |c_i|^2}$$

for all sets of sequences $\{c_i\} \subseteq l^2(N)$. So by the remark after Theorem 3.1, $\{g_i\}_{i=m+1}^{\infty}$ is a Riesz basis for $\overline{\text{span}}\{g_i\}_{i=m+1}^{\infty}$. If we define the operator T on $\mathcal{H}$ by

$$Tf_i = f_i, \ n < i \leq m, \quad Tf_i = g_i, \quad i \geq m+1$$

(extended by linearity), then we have an invertible operator on $\mathcal{H}$. The argument is that every $f \in \mathcal{H}$ has a representation $f = \sum_{i=n+1}^{\infty} c_i f_i$, leading to

$$||(I-T)f|| = \left\| \sum_{i=m+1}^{\infty} c_i(f_i - g_i) \right\|$$

$$\leq \mu \sqrt{\sum_{i=m+1}^{\infty} |c_i|^2} \leq \frac{\mu}{\sqrt{A}} \left\| \sum_{i=m+1}^{\infty} c_i f_i \right\| \leq \frac{\mu}{\sqrt{A}} \cdot ||f||.$$

As a consequence,

$$\text{codim}(\overline{\text{span}}\{g_i\}_{i=m+1}^{\infty}) = \ \text{codim}(\overline{\text{span}}\{f_i\}_{i=m+1}^{\infty}) = m - n.$$

Take $m-n$ independent elements $\{g_{i_k}\}_{k=1}^{m-n}$ outside $\overline{\text{span}}\{g_i\}_{i=m+1}^{\infty}$. Then $\{g_{i_k}\}_{k=1}^{m-n} \cup \{g_i\}_{i=m+1}^{\infty}$ is a frame for $\overline{\text{span}}\{\{g_{i_k}\}_{k=1}^{m-n} \cup \{g_i\}_{i=m+1}^{\infty}\} = \mathcal{H}$, since only finitely many elements have been taken away from the frame $\{g_i\}_{i=1}^{\infty}$. If $\sum_{k=1}^{m-n} c_k g_{i_k} + \sum_{i=m+1}^{\infty} c_i g_i = 0$ now, then all coefficients are zero; first,

$$\sum_{k=1}^{m-n} c_k g_{i_k} = -\sum_{i=m+1}^{\infty} c_i g_i = 0$$

(if the sums were not equal to zero we could delete an element g_{i_k} and still have a frame for $\mathcal{H}$ contradicting the fact that $\text{codim}(\overline{\text{span}}\{g_i\}_{i=m+1}^{\infty}) = m - n$) and since $\{g_{i_k}\}_{k=1}^{m-n}$ is an independent set and $\{g_i\}_{i=m+1}^{\infty}$ a Riesz basis, all coefficients must be zero. So $\{g_{i_k}\}_{k=1}^{m-n} \cup \{g_i\}_{i=m+1}^{\infty}$ is a Riesz basis, i.e., $\{g_i\}_{i=1}^{\infty}$ also has excess n.

Now suppose that $\{f_i\}_{i=1}^{\infty}$ has infinite excess. Let $\{f_i\}_{i\in I}$ be a subset which is a Riesz basis. Then the corresponding set $\{g_i\}_{i\in I}$ spans a space of finite codimension, i.e., $\text{codim}(\overline{\text{span}}\{g_i\}_{i\in I}) < \infty$. This follows by the same compactness argument as we used in the finite excess case, which shows that there exist finitely many $f_i, i \in I$ with the property that if we take them away then we obtain a family which spans a space with the same codimension as the corresponding space of g_i's. Now take a finite family $\{g_i\}_{i\in J}$ such that $\{g_i\}_{i\in I\cup J}$ is total. Since $\{f_i\}_{i\in I\cup J}$ is a frame with finite excess, the finite excess result gives that $\{g_i\}_{i\in I\cup J}$ is a frame containing a Riesz basis, implying that $\{g_i\}_{i=1}^{\infty}$ has infinite excess. □

We can express the result in the following way: define an equivalence relation $\sim$ on the set of frames for $\mathcal{H}$ by

$$\{f_i\}_{i=1}^{\infty} \sim \{g_i\}_{i=1}^{\infty} \Leftrightarrow K \text{ is compact as an operator from } l^2(N) \text{ into } \mathcal{H}.$$

The equivalence relation partitions the set of frames into equivalence classes. If a frame contains a Riesz basis, then every frame in its equivalence class contains a Riesz basis, and the frames have the same excess.

Let us go back to Theorem 3.3. If $\{g_i\}_{i=1}^{\infty}$ is not total, we still know (from Theorem 3.1) that $\{g_i\}_{i=1}^{\infty}$ is a frame for its closed span. By checking the proof of Theorem 3.3 we obtain the following corollary.

COROLLARY 3.4. *Suppose that $\{f_i\}_{i=1}^{\infty}$ is a frame containing a Riesz basis and that K is compact as a mapping from $\ell^2(N)$ into $\mathcal{H}$. Then $\{g_i\}_{i=1}^{\infty}$ is a frame for its closed linear span, and it contains a Riesz basis for this space. The excess referring to $\overline{\text{span}}\{g_i\}_{i=1}^{\infty}$ is equal to the excess of $\{f_i\}_{i=1}^{\infty}$ referring to $\mathcal{H}$, plus the dimension of the orthogonal complement of $\overline{\text{span}}\{g_i\}_{i=1}^{\infty}$ in $\mathcal{H}$.*

Now we want to study the excess property of perturbations in the sense of Theorem 3.1 (b). We need a result, which might be interesting in itself. To motivate it, consider a near-Riesz basis $\{f_i\}_{i=1}^{\infty}$ containing a Riesz basis $\{f_i\}_{i\in I}$. Unfortunately, the lower bound for $\{f_i\}_{i\in I}$ can be arbitrarily small compared to the lower bound A of $\{f_i\}_{i=1}^{\infty}$. Our result states that if we are willing to delete sufficiently (still finitely) many elements, then we can obtain a family which is a Riesz basis for its closed span and which has a lower bound as close to A as we want.

PROPOSITION 3.5. *Let $\{f_i\}_{i=1}^{\infty}$ be a near-Riesz basis with lower bound A. Given $\epsilon > 0$, there exists a finite set $J \subseteq N$ such that $\{f_i\}_{i\in N-J}$ is a Riesz basis for its closed span, with lower bound $A-\epsilon$.*

Proof. As in the proof of Theorem 3.3, write $\{f_i\}_{i=1}^{\infty} = \{f_i\}_{i=1}^{n} \cup \{f_i\}_{i=n+1}^{\infty}$, where $\{f_i\}_{i=n+1}^{\infty}$ is a Riesz basis for $\mathcal{H}$. Let $d(\cdot,\cdot)$ denote the distance inside $\mathcal{H}$ (i.e., $d(f,E) = \inf_{g\in E}||f-g||$ for $f \in \mathcal{H}, E \subseteq \mathcal{H}$) and choose a number $m > n$ such that

$$d(f_j,\ \text{span}\{f_i\}_{i=n+1}^{m}) < \sqrt{\frac{\epsilon}{n}},\quad j = 1, ..., n.$$

We want to show that $\{f_i\}_{i=m+1}^{\infty}$ is a Riesz basis for its closed span, with lower bound $A-\epsilon$. Let P denote the orthogonal projection onto $\overline{\text{span}}\{f_i\}_{i=n+1}^{m}$. Since $||\sum c_i f_i|| \geq ||\sum c_i(I-P)f_i||$ for all sequences, it suffices to show that $\{(I-P)f_i\}_{i=m+1}^{\infty}$ satisfies the lower Riesz basis condition with bound $A-\epsilon$. Let $f \in (I-P)\mathcal{H}$. Then

$$\sum_{i=m+1}^{\infty} |\langle f, (I-P)f_i\rangle|^2 = \sum_{i=1}^{\infty} |\langle f, (I-P)f_i\rangle|^2 - \sum_{i=1}^{n} |\langle f, (I-P)f_i\rangle|^2$$
$$\geq A||f||^2 - \sum_{i=1}^{n} ||f||^2 \cdot ||(I-P)f_i||^2 \geq (A-\epsilon)||f||^2.$$

To conclude that $\{(I-P)f_i\}_{i=m+1}^{\infty}$ has lower Riesz basis bound $A-\epsilon$, we only have to show that $\{(I-P)f_i\}_{i=m+1}^{\infty}$ is ω-independent. But if $\sum_{i=m+1}^{\infty} c_i(I-P)f_i = 0$, then $\sum_{i=m+1}^{\infty} c_i f_i = P\sum_{i=m+1}^{\infty} c_i f_i$, implying that both sides are equal to zero, since $P\sum_{i=m+1}^{\infty} c_i f_i \in \text{span}\{f_i\}_{i=n+1}^{m}$ and $\{f_i\}_{i=n+1}^{\infty}$ is ω-independent. Therefore $c_i = 0$ for all i. ∎

THEOREM 3.6. *Let $\{f_i\}_{i=1}^{\infty}$ be a frame for $\mathcal{H}$ with bounds A, B. Let $\{g_i\}_{i=1}^{\infty} \subseteq \mathcal{H}$ and assume that there exist $\lambda, \mu \geq 0$ such that $\lambda + \frac{\mu}{\sqrt{A}} < 1$ and*

$$\left\|\sum c_i(f_i - g_i)\right\| \leq \lambda \cdot \left\|\sum c_i f_i\right\| + \mu \cdot \sqrt{\sum |c_i|^2}$$

for all finite sequences $\{c_i\}$. Then

$$\{f_i\}_{i=1}^{\infty} \text{ is a near-Riesz basis} \Leftrightarrow \{g_i\}_{i=1}^{\infty} \text{ is a near-Riesz basis},$$

in which case $\{f_i\}_{i=1}^{\infty}$ and $\{g_i\}_{i=1}^{\infty}$ have the same excess.

Proof. First assume that $\{f_i\}_{i=1}^{\infty}$ is a near-Riesz basis with excess n. Let m be chosen as in the proof of Proposition 3.5, corresponding to an ϵ satisfying the condition $\lambda + \frac{\mu}{\sqrt{A-\epsilon}} < 1$. Let Q denote the orthogonal projection onto $\overline{\text{span}}\{f_i\}_{i=m+1}^{\infty}$. Then every element $f \in \mathcal{H}$ can be written $f = (I-Q)f + Qf = (I-Q)f + \sum_{i=m+1}^{\infty} c_i f_i$ for some coefficients c_i. Now define an operator $T : \mathcal{H} \to \mathcal{H}$ by

$$Tf = f, \ f \in \overline{\text{span}}\{f_i\}_{i=m+1}^{\infty}{}^{\perp}, \ \ Tf_i = g_i, \ i \geq m+1.$$

T is bounded. Given $f \in \mathcal{H}$ we choose a representation as above. Then

$$||(I-T)f|| = \left\| \sum_{i=m+1}^{\infty} c_i(f_i - g_i) \right\| \leq \lambda \cdot \left\| \sum_{i=m+1}^{\infty} c_i f_i \right\| + \mu \cdot \sqrt{\sum_{i=m+1}^{\infty} |c_i|^2}$$

$$\leq (\lambda + \frac{\mu}{\sqrt{A-\epsilon}}) \left\| \sum_{i=m+1}^{\infty} c_i f_i \right\| = \left(\lambda + \frac{\mu}{\sqrt{A-\epsilon}} \right) ||Qf|| \leq \left(\lambda + \frac{\mu}{\sqrt{A-\epsilon}} \right) ||f||.$$

It follows that T is an isomorphism of $\mathcal{H}$ onto $\mathcal{H}$. So $\{g_i\}_{i=m+1}^{\infty}$ is a Riesz basis for its closed span, and

$$\dim(\overline{\text{span}}\{g_i\}_{i=m+1}^{\infty}{}^{\perp}) = \dim(\overline{\text{span}}\{f_i\}_{i=m+1}^{\infty}{}^{\perp}).$$

As a consequence, $\{f_i\}_{i=1}^{\infty}$ and $\{g_i\}_{i=1}^{\infty}$ have the same excess.

Now assume that $\{g_i\}_{i=1}^{\infty}$ is a near-Riesz basis. By reindexing we may again assume that $\{g_i\}_{i=n+1}^{\infty}$ is a Riesz basis for $\mathcal{H}$. Define a bounded operator $W : \mathcal{H} \to \mathcal{H}$ by $Wf := \sum_{i=1}^{\infty} \langle f, S^{-1}f_i \rangle g_i$. Then as in the original proof from [7], one proves that W is an isomorphism of $\mathcal{H}$ onto $\mathcal{H}$. If we define $W_n : \mathcal{H} \to \mathcal{H}$ by $W_n f = \sum_{i=n+1}^{\infty} \langle f, S^{-1}f_i \rangle g_i$, then this operator has a range with finite codimension in $\mathcal{H}$, which we will write as

$$\text{codim}_{\mathcal{H}}(R_{W_n}) < \infty.$$

Now let $\{e_i\}_{i=1}^{\infty}$ be the natural basis for $l^2(N)$; i.e., e_i is the sequence with 1 in the ith entry, otherwise 0. There exists a bounded invertible operator $V : \mathcal{H} \to \overline{\text{span}}\{e_i\}_{i=n+1}^{\infty}$ such that $Vg_i = e_i$ for $i \geq n+1$, and clearly

$$\text{codim}_{\overline{\text{span}}\{e_i\}_{i=n+1}^{\infty}}(R_{VW_n}) < \infty.$$

Observe that $VW_n f = \sum_{i=n+1}^{\infty} \langle f, S^{-1}f_i \rangle e_i = \{\langle f, S^{-1}f_i \rangle\}_{i=n+1}^{\infty}$. So

$$(VW_n)^*\{c_i\} = \sum_{i=n+1}^{\infty} c_i S^{-1} f_i = S^{-1} \sum_{i=n+1}^{\infty} c_i f_i.$$

Since $R_{VW_n}^{\perp} = N_{(VW_n)^*}$ has finite dimension, $\{c_i\}_{i=n+1}^{\infty} \longmapsto \sum_{i=n+1}^{\infty} c_i f_i$ also has a finite dimensional kernel. Therefore

$$T : \ell^2(N) \to \mathcal{H}, \ \ T\{c_i\}_{i=1}^{\infty} = \sum_{i=1}^{\infty} c_i f_i$$

has a finite dimensional kernel, and now the theorem of Holub implies that $\{f_i\}_{i=1}^{\infty}$ is a near-Riesz basis. By the first part of the theorem the two frames $\{f_i\}_{i=1}^{\infty}$ and $\{g_i\}_{i=1}^{\infty}$ now have the same excess, and the proof is complete. □

Example 2. Let us use Theorem 3.6 to give another argument in Example 1. We consider $\{g_i\}_{i=1}^{\infty}$ as a perturbation of the frame $\{f_i\}_{i=1}^{\infty}$, where

$$f_1 = e_1, \qquad f_i = e_{i-1}, \quad i \geq 2.$$

$\{f_i\}_{i=1}^{\infty}$ is a frame with excess 1 and bounds $A = 1, B = 2$. Since

$$\left\|\sum_{i=1}^{\infty} c_i(f_i - g_i)\right\| = \left\|\sum_{i=2}^{\infty} c_i \frac{1}{i} e_i\right\| \le \frac{1}{2}\left[\sum_{i=1}^{\infty} |c_i|^2\right]^{1/2} \quad \forall \{c_i\}_{i=1}^{\infty} \in \ell^2(N),$$

we conclude by Theorems 3.1 and 3.6 that $\{g_i\}_{i=1}^{\infty}$ is a frame with bounds $(1 - \frac{1}{2})^2 = \frac{1}{4}$, $2(1 + \frac{1}{2\sqrt{2}})^2$, and excess 1.

Unfortunately, the requirement that $\{f_i\}_{i=1}^{\infty}$ has finite excess is needed in Theorem 3.6. In fact we are able to construct examples, where $\{f_i\}$ is a tight frame with infinite excess and $\{g_i\}$ does not contain a Riesz basis but where the perturbation condition is satisfied. Let us shortly describe how one can do this. Define $\{f_i^n\}_{i=1,n=1}^{n+1,\infty}$ as in Lemma 2.7. Given $\epsilon > 0$, let

$$g_i^n = e_{\frac{(n-1)n}{2}+i} - \frac{1-\epsilon}{n}\sum_{j=1}^{n} e_{\frac{(n-1)n}{2}+j}, \quad 1 \le i \le n,$$

$$g_{n+1}^n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} e_{\frac{(n-1)n}{2}+j}.$$

Now, given a sequence $\{c_i^n\}$ we have

$$\left\|\sum c_i^n(f_i^n - g_i^n)\right\| = \epsilon \cdot \left\|\sum_{n=1}^{\infty}\left[\sum_{i=1}^{n} c_i^n\right]\frac{1}{n}\sum_{j=1}^{\infty} e_{\frac{(n-1)n}{2}+j}\right\|$$

$$\le \epsilon\sqrt{\sum_{n=1}^{\infty}\left|\sum_{i=1}^{n} c_i^n \frac{1}{\sqrt{n}}\right|^2} \le \epsilon\sqrt{\sum |c_i^n|^2}. \tag{3}$$

By Lemma 2.5, $\{f_i^n\}_{i=1,n=1}^{n,\infty}$ is a frame with bounds 1. If we choose $\epsilon < 1$, then the perturbation condition is satisfied with $\lambda = 0, \mu = \epsilon$, implying that $\{g_i^n\}_{i=1,n=1}^{n+1,\infty}$ is a frame with bounds $(1-\epsilon)^2, (1+\epsilon)^2$.

Claim. $\{g_i^n\}_{i=1,n=1}^{n,\infty}$ is a Riesz basis for $\mathcal{H}$. We only need to prove that $\{g_i^n\}_{i=1,n=1}^{n,\infty}$ satisfies the lower Riesz basis condition. Given a sequence $\{c_i^n\}$ we have

$$\left\|\sum_{n=1}^{\infty}\sum_{i=1}^{n} c_i^n g_i^n\right\| \ge \left\|\sum_{n=1}^{\infty}\sum_{i=1}^{n} c_i^n e_{\frac{(n-1)n}{2}+i}\right\| - (1-\epsilon)\left\|\sum_{n=1}^{\infty}\left(\sum_{i=1}^{n} c_i^n\right)\frac{1}{n}\sum_{i=1}^{n} e_{\frac{(n-1)n}{2}+i}\right\|$$

$$\ge \sqrt{\sum |c_i^n|^2} - (1-\epsilon)\sqrt{\sum_{n=1}^{\infty}|\sum_{i=1}^{n} c_i^n \frac{1}{\sqrt{n}}|^2} \ge \epsilon\sqrt{\sum |c_i^n|^2}.$$

So actually we have an example where $\{f_i^n\}_{i=1,n=1}^{n+1,\infty}$ does not contain a Riesz basis but the perturbed family does. To obtain the example we were looking for, we use the fact that $\{g_i^n\}$ has the lower bound $(1-\epsilon)^2$. By (3) above we can consider $\{f_i^n\}$ as a perturbation of $\{g_i^n\}$ if $\frac{\epsilon}{1-\epsilon} < 1$, i.e., if $\epsilon < \frac{1}{2}$. So we get our example by choosing $\epsilon < 1/2$ and switching the roles of $\{f_i^n\}$ and $\{g_i^n\}$.

Example 3. A Weyl–Heisenberg frame is a frame for $L^2(R)$ of the form

$$\{f_{m,n}\}_{m,n\in Z} = \{e^{imbx} f(x - na)\}_{m,n\in Z},$$

where $f \in L^2(R),\ a, b > 0$. It is well known that $\{f_{m,n}\}_{m,n\in Z}$ is a frame for $L^2(R)$ if f has support in an interval of length $1/b$ and

$$\exists A, B > 0 :\ A \leq \sum_{n\in Z} |f(x - na)|^2 \leq B,\ \text{a.e.}$$

This can only be satisfied if $ab \leq 1$. The case $ab = 1$ implies that $\{f_{m,n}\}_{m,n\in Z}$ is a Riesz basis, cf. [2, 3]. Heil [11, p. 139] has shown that if $ab < 1$, then there exist finite sets $F \subseteq Z\times Z$ of arbitrarily large cardinality such that $\{f_{m,n}\}_{(m,n)\in Z^2-F}$ is a frame. That is, if $\{f_{m,n}\}_{m,n\in Z}$ contains a Riesz basis, then $\{f_{m,n}\}_{m,n\in Z}$ has infinite excess.

Acknowledgments. The second author would like to thank Chris Heil for fruitful discussions on the subject.

REFERENCES

[1] A. Aldroubi, *Portraits of frames*, Proc. Amer. Math. Soc., 123 (1995), pp. 1661–1668.
[2] J. Benedetto and D. Walnut, *Gabor frames for L^2 and related spaces*, in Wavelets: Mathematics and Applications, J. Benedetto and M. Frazier, eds., CRC Press, Boca Raton, FL, 1993, pp. 97–162.
[3] J. Benedetto, C. Heil, and D. Walnut, *Differentiation and the Balian–Low theorem*, J. Fourier Anal. Appl., 1 (1995), pp. 355–402.
[4] P. G. Casazza and O. Christensen, *Hilbert space frames containing a Riesz basis and Banach spaces which have no subspace isomorphic to c_0*, J. Math. Anal. Appl., 202 (1996), pp. 940–950.
[5] O. Christensen, *Frames and the projection method*, Appl. Comp. Harm. Anal., 1 (1993), pp. 50–53.
[6] O. Christensen, *Frames containing Riesz bases and approximation of the frame coefficients using finite dimensional methods*, J. Math. Anal. Appl., 199 (1996), pp. 256–270.
[7] O. Christensen, *A Paley–Wiener theorem for frames*, Proc. Amer. Math. Soc., 123 (1995), pp. 2199–2202.
[8] O. Christensen and C. Heil, *Perturbation of Banach frames and atomic decompositions*, Math. Nach., 185 (1997), pp. 33–47.
[9] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341–366.
[10] J. Holub, *Pre-frame operators, Besselian frames and near-Riesz bases*, Proc. Amer. Math. Soc., 122 (1994), pp. 779–785.
[11] C. Heil, *Wiener Amalgam Spaces in Generalized Harmonic Analysis and Wavelet Analysis*, Ph.D. thesis, University of Maryland, College Park, MD, 1990.
[12] C. Heil and D. Walnut, *Continuous and discrete wavelet transforms*, SIAM Rev., 31 (1989), pp. 628–666.
[13] J. Lindenstrauss and L. Tzafriri, *Classical Banach Spaces* 1, Springer, New York, 1977.
[14] K. Seip, *On the connection between exponential loss and certain related sequences in $L^2(-\pi, \pi)$*, J. Funct. Anal., 130 (1995), pp. 131–160.
[15] I. Singer, *Bases in Banach Spaces* 1, Springer, New York, 1970.
[16] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 279–292, January 1998

016

HARMONIOUS EXTENSIONS*

E. LE GRUYER† AND J. C. ARCHER†

Abstract. We give a class of optimally stable solutions to Tietze's extension problem in general, metrically convex compact metric spaces. The extensions which solve this problem are the unique stationary states of nonlinear processes of regularization.

Key words. stability, processes of regularization, extension of functionals

AMS subject classifications. 41A05, 46A22, 47B50, 65D05

PII. S0036141095294067

1. Problem statement. In this article, we give a class of optimally stable solutions to Tietze's extension problem in general, metrically convex compact metric spaces. The extensions which solve this problem are the unique stationary states of nonlinear processes of regularization. The title is suggested by the formal analogy between the processes that we describe and the process of harmonic regularization (the classical process of diffusion). This analogy is briefly indicated at the end of this section and in section 4 of the paper.

Let (E, d) denote any metric space, and let f be any scalar-valued continuous function whose domain is a closed nonempty subset of E. Since Tietze's original result, which produced a continuous extension $\mathcal{E}(f)$ of f, several solutions have been proposed to improve the quality of the extension. Kakutani (in separable metric spaces) and Dugundji [2] (in a more general case than the metric case) have proposed linear and positive schemes $f \mapsto \mathcal{E}(f)$. These schemes are optimally data-value stable (DV-stable); that is,

$$\|\mathcal{E}(f) - \mathcal{E}(g)\|_{\infty,E} \leq D\|f - g\|_{\infty,A},$$

with $D = 1$, for any pair of scalar-valued continuous functions f, g with common domain A. Here the symbol $\|.\|_\infty$ denotes the usual supremum norm. When $E = \mathbb{R}^n$, linear optimally DV-stable schemes exist which are, moreover, Ω-stable. That is, they satisfy

$$\hat{\omega}(\mathcal{E}(f)) \leq C\hat{\omega}(f),$$

where C denotes a constant which does not depend on f and $\hat{\omega}(g)$ denotes the concave modulus of continuity of g. These schemes are obtained [1] by both improving the original construction of Whitney [7] and a result of Glaeser [3]. It is well known that, in the multidimensional case ($n > 1$), these linear schemes cannot be optimally Ω-stable ($C > 1$).

By improving a scheme of Mc Shane [5], we have obtained an optimally Ω-stable ($C = 1$) extension scheme $\mathcal{E}$, which is, moreover, DV-stable with $D \leq 3$, and also data-site stable [4]. Let us recall that $\mathcal{E}$ is defined by $\mathcal{E}(f)(x) := \sup_{a \in dom(f)}(f(a) - \hat{\omega}(f; d(x, a)))$. It can be shown that this scheme $\mathcal{E}$ is self-reproducing. That is,

*Received by the editors October 31, 1995; accepted for publication (in revised form) July 17, 1996.

http://www.siam.org/journals/sima/29-1/29406.html

†Institut National des Sciences Appliquées, 20 Avenue des Buttes de Coësmes, 35043 Rennes cedex, France (legruyer@perceval.univ-rennes1.fr, archer@perceval.univ-rennes1.fr).

$\mathcal{E}(\mathcal{E}(f)|B) = \mathcal{E}(f)$ for any f as above and any closed subset B of E containing $\mathrm{dom}(f)$ (the symbol "|" denotes restriction to). This scheme is not a local scheme because the modulus of continuity of a function is not a local notion. The concept of locality, which is, in smooth spaces, closely related to PDE (see section 4), has the following simple formulation: an extension scheme $\mathcal{E}$ is said to be *local* if $\mathcal{E}(\mathcal{E}(f)|\partial B)|B = \mathcal{E}(f)|B$ for any f as above and for any closed and bounded subset B of E with boundary ∂B such that the interior of B does not intersect $\mathrm{dom}(f)$.

In this paper, we produce, when (E, d) is any metrically convex compact metric space, a class of extension schemes which are both optimally Ω-stable and optimally DV-stable. Moreover, we establish (Proposition 3.9) the quasi-data-site stability, the quasi-locality, and the quasi-self-reproduction of some schemes of this class. For each extension scheme $\mathcal{K}$ of our class, the extension $\mathcal{K}(f)$ of f is obtained as follows. Starting with any continuous extension f_0 of f (Tietze's theorem ensures that such an extension exists), we define inductively a sequence $(f_n)_{n\in\mathbb{N}}$ of continuous extensions of f by

$$f_{n+1}(x) := \tilde{f}_n(x) := \frac{1}{2}\sup_{u\in D(x)} f_n(u) + \frac{1}{2}\inf_{u\in D(x)} f_n(u), x \in E,$$

where $D(x) :=$ ball centered at x of radius $r(x)$ satisfying $r(x) = 0 \iff x \in \mathrm{dom}(f)$; $|r(x) - r(y)| \le d(x, y), x, y \in E$.

The extension $\mathcal{K}(f)$ of f is the limit of this sequence, that is, the stationary state of the process of regularization $g \to \tilde{g}$. This stationary state does not depend on the initial state f_0.

Let us note, when E is a Euclidean space, the formal analogy between the process $g \mapsto \tilde{g}$ and the process of harmonic regularization $g \mapsto \hat{g}$ defined by

$$\hat{g}(x) = \int_{y\in D(x)} g(y)dy \Big/ \int_{y\in D(x)} dy, x \in E,$$

for which it is known [6] that the stationary states are harmonic functions.[1] It is this analogy which has led us to call the processes $g \mapsto \tilde{g}$ processes of harmonious regularization. It does not seem impossible to us that the processes of harmonious regularization, which regularize by ordering (the processes of harmonic regularization regularize by homogenizing) could occur in some model of the morphogenesis of ordered states of materials.

From a technical point of view, we prove the existence of harmonious extensions with the help of Ascoli's theorem and of Schauder's fixed point theorem. The nonlinearity of the harmonious extension schemes makes the proof of the uniqueness more difficult than in the harmonic case (see Theorem 3.3). Schauder's theorem is insufficient to prove the convergence of our processes to their stationary states: our proof needs an analysis of the story of the processes of harmonious regularization (see Theorem 3.5).

2. Preliminaries. Let us first recall some definitions. We call *concave modulus of continuity* any mapping $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ which satisfies the following:

(i) $\omega(0) = 0$ and ω is continuous at 0;

(ii) ω is increasing: $h_1 \le h_2 \Rightarrow \omega(h_1) \le \omega(h_2)$;

[1] We are indebted to Prof. Y. Guivarc'h, who brought the result of W.A. Veech to our attention.

(iii) ω is concave: $\forall(\lambda_i)_{i=1,\dots,n}, 0 \le \lambda_i \le 1, \sum_{i=1}^n \lambda_i = 1$,

$$\forall(h_i)_{i=1,\dots,n}, \sum_{i=1}^n \lambda_i \omega(h_i) \le \omega\left(\sum_{i=1}^n \lambda_i h_i\right).$$

Let (E,d) be any metric space and let f be any function from E to $\mathbb{R}$. We say that f is Ω-continuous if there exists a concave modulus of continuity ω such that, for any $x, y \in E$,

$$|f(y) - f(x)| \le \omega(d(x,y)).$$

For such a function f, the lower bound of those concave moduli of continuity which satisfy the inequality above is still a concave modulus of continuity which satisfies this inequality. We denote it by $\hat{\omega}(f)$.

Any Ω-continuous function is uniformly continuous. Any bounded and uniformly continuous function is Ω-continuous. Therefore, any continuous function from a compact metric space to $\mathbb{R}$ is Ω-continuous.

A metric space (E,d) is said to be *metrically convex* if, for any $x, y \in E$ and for any real r, $0 \le r \le d(x,y)$, there exists some $z \in E$ such that $d(x,z) = r$ and $d(x,z) + d(z,y) = d(x,y)$.

Moreover, let us recall the following formulas:

$$\sup_{i\in I} x_i - \sup_{j\in J} y_j = \sup_{i\in I} \inf_{j\in J} (x_i - y_j); \tag{1}$$

$$\inf_{i\in I} x_i - \inf_{j\in J} y_j = \sup_{j\in J} \inf_{i\in I} (x_i - y_j); \tag{2}$$

$$\sup_{i\in I} x_i - \sup_{i\in I} y_i \le \sup_{i\in I} (x_i - y_i); \tag{3}$$

$$\inf_{i\in I} x_i - \inf_{i\in I} y_i \le \sup_{i\in I} (x_i - y_i). \tag{4}$$

The main results of this paper are contained in Theorems 3.3 and 3.5 below. From now on, (E,d) denotes any metrically convex compact metric space, A any closed nonempty subset of E, and r any mapping from E to $\mathbb{R}^+$ which satisfies the following:

(i) $r(x) = 0$ iff $x \in A$;

(ii) $|r(x) - r(y)| \le d(x,y), x, y \in E$.

Let us note here that such mappings r exist:

(i) $r(x) := \rho d(x,A), 0 < \rho \le 1$;

(ii) $r(x) := \inf(h, d(x,A)), h > 0$.

We denote by $D(x)$ the ball of center x, radius $r(x)$:

$$D(x) := \{y \in E : d(x,y) \le r(x)\}.$$

We start with a geometrical lemma.

LEMMA 2.1. *For any $x, y \in E$, we have*

$$\frac{1}{2} \sup_{u\in D(x)} \inf_{v\in D(y)} d(u,v) + \frac{1}{2} \sup_{v\in D(y)} \inf_{u\in D(x)} d(u,v) \le d(x,y).$$

Proof. Let us first establish that, for any $x, v \in E$ such that $d(x,v) \geq r(x)$, we have

$$\inf_{u \in D(x)} d(u,v) = d(x,v) - r(x).$$

To prove this inequality, we note that, by convexity, there exists $z \in E$ such that $d(x,z) = r(x)$ and $d(z,v) = d(x,v) - r(x)$. Thus we have $z \in D(x)$. Moreover, for any $u \in D(x)$, we have, by the triangle inequality,

$$d(x,v) \leq d(x,u) + d(u,v) \leq r(x) + d(u,v).$$

We infer that $d(u,v) \geq d(x,v) - r(x) = d(z,v)$ and, therefore, that

$$\inf_{u \in D(x)} d(u,v) = d(z,v) = d(x,v) - r(x),$$

which is the stated equality.

Now let $x, y \in E$ be such that there exists $v \in D(y)$, $d(x,v) \geq r(x)$. Using the equality above, the triangle inequality, and the definition of $D(y)$, we have

$$\begin{aligned} \sup_{v \in D(y)} \inf_{u \in D(x)} d(u,v) &\leq \sup_{v \in D(y)} d(x,v) - r(x) \\ &\leq \sup_{v \in D(y)} (d(x,y) + d(y,v)) - r(x) \\ &\leq d(x,y) + r(y) - r(x). \end{aligned}$$

We are now ready to prove the inequality of Lemma 2.1.

First case: $D(x) \subset D(y)$. In this case we have

$$\sup_{u \in D(x)} \inf_{v \in D(y)} d(u,v) = 0$$

and, by the inequality just established,

$$\sup_{v \in D(y)} \inf_{u \in D(x)} d(u,v) \leq d(x,y) + r(y) - r(x).$$

The result follows in this case since, by hypothesis, we have $|r(x) - r(y)| \leq d(x,y)$.

Second case: $D(y) \subset D(x)$. This case is similar to the first one.

Third case: $D(y) \not\subset D(x)$ and $D(x) \not\subset D(y)$. In this case we can apply twice the inequality previously established:

$$\begin{aligned} \frac{1}{2} \sup_{u \in D(x)} \inf_{v \in D(y)} d(u,v) + \frac{1}{2} \sup_{v \in D(y)} \inf_{u \in D(x)} d(u,v) &\leq \frac{1}{2}(d(x,y) + r(y) - r(x)) \\ &\quad + \frac{1}{2}(d(x,y) + r(x) - r(y)) \\ &\leq d(x,y), \end{aligned}$$

which is the stated result. ∎

3. Main results.

DEFINITION 3.1. *For any bounded function f from E to $\mathbb{R}$ we define a new bounded function $\tilde{f}$ from E to $\mathbb{R}$, called the harmonious regularization of f, by*

$$\tilde{f}(x) = \frac{1}{2}\sup_{u\in D(x)} f(u) + \frac{1}{2}\inf_{u\in D(x)} f(u), x \in E.$$

PROPOSITION 3.2. (i) *For any uniformly continuous function f from E to $\mathbb{R}$, we have*

$$\hat{\omega}(\tilde{f}) \leq \hat{\omega}(f);$$

(ii) *For any bounded functions f, g from E to $\mathbb{R}$, we have*

$$\|\tilde{f} - \tilde{g}\|_{\infty,E} \leq \|f - g\|_{\infty,E}.$$

Proof. (i) Let $x, y \in E$. We have

$$\begin{aligned}
\tilde{f}(x) - \tilde{f}(y) &= \frac{1}{2}\left(\sup_{u\in D(x)} f(u) + \inf_{u\in D(x)} f(u)\right) - \frac{1}{2}\left(\sup_{v\in D(y)} f(v) + \inf_{v\in D(y)} f(v)\right) \\
&= \frac{1}{2}\sup_{u\in D(x)}\inf_{v\in D(y)}(f(u) - f(v)) + \frac{1}{2}\sup_{v\in D(y)}\inf_{u\in D(x)}(f(u) - f(v)) \\
&\leq \frac{1}{2}\sup_{u\in D(x)}\inf_{v\in D(y)}\hat{\omega}(f; d(u,v)) + \frac{1}{2}\sup_{v\in D(y)}\inf_{u\in D(x)}\hat{\omega}(f; d(u,v)) \\
&\leq \frac{1}{2}\hat{\omega}\left(f; \sup_{u\in D(x)}\inf_{v\in D(y)} d(u,v)\right) + \frac{1}{2}\hat{\omega}\left(f; \sup_{v\in D(y)}\inf_{u\in D(x)} d(u,v)\right) \\
&\leq \hat{\omega}\left(f; \frac{1}{2}\sup_{u\in D(x)}\inf_{v\in D(y)} d(u,v) + \frac{1}{2}\sup_{v\in D(y)}\inf_{u\in D(x)} d(u,v)\right) \\
&\leq \hat{\omega}(f; d(x,y)).
\end{aligned}$$

The first equality is a consequence of the definition of $\tilde{f}$, and the second comes from formulas (1) and (2). The first inequality follows from the definition of $\hat{\omega}(f)$, the second from the monotonicity of $\hat{\omega}(f)$, the third from the concavity of $\hat{\omega}(f)$, and the last from Lemma 2.1 and from the monotonicity of $\hat{\omega}(f)$. We finish the proof by an exchange of the roles of x and y.

(ii) This assertion follows immediately from formulas (3) and (4). □

We are now ready to establish the following theorem.

THEOREM 3.3. *Any continuous function f from A to $\mathbb{R}$ has a unique continuous extension $\mathcal{K}(f)$ from E to $\mathbb{R}$ which satisfies the functional equation*

$$g(x) = \frac{1}{2}\sup_{u\in D(x)} g(u) + \frac{1}{2}\inf_{u\in D(x)} g(u). \tag{5}$$

Proof. Existence: let $\mathcal{C}^0(E,\mathbb{R})$ be the Banach space of all continuous mappings from E to $\mathbb{R}$ with the norm $\|.\|_{\infty,E}$ of the uniform convergence. Let

$$K := \{g \in \mathcal{C}^0(E,\mathbb{R}) : g \text{ extends } f, \|g\|_{\infty,E} \leq \|f\|_{\infty,A}, \hat{\omega}(g) \leq \hat{\omega}(f)\}.$$

Using the result of Mc Shane ([5, Theorem 2 and Corollary 2]), K is a nonempty subset of $\mathcal{C}^0(E,\mathbb{R})$. Let $g_1, g_2 \in K, 0 \leq \lambda \leq 1, g := (1-\lambda)g_1 + \lambda g_2$. It is immediate that g is a continuous extension of f and that

$$\|g\|_{\infty,E} \leq \|f\|_{\infty,A}.$$

Moreover, for any $x, y \in E$, we have

$$|g(x) - g(y)| \leq (1-\lambda)\hat{\omega}(g_1; d(x,y)) + \lambda\hat{\omega}(g_2; d(x,y)) \leq \hat{\omega}(f; d(x,y));$$

that is, $\hat{\omega}(g) \leq \hat{\omega}(f)$. We infer that K is a convex subset of $\mathcal{C}^0(E,\mathbb{R})$. This set K is closed and, by Ascoli's theorem, it is, moreover, a compact subset of $\mathcal{C}^0(E,\mathbb{R})$. Using Proposition 3.2, we have $\|\tilde{g}\|_{\infty,E} \leq \|g\|_{\infty,E}$ and $\hat{\omega}(\tilde{g}) \leq \hat{\omega}(g)$. Therefore, the operator $g \mapsto \tilde{g}$ maps K into K. Using Proposition 3.2 (ii), this operator is a continuous mapping. The proof of the existence follows now from Schauder's fixed point theorem.

Uniqueness: let g and h be two continuous extensions of f which satisfy the functional equation (5). Let us set

$$\Delta := \sup_{x\in E}(g(x) - h(x)), \quad F := \{x \in E : g(x) - h(x) = \Delta\},$$

$$M := \sup_{x\in F} g(x), \quad G := \{x \in F : g(x) = M\}.$$

The set G is nonempty because E is compact and because g and h are continuous mappings. Let us first show that $D(x) \subset F$ for any $x \in G$. Let us assume, by way of a contradiction, that there exists $x \in G$ such that $D(x) \not\subset F$. Since $G \subset F$ and since g and h are two extensions of f which satisfy (5), we have

$$\Delta = g(x) - h(x) = \frac{1}{2}\sup_{z\in D(x)} g(z) - \frac{1}{2}\sup_{z\in D(x)} h(z) + \frac{1}{2}\inf_{z\in D(x)} g(z) - \frac{1}{2}\inf_{z\in D(x)} h(z).$$

Since, by formula (4) and by definition of Δ we have

$$\inf_{z\in D(x)} g(z) - \inf_{z\in D(x)} h(z) \leq \sup_{z\in D(x)} (g(z) - h(z)) \leq \Delta,$$

we infer that

$$\Delta \leq \sup_{z\in D(x)} g(z) - \sup_{z\in D(x)} h(z). \tag{6}$$

Then let $y \in D(x)$ such that $g(y) = \sup_{z\in D(x)} g(z)$. Since $g(x) = M$ and $x \in D(x)$, we infer that $g(y) \geq M$.

First case: $g(y) > M$. In this case, we have $y \notin F$ because, by definition of F and M, we have $g(z) \leq M$ for any $z \in F$. But by (6), we also have

$$\Delta \leq g(y) - \sup_{z\in D(x)} h(z) \leq g(y) - h(y) \leq \Delta,$$

that is, $\Delta = g(y) - h(y)$, from which we infer $y \in F$, a contradiction.

Second case: $g(y) = M$. In this case, as g extends f and satisfies (5), we have

$$g(x) = \frac{1}{2} \sup_{z \in D(x)} g(z) + \frac{1}{2} \inf_{z \in D(x)} g(z).$$

As $x \in G$, we have $g(x) = M$. Therefore, $\inf_{z \in D(x)} g(z) = M$, and it follows that for any $z \in D(x)$ we have $g(z) = M$. Since, by hypothesis, $D(x)$ is not contained in F, let $t \in D(x)$, $t \notin F$. We have $g(t) = M$. Applying (6) again, we obtain

$$\Delta \leq g(t) - \sup_{z \in D(x)} h(z) \leq g(t) - h(t) \leq \Delta.$$

Now, as in the first case, we obtain $\Delta = g(t) - h(t)$; therefore, $t \in F$, a contradiction.

We have therefore proved that, for any $x \in G$, $D(x)$ is a subset of F. Hence, for any $u \in D(x)$, we have $g(u) \leq M$. Applying (5) again, we infer that $g(u) = M$ for any $u \in D(x)$. In other words, for any $x \in G$, $D(x)$ is a subset of G.

Now let us show that $G \bigcap A \neq \emptyset$. Let us assume, again by way of contradiction, that the distance $d(G, A)$ from G to A is strictly positive, and let $x \in G$, $a \in A$ such that $d(x, a) = d(G, A) > 0$. Using the hypotheses on the mapping r, we have $r(a) = 0$, $r(x) > 0$, and $r(x) \leq d(x, A) \leq d(x, a)$. By convexity of (E, d), there exists $z \in E$ such that

(i) $d(x, z) = r(x)$;

(ii) $d(z, a) = d(x, a) - r(x)$.

From (i), we have $z \in D(x)$. Since $D(x)$ is a subset of G, we infer that $z \in G$. From (ii), we have $d(z, a) < d(G, A)$, a contradiction.

Now, as $G \subset F$, we also have $F \bigcap A \neq \emptyset$. As g and h are both extensions of f, we infer that $\Delta = 0$, that is, $g \leq h$. By exchanging the roles of g and h , we obtain $g = h$, which is the stated result. □

Remark 3.4. (1) A corollary of the proof of Theorem 3.3 (uniqueness) is that the harmonious extension schemes $\mathcal{K}$ satisfy the maximum principle.

(i) $\sup_{z \in E} \mathcal{K}(f)(z) = \sup\{f(a) : a \in \mathrm{dom}(f)\}$.

(ii) If there exists x belonging to $E - \mathrm{dom}(f)$ such that

$$\mathcal{K}(f)(x) = \sup_{z \in E} \mathcal{K}(f)(z),$$

then $\mathcal{K}(f)$ is constant in a neighborhood of x. Here, as usual, f denotes any scalar-valued continuous function whose domain is a closed nonempty subset of E.

(2) Another corollary of the proof of Theorem 3.3 is that

$$\hat{\omega}(\mathcal{K}(f)) \leq \hat{\omega}(f).$$

The properties of stability of the extension scheme $\mathcal{K}$ are formulated in Theorem 3.5 below. Hypotheses and notations are those of Theorem 3.3.

THEOREM 3.5. (i) *For any continuous function f_0 from E to $\mathbb{R}$ which extends f, the sequence $(f_n)_{n \in \mathbb{N}}$ inductively defined by $f_{n+1} = \tilde{f}_n$ converges to $\mathcal{K}(f)$.*

(ii) *If g is any continuous function from A to $\mathbb{R}$, then we have*

$$\|\mathcal{K}(f) - \mathcal{K}(g)\|_{\infty, E} \leq \|f - g\|_{\infty, A}.$$

Proof. To show (i), it is sufficient to prove that $\lim_{n \to \infty} D_n = 0$, where $D_n := \sup_{x \in E} |f_{n+1}(x) - f_n(x)|$.

Indeed, by definition of the harmonious regularization and from Proposition 3.2, the sequence $(f_n)_{n\in\mathbb{N}}$ is equicontinuous and equibounded. Therefore, by Ascoli's theorem, there exists a subsequence $(f_{\varphi(n)})_{n\in\mathbb{N}}$ which converges to a continuous extension of f denoted by g. Since, by Proposition 3.2(ii), the operator of harmonious regularization is continuous, we infer that

$$\lim_{n\to\infty} \tilde{f}_{\varphi(n)} = \tilde{g}.$$

Now, as $\tilde{f}_{\varphi(n)} = f_{\varphi(n)+1}$, the convergence to 0 of sequence D_n implies that

$$\lim_{n\to\infty} f_{\varphi(n)} = \lim_{n\to\infty} \tilde{f}_{\varphi(n)},$$

that is, $\tilde{g} = g$. From Theorem 3.3, we have $g = \mathcal{K}(f)$.

If the sequence $(f_n)_{n\in\mathbb{N}}$ did not converge to $\mathcal{K}(f)$, there would exist $\epsilon > 0$ and a subsequence $(f_{\psi(n)})_{n\in\mathbb{N}}$ such that $\|f_{\psi(n)} - \mathcal{K}(f)\|_{\infty,E} > \epsilon$ for any $n \in \mathbb{N}$. Using Proposition 3.2 and Ascoli's theorem again, we could, by the same argument, obtain a new subsequence converging to a continuous extension of f satisfying (5) and distinct from $\mathcal{K}(f)$, contradicting Theorem 3.3.

To prove that $\lim_{n\to\infty} D_n = 0$, we set

$$\Delta_n^+ := \sup_{x\in E}(f_{n+1}(x) - f_n(x)),$$

$$\Delta_n^- := \sup_{x\in E}(f_n(x) - f_{n+1}(x)).$$

We notice that the sequences $(\Delta_n^+)_{n\in\mathbb{N}}$ and $(\Delta_n^-)_{n\in\mathbb{N}}$ are positive because the functions f_n are extensions of f.

As $0 \le D_n \le \sup(\Delta_n^+, \Delta_n^-)$, it is sufficient to prove that $\lim_{n\to\infty} \Delta_n^+ = 0$ and $\lim_{n\to\infty} \Delta_n^- = 0$.

Let us show that these sequences are decreasing. From the definition of f_{n+1} and f_{n+2}, we have

$$\Delta_{n+1}^+ = \sup_{x\in E}\left(\frac{1}{2}\sup_{z\in D(x)} f_{n+1}(z) - \frac{1}{2}\sup_{z\in D(x)} f_n(z) + \frac{1}{2}\inf_{z\in D(x)} f_{n+1}(z) - \frac{1}{2}\inf_{z\in D(x)} f_n(z)\right).$$

Therefore, by formulas (3) and (4), we have

$$\sup_{z\in D(x)} f_{n+1}(z) - \sup_{z\in D(x)} f_n(z) \le \Delta_n^+$$

and

$$\inf_{z\in D(x)} f_{n+1}(z) - \inf_{z\in D(x)} f_n(z) \le \Delta_n^+.$$

It follows that $\Delta_{n+1}^+ \le \Delta_n^+$. The proof is similar for the sequence $(\Delta_n^-)_{n\in\mathbb{N}}$.

Now the proof of $\lim_{n\to\infty} D_n = 0$ and, therefore, the proof of (i) will be an immediate consequence of the following two lemmas.

LEMMA 3.6. *Let $\delta > 0$ and let $(u_n)_{n\in\mathbb{N}}$ be a decreasing sequence of positive numbers satisfying, for any integer n and for any strictly positive integer p,*

$$u_{n+p} \le 2^{-p}((2^p - p)u_n + \delta).$$

Then

$$\lim_{n\to\infty} u_n = 0.$$

From now on, we set $\delta := \hat{\omega}(f_0; \operatorname{diam}(E))$, where $\operatorname{diam}(E)$ denotes the diameter of the compact metric space E. Using Proposition 3.2(i), we have

$$|f_n(x) - f_n(y)| \le \delta$$

for any $n \in \mathbb{N}$, $x, y \in E$. □

LEMMA 3.7. *For any integer n and any strictly positive integer p, we have*

(i) $\Delta^+_{n+p} \le 2^{-p}((2^p - p)\Delta^+_n + \delta)$,

(ii) $\Delta^-_{n+p} \le 2^{-p}((2^p - p)\Delta^-_n + \delta)$.

Proof of Lemma 3.6. Let $p > 0$ and $a_p = 1 - 2^{-p}p$. By a recursive application of the hypothesis, we obtain, for any $k \in \mathbb{N}$, $k > 0$,

$$\begin{aligned} u_{kp} &\le a_p u_{(k-1)p} + \delta 2^{-p} \\ &\quad \cdots \\ &\le (a_p)^k u_0 + (1 + \cdots + (a_p)^{k-1})\delta 2^{-p} \\ &\le (a_p)^k u_0 + (1/(1 - a_p))\delta 2^{-p}. \end{aligned}$$

That is,

$$u_{kp} \le (a_p)^k u_0 + \delta/p.$$

The sequence $(u_n)_{n\in\mathbb{N}}$ is positive and decreasing and therefore converges. Thus its subsequence $(u_{kp})_{k\in\mathbb{N}}$ also converges to the same limit. Since $0 \le a_p < 1$, the inequality above shows that this limit is positive and smaller than δ/p. Lemma 3.6 follows since p can be chosen arbitrarily large. □

Proof of Lemma 3.7. The proofs of (i) and (ii) are similar. Let us show (i).

Let $n \in \mathbb{N}$ and $x_0 \in E$ such that $f_{n+1}(x_0) - f_n(x_0) = \Delta^+_n$. We define inductively two sequences $(x_k)_{k=0,\dots,n}$ and $(y_k)_{k=0,\dots,n-1}$ of elements of E by

(a) $y_0 := x_0$,

(b) x_{k+1} is chosen in $D(x_k)$ such that $f_{n-k}(x_{k+1}) = \sup_{z\in D(x_k)} f_{n-k}(z)$,

(c) y_{k+1} is chosen in $D(y_k)$ such that $f_{n-k-1}(y_{k+1}) = \inf_{z\in D(y_k)} f_{n-k-1}(z)$.

Now let p, k be integers, $1 \le p \le n$, $0 \le k \le p - 1$.

By definition of f_{n-k+1}, x_k, and x_{k+1}, we have

$$f_{n-k+1}(x_k) = \frac{1}{2}(f_{n-k}(x_{k+1}) + \inf_{z\in D(x_k)} f_{n-k}(z)).$$

Using this equality inductively for $k = 0, \dots, p - 1$, we obtain

$$f_{n+1}(x_0) = 2^{-p}\left(f_{n+1-p}(x_p) + \sum_{k=0}^{p-1} 2^{p-1-k} \inf_{z\in D(x_k)} f_{n-k}(z)\right).$$

Similarly, we also have

$$f_n(y_0) = 2^{-p}\left(f_{n-p}(y_p) + \sum_{k=0}^{p-1} 2^{p-1-k} \sup_{z\in D(y_k)} f_{n-k-1}(z)\right).$$

Now we write Δ_n^+ in the following form:

$$\Delta_n^+ = 2^{-p}\left(f_{n+1-p}(x_p) - f_{n-p}(y_p) + \inf_{z\in D(x_0)} f_n(z) - \sup_{z\in D(x_0)} f_{n-1}(z) + R_1 + S_1\right),$$

where

$$R_1 := \sum_{i=1}^{p-1} 2^{p-1-i} \inf_{z\in D(x_i)} f_{n-i}(z) - (2^{p-1}-1) \sup_{z\in D(x_0)} f_{n-1}(z),$$

$$S_1 := (2^{p-1}-1) \inf_{z\in D(y_0)} f_n(z) - \sum_{i=1}^{p-1} 2^{p-1-i} \sup_{z\in D(y_i)} f_{n-1-i}(z).$$

By the definitions of Δ_{n-p}^+ and δ, we have

$$\begin{aligned} f_{n+1-p}(x_p) - f_{n-p}(y_p) &= f_{n+1-p}(x_p) - f_{n-p}(x_p) + f_{n-p}(x_p) - f_{n-p}(y_p) \\ &\le \Delta_{n-p}^+ + \delta. \end{aligned}$$

Moreover, using $x_0 \in D(x_0)$ and the definition of f_n, we have

$$\begin{aligned} \inf_{z\in D(x_0)} f_n(z) - \sup_{z\in D(x_0)} f_{n-1}(z) &\le f_n(x_0) - \sup_{z\in D(x_0)} f_{n-1}(z) \\ &\le \frac{1}{2} \sup_{z\in D(x_0)} f_{n-1}(z) + \frac{1}{2} \inf_{z\in D(x_0)} f_{n-1}(z) - \sup_{z\in D(x_0)} f_{n-1}(z) \\ &\le \frac{1}{2}\left(\inf_{z\in D(x_0)} f_{n-1}(z) - \sup_{z\in D(x_0)} f_{n-1}(z)\right) \\ &\le 0. \end{aligned}$$

Therefore,

$$\Delta_n^+ \le 2^{-p}(\Delta_{n-p}^+ + \delta + R_1 + S_1). \tag{7}$$

Now it remains to bound R_1 and S_1. They are the first terms of sequences $(R_k)_{k=1,\dots,p-1}$ and $(S_k)_{k=1,\dots,p-1}$ defined by

$$R_k := \sum_{i=k}^{p-1} 2^{p-1-i} \inf_{z\in D(x_i)} f_{n-i}(z) - (2^{p-k}-1) \sup_{z\in D(x_{k-1})} f_{n-k}(z),$$

$$S_k := (2^{p-k}-1) \inf_{z\in D(y_{k-1})} f_{n+1-k}(z) - \sum_{i=k}^{p-1} 2^{p-1-i} \sup_{z\in D(y_i)} f_{n-1-i}(z).$$

Sublemma. *We have* (i)

$$R_k \le \left(2^{p-1-k} - \frac{1}{2}\right)\Delta_{n-1-k}^+ + R_{k+1}, k = 1, \dots, p-2, \tag{8}$$

$$R_{p-1} \le 0, \tag{9}$$

and (ii)

$$S_k \le \left(2^{p-1-k} - \frac{1}{2}\right)\Delta^+_{n-1-k} + S_{k+1}, k = 1, \dots, p-2, \tag{10}$$

$$S_{p-1} \le \frac{1}{2}\Delta^+_{n-p}. \tag{11}$$

Using part (i) of this sublemma, we have, therefore,

$$R_1 \le \left(2^{p-2} - \frac{1}{2}\right)\Delta^+_{n-2} + \left(2^{p-3} - \frac{1}{2}\right)\Delta^+_{n-3} + \cdots + \left(2 - \frac{1}{2}\right)\Delta^+_{n+1-p}.$$

Since sequence Δ^+_n is decreasing, we infer that

$$R_1 \le \left(2^{p-2} - \frac{1}{2} + 2^{p-3} - \frac{1}{2} + \cdots + 2 - \frac{1}{2}\right)\Delta^+_{n+1-p};$$

that is,

$$R_1 \le \left(2^{p-1} - \frac{p}{2} - 1\right)\Delta^+_{n-p}. \tag{12}$$

Similarly, using (10) and (11), we obtain

$$S_1 \le \left(2^{p-1} - \frac{p}{2} - \frac{1}{2}\right)\Delta^+_{n-p}. \tag{13}$$

Combining (7), (12) and (13), we arrive at

$$\Delta^+_n \le 2^{-p}((2^p - p)\Delta^+_{n-p} + \delta).$$

The stated result follows by the translation of n to $n + p$.

Proof of the sublemma. First let us show (8). Since $x_k \in D(x_{k-1})$, we have

$$\begin{aligned} -\sup_{z\in D(x_{k-1})} f_{n-k}(z) &\le -f_{n-k}(x_k) \\ &\le -\frac{1}{2}\sup_{z\in D(x_k)} f_{n-k-1}(z) - \frac{1}{2}\inf_{z\in D(x_k)} f_{n-k-1}(z). \end{aligned}$$

Using this inequality, we can bound R_k as follows:

$$\begin{aligned} R_k \le R_{k+1} &+ \left(2^{p-1-k} - \frac{1}{2}\right)\left[\inf_{z\in D(x_k)} f_{n-k}(z) - \inf_{z\in D(x_k)} f_{n-k-1}(z)\right] \\ &+ \frac{1}{2}\left(\inf_{z\in D(x_k)} f_{n-k}(z) - \sup_{z\in D(x_k)} f_{n-k-1}(z)\right). \end{aligned}$$

The expression in square brackets in this last inequality is directly bounded by Δ^+_{n-k-1}, and the last term is negative (by the argument used to prove inequality (7)).

Therefore, we obtain

$$R_k \le R_{k+1} + \left(2^{p-1-k} - \frac{1}{2}\right)\Delta^+_{n-k-1},$$

which is the stated inequality (8).

Moreover, since $x_{p-1} \in D(x_{p-1})$, we have

$$\begin{aligned} R_{p-1} &= \inf_{z \in D(x_{p-1})} f_{n-p+1}(z) - \sup_{z \in D(x_{p-2})} f_{n-p+1}(z) \\ &\leq f_{n-p+1}(x_{p-1}) - \sup_{z \in D(x_{p-2})} f_{n-p+1}(z). \end{aligned}$$

As, by construction, we have $x_{p-1} \in D(x_{p-2})$, we infer inequality (9).

The proof of (10) is similar to that of (8). It remains to bound S_{p-1}. We have

$$S_{p-1} = \inf_{z \in D(y_{p-2})} f_{n-p+2}(z) - \sup_{z \in D(y_{p-1})} f_{n-p}(z).$$

As $y_{p-1} \in D(y_{p-2})$, we have

$$S_{p-1} \leq f_{n-p+2}(y_{p-1}) - \sup_{z \in D(y_{p-1})} f_{n-p}(z).$$

Using the definition of $f_{n-p+2}(y_{p-1})$, we can write

$$\begin{aligned} S_{p-1} &\leq \frac{1}{2} \left(\sup_{z \in D(y_{p-1})} f_{n-p+1}(z) - \sup_{z \in D(y_{p-1})} f_{n-p}(z) \right) \\ &\quad + \frac{1}{2} \left(\inf_{z \in D(y_{p-1})} f_{n-p+1}(z) - \sup_{z \in D(y_{p-1})} f_{n-p}(z) \right) \\ &\leq \frac{1}{2} \Delta^+_{n-p} + \frac{1}{2} \left(f_{n-p+1}(y_{p-1}) - \sup_{z \in D(y_{p-1})} f_{n-p}(z) \right). \end{aligned}$$

Using now the definition of $f_{n-p+1}(y_{p-1})$, we have

$$\begin{aligned} f_{n-p+1}(y_{p-1}) &= \frac{1}{2} \sup_{z \in D(y_{p-1})} f_{n-p}(z) + \frac{1}{2} \inf_{z \in D(y_{p-1})} f_{n-p}(z) \\ &\leq \sup_{z \in D(y_{p-1})} f_{n-p}(z). \end{aligned}$$

Finally, we have $S_{p-1} \leq (1/2)\Delta^+_{n-p}$, which is inequality (11).

We have now finished the proof of the sublemma, of Lemma 3.2, and, therefore, of part (i) of Theorem 3.5.

Proof of Theorem 3.5(ii). Using a result of Dugundji [2], let f_0 and g_0 be two continuous functions from E to $\mathbb{R}$ which extend, respectively, f and g and are such that $\|f_0 - g_0\|_{\infty,E} \leq \|f - g\|_{\infty,A}$. Using Proposition 3.2, we have $\|f_n - g_n\|_{\infty,E} \leq \|f - g\|_{\infty,A}$, $n \in \mathbb{N}$. We obtain the stated result by letting n tend to ∞. □

Remark 3.8. By Theorem 3.5(i), we have a new proof of the existence of a continuous solution of functional equation (5) which extends f. This proof gives some information on the speed of convergence of the process of harmonious regularization.

From now on, we shall consider only those harmonious extension schemes $\mathcal{K}$ for which the radius $r(x)$ of the ball $D(x)$ used in the description of $\mathcal{K}$ has the following form: $r(x) = \inf(h, d(x, A))$, $h > 0$, h independent of x. As the set A of data sites will not remain fixed, we shall use $r(A, x)$, $D(A, x)$ instead of $r(x)$, $D(x)$. As outlined in section 1, we prove in Proposition 3.9 below further properties of those $\mathcal{K}$. We shall denote by δ the usual Hausdorff distance between compact nonempty subsets of E.

PROPOSITION 3.9. *For any scalar-valued continuous function f whose domain* $\mathrm{dom}(f)$ *is a nonempty closed subset of E, we have the following properties:*

(i) *for any closed subset B of E whose interior does not intersect* $\mathrm{dom}(f)$*, we have*

$$\|\mathcal{K}(\mathcal{K}(f)|\partial B) - \mathcal{K}(f)\|_{\infty,B} \leq 2\hat{\omega}(f;h),$$

where ∂B denotes the boundary of B;

(ii) *for any closed subset A of E containing* $\mathrm{dom}(f)$*, we have*

$$\|\mathcal{K}(\mathcal{K}(f)|A) - \mathcal{K}(f)\|_{\infty,E} \leq 2\hat{\omega}(f;h);$$

(iii) *for any closed nonempty closed subsets A,B of* $\mathrm{dom}(f)$*, we have*

$$\|\mathcal{K}(f|A) - \mathcal{K}(f|B)\|_{\infty,E} \leq 4(\hat{\omega}(f;\delta(A,B)) + \hat{\omega}(f;h)).$$

Proof. Let us show (i). For convenience, let us set $f_1 := \mathcal{K}(\mathcal{K}(f)|\partial B)$, $f_2 := \mathcal{K}(f)$, $D_1(x) := D(\partial B, x)$, $D_2(x) := D(\mathrm{dom}(f), x)$, and let us denote by $r_j(x)$ the radius of $D_j(x)$. Note that, as the interior of B does not intersect $\mathrm{dom}(f)$, we have $D_1(x) \subset D_2(x)$ for any $x \in B$. Setting

$$\Delta_{12} := \sup_{x\in B}(f_1(x) - f_2(x)), \quad \Delta_{21} := \sup_{x\in B}(f_2(x) - f_1(x)),$$

we must show that Δ_{12} and Δ_{21} are smaller than $2\hat{\omega}(f;h)$. As these two cases are similar, let us show, without loss of generality, that $\Delta_{12} \leq 2\hat{\omega}(f;h)$. First of all, let us show that Δ_{12} is attained in the strip $K := \{x \in B : d(x,\partial B) \leq h\}$. This result is immediate if $B = K$. Otherwise, we note that, by definition of the function r, $D_1(y) = D_2(y)$ for any $y \in B\backslash K$. From this, the proof of the uniqueness of the harmonious extension in Theorem 3.3 shows that Δ_{12} is attained for an element x of B such that $d(x,\partial B) = h$, that is, for an element x of K. Now, as $\mathcal{K}$ is a harmonious extension scheme, we can write, for such an $x \in K$,

$$\Delta_{12} := f_1(x) - f_2(x) = \frac{1}{2}Q_1 + \frac{1}{2}Q_2,$$

where

$$Q_1 := \sup_{z\in D_1(x)} f_1(z) - \inf_{z\in D_2(x)} f_2(z), \quad Q_2 := \inf_{z\in D_1(x)} f_1(z) - \sup_{z\in D_2(x)} f_2(z).$$

As $x \in K$, we have $D_1(x) \cap \partial B \neq \emptyset$, since $r_1(x) = d(x,\partial B)$ in this case. Choosing $c \in D_1(x) \cap \partial B$, we obtain $Q_2 \leq \mathcal{K}(\mathcal{K}(f)|\partial B)(c) - \mathcal{K}(f)(c)$ and, therefore, $Q_2 \leq 0$, because $\mathcal{K}$ is an extension scheme. On the other hand, we can write $Q_1 = R_1 + S_1$, where

$$R_1 := \sup_{z\in D_1(x)} f_1(z) - \inf_{z\in D_2(x)} f_1(z), \quad S_1 := \inf_{z\in D_2(x)} f_1(z) - \inf_{z\in D_2(x)} f_2(z).$$

We immediately have $S_1 \leq \Delta_{12}$. Using the optimal Ω-stability of $\mathcal{K}$ (and $D_1(x) \subset D_2(x)$), we have also $R_1 \leq \hat{\omega}(f;2r_2(x))$. Since, by the definition of the function r, we have $r_2(x) \leq h$, we obtain

$$\Delta_{12} \leq \frac{1}{2}(2\hat{\omega}(f;h) + \Delta_{12}),$$

that is, $\Delta_{12} \leq 2\hat{\omega}(f;h)$, which is the desired inequality. The proofs of (ii) and (iii) use similar arguments. ∎

4. Final remark.

Remark 4.1. It is known that the processes of harmonic regularization are, in Euclidean space $\mathbb{R}^N$, associated with the heat equation

$$\frac{\partial u}{\partial t} = \sum_{i=1}^{N} \frac{\partial^2 u}{\partial x_i^2}.$$

An elementary calculation shows that the process of harmonious regularization considered in Proposition 3.9 above, for infinitesimal h, is possibly connected with the following PDE:

$$\frac{\partial u}{\partial t} = \left(\sum_{i,j=1}^{N} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} \frac{\partial^2 u}{\partial x_i \partial x_j} \right) / \sum_{i=1}^{N} \left(\frac{\partial u}{\partial x_i} \right)^2.$$

Acknowledgment. The authors would like to gratefully acknowledge the referees for their comments and suggestions.

REFERENCES

[1] J. C. Archer and E. Le Gruyer, *On the Whitney's extension theorem*, Bull. Sci. Math., 119 (1995), pp. 235–266.
[2] J. Dugundji, *An extension of Tietze's theorem*, Pacific J. Math., 1 (1951), pp. 353–367.
[3] G. Glaeser, *Etude de quelques algèbres tayloriennes*, J. Anal. Math. Jerusalem, 6 (1958), pp. 1–124.
[4] E. Le Gruyer and J. C. Archer, *Stability and convergence of extension schemes to continuous functions in general metric spaces*, SIAM J. Math. Anal., 27 (1996), pp. 274–285.
[5] E. J. Mc Shane, *Extension of range of functions*, Bull. Amer. Math. Soc., 40 (1934), pp. 837–842.
[6] W. A. Veech, *A zero-one law for a class of random walks and a converse to Gauss' mean value theorem*, Ann. of Math., 97 (1973), pp. 189–216.
[7] H. Whitney, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63–89.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 293–308, March 1998

BEHAVIORS OF SOLUTIONS FOR THE BURGERS EQUATION WITH BOUNDARY CORRESPONDING TO RAREFACTION WAVES*

TAI-PING LIU†, AKITAKA MATSUMURA‡, AND KENJI NISHIHARA§

Abstract. We investigate the asymptotic behaviors of solutions of the initial-boundary value problem to the generalized Burgers equation $u_t + f(u)_x = u_{xx}$ on the half-line with the conditions $u(0,t) = u_-, \quad u(\infty,t) = u_+$, where the corresponding Cauchy problem admits the rarefaction wave as an asymptotic state. In the present problem, because of the Dirichlet boundary, the asymptotic states are divided into five cases dependent on the signs of the characteristic speeds $f'(u_\pm)$ of the boundary state $u_- = u(0)$ and the far field state $u_+ = u(\infty)$. In all cases both global existence of the solution and the asymptotic behavior are shown without smallness conditions. New wave phenomena are observed. For instance, when $f'(u_-) < 0 < f'(u_+)$, the solution behaves as the superposition of (a part of) a viscous shock wave as boundary layer and a rarefaction wave propagating away from the boundary.

Key words. rarefaction wave, viscous shock wave, asymptotic behavior

AMS subject classifications. 35L60, 35L65

PII. S0036141096306005

1. Introduction. We consider the initial-boundary value problem (IBVP) on the half-line $\mathbf{R}_+ = (0,\infty)$ for scalar viscous conservation laws:

$$\text{(IBVP)} \qquad \begin{cases} u_t + f(u)_x = u_{xx}, \quad x \in \mathbf{R}_+, \quad t > 0, \\ u(0,t) = u_-, \quad t > 0, \\ u(x,0) = u_0(x) = \begin{cases} = u_- & x = 0, \\ \to u_+ & x \to \infty, \end{cases} \end{cases}$$

where $u_\pm$ are given constants. Here, we study the case that the corresponding Riemann problem

$$\begin{cases} u_t + f(u)_x = 0, \quad x \in \mathbf{R}, \quad t > 0, \\ u(x,0) = u_0^R(x) := \begin{cases} u_- & x < 0, \\ u_+ & x > 0 \end{cases} \end{cases} \tag{1.1}$$

yields the rarefaction wave solution

$$r^R(x/t) = \begin{cases} u_- & x \le f'(u_-)t, \\ (f')^{-1}(x/t) & f'(u_-)t \le x \le f'(u_+)t, \\ u_+ & x \ge f'(u_+)t. \end{cases} \tag{1.2}$$

This is the case when either

$$f''(u) > 0 \quad \text{for} \quad u \quad \text{under consideration, and} \quad u_- < u_+ \tag{1.3}$$

*Received by the editors June 12, 1996; accepted for publication January 6, 1997.
http://www.siam.org/journals/sima/29-2/30600.html

†Department of Mathematics, Stanford University, CA 94305 (liu@cauchy.stanford.edu). This author was supported in part by Army Research grant DAAH04-94-G-0045, NSF grant DMS-9216275-001, and Navy Research grant N00014-95-1-0468.

‡Department of Mathematics, Osaka University, Osaka 560, Japan. This author was supported in part by Grant-in-Aid for Scientific Reseach (A) of the Ministry of Education, Science, Sports, and Culture.

§School of Political Science and Economics, Waseda University, Tokyo 169-50, Japan. This author was supported in part by Waseda University grant for Special Research Project 95A-220.

or

$$f''(u) < 0 \quad \text{for} \quad u \quad \text{under consideration, and} \quad u_- > u_+.$$

Without loss of generality we assume (1.3). The problem where the corresponding Cauchy problem has a viscous shock wave has been investigated by Yu [13], Liu and Yu [7], and Liu and Nishihara [6].

Since the solution of (IBVP) has a boundary at $x = 0$, the signs of the characteristic speeds $f'(u_\pm)$ divide the asymptotic state into five cases:

$$\begin{array}{ll} (1) & f'(u_-) < f'(u_+) < 0, \\ (2) & f'(u_-) < f'(u_+) = 0, \\ (3) & f'(u_-) < 0 < f'(u_+), \\ (4) & 0 = f'(u_-) < f'(u_+), \\ (5) & 0 < f'(u_-) < f'(u_+). \end{array}$$

We also assume

$$f(0) = f'(0) = 0 \tag{1.4}$$

without loss of generality. Then, the graph of f in each case is shown in Figure 1.1.

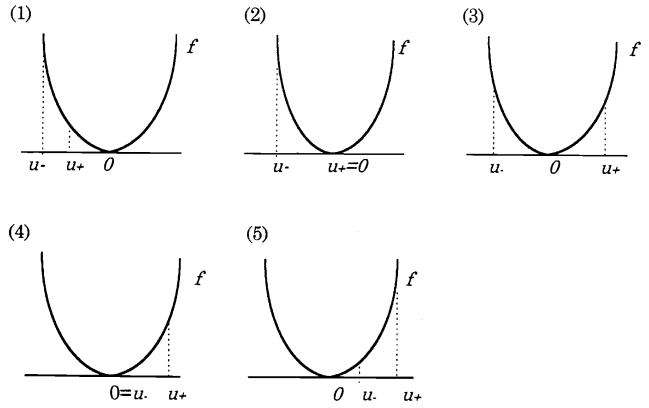

FIG. 1.1.

In the cases (1) and (2), the IBVP has a stationary solution $\phi = \phi_i(x), i = 1, 2$, respectively.

LEMMA 1.1. *Suppose that $f \in C^2$, and* (1.3) *and* (1.4) *hold. When* (1) $f'(u_-) < f'(u_+) < 0$ *or* (2) $f'(u_-) < f'(u_+) = 0$, *the boundary value problem of the ordinary differential equation*

$$\begin{cases} f(\phi)_x = \phi_{xx}, \quad x \in \mathbf{R}_+, \\ \phi(0) = u_-, \quad \phi(+\infty) = u_+ \end{cases} \tag{1.5}$$

has a unique solution $\phi_i \in C^3([0, \infty))(i = 1, 2)$, respectively, which satisfies

$$\phi_i'(x) > 0, \quad i = 1, 2, \tag{1.6}$$

$$\begin{cases} |\phi_1(x) - u_+| \le C_1 \exp(-|f'(u_+)|x) \\ |\phi_2(x) - u_+| \le C_2(1+x)^{-1} \end{cases} \tag{1.7}$$

for some constants C_1 *and* C_2.

Remark 1.1. If we take the extension $\hat{f}$ of f as Figure 1.2, then there is a unique viscous shock wave $\hat{\phi}_i(x+x_0)$ up to a shift such that

$$\begin{cases} \hat{f}(\hat{\phi})_x = \hat{\phi}_{xx}, \quad x \in \mathbf{R}, \\ \hat{\phi}(-\infty) = u_*, \quad \hat{\phi}(+\infty) = u_+. \end{cases}$$

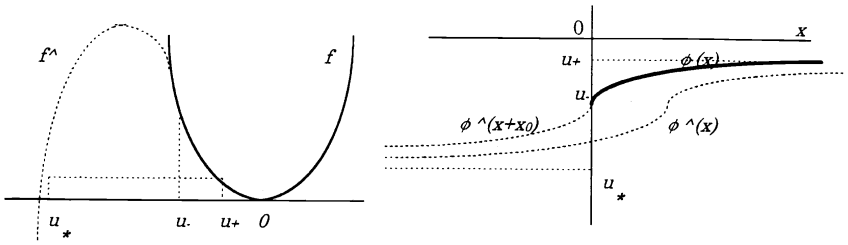

FIG. 1.2.

In fact, both the Rankine–Hugoniot condition $\frac{\hat{f}(u_+)-\hat{f}(u_*)}{u_+-u_*} = s = 0$ and the Oleinik entropy condition $-s(\hat{\phi}-u_+)+\hat{f}(\hat{\phi})-\hat{f}(u_+) < 0$ for $u_* < \hat{\phi} < u_+$ hold. Therefore, we can take $\phi_i(x) = \hat{\phi}(x+x_0)|_{\mathbf{R}_+}$, the unique profile with $\hat{\phi}(x_0) = u_-$. We also note that $\phi_i(x)(i=1,2)$ is, respectively, a part of a viscous shock wave which is nondegenerate in the case (1), degenerate in the case (2). For the details, see Liu and Nishihara [6].

Denoting the usual Lebesgue space and Sobolev space by $L^2 = L^2(\mathbf{R}_+)$ and $H^1 = H^1(\mathbf{R}_+)$, respectively, we have the first main theorem.

THEOREM 1.2 (in the case of $f'(u_-) < f'(u_+) \le 0$). *Suppose that* (1.3) *and* (1.4) *hold and that* $u_0 - \phi \in H^1$, *where* $\phi = \phi_i(i=1,2)$ *is a stationary solution obtained in Lemma* 1.1. *Then there exists a unique global solution* u *of* IBVP *such that*

$$u - \phi \in C([0,\infty); H^1), \quad (u-\phi)_x \in L^2(0,\infty; H^1)$$

and, moreover,

$$\sup_{\mathbf{R}_+} |u(x,t) - \phi(x)| \to 0 \quad as \quad t \to \infty.$$

Decay rates of $u - \phi$ are also obtained in the next section.

Next, we consider the cases (4) and (5) where the asymptotic state $\psi^R(x,t)$ is the restriction to $\mathbf{R}_+$ of the rarefaction wave $r^R(x/t)$ given by (1.2):

$$\psi^R(x,t) = r(x/t)|_{\mathbf{R}_+}. \tag{1.8}$$

THEOREM 1.3 (in the case of $0 \le f'(u_-) < f'(u_+)$). *Suppose that* (1.3) *and* (1.4) *hold and that* $u_0 - \psi^R(\cdot,0) \in H^1$. *Then, the IBVP has a unique global solution* $u(x,t)$ *which satisfies*

$$u - \psi^R \in C([0,\infty); H^1), \quad (u-\psi^R)_x, u_{xx} \in L^2(\mathbf{R}_+ \times \mathbf{R}_+),$$

and

$$\sup_{\mathbf{R}_+} |u(x,t) - \psi^R(x,t)| \to 0 \quad as \quad t \to \infty.$$

Finally, consider the case (3), which is the combination of both the cases (2) and (4). The asymptotic state is the superposition of $\phi = \phi_2(x)$ and $\psi^R = \psi_4^R(x,t)$, where ϕ_2 is the stationary solution connecting $u_-(<0)$ to 0 and ψ_4^R is the rarefaction wave connecting 0 to $u_+(>0)$. See Figure 1.3.

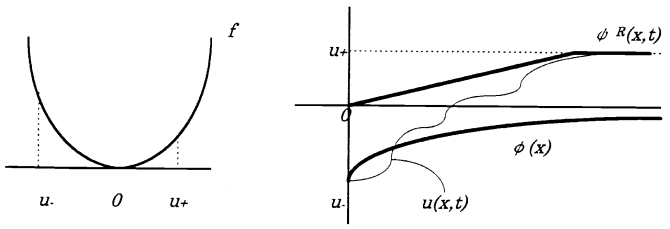

FIG. 1.3.

The following is our main theorem.

THEOREM 1.4 (in the case of $f'(u_-) < 0 < f'(u_+)$). *Suppose that* (1.3) *and* (1.4) *hold and that* $u_0 - \phi_2(\cdot) - \psi_4(\cdot,0) \in H^1$. *Then, there exists a unique global solution* $u(x,t)$ *of* (IBVP) *such that*

$$u - \phi_2 - \psi_4^R \in C([0,\infty); H^1),$$

$$(u - \phi_2 - \psi_4^R)_x, (u - \phi_2)_{xx} \in L^2(\mathbf{R}_+ \times \mathbf{R}_+),$$

and

$$\sup_{\mathbf{R}_+} |u(x,t) - \phi_2(x) - \psi_4^R(x,t)| \to 0 \quad as \quad t \to \infty.$$

Remark 1.2. As noted in Remark 1.1, ϕ_2 is a part of a viscous shock wave. Hence, the superposition of the viscous shock wave and the rarefaction wave constitutes our asymptotic state. As far as the authors know, there are few results on the stability of the superposition of different types of nonlinear waves (cf. Liu [5] for another example of such a superposition, but in an entirely different setting). In the Cauchy problem there is the question of determining the location of viscous shock waves. In the present case, the location is uniquely determined by the boundary.

Remark 1.3. Since $\phi_2(x), \psi_4^R(x,t)|_{x=t^\alpha} \to 0$ as $t \to \infty$ for $0 < \alpha < 1$, the solution $u(x,t)$ in Theorem 1.3 behaves like

$$u(x,t) \sim \begin{cases} \phi_2(x), & 0 < x \le Ct^\alpha, \\ \psi_4^R(x,t), & x \ge Ct^\alpha, \end{cases}$$

as $t \to \infty$. The asymptotic rate we will obtain is optimal when $\alpha = 1/2$.

Our plan of this paper is as follows. After stating the notation, in section 2 we investigate the cases (1) and (2), which correspond to the viscous shock waves. The cases (4) and (5) corresponding to the rarefaction waves are investigated in section 3. In the final section, we consider the case (3), which is the main part of this paper.

Notation. By $c_i, C_i (i \in \mathbf{Z}_+)$, or simply c, C, we denote several positive constants without confusion, where $\mathbf{Z}_+$ is a set of positive integers. We also denote $f(x) \sim g(x)$ as $x \to a$ when $C^{-1}g < f < Cg$ in a neighborhood of a. For function spaces, as stated above, $L^2 = L^2(\mathbf{R}_+)$ and $H^1 = H^1(\mathbf{R}_+)$ denote the usual Lebesgue space and Sobolev space with norms $||\cdot||$ and $||\cdot||_1$, respectively. For the weight function $w(x)$, L^2_w denotes the space of measurable functions f satisfying $\sqrt{w}f \in L^2$ with the norm $|f|_w = (\int_0^\infty w(x)|f(x)|^2dx)^{1/2}$. In the present paper we will use the weight function $w(x) = \langle x\rangle^\beta = (1+x^2)^{\beta/2}$, $x \geq 0$. The space $L^2_{\langle x\rangle^\beta}$ is written simply by L^2_β with norm $||\cdot||_\beta$.

2. Convergence to a viscous shock wave.

2.1. Reformulation of problem. We restate our problem

$$\text{(IBVP)} \qquad \begin{cases} u_t + f(u)_x = u_{xx}, \quad x \in \mathbf{R}_+, \quad t > 0, \\ \\ u(0,t) = u_-, \quad t > 0, \\ u(x,0) = u_0(x) = \begin{cases} = u_- & x = 0, \\ \to u_+ & x \to +\infty, \end{cases} \end{cases}$$

with $f''(u) > 0$ and $u_- < u_+ \leq 0$.

The stability theorem for the viscous shock waves to the equivalent problem has been obtained by Liu and Nishihara [5], in which the flux function f is not necessarily assumed to be convex or concave. Here, because of the convexity of f, we will obtain sharper results.

Putting $\phi = \phi_i(x)(i = 1, 2)$ and

$$u(x,t) = \phi(x) + v(x,t), \tag{2.1}$$

(IBVP) can be reformulated as

$$\begin{cases} v_t + (f(\phi+v) - f(\phi))_x = v_{xx}, \quad x \in \mathbf{R}_+, \quad t > 0, \\ \\ v(0,t) = 0, \\ \\ v(x,0) = v_0(x) := u_0(x) - \phi(x). \end{cases} \tag{2.2}$$

THEOREM 2.1 (in the case of $f'(u_-) < f'(u_+) \leq 0$). *Assume that the same conditions as those in Theorem* 1.1 *hold and that* $v_0 \in H^1$*; then there exists a unique solution* v *of* (2.2) *which satisfies*

$$v \in C([0,\infty); H^1), \quad v_x \in L^2(0,\infty; H^1),$$

$$\sup_{\mathbf{R}_+} |v(x,t)| \to 0 \quad \textit{as} \quad t \to \infty.$$

Theorem 1.1 is a direct consequence of Theorem 2.1. The combination of the local existence and a priori estimates proves Theorem 2.1.

We define the solution space by

$$X(0,T) = \{v \in C([0,T); H^1); \quad v_x \in L^2(0,T; H^1) \quad \text{with}$$

$$\partial_x^m v(x,t)|_{x=0} < +\infty \quad \text{for} \quad t \in (0,T) \quad \text{and} \quad m \in \mathbf{Z}_+\}.$$

PROPOSITION 2.2 (local existence). *For any $v_0 \in H^1$ with $||v_0||_1 \le M$, there exists a positive time T_0 and a unique solution $v \in X(0,T_0)$ of (2.2) satisfying $\sup_{0<\tau<T_0} ||v(\cdot,t)||_1 \le 2M$.*

The equation (2.2) is rewritten as an integral equation

$$v(x,t) = \int_0^\infty G(x,y;t)v_0(y)dy + \int_0^t \int_0^\infty G(x,y;t-\tau)(f(\phi+v)-f(\phi))_y dyd\tau, \tag{2.3}$$

where

$$G(x,y;t) = \frac{1}{\sqrt{4\pi t}}\left(e^{-\frac{(x-y)^2}{4t}} - e^{-\frac{(x+y)^2}{4t}}\right).$$

Making use of (2.3), Proposition 2.2 is proved in a standard way.

PROPOSITION 2.3 (a priori estimate). *Suppose that v is a solution of (2.2) in $X(0,T)$ for a positive constant T. Then, there exists a positive constant $C = C(||v_0||_1)$, independent of T, such that the solution v satisfies the estimate*

$$\sup_{0<\tau<t} ||v(\tau)||_1^2 + \int_0^t (||\sqrt{\phi_x}v(\tau)||^2 + ||v_x(\tau)||_1^2)d\tau \le C||v_0||_1^2. \tag{2.4}$$

2.2. A priori estimate. We devote ourselves to the proof of Proposition 2.2.

Let v be a solution of (2.2) in $X(0,T)$. First, multiply $(2.2)_1$ by v; then we have a divergence form

$$\begin{aligned}&(\tfrac{1}{2}v^2)_t + \left\{(f(\phi+v)-f(\phi))v - \left(\int_\phi^{\phi+v} f(s)ds - f(\phi)v\right) - v_x v\right\}_x \\ &\quad +(f(\phi+v)-f(\phi)-f'(\phi)v)\phi_x + v_x^2 = 0.\end{aligned} \tag{2.5}$$

Since $v_0 \in H^1$, $\sup_{\mathbf{R}_+} |v_0(x)| \le C_0$, and so

$$\sup_{\mathbf{R}_+} |v(x,t)| \le C_0, \quad 0 \le t < T, \tag{2.6}$$

due to the maximum principle of the parabolic equation. Hence,

$$(f(\phi+v)-f(\phi)-f'(\phi)v)\phi_x \ge \frac{c_0}{2}\phi_x v^2, \tag{2.7}$$

where $c_0 := \min_{u_- - C_0 \le u \le u_+ + C_0} f''(u) > 0$. Integrating (2.5) over $\mathbf{R}_+ \times (0,t)$ and using (2.7), we have

$$||v(t)||^2 + \int_0^t (||\sqrt{\phi_x}v(\tau)||^2 + ||v_x(\tau)||^2)d\tau \le C||v_0||^2. \tag{2.8}$$

Next, differentiate (2.2) in x and multiply the resultant equation by v_x to obtain

$$\begin{aligned}&(\tfrac{1}{2}v_x^2)_t + \{(f'(\phi+v)-f'(\phi))\phi_x v_x + \tfrac{1}{2}f'(\phi+v)v_x^2 - v_{xx}v_x\}_x \\ &\quad -(f'(\phi+v)-f'(\phi))\phi_x v_{xx} + \tfrac{1}{2}f''(\phi+v)(\phi_x+v_x)v_x^2 + v_{xx}^2 = 0.\end{aligned} \tag{2.9}$$

The integration of (2.9) over $\mathbf{R}_+ \times (0,t)$ yields

$$\begin{aligned} &\tfrac{1}{2}||v_x(t)||^2 + \int_0^t \left\{ \left(-\frac{f'(u_-)}{2} v_x^2 + v_{xx} v_x \right) |_{x=0} + ||v_{xx}(\tau)||^2 \right\} d\tau \\ &\le \tfrac{1}{2}||v_{0x}||^2 + C \int_0^t \int_0^\infty (\phi_x |v v_{xx}| + v_x^2 + |v_x|^3) dx d\tau. \end{aligned} \tag{2.10}$$

Here we have used (2.6). Since the equation (2.2) implies

$$f'(u_-) v_x(0,t) = v_{xx}(0,t), \tag{2.11}$$

we can estimate the integral on the boundary as follows:

$$\begin{aligned} &\left| \int_0^t \left(-\frac{f'(u_-)}{2} v_x^2 + v_{xx} v_x \right) \right|_{x=0} d\tau \le C \int_0^t v_x(0,\tau)^2 d\tau \\ &\le \tfrac{1}{4}||v_{xx}(\tau)||^2 d\tau + C \int_0^t ||v_x(\tau)||^2 d\tau. \end{aligned} \tag{2.12}$$

The last term of (2.10) is estimated as follows:

$$\begin{aligned} &C \int_0^t \int_0^\infty |v_x|^3 dx d\tau \le C \int_0^t ||v_{xx}(\tau)||^{1/2} ||v_x(\tau)||^{5/2} d\tau \\ &\le \tfrac{1}{2} \int_0^t ||v_{xx}(\tau)||^2 + C \sup_{0<\tau<t} ||v_x(\tau)||^{4/3} \cdot \int_0^t ||v_x(\tau)||^2 d\tau. \end{aligned} \tag{2.13}$$

It follows from (2.10)–(2.13) and (2.8) that

$$\sup_{0<\tau<t} ||v_x(\tau)||^2 + \int_0^t ||v_{xx}(\tau)||^2 d\tau \le C||v_0||_1^2 + C||v_0||^2 \cdot \sup_{0<\tau<t} ||v_x(\tau)||^{4/3},$$

which yields

$$\sup_{0<\tau<t} ||v_x(\tau)||^2 + \int_0^t ||v_{xx}(\tau)||^2 d\tau \le C||v_0||_1^2. \tag{2.14}$$

The combination (2.8) with (2.14) proves the estimate (2.4).

2.3. Convergence rates to viscous shock wave. First, consider the case (1), that is, the nondegenerate shock case. Note that $f'(u) \le -c_1 < 0$ for $u_+ \le u \le u_- < 0$. By Theorem 1.2

$$\sup_{\mathbf{R}_+} |v(x,t)| \to 0 \quad \text{as} \quad t \to \infty,$$

and hence, for any $\varepsilon > 0$, there is a positive time $t_1 = t_1(\varepsilon)$ such that

$$\sup_{\mathbf{R}_+} |v(x,t)| \le \varepsilon \quad \text{for} \quad t \ge t_1. \tag{2.15}$$

Multiplying (2.5) by $\langle x\rangle^\beta = (1+x)^{\beta/2}$, we have

$$
\begin{aligned}
&(\tfrac{\langle x\rangle^\beta}{2}v^2)_t + (\langle x\rangle^\beta\{\cdots\})_x \\
&-\beta x\langle x\rangle^{\beta-2}\left\{(f(\phi+v)-f(\phi))v - \left(\int_\phi^{\phi+v} f(s)ds - f(\phi)v\right) - v_x v\right\} \qquad (2.16)\\
&+\langle x\rangle^\beta(f(\phi+v)-f(\phi)-f'(\phi)v)\phi_x + \langle x\rangle^\beta v_x^2 = 0.
\end{aligned}
$$

By virtue of (2.15), for sufficiently small $\varepsilon_0 > 0$ and $t \ge t_1(\varepsilon_0)$,

$$
\begin{aligned}
&-\left\{(f(\phi+v)-f(\phi))v - \left(\int_\phi^{\phi+v} f(s)ds - f(\phi)v\right)\right\} \qquad (2.17)\\
&= -(f'(\phi+\theta_1 v) - \tfrac{1}{2}f'(\phi+\theta_2 v))v^2 \ge \tfrac{c_1}{2}v^2, \quad \text{where} \quad |\theta_i| < 1, \quad i = 1,2.
\end{aligned}
$$

The integration of (2.16) over $(0,\infty)\times(0,t)$, $t \le t_1(\varepsilon_0)$, gives

$$
\begin{aligned}
&\int_0^\infty \langle x\rangle^\beta v(x,t)^2 dx + \int_0^t\int_0^\infty \langle x\rangle^\beta(\phi_x v^2 + v_x^2)dxd\tau \qquad (2.18)\\
&\le C\left(\int_0^\infty \langle x\rangle^\beta v_0(x)^2 dx + \int_0^t\int_0^\infty \langle x\rangle^{\beta-1}(v^2 + |vv_x|)dxd\tau\right),
\end{aligned}
$$

and hence

$$
|v(t)|_\beta^2 + \int_0^t |v_x(\tau)|_\beta^2 d\tau \le C(t_1)|v_0|_\beta^2, \quad t \le t_1(\varepsilon_0). \qquad (2.19)
$$

Multiplying (2.16) by $(1+t-t_1)^\gamma$ and integrating the resultant equation over $(0,\infty)\times[t_1,t)$, we have

$$
\begin{aligned}
&(1+t-t_1)^\gamma|v(t)|_\beta^2 + \int_{t_1}^t (1+t-t_1)^\gamma(\beta|v(\tau)|_{\beta-1}^2 + |v_x(\tau)|_\beta^2)d\tau \\
&\le C(|v(t_1)|_\beta^2 + \gamma\int_{t_1}^t (1+\tau-t_1)^{\gamma-1}|v(\tau)|_\beta^2 d\tau) \qquad (2.20)\\
&+C\int_{t_1}^t (1+\tau-t_1)^\gamma\int_0^\infty \langle x\rangle^{\beta-1}|vv_x|dxd\tau.
\end{aligned}
$$

The final term of (2.20) is estimated as follows:

$$
\begin{aligned}
&C\int_{t_1}^t (1+\tau-t_1)^\gamma\int_0^\infty \langle x\rangle^{\beta-1}|vv_x|dxd\tau \qquad (2.21)\\
&\le \tfrac{\beta}{2}\int_{t_1}^t (1+\tau-t_1)^\gamma|v(\tau)|_{\beta-1}^2 d\tau + C\int_{t_1}^t (1+\tau-t_1)^\gamma\left(\int_0^R + \int_R^\infty\right)\langle x\rangle^{\beta-1}v_x^2 dxd\tau \\
&\le \tfrac{1}{2}\int_{t_1}^t (1+\tau-t_1)^\gamma(\beta|v(\tau)|_{\beta-1}^2 + |v_x(\tau)|_\beta^2)d\tau + C_R\int_{t_1}^t (1+\tau-t_1)^\gamma||v_x(\tau)||^2 d\tau
\end{aligned}
$$

for sufficiently large R. Combining (2.19)–(2.21), we have

$$
\begin{aligned}
&(1+t-t_1)^\gamma |v(t)|_\beta^2 + \int_{t_1}^t (1+\tau-t_1)^\gamma (\beta |v(\tau)|_{\beta-1}^2 + |v_x(\tau)|_\beta^2) d\tau \\
&\le C(|v_0|_\beta^2 + \gamma \int_{t_1}^t (1+\tau-t_1)^{\gamma-1} |v(\tau)|_\beta^2 d\tau + \int_{t_1}^t (1+\tau-t_1)^\gamma ||v_x(\tau)||^2 d\tau).
\end{aligned} \tag{2.22}
$$

This basic weighted energy estimate leads to the following lemma, using the procedure of reduction of Kawashima and Matsumura [2] and Matsumura and Nishihara [10].

LEMMA 2.4. *Suppose* (1.3), (1.4), *and* $f'(u_-) < f'(u_+) < 0$. *If* $v_0 \in H^1 \cap L^2_\alpha$, *then the solution* v *of* (2.2) *satisfies*

$$
(1+t)^\gamma ||v(t)||^2 + \int_0^t (1+\tau)^\gamma ||v_x(\tau)||^2 d\tau \le C|v_0|_\alpha^2 \tag{2.23}
$$

for any $\gamma \le \alpha$ *(*α*:integer) or* $\gamma < \alpha$ *(*α*:noninteger).*

For the derivative of v in x, we can easily show the similar estimate. Thus we obtain the following theorem.

THEOREM 2.5 (rate of asymptotics for $f'(u_-) < f'(u_+) < 0$). *Suppose that* (1.3) *and* (1.4) *hold and that* $f'(u_-) < f'(u_+) < 0$. *If* $u_0 - \phi_1 \in H^1 \cap L^2_\alpha$, *then the solution* $u(x,t)$ *of* (IBVP) *satisfies*

$$
\sup_{\mathbf{R}_+} |u(x,t) - \phi_1(x)| \le C_\varepsilon (1+t)^{-\frac{\alpha}{2}+\varepsilon} (||u_0 - \phi_1||_1 + |u_0 - \phi_1|_\alpha), \tag{2.24}
$$

where $\varepsilon = 0$ *if* α *is integer and* $\varepsilon > 0, C_\varepsilon \to \infty (\varepsilon \to 0)$ *if* α *is not integer.*

Remark 2.1. Nishikawa [12] has recently improved the result of Matsumura and Nishihara [10], so we can take $\varepsilon = 0$ even for noninteger α by the same method as his.

Remark 2.2. If $V_0 := \int_0^x v_0(y)dy = \int_0^x (u_0(y) - \phi_1(y))dy \in H^2$ is sufficiently small, then the reformulated problem

$$
\begin{cases}
V_t + f'(\phi_1)V_x - V_{xx} = -(f(\phi_1 + V_x) - f(\phi_1) - f'(\phi_1)V_x), \quad x \in \mathbf{R}_+, \quad t > 0, \\
V(x,0) = V_0(x), \\
V_x(0,t) = 0
\end{cases}
$$

has a unique global solution $V(x,t) \in C([0,\infty); H^2)$ satisfying

$$
\sup_{\mathbf{R}_+} |V_x(x,t)| = \sup_{\mathbf{R}_+} |u(x,t) - \phi_1(x)| \to 0 \quad \text{as} \quad t \to \infty \tag{2.25}
$$

(see Liu and Nishihara [6]). Moreover, if $V_0 \in L^2_\alpha$, then they conclude

$$
\sup_{\mathbf{R}_+} |u(x,t) - \phi_1(x)| \le C_\varepsilon (1+t)^{-\frac{\alpha}{2}+\varepsilon}. \tag{2.26}
$$

Here, ε and C_ε are same as those in Theorem 2.1. In the present case the assumptions $V_0 \in L^2_\alpha$ and $V_{0x} \in L^2_{\alpha+1}$, which seem to be reasonable, improve the rate of asymptotics as

$$
\sup_{\mathbf{R}_+} |u(x,t) - \phi_1(x)| \le C_\varepsilon (1+t)^{-\frac{\alpha+1}{2}+\varepsilon}. \tag{2.27}
$$

See also Nishihara and Rajopadhye [11].

Remark 2.3. If $\int_0^\infty e^{\delta x}|u_0(x)-\phi_1(x)|^2dx<+\infty$ for some $\delta>0$, then it also holds that the solution $u(x,t)$ of (IBVP) satisfies

$$\sup_{\mathbf{R}_+}|u(x,t)-\phi_1(x)|\le Ce^{-\delta' t}$$

for some positive constant δ'.

Similar considerations for the degenerate shock case (2) corresponding to Remark 2.2 are still available. However, since $f'(\phi_2(x)+v(x,t))$ is not uniformly negative, the procedure for the case (1) in this section is not directly applicable. Without going into the details we obtain the following decay properties for small data in the case (2) by the same arguments as in [10].

THEOREM 2.6 (rate of asymptotics for $f'(u_-)<f'(u_+)=0$). *Suppose that* (1.3) *and* (1.4) *hold and that* $f'(u_-)<f'(u_+)=0$. *If* $V_0:=\int_0^x(u_0(y)-\phi_2(y))dy\in H^2\cap L^2_\alpha$ *and* $V_{0x}\in L^2_{\alpha+1}$ *for* $\alpha<2$ *are small, then the solution* $u(x,t)$ *of* (IBVP) *satisfies*

$$\sup_{\mathbf{R}_+}|u(x,t)-\phi_2(x)|\le C_\varepsilon(1+t)^{-\frac{\alpha+2}{4}+\varepsilon},$$

where $\varepsilon>0$ *and* $C_\varepsilon\to\infty$ *as* $\varepsilon\to\infty$.

3. Convergence to rarefaction wave.

3.1. Reformulated problem in the case of $0\le f'(u_-)<f'(u_+)$. In this section we consider the cases (4) and (5), that is, $0\le f'(u_-)<f'(u_+)$. Since the asymptotic state $\psi^R(x,t)=r^R(x/t)|_{\mathbf{R}_+}$ is not smooth, we construct the smooth approximation $\psi=\psi_i(x,t)(i=4,5)$.

For the case (4) we prepare the following lemma.

LEMMA 3.1. *Suppose that* (1.3) *and* (1.4) *hold and that* $0=f'(u_-)<f'(u_+)$. *Let* $w(x,t)$ *be a unique smooth solution of the Cauchy problem*

$$\text{(3.1)}\qquad \begin{cases} w_t+ww_x=0, \quad x\in\mathbf{R}, \quad t>0,\\ \\ w(x,0)=w_{40}(x)=w_+\cdot\kappa_q\displaystyle\int_0^x(1+y^2)^{-q}dy, \quad q>1/2, \end{cases}$$

where $w_+=f'(u_+)>0$ *and* $\kappa_q\int_0^\infty(1+y^2)^{-q}dy=1$. *Then,* $\psi=\psi_4(x,t):=(f')^{-1}(w(x,t))|_{\mathbf{R}_+}$ *satisfies*

$$\text{(3.2)}\qquad \begin{cases} \psi_t+f(\psi)_x=0, \quad x\in\mathbf{R}_+, \quad t>0,\\ \\ \psi(0,t)=0(=u_-),\\ \psi(x,0)=\psi_{40}(x):=(f')^{-1}(w_{40}(x))=\begin{cases} =0, & x=0,\\ \to u_+, & x\to+\infty \end{cases} \end{cases}$$

and the following:

(i) $0=u_-\le\psi(x,t)<u_+, \quad \psi_x(x,t)>0, \quad (x,t)\in\mathbf{R}_+\times(0,\infty)$.

(ii) *For any* $1\le p\le\infty$ *there exists a constant* $C_{p,q}$ *such that*
$||\psi_x(t)||_{L^p}\le C_{p,q}\min(u_+,u_+^{1/p}t^{-1+\frac{1}{p}})$,
$||\psi_{xx}(t)||_{L^p}\le C_{p,q}\min(u_+,u_+^{-\frac{p-1}{2pq}}t^{-1-\frac{p-1}{2pq}})$.

(iii) $\lim_{t\to\infty}\sup_{\mathbf{R}_+}|\psi(x,t)-\psi_4^R(x,t)|=0$.

The proof of Lemma 3.1 is given by the characteristic curve method. See Matsumura and Nishihara [8], [9].

Set

$$u(x,t) = \psi_4(x,t) + v(x,t); \tag{3.3}$$

then the perturbation v satisfies

$$\begin{cases} v_t + (f(\psi_4 + v) - f(\psi_4))_x - v_{xx} = \psi_{4xx}, \quad x \in \mathbf{R}_+, \quad t > 0, \\ \\ v(0,t) = 0, \\ \\ v(x,0) = v_0(x) := v_0(x) - \psi_{40}(x). \end{cases} \tag{3.4}$$

Our theorem for the case (4) is the following.

THEOREM 3.2 (in the case of $0 = f'(u_-) < f'(u_+)$). *Suppose that* (1.3) *and* (1.4) *hold and that* $0 = f'(u_-) < f'(u_+)$. *If* $v_0 \in H^1$, *then there exists a unique solution* v *of* (3.4) *such that*

$$v \in C([0,\infty); H^1), \quad v_x \in L^2(0,\infty; H^1) \tag{3.5}$$

and, moreover,

$$\sup_{\mathbf{R}_+} |v(x,t)| \to 0 \quad \textit{as} \quad t \to \infty. \tag{3.6}$$

Theorem 1.2, in the case of $0 = f'(u_-) < f'(u_+)$, is a direct consequence of Theorem 3.2, Lemma 3.1 (iii), and (3.3).

In the case of $0 < f'(u_-) < f'(u_+)$, we prepare the following lemma in place of Lemma 3.1.

LEMMA 3.3. *Suppose that* (1.3) *and* (1.4) *hold and that* $0 < f'(u_-) < f'(u_+)$. *Let* $w(x,t)$ *be a unique global solution of the Cauchy problem*

$$\begin{cases} w_t + ww_x = 0, \quad x \in \mathbf{R}, \quad t > 0, \\ \\ w(x,0) = w_{50}(x) := \frac{w_+ + w_-}{2} + \frac{w_+ - w_-}{2}\kappa_q \displaystyle\int_0^x (1+y^2)^{-q} dy, \quad q > 3/2, \end{cases} \tag{3.7}$$

where $w_\pm = f'(u_\pm) > 0$ *and* $\kappa_q \int_0^\infty (1+y^2)^{-q} dy = 1$. *Then,* $\psi = \psi_5(x,t) := (f')^{-1}(w(x,t))|_{\mathbf{R}_+}$ *satisfies*

$$\begin{cases} \psi_t + f(\psi)_x = 0, \quad x \in \mathbf{R}_+, \quad t > 0, \\ \\ \psi(x,0) = \psi_{50}(x) := (f')^{-1}(w_{50}(x)), \end{cases} \tag{3.8}$$

and the following:

(i) $0 < u_- < \psi(x,t) < u_+, \quad \psi_x(x,t) > 0, \quad (x,t) \in \mathbf{R}_+ \times (0,\infty)$.

(ii) *For any* $1 \le p \le \infty$ *there exists a constant* $C_{p,q}$ *such that*
$||\psi_x(t)||_{L^p} \le C_{p,q} \min(|u_+ - u_-|, |u_+ - u_-|^{1/p} t^{-1+\frac{1}{p}})$,
$||\psi_{xx}(t)||_{L^p} \le C_{p,q} \min(|u_+ - u_-|, |u_+ - u_-|^{-\frac{p-1}{2pq}} t^{-1-\frac{p-1}{2pq}})$.

(iii) *For some constant* C_q
$|\psi(0,t) - u_-| \le C_q |u_+ - u_-| (1 + (|u_+ - u_-|t)^2)^{-q/3}$,
$|\psi_x(0,t)| \le C_q |u_+ - u_-| (1 + (|u_+ - u_-|t)^2)^{-q/2}$.

(iv) $\lim_{t\to\infty} \sup_{\mathbf{R}_+} |\psi(x,t) - \psi_5^R(x,t)| = 0$.

For the proof see Matsumura and Nishihara [8], [9].

The main difference between Lemmas 3.1 and 3.3 is the boundary value $\psi(0,t)$. The perturbation

$$v(x,t) = u(x,t) - \psi_5(x,t) \tag{3.9}$$

has a "boundary layer $u_- - \psi_5(0,t)$" at $x = 0$:

$$\begin{cases} v_t + (f(\psi_5 + v) - f(\psi_5))_x - v_{xx} = \psi_{5xx}, \quad x \in \mathbf{R}_+, \quad t > 0, \\ v(0,t) = u_- - \psi_5(0,t), \\ v(x,0) = v_0(x) := u_0(x) - \psi_{50}(x). \end{cases} \tag{3.10}$$

However, Lemma 3.3 (iii) shows that $u_- - \psi_5(0,t) \in L^1(\mathbf{R}_+)$ and $\psi_{5x}(0,t) \in L^1(\mathbf{R}_+)$ in t, from which we have the following theorem.

THEOREM 3.4 (in the case of $0 < f'(u_-) < f'(u_+)$). *Suppose that* (1.3) *and* (1.4) *hold and that* $0 < f'(u_-) < f'(u_+)$. *If* $v_0 \in H^1$, *then there exists a unique solution* v *of* (3.10) *such that*

$$v \in C([0,\infty); H^1), \quad v_x \in L^2(0,\infty; H^1) \tag{3.11}$$

and, moreover,

$$\sup_{\mathbf{R}_+} |v(x,t)| \to \infty \quad \textit{as} \quad t \to \infty. \tag{3.12}$$

We note that (3.12), together with Lemma 3.3 (iii) and (iv), yields that $\sup_{\mathbf{R}_+} |u(x,t) - \psi_5^R(x,t)| \to 0$ as $t \to \infty$. Hence, Theorem 1.2 in the case of $0 < f'(u_-) < f'(u_+)$ follows from Theorem 3.2.

In the next section we devote ourselves to the proof of Theorem 3.2. The proof of Theorem 3.4 is a little bit more complicated than that of Theorem 3.2. However, it is along the same line and is omitted.

3.2. Proof of Theorem 3.2. We can easily show the local existence of the solution of (3.4) in the solution space

$$X_4(0,T) = \{v \in C([0,T]; H^1), \quad v_x \in L^2(0,T; H^1) \quad \text{and}$$

$$\partial_x^m v(x,t)|_{x=0} < +\infty \quad \text{for} \quad t \in (0,T] \quad \text{and} \quad m \in \mathbf{Z}_+\}.$$

It remains to show the a priori estimates.

PROPOSITION 3.5 (a priori estimate). *Suppose* v *is a solution of* (3.4) *in* $X_4(0,T)$. *Then there exists a positive constant* C, *independent of* T, *satisfying*

$$||v(t)||_1^2 + \int_0^t (||\sqrt{\psi_4(\tau)}v(\tau)||^2 + ||v_x(\tau)||_1^2)d\tau \le C(||v_0||_1^2 + 1). \tag{3.13}$$

Remark 3.1. By virtue of (3.13) we note that

$$\int_0^t v_x(0,\tau)^2 d\tau \le C \int_0^t ||v_x(\tau)||||v_{xx}(\tau)||d\tau \le C(||v_0||_1^2 + 1).$$

Proof. Multiplying (3.4) by v and integrating the resultant equation over $\mathbf{R}_+ \times (0,t)$, we have

$$\begin{aligned}&\tfrac{1}{2}||v(t)||^2 + \int_0^t \left\{ \int_0^\infty (f(\psi_4+v) - f(\psi_4) - f'(\psi_4)v)\psi_{4x}dx + ||v_x(\tau)||^2 \right\} d\tau \\ &\qquad \le \tfrac{1}{2}||v_0||^2 + \int_0^t ||\psi_{4xx}(\tau)||||v(\tau)||d\tau.\end{aligned} \tag{3.14}$$

The maximum principle for v and the Gronwall inequality with (3.14) yield

$$||v(t)||^2 + \int_0^t (||\sqrt{\psi_4(\tau)}v(\tau)||^2 + ||v_x(\tau)||^2)d\tau \le C(||v_0||^2 + 1). \tag{3.15}$$

Next, to estimate the derivative of v, we have the relation at the boundary

$$v_{xx}(0,t) = -\psi_{xx}(0,t) \tag{3.16}$$

from the equation (3.4). With (3.16) we calculate $\int_0^t \int_0^\infty \frac{\partial}{\partial x}(3.4)_1 \cdot v_x dx d\tau$ and obtain

$$\begin{aligned}&\tfrac{1}{2}||v_x(t)||^2 + \int_0^t ||v_{xx}(\tau)||^2 d\tau \\ &\le \tfrac{1}{2}||v_{0x}||^2 + C\int_0^t \int_0^\infty |\psi_{4x}vv_{xx} + v_xv_{xx} + \psi_{4xx}v_{xx}|dxd\tau \\ &\qquad + \int_0^t (-v_{xx}(0,\tau) - \psi_{4xx}(0,\tau))v_x(0,\tau)d\tau \\ &\le C\left(||v_{0x}||^2 + 1 + \int_0^t (||\sqrt{\psi_4(\tau)}v(\tau)||^2 + ||v_x(\tau)||^2)d\tau \right).\end{aligned} \tag{3.17}$$

Combining (3.15) with (3.17), we conclude (3.13). □

4. Asymptotics to superposition of nonlinear waves.

4.1. Reformulation of the problem. Referring to the preceding sections, we take

$$\Phi_3(x,t) := \phi_2(x) + \psi_4(x,t) \tag{4.1}$$

as an asymptotic state at $t = \infty$, instead of $\phi_2(x) + \psi_4^R(x,t)$, where ϕ_2 and ψ_4 are, respectively, given in Lemmas 1.1 and 3.1.

The perturbation

$$v(x,t) = u(x,t) - \Phi_3(x,t) = u(x,t) - \phi_2(x) - \psi_4(x,t) \tag{4.2}$$

satisfies the reformulated problem

$$\begin{cases} v_t + (f(\Phi_3 + v) - f(\Phi_3))_x - v_{xx} = F, \\ v(0,t) = 0, \\ v(x,0) = v_0(x) := u_0(x) - \phi_2(x) - \psi_4(x,0), \end{cases} \tag{4.3}$$

where

$$F = -(f'(\phi_2+\psi_4)-f'(\phi_2))\phi_{2x} - (f'(\phi_2+\psi_4)-f'(\psi_4))\psi_{4x} + \psi_{4xx}. \tag{4.4}$$

Our final theorem is the following.

THEOREM 4.1 (in the case of $f'(u_-) < 0 < f'(u_+)$). *Suppose that* (1.3) *and* (1.4) *hold and that* $f'(u_-) < 0 < f'(u_+)$. *If* $v_0 \in H^1$, *then there exists a unique solution* v *of* (4.3) *which satisfies*

$$v \in C([0,\infty); H^1), \quad v_x \in L^2(0,\infty; H^1), \tag{4.5}$$

and

$$\sup_{\mathbf{R}_+} |v(x,t)| \to 0 \quad \textit{as} \quad t \to \infty. \tag{4.6}$$

The main Theorem 1.3 is a direct consequence of Theorem 4.1. As for the above theorems, Theorem 4.1 is proved by the local existence theorem with the a priori estimates. The solution space is

$$X_3(0,T) = \{v \in C([0,T]; H^1); v_x \in L^2(0,T; H^1) \quad \text{and}$$

$$\partial_x^m v(x,t)|_{x=0} < +\infty \quad \text{for} \quad t \in (0,T] \quad \text{and} \quad m \in \mathbf{Z}_+\}.$$

We will devote ourselves to the a priori estimates in the next section.

PROPOSITION 4.2 (a priori estimates). *Suppose that* $v \in X_3(0,T)$ *is a solution of* (4.3). *Then there exists a positive constant* C, *independent of* T, *satisfying*

$$||v(t)||_1^2 + \int_0^t (||\sqrt{\Phi_3(\tau)}v(\tau)||^2 + ||v_x(\tau)||_1^2)d\tau \le C(||v_0||_1^2 + 1). \tag{4.7}$$

4.2. Proof of Proposition 4.2. Multiplying $(4.3)_1$ by v and integrating the resultant equation over $\mathbf{R}_+$, we have

$$\frac{1}{2}\frac{d}{dt}||v(t)||^2 + \int_0^\infty (f(\Phi+v) - f(\Phi) - f'(\Phi)v)\Phi_x dx + ||v_x(t)||^2 = \int_0^\infty Fv dx. \tag{4.8}$$

(We drop the suffices "2", "3", and "4".) Since $\Phi_x = \phi_x + \psi_x > 0$ and $f''(\Phi+v) \ge c_0 > 0$ by the maximum principle, (4.8) gives

$$\frac{d}{dt}||v(t)||^2 + ||\sqrt{\Phi_x(t)}v(t)||^2 + ||v_x(t)||^2 \le C\left|\int_0^\infty Fv dx\right|. \tag{4.9}$$

We estimate the last term of (4.9) using (4.4). First,

$$\begin{aligned} &C|-\int_0^\infty (f'(\phi+\psi)-f'(\phi))\phi_x v dx| \\ &\le C\int_0^\infty \psi\phi_x|v|dx = \int_0^{f'(u_+)t} + \int_{f'(u_+)t}^\infty := I_1 + I_2. \end{aligned} \tag{4.10}$$

By virtue of $\phi < 0, \psi > 0$, and Lemmas 1.1 and 3.1(ii), we have

$$
\begin{aligned}
I_1 &\le C \sup_{\mathbf{R}_+} |v| \cdot \{[\phi\psi]_0^{f'(u_+)t} + \int_0^{f'(u_+)t} (-\phi)\psi_x dx\} \\
&\le C||v(t)||^{1/2}||v_x(t)||^{1/2}(1+t)^{-1} \int_0^{f'(u_+)t} \frac{dx}{1+x} \\
&\le \tfrac{1}{8}||v_x(t)||^2 + C\{(1+t)^{-1}\log(2+t)\}^{4/3}(||v(t)||^2+1),
\end{aligned} \tag{4.11}
$$

and

$$
\begin{aligned}
I_2 &\le C \sup_{\mathbf{R}_+} \cdot u_+ \int_{f'(u_+)t}^{\infty} \phi_x(x)dx \le C||v(t)||^{1/2}||v_x(t)||^{1/2}(1+t)^{-1} \\
&\le \tfrac{1}{8}||v_x(t)||^2 + C(1+t)^{-4/3}(||v(t)||^2+1).
\end{aligned} \tag{4.12}
$$

Secondly, in a similar fashion to (4.11) and (4.12),

$$
\begin{aligned}
&C|-\int_0^{\infty} (f'(\phi+\psi)-f'(\psi))\psi_x v dx| \\
&\le \tfrac{1}{4}||v_x(t)||^2 + C\{(1+t)^{-4/3} + ((1+t)^{-1}\log(1+t))^{4/3}\}(||v(t)||^2+1).
\end{aligned} \tag{4.13}
$$

Thirdly,

$$
\left|\int_0^{\infty} \psi_{xx} v dx\right| \le ||\psi_{xx}(t)||||v(t)||. \tag{4.14}
$$

Substituting (4.10)–(4.14) into (4.9) and integrating it over $(0,t)$, we have the desired estimate

$$
||v(t)||^2 + \int_0^t (||\sqrt{|\Phi_x(\tau)|}v(\tau)||^2 + ||v_x(\tau)||^2)d\tau \le C(||v_0||^2+1). \tag{4.15}
$$

Here we have used the Gronwall inequality.

Next, differentiate $(4.4)_1$ in x:

$$
v_{xt} + (f'(\Phi+v)v_x)_x - v_{xxx} = F_x - \{(f'(\Phi+v)-f'(\Phi))\Phi_x\}_x. \tag{4.16}
$$

By the equation (4.3) with (4.4) we have the relation at the boundary

$$
-f'(u_-)v_x(0,t) + v_{xx}(0,t) = f'(u_-)\psi_x(0,t) - \psi_{xx}(0,t). \tag{4.17}
$$

Multiplying (4.16) by v_x and integrating it over $(0,\infty)$, we have

$$
\begin{aligned}
&\tfrac{1}{2}\tfrac{d}{dt}||v_x(t)||^2 + (-\tfrac{1}{2}f'(u_-)v_x^2 + v_{xx}v_x)|_{x=0} + ||v_{xx}(t)||^2 \\
&\le (f'(u_-)\psi_x(0,t) - \psi_{xx}(0,t))v_x(0,t) \\
&\quad + C\int_0^{\infty} \{(1+|v_x|)v_x^2 + (|F| + \Phi_x|v|)|v_{xx}|\}dx.
\end{aligned} \tag{4.18}
$$

By virtue of (4.17) and $v_x(0,t)^2 \le C||v_x(t)||||v_{xx}(t)||$, the integration of (4.18) over $(0,t)$ yields

$$||v_x(t)||^2 + \int_0^t ||v_{xx}(\tau)||^2 d\tau$$

$$\le \{||v_{0x}||^2 + 1 + \int_0^\infty (||\sqrt{\Phi_x(\tau)}v(\tau)||^2 + ||v_x(\tau)||^2)d\tau\}$$

$$+C\int_0^\infty ||v_x(0,\tau)||^2 d\tau \cdot \sup_{0<\tau<t} ||v_x(0,\tau)||^{4/3}$$

and hence, by (4.15),

$$\sup_{0<\tau<t} ||v_x(\tau)||^2 + \int_0^t ||v_{xx}(\tau)||^2 d\tau \le C(||v_0||_1^2 + 1). \tag{4.19}$$

The combination of (4.15) and (4.19) gives a desired estimate (4.7), which completes the proof of Proposition 4.2.

REFERENCES

[1] A. M. Il'in and O. A. Oleinik, *Asymptotic behavior of the solution of the Cauchy problem for certain quasilinear equations for large time*, Mat. Sbornik, 51 (1960), pp. 191–216 (in Russian).

[2] S. Kawashima and A. Matsumura, *Asymptotic stability of traveling wave solutions of systems for one-dimensional gas motion*, Comm. Math. Phys., 101 (1985), pp. 979–127.

[3] S. Kawashima and A. Matsumura, *Stability of shock profiles in viscoelasticity with non-convex constitutive relations*, Comm. Pure Appl. Math., 47 (1994), pp. 1547–1569.

[4] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[5] T.-P. Liu, *Riemann problem for viscous conservation laws*, to appear.

[6] T.-P. Liu and K. Nishihara, *Asymptotic behavior for scalar viscous conservation laws with boundary effect*, J. Differential Equations, 133 (1997), pp. 296–320.

[7] T.-P. Liu and S.-H. Yu, *Propagation of stationary viscous Burgers shock under the effect of boundary*, Arch. Rational Mech. Anal., 139 (1997), pp. 57–82.

[8] A. Matsumura and K. Nishihara, *Asymptotics toward the rarefaction waves of the solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 3 (1986), pp. 1–13.

[9] A. Matsumura and K. Nishihara, *Global stability of the rarefaction wave of a one-dimensional model system for compressible viscous gas*, Comm. Math. Phys., 144 (1992), pp. 325–335.

[10] A. Matsumura and K. Nishihara, *Asymptotic stability of traveling waves for scalar viscous conservation laws with non-convex nonlinearity*, Comm. Math. Phys., 165 (1994), pp. 83–96.

[11] K. Nishihara and S. V. Rajopadhye, *Asymptotic behvior of solutions to the Korteweg-de Vries-Burgers equation*, Differential Integral Equations, to appear.

[12] M. Nishikawa, *Convergence rate to the traveling wave for viscous conservation laws*, Funkcial. Ekvac., to appear.

[13] S.-H. Yu, *The Asymptotic Behavior of the Burgers Equation on the Quarter Plane*, preprint.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 309–320, March 1998

002

SOME OVERDETERMINED BOUNDARY VALUE PROBLEMS WITH ELLIPTICAL FREE BOUNDARIES*

ANTOINE HENROT† AND GÉRARD A. PHILIPPIN‡

Abstract. In this paper we study three different overdetermined boundary value problems in $\mathbf{R}^2$: a problem of torsion, a problem of electrostatic capacity, and a problem of polarization. In each case we prove that a solution exists if and only if the free boundary is an ellipse. The techniques we use rely on classical complex function theory, maximum principle, and some topological argument.

Key words. overdetermined partial differential equation, free boundary, conformal map, ellipse

AMS subject classifications. 35N10, 35R35, 30E25

PII. S0036141096307217

1. Introduction.

In his reference book on potential theory, O. D. Kellogg [Ke] has established that the density of charge at any point of an ellipsoid is proportional to the distance from the center to the tangent plane at that point.

The above mentioned property may also be formulated as follows:

$$|\nabla u| = \text{const.}h \quad \text{on } \partial\Omega, \tag{1}$$

where u is the electrostatic potential of the ellipsoid

$$\Omega := \left\{ \mathbf{x} = (x, y, z) \in \mathbf{R}^3 \,\middle|\, \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} < 1 \right\},$$

defined as the solution of the boundary value problem

$$\Delta u = 0 \quad \text{in } \mathbf{R}^3 \backslash \bar{\Omega}, \tag{2}$$

$$u = 1 \quad \text{on } \partial\Omega, \tag{3}$$

$$u = O\left(\frac{1}{r}\right) \text{ as } r \to \infty, \tag{4}$$

and $h = \mathbf{x}.\mathbf{n}$ is the scalar product of $\mathbf{x}$ with the exterior normal vector of $\partial\Omega$. This property leads to the following question: is the overdetermined boundary value problem (1)–(4) solvable only if Ω is an ellipsoid?

We note that the standard methods of investigation in the topics of overdetermined problems, such as, e.g., Serrin's moving plane method [Se] or even the maximum principle approach (see, e.g., [We]), may not be appropriate to characterize ellipsoids!

In this paper we are not going to investigate such challenging overdetermined problems in $\mathbf{R}^3$. However we shall analyze similar problems in $\mathbf{R}^2$, in order to take advantage of such powerful tools as the conformal mapping techniques.

*Received by the editors July 22, 1996; accepted for publication (in revised form) December 10, 1996.

http://www.siam.org/journals/sima/29-2/30721.html

†Equipe de Mathématiques, UMR CNRS 6623, Université de Franche-Comté Route de Gray, 25030 Besançon Cedex, France (henrot@math.univ-fcomte.fr). This author would like to thank the University Laval and Michel Fortin for a visiting grant.

‡Département de Mathématiques et de Statistique, Université Laval, Québec G1K 7P4, Canada (gphilip@mat.ulaval.ca).

The second section of the paper deals with some overdetermined Saint-Venant problems. The third section is dedicated to the analogue of problem (1)–(4) in $\mathbf{R}^2$, and the fourth section to some overdetermined polarization problem. In the fifth section, we collect some related open problems with ellipsoidal free boundaries in higher dimension.

2. An overdetermined Saint-Venant problem. Let u be the solution of the Saint-Venant problem

$$\Delta u = -2 \quad \text{in } \Omega, \tag{5}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{6}$$

where Ω is a simply connected bounded domain in $\mathbf{R}^2$. Assuming Ω convex, Makar-Limanov in 1971 [Ma-Li] has established the convexity of the level lines $\{u = \text{const.}\}$ of problem (5), (6). His method is based on the auxiliary function

$$M := u_{,ij}u_{,i}u_{,j} - |\nabla u|^2\Delta u + u[(\Delta u)^2 - u_{,ij}u_{,ij}]. \tag{7}$$

In (7) and in the rest of the paper a comma followed by indices indicates partial differentiation, and we adopt the usual summation convention on repeated indices. Using an identity derived in [Ph-Po] we compute

$$M_{,k} = u_{,ijk}u_{,i}u_{,j} - 2uu_{,ij}u_{,ijk} \quad \text{in } \Omega, \tag{8}$$

$$\Delta M = -2uu_{,ijk}u_{,ijk} \le 0 \quad \text{in } \Omega. \tag{9}$$

Moreover, we have on $\partial\Omega$

$$M = K|\nabla u|^3, \tag{10}$$

where K is the curvature of $\partial\Omega$. It follows from (9), (10) that M is positive in Ω if Ω is convex, which implies the convexity of the level lines $\{u = \text{const.}\}$.

Another consequence of (9) is that M is constant in Ω if and only if Ω is an ellipse. Indeed we have equality in (9) if and only if $u_{,ijk} = 0,\ i, j, k = 1, 2$, i.e., if and only if u is a quadratic polynomial in the two variables (x, y).

This section deals with problem (5), (6) overdetermined by the further boundary condition

$$K|\nabla u|^3 = c = \text{const.} > 0 \text{ on } \partial\Omega. \tag{11}$$

From the above remark we infer that (11) is satisfied if Ω is an ellipse. The next statement asserts that ellipses are the only domains for which condition (11) is satisfied.

THEOREM 1. *The overdetermined problem* (5), (6), (11) *is solvable only if* Ω *is an ellipse.*

Obviously we have to show the following implication:

$$M = \text{const. on } \partial\Omega \Longrightarrow M = \text{const. throughout } \Omega. \tag{12}$$

Let us assume that (11) is satisfied, i.e., that $M = \text{const.}$ on $\partial\Omega$. The outward normal derivative $\frac{\partial M}{\partial n}$ must be everywhere nonpositive on $\partial\Omega$ since M takes its minimum there. The conclusion of Theorem 1 would then follow if we succeed to construct a point $P \in \partial\Omega$ at which $\frac{\partial M}{\partial n} = 0$ in view of Hopf's second maximum principle [Pr-We]. On the other hand if the conclusion of Theorem 1 is incorrect, there would exist a

domain Ω for which $\frac{\partial M}{\partial n} < 0$ everywhere on $\partial\Omega$. Our proof of Theorem 1 will consist of showing that the two conditions $M = \text{const.}$ on $\partial\Omega$, $\frac{\partial M}{\partial n} < 0$ on $\partial\Omega$ lead to a contradiction.

Let $\mathbf{n} := (n_1, n_2) = (\cos\theta, \sin\theta)$, $\theta \in [0, 2\pi)$ be the exterior normal vector on $\partial\Omega$. Let s denote the arc length on $\partial\Omega$ oriented in the anticlockwise direction. Using the relations

$$u_{xxx} = -u_{xyy}, \qquad u_{yyy} = -u_{xxy}, \tag{13}$$

$$\mathbf{n} = -\frac{\nabla u}{|\nabla u|}, \tag{14}$$

and (8), we compute

$$\begin{aligned} \frac{\partial M}{\partial n} &= \mathbf{n}\cdot\nabla M = -|\nabla u|^{-1} u_{,ijk} u_{,i} u_{,j} u_{,k} \\ &= |\nabla u|^{-1}\left\{u_{xxy}u_y[u_y^2 - 3u_x^2] + u_{xyy}u_x[u_x^2 - 3u_y^2]\right\}, \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{\partial M}{\partial s} &= 0 = -M_x n_2 + M_y n_1 \\ &= |\nabla u|^{-1}\left\{u_{xxy}u_x[u_x^2 - 3u_y^2] + u_{xyy}u_y[3u_x^2 - u_y^2]\right\}. \end{aligned} \tag{16}$$

The overdetermined condition (11) implies that Ω is strictly convex with a smooth boundary $\partial\Omega$. It follows then that the angle θ between $\mathbf{n}$ and the x-axis is strictly increasing as a function of s on $\partial\Omega$. Let $(x(\theta), y(\theta))$ be a parametric representation of $\partial\Omega$. From (14), (11) we compute

$$\begin{aligned} (x'(s), y'(s)) &= (\dot{x}(s), \dot{y}(s))\frac{ds}{d\theta} = (\dot{x}(s), \dot{y}(s))K^{-1} \\ &= (u_{y,}, -u_x)|\nabla u|^{-1}K^{-1} = |\nabla u|^2 c^{-1}(u_{y,}, -u_x), \end{aligned} \tag{17}$$

where a prime stands for $\frac{d}{d\theta}$, a dot stands for $\frac{d}{ds}$.

Let us now consider the pair of periodic functions $\varphi_1(\theta), \varphi_2(\theta)$ defined on $\partial\Omega$ as

$$\varphi_1(\theta) := u_{xx}(x(\theta), y(\theta)), \tag{18}$$

$$\varphi_2(\theta) := -u_{xy}(x(\theta), y(\theta)), \tag{19}$$

$\theta \in [0, 2\pi)$. Using (13), (17) we compute

$$\varphi_1'(\theta) = u_{xxx}x' + u_{xxy}y' = -\frac{|\nabla u|^2}{c}\{u_{xxy}u_x + u_{xyy}u_y\}, \tag{20}$$

$$\varphi_2'(\theta) = -u_{xyx}x' - u_{xyy}y' = -\frac{|\nabla u|^2}{c}\{u_{xxy}u_y - u_{xyy}u_x\}. \tag{21}$$

Using the identities

$$\cos 2\theta = \cos^2\theta - \sin^2\theta = |\nabla u|^{-2}(u_x^2 - u_y^2), \tag{22}$$

$$\sin 2\theta = 2\sin\theta\cos\theta = 2|\nabla u|^{-2}u_x u_y\ , \tag{23}$$

together with (20), (21) we obtain

$$\varphi_1'\cos 2\theta - \varphi_2'\sin 2\theta = \frac{1}{c}\{u_{xxy}u_x(3u_y^2 - u_x^2) + u_{xyy}u_y(u_y^2 - 3u_x^2)\}, \tag{24}$$

$$\varphi_1'\sin 2\theta + \varphi_2'\cos 2\theta = \frac{1}{c}\{u_{xxy}u_y(u_y^2 - 3u_x^2) + u_{yyx}u_x(u_x^2 - 3u_y^2)\}. \tag{25}$$

Using (15), (16), and assuming $M = \text{const.}$ on $\partial\Omega$, $\frac{\partial M}{\partial n} < 0$ on $\partial\Omega$, we can rewrite (24), (25) as

$$(\varphi_1'(\theta), \varphi_2'(\theta)) = a(\theta)(\sin 2\theta, \cos 2\theta), \tag{26}$$

with

$$a(\theta) := \frac{1}{c}|\nabla u|\frac{\partial M}{\partial n} < 0 \qquad \forall \theta \in [0, 2\pi). \tag{27}$$

From (26), (27) we conclude that the closed curve Γ given by the parametric representation $(\varphi_1(\theta), \varphi_2(\theta)), \theta \in [0, 2\pi)$ has its tangent which makes two revolutions around the origin in the negative direction.

In other words we have

$$\text{turn } (\Gamma) = -2, \tag{28}$$

where turn (Γ) is the turning number associated with Γ. We refer to [Be-Go] for a precise definition of turn (Γ) as the degree of the unit tangent map of Γ.

We shall now compute turn (Γ) using a different approach and obtain a contradiction. To this purpose we consider the auxiliary complex-valued function $f(z)$ defined in Ω as

$$f(z) := u_{xx} - iu_{xy}, \tag{29}$$

with $z := x + iy \in \bar{\Omega}$. We note that u_{xx} and $-u_{xy}$ are harmonic conjugate functions in view of (5), (13); i.e., $f(z)$ is analytic and maps $\partial\Omega$ onto Γ. Let the unit disc $|\zeta| < 1$ in the complex ζ-plane be mapped conformally onto Ω and denote by $z = \phi(\zeta)$ the univalent function associated to this mapping. We obtain a parametric representation of Γ in terms of $\psi \in [0, 2\pi)$ by setting

$$\varphi_1(\psi) + i\varphi_2(\psi) = f(\phi(e^{i\psi})). \tag{30}$$

The hodograph of Γ given by the analytic function

$$\frac{d}{d\psi}(\varphi_1 + i\varphi_2) = ie^{i\psi}\phi'(e^{i\psi})f'(\phi(e^{i\psi})) \tag{31}$$

is a closed curve homotopic to the unit tangent map of Γ (parametrized by $\psi \mapsto \frac{ie^{i\psi}\phi'(e^{i\psi})f'(\phi(e^{i\psi}))}{|\phi'(e^{i\psi})f'(\phi(e^{i\psi}))|}$) that does not go through the origin since we have

$$f'(z) = u_{xxx} - iu_{xxy} = -(u_{xyy} + iu_{xxy}) \neq 0 \qquad \forall z \in \partial\Omega \tag{32}$$

by assumption, and since $\phi'(\zeta) \neq 0$ in $\{|\zeta| < 1\}$ by conformality. The number turn (Γ) may therefore be evaluated by the integral

$$\text{turn } (\Gamma) = \frac{1}{2i\pi}\int_{|\zeta|=1} \frac{F'(\zeta)d\zeta}{F(\zeta)} \tag{33}$$

with

$$F(\zeta) := i\zeta\phi'(\zeta)f'(\phi(\zeta)); \tag{34}$$

turn (Γ) coincides therefore with the number of zeros (counted with their multiplicity) of $F(\zeta)$ in the disc $|\zeta| < 1$ and is therefore ≥ 1, in contradition to (28). This achieves the proof of Theorem 1.

3. An overdetermined boundary value problem for the electrostatic potential in $\mathbf{R}^2$. Let K be a simply connected bounded domain in $\mathbf{R}^2$, and let u be the solution of the following boundary value problem:

$$\Delta u = 0 \quad \text{in } \Omega := \mathbf{R}^2 \backslash \bar{K}, \tag{35}$$

$$u = 1 \quad \text{on } \partial K, \tag{36}$$

$$\lim_{r\to\infty} \frac{u}{\log r} = 1, \tag{37}$$

where $r := \sqrt{x^2+y^2}$ is the distance from the point (x,y) to the origin $O \in K$. We note that u takes its minimum value on ∂K. Moreover, if K is an ellipse, $u(x,y)$ can be computed in terms of elliptic functions and the following two relations hold on $\partial\Omega$:

$$x|\nabla u| = \epsilon\, n_1, \tag{38}$$

$$|\nabla u| = \delta h, \tag{39}$$

where ϵ is a negative constant, δ is a positive constant, $\mathbf{n} := (n_1, n_2)$ denotes the normal vector on ∂K oriented inside K, and $h := -x_i n_i$ is the distance from the tangent of ∂K to the origin $O \in K$. In this section we establish the following result.

THEOREM 2. *The boundary value problem* (35)–(37) *overdetermined either by* (38) *or by* (39) *is solvable only if K is an ellipse.*

Let $\hat{\Omega}$ be the exterior of the unit disc $|\zeta| < 1$ of the complex ζ-plane. This domain $\hat{\Omega}$ can be mapped conformally onto Ω by means of a univalent complex-valued function $\zeta \to z = \phi(\zeta)$ of the following form:

$$\phi(\zeta) = a\zeta + \sum_{n=0}^{\infty} b_n \zeta^{-n}, \tag{40}$$

where a can be chosen real positive. Let $\hat{u}$ be the solution of problem (35)–(37) associated with $\hat{\Omega}$. We have obviously

$$\hat{u}(\zeta) = 1 + \log \rho, \tag{41}$$

with $\rho := |\zeta|$, $\zeta := \rho e^{i\psi} \in \hat{\Omega}$. We then obtain the solution $u(z)$ of (35)–(37) in Ω by conformal transplantation of $\hat{u}(\zeta)$. This leads to

$$\begin{aligned} u(z) &:= \hat{u}(\phi^{-1}(z)) = \hat{u}(\zeta) = 1 + \log|\zeta| \\ &= 1 + Re \log \zeta = 1 + Re\{\log \phi^{-1}(z)\}. \end{aligned} \tag{42}$$

The complex gradient $\nabla u := u_x + iu_y$ is then given by

$$\nabla u = \overline{(\log \phi^{-1}(z))'} = 1/\ \overline{\phi^{-1}(z)\phi'(\phi^{-1}(z))}. \tag{43}$$

Moreover, the complex unit normal vector $n := n_1 + in_2$ is obtained by differentiating the parametric representation $\phi(e^{i\psi})$ of $\partial\Omega$ with respect to the parameter ψ. This leads to

$$n = n_1 + in_2 = -e^{i\psi} \frac{\phi'(e^{i\psi})}{|\phi'(e^{i\psi})|}. \tag{44}$$

Combining (43) and (44), we obtain the following expression for the normal derivative $\frac{\partial u}{\partial n}$ on $\partial\Omega$ in terms of the mapping function $\phi(\zeta)$:

$$\left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = Re\left[(u_x - iu_y)(n_1 + in_2)\right] = -\frac{1}{|\phi'(e^{i\psi})|}. \tag{45}$$

Since $u = 1$ on $\partial\Omega$, we have

$$|\nabla u|_{\partial\Omega} = -\left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = \frac{1}{|\phi'(e^{i\psi})|}. \tag{46}$$

With (46) we can rewrite the overdetermined conditions (38) and (39) as follows:

$$Re\{\phi(e^{i\psi})\} = -\epsilon Re\{e^{i\psi}\phi'(e^{i\psi})\}, \quad \psi \in [0, 2\pi), \tag{47}$$

and

$$Re\left\{e^{i\psi}\phi'(e^{i\psi})\,\overline{\phi(e^{i\psi})}\right\} = \frac{1}{\delta} > 0. \tag{48}$$

In the next step we shall show that either (47) or (48) implies separately that ϕ has the form

$$\phi(\zeta) = a\zeta + b_0 + b_1\zeta^{-1}, \tag{49}$$

from which we conclude that K must be an ellipse.

First case (47). Obviously we can select $b_0 = 0$ in (40) since this choice affects the mapping by a translation only. Making use of (40) with $b_n = \beta_n + i\gamma_n$, $n = 1, 2, \ldots$, condition (47) gives

$$\begin{aligned} &a\cos\psi + \sum_1^\infty (\beta_n \cos n\psi + \gamma_n \sin n\psi) \\ &\qquad = -\epsilon\left\{a\cos\psi - \sum_1^\infty n(\beta_n \cos n\psi + \gamma_n \sin n\psi)\right\}, \end{aligned} \tag{50}$$

from which we obtain

$$\beta_1 = \frac{a(1+\epsilon)}{\epsilon - 1}, \quad \beta_n = 0 \quad \forall n \geq 2, \tag{51}$$

$$\gamma_n = 0 \qquad \forall n \geq 1. \tag{52}$$

This leads to the desired form (49) of $\phi(\zeta)$.

Second case (48). The analysis of this case is more complicated. The condition (48) may be rewritten under the form

$$e^{i\psi}\,\overline{\phi(e^{i\psi})}\phi'(e^{i\psi}) + e^{-i\psi}\phi(e^{i\psi})\,\overline{\phi'(e^{i\psi})} = \text{const}. \tag{53}$$

Again we write

$$\phi(\zeta) := a\zeta + \sum_1^\infty b_n\zeta^{-n} \tag{54}$$

without loss of generality. Let us now introduce the associated function

$$\tilde{\phi}(\zeta) := \overline{\phi\left(1/\bar{\zeta}\right)} = a\zeta^{-1} + \sum_{1}^{\infty} \overline{b_n}\zeta^n. \tag{55}$$

With

$$\tilde{\phi}(e^{i\psi}) = \overline{\phi(e^{i\psi})}, \tag{56}$$

$$\tilde{\phi}'(\zeta) = -\frac{1}{\zeta^2}\overline{\phi'\left(1/\bar{\zeta}\right)}, \tag{57}$$

the condition (53) may be rewritten as

$$e^{i\psi}\tilde{\phi}(e^{i\psi})\phi'(e^{i\psi}) - e^{i\psi}\phi(e^{i\psi})\tilde{\phi}'(e^{i\psi}) = \text{const.} = c. \tag{58}$$

Moreover, we know a priori that the overdetermined condition implies that the boundary $\partial\Omega$ is analytic. This follows from Theorem 1.4 in Chapter 2 of Friedman's book [Fr]. The function ϕ analytic in $|\zeta| > 1$ may therefore be analytically extended inside the unit disc in a neighborhood of $|\zeta| = 1$, and the function $\tilde{\phi}$ analytic in $|\zeta| < 1$ may be analytically extended outside the unit disc in a neighborhood of $|\zeta| = 1$. There exists therefore a two-sided neighborhood $\omega := \{1 - \epsilon < |\zeta| < 1 + \epsilon\}$ of the unit circle in which ϕ and $\tilde{\phi}$ are analytic. In ω we have the identity

$$\tilde{\phi}(\zeta)\phi'(\zeta) - \tilde{\phi}'(\zeta)\phi(\zeta) = \frac{c}{\zeta} \tag{59}$$

in view of (58). Differentiating (59) we obtain

$$\tilde{\phi}\phi'' - \tilde{\phi}''\phi = -\frac{c}{\zeta^2}. \tag{60}$$

Combining (59) and (60) leads to

$$\tilde{\phi}\phi' - \tilde{\phi}'\phi + \zeta(\tilde{\phi}\phi'' - \tilde{\phi}''\phi) = 0, \tag{61}$$

which may be rewritten as

$$\zeta\,\frac{\phi' + \zeta\phi''}{\phi} = \zeta\,\frac{\tilde{\phi}' + \zeta\tilde{\phi}''}{\tilde{\phi}} \quad \text{in } \omega, \tag{62}$$

where the factors ζ on both sides of (62) make the right-hand side analytic at the origin since we have

$$\lim_{\zeta\to 0} \zeta\,\frac{\tilde{\phi}' + \zeta\tilde{\phi}''}{\tilde{\phi}} = 1 \tag{63}$$

as a consequence of (55), and the left-hand side bounded at infinity since we have

$$\lim_{\zeta\to\infty} \zeta\,\frac{\phi' + \zeta\phi''}{\phi} = 1 \tag{64}$$

as a consequence of (54). The identity (62) shows in fact that the function defined by $\zeta\frac{\phi'+\zeta\phi''}{\phi}$ outside the unit disc and by $\zeta\frac{\tilde{\phi}'+\zeta\tilde{\phi}''}{\tilde{\phi}}$ inside the closed unit disc is analytic

and bounded in the whole complex ζ-plane. It then follows from the classical theorem of Liouville that this function is constant (and equal to one, thanks to (64)). This fact leads to the following differential equation for $\phi(\zeta)$:

$$\zeta^2\phi'' + \zeta\phi' - \phi = 0 \qquad \forall|\zeta| > 1. \tag{65}$$

The differential equation (65) has the general solution

$$\phi(\zeta) = a\zeta + b\zeta^{-1}. \tag{66}$$

This achieves the proof of Theorem 2.

4. An overdetermined polarization problem. Let K be a simply connected bounded domain in $\mathbf{R^2}$. We consider the polarization problem in the x direction defined in $\Omega := \mathbf{R}^2\backslash\bar{K}$ as

$$\Delta u = 0 \quad \text{in } \Omega := \mathbf{R}^2\backslash\bar{K}, \tag{67}$$

$$u = x + b \quad \text{on } \partial\Omega, \tag{68}$$

$$u = O(r^{-1}) \quad \text{as } r := \sqrt{x^2+y^2} \to \infty. \tag{69}$$

In (68), b is a constant that is uniquely determined by the condition (69). In this section we establish the following result.

THEOREM 3. *Let u be the solution of problem* (67)–(69). *Assume moreover that u satisfies the further boundary condition*

$$\frac{\partial u}{\partial n} = cn_1 \quad \text{on } \partial\Omega, \tag{70}$$

where $\mathbf{n} := (n_1, n_2)$ is the unit normal vector on $\partial\Omega$ directed inside K and where c is a constant. Then K must be an ellipse whose axes are parallel to the axes of coordinates.

Before proving Theorem 3 we note that the constant c in (70) cannot be given arbitrarily. In fact c depends only on the geometry of K and is given by

$$c = -\frac{\int_{\partial\Omega} n_2^2\, \mathbf{x}.\mathbf{n}\, ds}{\int_{\partial\Omega} n_1^2\, \mathbf{x}.\mathbf{n}\, ds} < 0. \tag{71}$$

For the proof of (71) we compute

$$\int_\Omega |\nabla u|^2 dx = \int_{\partial\Omega} u\frac{\partial u}{\partial n} ds = c\int_{\partial\Omega} x n_1 ds = -cA, \tag{72}$$

where A is the area of K. (72) implies that c is negative. Moreover, from (70) we have

$$\frac{\partial(u-x)}{\partial n} = (c-1)n_1 \quad \text{on } \partial\Omega. \tag{73}$$

Inserting (72), (73) into Rellich's identity

$$\int_\Omega |\nabla u|^2 dx + A = -\frac{1}{2}\int_{\partial\Omega} \left(\frac{\partial(u-x)}{\partial n}\right) \mathbf{x}.\mathbf{n}\, ds \tag{74}$$

that is also derived in [Pa-Ph], we obtain (71).

For the proof of Theorem 3 we proceed again by conformal transplantation. Let $\hat{\Omega}$ be the exterior of the unit disc $|\zeta| < 1$ of the complex ζ-plane; $\hat{\Omega}$ may be mapped conformally onto Ω by means of a univalent complex-valued function $\zeta \to z = \phi(\zeta)$ of the following form:

$$\phi(\zeta) = a\zeta + \sum_{n=0}^{\infty} b_n \zeta^{-n}, \tag{75}$$

where a may be chosen real positive without loss of generality. Let $\hat{u}(\zeta)$ be the transplanted function defined in $\hat{\Omega}$ as

$$\hat{u}(\zeta) := u(\phi(\zeta)). \tag{76}$$

This function $\hat{u}(\zeta)$ satisfies the following boundary value problem:

$$\Delta \hat{u} = 0 \quad \text{in } \hat{\Omega}, \tag{77}$$

$$\hat{u}(e^{i\psi}) = u(\phi(e^{i\psi})) = Re(\phi(e^{i\psi})) + b, \quad \psi \in [0, 2\pi], \tag{78}$$

$$\hat{u} = O\left(\frac{1}{|\zeta|}\right) \text{ as } |\zeta| \to \infty. \tag{79}$$

In (78) we have used the notation $\zeta = \rho e^{i\psi}$. From (79) we see that $b = -Re(b_0)$. The solution of (77)–(79) is easily computable. We find

$$\begin{aligned} \hat{u}(\zeta) &= \hat{u}(\rho e^{i\psi}) \\ &= \frac{1}{2}\left\{ \frac{\bar{a} + b_1}{\rho} e^{-i\psi} + \frac{a + \bar{b}_1}{\rho} e^{i\psi} + \sum_{2}^{\infty} \frac{b_n e^{-in\psi} + \bar{b}_n e^{in\psi}}{\rho^n} \right\}. \end{aligned} \tag{80}$$

Since we have $a = \bar{a}$, (80) may be rewritten as

$$\hat{u}(\zeta) = Re\left(\frac{a}{\zeta}\right) + Re\left(\sum_{1}^{\infty} b_n \zeta^{-n}\right) = Re\left\{\frac{a}{\zeta} + \phi(\zeta) - a\zeta - b_0\right\} \tag{81}$$

in view of (75), from which we compute

$$u(z) = \hat{u}(\phi^{-1}(z)) = Re\left\{\frac{a}{\phi^{-1}(z)} + z - a\phi^{-1}(z) - b_0\right\}. \tag{82}$$

Using the identity $\nabla\{Re\, f(z)\} = \overline{f'(z)}$ with $f(z)$ analytic, we compute from (82)

$$\overline{\nabla u(z)} = u_x - iu_y = 1 - \frac{a}{\phi'(\phi^{-1}(z))[\phi^{-1}(z)]^2} - \frac{a}{\phi'(\phi^{-1}(z))}\,. \tag{83}$$

The complex-valued normal vector $n := n_1 + in_2$ of $\partial\Omega$ may also be expressed in terms of the parametrization $\psi \to \phi(e^{i\psi})$, $\psi \in [0, 2\pi)$ of $\partial\Omega$. We obtain

$$n = n_1 + in_2 = -e^{i\psi}\, \frac{\phi'(e^{i\psi})}{|\phi'(e^{i\psi})|}. \tag{84}$$

From (83), (84) we compute

$$\begin{aligned} \frac{\partial u(\phi(e^{i\psi}))}{\partial n} &= Re\{\overline{\nabla u}.n\} \\ &= Re\left\{-\frac{e^{i\psi}\phi'(e^{i\psi})}{|\phi'(e^{i\psi})|} + \frac{a}{|\phi'(e^{i\psi})|e^{i\psi}} + \frac{ae^{i\psi}}{|\phi'(e^{i\psi})|}\right\}. \end{aligned} \tag{85}$$

The overdetermined condition (70) may therefore be rewritten as

$$Re\{e^{i\psi}\phi'(e^{i\psi}) - ae^{-i\psi} - ae^{i\psi}\} = Re\{ce^{i\psi}\phi'(e^{i\psi})\}. \tag{86}$$

From (75) we compute

$$e^{i\psi}\phi'(e^{i\psi}) = ae^{i\psi} - \sum_{1}^{\infty} nb_n e^{-in\psi}. \tag{87}$$

From (86), (87) we obtain by identification

$$\begin{aligned} &-a\cos\psi - Re(b_1)\cos\psi - Im(b_1)\sin\psi \\ &= c(a\cos\psi - Re(b_1)\cos\psi - Im(b_1)\sin\psi) \end{aligned} \tag{88}$$

and

$$Re(nb_n e^{-in\psi}) = cRe(nb_n e^{-in\psi}), \quad n \geq 2. \tag{89}$$

Equation (88) gives

$$b_1 = \frac{c+1}{c-1}\, a. \tag{90}$$

Since $c \neq 1$, we obtain from (89)

$$b_n = 0, \quad n \geq 2. \tag{91}$$

The above computation shows that the mapping function $\phi(\zeta)$ must have the particular form

$$\phi(\zeta) = a\zeta + b_0 + \frac{c+1}{c-1}\,\frac{a}{\zeta}, \tag{92}$$

which implies the desired result. We finally note that K is a disc if $c = -1$.

5. Some open problems. In this section, we want to indicate some open problems which generalize in three dimensions the problems studied in the previous sections. In what follows, Ω_0 will denote the ellipsoid

$$\Omega_0 = \left\{ \mathbf{x} = (x,y,z) \,\Big/\, \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} < 1 \right\},$$

with $a \geq b \geq c > 0$. The distance h from the tangent plane of $\partial\Omega_0$ to the origin is given by

$$h = \mathbf{x}.\mathbf{n} = \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right)^{-1/2}, \tag{93}$$

and the Gaussian curvature G of $\partial\Omega_0$ is given by

$$G = h^4/a^2b^2c^2. \tag{94}$$

Of course, the problems described below have analogous N-dimensional versions which could easily be formulated.

The Saint-Venant problem. Let u be the solution of the Saint-Venant problem (5), (6) in Ω_0. We have

$$u(x,y,z) = A\left(\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1\right), \tag{95}$$

$$|\nabla u|^2 = 4A^2\left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4}\right), \tag{96}$$

with $A = -2(\frac{2}{a^2} + \frac{2}{b^2} + \frac{2}{c^2})^{-1}$. It follows from (94), (96) that

$$|\nabla u|^4\, G = \text{const.} \quad \text{on } \partial\Omega_0. \tag{97}$$

It is then natural to state the following problem.

Open problem 1. Let u be the solution of the Saint-Venant problem (5), (6) in Ω, where Ω is a regular bounded, simply connected domain in $\mathbf{R}^3$. Assume moreover that the overdetermined condition (97) is satisfied where G is the Gaussian curvature of $\partial\Omega$. Then prove that Ω is an ellipsoid.

The electrostatic potential problem. Let u be the electrostatic potential of the ellipsoid Ω_0, i.e., the solution of the boundary value problem

$$\Delta u = 0 \quad \text{in } \Omega := \mathbf{R}^3 \backslash \bar{\Omega}_0, \tag{98}$$

$$u = 1 \quad \text{on } \partial\Omega_0, \tag{99}$$

$$|\nabla u| = O(r^{-2}) \quad \text{as } r \to \infty. \tag{100}$$

Using ellipsoidal coordinates, we are able to write u explicitly as

$$u(x,y,z) = \int_\lambda^\infty \frac{dt}{\{(a^2+t)(b^2+t)(c^2+t)\}^{1/2}}, \tag{101}$$

where λ is the largest root of the equation

$$\frac{x^2}{a^2+\lambda} + \frac{y^2}{b^2+\lambda} + \frac{z^2}{c^2+\lambda} = 1.$$

We refer to [Ke] for the derivation of (101). From (101), it follows that

$$x|\nabla u| = \epsilon\, n_1, \tag{102}$$

$$|\nabla u| = \delta h \tag{103}$$

on $\partial\Omega_0$ as in the 2-dimensional case, where ϵ and δ are constants. Therefore, we can state the following problem.

Open problem 2. Let K be a regular simply connected compact set in $\mathbf{R}^3$. Let u be the electrostatic potential defined as the solution of (98)–(100) in $\mathbf{R}^3 \setminus K$. Assume moreover that u satisfies either the overdetermined condition (102) or (103). Then, prove that K is an ellipsoid.

The polarization problem. Let u be the solution of the polarization problem for the ellipsoid

$$\Delta u = 0 \quad \text{in } \Omega := \mathbf{R}^3 \backslash \bar{\Omega}_0, \tag{104}$$

$$u = x + b \quad \text{on } \partial\Omega_0, \tag{105}$$

$$|\nabla u| = O(r^{-3}) \quad \text{as } r \to \infty. \tag{106}$$

Using ellipsoidal coordinates, we are again able to write u explicitly as (see [Sc-Sz])

$$u(x,y,z) = \frac{2x}{p}\int_{\lambda}^{\infty} \frac{dt}{(a^2+t)\{(a^2+t)(b^2+t)(c^2+t)\}^{1/2}}, \tag{107}$$

where p is a constant determined by the boundary condition. It follows from (107) that

$$\frac{\partial u}{\partial n} = \text{const.} n_1 \tag{108}$$

on $\partial\Omega_0$ as in the 2-dimensional case.

Open problem 3. Let K be a regular, simply connected compact set in $\mathbf{R}^3$. Let u be the solution of the polarization problem (104)–(106) in $\mathbf{R}^3 \setminus K$. Assume moreover that u satisfies the overdetermined condition (108). Then, prove that K is an ellipsoid.

REFERENCES

[Be-Go] M. Berger and B. Gostiaux, *Differential Geometry: Manifolds, Curves and Surfaces*, Graduate Texts in Mathematics 115, Springer-Verlag, Berlin, New York, 1988.

[Fr] A. Friedman, *Variational Principles and Free Boundary Problems*, Wiley, New York, 1982.

[Ke] O. D. Kellogg, *Foundations of Potential Theory*, Dover, New York, 1953.

[Ma-Li] L. G. Makar-Limanov, *Solutions of Dirichlet problem for the equation $\Delta u = -1$ in a convex region*, Math. Notes Acad. Sci. URSS, 9 (1971), pp. 52–53.

[Pa-Ph] L. E. Payne and G. A. Philippin, *Isoperimetric inequalities for polarization and virtual mass*, J. Anal. Math., 47 (1986), pp. 255–267.

[Ph-Po] G. A. Philippin and G. Porru, *Isoperimetric inequalities and overdetermined problems for the Saint-Venant equation*, New Zealand J. Math., 25 (1996), pp. 217–227.

[Pr-We] M. H. Protter and H. F. Weinberger, *Maximum Principles and their Applications*, Math. Acad. Press, London, UK, 1981.

[Sc-Sz] M. Schiffer and G. Szégö, *Virtual mass and polarization*, Trans. Amer. Math. Soc., 67 (1949), pp. 130–205.

[Se] J. Serrin, *A symmetry problem in potential theory*, Arch. Rational Mech. Anal., 43 (1971), pp. 304–318.

[We] H. F. Weinberger, *Remark on the preceding paper of Serrin*, Arch. Rational Mech. Anal., 43 (1971), pp. 319–320.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 321–342, March 1998

003

ON A FOURTH-ORDER DEGENERATE PARABOLIC EQUATION: GLOBAL ENTROPY ESTIMATES, EXISTENCE, AND QUALITATIVE BEHAVIOR OF SOLUTIONS*

ROBERTA DAL PASSO†, HARALD GARCKE‡, AND GÜNTHER GRÜN§

Abstract. By means of energy and entropy estimates, we prove existence and positivity results in higher space dimensions for degenerate parabolic equations of fourth order with nonnegative initial values. We discuss their asymptotic behavior for $t \to \infty$ and give a counterexample to uniqueness.

Key words. fourth-order degenerate parabolic equations, existence, regularity, long-time behavior, thin films

AMS subject classifications. 35K35, 35K55, 35K65, 35B40, 35B65, 76D08

PII. S0036141096306170

1. Introduction. In this paper we will present new results on existence, (non) uniqueness, positivity, and asymptotic behavior in higher space dimensions of weak solutions to degenerate parabolic equations of fourth order of the form

$$\begin{aligned} u_t + \operatorname{div}\big(m(u)\nabla\Delta u\big) &= 0 \quad \text{in } \Omega \times (0,T), \\ \frac{\partial u}{\partial \nu} = \frac{\partial}{\partial \nu}\Delta u &= 0 \quad \text{on } \partial\Omega \times [0,T], \\ u(0,\cdot) &= u_0(\,\cdot\,) \quad \text{in } \Omega. \end{aligned} \tag{1.1}$$

We assume that the nonnegative diffusion coefficient m vanishes at zero and has at most polynomial growth. We denote by n its growth exponent near zero.
Equation (1.1) can be seen as the archetype of a class of parabolic equations of higher order which appear in material sciences and fluid dynamics. For instance, in lubrication theory (cf. [3], [8] and the references therein), u describes the height of a viscous droplet spreading on a plain, solid surface; in the Cahn–Hilliard model of phase separation for binary mixtures, u plays the role of the concentration of one component (cf. [10]), and in a plasticity model (cf. [13] and the references therein) u stands for the density of dislocations.

Crucial for these applications is the fact that it is possible to construct solutions of (1.1) which preserve nonnegativity as has been proved for space dimension $N = 1$ by Bernis and Friedman [6] and for higher space dimensions in the papers by Grün [13] and by Elliott and Garcke [10]. This behavior is in strong contrast to that of classical solutions to linear parabolic equations of fourth order which in general become

*Received by the editors June 12, 1996; accepted for publication (in revised form) January 6, 1997. All authors were supported by the Vigoni Exchange Programme *Mathematical Models for the Physical Phenomena of Reaction Diffusion and Transition of Phase*, Project 6243. The first two authors were partially supported by DFG through SFB256 *Nichtlineare Partielle Differentialgleichungen*, and the third author was partially supported by the European Union through the Training and Mobility of Researchers Programme grant ERB4001GT950586.

http://www.siam.org/journals/sima/29-2/30617.html

†Dipartimento di Matematica, Università di Tor Vergata, Via della Ricerca Scientifica, 00133 Roma, Italy (dalpasso@mat.utovrm.it).

‡Universität Bonn, Institut für Angewandte Mathematik, Wegelerstr. 6, 53115 Bonn, Germany (harald@iam.uni-bonn.de).

§Universität Bonn, Institut für Angewandte Mathematik, Beringstr. 6, 53115 Bonn, Germany (gg@iam.uni-bonn.de).

negative even in the case of strictly positive initial values. Moreover, the publications of Beretta, Bertsch, and Dal Passo [2] and of Bertozzi and Pugh [8], who study this equation in space dimension $N = 1$, reveal a rich structure of qualitative behavior of solutions depending on the diffusion growth exponent n. To put it concisely, the larger n is, the stronger is the tendency of solutions to stay positive and the weaker is the regularity at the boundary of the set where u vanishes.

Recently Bernis [4], [5] showed for the special case $m(u) = u^n$ with $0 < n < 3$ that in space dimension $N = 1$ solutions to (1.1) have the property of finite speed of propagation. More precisely, this means that the interface separating the regions where u is positive and where u is equal to zero moves with finite velocity as time progresses. To obtain these results Bernis used local versions of entropy estimates first derived in [2].

While in the case of higher space dimensions existence results up to now were restricted to the cases $1 \le n < 2$ if $u_0 \ge 0$ arbitrary, and $n \ge 1$ if u_0 is strictly positive (cf. [13] and [10]), the results presented here will overcome this restriction and assure the existence for $\frac{1}{8} < n < 3$ and arbitrary nonnegative initial values u_0. It turns out that in general we cannot expect solutions to have $L^2(0,T;H^2(\Omega))$-regularity. In previous works (cf. [13] and [10]), this property has been a major ingredient in the definition of solution. Thus, we are forced to introduce a new solution concept that—to put in concisely—differs from the previous one in such a way that in the corresponding weak formulation of (1.1) derivatives of u of higher order than 1 do not appear. For the technical details, we refer the reader to the statement of Definition 3.1.

Let us point out that the growth exponent $n = 3$ seems to be a border case in the theory of degenerate fourth-order parabolic equations. This already has been indicated by results of Bernis, Peletier, and Williams [7], who showed that source-type solutions with finite mass only exist for $0 < n < 3$.

Technically, the restriction to values of $n < 3$ in this paper is due to the fact that entropy estimates for compactly supported initial data are not achievable if $n \ge 3$. For similar reasons the results of [4], [5] and many of the results of [2] and [8] are restricted to $n < 3$.

Our work will be based upon a refinement of those entropy estimates which have been used in [13] and [10]; they generalize results of the papers [2] or [8] to space dimensions $N = 2, 3$, which are the relevant ones for applications. Since in the case of higher space dimensions it is still not known whether solutions of (1.1) are in $L^\infty(\Omega_T)$ or whether they are strictly bounded away from zero in case of positive initial values, a more careful approximation process has to be applied than for space dimension $N = 1$.

Another important ingredient in the higher dimensional case which may be of independent interest is the generalization of the formula of integration by parts

$$\int_0^1 f'(u)u_x^2 u_{xx}\,dx = -\tfrac{1}{3}\int_0^1 f''(u)u_x^4\,dx$$

for appropriately smooth functions with $u_x(0) = u_x(1) = 0$ to higher space dimensions, which will be essential in order to obtain the entropy estimates (cf. Lemma 2.3).

Confining ourselves at the moment to entropy estimates global in space, in a forthcoming paper we wish to derive local versions of the entropy estimates and we hope to show results analogous to those of Bernis [4], [5] in higher space dimensions.

Let us briefly describe the outline of this paper.

In section 2 we state the refined entropy estimate first for auxiliary problems with positive initial values and sufficiently large diffusion growth exponents n. In a second step, we extend this result to arbitrary positive values of n by use of an appropriate approximation method.

Section 3 contains the main results of this paper. In Definition 3.1 we introduce the solution concept, in the framework of which we can prove existence of solutions for $\frac{1}{8} < n < 3$ and arbitrary, nonnegative initial values $u_0 \in H^1(\Omega)$ (cf. Theorem 3.2). As a consequence of the a priori estimates derived so far we improve the results of [13] about positivity of solutions (cf. Theorem 3.4) and show convergence to the mean value for $t \to \infty$ with respect to the H^1-norm (cf. Theorem 3.5). The latter result can be used in order to discuss the problem of uniqueness in the framework of the solution concept. By constructing steady state solutions with compact support which do not satisfy the entropy estimates but nevertheless solve the equation in the sense of the solution concept, it becomes evident that we cannot expect uniqueness of solutions without imposing regularity properties at the boundary of the set where u vanishes. Whether these regularity properties are already sufficient for uniqueness still remains an open problem.

Notation. In the whole paper we assume that $\Omega \subset \mathbb{R}^N$ ($N \in \{2, 3\}$) is an open and bounded domain with boundary of class $C^{1,1}$ (or $C^{0,1}$ if Ω is convex) which is piecewise smooth. We denote by I the time interval $(0, T)$, and Ω_T stands for the space-time cylinder $\Omega \times (0, T)$. We denote by ν the unit outer normal vector to $\partial\Omega$, and $II(.)$ is the second fundamental form of $\partial\Omega$. By $H^2_*(\Omega)$ we denote $\{u \in H^2(\Omega) : \frac{\partial}{\partial\nu} u = 0 \text{ on } \partial\Omega\}$. We will use as abbreviation the notation $u \in L^{p-}(\Omega)$ to indicate that $u \in L^q(\Omega)$ for all $q < p$. Furthermore, $\mathcal{L}^N$ denotes the N-dimensional Lebesgue measure and $\mathcal{H}^{N-1}$ denotes the $(N-1)$-dimensional Hausdorff measure. For vectors $v, w \in \mathbb{R}^N$ and symmetric matrices $A \in \mathbb{R}^{N\times N}$, we write $\langle v, A, w\rangle$ instead of $\sum_{i,j} v_i A_{ij} w_j$, and $\langle \cdot, \cdot \rangle$ stands for the standard scalar product on $\mathbb{R}^N$.

2. Entropy estimates. In this section we will present the entropy estimates essential for the qualitative results. For mainly technical reasons we shall confine ourselves at first to a special case of problem (1.1) that is characterized by the following additional conditions on the diffusion coefficient m and on the initial data u_0.

(A1) The diffusivity $m \in C^1(\mathbb{R}^+_0) \cap W^{1,\infty}(\mathbb{R}^+_0)$ can be written as $m(\tau) = \tau^n \cdot f(\tau)$ ($\tau \in \mathbb{R}^+_0$) with a positive function f such that $\|f\|_{C^{1,1}(\mathbb{R}^+_0, \mathbb{R}^+_0)} < \infty$. Furthermore, we assume that m is uniformly bounded from below by a positive constant for sufficiently large values of τ.

(A2) The growth exponent n satisfies

$$n > \begin{cases} 4 & \text{if } N = 2, \\ 8 & \text{if } N = 3. \end{cases}$$

(A3) The initial data $u_0 \in H^1(\Omega)$ are strictly positive; i.e., there exists a constant $\delta > 0$ such that $u_0 \geq \delta > 0$.

In [10] and [13], it has been proved that under the assumptions (A1)–(A3) there exists a pair of functions

$$(u, J) \in H^1\big(I; (H^1(\Omega))'\big) \cap L^\infty\big(I; H^1(\Omega)\big) \cap L^2\big(I; H^2_*(\Omega)\big) \times L^2(\Omega_T, \mathbb{R}^N)$$

which solves (1.1) in the following weak sense:

$$u_t = -\operatorname{div} J \quad \text{in } L^2\big(0, T; (H^1(\Omega))'\big) \tag{2.1}$$

and

$$\int_{\Omega_T} J \cdot \eta = -\int_{\Omega_T} \Delta u \, \mathrm{div}(m(u)\eta) \tag{2.2}$$

for all $\eta \in L^2(0,T;H^1(\Omega,\mathbb{R}^n)) \cap L^\infty(\Omega_T,\mathbb{R}^n)$.

Moreover, for a positive constant C the following estimate holds true:

$$\int_\Omega u(T)^{2-n} dx + \int_{\Omega_T} |\Delta u|^2 \, dxdt \le C \int_\Omega u_0^{2-n} dx \,. \tag{2.3}$$

In particular, this implies that u is strictly positive for almost every $t \in I$ (cf. [13]). This positivity property is the key to the following lemma assuring that under the assumptions (A1)–(A3) $\nabla\Delta u(t) \in L^2(\Omega)$ for almost every $t \in I$, and therefore $J(t) = m(u(t))\nabla\Delta u(t)$ in $L^2(\Omega)$ for these t. As a further consequence, we show that the L^2-norm of ∇u is monotonically decreasing, which will be important to obtain our results regarding asymptotic behavior.

LEMMA 2.1. *Assume* (A1)–(A3) *and let* u *be the weak solution of* (1.1) *constructed as described above. Then for almost all* $t \in [0,T]$ *we have* $u(t) \in C^\beta(\Omega)$ *(for* $\beta > 0$ *appropriately small),* $u(t)$ *is strictly positive, and*

$$J(t,\cdot) = m\big(u(t,\cdot)\big)\nabla\Delta u(t,\cdot) \in L^2(\Omega). \tag{2.4}$$

In addition, for almost every $t_1, t_2 \in I$ *the following estimate is true:*

$$\frac{1}{2}\int_\Omega |\nabla u(t_2)|^2 \, dx + \int_{t_1}^{t_2}\int_\Omega m(u)\,|\nabla\Delta u|^2 \, dxdt \le \frac{1}{2}\int_\Omega |\nabla u(t_1)|^2 \, dx \,. \tag{2.5}$$

The proof of Lemma 2.1 is contained in the Appendix at the end of this paper.

We now state the main entropy estimate. As in papers [2] and [8], we define the entropy to be $G_\alpha(t) = \int_A^t \int_A^s \frac{\tau^{\alpha+n-1}}{m(\tau)} \, d\tau ds$. Here, A is an arbitrary but fixed positive constant.

PROPOSITION 2.2. *Assume* (A1)–(A3) *and let* u *be the weak solution of* (1.1) *constructed by the method of* [10] *and* [13]. *Let* α *and* γ *be real numbers satisfying* $\frac{1}{2} < \alpha + n < 2$,

$$\frac{t+1-\sqrt{(t-2)(1-2t)}}{3} \le \gamma \le \frac{t+1+\sqrt{(t-2)(1-2t)}}{3} \qquad \text{with } \ t := \alpha + n \tag{2.6}$$

and $\gamma \in (\frac{1}{3}, 1)$ *if* $\alpha + n < 1$. *Then we have*

$$\begin{aligned} \int_\Omega G_\alpha\big(u(T)\big)\, dx + C_1^{-1}\int_{\Omega_T} u^{\alpha+n-3}\,|\nabla u|^4 \, dxdt \\ + C_1^{-1}\int_{\Omega_T} u^{\alpha+n+1-2\gamma}\left\{\tfrac{2}{3}\left|D^2 u^\gamma\right|^2 + \tfrac{1}{3}\,|\Delta u^\gamma|^2\right\} dxdt \\ \le \int_\Omega G_\alpha(u_0)\, dx + C_2\int_{\Omega_T} u^{\alpha+n+1}\, dxdt. \end{aligned} \tag{2.7}$$

Here, C_1 *and* C_2 *are constants only depending on the domain* Ω*; in particular,* C_2 *becomes zero if* Ω *is convex.*

Proof. The proof essentially consists of three parts. In the first one we introduce regularized versions of $G'_\alpha(u)$ which we use as test functions in the weak formulation of

equation (1.1). The main difficulty is to control the term containing spatial derivatives which we will henceforth call "elliptic part." To estimate this elliptic part, we at first only formulate a key inequality which allows to pass to the limit with some regularizing parameters (cf. part 2) and thus to establish the result. The last part will be devoted to the detailed verification of that particular inequality.

Part 1. Consider for positive parameters A, σ the functions

$$g^-_{\alpha\sigma}(s) := \int_A^s \frac{(\tau+\sigma)^{\alpha+n-1}}{m(\tau)}\,d\tau \tag{2.8}$$

and

$$g^+_{\alpha\sigma}(s) := \int_A^s \frac{\tau^{\alpha+n-1}}{m(\tau)(1+\sigma\tau^{\alpha+n-1})}\,d\tau\ . \tag{2.9}$$

If $\alpha+n<1$, we choose $g^-_{\alpha\sigma}(u+\varepsilon)$ as the test function in (2.1), (2.2); otherwise we choose $g^+_{\alpha\sigma}(u+\varepsilon)$ and obtain, using Lemma 2.1,

$$\int_0^T \left\langle u_t, g^\pm_{\alpha\sigma}(u+\varepsilon)\right\rangle\,dt = \int_0^T\int_\Omega \big(m(u)\nabla\Delta u, \nabla g^\pm_{\alpha\sigma}(u+\varepsilon)\big)\,dxdt. \tag{2.10}$$

The nonnegativity of u, the shift by ε, and the $L^\infty\big(I;H^1(\Omega)\big)\cap L^2\big(I;H^2(\Omega)\big)$-regularity of u guarantee the admissibility of these test functions. We notice that the functions $G^\pm_{\alpha\sigma}$, defined by

$$G^\pm_{\alpha\sigma}(s) := \int_A^s g^\pm_{\alpha\sigma}(\tau)\,d\tau, \tag{2.11}$$

are nonnegative and convex, that for fixed $\varepsilon>0$ their derivative $g^\pm_{\alpha\sigma}$ has at most linear growth on $[\varepsilon,\infty)$, and that for this reason we obtain (cf. [13, Lemma 2.6])

$$\int_0^T \left\langle u_t, g^\pm_{\alpha\sigma}(u+\varepsilon)\right\rangle\,dt = \int_\Omega G^\pm_{\alpha\sigma}(u+\varepsilon)(T,x)\,dx - \int_\Omega G^\pm_{\alpha\sigma}(u_0+\varepsilon)(x)\,dx. \tag{2.12}$$

In what follows, we shall make the calculations explicit only for $g^-_{\alpha\sigma}$ (i.e., $\alpha+n<1$). But by minor modifications the same strategy will work also in the case $\alpha+n\ge 1$.

Part 2. For almost all $t\in I$ the following inequality offers an estimate for the elliptic part (we set $\beta_{\varepsilon\sigma}(u) := (u+\varepsilon+\sigma)^{\alpha+n-1}$ for short):

(2.13)

$$\begin{aligned}
&-\int_\Omega m(u)\nabla\Delta u\nabla g^-_{\alpha\sigma}(u+\varepsilon)\\
&\ge \int_\Omega S_1(u)\left[D^2\big((u+\varepsilon)^\gamma\big)\right]^2 + \int_\Omega S_2(u)\left[\Delta\big((u+\varepsilon)^\gamma\big)\right]^2 + \int_\Omega S_3(u)\,|\nabla u|^4\\
&\quad -\int_\Omega\Big((\delta_1+\delta_3)(u+\varepsilon+\sigma)^{\alpha+n-3}\,|\nabla u|^4 + \delta_2(u+\varepsilon+\sigma)^{\alpha+n-1}(u+\varepsilon)^{-2}|\nabla u|^4\Big)\\
&\quad -\delta_3\int_\Omega\left|D^2(u+\varepsilon+\sigma)^{\frac{\alpha+n+1}{2}}\right|^2 - \frac{C}{\delta_1}\int_\Omega \beta_{\varepsilon\sigma}(u)\left(\frac{m(u)}{m(u+\varepsilon)}-1\right)|\Delta u|^2\\
&\quad -\frac{C}{\delta_2}\int_\Omega \beta_{\varepsilon\sigma}(u)|\Delta u|^2\left(\frac{\varepsilon}{u+\varepsilon}\right)^2 - C_{\delta_3}\int_\Omega (u+\varepsilon+\sigma)^{\alpha+n+1}\\
&\quad = \mathrm{I}+\mathrm{II}+\cdots+\mathrm{VIII}
\end{aligned}$$

with $\delta_i > 0$ arbitrary, $C > 0$ independent of t, and

$$\begin{aligned}
S_1(u) &= \tfrac{2}{3}\gamma^{-2}\beta_{\varepsilon\sigma}(u)\cdot(u+\varepsilon)^{2-2\gamma},\\
S_2(u) &= \tfrac{1}{3}\gamma^{-2}\beta_{\varepsilon\sigma}(u)\cdot(u+\varepsilon)^{2-2\gamma},\\
S_3(u) &= (1-\gamma)\left(\gamma-\tfrac{1}{3}\right)(u+\varepsilon+\sigma)^{\alpha+n-1}\cdot(u+\varepsilon)^{-2}\\
&\quad -\tfrac{1}{3}(\alpha+n-1)(\alpha+n-2)(u+\varepsilon+\sigma)^{\alpha+n-3}\\
&\quad +\tfrac{2}{3}(1-\gamma)(1-\alpha-n)\frac{(u+\varepsilon+\sigma)^{\alpha+n-2}}{(u+\varepsilon)}.
\end{aligned}$$

Before we prove the above inequality in part 3, we show that this inequality gives the assertion of Proposition 2.2.

Consequences of inequality (2.13). Observing that the first and third term in $S_3(u)$ are positive and that $\sigma > 0$, we have that $S_3(u) > 0$ if

$$c(\alpha,n,\gamma) := (1-\gamma)\left(\gamma-\tfrac{1}{3}\right) - \tfrac{1}{3}(\alpha+n-1)(\alpha+n-2) + \tfrac{2}{3}(1-\gamma)(1-\alpha-n)$$

is positive, i.e., if (2.6) is true.

Then, choosing $\delta_1, \delta_2, \delta_3$ sufficiently small, the terms IV and V in (2.13) can be absorbed in I, II, and III. In addition, we have the following:

- The integrand on the left-hand side is in $L^1(\Omega_T)$.
- As a consequence of the $L^2(I;H^2(\Omega))$-regularity of u, the integrands in the terms VI, VII, and VIII are uniformly bounded in $L^1(\Omega_T)$.
- The integrands in the terms I, II, and III are nonnegative.

This allows us to integrate over the whole space-time cylinder and to obtain from (2.10), (2.12), and (2.13) the existence of positive constants $\tilde{C}_1$ and $\tilde{C}_2$ independent of ε and σ such that

(2.14)
$$\begin{aligned}
&\int_\Omega G_{\alpha\sigma}(u+\varepsilon)_{|T} + \tilde{C}_1^{-1}\int_{\Omega_T}(u+\varepsilon+\sigma)^{\alpha+n-3}\,|\nabla u|^4\\
&\qquad + \tilde{C}_1^{-1}\int_{\Omega_T}\beta_{\varepsilon\sigma}(u)\cdot(u+\varepsilon)^{2-2\gamma}\left\{\tfrac{2}{3}\left|D^2\big((u+\varepsilon)^\gamma\big)\right|^2 + \tfrac{1}{3}\left|\Delta\big((u+\varepsilon)^\gamma\big)\right|^2\right\}\\
&\qquad \le \int_\Omega G_{\alpha\sigma}(u_0+\varepsilon) + \tilde{C}_2\Bigg\{\int_{\Omega_T}(u+\varepsilon+\sigma)^{\alpha+n+1}\\
&\qquad\qquad + \int_{\Omega_T}\beta_{\varepsilon\sigma}(u)\,|\Delta u|^2\left\{\frac{m(u)}{m(u+\varepsilon)} - 1 + \left(\frac{\varepsilon}{u+\varepsilon}\right)^2\right\}\Bigg\}.
\end{aligned}$$

Passage to the limit $\varepsilon \to 0$ *in inequality* (2.14). From the $L^2(I;H^2(\Omega))\cap L^\infty(I;H^1(\Omega))$-regularity of u, the uniform boundedness of $\beta_{\varepsilon\sigma}(u)$, and Lebesgue's theorem, we infer that the second term on the right-hand side converges to $\int_{\Omega_T} u^{\alpha+n+1}$ and that the last term on the right-hand side tends to zero. Since $u_0 \ge \delta > 0$ and $u_0 \in H^1(\Omega)$, a further application of Lebesgue's theorem gives that $\int_\Omega G_{\alpha\sigma}(u_0+\varepsilon)$ converges to $\int_\Omega G_{\alpha\sigma}(u_0)$ for $\varepsilon \to 0$.

Since the third term on the left-hand side can also be written in the form $\int_{\Omega_T}|\nabla(u+\varepsilon+\sigma)^{\frac{\alpha+n+1}{4}}|^4$, it is obvious that $\int_{\Omega_T}|\nabla(u+\sigma)^{\frac{\alpha+n+1}{4}}|^4$ is dominated by $\liminf_{\varepsilon\to 0}\int_{\Omega_T}(u+\varepsilon+\sigma)^{\alpha+n-3}\,|\nabla u|^4$. Similarly, the second term on the left-hand side can be handled,

and by use of Fatou's lemma we derive

(2.15)
$$\int_\Omega G_{\alpha\sigma}\big(u(T)\big)\,dx + C_1^{-1}\int_{\Omega_T}(u+\sigma)^{\alpha+n-1}\cdot u^{2-2\gamma}\left\{\tfrac{2}{3}\left|D^2u^\gamma\right|^2+\tfrac{1}{3}\left|\Delta u^\gamma\right|^2\right\}dxdt$$
$$+C_1^{-1}\int_{\Omega_T}(u+\sigma)^{\alpha+n-3}\left|\nabla u\right|^4\,dxdt$$
$$\le\int_\Omega G_{\alpha\sigma}(u_0)\,dx + C_2\int_{\Omega_T}(u+\sigma)^{\alpha+n+1}\,dxdt\,.$$

Passage to the limit $\sigma\to 0$ in inequality (2.15). An application of Fatou's lemma, the monotone convergence theorem, and the same methods to estimate limits of gradients of powers of u as used before gives the following result:

(2.16)
$$\int_\Omega G_{\alpha}\big(u(T)\big)\,dx + C_1^{-1}\int_{\Omega_T}u^{\alpha+n+1-2\gamma}\left\{\tfrac{2}{3}\left|D^2u^\gamma\right|^2+\tfrac{1}{3}\left|\Delta u^\gamma\right|^2\right\}dxdt$$
$$+C_1^{-1}\int_{\Omega_T}u^{\alpha+n-3}\left|\nabla u\right|^4\,dxdt$$
$$\le\int_\Omega G_{\alpha}(u_0)\,dx\,+C_2\int_{\Omega_T}u^{\alpha+n+1}\,dxdt\,.$$

Part 3. (Proof of inequality (2.13)). Integrating by parts and using $\frac{\partial u}{\partial\nu}=0$ on $\partial\Omega$, we obtain

$$-\int_\Omega m(u)\,\nabla\Delta u\,\nabla g^-_{\alpha\sigma}(u+\varepsilon)\,dx=\int_\Omega \operatorname{div}\left(\frac{m(u)}{m(u+\varepsilon)}\,\beta_{\varepsilon\sigma}(u)\,\nabla u\right)\Delta u\,dx$$
$$=\int_\Omega\left(\frac{m(u)}{m(u+\varepsilon)}\,\beta_{\varepsilon\sigma}(u)\right)\left|\Delta u\right|^2\,dx+\int_\Omega\left(\frac{m(u)}{m(u+\varepsilon)}\,\beta_{\varepsilon\sigma}(u)\right)_u\left|\nabla u\right|^2\,\Delta u\,dx=:I_1+I_2\,.$$

For I_1 and I_2 we calculate

$$I_1=\int_\Omega\beta_{\varepsilon\sigma}(u)\left|\Delta u\right|^2\,dx+\int_\Omega\beta_{\varepsilon\sigma}(u)\left(\frac{m(u)}{m(u+\varepsilon)}-1\right)\left|\Delta u\right|^2\,dx=I_1^1+I_1^2,$$
$$I_2=\int_\Omega\left\{\left(\frac{m'(u)}{m(u+\varepsilon)}-\frac{m(u)m'(u+\varepsilon)}{m(u+\varepsilon)^2}\right)\beta_{\varepsilon\sigma}(u)\right.$$
$$\left.+\frac{m(u)}{m(u+\varepsilon)}(\alpha+n-1)(u+\varepsilon+\sigma)^{\alpha+n-2}\right\}\left|\nabla u\right|^2\Delta u\,dx$$
$$=I_2^1+I_2^2.$$

Observing the existence of a constant C independent of ε with the property

$$\left|\frac{m'(u)}{m(u+\varepsilon)}-\frac{m(u)m'(u+\varepsilon)}{m(u+\varepsilon)^2}\right|\le C\cdot\left(\frac{u}{u+\varepsilon}\right)^{n-1}\frac{\varepsilon}{(u+\varepsilon)^2},$$

the term I_2^1 can be estimated as follows:

$$I_2^1\ge -C\int_\Omega\beta_{\varepsilon\sigma}(u)(u+\varepsilon)^{-1}\cdot\left|\nabla u\right|^2\cdot\left|\Delta u\right|\frac{\varepsilon}{u+\varepsilon}\,dx$$
$$\ge-\delta_2\int_\Omega\beta_{\varepsilon\sigma}(u)(u+\varepsilon)^{-2}\cdot\left|\nabla u\right|^4\,dx-\frac{\tilde C}{\delta_2}\int_\Omega\beta_{\varepsilon\sigma}(u)\left|\Delta u\right|^2\cdot\left(\frac{\varepsilon}{u+\varepsilon}\right)^2\,dx.$$

We now write I_2^2 as

$$\begin{aligned} I_2^2 &= (\alpha+n-1)\int_\Omega (u+\varepsilon+\sigma)^{\alpha+n-2}\,|\nabla u|^2\,\Delta u\,dx \\ &\quad +(\alpha+n-1)\int_\Omega (u+\varepsilon+\sigma)^{\alpha+n-2}\left(\frac{m(u)}{m(u+\varepsilon)}-1\right)|\nabla u|^2\,\Delta u\,dx \\ &= R_1+R_2. \end{aligned}$$

R_2 can be estimated as

$$\begin{aligned} R_2 \geq &-\delta_1\int_\Omega (u+\varepsilon+\sigma)^{\alpha+n-3}\,|\nabla u|^4\,dx \\ &-\frac{C}{\delta_1}\int_\Omega (u+\varepsilon+\sigma)^{\alpha+n-1}\left(\frac{m(u)}{m(u+\varepsilon)}-1\right)^2|\Delta u|^2\,dx. \end{aligned}$$

To proceed further it will be worthwhile to state a formula of integration by parts that generalizes the one-dimensional formula

$$\int_0^1 f'(u)\cdot u_x^2\cdot u_{xx}\,dx = -\tfrac{1}{3}\int_0^1 f''(u)\cdot u_x^4\,dx \qquad \text{if } u_x(0)=u_x(1)=0$$

to higher space dimensions.

LEMMA 2.3. *Let $f\in W^{2,\infty}(\mathbb{R})$ and $u\in H^2_*(\Omega)$. Then we have*

$$\begin{aligned} \int_\Omega f'(u)\,|\nabla u|^2\,\Delta u\,dx = &-\tfrac{1}{3}\int_\Omega f''(u)\,|\nabla u|^4\,dx \\ &+\tfrac{2}{3}\left(\int_\Omega f(u)\left|D^2u\right|^2\,dx-\int_\Omega f(u)\,|\Delta u|^2\,dx\right) \\ &+\tfrac{2}{3}\int_{\partial\Omega} f(u)II\big(\nabla u\big)\,d\mathcal{H}^{N-1}. \end{aligned}$$

Here, $II(\cdot)$ denotes the second fundamental form of $\partial\Omega$.

For a proof of Lemma 2.3 we refer to the Appendix.

Now, setting $f(u)=\beta_{\varepsilon\sigma}(u)$ and using the identity

$$D_{x_ix_j}v = \gamma^{-1}v^{1-\gamma}\,D_{x_ix_j}v^\gamma-(\gamma-1)\,v^{-1}\,D_{x_i}v\,D_{x_j}v$$

for $v=u+\varepsilon$, R_1 reads as

$$\begin{aligned} R_1 = &-\tfrac{1}{3}\int_\Omega \beta''_{\varepsilon\sigma}(u)\,|\nabla u|^4+\tfrac{2}{3}\gamma^{-2}\int_\Omega \beta_{\varepsilon\sigma}(u)\,(u+\varepsilon)^{2-2\gamma}\left\{\left[D^2(u+\varepsilon)^\gamma\right]^2-\left[\Delta(u+\varepsilon)^\gamma\right]^2\right\} \\ &+2(\gamma-1)\int_\Omega \beta_{\varepsilon\sigma}(u)\,(u+\varepsilon)^{-1}\,\Delta u\,|\nabla u|^2+\tfrac{2}{3}(\gamma-1)\int_\Omega\left((u+\varepsilon)^{-1}\,\beta_{\varepsilon\sigma}(u)\right)_u|\nabla u|^4 \\ &+\tfrac{2}{3}\int_{\partial\Omega}\beta_{\varepsilon\sigma}(u)\,II\big(\nabla u\big)\,d\mathcal{H}^{N-1}. \end{aligned}$$

For I_1^1 we obtain

$$\begin{aligned} I_1^1 = &\gamma^{-2}\int_\Omega \beta_{\varepsilon\sigma}(u)\,(u+\varepsilon)^{2-2\gamma}\left[\Delta\big((u+\varepsilon)^\gamma\big)\right]^2-2(\gamma-1)\int_\Omega \beta_{\varepsilon\sigma}(u)\,(u+\varepsilon)^{-1}\,\Delta u\,|\nabla u|^2 \\ &-(\gamma-1)^2\int_\Omega \beta_{\varepsilon\sigma}(u)\,(u+\varepsilon)^{-2}\,|\nabla u|^4\ . \end{aligned}$$

Summing up, we finally arrive at

$$
\begin{aligned}
I_1^1 + R_1 = &\int_\Omega S_1(u) \left[D^2\big((u+\varepsilon)^\gamma\big)\right]^2 \, dx + \int_\Omega S_2(u) \left[\Delta\big((u+\varepsilon)^\gamma\big)\right]^2 \, dx \\
&+ \int_\Omega S_3(u) \, |\nabla u|^4 \, dx + \tfrac{2}{3} \int_{\partial\Omega} \beta_{\varepsilon\sigma}(u) \, II\big(\nabla u\big) \, d\mathcal{H}^{N-1}
\end{aligned} \tag{2.17}
$$

with $S_1(u), S_2(u)$, and $S_3(u)$ as in (2.13). Collecting all the terms, (2.13) will be established, provided we can estimate the boundary term accordingly.

Let us remark that convexity of Ω implies the nonnegativity of the last term on the right-hand side in (2.17). Hence in this case we can neglect it for our a priori estimates, and in particular the last term on the right-hand side of (2.13) cancels out. Let us now concentrate on the case that Ω is not convex. First we state an interpolation inequality that easily can be proved by contradiction using the compactness of the imbedding $H^1(\Omega) \to L^2(\partial\Omega)$ which holds true for domains with Lipschitz boundary.

LEMMA 2.4. *Let Ω be of class $C^{0,1}$. For each $\varepsilon > 0$ there exists a constant $C_\varepsilon < \infty$ such that we have, for all $v \in H^2(\Omega)$,*

$$
\|\nabla v\|_{L^2(\partial\Omega)} \le \varepsilon \left\|D^2 v\right\|_{L^2(\Omega)} + C_\varepsilon \|v\|_{L^2(\Omega)} .
$$

Since $\partial\Omega$ is of class $C^{1,1}$ we can conclude that the second fundamental form of $\partial\Omega$ is uniformly bounded. Thus we have to estimate $\int_{\partial\Omega} \beta_{\varepsilon\sigma}(u) \, |\nabla u|^2 \, d\mathcal{H}^{N-1}$. Choosing a function $\Phi_{\varepsilon\sigma} : \mathbb{R} \to \mathbb{R}$ with $\Phi'_{\varepsilon\sigma} = \sqrt{\beta_{\varepsilon\sigma}}$ and applying Lemma 2.4, we obtain, after straightforward calculations,

$$
\int_{\partial\Omega} \beta_{\varepsilon\sigma}(u) \, |\nabla u|^2 \, d\mathcal{H}^{N-1} \le \delta_3 \int_\Omega \left|D^2 \Phi_{\varepsilon\sigma}(u)\right|^2 \, dx + C_{\delta_3} \int_\Omega |\Phi_{\varepsilon\sigma}(u)|^2 \, dx.
$$

Using the relation $\Phi_{\varepsilon\sigma}(u) = \text{const} \cdot (u + \varepsilon + \sigma)^{\frac{\alpha+n+1}{2}}$, we end up with

$$
\begin{aligned}
-\int_{\partial\Omega} \beta_{\varepsilon\sigma}(u) \, |\nabla u|^2 \, d\mathcal{H}^{N-1} \ge &- \delta_3 \int_\Omega \left|D^2 (u+\varepsilon+\sigma)^{\frac{\alpha+n+1}{2}}\right|^2 \, dx \\
&- C_{\delta_3} \int_\Omega |u+\varepsilon+\sigma|^{\alpha+n+1} \, dx.
\end{aligned}
$$

This proves estimate (2.13) and therefore Proposition 2.2. ∎

Our next goal is to establish entropy estimates in the spirit of Proposition 2.2 for diffusion coefficients m which are bounded from below by positive constants for large values of τ. We distinguish two cases:

(i)

$$
m \in C^1(\mathbb{R}^+) \cap W^{1,\infty}(1,\infty) \text{ with } n > 0 \text{ arbitrary} \tag{2.18}
$$

and (ii)

$$
\begin{aligned}
m(\tau) = |\tau|^n \cdot f(\tau) \qquad &\text{with } n > 0 \text{ arbitrary if } N = 2, \\
&0 < n < 4 \qquad \text{if } N = 3,
\end{aligned} \tag{2.19}
$$

$$
f \in C^2(\mathbb{R}, \mathbb{R}^+) \cap L^\infty(\mathbb{R}, \mathbb{R}^+). \tag{2.20}
$$

We shall proceed as follows: for a special choice of mobilities $(m_\delta)_{\delta>0}$ which approach m from below we obtain entropy estimates by use of Proposition 2.2.

A compactness lemma which offers all the convergence properties necessary for passing to the limit with $\delta \searrow 0$ will be crucial. It reads as follows.

LEMMA 2.5. *Let $(u_\delta, J_\delta)_{\delta\searrow 0}$ be a family of pairs of functions having the following properties:*

(i) $(u_\delta)_t = -\operatorname{div} J_\delta$ *in* $L^2(I;(W^{1,q}(\Omega))')$ $\quad \forall q > \frac{4N}{2N+(2-N)\cdot n}$,

(ii) $(J_\delta)_{\delta\searrow 0}$ *is uniformly bounded in* $L^2\big(I;L^{q'}(\Omega)\big)$ $\quad \forall q' < \frac{4N}{2N+(N-2)\cdot n}$,

(iii) $(u_\delta)_{\delta\searrow 0}$ *are nonnegative and for a number* $\beta \in \big(\frac{3}{4},\frac{3}{2}\big)$ *we have* $(u_\delta^\beta)_{\delta\searrow 0}$ *is uniformly bounded in* $L^2\big(I;H^2(\Omega)\big)$,

(iv) $(u_\delta)_{\delta\searrow 0}$ *is uniformly bounded in* $L^\infty\big(I;H^1(\Omega)\big)$.

Under the assumption that $q < \frac{2N}{N-2}$, the family $(u_\delta^\beta)_{\delta\searrow 0}$ is relatively compact in $L^2\big(I;H^1(\Omega)\big)$.

Remark. In dimension $N = 2$ there exists for every $n > 0$ a real number q in agreement with (i) which satisfies $q < \frac{2N}{N-2}$. In dimension $N = 3$ we need the condition $n < 4$, unless $(J_\delta)_{\delta\searrow 0}$ is uniformly bounded in $L^2(\Omega_T)$. The assumption $\beta \in (\frac{3}{4},\frac{3}{2})$ is not the most general condition one could impose, but it is nevertheless sufficient for our applications.

Proof of Lemma 2.5. It mainly consists of three steps.

(1) There is a constant $C > 0$ independent of $\delta > 0$ such that

$$\int_0^{T-h}\int_\Omega \big(u_\delta^\beta(t+h,x)-u_\delta^\beta(t,x)\big)\big(u_\delta(t+h,x)-u_\delta(t,x)\big)\,dxdt \le C\cdot h. \tag{2.21}$$

To prove this assertion we choose as the test function in the weak formulation of $(u_\delta)_t = -\operatorname{div} J_\delta$ the function $\phi(t,x) = \big(u_\delta^\beta(s+h,x)-u_\delta^\beta(s,x)\big)\cdot\chi_{[s,s+h]}(t)$ which is admissible. Integrating over $[0,T]$ with respect to t and over $[0,T-h]$ with respect to s, we arrive at

$$\begin{aligned}
&\left|\int_0^{T-h}\int_\Omega \big(u_\delta^\beta(s+h,x)-u_\delta^\beta(s,x)\big)\big(u_\delta(s+h,x)-u_\delta(s,x)\big)\,dsdx\right| \\
&\le \left|\int_0^{T-h}\int_s^{s+h}\int_\Omega J_\delta(t,x)\cdot\nabla\big(u_\delta^\beta(s+h,x)-u_\delta^\beta(s,x)\big)\,dxdtds\right| \\
&\le \int_0^{T-h}\int_0^h \|J_\delta(s+\tau,\cdot)\|_{q'}\cdot\Big\|\nabla\big(u_\delta^\beta(s+h,.)-u_\delta^\beta(s,.)\big)\Big\|_q\,d\tau\,ds \\
&\le 2\int_0^h \|J_\delta\|_{L^2(L^{q'})}\cdot\big\|\nabla u^\beta\big\|_{L^2(L^q)}\,d\tau \;\le 2\cdot C\cdot h.
\end{aligned}$$

(2) A subsequence of $(u_\delta^\beta)_{\delta\searrow 0}$ converges to u^β strongly in $L^1(\Omega_T)$.

In order to derive (2) we observe that

(a) $u_\delta \rightharpoonup u$ in $L^2\big(I;H^1(\Omega)\big)$ according to (iv),

(b) $(u_\delta^{\beta+1})_{\delta\searrow 0}$ is uniformly bounded in $L^\infty\big(I;L^1(\Omega)\big)$ (cf. iv),

(c) (2.21) holds true.

With the notation $b(u) = u^\beta$ and $B(u) = \frac{\beta}{\beta+1}u^{\beta+1}$ we can apply the following lemma (for a proof, see Alt and Luckhaus [1, Lemma 1.9]) in order to establish the result.

LEMMA. *Suppose that $(u_\delta)_{\delta\searrow 0}$ is a sequence which converges weakly to u in the*

space $L^r((0,T);W^{1,r}(\Omega))$ *and satisfies the estimates*

$$\frac{1}{h}\int_0^{T-h}\int_\Omega (b(u_\delta(t+h))-b(u_\delta(t)))(u_\delta(t+h)-u_\delta(t))dt \le C$$

and

$$\int_\Omega B(u_\delta(t)) \le C \quad \textit{for } 0<t<T,$$

uniformly with respect to δ. *Then* $b(u_\delta)\to b(u)$ *in* $L^1(\Omega_T)$ *and* $B(u_\delta)\to B(u)$ *almost everywhere.*

Let us remark that in the case $q=2$ this convergence result could have been deduced directly by using $u_t \in L^2(I;(H^1(\Omega))')$ and $u\in L^\infty(I;H^1(\Omega))$.

(3) From Fréchet–Kolmogorov's theorem and (2) we infer that

$$\lim_{h\to 0}\left\|u_\delta^\beta(\cdot+h,\cdot)-u_\delta^\beta(\cdot,\cdot)\right\|_{L^1\left((0,T-h),L^1(\Omega)\right)}=0$$

uniformly for $\delta>0$.

We then apply the following theorem due to J. Simon.

THEOREM (see [14, p. 84]). *Let* $X\subset B\subset Y$ *with compact imbedding* $X\hookrightarrow B$ *and* $1\le p\le\infty$.

If $F\subset L^p(I;X)$ *is bounded and* $\|f(\cdot+h,\cdot)-f(\cdot)\|_{L^p(0,T-h,Y)}\to 0$ *uniformly for* $f\in F$ *as* $h\to 0$, *then* F *is relatively compact in* $L^p(I;B)$.

With the choice $p=1$, $X=H^2(\Omega)$, $B=H^1(\Omega)$, and $Y=L^1(\Omega)$, we obtain at first that $(u_\delta^\beta)_{\delta\searrow 0}$ is relatively compact in $L^1\big(I;H^1(\Omega)\big)$, and then with (iv) we immediately obtain the assertion. □

Let us now specify the auxiliary diffusion coefficients m_δ which we want to use in order to obtain entropy estimates analogous to Proposition 2.2. We consider two cases. At first we study the case when m is bounded, but we allow the growth exponent near zero to be arbitrary. Having shown convergence of approximating solutions in this case, we are in a position to investigate situations where m has polynomial growth at infinity. Thus, the auxiliary diffusion coefficients read as follows.

(1) If m is bounded (like in (2.18)), we choose

$$m_\delta^{(1)}(\tau)=\frac{\tau^s m(\tau)}{\delta m(\tau)+\tau^s}\ , \tag{2.22}$$

with sufficiently large s to apply Proposition 2.2.

(2) If m has polynomial growth, i.e.,

$$m(\tau)=|\tau|^n\cdot f(\tau) \qquad \text{with } n\in\begin{cases}(0,\infty) & \text{if } N=2,\\ (0,4) & \text{if } N=3,\end{cases} \tag{2.23}$$

and f as in (2.20), then we choose

$$m_\delta^{(2)}(\tau)=\frac{\tau^n}{1+\delta\,|\tau|^n}\cdot f(\tau).$$

This leads to the following auxiliary problems:

$$P_\delta^i\begin{cases}(u_\delta^{(i)})_t+\operatorname{div}\big(m_\delta^{(i)}(u_\delta^{(i)})\nabla\Delta u_\delta^{(i)}\big)=0 & \text{in } \Omega_T,\\ \frac{\partial}{\partial\nu}u_\delta^{(i)}=\frac{\partial}{\partial\nu}\Delta u_\delta^{(i)}=0 & \text{on } \partial\Omega\times[0,T],\\ u_{\delta 0}^{(i)}=u_0+\delta^{\Theta_i} & \text{in } \Omega,\end{cases} \tag{2.24}$$

with $0 < \Theta_1 < \frac{1}{s-\alpha-n-1}$ and $0 < \Theta_2$.

Using the notation

$$
\begin{aligned}
G^{(i)}_{\alpha\delta}(t) &:= \int_A \int_A \frac{\tau^{\alpha+n-1}}{m^{(i)}_\delta(\tau)}\, d\tau ds \qquad \text{and} \\
G^{(i)}_{\alpha}(t) &:= \int_A \int_A \frac{\tau^{\alpha+n-1}}{m^{(i)}(\tau)}\, d\tau ds,
\end{aligned} \tag{2.25}
$$

we have the following proposition.

PROPOSITION 2.6. *Let $n_i > 0$ be the growth exponent of $m^{(i)}_\delta$ near zero and let $q_i > \frac{4N}{2N+(2-N)n_i}$. Assume that there exist constants $C_i > 0$ and α_i with $\frac{1}{2} < \alpha_i + n_i < 2$ such that $\int_\Omega G^{(i)}_\alpha(u_0)\, dx < C_i$. In the case $i = 2$ we furthermore require that the pair (N, n_2) satisfy the conditions in (2.23). Then for a subsequence $(u^{(i)}_\delta)_{\delta \searrow 0}$ of solutions to the auxiliary problems P^i_δ the following convergence properties hold true:*

(i) $u^{(i)}_\delta \overset{*}{\rightharpoonup} u^{(i)}$ *in* $L^\infty\big(I; H^1(\Omega)\big)$,

(ii) $J^{(i)}_\delta \rightharpoonup J^{(i)}$ *in* $L^2\big(I; L^2(\Omega)\big)$ *if* $i = 1$ *or*
in $L^2\big(I; L^{q_i'}(\Omega)\big)$ *if* $i = 2$, *respectively*,

(iii) $\big(u^{(i)}_\delta\big)_t \rightharpoonup \big(u^{(i)}\big)_t$ *in* $L^2\big(I; \big(H^1(\Omega)\big)'\big)$ *if* $i = 1$ *or*
in $L^2\big(I; \big(W^{1,q_i}(\Omega)\big)'\big)$ *if* $i = 2$, *respectively*,

(iv) $\big(u^{(i)}_\delta\big)^{\frac{\alpha+n+1}{2}} \rightharpoonup \big(u^{(i)}\big)^{\frac{\alpha+n+1}{2}}$ *in* $L^2\big(I; H^2(\Omega)\big)$,

(v) $\big(u^{(i)}_\delta\big)^{\frac{\alpha+n+1}{4}} \rightharpoonup \big(u^{(i)}\big)^{\frac{\alpha+n+1}{4}}$ *in* $L^4\big(I; W^{1,4}(\Omega)\big)$,

(vi) $\big(u^{(i)}_\delta\big)^{\frac{\alpha+n+1}{2}} \to \big(u^{(i)}\big)^{\frac{\alpha+n+1}{2}}$ *strongly in* $L^2\big(I; H^1(\Omega)\big)$,

where q_i' denotes the conjugate exponent to q_i.

For the limiting function $u^{(i)}$ the following estimate is valid with constants C_1 and C_2 which only depend on the domain Ω (in particular, C_2 is zero if Ω is convex):

$$
\begin{aligned}
\sup_{t\in[0,T]} \int_\Omega G^{(i)}_\alpha\big(u^{(i)}(t)\big) &+ C_1^{-1} \int_{\Omega_T} \Big| D^2 (u^{(i)})^{\frac{\alpha+n+1}{2}} \Big|^2 + C_1^{-1} \int_{\Omega_T} \Big| \nabla (u^{(i)})^{\frac{\alpha+n+1}{4}} \Big|^4 \\
&\le \int_\Omega G^{(i)}_\alpha(u_0) + C_2 \int_{\Omega_T} (u^{(i)})^{\alpha+n+1}\,.
\end{aligned} \tag{2.26}
$$

Proof. We present it in detail only for the case $i = 1$, drop here the superscript (i) and indicate the main modification necessary for the other case. For a given α satisfying $\frac{1}{2} < \alpha + n < 2$ we choose $\tilde\alpha := \alpha + (n - s)$ and apply Proposition 2.2 with $\frac{1}{2} < \tilde\alpha + s < 2$ to u_δ. This gives after rewriting in terms of α and n:

$$
\begin{aligned}
\sup_{t\in[0,T]} \int_\Omega G_{\alpha\delta}\big(u_\delta(t,x)\big) &+ C_1^{-1} \int_{\Omega_T} u_\delta^{\alpha+n+1-2\gamma} \big| D^2 u_\delta^\gamma \big|^2 + C_1^{-1} \int_{\Omega_T} u_\delta^{\alpha+n-3} \big| \nabla u_\delta \big|^4 \\
&\le \int_\Omega G_{0\delta}\big(u_{0\delta}(x)\big) + C_2 \int_{\Omega_T} u_\delta^{\alpha+n+1}\,.
\end{aligned} \tag{2.27}
$$

Using now the identity

$$
D_{x_i x_j} u^\gamma = \gamma(\gamma-1)\, u^{\gamma-2} D_{x_i} u D_{x_j} u + \gamma u^{\gamma-1} D_{x_i x_j} u
$$

and Young's inequality, we observe after straightforward calculations the existence of a new constant C_1 independent of δ, γ such that

$$\begin{aligned}\sup_{t\in[0,T]}\int_\Omega G_{\alpha\delta}\big(u_\delta(t,x)\big) &+ C_1^{-1}\int_{\Omega_T}\Big|D^2 u_\delta^{\frac{\alpha+n+1}{2}}\Big|^2 + C_1^{-1}\int_{\Omega_T}\Big|\nabla u_\delta^{\frac{\alpha+n+1}{4}}\Big|^4 \\ &\le \int_\Omega G_{\alpha\delta}\big(u_{0\delta}(x)\big) + C_2\int_{\Omega_T} u_\delta^{\alpha+n+1}\ .\end{aligned} \tag{2.28}$$

From estimate (2.5) we infer the validity of (i), (ii), and (iii). Combining these results with the inequality

$$\int_\Omega G_{\alpha\delta}\big(u_{0\delta}(x)\big)\,dx \le \int_\Omega G_\alpha\big(u_0(x)\big)\,dx + o_\delta(1),$$

we obtain by use of Lemma 2.5 the validity of (iv)–(vi). Writing

$$G_{\alpha\delta}(t) = \int_A^t\int_A^s G_\alpha(\tau)d\tau ds + \delta\int_A^t\int_A^s \tau^{\alpha+n-s-1}d\tau ds,$$

observing that the second term on the right is nonnegative, and using both Fatou's lemma and the convergence of u_δ to u pointwise almost everywhere, we end up with (2.26).

For the case $i = 2$ we only have to convince ourselves that J_δ is uniformly bounded in $L^2\big(I; L^{q'}(\Omega)\big)$ for $q' < \frac{4N}{2N+(N-2)n}$. This can be seen by use of Hölder's inequality and the uniform boundedness of the quantities

$$\int_{\Omega_T}\big(m_\delta(u_\delta^\varepsilon)+\varepsilon\big)\,|\nabla\Delta u_\delta^\varepsilon|^2\,dxdt$$

which occur during the proof of existence of solutions to degenerate problems by use of nondegenerate auxiliary problems (cf. [13], [10]). □

Remark. For $N = 3$ the Sobolev imbedding theorem implies that the limit u from Proposition 2.6 belongs to $L^\infty\big(I; L^6(\Omega)\big)$. Using the fact that $\nabla u^{\frac{\alpha+n+1}{4}} \in L^4(\Omega_T)$, we can apply interpolation theory (cf. DiBenedetto [9, Proposition 3.2]) to get L^p-regularity of u for all $p \le \alpha + n + 9$.

As a further consequence of Proposition 2.6 we obtain the following corollary.

COROLLARY 2.7. *Under the same assumptions as in Proposition* 2.6*, for a subsequence of* $(u_\delta^{(i)})_{\delta\searrow 0}$ *the following is true:*

$$\nabla(u_\delta^{(i)})^{\frac{\alpha+n+1}{4}} \longrightarrow \nabla(u^{(i)})^{\frac{\alpha+n+1}{4}} \ \textit{strongly in } L^{4-}\big([u^{(i)}>0]\big) \textit{ and pointwise a.e.}, \tag{2.29}$$

$$\nabla u_\delta^{(i)} \longrightarrow \nabla u^{(i)} \ \textit{strongly in } L^2(\Omega_T). \tag{2.30}$$

Proof. With the help of Vitali's theorem, relation (2.29) follows from point (v) of Proposition 2.6 and the convergence pointwise almost everywhere of $(\nabla(u_\delta^{(i)})^{\frac{\alpha+n+1}{2}})_{\delta\searrow 0}$ on the set $[u^{(i)} > 0]$.

Concerning the convergence behavior of $(\nabla u_\delta^{(i)})_{\delta\searrow 0}$, we observe—combining (2.29), the identity $|\nabla v|^2 = (\frac{4}{\alpha+n+1})^2 v^{\frac{3-\alpha-n}{2}} |\nabla v^{\frac{\alpha+n+1}{4}}|^2$, and the points (v) and (vi) of Proposition 2.6—that for δ tending to zero $||u_\delta^{(i)}||_{L^2(\Omega_T)}$ converges to $||u^{(i)}||_{L^2(\Omega_T)}$.

Now we recall the following well-known lemma.

LEMMA. *Assume that a sequence $(u_n)_{n\in\mathbb{N}}$ weakly converges to u in a Hilbert space X. If additionally $(\|u_n\|_X)_{n\in\mathbb{N}}$ converges to $\|u\|_X$, then $(u_n)_{n\in\mathbb{N}}$ strongly converges to u in X.* □

Hence, (2.30) can easily be established.

3. Existence, qualitative behavior, and (non)uniqueness of solutions. In this section we improve the existence results obtained previously in [13] and [10]. In particular, it will be possible to treat initial values with compact support in the case $2 \le n < 3$ which is important in lubrication theory. Moreover, we are able to extend the range of allowed diffusion growth exponents beneath 1 to $\frac{1}{8}$ and we propose a solution concept also for diffusion coefficients with polynomial growth.

Let us begin with the following assumption (A4) which basically requires that m (m', m'', respectively) have at most polynomial growth with the exponents n ($n-1, n-2$, respectively).

(A4) The diffusion coefficient m and the first two derivatives can be written as $m(\tau) = \tau^n \cdot f_0(\tau)$, $m'(\tau) = \tau^{n-1}\cdot f_1(\tau)$, and $m''(\tau) = \tau^{n-2}\cdot f_2(\tau) (\tau \in \mathbb{R}_0^+, i = 0,1,2)$ with functions f_i such that $\|f_i\|_{C^{2-i}(\mathbb{R}_0^+,\mathbb{R}_0^+)} < \infty$ $(i = 0,1,2)$. Furthermore, we assume that f_0 is positive and that m is bounded from below by a positive constant for large values of τ.

Let us present our concept of solution.

DEFINITION 3.1. *Let $N \ge 2$, $n > 0$, and m satisfy* (A4). *Let (u, J) be the element of $L^\infty(I; H^1(\Omega)) \cap H^1(I; (W^{1,q}(\Omega))') \times L^2(I; L^{q'}(\Omega; \mathbb{R}^N))$ where q satisfies one of the following properties:*

(i) *$q = 2$ if $m \in L^\infty(\mathbb{R})$,*

(ii) *$q > \frac{4N}{2N+(2-N)n}$ (and $n < \frac{2N}{N-2}$ if $N \ge 3$).*

We call the pair (u, J) the solution of (1.1) *if*

$$u_t = -\operatorname{div} J \quad \text{in } L^2\big(I; \big(W^{1,q}(\Omega)\big)'\big)\ , \tag{3.1}$$

$m''(u)\,|\nabla u|^3$ is in $L^1([u > 0])$, and J satisfies the relation $J = m(u)\nabla\Delta u$ in the following weak sense:

$$\begin{aligned}\int_{\Omega_T} J\cdot\eta\,dxdt = {}& \tfrac{1}{2}\int_{[u>0]} m''(u)\ |\nabla u|^2\ \nabla u\eta + \tfrac{1}{2}\int_{[u>0]} m'(u)\ |\nabla u|^2\ \operatorname{div}\eta \\ &+ \int_{[u>0]} m'(u)\big\langle \nabla u, D\eta, \nabla u\big\rangle + \int_{\Omega_T} m(u)\,\nabla u\nabla\operatorname{div}\eta \\ &\forall \eta \in L^\infty\big(I; W^{2,\infty}(\Omega;\mathbb{R}^N)\big) \text{ such that } \eta\cdot\nu = 0 \text{ on } \partial\Omega\ .\end{aligned} \tag{3.2}$$

(Here, q' denotes the conjugate exponent to q.)

Remark. Property (ii) is only necessary as long as there does not exist a positive result about boundedness of solutions.

Our main existence result reads as follows.

THEOREM 3.2 (existence of regular solutions). *Assume that the diffusion coefficient satisfies* (A4) *and that the initial value $u_0 \in H^1(\Omega)$ is nonnegative and satisfies*

$$\int_\Omega G_\alpha(u_0)\,dx < \infty$$

for a certain constant $\alpha \in (\frac{1}{2}-n, 2-n)$. *If one of the following combinations is true,*

(i) $N = 2$ *and* $n > \frac{1}{8}$,

(ii) $N = 3$, $m \in L^\infty(\mathbb{R})$, *and* $n > \frac{1}{8}$,

(iii) $N = 3$, $\frac{1}{8} < n < 4$, *and* $\alpha > -2$,

then there exists a solution (u, J) *to equation* (1.1) *in the sense of Definition* 3.1.

In particular, u *has the following additional regularity properties:*

$$u^{\frac{\alpha+n+1}{4}} \in L^4(I; W^{1,4}(\Omega)),$$
$$u^{\frac{\alpha+n+1}{2}} \in L^2(I; H^2(\Omega)).$$

Remark.

(i) For diffusion growth coefficients $0 < n < 3$, it is always possible to find a real number $\alpha_0 \in (\frac{1}{2}-n, 2-n)$ such that $\alpha_0 + 1 > 0$. Thus $\int_\Omega G_{\alpha_0}(u_0)dx < \infty$ for arbitrary nonnegative initial data $u_0 \in H^1(\Omega)$, i.e., for arbitrary $n \in (\frac{1}{8}, 3)$ and arbitrary nonnegative initial data $u_0 \in H^1(\Omega)$, a regular solution to (1.1) does exist.

(ii) Let us emphasize that in order to give a meaning to the first three integrands on the right-hand side of (3.2), we make essential use of the identities $u^{n-2}|\nabla u|^3 = \big(\frac{4}{\alpha+n+1}\big)^3 u^{\frac{n+1-3\alpha}{4}} \Big|\nabla u^{\frac{\alpha+n+1}{4}}\Big|^3$ and $u^{n-1}|\nabla u|^2 = \big(\frac{4}{\alpha+n+1}\big)^2 u^{\frac{n+1-\alpha}{2}} \Big|\nabla u^{\frac{\alpha+n+1}{4}}\Big|^2$.

Proof of Theorem 3.2. Let us begin with the case $m \in L^\infty(\mathbb{R})$ and $N = 2$ or 3. For the ease of presentation we now drop the superscript (i). From Proposition 2.6 and Corollary 2.7 we infer the following convergence behavior for a subsequence $(u_\delta, J_\delta)_{\delta\to 0}$ of solutions to the auxiliary problems P_δ^1:

(i) $u_{\delta t} \rightharpoonup u_t$ in $L^2\big(I; \big(H^1(\Omega)\big)'\big)$,

(ii) $\nabla u_\delta \longrightarrow \nabla u$ strongly in $L^2(\Omega_T)$,

(iii) $J_\delta \rightharpoonup J$ in $L^2(\Omega_T)$,

(iv) $\nabla u_\delta^{\frac{\alpha+n+1}{4}} \rightharpoonup \nabla u^{\frac{\alpha+n+1}{4}}$ in $L^4(\Omega_T)$,

(v) $\nabla u_\delta^{\frac{\alpha+n+1}{4}} \longrightarrow \nabla u^{\frac{\alpha+n+1}{4}}$ strongly in $L^{4-}\big([u > 0]\big)$ and pointwise a.e.

From (i) and (iii) relation (3.1) follows immediately. To proceed with the identification of J, we use the formula

$$\begin{aligned}\int_\Omega m_\delta(u_\delta)\,\nabla\Delta u_\delta \eta \;=\; &\tfrac{1}{2}\int_\Omega m_\delta''(u_\delta)\;|\nabla u_\delta|^2\;\nabla u_\delta\,\eta \;+\; \tfrac{1}{2}\int_\Omega m_\delta'(u_\delta)\;|\nabla u_\delta|^2\;\operatorname{div}\eta \\ &+ \int_\Omega m_\delta'(u_\delta)\langle\nabla u_\delta, D\eta, \nabla u_\delta\rangle \;+\; \int_\Omega m_\delta(u_\delta)\,\nabla u_\delta\nabla\operatorname{div}\eta,\end{aligned} \tag{3.3}$$

which is valid for $m_\delta \in C^2(\mathbb{R}) \cap W^{2,\infty}(\mathbb{R})$, $u_\delta \in H^2_*(\Omega)$ with $\nabla\Delta u_\delta \in L^2(\Omega)$ and $\eta \in W^{2,\infty}(\Omega)$ with $\eta \cdot \nu \equiv 0$ on $\partial\Omega$ as will be proved in the Appendix.

By our choice of m_δ we can identify $J_\delta(t)$ for a.e. $t \in I$ with $m_\delta(u_\delta)\,\nabla\Delta u_\delta$; thus J_δ can be related to $m_\delta(u_\delta)\,\nabla\Delta u_\delta$ in the sense of (3.2). Let us now pass to the limit $\delta \searrow 0$ on the right-hand side of (3.3). As the third term on the right-hand side of (3.3) qualitatively shows the same behavior as the second one and as the fourth term can easily be handled by using convergence property (ii), we will discuss in detail only the first and the second term. Writing $m_\delta'(\tau)$, $m_\delta''(\tau)$ as

$$m_\delta'(\tau) = \tau^{n-1} f_{1,\delta}(\tau) \text{ and } m_\delta''(\tau) = \tau^{n-2} f_{2,\delta}(\tau),$$

we observe after straightforward calculations that $f_{i,\delta}$ are uniformly bounded in $L^\infty(\mathbb{R}_0^+)$ $(i=1,2)$ and that on each compact subset of $(0,\infty)$, $f_{i,\delta}$ converges to f_i uniformly. For α with $\frac{1}{2} < \alpha + n < 2$ we write $\int_{\Omega_T} m_\delta''(u_\delta)\,|\nabla u|^2\,\nabla u \eta\,dxdt$ as

$$\int_{\Omega_T} u_\delta^{\frac{n+1-3\alpha}{4}} \left|\nabla u_\delta^{\frac{\alpha+n+1}{4}}\right|^2 \nabla u_\delta^{\frac{\alpha+n+1}{4}} \cdot f_{2,\delta}(u_\delta)\,\eta\,dxdt.$$

Condition $n > \frac{1}{8}$ ensures that α can be modified in such a way that $\frac{n+1-3\alpha}{4}$ is positive, and in addition $\int_\Omega G_\alpha(u_0)dx$ is bounded.

Using the properties of $f_{2,\delta}$, the regularity of u_δ, and $n > 1/8$, we observe that the term in u_δ converges weakly in $L^{1+\sigma}(\Omega_T)$ (for values of σ sufficiently small) to a function $\tilde{\beta}$. From (v) and the pointwise convergence of $f_{2,\delta}(u_\delta)$ we infer $\tilde{\beta} = m''(u)\,|\nabla u|^2 \cdot \nabla u$ on the set $[u>0]$.

On the set $[u=0]$ we argue as follows: for each $\varepsilon > 0$ we find by Egorov's theorem a set $S_\varepsilon \subset [u=0]$ with $\mathcal{L}^{N+1}(S_\varepsilon) < \varepsilon$ such that u_δ uniformly converges to u on the subset $[u=0] \setminus S_\varepsilon$. Thus we can estimate:

$$\begin{aligned}
&\left|\int_{[u=0]} m_\delta''(u_\delta)\,|\nabla u_\delta|^2\,\nabla u_\delta \cdot \eta\,dxdt\right| \\
&\quad\le \int_{S_\varepsilon \cup \left([u=0]\setminus S_\varepsilon\right)} u_\delta^{\frac{n+1-3\alpha}{4}} \cdot \left|\nabla u_\delta^{\frac{\alpha+n+1}{4}}\right|^3 \cdot |f_{2,\delta}(u_\delta)| \cdot |\eta|\;dxdt \\
&\quad\le o_\delta(1) \int_{[u=0]\setminus S_\varepsilon} \left|\nabla u_\delta^{\frac{\alpha+n+1}{4}}\right|^3 dxdt \\
&\qquad + \left(\int_{S_\varepsilon} u_\delta^{n+1-3\alpha}\,dxdt\right)^{\frac{1}{4}} \left(\int_{S_\varepsilon} \left|\nabla u_\delta^{\frac{\alpha+n+1}{4}}\right|^4 dxdt\right)^{\frac{3}{4}} \|f_{2,\delta}\|_\infty \cdot \|\eta\|_\infty .
\end{aligned} \tag{3.4}$$

By Vitali's theorem the second term in (3.4) converges to zero when $\mathcal{L}^{N+1}(S_\varepsilon) \to 0$. This gives the convergence of the first term on the right-hand side of (3.3).

For the second term on the right-hand side of (3.3) we write

$$\int_{\Omega_T} u_\delta^{\frac{-\alpha+n+1}{2}} \left|\nabla u_\delta^{\frac{\alpha+n+1}{4}}\right|^2 \cdot f_{1,\delta}(u_\delta) \cdot \operatorname{div}\eta\,dxdt$$

and use exactly the same technique as before in order to identify the limit. Putting everything together, the validity of (3.2) is established for m bounded. Eventually, we remark that the case of an unbounded diffusivity m which grows at most polynomially can be handled similarly, provided we use the restrictions on n in order to guarantee applicability of Hölder's inequality in the analogue to (3.4). Using the remark at the end of the first section, we see that this is possible if $\alpha > -2$. □

Let us point out that the regularity properties of u stated in Theorem 3.2 imply the following result about the behavior of the normal derivative of a solution at the boundary of $\operatorname{supp}(u)$.

COROLLARY 3.3 (regularity at the free boundary). *Let* $\hat{P} := \{t \in I : ||u(t,.)^{\frac{\alpha+n+1}{4}}||_{W^{1,4}(\Omega)} < \infty\}$ *with* α *as in Theorem* 3.2. *Then the following results are true:*

(i) *Under the assumption that* $\partial[supp(u(t,.)]$ *is an* $(N-1)$*-rectifiable set, we have for* $t \in \hat{P}$ *and for* $\mathcal{H}^{N-1}$*−almost every* $x \in \partial[supp(u(t,.))]$ *that the normal derivative* $\frac{\partial}{\partial \nu} u(t,x)$ *exists and that it is equal to zero.*

(ii) *If $N = 2$ and $n < 2$, then for arbitrary $t \in \hat{P}$ and for all $x \in [u(t,.) = 0]$ we have that $\nabla u(t,x)$ vanishes.*

Proof. Let us begin with the proof of (i). Assuming first that $\partial[\mathrm{supp}(u(t,.))]$ consists of finitely many portions of hyperplanes, the result follows from the $W^{1,4}$-regularity of $u^{\frac{\alpha+n+1}{4}}$, the nonnegativity of u, and the fact that $W^{1,p}$-functions are absolutely continuous along almost all line segments (cf. Theorem 2.1.4 of [15]). By flattening each $C^{0,1}-$portion of $\partial[\mathrm{supp}(u(t,.)]$, the general case can be established by straightforward calculations.

In order to prove (ii), we observe that if $N = 2$ and $x \in [u(t,.) = 0]$, there exists a positive constant C depending only on Ω and the $W^{1,4}(\Omega)$-norm of $u(t,.)$ such that for all $y \in \Omega$ we have $u(t,y) \le C\,|y-x|^{\frac{2}{\alpha+n+1}}$. If $n < 2$, α can be chosen in such a way that $\frac{2}{\alpha+n+1} > 1$. This completes the proof. □

The strengthened entropy estimates also enable us to improve results concerning positivity properties of solutions. Combining (2.26) and the techniques of ([13, Theorems 1.2 and 1.3]), we arrive at the following theorem.

THEOREM 3.4 (positivity properties). *Let u be a solution of* (1.1) *in the sense of Theorem* 3.2 *and assume the initial value satisfies $\int_\Omega u_0^{3/2-n}\,dx < \infty$.*

(i) *If $n \ge 3/2$ there does not exist a subset $E \subset \Omega$ with $\mathcal{L}^N(E) > 0$ and a time t_0 such that $\int_E u(x,t_0)\,dx = 0$.*

(ii) *If $N = 2$ and $n > 3$ or if $N = 3$ and $n > 6$, the solution u is for almost every $t \in [0,T]$ strictly positive in Ω.*

Let us now discuss the asymptotic behavior of a function u which solves (1.1) in the sense of Theorem 3.2 and in particular satisfies a priori estimate (2.26). We obtain the following theorem.

THEOREM 3.5 (convergence to the mean value). *Let $n > \frac{1}{8}$ be the diffusion growth exponent. Suppose Ω is convex and u solves* (1.1) *in the sense of Theorem* 3.2 *and satisfies the a priori estimate* (2.26) *for an α with $\alpha + n > 1$.*

Then

$$\lim_{t\to\infty} u(t) = \frac{1}{|\Omega|}\int_\Omega u_0(x)\,dx \qquad \text{in } H^1(\Omega).$$

Proof. Note that for convex Ω the constant C_2 in Proposition 2.2 is equal to zero. Thus there is an increasing sequence $(t_k)_{k\in\mathbb{N}}$ tending to infinity with the property

(i) $\int_\Omega \left|D^2 u^{\frac{\alpha+n+1}{2}}\right|^2 (t_k)\,dx \searrow 0,$

(ii) $\int_\Omega \left|\nabla u^{\frac{\alpha+n+1}{4}}\right|^4 (t_k)\,dx \searrow 0,$

which implies that both $\left(\nabla u^{\frac{\alpha+n+1}{2}}(t_k)\right)_{k\in\mathbb{N}}$ and $\left(u^{\frac{\alpha+n+1}{4}}(t_k)\right)_{k\in\mathbb{N}}$ converge to a constant in the corresponding norms. Hence $u^{\frac{\alpha+n+1}{2}}$ also converges to a constant with respect to the L^2-norm and thus

$$\int_\Omega \left|\nabla u^{\frac{\alpha+n+1}{2}}(t_k)\right|^2 dx \searrow 0\ .$$

Therefore,

$$\begin{aligned}\int_\Omega |\nabla u|^2 (t_k)\,dx &\le \int_{[u\ge 1]} u^{\alpha+n-1}\,|\nabla u|^2 (t_k)\,dx + \int_{[u<1]} u^{\frac{\alpha+n-3}{2}}\,|\nabla u|^2 (t_k)\,dx \\ &= o_k(1)\ .\end{aligned}$$

Now using the monotonicity formula (2.5) and the strong convergence of ∇u_δ to ∇u with respect to the $L^2(\Omega_T)$-norm (cf. (2.30)), we infer for almost every $t_1, t_2 \in I$

$$\int_\Omega |\nabla u(t_2)|^2 \, dx - \int_\Omega |\nabla u(t_1)|^2 \, dx \leq -2 \liminf_{\delta \to 0} \int_{t_1}^{t_2} \int_\Omega m(u_\delta) \, |\nabla \Delta u_\delta|^2 \, dxdt;$$

i.e., $\int_\Omega |\nabla u(t,x)|^2 \, dx$ is nonincreasing in t. Thus the result follows just by application of Poincaré's inequality for functions with mean value zero. □

Remark. (1) If $\frac{1}{8} < n < 3$ we get that for all nonnegative initial data $u_0 \in H^1(\Omega)$ there exists an α such that $\alpha + n > 1$ and such that the a priori estimate (2.26) is satisfied. This means that in particular all solutions with compactly supported initial data converge to the mean.

(2) If $n \geq 3$ the same convergence behavior holds true if we impose an additional condition on the initial value; namely, there is a number $\sigma > n - 3$ such that

$$\int_\Omega u_0^{-\sigma} \, dx < \infty \ .$$

This implies that $\int_\Omega G_\alpha(u_0)$ is bounded for an α with $\alpha + n > 1$.

(3) If Ω is not convex, the second term on the right-hand side of (2.26) cannot be neglected any longer, and thus the method of proof above cannot be applied. Nevertheless, by using the boundedness of

$$\int_0^\infty \int_\Omega |\Delta u|^2 \, dxdt$$

the result still can be established if a priori estimate (2.3) holds true, i.e., if $n < 2$ or if $n \geq 2$ and the initial value is strictly positive. Whether it is also true in the case when $n \geq 2$ and initial values have compact support still remains an open question.

Let us now construct steady state solutions with compact support which solve (1.1) in the sense of equations (3.1) and (3.2) for $n > 1$. In Theorem 3.5 we have already proved that for arbitrary, nonnegative initial values $u_0 \in H^1(\Omega)$ with compact support there exist solutions which converge for $t \to \infty$ to the mean value with respect to the H^1-norm. This illustrates in particular that for values of $n \in (1,3)$ we cannot expect results on uniqueness without imposing additional regularity properties at the free boundary in the spirit of estimate (2.26).

LEMMA 3.6 (steady state solutions with compact support). *Let $\Omega' \subset\subset \Omega$ be a subdomain with smooth boundary. Let $\hat{u}$ be the solution of*

$$\begin{aligned} -\Delta \hat{u} &= 1 \ \text{in } \Omega', \\ \hat{u} &= 0 \ \text{on } \partial\Omega'. \end{aligned}$$

Combining the function u defined by

$$u = \begin{cases} \hat{u} \ \text{in } \Omega', \\ 0 \ \text{in } \Omega \setminus \Omega' \end{cases}$$

with $J \equiv 0$, we obtain a weak solution to (1.1) *for values of $n > 1$ in the sense of Definition* 3.1.

Proof. From elliptic regularity theory (cf. Gilbarg and Trudinger [12, Chapter 8]) we infer

$$\hat{u} \in H^2(\Omega') \cap C^\infty(\Omega') \cap W^{1,\infty}(\Omega')$$

and

$$(*) \qquad \int_{\Omega'} \nabla\hat{u}\,\nabla\psi = \int_{\Omega'} \psi \qquad \forall \psi \in W_0^{1,1+\varepsilon}(\Omega').$$

In addition $\hat{u}$ is positive in Ω' (cf. [12, Theorem 8.19]).

Since J is equal to zero on Ω, it will be sufficient to prove that the right-hand side in (3.2) vanishes. We have

$$\begin{aligned} RHS &= \tfrac{1}{2}\int_{\Omega'\times I} m''(\hat{u})\,|\nabla\hat{u}|^2\,\nabla\hat{u}\,\eta + \int_{\Omega'\times I} m'(u)\langle\nabla\hat{u}, D\eta, \nabla\hat{u}\rangle \\ &\quad -\tfrac{1}{2}\int_{\Omega'\times I} m'(\hat{u})\,\operatorname{div}\eta\,\nabla\hat{u}\,\nabla\hat{u} + \int_{\Omega'\times I} \nabla\underbrace{\big(m(\hat{u})\,\operatorname{div}\eta\big)}_{\in W_0^{1,1+\varepsilon}(\Omega')}\cdot\nabla\hat{u} \\ &= I + II + III + IV. \end{aligned}$$

(Since $\hat{u} \in H^2(\Omega') \cap W^{1,\infty}(\Omega')$ and $n > 1$, the boundedness of all the integrals in the equation above is a simple consequence of Hölder's inequality.)

Let ν' be the unit outer normal vector on $\partial\Omega'$. Using the relation $(*)$, we obtain

$$\begin{aligned} IV &= \int_{\Omega'\times I} m(\hat{u})\,\operatorname{div}\eta = -\int_{\Omega'_T} \nabla m(\hat{u})\,\eta + \int_{\partial\Omega'\times I} \underbrace{m(\hat{u})\eta\cdot\nu'\,d\mathcal{H}^N}_{=0} \\ &= -\int_{\Omega'_T} m'(u)\,\nabla u\eta \overset{(*)}{=} -\int_{\Omega'_T} \nabla\big(m'(u)\nabla u\,\eta\big)\cdot\nabla u \\ &= -\int_{\Omega'_T} m''(u)\,|\nabla u|^2\,\eta\,\nabla u - \int_{\Omega'_T} m'(u)\langle\nabla u, D\eta, \nabla u\rangle - \int_{\Omega'_T} m'(u)\langle\eta, D^2u, \nabla u\rangle \\ &= IV_1 + IV_2 + IV_3. \end{aligned}$$

For IV_3 we compute:

$$\begin{aligned} IV_3 &= -\tfrac{1}{2}\int_{\Omega'_T} m'(u)\eta\cdot\nabla\,|\nabla u|^2 \\ &= \tfrac{1}{2}\int_{\Omega'_T} \operatorname{div}\big(m'(u)\,\eta\big)\cdot|\nabla u|^2 - \tfrac{1}{2}\underbrace{\int_{\partial\Omega'\times I} m'(u)\,|\nabla u|^2\,\eta\cdot\nu'\,d\mathcal{H}^N}_{=0 \text{ (since } m'(u)=0 \text{ on } \partial\Omega')} \\ &= \tfrac{1}{2}\int_{\Omega'_T} m''(u)\,|\nabla u|^2\,\eta\,\nabla u + \tfrac{1}{2}\int_{\Omega'_T} m'(u)\,\operatorname{div}\eta\,|\nabla u|^2\,. \end{aligned}$$

Summing up, we obtain $I + II + III = -IV$, which proves the claim. □

Remark. By linearity it is clear that we can adjust $\hat{u}$ in such a way that each positive mean value can be reached. We just take

$$\begin{aligned} \Delta\hat{u} &= \alpha \qquad \text{in } \Omega', \\ \hat{u} &= 0 \qquad \text{on } \partial\Omega', \end{aligned}$$

with $\alpha > 0$ arbitrary.

Appendix.

A. Proof of Lemma 2.1 (*proof of* (2.4)). Let $(u_\varepsilon)_{\varepsilon\searrow 0}$ be a sequence of solutions to auxiliary problems with nondegenerate mobility $m_\varepsilon(\tau) = m(\tau) + \varepsilon$ and initial value u_0 (cf. [13] and [10]). Combining the Aubin–Lions lemma and the uniform boundedness of u_ε (or of $(u_\varepsilon)_t$) in $L^2\big(I; H^2(\Omega)\big)$ (or in $L^2\big(I; (H^1(\Omega))'\big)$, respectively), we notice that a subsequence $(u_\varepsilon)_{\varepsilon\searrow 0}$ strongly converges to u in $L^2\big(I; C^\beta(\Omega)\big)$ for sufficiently small, positive β and that we can extract a subsequence such that $u_\varepsilon(t) \to u(t)$ in $C^\beta(\Omega)$ for almost all $t \in I$. Using the positivity result in Theorem 1.3 of [13], we observe that there is a set $P \subset I$ with $\mathcal{L}^1(I \setminus P) = 0$ such that for all $t \in P$, $u(t)$ is in $C^\beta(\Omega)$ and is strictly positive on Ω.

It will be sufficient to show that the one-dimensional measure of the set

$$E = \left\{ t \in P : \liminf_{\varepsilon\to 0} \int m_\varepsilon\big(u_\varepsilon(t)\big) \left|\nabla \Delta u_\varepsilon\right|^2 (t)\, dx = \infty \right\}$$

vanishes. Defining $K_\varepsilon(t) := \int_\Omega m_\varepsilon\big(u_\varepsilon(t)\big) \left|\nabla\Delta u_\varepsilon(t)\right|^2 dx$ and

$$[A]_n = \begin{cases} A & \text{if } A < n, \\ n & \text{otherwise} \end{cases}$$

and using the positivity of $u(t,\cdot)$ for almost every t, we have, by Lebesgue's theorem,

$$C \geq \int_E \int_\Omega m_\varepsilon\big(u_\varepsilon(t,x)\big) \left|\nabla\Delta u_\varepsilon(t,x)\right|^2 dxdt \geq \int_E [K_\varepsilon(t)]_n \, dt \overset{\varepsilon\to 0}{\longrightarrow} n \cdot |E| .$$

As n can be chosen arbitrarily, this implies $\mathcal{L}^1(E) = 0$. For fixed $t \in P \setminus E$ we can select a subsequence of $(\nabla\Delta u_\varepsilon(t,.))_{\varepsilon\to 0}$ which weakly converges in $L^2(\Omega)$. Then (2.2) implies that $J(t) = m(u(t))\nabla\Delta u(t) \quad$ in $L^2(\Omega)$ for all $t \in P \setminus E$.

Proof of (2.5). Choosing $\psi(t,x) = \chi_{[t_1,t_2]}\Delta u_\varepsilon$ as the test function in the weak formulation

$$\int_0^T \langle (u_\varepsilon)_t, \psi\rangle dt - \int_{\Omega_T} m_\varepsilon(u_\varepsilon)\nabla\Delta u_\varepsilon \nabla\psi dxdt = 0,$$

$$\psi \in L^2(I; H^1(\Omega)) \text{ arbitrary}$$

for auxiliary problems with nondegenerate mobility as described above, we immediately obtain for almost all $t_1, t_2 \in I$

$$\frac{1}{2}\int_\Omega |\nabla u_\varepsilon(t_2)|^2 dx + \int_{t_1}^{t_2}\int_\Omega m_\varepsilon(u_\varepsilon) \left|\nabla\Delta u_\varepsilon\right|^2 dxdt \leq \frac{1}{2}\int_\Omega |\nabla u_\varepsilon(t_1)|^2 dx.$$

Now observing that ∇u_ε converges to ∇u strongly in $L^{\frac{2N+4}{N}-}(\Omega_T)$ (cf., e.g., [13, Lemma 2.8]) and that $m_\varepsilon^{\frac{1}{2}}(u_\varepsilon)\nabla\Delta u_\varepsilon$ in $L^2(\Omega_T)$ weakly converges to a function β which for almost every $t \in I$ can be identified with $m(u)\nabla\Delta u$, the result follows by the lower semicontinuity of the norm under weak convergence.

B. Proof of Lemma 2.3 and (3.3). The following result will be essential.

LEMMA B.1. *Let $\Omega \subset \mathbb{R}^N$ be a domain with piecewise smooth boundary of class $C^{0,1}$. For every vector field $\eta \in H^2(\Omega;\mathbb{R}^N)$ which is tangential on $\partial\Omega$ we have*

$$(D\eta \cdot \nu)_{\|} = -d\nu \cdot \eta \tag{B.1}$$

a.e. on $\partial\Omega$. Here, $(\cdot)_{\|}$ denotes the tangential component of a vector field.

Proof. Since $\eta \cdot \nu = 0$, we have for arbitrary smooth curves $c:\,[0,1]\longrightarrow\partial\Omega$

$$0 = \frac{d}{dt}\left\langle \eta_{|c(t)}, \nu_{|c(t)}\right\rangle = \langle D\eta_{|c(t)}\dot c(t), \nu_{|c(t)}\rangle + \left\langle \eta_{|c(t)}, d\nu_{|c(t)}\cdot \dot c(t)\right\rangle$$

for almost every $x = c(t) \in \partial\Omega$. As $d\nu : T_{c(t)}S\longrightarrow T_{c(t)}S$ is a self-adjoint, linear mapping (cf. [11, Chapter V]), we obtain

$$\left\langle \eta_{|c(t)}, d\nu_{|c(t)}\cdot \dot c(t)\right\rangle = \left\langle d\nu_{|c(t)}\cdot \eta_{|c(t)}, \dot c(t)\right\rangle,$$

which proves the lemma. □

Proof of Lemma 2.3. We assume $\nabla\Delta u \in L^2(\Omega)$. The result for functions with weaker regularity can be proved by an approximation argument. Successive integration by parts gives

$$\begin{aligned}\int_\Omega f'(u)\,|\nabla u|^2\,\Delta u &= -\int_\Omega f(u)\,|\Delta u|^2 - \int_\Omega f(u)\,\nabla\Delta u\,\nabla u\\ &= -\int_\Omega f(u)\,|\Delta u|^2 + \int_\Omega f'(u)\langle\nabla u, D^2u, \nabla u\rangle + \int_\Omega f(u)\,\left|D^2u\right|^2\\ &\qquad - \int_{\partial\Omega} f(u)\langle\nabla u, D^2u, \nu\rangle\, d\mathcal{H}^{N-1}.\end{aligned}$$

Using now the identity $\langle\nabla u, D^2u, \nabla u\rangle = \frac12\langle\nabla\,|\nabla u|^2, \nabla u\rangle$, we obtain ($\frac{\partial}{\partial\nu}u = 0$ on $\partial\Omega$!):

$$\int_\Omega f'(u)\langle\nabla u, D^2u, \nabla u\rangle = -\frac12\int_\Omega f''(u)\,|\nabla u|^4 - \frac12\int_\Omega f'(u)\,|\nabla u|^2\,\Delta u.$$

Putting everything together and using formula (B.1), the result can be established easily. □

Proof of (3.3). Integration by parts shows

$$\begin{aligned}\int_\Omega m(u)\,\nabla\Delta u\,\eta\,dx &= \tfrac12\int_\Omega m''(u)\,|\nabla u|^2\,\nabla u\eta\,dx + \tfrac12\int_\Omega m'(u)\,|\nabla u|^2\,\mathrm{div}\,\eta\,dx\\ &\quad + \int_\Omega m'(u)\langle\nabla u, D\eta, \nabla u\rangle\,dx + \int_\Omega m(u)\,\nabla u\nabla\,\mathrm{div}\,\eta\,dx\\ &\quad - \int_{\partial\Omega} m(u)\big\{\langle\nabla u, D\eta, \nu\rangle - \langle\eta, D^2u, \nu\rangle\big\}\,d\mathcal{H}^{N-1}.\end{aligned}$$

Since both η and ∇u are tangential vector fields on $\partial\Omega$, we can apply (B.1) twice and obtain the result. □

Acknowledgments. It is a pleasure to thank Michiel Bertsch for deep and fruitful discussions.

REFERENCES

[1] H. W. Alt and S. Luckhaus, *Quasilinear elliptic parabolic differential equations*, Math. Z., 183 (1983), pp. 311–338.

[2] E. Beretta, M. Bertsch, and R. Dal Passo, *Nonnegative solutions of a fourth order nonlinear degenerate parabolic equation*, Arch. Rational Mech. Anal., 129 (1995), pp. 175–200.

[3] F. Bernis, *Viscous flows, fourth order nonlinear degenerate parabolic equations and singular elliptic problems*, in Free Boundary Problems: Theory and Applications, J. I. Diaz, M. A. Herrero, A. Linan, and J. L.Vazquez, eds., Pitman Research Notes in Mathematics 323, Longman, Harlow, 1995, pp. 40–56.

[4] F. Bernis, *Finite speed of propagation and continuity of the interface for thin viscous flows*, Adv. Differential Equations, 1 (1996), pp. 337–368.

[5] F. Bernis, *Finite speed of propagation for thin viscous flows when* $2 \le n < 3$, C. R. Acad. Sci. Paris Sér.I Math., 322 (1996), pp. 1169–1174.

[6] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.

[7] F. Bernis, L. A. Peletier, and S. M. Williams, *Source type solutions of a fourth order nonlinear degenerate parabolic equation*, Nonlinear Anal., 18 (1992), pp. 217–234.

[8] A. L. Bertozzi and M. Pugh, *The lubrication approximation for thin viscous films: Regularity and long time behaviour of weak solutions*, Comm. Pure Appl. Math., 49 (1996), pp. 85–123.

[9] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, Berlin, 1993.

[10] C. M. Elliott and H. Garcke, *On the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.

[11] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian Geometry*, Springer-Verlag, Berlin, 1987.

[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[13] G. Grün, *Degenerate parabolic differential equations of fourth order and a plasticity model with nonlocal hardening*, Z. Anal. Anwendungen, 14 (1995), pp. 541–574.

[14] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Annali di Mathematica Pura ed Applicata, 146 (1987), pp. 65–96.

[15] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 343–379, March 1998

004

BOUNDARY LAYERS IN THE HOMOGENIZATION OF A SPECTRAL PROBLEM IN FLUID–SOLID STRUCTURES*

GRÉGOIRE ALLAIRE† AND CARLOS CONCA‡

Abstract. This paper is devoted to the asymptotic analysis of the spectrum of a mathematical model that describes the vibrations of a coupled fluid–solid periodic structure. In a previous work [*Arch. Rational Mech. Anal.*, 135 (1996), pp. 197–257] we proved by means of a Bloch wave homogenization method that, in the limit as the period goes to zero, the spectrum is made of three parts: the macroscopic or homogenized spectrum, the microscopic or Bloch spectrum, and a third component, the so-called *boundary layer spectrum*. While the two first parts were completely described as the spectrum of some limit problem, the latter was merely defined as the set of limit eigenvalues corresponding to sequences of eigenvectors concentrating on the boundary. It is the purpose of this paper to characterize explicitly this boundary layer spectrum with the help of a family of limit problems revealing the intimate connection between the periodic microstructure and the boundary of the domain. We therefore obtain a "completeness" result, i.e., a precise description of all possible asymptotic behaviors of sequences of eigenvalues, at least for a special class of polygonal domains.

Key words. homogenization, Bloch waves, spectral analysis, boundary layers, fluid–solid structures

AMS subject classification. 35B40

PII. S0036141096304328

1. Introduction.

1.1. Setting of the problem. This paper is devoted to the study of some boundary layer phenomena which arise in the asymptotic analysis of the spectrum of a mathematical model describing the vibrations of a coupled periodic system of solid tubes immersed in a perfect incompressible fluid. This simple model is due to Planchard, who studied it intensively (see [31], [32]). Since we introduced it at length in section 1.2 of our previous work [3] we content ourselves with briefly recalling the statement of this problem.

We consider a periodic bounded domain Ω_ϵ obtained from a fixed bounded open set Ω in $\mathbb{R}^N$ by removing a collection of identical, periodically distributed holes $(T_p^\epsilon)_{1\le p\le n(\epsilon)}$. The distance between adjacent holes as well as their size are both of the order of ϵ, the size of the period which is a small parameter going to zero. Correspondingly, the number of holes $n(\epsilon)$ is of the order of ϵ^{-N}, where N is the spatial dimension. More precisely, let us first define the standard unit cell $Y = (0;1)^N$ which, upon rescaling to size ϵ, becomes the period in Ω. Let T be a smooth, simply connected, closed subset of Y, assumed to be *strictly included* in Y (i.e., T does not touch the boundary of Y). The set T represents the reference tube (or rod) and the unit fluid cell is defined as

$$Y^* = Y \setminus T.$$

*Received by the editors May 28, 1996; accepted for publication (in revised form) December 10, 1996. The second author is partially supported by the Chilean programme of presidential chairs in science and by Fondecyt under grant 197-0734.

http://www.siam.org/journals/sima/29-2/30432.html

†Commissariat à l'Energie Atomique, DRN/DMT/SERMA, C.E. Saclay, 91191 Gif sur Yvette, France and Laboratoire d'Analyse Numérique, Université Paris 6 (allaire@ann.jussieu.fr).

‡Departamento de Ingeniería Matemática, Universidad de Chile, Casilla 170/3, Correo 3, Santiago, Chile (cconca@dim.uchile.cl).

For each value of the small positive parameter ϵ, the fluid domain Ω_ϵ is obtained from the reference domain Ω by removing a periodic arrangement of tubes ϵT with period ϵY. Denoting by (T_p^ϵ) the family of all translates of ϵT by vectors ϵp (where p is a multi-index in $\mathbb{Z}^N$) and by (Y_p^ϵ) the corresponding family of cells, we define

$$\Omega_\epsilon = \Omega \setminus \bigcup_{p=1}^{n(\epsilon)} T_p^\epsilon. \tag{1}$$

Although p is a multi-index in $\mathbb{Z}^N$, for simplicity we denote its range by $1 \le p \le n(\epsilon)$. To obtain the fluid domain Ω_ϵ in (1), we remove from the original domain Ω only those tubes T_p^ϵ which belong to a cell Y_p^ϵ completely included in Ω. This has the effect that no tube meets the boundary $\partial\Omega$. Analogously, (Γ_p^ϵ) denotes the family of tubes boundaries (∂T_p^ϵ).

We are interested in the following spectral problem in Ω: find the eigenvalues λ_ϵ and the corresponding normalized eigenvectors u_ϵ, solutions of

$$\begin{cases} -\Delta u_\epsilon = 0 & \text{in } \Omega_\epsilon, \\ \lambda_\epsilon \frac{\partial u_\epsilon}{\partial n} = \epsilon^{-N} \vec{n} \cdot \displaystyle\int_{\Gamma_p^\epsilon} u_\epsilon \vec{n} ds & \text{on } \Gamma_p^\epsilon \text{ for } 1 \le p \le n(\epsilon), \\ u_\epsilon = 0 & \text{on } \partial\Omega, \end{cases} \tag{2}$$

where $\vec{n}$ denotes the exterior unit normal to Ω_ϵ.

The homogenization of this model has already attracted the attention of several authors (see [1], [14], [16], [17]). Even though it is a spectral problem involving the Laplace operator, it is easily seen to admit only finitely many eigenvalues, exactly $Nn(\epsilon)$ (the number of tubes times the number of degrees of freedom in their displacements). To this end, a finite-dimensional operator S_ϵ is introduced, which acts on the family of tube displacements $\vec{s} = (\vec{s}_p)_{1 \le p \le n(\epsilon)}$ with $\vec{s}_p \in \mathbb{R}^N$,

$$\begin{aligned} S_\epsilon \,:\, \mathbb{R}^{Nn(\epsilon)} &\longrightarrow \mathbb{R}^{Nn(\epsilon)}, \\ (\vec{s}_p)_{1 \le p \le n(\epsilon)} &\mapsto \left(\frac{1}{\epsilon^N} \int_{\Gamma_p^\epsilon} u_\epsilon \vec{n} ds \right)_{1 \le p \le n(\epsilon)}, \end{aligned} \tag{3}$$

where the fluid potential u_ϵ is now the unique solution in $H^1(\Omega_\epsilon)$ of

$$\begin{cases} -\Delta u_\epsilon = 0 & \text{in } \Omega_\epsilon, \\ \frac{\partial u_\epsilon}{\partial n} = \vec{s}_p \cdot \vec{n} & \text{on } \Gamma_p^\epsilon \text{ for } 1 \le p \le n(\epsilon), \\ u_\epsilon = 0 & \text{on } \partial\Omega. \end{cases} \tag{4}$$

According to [17], S_ϵ is self-adjoint, positive definite, and its spectrum, denoted by $\sigma(S_\epsilon)$, coincides with the set of eigenvalues of (2). Of course, since S_ϵ acts in a finite-dimensional space, $\sigma(S_\epsilon)$ is made up of $Nn(\epsilon)$ real numbers. It has been further proved that all eigenvalues of S_ϵ are uniformly bounded away from zero and from infinity (see, e.g., Proposition 1.2.1 and Lemma 1.2.2 in [3]). As the period ϵ goes to zero, $\sigma(S_\epsilon)$, considered as a subset of $\mathbb{R}^+$, converges to a limit set σ_∞ which, by definition, is the set of all cluster points of (sub)sequences of eigenvalues of S_ϵ

$$\sigma_\infty = \{\lambda \in \mathbb{R}^+ \mid \exists \text{ a subsequence } \lambda_{\epsilon'} \in \sigma(S_{\epsilon'}) \text{ such that } \lambda_{\epsilon'} \to \lambda\}.$$

Finding an adequate characterization of the limit set σ_∞ was the main goal of our previous paper [3]. A positive answer to this problem is given in the present article for a special class of polygonal domains.

1.2. Survey of the previous results. The characterization of σ_∞ amounts to studying the asymptotic behavior of the spectral problem (2), or, in other words, to homogenize (2) as the parameter ϵ goes to zero. To our knowledge, this can be done, at least, using two different approaches: the classical *homogenization process* for periodic structures (see, e.g., the reference books [7], [8], [24], [28], [35]) or the so-called *Bloch wave method* (also called the *nonstandard homogenization procedure* in [16]; see [8], [33], [34], [36] for an introduction to Bloch waves in spectral analysis). The former naturally yields the *homogenized or macroscopic spectrum* of (2), while the latter is associated with the so-called *Bloch or microscopic spectrum*.

Historically the second approach was the first applied to problem (2) by C. Conca, M. Vanninathan, and their coworkers [1], [15], [16], [17]. The key point in this method is to rescale the ϵ-network of tubes to size 1 and, therefore, as ϵ goes to zero, to obtain an infinite limit domain containing a periodic array of unit tubes. Then, the limit problem is amenable to the celebrated Bloch wave decomposition (also known as the Floquet decomposition; see the original work of F. Bloch [11] or the first mathematical papers [19], [30], [36] or the books [8], [33]). The spectrum of this limit problem is called the *Bloch spectrum*.

Although it seems the easiest to apply, the first approach (i.e., the classical homogenization) has only been recently applied to problem (2) in our previous article [3]. By homogenizing the operator S_ϵ with the help of the *two-scale convergence* (see [2], [29]), a homogenized equation is obtained in the domain Ω. Its spectrum is called the *homogenized spectrum*. It turns out that the homogenized spectrum is completely different from the Bloch spectrum, and therefore both approaches are complementary. This is possible since in neither case the underlying sequences of linear operators converge uniformly to their limit which are noncompact operators. In addition to this homogenization result, our paper [3] provides a unified theory for both approaches that we called the *Bloch wave homogenization* method. We refer to [3] for more details (see also [4], [5]), and we simply recall our main results.

The homogenization of model (2) amounts to analyzing the convergence of the sequence of operators S_ϵ. Since these operators are defined on a space which varies with ϵ, we extend them to the fixed space $[L^2(\Omega)^N]^{K^N}$, where K is an arbitrary positive integer. Denoting by S_ϵ^K this extension, it will be amenable to a standard asymptotic analysis, while keeping essentially the same spectrum as S_ϵ. Following the lead of Planchard [32], the reference cell of our homogenization procedure is KY instead of simply Y (this technique is referred to as homogenization by packets in [32]). To give a precise definition of S_ϵ^K we introduce two linear maps: a projection P_ϵ^K from $[L^2(\Omega)^N]^{K^N}$ into $\mathbb{R}^{Nn(\epsilon)}$ and an extension E_ϵ^K from $\mathbb{R}^{Nn(\epsilon)}$ into $[L^2(\Omega)^N]^{K^N}$ such that $S_\epsilon^K = E_\epsilon^K S_\epsilon P_\epsilon^K$. To do so, some notation is required concerning the two indices p (indexing constant vectors in $\mathbb{R}^{Nn(\epsilon)}$) and j (indexing vector functions in $[L^2(\Omega)^N]^{K^N}$).

DEFINITION 1.1. *Let KY be the reference cell $(0,K)^N$ which is made of K^N subcells Y_j of the type $(0,1)^N$ containing a single tube T_j. The multi-integer $j = (j_1,\dots,j_N)$ which enumerates all the tubes in KY takes its values in $\{0,1,\dots,K-1\}^N$ (we use the notation $0 \le j \le K-1$). Let $p = (p_1,\dots,p_N)$ be the multi-integer which enumerates all the tubes in Ω_ϵ (see (1)). We define a third multi-integer $\ell = (\ell_1,\dots,\ell_N)$ which enumerates all the periodic reference cells $\epsilon(KY)$ in Ω_ϵ (its range is denoted by $1 \le \ell \le n_K(\epsilon)$). These three indices are assumed to be related by the*

following one-to-one map:

$$\ell_m = E\left(\frac{p_m}{K}\right), \quad j_m = p_m - K\ell_m \quad \forall m = 1, ..., N, \tag{5}$$

where $E(\cdot)$ denotes the integer-part function.

Then, P_ϵ^K and E_ϵ^K are defined by

$$\begin{aligned} P_\epsilon^K : [L^2(\Omega)^N]^{K^N} &\longrightarrow \mathbb{R}^{Nn(\epsilon)}, \\ (\vec{s}_j(x))_{0\le j\le K-1} &\longrightarrow \left(\vec{s}_p = \tfrac{1}{|\epsilon(KY)_\ell|}\int_{\epsilon(KY)_\ell} \vec{s}_j(x)dx\right)_{1\le p\le n(\epsilon),} \end{aligned} \tag{6}$$

$$\begin{aligned} E_\epsilon^K : \mathbb{R}^{Nn(\epsilon)} &\longrightarrow [L^2(\Omega)^N]^{K^N}, \\ (\vec{s}_p)_{1\le p\le n(\epsilon)} &\longrightarrow \left(\vec{s}_j(x) = \sum_\ell \chi_{\epsilon(KY)_\ell}(x)\vec{s}_p\right)_{0\le j\le K-1,} \end{aligned} \tag{7}$$

where p is related to (ℓ, j) by formula (5). One can easily check that the adjoint $(P_\epsilon^K)^*$ of P_ϵ^K is nothing but $(\epsilon K)^{-N}E_\epsilon^K$ and that $P_\epsilon^K E_\epsilon^K$ is equal to the identity in $\mathbb{R}^{Nn(\epsilon)}$. Therefore, S_ϵ^K is also self-adjoint compact and its spectrum is exactly that of S_ϵ, plus the new eigenvalue 0 which has infinite multiplicity.

The homogenization of the extended operator S_ϵ^K is now amenable to the two-scale convergence method [2], [29]. However, the limit operator S^K has a complicated form which can be simplified by using the following discrete Bloch wave decomposition (see [1]).

LEMMA 1.2. *For any family $(\vec{s}_j)_{0\le j\le K-1}$ of vectors in $\mathbb{C}^N$, let $\vec{s}(y)$ be the following KY-periodic function, piecewise constant in each subcell Y_j:*

$$\vec{s}(y) = \sum_{j=0}^{K-1} \vec{s}_j \chi_{Y_j}(y) \quad \forall y \in KY.$$

There exists a unique family of constant vectors $(\vec{t}_j)_{0\le j\le K-1}$ in $\mathbb{C}^N$ such that

$$\vec{s}(y) = \sum_{j=0}^{K-1} \vec{t}_j e^{2\pi\imath \frac{j}{K}\cdot E(y)} \quad \forall y \in KY, \tag{8}$$

where $E(\cdot)$ denotes the integer-part function. Moreover, the Bloch wave decomposition operator $\mathcal{B}$, defined by $\mathcal{B}(\vec{s}_j) = K^{N/2}(\vec{t}_j)$, is an isometry on $(\mathbb{C}^N)^{K^N}$.

The first main result in [3] (see Theorem 3.2.1) is the following theorem.

THEOREM 1.3. *The sequence $S_\epsilon^K = E_\epsilon^K S_\epsilon P_\epsilon^K$ converges strongly to a limit S^K; i.e., for any family $(\vec{s}_j(x))_{0\le j\le K-1}$, $S_\epsilon^K(\vec{s}_j)$ converges strongly to $S^K(\vec{s}_j)$ in $[L^2(\Omega)^N]^{K^N}$. Furthermore, the limit operator S^K is given by*

$$S^K = \mathcal{B}^* T^K \mathcal{B}, \text{ with } T^K = \operatorname{diag}\left[(T_j^K)_{0\le j\le K-1}\right], \tag{9}$$

where the entries T_j^K are self-adjoint continuous but noncompact operators in $L^2(\Omega)^N$, defined by

$$T_j^K \vec{t}_j = \begin{cases} (A(0) - I)\nabla u - (A(0) - |Y^*|I)\vec{t}_0 & \text{if } j = 0, \\ A(\frac{j}{K})\vec{t}_j & \text{if } j \neq 0, \end{cases} \tag{10}$$

where I is the identity matrix and u is the unique solution of the homogenized problem

$$\begin{cases} -\mathrm{div}(A(0)\nabla u) = \mathrm{div}((I - A(0))\vec{t}_0) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{11}$$

and, for $\theta \in [0,1]^N$, $A(\theta)$ is the Bloch homogenized matrix with components $(A_{mm'}(\theta))_{1\le m,m'\le N}$ defined by

$$\bar{A}_{mm'}(\theta) = \int_{Y^*} \nabla w_m^\theta(y) \cdot \nabla \bar{w}_{m'}^\theta(y) dy, \tag{12}$$

where $(w_m^\theta)_{1\le m\le N}$ are solutions of the so-called cell problem at the Bloch frequency θ:

$$\begin{cases} -\Delta w_m^\theta = 0 & \text{in } Y^*, \\ (\nabla w_m^\theta - \vec{e}_m)\cdot \vec{n} = 0 & \text{on } \partial T, \\ y \to e^{-2\pi\imath\theta\cdot y} w_m^\theta(y) \ Y^*\text{-periodic.} \end{cases} \tag{13}$$

The first component T_0^K of the limit operator T^K is the same for all K and is denoted by S in what follows. It is called the macroscopic or homogenized limit of S_ϵ ((11) is also called the homogenized equation). The spectrum $\sigma(S)$ is essential and has been explicitly characterized in Theorems 2.1.4 and 2.1.5 of [3]. The other components of T^K are simple linear multiplication operators that represent the microscopic or Bloch limit behavior of the sequence S_ϵ^K.

According to Proposition 3.2.6 in [3], the matrix $A(\theta)$ is Hermitian and positive definite for any value of θ. Furthermore, it is a continuous function of θ, except at the origin $\theta = 0$. Nevertheless, it is continuous at the origin along rays of constant direction (see Proposition 3.4.4 in [3]). Denoting by $0 < \lambda_1(\theta) \le \lambda_2(\theta) \le \cdots \le \lambda_N(\theta)$ its eigenvalues, we can define the so-called *Bloch spectrum* by

$$\sigma_{Bloch} = \bigcup_{m=1}^{N} \overline{\lambda_m(]0,1[^N)},$$

where $\overline{\lambda_m(]0,1[^N)}$ denotes the closure of the image of $]0,1[^N$ under the maps $\lambda_m(\cdot)$. We deduce our second main result.

THEOREM 1.4. *The strong convergence of S_ϵ^K to the limit operator S^K implies the lower semicontinuity of the spectrum*

$$\sigma(S^K) \subset \lim_{\epsilon\to 0} \sigma(S_\epsilon^K).$$

By letting K go to infinity, we obtain

$$\sigma(S) \cup \sigma_{Bloch} \subset \lim_{\epsilon\to 0} \sigma(S_\epsilon). \tag{14}$$

REMARK 1.5. *As a matter of fact, the Bloch spectrum σ_{Bloch} and the homogenized spectrum $\sigma(S)$ do not coincide. Therefore, both type of limit problems (macroscopic* (11) *and microscopic* (13)*) are complementary. As already mentioned, the Bloch spectrum has already been characterized by C. Conca and M. Vanninathan in* [17] *by means of a different method, the so-called nonstandard homogenization procedure (see also the book* [16]*).*

The question is now to see whether the inclusion in (14) is actually an equality, i.e., if our asymptotic analysis is complete. It turns out that the homogenized and the Bloch spectra are usually not enough to describe σ_∞ because the interaction between the boundary $\partial\Omega$ and the microstructure is not taken into account in our analysis. More precisely, there may well exist sequences of eigenvectors of (2) which concentrate near the boundary $\partial\Omega$ of Ω. They behave as *boundary layers* in the sense that they converge strongly to zero locally inside the domain. Clearly the oscillations of these eigenvectors cannot be captured by the usual homogenization method; neither are they filtered in the Bloch spectrum which is insensitive to the boundary.

Nevertheless, the third main result of our previous paper [3] shows that for any other type of sequences of eigenvectors (not concentrating on the boundary), the limits of the corresponding sequences of eigenvalues belong to $\sigma(S) \cup \sigma_{Bloch}$. More exactly, introducing the subset of σ_∞

$$\begin{aligned} \sigma_{boundary} = \quad & \{\lambda \in \mathbb{R} \mid \exists(\lambda_{\epsilon'}, \vec{s}^{\epsilon'}) \text{ such that } S^1_{\epsilon'}\vec{s}^{\epsilon'} = \lambda_{\epsilon'}\vec{s}^{\epsilon'},\ \lambda_{\epsilon'} \to \lambda, \\ & \|\vec{s}^{\epsilon'}\|_{L^2(\Omega)^N} = 1, \text{ and } \forall\omega \text{ with } \overline{\omega} \subset \Omega,\ \|\vec{s}^{\epsilon'}\|_{L^2(\omega)^N} \to 0\}, \end{aligned} \tag{15}$$

where ϵ' is a subsequence of ϵ and S^1_ϵ is the extension to $L^2(\Omega)^N$ of S_ϵ, we proved the following theorem (see Theorem 3.2.9 in [3]).

THEOREM 1.6. *The limit set of the spectrum of the operator S_ϵ is precisely made of three parts; the homogenized, the Bloch, and the boundary layer spectrum*

$$\lim_{\epsilon\to 0} \sigma(S_\epsilon) = \sigma_\infty = \sigma(S) \cup \sigma_{Bloch} \cup \sigma_{boundary}.$$

The proof of this *completeness result* is the focus of section 3.4 in [3]. It involves a new type of default measure for weakly converging sequences of eigenvectors of S_ϵ, the so-called *Bloch measures* which quantify its amplitude and direction of oscillations.

Of course the definition of $\sigma_{boundary}$ is not satisfactory, since it does not characterize that part of the limit set σ_∞ as the spectrum of some limit operator associated with the boundary $\partial\Omega$. In particular, it is not clear whether $\sigma_{boundary}$ is empty or included in $\sigma(S) \cup \sigma_{Bloch}$. It is the purpose of the present paper to characterize explicitly $\sigma_{boundary}$, at least for special rectangular domains Ω and associated sequences of parameters ϵ.

REMARK 1.7. *By their very definitions, the limit spectrum σ_∞ and the boundary layer spectrum $\sigma_{boundary}$ depend a priori on the choice of the sequence of small parameters ϵ. On the contrary, the homogenized spectrum $\sigma(S)$ and the Bloch spectrum σ_{Bloch} are independent of the sequence ϵ. We believe that $\sigma_{boundary}$ is actually strongly dependent on the sequence ϵ. In particular, we shall characterize it only for a specific sequence ϵ. We thank C. Castro and E. Zuazua for clarifying discussions on this topic* [12].

1.3. Presentation of the main new results. There are mainly two new results in this paper which correspond to the next two sections. First, in section 2 we introduce a new class of limit problems involving the interaction between the tubes array and the domain boundary. We assume that the domain Ω is cylindrical;

$$\Omega = \Sigma\times]0; L[, \tag{16}$$

where Σ is an open bounded set in $\mathbb{R}^{N-1}$ and $L > 0$ is a positive length. A generic point x in $\mathbb{R}^N$ is denoted by $x = (x', x_N)$ with $x' \in \mathbb{R}^{N-1}$ and $x_N \in \mathbb{R}$ (x_N is the coordinate along the axis of Ω). Let us define a semi-infinite band

$$G = Y'\times]0; +\infty[,$$

where $Y' =]0,1[^{N-1}$ is the unit cell in $\mathbb{R}^{N-1}$. This new "boundary layer" limit problem takes place in the fluid part of G, denoted by G^* and defined by

$$G^* \;=\; G \setminus \bigcup_{q\geq 1} T_q,$$

where (T_q) is the infinite collection of tubes periodically disposed in G. With each tube T_q is associated a displacement $\vec{s}_q \in \mathbb{R}^N$. We denote by ℓ^2 the space of families $(\vec{s}_q)_{q\geq 1}$ such that $\sum_{q\geq 1} |\vec{s}_q|^2$ is finite. Introducing a Bloch parameter $\theta' \in [0,1]^{N-1}$, we define a "boundary layer" operator $d_{\theta'}$ by

$$\begin{aligned} & d_{\theta'} \,:\, \ell^2 \longrightarrow \ell^2, \\ & (\vec{s}_q)_{q\geq 1} \;\mapsto\; \left(\int_{\Gamma_q} u_{\theta'} \vec{n} ds \right)_{q\geq 1}, \end{aligned} \tag{17}$$

where $u_{\theta'}(y)$ is the unique solution of

$$\begin{cases} -\Delta u_{\theta'} = 0 & \text{in } G^*, \\ \frac{\partial u_{\theta'}}{\partial n} = \vec{s}_q \cdot \vec{n} & \text{on } \Gamma_q,\ q \geq 1, \\ u_{\theta'} = 0 & \text{if } y_N = 0, \\ y' \mapsto e^{-2\pi\imath\theta'\cdot y'} u_{\theta'}(y', y_N) & Y'\text{-periodic.} \end{cases}$$

Our first result (see Theorem 2.18) is concerned with the continuity of the spectrum of $d_{\theta'}$, considered as a subset of $\mathbb{R}$, with respect to the Bloch parameter θ'.

THEOREM 1.8. *For all $\theta' \in [0,1]^{N-1}$, $d_{\theta'}$ is a self-adjoint continuous but non-compact operator in ℓ^2. Its spectrum $\sigma(d_{\theta'})$ depends continuously on θ', except at $\theta' = 0$. Defining the* boundary layer spectrum *associated with the surface Σ*

$$\sigma_\Sigma \stackrel{\text{def}}{=} \bigcup_{\theta' \in]0,1[^{N-1}} \sigma(d_{\theta'}) \cup \sigma(d_0),$$

we have

$$\sigma_\Sigma \subset \lim_{\epsilon \to 0} \sigma(S_\epsilon).$$

In general, $\sigma(d_{\theta'})$ is not included in the previously found limit spectrum $\sigma(S) \cup \sigma_{Bloch}$ (see Proposition 2.17). Therefore, the new class of limit problems defined by (17) is not redundant with the homogenized or the Bloch limit problems. Our main tool for proving this theorem is a variant of the two-scale convergence adapted to boundary layers, using test functions which oscillate periodically in the directions parallel to the boundary Σ and decay asymptotically fast in the normal direction to Σ (see section 2.1). Remark that the above result holds for any cylindrical domain of the type (16) and for any sequence of periods ϵ going to zero.

Section 3 is devoted to our second main result which requires additional assumptions on the geometry of the domain and on the sequence of periods ϵ. More precisely, we now assume that Ω is a rectangle with integer dimensions

$$\Omega \;=\; \prod_{i=1}^{N}]0; L_i[\quad \text{and} \quad L_i \in \mathbb{N}^* \tag{18}$$

and that the sequence ϵ is exactly

$$\epsilon_n \;=\; \frac{1}{n},\; n \in \mathbb{N}^*.$$

These assumptions imply that, for any ϵ_n, the domain Ω is the union of a finite number of *entire* cells of size ϵ_n. Then, the above analysis of the boundary layer spectrum σ_Σ can be achieved for any face Σ of the rectangle Ω. Of course a completely similar analysis can be done for all the lower dimensional manifolds (edges, corners, etc.) of which the boundary of Ω is made up. For each type of manifold, a different family of limit problems arise which are straightforward generalizations of (17). For example, in two space dimensions, the corners of Ω give rise to a limit problem in the quarter of space $\mathbb{R}^+ \times \mathbb{R}^+$ filled with a periodic array of tubes (see section 3.3). Finally, we prove a completeness result (see Theorem 3.1).

THEOREM 1.9. *The limit set of the spectrum of the operator S_{ϵ_n} is precisely made of three parts; the homogenized, the Bloch, and the union of all boundary layer spectra, as defined in Theorem* 1.8*,*

$$\lim_{\epsilon_n \to 0} \sigma(S_{\epsilon_n}) = \sigma(S) \cup \sigma_{Bloch} \cup \sigma_{\partial\Omega},$$

with the notation

$$\sigma_{\partial\Omega} \;=\; \bigcup_{\Sigma \subset \partial\Omega} \sigma_\Sigma,$$

where the union is over all hypersurfaces and lower dimensional manifolds composing the boundary $\partial\Omega$.

REMARK 1.10. *The difference between the above completeness theorem and Theorem* 1.6 *is that, here, the boundary layer spectrum $\sigma_{\partial\Omega}$ is explicitly defined for the specific sequence of parameters ϵ_n as the spectrum of a family of limit operators, while, in our previous result, the boundary layer spectrum $\sigma_{boundary}$ was indirectly defined for any sequence ϵ but not explicitly characterized.*

We conclude this introduction by giving a few references to related works on boundary layers in homogenization and by a short discussion on numerical studies concerning problem (2). Apart from the classical books [7, Chapter 7] and [26], we refer mainly to the papers [6], [9], [10], and [27]. Planchard's model has already been studied numerically. The Bloch eigenvalues $\lambda_i(\theta)$ were computed by F. Aguirre in a two-dimensional example. A brief account of his work is given in [1]. On the other hand, direct numerical computations of the entire spectrum $\sigma(S_\epsilon)$ (for a fixed value of ϵ, and without using homogenization) have been reported in [23]. To our knowledge, these are the only available numerical results concerning a large tube array (see also [21], [22]). Of course, these results are consistent with Theorem 1.9 describing the asymptotic behavior of $\sigma(S_\epsilon)$. In particular, some vibration modes displayed in [23] are numerical evidence that $\sigma_{\partial\Omega}$ is not empty; i.e., there exist eigenvectors which are localized near the boundary or the corners of Ω.

2. Boundary layer homogenization. In this section we assume that Ω is a cylindrical bounded open set in $\mathbb{R}^N$ in the sense that it is defined by

$$\Omega = \Sigma \times]0; L[, \tag{19}$$

where Σ is an open bounded set in $\mathbb{R}^{N-1}$ and $L > 0$ is a positive length. With no loss of generality, we assume that the axis of the cylindrical domain Ω is parallel to the Nth

canonical direction. Therefore, a generic point x in Ω is denoted by $x = (x', x_N)$ with $x' \in \Sigma$ and $x_N \in]0; L[$. The goal of this section is to analyze the asymptotic behavior of that part of the spectrum $\sigma(S_\epsilon)$ which corresponds to eigenvectors concentrating on the boundary $\Sigma \times \{0\}$, under the sole geometric assumption (19) (in particular, no restrictions are made on the sequence ϵ which goes to zero).

2.1. Two-scale convergence for boundary layers. We begin by adapting the classical two-scale convergence method of Allaire [2] and Nguetseng [29] to the case of boundary layers, that is, sequences of functions in Ω which concentrate near the boundary $\Sigma \times \{0\}$. This method of "two-scale convergence for boundary layers" will allow us to understand this phenomenon of concentration of oscillations near the boundary. The usual two-scale convergence relies on periodically oscillating test functions with a unit period $Y =]0,1[^N$. Here, we use test functions which oscillate only in the directions parallel to the boundary Σ (with period $Y' =]0,1[^{N-1}$) and which simply decay in the Nth direction orthogonal to Σ.

Let us define a semi-infinite band $G = Y' \times]0; +\infty[$, where $Y' =]0,1[^{N-1}$ is the unit cell in $\mathbb{R}^{N-1}$. A generic point y is denoted by $y = (y', y_N)$ with $y' \in Y'$ and $y_N \in]0; +\infty[$. We introduce the space $L^2_\#(G)$ of square integrable functions in G which are periodic in the $(N-1)$ first variables, i.e.,

$$L^2_\#(G) = \{\phi(y) \in L^2(G) \mid y' \mapsto \phi(y', y_N) \text{ is } Y'\text{-periodic}\}.$$

We also denote by $C(\overline{\Sigma})$ the space of continuous functions on the closure of Σ, a compact set in $\mathbb{R}^{N-1}$.

Combining the concentration effect in y_N and the periodic oscillations in Y', the following convergence result is obtained for a sequence $\phi(\frac{x}{\epsilon})$ when ϕ belongs to $L^2_\#(G)$ (further modulated by $x' \in \Sigma$).

LEMMA 2.1. *Let $\varphi(x', y) \in L^2_\#\left(G; C(\overline{\Sigma})\right)$. Then*

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_\Omega \left|\varphi\left(x', \frac{x}{\epsilon}\right)\right|^2 dx = \frac{1}{|Y'|} \int_\Sigma \int_G |\varphi(x', y)|^2 dx' dy.$$

REMARK 2.2. *Remark that, in the left-hand side of the above equation, the second argument of φ is x/ϵ and not only x'/ϵ. This implies that there is a concentration effect near 0 in the x_N variable since φ is not periodic in this direction. This, in turn, explains the $1/\epsilon$ scaling in front of the left-hand side, in order to get a nonzero limit.*

As usual in the context of two-scale convergence, the above result is not specific to the space $L^2_\#\left(G; C(\overline{\Sigma})\right)$, which could be replaced, for example, by $L^2\left(\Sigma; C_{c\#}(\bar{G})\right)$, where $C_{c\#}(\bar{G})$ is the space of continuous functions in G, periodic in y' of period Y', and with bounded support in y_N.

In view of Lemma 2.1, we define a notion of "two-scale convergence for boundary layers."

DEFINITION 2.3. *Let $(u_\epsilon)_{\epsilon>0}$ be a sequence in $L^2(\Omega)$. It is said to* two-scale converge in the sense of boundary layers *on Σ if there exists $u_0(x', y) \in L^2(\Sigma \times G)$ such that*

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_\Omega u_\epsilon(x) \varphi\left(x', \frac{x}{\epsilon}\right) dx = \frac{1}{|Y'|} \int_\Sigma \int_G u_0(x', y) \varphi(x', y) dx' dy$$

for all smooth functions $\varphi(x', y)$ defined in $\Sigma \times G$ such that $y' \mapsto \varphi(x', y', y_N)$ is Y'-periodic and φ has a bounded support in $\Sigma \times G$.

This definition makes sense because of the following compactness theorem which generalizes the usual two-scale convergence compactness theorem in [2], [29].

THEOREM 2.4. *Let $(u_\epsilon)_{\epsilon>0}$ be a sequence in $L^2(\Omega)$ such that there exists a constant C, independent of ϵ, for which*

$$\frac{1}{\sqrt{\epsilon}}\|u_\epsilon\|_{L^2(\Omega)} \le C.$$

There exists a subsequence, still denoted by ϵ, and a limit function $u_0(x',y) \in L^2(\Sigma \times G)$ such that

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_\Omega u_\epsilon(x)\varphi\left(x',\frac{x}{\epsilon}\right)dx = \frac{1}{|Y'|}\int_\Sigma\int_G u_0(x',y)\varphi(x',y)dx'dy \tag{20}$$

for all functions $\varphi(x',y) \in L^2_\#\left(G;C(\overline{\Sigma})\right)$.

Remark that Theorem 2.4 does not apply to sequences which are merely bounded in $L^2(\Omega)$ but also converge strongly to zero in $L^2(\Omega)$ as the square root of ϵ. Of course, this is the case for a sequence of the type $\varphi(x',\frac{x}{\epsilon})$, where $\varphi(x',y)$ is as in Lemma 2.1; then, the limit is nothing but $\varphi(x',y)$ itself.

It is not difficult to check that the L^2-norm is weakly lower semicontinuous with respect to the two-scale convergence (see Proposition 1.6 in [2]); i.e., in the present situation

$$\lim_{\epsilon\to 0}\frac{1}{\sqrt{\epsilon}}\|u_\epsilon\|_{L^2(\Omega)} \ge \frac{1}{|Y'|^{1/2}}\|u_0\|_{L^2(\Sigma\times G)}.$$

The next proposition asserts a corrector-type result when the above inequality is actually an equality.

PROPOSITION 2.5. *Let $(u_\epsilon)_{\epsilon>0}$ be a sequence in $L^2(\Omega)$ which two-scale converges in the sense of boundary layers to a limit $u_0(x',y) \in L^2(\Sigma\times G)$. Assume further that it two-scale converges* strongly*, that is,*

$$\lim_{\epsilon\to 0}\frac{1}{\sqrt{\epsilon}}\|u_\epsilon\|_{L^2(\Omega)} = \frac{1}{|Y'|^{1/2}}\|u_0\|_{L^2(\Sigma\times G)}.$$

Then,

(i) *for any sequence $(v_\epsilon)_{\epsilon>0}$ in $L^2(\Omega)$ which two-scale converges in the sense of boundary layers to a limit $v_0(x',y) \in L^2(\Sigma\times G)$, one has*

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_\Omega u_\epsilon v_\epsilon dx = \frac{1}{|Y'|}\int_\Sigma\int_G u_0(x',y)v_0(x',y)dx'dy;$$

(ii) *if $u_0(x',y)$ is smooth, say $u_0 \in L^2_\#\left(G;C(\overline{\Sigma})\right)$, then*

$$\lim_{\epsilon\to 0}\frac{1}{\sqrt{\epsilon}}\left\|u_\epsilon(x) - u_0\left(x',\frac{x}{\epsilon}\right)\right\|_{L^2(\Omega)} = 0.$$

In order to investigate the convergence of sequences of functions in $H^1_0(\Omega)$, we first have to define adequate functional spaces for the two-scale limit. Let $C^\infty_{c\#}(G)$ be the space of smooth functions in $\overline{G}$ which are Y'-periodic in y' and have a compact support in y_N (i.e., they vanish for sufficiently large and small y_N but not necessarily on the whole ∂G). Let $H^1_{0\#}(G)$ be the Sobolev space obtained by completion of $C^\infty_{c\#}(G)$ with

respect to the $H^1(G)$-norm. We denote by $H^1_{0\#,loc}(G)$ the space of functions which are "locally" in $H^1_{0\#}(G)$, i.e., which coincide with a function of $H^1_{0\#}(G)$ in any compact set of $\overline{G}$. We define a Deny–Lions-type space (cf. [18]) $D^1_{0\#}(G)$ as the completion of $C^\infty_{c\#}(G)$ with respect to the $L^2(G)^N$-norm of the gradient

$$D^1_{0\#}(G) = \Big\{\psi(y) \in H^1_{0\#,loc}(G) \mid \exists \quad \psi_n \in C^\infty_{c\#}(G) \quad \text{such that} \quad \lim_{n\to+\infty} \|\nabla(\psi-\psi_n)\|_{L^2(G)^N} = 0\Big\}. \tag{21}$$

It is easily seen that a function in $D^1_{0\#}(G)$ vanishes when $y_N = 0$ but does not necessarily go to 0 when y_N goes to infinity since $D^1_{0\#}(G)$ contains functions which grow like y_N^α at infinity with $\alpha < 1/2$. We are now in a position to state our next result.

PROPOSITION 2.6. *Let $(u_\epsilon)_{\epsilon>0}$ be a sequence in $H^1_0(\Omega)$ such that there exists a constant C, independent of ϵ, for which*

$$\frac{1}{\sqrt{\epsilon}}\left(\|u_\epsilon\|_{L^2(\Omega)} + \|\nabla u_\epsilon\|_{L^2(\Omega)^N}\right) \le C.$$

Then, there exists a subsequence, still denoted by ϵ, and a limit $u_0(x',y) \in L^2(\Sigma; D^1_{0\#}(G))$ such that

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_\Omega u_\epsilon(x)\varphi\left(x',\frac{x}{\epsilon}\right)dx = 0,$$

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_\Omega \nabla u_\epsilon(x)\cdot\psi\left(x',\frac{x}{\epsilon}\right)dx = \frac{1}{|Y'|}\int_\Sigma\int_G \nabla_y u_0(x',y)\cdot\psi(x',y)dx'dy$$

for any functions $\varphi \in L^2_\#\left(G; C(\overline{\Sigma})\right)$ and $\psi \in L^2_\#\left(G; C(\overline{\Sigma})^N\right)$.

Remark that, in Proposition 2.6, the two-scale limit $u_0(x',y)$ does not belong to $L^2(\Sigma; H^1(G))$ as could be expected. The reason is that only $\nabla_y u_0 \in L^2(\Sigma\times G)$, while u_0 itself has no reason to belong to $L^2(\Sigma\times G)$. Since the proofs of the above results are very similar to those of the usual two-scale convergence theory, we simply sketch the proofs of Lemma 2.1, Theorem 2.4, and Proposition 2.6.

Proof of Lemma 2.1. Let us first assume that $\varphi(x',y) \in L^2_\#\left(G; C(\overline{\Sigma})\right)$ has bounded support in y_N; i.e., there exists $M > 0$ such that

$$\varphi(x',y) = 0 \text{ if } y_N \ge M.$$

Then, by the change of variables $y_N = x_N/\epsilon$ and for sufficiently small ϵ, we have

$$\begin{aligned}\tfrac{1}{\epsilon}\textstyle\int_\Omega |\varphi(x',\frac{x}{\epsilon})|^2dx &= \tfrac{1}{\epsilon}\textstyle\int_0^L\int_\Sigma |\varphi(x',\frac{x'}{\epsilon},\frac{x_N}{\epsilon})|^2dx'dx_N\\ &= \textstyle\int_0^{L/\epsilon}\int_\Sigma |\varphi(x',\frac{x'}{\epsilon},y_N)|^2dx'dy_N\\ &= \textstyle\int_0^{M}\int_\Sigma |\varphi(x',\frac{x'}{\epsilon},y_N)|^2dx'dy_N.\end{aligned} \tag{22}$$

The usual convergence result for oscillating functions in $\mathbb{R}^{N-1}$ (see, e.g., [2] and references therein) yields that for almost everywhere $y_N \in (0;M)$

$$\lim_{\epsilon\to 0}\int_\Sigma\left|\varphi\left(x',\frac{x'}{\epsilon},y_N\right)\right|^2dx' = \frac{1}{|Y'|}\int_\Sigma\int_{Y'}|\varphi(x',y',y_N)|^2dx'dy'$$

and that

$$\int_\Sigma \left|\varphi\left(x', \frac{x'}{\epsilon}, y_N\right)\right|^2 dx' \le |\Sigma| \int_{Y'} \max_{x'\in\overline{\Sigma}} |\varphi(x', y', y_N)|^2 dy'.$$

Therefore, applying the Lebesgue theorem, we deduce that

$$\lim_{\epsilon\to 0} \int_0^M \int_\Sigma \left|\varphi\left(x', \frac{x'}{\epsilon}, y_N\right)\right|^2 dx' dy_N = \frac{1}{|Y'|} \int_\Sigma \int_G |\varphi(x', y', y_N)|^2 dx' dy.$$

The density of such functions $\varphi(x', y)$ in $L^2_\#\left(G; C(\overline{\Sigma})\right)$ implies the desired result for any function in $L^2_\#\left(G; C(\overline{\Sigma})\right)$.

Proof of Theorem 2.4. Using the assumed uniform bound on u_ϵ, by the Schwarz inequality we obtain

$$\left|\frac{1}{\epsilon}\int_\Omega u_\epsilon(x)\varphi\left(x', \frac{x}{\epsilon}\right) dx\right| \le C\Big(\frac{1}{\epsilon}\int_\Omega \Big|\varphi\Big(x', \frac{x}{\epsilon}\Big)\Big|^2 dx\Big)^{\frac{1}{2}}.$$

Passing to the limit, up to a subsequence, which may depend on φ in the left-hand side and using Lemma 2.1 in the right-hand side, yield

$$\left|\lim_{\epsilon\to 0} \frac{1}{\epsilon}\int_\Omega u_\epsilon(x)\varphi\left(x', \frac{x}{\epsilon}\right) dx\right| \le C\Big(\int_\Sigma \int_G |\varphi(x', y)|^2 dx' dy\Big)^{\frac{1}{2}}. \tag{23}$$

Since $L^2_\#\left(G; C(\overline{\Sigma})\right)$ is separable, varying φ over a dense countable subset, by a standard diagonalization process, we can extract a subsequence of ϵ such that (23) is valid for all functions φ in this subset. By density, we conclude that the limit in the left side of (23), as a function of φ, defines a continuous linear form in $L^2(\Sigma\times G)$. Then, the classical Riesz representation theorem immediately implies the existence of a function $u_0(x, y) \in L^2(\Sigma\times G)$ which satisfies (20). This finishes the proof of Theorem 2.4.

Proof of Proposition 2.6. By application of Theorem 2.4, up to a subsequence, there exist two limits $u(x', y) \in L^2(\Sigma\times G)$ and $\xi^0(x', y) \in L^2(\Sigma\times G)^N$ such that u_ϵ and ∇u_ϵ two-scale converge in the sense of boundary layers to these respective limits; i.e.,

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\int_\Omega u_\epsilon(x)\varphi\left(x', \frac{x}{\epsilon}\right) dx = \frac{1}{|Y'|}\int_\Sigma \int_G u(x', y)\varphi(x', y) dx' dy, \tag{24}$$

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\int_\Omega \nabla u_\epsilon(x)\cdot\psi\left(x', \frac{x}{\epsilon}\right) dx = \frac{1}{|Y'|}\int_\Sigma \int_G \xi_0(x', y)\cdot\psi(x', y) dx' dy \tag{25}$$

for any functions $\varphi \in L^2_\#\left(G; C(\overline{\Sigma})\right)$ and $\psi \in L^2_\#\left(G; C(\overline{\Sigma})^N\right)$. Integrating by parts in (25), we obtain

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\int_\Omega u_\epsilon(x)\mathrm{div}_y\psi\left(x', \frac{x}{\epsilon}\right) dx = 0.$$

In view of (24), this implies that

$$\frac{1}{|Y'|}\int_\Sigma \int_G u(x', y)\mathrm{div}_y\psi(x', y) dx' dy = 0.$$

Another integration by parts yields that $u(x', y)$ does not depend on y. On the other hand, it belongs to $L^2(\Sigma \times G)$ and G is unbounded. Since the only constant which belongs to $L^2(G)$ is zero, we deduce that $u = 0$. Now, specializing (25) to test functions ψ such that $\mathrm{div}_y \psi = 0$ and integrating by parts, we also obtain that

$$\frac{1}{|Y'|} \int_\Sigma \int_G \xi_0(x', y) \cdot \psi(x', y) dx' dy = 0.$$

As is well known, the orthogonal of divergence-free fields is exactly the set of gradients (see Proposition 1.14 in [2] for a precise statement and references). Therefore, there exists a function $u_0(x', y)$ in $L^2(\Sigma; D^1_{0\#}(G))$ such that $\xi_0 = \nabla_y u_0$ (we use the space $D^1_{0\#}(G)$ since u_0 has no reason to belong to $L^2(\Sigma \times G)$).

2.2. Convergence analysis. Recall that the original operator S_ϵ, defined by (3), acts in the space $\mathbb{R}^{Nn(\epsilon)}$ which depends on ϵ and that our strategy was to extend S_ϵ to a fixed space where a convergence analysis is possible. So far, the domain $\Omega = \Sigma \times]0, L[$ was considered periodic of period ϵY. Nevertheless, from now on, Ω is seen as a periodic domain with a new period G^K_ϵ defined by

$$G^K_\epsilon \stackrel{\text{def}}{=}]0; \epsilon K[^{N-1} \times]0; L[,$$

with K an integer larger than 1. We shall construct an extension of S_ϵ well suited for the previous two-scale convergence "in the sense of boundary layers" with such a period G^K_ϵ.

REMARK 2.7. *As already mentioned, we make no special hypothesis on the sequence of small parameters ϵ. However, the periodic arrangement of tubes in Ω is required to be aligned with Σ in such a way that the first row of periodic cells ϵY has a boundary which coincides with $\Sigma \times \{0\}$. In other words, the first layer of tubes close to Σ is at a fixed distance $\frac{\epsilon}{2}$ of $\Sigma \times \{0\}$ (see Figure 1).*

By a rescaling of ratio ϵ, this new period G^K_ϵ corresponds to a finite length truncation of the new reference cell

$$G^K \stackrel{\text{def}}{=} KG =]0; K[^{N-1} \times]0; +\infty[= KY' \times]0; +\infty[.$$

In the reference cell G^K (see Figure 2) we put infinitely many layers of tubes in the Nth direction, each layer being made of K^{N-1} tubes. The tubes in G^K are denoted by T_j, where $j = (j', j_N)$ is a multi-index such that $j_N \geq 1$ is an integer, which labels the corresponding layer in G^K, and j' is a multi-integer in $\{0, 1, \ldots, K-1\}^{N-1}$, which locates the tube T_j in its layer j_N. The fluid part in G^K is denoted by G^{*K}, i.e.,

$$G^{*K} \;=\; G^K \setminus \bigcup_{\substack{0 \leq j' \leq K-1 \\ 1 \leq j_N}} T_j.$$

To each tube T_j in G^K we associate the subcell Y_j and the fluid subcell $Y^*_j = Y_j \setminus T_j$ analogous to Y and Y^*, respectively (see Figure 2). The main idea is to attach to each tube T_j in G^K a different displacement function $\vec{s}(x')$, depending only on the variable $x' \in \Sigma$, such that the family $(\vec{s}_j(x'))_{\substack{0 \leq j' \leq K-1 \\ 1 \leq j_N}}$ belongs to the space $L^2(\Sigma; \ell^2_K)$, where ℓ^2_K is the Hilbert space defined by

$$\ell^2_K = \left\{ (\vec{s}_j)_{\substack{0 \leq j' \leq K-1 \\ 1 \leq j_N}} \;\middle|\; \vec{s}_j \in \mathbb{C}^N, \quad \sum_{\substack{0 \leq j' \leq K-1 \\ 1 \leq j_N}} |\vec{s}_j|^2 < +\infty \right\}.$$

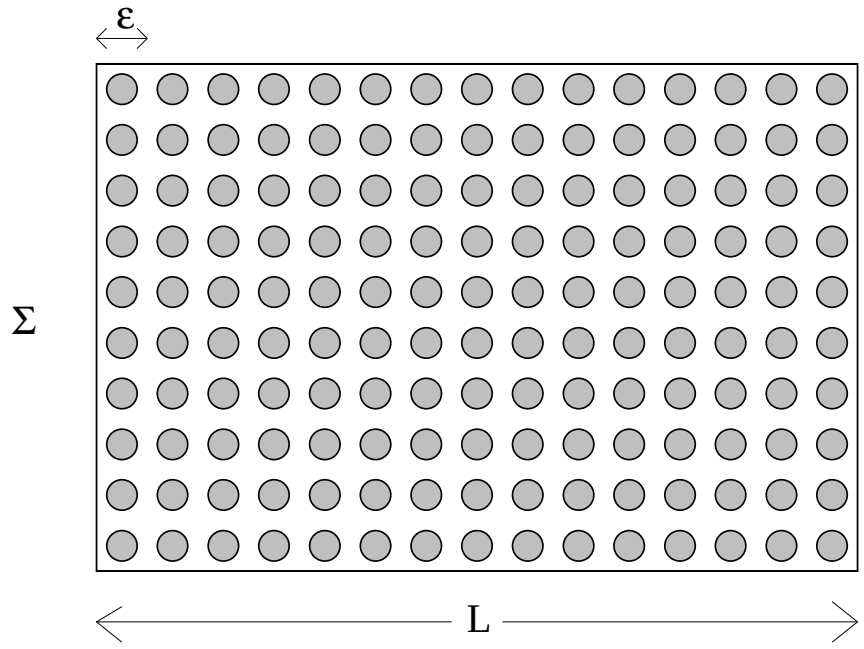

FIG. 1. *Cylindrical domain* $\Omega = \Sigma \times (0, L)$.

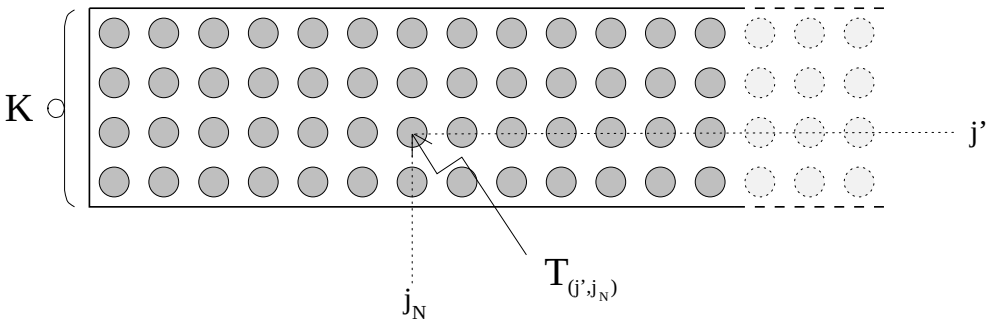

FIG. 2. *Reference cell* G^K.

Remark that this definition of ℓ^2_K implies a decay of the displacement function $\vec{s}_j$ as j_N goes to $+\infty$. Note also that each family $(\vec{s}_j(x')) \in L^2(\Sigma; \ell^2_K)$ can be identified with a function $\vec{s}(x', y) \in L^2(\Sigma \times G^K)$ which is constant in each subcell Y_j.

We now introduce the extended operator B^K_ϵ defined in $L^2(\Sigma; \ell^2_K)$ by

$$B^K_\epsilon = E^K_\epsilon S_\epsilon P^K_\epsilon,$$

where P^K_ϵ and E^K_ϵ are, respectively, projection and extension operators between $\mathbb{R}^{Nn(\epsilon)}$ and $L^2(\Sigma; \ell^2_K)$. To define precisely P^K_ϵ and E^K_ϵ we need the following notation.

DEFINITION 2.8. *Let $j = (j', j_N)$ denote the multi-index which enumerates all tubes in the periodic reference cell G^K. We use the notation $0 \le j' \le K - 1$ to indicate that j' varies in $\{0, 1, \ldots, K-1\}^{N-1}$ and $j_N \ge 1$ to indicate that j_N takes any positive integer value. Let $p = (p_1, \ldots, p_N)$ be the multi-integer which enumerates all the tubes in Ω (see Definition 1). The index p is such that the tube T^ϵ_p is located*

in the cell whose origin lies at the point $\epsilon p \in \Omega$. To describe its range we use the notation $1 \leq p \leq n(\epsilon)$, where $n(\epsilon)$ is the total numbers of tubes in Ω. We define a third multi-integer $\ell' = (\ell_1, \ldots, \ell_{N-1})$ which enumerates all the periodic reference cells $G^K_{\epsilon,\ell'}$ covering Ω (each being identical, up to a translation, to G^K_ϵ). For simplicity its range is denoted by $1 \leq \ell' \leq n_K(\epsilon)$. These three indices are assumed to be related by the following one-to-one relationship:

$$\left\{ \begin{array}{l} \ell_m = E(\frac{p_m}{K}), \quad j_m = p_m - K\ell_m \quad \textit{for} \quad 1 \leq m \leq N-1, \\ j_N = p_N, \end{array} \right. \tag{26}$$

where E denotes the integer-part function. This yields a one-to-one map between the tubes (T^ϵ_p) and their location in the cell $G^K_{\epsilon,\ell'}$ at the position j' in the layer j_N.

Then, we define a projection

$$\begin{array}{rcl} P^K_\epsilon \,:\, L^2(\Sigma; \ell^2_K) & \longrightarrow & \mathbb{R}^{Nn(\epsilon)}, \\ (\vec{s}_j(x'))_{\substack{0 \leq j' \leq K-1 \\ 1 \leq j_N}} & \mapsto & (\vec{s}_p)_{1 \leq p \leq n(\epsilon)} \end{array} \tag{27}$$

given by

$$\vec{s}_p = \frac{1}{|\epsilon K Y'|} \int_{(\epsilon K Y')_{\ell'}} \vec{s}_j(x') dx',$$

where (p, j, ℓ') are related by formula (26) and $(\epsilon K Y')_{\ell'}$ is the cross section of the cell $G^K_{\epsilon,\ell'}$.

We also define an extension

$$\begin{array}{rcl} E^K_\epsilon \,:\, \mathbb{R}^{Nn(\epsilon)} & \longrightarrow & L^2(\Sigma; \ell^2_K), \\ (\vec{s}_p)_{1 \leq p \leq n(\epsilon)} & \mapsto & (\vec{s}_j(x'))_{\substack{0 \leq j' \leq K-1 \\ 1 \leq j_N}} \end{array} \tag{28}$$

given by

$$\vec{s}_j(x') = \sum_{\ell'} \chi_{(\epsilon K Y')_{\ell'}}(x') \vec{s}_p,$$

where (p, j, ℓ') are related by formula (26) and $\chi_{(\epsilon K Y')_{\ell'}}(x')$ is the characteristic function of $(\epsilon K Y')_{\ell'}$. By convention, $\vec{s}_p$ is taken equal to 0 if the values of j and ℓ' correspond to a cell truncated by the boundary $\partial\Omega$ which therefore contains no tube.

One can easily check that P^K_ϵ and E^K_ϵ are adjoint operators (up to a multiplicative constant) and that the product $P^K_\epsilon E^K_\epsilon$ is nothing but the identity in $\mathbb{R}^{Nn(\epsilon)}$. Therefore, the spectrum of B^K_ϵ consists of that of S_ϵ and zero as an eigenvalue of infinite multiplicity. We summarize these results in the next lemma, the proof of which is safely left to the reader.

LEMMA 2.9. *The operators P^K_ϵ and E^K_ϵ satisfy the following properties;*

1. $(P^K_\epsilon)^\star = (\epsilon K)^{-(N-1)} E^K_\epsilon$,
2. $(E^K_\epsilon)^\star = (\epsilon K)^{(N-1)} P^K_\epsilon$,
3. $P^K_\epsilon E^K_\epsilon = Id_{\mathbb{R}^{Nn(\epsilon)}}$.

Therefore, the extended operator $B^K_\epsilon = E^K_\epsilon S_\epsilon P^K_\epsilon$ is self-adjoint and compact in $L^2(\Sigma; \ell^2_K)$. Its spectrum is

$$\sigma(B^K_\epsilon) = \sigma(S_\epsilon) \bigcup \{0\}.$$

The convergence analysis of this sequence of extended operators B_ϵ^K is amenable to the two-scale convergence method in the sense of boundary layers (as introduced in the previous section). It turns out that the corresponding limit operator B^K has a complicated form which can be considerably simplified by introducing the so-called Bloch wave decomposition. However, we emphasize that this decomposition will affect only the $(N-1)$ first variables and not the last one, orthogonal to the boundary Σ.

LEMMA 2.10. *Given a family $(\vec{s}_j)_{\substack{0\le j'\le K-1\\ 1\le j_N}}$ in ℓ^2_K, there exists a unique family $(\vec{t}_j)_{\substack{0\le j'\le K-1\\ 1\le j_N}}$ in ℓ^2_K such that, for any fixed j_N,*

$$\sum_{0\le j'\le K-1} \vec{s}_j \chi_{Y_{j'}}(y') = \sum_{0\le j'\le K-1} \vec{t}_j e^{2\pi\imath \frac{j'}{K}\cdot E(y')},$$

where $E(\cdot)$ denotes the integer part function and $(Y_{j'})_{0\le j'\le K-1}$ is the family of subcells of KY'. Moreover, Parseval's identity holds true; i.e., for any fixed j_N,

$$\sum_{0\le j'\le K-1} |\vec{s}_j|^2 \;=\; K^{N-1} \sum_{0\le j'\le K-1} |\vec{t}_j|^2.$$

The proof of Lemma 2.10 is standard (see, e.g., [1]). Remark that ℓ^2_K is isomorphic to $(\ell^2_1)^{K^{N-1}}$ by identifying an element $(\vec{s}_j)_{\substack{0\le j'\le K-1\\ 1\le j_N}}$ of ℓ^2_K as a collection of K^{N-1} elements $(\vec{s}_{(j',j_N)})_{j_N\ge 1}$ of ℓ^2_1. Therefore, in Lemma 2.10, one could replace ℓ^2_K by $(\ell^2_1)^{K^{N-1}}$. Let us define a linear map $\mathcal{B}'$

$$\begin{array}{rccl} \mathcal{B}' : & \ell^2_K & \longrightarrow & (\ell^2_1)^{K^{N-1}}, \\ & (\vec{s}_j) & \mapsto & (K^{\frac{N-1}{2}}\vec{t}_{(j',j_N)}), \end{array} \tag{29}$$

where the vectors $\vec{s}_j$ and $\vec{t}_j$ are related as in Lemma 2.10. This Bloch decomposition $\mathcal{B}'$ (the prime indicates that it concerns only the first $(N-1)$ variables) is easily seen to be an isometry from ℓ^2_K to $(\ell^2_1)^{K^{N-1}}$; namely, $(\mathcal{B}')^\star = (\mathcal{B}')^{-1}$.

We are now in a position to state the main result on the asymptotic behavior of B_ϵ^K.

THEOREM 2.11. *For each fixed $K\ge 1$, as ϵ goes to 0, the sequence B_ϵ^K converges strongly to a limit B^K into $L^2(\Sigma;\ell^2_K)$; i.e., for any function $\vec{s}(x')\in L^2(\Sigma;\ell^2_K)$ we have*

$$B_\epsilon^K \vec{s}(x') \longrightarrow B^K \vec{s}(x') \quad \text{in } L^2(\Sigma;\ell^2_K) \text{ strongly.}$$

By using the Bloch decomposition $\mathcal{B}'$ defined in (29)*, the operator B^K can be diagonalized*

$$B^K = (\mathcal{B}')^\star D^K \mathcal{B}' \quad \text{with} \quad D^K = \operatorname{diag}(D^K_{j'})_{0\le j'\le K-1},$$

where the entries $D^K_{j'}$ are self-adjoint continuous (but not compact) operators in $L^2(\Sigma;\ell^2_1)$ defined, for any $(\vec{s}_{j_N}(x'))_{j_N\ge 1}\in L^2(\Sigma;\ell^2_1)$, by

$$D^K_{j'}(\vec{s}_{j_N}(x')) = \left(\int_{\Gamma_{j_N}} u_{j'}\vec{n}ds\right)_{j_N\ge 1},$$

where $u_{j'}(y)$ is the unique solution of

$$(30)\qquad \begin{cases} -\Delta_y u_{j'} = 0 & \text{in } G^*, \\ \frac{\partial u_{j'}}{\partial n} = \vec{s}_{j_N} \cdot \vec{n} & \text{on } \Gamma_{j_N},\ j_N \geq 1, \\ u_{j'} = 0 & \text{on } y_N = 0, \\ y' \mapsto e^{-2\pi\imath \frac{j'}{K}\cdot y'} u_{j'}(y', y_N) & Y' - \text{periodic}, \end{cases}$$

where G^* is the fluid part of the semi-infinite band G (see Figure 2).

REMARK 2.12. *Of course, the solution $u_{j'}$ of* (30) *depends also on the variable $x' \in \Sigma$ since each displacement $\vec{s}_{j_N}(x')$ depends on x'. Nevertheless, x' plays the role of a parameter, since* (30) *is a partial differential equation in the variable y only. The limit problem* (30) *admits a unique solution $u_{j'}(x', y)$ in the space $L^2(\Sigma; D^1_{j',\#}(G^*))$, where $D^1_{j',\#}(G^*)$ is a Deny–Lions-type space. More precisely, it is defined as $D^1_{0\#}(G)$ in* (21)*, the only difference being that functions in $D^1_{j',\#}(G^*)$ satisfy a $(e^{2\pi\imath \frac{j'}{K}}, Y')$ periodicity condition in y', instead of the usual Y' periodicity. Recall that a function $w(y)$ satisfying the periodicity condition of the limit problem* (30) *is said to be $(e^{2\pi\imath \frac{j'}{K}}, Y')$-periodic in y' because such a function also satisfies the following (generalized) periodicity condition:*

$$w(y + (k', 0)) = e^{2\pi\imath \frac{j'\cdot k'}{K}} w(y) \quad \forall y = (y', y_N) \text{ and } \forall k' \in \mathbb{Z}^{N-1}.$$

For more details on this class of functions, we refer to [1], [16].

The key of the proof of Theorem 2.11 is the following homogenization result for the fluid potential when the displacements of the tubes are given in terms of the projection operator P^K_ϵ. Remark that, in view of definition (27) of P^K_ϵ, such a family of displacements concentrates near the boundary $\Sigma \times \{0\}$ as ϵ goes to 0.

PROPOSITION 2.13. *For any $\vec{s}(x') \in L^2(\Sigma; \ell^2_K)$ let us define $u_\epsilon = u_\epsilon(\vec{s})$ as the unique solution in $H^1(\Omega_\epsilon)$ of*

$$(31)\qquad \begin{cases} -\Delta u_\epsilon = 0 & \text{in } \Omega_\epsilon, \\ \frac{\partial u_\epsilon}{\partial n} = \left(P^K_\epsilon \vec{s}(x')\right)_p \cdot \vec{n} & \text{on } \Gamma^\epsilon_p,\ 1 \leq p \leq n(\epsilon), \\ u_\epsilon = 0 & \text{on } \partial\Omega. \end{cases}$$

Then, u_ϵ two scale converges in the sense of boundary layers to 0 and ∇u_ϵ two-scale converges in the sense of boundary layers to $\nabla_y u_0(x', y)$, where $u_0(x', y)$ is the unique solution in $L^2(\Sigma, D^1_{0\#}(G^{*K}))$ of

$$(32)\qquad \begin{cases} -\Delta_y u_0 = 0 & \text{in } G^{*K}, \\ \frac{\partial u_0}{\partial n} = \vec{s}_j \cdot \vec{n} & \text{on } \Gamma_j, \\ u_0 = 0 & \text{if } y_N = 0, \\ y' \mapsto u_0(x', y', y_N) & KY'\text{-periodic}, \end{cases}$$

and ∇u_ϵ two-scale converges strongly, i.e.,

$$(33)\qquad \lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_{\Omega_\epsilon} |\nabla u_\epsilon|^2 dx = \frac{1}{|KY'|} \int_\Sigma \int_{G^K} |\nabla_y u_0|^2 dx' dy.$$

Moreover, if $\vec{s}^\epsilon(x')$ is a sequence which converges weakly to a limit $\vec{s}(x')$ in $L^2(\Sigma; \ell^2_K)$, then the sequence of associated solutions $u_\epsilon(\vec{s}^\epsilon)$ two-scale converges in the sense of boundary layers to 0 and $\nabla u_\epsilon(\vec{s}^\epsilon)$ two-scale converges in the sense of boundary layers to $\nabla u_0(x', y)$, where u_0 is still the solution of (32).

REMARK 2.14. *A priori, the solution u_ϵ of* (31) *is defined only in the fluid domain Ω_ϵ which is a varying set as ϵ goes to* 0. *However, it is a standard matter (see* [13]*) to build an extension operator X_ϵ acting from $H^1(\Omega_\epsilon)$ into $H^1(\Omega)$ such that, for any $v \in H^1(\Omega_\epsilon)$,*

$$X_\epsilon v = v \text{ in } \Omega_\epsilon \text{ and } \|X_\epsilon v\|_{H^1(\Omega)} \le C\|v\|_{H^1(\Omega_\epsilon)},$$

where C is a positive constant independent of ϵ. In what follows, we shall always identify functions in $H^1(\Omega_\epsilon)$ (as u_ϵ) with their extension in $H^1(\Omega)$ (as $X_\epsilon u_\epsilon$).

To prove Proposition 2.13 we need two technical lemmas.

LEMMA 2.15. *The extension and projection operators E_ϵ^K and P_ϵ^K satisfy the following estimates:*

(i) $\|P_\epsilon^K \vec{s}(x')\|_{\mathbb{R}^{Nn(\epsilon)}} \le C\epsilon^{-\frac{N-1}{2}} \|\vec{s}(x')\|_{L^2(\Sigma;\ell^2_K)}$,

(ii) $\|E_\epsilon^K(\vec{s}_p)\|_{L^2(\Sigma;\ell^2_K)} \le C\epsilon^{\frac{N-1}{2}} \|(\vec{s}_p)_{1\le p\le n(\epsilon)}\|_{\mathbb{R}^{Nn(\epsilon)}}$,

where C is a constant independent of ϵ and the norms are defined by

$$\|(\vec{s}_p)_{1\le p\le n(\epsilon)}\|^2_{\mathbb{R}^{Nn(\epsilon)}} = \sum_{1\le p\le n(\epsilon)} |\vec{s}_p|^2,$$

$$\|\vec{s}(x')\|^2_{L^2(\Sigma;\ell^2_K)} = \int_\Sigma \sum_{\substack{0\le j'\le K-1\\ 1\le j_N}} |\vec{s}_j(x')|^2 dx'.$$

Proof. Let us prove (i) (the other inequality (ii) has a similar proof). By definition of P_ϵ^K,

$$\|P_\epsilon^K \vec{s}(x')\|^2_{\mathbb{R}^{Nn(\epsilon)}} = \sum_{1\le p\le n(\epsilon)} \Big(\frac{1}{|\epsilon KY'|}\int_{(\epsilon KY')_{\ell'}} \vec{s}_j(x')dx'\Big)^2,$$

where (p, j, ℓ') are related by formula (26). Applying the Cauchy–Schwarz inequality and summing over ℓ' yield

$$\begin{aligned} \|P_\epsilon^K \vec{s}(x')\|^2_{\mathbb{R}^{Nn(\epsilon)}} &\le \sum_{1\le p\le n(\epsilon)} \tfrac{1}{|\epsilon KY'|}\textstyle\int_{(\epsilon KY')_{\ell'}} |\vec{s}_j(x')|^2 dx' \\ &\le \tfrac{1}{(K\epsilon)^{N-1}}\textstyle\int_\Sigma \sum_j |\vec{s}_j(x')|^2 dx', \end{aligned} \tag{34}$$

which is the desired result.

LEMMA 2.16. *Let $\vec{s}^\epsilon(x')$ be a sequence of functions which converges weakly to $\vec{s}(x')$ in $L^2(\Sigma;\ell^2_K)$. Define a piecewise constant function*

$$\vec{a}^\epsilon(x) = \sum_{\ell'}\sum_j \Big(\frac{1}{|\epsilon KY'|}\int_{(\epsilon KY')_{\ell'}} \vec{s}^\epsilon_j(x')dx'\Big)\chi_{Y^\epsilon_{j\ell'}}(x),$$

where $\chi_{Y^\epsilon_{j\ell'}}(x)$ is the characteristic function of the jth subcell of the periodic cell $G^K_{\epsilon,\ell'}$. Then, $\vec{a}^\epsilon$ two-scale converges in the sense of boundary layers to a limit $\vec{a}^0(x,y) \in L^2(\Sigma \times G^K)$ defined by

$$\vec{a}^0(x,y) = \sum_j \vec{s}_j(x')\chi_{Y_j}(y),$$

where $\chi_{Y_j}(y)$ *is the characteristic function of the jth subcell of the reference cell* G^K. *Moreover, if* $\vec{s}^\epsilon(x')$ *converges* strongly *to* $\vec{s}(x')$ *in* $L^2(\Sigma;\ell^2_K)$, *then* $\vec{a}^\epsilon$ *two-scale converges* strongly *to* $\vec{a}^0$ *in the sense of boundary layers, i.e.,*

$$\lim_{\epsilon\to 0}\frac{1}{\sqrt{\epsilon}}\|\vec{a}^\epsilon(x)\|_{L^2(\Omega)}=\frac{1}{K^{\frac{N-1}{2}}}\|\vec{a}^0(x',y)\|_{L^2(\Sigma\times G^K)}.$$

Proof. The proof is very similar to that of Lemma 3.3.2 in our previous work [3], so we briefly sketch it. Let $\vec{\varphi}(x',y)$ be a suitable smooth test function defined on $\Sigma\times G^K$ with values in $\mathbb{R}^N$ such that $y'\to\vec{\varphi}(x',y',y_N)$ is KY'-periodic and $\vec{\varphi}$ vanishes for sufficiently large y_N. We check the definition of two-scale convergence:

$$\begin{aligned}&\tfrac{1}{\epsilon}\textstyle\int_\Omega \vec{a}^\epsilon(x)\cdot\vec{\varphi}(x',\frac{x}{\epsilon})dx\\&=\tfrac{1}{\epsilon}\sum_{\ell,j}\Big(\tfrac{1}{(\epsilon K)^{N-1}}\textstyle\int_{\epsilon(KY')_{\ell'}}\vec{s}^\epsilon_j(x')dx'\Big)\cdot\int_{Y^\epsilon_{j\ell'}}\vec{\varphi}(x',\tfrac{x}{\epsilon})dx\\&=\tfrac{1}{K^{N-1}}\textstyle\sum_j\int_\Sigma\vec{s}^\epsilon_j(x')\cdot\Big[\sum_{\ell'}\Big(\tfrac{1}{\epsilon^N}\int_{Y^\epsilon_{j\ell'}}\vec{\varphi}(x',\tfrac{x}{\epsilon})dx\Big)\chi_{\epsilon(KY')_{\ell'}}(x')\Big]dx'.\end{aligned}$$

It is easily seen that for each fixed j the term between brackets converges strongly to $\int_{Y_j}\vec{\varphi}(x',y)dy$ in $L^2(\Sigma)^N$. Remark that the sum in j is finite since $\vec{\varphi}$ has a bounded support in G^K. Thus we can pass to the limit and obtain the desired result

$$\frac{1}{K^{N-1}}\sum_j\int_\Sigma\vec{s}_j(x')\cdot\left(\int_{Y_j}\vec{\varphi}(x',y)dy\right)dx'.$$

If $\vec{s}^\epsilon_j$ converges strongly to $\vec{s}_j$, the strong two-scale convergence of $\vec{a}^\epsilon(x)$ is obtained by a similar proof, replacing in the above computation the test function $\vec{\varphi}$ by $\vec{a}^\epsilon(x)$.

Proof of Proposition 2.13. Multiplying (31) by u_ϵ and integrating by parts, we get

$$\begin{aligned}\textstyle\int_{\Omega_\epsilon}|\nabla u_\epsilon|^2dx&=\sum_{1\le p\le n(\epsilon)}\big(P^K_\epsilon\vec{s}\big)_p\cdot\textstyle\int_{\Gamma^\epsilon_p}u_\epsilon\vec{n}ds\\&\le\|\big(P^K_\epsilon\vec{s}\big)\|_{\mathbb{R}^{Nn(\epsilon)}}\left\|\Big(\textstyle\int_{\Gamma^\epsilon_p}u_\epsilon\vec{n}ds\Big)\right\|_{\mathbb{R}^{Nn(\epsilon)}}.\end{aligned}\tag{35}$$

An easy calculation (see Lemma 2.2.3 in [3] if necessary) shows that

$$\left\|\Big(\int_{\Gamma^\epsilon_p}u_\epsilon\vec{n}ds\Big)\right\|^2_{\mathbb{R}^{Nn(\epsilon)}}\le C\epsilon^N\|\nabla u_\epsilon\|^2_{L^2(\Omega_\epsilon)^N},$$

and hence, using Lemma 2.15 we conclude that

$$\int_{\Omega_\epsilon}|\nabla u_\epsilon|^2dx\le C\epsilon\|\vec{s}(x')\|^2_{L^2(\Sigma;\ell^2_K)}.$$

A standard Poincaré inequality in Ω yields the same estimate for u_ϵ in $L^2(\Omega_\epsilon)$:

$$\int_{\Omega_\epsilon}|u_\epsilon|^2dx\le C\epsilon\|\vec{s}(x')\|^2_{L^2(\Sigma;\ell^2_K)}.$$

We now apply the method of two-scale convergence for the asymptotic analysis of the sequence u_ϵ, using test functions with G^K as the periodic cell (since we decided to consider G^K to be the reference cell and not G). By virtue of Proposition 2.6 there exists a subsequence of u_ϵ and a limit function $u_0(x', y)$ in $L^2(\Sigma; D^1_{0\#}(G^K))$ such that $(u_\epsilon, \nabla u_\epsilon)$ two-scale converge in the sense of boundary layers to $(0, \nabla_y u_0)$. Let $\varphi(x', y)$ be a smooth function in $L^2(\Sigma; D^1_{0\#}(G^K))$. Multiplying the equation (31) by $\varphi(x', \frac{x}{\epsilon})$ we obtain

$$\frac{1}{\epsilon}\int_\Omega \chi_{\Omega_\epsilon}(x)\nabla u_\epsilon \cdot \nabla_y \varphi\left(x', \frac{x}{\epsilon}\right) dx + \int_\Omega \chi_{\Omega_\epsilon}(x)\nabla u_\epsilon \nabla_{x'}\varphi\left(x', \frac{x}{\epsilon}\right) dx$$

$$= \sum_{1\le p\le n(\epsilon)} \left(P^K_\epsilon \vec{s}\right)_p \cdot \int_{\Gamma^\epsilon_p} \varphi\left(x', \frac{x}{\epsilon}\right)\vec{n} ds$$

$$= \frac{1}{\epsilon}\int_\Omega \left(\chi_{\Omega_\epsilon}(x) - 1\right)\vec{a}^\epsilon(x) \cdot \left(\nabla_y\varphi\left(x', \frac{x}{\epsilon}\right) + \epsilon\nabla_{x'}\varphi\left(x', \frac{x}{\epsilon}\right)\right) dx,$$

where $\chi_{\Omega_\epsilon}(x)$ is the periodic characteristic function of Ω_ϵ and $\vec{a}^\epsilon(x)$ is a piecewise constant function defined as in Lemma 2.16 by

$$\vec{a}^\epsilon = \sum_{\ell'}\sum_j \Big(\frac{1}{|\epsilon KY'|}\int_{(\epsilon KY')_{\ell'}} \vec{s}_j(x')dx'\Big)\chi_{Y^\epsilon_{j\ell'}}(x).$$

Remark that both terms involving $\nabla_{x'}\varphi$ go to zero with ϵ. Applying Lemma 2.16, we pass to the two-scale limit in the remaining terms to get

$$\frac{1}{|KY'|}\int_\Sigma\int_{G^{*K}} \nabla_y u_0(x', y)\cdot\nabla_y\varphi(x', y)dx'dy = \frac{-1}{|KY'|}\int_\Sigma\sum_j\int_{T_j}\vec{s}_j(x')\cdot\nabla_y\varphi(x', y)dx'dy$$

which is nothing but the variational formulation of the limit equation (32). A standard application of the Lax–Milgram lemma yields uniqueness of the solution u_0 in $L^2(\Sigma; D^1_{0\#}(G^K))$. Thus the entire sequence u_ϵ converges to the same limit u_0.

The proof of the energy convergence (33) is standard by passing to the two-scale limit in the right-hand side of (35) since $\vec{a}^\epsilon$ two-scale converges strongly in the sense of Proposition 2.5 (see Proposition 2.2.4 in [3]).

To prove the two-scale convergence of $u_\epsilon(\vec{s}^\epsilon)$ to u_0, when $\vec{s}^\epsilon$ converges weakly to $\vec{s}$ in $L^2(\Sigma; \ell^2_K)$, it suffices to repeat the same above arguments since Lemma 2.16 asserts that $\vec{a}^\epsilon$ two-scale converges to $\vec{a}^0$ even if $\vec{s}^\epsilon$ converges weakly. Note that in this case we do not have the energy convergence.

Proof of Theorem 2.11. Let $\vec{s}(x') \in L^2(\Sigma; \ell^2_K)$ and $\vec{t}^\epsilon$ be a sequence which converges weakly to $\vec{t}$ in $L^2(\Sigma; \ell^2_K)$. Our goal is to prove that

$$\lim_{\epsilon\to 0}\left\langle B^K_\epsilon \vec{s}(x'), \vec{t}^\epsilon(x')\right\rangle_{L^2(\Sigma;\ell^2_K)} = \left\langle B^K\vec{s}(x'), \vec{t}(x')\right\rangle_{L^2(\Sigma;\ell^2_K)}.$$

By definition of B_ϵ^K, we have

$$\begin{aligned}\left\langle B_\epsilon^K \vec{s}(x'), \vec{t^\epsilon}(x')\right\rangle_{L^2(\Sigma;\ell^2_K)} &= \left\langle E_\epsilon^K S_\epsilon P_\epsilon^K \vec{s}(x'), \vec{t^\epsilon}(x')\right\rangle_{L^2(\Sigma;\ell^2_K)} \\ &= (\epsilon K)^{N-1}\left\langle S_\epsilon P_\epsilon^K \vec{s}(x'), P_\epsilon^K \vec{t^\epsilon}(x')\right\rangle_{\mathbb{R}^{Nn(\epsilon)}} \\ &= (\epsilon K)^{N-1}\sum_{1\le p\le n(\epsilon)} \tfrac{1}{\epsilon^N}\left(\textstyle\int_{\Gamma_p^\epsilon} u_\epsilon(\vec{s})\vec{n}ds\right)\cdot (P_\epsilon^K \vec{t^\epsilon})_p \\ &= \tfrac{K^{N-1}}{\epsilon}\textstyle\int_{\Omega_\epsilon} \nabla u_\epsilon(\vec{s})\cdot\nabla u_\epsilon(\vec{t^\epsilon})dx.\end{aligned}$$

By Proposition 2.13 we know that $\nabla u_\epsilon(\vec{s})$ two-scale converges *strongly* in the sense of boundary layers to $\nabla_y u_0(\vec{s})$ while $\nabla u_\epsilon(\vec{t^\epsilon})$ two-scale converges weakly to $\nabla_y u_0(\vec{t})$. By virtue of Proposition 2.5 we can pass to the limit in the product and we get

$$\lim_{\epsilon\to 0}\left\langle B_\epsilon^K \vec{s}(x'), \vec{t^\epsilon}(x')\right\rangle_{L^2(\Sigma;\ell^2_K)} = \int_\Sigma\int_{G^{*K}} \nabla_y u_0(\vec{s})\cdot\nabla_y u_0(\vec{t})dx'dy,$$

where $u_0(\vec{s})$ and $u_0(\vec{t})$ are solutions of the homogenized problem (32) with $\vec{s}$ and $\vec{t}$, respectively, as the right-hand side. A simple integration by parts shows that

$$\int_\Sigma\int_{G^{*K}} \nabla_y u_0(\vec{s})\cdot\nabla_y u_0(\vec{t})dx'dy \;=\; \left\langle B^K \vec{s}(x'), \vec{t}(x')\right\rangle_{L^2(\Sigma;\ell^2_K)},$$

where the limit operator B^K is defined by

$$B^K\vec{s}(x') = \left(\int_{\Gamma_j} u_0(\vec{s})\vec{n}ds\right)_{\substack{0\le j'\le K-1\\ 1\le j_N}}. \tag{36}$$

This proves the strong convergence of B_ϵ^K to B^K on $L^2(\Sigma;\ell^2_K)$. Obviously, B^K is self-adjoint and continuous but not compact since x' plays the role of a parameter in the definition of B^K.

It remains to diagonalize B^K with the help of the Bloch decomposition $\mathcal{B}'$. This diagonalization process has already been exposed in section 3.3 of our previous paper [3] in a slightly different context. For the sake of brevity, we do not repeat this standard argument here. Let us simply indicate the three main steps of this Bloch diagonalization. First, we apply the operator $\mathcal{B}'$ to $\vec{s}(x') = (\vec{s}_j(x'))_{\substack{0\le j'\le K-1\\ j_N\ge 1}}$ which gives the Bloch decomposition of $\vec{s}(x')$ with respect to the multi-index j' (not including j_N). Secondly, plugging this Bloch decomposition in the limit equation (32) (which holds in G^{*K}) and using a similar Bloch decomposition of $u_0(\vec{s})$, we decompose (32) in a family of K^{N-1} equations defined in a single reference cell G^*. In a third step, applying again the Bloch decomposition $\mathcal{B}'$ to formula (36) yields the desired diagonalization of B^K.

2.3. Analysis of the limit spectrum. In this section we analyze the spectrum of the limit operator B^K and, from the strong convergence of B_ϵ^K to B^K, we deduce the lower semicontinuous convergence of the spectrum $\sigma(S_\epsilon)$ to the limit spectrum $\sigma(B^K)$. Recall that for any $K\ge 1$, the extended operator B_ϵ^K has a spectrum given by

$$\sigma(B_\epsilon^K) = \sigma(S_\epsilon)\cup\{0\}.$$

Since B_ϵ^K converges strongly to B^K in $L^2(\Sigma;\ell_K^2)$, by virtue of Proposition 2.1.11 in [3], we have

$$\sigma(B^K) \subset \sigma_\infty = \lim_{\epsilon\to 0} \sigma(S_\epsilon).$$

From Rellich's theorem, the strong convergence of the spectral family associated with B_ϵ^K to that of B^K is also easily deduced (see Theorem 3.2.5 in [3]). This gives some (partial) information on the convergence of eigenvectors that we shall not use below.

In view of Theorem 2.11,

$$B^K = (\mathcal{B}')^{-1} D^K \mathcal{B}' \quad \text{with} \quad D^K = \operatorname{diag}(D_{j'}^K)_{0\le j'\le K-1},$$

where each $D_{j'}^K$ is a self-adjoint continuous operator in $L^2(\Sigma;\ell_1^2)$. Since $\mathcal{B}'$ is an isometry, we have

$$\sigma(B^K) = \bigcup_{0\le j'\le K-1} \sigma(D_{j'}^K).$$

By the very definition of $D_{j'}^K$, the macroscopic variable $x' \in \Sigma$ plays the role of a parameter. Therefore, for any fixed value of x', $D_{j'}^K$ can be identified with an operator $d_{\frac{j'}{K}}$ acting in ℓ_1^2 which does not depend on x'. Introducing the Bloch parameter $\theta' = \frac{j'}{K} \in [0,1]^{N-1}$, this new operator $d_{\theta'}$ is defined by

$$\begin{aligned} &d_{\theta'} : \ell_1^2 \longrightarrow \ell_1^2, \\ &(\vec{s}_q)_{q\ge 1} \mapsto \left(\int_{\Gamma_q} u_{\theta'} \cdot \vec{n} ds \right)_{q\ge 1}, \end{aligned} \tag{37}$$

where $u_{\theta'}(y)$ is the unique solution of

$$\begin{cases} -\Delta u_{\theta'} = 0 & \text{in } G^*, \\ \frac{\partial u_{\theta'}}{\partial n} = \vec{s}_q \cdot \vec{n} & \text{on } \Gamma_q,\ q \ge 1, \\ u_{\theta'} = 0 & \text{if } y_N = 0, \\ y' \mapsto e^{-2\pi i \theta'\cdot y'} u_{\theta'}(y', y_N) & Y'\text{-periodic.} \end{cases}$$

In (37) the positive integer q is nothing but the index j_N introduced in Definition 2.8. Clearly, we have

$$\sigma(D_{j'}^K) = \sigma(d_{\frac{j'}{K}}).$$

As is well known, the spectrum of a self-adjoint operator can be decomposed in its *discrete* part, made of, at most, a countable number of isolated eigenvalues of finite multiplicities, and its *essential* part, for which the Weyl criterion applies (see, e.g., [25], [33], [34]). The next proposition characterizes the spectrum of $d_{\theta'}$.

PROPOSITION 2.17. *For all $\theta' \in [0,1]^{N-1}$, $d_{\theta'}$ is a self-adjoint continuous but noncompact operator in ℓ_1^2. Labeling the eigenvalues of the discrete spectrum $\sigma_{disc}(d_{\theta'})$ by decreasing order, each discrete eigenvalue is piecewise continuous in θ'. The essential spectrum is given by*

$$\sigma_{ess}(d_{\theta'}) = \bigcup_{\theta_N\in[0,1]} \sigma(A(\theta', \theta_N)),$$

where $A(\theta', \theta_N)$ is the Bloch homogenized matrix, defined by (12)*, which is continuous in $\theta \in]0,1[^N$ but discontinuous at $\theta = 0$. Moreover, the entire spectrum $\sigma(d_{\theta'})$, considered as a subset of $\mathbb{R}^+$, depends continuously on θ', except at $\theta' = 0$.*

Because we use the usual convenient labeling of the discrete eigenvalues by decreasing order, we can merely prove that they are piecewise continuous. This is due to the fact that, when θ' varies, an analytical branch (if any) of discrete eigenvalues may merge into the essential spectrum: this yields a "jump" in the labeling of discrete eigenvalues. Therefore, one cannot hope to prove a global continuity of these eigenvalues with such an ordering.

Let us postpone for a moment the proof of Proposition 2.17 and define the so-called *boundary layer spectrum* associated with the surface Σ:

$$\sigma_\Sigma \stackrel{\text{def}}{=} \bigcup_{\theta' \in]0,1[^{N-1}} \sigma(d_{\theta'}) \cup \sigma(d_0). \tag{38}$$

By virtue of Proposition 2.17, we have

$$\sigma_{Bloch} \subset \sigma_\Sigma. \tag{39}$$

Therefore σ_Σ also has a band structure since it includes the Bloch spectrum, but it may include new bands of eigenvalues of $\sigma_{disc}(d_{\theta'})$. It also contains the isolated eigenvalues of $\sigma_{disc}(d_0)$. Therefore σ_Σ can contain elements which are not included in the previous limit spectrum $\sigma(S) \cup \sigma_{Bloch}$ (see section 1.2). The continuity of $\sigma(d_{\theta'})$ with respect to θ' ensures that σ_Σ is the closure of the union of all spectra $\sigma(d_{\theta'})$ with θ' rational.

$$\overline{\bigcup_{K \geq 1} \bigcup_{0 \leq j' \leq K-1} \sigma(d_{\frac{j'}{K}})} = \sigma_\Sigma.$$

We summarize our results in the following theorem.

THEOREM 2.18. *The boundary layer spectrum associated to Σ is included in the limit spectrum*

$$\sigma_\Sigma \subset \sigma_\infty.$$

REMARK 2.19. *Of course σ_Σ is not the complete boundary layer spectrum since it is concerned only with that part of the spectrum concentrating near Σ. A completely similar analysis has to be done for all the $(N-1)$-dimensional surfaces and all other lower dimensional manifolds (edges, corners, etc.) of which the boundary of Ω is made up. Then, we shall prove in the next section that the union of all these contributions, the so-called boundary layer spectrum, plus the usual homogenized spectrum and the Bloch spectrum, is equal to σ_∞, at least when Ω is made up only of entire cells ϵY.*

Proof of Proposition 2.17. Let us first prove that the essential spectrum of $d_{\theta'}$ is included in the Bloch spectrum, and, more precisely,

$$\sigma_{ess}(d_{\theta'}) = \bigcup_{0 \leq \theta_N \leq 1} \sigma(A(\theta', \theta_N)),$$

where $A(\theta)$ is the usual Bloch homogenized matrix defined in (12). In particular, this proves that $\sigma_{ess}(d_{\theta'}) \neq \{0\}$, so $d_{\theta'}$ is not compact.

Let $\lambda(\theta)$ be an eigenvalue of $A(\theta)$ and $u(\theta)$ be the associated potential solution of

$$\begin{cases} -\Delta_y u(\theta) = 0 & \text{in } Y^*, \\ \frac{\partial u(\theta)}{\partial n} = \lambda^{-1}(\theta) \int\limits_{\Gamma} u(\theta)\vec{n} ds & \text{on } \Gamma, \\ y \mapsto e^{-2\pi\imath\theta\cdot y} u(\theta, y) & Y\text{-periodic.} \end{cases}$$

We construct a Weyl sequence u_n associated with the spectral value $\lambda(\theta)$ by

$$u_n = \frac{u(\theta)\psi_n}{\|u(\theta)\psi_n\|_{L^2(G^*)}},$$

where $\psi_n(y_N)$ is a cut-off function defined by

$$\begin{cases} \psi_n(y_N) = & y_N & \text{when } 0 \le y_N \le 1, \\ \psi_n(y_N) = & 1 & \text{when } 1 \le y_N \le n, \\ \psi_n(y_N) = & n+1-y_N & \text{when } n \le y_N \le n+1, \\ \psi_n(y_N) = & 0 & \text{when } y_N \ge n+1. \end{cases}$$

By definition, $\|u_n\|_{L^2(G^*)} = 1$ and $\lim_{n\to+\infty} \|u(\theta)\psi_n\|_{L^2(G^*)} = +\infty$. Then, it is easily checked that, for any $\varphi \in D^1_{0\#}(G)$ (the Deny–Lions-type space defined in (21)),

$$\int_{G^*} \nabla u_n \cdot \nabla\varphi dy = \frac{1}{\lambda(\theta)} \sum_{q\ge 1} \Big(\int_{\Gamma_q} u_n \vec{n} ds \Big) \cdot \Big(\int_{\Gamma_q} \varphi \vec{n} ds \Big) + \langle r_n, \varphi \rangle ,$$

where r_n is a negligible remainder term in the sense that

$$\lim_{n\to+\infty} \frac{\langle r_n, \varphi\rangle}{\|\nabla\varphi\|_{L^2(G^*)^N}} = 0.$$

Furthermore, $\vec{s}_n = (\int_{\Gamma_q} u_n \vec{n} ds)_{q\ge 1}$ converges weakly to 0 in ℓ^2_1 since

$$\lim_{n\to+\infty} \|u(\theta)\psi_n\|_{L^2(G^*)} = +\infty.$$

Therefore, $\vec{s}_n$ is a Weyl sequence associated with $\lambda(\theta)$ for the operator $d_{\theta'}$. This proves that $\lambda(\theta) \in \sigma_{ess}(d_{\theta'})$. To prove the converse inclusion,

$$\sigma_{ess}(d_{\theta'}) \subset \bigcup_{0\le\theta_N\le 1} \sigma(A(\theta', \theta_N)),$$

we consider a Weyl sequence $\vec{s}_n$ for a spectral value $\lambda \in \sigma_{ess}(d_{\theta'})$. Let u_n be the associated potential solution, i.e.,

$$\begin{cases} -\Delta u_n = 0 & \text{in } G^*, \\ \frac{\partial u_n}{\partial n} = (\vec{s}_n)_q \cdot \vec{n} & \text{on } \Gamma_q,\ q \ge 0, \\ u_n = 0 & \text{if } y_N = 0, \\ y' \mapsto e^{-2\pi\imath\theta'\cdot y'} u_n(y', y_N) & Y'\text{-periodic.} \end{cases} \tag{40}$$

Since $\|\vec{s}_n\|_{\ell^2_1} = 1$ and $\vec{s}_n \rightharpoonup 0$ in ℓ^2_1 weakly, it is easily seen that u_n converges to 0 weakly in $H^1(G^*)$. Furthermore, since the weak convergence to 0 of $\vec{s}_n$ implies that its components $(\vec{s}_n)_q$ go to 0 for fixed q, it is not difficult to check that, for any compact set $\mathcal{K}$ of G^*, u_n converges strongly to 0 in $H^1(\mathcal{K})$ (multiply equation (40)

by ϕu_n where ϕ is equal to 1 in $\mathcal{K}$ and is compactly supported away from infinity). Introducing a sequence

$$v_n = \frac{\psi u_n}{\|\psi u_n\|_{L^2(G^*)}},$$

where $\psi(y_N)$ is a cut-off function defined by

$$\begin{cases} \psi(y_N) = & 0 & \text{for } y_N \leq 0, \\ \psi(y_N) = & y_N & \text{for } 0 \leq y_N \leq 1, \\ \psi(y_N) = & 1 & \text{for } y_N \geq 1, \end{cases}$$

it is straightforward to prove that

$$\int_{B^*} \nabla v_n \cdot \nabla \varphi dx = \frac{1}{\lambda} \sum_{q \in \mathbb{Z}} \Big(\int_{\Gamma_q} v_n \vec{n} ds \Big) \cdot \Big(\int_{\Gamma_q} \varphi \vec{n} ds \Big) + \langle r_n, \varphi \rangle$$

for any $\varphi \in D^1_{\#}(B^*)$, where B^* is the infinite band $Y' \times]-\infty; +\infty[$ perforated by the periodic arrangement of tubes $(T_q)_{q \in \mathbb{Z}}$, and r_n is another negligible remainder term such that

$$\lim_{n \to +\infty} \frac{\langle r_n, \varphi \rangle}{\|\nabla \varphi\|_{L^2(B^*)^N}} = 0.$$

Therefore,

$$\vec{t}_n = \left(\int_{\Gamma_q} v_n \vec{n} ds \right)_{q \in \mathbb{Z}}$$

is a Weyl sequence for an operator similar to $d_{\theta'}$ but defined in the whole infinite band B^* instead of the semi-infinite band G^*. A standard Bloch decomposition with respect to the variable y_N yields that λ belongs to $\bigcup_{0 \leq \theta_N \leq 1} \sigma(A(\theta', \theta_N))$.

To conclude the proof of Proposition 2.17, it remains to prove that the isolated eigenvalues of finite multiplicity $\lambda(\theta') \in \sigma_{disc}(d_{\theta'})$ are piecewise continuous with respect to θ'. Let θ'_n be a sequence converging to θ' in $]0,1[^{N-1}$. Obviously, the sequence of continuous operators $d_{\theta'_n}$ uniformly converges to $d_{\theta'}$ in ℓ^2_1. Now, let us invoke a classical theorem (see, e.g., Theorem 3.1., Chapter I.3 in [20]) which states that for any closed curve γ in the complex plane, which encloses a finite number of eigenvalues of $\sigma_{disc}(d_{\theta'})$ and does not intersect $\sigma(d_{\theta'})$, there exists n_0 such that for any $n \geq n_0$, the curve γ contains the same number of eigenvalues (including multiplicities) of $\sigma_{disc}(d_{\theta'_n})$ and does not intersect $\sigma(d_{\theta'_n})$. This is nothing but the local continuity of the eigenvalues of $\sigma_{disc}(d_{\theta'})$ (enumerated, for example, in decreasing order). Remark that the continuity of the pth eigenvalue of $\sigma_{disc}(d_{\theta'})$ breaks down only when one of the previous eigenvalues (with label between 1 and $p-1$) meets the essential spectrum $\sigma_{ess}(d_{\theta'})$ as θ' varies. In any case, since $\sigma_{ess}(d_{\theta'})$ depends continuously on $\theta' \neq 0$, this proves that the entire spectrum $\sigma(d_{\theta'})$ depends also continuously on $\theta' \neq 0$. The lack of continuity for $\sigma(d_{\theta'})$ at $\theta' = 0$ is a phenomenon already explained in our previous work (see Proposition 3.3.4 in [3]).

REMARK 2.20. *When the tube T is symmetric in Y (in other words, by reflexion with respect to the hyperplane $[y_N = 0]$, G^* yields the infinite periodic array of tubes B^*), it can readily be checked that there is no isolated eigenvalue of finite multiplicity*

for $d_{\theta'}$; i.e., $\sigma_{disc}(d_{\theta'}) = \emptyset$ for all $\theta' \in [0,1]^{N-1}$. If this were not the case, by symmetry an eigenvalue of $\sigma_{disc}(d_{\theta'})$ would also be an eigenvalue of finite multiplicity for a similar operator in the infinite band B^*, which is impossible since by translation there exists an infinite number of eigenvectors.

We conclude this section by proving that the eigenvectors corresponding to isolated eigenvalues of finite multiplicity of $d_{\theta'}$ are localized in the vicinity of the boundary $[y_N = 0]$ since they decay exponentially at infinity.

PROPOSITION 2.21. *Let λ be an eigenvalue in $\sigma_{disc}(d_{\theta'})$ and let $(\vec{s}_q)_{q\geq 1}$ be a corresponding eigenvector. There exists a positive constant $\alpha > 0$ such that $(e^{\alpha p}\vec{s}_q)_{q\geq 1}$ belongs to ℓ^2_1.*

Proof. The argument is by contradiction of the Weyl property for eigenvalues in the essential spectrum. For $\lambda \in \sigma_{disc}(d_{\theta'})$, let $\vec{s} = (\vec{s}_q)_{q\geq 1}$ be a corresponding normalized eigenvector and $u(y)$ the corresponding potential, solution of

$$\left\{\begin{array}{ll} -\Delta u = 0 & \text{in } G^*, \\ \frac{\partial u}{\partial n} = \vec{s}_q \cdot \vec{n} & \text{on } \Gamma_q,\ q \geq 1, \\ u = 0 & \text{if } y_N = 0, \\ y' \mapsto e^{-2\pi\imath\theta'\cdot y'} u(y', y_N) & Y'\text{-periodic.} \end{array}\right. \tag{41}$$

By definition, for all $q \geq 1$, it satisfies

$$\int_{\Gamma_q} u \cdot \vec{n} ds \ = \ \lambda \vec{s}_q.$$

Let us define a sequence $(\vec{s}^{\,n})_{n\geq 0}$ in ℓ^2_1 by

$$\vec{s}^{\,n} = (\vec{s}^{\,n}_q)_{q\geq 1} \text{ with } \vec{s}^{\,n}_q = \left\{\begin{array}{ll} 0 & \text{if } q < n, \\ \frac{\vec{s}_q}{\sqrt{\sum_{p=n}^{\infty} |\vec{s}_p|^2}} & \text{if } q \geq n. \end{array}\right.$$

It is easily seen that $\vec{s}^{\,n}$ converges weakly to 0 in ℓ^2_1 with $\|\vec{s}^{\,n}\|_{\ell^2_1} = 1$. However, since λ does not belong to the essential spectrum of $d_{\theta'}$, any subsequence of $\vec{s}^{\,n}$ cannot be a Weyl sequence for λ. This implies the existence of a positive constant C and an integer n_0 such that, for any $n \geq n_0$,

$$\|d_{\theta'}\vec{s}^{\,n} - \lambda \vec{s}^{\,n}\|_{\ell^2_1} \geq C > 0. \tag{42}$$

As usual $u_n(y)$ is the potential associated with $\vec{s}^{\,n}$ through an equation similar to (41). We introduce a smooth cut-off function $\psi_n(y_N)$ such that $\psi_n = 0$ on all tubes T_q for $q < n$, and $\psi_n = 1$ on all tubes T_q for $q \geq n$. Let us denote by ω_n the bounded support of $\nabla\psi_n$ which lies between T_{n-1} and T_n. Introducing an approximation v_n of the potential u_n, defined by

$$v_n(y) = \frac{\psi_n(y_N)\,(u(y) - c_n)}{\sqrt{\sum_{p=n}^{\infty} |\vec{s}_p|^2}} \text{ with } c_n = \frac{1}{|\omega_n|}\int_{\omega_n} u(y)dy,$$

we write

$$d_{\theta'}\vec{s}^{\,n} \ = \ \lambda \vec{s}^{\,n} + \left(\int_{\Gamma_q} (u_n - v_n)\cdot \vec{n} ds\right)_{q\geq 1}.$$

From (42) we deduce

$$\|\nabla(u_n - v_n)\|_{L^2(G^*)^N} \geq C > 0 \text{ for } n \geq n_0.$$

Using the equations for u and u_n, a simple computation yields

$$\int_{G^*} |\nabla(u_n - v_n)|^2 dy = \int_{G^*} \nabla\psi_n \cdot \frac{(u_n - v_n)\nabla u - (u - c_n)\nabla(u_n - v_n)}{\sqrt{\sum_{p=n}^{\infty} |\vec{s}_p|^2}}. \tag{43}$$

Remark that the integral in the right-hand side reduces to ω_n since $\nabla\psi_n$ has bounded support in ω_n. Applying the Poincaré–Wirtinger inequality in ω_n to $(u-c_n)$ and $(u_n - v_n)$ (this last term has not zero average in ω_n, but (43) is invariant by substraction of a constant to $(u_n - v_n)$), we obtain from (43)

$$\|\nabla(u_n - v_n)\|_{L^2(G^*)^N} \leq C \frac{\|\nabla u\|_{L^2(\omega_n)^N}}{\sqrt{\sum_{p=n}^{\infty} |\vec{s}_p|^2}},$$

which implies

$$\sum_{p=n}^{\infty} |\vec{s}_p|^2 \leq C \|\nabla u\|^2_{L^2(\omega_n)^N}. \tag{44}$$

On the other hand, multiplying equation (41) by $\psi_n(u - c_n)$ and integrating by parts gives

$$\int_{G^*} \psi_n |\nabla u|^2 dy + \int_{G^*} (u - c_n)\nabla u \cdot \nabla\psi_n dy = \lambda \sum_{p=n}^{\infty} |\vec{s}_p|^2.$$

Applying again the Poincaré–Wirtinger inequality in ω_n to $(u - c_n)$ yields

$$\int_{G^*} \psi_n |\nabla u|^2 dy \leq \lambda \sum_{p=n}^{\infty} |\vec{s}_p|^2 + C \|\nabla u\|^2_{L^2(\omega_n)^N}. \tag{45}$$

Let us denote by G_n the subset of G^* defined by $G_n = \{y \in G^* | y_N > n\}$. From (44) and (45) we deduce

$$\|\nabla u\|^2_{L^2(G_{n+1})^N} \leq C \|\nabla u\|^2_{L^2(\omega_n)^N} \leq C \left(\|\nabla u\|^2_{L^2(G_n)^N} - \|\nabla u\|^2_{L^2(G_{n+1})^N} \right),$$

which implies, for $n \geq n_0$,

$$\|\nabla u\|^2_{L^2(G_n)^N} \leq \left(\frac{C}{1 + C} \right)^{n - n_0} \|\nabla u\|^2_{L^2(G_{n_0})^N}. \tag{46}$$

It is easily seen that (46) implies the desired result.

3. Completeness of the boundary layer spectrum. In this section we assume that Ω is a rectangle with integer dimensions, i.e.,

$$\Omega = \prod_{i=1}^{N}]0; L_i[\quad \text{and} \quad L_i \in \mathbb{N}^*. \tag{47}$$

The sequence of small parameters ϵ is also assumed to be

$$\epsilon_n \;=\; \frac{1}{n},\; n \in \mathbb{N}^*. \tag{48}$$

Remark that all the previous results in this paper hold for any type of sequence ϵ going to zero. From now on, we restrict ourselves to the sequence ϵ_n since, for any $n \geq 1$, the domain Ω is the union of a finite number of *entire periodic cells* $Y_p^{\epsilon_n}$. However, to simplify the notation, we shall not indicate the dependence on n and simply denote by ϵ the particular sequence defined in (48).

Remark that the assumption on the geometry of Ω can be slightly relaxed. Any polygonal domain with faces parallel to the axis (i.e., the normal is everywhere one of the basis vectors) and having vertex with integer coordinates could equally be considered.

3.1. Presentation of the main result. This section is devoted to the so-called completeness of the limit spectrum. Recall that in our previous work [3] we proved that

$$\sigma_\infty \;=\; \sigma(S) \cup \sigma_{Bloch} \cup \sigma_{boundary}, \tag{49}$$

where $\sigma_{boundary}$ is defined in (15). In section 2, we proved that

$$\sigma_\infty \;\supset\; \sigma_\Sigma,$$

where σ_Σ is the boundary layer spectrum associated with the surface Σ, defined by (38). Remark that, *due to our hypotheses on the domain* Ω *and on the sequence* ϵ, the surface Σ can be any of the faces of Ω defined by

$$\prod_{\substack{j=1\\ j\neq i}}^{N}]0; L_j[\times\{0\} \quad \text{or} \quad \prod_{\substack{j=1\\ j\neq i}}^{N}]0; L_j[\times\{L_i\} \quad \text{for } 1 \leq i \leq N.$$

Of course, the analysis of section 2 can be repeated for any other lower dimensional manifolds (edges, corners, etc.) which compose the boundary of Ω. For $0 \leq m \leq N-1$, let us define the m-dimensional parts of $\partial\Omega$ as

$$\Sigma_{m,\tau} = \prod_{j=1}^{m}]0; L_{\tau(j)}[\times \prod_{j=m+1}^{N} \{x_{\tau(j)} = 0 \text{ or } L_{\tau(j)}\},$$

where τ is any permutation of the numbers $\{1, 2, \ldots, N\}$. There are $2^{N-m}C_N^{N-m}$ m-dimensional manifolds of the type $\Sigma_{m,\tau}$. A simple adaptation of the two-scale convergence in the sense of boundary layers for such manifolds allows us to prove that, for any m and τ,

$$\sigma_\infty \;\supset\; \sigma_{\Sigma_{m,\tau}},$$

where $\sigma_{\Sigma_{m,\tau}}$ is the spectrum of a family of limit problems posed, not in a semi-infinite band as in section 2, but rather in a periodic domain bounded in the variables $x_{\tau(1)}, \ldots, x_{\tau(m)}$ and unbounded with respect to the other variables (see section 3.3 for the case of corners in two space dimension). Eventually, defining the union of all these spectra

$$\sigma_{\partial\Omega} = \bigcup_{m,\tau} \sigma_{\Sigma_{m,\tau}}, \tag{50}$$

we deduce from Theorem 2.18 and from the geometric assumptions (47), (48) that

$$\sigma_\infty \supset \sigma_{\partial\Omega}. \tag{51}$$

Comparing our results (49) and (51), a completeness result amounts to link the two definitions of the boundary layer spectrum $\sigma_{\partial\Omega}$ and $\sigma_{boundary}$.

THEOREM 3.1. *For the sequence ϵ_n defined by* (48)*, the boundary layer spectrum satisfies*

$$\sigma_{boundary} \subset \sigma_{\partial\Omega}.$$

Therefore, the limit spectrum of the sequence S_{ϵ_n} is precisely made of three parts; the homogenized, the Bloch, and the boundary layer spectrum

$$\lim_{\epsilon_n \to 0} \sigma(S_{\epsilon_n}) = \sigma(S) \cup \sigma_{Bloch} \cup \sigma_{\partial\Omega},$$

where the boundary layer spectrum $\sigma_{\partial\Omega}$ is explicitly defined by (50).

REMARK 3.2. *Remark that Theorem* 3.1 *does not state that $\sigma_{boundary}$, defined by* (15)*, and $\sigma_{\partial\Omega}$ coincide. Indeed, we have shown in* (39) *that $\sigma_{\partial\Omega}$ contains the Bloch spectrum. It is not clear whether $\sigma_{boundary}$ contains the Bloch spectrum too. The comparison of $\sigma_{\partial\Omega}$ and $\sigma_{boundary}$ is definitely a very difficult question. We suspect that if the definition of $\sigma_{boundary}$ is modified in such a way that it contains only limit eigenvalues corresponding to sequences of eigenvectors which decay exponentially fast away from the boundary, then it may coincide with that part of $\sigma_{\partial\Omega}$ made of discrete eigenvalues (which also have exponentially decreasing corresponding eigenvectors).*

To prove this completeness result, we need an intermediate result in the spirit of section 2.

THEOREM 3.3. *As in section* 2*, let Ω be a domain defined by*

$$\Omega = \Sigma \times]0; L[,$$

with Σ a bounded open set in $\mathbb{R}^{N-1}$ and $L > 0$. Recall that S^1_ϵ is the extension of S_ϵ to $L^2(\Omega)^N$. Consider a sequence of eigenvalues λ_ϵ and eigenvectors $\vec{s}^\epsilon$ such that

$$S^1_\epsilon \vec{s}^\epsilon = \lambda_\epsilon \vec{s}^\epsilon \quad \textit{with} \quad \|\vec{s}^\epsilon\|_{L^2(\Omega)^N} = 1 \quad \textit{and} \quad \lim_{\epsilon \to 0} \lambda_\epsilon = \lambda.$$

Assume that for all subset ω such that $\overline{\omega} \subset \Omega$, we have

$$\lim_{\epsilon \to 0} \|\vec{s}^\epsilon\|_{L^2(\omega)^N} = 0. \tag{52}$$

Assume further that there exists an $(N-1)$-dimensional open set σ, with $\overline{\sigma} \subset \Sigma$, a positive number l, with $0 < l < L$, and a positive constant c such that

$$\lim_{\epsilon \to 0} \|\vec{s}^\epsilon\|_{L^2(\sigma \times]0,l[)^N} \geq c > 0. \tag{53}$$

Then, λ belongs to the boundary layer spectrum associated with the surface Σ

$$\lambda \in \sigma_\Sigma,$$

where σ_Σ is defined by (38).

The proof of Theorem 3.3 is the focus of the next section. If we admit it for the moment, as well as its generalizations concerning all other manifolds $\Sigma_{m,\tau}$ making up the boundary $\partial\Omega$, we are in a position to complete the following proof.

Proof of Theorem 3.1. Let $\lambda \in \sigma_{boundary}$. By definition there exists a subsequence (still denoted by ϵ), eigenvalues λ_ϵ, and eigenvectors $\vec{s}^\epsilon$ of S^1_ϵ such that

$$S^1_\epsilon \vec{s}^\epsilon = \lambda_\epsilon \vec{s}^\epsilon \quad \text{with} \quad \|\vec{s}^\epsilon\|_{L^2(\Omega)^N} = 1 \quad \text{and} \quad \lim_{\epsilon\to 0} \lambda_\epsilon = \lambda,$$

and, for all subset ω satisfying $\overline{\omega} \subset \Omega$,

$$\lim_{\epsilon\to 0} \|\vec{s}^\epsilon\|_{L^2(\omega)^N} = 0.$$

If there exists an $(N-1)$-dimensional open subset σ_i, compactly embedded in $\prod_{\substack{j=1\\ j\neq i}}^N]0; L_j[$, a positive length $0 < l_i < L_i$, a positive constant c, and another subsequence (still denoted by ϵ) such that

$$\lim_{\epsilon\to 0} \|\vec{s}^\epsilon\|_{L^2(\sigma_i\times]0,l_i[)^N} \geq c > 0 \quad \text{or} \quad \lim_{\epsilon\to 0} \|\vec{s}^\epsilon\|_{L^2(\sigma_i\times]l_i,L_i[)^N} \geq c > 0, \tag{54}$$

then, by application of Theorem 3.3, the limit eigenvalue belongs to $\sigma_{\partial\Omega}$ as desired.

If (54) does not hold true for any such σ_i, l_i, c, and subsequence ϵ, it implies that the L^2-norm of $\vec{s}^\epsilon$ concentrates near the lower dimensional edges of the rectangle Ω. In this case, we repeat the above argument with an $(N-2)$-dimensional open set included in one of the set $\Sigma_{N-2,\tau}$, and so on up to the 0-dimensional set made of one of the vertices of Ω. A tedious but simple induction argument on the dimension m shows that there exists at least a dimension $0 \leq m \leq N-1$, a permutation τ, positive lengths $(l_{\tau(j)})_{m+1\leq j\leq N}$, a positive constant c, and a subsequence ϵ such that

$$\lim_{\epsilon\to 0} \|\vec{s}^\epsilon\|_{L^2(\omega)^N} \geq c > 0,$$

with $\omega \subset \Omega$ of the type

$$\omega = \sigma \times \prod_{j=m+1}^N \big(]0, l_{\tau(j)}[\text{ or }]l_{\tau(j)}, L_{\tau(j)}[\big) \quad \text{and} \quad \overline{\sigma} \subset \prod_{j=1}^m]0; L_{\tau(j)}[.$$

Then, applying an adequate generalization of Theorem 3.3, this proves that the limit eigenvalue belongs to $\sigma_{\partial\Omega}$.

3.2. Proof of the completeness. This section is devoted to the proof of Theorem 3.3 which is divided in several lemmas and propositions. Let us begin by recalling the definition of the associated potential u_ϵ, solution of

$$\begin{cases} -\Delta u_\epsilon = 0 & \text{in } \Omega_\epsilon, \\ \frac{\partial u_\epsilon}{\partial n} = \vec{s}^\epsilon_p \cdot \vec{n} & \text{on } \Gamma^\epsilon_p, \quad 1 \leq p \leq n(\epsilon), \\ u_\epsilon = 0 & \text{on } \partial\Omega. \end{cases} \tag{55}$$

The spectral equation $\widetilde{S}_\epsilon \vec{s}^\epsilon = \lambda_\epsilon \vec{s}^\epsilon$ implies that

$$\left(\int_{\Gamma^\epsilon_p} u_\epsilon \vec{n} \right)_{1\leq p\leq n(\epsilon)} = \lambda_\epsilon \vec{s}^\epsilon_p. \tag{56}$$

By assumption (52), for all subsets ω such that $\overline{\omega} \subset \Omega$, we have

$$\lim_{\epsilon\to 0} \|\vec{s}^\epsilon\|_{L^2(\omega)^N} = 0.$$

In other words, all the energy of the eigenvectors $\vec{s}^\epsilon$ concentrates near the boundary $\partial\Omega$. This concentration effect has important consequences on the associated potential u_ϵ.

LEMMA 3.4. *The sequence u_ϵ defined in* (55) *converges to* 0 *in $H_0^1(\Omega)$ weakly and strongly in $L^2(\Omega)$. Furthermore, u_ϵ converges strongly to* 0 *in $H^1_{loc}(\Omega)$.*

Proof. Multiplying equation (55) by a test function $v \in H_0^1(\Omega)$ yields

$$\int_{\Omega_\epsilon} \nabla u_\epsilon \cdot \nabla v dx = \sum_{p=1}^{n(\epsilon)} \vec{s}_\epsilon^p \cdot \Big(\int_{\Gamma_p^\epsilon} v\vec{n} ds \Big) = \int_\Omega \vec{s}^\epsilon(x) \cdot \vec{z}^\epsilon(x) dx,$$

where

$$\vec{z}^\epsilon(x) = -\sum_{p=1}^{n(\epsilon)} \frac{1}{\epsilon^N} \Big(\int_{T_p^\epsilon} \nabla v(x) dx \Big) \chi_{Y_p^\epsilon}(x).$$

It is easily seen that $\vec{z}^\epsilon$ converges strongly to $-\frac{|T|}{|Y|}\nabla v(x)$ in $L^2(\Omega)^N$. Since $\vec{s}^\epsilon$ converges weakly to 0 in $L^2(\Omega)^N$ by virtue of (52), we deduce that u_ϵ converges to 0 weakly in $H_0^1(\Omega)$ and, by the Rellich theorem, strongly in $L^2(\Omega)$. Finally, for any open set ω such that $\overline{\omega} \subset \Omega$, let φ be a smooth function with compact support in Ω and equal to 1 on ω. Multiplying (55) by $\varphi^2 u_\epsilon$ and integrating by parts leads to

$$\int_{\Omega_\epsilon} \varphi^2 |\nabla u_\epsilon|^2 dx = -2 \int_{\Omega_\epsilon} \varphi u_\epsilon \nabla\varphi \cdot \nabla u_\epsilon dx + \sum_{p=1}^{n(\epsilon)} \vec{s}_p^\epsilon \cdot \Big(\int_{\Gamma_p^\epsilon} \varphi^2 u_\epsilon \vec{n} ds \Big). \tag{57}$$

Since u_ϵ converges weakly to 0 in $H_0^1(\Omega)$, the first term in the right-hand side of (57) goes to 0 with ϵ. In view of (56), the second term is bounded by

$$\|\varphi\|^2_{L^\infty(\Omega)} \|\vec{s}^\epsilon\|^2_{L^2(\mathrm{supp}(\varphi))},$$

which goes to 0 by virtue of the assumption (52). Therefore, we deduce from (57) that ∇u_ϵ converges strongly to 0 in $L^2(\omega)^N$. This concludes the proof of Lemma 3.4.

By assumption (53), there exists an $(N-1)$-dimensional open set σ, with $\overline{\sigma} \subset \Sigma$, such that the sequence of eigenvectors concentrates partly near σ. By translation, one can always assume that the origin lies inside σ. The strategy of the proof is to rescale the domain Ω by the change of variables $y = \frac{x}{\epsilon}$ and then to transform the sequence of eigenvectors $\vec{s}^\epsilon$ in a Weyl sequence for a limit operator. The limit domain will be $\mathbb{R}^N_+ = \{y \in \mathbb{R}^N | y_N > 0\}$ since we have carefully choose the origin to belong to σ. The limit fluid domain is denoted by $G^{*\infty}$, which is defined by

$$G^{*\infty} = \mathbb{R}^N_+ \setminus \bigcup_{j \in \mathbb{Z}^N_+} T_j,$$

where T_j denotes the tube j placed in the subcell Y_j (centered at the point $j = (j', j_N)$ with $j' \in \mathbb{Z}^{N-1}$ and $j_N \in \mathbb{Z}_+$). In this limit domain we define a limit operator B^∞, which acts from ℓ^∞_2 in itself, by

$$B^\infty \vec{s} = \left(\int_{\Gamma_j} u\vec{n} ds \right)_{j \in \mathbb{Z}^N_+} \qquad \forall \vec{s} \in \ell^\infty_2,$$

where $u(y)$ is the unique solution in $D_0^1(G^{*\infty})$ of

$$\left\{\begin{array}{ll} -\Delta u = 0 & \text{in } G^{*\infty}, \\ \frac{\partial u}{\partial n} = \vec{s}_j \cdot \vec{n} & \text{on } \Gamma_j,\ j \in \mathbb{Z}_+^N, \\ u = 0 & \text{on } \mathbb{R}^{N-1} \times \{0\}. \end{array}\right. \tag{58}$$

Recall that elements in $D_0^1(G^{*\infty})$ are restrictions to $G^{*\infty}$ of functions wnin $D_0^1(\mathbb{R}_+^N)$ which, in its turn, is the closure, with respect to the L^2-norm of the gradient, of smooth functions with compact support in $\mathbb{R}_+^N$.

REMARK 3.5. *The limit domain $G^{*\infty}$ is nothing but the limit as K goes to infinity of the domain G^{*K} defined in section* 2.2. *By the same token, the Hilbert space ℓ_2^∞ is the limit of ℓ_2^K (it is also equal to $\ell_2(\mathbb{Z}_+^N;\mathbb{C}^N)$). In some sense the limit operator B^∞ is also the limit of the operator B^K defined in Theorem* 2.11.

Let φ be a smooth function, equal to 1 in $\omega = \sigma\times]0, L[$, with compact support in $\Sigma\times]-L; L[$ (i.e., φ vanishes on all faces of Ω except on that defined by $x_N = 0$). We use φ to localize the sequence of eigenvectors $\vec{s}^\epsilon$ in a vicinity of ω. Let us define a sequence $\vec{t}^{\,\epsilon}$ by

$$\vec{t}^{\,\epsilon} = E_\epsilon^1 P_\epsilon^1(\varphi(x)\vec{s}^\epsilon(x)),$$

where $E_\epsilon^1 P_\epsilon^1$ is the projection operator in $L^2(\Omega)^N$ on piecewise constant functions (cf. their definitions (27) and (28)).

Remark that, by assumption (53), the sequence $\vec{t}^{\,\epsilon}$ satisfies

$$\lim_{\epsilon\to 0} \|\vec{t}^{\,\epsilon}\|_{L^2(\omega)^N} \geq c > 0.$$

Let us define $G_\epsilon^{*\infty}$ as $G^{*\infty}$ rescaled to size ϵ. Let v_ϵ be the potential in $G_\epsilon^{*\infty}$ associated with $\vec{t}^{\,\epsilon}$, defined by

$$\left\{\begin{array}{ll} -\Delta v_\epsilon = 0 & \text{in } G_\epsilon^{*\infty}, \\ \frac{\partial v_\epsilon}{\partial n} = \vec{t}_p^{\,\epsilon} \cdot \vec{n} & \text{on } \Gamma_p^\epsilon,\ p \in \mathbb{Z}_+^N, \\ v_\epsilon = 0 & \text{on } \mathbb{R}^{N-1} \times \{0\}. \end{array}\right. \tag{59}$$

LEMMA 3.6. *The sequence v_ϵ defined by* (59) *converges to zero in $D_0^1(\mathbb{R}_+^N)$ weakly and in $H_{loc}^1(\mathbb{R}_+^N)$ strongly.*

Proof. The argument is similar to that of Lemma 3.4, except that the Rellich theorem applies only for compact sets in $\overline{\mathbb{R}_+^N}$.

LEMMA 3.7. *The difference $w_\epsilon = v_\epsilon - \varphi u_\epsilon$ converges strongly to zero in $D_0^1(\mathbb{R}_+^N)$.*

Proof. A simple calculation provides the following key identity:

$$\int_{\mathbb{R}_+^N} |\nabla w_\epsilon|^2 = \int_{\mathbb{R}_+^N} \nabla v_\epsilon \cdot \nabla w_\epsilon - \int_{\mathbb{R}_+^N} \nabla u_\epsilon \cdot \nabla(\varphi w_\epsilon) - \int_{\mathbb{R}_+^N} \nabla\varphi \cdot (u_\epsilon \nabla w_\epsilon - w_\epsilon \nabla u_\epsilon). \tag{60}$$

By virtue of Lemmas 3.4 and 3.6, u_ϵ and w_ϵ converge to zero strongly in L^2 of the support of φ. Therefore, the last term in (60) goes to zero with ϵ. On the other hand, an integration by parts yields

$$\int_{G_{\epsilon\infty}^*} \nabla v_\epsilon \cdot \nabla w_\epsilon - \int_{\Omega_\epsilon} \nabla u_\epsilon \cdot \nabla(\varphi w_\epsilon) = \sum_{p=1}^{n(\epsilon)} \left[\vec{t}_p^{\,\epsilon} \cdot \Big(\int_{\Gamma_p^\epsilon} w_\epsilon \vec{n} \Big) - \vec{s}_p^{\,\epsilon} \cdot \Big(\int_{\Gamma_p^\epsilon} \varphi w_\epsilon \vec{n} \Big) \right].$$

Since $\vec{t}_p^{\,\epsilon} = \frac{1}{\epsilon^N}\int_{Y_p^\epsilon} \varphi \vec{s}^{\,\epsilon}_p dx$ and $|\varphi(x) - \varphi(x_p^\epsilon)| \le \epsilon \|\nabla\varphi\|_{L^\infty}$, where x_p^ϵ is the center of the cube Y_p^ϵ which contains x, we obtain

$$\left| \int_{G^*_{\epsilon\infty}} \nabla v_\epsilon \cdot \nabla w_\epsilon - \int_{\Omega_\epsilon} \nabla u_\epsilon \cdot \nabla(\varphi w_\epsilon) \right| \le \epsilon \|\nabla\varphi\|_{L^\infty} \|\vec{s}^{\,\epsilon}\|_{L^2(\Omega)^N} \|\nabla w_\epsilon\|_{L^2(\mathbb{R}^+_N)^N},$$

which gives the desired result.

LEMMA 3.8. *From Lemma* 3.7 *we deduce the following approximation result for the displacement vector* $\vec{t}^{\,\epsilon}$*:*

$$\lim_{\epsilon\to 0} \sum_{p\in\mathbb{Z}^N_+} \epsilon^N \left| \frac{1}{\epsilon^N} \int_{\Gamma_p^\epsilon} v_\epsilon \vec{n} ds - \lambda_\epsilon \vec{t}_p^{\,\epsilon} \right|^2 = 0.$$

Proof. We have

$$\begin{aligned} \epsilon^N \sum_{p\in\mathbb{Z}^N_+} \left| \tfrac{1}{\epsilon^N} \textstyle\int_{\Gamma_p^\epsilon} (v_\epsilon - \varphi u_\epsilon)\vec{n} ds \right|^2 &\le \sum_{p\in\mathbb{Z}^N_+} \|\nabla(v_\epsilon - \varphi u_\epsilon)\|^2_{L^2(T_p^\epsilon)^N} \\ &\le \|\nabla(v_\epsilon - \varphi u_\epsilon)\|^2_{L^2(\mathbb{R}^N_+)^N}, \end{aligned} \tag{61}$$

which goes to zero as $\epsilon \to 0$ by virtue of Lemma 3.7. Furthermore,

$$\epsilon^N \sum_{p\in\mathbb{Z}^N_+} \left| \frac{1}{\epsilon^N} \int_{\Gamma_p^\epsilon} \varphi u_\epsilon \vec{n} ds - \lambda_\epsilon \vec{t}_p^{\,\epsilon} \right|^2 \le \epsilon \|\nabla\varphi\|_{L^\infty} \|\nabla u_\epsilon\|_{L^2(\Omega)^N}$$

since, $\vec{s}^{\,\epsilon}$ being constant in each cell Y_p^ϵ,

$$\frac{1}{\epsilon^N} \int_{\Gamma_p^\varepsilon} \varphi u_\epsilon \vec{n} ds = \frac{1}{\varepsilon^N} \int_{\Gamma_p^\epsilon} \Big(\varphi(s) - \frac{1}{\varepsilon^N} \int_{Y_p^\epsilon} \varphi(t) dt \Big) u_\epsilon \vec{n} ds + \lambda_\epsilon (P^1_\epsilon \varphi \vec{s}^{\,\epsilon})_p.$$

Summing these two estimates yields the desired result.

Now, let us define a sequence $\vec{\tau}^{\,\epsilon}$ in ℓ^∞_2 by

$$\vec{\tau}^{\,\epsilon} = \epsilon^{N/2} (\vec{t}_p^{\,\epsilon})_{p\in\mathbb{Z}^N_+},$$

which plays the role of a Weyl sequence for the limit operator B^∞.

PROPOSITION 3.9. *The sequence* $\vec{\tau}^{\,\epsilon}$ *satisfies*

$$\lim_{\epsilon\to 0} \|\vec{\tau}^{\,\epsilon}\|_{\ell^\infty_2} \ge c > 0,$$

and

$$B^\infty \vec{\tau}^{\,\epsilon} = \lambda \vec{\tau}^{\,\epsilon} + \vec{r}^{\,\epsilon}, \tag{62}$$

where $\vec{r}^{\,\epsilon}$ *is a remainder term which goes to zero strongly in* ℓ^∞_2*.*

Proof. A simple rescaling in (59) shows that $\tilde{v}_\epsilon(y) = \epsilon^{\frac{N}{2}-1} v_\epsilon(\epsilon y)$ is the unique solution in $D^1_0(G^{*\infty})$ of

$$\begin{cases} -\Delta \widetilde{v}_\epsilon = 0 & \text{in } G^{*\infty}, \\ \frac{\partial \widetilde{v}_\epsilon}{\partial n} = \vec{\tau}_p^{\,\epsilon} \cdot \vec{n} & \text{on } \Gamma_p,\ p \in \mathbb{Z}^N_+, \\ \widetilde{v}_\epsilon = 0 & \text{on } \mathbb{R}^{N-1} \times \{0\}. \end{cases} \tag{63}$$

Furthermore, $\|\nabla_y \widetilde{v}_\epsilon\|_{L^2(G^{*\infty})^N} = \|\nabla_x v_\epsilon\|_{L^2(G_\epsilon^{*\infty})^N}$. By definition,

$$B^\infty \vec{\tau}^\epsilon = \left(\int_{\Gamma_p} \widetilde{v}_\epsilon \vec{n} ds \right)_{p \in \mathbb{Z}_+^N} = \epsilon^{-\frac{N}{2}} \left(\int_{\Gamma_p^\epsilon} v_\epsilon \vec{n} ds \right)_{p \in \mathbb{Z}_+^N}.$$

Defining $\vec{r}^\epsilon = (\vec{r}_p^\epsilon)_{p \in \mathbb{Z}_+^N}$ by

$$\vec{r}_p^\epsilon = \epsilon^{\frac{N}{2}} \left(\frac{1}{\epsilon^N} \int_{\Gamma_p^\epsilon} v_\epsilon \vec{n} ds - \lambda_\epsilon \vec{t}_p^\epsilon \right),$$

we get

$$B^\infty \vec{\tau}^\epsilon = \lambda_\epsilon \vec{\tau}^\epsilon + \vec{r}^\epsilon,$$

which, by virtue of Lemma 3.8, is the desired result.

To conclude the proof of Theorem 3.3, we remark that either $\vec{\tau}^\epsilon$ converges weakly in ℓ_2^∞ to a nonzero limit $\vec{\tau}$ (up to a subsequence) or $\vec{\tau}^\epsilon$ converges weakly to $\vec{0}$ in ℓ_2^∞. In the first case, passing to the limit as ϵ goes to 0, we obtain that $\vec{\tau} \neq \vec{0}$ is an eigenvector of B^∞ for λ (the limit of the sequence λ_ϵ). In the latter case, this proves that $\vec{\tau}^\epsilon$ is a Weyl sequence for the spectral value λ which belongs to the essential spectrum of B^∞. Now, it is a standard matter (see, e.g., [15], [16]) to show, by a Bloch wave decomposition analogous to that of section 2.3, that the spectrum of B^∞ is nothing but $\lim_{K \to +\infty} \sigma(B^K)$, i.e., the boundary layer spectrum associated with the face Σ of Ω.

REMARK 3.10. *Let us remark that Theorem* 3.3 *is valid for any choice of the sequence ϵ and not only for the particular sequence ϵ_n defined in* (48). *The interested reader will not fail to notice that the present proof of the completeness result is different from that of our previous work* [3]. *In this paper, we used the concept of Bloch measures in order to prove a similar completeness result by means of an energetic method. Here, we propose a new proof (in a slightly different context), based on a rescaling argument, which is simpler, although less precise, and which could equally be applied in* [3].

3.3. Analysis of the corner spectrum. In section 3.1 the boundary layer spectrum $\sigma_{\partial\Omega}$ was defined as the union of all spectra of the type σ_Σ, where Σ is any lower dimensional manifold composing the boundary $\partial\Omega$. When Σ is an $(N-1)$-dimensional hyperplane, a complete derivation of σ_Σ has been given in section 2. However, for lower dimensional manifold we have been a little cavalier in saying that the analysis of section 2 can be easily generalized to the case of edges, corners, and so on. The purpose of this section is to briefly indicate some details of this generalization when analyzing the *corner spectrum*. Since the physical problem of interest is truly two-dimensional, we restrict ourselves to the case of corners of the plane square domain Ω (this has the advantage of simplifying the exposition).

Therefore, our domain Ω is now a rectangle with integer dimensions, i.e.,

$$\Omega =]0; L_1[\times]0; L_2[.$$

We describe the limit spectrum associated with the corner located at the origin. We introduce the space ℓ_+^2 of displacements defined by

$$\ell_+^2 = \left\{ (\vec{s}_j)_{j=(j_1,j_2)\ j_1 \geq 1, j_2 \geq 1} \ \middle|\ \vec{s}_j \in \mathbb{R}^2, \quad \sum_{j_1,j_2=1}^{+\infty} |\vec{s}_{j_1,j_2}|^2 < +\infty \right\}.$$

Remark that this definition of ℓ^2_+ implies a decay of the displacement $\vec{s}_j$ as j_1 or j_2 goes to $+\infty$.

We extend the operator S_ϵ to the larger space ℓ^2_+ by the following formula:

$$C_\epsilon = E_\epsilon S_\epsilon P_\epsilon,$$

where P_ϵ and E_ϵ are, respectively, projection and extension operators between $\mathbb{R}^{Nn(\epsilon)}$ and ℓ^2_+. Their definition is very simple. Recall that a tube T^ϵ_j in Ω is located in a cell Y^ϵ_j whose origin is ϵj. We denote the range of all indices j such that T^ϵ_j is included in Ω by $1 \le j \le n(\epsilon)$. The projection is defined by

$$\begin{array}{lll} P_\epsilon : \ell^2_+ & \longrightarrow & \mathbb{R}^{Nn(\epsilon)}, \\ (\vec{s}_j)_{j=(j_1,j_2)\ j_1\ge 1, j_2\ge 1} & \mapsto & (\vec{s}_j)_{1\le j\le n(\epsilon)}, \end{array}$$

and the extension by

$$\begin{array}{lll} E_\epsilon : \mathbb{R}^{Nn(\epsilon)} & \longrightarrow & \ell^2_+, \\ (\vec{s}_j)_{1\le j\le n(\epsilon)} & \mapsto & (\vec{t}_j)_{j=(j_1,j_2)\ j_1\ge 1, j_2\ge 1}, \end{array}$$

with $\vec{t}_j = \vec{s}_j$ if $1 \le j \le n(\epsilon)$ and $\vec{t}_j = 0$ otherwise.

One can easily check that P_ϵ and E_ϵ are adjoint operators and that the product $P_\epsilon E_\epsilon$ is equal to the identity in $\mathbb{R}^{Nn(\epsilon)}$. Therefore, the spectrum of C_ϵ consists of that of S_ϵ and zero as an eigenvalue of infinite multiplicity.

The convergence analysis of C_ϵ is much simpler than that in section 2 because ℓ^2_+ is not a space of periodically oscillating displacements. There is no need to introduce any notion of two-scale convergence for corner boundary layers. A simple rescaling argument is enough. More precisely, denoting by Q_+ the first quadrant in the plane

$$Q_+ \ =]0;+\infty[\times]0;+\infty[,$$

we replace the two-scale convergence by the weak convergence in $L^2(Q_+)$: with each bounded sequence $u_\epsilon(x)$ in $L^2(\Omega)$, we associate the rescaled sequence $v_\epsilon(y)$ defined by

$$v_\epsilon(y) = \begin{cases} \epsilon^2 u_\epsilon(\epsilon y) & \text{if } \epsilon y \in \Omega, \\ 0 & \text{otherwise,} \end{cases}$$

which is also bounded in $L^2(Q_+)$.

Then, a similar analysis to that of section 2 shows that the sequence of operators C_ϵ converges strongly in $\mathcal{L}(\ell^2_+)$ to a limit operator C_∞ defined by

$$C_\infty : \ell^2_+ \longrightarrow \ell^2_+ \tag{64}$$

$$(\vec{s}_j)_{j=(j_1,j_2)\ j_1\ge 1, j_2\ge 1} \ \mapsto \ \left(\int_{\Gamma_j} u\vec{n} ds \right)_{j=(j_1,j_2)\ j_1\ge 1, j_2\ge 1}, \tag{65}$$

where $u(y)$ is the unique solution of

$$\begin{cases} -\Delta u = 0 & \text{in } Q^*_+ = Q_+ \setminus \bigcup_j T_j, \\ \frac{\partial u}{\partial n} = \vec{s}_j \cdot \vec{n} & \text{on } \Gamma_j,\ j_1 \ge 1, j_2 \ge 1, \\ u = 0 & \text{on } \partial Q_+, \\ \lim\limits_{|y|\to+\infty} u(y) = 0. \end{cases}$$

Clearly, C_∞ is a self-adjoint noncompact operator acting in ℓ^2_+. As in Proposition 2.17, one can prove that the essential spectrum of C_∞ is precisely the Bloch spectrum. However, the discrete spectrum of C_∞ may contain new eigenvalues which correspond to eigenvectors localized in the corner of Q_+.

Acknowledgments. The authors would like to thank one of the anonymous referees for several sharp comments which enabled them to improve and revise the original version.

REFERENCES

[1] F. Aguirre and C. Conca, *Eigenfrequencies of a tube bundle immersed in a fluid*, Appl. Math. Optim., 18 (1988), pp. 1–38.
[2] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.
[3] G. Allaire and C. Conca, *Bloch wave homogenization for a spectral problem in fluid-solid structures*, Arch. Rational Mech. Anal., 135 (1996), pp. 197–257.
[4] G. Allaire and C. Conca, *Analyse asymptotique spectrale de l'équation des ondes. Homogénéisation par ondes de Bloch*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 293–298.
[5] G. Allaire and C. Conca, *Analyse asymptotique spectrale de l'équation des ondes. Complétude du spectre de Bloch*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 557–562.
[6] I. Babuška, *Solution of interface problems by homogenization* I, II, III, SIAM J. Math. Anal., 7 (1976), pp. 603–634, pp. 635–645, and 8 (1977), pp. 923–937.
[7] N. Bakhvalov and G. Panasenko, *Homogenization: Averaging Processes in Periodic Media*, Mathematics and its Applications 36, Kluwer Academic Publishers, Dordrecht, 1990.
[8] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, 1978.
[9] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Boundary layer analysis in homogenization of diffusion equations with Dirichlet conditions in the half space*, in Proc. Internat. Symposium SDE, K. Ito, ed., J. Wiley, New York, 1978, pp. 21–40.
[10] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Boundary layers and homogenization of transport processes*, Publ. Res. Inst. Math. Sci., 151 (1979), pp. 53–157.
[11] F. Bloch, *Uber die Quantenmechanik der Electronen in Kristallgittern*, Z. Phys., 52 (1928), pp. 555–600.
[12] C. Castro and E. Zuazua, *Une remarque sur l'analyse asymptotique spectrale en homogénéisation*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 1043–1048.
[13] D. Cioranescu and J. Saint Jean Paulin, *Homogenization in open sets with holes,* J. Math. Anal. Appl., 71 (1979), pp. 590–607.
[14] C. Conca, J. Planchard, B. Thomas, and M. Vanninathan, *Problèmes mathématiques en couplage fluide-structure. Applications aux faisceaux tubulaires*, Collection EDF/DER 85, Eyrolles, Paris, 1994.
[15] C. Conca, J. Planchard, and M. Vanninathan, *Limiting behaviour of a spectral problem in fluid-solid structures*, Asymptotic Anal., 6 (1993), pp. 365–389.
[16] C. Conca, J. Planchard, and M. Vanninathan, *Fluids and Periodic Structures*, Research in Applied Mathematics 38, J. Wiley and Masson, Paris, 1995.
[17] C. Conca and M. Vanninathan, *A spectral problem arising in fluid-solid structures*, Comput. Methods Appl. Mech. Engrg., 69 (1988), pp. 215–242.
[18] J. Deny and J. L. Lions, *Les espaces du type de Beppo-Levi*, Ann. Inst. Fourier, 5 (1953–54), pp. 305–370.
[19] I. M. Gelfand, *Expansion in a series of eigenfunctions of an equation with periodic coefficients*, Dokl. Akad. Nauk. SSSR, 73 (1950), pp. 1117–1120.
[20] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Translations of Mathematical Monographs 18, AMS, Providence, RI, 1969.
[21] M. Ibnou Zahir, *Problèmes de Valeurs Propres avec des Conditions aux Limites Non Locales*, Thesis, Université Pierre et Marie Curie, Paris, 1984.
[22] M. Ibnou Zahir and J. Planchard, *Natural frequencies of a tube bundle in an incompressible fluid*, Comput. Methods Appl. Mech. Engrg., 41 (1983), pp. 47–68.
[23] M. Ibnou Zahir and J. Planchard, *Fréquences et modes propres d'un faisceau de* 100 *tubes rigides supportés élastiquement et placés dans un fluide incompressible*, Internal report HI 4419-07, Electricité de France, Direction des Etudes et Recherches, 1983.
[24] V. Jikov, S. Kozlov, and O. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[25] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.

[26] J. L. Lions, *Some Methods in the Mathematical Analysis of Systems and their Control*, Science Press, Beijing, Gordon and Breach, New York, 1981.
[27] S. Moskow and M. Vogelius, *First order corrections to the homogenized eigenvalues of a periodic composite medium. A convergence proof*, Proc. Roy. Soc. Edinburgh, to appear.
[28] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modeling of Composite Materials, R. V. Kohn, ed., Series Progress in Nonlinear Differential Equations and their Applications, Birkhauser, Boston, MA, 1997, pp. 21–44.
[29] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[30] F. Odeh and J. Keller, *Partial differential equations with periodic coefficients and Bloch waves in crystals*, J. Math. Phys., 5 (1964), pp. 1499–1504.
[31] J. Planchard, *Eigenfrequencies of a tube-bundle placed in a confined fluid*, Comput. Methods Appl. Mech. Engrg., 30 (1982), pp. 75–93.
[32] J. Planchard, *Global behaviour of large elastic tube-bundles immersed in a fluid*, Comput. Mech., 2 (1987), pp. 105–118.
[33] M. Reed and B. Simon, *Methods of Modern Mathematical Physics,* IV *Analysis of Operators*, Academic Press, New York, 1978.
[34] J. Sánchez-Hubert and E. Sánchez-Palencia, *Vibration and Coupling of Continuous Systems. Asymptotic Methods*, Springer-Verlag, Berlin, 1989.
[35] E. Sánchez-Palencia, *Non Homogeneous Media and Vibration Theory,* Lecture Notes in Physics 127, Springer-Verlag, Berlin, 1980.
[36] C. Wilcox, *Theory of Bloch waves*, J. Anal. Math., 33 (1978), pp. 146–167.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 380–394, March 1998

005

SMOOTHNESS BETWEEN COEFFICIENTS AND BOUNDARY VALUES FOR THE WAVE EQUATION*

GANG BAO†

Abstract. In this work, a density determination problem of an acoustic medium from point source response is studied. The continuity and differentiability are examined for the *forward map* which maps the density to the pressure field. Estimates for both the Lipschitz continuity and the Frechét differentiability are obtained. The continuity of the *linearized* forward map is also obtained. These estimates are crucial for linearization of the nonlinear forward map.

Key words. density determination, regularity of the forward map, continuity of the linearized forward map, microlocal cut-off

AMS subject classifications. 35R30, 86A22

PII. S0036141096299363

1. Introduction and statement of the results. Many wave propagation problems with a point energy source may be modeled by the following linear acoustic wave equation:

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \Delta - \nabla\sigma\cdot\nabla\right)u = \delta(x,t)\ , \tag{1.1}$$

$$u = 0, \quad t < 0\ , \tag{1.2}$$

where the mechanical properties of the medium are described by $\sigma = \sigma(x)$ the logarithm of the density, $c = c(x)$ the sound velocity of the medium, and $u = u(x,t)$ the excess pressure. This work deals with the simplest case where the velocity is assumed to be a fixed constant, say $c = 1$. Also, the problem is set in $\mathbf{R}^{n+1}$ for $n \geq 2$.

Define the *forward map* F as

$$F : \sigma \to (\phi u)\mid_{x_n=0}\ , \tag{1.3}$$

where $\phi \in C_0^\infty(\mathbf{R}^n)$ is supported inside the conoid $\{t > |x|\}$ and near$\{x_n = 0\}$. The cut-off function $\phi(x,t)$ insures that the restriction of distribution u to the hypersurface $\{x_n = 0\}$ is well defined even when the equation (1.1) has a singular right-hand side. It is easily seen that even though the governing equation is linear, the forward map is nonlinear. Thus, solving the inverse problem may be viewed as inverting the nonlinear functional relation F.

In this paper we continue our study of the forward map F. In its predecessor [4], we obtained an a priori upper bound for the linearized forward map with respect to a smooth reference density. Our goal in this work is to establish conditions under which the forward map F is continuous and differentiable and the linearized forward map is continuous. These aspects of forward map are crucial in the study of the inverse problem which arises in reflection seismology, oil exploration, ground-penetrating radar, etc. Here, the inverse problem of interest is to determine the coefficients σ from the measured data of the excess pressure u on a time-like hypersurface $\{x_n = 0\}$. More

*Received by the editors February 26, 1996; accepted for publication (in revised form) February 18, 1997. This research was partially supported by NSF grants DMS 9501099, DMS 9705139, and a Research Development Award (University of Florida).
http://www.siam.org/journals/sima/29-2/29936.html
†Department of Mathematics, University of Florida, Gainesville, FL 32611 (bao@math.ufl.edu).

specifically, given data $F_{data}(x', t)$ (point source response, $x = (x', x_n)$) obtained from many receivers on the earth, find the density $\sigma(x)$ so that

$$F(\sigma) = F_{data},$$

or minimize the error $(F_{data} - F(\sigma))$ in some suitable norm. Two questions arise immediately:

- Since the forward map F is nonlinear, one naturally considers its linearization. So far, most progress has been made through the study of the linearization. However, it seems that very little work has been done to justify this commonly used procedure. Does the linearization of F provide useful information in recovering the density?
- The large size of the typical data set demands a fast means of solving the minimization problem. A natural candidate would be some Gauss–Newton-like method. When can one formulate such an algorithm?

To answer either one of the above questions requires the understanding of regularity of the forward map. In addition, local properties of the inverse problem may be obtained by examining the differentiability of the forward map. Also, continuity properties are crucial in the study of the linearized forward map with respect to a nonsmooth reference density.

Questions on regularity of the forward map have been studied extensively for layered problems where c and σ depend only on x_n. For a similar problem with the impulse-response instead of point source, Symes showed in [15] (for the constant wave speed case) that the forward map is a C^1-*diffeomorphism* by applying the method of geometrical optics together with energy estimates; see [15] for other references. The situation becomes more difficult for the multidimensional (nonlayered) problems. See Symes [14, 16], Sacks and Symes [13], Rakesh [11], and Bao and Symes [4] for some partial results. The difficulties are essentially due to the ill-posed nature of the time-like hyperbolic Cauchy problem and the presence of nonsmooth coefficients. Recently, in the study of this class of inverse problems and other close related problems [2, 3, 4], we have employed nonsmooth microlocal analysis techniques to obtain the optimal time-like trace regularity under weaker hypotheses on the coefficients. In this work, using these techniques, we establish new estimates on the regularity of the forward map and the continuity of its linearization.

We next state the main results of this paper.

Let σ_1 and σ_0 be the densities corresponding to the excess pressure u_1 and u_0, respectively; we have from (1.1) and (1.2) that

$$\begin{aligned} (\Box - \nabla\sigma_1 \cdot \nabla)\tilde{u} &= \nabla\delta\sigma \cdot \nabla u_0 \ , \\ \tilde{u} &= 0, \quad t < 0 \ , \end{aligned} \tag{1.4}$$

where $\Box = \partial_t^2 - \Delta$, $\tilde{u} = u_1 - u_0$, and $\delta\sigma = \sigma_1 - \sigma_0$. Moreover,

$$(\phi\tilde{u})|_{x_n=0} = (\phi u_1)|_{x_n=0} - (\phi u_0)|_{x_n=0} \ .$$

For a real number α, we denote by $[\alpha]$ the smallest integer that is not smaller than α. In the following statements of theorems, we always assume that

$$[l + (n-1)/2] > 1 + n/2 \text{ and } \ \tau > [l + (n-1)/2] + [(n-1)/2] + n/2 + 2 \ .$$

Let $M_\tau > 0$ and define

$$\mathcal{M}_\tau = \{\sigma \in C_0^\infty\{x_n > 0\}, \ \ ||\sigma||_\tau < M_\tau\}.$$

We also assume that the density and its perturbations are supported in the half-space $\{x_n > 0\}$.

THEOREM 1.1. *There exists a constant C depending on M_τ and the support of ϕ so that for σ_1 and $\sigma_0 \in \mathcal{M}_\tau$,*

$$||F(\sigma_1) - F(\sigma_0)||_l \leq C||\sigma_1 - \sigma_0||_{[l+\frac{n-1}{2}]} \, . \tag{1.5}$$

The following results concern the differentiability of the forward map. The formal linearization DF of the forward map F, with respect to the reference state (σ_0, u_0), is defined by the linearized problem

$$\begin{array}{c} (\Box - \nabla\sigma_0 \cdot \nabla)\delta u = \nabla\delta\sigma \cdot \nabla u_0 \, , \\ \delta u = 0, \quad t < 0, \end{array} \tag{1.6}$$

and

$$DF(\sigma_0)\delta\sigma = (\phi\delta u)|_{x_n=0} \, . \tag{1.7}$$

Recall that $\tilde{u}$ solves (1.4) and

$$F(\sigma_1) - F(\sigma_0) = (\phi\tilde{u})|_{x_n=0} \, ; \tag{1.8}$$

then

$$F(\sigma_1) - F(\sigma_0) - DF(\sigma_0)\delta\sigma = (\phi(\tilde{u} - \delta u))|_{x_n=0} \, , \tag{1.9}$$

where

$$\begin{array}{c} (\Box - \nabla\sigma_1 \cdot \nabla)(\tilde{u} - \delta u) = \nabla\delta\sigma \cdot \nabla\delta u \, , \\ \tilde{u} - \delta u = 0, \quad t < 0 \, . \end{array} \tag{1.10}$$

THEOREM 1.2. *There exists a constant C depending on M_τ and the support of ϕ so that for σ_1 and $\sigma_0 \in \mathcal{M}_\tau$,*

$$||F(\sigma_1) - F(\sigma_0) - DF(\sigma_0)\delta\sigma||_l \leq C||\delta\sigma||_{\tau+1}||\delta\sigma||_{[l+\frac{n-1}{2}]} \, .$$

THEOREM 1.3. *There exists a constant C depending on $M_{\tau+1}$ and the support of ϕ so that for σ_1 and $\sigma_0 \in \mathcal{M}_{\tau+1}$,*

$$||F(\sigma_1) - F(\sigma_0) - DF(\sigma_0)\delta\sigma||_l \leq C||\delta\sigma||_\tau||\delta\sigma||_{[l+\frac{n-1}{2}]} \, .$$

Note that the results of Theorems 1.2 and 1.3 do not imply that of Theorem 1.1 since the estimates depend on higher norms of the coefficients.

Finally, we present a continuity estimate for the linearized forward map.

THEOREM 1.4. *There exists a constant C depending on $M_{\tau+1}$ and the support of ϕ so that for σ_1, $\sigma_0 \in \mathcal{M}_{\tau+1}$, and $\eta \in C_0^\infty\{x_n > 0\}$,*

$$||DF(\sigma_1)\eta - DF(\sigma_0)\eta||_l \leq C\left[||\sigma_1 - \sigma_0||_{[l+\frac{n-1}{2}]} \, ||\eta||_{\tau+1} + ||\eta||_{[l+\frac{n-1}{2}]} \, ||\sigma_1 - \sigma_0||_{\tau+1}\right] \, .$$

In particular, DF extends to a Lipschitz continuous map:

$$\begin{array}{c} \sigma \to DF(\sigma) \\ H^{\tau+1}_{comp}(\mathbf{R}^{n-1} \times [0,\infty)) \to \mathcal{L}[H^{\tau+1}_{comp}(\mathbf{R}^{n-1} \times [0,\infty)), \; H^l(\mathbf{R}^{n-1} \times [0,\infty)]. \end{array}$$

Remarks. Our techniques in the proofs are similar to those in [4], where we showed that the linearized forward map $DF(\sigma_0)$ is a bounded map for a smooth reference density σ_0 (Theorem 1.1 in [4]).

In this work, we only study the density determination problem. A more interesting problem is to study the dependence of the boundary values of the pressure field on the velocity c. At present, no regularity result for the multidimensional velocity inversion problem is available. In this case, additional difficulties will occur because of the nonsmooth principal parts in the wave equation. We believe the general ideas described here may be extended to study the velocity inversion problem. Some recent progress has been made in [3] by establishing results on trace regularity and propagation of singularities. The reader is referred to Lewis and Symes [10] for regularity results in the one-dimensional case.

We assume that the reader is familiar with the basic calculus of *pseudodifferential operators* (" ψ*.d.o.* ") as stated in Taylor [17]. Throughout, C serves as a generalized positive constant the precise value of which is not needed.

2. Preliminaries. We recapitulate some useful results in this section. Let us begin with the algebraic properties of the Sobolev spaces and microlocal Sobolev spaces. Here H^s is the standard L^2-type Sobolev space and H^s_{loc} is the corresponding local Sobolev space.

PROPOSITION 2.1 (generalized Schauder's lemma). *If $u \in H^{s_1}(\mathbf{R}^n)$ and $v \in H^{s_2}(\mathbf{R}^n)$, with $s_1 + s_2 \geq 0$, then*

$$uv \in H^{min(s_1, s_2, s_1+s_2-n/2+\delta)} \quad \textit{for any } \delta > 0 \, .$$

Here we also state an extended Rauch lemma established in [2].

LEMMA 2.2. *Suppose that $\Omega = \Omega_0 \times \Omega_1 \subset\subset \mathbf{R}^{n_0} \times \mathbf{R}^{n-n_0}$ $(1 \leq n_0 \leq n)$, $\Omega' \subset\subset \Omega$, $n_0/2 < s$, $0 \leq l \leq s$, q, and $q < l + s - n_0/2$. Suppose that $Q \in S^0(\Omega')$, $\tilde{Q} \in S^0(\Omega)$ elliptic on $ES(Q)$, and $Q_0 \in S^0(\Omega_0)$ satisfies that $(x, y, \xi, \eta) \in ES(\tilde{Q}), \xi \neq 0 \Longrightarrow (x, \xi) \in ES(Q_0)$. Then there exists a constant $C > 0$ so that for $u \in C_0^\infty(\Omega_0)$ and $v(x, y) \in C_0^\infty(\Omega)$,*

$$||Quv||_{q,\Omega'} \leq C(||u||_{s,\Omega_0} + ||Q_0 u||_{q,\Omega_0})(||v||_{l,\Omega} + ||\tilde{Q}v||_{q,\Omega}).$$

Remark. The lemma (for $n_0 = n$) implies the original Rauch lemma in [12].

Let u_0 solve the model problem (1.1) and (1.2) for $\sigma = \sigma_0$.

LEMMA 2.3 (see [4, Theorem 3.1]). *Suppose that $1 + n/2 < s$ and $\sigma_0 \in H^s(\mathbf{R}^n)$. Then for $l < s - n + 1/2$*

$$\partial_t^l u_0 \in L^2_{loc}(U) \, ,$$

where $U = \{\mathbf{R}^n \times (0, T_1)\} \cap \{t > |x|\}$ $(T_1 > 0)$. And for $\phi \in C_0^\infty(U)$, the following estimate holds:

$$||\phi \partial_t^l u_0|| \leq C \, , \tag{2.1}$$

where the constant C depends on ϕ and $||\sigma_0||_s$.

Remark. For an explicit construction of the operator B, we refer the reader to [4].

Let $\Pi : T^*(\Omega_0) \to \Omega_0$ denote the projection of $T^*(\Omega_0)$ onto its base space.

LEMMA 2.4 (see [4, Lemma 3.1]). *Suppose that ψ, $\phi \in C_0^\infty(\mathbf{R}^{n+1})$, $s > n/2$, $k < s + 2 - n/2$, $\sigma_0 \in H^{s+1}_{comp}(\mathbf{R}^n)$, β is a null bicharacteristic strip for $\Box$, and*

$\Omega \subset\subset \mathbf{R}^{n+1}$ *satisfies* $\Pi\beta \cap \bar{\Omega} = \emptyset$. *Then there exist* $B \in OPS^0$ *with* $ES(B)$ *supported in an arbitrarily small conic neighborhood of* β *and* $C > 0$ *so that any* $w \in H^1_{loc}(\mathbf{R}^{n+1})$ *vanishing for large* t *and satisfying*

$$\Box w - \nabla\sigma_0 \cdot \nabla w = f \in L^2(\Omega), \ \ \mathrm{supp}(f) \subset \Omega \ ,$$

satisfies in addition

$$||\psi B \phi w||_k \leq C||f||_0.$$

LEMMA 2.5 (see [4, Lemma 3.2]). *Suppose that* ψ, $\phi \in C_0^\infty(\mathbf{R}^{n+1})$, $s > n/2$, $k < s - n/2 + 2$, $\sigma_0 \in H^{s+1}_{comp}(\mathbf{R}^n)$, P *is a* ψ*.d.o. of order zero such that a conic neighborhood of its essential support* $(ES(P))$ *is contained in the microlocal elliptic region of* $\Box$, *and* $\Omega \subset\subset \mathbf{R}^{n+1}$ *satisfies* $\Pi ES(P) \cap \bar{\Omega} = \emptyset$. *Then there exists a constant* $C > 0$ *so that any* $w \in H^1_{loc}(\mathbf{R}^{n+1})$ *vanishing for large* t *and satisfying*

$$\Box w - \nabla\sigma_0 \cdot \nabla w = f \in L^2(\Omega)$$

satisfies

$$||\psi P \phi w||_k \leq C||f||_0,$$

where the constant C *depends on* σ_0, k, P, ϕ, *and* ψ, *but not on* w.

We need the anisotropic Sobolev spaces $\mathcal{H}^{m,s}(\mathbf{R}^{n+1})$ introduced originally by Hörmander [9]:

$$\mathcal{H}^{m,s}(\mathbf{R}^k) = \{f \in \mathcal{D}', \ D^\alpha_{x',x_k} f \in L^2(\mathbf{R}^k) \ \forall \ \alpha = (\alpha_1, \alpha_1, \ldots, \alpha_k), |\alpha| \leq m+s, \alpha_k \leq m\},$$

where $D^\alpha_{x',x_k} = D^{\alpha_1,\ldots,\alpha_{k-1}}_{x'} D^{\alpha_k}_{x_k}$.

PROPOSITION 2.6 (see [9]). *Suppose* $m > 1/2$ *and* $m + s > k/2$. *Then*

$$\mathcal{H}^{m,s} \subset L^\infty(\mathbf{R}^k) \cap C^0(\mathbf{R}^k) \ \ \textit{continuous inclusion} \, .$$

We also need a conormal regularity result for the wave equation in the following form:

$$\begin{aligned} &(\partial_t^2 - \Delta - \nabla\sigma_0 \cdot \nabla)u(x,t) = a(x)S(t - r(x)) \ , \\ &\quad u = 0, \quad t < 0 \ , \end{aligned} \tag{2.2}$$

where again $r(x) = |x|$.

Introduce the standard polar coordinates with variables $r = |x|$, θ_1, θ_2, $\ldots$, θ_{n-1}. Denote $T_1 = \partial_t + \partial_r$, $T_{i+1} = \partial_{\theta_i}$ for $i = 1, 2, \ldots, n-1$, and $T_{n+1} = \partial_t - \partial_r$. Then T_i $(i = 1, 2, \ldots, n)$ form a basis of the tangent space to the characteristic surface $\{t = |x|\}$.

LEMMA 2.7 (see [4, Theorem 5.1]). *Suppose that, in* (2.12), $S \in H^{m-1}_{loc}$ *and* $a(x)$ *is a smooth function. Suppose also that* $k \geq 0$, $p \geq m+k$, $p > k+n/2+1$, $q \geq m+k-1$, *and* $q > k + n/2$. *Then for* $\{i_1, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$ *and* $\phi(x,t) \in C_0^\infty(\mathbf{R}^{n+1})$,

$$T_{i_1} \cdots T_{i_k} u \in H^m_{loc} \ \textit{or} \ \ u \in H^{m,k}_{loc}.$$

In addition,

$$||\phi T_{i_1} \cdots T_{i_k} u||_m \leq C||a||_q$$

with the constant C *depending on* $||\sigma_0||_p$.

3. Progressing wave expansion. Consider a problem related to the model problem (1.1)(1.2)

$$\begin{aligned} (\Box - \nabla\sigma_0 \cdot \nabla)v_0 &= \delta^{(-\frac{n-1}{2})}(t)\delta(x) \ , \\ v_0 &= 0, \quad t < 0 \ . \end{aligned} \tag{3.1}$$

Hadamard's construction leads to the progressing wave expansion for v_0,

$$v_0 = \sum_{k=0}^{m-1} b_k(x) S_k(t - r(x)) + R_m(x,t), \tag{3.2}$$

where $r(x) = |x|$, $S_0(\cdot) = H(\cdot)$ is the Heaviside function, $S_k' = S_{k-1}$ $(k \geq 1)$, and $\{b_k\}$ $(k = 1, \ldots, m-1)$ solve the transport equations

$$2\nabla r \cdot \nabla b_0 + (\Delta r + \nabla r \cdot \nabla\sigma_0) b_0 = 0, \tag{3.3}$$

$$2\nabla r \cdot \nabla b_k + (\Delta r + \nabla r \cdot \nabla\sigma_0) b_k = \Delta b_{k-1} + \nabla\sigma_0 \cdot \nabla b_{k-1} \ . \tag{3.4}$$

Moreover, the remainder term R_m satisfies

$$\begin{aligned} (\Box - \nabla\sigma_0 \cdot \nabla)R_m &= (\Delta + \nabla\sigma_0 \cdot \nabla) b_{m-1} S_{m-1}(t - r(x)) \ , \\ R_m &= 0 \ , \quad t < 0 \ . \end{aligned} \tag{3.5}$$

Away from the origin, R_m is more regular. We refer the reader to [8] or [7] for a general discussion on the method of progressing wave expansions.

Similarly, integrating (1.6) in t gives

$$\begin{aligned} (\Box - \nabla\sigma_0 \cdot \nabla)\delta v &= \nabla\delta\sigma \cdot \nabla v_0 \ , \\ \delta v &= 0, \quad t < 0 \ , \end{aligned} \tag{3.6}$$

where $\delta u = \partial_t^{\frac{n-1}{2}} \delta v$.

The progressing wave expansion for δv takes the following form:

$$\delta v = \sum_{k=0}^{m-1} h_k(x) S_k(t - r(x)) + R_m^\delta(x,t), \tag{3.7}$$

where

$$2\nabla r \cdot \nabla h_0 + (\Delta r + \nabla r \cdot \nabla\sigma_0) h_0 = -b_0 \nabla\delta\sigma \cdot \nabla r, \tag{3.8}$$

$$\begin{aligned} 2\nabla r \cdot \nabla h_k + (\Delta r + \nabla r \cdot \nabla\sigma_0) h_k = (\Delta + \nabla\sigma_0 \cdot \nabla) h_{k-1} \\ + \nabla\delta\sigma \cdot (\nabla b_{k-1} - b_k \nabla r) \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} (\Box - \nabla\sigma_0 \cdot \nabla)R_m^\delta &= \nabla\delta\sigma \cdot \nabla R_m + [\nabla\delta\sigma \cdot \nabla b_{m-1} \\ &\quad + (\Delta + \nabla\sigma_0 \cdot \nabla) h_{m-1}] S_{m-1}(t - r(x)) \ , \\ R_m^\delta &= 0, \quad t < 0 \ . \end{aligned} \tag{3.10}$$

Let us first simplify the transport equations by introducing two functions α and α_0 such that

$$\nabla r \cdot \nabla\alpha_0 = \Delta r/2 \ \text{ and } \alpha = \sigma_0/2 + \alpha_0 \ .$$

It is obvious that away from the origin α is nothing more than a smooth perturbation of $\sigma_0/2$.

We can then rewrite the transport equations as

$$2\nabla r \cdot \nabla(b_0 e^{\alpha}) = 0, \tag{3.11}$$

$$2\nabla r \cdot \nabla(b_k e^{\alpha}) = e^{\alpha}(\Delta b_{k-1} + \nabla\sigma_0 \cdot \nabla b_{k-1}) \tag{3.12}$$

and

$$2\nabla r \cdot \nabla(h_0 e^{\alpha}) = -e^{\alpha} b_0 \nabla\delta\sigma \cdot \nabla r, \tag{3.13}$$

$$2\nabla r \cdot \nabla(h_k e^{\alpha}) = e^{\alpha}[(\Delta + \nabla\sigma_0 \cdot \nabla)h_{k-1} + \nabla\delta\sigma \cdot (\nabla b_{k-1} - b_k \nabla r)]\ . \tag{3.14}$$

Let V denote the vector field $\nabla r \cdot \nabla$ and $Char(\nabla r \cdot \nabla) = \{(x,\xi) \in T^*(\mathbf{R}^n),\ \nabla r \cdot \xi = 0\}$.

LEMMA 3.1 (see [4, Lemma 4.1]). *Assume that u is a smooth function with $\mathrm{supp}(u) \subset \{|x| > \delta\}$ for some $\delta > 0$. Assume also that $\phi \in C_0^{\infty}(\mathbf{R}^n)$ with $\phi = 1$ on Ω and $\mathrm{supp}(\phi) \subset\subset \Omega'$, where Ω and Ω' are bounded open sets in $\mathbf{R}^n$. Then there exist a $Q \in OPS^0$ which is elliptic on $Char(V)$, $[Q,V] \in OPS^{-\infty}$, and $\phi' \in C_0^{\infty}(\mathbf{R}^n)$, $\phi' > 0$ on $\mathrm{supp}(\phi)$, such that for $s \in \mathbf{R}$ the estimates*

$$||u||_{s,\Omega} \le C||Q\phi' V u||_{s,\Omega'} + C||Vu||_{s-1,\Omega'} + C||u||_{\tau,\Omega'}, \tag{3.15}$$

$$||Q\phi u||_{s,\Omega} \le C||Q\phi V u||_{s,\Omega} + C||u||_{\tau,\Omega'} \tag{3.16}$$

hold for any $\tau \in \mathbf{R}$, where the constants are independent of u.

PROPOSITION 3.2 (see [4, Theorem 4.1]). *Suppose Ω is a bounded open subset of $\mathbf{R}^n$. Then for $i = 0, 1, \ldots, m-1$,*

$$||b_i||_{k,\Omega} \le C_{i,k},$$

where the constants $C_{i,k}$ depend on $||\sigma_0||_{s_i+2i}$ with $s_i > n/2 + 1$, $s_i \ge k$.

LEMMA 3.3. *Suppose that Ω is a bounded open subset of $\mathbf{R}^n$. Then for $i = 0, 1, \ldots, m-1$,*

$$||h_i||_{k,\Omega} \le C||\delta\sigma||_{k+2i+1},$$

where the constant depends on $||\sigma_0||_{p+2i}$ with $p > n/2 + 1$ and $p \ge k$.

Remark. An alternative way to write the equation (3.13) is

$$\nabla r \cdot \nabla(2h_0 e^{\alpha} + e^{\alpha} b_0 \delta\sigma) = \nabla r \cdot \nabla(e^{\alpha} b_0)\delta\sigma.$$

Thus the right-hand side involves the first-order derivative of b_0 but no derivative of $\delta\sigma$. Using this fact, similar estimates will then yield that

$$||h_i||_{k,\Omega} \le C||\delta\sigma||_{k+2i},$$

but the constant C depends on $||\sigma_0||_{p+2i+1}$ with $p > n/2$ and $p \ge k$.

The proof of Lemma 3.3 may be given by applying Lemma 3.1 and the regularity result for b_i (Proposition 3.2) to the energy estimates of h_i by using the assumption that $\delta\sigma = 0$ near $\{x_n = 0\}$. We shall omit the proof.

4. On the differentiability. Here we shall prove Theorem 1.2. The proof of Theorem 1.3 may be given by repeating the same arguments and using the remark after Lemma 3.3. Also, we shall omit the proof of Theorem 1.1 since it also follows the same arguments.

Due to the consideration of the simple principal part in the progressing wave expansions, we integrate (1.10) in t to get

$$\begin{aligned} (\Box - \nabla\sigma_1 \cdot \nabla)(v - \delta v) &= \nabla\delta\sigma \cdot \nabla\delta v, \\ v - \delta v &= 0, \quad t < 0 \, , \end{aligned} \tag{4.1}$$

where $\tilde{u} = \partial_t^{\frac{n-1}{2}} v$, $\delta u = \partial_t^{\frac{n-1}{2}} \delta v$, and v_0 solves

$$\begin{aligned} (\Box - \nabla\sigma_0 \cdot \nabla)v_0 &= \delta^{(-\frac{n-1}{2})}(t)\delta(x) \, , \\ v_0 &= 0, \quad t < 0 \, . \end{aligned} \tag{4.2}$$

Then, for $l \in \mathbf{R}$,

$$\begin{aligned} ||F(\sigma_1) - F(\sigma_0) - DF(\sigma_0)\delta\sigma||_l &= ||(\phi(\tilde{u} - \delta u))\,|_{x_n=0}\,||_l \\ &\le C||(\phi(v - \delta v))\,|_{x_n=0}\,||_{l_1}, \end{aligned} \tag{4.3}$$

where l_1 denotes $[l + (n-1)/2]$. In this way, we have reduced the continuity analysis to a time-like trace regularity estimate for $v - \delta v$.

Since $\phi \in C_0^\infty$ is supported near the time-like trace surface $\{x_n = 0\}$, a trace regularity result [2, Theorem 3.1] implies that

$$||\phi(v - \delta v)|_{x_n=0}||_{l_1} \le C||\phi_0(v - \delta v)||_{l_1}, \tag{4.4}$$

where C depends on the H^{l_1+1}-norm of σ_1 and $\phi_0 \in C_0^\infty$ supported near $\mathrm{supp}(\phi)$. Thus the estimate of $||\phi(v-\delta v)|_{x_n=0}||_{l_1}$ may be reduced to estimating $||\phi_0(v-\delta v)||_{l_1}$.

Because of the assumptions that $\delta\sigma$ is supported away from $\{x_n = 0\}$ and ϕ_0 is supported near the time-like surface, a result in [4, Proposition 6.1] implies that it suffices to estimate the t-derivatives of $v - \delta v$. We study the regularity of $v - \delta v$ through a dual problem. It is then convenient to look at the symmetric form. For $\rho_1(x) = e^{-\sigma_1}$,

$$\begin{aligned} \Box_1(v - \delta v) = \left[\frac{1}{\rho_1}\partial_t^2 - \nabla \cdot \left(\frac{1}{\rho_1}\nabla\right)\right](v - \delta v) &= \frac{1}{\rho_1}\nabla\delta\sigma \cdot \nabla\delta v \, , \\ v - \delta v = 0, \quad t < 0 \, . \end{aligned} \tag{4.5}$$

Introduce a dual problem to (4.5)

$$\begin{aligned} \Box_1' w = \left[\frac{1}{\rho_1}\partial_t^2 - \nabla \cdot \left(\frac{1}{\rho_1}\nabla\right)\right] w &= \Psi \, , \\ w = 0, \quad t \gg T_0 \, , \end{aligned} \tag{4.6}$$

where $\Psi \in C_0^\infty(\Omega_0)$ with $\Omega_0 \subseteq \{|x_n| < \epsilon\} \cap \{t \in (0, T_0), t > |x| + \epsilon\}$ for some $\epsilon > 0$.

Thus to obtain

$$||\partial_t^{l_1}(v - \delta v)||_{0,\Omega_0} \le C||\delta\sigma||_{\tau+1}\,||\delta\sigma||_{l_1} \, , \tag{4.7}$$

it suffices to show that for any $\Psi \in C_0^\infty(\Omega_0)$

$$|(\partial_t^{l_1}(v - \delta v), \Psi)| \le C||\delta\sigma||_{\tau+1}||\delta\sigma||_{l_1}\,||\Psi||_0 \, . \tag{4.8}$$

Integration by parts gives

$$\begin{aligned}(\partial_t^{l_1}(v-\delta v),\Psi) &= (\rho_1^{-1}\nabla\delta\sigma\cdot\nabla\partial_t^{l_1}\delta v, w)\\ &= (\rho_1^{-1}\nabla\delta\sigma\cdot\nabla\delta v, \partial_t^{l_1}(\phi_1 w)),\end{aligned} \tag{4.9}$$

where $\phi_1 \in C_0^\infty(\mathbf{R}^{n+1})$, ϕ_1 is supported in a small neighborhood of $\mathrm{supp}(w) \cap \mathrm{supp}(\nabla\delta\sigma\cdot\nabla\partial_t^{l_1}\delta v)$, and $\phi_1 = 1$ on $\mathrm{supp}(w)\cap\mathrm{supp}(\nabla\delta\sigma\cdot\nabla\partial_t^{l_1}\delta v)$.

Construct Q_0, Q_1, and $Q_2 \in OPS^0$ such that

- $\mathrm{supp}(q_0)$ is strictly contained in the light cone $\{t \geq r(x)\}$; q_1 and q_2 are supported near the light cone;
- $ES(Q_1) \subseteq$ a conic neighborhood of $Char(\partial_t + \nabla r\cdot\nabla)$;
- $ES(Q_2) \subseteq \{\omega + \nabla r\cdot\xi \neq 0\}$;
- $Q_0 + Q_1 + Q_2 = I$.

Hence

$$\partial_t^{l_1}(\phi_1 w) = \partial_t^{l_1}Q_0(\phi_1 w) + \partial_t^{l_1}Q_1(\phi_1 w) + \partial_t^{l_1}Q_2(\phi_1 w)\ .$$

Therefore

$$(\partial_t^{l_1}(v-\delta v),\Psi) = E_0 + E_1 + E_2\ , \tag{4.10}$$

where $E_j = (\rho_1^{-1}\nabla\delta\sigma\cdot\nabla\delta v, \partial_t^{l_1}Q_j(\phi_1 w))$ for $j = 0, 1, 2$.

Integration by parts gives

$$\begin{aligned}|E_0| &= \left|\sum_{j=1}^n (\rho_1^{-1}\partial_{x_j}\delta\sigma\partial_{x_j}\delta v, \partial_t^{l_1}Q_0(\phi_1 w))\right|\\ &= \left|\left(\partial_t^{l_1}\delta v, \sum_{j=1}^n \partial_{x_j}\phi_2\rho_1^{-1}\partial_{x_j}\delta\sigma Q_0(\phi_1 w)\right)\right|\ ,\end{aligned}$$

where $\phi_2 \in C_0^\infty(\mathbf{R}^{n+1})$ is supported strictly inside the light cone and $\phi_2 = 1$ on $\mathrm{supp}(q_0)\cap\mathrm{supp}(\phi_1)$.

Introduce another dual problem

$$\begin{aligned}&\Box_0' U = \left[\frac{1}{\rho_0}\partial_t^2 - \nabla\cdot\left(\frac{1}{\rho_0}\nabla\right)\right]U = -\sum_{j=1}^n \partial_{x_j}\left[\phi_2\frac{1}{\rho_1}\partial_{x_j}\delta\sigma Q_0(\phi_1 w)\right]\ ,\\ &\quad U = 0\ ,\quad t \gg T_0\ .\end{aligned} \tag{4.11}$$

Then

$$|E_0| = |(\partial_t^{l_1}\delta v, \Box_0' U)| = |(\partial_t^{l_1}\delta v, \Box_0'\phi_3 U)|\ ,$$

where $\phi_3 \in C_0^\infty$ supported strictly inside the characteristic surface and $\phi_3 = 1$ on $\mathrm{supp}(\phi_2)$.

Using the equation for δv, we have by integration by parts

$$\begin{aligned}|E_0| &= |(\nabla\delta\sigma\cdot\nabla\partial_t^{l_1}v_0, \phi_3 U)|\\ &= \left|\left(\partial_t^{l_1}v_0, \sum_{j=1}^n \partial_{x_j}(\phi_3 U\partial_{x_j}\delta\sigma)\right)\right|\\ &\leq ||\partial_t^{l_1}v_0||\ \left\|\sum_{j=1}^n \partial_{x_j}(\phi_3 U\partial_{x_j}\delta\sigma)\right\|\ ,\end{aligned}$$

where the integral is over the set strictly inside the characteristic surface. It follows from Lemma 2.3 that $||\partial_t^{l_1} v_0|| \leq C$ with the constant C depending on $||\sigma_0||_s$ for $s > l_1 + n/2$.

Thus

$$|E_0| \leq C||\phi_3 U \nabla \delta\sigma||_1 \leq C||\phi_3 U||_1 \, ||\delta\sigma||_{s_1+1}$$

for some $s_1 > n/2$.

Applying the method of energy estimates to the second dual problem yields further that

$$\begin{aligned} ||\phi_3 U||_1 &\leq C||\phi_2 \rho_1^{-1} \nabla \delta\sigma Q_0(\phi_1 w)||_1 \\ &\leq C||\delta\sigma||_{s_1+1}||\phi_1 w||_1 \\ &\leq C||\delta\sigma||_{s_1+1}||\Psi|| \, , \end{aligned}$$

where the constant C depends on $||\rho_1||_{s_1}$.

Therefore

$$|E_0| \leq C||\delta\sigma||^2_{s_1+1}||\Psi|| \, . \tag{4.12}$$

We next estimate the term E_1. According to the progressing wave expansion (3.7),

$$\nabla\delta\sigma \cdot \nabla\delta v = -h_0 \nabla\delta\sigma \cdot \nabla r \; \delta(t-r) + \nabla\delta\sigma \cdot \nabla h_0 \; H(t-r) + \nabla\delta\sigma \cdot \nabla R_1^\delta \, , \tag{4.13}$$

where $R_1^\delta(x,t)$ solves

$$\begin{aligned} (\Box - \nabla\sigma_0 \cdot \nabla) R_1^\delta &= \nabla\delta\sigma \cdot \nabla R_1 + \nabla\delta\sigma \cdot \nabla b_0 + (\Delta + \nabla\sigma_0 \cdot \nabla) h_0 \; H(t - r(x)) \, , \\ R_1^\delta &= 0, \quad t < 0 \, . \end{aligned} \tag{4.14}$$

Consequently

$$\begin{aligned} E_1 = &(-\psi\rho_1^{-1} h_0 \nabla\delta\sigma \cdot \nabla r \; \delta(t-r) + \psi\rho_1^{-1} \nabla\delta\sigma \cdot \nabla h_0 \; H(t-r), \partial_t^{l_1} Q_1(\phi_1 w)) \\ &+(\phi_1 \rho_1^{-1} \nabla\delta\sigma \cdot \nabla R_1^\delta, \partial_t^{l_1} Q_1(\phi_1 w)), \end{aligned}$$

with $\psi \in C_0^\infty(\mathbf{R}^n)$, $\psi(x) = 1$ on the x-projection of $\text{supp}(\nabla\delta\sigma \cdot \nabla\delta v) \cap \text{supp}(\partial_t^{l_1} Q_1(\phi_1 w))$.

Hence

$$\begin{aligned} |E_1| \leq &|(-\psi\rho_1^{-1} h_0 \nabla\delta\sigma \cdot \nabla r, (\partial_t^{l_1} Q_1(\phi_1 w))_{t=r})| + |(\psi\rho_1^{-1} \nabla\delta\sigma \cdot \nabla h_0, (\partial_t^{l_1 - 1} Q_1(\phi_1 w))_{t=r})| \\ &+|(\phi_1 \rho_1^{-1} \nabla\delta\sigma \cdot \nabla R_1^\delta, \partial_t^{l_1} Q_1(\phi_1 w))| \, . \end{aligned}$$

An application of the trace theorem yields that

$$\begin{aligned} |E_1| \leq &||\psi h_0 \rho_1^{-1} \nabla\delta\sigma \cdot \nabla r|| \; ||Q_1(\phi_1 w)||_{l_1 + 1/2} + ||\psi\rho_1^{-1} \nabla\delta\sigma \cdot \nabla h_0|| \; ||Q_1(\phi_1 w)||_{l_1 - 1/2} \\ &+||\phi_1 \rho_1^{-1} \nabla\delta\sigma \cdot \nabla R_1^\delta|| \; ||Q_1(\phi_1 w)||_{l_1} \, . \end{aligned}$$

By the construction of Q_1, the results on propagation of singularities (Lemmas 2.4 and 2.5) give that

$$||Q_1 \phi_1 w||_{l_1 + 1/2} \leq C||\Psi||_0 \, ,$$

where C depends on $||\sigma_1||_p$ for $p > l_1 + (n-1)/2$. Hence

$$\begin{aligned}|E_1| &\le C(||\psi h_0 \nabla \delta\sigma \cdot \nabla r|| + ||\psi \nabla \delta\sigma \cdot \nabla h_0|| + ||\phi_1 \nabla \delta\sigma \cdot \nabla R_1^\delta||)\ ||\Psi|| \\ &\le C||\delta\sigma||_{l_1}||\psi h_0||_1||\Psi|| + C||\phi_1 \nabla \delta\sigma \cdot \nabla R_1^\delta||\ ||\Psi||,\end{aligned}$$

with C depending on $||\sigma_1||_s$, $s > l_1 + (n-1)/2$.

Thus the problem has been reduced to estimate

$$||\phi_1 \nabla \delta\sigma \cdot \nabla R_1^\delta|| \le C||\delta\sigma||_{l_1}||\tilde{\phi_1} R_1^\delta||_1$$

for some $\tilde{\phi_1} \in C_0^\infty(\mathbf{R}^{n+1})$.

From (4.14), the method of energy estimates gives

$$||\tilde{\phi_1} R_1^\delta||_1 \le C||\delta\sigma||_{s+1}\ ,$$

where $s > n/2$ and the constant C depends on $||\sigma_0||_{s+1}$.

Therefore

$$|E_1| \le C||\delta\sigma||_{s+1}||\delta\sigma||_{l_1}||\Psi||\ . \tag{4.15}$$

We now estimate the term E_2.

Since $\partial_t + \nabla r \cdot \nabla$ is elliptic on $ES(\partial_t^{l_1} Q_2)$, the calculus of $\psi.d.o.$ implies that there exists a $\psi.d.o.$ $\tilde{Q}_2$ of order zero such that

$$\partial_t^{l_1 - i} Q_2 = (\partial_t + \nabla r \cdot \nabla)^{l_1 - i} \tilde{Q}_2 + K_i \text{ for } i = 0, 1, 2\ , \tag{4.16}$$

where K_i are smoothing operators.

Once again, from the progressing wave expansion

$$\nabla \delta\sigma \cdot \nabla \delta v = \sum_{i=0}^{2} d_i(x) S_{i-1}(t - r(x)) + \nabla \delta\sigma \cdot \nabla R_2^\delta(x, t)\ , \tag{4.17}$$

where $S_{-1}(t - r(x)) = \delta(t - r(x))$, $S_0(t - r(x)) = H(t - r(x))$, $S_1' = H$, and

$$d_0 = -h_0 \nabla \delta\sigma \cdot \nabla r\ , \tag{4.18}$$

$$d_1 = \nabla \delta\sigma \cdot \nabla h_0 - h_1 \nabla \delta\sigma \cdot \nabla r\ , \tag{4.19}$$

$$d_2 = \nabla \delta\sigma \cdot \nabla h_1\ . \tag{4.20}$$

Here R_2^δ solves the equation (3.10) with $m = 2$.

Denote $T_1 = \partial_t + \nabla r \cdot \nabla$. We can then rewrite E_2 as, for some $\psi \in C_0^\infty(\mathbf{R}^n)$,

$$\begin{aligned}E_2 &= \sum_{i=0}^{2} (\psi \rho_1^{-1} d_i S_{i-1}, \partial_t^{l_1} Q_2(\phi_1 w)) + (\phi_1 \rho_1^{-1} \nabla \delta\sigma \cdot \nabla R_2^\delta, \partial_t^{l_1} Q_2(\phi_1 w)) \\ &= \sum_{i=0}^{2} (\psi \rho_1^{-1} d_i, (\partial_t^{l_1 - i} Q_2(\phi_1 w))_{t=r}) + (\phi_1 \rho_1^{-1} \nabla \delta\sigma \cdot \nabla R_2^\delta, \partial_t^{l_1} Q_2(\phi_1 w))\ .\end{aligned}$$

Notice that

$$(\partial_t^{l_1 - i} Q_2(\phi_1 w))_{t=r} = (\nabla r \cdot \nabla)^{l_1 - i} (\tilde{Q}_2(\phi_1 w))_{t=r}\ .$$

We may use (4.16) and integration by parts to get

$$
\begin{aligned}
|E_2| &\le \sum_{i=0}^{2} ||(\nabla r\cdot\nabla+\Delta r)^{l_1-i-1}\psi\rho_1^{-1}d_i||\ ||(T_1\tilde{Q}_2(\phi_1 w))_{t=r})|| \\
&\quad +||(T_1+\Delta r)^{l_1-1}\phi_1\rho_1^{-1}\nabla\delta\sigma\cdot\nabla R_2^{\delta}||\ ||T_1\tilde{Q}_2(\phi_1 w)|| \\
&\le C\sum_{i=0}^{2} ||\psi\rho_1^{-1}d_i||_{l_1-i-1}||\tilde{Q}_2(\phi_1 w)_{t=r}||_1 + ||(T_1+\Delta r)^{l_1-1}\phi_1\rho_1^{-1}\nabla\delta\sigma\cdot\nabla R_2^{\delta}||\ ||\phi_1 w||_1\ .
\end{aligned}
$$

Applying the generalized Schauder's lemma and the assumption that $l_1-1>n/2$ as well as Lemma 3.3, we have from (4.18)–(4.20) that

$$
\begin{aligned}
&||\psi\rho_1{}^{-1}d_0||_{l_1-1}\le C||\delta\sigma||_{l_1}||\psi_0 h_0||_{l_1-1}\le C_0||\delta\sigma||_{l_1}^2\ , \\
&||\psi\rho_1{}^{-1}d_1||_{l_1-2}\le C||\delta\sigma||_{l_1}(||\psi_1 h_0||_{l_1-1}+||\psi_1 h_1||_{l_1-2})\le C_1||\delta\sigma||_{l_1}^2\ , \\
&||\psi\rho_1{}^{-1}d_2||_{l_1-3}\le C||\delta\sigma||_{l_1}||\psi_2 h_1||_{l_1-2}\le C_1||\delta\sigma||_{l_1}^2\ ,
\end{aligned}
$$

where $\psi_i\in C_0^{\infty}(\mathbf{R}^n)$. By Lemma 3.3, the constants C_0 and C_1 depend on $||\sigma_0||_{s_0}$ and $||\sigma_0||_{s_1}$, respectively, where $s_0>n/2$, $s_1>2+n/2$ and $s_0\ge l_1-1$, $s_1\ge l_1$. The constants also depend on $||\sigma_1||_{l_1-1}$.

Since $ES(\tilde{Q}_2)\subset\{\omega+\nabla r\cdot\xi\neq 0\}$, a trace regularity result (Corollary 2 in [1]) implies the existence of a $\phi_1'\in C_0^{\infty}(\mathbf{R}^{n+1})$ such that

$$
||(\tilde{Q}_2\phi_1 w)_{t=r(x)}||_1\le C||\phi_1' w||_1\le C||\Psi||_0\ .
$$

Therefore

$$
|E_2|\le C||\delta\sigma||_{l_1}^2||\Psi||+C||(T_1+\Delta r)^{l_1-1}\phi_1\nabla\delta\sigma\cdot\nabla R_2^{\delta}||\ ||\Psi||\ . \tag{4.21}
$$

Thus it suffices to estimate $||(T_1+\Delta r)^{l_1-1}\phi_1\nabla\delta\sigma\cdot\nabla R_2^{\delta}||$ or essentially to estimate, for $\phi\in C_0^{\infty}(\mathbf{R}^{n+1})$,

$$
||\phi T_1^{l_1-1}\nabla\delta\sigma\cdot\nabla R_2^{\delta}||\ . \tag{4.22}
$$

By Lipschitz's formula, it is sufficient to estimate

$$
\begin{aligned}
&\sum_{i=0}^{l_1-1}||\phi T_1^{l_1-1-i}\nabla\delta\sigma\cdot T_1^{i}\nabla R_2^{\delta}|| \\
&\quad =||\phi\nabla\delta\sigma\cdot T_1^{l_1-1}\nabla R_2^{\delta}||+\sum_{i=0}^{l_1-2}||\phi T_1^{l_1-1-i}\nabla\delta\sigma\cdot T_1^{i}\nabla R_2^{\delta}||\ .
\end{aligned} \tag{4.23}
$$

Note that although $R_2^{\delta}\in H_{loc}^2$ from the progressing wave expansion, the terms on the right-hand side of the above expression are still well defined due to the conormal properties of R_2^{δ}. We have by using the assumption $l_1-1>n/2$ and Schauder's lemma

$$
||\phi\nabla\delta\sigma\cdot T_1^{l_1-1}\nabla R_2^{\delta}||\le C||\delta\sigma||_{l_1}||\phi T_1^{l_1-1}\nabla R_2^{\delta}||.
$$

An application of Lemmas 2.7 and 3.3 yields that $R_2^{\delta}\in H_{loc}^{1,l_1-1}$ and

$$
||\phi T_1^{l_1-1}\nabla R_2^{\delta}||\le C||\delta\sigma||_q
$$

with $q > l_1 + 4 + n/2$ and the constant C depending on $||\sigma_0||_{l_1+3}$.

Thus

$$||\phi\nabla\delta\sigma \cdot T_1^{l_1-1}\nabla R_2^\delta|| \le C||\delta\sigma||_q||\delta\sigma||_{l_1}.$$

To estimate the second term on the right-hand side of (4.24) requires a different approach. We shall show that $\phi T_1^i \nabla R_2^\delta$ $(0 \le i \le l_1 - 2)$ is bounded by $C||\delta\sigma||_{\tau+1}$. Indeed if this is the case, then

$$\begin{aligned}\sum_{i=0}^{l_1-2} ||\phi T_1^{l_1-1-i}\nabla\delta\sigma \cdot T_1^i\nabla R_2^\delta|| &\le C\sum_{i=0}^{l_1-2} ||T_1^{l_1-1-i}\nabla\delta\sigma||\ ||\delta\sigma||_{l_1} \\ &\le C||\delta\sigma||_{\tau+1}||\delta\sigma||_{l_1}.\end{aligned}$$

According to Proposition 2.6, it suffices to show that for some integer $s > (n-1)/2$

$$R_2^\delta \in H_{loc}^{2,s+l_1-2}\ .$$

This can be done by an application of Lemmas 2.7 and 3.3. In fact, by applying Lemma 2.7 and Proposition 2.6 to the equation (3.10) (for $m = 2$), the term $\phi T_1^i \nabla R_2^\delta$ is bounded by

$$C(||\phi\nabla\delta\sigma \cdot \nabla R_2||_q + ||\psi\nabla\delta\sigma \cdot \nabla b_1 + \psi(\Delta + \nabla\sigma_0 \cdot \nabla)h_1||_q),$$

where the constant depends on $||\sigma_0||_p$ with $p \ge s + l_1$ and $p > s + l_1 + n/2 - 1$, and $q \ge s + l_1 - 1$, $q > s + l_1 + n/2 - 2$ for $\phi, \psi \in C_0^\infty$. We then use Lemma 3.3 and the regularity result for b_1 (Proposition 3.2) to obtain that

$$|\phi T_1^i \nabla R_2^\delta| \le C||\delta\sigma||_{q+5},$$

where the constant depends on $||\sigma_0||_{q+4}$ with $q + 4 > l_1 + [(n-1)/2] + n/2 + 2$.

Combining the discussions above, we finally obtain

$$|E_2| \le C||\delta\sigma||_{\tau+1}||\delta\sigma||_{l_1}||\Psi|| \tag{4.24}$$

with the constant C depending on $||\sigma_0||_\tau$.

5. Continuity of the linearized map. We now sketch the proof of Theorem 1.4. From the definition of the linearized forward map (1.7), we have

$$[DF(\sigma_1) - DF(\sigma_0)](\eta) = (\phi(\delta u_1 - \delta u_0))|_{x_n=0}\ ,$$

where for $i = 0, 1$

$$\begin{aligned}(\Box - \nabla\sigma_i \cdot \nabla)\delta u_i &= \nabla\eta \cdot \nabla u_i\ , \\ \delta u_i &= 0 \quad t < 0\ .\end{aligned} \tag{5.1}$$

Thus

$$\begin{aligned}||[DF(\sigma_1) - DF(\sigma_0)](\eta)||_l &\le ||(\phi(\delta u_1 - \delta u_0))|_{x_n=0}||_l \\ &\le C||(\phi(\delta v_1 - \delta v_0))|_{x_n=0}||_{l_1}\ ,\end{aligned} \tag{5.2}$$

where $\delta u_i = \partial_t^{[(n-1)/2]}\delta v_i$,

$$\begin{aligned}(\Box - \nabla\sigma_i \cdot \nabla)\delta v_i &= \nabla\eta \cdot \nabla v_i\ , \\ \delta v_i &= 0, \quad t < 0\ ,\end{aligned} \tag{5.3}$$

and

$$(\Box - \nabla\sigma_i \cdot \nabla)v_i = \delta^{(-\frac{n-1}{2})}(t)\delta(x) \ , \quad v_i = 0, \quad t < 0 \ . \tag{5.4}$$

From (5.3), similar to (4.5), we have ($\rho_1(x) = e^{-\sigma_1}$)

$$\Box_1(\delta v_1 - \delta v_0) = \rho_1^{-1}\nabla\eta \cdot \nabla(\delta v_1 - \delta v_0) - \rho_1^{-1}\nabla(\sigma_1 - \sigma_0) \cdot \nabla\delta v_0 \ , \tag{5.5}$$

$$\delta v_1 - \delta v_0 = 0. \tag{5.6}$$

Once again, by considering the dual problem (4.6), we wish to estimate

$$|(\partial_t^{l_1}(\delta v_1 - \delta v_0), \Psi)|/||\Psi||_0 \ .$$

Formal integration by parts yields

$$\begin{aligned}(\partial_t^{l_1}(\delta v_1 - \delta v_0), \Psi) &= (\rho_1^{-1}\nabla\eta \cdot \partial_t^{l_1}(v_1 - v_0) - \rho_1^{-1}\nabla(\sigma_1 - \sigma_0) \cdot \nabla\partial_t^{l_1}\delta v_0, w) \\ &= (\rho_1^{-1}\nabla\eta \cdot \nabla(v_1 - v_0), \partial_t^{l_1}(\phi_1 w)) - (\rho_1^{-1}\nabla\eta \cdot \nabla\delta v_0, \partial_t^{l_1}(\phi_1 w)).\end{aligned} \tag{5.7}$$

It suffices to estimate the terms in (5.7). The second term may be estimated the same way as in section 4. We obtain

$$|(\rho_1^{-1}\nabla\eta \cdot \nabla\delta v_0, \partial_t^{l_1}(\phi_1 w))| \le C||\sigma_1 - \sigma_0||_{l_1} \ ||\eta||_{\tau+1} \ ||\Psi||_0 \ . \tag{5.8}$$

To estimate the first term, recall that $v_1 - v_0$ solves

$$(\Box - \nabla\sigma_1 \cdot \nabla)(v_1 - v_0) = \nabla(\sigma_1 - \sigma_0) \cdot \nabla v_0 \ , \quad v_1 - v_0 = 0, \quad t < 0 \ . \tag{5.9}$$

Hence similar estimates as in section 4 yield

$$|(\rho_1^{-1}\nabla\eta \cdot \nabla(v_1 - v_0), \partial_t^{l_1}(\phi_1 w))| \le C||\eta||_{l_1} \ ||\sigma_1 - \sigma_0||_{\tau+1} \ ||\Psi||_0 \ . \tag{5.10}$$

Theorem 1.4 may be proved by combining (5.7)–(5.10).

Acknowledgment. I thank Prof. William W. Symes for providing me with many useful ideas.

REFERENCES

[1] G. Bao and W. Symes, *A trace theorem for solutions of linear partial differential equations*, Math. Methods Appl. Sci., 14 (1991), pp. 553–562.
[2] G. Bao and W. Symes, *Trace regularity result for a second order hyperbolic equation with nonsmooth coefficients*, J. Math. Anal. Appl., 174 (1993), pp. 370–389.
[3] G. Bao and W. Symes, *Time like trace regularity of the wave equation with a nonsmooth principal part*, SIAM J. Math. Anal., 26 (1995), pp. 129–146.
[4] G. Bao and W. Symes, *On the sensitivity of solutions of hyperbolic equation to the coefficients*, Comm. Partial Differential Equations, 21 (1996), pp. 395–422.
[5] M. Beals, *Propagation and Interaction of Singularities in Nonlinear Hyperbolic Problems*, Birkhäuser Boston, Cambridge, MA, 1989.
[6] M. Beals and M. Reed, *Propagation of singularities for hyperbolic pseudodifferential operators and applications to nonlinear problems*, Comm. Pure Appl. Math., 35 (1982), pp. 169–184.
[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2, Wiley-Interscience, New York, 1962.
[8] F. G. Friedlander, *Sound Pulses*, Cambridge University Press, Cambridge, UK, 1958.
[9] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, New York, 1969.

[10] R. M. Lewis and W. Symes, *On the relation between the velocity coefficient and boundary value for solutions of the one-dimensional wave equation*, Inverse Problems, 7 (1991), pp. 597–631.
[11] Rakesh, *A linearized inverse problem for the wave equation*, Comm. Partial Differential Equations, 13 (1988), pp. 573–601.
[12] J. Rauch, *Singularities of solutions to semilinear wave equations*, J. Math. Pures Appl., 58 (1979), pp. 299–308.
[13] P. Sacks and W. Symes, *Uniqueness and continuous dependence for a multidimensional hyperbolic inverse problem*, Comm. Partial Differential Equations, 10 (1985), pp. 635–676.
[14] W. Symes, *Some aspects of inverse problems in several dimensional wave propagation*, in Proc. Conference on Inverse Problems, SIAM-AMS Proceedings 14, D. W. McLaughlin, ed., Amer. Math. Soc., Providence, RI, 1983.
[15] W. Symes, *On the relation between coefficient and boundary values for solutions of Webster's horn equation*, SIAM J. Math. Anal., 17 (1986), pp. 1400–1420.
[16] W. Symes, *Linearization stability for an inverse problem in several-dimensional wave propagation*, SIAM J. Math. Anal., 17 (1986), pp. 132–151.
[17] M. Taylor, *Pseudo-Differential Operators*, Princeton University Press, Princeton, NJ, 1981.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 395–423, March 1998

006

TIME-HARMONIC ELECTROMAGNETIC FIELDS IN THIN CHIRAL CURVED LAYERS*

H. AMMARI† AND J. C. NÉDÉLEC†

Abstract. In this paper, we prove existence and uniqueness of the solution to the diffraction problem of a plane electromagnetic field by a chiral curved layer covering a perfectly conducting object. Approximative impedance conditions are given for thin chiral curved layers and optimal error estimates are obtained.

Key words. integral equations, chiral media, approximative impedance conditions, Drude–Born–Fedorov equations, error estimates

AMS subject classifications. 35C15, 35Q60, 45B05, 45L05

PII. S0036141096305504

1. Introduction.

The interaction of electromagnetic fields at microwave frequencies with the optically active chiral media has attracted the attention of the electromagnetic community due to its potential applications in the field of antennas, microwave devices, waveguide propagation, scattering, etc. The original investigations of optical activity, or the effect of chirality on the polarization of light, also date back to the nineteenth century. Arago [5], Biot [7], Cauchy [12], Pasteur [23], and Fresnel [15] all examined optical activity in solid and liquid chiral medium. The concept of chirality has also played an increasingly important role in chemistry (see Prelog [24]). Chiral media can be characterized by a generalized set of constitutive relations in which the electric and magnetic fields are coupled. The coupling strength is given by the magnitude of a quantity known as the chirality admittance which determines the bulk electromagnetic properties of these materials. In scattering applications it may be possible to use chiral material to coat a scatterer and thus control its scattering properties more efficiently than a dielectric coating due to the extra degree of freedom offered by the presence of chiral parameter. Antennas coated with chiral materials may have significantly interesting radiation characteristics, and surface-relief gratings constructed using chiral material may find application in integrated optics and holography, for instance.

In this paper, we prove uniqueness and existence of the solution to the diffraction problem by a layer of chiral material covering a perfectly conducting object. Assuming that the thickness of the layer is sufficiently small compared to the wavelength of the incident radiation, we derive impedance boundary conditions which can be used, along with the radiation condition, to compute the scattered field without requiring detailed modeling of the field quantities inside the chiral coating. This procedure permits accurate numerical calculations of the scattered wave for chiral coatings as in the achiral case [13]. The starting points of the present paper are the Maxwell equations and the constitutive relations for the chiral layer. Different expressions exist for the constitutive relations. The Drude–Born–Fedorov constitutive equations are used here. For an interesting explanation of these equations and various physical

*Received by the editors June 24, 1996; accepted for publication (in revised form) February 6, 1997.

http://www.siam.org/journals/sima/29-2/30550.html

†Centre de Mathématiques Appliquées, CNRS URA 756, École Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr, nedelec@barbes.polytechnique.fr).

aspects of the propagation inside chiral media, we refer to [14], [8], [9], [10], [6], [17], [18], [20], [21], [22], [25], [16], and the references therein. It is noted that this list of papers is by no means complete, but it is representative of the work that is presently being conducted in this area. All these papers have come from physicists and the engineering community. To our knowledge, the existence and the uniqueness of the solution to the diffraction problem by a chiral layer has never been established. The present paper is devoted to presenting some results and approaches dealing with the rigorous scattering by a chiral medium.

2. Chirality and reduction of scattering. Let δ be a strictly positive parameter. Let $\Omega^i \subset \mathbb{R}^3$ be an open bounded set of class $\mathcal{C}^\infty$, with $\Gamma = \partial\Omega^i$ its boundary, n the outward normal. Let Ω_1^δ be the chiral layer, $\Omega_2^\delta = \Omega^e \setminus \overline{\Omega_1^\delta}$, Γ^δ the interface between the chiral material Ω_1^δ and the exterior domain Ω_2^δ. We assume that Γ^δ is parallel to the curve Γ. Let $\Omega_{2,R}^\delta = \Omega_2^\delta \cap \{r < R\}$ and $\Sigma_R = \{r = R\}$. For $x \in \Omega^e = \mathbb{R}^3 \setminus \overline{\Omega^i}$ sufficiently close to Γ^δ, we denote by x_{Γ^δ} the orthogonal projection of x on Γ^δ, $s = -(x - x_{\Gamma^\delta}) \cdot n(x_{\Gamma^\delta})$, where $n(x_{\Gamma^\delta})$ is the outward normal to Γ^δ. (x_{Γ^δ}, s) is a parameterization of the neighborhood of Γ^δ:

$$x(x_{\Gamma^\delta}, s) = x_{\Gamma^\delta} + s\, n(x_{\Gamma^\delta}).$$

The curve Γ is such that

$$\Gamma = \Big\{ x_{\Gamma^\delta} - \delta n(x_{\Gamma^\delta}), x_{\Gamma^\delta} \in \Gamma^\delta \Big\}.$$

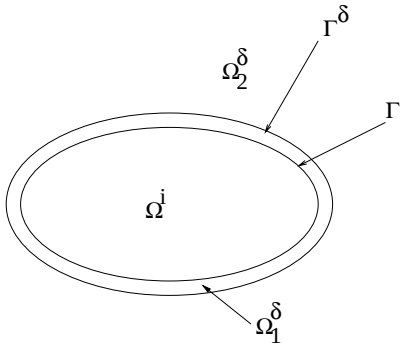

FIG. 2.1. *The chiral layer.*

We shall need some differential operators on the curve Γ^δ that we will introduce. The normal vector on Γ^δ is the opposite of the gradient of the function s. We use this property to define

$$n(x) = -\overrightarrow{\text{grad}}\, s(x) = n(x_{\Gamma^\delta}).$$

The mean curvature is defined as $2\, c(x_{\Gamma^\delta}) = \text{div}\, n$. We also introduce the family of parallel curves Γ_s^δ :

$$\Gamma_s^\delta = \Big\{ \text{ } x \in \Omega^e \text{ such that } x = x_{\Gamma^\delta} + s\, n(x_{\Gamma^\delta}) \text{ , } x_{\Gamma^\delta} \in \Gamma^\delta \text{ } \Big\}.$$

The vector field n is the field of the normal vectors to the curves Γ_s^δ. With any function f defined on the curve Γ^δ, we associate the function $\tilde{f}$ defined for x near Γ^δ

by

$$\tilde{f}(x) \,=\, f(x_{\Gamma^\delta}).$$

We also define $\tilde{u}$, for any smooth tangential vector field u on Γ^δ, as

$$\tilde{u}(x_{\Gamma^\delta} + sn(x_{\Gamma^\delta})) \,=\, u(x_{\Gamma^\delta}) - s\,\mathcal{R}\,u(x_{\Gamma^\delta}),$$

where $\mathcal{R}$ is the tensor of curvature at the point x_{Γ^δ}. It is defined as $\mathcal{R} \,=\, \overrightarrow{\operatorname{grad}}\, n$. The tangential gradient of the tangential field u is

$$\overrightarrow{\operatorname{grad}}_{\Gamma^\delta} u \,=\, \overrightarrow{\operatorname{grad}}\,\tilde{u}\,|_{\Gamma^\delta}\ .$$

This vector belongs to the tangent to the curve Γ^δ. For any smooth tangential vector field u on Γ^δ, let us define $\operatorname{div}_{\Gamma^\delta} u$ as

$$\operatorname{div}_{\Gamma^\delta} u \,=\, \operatorname{div}\tilde{u}\,|_{\Gamma^\delta}\ .$$

Let $\overrightarrow{\operatorname{curl}}_{\Gamma^\delta}$ denote the tangential rotational operator on Γ^δ. It is such that

$$\overrightarrow{\operatorname{curl}}_{\Gamma^\delta} u \,=\, \overrightarrow{\operatorname{grad}}_{\Gamma^\delta} u \,\wedge\, n$$

for any function u defined on the curve Γ^δ. The scalar tangential rotational operator is defined as

$$\operatorname{curl}_{\Gamma^\delta} u \,=\, \operatorname{curl}\tilde{u} \,\cdot\, n\,|_{\Gamma^\delta},$$

where u is a tangential vector to Γ^δ and $\tilde{u}(x) \,=\, u(x_{\Gamma^\delta})$. It is well known that

$$\operatorname{curl}_{\Gamma^\delta} u \,=\, \operatorname{div}_{\Gamma^\delta}(u \wedge n). \tag{2.1}$$

Finally, we recall the following formula.

LEMMA 2.1. *Let u be a smooth vectorial function defined in Ω; we have*

$$\tau_0(\operatorname{div} u) \,=\, \operatorname{div}_{\Gamma^\delta} u_{\Gamma^\delta} + 2\,c(x_{\Gamma^\delta})\; u \,\cdot\, n + \tau_1(u\,.\,n) \tag{2.2}$$

and

$$\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}(u \,\cdot\, n) \,=\, \tau_1(u_{\Gamma^\delta}) + n \wedge \operatorname{curl} u + \mathcal{R}(u_{\Gamma^\delta}), \tag{2.3}$$

where τ_0 and τ_1 are, respectively, the first and the second trace and u_{Γ^δ} is the tangential component of u.

Optical activity of the chiral layer Ω_1^δ of thickness δ can be explained by the direct substitution of the Drude–Born–Fedorov constitutive equations

$$\begin{cases} D^\delta &=& \varepsilon^\delta\Big(E^\delta + \beta^\delta \overrightarrow{\operatorname{curl}}\, E^\delta\Big), \\ B^\delta &=& \mu^\delta\Big(H^\delta + \beta^\delta \overrightarrow{\operatorname{curl}}\, H^\delta\Big) \end{cases} \tag{2.4}$$

into Maxwell's equations

$$\begin{cases} \overrightarrow{\operatorname{curl}}\, E^\delta &=& i\omega\, B^\delta, \\ \overrightarrow{\operatorname{curl}}\, H^\delta &=& -i\omega\, D^\delta. \end{cases} \tag{2.5}$$

It follows that the equivalent monochromatic constitutive relations

$$\left\{\begin{array}{rcl}\Big(1-(k^\delta\beta^\delta)^2\Big)\,D^\delta &=& \varepsilon^\delta E^\delta+\dfrac{i\beta^\delta}{\omega}(k^\delta)^2\,H^\delta,\\ \Big(1-(k^\delta\beta^\delta)^2\Big)\,B^\delta &=& \mu^\delta H^\delta-\dfrac{i\beta^\delta}{\omega}(k^\delta)^2\,E^\delta,\end{array}\right. \tag{2.6}$$

where the function ε^δ is the electric permittivity, μ^δ is the magnetic permeability, β^δ is the chirality admittance, and

$$k^\delta \,=\, \omega\sqrt{\varepsilon^\delta\,\mu^\delta} \tag{2.7}$$

is simply a shorthand notation and does not represent any wavenumber inside the chiral medium. Throughout what follows we assume that $|k^\delta\beta^\delta|\neq 1$. We also assume that $\varepsilon^\delta=\varepsilon_2,\mu^\delta=\mu_2$, and $\beta^\delta=0$ outside the chiral layer Ω_1^δ where ε_2 and μ_2 are two strictly positive constants. Now, combining (2.5) and (2.6) gives

$$\left\{\begin{array}{rcl}\overrightarrow{\rm curl}\,E^\delta &=& (\gamma^\delta)^2\,\beta^\delta E^\delta+i\omega\mu^\delta\left(\dfrac{\gamma^\delta}{k^\delta}\right)^2\,H^\delta,\\ \overrightarrow{\rm curl}\,H^\delta &=& (\gamma^\delta)^2\,\beta^\delta\,H^\delta-i\omega\varepsilon^\delta\left(\dfrac{\gamma^\delta}{k^\delta}\right)^2\,E^\delta.\end{array}\right. \tag{2.8}$$

In these equations, the parameter γ^δ is defined as

$$(\gamma^\delta)^2 \,=\, \frac{(k^\delta)^2}{1-(\,k^\delta\,\beta^\delta\,)^2}. \tag{2.9}$$

We are interested in the scattering problem for a plane incident field E^{in} defined as $E^{in}(x)=p\,e^{i\,\omega\,q\,.\,x}$, where the vectors $p,q\in\mathbb{R}^3$ must satisfy $q\,.\,q=\mu_2\varepsilon_2$ and $q\,.\,p=0$. Then the pair $(E^{in},\overrightarrow{\rm curl}E^{in}/i\omega\mu_2)$ satisfies the Maxwell equations in Ω_2^δ and our scattering problem can be formulated as follows.

$$\left\{\begin{array}{l}\overrightarrow{\rm curl}\,\dfrac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\rm curl}\,E^\delta-\omega^2\,\beta^\delta\varepsilon^\delta\overrightarrow{\rm curl}\,E^\delta-\omega^2\,\overrightarrow{\rm curl}\,(\beta^\delta\varepsilon^\delta)\,E^\delta\\ \qquad -\omega^2\,\varepsilon^\delta\,E^\delta\,=\,0\ \text{in}\ \Omega^e,\\ n\,\wedge\,E^\delta\,=\,0\quad\text{on}\ \Gamma,\\ E^\delta-E^{in}\ \text{satisfies the classical radiation condition.}\end{array}\right. \tag{2.10}$$

3. Uniqueness and existence results for the diffraction problem by the chiral layer. This section is devoted to proving the existence and uniqueness of a field E^δ satisfying the equations (2.10). We use the technique of boundary integral equations. We shall reduce the Drude–Born–Fedorov equations (2.10) to a boundary integral equation on $\Gamma^{\delta,h}=\{x=(x_{\Gamma^\delta},s=h)\}$. First, we consider the following boundary-value problems:

$$\left\{\begin{array}{l}\overrightarrow{\rm curl}\,\dfrac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\rm curl}\,e^\delta-\omega^2\,\beta^\delta\varepsilon^\delta\overrightarrow{\rm curl}\,e^\delta-\omega^2\,\overrightarrow{\rm curl}\,\beta^\delta\varepsilon^\delta\,e^\delta\\ \qquad -\omega^2\,\varepsilon^\delta\,e^\delta\,=\,0\ \text{in}\ \Omega_1^{\delta,h},\\ n\,\wedge\,e^\delta\,=\,0\quad\text{on}\ \Gamma,\\ n\,\wedge\,e^\delta\,=\,g^\delta\quad\text{on}\ \Gamma^{\delta,h}=\Big\{x=(x_{\Gamma^\delta},s=h)\Big\},\end{array}\right. \tag{3.1}$$

and

$$\text{(3.2)}\qquad \begin{cases} \overrightarrow{\text{curl}}\,\overrightarrow{\text{curl}} v^\delta - \omega^2\,\varepsilon_2\mu_2\,v^\delta \;=\; 0 \quad \text{in } \Omega_2^{\delta,h}, \\ n \wedge v^\delta \;=\; g^\delta \quad \text{on } \Gamma^{\delta,h} (=\partial\Omega_2^{\delta,h}), \\ v^\delta \text{ satisfies the outgoing radiation condition at infinity,} \end{cases}$$

where

$$g^\delta \in TH^{1/2}(\text{div}, \Gamma^{\delta,h}) = \left\{ \, c \in (H^{1/2}(\Gamma^{\delta,h}))^3, n \cdot c = 0 \text{ on } \Gamma^{\delta,h}, \text{div}_{\Gamma^{\delta,h}}\, c \in H^{1/2}(\Gamma^{\delta,h}) \, \right\}$$

is a known vector field. h is an arbitrarily strictly positive constant which must be small enough. $\Omega_1^{\delta,h}$ and $\Omega_2^{\delta,h}$ are given by $\Omega_1^{\delta,h} = \{(x_{\Gamma^\delta}, s), 0 < s < h\}$, and $\Omega_2^{\delta,h} = \mathbb{R}^3 \setminus \overline{\Omega_1^{\delta,h} \cup \Omega^i}$.

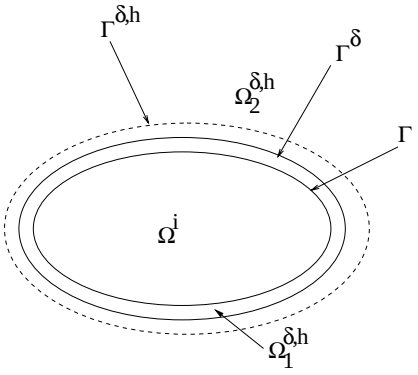

FIG. 3.1. *The chiral layer.*

The weak form of the Drude–Born–Fedorov equations (3.1) in a neighborhood of the interface Γ^δ gives the following jump relations on Γ^δ:

$$\text{(3.3)}\qquad \begin{cases} \left(e_1^\delta \,-\, e_2^\delta \right) \wedge n \;=\; 0, \\ \left(\overrightarrow{\text{curl}}\, e_1^\delta - \dfrac{\mu_1}{\mu_2} \overrightarrow{\text{curl}}\, e_2^\delta \right) \wedge n \;=\; 0, \end{cases}$$

where $e_1^\delta = e^\delta\,|_{\Omega_1^\delta}$ and $e_2^\delta = e^\delta\,|_{\Omega_1^{\delta,h} \setminus \overline{\Omega_1^\delta}}$, if we assume that the jump of β^δ across the interface Γ^δ is null.

Let

$$X^\delta = \left\{ \, H(\overrightarrow{\text{curl}}, \Omega_1^{\delta,h}), w^\delta \wedge n = 0 \text{ on } \partial\Omega_1^{\delta,h} = \Gamma \cup \Gamma^{\delta,h} \, \right\},$$

where

$$H(\overrightarrow{\text{curl}}, \Omega_1^{\delta,h}) = \left\{ \, w^\delta \in (L^2(\Omega_1^{\delta,h}))^3, \overrightarrow{\text{curl}}\, w^\delta \in (L^2(\Omega_1^{\delta,h})^3 \, \right\}.$$

Let $\tilde{g}^\delta$ be in $(H^1(\Omega_1^{\delta,h}))^3$ such that $n \wedge \tilde{g}^\delta = g^\delta$ on $\Gamma^{\delta,h}$ and $n \wedge \tilde{g}^\delta = 0$ on Γ. We define $\tilde{e}^\delta$ by

$$\tilde{e}^\delta = e^\delta - \tilde{g}^\delta.$$

It is easily seen that

$$
\begin{aligned}
&\overrightarrow{\operatorname{curl}}\,\frac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\operatorname{curl}}\,\tilde{e}^\delta-\omega^2\,\beta^\delta\varepsilon^\delta\overrightarrow{\operatorname{curl}}\,\tilde{e}^\delta-\omega^2\,\overrightarrow{\operatorname{curl}}\,\beta^\delta\varepsilon^\delta\,\tilde{e}^\delta-\omega^2\,\varepsilon^\delta\,\tilde{e}^\delta\\
&=-\overrightarrow{\operatorname{curl}}\,\frac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\operatorname{curl}}\,\tilde{g}^\delta+\omega^2\,\beta^\delta\varepsilon^\delta\overrightarrow{\operatorname{curl}}\,\tilde{g}^\delta+\omega^2\,\overrightarrow{\operatorname{curl}}\,\beta^\delta\varepsilon^\delta\,\tilde{g}^\delta\\
&+\omega^2\,\varepsilon^\delta\,\tilde{g}^\delta\ \text{in}\ (\mathcal{D}'(\Omega_1^{\delta,h}))^3,
\end{aligned}
\tag{3.4}
$$

where $\mathcal{D}'(\Omega_1^{\delta,h})$ is the space of all distributions in $\Omega_1^{\delta,h}$. If we multiply (3.4) by a test function $v^\delta\in X^\delta$, integrate over $\Omega_1^{\delta,h}$, and use integration by parts and the jump relations (3.3), we obtain

$$
\begin{aligned}
&\int_{\Omega_1^{\delta,h}}\frac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\operatorname{curl}}\,\tilde{e}^\delta\cdot\overrightarrow{\operatorname{curl}}\,v^\delta+\omega^2\int_{\Omega_1^{\delta,h}}\beta^\delta\varepsilon^\delta\overrightarrow{\operatorname{curl}}\,\tilde{e}^\delta\cdot v^\delta\\
&-\omega^2\int_{\Omega_1^{\delta,h}}\beta^\delta\varepsilon^\delta\,\tilde{e}^\delta\cdot\overrightarrow{\operatorname{curl}}\,v^\delta-\omega^2\int_{\Omega_1^{\delta,h}}\varepsilon^\delta\,\tilde{e}^\delta\cdot v^\delta\\
&=-\int_{\Omega_1^{\delta,h}}\frac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\operatorname{curl}}\,\tilde{g}^\delta\cdot\overrightarrow{\operatorname{curl}}\,v^\delta-\omega^2\int_{\Omega_1^{\delta,h}}\beta^\delta\varepsilon^\delta\overrightarrow{\operatorname{curl}}\,\tilde{g}^\delta\cdot v^\delta\\
&+\omega^2\int_{\Omega_1^{\delta,h}}\beta^\delta\varepsilon^\delta\,\tilde{g}^\delta\cdot\overrightarrow{\operatorname{curl}}\,v^\delta+\omega^2\int_{\Omega_1^{\delta,h}}\varepsilon^\delta\,\tilde{g}^\delta\cdot v^\delta.
\end{aligned}
\tag{3.5}
$$

Our variational problem is then to find the vector field $e^\delta\in H(\overrightarrow{\operatorname{curl}},\Omega_1^{\delta,h})$ such that (3.5) holds for all $v^\delta\in X^\delta$. Because the injection $X^\delta\hookrightarrow(L^2(\Omega^{\delta,h}))^3$ is not compact, we cannot apply the Fredholm theory to (3.5) to show the existence and uniqueness of a solution of this problem in X^δ. We shall prove Lemma 3.1 by using the generalized Lax–Milgram lemma of Babuška and Brezzi [11]. We discuss the unique solvability of (3.5) and (3.2). We may state and prove the following results.

LEMMA 3.1. *Assume that* (3.1) *has at most one solution. Then the variational problem* (3.5) *has a unique solution in* $H(\overrightarrow{\operatorname{curl}},\Omega_1^{\delta,h})$.

LEMMA 3.2. *The boundary-value problem* (3.2) *has a unique solution in* $H^{1,loc}(\Omega_2^{\delta,h})$.

It is well known that (3.2) has exactly one solution (cf. [2] for instance). Now, our aim is to sketch a proof of Lemma 3.1. By essentially the same arguments as in Abboud [1], we can establish the Fredholm alternative for the variational problem (3.5). We seek for $e^\delta\in H(\overrightarrow{\operatorname{curl}},\Omega_1^{\delta,h})$ and we decompose it into a field and a gradient:

$$
e^\delta=u^\delta+\overrightarrow{\operatorname{grad}}p^\delta+\tilde{g}^\delta,
\tag{3.6}
$$

where $u^\delta\in H(\overrightarrow{\operatorname{curl}},\Omega_1^{\delta,h})$ satisfies

$$
\operatorname{div}(\varepsilon^\delta\,u^\delta-\beta^\delta\,\varepsilon^\delta\,\overrightarrow{\operatorname{curl}}\,u^\delta)=0\quad\text{in}\ \Omega_1^{\delta,h},
\tag{3.7}
$$

$u^\delta\wedge n=0$ on $\Gamma\cup\Gamma^{\delta,h}$ and $p^\delta\in H_0^1(\Omega_1^{\delta,h})$. (3.5) will lead to a saddle point problem involving the form a^δ defined as

$$
\begin{aligned}
a^\delta(u^\delta,w^\delta)=&\int_{\Omega_1^{\delta,h}}\frac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\operatorname{curl}}\,u^\delta\cdot\overrightarrow{\operatorname{curl}}\,\overline{w}^\delta\\
&+\omega^2\int_{\Omega_1^{\delta,h}}\beta^\delta\,\varepsilon^\delta\,u^\delta\cdot\overrightarrow{\operatorname{curl}}\,\overline{w}^\delta\\
&-\omega^2\int_{\Omega_1^{\delta,h}}\beta^\delta\,\varepsilon^\delta\,\overrightarrow{\operatorname{curl}}\,u^\delta\cdot\overline{w}^\delta-\omega^2\int_{\Omega_1^{\delta,h}}\varepsilon^\delta\,u^\delta\cdot\overline{w}^\delta
\end{aligned}
$$

for all $w^\delta \in X^\delta$ and the form b^δ given as

$$\begin{aligned} b^\delta(w^\delta, q^\delta) &= a^\delta(\overrightarrow{\text{grad}} q^\delta, w^\delta) \\ &= \omega^2 \int_{\Omega_1^{\delta,h}} \beta^\delta \, \varepsilon^\delta \, \overrightarrow{\text{grad}} q^\delta \cdot \overrightarrow{\text{curl}}\, \overline{w}^\delta - \omega^2 \int_{\Omega_1^{\delta,h}} \varepsilon^\delta \, \overrightarrow{\text{grad}} q^\delta \cdot \overline{w}^\delta \end{aligned}$$

for all $w^\delta \in X^\delta$ and $q^\delta \in H_0^1(\Omega_1^{\delta,h})$. Now we show how to split the variational problem (3.5) into two distinct equations. First, by multiplying (3.7) by $q \in H_0^1(\Omega_1^{\delta,h})$ and integrating by parts over $\Omega_1^{\delta,h}$, we obtain

$$b^\delta(u^\delta, q^\delta) = 0 \quad \text{for all } q^\delta \in H_0^1(\Omega_1^{\delta,h}).$$

Second, multiplying (3.4) by $w \in X^\delta$ and integrating by parts over $\Omega_1^{\delta,h}$ give by using the decomposition (3.6) of the electric field e^δ

$$a^\delta(u^\delta, w^\delta) + b^\delta(w^\delta, p^\delta) = -a^\delta(\tilde{g}^\delta, w^\delta) \quad \text{for all } w^\delta \in X^\delta.$$

Then, we obtain

$$\text{(3.8)} \qquad \begin{cases} a^\delta(u^\delta, w^\delta) + b^\delta(w^\delta, p^\delta) = -a^\delta(\tilde{g}^\delta, w^\delta) & \text{for all } w^\delta \in X^\delta, \\ b^\delta(u^\delta, q^\delta) = 0 & \text{for all } q^\delta \in H_0^1(\Omega_1^{\delta,h}). \end{cases}$$

It is clear that there exist two strictly positive constants C_1 and C_2 such that

$$|a^\delta(u^\delta, u^\delta)| \geq C_1 \, ||\overrightarrow{\text{curl}}\, u^\delta||^2_{(L^2(\Omega_1^{\delta,h}))^3} - C_2 \, ||u^\delta||^2_{(L^2(\Omega_1^{\delta,h}))^3}.$$

It is easily shown that the form b^δ satisfies the inf-sup condition: for each q^δ in $H_0^1(\Omega_1^{\delta,h})$, it is clear that $\overrightarrow{\text{grad}}\, q^\delta$ satisfies $\overrightarrow{\text{grad}} q^\delta \wedge n = 0$ on $\Gamma \cup \Gamma^\delta$ and there exists a strictly positive constant C_3 such that

$$|b^\delta(\overrightarrow{\text{grad}} q^\delta, q^\delta)| \geq C_3 \; ||q^\delta||^2_{H_0^1(\Omega_1^{\delta,h})}.$$

On the other hand, if we assume that $\overrightarrow{\text{grad}}(\beta^\delta \varepsilon^\delta) \in (L^\infty(\Omega_1^{\delta,h}))^3$ then the imbedding $N^\delta \hookrightarrow (L^2(\Omega_1^{\delta,h}))^3$ is compact, where

$$N^\delta = \Big\{ u^\delta \in H(\overrightarrow{\text{curl}}, \Omega_1^{\delta,h}), \text{div}\,(\, \varepsilon^\delta \, u^\delta - \beta^\delta \, \varepsilon^\delta \, \overrightarrow{\text{curl}}\, u^\delta) = 0, u^\delta \wedge n = 0 \text{ on } \partial\Omega_1^{\delta,h} \Big\}$$

is the kernel of the form b^δ. Therefore, a^δ is a compact perturbation of a coercive form on the kernel N^δ of b^δ. It follows that the Fredholm alternative holds, and then uniqueness implies existence of the solution of (3.8). In order to complete the proof of Lemma 3.1 we should prove the following decomposition lemma which may be considered as a generalization of the Hodge decomposition.

LEMMA 3.3. *Let $\tilde{e}^\delta$ be in*

$$X^\delta = \Big\{ v^\delta \in H(\overrightarrow{\text{curl}}, \Omega_1^{\delta,h}), v^\delta \wedge n = 0 \text{ on } \partial\Omega_1^{\delta,h} = \Gamma^{\delta,h} \cup \Gamma \Big\}.$$

Then, there exist $u^\delta \in N^\delta$ and $p^\delta \in H_0^1(\Omega_1^{\delta,h})$ such that

$$\text{(3.9)} \qquad \tilde{e}^\delta = u^\delta + \nabla p^\delta \quad \text{in } \Omega_1^{\delta,h}.$$

Proof. We define f^δ by

$$f^\delta \;=\; \varepsilon^\delta \beta^\delta \,\overrightarrow{\mathrm{curl}}\,\tilde{e}^\delta - \varepsilon^\delta\,\tilde{e}^\delta.$$

$\tilde{e}^\delta \in H(\overrightarrow{\mathrm{curl}}, \Omega^{\delta,h})$ yields $f^\delta \in (L^2(\Omega^{\delta,h}))^3$. It is well known that the transmission problem

$$\begin{cases} \operatorname{div} \varepsilon^\delta \nabla \varphi^\delta_{c_1,c_2} \;=\; \operatorname{div} f^\delta \quad \text{in } \Omega_1^{\delta,h}, \\ \varphi^\delta_{c_1,c_2} \;=\; c_1 \quad \text{on } \Gamma^{\delta,h}, \\ \varphi^\delta_{c_1,c_2} \;=\; c_2 \quad \text{on } \Gamma \end{cases}$$

has a unique solution in $H^1(\Omega^{\delta,h})$ for any $(c_1, c_2) \in \mathbb{C} \times \mathbb{C}$. Let p^δ be defined by

$$p^\delta \;=\; \varphi^\delta_{0,0} \in H^1_0(\Omega_1^{\delta,h})$$

and u^δ given by

$$u^\delta \;=\; \tilde{e}^\delta - \nabla\,\varphi^\delta_{0,0}.$$

It is clear that u^δ is in N^δ; so, the decomposition (3.9) holds. □

Now, using Lemma 3.3, the uniqueness of a solution in $H(\overrightarrow{\mathrm{curl}}, \Omega_1^{\delta,h})$ to (3.1) gives the uniqueness of a solution to (3.8). The proof of Lemma 3.1 is then over.

Throughout what follows we assume that there exists $h > 0$ such that 0 is not an eigenvalue of the Drude–Born–Fedorov operator

$$\overrightarrow{\mathrm{curl}}\,\frac{1}{\mu^\delta}\,(1 - \omega^2 \varepsilon^\delta \mu^\delta (\beta^\delta)^2)\,\overrightarrow{\mathrm{curl}}\,\cdot - \omega^2\,\beta^\delta\,\varepsilon^\delta\,\overrightarrow{\mathrm{curl}}\,\cdot - \omega^2\,\overrightarrow{\mathrm{curl}}\,\beta^\delta\,\varepsilon^\delta\,\cdot - \omega^2\,\varepsilon^\delta\,\cdot \quad \text{in } \Omega_1^{\delta,h}$$

with the boundary condition $n \wedge \cdot = 0$ on $\Gamma^{\delta,h} \cup \Gamma$. The existence of such h can be verified, but it is not our purpose here.

Now, in order to reduce the diffraction problem by a chiral layer covering a perfectly conducting object to a boundary integral equation on $\Gamma^{\delta,h}$, we consider the two pseudodifferential operators on $\Gamma^{\delta,h}$ defined as

$$T_1(\beta,\omega)\;: g^\delta \in TH^{1/2}(\operatorname{div}, \Gamma^{\delta,h}) \longmapsto \overrightarrow{\mathrm{curl}}\, e^\delta \wedge n \in TH^{1/2}(\operatorname{div}, \Gamma^{\delta,h}), \tag{3.10}$$

where e^δ is the unique solution of (3.1), and

$$T_2(\omega)\;: g^\delta \in TH^{1/2}(\operatorname{div}, \Gamma^{\delta,h}) \longmapsto \overrightarrow{\mathrm{curl}}\, v^\delta \wedge n \in TH^{1/2}(\operatorname{div}, \Gamma^{\delta,h}), \tag{3.11}$$

where v^δ is the unique solution of (3.2) satisfying the radiation condition. We have to show that $T_1(\beta,\omega)\,(g^\delta)$ and $T_2(\omega)\,(g^\delta)$ are in $TH^{1/2}(\operatorname{div}, \Gamma^{\delta,h})$ for all g^δ in $TH^{1/2}(\operatorname{div}, \Gamma^{\delta,h})$. Let us define h^δ in $\Omega_1^{\delta,h}$ by

$$\overrightarrow{\mathrm{curl}}\, e^\delta \;=\; (\gamma^\delta)^2\,\beta^\delta\, e^\delta + i\omega\,\mu^\delta \left(\frac{\gamma^\delta}{k^\delta}\right)^2 h^\delta.$$

(3.1) yields a magnetic boundary-value problem for h^δ, and a solution can be found through a variational method similar to (3.5). e^δ and h^δ are in $(H^1(\Omega_1^{\delta,h} \cap \Omega_2^{\delta,h'}))^3$ for all $0 < h' < h$. By the trace theorem, we find that

$$\overrightarrow{\mathrm{curl}}\, e^\delta \wedge n \in TH^{1/2}(\Gamma^{\delta,h}),$$

and furthermore,

$$\begin{aligned}\mathrm{div}_{\Gamma^{\delta,h}}\left(\overrightarrow{\mathrm{curl}}\, e^\delta \wedge n\right) &= \overrightarrow{\mathrm{curl}}\,\overrightarrow{\mathrm{curl}}\, e^\delta \cdot n \\ &= \omega^2\varepsilon_2\mu_2\, e^\delta \cdot n\,|_{\Gamma^{\delta,h}} \in H^{1/2}(\Gamma^{\delta,h}).\end{aligned}$$

Similarly, we define w^δ by

$$\overrightarrow{\mathrm{curl}}\, v^\delta = i\omega\mu_2\, w^\delta.$$

w^δ is the unique outgoing solution of the Maxwell equations

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,\overrightarrow{\mathrm{curl}}\, w^\delta - \omega^2\varepsilon_2\mu_2\, w^\delta = 0 \quad \text{in } \Omega_2^{\delta,h}, \\ i\omega\mu_2 w^\delta \cdot n = \mathrm{div}_{\Gamma^{\delta,h}}\, g^\delta \in H^{1/2}(\Gamma^{\delta,h}), \quad \text{on } \Gamma^{\delta,h}. \end{cases}$$

It is well known that, under the regularity assumption $g^\delta \in H^{1/2}(\Gamma^{\delta,h})$, v^δ and w^δ are in $(H^{1,loc}(\Omega_2^{\delta,h}))^3$. Thus we have

$$\overrightarrow{\mathrm{curl}}\, v^\delta \wedge n = i\omega\mu_2\, w^\delta \wedge n \in TH^{1/2}(\Gamma^{\delta,h})$$

and

$$\mathrm{div}_{\Gamma^{\delta,h}}\left(\overrightarrow{\mathrm{curl}}\, v^\delta \wedge n\right) = \omega^2\varepsilon_2\mu_2\, v^\delta \cdot n\,|_{\Gamma^{\delta,h}} \in H^{1/2}(\Gamma^{\delta,h}).$$

The jump relations lead to the boundary integral equation

$$T_1(\beta,\omega)(g^\delta) - T_2(\omega)(g^\delta) = g^{in} \quad \text{on } \Gamma^{\delta,h}, \tag{3.12}$$

where

$$g^{in} = T_2(\omega)(n \wedge E^{in}) - \overrightarrow{\mathrm{curl}}\, E^{in} \wedge n$$

and

$$g^\delta = n \wedge E^\delta;$$

E^δ satisfies (2.10). Using Lemmas 3.1 and 3.2, it is clear that the unique solvability of (2.8) is equivalent to the unique solvability of (3.12). To discuss existence and uniqueness of solutions of (3.12) let us introduce the following operators:

$$L_0 = T_1(\beta = 0, \omega) - T_2(\omega),$$

$$L_1 = T_1(\beta,\omega) - T_2(\omega),$$

and

$$L_t = t\, L_1 + (1-t)\, L_0 \quad \text{for all } 0 \le t \le 1.$$

The diffraction problem of Maxwell's equations by a layer of dielectric achiral material has a unique solution (cf. [2]). Therefore, we have the following result.

LEMMA 3.4. *The operator* $L_0 : TH^{1/2}(\mathrm{div}, \Gamma^{\delta,h}) \mapsto TH^{1/2}(\mathrm{div}, \Gamma^{\delta,h})$ *is an isomorphism.*

Proof. First we prove the injectivity of the operator L_0. Let $g \in TH^{1/2}(\Gamma^{\delta,h})$ be a solution to

$$L_0 g = 0. \tag{3.13}$$

Let E_1 be the unique solution of the Maxwell equations

$$\begin{cases} \overrightarrow{\text{curl}}\, \dfrac{1}{\mu^\delta}\, \overrightarrow{\text{curl}}\, E_1 - \omega^2 \varepsilon^\delta\, E_1 = 0 \quad \text{in } \Omega_1^{\delta,h}, \\ n \wedge E_1 = 0 \quad \text{on } \Gamma, \\ n \wedge E_1 = g \quad \text{on } \Gamma^{\delta,h}. \end{cases}$$

Let E_2 be given as the unique solution to the following problem:

$$\begin{cases} \overrightarrow{\text{curl}}\, \overrightarrow{\text{curl}}\, E_2 - \omega^2 \varepsilon_2 \mu_2\, E_2 = 0 \quad \text{in } \Omega_2^{\delta,h}, \\ n \wedge E_2 = g \quad \text{on } \Gamma^{\delta,h}, \\ E_2 \text{ satisfies the outgoing radiation condition at infinity.} \end{cases}$$

The field E_2 is in $(H^{1,loc}(\Omega_2^\delta))^3$. We define the field E by

$$E = \begin{cases} E_1 \quad \text{in } \Omega_1^{\delta,h}, \\ E_2 \quad \text{in } \Omega_2^{\delta,h}. \end{cases}$$

$L_0 g = 0$ implies that the jump of $\overrightarrow{\text{curl}}\, E \wedge n$ across $\Gamma^{\delta,h}$ is null. It follows that the field E satisfies the Maxwell equations

$$\begin{cases} \overrightarrow{\text{curl}}\, \dfrac{1}{\mu^\delta}\, \overrightarrow{\text{curl}}\, E - \omega^2 \varepsilon^\delta\, E = 0 \quad \text{in } \Omega^e, \\ n \wedge E = 0 \quad \text{on } \Gamma, \\ E \text{ satisfies the outgoing radiation condition at infinity,} \end{cases}$$

in a weak sense. We deduce from [2] that $E = 0$. Then, it follows that

$$g = n \wedge E|_{\Gamma^{\delta,h}} = 0.$$

Second, we verify the surjectivity of the operator L_0. Let g^{in} be in $TH^{1/2}(\text{div}, \Gamma^{\delta,h})$. Let $E^{in} \in (H^{1,loc}(\Omega_2^{\delta,h}))^3$ be defined as the unique solution of the following problem:

$$\begin{cases} \overrightarrow{\text{curl}}\, \overrightarrow{\text{curl}}\, E^{in} - \omega^2 \varepsilon_2 \mu_2\, E^{in} = 0 \quad \text{in } \Omega_2^{\delta,h}, \\ T_2(\omega)(n \wedge E^{in}) - \overrightarrow{\text{curl}}\, E^{in} \wedge n = g^{in} \quad \text{on } \Gamma^{\delta,h}, \\ E^{in} \text{ satisfies the downgoing radiation condition at infinity.} \end{cases}$$

The uniqueness of E^{in} follows from the fact that the boundary condition

$$T_2(\omega)(n \wedge E^{in}) - \overrightarrow{\text{curl}}\, E^{in} \wedge n = 0 \quad \text{on } \Gamma^{\delta,h}$$

implies (from the definition of the operator $T_2(\omega)$) that E^{in} satisfies the outgoing radiation condition. But it also satisfies the downgoing radiation condition, and then

it is a trivial solution. To show the existence of a solution, we make use of the boundary integral equation method. If we make an ansatz for E^{in} in the form

$$E^{in}(x) = -\omega\sqrt{\varepsilon_2\mu_2}\overrightarrow{\mathrm{curl}}\int_{\Gamma^{\delta,h}} a(y)\frac{e^{-i\omega\sqrt{\varepsilon_2\mu_2}|x-y|}}{4\pi|x-y|}ds(y) \quad \text{in } \Omega_2^{\delta,h},$$

where a is a vector in $TH^{1/2}(\mathrm{div},\Gamma^{\delta,h})$, we obtain that a satisfies the following integral equation on $\Gamma^{\delta,h}$:

$$-i\sqrt{\frac{\varepsilon_2}{\mu_2}}n\wedge\overrightarrow{\mathrm{curl}}\overrightarrow{\mathrm{curl}}\int_{\Gamma^{\delta,h}} a(y)\left(\frac{e^{-i\omega\sqrt{\varepsilon_2\mu_2}|x-y|}}{4\pi|x-y|}+\frac{e^{i\omega\sqrt{\varepsilon_2\mu_2}|x-y|}}{4\pi|x-y|}\right)ds(y) = \frac{1}{i\omega\mu_2}\,g^{in},$$

which is of Fredholm type. The existence of E^{in} follows then from the uniqueness. From [2], the Maxwell system

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,E \;=\; i\omega\mu^{\delta}\,H \quad \text{in } \Omega^e, \\ \overrightarrow{\mathrm{curl}}\,H \;=\; -i\omega\varepsilon^{\delta}\,E \quad \text{in } \Omega^e, \\ E-E^{in} \text{ satisfies the outgoing radiation condition at infinity} \end{cases}$$

with the boundary condition $E\wedge n = 0$ on Γ has a unique solution

$$(E,H)\in\Big(H^{1,loc}(O)\Big)^3\times\Big(H^{1,loc}(O)\Big)^3 \quad \text{for all } \overline{O}\subset\Omega_2^{\delta,h}.$$

By construction, the tangential field $n\wedge E\in TH^{1/2}(\Gamma^{\delta,h})$ satisfies

$$L_0(n\wedge E) \;=\; -g^{in}.$$

In order to complete the proof of this lemma, we should verify that $\mathrm{div}_{\Gamma^{\delta,h}}(n\wedge E)$ is in $H^{1/2}(\Gamma^{\delta,h})$. Using the identity

$$\mathrm{div}_{\Gamma^{\delta,h}}(n\wedge E) \;=\; -n\cdot\overrightarrow{\mathrm{curl}}\,E|_{\Gamma^{\delta,h}},$$

we obtain

$$\mathrm{div}_{\Gamma^{\delta,h}}(n\wedge E) \;=\; -i\omega\mu_2\,n\cdot H\in H^{1/2}(\Gamma^{\delta,h}),$$

and then the proof is complete. □

Our next result plays an important role in the proof of the existence and uniqueness results for the Drude–Born–Fedorov equations (2.8).

LEMMA 3.5.

$$\Im m\,(L_t\,g, n\wedge g) \;=\; 0 \Longrightarrow g \;=\; 0 \quad \text{for all } t\in\,[0,1\,]. \tag{3.14}$$

Proof. Multiplying (3.1) by $\overline{e}^{\delta}$ and integrating by parts, we obtain

$$\Im m\,(T_1(\beta,\omega)\,g^{\delta}, n\wedge g^{\delta}) \;=\; 0.$$

Now, multiplying (3.2) by $\overline{v}^{\delta}$ and integrating by parts over $\Omega_{2,R}^{\delta,h} = \Omega_2^{\delta,h}\cap\{r<R\}$ yield

$$\Im m\,(T_2(\omega)\,g^{\delta}, n\wedge g^{\delta}) \;=\; \Im m\,\Big(T_R(\omega)(n\wedge(n\wedge v^{\delta})), n\wedge(n\wedge v^{\delta})\Big),$$

where $T_R(\omega)$ is the pseudodifferential operator on Σ_R given by

$$-n \wedge (n \wedge v^\delta)|_{\Sigma_R} \longmapsto \overrightarrow{\mathrm{curl}}\, v^\delta \wedge n|_{\Sigma_R}.$$

Here v^δ is the solution to the Maxwell equations on $\Omega_2^{\delta,h}$ satisfying the radiation condition. From the property (cf. Appendix A)

$$\Im m\,(T_R(\omega)(n \wedge (n \wedge v^\delta)), n \wedge (n \wedge v^\delta)) = 0 \Longrightarrow n \wedge v^\delta = 0$$

and by the Cauchy–Kowaleska uniqueness theorem, (3.14) holds for all t in $[0,1]$. □

We write the operator L_1 as

$$L_1 = L_0 + \Big(T_1(\beta,\omega) - T_1(\beta=0,\omega)\Big).$$

We show that equation (3.12) is uniquely solvable in $TH^{1/2}(\mathrm{div},\Gamma^{\delta,h})$.

LEMMA 3.6. *Assume that the Maxwell equations*

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,\overrightarrow{\mathrm{curl}}\, u - \omega^2\,\varepsilon_2\,\mu_2\,u = 0 & \text{in } \Omega_1^{\delta,h}, \\ n \wedge u = 0 & \text{on } \Gamma, \\ n \wedge u = 0 & \text{on } \Gamma^{\delta,h} \end{cases}$$

have only the trivial solution in $(H^1(\Omega_1^{\delta,h}))^3$. *Then the operator* $T_1(\beta=0,\omega)-T_1(\beta,\omega)$ *is compact in* $TH^{1/2}(\mathrm{div},\Gamma^{\delta,h})$.

Proof. Let g be in $TH^{1/2}(\mathrm{div},\Gamma^{\delta,h})$. We define v_0 as the unique solution of Maxwell's equations

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,\overrightarrow{\mathrm{curl}}\, v_0 - \omega^2\,\varepsilon_2\,\mu_2\,v_0 = 0 & \text{in } \Omega_1^{\delta,h}, \\ n \wedge v_0 = 0 & \text{on } \Gamma, \\ n \wedge v_0 = g & \text{on } \Gamma^{\delta,h} \end{cases}$$

and v_1 as the unique solution of the following boundary-value problem:

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,\dfrac{1}{\mu^\delta}\,(1-\omega^2\varepsilon^\delta\mu^\delta(\beta^\delta)^2)\,\overrightarrow{\mathrm{curl}}\, v_1 - \omega^2\,\beta^\delta\,\varepsilon^\delta\,\overrightarrow{\mathrm{curl}}\, v_1 \\ -\omega^2\,\overrightarrow{\mathrm{curl}}\,\beta^\delta\,\varepsilon^\delta\,v_1 - \omega^2\,\varepsilon^\delta\,v_1 = 0 & \text{in } \Omega_1^{\delta,h}, \\ n \wedge v_1 = 0 & \text{on } \Gamma, \\ n \wedge v_1 = g & \text{on } \Gamma^{\delta,h}. \end{cases}$$

Therefore,

$$\Big(T_1(\beta=0,\omega) - T_1(\beta,\omega)\Big)\,g = \overrightarrow{\mathrm{curl}}\, v \wedge n,$$

where $v = v_0 - v_1$. v satisfies

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,\overrightarrow{\mathrm{curl}}\, v - \omega^2\,\varepsilon_2\,\mu_2\,v = w & \text{in } \Omega_1^{\delta,h}, \\ n \wedge v = 0 & \text{on } \Gamma, \\ n \wedge v = 0 & \text{on } \Gamma^{\delta,h}, \end{cases}$$

where

$$w = \begin{cases} \omega^2 \varepsilon_2 \mu_2 v_1 - 2(\gamma_1)^2 \beta \overrightarrow{\text{curl}}\, v_1 - (\gamma_1)^2 v_1 & \text{in } \Omega_1^\delta, \\ 0 \quad \text{in } \Omega_1^{\delta,h} \setminus \overline{\Omega_1^\delta}. \end{cases}$$

Since $w = 0$ in a neighborhood of $\Gamma^{\delta,h}$, it follows under the regularity assumption of $\Gamma^{\delta,h}$ that

$$T_1(\beta = 0, \omega) = T_1(\beta, \omega) \bmod OPS^{-\infty}(\Gamma^{\delta,h})$$

(cf. [26]), so the proof of the lemma is over. □

Since the operator L_0 is invertible, the operator L_1 from $TH^{1/2}(\text{div}, \Gamma^{\delta,h})$ into itself satisfies then the Fredholm alternative. From the uniqueness Lemma 3.5, it follows that the mapping L_1 maps $TH^{1/2}(\text{div}, \Gamma^{\delta,h})$ onto $TH^{1/2}(\text{div}, \Gamma^{\delta,h})$. Thus, we have the following theorem.

THEOREM 3.7. *The boundary-value problem* (2.10) *has a unique solution.*

Proof. If E is a solution of (2.10) then $n \wedge E \in TH^{1/2}(\text{div}, \Gamma^{\delta,h})$ is a solution of the integral equation $L_1 u = g^{in}$. Conversely, let g be the unique solution of the above integral equation. Let E be defined by

$$E = \begin{cases} E_1 \quad \text{in } \Omega_1^{\delta,h}, \\ E_2 + E^{in} \quad \text{in } \Omega_2^{\delta,h}, \end{cases}$$

where E_1 is the unique solution of

$$\begin{cases} \overrightarrow{\text{curl}}\, \dfrac{1}{\mu^\delta} (1 - \omega^2 \varepsilon \mu^\delta (\beta^\delta)^2) \overrightarrow{\text{curl}}\, E_1 - \omega^2 \beta^\delta \varepsilon^\delta \overrightarrow{\text{curl}}\, E_1 \\ -\omega^2 \overrightarrow{\text{curl}}\, \beta^\delta \varepsilon^\delta E_1 - \omega^2 \varepsilon^\delta E_1 = 0 \quad \text{in } \Omega_1^{\delta,h}, \\ n \wedge E_1 = 0 \quad \text{on } \Gamma, \\ n \wedge E_1 = g \quad \text{on } \Gamma^{\delta,h}, \end{cases}$$

and E_2 is the unique solution of Maxwell's equations

$$\begin{cases} \overrightarrow{\text{curl}}\, \overrightarrow{\text{curl}} v^\delta - \omega^2 \varepsilon_2 \mu_2 E_2 = 0 \quad \text{in } \Omega_2^{\delta,h}, \\ n \wedge E_2 = g \quad \text{on } \Gamma^{\delta,h}, \\ E_2 \text{ satisfies the outgoing radiation condition at infinity.} \end{cases}$$

It follows that E satisfies equations (2.10); the proof of the theorem is over. □

4. Generalized impedance boundary conditions. In this section, we assume that Ω_1^δ and Ω_2^δ are filled with materials in such a way that the magnetic permeability μ^δ satisfies

$$\mu^\delta = \begin{cases} \mu_1 \quad \text{in } \Omega_1^\delta, \\ \mu_2 \quad \text{in } \Omega_2^\delta, \end{cases} \tag{4.1}$$

the dielectric coefficient ε^δ satisfies

$$\varepsilon^\delta = \begin{cases} \varepsilon_1 \quad \text{in } \Omega_1^\delta, \\ \varepsilon_2 \quad \text{in } \Omega_2^\delta, \end{cases} \tag{4.2}$$

and the chirality admittance β^δ satisfies

$$\beta^\delta = \begin{cases} \beta_1 = \beta & \text{in } \Omega_1^\delta, \\ \beta_2 = 0 & \text{in } \Omega_2^\delta, \end{cases} \tag{4.3}$$

where $\mu_1, \mu_2, \varepsilon_1, \varepsilon_2$ are positive constants such that $\mu_1 \neq \mu_2, \varepsilon_1 \neq \varepsilon_2$ and β is a real constant.

Setting

$$E_j^\delta = E^\delta|_{\Omega_j^\delta}, \quad (j = 1, 2),$$

the equations (2.8) yield

$$\overrightarrow{\operatorname{curl}}\,\overrightarrow{\operatorname{curl}} E_1^\delta - 2\gamma_1^2\,\beta\,\overrightarrow{\operatorname{curl}}\,E_1^\delta - \gamma_1^2\,E_1^\delta = 0 \quad \text{in } \Omega_1^\delta \tag{4.4}$$

and

$$\overrightarrow{\operatorname{curl}}\,\overrightarrow{\operatorname{curl}} E_2^\delta - \omega^2\,\varepsilon_2\,\mu_2\,E_2^\delta = 0 \quad \text{in } \Omega_2^\delta. \tag{4.5}$$

The weak form of the Maxwell equations in a neighborhood of Γ^δ gives the following jump relations on Γ^δ:

$$\begin{cases} \left(E_1^\delta - E_2^\delta\right) \wedge n = 0, \\ \left(\overrightarrow{\operatorname{curl}}\,E_1^\delta - \dfrac{1}{\mu_c}\overrightarrow{\operatorname{curl}}\,E_2^\delta\right) \wedge n = \beta\,(\gamma_1)^2\,E_1^\delta \wedge n, \end{cases} \tag{4.6}$$

where n is the normal to Γ^δ and μ_c defined by $\mu_c = \mu_2\,(1 - \omega^2\varepsilon_1\mu_1\beta^2)/\mu_1$.

The purpose of this section is to derive approximative impedance conditions useful to performing numerical calculations of the scattered field by the chiral layer. Our program is as follows. First, we derive these approximative impedance conditions. Then, we establish the existence and uniqueness of a solution $\mathcal{E}^\delta$ to the Maxwell equations in Ω_2^δ satisfying the radiation condition with these new boundary conditions on Γ^δ. Finally, by using a simple continuity argument, error estimates between $E_2^\delta - E^{in}$ and $\mathcal{E}^\delta$ are proved.

Using Lemma 2.1, it follows that

$$\overrightarrow{\operatorname{curl}}\,E_1^\delta \wedge n = \frac{\partial}{\partial s}E_{\Gamma_s^\delta}^\delta - \overrightarrow{\operatorname{grad}}_{\Gamma_s^\delta}(E_1^\delta \cdot n) + \mathcal{R}(E_{\Gamma_s^\delta}^\delta). \tag{4.7}$$

The remarkable identity (2.1) yields

$$\overrightarrow{\operatorname{curl}}\,H_1^\delta \cdot n = \operatorname{curl}_{\Gamma_s^\delta} H_{\Gamma_s^\delta}^\delta = \operatorname{div}_{\Gamma_s^\delta}(H_1^\delta \wedge n). \tag{4.8}$$

Since $E_1^\delta \wedge n = 0$ on Γ^δ, by integrating the identity (4.7) we obtain

$$E_{\Gamma^\delta}^\delta(x_{\Gamma^\delta}) = \int_{-\delta}^0 \overrightarrow{\operatorname{curl}}\,E_1^\delta \wedge n\,ds + \int_{-\delta}^0 \overrightarrow{\operatorname{grad}}_{\Gamma_s^\delta}(E_1^\delta \cdot n)\,ds - \int_{-\delta}^0 \mathcal{R}(E_{\Gamma_s^\delta}^\delta)\,ds. \tag{4.9}$$

(2.8) gives

$$\overrightarrow{\operatorname{curl}}\,E_1^\delta \wedge n = (\gamma_1)^2\,\beta\,E_1^\delta \wedge n + i\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2 H_1^\delta \wedge n. \tag{4.10}$$

Combining (4.9) and (4.10) gives

$$\begin{aligned} E^\delta_{\Gamma^\delta}(x_{\Gamma^\delta}) &= \delta\,(\gamma_1)^2\,\beta\,E^\delta_1 \wedge n + i\delta\,\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2 H^\delta_1 \wedge n \\ &\quad +\delta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}(E^\delta_1(x_{\Gamma^\delta})\cdot n) + 0(\delta^2). \end{aligned} \tag{4.11}$$

Now, from

$$\operatorname{div}_{\Gamma^\delta}(H^\delta_1 \wedge n) = (\gamma_1)^2\beta\,H^\delta_1\cdot n - i\omega\varepsilon_1\left(\frac{\gamma_1}{k_1}\right)^2 E^\delta_1\cdot n \tag{4.12}$$

and

$$\operatorname{div}_{\Gamma^\delta}(E^\delta_1 \wedge n) = (\gamma_1)^2\beta\,E^\delta_1\cdot n + i\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2 H^\delta_1\cdot n, \tag{4.13}$$

it is easily seen that

$$E^\delta_1\cdot n = \frac{i}{\omega\varepsilon_1}\operatorname{div}_{\Gamma^\delta}(H^\delta_1\wedge n) - \beta\operatorname{div}_{\Gamma^\delta}(E^\delta_1\wedge n). \tag{4.14}$$

It follows that

$$\begin{aligned} -n\wedge(n\wedge E^\delta_2) &= i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2 H^\delta_2\wedge n + \frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(H^\delta_2\wedge n) \\ &\quad -\delta\beta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(E^\delta_2\wedge n) + 0(\delta^2). \end{aligned} \tag{4.15}$$

In the case of achiral thin layer of dielectric material, (4.15) is reduced to the impedance boundary condition obtained by Engquist and Nédélec [13]:

$$-n\wedge(n\wedge E^\delta_2) = i\delta\omega\mu_1\,H^\delta_2\wedge n + \frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(H^\delta_2\wedge n) + 0(\delta^2). \tag{4.16}$$

Our aim is to prove the existence and uniqueness of $\mathcal{E}^\delta$ and $\mathcal{H}^\delta$ solutions to the Maxwell equations

$$\begin{cases} \overrightarrow{\operatorname{curl}}\,\mathcal{E}^\delta = i\omega\,\mu_2\,\mathcal{H}^\delta & \text{in } \Omega^\delta_2, \\ \overrightarrow{\operatorname{curl}}\,\mathcal{H}^\delta = -i\omega\,\varepsilon_2\,\mathcal{E}^\delta & \text{in } \Omega^\delta_2, \end{cases} \tag{4.17}$$

with the new boundary conditions given as

$$\begin{aligned} &n\wedge(n\wedge\mathcal{E}^\delta) + i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\mathcal{H}^\delta\wedge n + \frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(\mathcal{H}^\delta\wedge n) \\ &-\delta\beta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(\mathcal{E}^\delta\wedge n) = g^\delta \quad \text{on } \Gamma^\delta, \end{aligned} \tag{4.18}$$

where g^δ is a known smooth vector field and $\mathcal{E}^\delta$ and $\mathcal{H}^\delta$ satisfy the radiation condition at infinity. First, we prove the following uniqueness result.

LEMMA 4.1. *Assume $\beta = 0$. Then, the Maxwell system* (4.17) *with the boundary conditions* (4.18) *has at most one solution.*

Proof. Let $\mathcal{E}^\delta$ and $\mathcal{H}^\delta$ satisfy (4.17) with the following boundary conditions on Γ^δ:

$$n \wedge (n \wedge \mathcal{E}^\delta) + i\delta\omega\mu_1 \mathcal{H}^\delta \wedge n + \frac{i\delta}{\omega\varepsilon_1} \overrightarrow{\mathrm{grad}}_{\Gamma^\delta} \mathrm{div}_{\Gamma^\delta} (\mathcal{H}^\delta \wedge n) = 0 \quad \text{on } \Gamma^\delta, \tag{4.19}$$

and the radiation condition at infinity. Direct computation yields

$$\int_{\Omega^\delta_{2,R}} |\overrightarrow{\mathrm{curl}}\, \mathcal{E}^\delta|^2 - \omega^2 \varepsilon_2\mu_2 \int_{\Omega^\delta_{2,R}} |\mathcal{E}^\delta|^2 + \delta\omega^2\mu_1\,\mu_2 \int_{\Gamma^\delta} |\mathcal{H}^\delta \wedge n|^2$$
$$-\frac{\delta\,\mu_2}{\varepsilon_1} \int_{\Gamma^\delta} |\mathrm{div}_{\Gamma^\delta} (\mathcal{H}^\delta \wedge n)|^2 + i\omega\,\mu_2 \int_{\Sigma_R} (\mathcal{H}^\delta \wedge n) \cdot \overline{\mathcal{E}}^\delta = 0.$$

By taking the imaginary part in the above expression, we obtain

$$\Re e\Big(\int_{\Sigma_R} (\mathcal{H}^\delta \wedge n) \cdot \overline{\mathcal{E}}^\delta \Big) = 0. \tag{4.20}$$

Classical arguments, based on the Cauchy–Kowaleska uniqueness theorem and the well-known properties of the pseudodifferential operator $T_R(\omega)$ (cf. Appendix A), ensure that $\mathcal{E}^\delta = 0$ and $\mathcal{H}^\delta = 0$ in $\Omega^\delta_{2,R}$, and the proof of the uniqueness of a solution to (4.17) with the boundary conditions (4.19) is complete. □

Lemma 4.1 provides a proof of the uniqueness of the solution to the Maxwell system (4.17) with the approximative impedance condition written by Engquist and Nédélec.

LEMMA 4.2. *The Maxwell system* (4.17) *with the boundary conditions*

$$\begin{aligned} &n \wedge (n \wedge \mathcal{E}^\delta) + i\delta\omega\mu_1 \left(\frac{\gamma_1}{k_1}\right)^2 \mathcal{H}^\delta \wedge n + \frac{i\delta}{\omega\varepsilon_1} \overrightarrow{\mathrm{grad}}_{\Gamma^\delta} \mathrm{div}_{\Gamma^\delta} (\mathcal{H}^\delta \wedge n) \\ &-\delta\beta \overrightarrow{\mathrm{grad}}_{\Gamma^\delta} \mathrm{div}_{\Gamma^\delta} (\mathcal{E}^\delta \wedge n) = 0 \quad \text{on } \Gamma^\delta \end{aligned} \tag{4.21}$$

has nontrivial solutions only if μ_1 belongs to a countable set of exceptional values.

Proof. We introduce the scalar three-dimensional fundamental solution corresponding to ω as

$$\Phi(x,y) = \frac{e^{i\omega\sqrt{\varepsilon_2\mu_2}|x-y|}}{4\pi|x-y|}, \quad x \neq y.$$

Let

$$u^\delta = \mathcal{H}^\delta \wedge n.$$

The electric field $\mathcal{E}^\delta$ has the following integral representation:

$$\begin{aligned} \mathcal{E}^\delta(x) &= -\frac{i}{\omega\varepsilon_2} \overrightarrow{\mathrm{grad}} \int_{\Gamma^\delta} \Phi(x-y)\, \mathrm{div}_{\Gamma^\delta} u^\delta(y)\, d\gamma(y) \\ &\quad -i\omega\mu_2 \int_{\Gamma^\delta} \Phi(x-y)\, u^\delta(y)\, d\gamma(y) \\ &\quad + \overrightarrow{\mathrm{curl}} \int_{\Gamma^\delta} n(y) \wedge \mathcal{E}^\delta(y)\, \Phi(x-y)\, d\gamma(y). \end{aligned} \tag{4.22}$$

From

$$\begin{aligned}\overrightarrow{\mathrm{curl}}\,\mathcal{E}^\delta\cdot n &= -\mathrm{curl}_{\Gamma^\delta}\Big(n\wedge(n\wedge\mathcal{E}^\delta)\Big)\\ &= i\delta\,\omega\,\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\mathrm{curl}_{\Gamma^\delta}\,u^\delta,\end{aligned}$$

we deduce that

$$\begin{aligned}n\wedge\mathcal{E}^\delta &= i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2 n\wedge u^\delta+\frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\mathrm{div}_{\Gamma^\delta}\,u^\delta\\ &\quad +i\delta^2\,\beta\,\omega\,\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\mathrm{curl}_{\Gamma^\delta}\,u^\delta,\end{aligned}\tag{4.23}$$

and then

$$\mathrm{div}_{\Gamma^\delta}\,(n\wedge\mathcal{E}^\delta)=i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\mathrm{div}_{\Gamma^\delta}\,(n\wedge u^\delta).$$

We rewrite the approximative impedance conditions (4.21) as

$$\begin{aligned}&n\wedge(n\wedge\mathcal{E}^\delta)+i\delta\,\omega\,\mu_1\left(\frac{\gamma_1}{k_1}\right)^2 u^\delta+\frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\mathrm{div}_{\Gamma^\delta}\,(u^\delta)\\ &-i\delta^2\omega\mu_1\beta\left(\frac{\gamma_1}{k_1}\right)^2\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,(\mathrm{div}_{\Gamma^\delta}\,(u^\delta\wedge n))=0.\end{aligned}\tag{4.24}$$

Now, multiplying (4.24) by v^δ and integrating by parts on Γ^δ yields

$$\begin{aligned}&\int_{\Gamma^\delta}\mathcal{E}^\delta(x)\cdot v^\delta(x)\,d\gamma(x)=i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\int_{\Gamma^\delta}u^\delta(x)\cdot v^\delta(x)\,d\gamma(x)\\ &-\frac{i\delta}{\omega\varepsilon_1}\int_{\Gamma^\delta}\mathrm{div}_{\Gamma^\delta}u^\delta(x)\,\mathrm{div}_{\Gamma^\delta}\,v^\delta(x)\,d\gamma(x)\\ &+i\delta^2\,\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}\mathrm{div}_{\Gamma^\delta}\,v^\delta(x)\,\mathrm{div}_{\Gamma^\delta}\,(u^\delta(x)\wedge n(x))\,d\gamma(x).\end{aligned}$$

From the integral representation (4.22) it is easily seen that

$$\begin{aligned}&\frac12\int_{\Gamma^\delta}\mathcal{E}^\delta(x)\cdot v^\delta(x)\,d\gamma(x)=\frac{i}{\omega\varepsilon_2}\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,\mathrm{div}_{\Gamma^\delta}\,u^\delta(y)\,\mathrm{div}_{\Gamma^\delta}\,v^\delta(x)\,d\gamma(y)\,d\gamma(x)\\ &-i\omega\mu_2\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,u^\delta(y)\cdot v^\delta(x)\,d\gamma(y)\,d\gamma(x)\\ &+\int_{\Gamma^\delta}\Big(\overrightarrow{\mathrm{curl}}\int_{\Gamma^\delta}n(y)\wedge\mathcal{E}^\delta(y)\,\Phi(x-y)\,d\gamma(y)\Big)\cdot v^\delta(x)\,d\gamma(x).\end{aligned}\tag{4.25}$$

Then, we obtain that

$$
\begin{aligned}
&\frac{2i}{\omega\varepsilon_2}\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,\mathrm{div}_{\Gamma^\delta}\,u^\delta(y)\,\mathrm{div}_{\Gamma^\delta}\,v^\delta(x)\,d\gamma(y)\,d\gamma(x)\\
&-2i\omega\mu_2\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,u^\delta(y)\,\cdot\,v^\delta(x)\,d\gamma(y)\,d\gamma(x)\\
&+2\int_{\Gamma^\delta}\Big(\overrightarrow{\mathrm{curl}}\int_{\Gamma^\delta}n(y)\,\wedge\,\mathcal{E}^\delta(y)\,\Phi(x-y)\,d\gamma(y)\Big)\,\cdot\,v^\delta(x)\,d\gamma(x)\\
&=i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\int_{\Gamma^\delta}u^\delta(x)\,\cdot\,v^\delta(x)\,d\gamma(x)-\frac{i\delta}{\omega\varepsilon_1}\int_{\Gamma^\delta}\mathrm{div}_{\Gamma^\delta}u^\delta(x)\mathrm{div}_{\Gamma^\delta}\,v^\delta(x)\,d\gamma(x)\\
&+i\delta^2\,\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}\mathrm{div}_{\Gamma^\delta}\,v^\delta(x)\,\mathrm{div}_{\Gamma^\delta}\,(u^\delta(x)\,\wedge\,n(x))\,d\gamma(x).
\end{aligned}
\tag{4.26}
$$

We seek u^δ in $TH^1(\Gamma^\delta)\,=\,\{c\in(H^1(\Gamma^\delta))^3, c\,.\,n\,=\,0\,\}$ such that (4.26) holds for all tangential fields $v^\delta\,\in\,TH^1(\Gamma^\delta)$. We use the following decomposition of the space $TH^1(\Gamma^\delta)$.

LEMMA 4.3. $TH^1(\Gamma^\delta)\,=\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,H^2(\Gamma^\delta)\,\oplus\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,H^2(\Gamma^\delta)$.

The proof of this lemma is exactly the same as in [19] for the space

$$
TH^{-1/2}(\mathrm{div},\Gamma^\delta)\,=\Big\{c\in(H^{-1/2}(\Gamma^\delta))^3, c\,\cdot\,n=0, \mathrm{div}_{\Gamma^\delta}\,c\in H^{-1/2}(\Gamma^\delta)\Big\}.
$$

We decompose the tangential fields u^δ and v^δ as follows:

$$
\left\{\begin{array}{l}
u^\delta\;=\;\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi^\delta+\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi^\delta,\\
v^\delta\;=\;\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi^\delta_t+\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi^\delta_t.
\end{array}\right.
$$

(4.26) leads to the following equation:

$$
\begin{aligned}
&\frac{i\delta}{\omega\varepsilon_1}\int_{\Gamma^\delta}\Delta_{\Gamma^\delta}\,\psi^\delta(x)\,\Delta_{\Gamma^\delta}\,\psi^\delta_t(x)\,d\gamma(x)\\
&+\frac{2i}{\omega\varepsilon_2}\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,\Delta_{\Gamma^\delta}\,\psi^\delta(y)\,\Delta_{\Gamma^\delta}\,\psi^\delta_t(x)\,d\gamma(y)\,d\gamma(x)\\
&-2i\omega\mu_2\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi^\delta(y)\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi^\delta_t(x)\,d\gamma(y)\,d\gamma(x)\\
&-2i\omega\mu_2\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi^\delta(y)\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi^\delta_t(x)\,d\gamma(y)\,d\gamma(x)\\
&+2\int_{\Gamma^\delta}(\mathcal{R}(x)-2c(x))\Big\{\int_{\Gamma^\delta}\Big(i\delta\omega\mu_1\,(-\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\psi^\delta(y)+\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi^\delta(y))
\end{aligned}
$$

$$+\frac{i\delta}{\omega\varepsilon_1}\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\psi^\delta(y)\Big)\,\Phi(x-y)\,d\gamma(y)\,\wedge\,n(x)\Big\}\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(x)$$

$$-i\delta^2\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}(\mathcal{R}(x)-2c(x))\Big(\int_{\Gamma^\delta}\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\varphi^\delta(y)$$

$$\Phi(x-y)\,d\gamma(y)\,\wedge\,n(x)\Big)\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(x)$$

$$+i\delta\omega\mu_1\int_{\Gamma^\delta}\int_{\Gamma^\delta}\frac{\partial\Phi}{\partial n_x}(x-y)\,(-\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\psi^\delta(y)+\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi^\delta(y))\,\cdot\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(y)\,d\gamma(x)$$

$$+\frac{i\delta}{\omega\varepsilon_1}\int_{\Gamma^\delta}\int_{\Gamma^\delta}\frac{\partial\Phi}{\partial n_x}(x-y)\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\psi^\delta(y)\,\cdot\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(y)\,d\gamma(x)$$

$$-i\delta^2\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}\int_{\Gamma^\delta}\frac{\partial\Phi}{\partial n_x}(x-y)\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\varphi^\delta(y)\,\cdot\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(y)\,d\gamma(x)$$

$$+i\delta^2\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}\Delta_{\Gamma^\delta}\,\varphi^\delta(x)\,\Delta_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(x)$$

$$-i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\int_{\Gamma^\delta}\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi^\delta(x)\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(x)\,=\,0.$$

By multiplying (4.23) by $\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi_t^\delta$, integrating by parts, and using (4.25), we obtain

$$i\delta^2\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}\Delta_{\Gamma^\delta}\,\varphi^\delta(x)\,\cdot\,\Delta_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(x)$$

$$-i\delta\omega\mu_1\int_{\Gamma^\delta}\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi^\delta(x)\,\cdot\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(x)$$

$$=\frac{2i}{\omega\varepsilon_2}\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,\Delta_{\Gamma^\delta}\,\psi^\delta(y)\,\Delta_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(x)\,d\gamma(y)$$

$$-2i\omega\mu_2\int_{\Gamma^\delta}\int_{\Gamma^\delta}\Phi(x-y)\,(\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi^\delta(y)+\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi^\delta(y))\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(x)\,d\gamma(y)$$

$$+2\int_{\Gamma^\delta}(\mathcal{R}(x)-2c(x))\Big\{\int_{\Gamma^\delta}\Big(i\delta\omega\mu_1\,(-\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\psi^\delta(y)+\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi^\delta(y))$$

$$+\frac{i\delta}{\omega\varepsilon_1}\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\psi^\delta(y)\Big)\,\Phi(x-y)\,d\gamma(y)\,\wedge\,n(x)\Big\}\,\cdot\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(x)$$

$$-i\delta^2\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}(\mathcal{R}(x)-2c(x))\Big(\int_{\Gamma^\delta}\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\varphi^\delta(y)$$

$$\Phi(x-y)\,d\gamma(y)\,\wedge\,n(x)\Big)\,\cdot\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(x)$$

$$-i\delta\omega\mu_1\int_{\Gamma^\delta}\int_{\Gamma^\delta}\frac{\partial\Phi}{\partial n_x}(x-y)\,(-\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\psi^\delta(y)+\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi^\delta(y))\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(y)\,d\gamma(x)$$

$$-\frac{i\delta}{\omega\varepsilon_1}\int_{\Gamma^\delta}\int_{\Gamma^\delta}\frac{\partial\Phi}{\partial n_x}(x-y)\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\psi^\delta(y)\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(y)\,d\gamma(x)$$

$$+i\delta^2\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\beta\int_{\Gamma^\delta}\int_{\Gamma^\delta}\frac{\partial\Phi}{\partial n_x}(x-y)\,\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\Delta_{\Gamma^\delta}\,\varphi^\delta(y)\,\cdot\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(y)\,d\gamma(x).$$

Straightforward computations give that the system of the above two equations leads

to a compact perturbation of the coercive form

$$\begin{pmatrix} \dfrac{1}{\omega\varepsilon_1}\displaystyle\int_{\Gamma^\delta} \Delta_{\Gamma^\delta}\,\psi^\delta(x)\,\Delta_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(x) & \delta\omega\mu_1\left(\dfrac{\gamma_1}{k_1}\right)^2\beta\displaystyle\int_{\Gamma^\delta}\Delta_{\Gamma^\delta}\,\varphi^\delta(x)\,\Delta_{\Gamma^\delta}\,\psi_t^\delta(x)\,d\gamma(x) \\ 0 & \omega\mu_1\delta\left(\dfrac{\gamma_1}{k_1}\right)^2\beta\displaystyle\int_{\Gamma^\delta}\Delta_{\Gamma^\delta}\,\varphi^\delta(x)\,\Delta_{\Gamma^\delta}\,\varphi_t^\delta(x)\,d\gamma(x) \end{pmatrix}$$

on $H^2(\Gamma^\delta)/\mathbb{C}\times H^2(\Gamma^\delta)/\mathbb{C}$.

Now, we write the approximative impedance conditions (4.24) in the following form:

$$\begin{aligned} &n\wedge(n\wedge\mathcal{E}^\delta)+i\delta\,\omega\,u^\delta+\frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(u^\delta)\\ &-\delta\beta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\left(\overrightarrow{\operatorname{curl}}\,\mathcal{E}^\delta\cdot n\right)=-i\delta\,\omega\left(\mu_1\left(\frac{\gamma_1}{k_1}\right)^2-1\right)u^\delta. \end{aligned} \tag{4.27}$$

The above computations show that u^δ appears in (4.27) as an eigenvector of a compact perturbation of an invertible operator and

$$-i\delta\,\omega\left(\mu_1\left(\frac{\gamma_1}{k_1}\right)^2-1\right)$$

is its eigenvalue. Since the operator

$$\begin{aligned} &u^\delta\longmapsto n\wedge(n\wedge\mathcal{E}^\delta)+i\delta\,\omega\,u^\delta+\frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(u^\delta)\\ &-\delta\beta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\left(\overrightarrow{\operatorname{curl}}\,\mathcal{E}^\delta\cdot n\right) \end{aligned}$$

is independent of μ_1, the proof of the uniqueness Lemma 4.2 is then over. □

The proof of Lemma 4.2 also shows the existence of solutions to our boundary-value problem. Now, we use the boundary integral equation method to obtain optimal error estimates. We make an ansatz for $\mathcal{E}^\delta$ and $\mathcal{H}^\delta$ in the form

$$\begin{cases} \mathcal{E}^\delta(x)=\omega\sqrt{\varepsilon_2\mu_2}\,\overrightarrow{\operatorname{curl}}\displaystyle\int_{\Gamma^\delta}a(y)\,\Phi(x,y)\,ds(y) & \text{in } \Omega_2^\delta,\\ \mathcal{H}^\delta(x)=\dfrac{1}{i\omega\mu_2}\,\overrightarrow{\operatorname{curl}}\,\mathcal{E}^\delta(x) & \text{in } \Omega_2^\delta, \end{cases} \tag{4.28}$$

where a is a vector function in $TH^2(\Gamma^\delta)=\{c\in(H^2(\Gamma^\delta))^3, c\cdot n=0\}$. From the properties of Φ, we see that $(\mathcal{E}^\delta,\mathcal{H}^\delta)$ satisfy the Maxwell equations (4.17) and the radiation condition. The tangential components of $\mathcal{E}^\delta$ and $\mathcal{H}^\delta$ on Γ^δ take the form

$$\begin{aligned} n\wedge\mathcal{E}^\delta(x) &= \omega\sqrt{\varepsilon_2\mu_2}\,n\wedge\int_{\Gamma^\delta}\overrightarrow{\operatorname{curl}}_x(a(y)\,\Phi(x,y))\,ds(y)\\ &\quad+\tfrac{\omega}{2}\sqrt{\varepsilon_2\mu_2}\,a(x),\quad x\in\Gamma^\delta,\\ n\wedge\mathcal{H}^\delta(x) &= -i\sqrt{\frac{\varepsilon_2}{\mu_2}}\,n\wedge\overrightarrow{\operatorname{curl}}\,\overrightarrow{\operatorname{curl}}\int_{\Gamma^\delta}a(y)\,\Phi(x,y)\,ds(y),\quad x\in\Gamma^\delta. \end{aligned} \tag{4.29}$$

The boundary conditions (4.29) for $\mathcal{E}^\delta$ and $\mathcal{H}^\delta$ will lead to an integral equation on Γ^δ for the unknown vector a. In order to write it, we introduce the following boundary

operators:

$$
\begin{aligned}
Q(a)(x) &= n \wedge a(x), \quad x \in \Gamma^\delta,\\
M(a)(x) &= n \wedge \overrightarrow{\operatorname{curl}} \int_{\Gamma^\delta} a(y)\,\Phi(x,y)\,ds(y), \quad x \in \Gamma^\delta,\\
N(a)(x) &= n \wedge \int_{\Gamma^\delta} a(y)\,\Phi(x,y)\,ds(y), \quad x \in \Gamma^\delta,\\
P(a)(x) &= n \wedge \overrightarrow{\operatorname{curl}}\,\overrightarrow{\operatorname{curl}} \int_{\Gamma^\delta} a(y)\,\Phi(x,y)\,ds(y), \quad x \in \Gamma^\delta.
\end{aligned}
$$

Then, (4.18) leads to

$$
\begin{aligned}
&\frac{1}{2}\,\delta\,\beta\,\sqrt{\varepsilon_2\mu_2}\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} a - \frac{\delta}{\omega}\sqrt{\frac{\varepsilon_2}{\mu_2}}\,\frac{1}{\omega\varepsilon_1}\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} Pa\\
&+\delta\,\beta\,\sqrt{\varepsilon_2\mu_2}\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} Ma + \delta\,\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\sqrt{\frac{\varepsilon_2}{\mu_2}}\,Pa\\
&+\frac{\sqrt{\varepsilon_2\mu_2}}{2}\,Qa + \sqrt{\varepsilon_2\mu_2}\,QMa = g^\delta \in TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta),
\end{aligned}
\tag{4.30}
$$

where $TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta) = \{c \in (L^2(\Gamma^\delta))^3, c\cdot n = 0, \operatorname{curl}_{\Gamma^\delta} c \in H^1(\Gamma^\delta)\}$. Thus we have the following theorem.

THEOREM 4.4. *The vector function $a \in TH^2(\Gamma^\delta)$ is a solution of the integral equation* (4.30) *if and only if the fields $(\mathcal{E}^\delta,\mathcal{H}^\delta)$ satisfy the boundary conditions* (4.18).

In order to discuss the solvability of the integral equation (4.30), we compute $\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} P\,a, \overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} M\,a$, $\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} N\,a$, $\operatorname{curl}_{\Gamma^\delta} P\,a$ and $\operatorname{curl}_{\Gamma^\delta} QM\,a$. Direct computation yields

$$
\begin{aligned}
\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} P\,a &= \overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(a\wedge n) - \omega^2\sqrt{\varepsilon_2\mu_2}\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} N\,a,\\
&= \overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{curl}_{\Gamma^\delta} a - \omega^2\sqrt{\varepsilon_2\mu_2}\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} N\,a,
\end{aligned}
\tag{4.31}
$$

$$
\begin{aligned}
\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} M\,a &= -\frac{\partial}{\partial s}\Big(n\wedge\Big(n\wedge\overrightarrow{\operatorname{curl}}\overrightarrow{\operatorname{curl}}\int_{\Gamma^\delta} a\,\Phi\Big)\Big) + \omega^2\,\varepsilon_2\,\mu_2\,n\wedge\overrightarrow{\operatorname{curl}}\int_{\Gamma^\delta} a\,\Phi\\
&\quad -n\wedge\overrightarrow{\operatorname{curl}}\,a,\\
&= -i\sqrt{\frac{\mu_2}{\varepsilon_2}}\,\frac{\partial}{\partial s}\Big(n\wedge(n\wedge\mathcal{H}^\delta)\Big) + \omega\sqrt{\varepsilon_2\,\mu_2}\,n\wedge\mathcal{E}^\delta - n\wedge\overrightarrow{\operatorname{curl}}\,a,
\end{aligned}
\tag{4.32}
$$

$$
\begin{aligned}
\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta} N\,a &= \overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\Big(\overrightarrow{\operatorname{curl}}\int_{\Gamma^\delta} a\,\Phi\cdot n\Big),\\
&= \frac{1}{\omega\sqrt{\varepsilon_2\mu_2}}\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}(\mathcal{E}^\delta\cdot n),
\end{aligned}
\tag{4.33}
$$

$$
\begin{aligned}
\operatorname{curl}_{\Gamma^\delta} P\,a &= \operatorname{curl}_{\Gamma^\delta}\Big(n\wedge\Big(\mathcal{R} - 2c - \frac{\partial}{\partial n}\Big)\\
&\quad\Big\{-\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\Big(\int_{\Gamma^\delta} a(y)\cdot(n(x)-n(y))\Phi(x-y)\,ds(y)\Big)\\
&\quad -\mathcal{R}\Big(n\wedge\Big(n\wedge\int_{\Gamma^\delta} a\Phi\Big)\Big) + \frac{\partial}{\partial n}\Big(n\wedge\Big(n\wedge\int_{\Gamma^\delta} a\Phi\Big)\Big)\Big\}\Big)
\end{aligned}
\tag{4.34}
$$

$$\text{(4.35)}\quad \mathrm{curl}_{\Gamma^\delta} QM\,a \;=\; \mathrm{curl}_{\Gamma^\delta}\Big(n\wedge\int_{\Gamma^\delta}(\mathcal{R}-2c)(a\wedge n)\Phi-\int_{\Gamma^\delta} a\wedge n\frac{\partial\Phi}{\partial n_x}\Big).$$

Now, we write the integral equation (4.30) as follows:

$$\Big(L^\delta+K^\delta\Big)a\;=\;g^\delta, \tag{4.36}$$

where

$$L^\delta\,a\;=\;\frac{1}{2}\delta\,\beta\,\sqrt{\varepsilon_2\mu_2}\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\mathrm{div}_{\Gamma^\delta}\,a-\frac{\delta}{\omega}\sqrt{\frac{\varepsilon_2}{\mu_2}}\,\frac{1}{\omega\varepsilon_1}\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\mathrm{div}_{\Gamma^\delta}\,(a\wedge n) \tag{4.37}$$

and

$$\begin{aligned} K^\delta\,a \;&=\; \delta\,\frac{\varepsilon_2}{\varepsilon_1}\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\mathrm{div}_{\Gamma^\delta}\,N\,a+\delta\,\beta\,\sqrt{\varepsilon_2\mu_2}\,\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\mathrm{div}_{\Gamma^\delta}\,M\,a\\ &\quad+\sqrt{\varepsilon_2\mu_2}\,QM\,a+\delta\,\mu_1\,\left(\frac{\gamma_1}{k_1}\right)^2\sqrt{\frac{\varepsilon_2}{\mu_2}}\,P\,a+\frac{\sqrt{\varepsilon_2\mu_2}}{2}\,Q\,a. \end{aligned} \tag{4.38}$$

We treat this equation in the spaces

$$L^\delta+K^\delta\;:TH^2(\Gamma^\delta)\longmapsto TL^2(\mathrm{curl}_{\Gamma^\delta},\Gamma^\delta).$$

From (4.31)–(4.35), it is clear that $\mathrm{curl}_{\Gamma^\delta}\,K^\delta\,a\in H^2(\Gamma^\delta)$ and

$$K^\delta\,a\in TH^1(\Gamma^\delta)=\Big\{c\in(H^1(\Gamma^\delta))^3, c\cdot n=0\Big\}$$

for $a\in TH^2(\Gamma^\delta)$. The operator K^δ is then compact between these spaces. We show that the operator L^δ is an isomorphism from $TH^2(\Gamma^\delta)$ onto $TL^2(\mathrm{curl}_{\Gamma^\delta},\Gamma^\delta)$. By using the decomposition

$$a\;=\;\overrightarrow{\mathrm{curl}}_{\Gamma^\delta}\,\varphi+\overrightarrow{\mathrm{grad}}_{\Gamma^\delta}\,\psi,$$

the integral equation

$$L^\delta\,a\;=\;g^\delta$$

is equivalent to the following elliptic system:

$$\begin{cases} \dfrac{\sqrt{\varepsilon_2\mu_2}}{2}\,\Delta_{\Gamma^\delta}\,\psi\;=\;-\mathrm{curl}_{\Gamma^\delta}\,g^\delta\;\in H^1(\Gamma^\delta),\\ \sqrt{\dfrac{\varepsilon_2}{\mu_2}}\,\dfrac{\delta}{\omega^2\varepsilon_1}\,\Delta^2_{\Gamma^\delta}\,\varphi\;=\;-\dfrac{1}{2}\delta\,\beta\,\sqrt{\varepsilon_2\mu_2}\,\Delta^2_{\Gamma^\delta}\,\psi+\mathrm{div}_{\Gamma^\delta}\,g^\delta\;\in H^{-1}(\Gamma^\delta), \end{cases}$$

which has a unique solution $(\varphi,\psi)\in(H^3(\Gamma^\delta)/\mathbb{C})^2$. The integral equation $L^\delta\,a=g^\delta$ then has a unique solution in $TH^2(\Gamma^\delta)$.

Applying these results to (4.30) yields

$$a+(L^\delta)^{-1}\,K^\delta a\;=\;(L^\delta)^{-1}\,g^\delta, \tag{4.39}$$

which is a Fredholm equation of the second kind in $TH^2(\Gamma^\delta)$. Now, under the assumption that μ is not is the set of exceptional values introduced in Lemma 4.2, we can prove the following theorem.

THEOREM 4.5. *There exists a unique solution to the integral equation* (4.30) *in* $TH^2(\Gamma^\delta)$.

From Lemma 4.2 the boundary-value problem (4.17)–(4.18) itself has at most one solution. To show the uniqueness of the solution to the integral equation (4.30), let us consider $a \in TH^2(\Gamma^\delta)$ a solution of (4.30) and define

$$\mathcal{E}^\delta(x) = \omega \sqrt{\varepsilon_2 \mu_2}\, \overrightarrow{\operatorname{curl}} \int_{\Gamma^\delta} a(y)\, \Phi(x,y)\, ds(y) \quad \text{in } \Omega_2^\delta,$$
$$\mathcal{H}^\delta(x) = \frac{1}{i\omega\mu_2}\, \overrightarrow{\operatorname{curl}}\, \mathcal{E}^\delta(x) \quad \text{in } \Omega_2^\delta.$$

Standard arguments state that $\mathcal{E}^\delta$ and $\mathcal{H}^\delta$ are solutions of the boundary-value problem (4.17)–(4.18). Hence, there exists a unique solution to the integral equation (4.30) in $TH^2(\Gamma^\delta)$.

The uniform continuity of the operator $(\,(L^\delta)^{-1}\, K^\delta + I\,)^{-1}\,(L^\delta)^{-1}$ from $TL^2(\operatorname{curl}_{\Gamma^\delta}, \Gamma^\delta)$ onto $TH^2(\Gamma^\delta)$ with respect to δ gives the existence of a strictly positive constant C independent of δ such that the unique solution of the equation (4.39) satisfies the following estimate:

$$||a||_{TH^2(\Gamma^\delta)} \leq C\, ||g^\delta||_{TL^2(\operatorname{curl}_{\Gamma^\delta}, \Gamma^\delta)}.$$

From the integral representations (4.28), it is easily established that for each $O \subset\subset \Omega_2^\delta$, there exists a strictly positive constant C independent of δ such that

$$||\mathcal{E}^\delta||_{H(\overrightarrow{\operatorname{curl}}, O)} + ||\mathcal{H}^\delta||_{H(\overrightarrow{\operatorname{curl}}, O)} \leq C\, ||a||_{TH^2(\Gamma^\delta)}.$$

Therefore, we obtain the following result.

THEOREM 4.6. *For each* $O \subset\subset \Omega_2^\delta$, *there exists a positive constant* C *independent of* δ *such that*

$$||\mathcal{E}^\delta||_{H(\overrightarrow{\operatorname{curl}}, O)} + ||\mathcal{H}^\delta||_{H(\overrightarrow{\operatorname{curl}}, O)} \leq C\, ||g^\delta||_{TL^2(\operatorname{curl}_{\Gamma^\delta}, \Gamma^\delta)}. \tag{4.40}$$

Now, let $\mathcal{E}^\delta, \mathcal{H}^\delta$ be the solution of the Maxwell equations with the new boundary conditions corresponding to

$$g^\delta = n \wedge (n \wedge E^{in}) + i\delta\omega\mu_1 \left(\frac{\gamma_1}{k_1}\right)^2 H^{in} \wedge n + \frac{i\delta}{\omega\varepsilon_1}\, \overrightarrow{\operatorname{grad}}_{\Gamma^\delta} \operatorname{div}_{\Gamma^\delta} (H^{in} \wedge n)$$
$$-\delta\beta\, \overrightarrow{\operatorname{grad}}_{\Gamma^\delta} \operatorname{div}_{\Gamma^\delta} (E^{in} \wedge n)$$

in (4.18) ($H^{in} = -i\omega\, \varepsilon_2\, \overrightarrow{\operatorname{curl}}\, E^{in}$). By Theorem 4.4, we obtain the following error estimate.

THEOREM 4.7. *For each* $O \subset\subset \Omega_2^\delta$, *there exists a positive constant* C *independent of* δ *such that*

$$||\mathcal{E}^\delta - E_2^{sc,\delta}||_{H(\overrightarrow{\operatorname{curl}}, O)} + ||\mathcal{H}^\delta - H_2^{sc,\delta}||_{H(\overrightarrow{\operatorname{curl}}, O)} \leq C\, \delta^2, \tag{4.41}$$

where $E_2^{sc,\delta} = E_2^\delta - E^{in}$ *and* $H_2^{sc,\delta} = H_2^\delta - H^{in}$ *are the unique solutions of* (2.8) *in* Ω_2^δ *satisfying the radiation condition at infinity.*

Proof. We first recall that the tangential component of the electric field on Γ^δ is given by

$$\begin{aligned} E^\delta_{\Gamma^\delta}(x_{\Gamma^\delta}) &= (\gamma_1)^2\,\beta\int_{-\delta}^0 E_1^\delta\wedge n\,ds + i\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\int_{-\delta}^0 H_1^\delta\wedge n\,ds \\ &\quad + \frac{i}{\omega\varepsilon_1}\int_{-\delta}^0 \overrightarrow{\operatorname{grad}}_{\Gamma^\delta_s}\,\operatorname{div}_{\Gamma^\delta_s}(H_1^\delta\wedge n)\,ds - i\beta\int_{-\delta}^0 \overrightarrow{\operatorname{grad}}_{\Gamma^\delta_s}\,\operatorname{div}_{\Gamma^\delta_s}(E_1^\delta\wedge n)\,ds \\ &\quad - \int_{-\delta}^0 \mathcal{R}(E^\delta_{\Gamma^\delta_s})\,ds. \end{aligned}$$

Let us define

$$\begin{cases} e^\delta = \mathcal{E}^\delta - E_2^{sc,\delta}, \\ h^\delta = \mathcal{H}^\delta - H_2^{sc,\delta}. \end{cases}$$

It is clear that (e^δ, h^δ) satisfy the Maxwell equations

$$\begin{cases} \overrightarrow{\operatorname{curl}}\,e^\delta = i\omega\,\mu_2\,h^\delta \quad \text{in } \Omega_2^\delta, \\ \overrightarrow{\operatorname{curl}}\,h^\delta = -i\omega\,\varepsilon_2\,e^\delta \quad \text{in } \Omega_2^\delta, \end{cases}$$

with the boundary conditions given by

$$\begin{aligned} &n\wedge(n\wedge e^\delta) + i\delta\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2 h^\delta\wedge n + \frac{i\delta}{\omega\varepsilon_1}\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(h^\delta\wedge n) \\ &-\delta\beta\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(e^\delta\wedge n) = w^\delta \quad \text{on } \Gamma^\delta, \end{aligned}$$

where

$$\begin{aligned} w^\delta &= \frac{i}{\omega\varepsilon_1}\Big(\int_{-\delta}^0 \overrightarrow{\operatorname{grad}}_{\Gamma^\delta_s}\operatorname{div}_{\Gamma^\delta_s}(H_1^\delta\wedge n)\,ds - \delta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(H_1^\delta(x_{\Gamma^\delta})\wedge n)\Big) \\ &\quad -i\beta\Big(\int_{-\delta}^0 \overrightarrow{\operatorname{grad}}_{\Gamma^\delta_s}\operatorname{div}_{\Gamma^\delta_s}(E_1^\delta\wedge n)\,ds - \delta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(E_1^\delta(x_{\Gamma^\delta})\wedge n)\Big) \\ &\quad -\int_{-\delta}^0 \mathcal{R}(E^\delta_{\Gamma^\delta_s})\,ds + (\gamma_1)^2\beta\int_{-\delta}^0 E_1^\delta\wedge n\,ds \\ &\quad +i\omega\mu_1\left(\frac{\gamma_1}{k_1}\right)^2\Big(\int_{-\delta}^0 H_1^\delta\wedge n\,ds - \delta\,H_1^\delta\wedge n\Big). \end{aligned}$$

From Theorem 4.4, it follows that for each $O\subset\subset\Omega_2^\delta$, there exists a positive constant C independent of δ such that

$$||e^\delta||_{H(\overrightarrow{\operatorname{curl}},O)} + ||h^\delta||_{H(\overrightarrow{\operatorname{curl}},O)} \le C\,||w^\delta||_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)}. \tag{4.42}$$

Now, in order to prove the optimal error estimate (4.41), we should estimate the following quantities:

$$\Big|\int_{-\delta}^0 \overrightarrow{\operatorname{grad}}_{\Gamma^\delta_s}\operatorname{div}_{\Gamma^\delta_s}(H_1^\delta\wedge n)\,ds - \delta\,\overrightarrow{\operatorname{grad}}_{\Gamma^\delta}\operatorname{div}_{\Gamma^\delta}(H_1^\delta(x_{\Gamma^\delta})\wedge n)\Big|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)},$$

$$\left| \int_{-\delta}^{0} \overrightarrow{\operatorname{grad}}_{\Gamma_s^\delta} \operatorname{div}_{\Gamma_s^\delta} (E_1^\delta \wedge n)\, ds - \delta \overrightarrow{\operatorname{grad}}_{\Gamma^\delta} \operatorname{div}_{\Gamma^\delta} (E_1^\delta(x_{\Gamma^\delta}) \wedge n) \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)},$$

$$\left| \int_{-\delta}^{0} \mathcal{R}(E_{\Gamma_s^\delta}^\delta)\, ds \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)}, \left| \int_{-\delta}^{0} E_1^\delta \wedge n\, ds \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)},$$

and

$$\left| \int_{-\delta}^{0} H_1^\delta \wedge n\, ds - \delta\, H_1^\delta \wedge n \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)}.$$

We need the following lemma.

LEMMA 4.8. *Let u be a smooth vectorial function defined in Ω_1^δ. We have*

$$\left| \int_{-\delta}^{0} \overrightarrow{\operatorname{grad}}_{\Gamma_s^\delta} \operatorname{div}_{\Gamma_s^\delta} (u \wedge n)\, ds - \delta \overrightarrow{\operatorname{grad}}_{\Gamma^\delta} \operatorname{div}_{\Gamma^\delta} (u(x_{\Gamma^\delta}) \wedge n) \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)} \leq C\,\delta^2,$$

where C is a strictly positive constant independent of δ.

Formally, since the derivatives of the tangential components of the magnetic and the electric field are bounded uniformly with respect to δ, it follows that there exists a strictly positive constant C such that

$$\left| \int_{-\delta}^{0} \mathcal{R}(E_{\Gamma_s^\delta}^\delta)\, ds \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)} \leq C\,\delta^2,$$

$$\left| \int_{-\delta}^{0} E_1^\delta \wedge n\, ds \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)} \leq C\,\delta^2,$$

and

$$\left| \int_{-\delta}^{0} H_1^\delta \wedge n\, ds - \delta\, H_1^\delta \wedge n \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)} \leq C\,\delta^2.$$

Lemma 4.2 yields

$$\left| \int_{-\delta}^{0} \overrightarrow{\operatorname{grad}}_{\Gamma_s^\delta} \operatorname{div}_{\Gamma_s^\delta} (H_1^\delta \wedge n)\, ds - \delta \overrightarrow{\operatorname{grad}}_{\Gamma^\delta} \operatorname{div}_{\Gamma^\delta} (H_1^\delta(x_{\Gamma^\delta}) \wedge n) \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)} \leq C\,\delta^2,$$

$$\left| \int_{-\delta}^{0} \overrightarrow{\operatorname{grad}}_{\Gamma_s^\delta} \operatorname{div}_{\Gamma_s^\delta} (E_1^\delta \wedge n)\, ds - \delta \overrightarrow{\operatorname{grad}}_{\Gamma^\delta} \operatorname{div}_{\Gamma^\delta} (E_1^\delta(x_{\Gamma^\delta}) \wedge n) \right|_{TL^2(\operatorname{curl}_{\Gamma^\delta},\Gamma^\delta)} \leq C\,\delta^2.$$

The proof of the error estimate is then complete by using (4.42). ∎

Finally, we note that the present method for obtaining the approximative impedance conditions is quite general, and it is applicable to coatings with arbitrary constitutive relations, such as dielectric layers with tensor permittivity and nonreciprocal coatings. Multiple layers may also be handled in a systematic manner.

5. Conclusions. In this paper, the diffraction problem by a chiral layer is formulated. The existence and the uniqueness of a solution to this problem is proved except for a possible discrete set of frequencies. In the case of a thin chiral layer, approximative impedance conditions are derived. The well-posedness of the new boundary-value problem obtained from these equivalent boundary conditions is established, and optimal error estimates between the unique solution of this new boundary-value problem and the solution of the diffraction problem by the chiral layer are obtained.

Appendix A. Let $(Y_l^m)_{-l\leq m\leq l}$ be an orthonormal sequence of spherical harmonics of order l on the unit sphere Σ normalized such that

$$\int_\Sigma Y_l^m \cdot \overline{Y}_{l'}^{m'} = \delta_{l,l'}^{m,m'}.$$

The basis functions for tangential fields on Σ_R are then

$$G_l^m = \frac{1}{\sqrt{l(l+1)}}\overrightarrow{\mathrm{grad}}_\Sigma Y_l^m \text{ and } R_l^m = n \wedge G_l^m \quad \text{for } -l\leq m\leq l,\ l\geq 1.$$

The tangential vector fields G_l^m and G_l^m are an orthonormal basis on the unit sphere (in the L^2 inner product). The multipolar vector basis for Maxwell's equations

$$\text{(A.1)} \qquad \begin{cases} \overrightarrow{\mathrm{curl}}\, E = i\omega\,\mu_2\, H, \\ \overrightarrow{\mathrm{curl}}\, H = -i\omega\,\varepsilon_2\, E \end{cases}$$

are given by

$$\text{(A.2)} \qquad \begin{cases} M_{j,l}^m(x) = \overrightarrow{\mathrm{curl}}\Big(x\, h_l^{(j)}(k|x|)\, Y_l^m\Big(\dfrac{x}{|x|}\Big)\Big), \\ N_{j,l}^m(x) = \dfrac{1}{ik}\overrightarrow{\mathrm{curl}}\, M_{j,l}^m(x), \end{cases}$$

where $j\in\{1,2\}, m=-l,\dots,l, l\geq 1, k=\omega\sqrt{\varepsilon_2\mu_2}$. $h_l^{(j)}$ denotes the spherical Hankel functions of the j kind and order l. It is well known that any solution of Maxwell's equations $(A.1)$ satisfying the outgoing radiation condition can be written for $|x|>R$ in the form

$$\text{(A.3)} \qquad E(x) = \sum_{l\geq 1}\sum_{m=-l}^{l}\Big(\alpha_l^m N_{1,l}^m(x) + \beta_l^m M_{1,l}^m(x)\Big).$$

In this appendix, we shall prove that if $\Im m \int_{\Sigma_R}(\overrightarrow{\mathrm{curl}}\, E \wedge n)\cdot\overline{E} = 0$ then $E\equiv 0$ in $|x|>R$. We need to express

$$\Im m \int_{\Sigma_R}\Big(\overrightarrow{\mathrm{curl}}\, E\wedge n\Big)\cdot\overline{E},$$

where E is a solution to Maxwell's equations $(A.1)$ satisfying the outgoing radiation condition at infinity in terms of the coefficients of the expansions for E. Using the expansion $(A.3)$ we obtain

$$\begin{aligned} \overrightarrow{\mathrm{curl}}\, E\wedge n &= \sum_{l\geq 1}\sum_{m=-l}^{l}\Big(\sqrt{l(l+1)}\,\beta_l^m\, h_l^{(1)}(kR)\, G_l^m\Big(\frac{x}{R}\Big)\Big) \\ &\quad + \frac{i}{kR}\sum_{l\geq 1}\sum_{m=-l}^{l}\Big(\sqrt{l(l+1)}\,\alpha_l^m\Big\{ h_l^{(1)}(kR) + kR(h_l^{(1)})'(kR)\Big\} R_l^m\Big(\frac{x}{R}\Big)\Big) \end{aligned}$$

on Σ_R. Now, from the definitions of G_l^m and R_l^m, we show that

$$-n\wedge\Big(n\wedge N_{1,l}^m(x)\Big) \,=\, -\frac{i}{kR}\sqrt{l(l+1)}\,\Big(h_l^{(1)}(kR)+kR(h_l^{(1)})'(kR)\Big)\,G_l^m\Big(\frac{x}{R}\Big),\quad \text{on } \Sigma_R$$

and

$$-n\wedge\Big(n\,\wedge\,M_{1,l}^m(x)\Big) \,=\, -\sqrt{l(l+1)}\,h_l^{(1)}(kR)\,R_l^m\Big(\frac{x}{R}\Big)\quad \text{on } \Sigma_R.$$

Direct computation yields

$$\int_{\Sigma_R}\Big(\overrightarrow{\operatorname{curl}}\,E\wedge n\Big)\cdot\overline{E} \,=\, \sum_{l\geq 1}\sum_{m=-l}^{l} l(l+1)\,\Big(|\alpha_l^m|^2+|\beta_l^m|^2\Big)\,\Im m\,\Big(h_l^{(1)}(kR)\,(h_l^{(1)})'(kR)\Big).$$

From

$$\Im m\; h_l^{(1)}(kR)\,\overline{(h_l^{(1)})'}(kR) \,<\, 0\quad \text{for all } l\geq 1,$$

proved in [2], we obtain that

$$\Im m\,\int_{\Sigma_R}\Big(\overrightarrow{\operatorname{curl}}\,E\,\wedge\,n\Big)\,\cdot\,\overline{E}\,=\,0\,\Longrightarrow\,E\,\equiv 0 \text{ in } |x|>R.$$

Appendix B. For the convenience of the reader we summarize in this appendix the vector identities used in this paper. Let Ω be a bounded domain with a smooth boundary $\partial\Omega$. n denotes its outward normal. Let u and v be two sufficiently smooth vectorial functions defined in Ω and w on $\partial\Omega$ such that $n\cdot w=0$. We have

$$\int_\Omega \overrightarrow{\operatorname{curl}}\,u\,\cdot\,v\,=\,\int_\Omega u\,\cdot\,\overrightarrow{\operatorname{curl}}\,v-\int_{\partial\Omega}\Big(u\,\wedge\,n\Big)\cdot\Big(n\,\wedge\,(n\,\wedge\,v)\Big).$$

$$\overrightarrow{\operatorname{curl}}\,\overrightarrow{\operatorname{curl}}\,u\,=\,\overrightarrow{\operatorname{grad}}\,\operatorname{div}u-\Delta\,u.$$

$$\overrightarrow{\operatorname{curl}}\,\overrightarrow{\operatorname{grad}}\,u\,=\,0.$$

$$\operatorname{div}\overrightarrow{\operatorname{curl}}\,u\,=\,0.$$

$$\operatorname{div}_{\partial\Omega}(u\,\wedge\,n)\,=\,n\,\cdot\,\overrightarrow{\operatorname{curl}}\,u|_{\partial\Omega}.$$

$$\int_{\partial\Omega}\overrightarrow{\operatorname{grad}}_{\partial\Omega}\operatorname{div}_{\partial\Omega}(u\,\wedge\,n)\,\cdot\,(v\,\wedge\,n)\,=\,-\int_{\partial\Omega}\operatorname{div}_{\partial\Omega}(u\,\wedge\,n)\operatorname{div}_{\partial\Omega}(v\,\wedge\,n).$$

$$\operatorname{div}_{\partial\Omega}\overrightarrow{\operatorname{curl}}_{\partial\Omega}\,w\,=\,0.$$

$$-\operatorname{curl}_{\partial\Omega}\overrightarrow{\operatorname{curl}}_{\partial\Omega}\,w\,=\,\Delta_{\partial\Omega}\,w\,=\,\operatorname{div}_{\partial\Omega}\overrightarrow{\operatorname{grad}}_{\partial\Omega}\,w.$$

$$\operatorname{curl}_{\partial\Omega}\overrightarrow{\operatorname{grad}}_{\partial\Omega}\,w\,=\,0.$$

$$\mathrm{div}_{\partial\Omega}\,(w \wedge n) \;=\; \mathrm{curl}_{\partial\Omega}\, w.$$

$$\overrightarrow{\mathrm{curl}}\, u \;=\; \mathrm{curl}_{\partial\Omega}\, u_{\partial\Omega}\, n + \overrightarrow{\mathrm{curl}}_{\partial\Omega}(u \cdot n) + (\mathcal{R} - 2c)(u \wedge n) - \frac{\partial}{\partial s}(u \wedge n),$$

where

$$u_{\partial\Omega} \;=\; -n \wedge (n \wedge u).$$

$$\begin{aligned}\overrightarrow{\mathrm{curl}}\,\overrightarrow{\mathrm{curl}}\, u \;=\;& (\mathrm{curl}_{\partial\Omega}\,\overrightarrow{\mathrm{curl}}_{\partial\Omega}\,(u \cdot n))\, n + \mathrm{div}_{\partial\Omega}\,(\mathcal{R}\, u_{\partial\Omega})\, n + \mathrm{div}_{\partial\Omega}\,\Big(\frac{\partial}{\partial n}\, u_{\partial\Omega}\Big)\\ &+\overrightarrow{\mathrm{curl}}_{\partial\Omega}\,\mathrm{curl}_{\partial\Omega}\, u_{\partial\Omega} + \Big(\mathcal{R} - 2c - \frac{\partial}{\partial s}\Big)\Big(-\overrightarrow{\mathrm{grad}}_{\partial\Omega}(u \cdot n)\\ &+\mathcal{R}\, u_{\partial\Omega} + \frac{\partial}{\partial s}\, u_{\partial\Omega}\Big).\end{aligned}$$

REFERENCES

[1] T. Abboud, *Formulation variationnelle des équations de Maxwell dans un réseau bipériodique de* $\mathbb{R}^3$, C. R. Acad. Sci. Paris Sér. I, 317 (1993), pp. 245–248.

[2] T. Abboud and J. C. Nédélec, *Electromagnetic waves in an inhomogeneous medium*, J. Math. Anal. Appl., 164 (1992), pp. 40–58.

[3] H. Ammari and J. C. Nédélec, *Sur les conditions d'impédance généralisées*, C. R. Acad. Sci. Paris Sér. I, 322 (1996), pp. 995–1000.

[4] H. Ammari and J. C. Nédélec, *Time harmonic electromagnetic fields in chiral media*, in Modern Mathematical Methods in Diffraction Theory and its Applications, E. Meister, Ed., Methoden Verfahren Math. Phys., 42 (1997), pp. 174–202.

[5] D. F. Arago, *Mémoire sur une modification remarquable qu'éprouvent les rayons lumineux dans leur passage à travers certains corps diaphanes, et sur quelques autres nouveaux phénomènes d'optique*, Mém. Sci. Math. Phys. Inst., 1811, pp. 93–134.

[6] S. Bassiri, C. H. Papas, and N. Engheta, *Electromagnetic wave propagation through a dielectric-chiral interface and through a chiral slab*, J. Opt. Soc. Amer. A, 5 (1988), pp. 1450–1459.

[7] J. B. Biot, *Mémoire sur un nouveau genre d'oscillation que les molécules de la lumière éprouvent en traversant certains cristaux*, Mém. Sci. Math. Phys. Inst., 1812, pp. 1–372.

[8] C. F. Bohren, *Scattering of electromagnetic waves by an optically active cylinder*, J. Colloi Interface Sci., 66 (1978), pp. 105–109.

[9] C. F. Bohren, *Scattering of electromagnetic waves by an optically active spherical shell,* J. Chem. Phys., 62 (1975), pp. 1466–1571.

[10] C. F. Bohren, *Light scattering by an optically active sphere*, Chem. Phys. Lett., 29 (1974), pp. 458–462.

[11] F. Brezzi, *On the existence and uniqueness of saddle-point problems arising from Lagrange multipliers*, RAIRO Anal. Numér., 8-R2 (1974), pp. 129–151.

[12] A. L. Cauchy, *Mémoire sur l'application de l'analyse mathématique à la recherche des lois générales observées par des physiciens, et, en particulier, sur les lois de polarisation circulaire*, C. R. Acad. Sci. Paris, 15 (1842), pp. 910–918.

[13] B. Engquist and J. C. Nédélec, *Effective Boundary Conditions for Acoustic and Electro-Magnetic Scattering in Thin Layers*, manuscript.

[14] F. I. Fedorov, *On the theory of optical activity in crystals.* I. *The law of conservation of energy and the optical activity tensors*, Opt. Spectrosc. (USSR), 6 (1959), pp. 237–240.

[15] A. Fresnel, *Extrait d'un mémoire sur la double réfraction particulière que présente le cristal de roche dans la direction de son axe*, Bull. Soc. Philomat., 1822, pp. 191–198.

[16] D. J. Hoppe and Y. Rahmat-Samii, *Higher order impedance boundary conditions revisited: Application to chiral coatings*, J. Elect. Waves Appl., 8 (1994), pp. 1303–1329.

[17] D. L. Jaggar, A. R. Michelson, and C. H. Papas, *On electromagnetic waves in chiral media*, Appl. Phys., 18 (1979), pp. 211–216.

[18] D. L. Jaggar and J. C. Liu, *Chiral layers on curved surfaces*, J. Elect. Waves Appl., 1992.

[19] A. de La Bourdonnaye, *Décomposition de $H_{\mathrm{div}}^{-1/2}(\Gamma)$ et nature de l'opérateur de Steklov-Poincaré du problème extérieur de l'électromagnétisme*, C. R. Acad. Sci. Paris Sér. I, 316 (1993), pp. 369–372.

[20] A. Lakhtakia, V. K. Varadan, and V. V. Varadan, *Time-Harmonic Electromagnetic Fields in Chiral Media*, Lecture Notes in Physics 355, Springer, Berlin, 1989.

[21] A. Lakhtakia, V. K. Varadan and V. V. Varadan, *Scattering and absorption characteristics of lossy dielectric, chiral, nonspherical objects*, Appl. Optics, 24 (1985), pp. 4146–4154.

[22] A. Lakhtakia, *Beltrami Fields in Chiral Media*, World Sci. Ser. Contemp. Chem. Phys. 2, World Sci. Publishing, River Edge, NJ, 1994.

[23] L. Pasteur, *Sur les relations qui peuvent exister entre la forme cristalline, la composante chimique et le sens de la polarisation rotatoire*, Ann. Chimie et Physique, 24 (1848), pp. 442–459.

[24] V. Prelog, *Chirality in Chemistry*, Nobel Lecture, Stockholm, Dec. 12, 1975.

[25] Z. X. Shen, *Scattering by a circular impedance cylinder coated with a composite anisotropic chiral sheath*, J. Elect. Waves Appl., 8 (1994), pp. 1625–1644.

[26] F. Trèves, *Introduction to Pseudo-Differential and Fourier Integral Operators*, Plenum Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 424–436, March 1998

007

BIFURCATION STRUCTURE OF STATIONARY SOLUTIONS OF A LOTKA–VOLTERRA COMPETITION MODEL WITH DIFFUSION*

YUKIO KAN-ON†

Abstract. In this paper, we consider a Lotka–Volterra competition model with diffusion which describes the dynamics of the population of two competing species, and establish the bifurcation structure of positive stationary solutions of the model. To do this, we shall employ the comparison principle and the bifurcation theory and study the spatial profile of positive stationary solutions.

Key words. Lotka–Volterra competition model, comparison principle, bifurcation theory

AMS subject classification. 35B32

PII. S0036141096305784

1. Introduction.

In order to understand the mechanism of phenomena which appear in various fields, we often use the system of reaction–diffusion equations

$$\begin{cases} \boldsymbol{u}_t = \varepsilon\, D\, \Delta \boldsymbol{u} + \boldsymbol{f}(\boldsymbol{u}), & x \in \Omega, \quad t > 0, \\ \frac{\partial}{\partial \nu} \boldsymbol{u} = 0, & x \in \partial\Omega, \quad t > 0, \end{cases} \tag{1.1}$$

with suitable initial condition, and discuss the existence of stationary solutions and their stability properties, where $\boldsymbol{u} \in \mathbf{R}^n$, ε is a positive constant, D is a diagonal matrix whose elements are positive, $\boldsymbol{f} : \mathbf{R}^n \to \mathbf{R}^n$ is a smooth function, Ω is a bounded domain in $\mathbf{R}^m$ with smooth boundary $\partial\Omega$, and ν is the outward unit normal vector on $\partial\Omega$.

When $n = 1$ holds, it is comparatively easy to study the existence of stationary solutions of (1.1) and their spatial profile by the analysis of motions in the phase plane, because the so-called *comparison principle* holds. Furthermore, it is well known that for suitable $f(u)$, the global attractor $\mathcal{A}$ of (1.1) is represented as $\mathcal{A} = \cup_{e \in \mathcal{E}} W^u(e)$, where $\mathcal{E}$ is the set of stationary solutions of (1.1) and $W^u(e)$ is an unstable manifold of (1.1) at $u = e$ (for example, see Chapter 4 in Hale [6]). This suggests that in order to understand the precise asymptotic behavior of solutions of (1.1), it is important to study the number of stationary solutions and their stability properties.

In general, the comparison principle does not always hold in case of $n \geq 2$. This leads to the considerable complexity for the study of the existence and stability of stationary solutions of (1.1). In this paper, to approach the problem for $n \geq 2$, we consider a Lotka–Volterra competition model with diffusion which describes the dynamics of the population $\boldsymbol{u} = (u, v)$ of two competing species, and assume the nonlinearity $\boldsymbol{f}(\boldsymbol{u}) = (f, g)(\boldsymbol{u})$ with

$$f(\boldsymbol{u}) = u\,(1 - u - c\,v), \quad g(\boldsymbol{u}) = v\,(a - b\,u - v),$$

which is most simple in the framework of Lotka–Volterra competition models, where a, b, and c are positive constants. We should note here that under the above nonlinearity,

*Received by the editors June 28, 1996; accepted for publication (in revised form) January 6, 1997.

http://www.siam.org/journals/sima/29-2/30578.html

†Department of Mathematics, Faculty of Education, Ehime University, Matsuyama 790, Japan (kanon@ed.ehime-u.ac.jp).

the comparison principle holds for (1.1) with respect to the order relation $\preceq_o$ which will be defined in the next section.

Many authors have studied the existence and stability of positive stationary solutions for a variety of Lotka–Volterra competition models including (1.1) (for instance, see Dancer [3] and Gui and Lou [5]). For the asymptotic behavior of solutions of (1.1), the following is well known (for example, see de Mottoni [4]):

(I) If $a < \min(b, 1/c)$, then $\lim_{t\to+\infty} \boldsymbol{u}(t,x) = (1,0)$.

(II) If $b < a < 1/c$, then

$$\lim_{t\to+\infty} \boldsymbol{u}(t,x) = \left(\frac{1-a\,c}{1-b\,c}, \frac{a-b}{1-b\,c}\right).$$

(III) If $1/c < a < b$, then both $(1,0)$ and $(0,a)$ are locally stable.

(IV) If $a > \max(b, 1/c)$, then $\lim_{t\to+\infty} \boldsymbol{u}(t,x) = (0,a)$.

Precisely speaking for case (III), Kishimoto and Weinberger [8] showed that any spatially inhomogeneous positive stationary solution is unstable if Ω is convex. On the other hand, Matano and Mimura [9] proved that there exist stable spatially inhomogeneous positive stationary solutions for suitably nonconvex domain Ω. (Along this line, we also refer to Mimura, Ei, and Fang [10] and Kan-on and Yanagida [7].) These results mean that there exist many positive stationary solutions of (1.1) and that their stability properties crucially depend on the shape of the domain Ω. We now address the following question: suppose that Ω is suitably fixed. How many positive stationary solutions does (1.1) have? In this paper, to answer this problem, we assume

$$\text{(A)} \qquad \Omega = (0,1), \quad D = I_2, \quad a = 1, \quad b = c > 1,$$

study the spatial profile of positive stationary solutions by employing the comparison principle and the bifurcation theory, and then establish the bifurcation structure of the positive stationary solutions of (1.1).

2. Statement of result. We set $X = \{\, \boldsymbol{u} \in C^2([0,1], \mathbf{R}^2) \mid \boldsymbol{u}_x(0) = 0,\ \boldsymbol{u}_x(1) = 0 \,\}$ and define the order relations $\preceq_s$ and $\preceq_o$ in the following manner:

$$(u_1, v_1) \preceq_s (u_2, v_2) \iff u_1 \le u_2,\ v_1 \le v_2,$$
$$(u_1, v_1) \preceq_o (u_2, v_2) \iff u_1 \le u_2,\ v_1 \ge v_2.$$

We denote by $\prec_s$ and $\prec_o$ the relations obtained from the above definition by replacing $\le$ with $<$. We shall say that $\boldsymbol{u}(x)$ is *positive* if $\boldsymbol{u}(x) \succ_s 0$ holds on $[0,1]$.

Setting $\bar{u} = 1/(1+b)$, $\bar{\boldsymbol{u}} = (\bar{u}, \bar{u})$, $\varepsilon_0 = +\infty$, and $\varepsilon_n = (b-1)/(n^2\,\pi^2\,(b+1))$ $(n \in \mathbf{N})$, we can easily check that spatially inhomogeneous positive stationary solutions of (1.1) bifurcate from $\boldsymbol{u} = \bar{\boldsymbol{u}}$ at $\varepsilon = \varepsilon_n$ for each $n \in \mathbf{N}$. In the next section, we shall study the geometrical position of the curve of such stationary solutions.

Here is the main result of this paper.

THEOREM 2.1. *Under the assumption* (A)*, the following holds for each* $n \in \mathbf{N}$*:*

(i) *For any* $\varepsilon \in (0, \varepsilon_n)$*, there exists a pair of spatially inhomogeneous positive stationary solutions* $\boldsymbol{u}_n^{\pm}(x,\varepsilon)$ *of* (1.1) *such that*

$$(-1)^j\, \boldsymbol{u}_{nx}^{-}(x,\varepsilon) \prec_o 0, \quad (-1)^j\, \boldsymbol{u}_{nx}^{+}(x,\varepsilon) \succ_o 0$$

are satisfied for each $0 \le j < n$ *and* $x \in (j/n, (j+1)/n)$*.*

(ii) *Both* $\boldsymbol{u}_n^{-}(.,\varepsilon)$ *and* $\boldsymbol{u}_n^{+}(.,\varepsilon)$ *are* C^1*-class functions from* $(0, \varepsilon_n)$ *to* X*.*

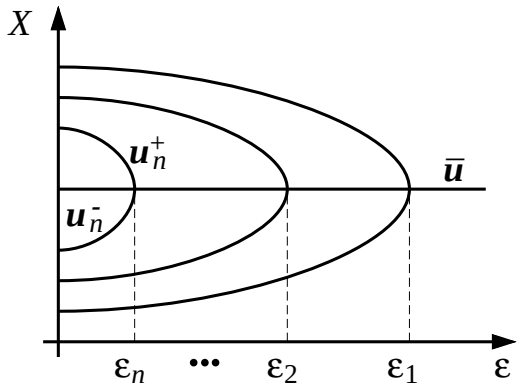

FIG. 1. *Bifurcation structure of positive stationary solutions.*

(iii) (1.1) *has no positive stationary solutions other than* $\bar{\boldsymbol{u}}$ *and* $\boldsymbol{u}_j^{\pm}(x,\varepsilon)$ $(1 \le j < n)$ *for any* $\varepsilon \in [\varepsilon_n, \varepsilon_{n-1})$.

Figure 1 indicates the bifurcation structure of positive stationary solutions of (1.1). We should note here that the secondary bifurcation of stationary solutions never occurs. In consideration of the result in Chafee and Infante [1], the above theorem means that the bifurcation structure of positive stationary solutions of (1.1) is similar to that of stationary solutions of

$$\begin{cases} u_t = \varepsilon\, u_{xx} + u\,(1-u^2), & x \in (0,1), \quad t > 0, \\ u_x = 0, & x = 0, 1, \quad t > 0. \end{cases}$$

We shall prove the above theorem in the following sections.

3. Bifurcation from spatially homogeneous solution. In this section, we calculate spatially inhomogeneous positive stationary solutions bifurcating from $\boldsymbol{u} = \bar{\boldsymbol{u}}$ at $\varepsilon = \varepsilon_n$ for each $n \in \mathbf{N}$. To do it, we set

$$\hat{\varepsilon}_n(\mu) = \varepsilon_n + \tilde{\varepsilon}_n(\mu)\,\mu^2, \quad \hat{\boldsymbol{u}}_n(x,\mu) = \bar{\boldsymbol{u}} + \mu\,\boldsymbol{e}\,\cos(n\,\pi\,x) + \mu^2\,\tilde{\boldsymbol{u}}_n(x,\mu), \tag{3.1}$$

where $\boldsymbol{e} = (1,-1)$ is an eigenvector of the 2×2-matrix $\boldsymbol{f}_{\boldsymbol{u}}(\bar{\boldsymbol{u}}) - \varepsilon_n\, n^2\, \pi^2\, I_2$ corresponding to the eigenvalue 0. Substituting $\varepsilon = \hat{\varepsilon}_n(\mu)$ and $\boldsymbol{u} = \hat{\boldsymbol{u}}_n(x,\mu)$ into (1.1), we have

$$\tilde{\varepsilon}_n(0) = -\frac{(11\,b-5)\,(b-1)}{2\,(5\,b-3)\,n^2\,\pi^2}$$

and then obtain $\tilde{\varepsilon}_n(0) < 0$ because of (A). We have the following by the bifurcation theory.

LEMMA 3.1 (see Crandall and Rabinowitz [2]). *For each* $n \in \mathbf{N}$, *there exist* $\delta_n > 0$, $\mu_n > 0$, *and* C^1*-class functions* $\hat{\varepsilon}_n(\mu)$, $\hat{\boldsymbol{u}}_n(.,\mu)$ *defined on* $(-\mu_n, \mu_n)$ *such that the following properties hold:*

(i) $\hat{\boldsymbol{u}}_n(x,\mu)$ *is a spatially inhomogeneous positive stationary solution of* (1.1) *with* $\varepsilon = \hat{\varepsilon}_n(\mu)$ *for each* $\mu \in (-\mu_n, 0) \cup (0, \mu_n)$.

(ii) $\hat{\varepsilon}_n(0) = \varepsilon_n$ *and* $\hat{\varepsilon}_n(\mu) < \varepsilon_n$ *for any* $\mu \in (-\mu_n, 0) \cup (0, \mu_n)$ *are satisfied.*

(iii) *Suppose that* $\boldsymbol{w}(x)$ *is a spatially inhomogeneous positive stationary solution of* (1.1) *for* $\varepsilon = \eta$. *If* $|\,\eta - \varepsilon_n\,| < \delta_n$ *and* $\|\,\boldsymbol{w} - \bar{\boldsymbol{u}}\,\|_X < \delta_n$ *hold, then there exists* $\rho \in (-\mu_n, \mu_n)$ *such that* $\eta = \hat{\varepsilon}_n(\rho)$ *and* $\boldsymbol{w} = \hat{\boldsymbol{u}}_n(.,\rho)$.

4. Property of positive stationary solutions. Let $\boldsymbol{u}_0(x) = (u_0, v_0)(x)$ be an arbitrary spatially inhomogeneous positive stationary solution of (1.1) for $\varepsilon > 0$. It is obvious from Lemma 3.1 that such a solution exists for some $\varepsilon > 0$. In this section, we study spatial profiles of $\boldsymbol{u}_0(x)$ which will be used in order to discuss the possibility of the secondary bifurcation in the next section.

By the boundary condition, we can regard $\boldsymbol{u}_0(x)$ as a periodic function with period 2 which satisfies $\boldsymbol{u}(x) = \boldsymbol{u}(-x)$ for any $x \in \mathbf{R}$. By the comparison principle, it is easy to show that $0 < u_0(x) + v_0(x)\,(\equiv K(x)) < 1$ holds for any $x \in \mathbf{R}$. Furthermore, we find that $H(x) = u_0(x) - v_0(x)$ satisfies

$$\begin{cases} 0 = \varepsilon\, H_{xx} + (1 - K(x))\, H, & x \in (0,1), \\ H_x(0) = 0, \quad H_x(1) = 0. \end{cases}$$

If $H(x) \gtrless 0$ holds for any $x \in [0,1]$, then we have

$$0 = \int_0^1 \{\, \varepsilon\, H_{xx}(x) + (1 - K(x))\, H(x) \,\}\, dx \gtrless 0.$$

This contradiction implies that there exists $x_0 \in [0,1]$ such that $H(x_0) = 0$ and $H(x) \neq 0$ for any $x \in [0, x_0)$.

We consider the case $x_0 = 0$. We obtain $H(x) = 0$ for any $x \in \mathbf{R}$ by virtue of uniqueness and then find that $u_0(x)$ satisfies

$$\begin{cases} 0 = \varepsilon\, u_{xx} + u\,\{\, 1 - (1+b)\, u \,\}, & x \in (0,1), \\ u_x(0) = 0, \quad u_x(1) = 0. \end{cases}$$

Since $u_0(x) > 0$ holds for any $x \in \mathbf{R}$, we obtain $u_0(x) = 1/(1+b)$ for any $x \in \mathbf{R}$. This contradicts that $\boldsymbol{u}_0(x)$ is spatially inhomogeneous. In a similar manner, we can derive a contradiction when $x_0 = 1$ holds. Hence we have $x_0 \in (0,1)$ and $H_x(x_0) \neq 0$. If $\varepsilon > 4\, x_0^2/\pi^2$ is satisfied, then we have

$$\begin{aligned} 0 &= \int_0^{x_0} \{\, \varepsilon\, H_{xx}(x) + (1 - K(x))\, H(x) \,\}\, \cos(\pi\, x/(2\, x_0))\, dx \\ &= \int_0^{x_0} \{\, 1 - K(x) - \varepsilon\, \pi^2/(4\, x_0^2) \,\}\, H(x)\, \cos(\pi\, x/(2\, x_0))\, dx \neq 0 \end{aligned}$$

because of integration by parts. This indicates that ε must satisfy $\varepsilon \leq 4\, x_0^2/\pi^2$.

We set

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

whose operation is the change of the role of u and v. Since $g(\boldsymbol{u}) = f(T\, \boldsymbol{u})$ is satisfied for any $\boldsymbol{u}$ because of (A), we find that $\boldsymbol{f}(T\, \boldsymbol{u}) = T\, \boldsymbol{f}(\boldsymbol{u})$ holds for any $\boldsymbol{u}$ and that $T\, \boldsymbol{u}_0(x)$ is a spatially inhomogeneous positive stationary solution of (1.1). Hence we see that

$$\boldsymbol{U}(x)\,(\equiv (U,V)(x)) = \boldsymbol{u}_0(x) - T\, \boldsymbol{u}_0(2\, x_0 - x)$$

is a periodic function with period 2 and that $\boldsymbol{U}(x)$ satisfies $\boldsymbol{U}(x_0) = 0$, $U_x(x_0) = V_x(x_0)$, and

$$0 = \varepsilon\, \boldsymbol{U}_{xx} + \boldsymbol{f}_{\boldsymbol{u}}^1(x)\, \boldsymbol{U}, \quad x \in \mathbf{R},$$

where

$$\boldsymbol{f}_{\boldsymbol{u}}^{1}(x)=\int_{0}^{1}\boldsymbol{f}_{\boldsymbol{u}}(\theta\,\boldsymbol{u}_{0}(x)+(1-\theta)\,T\,\boldsymbol{u}_{0}(2\,x_{0}-x))\,d\theta.$$

We assume that $U_x(x_0)\neq 0$ holds. It follows that there exists $x_1\in(x_0,x_0+2]$ such that $U(x_1)\,V(x_1)=0$ and $U(x)\,V(x)>0$ for any $x\in(x_0,x_1)$. If $U(x_1)=0$ is satisfied, then we have

$$\begin{aligned}0=\int_{x_0}^{x_1}(\varepsilon\,u_{0xx}(x)+f(\boldsymbol{u}_0(x)))\,v_0(2\,x_0-x)\,dx&=\varepsilon\,(u_0(x_1)\,U_x(x_1)-u_0(x_0)\,U_x(x_0))\\&\quad-\int_{x_0}^{x_1}(U(x)+b\,V(x))\,u_0(x)\,v_0(2\,x_0-x)\,dx\neq 0\end{aligned}$$

because of integration by parts. In a similar manner, we can derive a contradiction when $V(x_1)=0$ holds. Hence we have $\boldsymbol{U}_x(x_0)=0$. By uniqueness, we obtain $\boldsymbol{u}_0(x)=T\,\boldsymbol{u}_0(2\,x_0-x)$ for any $x\in\mathbf{R}$.

LEMMA 4.1. *Let $\boldsymbol{u}_0(x)=(u_0,v_0)(x)$ be an arbitrary spatially inhomogeneous positive stationary solution of (1.1) for $\varepsilon>0$. Then there exists a unique constant $x_0\in(0,1)$ such that $u_0(x)\neq v_0(x)$ for any $x\in[0,x_0)$ and*

$$\boldsymbol{u}_0(x)=\begin{cases}\boldsymbol{u}_0(r(x)) & \text{if } q(x)\equiv 0 \mod 4,\\ T\,\boldsymbol{u}_0(x_0-r(x)) & \text{if } q(x)\equiv 1 \mod 4,\\ T\,\boldsymbol{u}_0(r(x)) & \text{if } q(x)\equiv 2 \mod 4,\\ \boldsymbol{u}_0(x_0-r(x)) & \text{if } q(x)\equiv 3 \mod 4\end{cases}$$

for any $x\in\mathbf{R}$, where $q(x)\in\mathbf{Z}$ and $r(x)\in[0,x_0)$ are determined by $x=r(x)+q(x)\,x_0$.

The linearized operator $\mathcal{L}$ of (1.1) around $\boldsymbol{u}_0(x)$ can be represented as $\mathcal{L}\,\boldsymbol{u}=\varepsilon\,\boldsymbol{u}_{xx}+\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{u}_0(x))\,\boldsymbol{u}$. Since $\boldsymbol{u}_{0x}(x)$ is a nontrivial solution of $\mathcal{L}\,\boldsymbol{u}=0$ with $\boldsymbol{u}(0)=0$, we obtain $\boldsymbol{u}_{0xx}(0)\neq 0$. We set

$$\begin{pmatrix}f_u^0 & f_v^0\\ g_u^0 & g_v^0\end{pmatrix}(x)=\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{u}_0(x)).$$

From the functional form of $\boldsymbol{f}(\boldsymbol{u})$, we have $f_v^0(x)<0$ and $g_u^0(x)<0$ for any $x\in\mathbf{R}$.

LEMMA 4.2. *Let $(\varphi,\psi)(x)$ be an arbitrary solution of $\mathcal{L}\,\boldsymbol{u}=0$. Suppose that there exist x_2 and x_3 $(x_2<x_3)$ such that $\varphi(x_2)=0=\varphi(x_3)$ (resp., $\psi(x_2)=0=\psi(x_3)$) and $\varphi(x)>0$ (resp., $\psi(x)>0$) for any $x\in(x_2,x_3)$. Then $\psi(x)<0$ (resp., $\varphi(x)<0$) is satisfied for some $x\in(x_2,x_3)$.*

Proof. We only show the proof for the former case, because the latter can be proved in a similar manner. If $\psi(x)\geq 0$ holds for any $x\in(x_2,x_3)$, then we have

$$\begin{aligned}0=&\int_{x_2}^{x_3}(\varepsilon\,\varphi_{xx}(x)+f_u^0(x)\,\varphi(x)+f_v^0(x)\,\psi(x))\,u_0(x)\,dx\\=&\varepsilon\,(\varphi_x(x_3)\,u_0(x_3)-\varphi_x(x_2)\,u_0(x_2))-\int_{x_2}^{x_3}(\varphi(x)+b\,\psi(x))\,u_0(x)^2\,dx<0.\end{aligned}$$

This contradiction implies that the desired result holds. □

Since $\frac{d}{dx}(u_{0xx}\,v_{0xx})(0)=0$ and

$$\varepsilon\,\tfrac{d^2}{dx^2}(u_{0xx}\,v_{0xx})(0)=-f_v^0(0)\,v_{0xx}(0)^2-g_u^0(0)\,u_{0xx}(0)^2>0$$

hold if $u_{0xx}(0)\,v_{0xx}(0) = 0$ is satisfied, it follows from $\boldsymbol{u}_{0x}(2\,x_0) = 0$ that there exists $x_4 \in (0, 2\,x_0]$ such that $u_{0x}(x_4)\,v_{0x}(x_4) = 0$ and $u_{0x}(x)\,v_{0x}(x) \neq 0$ for any $x \in (0, x_4)$.

We consider the case $\boldsymbol{u}_{0x}(x_4) \neq 0$. It is obvious that

$$\boldsymbol{U}(x)\,(\equiv (U,V)(x)) = \boldsymbol{u}_0(x) - \boldsymbol{u}_0(2\,x_4 - x)$$

is a periodic function with period 2 and that $\boldsymbol{U}(x)$ satisfies $\boldsymbol{U}(x_4) = 0$, $\boldsymbol{U}_x(x_4) \neq 0$, $U_x(x_4)\,V_x(x_4) = 0$, $\boldsymbol{U}_{xx}(x_4) = 0$, and

$$\begin{aligned} \varepsilon\,U_{xxx}(x_4)\,V_x(x_4) &= -f_v^0(x_4)\,V_x(x_4)^2 > 0 \quad \text{if } U_x(x_4) = 0, \\ \varepsilon\,V_{xxx}(x_4)\,U_x(x_4) &= -g_u^0(x_4)\,U_x(x_4)^2 > 0 \quad \text{if } V_x(x_4) = 0. \end{aligned}$$

Hence we see that there exists $x_5 \in (x_4, x_4 + 2]$ such that $U(x_5)\,V(x_5) = 0$ and $U(x)\,V(x) > 0$ for any $x \in (x_4, x_5)$. We can derive a contradiction in the same manner as in the proof of Lemma 4.1. Therefore we obtain $\boldsymbol{u}_{0x}(x_4) = 0$. Since $\boldsymbol{u}_0(x) = \boldsymbol{u}_0(2\,x_4 - x)$ holds for any $x \in \mathbf{R}$, we find $H(2\,x_4 - x_0) = 0$. By Lemma 4.1, we have either $x_4 = x_0$ or $x_4 = 2\,x_0$. If $x_4 = x_0$ holds, then we obtain $H(x) = 0$ for any $x \in \mathbf{R}$ because of $H(x_0) = 0$ and $H_x(x_0) = 0$, which indicates that $\boldsymbol{u}_0(x)$ is a constant function. This contradiction implies that $x_4 = 2\,x_0$ holds. By Lemma 4.2 and the choice of x_4, we have $u_{0x}(x)\,v_{0x}(x) < 0$ for any $x \in (0, 2\,x_0)$.

LEMMA 4.3. *Let $\boldsymbol{u}_0(x) = (u_0, v_0)(x)$ be an arbitrary spatially inhomogeneous positive stationary solution of* (1.1) *for $\varepsilon > 0$, and let x_0 be the unique constant given in Lemma* 4.1. *Then there exists $\ell \in \mathbf{N}$ such that $2\,x_0\,\ell = 1$ holds. Furthermore, $u_{0x}(x)\,v_{0x}(x) < 0$ is satisfied for any $k \in \mathbf{Z}$ and $x \in (2\,k\,x_0, 2\,(k+1)\,x_0)$.*

For any spatially inhomogeneous positive stationary solution $\boldsymbol{u}(x)$ of (1.1), we denote by $\ell(\boldsymbol{u})$ the integer which is given by applying Lemma 4.3 to $\boldsymbol{u}(x)$.

COROLLARY 4.4. *Let $n \in \mathbf{N}$, and let $\hat{\boldsymbol{u}}_n(x,\mu) = (\hat{u}_n, \hat{v}_n)(x,\mu)$ be a C^1-class function given in Lemma* 3.1. *Then $\ell(\hat{\boldsymbol{u}}_n(.,\mu)) = n$ holds for any $\mu\,(\neq 0)$ in a neighborhood of $\mu = 0$.*

Proof. Since

$$(\hat{u}_n(x,\mu) - \hat{v}_n(x,\mu))/\mu = 2\,\cos(n\,\pi\,x) + O(\mu)$$

as $\mu \to 0$ by virtue of (3.1), we see from Lemmas 4.1 and 4.3 that the desired result holds. □

LEMMA 4.5. *Let $0 \leq \varepsilon_- < \varepsilon_+$, and let $\boldsymbol{u}(.,\varepsilon) = (u,v)(.,\varepsilon)$ be a continuous function on $(\varepsilon_-, \varepsilon_+)$ such that $\boldsymbol{u}(x,\varepsilon)$ is a spatially inhomogeneous positive stationary solution of* (1.1) *for each $\varepsilon \in (\varepsilon_-, \varepsilon_+)$. Then $\ell(\boldsymbol{u}(.,\varepsilon))$ is a constant function on $(\varepsilon_-, \varepsilon_+)$.*

Proof. We define $\ell_0 \in \mathbf{N}$ and $\eta_0 \in (\varepsilon_-, \varepsilon_+)$ by

$$\ell_0 = \ell(\boldsymbol{u}(.,\eta_0)) = \min_{\varepsilon \in (\varepsilon_-, \varepsilon_+)} \ell(\boldsymbol{u}(.,\varepsilon)).$$

By Lemma 4.3 and the continuity of $\boldsymbol{u}(.,\varepsilon)$, we find that for any $\kappa > 0$, there exists $\rho_1 > 0$ such that $u_x(x,\varepsilon)\,v_x(x,\varepsilon) < 0$ for any $\kappa \leq x \leq 1/\ell_0 - \kappa$ and $|\,\varepsilon - \eta_0\,| \leq \rho_1$. This means that $\ell(\boldsymbol{u}(.,\varepsilon)) = \ell_0$ holds for any ε in a neighborhood of $\varepsilon = \eta_0$. Hence we see that there is a maximal extended interval $J = (\hat{\varepsilon}_-, \hat{\varepsilon}_+)$ such that $\eta_0 \in \operatorname{Int} J$ and $\ell(\boldsymbol{u}(.,\varepsilon)) = \ell_0$ for any $\varepsilon \in J$. We assume $\varepsilon_- < \hat{\varepsilon}_-$. By the continuity of $\boldsymbol{u}(.,\varepsilon)$, we obtain $s_1\,\boldsymbol{u}_x(x,\hat{\varepsilon}_-) \succeq_o 0$ on $[0, 1/\ell_0]$ for some $s_1 \in \{\,-1, 1\,\}$ and $u_x(x_6,\hat{\varepsilon}_-)\,v_x(x_6,\hat{\varepsilon}_-) = 0$

for some $x_6 \in (0, 1/\ell_0)$. Since $f_v(\boldsymbol{u}(x, \hat{\varepsilon}_-)) < 0$ and $g_u(\boldsymbol{u}(x, \hat{\varepsilon}_-)) < 0$ are satisfied for any $x \in [0, 1]$, we have

$$u_{xx}(x_6, \hat{\varepsilon}_-) = 0, \quad 0 \le \hat{\varepsilon}_- \, s_1 \, u_{xxx}(x_6, \hat{\varepsilon}_-) = -f_v(\boldsymbol{u}(x_6, \hat{\varepsilon}_-)) \, s_1 \, v_x(x_6, \hat{\varepsilon}_-) \le 0$$

if $u_x(x_6, \hat{\varepsilon}_-) = 0$, and

$$v_{xx}(x_6, \hat{\varepsilon}_-) = 0, \quad 0 \ge \hat{\varepsilon}_- \, s_1 \, v_{xxx}(x_6, \hat{\varepsilon}_-) = -g_u(\boldsymbol{u}(x_6, \hat{\varepsilon}_-)) \, s_1 \, u_x(x_6, \hat{\varepsilon}_-) \ge 0$$

if $v_x(x_6, \hat{\varepsilon}_-) = 0$. These facts imply that $\boldsymbol{u}_x(x_6, \hat{\varepsilon}_-) = 0$ and $\boldsymbol{u}_{xx}(x_6, \hat{\varepsilon}_-) = 0$ hold. By uniqueness, we find that $\boldsymbol{u}(x, \hat{\varepsilon}_-)$ must be a constant function. This is a contradiction. Hence we obtain $\hat{\varepsilon}_- = \varepsilon_-$. Similarly, we can prove $\hat{\varepsilon}_+ = \varepsilon_+$. Therefore we have $\ell(\boldsymbol{u}(., \varepsilon)) = \ell_0$ for any $\varepsilon \in (\varepsilon_-, \varepsilon_+)$. □

5. Nonexistence of the eigenvalue 0. Let $\boldsymbol{u}_0(x) = (u_0, v_0)(x)$ be an arbitrary spatially inhomogeneous positive stationary solution of (1.1) for $\varepsilon > 0$, and let $\mathcal{L}$ be the linearized operator of (1.1) around $\boldsymbol{u} = \boldsymbol{u}_0(x)$. We denote by $\sigma(\mathcal{L})$ the set of spectra of $\mathcal{L}$ relative to X. In this section, we show that $0 \notin \sigma(\mathcal{L})$ holds.

We assume $0 \in \sigma(\mathcal{L})$ in order to derive a contradiction. Let $x_0 = 1/(2\,\ell(\boldsymbol{u}_0))$, and let $\phi(x) = (\phi^u, \phi^v)(x)$ be an arbitrary eigenfunction of $\mathcal{L}$ corresponding to the eigenvalue 0. By Lemma 4.1 and the boundary condition, we can regard $\phi(x)$ as a periodic function with period 2 which satisfies $\boldsymbol{u}(x) = \boldsymbol{u}(-x)$ for any $x \in \mathbf{R}$.

LEMMA 5.1. *Either* $\phi(2\,k\,x_0) \succ_o 0$, $\phi(2\,k\,x_0) = 0$, *or* $\phi(2\,k\,x_0) \prec_o 0$ *is satisfied for any* $k \in \mathbf{Z}$. *Furthermore if* $\phi^u(2\,k\,x_0) = 0$ *is satisfied for some* $k \in \mathbf{Z}$, *then* $\phi(x) = -\phi(4\,k\,x_0 - x)$ *holds for any* $x \in \mathbf{R}$.

Proof. We consider the case where both of $\phi^u(2\,k\,x_0)\,\phi^v(2\,k\,x_0) \ge 0$ and $\phi(2\,k\,x_0) \ne 0$ hold. We see from Lemma 4.1 that

$$\boldsymbol{U}(x)\,(\equiv (U, V)(x)) = \phi(x) + \phi(4\,k\,x_0 - x)$$

is a nontrivial solution of $\mathcal{L}\,\boldsymbol{u} = 0$ with period 2 and satisfies $U(2\,k\,x_0)\,V(2\,k\,x_0) \ge 0$, $\boldsymbol{U}(2\,k\,x_0) \ne 0$, and $\boldsymbol{U}_x(2\,k\,x_0) = 0$. Since $\frac{d}{dx}(U\,V)(2\,k\,x_0) = 0$ and

$$\varepsilon\,\tfrac{d^2}{dx^2}(U\,V)(2\,k\,x_0) = -f_v^0(2\,k\,x_0)\,V(2\,k\,x_0)^2 - g_u^0(2\,k\,x_0)\,U(2\,k\,x_0)^2 > 0$$

are satisfied if $U(2\,k\,x_0)\,V(2\,k\,x_0) = 0$, we find that there exists $x_7 \in (2\,k\,x_0,\, 2\,k\,x_0 + 2]$ such that the following properties hold:

(i) $U(x)\,V(x) > 0$ holds for any $x \in (2\,k\,x_0, x_7)$ and

(ii) $U(x_7)\,V(x_7) = 0$ is satisfied if $x_7 < 2\,k\,x_0 + 2$.

Since $\boldsymbol{u}_{0x}(2\,k\,x_0) = 0$ and $\boldsymbol{u}_{0x}(2\,k\,x_0 + 2) = 0$ are satisfied because of Lemma 4.1, we have

$$\begin{aligned} 0 =& \int_{2\,k\,x_0}^{x_7} (\varepsilon\,U_{xx}(x) + f_u^0(x)\,U(x) + f_v^0(x)\,V(x))\,u_0(x)\,dx \\ =& \varepsilon\,(U_x(x_7)\,u_0(x_7) - U(x_7)\,u_{0x}(x_7)) - \int_{2\,k\,x_0}^{x_7} (U(x) + b\,V(x))\,u_0(x)^2\,dx \ne 0 \end{aligned}$$

if $U(x_7) = 0$ or $x_7 = 2\,k\,x_0 + 2$, and

$$\begin{aligned} 0 =& \int_{2\,k\,x_0}^{x_7} (\varepsilon\,V_{xx}(x) + g_u^0(x)\,U(x) + g_v^0(x)\,V(x))\,v_0(x)\,dx \\ =& \varepsilon\,V_x(x_7)\,v_0(x_7) - \int_{2\,k\,x_0}^{x_7} (b\,U(x) + V(x))\,v_0(x)^2\,dx \ne 0 \end{aligned}$$

if $V(x_7) = 0$. These contradictions imply that the desired result holds. □

By the above lemma, we may assume $\phi^u(0) = 1$ and $\phi^v(0) < 0$ without loss of generality.

COROLLARY 5.2. *$\mathcal{N}(\mathcal{L}) = \mathrm{span}\{\,\phi\,\}$ is satisfied where $\mathcal{N}(\mathcal{L}) = \{\,\boldsymbol{u} \in X \mid \mathcal{L}\,\boldsymbol{u} = 0\,\}$.*

Proof. Let $\boldsymbol{u}(x) = (u, v)(x)$ be another eigenfunction of $\mathcal{L}$ corresponding to the eigenvalue 0. It is obvious that $\boldsymbol{U} = \boldsymbol{u} - u(0)\,\phi \in \mathcal{N}(\mathcal{L})$ holds. Applying the proof of Lemma 5.1 to $\boldsymbol{U}(x)$, we have $\boldsymbol{U}(0) = 0$. By uniqueness, we obtain $\boldsymbol{U}(x) = 0$ for any $x \in [0, 1]$, which implies that the desired result holds. □

LEMMA 5.3. *Let $\boldsymbol{u}(x) = (u, v)(x)$ be an arbitrary nontrivial solution of $\mathcal{L}\,\boldsymbol{u} = 0$ which satisfies $\boldsymbol{u}(x) = \boldsymbol{u}(x + 2)$ for any $x \in \mathbf{R}$. Then $u_x(\tau) \neq 0$ (resp., $v_x(\tau) \neq 0$) holds for any zero τ of $u(x)$ (resp., $v(x)$).*

Proof. We only show the proof for the former case, because the latter can be proved in a similar manner. We assume $u_x(\tau) = 0$ contrary to the conclusion.

We first consider the case $u_{xx}(\tau) = 0$. Then we obtain $v(\tau) = 0$ and

$$\varepsilon\, u_{xxx}(\tau)\, v_x(\tau) = -f_v^0(\tau)\, v_x(\tau)^2 > 0.$$

It follows from $\boldsymbol{u}(\tau) = 0$ that there exists $x_8 \in (\tau, \tau + 2]$ such that $u(x_8)\, v(x_8) = 0$ and $v_x(\tau)\, u(x) > 0$, $v_x(\tau)\, v(x) > 0$ for any $x \in (\tau, x_8)$. By Lemma 4.2, we see that there exists $x_9 \in (\tau, x_8)$ such that

$$v_x(\tau)\, v(x_9) < 0 \ \text{ if } u(x_8) = 0, \qquad v_x(\tau)\, u(x_9) < 0 \ \text{ if } v(x_8) = 0.$$

This contradiction implies that $u_{xx}(\tau) \neq 0$ holds. We see that there exist x_{10}^- and x_{10}^+ $(x_{10}^- < \tau < x_{10}^+)$ such that $u(x_{10}^\pm) = 0$ and $u_{xx}(\tau)\, u(x) > 0$ hold for any $x \in (x_{10}^-, \tau) \cup (\tau, x_{10}^+)$. By Lemma 4.2 and $v(\tau) = -\varepsilon\, u_{xx}(\tau)/f_v^0(\tau)$, it follows that there exist x_{11}^- and x_{11}^+ $(x_{10}^- < x_{11}^- < \tau < x_{11}^+ < x_{10}^+)$ such that $v(x_{11}^\pm) = 0$ and $u_{xx}(\tau)\, v(x) > 0$ for any $x \in (x_{11}^-, x_{11}^+)$. By Lemma 4.2, we have $u_{xx}(\tau)\, u(x) < 0$ for some $x \in (x_{11}^-, x_{11}^+)$. This contradiction implies that the desired result holds. □

We define $\{\,x_k^u\,\}_{k=1}^{\ell_u}$ (resp., $\{\,x_k^v\,\}_{k=1}^{\ell_v}$) by the set of zeros of $\phi^u(x)$ (resp., $\phi^v(x)$) on $(0, 1)$ which satisfy $x_k^u < x_{k+1}^u$ (resp., $x_k^v < x_{k+1}^v$) for each k. By Lemma 5.1, we obtain

$$0 < x_1^u < x_{\ell_u}^u < 1, \quad 0 < x_1^v < x_{\ell_v}^v < 1.$$

Setting $x_0^\nu = 0$ and $x_{\ell_\nu + 1}^\nu = 2 - x_{\ell_\nu}^\nu$ $(\nu = u,\, v)$, we have

$$(-1)^k\, \phi^u(x) > 0 \ \text{ on } (x_k^u, x_{k+1}^u), \quad (-1)^k\, \phi^v(x) < 0 \ \text{ on } (x_k^v, x_{k+1}^v) \tag{5.1}$$

for each k because of Lemma 5.3. By definition, it is obvious that

$$x_k^u \le x_k^v \le x_{k+1}^u \quad \text{or} \quad x_k^v \le x_k^u \le x_{k+1}^v \tag{5.2}$$

holds for $k = 0$. When $x_j^u \le x_j^v \le x_{j+1}^u$ holds, we have $x_{j+1}^u < x_{j+2}^v$ and $x_{j+1}^v < x_{j+2}^u$ by virtue of (5.1) and Lemma 4.2 and then obtain

$$\begin{aligned} &x_{j+1}^u \le x_{j+1}^v < x_{j+2}^u \quad \text{if } x_{j+1}^u \le x_{j+1}^v, \\ &x_{j+1}^v < x_{j+1}^u < x_{j+2}^v \quad \text{if } x_{j+1}^u > x_{j+1}^v, \end{aligned}$$

which implies that (5.2) holds for $k = j + 1$. In a similar manner, we can prove that (5.2) is satisfied for $k = j + 1$ when $x_j^v \le x_j^u \le x_{j+1}^v$ holds. By induction, we find that (5.2) holds for each $0 \le k \le \min(\ell_u, \ell_v)$. When $\ell_u < \ell_v$ is satisfied, we have

$$x_{\ell_u}^u \le x_{\ell_u+1}^v < x_{\ell_u+2}^v \le x_{\ell_v+1}^v \le x_{\ell_u+1}^u,$$

and then obtain $\phi^u(x)\,\phi^v(x) > 0$ for any $x \in (x^v_{\ell_u+1}, x^v_{\ell_u+2})$. This contradicts the fact of Lemma 4.2. In a similar manner, we can derive a contradiction when $\ell_v < \ell_u$ holds. Hence we have $\ell_u = \ell_v$.

We consider the case where there exists $k \in \mathbf{Z}$ such that $\phi^u(2\,k\,x_0) = 0$ holds. By Lemmas 5.1 and 5.3, we have $\phi(2\,k\,x_0) = 0$, $u_{0xx}(2\,k\,x_0) \neq 0$, and $\phi^u_x(2\,k\,x_0) \neq 0$. Setting

$$\boldsymbol{U}(x)\,(\equiv (U,V)(x)) = \boldsymbol{u}_{0x}(x) - \frac{u_{0xx}(2\,k\,x_0)}{\phi^u_x(2\,k\,x_0)}\,\phi(x),$$

we see that $\boldsymbol{U}(x)$ is a solution of $\mathcal{L}\,\boldsymbol{u} = 0$ and satisfies $\boldsymbol{U}(2\,k\,x_0) = 0$ and $U_x(2\,k\,x_0) = 0$. By Lemma 5.3, we obtain $\boldsymbol{U}(x) = 0$ for any $x \in \mathbf{R}$, which indicates $\phi(0) = 0$. This contradicts the fact $\phi^u(0) = 1$. Hence we have $\phi^u(2\,k\,x_0) \neq 0$ for any $k \in \mathbf{Z}$.

We consider the case where there exist j and k such that $2\,j\,x_0 < x^u_k < x^u_{k+1} < 2\,(j+1)\,x_0$. From Lemmas 4.2 and 4.3, we have $s_2\,\boldsymbol{u}_{0x}(x) \succ_o 0$ on $(2\,j\,x_0, 2\,(j+1)\,x_0)$ for some $s_2 \in \{-1, 1\}$, and $(-1)^k\,\phi^v(x_{12}) < 0$ for some $x_{12} \in (x^u_k, x^u_{k+1})$. Hence we see that there exists $0 \leq k' \leq \ell_v$ such that $x_{12} \in (x^v_{k'}, x^v_{k'+1})$ is satisfied. If $x^v_{k'} \leq 2\,j\,x_0$ holds, then we obtain $(-1)^k\,\phi^u(2\,j\,x_0) \geq 0$ by virtue of Lemma 5.1. By (5.1), we have $2\,j\,x_0 \leq x^u_{k-1}$ and $\phi^u(x)\,\phi^v(x) > 0$ for any $x \in (x^u_{k-1}, x^u_k)$. This contradicts the fact of Lemma 4.2. Therefore we obtain $2\,j\,x_0 < x^v_{k'}$. In a similar manner, we can prove $x^v_{k'+1} < 2\,(j+1)\,x_0$. Setting $x^-_{13} = \min(x^u_k, x^v_{k'})$ and $x^+_{13} = \max(x^u_{k+1}, x^v_{k'+1})$, we obtain

$$\begin{aligned}
&(-1)^k\,\phi^u(x) \leq 0 \quad \text{on } [x^-_{13}, x^u_k) \cup (x^u_{k+1}, x^+_{13}],\\
&(-1)^k\,\phi^v(x) \geq 0 \quad \text{on } [x^-_{13}, x^v_{k'}) \cup (x^v_{k'+1}, x^+_{13}]
\end{aligned}$$

because of Lemma 4.2. It follows that there exists $C_1 > 0$ such that

$$\boldsymbol{U}(x)\,(\equiv (U,V)(x)) = s_2\,\boldsymbol{u}_{0x}(x) - C_1\,(-1)^k\,\phi(x)$$

satisfies the following: (i) $\boldsymbol{U}(x)$ is a solution of $\mathcal{L}\,\boldsymbol{u} = 0$, (ii) $\boldsymbol{U}(x) \succeq_o 0$ holds for any $x \in [x^-_{13}, x^+_{13}]$, and (iii) there exists $x_{14} \in (x^-_{13}, x^+_{13})$ such that $U(x_{14})\,V(x_{14}) = 0$. Since $f^0_u(x) < 0$ and $g^0_u(x) < 0$ are satisfied for any $x \in \mathbf{R}$, we have

$$\begin{aligned}
&U_x(x_{14}) = 0, \quad 0 \leq \varepsilon\,U_{xx}(x_{14}) = -f^0_v(x_{14})\,V(x_{14}) \leq 0 \quad \text{if } U(x_{14}) = 0,\\
&V_x(x_{14}) = 0, \quad 0 \geq \varepsilon\,V_{xx}(x_{14}) = -g^0_u(x_{14})\,U(x_{14}) \geq 0 \quad \text{if } V(x_{14}) = 0.
\end{aligned}$$

These facts imply that $\boldsymbol{U}(x_{14}) = 0$ and $\boldsymbol{U}_x(x_{14}) = 0$ hold. By uniqueness, we obtain $\boldsymbol{U}(x) = 0$ for any $x \in \mathbf{R}$. This contradicts $U(0) = (-1)^{k+1}\,C_1 \neq 0$. Therefore we have

$$\#[\{\,x^u_k\,\}_{n=1}^{\ell_u} \cap (2\,j\,x_0, 2\,(j+1)\,x_0)] \leq 1$$

for any $0 \leq j < \ell(\boldsymbol{u}_0)$, where $\#A$ is the number of elements of the set A. In a similar manner, we can prove

$$\#[\{\,x^u_k\,\}_{n=1}^{\ell_u} \cap (2\,j\,x_0, 2\,(j+1)\,x_0)] \geq 1$$

for any $0 \leq j < \ell(\boldsymbol{u}_0)$. Hence we have

$$2\,j\,x_0 < x^u_{j+1} < 2\,(j+1)\,x_0 \tag{5.3}$$

for any $0 \le j < \ell(\boldsymbol{u}_0)$.

We see from Lemma 4.1 that $\boldsymbol{U}(x) = \phi(x + 4\,x_0) + \phi(4\,x_0 - x)$ is a solution of $\mathcal{L}\,\boldsymbol{u} = 0$ with $\boldsymbol{u}_x(0) = 0$. Since $\phi(x)$ is a periodic function with period 2, we obtain

$$\boldsymbol{U}(2 - x) = \phi(4\,x_0 - x + 2) + \phi(x + 4\,x_0 - 2) = \boldsymbol{U}(x)$$

for any $x \in \mathbf{R}$. Since $\boldsymbol{U} \in \mathcal{N}(\mathcal{L})$ holds because of $\boldsymbol{U}_x(1) = 0$, we find from Corollary 5.2 that there exists $C_2 \in \mathbf{R}$ such that $\boldsymbol{U}(x) = C_2\,\phi(x)$ for any $x \in \mathbf{R}$. Since $\phi^u(4\,k\,x_0) > 0$ and

$$2 \min_{0 \le j \le \ell(\boldsymbol{u}_0)} \phi^u(4\,j\,x_0) \le C_2\,\phi^u(4\,k\,x_0) \le 2 \max_{0 \le j \le \ell(\boldsymbol{u}_0)} \phi^u(4\,j\,x_0)$$

holds for any $0 \le k \le \ell(\boldsymbol{u}_0)$, we obtain $C_2 = 2$. By $\phi(x) = \phi(-x)$ for any $x \in \mathbf{R}$, we have

$$\begin{aligned}
\phi(4\,(k+2)\,x_0) - 2\,\phi(4\,(k+1)\,x_0) + \phi(4\,k\,x_0) = 0,\\
\phi_x(4\,(k+2)\,x_0) - 2\,\phi_x(4\,(k+1)\,x_0) + \phi_x(4\,k\,x_0) = 0
\end{aligned}$$

for any $k \in \mathbf{Z}$ and then obtain

$$\phi(4\,k\,x_0) = \phi(0) + k\,(\phi(4\,x_0) - \phi(0)), \quad \phi_x(4\,k\,x_0) = k\,\phi_x(4\,x_0)$$

for any $k \in \mathbf{Z}$. Since $\phi(x)$ is a periodic function, we have $\phi(4\,k\,x_0) = \phi(0)$ and $\phi_x(4\,k\,x_0) = 0$ for any $k \in \mathbf{Z}$. By uniqueness and Lemma 4.1, we obtain $\phi(x) = \phi(4\,x_0 - x)$ for any $x \in \mathbf{R}$. Hence we see that $\phi(x)$ is a nontrivial solution of

$$\begin{cases}
\mathcal{L}\,\boldsymbol{u} = 0, \quad x \in (0, 2\,x_0),\\
\boldsymbol{u}_x(0) = 0, \quad \boldsymbol{u}_x(2\,x_0) = 0.
\end{cases}$$

By Lemma 4.1 and $\boldsymbol{f}_{\boldsymbol{u}}(T\,\boldsymbol{u})T = T\,\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{u})$ for any $\boldsymbol{u}$, we have $T\,\phi(2\,x_0 - .) \in \mathcal{N}(\mathcal{L})$. Hence it follows from Corollary 5.2 that there exists $C_3 \in \mathbf{R}$ such that $\phi(x) = C_3\,T\,\phi(2\,x_0 - x)$ for any $x \in \mathbf{R}$, which implies $\phi(x) = C_3^2\,\phi(x)$ for any $x \in \mathbf{R}$. Since $\phi(x)$ is nontrivial, we have either $C_3 = 1$ or $C_3 = -1$. Since $0 > \phi^u(2\,x_0) = -\phi^v(0) > 0$ holds if $C_3 = -1$, we obtain $C_3 = 1$, which implies $\phi^u(x) = \phi^v(2\,x_0 - x)$ for any $x \in \mathbf{R}$.

We may assume $H(x) > 0$ for any $x \in [0, x_0)$. (If not, we use the below argument by replacing $[0, 2\,x_0]$ and $\phi(x)$ with $[2\,x_0, 4\,x_0]$ and $-\phi(x)$, respectively.) By Lemma 4.1, we have $H(x) < 0$ for any $x \in (x_0, 2\,x_0]$. Setting $x_{15}^- = \min(x_1^u, x_1^v)$, $x_{15}^+ = \max(x_1^u, x_1^v)$, $\Phi(x) = \phi^u(x) + \phi^v(x)$, and $\Psi(x) = \phi^u(x) - \phi^v(x)$, we can easily check that $x_{15}^+ = 2\,x_0 - x_{15}^-$ holds and that $(\Phi, \Psi)(x)$ is a nontrivial solution of

$$\begin{cases}
\varepsilon\,\Phi_{xx} = a_{11}(x)\,\Phi + a_{12}(x)\,\Psi,\\
\varepsilon\,\Psi_{xx} = a_{21}(x)\,\Phi + a_{22}(x)\,\Psi, \quad x \in (0, 2\,x_0),\\
\Phi_x = 0, \quad \Psi_x = 0, \quad x = 0, 2\,x_0,
\end{cases}$$

where

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}(x) = \begin{pmatrix} (b+1)\,K - 1 & (1-b)\,H \\ H & K - 1 \end{pmatrix}(x).$$

Since $\phi^u(x) = \phi^v(2\,x_0 - x)$ holds for any $x \in \mathbf{R}$, we have $\Phi(x) = \Phi(2\,x_0 - x)$ and $\Psi(x) = -\Psi(2\,x_0 - x)$ for any $x \in \mathbf{R}$. By $\phi^v(0) < 0 < \phi^u(0)$, we obtain $\Psi(x) > 0$ for

any $x \in [0, x_{15}^-)$ and $\Psi(x) < 0$ for any $x \in (x_{15}^+, 2\,x_0]$. Furthermore if $\Phi(x_0) \neq 0$ holds, then we have $\Phi(x_0)\,\Phi(x) > 0$ for any $x \in [x_{15}^-, x_{15}^+]$. Since $K(x)$ satisfies $0 < K(x) < 1$ for any $x \in \mathbf{R}$ and

$$0 = \varepsilon\,K_{xx} + \left(1 - \frac{b+1}{2}\,K\right)K + \frac{b-1}{2}\,H^2, \quad x \in \mathbf{R},$$

we obtain $2/(b+1) \leq K(x) < 1$ for any $x \in \mathbf{R}$ by virtue of the comparison principle. Hence we have $a_{11}(x) > 0$ and $a_{22}(x) < 0$ for any $x \in \mathbf{R}$.

First of all, we consider the case $\Phi(x_0) > 0$. Suppose that $x_{16} < 2\,x_0$ holds, where

$$x_{16} = \sup\{\, y \in [x_0, 2\,x_0] \,|\, \Phi(x) \geq 0 \text{ for any } x \in [x_0, y]\,\}\,(> x_{15}^+).$$

Setting

$$x_{17} = \sup\{\, y \in [x_{16}, 2\,x_0] \,|\, \Phi(x) \leq 0 \text{ for any } x \in [x_{16}, y]\,\}\,(\leq 2\,x_0),$$

we have $\Phi_x(x_{17}) \geq 0$ because of $\Phi(2\,x_0) = 0$. Since $\Phi(x)$ satisfies $\Phi(x_{16}) = 0$, $\Phi_x(x_{16}) \leq 0$, and

$$\varepsilon\,\Phi_{xx}(x) = a_{11}(x)\,\Phi(x) + a_{12}(x)\,\Psi(x) < 0, \quad x \in [x_{16}, x_{17}],$$

we obtain $x_{17} > x_{16}$ and $\Phi_x(x_{17}) < 0$. This contradiction implies that $\Phi(x) \geq 0\,(\not\equiv 0)$ holds for any $x \in [0, 2\,x_0]$. Since $H(x)$ satisfies

$$\begin{cases} \varepsilon\,H_{xx} - a_{22}(x)\,H = 0, \quad x \in (0, 2\,x_0), \\ H_x(0) = 0, \quad H_x(2\,x_0) = 0, \end{cases}$$

we have

$$0 < \int_0^{2\,x_0} H(x)^2\,\Phi(x)\,dx = \int_0^{2\,x_0} (\varepsilon\,\Psi_{xx}(x) - a_{22}(x)\,\Psi(x))\,H(x)\,dx = 0$$

because of integration by parts. This is a contradiction.

We next consider the case $\Phi(x_0) < 0$. It is obvious that $\Phi(x) < 0$ holds for any $x \in [x_{15}^-, x_{15}^+]$. If there exist x_{18}^- and x_{18}^+ $(x_{15}^- \leq x_{18}^- < x_{18}^+ \leq x_0)$ such that $\Psi(x_{18}^\pm) = 0$ and $\Psi(x) \leq 0$ for any $x \in [x_{18}^-, x_{18}^+]$, then we have

$$\begin{aligned} 0 > \int_{x_{18}^-}^{x_{18}^+} H(x)^2\,\Phi(x)\,dx &= \int_{x_{18}^-}^{x_{18}^+} (\varepsilon\,\Psi_{xx}(x) - a_{22}(x)\,\Psi(x))\,H(x)\,dx \\ &= \varepsilon\,(\Psi_x(x_{18}^+)\,H(x_{18}^+) - \Psi_x(x_{18}^-)\,H(x_{18}^-)) \geq 0 \end{aligned}$$

because of $H(x_{18}^\pm) \geq 0$. This fact and $\Psi(x_0) = 0$ show that $\Psi(x) \geq 0$ holds for any $x \in [0, x_0]$. Since

$$\varepsilon\,\Phi_{xx}(x) = a_{11}(x)\,\Phi(x) + a_{12}(x)\,\Psi(x) < 0$$

holds if both of $\Phi(x) < 0$ and $0 \leq x < x_0$ are satisfied, we obtain $\Phi(x) < 0$ and $\Phi_x(x) > 0$ for any $0 \leq x < x_0$ because of $\Phi_x(x_0) = 0$. This contradicts $\Phi_x(0) = 0$.

Finally, we consider the case $\Phi(x_0) = 0$. We have $x_{15}^- = x_{15}^+ = x_0$ because of $\Phi(x_0) = 2\,\phi^u(x_0)$. Since $(\Phi, \Psi)(x)$ is nontrivial, we obtain $\Psi_x(x_0) < 0$ and

$$\varepsilon\,\Phi_{xxxx}(x_0) = 2\,(1-b)\,H_x(x_0)\,\Psi_x(x_0) < 0.$$

We can derive a contradiction in a similar manner with the proof for the case where $\Phi(x_0) < 0$.

The above contradictions show that the following lemma holds.

LEMMA 5.4. *$0 \notin \sigma(\mathcal{L})$ holds.*

6. Proof of Theorem 2.1. Let $n \in \mathbf{N}$ be arbitrarily fixed. It follows from Lemmas 3.1, 4.5, 5.4, and Corollary 4.4 that there exists a maximal extended interval $J_n = (\eta_n, \varepsilon_n)$ such that the following property is satisfied:

(i) There exists a pair of C^1-class functions $\boldsymbol{u}_n^{\pm}(.,\varepsilon) = (u_n^{\pm}, v_n^{\pm})(.,\varepsilon)$ defined on J_n such that fact (i) in Theorem 2.1 holds for any $\varepsilon \in J_n$.

(ii) $\boldsymbol{u}_n^{\pm}(.,\varepsilon_n(\mu)) = \hat{\boldsymbol{u}}_n(.,\pm\mu)$ hold for any $0 \le \mu < \mu_n$.

By (1.1), we can easily obtain

$$\int_0^1 f(\boldsymbol{u}_n^{\pm}(x,\varepsilon))\,dx = 0, \quad \int_0^1 g(\boldsymbol{u}_n^{\pm}(x,\varepsilon))\,dx = 0$$

for each $\varepsilon \in J_n$. Since $\boldsymbol{u}_n^{\pm}(x,\varepsilon) \in [0,1] \times [0,1]$ hold for any x and ε because of the comparison principle, we see from the Ascoli–Arzela theorem that there exist stationary solutions $\boldsymbol{w}_0^-(x)$ and $\boldsymbol{w}_0^+(x)$ of (1.1) such that $\boldsymbol{u}_n^{\pm}(.,\varepsilon) \to \boldsymbol{w}_0^{\pm}$ as $\varepsilon \to \eta_n$ in X. Furthermore, we also find that $\boldsymbol{w}_0^-(x) \succeq_s 0$ and $\boldsymbol{w}_0^+(x) \succeq_s 0$ are satisfied for any $x \in [0,1]$.

We consider the case where there exists $x_{19} \in [0,1]$ such that $w^u(x_{19}) = 0$ holds, where $\boldsymbol{w}_0^+(x) = (w^u, w^v)(x)$. By $w_x^u(x_{19}) = 0$, we have $w^u(x) = 0$ for any $x \in [0,1]$ because of uniqueness. Hence we see that $w^v(x)$ is a nonnegative solution of

$$\begin{cases} 0 = \eta_n\, v_{xx} + v\,(1-v), & x \in (0,1), \\ v_x(0) = 0, \quad v_x(1) = 0, \end{cases}$$

which implies either (a) $w^v(x) \equiv 0$ on $[0,1]$ or (b) $w^v(x) \equiv 1$ on $[0,1]$. Since

$$\frac{f(\boldsymbol{u}_n^+(.,\varepsilon))}{u_n^+(.,\varepsilon)} \to \begin{cases} 1 & \text{for the case (a)}, \\ 1-b\,(<0) & \text{for the case (b)} \end{cases}$$

as $\varepsilon \to \eta_n$ holds in X, we have

$$\int_0^1 f(\boldsymbol{u}_n^+(x,\varepsilon))\,dx \begin{cases} > 0 & \text{for the case (a)}, \\ < 0 & \text{for the case (b)} \end{cases}$$

in a neighborhood of $\varepsilon = \eta_n$. This contradiction implies that $w^u(x) > 0$ holds for any $x \in [0,1]$. In a similar manner, we can prove $w^v(x) > 0$ for any $x \in [0,1]$. Hence we see that $\boldsymbol{w}_0^+(x)$ is positive. Analogously we can show that $\boldsymbol{w}_0^-(x)$ is also positive.

We consider the case where $\eta_n > 0$. By Corollary 4.4 and Lemma 4.5, we have $\ell(\boldsymbol{u}_n^{\pm}(.,\varepsilon)) = n$ for any $\varepsilon \in J_n$. By Lemma 5.4 and the implicit function theorem, it follows that $\boldsymbol{w}_0^-(x)$ and/or $\boldsymbol{w}_0^+(x)$ must be spatially homogeneous. Lemma 3.1 leads to the fact that $\eta_n = \varepsilon_m$ holds for some $m > n$. By Corollary 4.4, we see that $\ell(\boldsymbol{u}_n^-(.,\varepsilon)) = m$ or $\ell(\boldsymbol{u}_n^+(.,\varepsilon)) = m$ holds in a neighborhood of $\varepsilon = \eta_n$. This is a contradiction. Hence we have $\eta_n = 0$.

Let $\eta > 0$, and let $\boldsymbol{u}_0(x)$ be an arbitrary spatial inhomogeneous positive stationary solution of (1.1) for $\varepsilon = \eta$. We find from Lemma 5.4 and the implicit function theorem that there exists a maximal extended interval $J = (\varepsilon_-, \varepsilon_+)$ such that the following properties hold:

(i) $\eta \in \operatorname{Int} J$ is satisfied.

(ii) There exists a C^1-class function $\boldsymbol{w}(.,\varepsilon)$ defined on J such that $\boldsymbol{w}(x,\varepsilon)$ is a spatially inhomogeneous positive stationary solution of (1.1) for each $\varepsilon \in J$ and satisfies $\boldsymbol{w}(.,\eta) = \boldsymbol{u}_0$.

By Lemma 4.5, we have $\ell(\boldsymbol{w}(.,\varepsilon)) = \ell(\boldsymbol{u}_0)$ for any $\varepsilon \in J$, which implies $\varepsilon_+ \leq 1/(\pi^2\,\ell(\boldsymbol{u}_0)^2)$. Lemma 5.4, the implicit function theorem, and the above argument show that the limit $\lim_{\varepsilon\to\varepsilon_+} \boldsymbol{w}(x,\varepsilon)$ must be spatially homogeneous and positive. By Lemma 3.1 and Corollary 4.4, we have $\varepsilon_+ = \varepsilon_{\ell(\boldsymbol{u}_0)}$ and then obtain $\boldsymbol{u}_0 = \boldsymbol{u}^{\nu}_{\ell(\boldsymbol{u}_0)}(.,\eta)$ for some $\nu \in \{+,-\}$. Thus the proof is completed. □

REFERENCES

[1] N. Chafee and E. F. Infante, *A bifurcation problem for a nonlinear partial differential equation of parabolic type*, Applicable Anal., 4 (1974/75), pp. 17–37.

[2] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[3] E. N. Dancer, *On the existence and uniqueness of positive solutions for competing species model with diffusion*, Trans. Amer. Math. Soc., 326 (1991), pp. 829–859.

[4] P. de Mottoni, *Qualitative analysis for some quasi-linear parabolic systems*, Inst. Math. Pol. Acad. Sci. Zam, 190 (1979).

[5] C. Gui and Y. Lou, *Uniqueness and nonuniqueness of coexistence states in the Lotka-Volterra competition model*, Comm. Pure Appl. Math., 47 (1994), pp. 1571–1594.

[6] J. K. Hale, *Asymptotic Behavior of Dissipative Systems*, American Mathematical Society, Providence, RI, 1988.

[7] Y. Kan-on and E. Yanagida, *Existence of non-constant stable equilibria in competition-diffusion equations*, Hiroshima Math. J., 23 (1993), pp. 193–221.

[8] K. Kishimoto and H. F. Weinberger, *The spatial homogeneity of stable equilibria of some reaction-diffusion systems on convex domains*, J. Differential Equations, 58 (1985), pp. 15–21.

[9] H. Matano and M. Mimura, *Pattern formation in competition-diffusion systems in nonconvex domains*, Publ. Res. Inst. Math. Sci., 19 (1983), pp. 1049–1079.

[10] M. Mimura, S.-I. Ei, and Q. Fang, *Effect of domain shape on coexistence problems in a competition-diffusion system*, J. Math. Biol., 29 (1991), pp. 219–237.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 437–451, March 1998

008

ON THE PROPERTIES OF SOME NONLINEAR EIGENVALUES*

ANGELO ALVINO†, VINCENZO FERONE†, AND GUIDO TROMBETTI†

Abstract. We consider the eigenvalue problem for a nonlinear equation involving the p-Laplacian with homogeneous Dirichlet conditions. We give some properties of the first eigenvalue and of the corresponding eigenfunction. In particular we prove inequalities similar to the well-known Payne–Rayner inequality, which holds in the linear case.

Key words. nonlinear eigenvalues, Payne–Rayner inequality, symmetrization

AMS subject classifications. 35J65, 35B45, 49R05

PII. S0036141096302111

1. Introduction.

Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 2$. It is well known that the first eigenvalue of the problem,

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

is positive and simple; that is, all the corresponding eigenfunctions are scalar multiples of each other. Furthermore, if λ is the first eigenvalue of (1.1) then

$$\lambda = \min_{\substack{u \in W_0^{1,2}(\Omega) \\ u \not\equiv 0}} \frac{\displaystyle\int_\Omega |\nabla u|^2 \, dx}{\displaystyle\int_\Omega |u|^2 \, dx}. \tag{1.2}$$

Various comparison results have been given for λ and the associated eigenfunction. For example, when $n = 2$, in [PR1] the following inequality has been proven:

$$\int_\Omega u^2 \, dx \leq \frac{\lambda}{4\pi} \left(\int_\Omega |u| \, dx \right)^2,$$

where u is any eigenfunction of (1.1) corresponding to the first eigenvalue. Such a result has been extended in various directions. For example in [KJ] (see also [PR2]) a similar inequality has been proven for any dimension n, and in [C1], [C2], it has been shown that for any $0 < q < r < \infty$ one has

$$\|u\|_r \leq K(r, q, n, \lambda) \, \|u\|_q, \tag{1.3}$$

where K is a suitable constant. We observe that the inequalities we are referring to are isoperimetric; that is, they hold as equalities if and only if Ω is a ball.

*Received by the editors April 12, 1996; accepted for publication (in revised form) December 19, 1996. This work was partially supported by Italian MURST (40%).
http://www.siam.org/journals/sima/29-2/30211.html
†Dipartimento di Matematica e Applicazioni "R. Caccioppoli," Università di Napoli "Federico II", Complesso Monte S. Angelo, Via Cintia, 80126 Napoli, Italy (alvino@matna1.dma.unina.it, ferone@matna1.dma.unina.it, trombetti@matna1.dma.unina.it).

Our aim is to show that all the properties we have recalled about the first eigenvalue of (1.1) can be proven also for the first eigenvalue of the following nonlinear problem:

$$\begin{cases} -\Delta_p u = \lambda u|u|^{p-2} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.4}$$

where Δ_p denotes the p-Laplacian, that is, $\Delta_p u = \operatorname{div}\big(|\nabla u|^{p-2}\nabla u\big)$. The properties of the first eigenvalue and of the corresponding eigenfunctions of problem (1.4) have been studied by various authors (see, e.g., [Th], [A], [Sa], [B1], [B2], [L1], [L2]). For example, they prove that the first eigenvalue of (1.4) is positive, simple, and isolated.

We will consider the more general problem

$$\begin{cases} -\operatorname{div}\big((A\nabla u, \nabla u)^{(p-2)/2} A\nabla u\big) = \lambda u|u|^{p-2} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.5}$$

where $p > 1$ and $A = \{a_{ij}(x)\}$ is a symmetric matrix with measurable bounded coefficients satisfying an ellipticity condition. After having proven that the simplicity of the first eigenvalue of (1.5) is still true, by adapting the arguments of [L1], we give a Faber–Krahn inequality for the first eigenvalue λ_p of (1.5) (see [C1] for the case $p = 2$). In particular we will show that the first eigenvalue of (1.4) in a ball with the same measure as Ω is smaller or equal to λ_p and, furthermore, that equality holds if and only if Ω is a ball and $a_{ij}x_j = x_i$ for a.e. $x \in \mathbb{R}^n$, modulo translations.

It is well known that (see, e.g., [G]) any solution of (1.5) is bounded. In the last section of the paper we prove the following "reverse" inequality:

$$\|u\|_r \le \beta(r, q, p, n, \lambda_p)\, \|u\|_q, \tag{1.6}$$

where $0 < q < r \le \infty$. We remark that, as (1.3), the above inequality is isoperimetric. The proof of (1.6) is based on a comparison result between u and a suitable eigenfunction v of problem (1.4) defined in a ball B such that the corresponding first eigenvalue is equal to λ_p. This result is a consequence of a comparison result between solution and subsolutions of a one-dimensional problem of the following type:

$$\begin{cases} -\big(|\varphi'(s)|^{\gamma-2}\varphi'(s)\big)' = \lambda m(s)|\varphi(s)|^{\gamma-2}\varphi(s) & \text{in } (0, a), \\ \varphi(0) = \varphi(a) = 0, \end{cases}$$

where the weight $m(s)$ is a continuous function in the interval $(0, a)$ satisfying suitable assumptions. Obviously the comparison result between u and v is isoperimetric.

We finally recall that some estimates similar to (1.6) can be found also in [M].

2. General results about eigenvalues. Let Ω be a bounded domain in $\mathbb{R}^n$, $n \ge 2$. Let us consider the first eigenvalue λ_p of the problem

$$\begin{cases} -\operatorname{div}\big((A\nabla u, \nabla u)^{(p-2)/2} A\nabla u\big) = \lambda u|u|^{p-2} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where $p > 1$ and $A = \{a_{ij}(x)\}$ is a symmetric matrix with measurable bounded coefficients satisfying the ellipticity condition

$$a_{ij}(x)\xi_j\xi_i \ge |\xi|^2, \quad \text{a.e. } x \in \Omega\, \forall \xi \in \mathbb{R}^n.$$

It is well known that λ_p is the minimum of the Rayleigh quotient

$$\lambda_p = \min_{\substack{u\in W_0^{1,p}(\Omega)\\ u\not\equiv 0}} \frac{\displaystyle\int_\Omega (A\nabla u,\nabla u)^{p/2}\,dx}{\displaystyle\int_\Omega |u|^p\,dx} \tag{2.2}$$

and that the above minimization problem is equivalent to the weak form of (2.1) with $\lambda=\lambda_p$, that is,

$$\int_\Omega \big(A\nabla u,\nabla u\big)^{(p-2)/2}\big(A\nabla u,\nabla\varphi\big)dx = \lambda_p\int_\Omega u|u|^{p-2}\varphi\,dx \quad \forall\varphi\in W_0^{1,p}(\Omega). \tag{2.3}$$

Furthermore, the existence of a first eigenfunction and of the first eigenvalue can be proven by standard tools.

Using the methods in [L1] it is also possible to prove the following property for the eigenfunctions of (2.1).

THEOREM 2.1. *The first eigenvalue of* (2.1) *is simple; that is, all the associated eigenfunctions u are scalar multiples of each other.*

The proof of the above theorem is based on the following lemma, whose proof will be given in the Appendix (see also [L1]).

LEMMA 2.2. *Let $A=\{a_{ij}\}$ be a symmetric matrix such that $a_{ij}\xi_j\xi_i \ge |\xi|^2$ $\forall\xi\in\mathbb{R}^n$. Then we have:*

(i) *if $p\ge 2$:*

$$\big(A\xi_2,\xi_2\big)^{p/2}\ge \big(A\xi_1,\xi_1\big)^{p/2}+p\big(A\xi_1,\xi_1\big)^{\frac{p-2}{2}}\big(A\xi_1,\xi_2-\xi_1\big)+\frac{\big(A(\xi_2-\xi_1),\xi_2-\xi_1\big)^{p/2}}{2^{p-1}-1}$$

for every $\xi_1,\xi_2\in\mathbb{R}^n$;

(ii) *if $1<p<2$:*

$$\big(A\xi_2,\xi_2\big)^{p/2}\ge \big(A\xi_1,\xi_1\big)^{p/2}+p\big(A\xi_1,\xi_1\big)^{\frac{p-2}{2}}\big(A\xi_1,\xi_2-\xi_1\big) + c(p)\frac{\big(A(\xi_2-\xi_1),\xi_2-\xi_1\big)}{\big(\sqrt{(A\xi_1,\xi_1)}+\sqrt{(A\xi_2,\xi_2)}\big)^{2-p}}$$

for every $\xi_1,\xi_2\in\mathbb{R}^n$, where $c(p)$ is a constant which depends only on p.

Proof of Theorem 2.1. First of all we observe that any solution of (2.1) is Hölder continuous (see, e.g., [G]). Furthermore, if u minimizes (2.2), then $|u|$ does also; that is, $|u|$ is also an eigenfunction. By Harnack's inequality (see, e.g., [Se], [Tr]) it follows that either $|u|>0$ or $|u|\equiv 0$ in Ω. This means that if u is an eigenfunction then $u>0$ or $u<0$ in Ω.

Suppose now that u and v are two positive eigenfunctions relative to the first eigenvalue of (2.1). Clearly both u and v satisfy (2.3). We will prove that u and v are proportional. Following the arguments of [L1] we put $u_\varepsilon=u+\varepsilon$ and $v_\varepsilon=v+\varepsilon$ for $\varepsilon>0$. Then we use the test functions $\varphi=\big(u_\varepsilon^p-v_\varepsilon^p\big)/u_\varepsilon^{p-1}$ in the equation satisfied by u and $\varphi=\big(v_\varepsilon^p-u_\varepsilon^p\big)/v_\varepsilon^{p-1}$ in the equation satisfied by v. Adding the obtained equalities we have

$$\lambda_p\int_\Omega\left[\frac{u^{p-1}}{u_\varepsilon^{p-1}}-\frac{v^{p-1}}{v_\varepsilon^{p-1}}\right](u_\varepsilon^p-v_\varepsilon^p)dx$$

$$= \int_\Omega \big(A\nabla u_\varepsilon, \nabla u_\varepsilon\big)^{p/2} \left[1 + (p-1)\frac{v_\varepsilon^p}{u_\varepsilon^p}\right] dx + \int_\Omega \big(A\nabla v_\varepsilon, \nabla v_\varepsilon\big)^{p/2} \left[1 + (p-1)\frac{u_\varepsilon^p}{v_\varepsilon^p}\right] dx$$
$$- \int_\Omega p\big(A\nabla u_\varepsilon, \nabla v_\varepsilon\big) \left[\big(A\nabla u_\varepsilon, \nabla u_\varepsilon\big)^{(p-2)/2} \frac{v_\varepsilon^{p-1}}{u_\varepsilon^{p-1}} + \big(A\nabla v_\varepsilon, \nabla v_\varepsilon\big)^{(p-2)/2} \frac{u_\varepsilon^{p-1}}{v_\varepsilon^{p-1}}\right] dx.$$

Observing that $\big(A\nabla u_\varepsilon, \nabla u_\varepsilon\big)/u_\varepsilon^2 = \big(A\nabla(\log u_\varepsilon), \nabla(\log u_\varepsilon)\big)$, from Lemma 2.2 it follows that

$$\lambda_p \int_\Omega \left[\frac{u^{p-1}}{u_\varepsilon^{p-1}} - \frac{v^{p-1}}{v_\varepsilon^{p-1}}\right] (u_\varepsilon^p - v_\varepsilon^p) dx \geq 0. \tag{2.4}$$

On the other hand we also have

$$\lim_{\varepsilon \to 0^+} \int_\Omega \left[\frac{u^{p-1}}{u_\varepsilon^{p-1}} - \frac{v^{p-1}}{v_\varepsilon^{p-1}}\right] (u_\varepsilon^p - v_\varepsilon^p) dx = 0. \tag{2.5}$$

Let us now consider $p \geq 2$. Lemma 2.2 and the ellipticity condition allow us to write also

$$\lambda_p \int_\Omega \left[\frac{u^{p-1}}{u_\varepsilon^{p-1}} - \frac{v^{p-1}}{v_\varepsilon^{p-1}}\right] (u_\varepsilon^p - v_\varepsilon^p) dx \tag{2.6}$$
$$\geq \frac{1}{2^{p-1}-1} \int_\Omega (u_\varepsilon^p + v_\varepsilon^p) \Big(A\big(\nabla(\log u_\varepsilon - \log v_\varepsilon)\big), \nabla(\log u_\varepsilon - \log v_\varepsilon)\Big)^{p/2} dx$$
$$= \frac{1}{2^{p-1}-1} \int_\Omega \left(\frac{1}{u_\varepsilon^p} + \frac{1}{v_\varepsilon^p}\right) \Big(A\big(v_\varepsilon \nabla u_\varepsilon - u_\varepsilon \nabla v_\varepsilon\big), v_\varepsilon \nabla u_\varepsilon - u_\varepsilon \nabla v_\varepsilon\Big)^{p/2} dx$$
$$\geq \frac{1}{2^{p-1}-1} \int_\Omega \left(\frac{1}{u_\varepsilon^p} + \frac{1}{v_\varepsilon^p}\right) |v_\varepsilon \nabla u_\varepsilon - u_\varepsilon \nabla v_\varepsilon|^p dx \geq 0.$$

Passing to the limit in (2.6) and using (2.5), one obtains that u is proportional to v in Ω.

In the case $1 < p < 2$ one can use similar arguments. Instead of (2.6), by Lemma 2.2 one gets

$$\lambda_p \int_\Omega \left[\frac{u^{p-1}}{u_\varepsilon^{p-1}} - \frac{v^{p-1}}{v_\varepsilon^{p-1}}\right] (u_\varepsilon^p - v_\varepsilon^p) dx$$
$$\geq c(p) \int_\Omega \left(\frac{1}{u_\varepsilon^p} + \frac{1}{v_\varepsilon^p}\right) \frac{|v_\varepsilon \nabla u_\varepsilon - u_\varepsilon \nabla v_\varepsilon|^2}{\big(v_\varepsilon \sqrt{(Au_\varepsilon, u_\varepsilon)} + u_\varepsilon \sqrt{(Av_\varepsilon, v_\varepsilon)}\big)^{2-p}} dx \geq 0,$$

which gives the same conclusion. □

Clearly the above results also hold true when $n = 1$, for example, for the following eigenvalue problem in an interval $(0, a)$:

$$\begin{cases} -\big(|\varphi'(s)|^{\gamma-2}\varphi'(s)\big)' = \lambda |\varphi(s)|^{\gamma-2}\varphi(s) & \text{in } (0,a), \\ \varphi(0) = \varphi(a) = 0, \end{cases} \tag{2.7}$$

where $\gamma > 1$ (see, e.g., [O1], [O2]). We give here a version of Theorem 2.1 for a problem slightly more general than (2.7). More precisely we consider the problem

$$\begin{cases} -\big(|\varphi'(s)|^{\gamma-2}\varphi'(s)\big)' = \lambda m(s) |\varphi(s)|^{\gamma-2}\varphi(s) & \text{in } (0,a), \\ \varphi(0) = \varphi(a) = 0, \end{cases} \tag{2.8}$$

where the weight $m(s)$ is a continuous function in $(0,a)$ satisfying the condition

$$0 < m(s) \le \frac{c}{\big(\sigma(s)\big)^{\gamma-\alpha}}, \tag{2.9}$$

with c a positive constant, $\alpha > 0$, $\sigma(s) = \min(s, a-s)$. Once again the first eigenvalue λ_γ of problem (2.8) can be found as minimum of the Rayleigh quotient

$$\lambda_\gamma = \min_{\substack{\varphi \in W_0^{1,\gamma}(0,a) \\ \varphi \not\equiv 0}} \frac{\displaystyle\int_0^a |\varphi'(s)|^\gamma \, ds}{\displaystyle\int_0^a m(s)|\varphi(s)|^\gamma \, ds}. \tag{2.10}$$

The existence of a first eigenfunction and of the first eigenvalue can be proven by standard tools also in this case. It is enough to observe that, under assumption (2.9), the embedding of $W_0^{1,\gamma}(0,a)$ in the weighted space $L^\gamma_m(0,a)$ is compact (see for example [N]).

Using the methods of Theorem 2.1 it is possible to prove the following theorem.

THEOREM 2.3. *The first eigenvalue of* (2.8) *is simple; that is, all the associated eigenfunctions u are scalar multiples of each other.*

Proof. All the arguments in the proof of Theorem 2.1 work in the same way for problem (2.8) and for the corresponding weak formulation. So, let $u > 0$, $v > 0$, two eigenfunctions of (2.8), relative to the first eigenvalue. We have to verify the analogue of (2.5), that is,

$$\lim_{\varepsilon \to 0^+} \int_0^a \left[\left(\frac{u(s)}{u_\varepsilon(s)} \right)^{\gamma-1} - \left(\frac{v(s)}{v_\varepsilon(s)} \right)^{\gamma-1} \right] \big(u_\varepsilon^\gamma(s) - v_\varepsilon^\gamma(s) \big) m(s) \, ds = 0,$$

where $u_\varepsilon = u + \varepsilon$ and $v_\varepsilon = v + \varepsilon$, $\varepsilon > 0$. The above result can be obtained by simply observing that for $\eta \in (0, a/2)$ we have

$$\int_0^a f_\varepsilon(s) \, ds = \int_0^\eta f_\varepsilon(s) \, ds + \int_{a-\eta}^a f_\varepsilon(s) \, ds + \int_\eta^{a-\eta} f_\varepsilon(s) \, ds,$$

where

$$f_\varepsilon(s) = \left[\left(\frac{u(s)}{u_\varepsilon(s)} \right)^{\gamma-1} - \left(\frac{v(s)}{v_\varepsilon(s)} \right)^{\gamma-1} \right] \big(u_\varepsilon^\gamma(s) - v_\varepsilon^\gamma(s) \big) m(s).$$

The last integral goes to zero as $\varepsilon \to 0^+$, while, using (2.9) and well-known embedding results (see [N]), the first two integrals can be made arbitrarily close to zero by suitably choosing η. The claim is then proven. □

We are interested in the case

$$m(s) = \begin{cases} s^{\alpha-\gamma} & \text{if } s \in (0, a/2), \\ (a-s)^{\alpha-\gamma} & \text{if } s \in (a/2, a), \end{cases} \tag{2.11}$$

where $\alpha > 0$. Using the symmetry of the problem, the Pólya–Szegö principle, and Hardy–Littlewood inequality, it is easy to show that any positive eigenfunction of (2.8) with $m(s)$ as in (2.11) is symmetric with respect to $s = a/2$, and it is increasing in $(0, a/2)$, decreasing in $(a/2, a)$.

In particular such a result allows us to obtain information about the following one-dimensional problem:

$$\begin{cases} -\big(|\varphi'(s)|^{\gamma-2}\varphi'(s)\big)' = \mu\, s^{\alpha-\gamma}|\varphi(s)|^{\gamma-2}\varphi(s) & \text{in } (0,b), \\ \varphi(0) = \varphi'(b) = 0. \end{cases} \tag{2.12}$$

The first eigenvalue of such a problem is given by

$$\mu_\gamma = \min_{\substack{\varphi\in W^{1,\gamma}(0,b)\\ \varphi\not\equiv 0,\ \varphi(0)=0}} \frac{\displaystyle\int_0^b |\varphi'(s)|^\gamma\, ds}{\displaystyle\int_0^b |\varphi(s)|^\gamma s^{\alpha-\gamma}\, ds}. \tag{2.13}$$

Indeed any function which realizes the minimum in (2.13) suitably extended by symmetry in $(0,2b)$ can be used as the test function vanishing in 0 and $2b$ in the Rayleigh quotient

$$\frac{\displaystyle\int_0^{2b} |\varphi'(s)|^\gamma\, ds}{\displaystyle\int_0^{2b} |\varphi(s)|^\gamma m(s)\, ds}.$$

Conversely, if $\varphi \in W_0^{1,\gamma}(0,2b)$ minimizes the above quotient in $W_0^{1,\gamma}(0,2b)$ then the restriction of φ to $(0,b)$ can be used as test function in (2.13). A simple consequence of the above considerations is the following.

THEOREM 2.4. *The first eigenvalue of* (2.12) *is simple; that is, all the associated eigenfunctions u are scalar multiples of each other. Furthermore every positive eigenfunction of* (2.12) *associated with the first eigenvalue is strictly increasing in* $(0,b)$.

Now we state a useful comparison lemma which can be seen as the "nonlinear" version of analogous "linear" results (see, e.g., [Ba]).

LEMMA 2.5. *Let μ_γ be the first eigenvalue of* (2.12) *and let φ be a positive eigenfunction associated with μ_γ. Suppose that, for $c > b$, $\psi \in W^{1,\gamma}(0,c)$ is a nonnegative increasing function such that $\big(\psi'(s)^{\gamma-1}\big)$ is absolutely continuous in $[\varepsilon, c]$ for every $\varepsilon > 0$, and*

$$\begin{cases} -\big((\psi'(s))^{\gamma-1}\big)' \le \mu_\gamma\, s^{\alpha-\gamma}(\psi(s))^{\gamma-1} & \text{a.e. in } (0,c), \\ \psi(0) = \psi'(c) = 0. \end{cases} \tag{2.14}$$

If there exists $s_1 \in (0,b)$ such that $\varphi'(s_1) = \psi'(s_1)$ then

$$\psi(s) \le \varphi(s), \qquad \forall s \in (0, s_1). \tag{2.15}$$

Proof. First of all we prove that $\psi(s_1) \le \varphi(s_1)$. Suppose, ab absurdo, that $\psi(s_1) > \varphi(s_1)$. We put $w(s) = \psi(s) - \varphi(s)$. We define

$$z(s) = \begin{cases} \psi(s) & \text{if } s \in (0, s_1), \\ \varphi(s) + w(s_1) & \text{if } s \in [s_1, b]. \end{cases} \tag{2.16}$$

By construction, $\big(z'(s)\big)^{\gamma-1}$ is absolutely continuous in $[\varepsilon, c]$ for every $\varepsilon > 0$, and $z(s)$ satisfies

$$\begin{aligned}
&\big(z'(s)\big)^{\gamma-1} = \big(\varphi'(s)\big)^{\gamma-1}, \qquad z(s) > \varphi(s), \qquad s \in (s_1, b),\\
&\big(z'(s)\big)^{\gamma-1} = \big(\psi'(s)\big)^{\gamma-1}, \qquad z(s) = \psi(s), \qquad s \in (0, s_1).
\end{aligned}$$

This means that $z(s)$ satisfies the inequality

$$-\big((z'(s))^{\gamma-1}\big)' \le \mu_\gamma\, s^{\alpha-\gamma}(z(s))^{\gamma-1}, \quad \text{a.e. in } (0,b).$$

Multiplying the above inequality by $z(s)$ and integrating we obtain

$$\int_0^b |z'(s)|^\gamma ds \le \mu_\gamma \int_0^b |z(s)|^\gamma s^{\alpha-\gamma} ds,$$

where, taking into account the continuity of the embedding of $W^{1,\gamma}(0,b)$ into $L^\gamma_{s^{-\gamma}}(0,b)$, we have used the following:

$$\lim_{\varepsilon\to 0^+} \big(z'(\varepsilon)\big)^{\gamma-1} z(\varepsilon) = 0.$$

We have then proven

$$\min_{\substack{\varphi\in W^{1,p}(0,b)\\ \varphi\not\equiv 0\ \varphi(0)=0}} \frac{\displaystyle\int_0^b |\varphi'(s)|^p}{\displaystyle\int_0^b |\varphi(s)|^\gamma s^{\alpha-\gamma} ds} \le \mu_\gamma. \tag{2.17}$$

Using (2.13) and Theorem 2.4 we have that in (2.17) the equality sign has to hold and that $z(s)$ is proportional to $\varphi(s)$. Then, by (2.16), it has to coincide with $\varphi(s)$. But this is a contradiction. The claim is then proven.

In order to completely prove the assertion we observe that, if $\psi(s) > \varphi(s)$ for some $s \in (0, s_1)$, then a positive maximum $\tilde{s} \in (0, s_1)$ of $w(s)$ would exist. But this is forbidden because one can repeat the above arguments with $\tilde{s}$ in place of s_1. □

Remark 2.1. The assumption $c > b$ in Lemma 2.5 takes into account the only nontrivial case. Indeed, if $c = b$, then a function ψ which satisfies (2.14) has to coincide with the eigenfunction associated to the first eigenvalue of (2.12). This means that the hypothesis $\varphi'(s_1) = \psi'(s_1)$, $s_1 \in (0, b)$, implies $\varphi(s) = \psi(s)$.

About the case $c < b$ it is simple to observe that under such an assumption no function $\psi \not\equiv 0$ satisfying (2.14) exists. In fact, the existence of such a function would imply that the first eigenvalue of problem (2.13) in the interval $(0, b)$ is bigger or equal to the first eigenvalue of the same problem in $(0, c)$. This contradicts the fact that the first eigenvalue is a strictly decreasing function of the measure of the interval.

Remark 2.2. If one retraces the proof of Lemma 2.5 it is evident that it is not necessary that ψ satisfies (2.14) in an interval $(0, c)$. One can simply suppose that for $s_1 \in (0, b)$ the following condition holds:

$$\begin{cases} -\big((\psi'(s))^{\gamma-1}\big)' \le \mu_\gamma\, s^{\alpha-\gamma}(\psi(s))^{\gamma-1} & \text{a.e. in } (0, s_1),\\ \psi(0) = 0, \quad \psi'(s_1) = \varphi'(s_1). \end{cases}$$

3. Faber–Krahn inequality. In this section we will prove the following comparison result for the first eigenvalue of problem (2.1).

THEOREM 3.1. *If λ_p is the first eigenvalue of problem* (2.1) *then*

$$\lambda_p \geq \lambda_p^{\#},$$

where $\lambda_p^{\#}$ is the first eigenvalue of the problem

$$\begin{cases} -\mathrm{div}\big(|\nabla v|^{p-2}\nabla v\big) = \lambda v|v|^{p-2} & \text{in } \Omega^{\#}, \\ v = 0 & \text{on } \partial\Omega^{\#}, \end{cases} \tag{3.1}$$

and $\Omega^{\#}$ is the ball centered at the origin such that $|\Omega^{\#}| = |\Omega|$, where, here and in the following, we denote by $|E|$ the measure of any measurable set $E \subset \mathbb{R}^n$. Furthermore, $\lambda_p = \lambda_p^{\#}$ if and only if $\Omega = \Omega^{\#}$ and $a_{ij}(x)x_j = x_i$, a.e. in Ω, modulo translations.

Remark 3.1. In the case $p = 2$, $n = 2$ the above result is known as the Faber–Krahn theorem. In such a case it can be stated as *the membrane with lowest principal frequency is the circular one.*

Remark 3.2. We observe explicitly that, if B is a ball centered at the origin, then ν_p is the first eigenvalue of problem (3.1) in B if and only if the one-dimensional problem

$$\begin{cases} -\big(|\varphi'(s)|^{\gamma-2}\varphi'(s)\big)' = \mu\, s^{-\gamma(1-1/n)}|\varphi(s)|^{\gamma-2}\varphi(s) & \text{in } (0,|B|), \\ \varphi(0) = \varphi'(|B|) = 0, \end{cases}$$

where $\gamma = p' = p/(p-1)$, has first eigenvalue

$$\mu_\gamma = \left(\frac{\nu_p^{1/p}}{nC_n^{1/n}}\right)^{p'}.$$

Such a claim follows from the symmetry of problem (3.1). Because of the simplicity of the first eigenvalue, any eigenfunction v associated with the first eigenvalue of problem (3.1) in a ball B is spherically symmetric. If we put $B_s = \{x : C_n|x|^n < s\}$ and $V(s) = \int_{B_s} v^{p-1}(x)\,dx$, where v is a positive eigenfunction, we have

$$\begin{cases} -\Big(\big(V'(s)\big)^{1/(p-1)}\Big)' = \frac{\nu_p^{1/(p-1)}}{(nC_n^{1/n}s^{1-1/n})^{p/(p-1)}}\big(V(s)\big)^{1/(p-1)} & \text{in } (0,|B|), \\ V(0) = V'(|B|) = 0, \end{cases}$$

and the assertion follows.

Before giving the proof of Theorem 3.1 we recall a few definitions about rearrangements (for more details see, e.g., [Ba], [Ta1], [Ta3]). If $\varphi : \Omega \to \mathbb{R}$ is a measurable function we define the decreasing rearrangement of φ

$$\varphi^*(s) = \sup\{t > 0 : |\{x \in \Omega : |\varphi(x)| > t\}| > s\}$$

and the spherically symmetric decreasing rearrangement of φ

$$\varphi^{\#}(x) = \varphi^*(C_n|x|^n),$$

where C_n denotes the measure of the unit ball in $\mathbb{R}^n$.

The following result proven in [BZ] will be useful in the proof of Theorem 3.1.

THEOREM 3.2. *Let u be a positive function in $W_0^{1,p}(\Omega)$, $1 < p < \infty$. Suppose that the equality sign holds in the Pólya–Szegö inequality, that is,*

$$\int_\Omega |\nabla u|^p \, dx = \int_{\Omega^\#} |\nabla u^\#|^p \, dx.$$

If $|\{x : \nabla u^\# = 0\} \cap {u^\#}^{-1}(0, \operatorname{ess\,sup} u)| = 0$, then $u = u^\#$ a.e., modulo translations.

Proof of Theorem 3.1. Clearly $\lambda_p^\#$ can be written as

$$\lambda_p^\# = \min_{v \in W_0^{1,p}(\Omega^\#)} \frac{\displaystyle\int_{\Omega^\#} |\nabla v|^p \, dx}{\displaystyle\int_{\Omega^\#} |v|^p \, dx}.$$

If $u > 0$ is an eigenfunction associated with the first eigenvalue of (2.1), using the ellipticity condition and Pólya–Szegö principle, we have

$$\lambda_p = \frac{\displaystyle\int_\Omega (A\nabla u, \nabla u)^{p/2} \, dx}{\displaystyle\int_\Omega |u|^p \, dx} \geq \frac{\displaystyle\int_\Omega |\nabla u|^p \, dx}{\displaystyle\int_\Omega |u|^p \, dx} \geq \frac{\displaystyle\int_{\Omega^\#} |\nabla u^\#|^p \, dx}{\displaystyle\int_{\Omega^\#} |u^\#|^p \, dx} \geq \lambda_p^\#, \tag{3.2}$$

that is, the first part of the theorem.

About the case of equality we observe that if $\lambda_p = \lambda_p^\#$ then inequality (3.2) implies that $u^\#$ is an eigenfunction of problem (3.1) corresponding to the first eigenvalue. Furthermore, the following equality holds:

$$\int_\Omega |\nabla u|^p \, dx = \int_{\Omega^\#} |\nabla u^\#|^p \, dx, \tag{3.3}$$

because $\|u\|_p = \|u^\#\|_p$. On the other hand, the fact that $u^\#$ is an eigenfunction of (3.1) implies that

$$|\{x : 0 < u^\#(x) < \sup u, \ \nabla u^\#(x) = 0\}| = 0.$$

By Theorem 3.2 (3.3) then gives:

$$u^\#(x) = u(x), \qquad \Omega = \Omega^\#, \quad \text{modulo translations.}$$

The assertion about the coefficients a_{ij} is a consequence of the following observation (see also [ALT1], [Ke]). The vector $\nabla u = \nabla u^\#$ points in the direction of x. This means that if equality holds in (3.2) then

$$a_{ij}(x) x_j x_i = |x|^2 \quad \text{a.e. in } \Omega.$$

But if A is a symmetric matrix whose first eigenvalue is greater or equal to 1, then x is an eigenvector corresponding to the eigenvalue 1; that is the assertion. □

4. Payne–Rayner-type inequalities. Let us consider a fixed eigenfunction $u \not\equiv 0$ of problem (2.1), corresponding to the first eigenvalue, that is a solution of the problem

$$\begin{cases} -\operatorname{div}\big((A\nabla u, \nabla u)^{(p-2)/2} A\nabla u\big) = \lambda_p u |u|^{p-2} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

Let B be the ball centered at the origin such that the first eigenvalue of problem (3.1) in B is equal to λ_p, and let $v_q > 0$, $0 < q \le \infty$, be the corresponding eigenfunction such that

$$\|v_q\|_q = \|u\|_q. \tag{4.2}$$

In other words v_q satisfies (4.2) and solves the following problem:

$$\begin{cases} -\mathrm{div}\big(|\nabla v_q|^{p-2}\nabla v_q\big) = \lambda_p v_q |v_q|^{p-2} & \text{in } B, \\ \\ v_q = 0 & \text{on } \partial B. \end{cases} \tag{4.3}$$

A straightforward calculation shows that the ball B having the above properties is given by

$$B = \{x \in \mathbb{R}^n : |x| < (\kappa_p/\lambda_p)^{1/p}\},$$

where κ_p denotes the first eigenvalue of problem (4.3) in the unit ball. Making use of the characterization (2.2) and of Theorem 3.1 it is also clear that $|B| \le |\Omega|$.

We will prove the following comparison result between u and v_q.

THEOREM 4.1. *If u and v_q are defined as above we have*

(i) *if $0 < q < \infty$ then*

$$\int_0^s \big(u^*(t)\big)^q dt \le \int_0^s \big(v_q^*(t)\big)^q dt, \qquad s \in [0, |B|];$$

(ii) *if $q = \infty$ then*

$$u^*(s) \ge v_\infty^*(s), \qquad s \in [0, |B|].$$

If any of the above inequalities hold as equalities then $\Omega = B$, $u(x) = u^{\#}(x) = v_q(x)$ and $a_{ij}(x)x_j = x_i$, a.e. in Ω, modulo translations.

Before proving the theorem we quote a lemma whose proof is contained for example in [Ta1], [Ta2], [BFM].

LEMMA 4.2. *Let u be solution of* (4.1); *then the following inequality holds:*

$$-\big(u^*(s)\big)' \le \frac{\lambda_p^{1/(p-1)}}{(nC_n^{1/n}s^{1-1/n})^{p/(p-1)}} \left(\int_0^s \big(u^*(t)\big)^{p-1} dt\right)^{\frac{1}{p-1}} \qquad \text{a.e. in } (0, |\Omega|).$$

Proof of Theorem 4.1. If $|B| = |\Omega|$ then, because of Theorem 3.1, there is nothing to prove. So we will suppose $|B| < |\Omega|$.

Let us define

$$U(s) = \int_0^s \big(u^*(t)\big)^{p-1} dt, \qquad s \in [0, |\Omega|],$$

and

$$V(s) = \int_0^s \big(v_q^*(t)\big)^{p-1} dt, \qquad s \in [0, |B|].$$

Taking into account Lemma 4.2 and the definition of v_q we have that $U(s)$ and $V(s)$ satisfy the following conditions:

$$\begin{cases} -\Big(\big(U'(s)\big)^{1/(p-1)}\Big)' \le \frac{\lambda_p^{1/(p-1)}}{(nC_n^{1/n}s^{1-1/n})^{p/(p-1)}} \big(U(s)\big)^{1/(p-1)} & \text{a.e. in } (0, |\Omega|), \\ U(0) = U'(|\Omega|) = 0, \end{cases} \tag{4.4}$$

and

$$\text{(4.5)}\quad \begin{cases} -\Big(\big(V'(s)\big)^{1/(p-1)}\Big)' = \frac{\lambda_p^{1/(p-1)}}{(nC_n^{1/n}s^{1-1/n})^{p/(p-1)}}\big(V(s)\big)^{1/(p-1)} \text{ in } (0,|B|), \\ V(0) = V'(|B|) = 0. \end{cases}$$

We remark that (4.5) holds for every $0 < q \leq \infty$.

We start to consider the case $q = p-1$. First of all we observe that, by definition, (4.2) implies that $U(|B|) < U(|\Omega|) = V(|B|)$. We will prove that $U(s) \leq V(s)$ in $(0,|B|)$. Suppose ab absurdo that a positive maximum of $U(s) - V(s)$ exists in $s_1 \in (0,|B|)$. In such a point we have $U(s_1) > V(s_1)$ and $U'(s_1) = V'(s_1)$. Using (4.4), (4.5), and recalling Remark 3.2, we can apply Lemma 2.5 with $\varphi = V$, $\psi = U$, $b = |B|$, $c = |\Omega|$, $\gamma = p' = p/(p-1)$, $\mu_\gamma = \big(\lambda_p^{1/p}/nC_n^{1/n}\big)^{p'}$, $\alpha = p'/n$, obtaining a contradiction. The case $q = p-1$ is then proven.

In order to complete the proof of case (i) in theorem 4.1 we consider, for $0 < q < \infty$, the functions

$$U_q(s) = \int_0^s \big(u^*(t)\big)^q dt, \qquad s \in [0,|\Omega|],$$

and

$$V_q(s) = \int_0^s \big(v_q^*(t)\big)^q dt, \qquad s \in [0,|B|].$$

By definition (4.2) of v_q we have $U_q(|B|) < U_q(|\Omega|) = V_q(|B|)$. We will prove that $U_q(s) \leq V_q(s)$ in $(0,|B|)$. Suppose ab absurdo that a positive maximum of $U_q(s) - V_q(s)$ exists in $s_1 \in (0,|B|)$. In such a point we have $U_q(s_1) > V_q(s_1)$ and $U_q'(s_1) = V_q'(s_1)$; then $U'(s_1) = V'(s_1)$. Taking into account (4.4), (4.5), Remark 3.2, and Lemma 2.5, it follows that

$$\text{(4.6)}\qquad U(s) \leq V(s), \qquad s \in [0,s_1].$$

Now (4.4), (4.5), and (4.6) give

$$\text{(4.7)}\qquad u^*(s) \leq v_q^*(s) \qquad \text{a.e. in } (0,s_1).$$

Obviously (4.7) implies $U_q(s_1) \leq V_q(s_1)$; that is a contradiction. The proof of part (i) is then complete.

Part (ii) can be proven using similar arguments. We define

$$s_1 = \inf\{s \in (0,|B|) : u^*(t) \geq v_\infty^*(t) \quad \forall t \in (s,|B|]\}.$$

Taking into account the fact that $|B| < |\Omega|$ and $u^*(|B|) > v_\infty^*(|B|) = 0$, we have that s_1 is well defined and moreover $u^*(s_1) = v_\infty^*(s_1)$; that is, $U'(s_1) = V'(s_1)$ if $s_1 > 0$. Obviously in the case $s_1 = 0$ we immediately have the assertion. On the contrary, if ab absurdo $s_1 > 0$ then, by the above arguments, we have

$$\int_0^s \big(u^*(t)\big)^{p-1} dt \leq \int_0^s \big(v_\infty^*(t)\big)^{p-1} dt, \qquad s \in [0,s_1].$$

This inequality, together with (4.4) and (4.5), gives

$$-\big(u^*(s)\big)' \leq -\big(v_\infty^*(s)\big)' \qquad \text{a.e. in } (0,s_1).$$

Integrating between 0 and s and taking into account (4.2) we obtain

$$u^*(s) \geq v_\infty^*(s) \quad \text{in } (0, s_1),$$

which contradicts the assumption $s_1 > 0$.

Finally, the case of equality in (i) and (ii) follows immediately from Theorem 3.1. It is enough to observe that if, for example, equality holds in part (i), then $u^* = v_q^*$ and $|\Omega| = |B|$. This means that Theorem 3.1 applies. □

A consequence of Theorem 4.1 is the following reverse inequality.

THEOREM 4.3. *Let $u > 0$ be an eigenfunction of problem (4.1). Then, for $0 < q < r \leq \infty$, we have*

$$\|u\|_r \leq \beta(r, q, p, n, \lambda_p)\|u\|_q, \tag{4.8}$$

where $\beta(r, q, p, n, \lambda_p) = \frac{\|v\|_r}{\|v\|_q}$ and v is any nontrivial eigenfunction of (4.3). Furthermore, equality in (4.8) holds if and only if Ω is a ball centered at the origin, $u(x) = u^\#(x)$, and $a_{ij}(x)x_j = x_i$, a.e. in Ω, modulo translations.

Proof. Inequality (4.8) is an immediate consequence of Theorem 4.1. Indeed, choosing v_q as in (4.2) and defining $v_q^*(s) = 0$ for $s \in (|B|, |\Omega|)$, Theorem 4.1 implies that $u^q \in K(v_q^q)$, where, for $f \in L_+^1(\Omega)$ we denote by $K(f)$ the set of functions $\varphi \in L_+^1(\Omega)$ such that (see [ALT2])

$$\int_0^s \varphi^*(t)dt \leq \int_0^s f^*(t)dt \quad \forall s \in (0, |\Omega|), \qquad \text{and} \quad \int_0^{|\Omega|} \varphi^*(t)dt = \int_0^{|\Omega|} f^*(t)dt.$$

Using well-known properties about $K(f)$ (see, e.g., [ALT2]), one obtains

$$\|u\|_r \leq \|v_q\|_r = \frac{\|v_q\|_r}{\|v_q\|_q}\|u\|_q,$$

that is, the assertion.

About the case of equality in (4.8) it is clear that by Theorem 3.1 it will be sufficient to show that $|B| = |\Omega|$. We then suppose ab absurdo that $|B| < |\Omega|$. Let us first consider the case $r = \infty$. So we have

$$\|u\|_\infty = \frac{\|v\|_\infty}{\|v\|_q}\|u\|_q, \qquad 0 < q < \infty,$$

where v is a nontrivial eigenfunction of (4.3). In particular we can choose $v = v_q$ and then

$$\|u\|_\infty = \|v_q\|_\infty,$$

that is, $v_q = v_\infty$. Taking into account Theorem 4.1 we immediately have $u^*(s) = v_\infty^*(s)$, which gives the contradiction, again by Theorem 4.1.

The contradiction is achieved even in the case $0 < r < \infty$. Indeed, arguing as above, we have that equality in (4.8) implies

$$\|u\|_r = \|v_q\|_r.$$

We have already observed that, suitably defining $v_q^*(s)$ in $[|B|, |\Omega|]$, Theorem 4.1 says that $u^q \in K(v_q^q)$. The functional $F(\varphi) = \|\varphi\|_{r/q}$ is strictly convex on $K(v_q^q)$

and assumes its maximum $\|v_q^q\|_{r/q}$ on such a set. It is well known (see [ALT2]) that the maximum of $F(\varphi)$ is achieved only on extreme points of $K(v_q^q)$, that is, on rearrangements of v_q^q. This observation proves the assertion. □

Remark 4.1. As in [C2] one can observe that Lemma 4.2 holds for any solution of problem (2.1), not necessarily corresponding to the first eigenvalue. So the results stated in Theorems 4.1 and 4.3 hold true for every eigenfunction of problem (2.1).

Appendix.

Proof of Lemma 2.2. We start with the case $p \geq 2$. Let us put $\mathcal{A}(\xi) = \sqrt{(A\xi, \xi)}$. Simple arguments show that

$$\left|\mathcal{A}\left(\frac{\xi_1+\xi_2}{2}\right)\right|^p + \left|\mathcal{A}\left(\frac{\xi_1-\xi_2}{2}\right)\right|^p \leq \left[\mathcal{A}\left(\frac{\xi_1+\xi_2}{2}\right)^2 + \mathcal{A}\left(\frac{\xi_1-\xi_2}{2}\right)^2\right]^{p/2}$$

$$= \left[\frac{\mathcal{A}(\xi_1)^2 + \mathcal{A}(\xi_2)^2}{2}\right]^{p/2}.$$

By convexity of the function $f(t) = t^\alpha$, with $\alpha \geq 1$, we have the Clarkson-type inequality

$$\left|\mathcal{A}\left(\frac{\xi_1+\xi_2}{2}\right)\right|^p + \left|\mathcal{A}\left(\frac{\xi_1-\xi_2}{2}\right)\right|^p \leq \frac{\mathcal{A}(\xi_1)^p + \mathcal{A}(\xi_2)^p}{2}. \tag{A.1}$$

On the other hand the function $g(\xi) = \mathcal{A}(\xi)^p$ is strictly convex because $p \geq 2$ and $\mathcal{A}(\xi)^2$ is strictly convex. This means that

$$\mathcal{A}(\xi_2)^p \geq \mathcal{A}(\xi_1)^p + p\mathcal{A}(\xi_1)^{p-2}(A\xi_1, \xi_2 - \xi_1). \tag{A.2}$$

Using (A.1) and (A.2) we have

$$\mathcal{A}(\xi_2)^p \geq -\mathcal{A}(\xi_1)^p + 2\left|\mathcal{A}\left(\frac{\xi_1+\xi_2}{2}\right)\right|^p + 2\left|\mathcal{A}\left(\frac{\xi_1-\xi_2}{2}\right)\right|^p$$

$$\geq \mathcal{A}(\xi_1)^p + p\mathcal{A}(\xi_1)^{p-2}(A\xi_1, \xi_2 - \xi_1) + 2\left|\mathcal{A}\left(\frac{\xi_1-\xi_2}{2}\right)\right|^p.$$

We have then proven part (i) of Lemma 2.2 with the constant 2^{1-p} instead of $(2^{p-1}-1)^{-1}$. Using an iteration argument as in the appendix of [L1] one obtains the full claim.

In the case $1 < p < 2$ we consider the function

$$\phi(t) = \mathcal{A}(\xi_1 + t(\xi_2 - \xi_1))^p,$$

where $\mathcal{A}(\xi)$ is the previously defined function. If $\phi(t) = 0$ for some $t \in [0,1]$, then by ellipticity condition ξ_1 and ξ_2 are multiples of each other, with the possibility that one or both of them vanish. In such cases the claim of the lemma is evident. So we can suppose that $\phi(t) \neq 0$ in $[0,1]$. Then we can use the equality

$$\phi(1) = \phi(0) + \phi'(0) + \int_0^1 (1-t)\phi''(t)\,dt$$

to get

$$\mathcal{A}(\xi_1)^p = \mathcal{A}(\xi_1)^p + p\,\mathcal{A}(\xi_1)^{p-2}(A\xi_1, \xi_2 - \xi_1) + \int_0^1 (1-t)\phi''(t)\,dt. \tag{A.3}$$

A straightforward calculation gives

$$\begin{aligned}\phi''(t) = {} & p(p-2)\mathcal{A}(\xi_1 + t(\xi_2 - \xi_1))^{p-4}\big(A(\xi_1 + t(\xi_2 - \xi_1)), \xi_2 - \xi_1\big)^2 \\ & + p\mathcal{A}(\xi_1 + t(\xi_2 - \xi_1))^{p-2}\mathcal{A}(\xi_2 - \xi_1)^2.\end{aligned}$$

Now we observe that, because of the symmetry of the matrix A, we have, for every $\xi_1, \xi_2 \in \mathbb{R}^n$,

$$(A\xi_1, \xi_2)^2 \le (A\xi_1, \xi_1)(A\xi_2, \xi_2),$$

$$\mathcal{A}(\xi_1 + \xi_2) \le \mathcal{A}(\xi_1) + \mathcal{A}(\xi_2).$$

Using the above inequalities we obtain

$$\begin{aligned}\phi''(t) &\ge p(p-1)\mathcal{A}(\xi_1 + t(\xi_2 - \xi_1))^{p-2}\mathcal{A}(\xi_2 - \xi_1)^2 \\ &\ge p(p-1)\frac{\mathcal{A}(\xi_2 - \xi_1)^2}{\big(\mathcal{A}(\xi_1) + \mathcal{A}(\xi_2)\big)^{2-p}} \qquad t \in [0,1].\end{aligned}$$

This inequality and (A.3) give the assertion. □

Acknowledgments. We thank the referees who helped us in recovering a few misprintings and pointed out some useful references.

REFERENCES

[ALT1] A. Alvino, P. L. Lions, and G. Trombetti, *A remark on comparison results via symmetrization*, Proc. Roy. Soc. Edinburgh, 102A (1986), pp. 37–48.

[ALT2] A. Alvino, P. L. Lions, and G. Trombetti, *On optimization problems with prescribed rearrangements*, Nonlinear Anal., 13 (1989), pp. 185–220.

[A] A. Anane, *Simplicité et isolation de la première valeur propre du p-laplacien avec poids*, C. R. Acad. Sci. Paris, 305 (1987), pp. 725–728.

[Ba] C. Bandle, *Isoperimetric Inequalities and Applications*, Monographs and Studies in Math. 7, Pitman, London, 1980.

[BFM] M. F. Betta, V. Ferone, and A. Mercaldo, *Regularity for solutions of nonlinear elliptic equations*, Bull. Sci. Math., 118 (1994), pp. 539-567.

[B1] T. Bhattacharya, *Radial symmetry of the first eigenfunction for the p-laplacian in the ball*, Proc. Amer. Math. Soc., 104 (1988), pp. 169–174.

[B2] T. Bhattacharya, *Some results concerning the eigenvalue problem for the p-laplacian*, Ann. Acad. Sci. Fenn. Ser. A. I. Math., 14 (1989), pp. 325–343.

[BZ] J. E. Brothers and W. P. Ziemer, *Minimal rearrangements of Sobolev functions*, J. Reine Angew. Math., 384 (1988), pp. 153–179.

[C1] G. Chiti, *An isoperimetric inequality for the eigenfunctions of linear second order elliptic operators*, Boll. Un. Mat. Ital. A(6), 1 (1982), pp. 145–151.

[C2] G. Chiti, *A reverse Hölder inequality for the eigenfunctions of linear second order elliptic operators*, Z. Angew. Math. Phys., 33 (1982), pp. 143–148.

[G] E. Giusti, *Metodi diretti nel calcolo delle variazioni*, Unione Matematica Italiana, Bologna, 1994.

[Ke] S. Kesavan, *On a comparison theorem via symmetrization*, Proc. Roy. Soc. Edinburgh, 119A (1991), pp. 159–167.

[KJ] M. T. Kohler-Jobin, *Sur la premiére fonction propre d'une membrane: Une extension à N dimensions de l'inégalité isopérimétrique de Payne-Rayner*, Z. Angew. Math. Phys., 28 (1977), pp. 1137–1140.

[L1] P. Lindqvist, *On the equation div*$(|\nabla u|^{p-2}\nabla u) + \lambda |u|^{p-2}u = 0$, Proc. Amer. Math. Soc., 109 (1990), pp. 157–164. Addendum: Proc. Amer. Math. Soc., 116 (1992), pp. 583–584.

[L2] P. Lindqvist, *On non-linear Rayleigh quotients*, Potential Anal., 2 (1993), pp. 199–218.

[M] J. Mossino, *A generalization of the Payne-Rayner isoperimetric inequality*, Boll. Un. Mat. Ital. A(6), 2 (1983), pp. 335–342.

[N] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[O1] M. Ôtani, *Sur certaines équations différentielles ordinaires du second ordre aux inégalités du type Sobolev-Poincaré*, C. R. Acad. Sci. Paris, 296 (1983), pp. 415–418.

[O2] M. Ôtani, *A remark on certain nonlinear elliptic equations*, Proc. Fac. Sci. Tokai Univ., 19 (1984), pp. 23–28.

[PR1] L. E. Payne and M. E. Rayner, *An isoperimetric inequality for the first eigenfunction in the fixed membrane problem*, Z. Angew. Math. Phys., 23 (1972), pp. 13–15.

[PR2] L. E. Payne and M. E. Rayner, *Some isoperimetric norm bound for solutions of the Helmoltz equation*, Z. Angew. Math. Phys., 24 (1973), pp. 105–110.

[Sa] S. Sakaguchi, *Concavity properties of solutions to some degenerate quasilinear elliptic Dirichlet problems*, Ann. Scuola Norm. Sup. Pisa, 14 (1987), pp. 403-421.

[Se] J. Serrin, *Local behavior of solutions of quasi-linear equations*, Acta Math., 111 (1964), pp. 245–302.

[Ta1] G. Talenti, *Elliptic equations and rearrangements*, Ann. Scuola Norm. Sup. Pisa, Cl. Sci., 3 (1976), pp. 697–718.

[Ta2] G. Talenti, *Nonlinear elliptic equations, rearrangements of function and Orlicz spaces*, Ann. Mat. Pura Appl., 120 (1979), pp. 159-184.

[Ta3] G. Talenti, *Linear elliptic P.D.E.'s: Level sets, rearrangements and a priori estimates of solutions*, Boll. Un. Mat. Ital. B(6), 4 (1985), pp. 917–949.

[Th] F. de Thélin, *Sur l'espace propre associé à la première valeur propre du pseudo-laplacien*, C. R. Acad. Sci. Paris, 303 (1986), pp. 355–358.

[Tr] N. S. Trudinger, *On Harnack type inequalities and their application to quasilinear elliptic equations*, Comm. Pure Appl. Math., 20 (1967), pp. 721–747.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 452–480, March 1998

009

LINEAR PARABOLIC STOCHASTIC PDEs AND WIENER CHAOS*

R. MIKULEVICIUS† AND B. ROZOVSKII‡

Abstract. We study Cauchy's problem for a second-order linear parabolic stochastic partial differential equation (SPDE) driven by a cylindrical Brownian motion. Existence and uniqueness of a generalized (soft) solution is established in Sobolev, Hölder, and Lipschitz classes. We make only minimal assumptions, virtually identical to those common to similar deterministic problems. A stochastic Feynman–Kac formula for the soft solution is also derived. It is shown that the soft solution allows a Wiener chaos expansion and that the coefficients of this expansion can be computed recursively by solving a simple system of parabolic PDEs.

Key words. stochastic PDEs, Wiener chaos, soft solution, Feynman–Kac formula

AMS subject classifications. Primary, 60H15; Secondary, 35R60

PII. S0036141096299065

1. Introduction.

In this paper we study second-order linear SPDEs of the type

$$\begin{aligned} &du(t,x) = (\mathcal{L}u(t,x) + f(t,x))dt + (\mathcal{M}u(t,x) + g(t,x), dW_t)_Y, \\ &u(0,x) = \varphi(x),\ (t,x) \in (0,1] \times \mathbb{R}^d, \end{aligned} \tag{0.1}$$

where $\mathcal{L}u(t,x) = a^{ij}(t,x)u_{x_i,x_j} + b^i(t,x)u_{x_i} + c(t,x)u$, $\mathcal{M}u(t,x) = \sigma^i(t,x)u_{x_i} + h(t,x)u$, and W is a cylindrical Brownian motion in some Hilbert space Y. The coefficients of the operator $\mathcal{L}$ and $f(t,x)$ are real-valued functions while the coefficients of $\mathcal{M}$ and $g(t,x)$ are Y-valued. Equation (0.1) is assumed to be parabolic; i.e., the matrix (a^{ij}) is symmetric, and $A = (2a^{ij} - (\sigma^i, \sigma^j)_Y)$ is nonnegatively definite.

To motivate the study, we recall two important examples of equation (0.1).

1. *Backward diffusion equation* (see [11], [17], [29]). Let $X^{t,x}(s)$ be a diffusion process defined by the Ito equation

$$\begin{aligned} &dX^{t,x}(s) = b(X^{t,x}(s))ds + \sigma(X^{t,x}(s))dw_s, \quad s \in (t,1], \\ &X^{t,x}(t) = x, \quad x \in R^1, \text{ and } w_s \text{ is a one-dimensional Brownian motion.} \end{aligned}$$

It is well known (see, e.g., [29]) that assuming some smoothness of $b(x)$ and $\sigma(x)$ (which will be discussed later) one can show that the function $u(t,x) = X^{1-t,x}(1)$ is a solution of the equation

$$\begin{aligned} &du(t,x) = \left[\tfrac{1}{2}\sigma^2(x)u_{xx}(t,x) + b(x)u_x(t,x)\right]dt + \sigma(x)u_x(t,x)dw_1(t), \\ &u(0,x) = x\,, \text{ where } w_1(t) = w_1 - w_{1-t}. \end{aligned} \tag{0.2}$$

2. *Zakai equation* (see [32], [29]). Let us consider the nonlinear filtering problem in the following, rather common, setting. Assume that the signal process X_t is a diffusion process defined by the Ito equation

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t (\sigma(X_s)dw_s + \hat{\sigma}(X_s)d\hat{w}_s),$$

*Received by the editors February 21, 1996; accepted for publication (in revised form) January 21, 1997.

http://www.siam.org/journals/sima/29-2/29906.html

†Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania.

‡Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, CA 90089-1113 (rozovski@cams.usc.edu). This work was partially supported by ONR grant N00014-95-1-0229 and ARO grant DAAH04-95-1-0164.

where w and $\hat{w}$ are one-dimensional Brownian motions and X_0 is a random variable with density function $p(x)$. The Brownian motions w and $\hat{w}$ and the random variable X_0 are defined on the probability space $(\Omega, \mathcal{F}, P)$ and are assumed to be independent. The observation process is given by

$$y_t = \int_0^t h(X_s)ds + w_t \ .$$

It is a standard fact that for every function ψ such that $E|\psi(x_t)|^2 < \infty$, the optimal mean square estimate for $\psi(X_t)$, given the past of the observations $\mathcal{F}_t^y = \sigma(y_s, s \le t)$, is of the form

$$\hat{\psi}_t = \frac{E_{\tilde{P}}\left[\psi(X_t)\zeta(t)|\mathcal{F}_t^y\right]}{E_{\tilde{P}}\left[\zeta(t)|\mathcal{F}_t^y\right]},$$

where $\zeta(t) = \exp\{\int_0^t h(X_s)dy_s - \frac{1}{2}\int_0^t |h(X_s)|^2 ds\}$ and $d\tilde{P} = \zeta(1)^{-1}dP$.

Assuming some smoothness of the coefficients b, σ, and h, one can show that

$$E_{\tilde{P}}[\psi(X_t)\zeta(t)|\mathcal{F}_t^y] = \int \psi(x)u(t,x)dx \ , \tag{0.3}$$

where $u(t,x)$, usually referred to as the unnormalized filtering density, is a solution of the Zakai equation

$$\begin{aligned} &du(t,x) = \left[\tfrac{1}{2}\left(\left(\sigma^2(x) + \hat{\sigma}^2(x)\right)u(t,x)\right)_{xx} - \left(b(x)u(t,x)\right)_x\right]dt \\ &+ \left[h(x)u(t,x) - \left(\sigma(x)u(t,x)\right)_x\right]dy_t, \quad t > 0 \ , \\ &u(0,x) = p(x) \ . \end{aligned} \tag{0.4}$$

Since y is a Brownian motion on $(\Omega, \mathcal{F}, \tilde{P})$, equation (0.4) is obviously a particular case of equation (0.1).

The Cauchy problem (0.1) has been studied by many authors [27], [14], [2], etc. In the superparabolic case (the matrix A is uniformly nondegenerate) there exists a quite complete theory for this problem in Sobolev spaces $W^{n,2}(R^d)$ (see [29] and references therein) and in the spaces of Bessel potentials $H_p^s(R^d)$ (see [13]).

On the other hand, the existent theory for this problem in Hölder spaces, as well as the $W^{n,2}(R^d)$-theory in the degenerate case ($A \ge 0$), is not completely satisfactory. For example, in the latter case the existence of solutions to (0.1) in $W^{1,2}(R^d)$ is known only if $a^{ij} \in C_b^2(\mathbb{R}^d)$ and the remaining coefficients are assumed to have bounded derivatives of the first order (in x). By applying this result to the backward diffusion equation (0.2) one can see that the assumptions needed to make these equations meaningful are substantially stronger than those that ensure the existence of the diffusion process itself. The same applies to the Zakai equation. Indeed, the conditional expectation in the left-hand part of (0.3) is well defined if the coefficients $b, \sigma, \hat{\sigma}$, and h are continuous and bounded and $\sigma^2 + \hat{\sigma}^2 \ge \delta > 0$ (see [9], [30]), while in order to ensure the existence of a $W^{1,2}$-solution to the Zakai equation, one needs to assume additionally that σ and $\hat{\sigma}$ are three times differentiable, b and h are twice differentiable, and all the derivatives are bounded. Even in the superparabolic case, the Zakai equation has a solution in $W^{1,p}(R^d), p \ge 2$, only if σ has bounded derivatives in x (see [13]).

The objective of this article is to study the Cauchy problem (0.1) with nonsmooth coefficients. In particular, we would like to avoid the assumption on differentiability

of the coefficients of the operator $\mathcal{M}$ and the function g and simultaneously relax the superparabolicity assumption ($A^{ij}\xi_i\xi_j \geq \delta|\xi|^2$ for some $\delta > 0$, where $A^{ij} = 2a^{ij} - (\sigma^i, \sigma^j)_Y$).

These goals are hardly achievable within the scope of the standard (variational or semigroup) approaches to SPDEs. This is why we extend the notion of a solution to (0.1) by using the Cameron–Martin–Ito theory of (homogeneous) Wiener chaos (see [1], [16]). In a way, we treat the Wiener process W as an (infinite-dimensional) variable and solve equation (0.1) by separating (t, x) and W. More specifically, we begin with the introduction of a "skeleton" equation associated with (0.1). This equation could be formally obtained from (0.1) by replacing dW_t by $l(t)dt$, where $l(t)$ is a Y-valued deterministic function. Then we define a soft solution of equation (0.1) as an appropriately measurable and integrable function $u(t, x)$ such that its S-transform (Laplace–Wiener transform)

$$S^l u(t, x) = Eu(t, x) \exp\left\{\int_0^t \langle l(s), dW_s\rangle_Y - \frac{1}{2}\int_0^t \|l(s)\|_Y^2 ds\right\}$$

solves the skeleton equation for a sufficiently rich class of functions l.

We prove (see Theorem 1) existence and uniqueness of a soft solution to (0.1) in Sobolev spaces $W^{2,p}$, Hölder spaces $C^{2+\beta}$, and the Lipschitz space L under minimal assumptions on the coefficients and free forces. In particular, in the case of Sobolev spaces, we do not require any differentiability of the coefficients. In the case of Hölder and Lipschitz spaces the coefficients are Hölder and Lipschitz continuous (respectively). In the first two cases the matrix a is assumed to be uniformly nondegenerate and continuous. The superparabolic assumption is not required at any point in our exposition.

Of course, if the coefficients and the free forces are smooth enough, the soft solution coincides with the generalized (variational) solution in the sense of [29]. In the nonsmooth case the soft solution can be approximated by solutions of similar equations with smooth coefficients (Theorem 2).

It turns out (see Theorem 1) that the S-transform is invertible under the same assumptions that guarantee the existence and uniqueness of the soft solution, and the inverse is none other than the stochastic "Feynman–Kac" formula (in the terminology of [29], averaging over characteristic formula) for a solution of (0.1). In particular, this result eliminates the aforementioned gap between the assumptions which ensure the existence of a solution to (0.1) and those needed just to define the Feynman–Kac functional.

The Feynman–Kac formula is not always a convenient representation for a solution of problem (0.1). For example, the whole purpose of the Zakai equation is to provide a convenient recursive way to "compute" the unnormalized filter in the left-hand side of (0.3). It is readily checked that the Feynman–Kac formula for the Zakai equation is equivalent to this functional. So in this case, the Feynman–Kac interpretation of a soft solution does not serve the same purpose as the Zakai equation.

With this in mind, we developed another representation for a soft solution of (0.1) based on the Wiener chaos expansion. Specifically, we prove (Theorem 3) that a soft solution of the Cauchy problem (0.1) admits the following expansion in $L_2(\Omega)$:

$$u(t, x) = \sum_{\alpha\in I} \frac{1}{\sqrt{\alpha!}}\varphi_\alpha(t, x)\xi_\alpha(W), \tag{0.5}$$

where I is the set of all multi-indices of finite length, $\xi_\alpha(W)$ are Wick polynomials

of Wiener integrals $\int_0^t m_k(s)d(W_s, e_j)_Y$, where $\{m_k\}_{k\geq 1}$ and $\{e_j\}_{j\geq 1}$ are complete orthonormal systems in $L_2(0,1)$ and Y, respectively, and $\varphi_\alpha(t,x)$ are the deterministic coefficients in the Cameron–Martin orthogonal decomposition of $u(t,x)$. In section 3 we prove that $\{\varphi_\alpha\}_{\alpha\in I}$ is a solution of a simple recursive parabolic system of Kolmogorov-like deterministic equations (see [24]), referred to below as the S-system.

We prove (Theorem 3) that the S-system has a unique solution in Sobolev, Hölder, and Lipschitz classes and the expansion (0.5) holds under essentially the same minimal assumptions which were used to prove the existence of the soft solution. In some sense, the Wiener chaos expansion (0.5) could serve as another definition of the soft solution.

Of course, using Ito's decomposition theorem [7], one could rewrite the expansion (0.5) in terms of multiple Wiener integrals (Corollary 4). This expansion, while formally equivalent to (0.5), is not as flexible as the former (see [25]).

The Wiener chaos expansion is of special importance in nonlinear filtering. For one, the expansion (0.5) and the associated S-system are well defined and functional in many cases where the Zakai equation for the nonnormalized filtering density is not. Even in the case of smooth coefficients, when the Zakai equation is also well defined, the Wiener chaos expansion has some computational advantages. For example, an important feature of expansion (0.5) is that it separates observations and parameters, in that the Wick polynomials are completely defined by the observations process y_t, while the coefficients $\varphi_\alpha(t,x)$ are determined only by the coefficients of the signal process and the observation function h. This allows one to shift off-line the time-consuming computation related to solving PDEs (for more detail see [20]). We would like to remark that the method of soft solutions has clear-cut limitations. For example, it is not applicable if the coefficients of the equations are functions of W.

We conclude this introduction with some historical remarks.

The idea of using multiple Wiener integral expansions for solving stochastic differential equations was championed by Veretennikov and Krylov [31]; see also Zvonkin and Krylov [33], Isobe and Sato [6], and Léandre and Meyer [19]. (These works address ordinary Ito equations which are equivalent to soft solutions of the backward diffusion equation (0.2)).

Kunita [16] developed a similar representation for the Zakai equation with smooth coefficients; see also a related work by Ocone [26].

The notion of a soft solution was introduced in [24]. Similar constructions have also been used by several authors within the scope of white noise analysis (see, e.g., [28], [5], and references therein). Loosely speaking, the goal in these works was to study an appropriate skeleton equation and to prove that its solution is an S-transform of some element in the space of Hida distributions. Then this element is declared to be a generalized solution of the original equation.

Various particular cases of the Feynman–Kac formula can be found in the literature (see e.g., [5], [12], [15], [28], and the references therein).

The Wiener chaos expansion (0.5) and the associated S-system were first introduced in [23]; see also [24]. For applications of these expansions to nonlinear filtering, see [20] and references therein.

2. Soft solutions of SPDEs. Let $(\Omega, \mathcal{F}, P)$ be a probability space, Y a separable Hilbert space, and W a cylindrical Wiener process in Y. The latter means that we have a family of continuous martingales $W_t(f), f \in Y$, such that

$$\langle W(f), W(g)\rangle_t = t(f,g)_Y \quad \forall f, g \in Y\ .$$

In the future we will assume that $Y = L^2(U, \mathcal{U}, \kappa)$, where $(U, \mathcal{U}, \kappa)$ is a separable

measure space, $\kappa \geq 0$. This assumption is not essential in any way and is made only for the sake of convenience of notation.

Ito's stochastic integral for a stochastic process f in $L^2(U,\mathcal{U},\kappa)$ will be denoted

$$\int_0^t \int_U f(s,\eta) W(ds,d\eta) \ .$$

We denote $H = [0,1] \times R^d$ and suppose that the following measurable functions are given:

$$a : H \to \mathbb{R}^{d^2}, \quad b : H \to \mathbb{R}^d, \quad \sigma : H \times U \to \mathbb{R}^d, \quad c : H \to \mathbb{R},$$
$$h : H \times U \to \mathbb{R}, \quad f : H \to \mathbb{R}, \quad g : H \times U \to \mathbb{R}, \quad \varphi : \mathbb{R}^d \to \mathbb{R}$$

such that

$$|a| + |b| + |c| + \int_U (|\sigma|^2 + h^2) d\kappa \leq K < \infty \ .$$

It is also assumed that f, $\int_U g^2 d\kappa$, and φ are locally integrable.

The main objective of this paper is to study the equation

$$\begin{aligned} du(t,x) &= [\mathcal{L}u(t,x) + f(t,x)]dt \\ &+ \int_U (\mathcal{M}_{t,\eta} u(t,x) + g(t,x,\eta)) W(dt,d\eta), \\ u(0,x) &= \varphi(x), \end{aligned} \tag{1}$$

where $\mathcal{L}u = a^{ij}(t,x) u_{x_i x_j} + b^i(t,x) u_{x_i} + c(t,x)u$ and

$$\mathcal{M}u = \mathcal{M}_{t,\eta} u = \sigma^i(t,x,\eta) u_{x_i} + h(t,x,\eta) u \ .$$

Note that the above assumptions imply that $\mathcal{M}_{t,\eta} v$ is a Hilbert–Schmidt operator from $L^2(U,\mathcal{U},\kappa)$ into $L^2(H)$ for any $v \in W^{1,2}(\mathbb{R}^d)$. The case of a non-Hilbert–Schmidt operator μ is addressed in [25].

Now we shall further specify the structure of the matrix a. For this purpose let us introduce a measurable function $\hat{\sigma} : H \times U \to \mathbb{R}^d$. Everywhere below we assume that

$$\begin{aligned} &a^{ij}(\bar{x}) = \tfrac{1}{2} \int_U (\sigma^i(\bar{x},\eta)\sigma^j(\bar{x},\eta) + \hat{\sigma}^i(\bar{x},\eta)\hat{\sigma}^j(\bar{x},\eta)) d\kappa, \\ &\bar{x} \in H \ . \end{aligned} \tag{P}$$

This assumption asserts that equation (1) is parabolic in the sense of [29].

Obviously the Zakai equation and the backward diffusion equation discussed in the introduction are particular cases of equation (1). In both cases $U = \{1, \dots, d_1\}$ and κ is the counting measure on U, i.e.,

$$\int_A d\kappa = \sum_{i=1}^{d_1} 1_A(i) \ .$$

Let $C^{2+\beta}(H)$, $\beta \in (0,1)$, be the space of all continuous functions v on H having the finite norm

$$|v|_{2,\beta} = \sup_{t,x} |v(t,x)| + |D_x^2 v|_\beta \ ,$$

where

$$|v|_\beta = \sup_{t,x} |v(t,x)| + \sup_{t,x\neq y} \frac{|v(t,x)-v(t,y)|}{|x-y|^\beta} .$$

Denote $W^{2,p}(H), p>1$, the Sobolev class of p-integrable functions v on H having generalized space derivatives up to the second order with the finite norm

$$|v|_{2,p} = |v|_p + |D_x^2 v|_p ,$$

where $|v|_p = (\int_H |v|^p dt\, dx)^{1/p}$.

Denote by $L(H)$ the set of continuous functions v on H with finite norm

$$|v|_L = \sup_{t,x} |v(t,x)| + \sup_{t,x\neq y} \frac{|v(t,x)-v(t,y)|}{|x-y|} .$$

Below we consider three different sets of assumptions on the coefficients of equation (1):

(L) (a) $|f|_L + |b|_L + |c|_L \le K$;

(b)

$$\int_U \left\{|\sigma(\bar{x},\eta)-\sigma(\bar{y},\eta)|^2 + |\hat{\sigma}(\bar{x},\eta)-\hat{\sigma}(\bar{y},\eta)|^2 + |h(\bar{x},\eta)-h(\bar{y},\eta)|^2 + |g(\bar{x},\eta)-g(\bar{y},\eta)|^2\right\} d\kappa \le K|x-y|^2 ,$$
$$\forall \bar{x} = (t,x), \quad \bar{y} = (t,y) \in H ;$$
$$tr\ a + \int_U (|h|+|g|)^2 d\kappa \le K .$$

(C) For $\beta \in (0,1), \delta > 0$

(a) $|f|_\beta + |b|_\beta + |c|_\beta \le K$;

(b) $a^{ij}\xi_i\xi_j \ge \delta|\xi|^2 \qquad \forall \xi \in \mathbb{R}^d$;

(c) $\forall \bar{x} = (t,x), \bar{y} = (t,y) \in H$

$$\int_U \left\{|\sigma(\bar{x},\eta)-\sigma(\bar{y},\eta)|^2 + |\hat{\sigma}(\bar{x},\eta)-\hat{\sigma}(\bar{y},\eta)|^2 + |h(\bar{x},\eta)-h(\bar{y},\eta)|^2 + |g(\bar{x},\eta)-g(\bar{y},\eta)|^2\right\} d\kappa \le K|x-y|^{2\beta} ,$$
$$tr\, a + \int_U (|h|+|g|)^2 d\kappa \le K ;$$

(d) $\varphi \in C^{2,\beta}(\mathbb{R}^d)$.

(W) (a) $\lim_{|x-x'|\to 0} \sup_t \int_U \left\{|\sigma(t,x,\eta)-\sigma(t,x',\eta)|^2 + |\hat{\sigma}(t,x,\eta)-\hat{\sigma}(t,x',\eta)|^2\right\} d\kappa = 0$ and (b) of (C) holds;

(b) for $p > d$, $f \in L^p(H) \cap L^{2p}(H)$, $(\int_U |g|^2 d\kappa)^{1/2} \in L^p(H) \cap L^{2p}(H)$ and $\varphi \in W^{2,p}(\mathbb{R}^d) \cap W^{2,2p}(\mathbb{R}^d)$.

Throughout what follows, we assume that at least one of the assumptions (L), (C), or (W) holds.

Let $\mathcal{F}^W = \mathcal{F}_1^W$ be a P-completion of $\sigma(W_s(f), f \in L^2(U,d\kappa),\ s \in [0,1])$ and $\mathcal{F}_t^W = \mathcal{F}_t$ be a σ-algebra generated by $\cap_{r>t}\sigma(W_s(f), f \in L^2(U,d\kappa), s \in [0,r])$ and negligible sets of $\mathcal{F}^W$. For $l \in L^2([0,1]\times U, dtd\kappa)$, we define $q_t(l)$ as the unique solution of the equation

$$\begin{cases} dq_t(l) = q_t(l) \int_U l(t,\eta) W(dt,d\eta), t \in [0,1], \\ q_0(l) = 1. \end{cases}$$

Let $\mathcal{D}$ be the class of all $l \in L^2([0,1]\times U, dt\,d\kappa)$ such that $\int_U l(t,\eta)^2 d\kappa$ is bounded. Let[1] $\tilde{\mathcal{F}} = \mathcal{B}(H)\otimes\mathcal{F}^W, \tilde{\mathcal{F}}_t = \mathcal{F}_t\otimes\mathcal{B}(\mathbb{R}^d)$. For an $\tilde{\mathcal{F}}$-measurable and $\tilde{\mathcal{F}}_t$-adapted function $u(t,x)$ such that $E|u(t,x)|^2$ is locally bounded, we define its S^l-transform by

$$S^l u(t,x) = Eu(t,x)q_t(l) = Eu(t,x)q_1(l),\ l \in \mathcal{D}\ .$$

Simple (informal) computations yield that $u^l(t,x) = S^l u(t,x)$ verifies the skeleton equation.

$$\begin{cases} \partial_t u^l = \mathcal{L}u^l + f + \displaystyle\int_U l(\mathcal{M}u^l + g)d\kappa, t \in [0,1], \\ u^l(0,x) = \varphi(x). \end{cases} \tag{2}$$

This equation is central for our construction. It will be studied below in Hölder, Sobolev, and Lipschitz classes. The notion of a solution for the former two classes is well known (see, e.g., [3], [18], etc.). $L(H)$-solution, a solution in the Lipschitz class $L(H)$, is understood in the following sense.

DEFINITION 1. *A function $u \in L(H)$ is said to be an $L(H)$-solution of the Cauchy problem*

$$\begin{cases} \partial_t u = \mathcal{L}_u + f \quad in \quad H, \\ u(0,\cdot) = \varphi \end{cases} \tag{3}$$

if for every $y(x) \in C_0^\infty(\mathbb{R}^d)$ and all $t \in [0,1]$,

$$\begin{aligned} \int_{\mathbb{R}^d} u(t,x)y(x)dx = \int_{\mathbb{R}^d} \varphi(x)y(x)dx + \int_0^t \int_{\mathbb{R}^d} \big\{ -a^{ij}(s,x)u_{x_i}(s,x)y_{x_j}(x) \\ +[(b^i(s,x) - a^{ij}_{x_j}(s,x))u_{x_i}(s,x) + c(s,x)u(s,x) + f(s,x)]y(x)\big\}dx\,ds\ . \end{aligned} \tag{4}$$

It is a standard fact that if a function v is Lipschitz continuous on $\mathbb{R}^d$,

$$|v(x) - v(y)| \le C|x-y| \qquad \forall x, y \in \mathbb{R}^d\ ,$$

then v is differentiable almost everywhere (a.e.) on $\mathbb{R}^d$ and $\operatorname{essup}_x|v_x| \le C$. So, the equality (4) is well defined.

DEFINITION 2. *Let (C) (respectively, (W) or (L)) be satisfied. An $\tilde{\mathcal{F}}$-measurable $\mathcal{F}_t$-adapted function u is a $C^{2+\beta}$-soft (respectively, $W^{2,p}$-soft, L-soft) solution for* (1) *if, for each $l \in \mathcal{D}$, $u^l(t,x) = \mathcal{S}^l u(t,x)$ is a solution in the class $C^{2+\beta}(H)$ (respectively, $W^{2,p}(H)$ or $L(H)$) of the equation* (2).

To formulate the main result of this section we shall introduce some additional notation.

Let $\hat{W}$ be a cylindrical Wiener process in $L^2(U,\mathcal{U},\kappa)$ independent of W. Let us fix a number $s \le 1$ and set for $t \in [0,s]$, $W_t^s = W_s - W_{s-t}$ and $\hat{W}_t^s = \hat{W}_s - \hat{W}_{s-t}$.

Let $\mathcal{G} = \mathcal{G}_1^1$ be a P-completion of $\sigma(W_r^s, \hat{W}_r^s, r \in [0,1])$ *and* $\mathcal{G}_t^s$ be a σ-algebra generated by $\cap_{r>u}\sigma(W_v^s, \hat{W}_v^s, v \le r)$ and P-negligible sets of $\mathcal{G}_1^1$. Let $\mathcal{X} = C_{[0,1]}(\mathbb{R}^d)$ be the space of continuous $\mathbb{R}^d$-valued trajectories on $[0,1]$, equipped with the canonical process $X_t = X_t(w) = w(t)$, $w \in \mathcal{X}$, the canonical σ-algebra $\mathcal{C}_1^1 = \sigma(X_s, s \in [0,1])$, and the filtration $\mathcal{C}_t^s = \sigma(X_{s-r}, r \le t)$. On the product space $\bar{\Omega} = \Omega \times \mathcal{X}$ we define

[1]Here and below $\mathcal{B}$ denotes the Borel σ-algebra.

σ-algebras $\bar{\mathcal{G}} = \mathcal{G}_1^1 \otimes \mathcal{C}_1^1$, $\bar{\mathcal{G}}_t^s = \cap_{r>t} \mathcal{G}_r^s \otimes \mathcal{C}_r^s$, $\bar{\mathbf{G}}^s = (\bar{\mathcal{G}}_t^s)$. Denote by $B_{mc}(\bar{\Omega})$ the set of all bounded measurable functions f on $\bar{\Omega}$ such that $f(w,\cdot)$ is continuous on $\mathcal{X}$ for all $w \in \Omega$. Let $M_{mc}^1(\bar{\Omega})$ be the set of probability measures μ on $(\bar{\Omega}, \bar{\mathcal{G}})$ with the property $\mu|_\Omega = P$. This space is assumed to be endowed with the topology generated by the maps

$$\mu \to \mu(f), \mu \in M_{mc}^1(\bar{\Omega}), f \in B_{mc}(\bar{\Omega})$$

(see [8]).

DEFINITION 3. *For $(s,x) \in H$ define*[2]

$$\begin{aligned}\mathcal{S}(s,x,\sigma,\hat{\sigma},b) = \big\{\mu \in M_{mc}^1(\bar{\Omega}) : \quad & \mu\text{-a.e. } X_u = x \quad \forall u \geq s, \\ -dX_t = \int_U (\sigma(t,X_t,\eta) W(\overleftarrow{dt}, d\eta) + \hat{\sigma}(t,X_t,\eta)\hat{W}(\overleftarrow{dt}, d\eta)) & \\ + b(t,X_t)dt, t < s\big\} &\end{aligned}$$

A measure $\mu \in \mathcal{S}(s,x,\sigma,\hat{\sigma},b)$ is called a solution to $(s,x,\sigma,\hat{\sigma},b)$.

Remark 1. Let $\bar{P} \in M_{mc}^1(\bar{\Omega})$. Using the Ito formula one can check easily that $\bar{P} \in \mathcal{S}(s,x,\sigma,\hat{\sigma},b)$ if and only if

$$\begin{aligned} M_t = X_{s-t} - x - \int_{s-t}^s b(r,X_r)dr \in \mathcal{M}_{loc}(\bar{\mathbf{G}}^s, \bar{P}) \ , \\ X_u = x \qquad \forall u \geq s \ , \qquad \bar{P}\text{-a.e.} \\ \langle M, W^s(h) + \hat{W}^s(\hat{h})\rangle_t = \int_{s-t}^s \int_U \{\sigma(r,X_r,\eta)h(\eta) \\ + \hat{\sigma}(r,X_r,\eta)\hat{h}(\eta)\big\} d\kappa\, dr \quad \forall h, \hat{h} \in L^2(E, d\kappa) \end{aligned}$$

and

$$\begin{aligned}\langle M^i, M^j\rangle_t = \int_{s-t}^s \int_U \big\{\sigma^i(r,X_r,\eta)\sigma^j(r,X_r,\eta) \\ + \hat{\sigma}^i(r,X_r,\eta)\hat{\sigma}^j(r,X_r,\eta)\big\}\, d\kappa dr \ .\end{aligned}$$

Indeed, let $N_t = M_t - \int_{s-t}^s \int_U \sigma(r,X_r,\eta) W(\overleftarrow{dr}, d\eta) - \int_{s-t}^s \int_U \hat{\sigma}(r,X_r,\eta)\hat{W}(\overleftarrow{dr}, d\eta)$. Then

$$\begin{aligned}&\langle N^i\rangle_t = \langle M^i\rangle_t - 2\int_{s-t}^s \int_U \sigma^i(r,X_r,\eta)^2 d\kappa dr \\ &-2\int_{s-t}^s \int_U \hat{\sigma}^i(r,X_r,\eta)^2 d\kappa dr + \int_{s-t}^s \int_U \sigma^i(r,X_r,\eta)^2 d\kappa, dr \\ &+ \int_{s-t}^s \int_U \hat{\sigma}^i(r,X_r,\eta)^2 d\kappa dr = 0, i = 1,\dots,d \ .\end{aligned}$$

This proves the remark.

For $l \in \mathcal{D}$, set $B_l(t,x) = b(t,x) + \int_U \sigma(t,x,\eta) l(t,\eta) d\kappa$ and denote $B(t,x) = b(t,x) - \int_U \sigma(t,x,\eta) h(t,x,\eta), d\kappa$.

---

[2]The backward stochastic integral with respect to the cylindrical Wiener process W can be defined by the formula

$$\int_r^t \int_U f(s,\eta) W(\overleftarrow{ds}, d\eta) = \int_{1-t}^{1-r} \int_U f(1-s,\eta) W(ds, d\eta) \ .$$

For details see, e.g., [15] and [29].

THEOREM 1. *Assume (C) (respectively, (W) or (L)). Then there is a $C^{2+\beta}$-soft (respectively, $W^{2,p}$-soft or L-soft) solution u of (1) such that u^l is continuous for each $l \in \mathcal{D}$. For each $(s,x) \in H$ any two soft solutions having this property are P-indistinguishable.*

In addition, there exists a unique measure $P^B_{s,x} \in S(s,x,\sigma,\hat{\sigma},B)$ so that the map $s,x \to P^B_{s,x}$ is continuous and for every s,x,

$$u(s,x) = P^B_{s,x}\left[\int_0^s f(t,X_t)\rho(s,t)dt\right.$$

$$\left.+\varphi(X_0)\rho(s,0) + \int_0^s \int_U g(t,X_t,\eta)\rho(s,t)W(\overleftarrow{dt},d\eta)|\mathcal{F}^W_s\right] \qquad \text{P-a.s.,} \tag{5}$$

where

$$\rho(s,t) = \exp\left\{\int_t^s c(r,X_r)dr + \int_t^s \int_U h(r,X_r,\eta)W(\overleftarrow{dr},d\eta)\right.$$
$$\left.-\tfrac{1}{2}\int_t^s \int_U h(r,X_r,\eta)^2 d\kappa, dr\right\}.$$

Before proceeding with the proof of this statement we shall present some important auxiliary results.

LEMMA 1 (see [4]). *Let $\xi \in L^2(\Omega,\mathcal{F}^W,P)$ and $Eq_1(h)\xi = 0$ for each $h \in L^2([0,1]\times U, dt\, d\kappa)$, where*

$$q_t(h) = \exp\left\{\int_0^t \int_U h(s,\eta)W(ds,d\eta) - \tfrac{1}{2}\int_0^t \int_U h(s,\eta)^2 d\kappa\, ds\right\}.$$

Then $\xi = 0$ P-a.e.; i.e., the linear subspace generated by $q_1(h), h \in L^2([0,1]\times U$, $dt\, d\kappa)$ is dense in $L^2(\Omega,\mathcal{F}^W,P)$.

We will need two "auxiliary" sets of assumptions:

(A) For each $\bar{x} = (t,x), \bar{y} = (t,y) \in H$

$$\int_U \{|\sigma(\bar{x},\eta) - \sigma(\bar{y},\eta)|^2 + |\hat{\sigma}(\bar{x},\eta) - \hat{\sigma}(\bar{y},\eta)|^2\}d\kappa$$
$$+|b(\bar{x}) - b(\bar{y})|^2 \le K|x-y|^2,$$
$$|b| + \int_U (|\sigma| + |\hat{\sigma}|)^2 d\kappa \le K\,.$$

(B) (a) For some $\delta > 0$,

$$a^{ij}\xi_i\xi_j \ge \delta|\xi|^2 \qquad \forall \xi \in \mathbb{R}^d;$$

(b) $\lim_{|x-x'|\to 0}\sup_t \int_U \left\{|\sigma(t,x,\eta) - \sigma(t,x',\eta)|^2 + |\hat{\sigma}(t,x,\eta) - \hat{\sigma}(t,x',\eta)|^2\right\} d\kappa = 0$;

(c) $|a| + |b| \le K$.

LEMMA 2 (see [18], [10], [21], [22]). *Let assumption* (B) *be satisfied and $p > d$. Then there exists a constant $N = N(\delta,K,d,p)$ such that for each $(s,x) \in H$ and $\bar{P} \in S(s,x,\sigma,\hat{\sigma},b)$,*

$$\bar{P}\int_s^1 f(t,X_t)dt \le N|f|_p \quad \textit{for all} \quad f \in L^p(H)\,.$$

PROPOSITION 1. *Let (at least) one of the assumptions* (A) *and* (B) *be satisfied. Then for each* $(s,x) \in H$ *there exists a unique* $P_{s,x} \in S(s,x,\sigma,\hat{\sigma},b)$ *and the map* $(s,x) \to P_{s,x} \in M^1_{mc}(\bar{\Omega})$ *is continuous.*

Proof. 1. Assume (A). In this case the proof can be carried out along familiar lines and we confine ourselves to a short outline.

Iterating and using (A) in a standard way, we find a unique strong solution $X_t = X_t^{s,x}$ of the equation

$$\begin{aligned} -dX_t &= \int_U \left\{ \sigma(t,X_t,\eta) W(\overleftarrow{dt}, d\eta) + \hat{\sigma}(t,X_t,\eta)\hat{W}(\overleftarrow{dt}, d\eta) \right\} \\ &+ b(t,X_t)dt\ , \\ X_t &= x \quad \forall t \geq s\ . \end{aligned}$$

In addition, it is readily checked that $E \sup_t |X_t^{s,x}|^2$ is locally bounded on H.

Now, using Gronwall's inequality, we derive that

$$E \sup_t |X_t^{s,x} - X_t^{s',x'}|^2 \longrightarrow 0 \text{ as } (s',x') \to (s,x)\ . \tag{6}$$

Define $P_{s,x}(d\omega, dw) = \delta_{X^{s,x}_{\cdot}}(dw) P(d\omega) \in M^1_{mc}(\bar{\Omega})$, where δ_w is a Dirac measure on $\mathcal{X}$ at point w. Then obviously $P_{s,x} \in \mathcal{S}(s,x,\sigma,\hat{\sigma},b)$ is a unique solution to $(s,x,\sigma,\hat{\sigma},b)$. The continuity of the map $(s,x) \to P_{s,x}$ follows from (6).

2. Assume (B). First we will prove that a solution of the martingale problem $(s,x,\sigma,\hat{\sigma},b)$ is unique. This follows from Lemma 3 below.

LEMMA 3. *Let* $\bar{P}, \bar{P}' \in \mathcal{S}(s,x,\sigma,\hat{\sigma},b)$. *Then for every partition* $0 \leq s_1 \leq \cdots \leq s_n \leq 1$ *and for each bounded continuous* $g : (\mathbb{R}^d)^n \to \mathbb{R}$,

$$\bar{P}[g(X_{s_1}, \ldots, X_{s_n})|\mathcal{G}] = \bar{P}'[g(X_{s_1}, \ldots, X_{s_n})|\mathcal{G}] \qquad \text{P-a.e.} \tag{7}$$

Proof. The proof will be carried out by induction. Obviously (7) holds if $n = 1$ and $s_1 = 1$.

Let us fix the partition, $0 < s_1 \leq \cdots \leq s_n \leq 1$, and assume that for this partition (**) holds. It is sufficient to prove that for any $h,\ \hat{h} \in L^2([0,1] \times U, dt\, d\kappa)$ and $f \in C_0^\infty(H)$,

$$\begin{aligned} &\bar{P}\left[g(X_{s_1}, \ldots, X_{s_n}) \int_0^{s_1} f(t,X_t) dt q_0(h,\hat{h}) \right] \\ &= \bar{P}'\left[g(X_{s_1}, \ldots, X_{s_n}) \int_0^{s_1} f(t,X_t) dt q_0(h,\hat{h}) \right]\ , \end{aligned} \tag{8}$$

where

$$\begin{aligned} q_t(h,\hat{h}) &= \exp\left\{ \int_t^1 \int_U [hW(ds,d\eta) + \hat{h}\hat{W}(ds,d\eta)] \right. \\ &\left. -\tfrac{1}{2} \int_t^1 \int_U (h^2 + \hat{h}^2) d\kappa\, ds \right\}\ . \end{aligned}$$

Since f and $h, \hat{h}$ are arbitrary, it would follow by Lemma 1 that for each $s_0 \in [0, s_1)$, and $v \in C_0^\infty(\mathbb{R}^d)$,

$$\begin{aligned} &\bar{P}[g(X_{s_1}, \ldots, X_{s_n}) v(X_{s_0})|\mathcal{G}] \\ &= \bar{P}'[g(X_{s_1}, \ldots, X_{s_n}) v(X_{s_0})|\mathcal{G}] \qquad \text{P-a.e.} \end{aligned}$$

Thus for each $\Gamma_0, \dots, \Gamma_n \in \mathcal{B}(\mathbb{R}^d)$, $0 \le s_0 \le \cdots \le s_n \le 1$ we would have

$$\begin{aligned}&\bar{P}[X_{s_0} \in \Gamma_0, \dots, X_{s_n} \in \Gamma_n | \mathcal{G}] \\ &= \bar{P}'[X_{s_0} \in \Gamma_0, \dots, X_{s_n} \in \Gamma_n | \mathcal{G}] \qquad \text{P-a.e.}\end{aligned}$$

In order to prove (8) we take a continuous function $v \in W^{2,d+1}(H)$ solving the Cauchy problem

$$\begin{cases} \partial_t u = \tilde{\mathcal{L}} u + \tilde{\mathcal{M}}_u + f, \\ u(0, \cdot) = 0, \end{cases}$$

where $\tilde{\mathcal{L}} u = a^{ij} u_{x_i x_j} + b^i u_{x_i}$, $\tilde{\mathcal{M}} u = (\int_U (\sigma^i h + \hat{\sigma}^i \hat{h}) d\kappa) u_{x_i}$.

Such a solution exists according to [21] and the Sobolev imbedding theorem. By the Ito formula for backward stochastic differentials (see [29]) and Lemma 2 we have (see [22])

$$\begin{aligned}\bar{P}[g(X_{s_1}, \dots, X_{s_n}) \;\; & v(s_1, X_{s_1}) q_0(h, \hat{h})] \\ &= \bar{P}\left[g(X_{s_1}, \dots, X_{s_n}) \int_0^{s_0} [\tilde{\mathcal{M}} v(t, X_t) + f(t, X_t)] dt q_0(h, \hat{h}) \right] \\ &- \bar{P}\left[g(X_{s_1}, \dots, X_{s_n}) \int_0^{s_1} v_{x_i}(t, X_t) (\sigma^i(t, X_t, \eta) W(\overleftarrow{dt}, d\eta) \right. \\ &\left. + \; \hat{\sigma}^i(t, X_t, \eta) \hat{W}(\overleftarrow{dt}, d\eta)) q_0(h, \hat{h}) \right] \\ &= \bar{P}\left[g(X_{s_1}, \dots, X_{s_n}) \int_0^{s_1} f(t, X_t) dt q_0(h, \hat{h}) \right] .\end{aligned}$$

Then the uniqueness follows.

Let $w \in C_0^\infty(\mathbb{R}^d)$, $w \ge 0$, $\int w \, dx = 1$, $\operatorname{supp} w \subset \{x : |x| < 1\}$, $\epsilon > 0$, $w_\epsilon(x) = \epsilon^d w(x/\epsilon)$, $\sigma_\epsilon(t, x, y) = \sigma(t, \cdot, y) * w_\epsilon(x)$, $\hat{\sigma}_\epsilon(t, x, y) = \hat{\sigma}(t, \cdot, y) * w_\epsilon(x)$, $b_\epsilon(t, x) = b(t, \cdot) * w_\epsilon(x)$.

To prove the existence let us choose $\epsilon_n \to 0$ such that $a_n^{ij}(\bar{x}) \xi_i \xi_i \ge \delta/2 |\xi|^2$, where

$$\begin{aligned}&a_n^{ij}(\bar{x}) = \int_U (\sigma_n^i(\bar{x}, \eta) \sigma_n^j(\bar{x}, \eta) + \hat{\sigma}_n^i(\bar{x}, \eta) \hat{\sigma}_n^j(\bar{x}, \eta)) d\kappa \, , \\ &\sigma_n^i = \sigma_{\epsilon_n}^i, \hat{\sigma}_n^i = \hat{\sigma}_{\epsilon_n}^i, b_n = b_{\epsilon_n}, \quad \text{and} \\ &b_n \to b, \quad \int (|\sigma_n - \sigma|^2 + |\hat{\sigma}_n - \hat{\sigma}|^2) d\kappa \to 0, \\ &dtdx\text{-a.e.}\end{aligned}$$

Let $(s_n, x_n) \to (s, x)$, $\bar{P}^n \in S(s_n, x_n, \sigma_n, \bar{\sigma}_n, b_n)$. In the first part ((A) is satisfied) we proved already that for every n, $\bar{P}^n$ exists and is unique. Obviously, the set of measures $\{\bar{P}^n \circ X^{-1}, n \ge 1\}$ on $\mathcal{X}$ is relatively compact. Thus the set $\{\bar{P}^n, n \ge 1\}$ is also relatively compact in $M_{mc}^1(\bar{\Omega})$. In fact, it is shown in [8] that the former is necessary and sufficient for the latter.

We can assume that $\bar{P}^n \to \bar{P}$ in $M_{mc}^1(\bar{\Omega})$. For $r < t$, let $\bar{f}$ be a $\mathcal{C}_r$-measurable bounded continuous function and $H \in \mathcal{G}_r$. Then

$$\bar{P}^n 1_H \bar{f} (M_t^n - M_r^n) = 0,$$

where $M_t^n = X_t - \int_t^s b^n(r, X_r) dr$.

For each $m, K > 0$

$$\bar{P}^n \int_0^1 |b^n(r, X_r) - b^m(r, X_r)|dr$$
$$\leq N|(b^n - b^m)1_{\{|x|\leq K\}}|_p + C \sup_n \bar{P}^n(\sup_t |X_t| > K) .$$

Thus

$$\bar{P}1_H \bar{f}(M_t - M_s) = 0 .$$

Similarly, using the characterization given in Remark 1 we derive in a standard way that $\bar{P} \in S(s, x, \sigma, \hat{\sigma}, b)$. Since we have already proved that $\bar{P}$ is unique, the statement follows.

Now we shall discuss basic properties of $L(H)$-solutions to (4).

Set $\psi(x) = (1 + |x|^2)^{-d}$, $\hat{b}^j = b^j - [2a^{ij}\psi_{x_i} + a^{ij}_{x_i}]$, $\hat{c} = c - [\mathcal{L}\psi - 2a^{ij}\psi_{x_i}\psi_{x_j}]/\psi$, and $\hat{f} = \psi f$.

LEMMA 4. *Let u be an $L(H)$-solution of the Cauchy problem* (3). *Set $\bar{u}(t, x) = \psi(x)u(t, x)$. Then $\bar{u}(t, x)$ is a continuous map from $[0, 1]$ into $L^2(R^d)$. In addition $\bar{u}(t, x) \in L^2((0, 1); W^{1,2}(\mathbb{R}^d))$ and satisfies the "energy" equality*

$$\int_{\mathbb{R}^d} |\bar{u}(t, x)|^2 dx = \int_{\mathbb{R}^d} |\bar{u}(0, x)|^2 dx + 2 \int_0^t \int_{\mathbb{R}^d} \Big\{ - a^{ij}(s, x)\bar{u}_{x_i}(s, x)\bar{u}_{x_j}(s, x)$$
$$+ \Big[\hat{b}^i(s, x)\bar{u}_{x_i}(s, x) + \hat{c}(s, x)\bar{u}(s, x) + \hat{f}(s, x)\Big] \bar{u}(s, x)\Big\} dx\, ds . \tag{9}$$

Proof. The inclusion $\bar{u} \in W^{1,2}(\mathbb{R}^d)$ is obvious. It is also readily checked that $\bar{u}$ is a $(W^{1,2}(\mathbb{R}^d))$ generalized solution of the Cauchy problem

$$\frac{\partial \bar{u}}{\partial t} = \hat{\mathcal{L}}\bar{u} + \hat{f} \ , \ \bar{u}(0) = \psi\varphi, \tag{10}$$

where $\hat{\mathcal{L}}\bar{u} = (a^{ij}\bar{u}_{x_i})_{x_j} + \hat{b}^i\bar{u}_{x_i} + \hat{c}\bar{u}$. It follows from (4) and (10) that

$$\frac{\partial \bar{u}}{\partial t} \in L^2((0, 1); W^{-1,2}(\mathbb{R}^d)). \tag{11}$$

By the standard imbedding theorem (see, e.g., [18]), (4) and (11) give that there exists a version of function $\bar{u}$ which is continuous from [0,1] into $L^2(\mathbb{R}^d)$. It is well known (see, e.g., [18] or [29]) that if a generalized solution of (10) belongs to $L^2((0, 1); W^{1,2}(\mathbb{R}^d)) \cap C([0, 1]; L^2(\mathbb{R}^d))$, then it satisfies equality (9).

PROPOSITION 2. *For each $\bar{x} = (t, x), \bar{y} = (t, y) \in H$, let*

$$\int_U \left\{|\sigma(\bar{x}, \eta) - \sigma(\bar{y}, \eta)|^2 + |\hat{\sigma}(\bar{x}, \eta) - \hat{\sigma}(\bar{y}, \eta)|^2\right\} d\kappa \leq K|x - y|^2$$

and $|b|_L + |c|_L + \int(|\sigma| + |\hat{\sigma}|)^2 d\kappa \leq K$.

Then for each f, φ such that $|f|_L + |\varphi|_L < \infty$ there exists a unique $u \in L(H)$ solving (3). *Moreover,*

$$u(s, x) = E\left[\int_0^s \exp\left\{\int_t^s c(r, X_r^{s,x})dr\right\} f(t, X_t^{s,x})dt\right.$$
$$\left. + \exp\left\{\int_0^s c(r, X_r^{s,x})dr\right\} \varphi(X_0^{s,x})\right] , \tag{12}$$

where $X_t = X_t^{s,x}$ is a strong solution of the equation

$$\begin{cases} -dX_t &= \int_U (\sigma(t,X_t,\eta)W(\overleftarrow{dt},d\eta) + \hat{\sigma}(t,X_t,\eta)\hat{W}(\overleftarrow{dt},d\eta)) \\ & \quad +b(t,X_t)dt, t<s\ . \\ X_t &= x \quad \forall t \geq s\ . \end{cases}$$

Proof. 1. *The uniqueness.* It is enough to prove that a function $v \in L(H)$ that satisfies (10) with $f = \varphi = 0$ is identically zero.

By (9) we have

$$\|v(t)\|_0^2 = 2\int_0^t \int_{R^d} \left\{-a^{ij}v_{x_i}v_{x_j} + \hat{b}^i v_{x_i} v + \hat{c}v^2\right\} ds\, dx, \tag{13}$$

where $\|\cdot\|_0$ stands for the norm in $L^2(\mathbb{R}^d)$. Integrating by parts we notice that

$$\begin{aligned} 2\int_{R^d} \hat{b}^i v_{x_i} v dx = -\int (b^i - 2a^{ij}\psi_{x_j}\psi^{-1})_{x_i} v^2 dx \\ +2\int a^{ij}_{x_j} v_{x_i} v dx\ . \end{aligned} \tag{14}$$

Owing to the boundedness of the coefficients a, b, c and their first derivatives (for $a.a.x$), we derive from (13) and (14) that for all t,

$$\begin{aligned} \|v(t)\|_0^2 \leq 2\int_0^t \int_{R^d} \left\{-a^{ij}v_{x_i}v_{x_j} + a^{ij}_{x_j} v_{x_i} v\right\} dx\ ds \\ +c\int_0^t \|v(s)\|_0^2 ds\ . \end{aligned} \tag{15}$$

Again integrating by parts we get

$$\begin{aligned} \int_{R^d}\int_U (\sigma^i(\eta)\sigma^j(\eta))_{x_j} v_{x_i} v d\kappa dx = \int_{R^d}\int_U \sigma^i(\eta)\sigma^j_{x_j}(\eta) v_{x_i} v d\kappa dx \\ + \tfrac{1}{2}\int_{R^d}\int_U \sigma^i_{x_j}(\eta)\sigma^j(\eta)(v^2)_{x_i} dx = \int_{R^d}\int_U \sigma^i(\eta)\sigma^j_{x_j}(\eta) v_{x_i} v d\kappa dx \\ - \tfrac{1}{2}\int_{R^d}\int_U \sigma^i_{x_j}(\eta)\sigma^j_{x_i}(\eta) v^2 d\kappa dx + \tfrac{1}{2}\int_{R^d}\int_U \sigma^i_{x_i}(\eta)\sigma^j_{x_j}(\eta) v^2 d\kappa dx \\ + \int_{R^d}\int_U \sigma^i_{x_i}(\eta)\sigma^j(\eta) v v_{x_j} d\kappa dx = I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Obviously,

$$I_2 + I_3 \leq c\|v\|_0^2\ . \tag{16}$$

Applying the Hölder inequality and then the elementary inequality $2ab \leq a^2 + b^2$ we obtain

$$I_1 + I_4 \leq \int_{R^d}\int_U |\sigma^i(\eta) v_{x_i}|^2 d\kappa dx + c\|v\|_0^2\ . \tag{17}$$

Estimates (16) and (17) yield

$$\begin{aligned} \int_0^t \int_{R^d}\int_U (\sigma^i(\eta)\sigma^j(\eta))_{x_j} v_{x_i} v d\kappa dx ds \leq \int_0^t \int_{R^d}\int_U |\sigma^i(\eta) v_{x_i}|^2 d\kappa dx ds \\ +c\int_0^t \|v\|_0^2 ds\ . \end{aligned} \tag{18}$$

Similar arguments lead to the estimate

$$\int_0^t \int_{R^d} \int_U (\hat\sigma^i(\eta)\hat\sigma^j(\eta))_{x_j} v_{x_i} v d\kappa dx ds \le \int_0^t \int_{R^d} \int_U |\hat\sigma^i(\eta) v_{x_i}|^2 d\kappa dx ds + c\int_0^t \|v\|_0^2 ds\,. \tag{19}$$

Combining (18) and (19) we arrive at

$$2\int_0^t \int_{R^d} a^{ij}_{x_j} v_{x_i} v dx \le \int_0^t \int_{R^d} \int_U (|\sigma^i(\eta) v_{x_i}|^2 + |\hat\sigma^i(\eta) v_{x_i}|^2 + |\hat\sigma^i(\eta) v_{x_i}|^2) d\kappa dx ds + c\int_0^t \|v\|_0^2 ds = 2\int_0^t \int_{R^d} a^{ij} v_{x_i} v_{x_j} dx ds + c\int_0^t \|v\|_0^2 ds\,.$$

The latter inequality together with (15) gives

$$\|v(t)\|_0^2 \le c\int_0^t \|v(s)\|_0^2 ds\,,$$

and thus by the Gronwall lemma $\|v(t)\|_0 \equiv 0$.

2. *The existence.* Let $w \in C_0^\infty(\mathbb{R}^d)$, $w \ge 0$, $\mathrm{supp} w \subset \{x : |x| < 1\}$, $\int w dx = 1$, $w_\epsilon(x) = \epsilon^{-d} w(x/\epsilon)$. Define $b_\epsilon(t,\cdot) = b(t,\cdot) * w_\epsilon$, $c_\epsilon(t,\cdot) = c(t,\cdot) * w_\epsilon$, $\sigma_\epsilon(t,\cdot,\eta) = \sigma(t,\cdot,\eta) * w_\epsilon$, $\hat\sigma_\epsilon(t,\cdot,\eta) = \hat\sigma(t,\cdot,\eta) * w_\epsilon$, $f_\epsilon(t,\cdot) = f(t,\cdot) * w_\epsilon$, $\varphi_\epsilon = \varphi * w_\epsilon$. Consider the Cauchy problem

$$\begin{cases} -\partial_t u + \mathcal{L}^\epsilon u + f_\epsilon = 0 \text{ in } H, \\ u(0,\cdot) = \varphi_\epsilon, \end{cases} \tag{20}$$

where $\mathcal{L}^\epsilon$ is obtained by $\mathcal{L}$ by substituting $\sigma_\epsilon, \hat\sigma_\epsilon, b_\epsilon, c_\epsilon$ for $\sigma, \hat\sigma, b, c$. Then according to [29], we can find a unique smooth (with respect to x) solution u^ϵ to (20). For $(s,x) \in H$ denote by $X_t = X_t^{\epsilon,s,x}$ a solution to

$$\begin{cases} -dX_t &= \displaystyle\int_U (\sigma_\epsilon(t, X_t, \eta) W(\overleftarrow{dt}, d\eta) + \hat\sigma(t, X_t, \eta)\hat W(\overleftarrow{dt}, d\eta)) \\ &+ b(t, X_t) dt,\ t < s, \\ X_t &= x, \quad \forall t \ge s. \end{cases}$$

By the Ito formula we have

$$u^\epsilon(s,x) = E\left[\int_0^s \exp\left\{\int_t^s c_\epsilon(r, X_r^{\epsilon,s,x}) dr\right\} f_\epsilon(t, X_t^{\epsilon,s,x}) dt + \exp\left\{\int_0^s c_\epsilon(r, X_r^{\epsilon,s,x} dr\right\} \varphi_\epsilon(X_0^{\epsilon,s,x})\right]. \tag{21}$$

Now it is easy to see that

$$|u^\epsilon(s,x) - u^\epsilon(s',x))| \le C|s-s'|^{\frac{1}{2}} \tag{22}$$

$$|u^\epsilon(s,x) - u^\epsilon(s,x'))| \le C|x-x'|\,, \tag{23}$$

where the constant C does not depend on ϵ.

Thus we can find a sequence $\epsilon_k \downarrow 0$ such that $(u^k)(u^{\epsilon_k})$ converges to $u \in L(H)$ uniformly on compact subsets of H and (23) holds for u as well. So

$$|u_x^\epsilon| \le C, |u_x| \le C \ .$$

Since $u^k = u^{\epsilon_k}$ is a classical solution of (20) for every $y \in C_0^\infty(\mathbb{R}^d)$ we can rewrite (4) as

$$\int u^k(t,x)y(x)dx = \int \varphi_{\epsilon_k}(x)y(x)dx + \int_0^t \int \Big\{ - a^{ij}_{\epsilon_k}(s,x)u^k_{x_i}(s,x)y_{x_j}(x)$$

$$+[(b^i_{\epsilon_k}(s,x) - a^{ij}_{\epsilon_{k_{x_j}}}(s,x))u^k_{x_i}(s,x) + c_{\epsilon_k}(s,x)u^k(s,x) + f_{\epsilon_k}(s,x)]y(x)\Big\}dxds \ , \tag{24}$$

where

$$a^{ij}_{\epsilon_k}(s,x) = \tfrac{1}{2}\int_u (\sigma^i_{\epsilon_k}(s,x,\eta)\sigma^j_{\epsilon_k}(s,x,\eta) + \hat\sigma^i_{\epsilon_k}(s,x,\eta)\hat\sigma^j_{\epsilon_k}(s,x,\eta))d\kappa \ .$$

In order to pass to the limit as $h \to \infty$ we notice that for a bounded function $g(s,x)$ on H taking values in a Hilbert space V and any compact $K \subset \mathbb{R}^d$

$$\int_0^1 \int_K \|g_\epsilon(s,x) - g(s,x)\|^2 ds\, dx \overset{\epsilon \to 0}{\longrightarrow} 0 \ , \tag{25}$$

where $g_\epsilon(s,\cdot) = g(s,\cdot) * w_\epsilon$.

Also, if g_k, g are measurable functions on H such that $\sup_k |g_k| + \sup_k |g_k| \le C$ and for each compact K

$$\int_0^1 \int_K |g_k(s,x) - g(s,x)| \overset{k \to \infty}{\longrightarrow} 0 \ ,$$

then for every $\varphi \in C_0^\infty(\mathbb{R}^d)$,

$$\int_0^1 \int g_k(s,x)u^k_{x_l}(s,x)\varphi(x)dxds \overset{k \to \infty}{\longrightarrow} \int_0^1 \int g(s,x)u_{x_l}(s,x)\varphi(x)dxds \ . \tag{26}$$

Indeed, for each m

$$\overline{\lim}_k \left\{ \left| \int_0^1 |g_k(s,x)u^k_{x_l}(s,x)\varphi(x)ds\, dx - \int_0^1 \int g_m(s,x)u^k_{x_l}(s,x)\varphi(x)ds\, dx \right| \right.$$

$$\left. + \left| \int_0^1 \int g_k(s,x)u_{x_l}(s,x)\varphi(x)ds\, dx - \int_0^1 \int g_m(s,x)u_{x_l}(s,x)\varphi(x)d, dx \right| \right\}$$

$$\le c(m) \quad \text{and} \quad c(m) \overset{m \to \infty}{\longrightarrow} 0 \ .$$

On the other hand, for each m

$$\lim_k \int_0^1 \int g_m(s,x)u^k_{x_l}(s,x)\varphi(x)ds\, dx = -\lim_k \int_0^1 \int (g_m(s,x)\varphi(x))_{x_l}u^k(s,x)ds\, dx$$

$$= \int_0^1 \int (g_m(s,x)\varphi(x))_{x_l}u(s,x)ds\, dx = \int_0^1 \int g_m(s,x)\varphi(x)u_{x_l}(s,x)ds\, dx$$

and (26) follows.

Applying (25) for $V = L^2(U,\mathcal{U},d\kappa)$, $g = \sigma, \hat{\sigma}, \sigma_{x_l}, \hat{\sigma}_{x_l}$, we derive that for a compact $K \subset \mathbb{R}^d$

$$\int_0^1 \int_K (|a^{ij}_{\epsilon_k}(s,x) - a^{ij}(s,x)| + |a^{ij}_{\epsilon_k x_l}(s,x) - a^{ij}_{x_l}(s,x)|)ds\,dx \overset{k\to\infty}{\longrightarrow} 0\,.$$

Using the same argument for $V = \mathbb{R}$,

$$\int_0^1 \int_K (|c_{\epsilon_k}(s,x) - c(s,x)| + |\varphi_{\epsilon_k}(x) - \varphi(x)| + |f_{\epsilon_k}(s,x) - f(s,x)| \\ + |b^i_{\epsilon_k}(s,x) - b(s,x)|)ds\,dx \overset{k\to\infty}{\longrightarrow} 0\,.$$

Now using (26) we pass to the limit, as $k \to \infty$, in (24) and arrive at (4) for u, and (12) follows from (21). This completes the proof.

Proof of Theorem 1. 1. *The uniqueness.* The uniqueness is an immediate consequence of Lemma 1 and the corresponding uniqueness theorem for (2). For the class $L(H)$ the latter result follows from Proposition 1. Uniqueness of (2) in the classes $C^{2+\beta}(H)$ and $W^{2,p}(H)$ is well known; see, e.g., [3], [18], [21], [22], and a continuous version of a $W^{2,p}(H)$-solution exists by the Sobolev imbedding theorem.

Remark 2. Note that in the standard references (e.g., [3] or [18]) some smoothness of the coefficients of (3) in t is assumed. In fact this assumption is not necessary (see [21], [22]).

2. *The existence.* Obviously the assumption (L) implies (A) while (B) follows from either of the assumptions (C) or (W). This of course remains true if b in (A) and (B) is replaced by B or B^l for any $l \in \mathcal{D}$. So by Proposition 1 there exists a unique measure $P^B_{s,x} \in \mathcal{S}(s,x,\sigma,\hat{\sigma},B)$ and the map $(s,x) \to P^B_{s,x}$ is continuous. The right-hand part of (5) is well defined and we denote it by $v(s,x)$. Then $E|v(s,x)|^2$ is locally bounded by the Hölder inequality (and Lemma 2 if (W) is assumed). Fix $l \in \mathcal{D}$ and set $v^l(s,x) = Ev(s,x)q_s(l)$. Simple computation based on Girsanov's theorem shows that

$$v^l(s,x) = P^{B_l}_{s,x}\left[\int_0^s \exp\left\{\int_t^s \tilde{c}(r,X_r)dr\right\}(f(t,X_t) \right. \\ \left. + \int_U l(t,\eta)g(t,X_t,\eta)d\kappa)dt + \varphi(X_0)\exp\left\{\int_0^s \tilde{c}(r,X_r)dr\right\}\right], \tag{27}$$

where $\tilde{c}(t,x) = c(t,x) + \int_U l(t,\eta)h(t,x,\eta)d\kappa$ and $P^{B_l}_{s,x} \in \mathcal{S}(s,x,\sigma,\hat{\sigma},B^l)$.

Since the right-hand side of (27) solves equation (2) in the corresponding class and is continuous in (s,x) (see [3], [18], [21], [22] for $C^{2+\beta}(H)$, $W^{2,p}(H)$, and Proposition 1 for $L(H)$), we are done.

The following corollary is an obvious implication of (5) and the uniqueness property of soft solutions.

Corollary 1. *If $g = 0$, $f \geq 0$, and $\varphi \geq 0$ then the S-solution of equation* (1) *is nonnegative.*

Remark 3. Let the assumption (L) be satisfied. Then we can rewrite (5) as

$$u(s,x) = E\left[\int_0^s f(t,X^{s,x}_t)\tilde{\rho}(s,t)dt + \varphi(X^{s,x}_0)\tilde{\rho}(0,s) \right. \\ \left. + \int_0^s \int_U g(t,X^{s,x}_t,\eta)\tilde{\rho}(s,t)W(\overleftarrow{dt},d\eta)|\mathcal{F}^W_s\right], \tag{28}$$

where $X^{s,x}$ is a strong solution of the stochastic differential equation (SDE)

$$\begin{cases} -dX_t &= \int_U (\sigma(t,X_t,\eta)W(\overleftarrow{dt},d\eta)+\hat\sigma(t,X_t,\eta)\hat W(\overleftarrow{dt},d\eta)) \\ &+ B(t,X_t)dt, \quad t<s, \\ X_t &= x \quad \forall t\ge s\ , \end{cases}$$

and

$$\tilde\rho(s,t)=\exp\left\{\int_t^s c(r,X_r^{s,x})dr+\int_t^s\int_U h(r,X_r^{s,x},\eta)W(\overleftarrow{dr},d\eta)\right.$$
$$\left.-\tfrac12\int_t^s h(r,X_r^{s,x},\eta)^2 d\kappa\, dr\right\}\ .$$

Moreover, a simple computation shows that for each $p\ge 2$ there exists C such that

$$E|u(t,x)-u(t,y)|^p\le C|x-y|^p\ ,$$
$$E|u(t,x)-u(t',x)|^p\le C|t-t'|^{p/2}\ .$$

Now we will show that an S-solution can be obtained as a limit of strong solutions.

Let $w,\bar w\in C_0^\infty(\mathbb{R}^d), w\ge 0$, $0\le\bar w\le 1$, $\int w\,dx=1$, $\operatorname{supp} w\subset\{x:|x|<1\}$, and $\bar w(x)=1$ if $|x|\le 1$. Define $w_\epsilon(x)=\epsilon^{-d}w(x/\epsilon)$, $\bar w_\epsilon(x)=w(\epsilon x)$, $\sigma_\epsilon(t,\cdot,\eta)=\sigma(t,\cdot,\eta)*w_\epsilon, \hat\sigma_\epsilon(t,\cdot,\eta)=\hat\sigma(t,x\cdot,\eta)*w_\epsilon$, $b_\epsilon(t,\cdot)=b(t,\cdot)*w_\epsilon$, and $c_\epsilon(t,\cdot)=c(t,\cdot)*w_\epsilon$, $h_\epsilon(t,\cdot,\eta)=h(t,\cdot,\eta)*w_\epsilon$.

As before denote

$$\mathcal{L}^\epsilon u=a_\epsilon^{ij}u_{x_ix_j}+b_\epsilon^i u_{x_i}+c_\epsilon u\ ,$$
$$\mathcal{M}^\epsilon u=\mathcal{M}_\eta^\epsilon u=\sigma_\epsilon^i u_{x_i}+h_\epsilon u\ ,$$

where $a_\epsilon^{ij}=\frac12\int_U(\sigma_\epsilon^i\sigma_\epsilon^j+\hat\sigma_\epsilon^i\hat\sigma_\epsilon^j)d\kappa$.

Let $f^\epsilon(t,\cdot)=f(t,\cdot)*w_\epsilon\bar w_\epsilon, g^\epsilon(t,\cdot,\eta)=g(t,\cdot,\eta)*w_\epsilon\bar w_\epsilon$, $\varphi^\epsilon=\varphi*w_\epsilon\bar w_\epsilon$.

Consider the equation

$$\text{(29)}\quad\begin{cases} du(t,x)=(\mathcal{L}^\epsilon u(t,x)+f^\epsilon(t,x))dt+\int_U(\mathcal{M}_\eta^\epsilon u(t,x)+g^\epsilon(t,x))W(dt,d\eta), \\ u(0,\cdot)=\varphi^\epsilon\ . \end{cases}$$

If one of the conditions (C), (W), or (L) holds, then according to [29] (at this point the specific construction of approximations is used) there exists a unique classical solution $u=u^\epsilon$ to (29) and the following representation holds:

$$\text{(30)}\quad\begin{aligned} u^\epsilon(s,x)=E\Bigg[&\int_0^s f^\epsilon(t,X_t^{\epsilon,s,x})\rho_\epsilon(s,t)dt+\varphi^\epsilon(X_0^{\epsilon,s,x})\rho_\epsilon(0,s) \\ &+\int_0^s\int_U g^\epsilon(t,X_t^{\epsilon,s,x},\eta)\rho_\epsilon(s,t)W(\overleftarrow{dt},d\eta)|\mathcal{F}_s^W\Bigg]\ , \end{aligned}$$

where $X=X_t^{\epsilon,s,x}$ is a solution of

$$\begin{cases} -dX_t &= \int_U(\sigma_\epsilon(t,X_t,\eta)W(\overleftarrow{dt},d\eta)+\hat\sigma_\epsilon(t,X_t,\eta)\hat W(\overleftarrow{dt},d\eta)) \\ &+ B_\epsilon(t,X_t)dt, \\ X_t &= x \quad \forall t\ge s\ , \end{cases}$$

and

$$B_\epsilon = b_\epsilon - \int_U \sigma_\epsilon h_\epsilon d\kappa, \rho_\epsilon(s,t) = \exp\left\{\int_t^s c_\epsilon(r, X_r^{\epsilon,s,x})dr \right.$$
$$\left. + \int_t^s \int_U h_\epsilon(r, X_r^{\epsilon,s,x}, \eta) W(\overleftarrow{dr}, d\eta) - \tfrac{1}{2}\int_s^t \int_U h_\epsilon(r, X_r^{\epsilon,s,x}, \eta)^2 d\kappa dr \right\} .$$

THEOREM 2. (a) *Let (L) be satisfied and u be an L-soft solution of* (1). *Then for each $(s,x) \in H$, $u^\epsilon(s,x) \to u(s,x)$ in $L^2(\Omega, \mathcal{F}^W, P)$.*

(b) *Let (C) (respectively, (W)) be satisfied and u be a $C^{2+\beta}$-soft (respectively, $W^{2,p}$-soft) solution to* (1). *Then for each $(s,x) \in H$, $u^\epsilon(s,x) \to u(s,x)$ weakly in $L^2(\Omega, \mathcal{F}^W, P)$.*

Proof. Denote

$$\begin{aligned} F_\epsilon &= \int_0^s f_\epsilon(t, X_t^{\epsilon,s,x})\rho_\epsilon(s,t)dt \ , \\ G_\epsilon &= \int_0^s \int_U g^\epsilon(t, X_t^{\epsilon,s,x}, \eta)\rho_\epsilon(s,t) W(\overleftarrow{dt}, d\eta) \ , \\ \Phi_\epsilon &= \varphi^\epsilon(X_0^{\epsilon,s,x})\rho_\epsilon(0,s). \end{aligned}$$

Assume that f, $\int_E g^2 d\kappa, \varphi, D\varphi, D^2\varphi$ are bounded. Then for each $p \geq 1$

$$\sup_\epsilon E\left[|F_\epsilon|^p + |G_\epsilon|^p + |\Phi_\epsilon|^p\right] < \infty. \tag{31}$$

Let (L) be satisfied and $X_t = X_t^{s,t}$ be a solution of

$$\begin{cases} -dX_t = \displaystyle\int_U (\sigma(t, X_t, \eta) W(\overleftarrow{dt}, d\eta) + \hat{\sigma}(t, X_t, \eta)\hat{W}(\overleftarrow{dt}, d\eta)) \\ +B(t, X_t)dt, \\ X_t = x \qquad \forall\, t \geq s. \end{cases}$$

Then

$$E \sup_t |X_t^{\epsilon,s,x} - X_t^{s,x}|^2 \to 0 \ .$$

It is readily checked that for each $p \geq 1$

$$\sup_\epsilon E \sup_t \rho_\epsilon(s,t)^p < \infty \ .$$

Thus for $p \geq 1$,

$$E \sup_t |\rho_\epsilon(s,x) - \tilde{\rho}(s,x)|^p \overset{\epsilon \to}{\to} 0,$$

where $\tilde{\rho}(s,t) =$

$$= \exp\left\{\int_t^s c(r, X_r^{s,t})dr + \int_t^s \int_U h(r, X_r^{s,t}, \eta) W(\overleftarrow{dr}, d\eta) \right.$$
$$\left. -\tfrac{1}{2}\int_t^s \int_U h(r, X_r^{s,x}, \eta)^2 d\kappa, dr \right\}.$$

So for each $p \geq 1$ by (31)

$$E(|F_\epsilon - F|^p + |G_\epsilon - G|^p + |\Phi_\epsilon - \epsilon|^p) \overset{\epsilon \to}{\to} 0 \ ,$$

where

$$F = \int_0^s f(t, X_t^{s,x})\tilde{\rho}(s,t)dt\ ,$$
$$G = \int_0^s \int_U g(t, X_t^{s,x}, \eta)\tilde{\rho}(s,t)W(\overleftarrow{dt}, d\eta)\ ,$$
$$\Phi = \varphi(X_0^{s,x})\tilde{\rho}(0,s)\ .$$

Then $E[F_\epsilon + G_\epsilon + \Phi_\epsilon|\mathcal{F}_s^W] \to E[F + G + \Phi|\mathcal{F}_s^W]$ in $L^p(\Omega, \mathcal{F}^W, P)$. Thus passing to the limit in (30), we get (28) and the statement is proved in the (L) case.

Let (C) or (W) be satisfied and $f, \int g^2 d\kappa, \varphi, D\varphi, D^2\varphi$ be bounded. According to Lemma 1, there is a complete orthonormal system (CONS) in $L^2(\Omega, \mathcal{F}^W, P)$ consisting of linear combinations of $q_1(l), l \in \mathcal{D}$. Since $\sup_\epsilon E(|u^\epsilon(s,x)|^2 + |u(s,x)|^2) < \infty$, then in order to prove the weak convergence in $L^2(\Omega, \mathcal{F}^W, P)$ it is sufficient to demonstrate that

$$v_\epsilon^l(s,x) = Eu^\epsilon(s,x)q_1(l) \to Eu(s,x)q_1(l) = v^l(s,x)$$
$$\forall l \in \mathcal{D}\ .$$

Let $P_l^\epsilon \in \mathcal{S}(s,x,\sigma_\epsilon,\hat{\sigma}_\epsilon,B_l^\epsilon)$ and $P_l \in \mathcal{S}(s,x,\sigma,\hat{\sigma},B_l)$, where $B_l^\epsilon = b_\epsilon + \int_U \sigma_\epsilon l d\kappa$, $B_l = b + \int_U \sigma l d\kappa$. Then

$$v_\epsilon^l(s,x) = P_l^\epsilon \left[\int_0^s \exp\left\{\int_t^s \tilde{c}_\epsilon(r, X_r)dr\right\}(f^\epsilon(t, X_t)\right.$$
$$\left. + \int_U l(t,\eta)g^\epsilon(t, X_t, \eta)d\kappa)dt + \varphi^\epsilon(X_0)\exp\left\{\int_0^s \tilde{c}_\epsilon(t, X_t)dt\right\}\right]\ , \tag{32}$$

where $\tilde{c}_\epsilon = c_\epsilon + \int l h_\epsilon d\kappa$.

A similar formula holds for $v^l(s,x)$ if we omit ϵ everywhere.

Indeed, owing to Lemma 2 we can pass to the limit in (32). Indeed, let $\epsilon_n \to 0$. Then by [8] the set $\{P_l^{\epsilon_n}, n \geq 1\}$ is relatively compact. Assume that for some subsequence n_k, $P_l^{\epsilon_{n_k}} = P^k \to P_l$ as $k \to \infty$. Let $\tilde{c}_k = \tilde{c}_{\epsilon_{n_k}}, f^k = f^{\epsilon_{n_k}}$. Then for each m

$$A = P^k \int_0^s \exp\left\{\int_t^s \tilde{c}_k(r, X_r)dr\right\} f^k(t, X_t)dt$$
$$= P^k \int_0^s \exp\left\{\int_t^s \tilde{c}_m(r, X_r)dr\right\} f^m(t, X_t)dt$$
$$+ P^k \left[\int_0^s \exp\left\{\int_t^s \tilde{c}_k(r, X_r)dr\right\} f^k(t, X_t)dt\right.$$
$$\left. - \int_0^s \exp\left\{\int_t^s \tilde{c}_m(r, X_r)dr\right\} f^m(t, X_t)dt\right] = I_1 + I_2.$$

Now there exists a constant C such that

$$|I_2| \leq CP^k\left[\int_0^1 |\tilde{c}_k(r, X_r) - \tilde{c}_m(r, X_r)|dr\right.$$
$$\left. + \int_0^1 |f^k(r, X_r) - f^m(r, X_r)|dr\right] = I_{21} + I_{22}.$$

If (W) is satisfied, $I_{22} \le C'|f^k - f^m|_p \overset{k,m\to\infty}{\to} 0$. On the other hand, for each L

$$\begin{aligned} I_{21} &\le \bar{C}\Big(\sup_k P^k(\sup_t |X_t| > L) \\ &+P^k \int_0^1 |\tilde{c}_k(r, X_r) - \tilde{c}_m(r, X_r)|\bar{w}(X_r)dr \Big) \\ &\le \bar{c}\sup_k P^k(\sup_t |X_t| > L) \\ &+N|(\tilde{c}_k - \tilde{c}_m)\bar{w}_L|_p. \end{aligned}$$

Thus for each m, L,

$$\begin{aligned} B &= \lim_k \left| A - P_l \int_0^s \exp\left\{ \int_t^s \tilde{c}(r, X_r)dr \right\} f(t, X_t)dt \right| \\ &\le 2C'|f - f^m|_p + 2N|(\tilde{c} - \tilde{c}_m)\bar{w}_L|_p + C\sup_k P^k(\sup_t |X_t| \\ &> L) + P_l(\sup_t |X_t| > L) \ . \end{aligned}$$

Also $\overline{\lim_L}\ \overline{\lim_m} B = 0$. This proves that

$$\lim_k A = P_l \int_0^s \exp\left\{ \int_t^s \tilde{c}(r, X_r)dr \right\} f(t, X_t)dt \ .$$

Similarly

$$\begin{aligned} &P^k \left[\int_0^s \int_U l(t, \eta) g^{\epsilon_{n_k}}(t, X_t, \eta)d\kappa dt \right. \\ &\left. +\varphi^{\epsilon_{n_k}}(X_0) \exp\left\{ \int_0^s \tilde{c}_k(t, X_t)dt \right\} \right] \\ &\to P_l \left[\int_0^s \int_U l(t, \eta) g(t, X_t, \eta)d\kappa dt \right. \\ &\left. +\varphi(X_0) \exp\left\{ \int_0^s \tilde{c}_k(t, X_t)dt \right\} \right] \ . \end{aligned}$$

This completes the proof of the weak convergence in $L^2(\Omega, \mathcal{F}^W, P)$ given that either one of (C) and (W) is satisfied and f, $\int g^2 d\kappa, \varphi, D\varphi, D^2\varphi$ are bounded.

Now we drop the latter assumption and assume (W) in its original form. Define $f_n = f1_{\{|f|\le n\}}$, $g_n = g1_{\{\int g^2 d\kappa \le n\}}$, and take a sequence $\varphi_n \in C_0^\infty(\mathbb{R}^d)$ converging to φ in $W^{2,p}(\mathbb{R}^d)$. Let $u^{n,\epsilon}$ be a solution to problem (1) with f, g, φ replaced by f_n, g_n, φ_n. Then by Lemma 2

$$\lim_{n\to\infty} \sup_\epsilon E|u^{n,\epsilon}(s, x) - u^\epsilon(s, x)|^2 = 0 \ .$$

Let u^n be a solution of (1) with f, g, φ replaced by f_n, g_n, φ_n. Then we know already that for each n, $u^{n,\epsilon} \to u^n$ weakly in $L^2(\Omega, \mathcal{F}^W, P)$.

By Lemma 3

$$\lim_{n\to\infty} E|u^n(s, x) - u(s, x)|^2 = 0 \ .$$

3. Chaos expansions of soft solutions. Let us fix an arbitrary complete orthonormal system (m_k) in $L^2([0, 1] \times U, dtd\kappa)$ such that $m_k \in \mathcal{D} \quad \forall k \ge 1$. Let us also introduce random variables

$$\xi_k = \int_0^1 \int_U m_k(s, \eta) W(ds, d\eta)$$

and $H_n(\xi_k)$, where H_n is the nth Hermite polynomial[3] defined by

$$H_n(x) = (-1)^n \left(\frac{d^n}{dx^n} e^{-\frac{x^2}{2}} \right) e^{\frac{x^2}{2}}.$$

Let $\alpha = (\alpha_k)$ be an infinite multi-index, i.e., $\alpha_k \in \mathbb{N} = \{0, 1, 2, \ldots\}$, $k \geq 1$. We shall consider only such α that $|\alpha| = \sum_k \alpha_k < \infty$; i.e., only a finite number of α_k is not zero, and we denote by $\mathcal{J}$ the set of all such multi-indices. If $\alpha = (\alpha_k) \in \mathcal{J}$, the number $\alpha! = \Pi_k \alpha_k$ is well defined. Let $\xi_\alpha = \Pi_k H_{\alpha_k}(\xi_k)/\alpha!$, $\alpha = (\alpha_k) \in \mathcal{J}$. Now we recall the celebrated Cameron–Martin result.

THEOREM 3 (see [1], [4]). *The set $\{\xi_\alpha, \alpha \in \mathcal{J}\}$ is a CONS in $L^2(\Omega, \mathcal{F}^W, P)$; i.e., for each $\xi \in L^2(\Omega, \mathcal{F}^W, P)$,*

$$\xi = \sum_{\alpha \in \mathcal{J}} E[\xi \xi_\alpha] \xi_\alpha$$

and

$$E\xi^2 = \sum_{\alpha \in \mathcal{J}} (E[\xi \xi_\alpha])^2 \, .$$

Let $\mathcal{Z}$ be the set of real-valued sequences $z = (z_k)_{k \geq 1}$ such that only a finite number of z_k is not zero. For $\alpha = (\alpha_k) \in \mathcal{J}$, set $\partial^\alpha = \partial_z^\alpha = \Pi_k \frac{\partial^{\alpha_k}}{(\partial z_k)^{\alpha_k}}$ and $m = m_z = m(s, \eta, z) = \sum_k z_k m_k$ (obviously $m \in \mathcal{D}$). Set $p_t(z) = q_t(m_z)$.

COROLLARY 2. *If $\xi \in L^2(\Omega, \mathcal{F}_s^W, P)$ for some $s \in [0, 1]$, then*

(a) $\xi = \sum_{\alpha \in \mathcal{J}} \frac{1}{\sqrt{\alpha!}} \partial_z^\alpha E[\xi p_s(z)]|_{z=0} \xi_\alpha(s)$, *where* $\xi_\alpha(s) = \partial_z^\alpha p_s(z)|_{z=0}/\sqrt{\alpha!} = \Pi_k H_{\alpha_k}$ $(\int_0^s \int_U m_k(r, \eta) W(dr, d\eta))$;

(b) $E\xi^2 = \sum_{\alpha \in \mathcal{J}} \frac{1}{\alpha!} \left(\partial_z^\alpha E[\xi p_s(z)]|_{z=0} \right)^2$.

Proof. For $s = 1$, both (a) and (b) follow straightforwardly from Theorem 2 and the well-known formula

$$\xi_\alpha = \partial_z^\alpha p_1(z)|_{z=0}/\sqrt{\alpha!}$$

(see, e.g., [4]). Since $p_s(z)$ is an $\mathcal{F}_s^W$-martingale, it is readily checked that the general statement is a simple implication of the above special case $s = 1$.

Remark 4. If u is a soft solution to (1), we have by (2) the following equation for $u^z = u^{m_z} = S^{m_z} u$:

$$\begin{cases} \partial_t u^z = \mathcal{L} u^z + f + \displaystyle\int_U m_z (\mathcal{M} u^z + g) d\kappa, \\ u^z(0, \cdot) = \varphi \, . \end{cases} \tag{33}$$

Our objective is to expand a soft solution of (1) with respect to (ξ_α). In order to determine the coefficients of this expansion, we consider the recursive system of PDEs:

$$\begin{cases} \partial_t \varphi^\alpha = \mathcal{L} \varphi^\alpha + f 1_{\{|\alpha|=0\}} + \displaystyle\sum_k \alpha_k \int_U m_k (\mathcal{M} \varphi^{\alpha(k)} + g 1_{\{|\alpha|=1\}}) d\kappa, \\ \varphi^\alpha(0, x) = \varphi(x) 1_{\{|\alpha|=0\}}, \end{cases} \tag{34}$$

[3]The random variable $H_n(\xi_k)$ is often referred to as the Wick product (polynomial) and denoted $:(\xi_k)^n:$. This term and notation originated in physical literature (for more detail see [4], etc.).

where $\alpha = (\alpha_i) \in \mathcal{J}$ and by $\alpha(k)$ we denote a sequence $(\tilde{\alpha}_i)$ such that $\tilde{\alpha}_i = \alpha_i$ if $i \neq k$, and $\tilde{\alpha}_k = \max\{\alpha_k - 1, 0\}$. In other words one obtains the sequence $\alpha(k)$ by substituting $\max\{\alpha_k - 1, 0\}$ for α_k in α.

THEOREM 4. *Let (W) (respectively, (C), (L)) be satisfied. Then*

(a) *There exists a unique solution $\varphi^\alpha, \alpha \in \mathcal{J}$ of system (33) such that $\varphi^\alpha \in W^{2,p}(H) \cap W^{2,2p}(H)$ (respectively, $u^\alpha \in C^{2+\beta}(H)$, $L(H)$ for each $\alpha \in \mathcal{J}$).*

(b) *For each (s,x), the soft solution of (1) allows the Wiener chaos expansion:*

$$u(s,x) = \sum_{\alpha \in \mathcal{J}} u^\alpha(s,x)\xi_\alpha = \sum_{\alpha \in \mathcal{J}} u^\alpha(s,x)\xi_\alpha(s),$$

where $u^\alpha = \frac{1}{\sqrt{\alpha!}}\varphi^\alpha$.

Proof. Assume (L). In this case, the uniqueness follows from Proposition 1 by induction, and we only have to prove the existence.

Let u be a soft solution of (1) given by (5) or (28). We claim that $\varphi^\alpha = \partial_z^\alpha S^{m_z} u|_{z=0} = E[u\partial_z^\alpha p_1(z)|_{z=0}]$, $\alpha \in \mathcal{J}$, is a solution to (34). By Remark 3, $\varphi^\alpha \in L(H)$, $\alpha \in \mathcal{J}$. On the other hand, by definition (see Definition 1) (33) means that for each $\psi \in C_0^\infty(\mathbb{R}^d)$ and $t \in [0,s]$,

$$\begin{aligned}
&\int u^z(t,x)\psi(x)dx - \int \varphi(x)\psi(x)dx \\
&\quad = \int_0^t \int a^{ij}(r,x)u^z_{x_i}(r,x)\psi_{x_j}(x)dr\,dx \\
&\quad - \int_0^t \int a^{ij}_{x_j}(r,x)u^z_{x_i}(r,x)\psi(x)dx + \int_0^t \int c(r,x)u^z(r,x)\psi(x)dx\,dr \\
&\quad + \int_0^t \int b^i(r,x)u^z_{x_i}(r,x)\psi(x)dx\,dr + \int_0^t \int \int_U m_z(r,\eta)(\mathcal{M}u^z(r,x) \\
&\qquad + g(r,x,\eta))\psi(x)dx\,d\kappa\,dr\,.
\end{aligned} \tag{35}$$

Since according to Remark 3 we have Lipschitz estimates, it is possible to differentiate (35) with respect to z. Indeed, let (η_k) be a CONS in $L^2(\Omega, \mathcal{F}^W, P)$. Then $u(t,x) = \sum_k f^k(t,x)\eta_k$, $f^k(t,x) = E[u(t,x)\eta_k]$. Let

$$e_i = (\underbrace{0, \dots, 0}_{i-1}, 1, 0, \dots, 0).$$

Then by Remark 3 for each t, x, and positive h

$$\begin{aligned}
&\sum_k (f^k(t,x+he_i) - f^k(t,x))^2 = E|u(t,x+he_i) \\
&-u(t,x)|^2 \leq Ch^2.
\end{aligned}$$

Thus for each t, $f^k_{x_i}$ exists dx-a.e., is bounded, and coincides with the partial generalized derivative of f^k. Moreover, for each t

$$\lim_{h \to 0} \frac{f^k(t,x+he_i) - f^k(t,x)}{h} \to f^k_{x_i}(t,x)$$

dx-a.e. Let $u^n(t,x) = \sum_1^n f^k(t,x)\eta_k$. Then for each t dx-a.e.

$$\sum_1^n f^k_{x_i}(t,x)^2 = \lim_{h \to 0} \sum_1^n \frac{(f^k(t,x+he_i) - f^k(t,x))^2}{h^2} \leq C\,.$$

So for each t, dx-a.e.

$$\sum_1^\infty f^k_{x_i}(t,x)^2 \le C\ .$$

Therefore, for each t we have the following equalities dx-a.e. in $L^2(\Omega,\mathcal{F}^W,P)$:

$$\begin{aligned}&\lim_{h\to 0}\tfrac{u(t,x+he_i)-u(t,x)}{h}=\textstyle\sum_i f^k_{x_i}(t,x)\eta_k\\&=\lim_{n\to\infty}u^n_{x_i}(t,x);\end{aligned}$$

i.e., $u_{x_i}(t,x)$ exists in $L^2(\Omega,\mathcal{F}^W,P)$ for each t dx-a.e. and coincides with the generalized partial derivative. Moreover, $E|u_{x_i}(t,x)|^2\le K$ for each $t\,dx$-a.e. Thus

$$\begin{aligned}&u^z_{x_i}(t,x)=E[u_{x_i}(t,x)p_t(z)]\ ,\\&\partial^\alpha_z u^z_{x_i}(t,x)|_{z=0}=E[u_{x_i}(t,x)\partial^\alpha_z p_t(z)|_{z=0}]\ ,\end{aligned}$$

and

$$\varphi^\alpha_{x_i}(t,x)=E[u_{x_i}(t,x)\partial^\alpha_z p_t(z)|_{z=0}]$$

for each t dx-a.e.

It is readily checked now that by applying the differentiation operator $\partial^\alpha_z[\cdot]|_{z=0}$ to both sides of (35) we arrive at an integral equation equivalent to (34). This of course completes the proof of (a) in the class $L(H)$.

Since $u^\alpha(s,x)=\frac{1}{\sqrt{\alpha!}}\partial^\alpha_z E[u(s,x)p_s(z)]|_{z=0}$, part (b) of Theorem 4 follows from Corollary 2.

2. Now assume (W) or (C). In this case uniqueness follows by the same arguments as in 1.

Proof of the existence is also analogous to the one in 1. Only two steps require special justification. Specifically, one has to demonstrate that in this case, too, $\varphi^\alpha=\partial^\alpha_z S^{m_z}u|_{z=0}$, and also $\varphi^\alpha\in W^{2,p}(H)\cap W^{2,2p}(H)$ (respectively, $\varphi^\alpha\in C^{2+\beta}(H)$) for each $\alpha\in\mathcal{J}$. These two issues are addressed below.

If (W) or (C) is satisfied, by Remark 4 we have that u^z satisfies equation (33). Let $\mathcal{Z}_N=\{z=(z_k)\in\mathcal{Z}: z_k=0 \text{ if } k>N\}$, $\mathcal{J}_N=\{\alpha=(\alpha_k)\in\mathcal{J}:\alpha_k=0 \text{ if } k>N\}$, $N\ge 1$. Assume that for $|\alpha|=n$, $\partial^\alpha_z u^z=\partial^\alpha_z S^{m_z}u$ satisfies

$$\text{(36)}\qquad\begin{cases}\partial_t\partial^\alpha_z u^z=\mathcal{L}\partial^\alpha_z u^z+f1_{\{\alpha=0\}}\\+\displaystyle\int_U m_z\mathcal{M}\partial^\alpha_z u^z d\kappa+\int_U\alpha_k m_k(\mathcal{M}\partial^{\alpha(k)}_z u^z\\+g1_{\{|\alpha|=1\}})d\kappa\\\partial^\alpha u^z(0,x)=\varphi 1_{\{|\alpha|=0\}}\ .\end{cases}$$

If (W) is satisfied, we assume that

$$|\partial^\alpha_z u^z|_{2,p}\le C(N,R,|\alpha|)(|f|_p+|g|_p+|\varphi|_{2,p})\ ,$$

where $|\alpha|\le n$, $z\in\mathcal{Z}_N$, and $|z|\le R$, $\alpha\in\mathcal{J}_N$, $|g|_p=(\int_0^1\int_{\mathbb{R}^d}(\int_E g^2 d\kappa)^{\frac{p}{2}}dxdt)^{\frac{1}{p}}$.

If (C) is satisfied, we assume that

$$|\partial^\alpha_z u^z|_{2,\beta}\le C(N,R,\alpha)(|f|_\beta+|g|_\beta+|\varphi|_{2,\beta})\ ,$$

where $|\alpha| \le n$, $\alpha \in \mathcal{J}_N$, $z \in \mathcal{Z}_N$, $|z| \le R$,

$$|g|_\beta = \sup_{t,x} \left(\int |g|^2 d\kappa \right)^{\frac{1}{2}} + \sup_{t,x \ne y} \frac{\left(\int |g(t,x,\eta) - g(t,y,\eta)|^2 d\kappa \right)^{\frac{1}{2}}}{|x-y|^\beta} .$$

Let us fix $\alpha \in \mathcal{J}_N$ such that $\alpha = \alpha' + \gamma$, $|\alpha'| = n$, $|\gamma| = 1$, $\gamma = (\gamma_k)$, $\gamma_l \ne 0$.

So for $\partial_z^{\alpha'} u^z$ equation (36) is satisfied according to our assumption. Then we take $\epsilon > 0$ and consider

$$\Delta_\epsilon^{\alpha',\gamma} u^z = (\partial^{\alpha'} u^{z+\epsilon e_l} - \partial^{\alpha'} u^z)/\epsilon ,$$

where the lth component of e_l is 1 and the remaining components are zeros. Then using assumed estimates we can pass to the limit and obtain the equation (36) for $\partial_z^{\alpha'+\gamma} u^z$ and similar estimates for $|\alpha'| = n+1$. Thus our statement follows by induction if we start from $\alpha = 0$. If $|\alpha| = 0$, (36) follows from [21].

Remark 5. If (L) is assumed, the existence of $L(H)$-solutions of (34) can be proved in a more analytic way by approximating smoothly the "free forces" and the coefficients.

The rest of this section is dedicated to the derivation of a multiple Wiener integral version of the Wiener chaos expansion of Theorem 4.

First, we will demonstrate that a solution of the S-system (34) can be "explicitly solved" (represented as an integral of superpositions of the operators $\mathcal{M}_{s,\eta}$ and semigroups associated with the operator $\mathcal{L}$).

Denote by $T_{s,t} f$ the solution of the problem

$$\begin{cases} -\partial_t u + \mathcal{L} u = 0, \quad s \le t, \\ u(s,\cdot) = f. \end{cases}$$

Remark 6. If (C) or (W) are satisfied, then for each $(s,x) \in H$ there exists a unique $P_{s,x} \in \mathcal{S}(s,x,\sigma,\hat{\sigma},b)$ and by [22] for $f \in C^{2+\beta}$ or $W^{2,p}$, respectively,

$$T_{s,t} f(x) = P_{t,x} \exp\left\{ \int_s^t c(u, X_u) du \right\} f(X_s).$$

To each multi-index $\alpha = (\alpha_k) \in \mathcal{J}$ of length n (i.e., $|\alpha| = n$) we relate a set K_α. The elements of K_α are positive integers $k_i, i = 1, \dots, n$ such that each k is represented there by α_k-copies. This of course implies that if $\alpha_k = 0$ then k is not included in K_α. Let $\mathcal{P}^n$ be a permutation group of the set $\{1, \dots, n\}$, $s^n = (s_1, \dots, s_n)$, $\eta^n = (\eta_1, \dots, \eta_n)$, $ds^n = ds_1 \cdots ds_n, d\kappa^n = \kappa(d\eta_1) \cdots \kappa(d\eta_n)$, $U^n = U \times \cdots \times U$, n times.

PROPOSITION 3. *Let (C) or (W) be satisfied. Then for each multi-index $|\alpha| = n$, we have*

$$\begin{aligned} \varphi^\alpha(s,x) = \sum_{\sigma \in \mathcal{P}^n} \int_0^s \int_0^{s_n} \cdots \int_0^{s_2} \int_{U^n} T_{s_n,s} M_{s_n,\eta_n} \cdots \\ \cdots T_{s_1,s_2} (M_{s_1,\eta_1} \varphi^0(s_1,x) + g(s_1,x,\eta_1)) m_{k_{\sigma(n)}}(s_n,\eta_n) \cdots \\ \cdots m_{k_{\sigma(1)}}(s_1,\eta_1) ds^n d\kappa^n \qquad \text{if } n \ge 2 , \end{aligned} \tag{37}$$

and

$$\begin{aligned} \varphi^\alpha(s,x) = \int_0^s \int_U T_{t,s} m_{k_1}(t,\eta) (M_{t,\eta} \varphi^0(t,\eta) \\ + g(t,x,\eta)) dt\, d\kappa \quad \text{if } n = 1, \quad \text{where } K_\alpha = \{k_1, \dots, k_n\} . \end{aligned}$$

Proof. To begin with we have to check that all the integrals in (37) are well defined. This problem is addressed in Lemma 5 and Corollary 3 below.

LEMMA 5.

(a) *Let* (C) *be satisfied. Then there exists* C *such that for each* $s < t \leq 1$, $v \in C^{\beta}(\mathbb{R}^d)$.

$$\int_s^t \int_U T_{s,u}(\mathcal{M}_{u,\eta} T_{u,t} v)^2 d\kappa \, du \leq C T_{s,t} v^2 \, . \tag{38}$$

(b) *Let* (W) *be satisfied. Then there exists* C *such that for each* $s < t \leq 1$, $v \in L^p(\mathbb{R}^d) \cap L^{2p}(\mathbb{R}^d)$, *the estimate* (38) *holds.*

Proof. (a) Let $v \in C^{2,\beta}(\mathbb{R}^d)$. Then by [21] $l = T_{s,t} v \in C^{2+\beta}(H_t)$, where $H_t = [0,t] \times \mathbb{R}^d$. Fix $(s,x) \in H_t$, $s < t$. Let $P_{s,x} \in \mathcal{S}(s,x,\sigma,\hat{\sigma},b)$. Then $l^2 \in C^{2+\beta}(H_t)$. The needed estimate with some C independent of v follows readily from Remark 6 and the Ito formula for $l^2(t,X_t)\exp\{\int_s^t 2c(r,X_r)dr\}$. Using standard estimates for the fundamental solutions of parabolic equations (see [18, Ch. 1V, section 13], and [3]) we derive (38) for $v \in C^{\beta}(\mathbb{R}^d)$ by passing to the limit.

(b) Let $v \in W^{2,p}(\mathbb{R}^d) \cap W^{2,2p}(\mathbb{R}^d)$. Then $l = T_{s,t} v \in W^{2,p}(H_t) \cap W^{2,2p}(H_t)$ (see [21], [22]), and $l^2 \in W^{2,p}(H_t)$ $H_t = [0,t] \times \mathbb{R}^d$. Fix $(s,x) \in H_t$, $s < t$. Let $P_{s,x} \in \mathcal{S}(s,x,\sigma,\hat{\sigma},b)$. Inequality (37) with C independent of v follows by Remark 6 and the Ito formula for v^2 (see [22]). On the other hand, as shown in [31], there exists a constant $N_{s,t}$ such that $\sup_x |T_{s,t} f| \leq N_{s,t} |f|_p$. Thus we can derive (38) for general v by passing to the limit.

COROLLARY 3. *Let* (C) *or* (W) *be satisfied. Then for each* $(s,x) \in H$, $n > 2$.

$$\begin{aligned}
&\int_0^s \int_0^{s_n} \cdots \int_0^{s_2} \int_{U^n} \Big[T_{s_n,s} M_{s_n,\eta_n} \cdots T_{s_1,s_2} (M_{s_1,\eta_1} \varphi^0(s_1,x) \\
&+ g(s_1,x,\eta_1)) \Big]^2 ds_1 \cdots d_{s_n} \kappa(d\eta_1) \cdots \kappa(d\eta_n) \\
&+ \int_s^1 \int_U \Big[T_{t,s} (M_{t,\eta} \varphi^0(t,x) + g(t,x,\eta) \Big]^2 dt \kappa(d\eta) < \infty \, .
\end{aligned}$$

Proof. Applying Lemma 5 and using induction arguments one can see that it is sufficient to prove the inequality

$$\int_0^s T_{s,t} (M_{t,\eta} \varphi^0(t,x) + g(t,x))^2 dt \, d\kappa < \infty \, .$$

Obviously

$$\varphi^0(s,x) = T_{0,s} \varphi(x) + \int_0^s T_{s,t} f(t,x) dt \, ,$$

and the statement follows by Lemma 5 and the estimates in [21].

Now we can proceed with the derivation of formula (36).

By (34) we have

$$\begin{aligned}
\varphi^{\alpha}(s,x) = \int_s^1 \int_E \sum_k \alpha_k m_k(t,\eta) T_{s,t} (M_{t,\eta} \varphi^{\alpha(k)}(t,x) \\
+ g(t,x,\eta) 1_{|\alpha|=1}) dt \, d\kappa,
\end{aligned} \tag{39}$$

and the representation formula is obviously true for $|\alpha| = 1$. Assume that it is true for $|\alpha| = n$. Let $|\alpha| = n+1$, $K_\alpha = \{k_1, \dots, k_{n+1}\}$. If $\mathcal{P}_j^{n+1}$ is a permutation group of the set $\{1, \dots, n+1\} \setminus \{j\}$, it follows from (39) that

$$\varphi^\alpha(s,x) = \int_0^s \int_U \sum_{j=1}^{n+1} m_{k_j}(t,\eta) T_{t,s} \mathcal{M}_{t,\eta} \varphi^{\alpha(k_j)}(t,x) dt\, d\kappa$$

$$= \int_0^s \int_U \sum_{j=1}^{n+1} m_{k_j}(t,\eta) T_{t,s} \mathcal{M}_{t,\eta} \sum_{\sigma \in \mathcal{P}_j^{n+1}} \int_0^t \int_0^{s_{n+1}} \dots \int_0^{s_{j+1}} \int_0^{s_j} \dots$$

$$\int_0^{s_2} \int_{U^n} T_{s_{n+1},t} M_{s_{n+1},\eta_{n+1}} \cdots T_{s_{j-1},s_{j+1}} \mathcal{M}_{s_{j-1},\eta_{j-1}} \cdots T_{s_1,s_2}(g(s_1,x,\eta_1)$$

$$+ \mathcal{M}_{s_1,\eta_1} \varphi^0(s_1,x)) ds_j^{n+1} dt d\kappa_j^{n+1} ,$$

$$ds_j^{n+1} = ds_1 \cdots ds_{j-1} ds_{j+1} \cdots ds_{n+1} ,\ d\kappa_j^{n+1}$$
$$= \kappa(d\eta_1) \cdots \kappa(d\eta_{j-1}) \kappa(d\eta_{j+1}) \cdots \kappa(d\eta_{n+1}) .$$

By Corollary 3 all the integrals are well defined. Thus our statement follows by induction.

Remark 7. For $\alpha \in \mathcal{J}$, $|\alpha| = n$, $K_\alpha = \{k_1, \dots, k_n\}$, define

$$e_\alpha = \sum_{\sigma \in \mathcal{P}^n} m_{k_{\sigma(1)}} \otimes \cdots \otimes m_{k_{\sigma(n)}} / \sqrt{\alpha! |\alpha|!} .$$

Then we can rewrite (39) in the following way:

$$\varphi^\alpha(s,x) = \sqrt{\alpha! |\alpha|!} \int_0^s \int_0^{s_n} \dots \int_0^{s_2} \int_{U^n} T_{s_n,s} \mathcal{M}_{s_n,\eta_n} \dots$$

$$\dots T_{s_2,s_1}(\mathcal{M}_{s_1,\eta_1} \varphi^0(s_1,x) + g(s_1,x,\eta_1)) e_\alpha(s^n,\eta^n) ds^n d\kappa^n$$

$$= \sqrt{\alpha!} / \sqrt{|\alpha|!} \int_{([0,1]\times U)^n} G^n(s^n,\eta^n) ds^n d\kappa^n ,$$

where

$$G^n(s^n,\eta^n) = \textstyle\sum_{\sigma \in \mathcal{P}^n} T_{s_{\sigma(n)},s} \mathcal{M}_{s_{\sigma(n)},\eta_{\sigma(n)}} \dots$$

$$\dots T_{s_{\sigma(1)},s_{\sigma(2)}}(\mathcal{M}_{s_{\sigma(1)},\eta_{\sigma(1)}} \varphi^0(s_{\sigma(1)},x)$$

$$+ g(s_{\sigma(1)},x,\eta_{\sigma(1)}) 1_{\{s_{\sigma(1)} > \dots > s_{\sigma(n)} > s\}} .$$

The last equality here follows immediately from the fact that $(e_\alpha)_{|\alpha|=n}$ form a CONS for the symmetric part of $L^2(([0,1]\times U)^n, ds^n d\kappa^n)$.

COROLLARY 4. *For each $\alpha \in \mathcal{J}$ such that $|\alpha| = n$,*

$$\sum_{|\alpha|=n} \frac{1}{\sqrt{\alpha!}} \varphi^\alpha(s,x)\xi_\alpha = \int_0^s \int_0^{s_n} \dots \int_0^{s_2} \int_{U^n} T_{s_n,s} \mathcal{M}_{s_n,\eta_n} \cdots$$

$$T_{s_1,s_2}(\mathcal{M}_{s_1,\eta_1}\varphi^0(s_1,x) + g(s_1,x,\eta_1))W(ds_n,d\eta_n)\cdots W(ds_1,d\eta_1),$$

$$\sum_{|\alpha|=n} \frac{1}{\alpha!} \varphi^\alpha(s,x)^2 = \int_0^s \int_0^{s_n} \dots \int_0^{s_2} \int_{U^n} \{T_{s_n,s}\mathcal{M}_{s_n,\eta_n} \dots$$

$$\dots T_{s_1,s_2}(\mathcal{M}_{s_1,\eta_1}\varphi^0(s_1,x) + g(s_1,x,\eta_1))\}^2 \, ds^n d\kappa^n \quad \text{if } n > 2,$$

and

$$\sum_{|\alpha|=1} \frac{1}{\sqrt{\alpha!}} \varphi^\alpha(s,x)\xi_\alpha = \int_0^s \int_U T_{t,s}(\mathcal{M}_{t,\eta}\varphi^0(t,x) + g(t,x,\eta))W(dt,d\eta),$$

$$\sum_{|\alpha|=1} \frac{1}{\alpha!} \varphi^\alpha(s,x)^2 = \int_0^s \int_U T_{t,s}(\mathcal{M}_{t,\eta}\varphi^0(t,x) + g(t,x,\eta))^2 dt\, d\kappa\,.$$

Proof. It follows from Theorem 3.1 in [7] that for $|\alpha| = n$,

$$\xi_\alpha/\sqrt{\alpha!} = \int_0^1 \int_0^{s_n} \dots \int_0^{s_2} \int_{U^n} e_\alpha(s^n,\eta^n) W(ds_1,d\eta_1)\cdots W(ds_n,d\eta_n)$$

(e_α was defined in Remark 7). Since G^n is a symmetric function on $([0,1]\times U)^n$ we have the $L^2(([0,1]\times U)^n, ds^n d\kappa^n)$ expansion for G^n:

$$\begin{aligned} G^n &= \sum_{|\alpha|=n} e_\alpha \int_{([0,1]\times U)^n} G^n(s^n,\eta^n) e_\alpha(s^n,\eta_n) ds^n d\kappa^n \\ &= \sum_{|\alpha|=n} c_\alpha e_\alpha\,. \end{aligned}$$

Thus,

$$\sum_{|\alpha|=n} \frac{\varphi_\alpha(s,x)}{\sqrt{\alpha!}} \xi_\alpha = \sum_{|\alpha|=n} c_\alpha \int_0^1 \int_0^{s_n} \dots \int_0^{s_2} e_\alpha(s^n,\eta^n) W(ds_n,d\eta_n)\dots$$

$$\dots W(ds_1,d\eta_1) = \int_0^s \int_0^{s_n} \int_0^{s_i} \int_{U^n} (T_{s_n,s}\mathcal{M}_{s_n,\eta_n} \dots$$

$$\dots T_{s_1,s_2}(\mathcal{M}_{s_1,\eta_1}\varphi^0(s_1,x) + g(s_1,x,\eta_1)))W(ds_1,d\eta_1)\dots W(ds_n,d\eta_n)$$

The equality for $\sum_{|\alpha|=n} \frac{\varphi_\alpha(s,x)^2}{\alpha!}$ follows from the latter in a simple way.

REFERENCES

[1] R. H. Cameron and W. T. Martin, *The orthogonal development of non-linear functionals in a series of Fourier–Hermite functions*, Ann. Math., 48 (1947), pp. 385–392.

[2] G. DaPrato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[3] S. D. Eidelman, *Parabolic Systems*, Nauka, Moscow, 1964.

[4] T. Hida, H. H. Kuo, J. Potthoff, and L. Streit, *White Noise*, Kluwer Academic Publishers, Dordrecht, 1993.

[5] H. Holden, B. Oksendal, J. Uboe, and T. Zhang, *Stochastic Partial Differential Equations. A Modeling, White Noise Functional Approach*, Birkhäuser, Basel, 1996.

[6] E. Isobe and Sh. Sato, *Wiener–Hermite expansion of a process generated by an Ito stochastic differential equation*, J. Appl. Prob., 20 (1983), pp. 754–765.

[7] K. Ito, *Multiple Wiener integral*, J. Math. Soc. Japan, 3 (1951), pp. 157–169.

[8] J. Jacod and J. Memin, *Sur un type de convergence intermediaire entre la convergence en loi et la convergence in probabilite*, Lecture Notes in Math. 850, Springer-Verlag, Berlin, 1981.

[9] N. V. Krylov, *On Ito's stochastic integral equations*, Theory Probab. Appl., 14 (1969), pp. 330–336.

[10] N. V. Krylov, *Some estimates of the probability density of a stochastic integral*, Math. USSR Izv., 8 (1974), pp. 233–254.

[11] N. V. Krylov and B. L. Rozovskii, *On the First Integrals and Liouville Equations for Diffusion Processes*, Lecture Notes in Control and Inform. 36, Springer, 1981, pp. 117–125.

[12] N. V. Krylov and B. L. Rozovskii, *On characteristics of degenerate second order parabolic equations*, J. Soviet Math., 32 (1986), pp. 226–248.

[13] N. V. Krylov, *On L_p theory of stochastic partial differential equations*, SIAM J. Math. Anal., 27 (1996), pp. 313–340.

[14] N. V. Krylov and B. L. Rozovskii, *On the Cauchy problem for linear stochastic partial differential equations*, Math. USSR Izvestija, 11 (1977), pp. 1267–1284.

[15] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK, 1978.

[16] H. Kunita, *Cauchy problem for stochastic partial differential equations arising in non-linear filtering theory*, Systems Control Lett., 1 (1981), pp. 37–41.

[17] H. Kunita, *On backward stochastic differential equations*, Stochastics, 6 (1982), pp. 293–313.

[18] O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Uraltseva, *Linear and Quasilinear Equations of Parabolic Type*, Translations of Mathematical Monographs 23, American Mathematical Society, 1968.

[19] R. Leandre and P. A. Meyer, *Sur le development d'une diffusion en chaos de Wiener*, Sém Prob. XXII, Lecture Notes in Math. 1372, Springer, 1989, pp. 161–164.

[20] S. Lototsky, R. Mikulevicius, and B. L. Rozovskii, *Nonlinear filtering revisited: A spectral approach*, SIAM J. Optim. Control, 35 (1997), pp. 435–461.

[21] R. Mikulevicius and H. Pragarauskas, *On the Cauchy problem for certain integro-differential operators in Sobolev and Hoelder spaces*, Lithuanian Math. J., 32 (1992), pp. 238–264.

[22] R. Mikulevicius and H. Pragarauskas, *On the martingale problem associated with nondegenerate Levy operators*, Lithuanian Math. J., 32 (1992), pp. 297–311.

[23] R. Mikulevicius and B. L. Rozovskii, *Separation of observations and parameters in nonlinear filtering*, in Proc. 2nd IEEE Conference on Decision and Control, San Antonio, TX, IEEE Control System Society, 1993, pp. 1565–1559.

[24] R. Mikulevicius and B. L. Rozovskii, *Soft solutions of linear parabolic SPDE's and the Wiener Chaos expansion*, in Stochastic Analysis on Infinite Dimensional Spaces, H. Kunita and H. H. Kuo, eds., Pitman Research Notes in Mathematics 310, Longman Press, London, 1994.

[25] D. Nualart and B. L. Rozovskii, *Weighted Wiener Chaos and linear SPDE's driven by a space-time white noise*, J. Funct. Anal., 49 (1997), pp. 200–225.

[26] D. Ocone, *Multiple integral expansions for nonlinear filtering*, Stochastic, 10 (1983), pp. 1–30.

[27] E. Pardoux, *Equations aux dérivees partielles stochastiques non linéaries monotones. Etude de solutions de type Ito: Thése.*, Université de Paris Sud., Orsay, 1975.

[28] J. Potthoff, G. Vage, and H. Watanabe, *Generalized Solutions of Linear Parabolic Stochastic Partial Differential Equations*, Preprint 210/96, Mannheim University, 1996.

[29] B. L. Rozovskii, *Stochastic Evolution Systems*, Kluwer Academic Publishers, Norwell, MA, 1990.

[30] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, New York, 1979.
[31] A. Yu. Veretennikov and N. V. Krylov, *On explicit formulas for solutions of stochastic differential equations*, Math. USSR Sbornik, 29 (1976), pp. 229–256.
[32] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrsch., 11 (1969), pp. 230–243.
[33] A. K. Zvonkin and N. V. Krylov, *Strong solutions of stochastic differential equations*, in Proc. of the School-Seminar on the Theory of Random Processes, Druskininkai, Vilnius, 1975.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 481–510, March 1998

010

CONSTRUCTION OF MULTISCALING FUNCTIONS WITH APPROXIMATION AND SYMMETRY*

G. PLONKA† AND V. STRELA‡

Abstract. This paper presents a new and efficient way to create multiscaling functions with given approximation order, regularity, symmetry, and short support. Previous techniques were operating in time domain and required the solution of large systems of nonlinear equations. By switching to the frequency domain and employing the latest results of the multiwavelet theory we are able to elaborate a simple and efficient method of construction of multiscaling functions. Our algorithm is based on a recently found factorization of the refinement mask through the two-scale similarity transform (TST). Theoretical results and new examples are presented.

Key words. approximation order, symmetry, multiscaling functions, multiwavelets

AMS subject classifications. 41A25, 42A38, 39B62

PII. S0036141096297182

1. Introduction. This paper discusses the construction of *multiscaling functions* which generate a *multiresolution analysis* (MRA) and lead to *multiwavelets*. A standard (*scalar*) MRA assumes that there is only one scaling function $\phi(t)$ whose translates $\phi(t-k)$ $(k \in \mathbb{Z})$ constitute an L^2-stable basis of their span V_0 [D2, SN]. We move a step forward and allow several functions $\phi_0(t), \dots, \phi_{r-1}(t)$. The vector $\boldsymbol{\phi}(t) = [\phi_0(t) \cdots \phi_{r-1}(t)]^{\mathrm{T}}$ is called a *multiscaling function* if the integer translates $\phi_\nu(\cdot - k)$ $(k \in \mathbb{Z}, \nu = 0, \dots, r-1)$ form an L^2-stable basis of V_0 and if $\boldsymbol{\phi}(t)$ satisfies a *dilation equation*,

$$\boldsymbol{\phi}(t) = \sum_{n=0}^{N} \boldsymbol{P}_n \, \boldsymbol{\phi}(2t-n). \tag{1.1}$$

Here the coefficients $\boldsymbol{P}_n$ are $r \times r$ matrices instead of usual scalars. The multiscaling function $\boldsymbol{\phi}$ generates an MRA $\{V_j : j \in \mathbb{Z}\}$ of multiplicity r. The corresponding wavelet spaces W_j can be generated by a *multiwavelet* $\boldsymbol{w}(t) = [w_0(t) \cdots w_{r-1}(t)]^{\mathrm{T}}$ associated with $\boldsymbol{\phi}(t)$, satisfying a *wavelet equation*

$$\boldsymbol{w}(t) = \sum_{n=0}^{K} \boldsymbol{D}_n \, \boldsymbol{\phi}(2t-n). \tag{1.2}$$

Again, $\boldsymbol{D}_n$ are $r \times r$ matrices.

Multiwavelets naturally generalize the scalar wavelets. For $r = 1$, (1.1) is the well-studied refinement equation (see, e.g., [CDM, DL1, DL2]). However, multiwavelets have some completely new features arising from the matrix nature ($r > 1$) of the equation (1.1). They can simultaneously possess symmetry, orthogonality, and high approximation order which is not possible in the scalar case [SB, D2]. This suggests that in some applications multiwavelets may behave better than the scalar ones. The

*Received by the editors January 12, 1996; accepted for publication (in revised form) December 10, 1996.

http://www.siam.org/journals/sima/29-2/29718.html

†Fachbereich Mathematik, Universität Rostock, D-18051 Rostock, Germany (plonka@mathematik.uni-rostock.de).

‡Department of Mathematics, MIT, Cambridge, MA 02139 (strela@math.mit.edu).

results of first experiments [SHSTH, XGHS] confirm this conjecture and show that the multiwavelets are definitely worth studying.

One of the first multiwavelet constructions is due to Alpert and Rokhlin [AR]. They considered a multiscaling function whose components are polynomials of degree $r-1$ supported on $[0,1]$. The general theory of multiwavelets, based on the MRA of multiplicity r, is discussed in [GLT, GL].

Using fractal interpolation, Geronimo, Hardin, and Massopust succeeded to construct a continuous multiscaling function $\boldsymbol{\phi}(t) = [\phi_0(t)\ \phi_1(t)]^{\mathrm{T}}$ with short support, symmetry, and second approximation order [GHM]. The plot of this pair $\phi_0(t)$, $\phi_1(t)$ is presented in Figure 1.1.

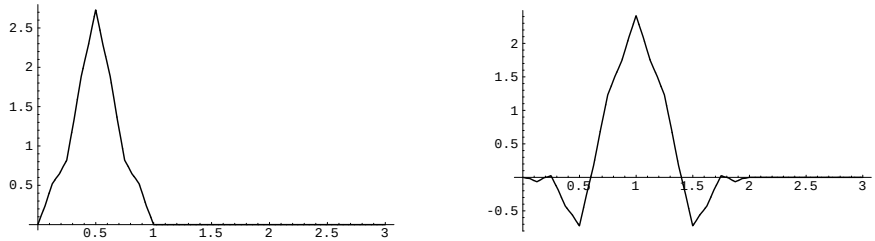

FIG. 1.1. *GHM symmetric orthogonal multiscaling function with approximation order* 2.

The results of [GHM] triggered many attempts to construct more examples [SS1, CL, DGHM] as well as the systematic study of multiscaling functions.

Properties of a refinable function can be formulated either in *time* or in *frequency* domain. In [SS2, HSS, L], conditions of orthogonality and approximation were established in the time domain. Also, a way to construct multiscaling functions with short support and low approximation order was found [SS1, HSS, CL]. Unfortunately, this method required the solution of a large system of nonlinear equations. We therefore switch to the frequency domain.

Working in the frequency domain, one faces the necessity of dealing with the Fourier transformation of (1.1),

$$\widehat{\boldsymbol{\phi}}(\omega) = \boldsymbol{P}\left(\frac{\omega}{2}\right)\,\widehat{\boldsymbol{\phi}}\left(\frac{\omega}{2}\right), \tag{1.3}$$

where $\widehat{\boldsymbol{\phi}} := [\widehat{\phi}_0, \ldots, \widehat{\phi}_{r-1}]^{\mathrm{T}}$, $\widehat{\phi}_\nu := \int_{-\infty}^{\infty} \phi_\nu(t)\, e^{-it\cdot}\, \mathrm{d}t$, and $\boldsymbol{P}(\omega)$ is the *refinement mask* corresponding to $\boldsymbol{\phi}(t)$,

$$\boldsymbol{P}(\omega) := \frac{1}{2}\sum_{n=0}^{N} \boldsymbol{P}_n\, e^{-i\omega n}. \tag{1.4}$$

In the scalar case, $P(\omega)$ is a *trigonometric polynomial*. In the vector case, $\boldsymbol{P}(\omega)$ becomes a *matrix of trigonometric polynomials*. To ensure certain approximation order, $\boldsymbol{P}(\omega)$ must satisfy necessary and sufficient conditions in the frequency domain. Those conditions were formulated and proved in [HSS, P3]. In [P3], it was also shown

that the vector $\boldsymbol{\phi}(t)$ can only provide approximation order m if the refinement mask $\boldsymbol{P}(\omega)$ can be factorized in the form

$$\boldsymbol{P}(\omega) = \frac{1}{2^m}\, \boldsymbol{C}_{m-1}(2\omega) \cdots \boldsymbol{C}_0(2\omega)\, \boldsymbol{P}^{(0)}(\omega)\, \boldsymbol{C}_0(\omega)^{-1} \cdots \boldsymbol{C}_{m-1}(\omega)^{-1}, \tag{1.5}$$

where $\boldsymbol{P}^{(0)}(\omega)$ is well defined and $\boldsymbol{C}_0(\omega), \ldots, \boldsymbol{C}_{m-1}(\omega)$ are matrices of a special form. The factorization (1.5) is not unique. With the help of the *two-scale similarity transformation* (TST), the whole set of possible factorizations can be described [S1].

The factorization (1.5) naturally generalizes the scalar case $r = 1$. As known, one scaling function with compact support and linearly independent integer translates provides approximation order m if and only if its refinement mask $P(\omega)$ has m zeros at $\omega = \pi$:

$$P(\omega) = \left(\frac{1 + e^{-i\omega}}{2}\right)^m q(\omega), \tag{1.6}$$

with $q(0) = 1$ and $q(\pi) \neq 0$. For $r = 1$, (1.6) coincides with (1.5) taking $\boldsymbol{P}^{(0)}(\omega) = q(\omega)$ and $\boldsymbol{C}_0(\omega) = \cdots = \boldsymbol{C}_{m-1}(\omega) = (1 - e^{i\omega})$. Daubechies connected the behavior of $q(\omega)$ in (1.6) with the decay properties of $\widehat{\phi}(\omega)$, and hence, she obtained estimates of smoothness of $\phi(t)$ [D1]. The factorization (1.5) plays the same role for a multiscaling function as (1.6) for a scalar one. In [CDP] and independently in [S1], it was shown how the factorization of the refinement mask $\boldsymbol{P}(\omega)$ leads to the decay of $\widehat{\boldsymbol{\phi}}(\omega)$. Similar results on regularity of refinable function vectors are presented in [Sh].

However, up to now, the factorization (1.5) has been shown to be necessary only. For the construction of multiscaling functions we need the sufficiency of a factorization (1.5) for approximation order m. In this paper, we show how, under mild conditions, the factorization of the refinement mask $\boldsymbol{P}(\omega)$ yields a solution of (1.1) with desired approximation properties. Using this result and the TST, a construction of multiscaling functions providing an arbitrary, fixed approximation order becomes simple. Description of the corresponding algorithm is our main purpose.

The outline of the paper is as follows. In section 2 we summarize previously known and new theoretical results on the symmetry of $\boldsymbol{\phi}(t)$, its approximation order, the factorization of the refinement mask $\boldsymbol{P}(\omega)$, and the TST. The main novelty of section 2 is the observation that the factorization of the refinement mask leads to the approximation order of the multiscaling functions (Theorem 2.6). Other remarkable new results are given in Theorems 2.7, 2.9, and Lemma 2.5.

In section 3, we present a new algorithm for the construction of a refinement mask $\boldsymbol{P}(\omega)$ with any given approximation order. We intensively study how the inner matrix $\boldsymbol{P}^{(0)}(\omega)$ and the transformation matrices $\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ should be chosen in order to obtain a smooth, symmetric multiscaling function with compact support. Several examples are given.

Section 4 contains the proof of Theorem 2.6.

2. Old and new theoretical results. In this section, we are going to present the results needed for the construction of symmetric multiscaling functions with given approximation order.

Let us start with definitions and notation. For a measurable function f over $\mathbb{R}$ and $m \in \mathbb{N}$ let

$$\|f\|_2 := \left(\int_{-\infty}^{\infty} |f(t)|^2 \,\mathrm{d}t\right)^{1/2}, \qquad \|f\|_{m,2} := \sum_{k=0}^{m} \|\mathrm{D}^k f\|_2.$$

Here and below $D := d/d\omega$ denotes the differentiation operator with respect to ω.

Let $W_2^m(\mathbb{R})$ be the usual Sobolev space with norm $\|\cdot\|_{m,2}$. For a vector $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$ of compactly supported functions, let $\mathcal{S} = \mathcal{S}(\boldsymbol{\phi})$ be the shift-invariant space spanned by the integer translates $\phi_\nu(t-k)$ $(\nu = 0,\dots,r-1,\ k \in \mathbf{Z})$. We say that $\boldsymbol{\phi}(t)$ provides *approximation order* m if for every $f \in W_2^m(\mathbb{R})$

$$\min\{\|f-s\| :\ s \in \mathcal{S}^h\} \le \mathrm{const}_{\mathcal{S}}\ h^m\, \|f\|_{W_2^m(\mathbb{R})},$$

where $\mathcal{S}^h$ is the scaled space $\mathcal{S}^h := \{s(\cdot/h) :\ s \in \mathcal{S}\}$.

A vector $\boldsymbol{v}$ of length r is said to be in $C_{2\pi}^m(\mathbb{R}^r)$ and, analogously, an $r \times r$ matrix $\boldsymbol{V}$ is in $C_{2\pi}^m(\mathbb{R}^{r\times r})$ if all its entries are 2π-periodic m times continuously differentiable functions.

2.1. Conditions of approximation. Assume that $\phi_\nu \in C(\mathbb{R}) \cap BV(\mathbb{R})$ $(\nu = 0,\dots,r-1)$ are compactly supported functions. Here $BV(\mathbb{R})$ denotes the set of functions of bounded variation. If the integer translates $\phi_\nu(\cdot - l)$ form a Riesz basis of $\mathcal{S}(\boldsymbol{\phi})$, then the following statements are equivalent (see [JL, P3]):

(i) The function vector $\boldsymbol{\phi}(t)$ provides approximation order m $(m \in \mathbb{N})$.

(ii) All algebraic polynomials of degree up to $m-1$ can be exactly reproduced by a linear combination of integer translates of $\phi_\nu(t)$.

(iii) $\boldsymbol{\phi}(t)$ satisfies the Strang–Fix conditions [SF] of order m; in other words, there is a finitely supported sequence of vectors $\{\boldsymbol{a}_l\}_{l\in\mathbb{Z}}$ such that $f(t) := \sum_{l\in\mathbb{Z}} \boldsymbol{a}_l^{\mathrm{T}}\, \boldsymbol{\phi}(t-l)$ satisfies the following conditions:

$$\widehat{f}(0) \ne 0; \qquad \mathrm{D}^n \widehat{f}(2\pi l) = 0 \quad (l \in \mathbb{Z}\setminus\{0\};\ n = 0,\dots,m-1).$$

Since condition (ii) yields vanishing moments for the corresponding multiwavelets it is often used in applications.

The approximation order of a refinable function vector $\boldsymbol{\phi}(t)$ satisfying (1.1) is intimately related with the properties of the refinement mask $\boldsymbol{P}(\omega)$ defined by (1.4). In the scalar case $(r = 1)$, when there is only one scaling function, $\boldsymbol{P}_n$ are real numbers and $P(\omega)$ is a scalar trigonometric polynomial. Then mth approximation order implies m zeros of $P(\omega)$ at $\omega = \pi$ [D2]. In the vector case, $\boldsymbol{P}(\omega)$ is a matrix, and the situation becomes more complicated. But still, similar conditions at the point $\omega = \pi$ hold.

THEOREM 2.1 (see [HSS, P3]). *Let $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$ be a refinable vector of compactly supported functions ϕ_ν. Further, assume that the integer translates $\phi_\nu(t-l)$ $(l \in \mathbb{Z})$ form a Riesz basis of $\mathcal{S}(\boldsymbol{\phi})$. Then $\boldsymbol{\phi}(t)$ provides approximation order m if and only if the refinement mask $\boldsymbol{P}(\omega)$ of $\boldsymbol{\phi}$ satisfies the following conditions: there are vectors $\boldsymbol{y}_k \in \mathbb{R}^r$; $\boldsymbol{y}_0 \ne \mathbf{0}$ $(k = 0,\dots,m-1)$ such that for $n = 0,\dots,m-1$,*

$$\sum_{k=0}^{n} \binom{n}{k} (\boldsymbol{y}_k)^T\, (2i)^{k-n}\, (D^{n-k}\boldsymbol{P})(0) = 2^{-n}\, (\boldsymbol{y}_n)^T, \tag{2.1}$$

$$\sum_{k=0}^{n} \binom{n}{k} (\boldsymbol{y}_k)^T\, (2i)^{k-n}\, (D^{n-k}\boldsymbol{P})(\pi) = \mathbf{0}^T. \tag{2.2}$$

Here $\mathbf{0}$ denotes the zero vector.

If a matrix $\boldsymbol{P}(\omega) \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ satisfies (2.1) and (2.2) for $n = 0,\dots,m-1$ with vectors $\boldsymbol{y}_0,\dots,\boldsymbol{y}_{m-1}$ $(\boldsymbol{y}_0 \ne \mathbf{0})$, then we shortly say that $\boldsymbol{P}(\omega)$ provides approximation order m with $\boldsymbol{y}_0,\dots,\boldsymbol{y}_{m-1}$. In order to prove that relations (2.1) and (2.2) imply

approximation order m, one only needs to assume that $\boldsymbol{y}_0^{\mathrm{T}}\,\widehat{\boldsymbol{\phi}}(0) \neq 0$. Riesz stability of integer translates $\phi_\nu(t-l)$ is *not* needed.

Remark. The result of Theorem 2.1 is a natural generalization of the scalar case. For $r = 1$, equations (2.1), (2.2) can be simplified to

$$P(0) = 1; \qquad \mathrm{D}^n P(\pi) = 0 \quad (n = 0, \ldots, m-1), \tag{2.3}$$

implying m zeros of $P(\omega)$ at $\omega = \pi$. Note that in the vector case, we need conditions in *two* points, $\omega = 0$ and $\omega = \pi$. Also, both eigenvalues and eigenvectors of $\boldsymbol{P}(0)$ and $\boldsymbol{P}(\pi)$ are important.

2.2. Two-scale similarity transform. A very useful research and construction tool in the theory of multiwavelets is the TST [S1]. We say that $\boldsymbol{Q}(\omega)$ is a TST of $\boldsymbol{P}(\omega)$ with the *transformation matrix* $\boldsymbol{M}(\omega) \in C_{2\pi}(\mathbb{R}^{r\times r})$ if

$$\boldsymbol{Q}(\omega) = \boldsymbol{M}(2\omega)\,\boldsymbol{P}(\omega)\,\boldsymbol{M}(\omega)^{-1}.$$

If $\boldsymbol{M}(\omega)$ is invertible for all $\omega \in \mathbb{R}$, then the TST is *nondegenerate*. It is easy to see that if $\boldsymbol{P}(\omega) \in C_{2\pi}(\mathbb{R}^{r\times r})$ is the refinement mask of $\boldsymbol{\phi} \in L^2(\mathbb{R}^r)$, then a nondegenerate TST with transformation matrix $\boldsymbol{M}(\omega) \in C_{2\pi}(\mathbb{R}^{r\times r})$ yields a matrix $\boldsymbol{Q}(\omega)$ which itself is a refinement mask of a refinable function vector $\boldsymbol{\psi} \in L^2(\mathbb{R}^r)$ such that $\widehat{\boldsymbol{\psi}}(\omega) = \boldsymbol{M}(\omega)\,\widehat{\boldsymbol{\phi}}(\omega)$:

$$\begin{aligned}\widehat{\boldsymbol{\psi}}(\omega) &= \boldsymbol{M}(\omega)\,\widehat{\boldsymbol{\phi}}(\omega) = \boldsymbol{M}(\omega)\,\boldsymbol{P}\left(\frac{\omega}{2}\right)\,\widehat{\boldsymbol{\phi}}\left(\frac{\omega}{2}\right)\\ &= \boldsymbol{M}(\omega)\,\boldsymbol{P}\left(\frac{\omega}{2}\right)\,\boldsymbol{M}\left(\frac{\omega}{2}\right)^{-1}\,\widehat{\boldsymbol{\psi}}\left(\frac{\omega}{2}\right) = \boldsymbol{Q}\left(\frac{\omega}{2}\right)\,\widehat{\boldsymbol{\psi}}\left(\frac{\omega}{2}\right).\end{aligned}$$

The following theorem shows that a nondegenerate TST preserves the approximation properties of a refinement mask.

THEOREM 2.2 (see [S1]). *Let a transformation matrix $\boldsymbol{M}(\omega) \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ be invertible for all $\omega \in \mathbb{R}$. Assume that $\boldsymbol{P} \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ provides approximation order m with vectors $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{m-1}$. Then $\boldsymbol{Q}(\omega) = \boldsymbol{M}(2\omega)\,\boldsymbol{P}(\omega)\,\boldsymbol{M}(\omega)^{-1}$ also provides approximation order m with vectors $\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{m-1}$, given by*

$$\boldsymbol{u}_k^T := \sum_{l=0}^{k} \binom{k}{l} i^{l-k}\,\boldsymbol{y}_l^T\,(D^{k-l}\boldsymbol{M}^{-1})(0) \qquad (k = 0, \ldots, m-1).$$

For more properties of the TST and the proof of Theorem 2.2 see [S1, S2].

2.3. Factorizations of the refinement mask. In the scalar case, the conditions of approximation (2.3) lead to a factorization of $P(\omega)$. A zero at $\omega = \pi$ means that $P(\omega)$ has a factor $(1 + e^{-i\omega})$. So $P(\omega)$ factorizes as in (1.6). This factorization plays the key role in the construction of regular scalar scaling functions [D2].

In the vector case, the conditions of approximation (2.1), (2.2) are more complicated, but still they imply a factorization of the *matrix* refinement mask $\boldsymbol{P}(\omega)$. This factorization opens a constructive way toward the creation of new multiscaling functions. But before starting with the factorization, we need to review some notation.

Let $r \in \mathbb{N}$ be fixed, and let $\boldsymbol{y} \in \mathbb{R}^r$ be a vector of length r. To start, assume that $\boldsymbol{y}$ is of the form

$$\boldsymbol{y} = [y_0 \cdots y_{l-1}\ 0 \cdots 0]^{\mathrm{T}}, \tag{2.4}$$

with $1 \le l \le r$ and $y_\nu \neq 0$ for $\nu = 0, \dots, l-1$. We introduce the direct sum of square matrices $\boldsymbol{A} \oplus \boldsymbol{B} := \mathrm{diag}(\boldsymbol{A}, \boldsymbol{B})$ and define the matrix $\boldsymbol{C_y}$ by

$$\boldsymbol{C_y}(\omega) := \widetilde{\boldsymbol{C}}\boldsymbol{y}(\omega) \oplus \boldsymbol{I}_{r-l}. \tag{2.5}$$

Here $\boldsymbol{I}_{r-l}$ denotes the $(r-l)\times(r-l)$ unit matrix, and for $l > 1$, $\widetilde{\boldsymbol{C}}\boldsymbol{y}(\omega)$ is defined by

$$\widetilde{\boldsymbol{C}}\boldsymbol{y}(\omega) := \begin{bmatrix} y_0^{-1} & -y_0^{-1} & 0 & \dots & 0 \\ 0 & y_1^{-1} & -y_1^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & y_{l-2}^{-1} & -y_{l-2}^{-1} \\ -e^{-i\omega}/y_{l-1} & 0 & \dots & 0 & y_{l-1}^{-1} \end{bmatrix}.$$

If $l = 1$, let $\widetilde{\boldsymbol{C}}\boldsymbol{y}(\omega) = (1 - e^{-i\omega})/y_0$.

For general $\boldsymbol{y} = [y_0 \cdots y_{r-1}]^{\mathrm{T}} \in \mathbb{R}^r$, $\boldsymbol{y} \neq \boldsymbol{0}$ we define $\boldsymbol{C_y} := (C_{j,k})_{j,k=0}^{r-1}$ by reshuffling rows and columns. More exactly, let $j_0 := \min\{j;\, y_j \neq 0\}$ and $j_1 := \max\{j; y_j \neq 0\}$. For all $j < j_1$ with $y_j \neq 0$ let $d_j := \min\{\mu :\, \mu > j,\, y_\mu \neq 0\}$. For $j_0 < j_1$, the entries of $\boldsymbol{C_y}$ are defined by

$$C_{j,k}(\omega) := \begin{cases} y_j^{-1} & y_j \neq 0 \quad \text{and} \quad j = k, \\ 1 & y_j = 0 \quad \text{and} \quad j = k, \\ -y_j^{-1} & y_j \neq 0 \quad \text{and} \quad d_j = k, \\ -e^{-i\omega}/y_{j_1} & j = j_1 \quad \text{and} \quad k = j_0, \\ 0 & \text{otherwise} \end{cases} \qquad (j,\, k = 0, \dots, r-1). \tag{2.6}$$

For $j_0 = j_1$, $\boldsymbol{C_y}$ is a diagonal matrix of the form

$$\boldsymbol{C_y}(\omega) := \mathrm{diag}(\underbrace{1, \dots, 1}_{j_0}, (1 - e^{-i\omega})/y_{j_0}, \underbrace{1, \dots, 1}_{r-1-j_0}). \tag{2.7}$$

It is easy to observe that $\boldsymbol{C_y}(\omega)$ is invertible for $\omega \neq 0$. In particular,

$$\det \boldsymbol{C_y}(\omega) = \left(\prod_{\substack{\nu=0 \\ y_\nu \neq 0}}^{r-1} y_\nu^{-1} \right) (1 - e^{-i\omega}). \tag{2.8}$$

Furthermore, $\boldsymbol{C_y}$ is chosen so that $\boldsymbol{y}^{\mathrm{T}} \boldsymbol{C_y}(0) = \boldsymbol{0}^{\mathrm{T}}$. We introduce

$$\boldsymbol{G_y}(\omega) := (1 - e^{-i\omega})\, \boldsymbol{C}_{\boldsymbol{y}}^{-1}(\omega)\ . \tag{2.9}$$

If $\boldsymbol{y}$ is of the form (2.4), then $\boldsymbol{G_y}(\omega) = \widetilde{\boldsymbol{G}}\boldsymbol{y}(\omega) \oplus (1 - e^{-i\omega})\, \boldsymbol{I}_{r-l}$ with

$$\widetilde{\boldsymbol{G}}\boldsymbol{y}(\omega) := \begin{bmatrix} y_0 & y_1 & y_2 & \dots & y_{l-1} \\ y_0 z & y_1 & y_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & y_{l-1} \\ y_0 z & y_1 z & \ddots & y_{l-2} & y_{l-1} \\ y_0 z & y_1 z & \dots & y_{l-2} z & y_{l-1} \end{bmatrix} \qquad (z := e^{-i\omega})\ .$$

Finally, let $\boldsymbol{e} = (e_\nu)_{\nu=0}^{r-1}$ corresponding to $\boldsymbol{y} = (y_\nu)_{\nu=0}^{r-1}$ be defined by

$$e_\nu := \begin{cases} 1 & \text{for } y_\nu \neq 0, \\ 0 & \text{for } y_\nu = 0 \end{cases} \qquad (\nu = 0, \dots, r-1). \tag{2.10}$$

Now we can proceed with the factorization of $\boldsymbol{P}(\omega)$.

THEOREM 2.3 (see [P3]). *Let $m > 1$ be fixed. Assume that $\boldsymbol{P} \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ provides approximation order m with vectors $\boldsymbol{y}_0, \dots, \boldsymbol{y}_{m-1}$ $(\boldsymbol{y}_0 \neq \boldsymbol{0})$. Then*

$$\widetilde{\boldsymbol{P}}(\omega) := 2\, \boldsymbol{C}_{\boldsymbol{y}_0}(2\omega)^{-1}\, \boldsymbol{P}(\omega)\, \boldsymbol{C}_{\boldsymbol{y}_0}(\omega), \tag{2.11}$$

with $\boldsymbol{C}_{\boldsymbol{y}_0}(\omega)$ defined by $\boldsymbol{y}_0$ via (2.6)–(2.7)*, provides approximation order at least $m-1$ with vectors $\widetilde{\boldsymbol{y}}_0, \dots, \widetilde{\boldsymbol{y}}_{m-2}$, given by*

$$(\widetilde{\boldsymbol{y}}_k)^T := \frac{1}{k+1} \sum_{\nu=0}^{k+1} \binom{k+1}{\nu} i^{\nu-k-1}\, (\boldsymbol{y}_\nu)^T\, (D^{k+1-\nu} \boldsymbol{C}_{\boldsymbol{y}_0})(0) \tag{2.12}$$

for $k = 0, \dots, m-2$. In particular $\widetilde{\boldsymbol{y}}_0 \neq \boldsymbol{0}$.

Moreover, if $\boldsymbol{e}$ corresponds to $\boldsymbol{y}_0$ in the sense of (2.10)*, then $\widetilde{\boldsymbol{P}}(\omega)$ in* (2.11) *satisfies $\widetilde{\boldsymbol{P}}(0)\, \boldsymbol{e} = \boldsymbol{e}$.*

Assume that $\boldsymbol{P} \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ provides approximation order m; then repeated application of Theorem 2.3 yields the desired factorization of $\boldsymbol{P}(\omega)$:

$$\boldsymbol{P}(\omega) = \frac{1}{2^m} \boldsymbol{C}_{\boldsymbol{x}_{m-1}}(2\omega) \cdots \boldsymbol{C}_{\boldsymbol{x}_0}(2\omega)\, \boldsymbol{P}^{(0)}(\omega)\, \boldsymbol{C}_{\boldsymbol{x}_0}(\omega)^{-1} \cdots \boldsymbol{C}_{\boldsymbol{x}_{m-1}}(\omega)^{-1}. \tag{2.13}$$

Here $\boldsymbol{P}^{(0)}(\omega) \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ and $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{m-1} \in \mathbb{R}^r$ are defined recursively by (2.12) [P3]. In particular, $\boldsymbol{x}_{m-1} = \boldsymbol{y}_0$ and, by (2.8),

$$\det \boldsymbol{P}(\omega) = \left(\frac{1 + e^{-i\omega}}{2^r} \right)^m \det \boldsymbol{P}^{(0)}(\omega).$$

However, the factorization (2.13) is not unique. The following theorem is a generalization of Theorem 2.3.

THEOREM 2.4 (see [S1]). *Let $m \geq 1$ be fixed, and let $\boldsymbol{P} \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ provide approximation order m with vectors $\boldsymbol{y}_0, \dots, \boldsymbol{y}_{m-1}$. Further, let $\boldsymbol{M} \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ satisfy the following conditions:*

1. *$\boldsymbol{M}(\omega)$ is invertible for all $\omega \neq 0$.*
2. *$\boldsymbol{M}(0)$ has a simple eigenvalue 0 with a corresponding left eigenvector $\boldsymbol{y}_0$ and $D(\det \boldsymbol{M})(0) \neq 0$.*

Then,

$$\widetilde{\boldsymbol{P}}(\omega) = 2\, \boldsymbol{M}(2\omega)^{-1}\, \boldsymbol{P}(\omega)\, \boldsymbol{M}(\omega) \tag{2.14}$$

provides approximation order at least $m-1$ with vectors $\boldsymbol{u}_0, \dots, \boldsymbol{u}_{m-2}$ $(m > 1)$ defined by

$$\boldsymbol{u}_k^T := \frac{1}{k+1} \sum_{j=0}^{k+1} \binom{k+1}{j} i^{j-k-1}\, \boldsymbol{y}_j^T\, (D^{k+1-j} \boldsymbol{M})(0) \qquad (k = 0, \dots, m-2).$$

In particular, $\boldsymbol{u}_0 \neq 0$. If $\boldsymbol{P}$ exactly provides approximation order $m = 1$, then $\widetilde{\boldsymbol{P}}(0)$ has the eigenvalue 1, but there exists no vector $\boldsymbol{y} \neq \boldsymbol{0}$ such that $\widetilde{\boldsymbol{P}}(\omega)$ satisfies (2.1), (2.2) *for $n = 0$.*

In [S1], this result was obtained directly, using similar ideas as in the proof of Theorem 2.3 in [P3]. Here we would like to give another proof which clearly shows the connection between the particular factorization matrix $\boldsymbol{C}_{\boldsymbol{y}_0}$ and general factorization matrices $\boldsymbol{M}$.

LEMMA 2.5. *Let $\boldsymbol{y} \in \mathbb{R}^r$ be a fixed nonzero vector, and let $\boldsymbol{M} \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ satisfy assumptions 1 and 2 of Theorem 2.4 (with $\boldsymbol{y}$ instead of $\boldsymbol{y}_0$). Further, let $\boldsymbol{C}_{\boldsymbol{y}}$ be an $r \times r$ matrix defined by $\boldsymbol{y}$ via (2.6)–(2.7). Then, there exists a matrix $\boldsymbol{M}_0(\omega) \in C_{2\pi}^{m-1}(\mathbb{R}^{r\times r})$ which is invertible for all $\omega \in \mathbb{R}$, and*

$$\boldsymbol{C}_{\boldsymbol{y}}(\omega)\, \boldsymbol{M}_0(\omega) = \boldsymbol{M}(\omega).$$

Proof. Let $\boldsymbol{G}_{\boldsymbol{y}}$ be the $r \times r$ matrix defined by $\boldsymbol{C}_{\boldsymbol{y}}$ via (2.9). Define $\boldsymbol{M}_0(\omega)$ as follows:

$$\boldsymbol{M}_0(\omega) := \begin{cases} \boldsymbol{C}_{\boldsymbol{y}}(\omega)^{-1}\, \boldsymbol{M}(\omega) & \text{for } \omega \neq 0, \\ (-i)\, \big((\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0)\, \boldsymbol{M}(0) + \boldsymbol{G}_{\boldsymbol{y}}(0)\, (\mathrm{D}\boldsymbol{M})(0)\big) & \text{for } \omega = 0. \end{cases}$$

Here, $\boldsymbol{M}_0(0)$ is found by the rule of l'Hospital from

$$M_0(0) = \lim_{\omega\to 0} \boldsymbol{C}_{\boldsymbol{y}}(\omega)^{-1}\, \boldsymbol{M}(\omega) = \lim_{\omega\to 0} \frac{1}{1-e^{-i\omega}}\, \boldsymbol{G}_{\boldsymbol{y}}(\omega)\, \boldsymbol{M}(\omega).$$

Since $\boldsymbol{C}_{\boldsymbol{y}}(\omega)$ is invertible for $\omega \neq 0$, the relation $\boldsymbol{C}_{\boldsymbol{y}}(\omega)\, \boldsymbol{M}_0(\omega) = \boldsymbol{M}(\omega)$ easily follows for $\omega \neq 0$. For $\omega = 0$, we find

$$\boldsymbol{C}_{\boldsymbol{y}}(0)\, \boldsymbol{M}_0(0) = (-i)\, \big(\boldsymbol{C}_{\boldsymbol{y}}(0)\, (\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0)\, \boldsymbol{M}(0) + \boldsymbol{C}_{\boldsymbol{y}}(0)\, \boldsymbol{G}_{\boldsymbol{y}}(0)\, (\mathrm{D}\boldsymbol{M})(0)\big)\,.$$

Observe that, by definition,

$$\boldsymbol{G}_{\boldsymbol{y}}(\omega)\, \boldsymbol{C}_{\boldsymbol{y}}(\omega) = \boldsymbol{C}_{\boldsymbol{y}}(\omega)\, \boldsymbol{G}_{\boldsymbol{y}}(\omega) = (1-e^{-i\omega})\, \boldsymbol{I}_r.$$

Hence,

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{y}}(0)\, \boldsymbol{G}_{\boldsymbol{y}}(0) &= \boldsymbol{0}_r, \\ (\mathrm{D}\boldsymbol{C}_{\boldsymbol{y}})(0)\, \boldsymbol{G}_{\boldsymbol{y}}(0) + \boldsymbol{C}_{\boldsymbol{y}}(0)\, (\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0) &= i\boldsymbol{I}_r. \end{aligned}$$

From the assumption $\boldsymbol{y}^{\mathrm{T}}\, \boldsymbol{M}(0) = \boldsymbol{0}^{\mathrm{T}}$ we have $\boldsymbol{G}_{\boldsymbol{y}}(0)\boldsymbol{M}(0) = \boldsymbol{0}_r$. Thus,

$$\boldsymbol{C}_{\boldsymbol{y}}(0)\, \boldsymbol{M}_0(0) = (-i)\, \big(i\boldsymbol{I}_r - (\mathrm{D}\boldsymbol{C}_{\boldsymbol{y}})(0)\, \boldsymbol{G}_{\boldsymbol{y}}(0)\big)\, \boldsymbol{M}(0) = \boldsymbol{M}(0).$$

We see that $\boldsymbol{C}_{\boldsymbol{y}}(\omega)\, \boldsymbol{M}_0(\omega) = \boldsymbol{M}(\omega)$ for all $\omega \in \mathbb{R}$. Since $\boldsymbol{C}_{\boldsymbol{y}}(\omega)$ and $\boldsymbol{M}(\omega)$ are invertible for $\omega \neq 0$, $\boldsymbol{M}_0(\omega)$ is also invertible for $\omega \neq 0$. Further, since $\mathrm{D}(\det \boldsymbol{C}_{\boldsymbol{y}})(0) \neq 0$ and $\mathrm{D}(\det \boldsymbol{M})(0) \neq 0$, it follows that

$$\det \boldsymbol{M}_0(0) = \lim_{\omega\to 0} \frac{\det \boldsymbol{M}(\omega)}{\det \boldsymbol{C}_{\boldsymbol{y}}(\omega)} = \frac{\mathrm{D}(\det \boldsymbol{M}(0))}{\mathrm{D}(\det \boldsymbol{C}_{\boldsymbol{y}}(0))} \neq 0.$$

Thus, $\boldsymbol{M}_0(0)$ is invertible. ∎

Proof of Theorem 2.4. Recall that by Theorem 2.2, a TST with an invertible transformation matrix does not change the approximation order of a refinement mask. Using the result of Lemma 2.5, we simply observe that a factorization step (2.14) with a matrix $\boldsymbol{M}(\omega)$ can be considered as a combination of factorization step (2.11) with $\boldsymbol{C}_{\boldsymbol{y}_0}(\omega)$ and a nondegenerate TST with the transform matrix $\boldsymbol{M}_0(\omega)$. ∎

While the matrices $\boldsymbol{C_y}$ are determined by a left eigenvector $\boldsymbol{y}$ of $\boldsymbol{C_y}(0)$ to the eigenvalue 0, we want to identify the matrices $\boldsymbol{M}$ with the help of right eigenvectors of $\boldsymbol{M}(0)$ to the eigenvalue 0. Letting $\boldsymbol{r}_0$ be a right eigenvector of $\boldsymbol{M}(0)$ in Theorem 2.4, we then have $\boldsymbol{M_{r_0}} := \boldsymbol{M}$. Hence, similar to (2.13), repeated application of Theorem 2.4 gives a general factorization of $\boldsymbol{P}(\omega)$:

$$\boldsymbol{P}(\omega) = \frac{1}{2^m} \boldsymbol{M_{r_{m-1}}}(2\omega) \cdots \boldsymbol{M_{r_0}}(2\omega)\, \boldsymbol{P}^{(0)}(\omega)\, \boldsymbol{M_{r_0}}(\omega)^{-1} \cdots \boldsymbol{M_{r_{m-1}}}(\omega)^{-1}. \tag{2.15}$$

2.4. Factorization implies approximation order. In this section we state the main theoretical results of the paper. First, let us again return for a moment to the scalar case ($r = 1$). In [St1], it was shown that the approximation order defines the number of factors $(1 + e^{-i\omega})$ in $P(\omega)$, and on the other hand each such factor increases the approximation order by one. Therefore, our next step is to prove the reverse of Theorems 2.3 and 2.4, or in other words, to show that the factorization (2.15) of the refinement mask yields approximation order m for the corresponding refinable function vector.

To this end, we need to introduce the "modified" Bernoulli numbers $\widetilde{B}_n$ ($n \in \mathbb{N}$), defined by the following relations:

$$\widetilde{B}_0 = 1, \qquad \sum_{l=0}^{n} \binom{n+1}{l} (-1)^l\, \widetilde{B}_l = 0, \tag{2.16}$$

or

$$\widetilde{B}_0 = 1, \qquad \widetilde{B}_n = \frac{(-1)^{n+1}}{n+1} \sum_{l=0}^{n-1} \binom{n+1}{l} (-1)^l\, \widetilde{B}_l \qquad (n \geq 1).$$

In particular,

$$\widetilde{B}_1 = \frac{1}{2}, \quad \widetilde{B}_2 = \frac{1}{6}, \quad \widetilde{B}_4 = -\frac{1}{30}.$$

Note that, apart from $\widetilde{B}_1$, the modified Bernoulli numbers coincide with the usual Bernoulli numbers B_n:

$$\widetilde{B}_n = B_n \quad (n \in \mathbb{N} \setminus \{1\}), \qquad \widetilde{B}_1 = -B_1.$$

This means that $\widetilde{B}_{2n+1} = B_{2n+1} = 0$ ($n \geq 1$), and we have

$$\sum_{l=0}^{n} \binom{n+1}{l} (-2)^l\, \widetilde{B}_l = \sum_{l=0}^{n} \binom{n+1}{l} 2^l\, B_l = \begin{cases} 1, & n = 0, \\ 2(-2^{n+1}+1)\, \widetilde{B}_{n+1}, & n \geq 1 \end{cases} \tag{2.17}$$

(see [AS]).

Now we are ready to state the main results of this section.

THEOREM 2.6. *Let $m \geq 1$ be a fixed integer and let $\widetilde{\boldsymbol{P}} \in C^m_{2\pi}(\mathbb{R}^{r\times r})$ be a refinement mask providing the approximation order m with $\widetilde{\boldsymbol{y}}_0, \ldots, \widetilde{\boldsymbol{y}}_{m-1} \in \mathbb{R}^r$ ($\widetilde{\boldsymbol{y}}_0 \neq \boldsymbol{0}$). Further, assume that there is a vector $\boldsymbol{e} \in \mathbb{R}^r$ ($\boldsymbol{e} \neq \boldsymbol{0}$), containing only the entries 0 or 1, such that $\widetilde{\boldsymbol{P}}(0)\, \boldsymbol{e} = \boldsymbol{e}$. Let $\boldsymbol{y} = (y_\nu)_{\nu=0}^{r-1} \in \mathbb{R}^r$ ($\boldsymbol{y} \neq \boldsymbol{0}$) be an arbitrary vector such that $\boldsymbol{e}$ corresponds to $\boldsymbol{y}$ in the sense of* (2.10). *Then the matrix $\boldsymbol{P}(\omega)$,*

$$\boldsymbol{P}(\omega) := \frac{1}{2}\, \boldsymbol{C_y}(2\omega)\, \widetilde{\boldsymbol{P}}(\omega)\, \boldsymbol{C_y}(\omega)^{-1}$$

with $\boldsymbol{C_y}$ defined by $\boldsymbol{y}$ via (2.6)–(2.7)*, provides approximation order at least $m+1$ with vectors $\boldsymbol{y}_0, \dots, \boldsymbol{y}_m$,*

$$\boldsymbol{y}_k^T := (-ik)\, \widetilde{\boldsymbol{y}}_{k-1}^T\, (D\boldsymbol{G_y})(0) + \sum_{l=0}^{k} \binom{k}{l} \widetilde{B}_{k-l}\, \widetilde{\boldsymbol{y}}_l^T\, \boldsymbol{G_y}(0) \quad (k = 0, \dots, m-1), \tag{2.18}$$

$$\begin{aligned} \boldsymbol{y}_m^T := (-im)\, \widetilde{\boldsymbol{y}}_{m-1}^T\, (D\boldsymbol{G_y})(0) + \sum_{l=0}^{m-1} \binom{m}{l} \widetilde{B}_{m-l}\, \widetilde{\boldsymbol{y}}_l^T\, \boldsymbol{G_y}(0) \\ - \frac{2^m}{2^m - 1} \sum_{k=0}^{m-1} \binom{m}{k} (2i)^{k-m}\, \widetilde{\boldsymbol{y}}_k^T\, (D^{m-k}\widetilde{\boldsymbol{P}})(0)\, \boldsymbol{G_y}(0), \end{aligned} \tag{2.19}$$

where $\widetilde{\boldsymbol{y}}_{-1} := \boldsymbol{0}$.

The proof of Theorem 2.6 is presented in section 4. In particular, we obtain from (2.18) that $\boldsymbol{y}_0^{\mathrm{T}} = \widetilde{\boldsymbol{y}}_0^{\mathrm{T}}\, \boldsymbol{G_y}(0) = (\sum_{\nu=0}^{r-1} \widetilde{y}_{0,\nu})\, \boldsymbol{y}^{\mathrm{T}}$ with $\widetilde{\boldsymbol{y}}_0 = (\widetilde{y}_{0,\nu})_{\nu=0}^{r-1}$. Observe that the technical assumption $\widetilde{\boldsymbol{P}}(0)\boldsymbol{e} = \boldsymbol{e}$ ensures that $\boldsymbol{C_y}$ has the same right eigenvector $\boldsymbol{e}$ to the eigenvalue 0 as $\widetilde{\boldsymbol{P}}(0)$ to the eigenvalue 1.

Again, we can generalize this result using the TST.

THEOREM 2.7. *Let $m \geq 1$ be a fixed integer and let $\widetilde{\boldsymbol{P}} \in C_{2\pi}^m(\mathbb{R}^{r\times r})$ be a refinement mask providing approximation order m with vectors $\widetilde{\boldsymbol{y}}_0, \dots, \widetilde{\boldsymbol{y}}_{m-1} \in \mathbb{R}^r$ $(\boldsymbol{y}_0 \neq \boldsymbol{0})$. Further, let $\boldsymbol{r}$ be a right eigenvector of $\widetilde{\boldsymbol{P}}(0)$ to the eigenvalue* 1*.*

Choose a matrix $\boldsymbol{M_r}(\omega) \in C_{2\pi}^m(\mathbb{R}^{r\times r})$ such that

1. *$\boldsymbol{M_r}(\omega)$ is invertible for all $\omega \in \mathbb{R},\ \omega \neq 0$.*
2. *$\boldsymbol{M_r}(0)$ has a simple eigenvalue* 0 *with $\boldsymbol{M_r}(0)\, \boldsymbol{r} = \boldsymbol{0}$.*
3. *$D(\det \boldsymbol{M_r})(0) \neq 0$.*

Let $\boldsymbol{u}$ be a left eigenvector of $\boldsymbol{M_r}(0)$ corresponding to the eigenvalue 0*. Then the matrix*

$$\boldsymbol{P}(\omega) := \frac{1}{2}\, \boldsymbol{M_r}(2\omega)\, \widetilde{\boldsymbol{P}}(\omega)\, \boldsymbol{M_r}(\omega)^{-1} \tag{2.20}$$

provides approximation order at least $m+1$ with vectors $\boldsymbol{u}_0, \dots, \boldsymbol{u}_m$, given by

$$\boldsymbol{u}_k^T := (-ik)\, \widetilde{\boldsymbol{u}}_{k-1}^T\, (D\boldsymbol{G_u})(0) + \sum_{l=0}^{k} \binom{k}{l} \widetilde{B}_{k-l}\, \widetilde{\boldsymbol{u}}_l^T\, \boldsymbol{G_u}(0) \qquad (k = 0, \dots, m-1),$$

$$\begin{aligned} \boldsymbol{u}_m^T := (-im)\, \widetilde{\boldsymbol{u}}_{m-1}^T\, (D\boldsymbol{G_u})(0) + \sum_{l=0}^{m-1} \binom{m}{l} \widetilde{B}_{m-l}\, \widetilde{\boldsymbol{u}}_l^T\, \boldsymbol{G_u}(0) \\ - \frac{2^m}{2^m - 1} \sum_{k=0}^{m-1} \binom{m}{k} (2i)^{k-m}\, \widetilde{\boldsymbol{u}}_k^T\, (D^{m-k}(\boldsymbol{M}_0(2\cdot)^{-1}\, \widetilde{\boldsymbol{P}}\, \boldsymbol{M}_0))(0)\, \boldsymbol{G_u}(0), \end{aligned}$$

where $\boldsymbol{M}_0(\omega)$ is an invertible matrix such that $\boldsymbol{C_u}(\omega)\, \boldsymbol{M}_0(\omega) = \boldsymbol{M_r}(\omega)$, $\widetilde{\boldsymbol{u}}_{-1} := \boldsymbol{0}$, and $\widetilde{\boldsymbol{u}}_k^T := \sum_{l=0}^{k} \binom{k}{l}\, i^{(l-k)}\, \widetilde{\boldsymbol{y}}_l^T\, (D^{k-l}\boldsymbol{M}_0^{-1})(0)$ for $k = 0, \dots, m-1$.

Proof. In [S1], it is shown that $\boldsymbol{P}(\omega)$ defined by (2.20) is in $C_{2\pi}^m(\mathbb{R}^{r\times r})$ and $\boldsymbol{P}(0)$ has a left eigenvector $\boldsymbol{u}$, corresponding to the eigenvalue 1:

$$\boldsymbol{u}^{\mathrm{T}}\, \boldsymbol{P}(0) = \boldsymbol{u}^{\mathrm{T}}.$$

Let $\boldsymbol{C_u}$ be defined by $\boldsymbol{u}$ via (2.6)–(2.7); then $\boldsymbol{u}^{\mathrm{T}}\,\boldsymbol{C_u}(0) = \boldsymbol{0}^{\mathrm{T}}$. By Lemma 2.5, there exists a regular matrix $\boldsymbol{M}_0 \in \boldsymbol{C}^m_{2\pi}(\mathbb{R}^{r\times r})$ such that

$$\boldsymbol{C_u}(\omega)\,\boldsymbol{M}_0(\omega) = \boldsymbol{M_r}(\omega).$$

Recall that the eigenvalue 0 of $\boldsymbol{C_u}(0)$ is simple, and we have $\boldsymbol{C_u}(0)\,\boldsymbol{e} = \boldsymbol{0}$, where $\boldsymbol{e}$ is connected with $\boldsymbol{u}$ via (2.10). Hence, from $\boldsymbol{M_r}(0)\,\boldsymbol{r} = \boldsymbol{0}$, it follows that $\boldsymbol{M}_0(0)\,\boldsymbol{r} = c\,\boldsymbol{e}$ with some constant $c \neq 0$. Since $\boldsymbol{M}_0(\omega)$ is invertible for all $\omega \in \mathbb{R}$, Theorem 2.2 implies that the matrix $\boldsymbol{M}_0(2\omega)\,\widetilde{\boldsymbol{P}}(\omega)\,\boldsymbol{M}_0(\omega)^{-1}$ also provides approximation order m. Furthermore,

$$\boldsymbol{M}_0(0)\,\widetilde{\boldsymbol{P}}(0)\,\boldsymbol{M}_0(0)^{-1}\,\boldsymbol{e} = \frac{1}{c}\,\boldsymbol{M}_0(0)\,\widetilde{\boldsymbol{P}}(0)\boldsymbol{r} = \frac{1}{c}\,\boldsymbol{M}_0(0)\boldsymbol{r} = \boldsymbol{e}.$$

Now, we are ready to apply Theorem 2.6 to the matrix $\boldsymbol{M}_0(2\omega)\,\widetilde{\boldsymbol{P}}(\omega)\,\boldsymbol{M}_0(\omega)^{-1}$, yielding that

$$\begin{aligned}\boldsymbol{P}(\omega) &= \frac{1}{2}\,\boldsymbol{M_r}(2\omega)\,\widetilde{\boldsymbol{P}}(\omega)\,\boldsymbol{M_r}(\omega)^{-1}\\ &= \frac{1}{2}\,\boldsymbol{C_u}(2\omega)\,\boldsymbol{M}_0(2\omega)\,\widetilde{\boldsymbol{P}}(\omega)\,\boldsymbol{M}_0(\omega)^{-1}\,\boldsymbol{C_u}(\omega)^{-1}\end{aligned}$$

provides approximation order at least $m+1$. The construction of $\boldsymbol{u}_k$ $(k = 0,\dots,m)$ follows from Theorems 2.6 and 2.2. □

Remarks. 1. Let us mention that a degenerate TST with $D(\det \boldsymbol{M})(0) \neq 0$ can change the approximation order only by one. This fact does not follow directly from Theorems 2.4 or 2.7. Only together do these theorems imply it.

2. In particular, we obtain that, in Theorem 2.7, the vector $\boldsymbol{u}_0$ is a multiple of $\boldsymbol{u}$, since $\boldsymbol{u}_0^{\mathrm{T}} = \widetilde{\boldsymbol{u}}_0^{\mathrm{T}}\,\boldsymbol{G_u}(0)$.

Repeated application of Theorem 2.7 yields the following corollary.

COROLLARY 2.8. *Suppose that a matrix $\boldsymbol{P}^{(0)}(\omega) \in C^{m-1}_{2\pi}(\mathbb{R}^{r\times r})$ is given. Moreover, let*

$$\boldsymbol{P}^{(0)}(0)\boldsymbol{r}_0 = \boldsymbol{r}_0, \quad \boldsymbol{x}_0^T\boldsymbol{P}^{(0)}(0) = \boldsymbol{x}_0^T, \quad \boldsymbol{x}_0^T\boldsymbol{P}^{(0)}(\pi) \neq 0$$

for some $\boldsymbol{x}_0, \boldsymbol{r}_0 \in \mathbb{R}^r$. For $n = 1,\dots,m$, construct the matrices

$$\boldsymbol{P}^{(n)}(\omega) := \frac{1}{2}\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(2\omega)\boldsymbol{P}^{(n-1)}(\omega)\boldsymbol{M}^{-1}_{\boldsymbol{r}_{n-1}}(\omega).$$

Here $\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(\omega)$ are chosen such that

1. *$\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(\omega)$ is invertible for all $\omega \neq 0$ and $D(\det \boldsymbol{M}_{\boldsymbol{r}_{n-1}})(0) \neq 0$;*
2. *$\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(0)$ has a simple eigenvalue 0 with a right eigenvector $\boldsymbol{r}_{n-1}$,*

$$\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(0)\boldsymbol{r}_{n-1} = \boldsymbol{0},$$

where $\boldsymbol{r}_{n-1}$ is the 1-eigenvector of $\boldsymbol{P}^{(n-1)}(0)$, i.e., $\boldsymbol{P}^{(n-1)}(0)\boldsymbol{r}_{n-1} = \boldsymbol{r}_{n-1}$.

Then there exist vectors $\boldsymbol{y}_0,\dots,\boldsymbol{y}_{m-1}$ $(\boldsymbol{y}_0 \neq \boldsymbol{0})$ such that the matrix $\boldsymbol{P}^{(m)}$

$$\boldsymbol{P}^{(m)}(\omega) := \frac{1}{2^m}\,\boldsymbol{M}_{\boldsymbol{r}_{m-1}}(2\omega)\cdots\boldsymbol{M}_{\boldsymbol{r}_0}(2\omega)\,\boldsymbol{P}^{(0)}(\omega)\,\boldsymbol{M}_{\boldsymbol{r}_0}(\omega)^{-1}\cdots\boldsymbol{M}_{\boldsymbol{r}_{m-1}}(\omega)^{-1}$$

provides approximation order m with $\boldsymbol{y}_0,\dots,\boldsymbol{y}_{m-1}$.

Corollary 2.8 opens an easy way to construct multiscaling functions with given approximation order. We discuss it in section 3. Note that the bulky formulas in Theorems 2.6 and 2.7 for $\boldsymbol{y}_k$ and $\boldsymbol{u}_k$ are only of theoretical interest. They will be used for the proof, but they need not be computed during the construction.

2.5. Regularity of multiscaling functions. In the scalar case, the approximation properties of the refinement mask are closely related with regularity of the scaling function. What happens in the vector case? To give an answer to this question we recall results from [CDP, S1].

Let $\boldsymbol{v}$ be a right eigenvector of $\boldsymbol{P}(0)$ for the eigenvalue 1. We introduce the spectral radius of $\boldsymbol{P}(0)$,

$$\rho(\boldsymbol{P}(0)) := \max\{|\lambda| : \boldsymbol{P}(0)\,\boldsymbol{x} = \lambda\,\boldsymbol{x},\ \boldsymbol{x} \neq \boldsymbol{0}\}.$$

Suppose that $\rho(\boldsymbol{P}(0)) < 2$. Then

$$\widehat{\boldsymbol{\Upsilon}}(\omega) := \lim_{n\to\infty} \Pi_{j=1}^{n} \boldsymbol{P}\left(\frac{\omega}{2^j}\right) \boldsymbol{v} \tag{2.21}$$

converges pointwise for all ω, and the convergence is uniform on compact sets (see [CDP]). Moreover, the following theorem holds.

THEOREM 2.9 (see [CDP]). *Let $\boldsymbol{P}$ be an $r \times r$ matrix of the form*

$$\boldsymbol{P}(\omega) = \frac{1}{2^m}\,\boldsymbol{C}_{\boldsymbol{x}_{m-1}}(2\omega)\cdots\boldsymbol{C}_{\boldsymbol{x}_0}(2\omega)\,\boldsymbol{P}^{(0)}(\omega)\,\boldsymbol{C}_{\boldsymbol{x}_0}(\omega)^{-1}\cdots\boldsymbol{C}_{\boldsymbol{x}_{m-1}}(\omega)^{-1},$$

where $\boldsymbol{C}_{\boldsymbol{x}_k}$ are defined by the vectors $\boldsymbol{x}_k \neq \boldsymbol{0}$ $(k = 0,\dots,m-1)$ via (2.6)–(2.7) *and $\boldsymbol{P}^{(0)}(\omega)$ is an $r \times r$ matrix with trigonometric polynomials as entries. Suppose that $\boldsymbol{P}^{(0)}(0)\boldsymbol{e}_0 = \boldsymbol{e}_0$, where $\boldsymbol{e}_0$ is defined by $\boldsymbol{x}_0$ via* (2.10). *Further, suppose that $\rho(\boldsymbol{P}^{(0)}(0)) < 2$, and let, for $k \geq 1$,*

$$\gamma_k := \frac{1}{k}\,\log_2 \sup_{\omega} \left\|\boldsymbol{P}^{(0)}\left(\frac{\omega}{2}\right)\cdots\boldsymbol{P}^{(0)}\left(\frac{\omega}{2^k}\right)\right\|. \tag{2.22}$$

Then there exists a constant $C > 0$ such that for all $\omega \in \mathbb{R}$,

$$\|\widehat{\boldsymbol{\Upsilon}}(\omega)\| \leq C\,(1+|\omega|)^{-m+\gamma_k},$$

where $\|\widehat{\boldsymbol{\Upsilon}}(\omega)\|$ denotes the Euclidean norm of $\widehat{\boldsymbol{\Upsilon}}(\omega) := (\widehat{\Upsilon}_\nu(\omega))_{\nu=0}^{r-1}$. Hence, if $\gamma_k < m-d$ $(d \in \mathbb{N})$, then Υ_ν $(\nu = 0,\dots,r-1)$ are $d-1$ times continuously differentiable.

If the conditions of Theorem 2.9 are satisfied and $\inf_{k\geq 1}\gamma_k < m-1$, then a compactly supported continuous solution $\boldsymbol{\Upsilon}(t)$ of (1.1) is unique in a wide class of functions. Further, the uniform convergence of the cascade algorithm (in time domain) is ensured (see [CDP, Sh]). Using the techniques from [S1] we obtain the following result.

COROLLARY 2.10. *Assume that for $n = 1,\dots,m$, $\boldsymbol{P}^{(n)}(\omega)$ is of the form*

$$\boldsymbol{P}^{(n)}(\omega) = \frac{1}{2^n}\,\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(2\omega)\cdots\boldsymbol{M}_{\boldsymbol{r}_0}(2\omega)\,\boldsymbol{P}^{(0)}(\omega)\,\boldsymbol{M}_{\boldsymbol{r}_0}(\omega)^{-1}\cdots\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(\omega)^{-1}.$$

Let $\boldsymbol{P}^{(0)}(\omega)$, $\boldsymbol{P}^{(n)}(\omega)$ and $\boldsymbol{M}_{\boldsymbol{r}_{n-1}}(\omega)$ $(n = 1,\dots,m)$ satisfy the assumptions of Corollary 2.8. *Further, suppose that $\rho(\boldsymbol{P}^{(0)}(0)) < 2$ and $\inf_{k\geq 1}\gamma_k < m-d$ $(d \in \mathbb{N})$, where γ_k is defined in* (2.22). *Then, $\boldsymbol{\Upsilon}(t)$ is a compactly supported $d-1$ times continuously differentiable solution of* (1.1) *with refinement mask $\boldsymbol{P}^{(m)}(\omega)$ providing approximation order at least m.*

Similar to the scalar case, the regularity of multiscaling functions depends both on the approximation order and the behavior of the residual $\boldsymbol{P}^{(0)}(\omega)$. Roughly speaking,

each approximation order adds one more derivative to the corresponding function vector, but the starting number of the derivatives depends on the $\boldsymbol{P}^{(0)}(\omega)$.

LEMMA 2.11. *Let $\boldsymbol{P}(\omega)$ be the refinement mask of a compactly supported continuously differentiable function vector $\boldsymbol{\phi} \in C^1(\mathbb{R}^r)$ providing approximation order at least 1; i.e., there exists a vector $\boldsymbol{y} \in \mathbb{R}^r$, $\boldsymbol{y} \neq \boldsymbol{0}$, such that*

$$\boldsymbol{y}^T \boldsymbol{P}(0) = \boldsymbol{y}^T, \qquad \boldsymbol{y}^T \boldsymbol{P}(\pi) = \boldsymbol{0}^T.$$

Further, assume that $\boldsymbol{P}(0)$ has a spectrum of the form $\{1, \mu_1, \ldots, \mu_{r-1}\}$ with each $\mu_\nu < 1/2$. Let $\boldsymbol{M}(\omega) \in C^1_{2\pi}(\mathbb{R})$ be an $r \times r$ matrix satisfying assumptions 1, 2 of Theorem 2.4 (with $\boldsymbol{y}$ instead of $\boldsymbol{y}_0$). Then

$$\widetilde{\boldsymbol{P}}(\omega) := 2\, \boldsymbol{M}(2\omega)^{-1}\, \boldsymbol{P}(\omega)\, \boldsymbol{M}(\omega) \tag{2.23}$$

is the refinement mask of a continuous function vector $\boldsymbol{\psi} = (\psi_\nu)_{\nu=0}^{r-1} \in C(\mathbb{R}^r) \cap L^1(\mathbb{R})$ and there is a constant $c_0 \in \mathbb{R}$ such that

$$\widehat{\boldsymbol{\phi}}(\omega) = \frac{c_0}{i\omega}\, \boldsymbol{M}(\omega) \widehat{\boldsymbol{\psi}}(\omega).$$

In particular, if $\boldsymbol{M} = \boldsymbol{C_y} \boldsymbol{M}_0$, with $\boldsymbol{C_y}$ defined by $\boldsymbol{y}$ as in (2.6)–(2.7) and a constant invertible matrix $\boldsymbol{M}_0$, then $\boldsymbol{\psi}$ is also compactly supported.

Proof. 1. Let us start with the case when $\widetilde{\boldsymbol{P}}(\omega) := 2\, \boldsymbol{C_y}(2\omega)^{-1}\, \boldsymbol{P}(\omega)\, \boldsymbol{C_y}(\omega)$ and $\boldsymbol{C_y}$ is defined by $\boldsymbol{y}$ as in (2.6)–(2.7). The assumptions on the spectrum of $\boldsymbol{P}(0)$ and the results of [CDP, S1] imply that $\rho(\widetilde{\boldsymbol{P}}(0)) = 1$, and 1 is a simple eigenvalue of $\widetilde{\boldsymbol{P}}(0)$. Hence, we can represent $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\psi}}$ in the form

$$\widehat{\boldsymbol{\phi}}(\omega) := \prod_{j=1}^{\infty} \boldsymbol{P}\left(\frac{\omega}{2^j}\right) \boldsymbol{a}, \qquad \widehat{\boldsymbol{\psi}}(\omega) := \prod_{j=1}^{\infty} \widetilde{\boldsymbol{P}}\left(\frac{\omega}{2^j}\right) \boldsymbol{b}, \tag{2.24}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are right eigenvectors of $\boldsymbol{P}(0)$ and $\widetilde{\boldsymbol{P}}(0)$, respectively. The convergence of the products in (2.24) is ensured by Theorem 3.2 in [CDP]. The observations in [P3] imply that $\widetilde{\boldsymbol{P}}(\omega)$ is a matrix of trigonometric polynomials ensuring a compactly supported solution $\boldsymbol{\psi}(t)$ of (1.1). The solutions $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are uniquely determined by (2.24) up to a constant factor [CDP, H, HC].

By the repeated substitution of (2.23) into (2.24) we get

$$\begin{aligned}
\widehat{\boldsymbol{\phi}}(\omega) &= \lim_{n\to\infty} \left(\prod_{j=1}^{n} \frac{1}{2}\, \boldsymbol{C_y}\left(\frac{2\omega}{2^j}\right) \widetilde{\boldsymbol{P}}\left(\frac{\omega}{2^j}\right) \boldsymbol{C_y}\left(\frac{\omega}{2^j}\right)^{-1} \right) \boldsymbol{a} \\
&= \boldsymbol{C_y}(\omega) \lim_{n\to\infty} \frac{1}{2^n} \left(\prod_{j=1}^{n} \widetilde{\boldsymbol{P}}\left(\frac{\omega}{2^j}\right) \right) \boldsymbol{C_y}\left(\frac{\omega}{2^n}\right)^{-1} \boldsymbol{a}.
\end{aligned}$$

Formula (2.9) gives

$$\widehat{\boldsymbol{\phi}}(\omega) = \lim_{n\to\infty} \frac{1}{2^n\,(1 - e^{-i\omega/2^n})}\, \boldsymbol{C_y}(\omega) \prod_{j=1}^{n} \widetilde{\boldsymbol{P}}\left(\frac{\omega}{2^j}\right) \boldsymbol{G_y}\left(\frac{\omega}{2^n}\right) \boldsymbol{a}.$$

2. Replacing $\boldsymbol{C_y}(\omega)$ and $\boldsymbol{C_y}(2\omega)^{-1}$ by $(1 - e^{-i\omega})\, \boldsymbol{G_y}(\omega)^{-1}$ and $(1 - e^{-2i\omega})^{-1}$ $\boldsymbol{G_y}(2\omega)$, respectively, we obtain from (2.23) (with $\boldsymbol{M} = \boldsymbol{C_y}$) that

$$\frac{1}{2}\,(1 + e^{-i\omega})\, \widetilde{\boldsymbol{P}}(\omega)\, \boldsymbol{G_y}(\omega) = \boldsymbol{G_y}(2\omega)\, \boldsymbol{P}(\omega).$$

In particular, for $\omega = 0$, it follows that $\widetilde{\boldsymbol{P}}(0)\,\boldsymbol{G_y}(0) = \boldsymbol{G_y}(0)\,\boldsymbol{P}(0)$. Hence, $\boldsymbol{G_y}(0)\,\boldsymbol{a}$ is a right eigenvector of $\widetilde{\boldsymbol{P}}(0)$, and there is a constant c_0 such that

$$\boldsymbol{G_y}(0)\,\boldsymbol{a} = c_0\,\boldsymbol{b}.$$

Observing that $\lim_{n\to\infty} 2^{-n}(1 - e^{-i\omega/2^n})^{-1} = (i\omega)^{-1}$, we get

$$\widehat{\boldsymbol{\phi}}(\omega) = \frac{c_0}{i\omega}\,\boldsymbol{C_y}(\omega) \prod_{j=1}^{\infty} \widetilde{\boldsymbol{P}}\left(\frac{\omega}{2^j}\right) \boldsymbol{b} = \frac{c_0}{i\omega}\,\boldsymbol{C_y}(\omega)\,\widehat{\boldsymbol{\psi}}(\omega).$$

Now take a refinement mask $\boldsymbol{P}(\omega)$ of a compactly supported function vector $\boldsymbol{\phi} \in C^1(\mathbb{R}^r)$ and an arbitrary matrix $\boldsymbol{M} \in \boldsymbol{C}^1_{2\pi}(\mathbb{R}^{r\times r})$ corresponding to $\boldsymbol{P}$ such that $\boldsymbol{M}$ satisfies conditions 1, 2 of Theorem 2.4 (with $\boldsymbol{y}$ instead of $\boldsymbol{y}_0$). Then, by Corollary 2.10,

$$\widetilde{\boldsymbol{P}}(\omega) = 2\,\boldsymbol{M}^{-1}_{\boldsymbol{y}}(2\omega)\,\boldsymbol{P}(\omega)\,\boldsymbol{M_y}(\omega)$$

is a refinement mask of a continuous function vector $\boldsymbol{\psi} \in C(\mathbb{R}^r)$. Using Lemma 2.5 we can prove the relation

$$\widehat{\boldsymbol{\phi}}(\omega) = \frac{c_0}{i\omega}\,\boldsymbol{M_y}(\omega)\,\widehat{\boldsymbol{\psi}}(\omega)$$

with an arbitrary chosen constant c_0 in the same manner as above. □

Using the spectral properties of transition operators, more results on regularity can be obtained [CDP, Sh, J].

2.6. Symmetry of multiscaling functions. In many applications, symmetry of the scaling functions is very desirable. Unfortunately, this property is very restrictive, and in the scalar case symmetry cannot be combined with orthogonality. In the vector case, there is more freedom, and the components of a refinable function vector can be symmetric and orthogonal at the same time. One such example was constructed in [GHM] and is shown in Figure 1.1. In this section we are going to discuss some results on symmetry of multiscaling functions. All details can be found in [S1].

We say that a refinable function vector $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$ *is symmetric* if all its components $\phi_\nu(t)$ are symmetric or antisymmetric. Symmetry implies some restrictions on a refinement mask $\boldsymbol{P}(\omega)$.

THEOREM 2.12 (see [S1]). *If there is a diagonal matrix*

$$\boldsymbol{E}(\omega) := \operatorname{diag}\left(\pm e^{-i2T_0\omega}, \ldots, \pm e^{-i2T_{r-1}\omega}\right)$$

such that the refinement mask $\boldsymbol{P}(\omega)$ of a refinable function vector $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$ satisfies

$$\boldsymbol{P}(\omega) = \boldsymbol{E}(2\omega)\boldsymbol{P}(-\omega)\boldsymbol{E}(\omega)^{-1}, \tag{2.25}$$

then $\boldsymbol{\phi}$ is symmetric. The constants T_ν occurring in $\boldsymbol{E}(\omega)$ are points of symmetry of the components $\phi_\nu(t)$, i.e., $\phi_\nu(T_\nu - t) = \pm\phi_\nu(T_\nu + t)$.

While constructing a vector of multiscaling functions using Corollary 2.8, it is reasonable to start with a symmetric one and try to preserve the symmetry at each step (see section 3). The following theorem specifies the factorization matrices $\boldsymbol{M}(\omega)$ which preserve the symmetry.

THEOREM 2.13 (see [S1]). *Suppose that all components $\widetilde{\phi}_\nu(t)$ of a refinable function vector $\widetilde{\boldsymbol{\phi}} = (\widetilde{\phi}_\nu)_{\nu=0}^{r-1}$ are symmetric (or antisymmetric) with points of symmetry $\widetilde{T}_\nu$ determining*

$$\widetilde{\boldsymbol{E}}(\omega) := \operatorname{diag}\left(\pm e^{-i2\widetilde{T}_0\omega}, \ldots, \pm e^{-i2\widetilde{T}_{r-1}\omega}\right). \tag{2.26}$$

Take a matrix $\boldsymbol{M}(\omega) \in C_{2\pi}(\mathbb{R}^{r\times r})$ satisfying assumptions 1, 2 of Theorem 2.4 and a matrix

$$\boldsymbol{E}(\omega) := \operatorname{diag}\left(\pm e^{-i2T_0\omega}, \ldots, \pm e^{-i2T_{r-1}\omega}\right) \tag{2.27}$$

such that

$$\boldsymbol{M}(\omega) = -\boldsymbol{E}(\omega)\boldsymbol{M}(-\omega)\widetilde{\boldsymbol{E}}^{-1}(\omega). \tag{2.28}$$

Then the new vector $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$, determined by $\widehat{\boldsymbol{\phi}}(\omega) = \frac{c_0}{i\omega}\boldsymbol{M}(\omega)\widehat{\widetilde{\boldsymbol{\phi}}}(\omega)$, is also symmetric and T_ν, $\nu = 0, \ldots, r-1$ are points of symmetry of its components.

Remark. Let us mention that if ϕ_ν has finite support l_ν, starting at point $t_1 \geq 0$, and T_ν is the point of symmetry of ϕ_ν, then $l_\nu \leq 2T_\nu$.

3. Construction of multiscaling functions. Finally we have reached the point where we can show how to construct refinement masks which yield multiscaling functions with desirable properties.

In the scalar case, there is no problem finding a mask providing any given order of accuracy. One can start with a trigonometric polynomial $P(\omega)$ such that $P(0) = 1$ and multiply by a power of $\frac{1}{2}(1 + e^{-i\omega})$ (see, e.g., [St1]). In the vector case, a TST with transformation matrix $\boldsymbol{M}(\omega)$ (as described in Theorem 2.7) plays the role of the factor $(1 + e^{-i\omega})$.

An algorithm for the construction of refinement masks, yielding multiscaling functions with given approximation order, can be obtained as a consequence of Corollary 2.8.

ALGORITHM 3.1. *Start with a matrix trigonometric polynomial $\boldsymbol{P}^{(n)}(\omega)$ providing approximation order $n \in \mathbb{N}_0$ such that $\rho(\boldsymbol{P}^{(n)}(0)) < 2$. Further, let $\boldsymbol{P}^{(n)}(0)$ possess an eigenvalue 1 with corresponding right eigenvector $\boldsymbol{r}_n$, i.e., $\boldsymbol{P}^{(n)}(0)\boldsymbol{r}_n = \boldsymbol{r}_n$.*

1. *Choose $\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ such that:*
   (a) $\det \boldsymbol{M}_{\boldsymbol{r}_n}(\omega) \neq 0$ *for* $\omega \neq 0$,
   (b) $D(\det \boldsymbol{M}_{\boldsymbol{r}_n})(0) \neq 0$,
   (c) $\boldsymbol{M}_{\boldsymbol{r}_n}(0)\boldsymbol{r}_n = \boldsymbol{0}$.
2. *Construct the matrix $\boldsymbol{P}^{(n+1)}(\omega)$:*

$$\boldsymbol{P}^{(n+1)}(\omega) := \frac{1}{2}\boldsymbol{M}_{\boldsymbol{r}_n}(2\omega)\boldsymbol{P}^{(n)}(\omega)\boldsymbol{M}_{\boldsymbol{r}_n}^{-1}(\omega).$$

3. *Find a right eigenvector $\boldsymbol{r}_{n+1}$ corresponding to the eigenvalue 1 of $\boldsymbol{P}^{(n+1)}(0)$.*
4. *Repeat steps 1, 2, 3 as many times as needed.*

By Theorem 2.7, the approximation order of $\boldsymbol{P}^{(n+1)}(\omega)$ is $n+1$, and $m-n$ cycles of Algorithm 3.1 are needed to get a refinement mask $\boldsymbol{P}^{(m)}$ providing approximation order m. In [S1], it was proven that $\boldsymbol{P}^{(n+1)}(0)$ has eigenvalue 1, so step 4 is consistent.

One can see that there are two matrices to be chosen in Algorithm 3.1, the starting matrix $\boldsymbol{P}^{(n)}(\omega)$ (only once in the beginning) and the transformation matrix $\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ (one on each cycle of the algorithm).

Corollary 2.10 shows that the regularity of the final function vector (determined by the refinement mask $\boldsymbol{P}^{(m)}(\omega)$) is governed by its approximation order m and by the properties of the starting matrix $\boldsymbol{P}^{(n)}(\omega)$.

The approximation order n implies that $\boldsymbol{P}^{(n)}$ can be factored:

$$\boldsymbol{P}^{(n)}(\omega) = \frac{1}{2^n}\, \boldsymbol{C}_{\boldsymbol{x}_{n-1}}(2\omega) \cdots \boldsymbol{C}_{\boldsymbol{x}_0}(2\omega)\, \boldsymbol{P}^{(0)}(\omega)\, \boldsymbol{C}_{\boldsymbol{x}_0}(\omega)^{-1} \cdots \boldsymbol{C}_{\boldsymbol{x}_{n-1}}(\omega)^{-1},$$

where $\boldsymbol{C}_{\boldsymbol{x}_k}$ are defined by vectors $\boldsymbol{x}_k \neq \boldsymbol{0}$ via (2.6)–(2.7). Further, the spectral radii of $\boldsymbol{P}^{(0)}(\omega)$ and $\boldsymbol{P}^{(k)}(0)$,

$$\boldsymbol{P}^{(k)}(\omega) := \frac{1}{2^k}\, \boldsymbol{C}_{\boldsymbol{x}_{k-1}}(2\omega) \cdots \boldsymbol{C}_{\boldsymbol{x}_0}(2\omega)\, \boldsymbol{P}^{(0)}(\omega)\, \boldsymbol{C}_{\boldsymbol{x}_0}(\omega)^{-1} \cdots \boldsymbol{C}_{\boldsymbol{x}_{k-1}}(\omega)^{-1} \quad (k \le n),$$

are related as follows [CDP, S1]:

$$\rho(\boldsymbol{P}^{(k)}(0)) = \max\{1,\, 2^{-k}\, \rho(\boldsymbol{P}^{(0)}(0))\} \quad (k = 0, \dots, n).$$

Let k_0 $(0 \le k_0 \le n)$ be the smallest integer such that $\rho(\boldsymbol{P}^{(k_0)}(0)) < 2$. Then by Theorem 2.9, it follows that the Fourier transformed solution vector $\widehat{\boldsymbol{\phi}}_n$ of (1.3), determined by $\boldsymbol{P}^{(n)}$, satisfies

$$\|\widehat{\boldsymbol{\phi}}_n(\omega)\| \le C\, (1 + |\omega|)^{-n+k_0+K_0},$$

where $K_0 := \inf_{l \ge 1} \gamma_l$, $\gamma_l = \frac{1}{l} \log_2 \sup_\omega \|\boldsymbol{P}^{(k_0)}(\frac{\omega}{2}) \cdots \boldsymbol{P}^{(k_0)}(\frac{\omega}{2^l})\|$. Thus, $m - n$ cycles of Algorithm 3.1 yield $\boldsymbol{P}^{(m)}$ providing a solution vector $\widehat{\boldsymbol{\phi}}_m(\omega)$ such that

$$\|\hat{\boldsymbol{\phi}}_m(\omega)\| \le C\, (1 + |\omega|)^{-m+k_0+K_0}.$$

So, if we want to get a multiscaling function with approximation order at least m and p derivatives, we need to apply $m_0 - n$ cycles of Algorithm 3.1, where m_0 is chosen such that $m_0 \ge \max\{m,\, k_0 + K_0 + p + 1\}$.

3.1. How to choose the transformation matrices $\boldsymbol{M}_{\boldsymbol{r}_n}(\boldsymbol{\omega})$. In the scalar case, $\boldsymbol{M}_{\boldsymbol{r}_k}(\omega) = (1 - e^{-i\omega})$ is fixed. In the vector case, we are flexible in the choice of $\boldsymbol{M}_{\boldsymbol{r}_k} \in C_{2\pi}(\mathbb{R}^{r \times r})$. Actually, only one eigenvalue and one eigenvector are restricted in $\boldsymbol{M}_{\boldsymbol{r}_k}(\omega)$. We can use this freedom to obtain multiscaling functions with desired properties.

Finite support. A refinement mask $\boldsymbol{P}^{(n+1)}(\omega)$ corresponds to a finitely supported scaling vector if all components of $\boldsymbol{P}^{(n+1)}(\omega)$ are trigonometric polynomials (algebraic polynomials in $z = e^{-i\omega}$) [MRV]. But

$$\boldsymbol{P}^{(n+1)}(\omega) = \frac{1}{2} \boldsymbol{M}_{\boldsymbol{r}_n}(2\omega) \boldsymbol{P}^{(n)}(\omega) \boldsymbol{M}_{\boldsymbol{r}_n}^{-1}(\omega) \tag{3.1}$$

contains $\boldsymbol{M}_{\boldsymbol{r}_n}(2\omega)$ and $\boldsymbol{M}_{\boldsymbol{r}_n}^{-1}(\omega)$ which generally are not matrices of trigonometric polynomials at the same time.

LEMMA 3.2. *Assume that $\boldsymbol{P}^{(n)}(\omega)$ is a matrix of trigonometric polynomials. If $\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ satisfies conditions* (a)–(c) *of Algorithm* 3.1, *$\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ is a matrix of trigonometric polynomials, and $\det \boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ is linear in $z = e^{-i\omega}$, then the components of $\boldsymbol{P}^{(n+1)}(\omega)$ in* (3.1) *are trigonometric polynomials.*

Proof. Let us use a well-known formula for an inverse matrix:

$$\boldsymbol{M}_{\boldsymbol{r}_n}^{-1}(\omega) = \frac{1}{\det \boldsymbol{M}_{\boldsymbol{r}_n}(\omega)} \boldsymbol{N}_{\boldsymbol{r}_n}(\omega). \tag{3.2}$$

Here the (i,j) element of the matrix $\boldsymbol{N}_{\boldsymbol{r}_n}(\omega)$ is the minor for the (j,i) element of $\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ (see [St2, p. 225]). In particular, $\boldsymbol{N}_{\boldsymbol{r}_n}(\omega)$ contains only trigonometric polynomials.

Since $\det \boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ is linear in z, and $\det \boldsymbol{M}_{\boldsymbol{r}_n}(0) = 0$, we have

$$\det \boldsymbol{M}_{\boldsymbol{r}_n}(\omega) = c_0(1 - e^{-i\omega}),$$

with a constant $c_0 \neq 0$, and according to (3.2),

$$\boldsymbol{P}^{(n+1)}(\omega) = \frac{1}{2\, c_0\, (1 - e^{-i\omega})} \boldsymbol{M}_{\boldsymbol{r}_n}(2\omega) \boldsymbol{P}^{(n)}(\omega) \boldsymbol{N}_{\boldsymbol{r}_n}(\omega). \tag{3.3}$$

It is easy to see that the components of $\boldsymbol{M}_{\boldsymbol{r}_n}(2\omega)\boldsymbol{P}^{(n)}(\omega)\boldsymbol{N}_{\boldsymbol{r}_n}(\omega)$ are trigonometric polynomials. In [S1], it was proven that $\boldsymbol{P}^{(n+1)}(0)$ is bounded. On the other hand, $(1 - e^{-i\omega})^{-1}$ is infinite at $\omega = 0$. Thus, all components of $\boldsymbol{M}_{\boldsymbol{r}_n}(2\omega)\boldsymbol{P}^{(n)}(\omega)\boldsymbol{N}_{\boldsymbol{r}_n}(\omega)$ must possess a root at $\omega = 0$ or, in other words, must be divisible by $(1 - e^{-i\omega})$. Hence, reducing $\boldsymbol{M}_{\boldsymbol{r}_n}(2\omega)\boldsymbol{P}^{(n)}(\omega)\boldsymbol{N}_{\boldsymbol{r}_n}(\omega)$ by $(1 - e^{-i\omega})$, we get a matrix trigonometric polynomial $\boldsymbol{P}^{(n+1)}(\omega)$. □

One way to choose $\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ satisfying the conditions of Algorithm 3.1 and Lemma 3.2 is given by Lemma 2.5. Take an arbitrary vector $\boldsymbol{y}_n = (y_{n,\nu})_{\nu=0}^{r-1}$ corresponding to $\boldsymbol{r}_n = (r_{n,\nu})_{\nu=0}^{r-1}$ in the sense that $y_{n,\nu} \neq 0$ if and only if $r_{n,\nu} \neq 0$ for $\nu = 0, \ldots, r-1$. Put

$$\boldsymbol{M}_{\boldsymbol{r}_n}(\omega) := \boldsymbol{C}_{\boldsymbol{y}_n}(\omega)\, \boldsymbol{R}_n$$

with $\boldsymbol{C}_{\boldsymbol{y}_n}(\omega)$ defined by $\boldsymbol{y}_n$ as in (2.6)–(2.7) and an arbitrary constant $r \times r$ matrix $\boldsymbol{R}_n$ with the only restriction

$$\boldsymbol{R}_n\, \boldsymbol{r}_n = \boldsymbol{e}_n,$$

where $\boldsymbol{e}_n$ corresponds to $\boldsymbol{r}_n$ via (2.10). Then $\boldsymbol{M}_{\boldsymbol{r}_n}(\omega)$ is linear in $z = e^{-i\omega}$ by construction and, by (2.8), $\det \boldsymbol{M}_{\boldsymbol{r}_n}$ is of the desired form. Moreover, we have $\boldsymbol{M}_{\boldsymbol{r}_n}(0)\, \boldsymbol{r}_n = \boldsymbol{C}_{\boldsymbol{y}_n}(0)\, \boldsymbol{R}_n\, \boldsymbol{r}_n = \boldsymbol{C}_{\boldsymbol{y}_n}(0)\, \boldsymbol{e}_n = \boldsymbol{0}$. A simple $\boldsymbol{R}_n$ satisfying the relation above is $\boldsymbol{R}_n := \operatorname{diag}(\widetilde{r}_{n,0}, \ldots, \widetilde{r}_{n,r-1})$, where

$$\widetilde{r}_{n,\nu} := \begin{cases} 1/r_{n,\nu} & r_{n,\nu} \neq 0, \\ 1 & r_{n,\nu} = 0. \end{cases}$$

Symmetry. A reasonable way to get symmetric multiscaling functions with high approximation order is to start with $\boldsymbol{P}^{(n)}(\omega)$, yielding a symmetric function vector with low approximation order, and to preserve symmetry on each cycle of Algorithm 3.1. It is remarkable that after each cycle, the number of symmetric and antisymmetric components of the multiscaling function changes, independent of the choice of $\boldsymbol{M}_{\boldsymbol{r}_n}$:

LEMMA 3.3. *Suppose that $\boldsymbol{M}(\omega)$ satisfies conditions* (a)–(c) *of Algorithm* 3.1 *and a* TST *with transformation matrix $\boldsymbol{M}(\omega)$ preserves the symmetry; i.e., $\boldsymbol{\phi} = (\phi_\nu)_{\nu=1}^{r-1}$, $\widetilde{\boldsymbol{\phi}} = (\widetilde{\phi}_\nu)_{\nu=1}^{r-1}$ are two symmetric multiscaling functions connected by the relation*

$$\widehat{\boldsymbol{\phi}}(\omega) = \frac{c_0}{i\omega}\, \boldsymbol{M}(\omega)\, \widehat{\widetilde{\boldsymbol{\phi}}}(\omega). \tag{3.4}$$

Then, for even r, the difference in the number of antisymmetric components in $\widetilde{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}$ is odd, and for odd r, this difference is even.

Proof. Let $\boldsymbol{P}^{(n)}(\omega)$ be the refinement mask of $\widetilde{\boldsymbol{\phi}}$ and $\boldsymbol{P}^{(n+1)}(\omega)$ the refinement mask of $\boldsymbol{\phi}$, and let $\boldsymbol{P}^{(n)}$ and $\boldsymbol{P}^{(n+1)}$ be related as in Algorithm 3.1, with $\boldsymbol{M}_{\boldsymbol{r}_n} := \boldsymbol{M}$. Then (3.4) is a consequence of Lemma 2.11. By Theorem 2.13 we have

$$\boldsymbol{M}(\omega) = -\boldsymbol{E}(\omega)\, \boldsymbol{M}(-\omega)\, \widetilde{\boldsymbol{E}}^{-1}(\omega), \tag{3.5}$$

where $\boldsymbol{E}(\omega)$, $\widetilde{\boldsymbol{E}}(\omega)$ are defined by the points of symmetry T_ν, $\widetilde{T}_\nu$ of ϕ_ν, $\widetilde{\phi}_\nu$ ($\nu = 0, \dots, r-1$) via (2.26), (2.27). Since $\boldsymbol{M}(\omega)$ satisfies the conditions of Algorithm 3.1, $\det \boldsymbol{M}(\omega)$ has a simple zero at $\omega = 0$ such that

$$f(e^{-i\omega}) := \det \boldsymbol{M}(\omega) = (1 - e^{-i\omega}) f_0(e^{-i\omega}), \quad f_0(1) \neq 0. \tag{3.6}$$

From (3.5), (2.26), and (2.27), it follows that

$$\begin{aligned} \det \boldsymbol{M}(\omega) = f(e^{-i\omega}) &= (-1)^r \det \boldsymbol{E}(\omega) \cdot \det \widetilde{\boldsymbol{E}}^{-1}(\omega) \cdot \det \boldsymbol{M}(-\omega) \\ &= e^{-2iT\omega} f(e^{i\omega}) (-1)^{N+r}, \end{aligned} \tag{3.7}$$

where $T = \sum_{\nu=0}^{r-1} (T_\nu - \widetilde{T}_\nu)$, and N is the difference in the number of antisymmetric functions in $\boldsymbol{\phi}$ and $\widetilde{\boldsymbol{\phi}}$. Let $z := e^{-i\omega}$; then by (3.6)

$$f(z) = (1 - z) f_0(z)$$

and by (3.7)

$$f(z) = z^{2T} (-1)^{N+r} f\left(\frac{1}{z}\right) = z^{2T} (-1)^{N+r} \left(1 - \frac{1}{z}\right) f_0\left(\frac{1}{z}\right).$$

Combining these two relations we find

$$(1 - z) f_0(z) = -(-1)^{N+r} z^{2T-1} (1 - z) f_0\left(\frac{1}{z}\right)$$

and hence

$$f_0(z) = (-1)^{N+r+1} z^{2T-1} f_0\left(\frac{1}{z}\right). \tag{3.8}$$

But (3.8) implies that, if $N + r + 1$ is odd, then $f_0(1) = 0$ and thus $D(\det \boldsymbol{M})(0) = 0$, which contradicts the assumptions. So $N + r + 1$ must be even and $N + r$ must be odd. ∎

3.2. Examples. In this final section, we employ Algorithm 3.1 for the construction of multiscaling functions with high approximation order and other desirable properties.

Example 1. In the first example, we are going to increase the approximation order of the refinement mask $\boldsymbol{P}^{(2)}(\omega)$ corresponding to the Geronimo–Hardin–Massopust (GHM) multiscaling function $\boldsymbol{\phi} := [\phi_0\ \phi_1]^{\mathrm{T}}$ (see Figure 1.1):

$$\boldsymbol{P}^{(2)}(\omega) = \frac{1}{20} \begin{bmatrix} 6 + 6e^{-i\omega} & 8\sqrt{2} \\ (-1 + 9e^{-i\omega} + 9e^{-2i\omega} - e^{-3i\omega})/\sqrt{2} & -3 + 10e^{-i\omega} - 3e^{-2i\omega} \end{bmatrix}.$$

The functions $\phi_0(t)$, $\phi_1(t)$ are continuous, symmetric, and provide second-order approximation. The integer translates $\phi_0(t-l)$, $\phi_1(t-l)$ ($l \in \mathbb{Z}$) are orthogonal. It is easy to see that a 1-eigenvector of $\boldsymbol{P}^{(2)}(0)$ is $\boldsymbol{r}_2 = [\sqrt{2}\ \ 1]^{\mathrm{T}}$:

$$\boldsymbol{P}^{(2)}(0)\boldsymbol{r}_2 = \frac{1}{20}\begin{bmatrix} 12 & 8\sqrt{2} \\ 8\sqrt{2} & 4 \end{bmatrix}\begin{bmatrix} \sqrt{2} \\ 1 \end{bmatrix} = \begin{bmatrix} \sqrt{2} \\ 1 \end{bmatrix}.$$

Let us apply one cycle of Algorithm 3.1 to $\boldsymbol{P}^{(2)}(\omega)$ with transformation matrix $\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)$ preserving symmetry and ensuring short support. Then, $\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)$ must satisfy the assumptions of Lemma 3.2 and the following relation:

$$\boldsymbol{M}_{\boldsymbol{r}_2}(\omega) = -\boldsymbol{E}(\omega)\,\boldsymbol{M}_{\boldsymbol{r}_2}(-\omega)\,\widetilde{\boldsymbol{E}}^{-1}(\omega) \tag{3.9}$$

(cf. Theorem 2.13). The first GHM scaling function is symmetric about $\widetilde{T}_0 = 1/2$, and the second is symmetric about $\widetilde{T}_1 = 1$; hence $\widetilde{\boldsymbol{E}}(\omega) = \operatorname{diag}(e^{-i\omega}, e^{-2i\omega})$. In order to get the supports of the new scaling functions as short as possible, we choose $T_0 = T_1 = 1$. Thus, let $\boldsymbol{E}(\omega) = \operatorname{diag}(-e^{-2i\omega}, e^{-2i\omega})$. We put

$$\boldsymbol{M}_{\boldsymbol{r}_2}(\omega) := \begin{bmatrix} 1+e^{-i\omega} & -2\sqrt{2} \\ 1-e^{-i\omega} & 0 \end{bmatrix};$$

then (3.9) is satisfied. Moreover,

$$\boldsymbol{M}_{\boldsymbol{r}_2}(0)\boldsymbol{r}_2 = \begin{bmatrix} 2 & -2\sqrt{2} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \sqrt{2} \\ 1 \end{bmatrix} = \boldsymbol{0},$$

$\det \boldsymbol{M}_{\boldsymbol{r}_2}(\omega) = 2\sqrt{2}(1-e^{-i\omega}) \neq 0$ for $\omega \neq 0$, $D(\det \boldsymbol{M}_{\boldsymbol{r}_2})(0) = i2\sqrt{2} \neq 0$, so $\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)$ satisfies all conditions of Algorithm 3.1. $\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)$ is a matrix of trigonometric polynomials and $\det \boldsymbol{M}_{\boldsymbol{r}_2}(\omega) = 2\sqrt{2}(1-e^{-i\omega})$ is linear in $z = e^{-i\omega}$, so by Lemma 3.2, finite support for the new scaling functions is ensured.

Now we perform step 3 of Algorithm 3.1 and compute $\boldsymbol{P}^{(3)}(\omega)$:

$$\begin{aligned}\boldsymbol{P}^{(3)}(\omega) &= \frac{1}{2}\boldsymbol{M}_{\boldsymbol{r}_2}(2\omega)\boldsymbol{P}^{(2)}(\omega)\boldsymbol{M}_{\boldsymbol{r}_2}^{-1}(\omega) \\ &= \frac{1}{40}\begin{bmatrix} -7+10e^{-i\omega}-7e^{-2i\omega} & 15(1-e^{-2i\omega}) \\ -4(1-e^{-2i\omega}) & 10(1+e^{-i\omega})^2 \end{bmatrix}.\end{aligned}$$

The resulting scaling functions are continuously differentiable and provide approximation order 3. They are plotted in Figure 3.1.

The mask $\boldsymbol{P}^{(3)}(\omega)$ corresponds to a dilation equation (1.1) with three matrix coefficients:

$$\boldsymbol{P}_0 = \frac{1}{40}\begin{bmatrix} -7 & 15 \\ -4 & 10 \end{bmatrix},\quad \boldsymbol{P}_1 = \frac{1}{40}\begin{bmatrix} 10 & 0 \\ 0 & 20 \end{bmatrix},\quad \boldsymbol{P}_2 = \frac{1}{40}\begin{bmatrix} -7 & -15 \\ 4 & 10 \end{bmatrix}.$$

We mention that the GHM dilation equation has four coefficients since GHM functions ϕ_0, ϕ_1 have different supports.

Observe that, in accordance with Lemma 3.3, one scaling function is symmetric and the other is antisymmetric. Moreover, the sum of the supports grows exactly by 1.

Unfortunately, the new functions are not orthogonal and for practical applications a biorthogonal multiscaling function should be constructed [DM, SS4].

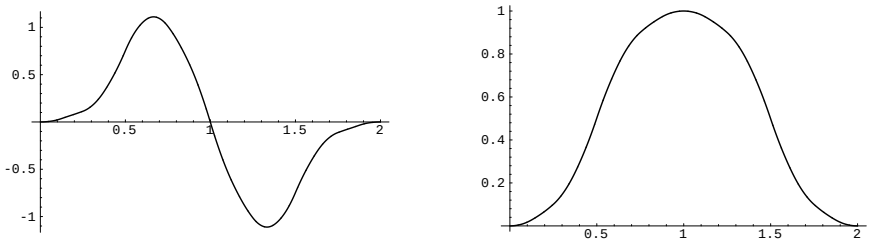

FIG. 3.1. *Symmetric multiscaling function with approximation order* 3.

Example 2. In the second example, we construct polynomial, symmetric multiscaling functions with two components, short support, and arbitrarily high approximation order. Let us start with the function vector $\boldsymbol{\phi}_2 := [\phi_{2,0}\ \phi_{2,1}]^{\mathrm{T}}$,

$$\phi_{2,0}(t) := \chi_{[0,1]}, \qquad \phi_{2,1}(t) := (1-2t)\,\chi_{[0,1]},$$

where $\chi_{[0,1]}$ denotes the characteristic function of $[0,\,1]$. The index 2 in $\boldsymbol{\phi}_2$ denotes the approximation order 2 provided by $\boldsymbol{\phi}_2$. Observe that both $\phi_{2,0}$ and $\phi_{2,1}$ are piecewise polynomials, but discontinuous; $\phi_{2,0}(1/2+t) = \phi_{2,0}(1/2-t)$ and $\phi_{2,1}(1/2+t) = -\phi_{2,1}(1/2-t)$. The vector $\boldsymbol{\phi}_2$ has the refinement mask

$$\boldsymbol{P}^{(2)}(\omega) := \frac{1}{4}\begin{bmatrix} 2+2z & 0 \\ 1-z & 1+z \end{bmatrix} \qquad (z := e^{-i\omega}),$$

with 1-eigenvector $\boldsymbol{r}_2 := [1\ \ 0]^{\mathrm{T}}$

$$\boldsymbol{P}^{(2)}(0)\,\boldsymbol{r}_2 = \frac{1}{4}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

We want to apply to $\boldsymbol{P}^{(2)}(\omega)$ one cycle of Algorithm 3.1 with a suitable transformation matrix $\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)$ which preserves symmetry and short support. We try to find $\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)$ satisfying the assumptions of Lemma 3.2 and such that

$$\boldsymbol{M}_{\boldsymbol{r}_2}(\omega) = -\operatorname{diag}(e^{-i\omega},\, e^{-2i\omega})\,\boldsymbol{M}_{\boldsymbol{r}_2}(-\omega)\,\operatorname{diag}(e^{-i\omega},\, -e^{-i\omega}).$$

Letting

$$\boldsymbol{M}_{\boldsymbol{r}_2}(\omega) := \begin{bmatrix} 0 & 2 \\ 1-z & -1-z \end{bmatrix} \quad (z = e^{-i\omega}),$$

we obtain, by applying Algorithm 3.1,

$$\boldsymbol{P}^{(3)}(\omega) = \frac{1}{2}\,\boldsymbol{M}_{\boldsymbol{r}_2}(2\omega)\,\boldsymbol{P}_2(\omega)\,\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)^{-1} = \frac{1}{8}\begin{bmatrix} 2(1+z) & 2 \\ 2z(1+z) & 1+4z+z^2 \end{bmatrix}.$$

The corresponding compactly supported function vector $\boldsymbol{\phi}_3 = [\phi_{3,0}\ \phi_{3,1}]^{\mathrm{T}}$ provides approximation order 3, since $\boldsymbol{M}_{\boldsymbol{r}_2}(\omega)$ satisfies all assumptions of Theorem 2.7. We

easily observe that

$$\phi_{3,0}(t) = \begin{cases} 2t(1-t) & t \in [0,1], \\ 0 & \text{otherwise}, \end{cases}$$
$$\phi_{3,1}(t) = \begin{cases} t^2 & t \in [0,1], \\ (2-t)^2 & t \in [1,2], \\ 0 & \text{otherwise}. \end{cases}$$

In particular, $\phi_{3,0}$ and $\phi_{3,1}$ are continuous functions. (This can also be seen by Corollary 2.10.)

Now we apply a second cycle of Algorithm 3.1 to $\boldsymbol{P}^{(3)}(\omega)$ in order to get a symmetric vector $\boldsymbol{\phi}_4$ of scaling functions $\phi_{4,0}$, $\phi_{4,1}$ with short support and approximation order 4. Observe that $\boldsymbol{P}^{(3)}(0)\,\boldsymbol{r}_3 = \boldsymbol{r}_3$ with $\boldsymbol{r}_3 := [1 \quad 2]^{\mathrm{T}}$, so the transformation matrix

$$\boldsymbol{M}_{\boldsymbol{r}_3}(\omega) := 3 \begin{bmatrix} 1-z & 0 \\ 1+z & -1 \end{bmatrix} \quad (z = e^{-i\omega})$$

satisfies the assumptions of Lemma 3.2, and we have

$$\boldsymbol{M}_{\boldsymbol{r}_3}(\omega) = -\operatorname{diag}(e^{-2i\omega}, -e^{-2i\omega})\, \boldsymbol{M}_{\boldsymbol{r}_3}(-\omega)\, \operatorname{diag}(e^{-i\omega}, e^{-2i\omega}).$$

We construct

$$\begin{aligned} \boldsymbol{P}^{(4)}(\omega) &= \frac{1}{2}\, \boldsymbol{M}_{\boldsymbol{r}_3}(2\omega)\, \boldsymbol{P}^{(3)}(\omega)\, \boldsymbol{M}_{\boldsymbol{r}_3}(\omega)^{-1} \\ &= \frac{1}{16} \begin{bmatrix} 4(1+z)^2 & -2(1-z)(1+z) \\ 3(1-z)(1+z) & -1+4z-z^2 \end{bmatrix}. \end{aligned}$$

The corresponding (compactly supported) functions $\phi_{4,0}$ and $\phi_{4,1}$ are again piecewise polynomials:

$$\phi_{4,0}(t) = \begin{cases} (-2t^3+3t^2) & t \in [0,1), \\ (2-t)^2\,(2t-1) & t \in [1,2], \\ 0 & \text{otherwise}, \end{cases}$$
$$\phi_{4,1}(t) = \begin{cases} t^2(3t-3) & t \in [0,1), \\ (2-t)^2\,(-3t-3) & t \in [1,2], \\ 0 & \text{otherwise}. \end{cases}$$

The functions $\phi_{4,0}$ and $\phi_{4,1}$ are symmetric, continuously differentiable functions. Note that $\phi_{4,0}$, $\phi_{4,1}$ are finite element functions studied in [SS3]. They are presented in Figure 3.2. Obviously, functions $\phi_{4,0}$ and $\phi_{4,1}$ are not orthogonal. For the construction of dual scaling functions and wavelets see [DM, SS4].

The procedure can be repeated as follows. Take

$$\boldsymbol{M}_{\boldsymbol{r}_{2k}}(\omega) := \begin{bmatrix} 0 & 2 \\ 1-z & -1-z \end{bmatrix} \qquad (k \in \mathbb{N},\ z = e^{-i\omega})$$

and

$$\boldsymbol{M}_{\boldsymbol{r}_{2k+1}}(\omega) := (2k+1) \begin{bmatrix} (1-z)/k & 0 \\ (1+z)/k & -2/(k+1) \end{bmatrix} \qquad (k \in \mathbb{N},\ z = e^{-i\omega})$$

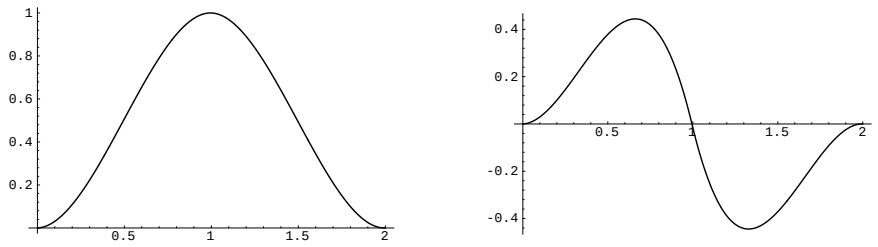

FIG. 3.2. *Polynomial multiscaling function with approximation order* 4.

and apply Algorithm 3.1 repeatedly with these transformation matrices. The refinement mask $\boldsymbol{P}^{(n)}$ ($n \in \mathbb{N}$; $n \geq 3$) then provides approximation order n; the corresponding multiscaling functions $\phi_{n,0}$ and $\phi_{n,1}$ are $(n-3)$-times continuously differentiable. If $n = 2k+1$ ($k \geq 1$), the corresponding multiscaling functions $\phi_{2k+1,0}$ and $\phi_{2k+1,1}$ are nothing but polynomial B-splines of order $2k+1$ with double knots, defined by the spline knots $0, 0, 1, 1, \ldots, k, k$ and $0, 1, 1, 2, 2, \ldots, k, k, k+1$, respectively. (This follows from a comparison with known recursion formulas for the refinement mask of B-splines vectors with multiple knots [P1, P2, P4]). In particular, $\operatorname{supp} \phi_{2k+1,0} = [0, k]$, $\operatorname{supp} \phi_{2k+1,1} = [0,\, k+1]$, and

$$\phi_{2k+1,0}(t) = \phi_{2k+1,0}(k-t), \qquad \phi_{2k+1,1}(t) = \phi_{2k+1,1}(k+1-t).$$

If $n = 2k$ ($k \geq 1$), the corresponding multiscaling functions $\phi_{2k,0}$ and $\phi_{2k,1}$ are nothing but polynomial B-splines of order $2k$, defined as the sum and the difference of the B-splines $N_{2k,0}$, $N_{2k,1}$ of order $2k$ with double knots, respectively. In other words, if $N_{2k,0}$ and $N_{2k,1}$ are defined by the spline knots $0, 0, \ldots, k-1, k-1, k$ and $0, 1, 1, \ldots, k-1, k, k$, then $\phi_{2k,0} = N_{2k,0} + N_{2k,1}$ and $\phi_{2k,1} = N_{2k,0} - N_{2k,1}$. In particular, $\operatorname{supp} \phi_{2k,0} = \operatorname{supp} \phi_{2k,1} = [0,\, k]$ and

$$\phi_{2k,0}(t) = \phi_{2k,0}(k-t), \qquad \phi_{2k,1}(t) = -\phi_{2k,1}(k-t).$$

Remark. For $r = 1$, the refinement mask $\boldsymbol{P}(\omega) = 2^{-m}(1+e^{-i\omega})^m$ determines the cardinal B-spline N_m of order m. Let $x_l := \lfloor l/r \rfloor$ ($l \in \mathbb{Z}$), where $\lfloor x \rfloor$ means the integer part of $x \in \mathbb{R}$. Then, the refinement mask

$$\boldsymbol{P}_m^r(\omega) := \frac{1}{2^m}\, \boldsymbol{C}_{\boldsymbol{x}_{m-1}}(2\omega) \cdots \boldsymbol{C}_{\boldsymbol{x}_0}(\omega)\, \boldsymbol{P}^{(0)}\, \boldsymbol{C}_{\boldsymbol{x}_0}(\omega)^{-1} \cdots \boldsymbol{C}_{\boldsymbol{x}_{m-1}}(\omega)^{-1}$$

with $\boldsymbol{C}_{\boldsymbol{x}_k}$ defined by the vector $\boldsymbol{x}_k := (x_{k+1}, \ldots, x_{k+r})^{\mathrm{T}}$ ($k = 0, \ldots m-1$) and

$$\boldsymbol{P}^{(0)} := \operatorname{diag}\left(2^{r-1}, \ldots, 2^0\right)$$

determines the vector of cardinal B-splines with r-fold knots [P1, P2, P4].

4. Proof of Theorem 2.6. Before starting the proof of Theorem 2.6 let us show some preliminary assertions. For a given $\widetilde{\boldsymbol{P}} \in C_{2\pi}^m(\mathbb{R}^{r\times r})$ and a nonzero vector $\boldsymbol{y} \in \mathbb{R}$, let the $r \times r$ matrix $\boldsymbol{P} \in C_{2\pi}^m(\mathbb{R}^{r\times r})$ be defined by

$$\boldsymbol{P}(\omega) = \frac{1}{2} \boldsymbol{C}_{\boldsymbol{y}}(2\omega)\, \widetilde{\boldsymbol{P}}(\omega)\, \boldsymbol{C}_{\boldsymbol{y}}(\omega)^{-1},$$

where $\boldsymbol{C_y}(\omega)$ is defined by $\boldsymbol{y}$ via (2.6)–(2.7). Hence, we have by (2.9)

$$(1-e^{-i\omega})\,\boldsymbol{G_y}(2\omega)\,\boldsymbol{P}(\omega)=\frac{1}{2}\,(1-e^{-2i\omega})\,\widetilde{\boldsymbol{P}}(\omega)\,\boldsymbol{G_y}(\omega),$$

i.e.,

$$\boldsymbol{G_y}(2\omega)\,\boldsymbol{P}(\omega)=\left(\frac{1+e^{-i\omega}}{2}\right)\,\widetilde{\boldsymbol{P}}(\omega)\,\boldsymbol{G_y}(\omega). \tag{4.1}$$

In the next lemma we compute $\boldsymbol{G_y}(2\omega)\,(\mathrm{D}^k\boldsymbol{P})(\omega)$ in terms of derivatives of $\widetilde{\boldsymbol{P}}(\omega)$ and lower derivatives of $\boldsymbol{P}(\omega)$.

LEMMA 4.1. *We have, for* $k\in\mathbb{N}$,

$$\begin{aligned}&\boldsymbol{G_y}(2\omega)\,(D^k\boldsymbol{P})(\omega)\\&=-\sum_{l=1}^{k}\binom{k}{l}2^l\,(D^l\boldsymbol{G_y})(2\omega)\,(D^{k-l}\boldsymbol{P})(\omega)+\left(\frac{1+e^{-i\omega}}{2}\right)\,(D^k\widetilde{\boldsymbol{P}})(\omega)\boldsymbol{G_y}(\omega)\\&+\frac{1}{2}\sum_{l=1}^{k}\binom{k}{l}(D^{k-l}\widetilde{\boldsymbol{P}})(\omega)\,(-i)^{l-1}\,\left([(2^l-1)e^{-i\omega}+1](D\boldsymbol{G_y})(\omega)-ie^{-i\omega}\,\boldsymbol{G_y}(\omega)\right).\end{aligned}$$

In particular,

$$\begin{aligned}\boldsymbol{G_y}(0)\,(D^k\boldsymbol{P})(0)&=-\sum_{l=1}^{k}\binom{k}{l}2^l\,(D^l\boldsymbol{G_y})(0)\,(D^{k-l}\boldsymbol{P})(0)+(D^k\widetilde{\boldsymbol{P}})(0)\,\boldsymbol{G_y}(0)\\&\quad+\frac{1}{2}\sum_{l=1}^{k}\binom{k}{l}(D^{k-l}\widetilde{\boldsymbol{P}})(0)\,(-i)^l\,[\boldsymbol{G_y}(0)+2^li(D\boldsymbol{G_y})(0)]\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{G_y}(0)\,(D^k\boldsymbol{P})(\pi)&=-\sum_{l=1}^{k}\binom{k}{l}2^l\,(D^l\boldsymbol{G_y})(0)\,(D^{k-l}\boldsymbol{P})(\pi)\\&+\frac{1}{2}\sum_{l=1}^{k}\binom{k}{l}(D^{k-l}\widetilde{\boldsymbol{P}})(\pi)\,(-i)^l\,[-\boldsymbol{G_y}(\pi)-(2^l-2)i(D\boldsymbol{G_y})(\pi)].\end{aligned}$$

Proof. From (4.1) it follows by differentiation that

$$\begin{aligned}&\sum_{l=0}^{k}\binom{k}{l}2^l(\mathrm{D}^l\boldsymbol{G_y})(2\omega)\,(\mathrm{D}^{k-l}\boldsymbol{P})(\omega)\\&=\sum_{l=0}^{k}\binom{k}{l}(\mathrm{D}^{k-l}\widetilde{\boldsymbol{P}})(\omega)\,\mathrm{D}^l\left(\left(\frac{1+e^{-i\cdot}}{2}\right)\boldsymbol{G_y}\right)(\omega).\end{aligned} \tag{4.2}$$

Observing that

$$\mathrm{D}^s\left(\frac{1+e^{-i\cdot}}{2}\right)(\omega)=\begin{cases}\frac{1+e^{-i\omega}}{2} & \text{for } s=0,\\ \frac{(-i)^s}{2}\,e^{-i\omega} & \text{for } s\geq 1\end{cases}$$

and

$$(\mathrm{D}^s \boldsymbol{G_y})(\omega) = \begin{cases} \boldsymbol{G_y}(\omega) & \text{for } s = 0, \\ (-i)^{s-1}\,(\mathrm{D}\boldsymbol{G_y})(\omega) & \text{for } s \geq 1, \end{cases} \tag{4.3}$$

it follows for $l > 0$ that

$$\begin{aligned} &\mathrm{D}^l \left(\left(\frac{1+e^{-i\cdot}}{2} \right) \boldsymbol{G_y} \right)(\omega) = \sum_{s=0}^{l} \binom{l}{s} \mathrm{D}^s \left(\frac{1+e^{-i\cdot}}{2} \right)(\omega)\, (\mathrm{D}^{l-s}\boldsymbol{G_y})(\omega) \\ &= \left(\frac{1+e^{-i\omega}}{2} \right) (-i)^{l-1}\,(\mathrm{D}\boldsymbol{G_y})(\omega) + \frac{(-i)^l}{2}\, e^{-i\omega}\, \boldsymbol{G_y}(\omega) \\ &\quad + \frac{e^{-i\omega}}{2} \sum_{s=1}^{l-1} \binom{l}{s} (-i)^{l-1}\,(\mathrm{D}\boldsymbol{G_y})(\omega) \\ &= \frac{(-i)^{l-1}}{2}\,(\mathrm{D}\boldsymbol{G_y})(\omega)\, \left((1+e^{-i\omega}) + e^{-i\omega}(2^l-2) \right) + \frac{(-i)^l}{2}\, e^{-i\omega}\, \boldsymbol{G_y}(\omega) \\ &= \frac{(-i)^{l-1}}{2} \left((2^l-1)e^{-i\omega} + 1 \right) (\mathrm{D}\boldsymbol{G_y})(\omega) + \frac{(-i)^l}{2}\, e^{-i\omega}\, \boldsymbol{G_y}(\omega). \end{aligned}$$

Hence,

$$\begin{aligned} &\sum_{l=0}^{k} \binom{k}{l} (\mathrm{D}^{k-l}\widetilde{\boldsymbol{P}})(\omega)\, \mathrm{D}^l \left(\left(\frac{1+e^{-i\cdot}}{2} \right) \boldsymbol{G_y} \right)(\omega) \\ &= \left(\frac{1+e^{-i\omega}}{2} \right) (\mathrm{D}^k \widetilde{\boldsymbol{P}})(\omega)\, \boldsymbol{G_y}(\omega) \\ &\quad + \frac{1}{2} \sum_{l=1}^{k} \binom{k}{l} (\mathrm{D}^{k-l}\widetilde{\boldsymbol{P}})(\omega)\, (-i)^{l-1} \left([(2^l-1)e^{-i\omega}+1]\,(\mathrm{D}\boldsymbol{G_y})(\omega) - ie^{-i\omega}\, \boldsymbol{G_y}(\omega) \right). \end{aligned}$$

Together with (4.2) the assertion of Lemma 4.1 follows. □

Proof of Theorem 2.6. By (2.18) for $k = 0$ we have $\boldsymbol{y}_0^{\mathrm{T}} = \widetilde{\boldsymbol{y}}_0^{\mathrm{T}}\, \boldsymbol{G_y}(0) = \widetilde{\boldsymbol{y}}_0^{\mathrm{T}}\, \boldsymbol{e}\, \boldsymbol{y}^{\mathrm{T}}$, where $\boldsymbol{e}$ corresponds to $\boldsymbol{y}$ via (2.10). Further, note that $\widetilde{\boldsymbol{P}}(0)\, \boldsymbol{e} = \boldsymbol{e}$ implies

$$\widetilde{\boldsymbol{P}}(0)\, \boldsymbol{G_y}(0) = \boldsymbol{G_y}(0). \tag{4.4}$$

By assumption, $\widetilde{\boldsymbol{P}}$ satisfies the conditions (2.1)–(2.2) for $n = 0, \ldots, m-1$ with $\widetilde{\boldsymbol{y}}_0, \ldots, \widetilde{\boldsymbol{y}}_{m-1}$. Hence, we get

$$\begin{aligned} &\frac{2^k}{2^k-1} \sum_{l=0}^{k-1} \binom{k}{l} (2i)^{l-k}\, \widetilde{\boldsymbol{y}}_l^{\mathrm{T}}\, (\mathrm{D}^{k-l}\widetilde{\boldsymbol{P}})(0)\, \boldsymbol{G_y}(0) \\ &= \frac{2^k}{2^k-1} \left(\frac{1}{2^k}\, \widetilde{\boldsymbol{y}}_k^{\mathrm{T}}\, \boldsymbol{E}(0) - \widetilde{\boldsymbol{y}}_k^{\mathrm{T}}\, \widetilde{\boldsymbol{P}}(0)\, \boldsymbol{E}(0) \right) = -\widetilde{\boldsymbol{y}}_k^{\mathrm{T}}\, \boldsymbol{E}(0) \end{aligned}$$

such that $\boldsymbol{y}_k^{\mathrm{T}}$ defined in (2.18)–(2.19) can be represented for $k = 0, \ldots, m$ in the form

$$\begin{aligned} \boldsymbol{y}_k^{\mathrm{T}} &= (-ik)\widetilde{\boldsymbol{y}}_{k-1}^{\mathrm{T}}\, (\mathrm{D}\boldsymbol{G_y})(0) + \sum_{l=0}^{k-1} \binom{k}{l} \widetilde{B}_{k-l}\, \widetilde{\boldsymbol{y}}_l^{\mathrm{T}}\, \boldsymbol{G_y}(0) \\ &\quad - \frac{2^k}{2^k-1} \sum_{l=0}^{k-1} \binom{k}{l} (2i)^{l-k}\, \widetilde{\boldsymbol{y}}_l^{\mathrm{T}}\, (\mathrm{D}^{k-l}\widetilde{\boldsymbol{P}})(0)\, \boldsymbol{E}(0). \end{aligned} \tag{4.5}$$

1. We have to show that $\boldsymbol{P}(\omega)$ satisfies equations (2.1)–(2.2) for $n = 0, \dots, m$ with $\boldsymbol{y}_0, \dots, \boldsymbol{y}_m$. That means, by (2.18) and (4.5), we have to show that for $n = 0, \dots, m$

$$\boldsymbol{A}_n(0) + \boldsymbol{B}_n(0) + \boldsymbol{C}_n(0) + \boldsymbol{D}_n(0) = 2^{-n}\, \boldsymbol{y}_n^{\mathrm{T}}$$

and

$$\boldsymbol{A}_n(\pi) + \boldsymbol{B}_n(\pi) + \boldsymbol{C}_n(\pi) + \boldsymbol{D}_n(\pi) = \boldsymbol{0}^{\mathrm{T}}$$

are satisfied with

$$\begin{aligned}
\boldsymbol{A}_n(\omega) &:= \sum_{l=0}^{n} \binom{n}{l} (2i)^{l-n}\, (-il)\, \widetilde{\boldsymbol{y}}_{l-1}^{\mathrm{T}}\, (\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0)\, (\mathrm{D}^{n-l}\boldsymbol{P})(\omega),\\
\boldsymbol{B}_n(\omega) &:= \sum_{l=0}^{n-1} \binom{n}{l} (2i)^{l-n} \sum_{s=0}^{l} \binom{l}{s} \widetilde{B}_{l-s}\, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, \boldsymbol{G}\boldsymbol{y}(0)\, (\mathrm{D}^{n-l}\boldsymbol{P})(\omega),\\
\boldsymbol{C}_n(\omega) &:= \sum_{s=0}^{n-1} \binom{n}{s} \widetilde{B}_{n-s}\, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, \boldsymbol{G}\boldsymbol{y}(0)\, \boldsymbol{P}(\omega),\\
\boldsymbol{D}_n(\omega) &:= -\frac{2^n}{2^n - 1} \sum_{s=0}^{n-1} \binom{n}{s} (2i)^{s-n}\, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, (\mathrm{D}^{n-s}\widetilde{\boldsymbol{P}})(0)\, \boldsymbol{G}\boldsymbol{y}(0)\, \boldsymbol{P}(\omega).
\end{aligned}$$

For $\omega = 0$ and $\omega = \pi$, we replace $\boldsymbol{G}\boldsymbol{y}(0)\, (\mathrm{D}^{n-l}\boldsymbol{P})(\omega)$ in $\boldsymbol{B}_n(\omega)$ by the corresponding expressions given in Lemma 4.1 and obtain

$$\boldsymbol{B}_n(\omega) = \boldsymbol{B}_n^0(\omega) + \boldsymbol{B}_n^1(\omega) + \boldsymbol{B}_n^2(\omega)$$

with

$$\begin{aligned}
\boldsymbol{B}_n^0(\omega) &:= -\sum_{l=0}^{n-1} \binom{n}{l} (2i)^{l-n} \sum_{s=0}^{l} \binom{l}{s} \widetilde{B}_{l-s}\, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \sum_{r=1}^{n-l} \binom{n-l}{r} 2^r\, (\mathrm{D}^r \boldsymbol{G}\boldsymbol{y})(0)\\
&\quad \times (\mathrm{D}^{n-l-r}\boldsymbol{P})(\omega),\\
\boldsymbol{B}_n^1(\omega) &:= \left(\frac{1 + e^{-i\omega}}{2}\right) \sum_{l=0}^{n-1} \binom{n}{l} (2i)^{l-n} \sum_{s=0}^{l} \binom{l}{s} \widetilde{B}_{l-s}\, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, (\mathrm{D}^{n-l}\widetilde{\boldsymbol{P}})(\omega)\, \boldsymbol{G}\boldsymbol{y}(\omega),\\
\boldsymbol{B}_n^2(\omega) &:= \frac{1}{2} \sum_{l=0}^{n-1} \binom{n}{l} (2i)^{l-n} \sum_{s=0}^{l} \binom{l}{s} \widetilde{B}_{l-s}\, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \sum_{r=1}^{n-l} \binom{n-l}{r} (\mathrm{D}^{n-l-r}\widetilde{\boldsymbol{P}})(\omega)\, (-i)^{r-1}\\
&\quad \times \left([(2^r - 1)e^{-i\omega} + 1](\mathrm{D}\boldsymbol{G}\boldsymbol{y})(\omega) - ie^{-i\omega}\, \boldsymbol{G}\boldsymbol{y}(\omega) \right).
\end{aligned}$$

2. First we show that for $\omega = 0$ and $\omega = \pi$,

$$\boldsymbol{A}_n(\omega) + \boldsymbol{B}_n^0(\omega) = \boldsymbol{0}^{\mathrm{T}}.$$

Note that $\binom{n}{s}\binom{n-s}{l-s} = \binom{n}{l}\binom{l}{s}$. Changing the order of summation over l and s and putting $r' := n - l - r$, it follows that

$$\begin{aligned}
\boldsymbol{B}_n^0(\omega) &= (-i) \sum_{s=0}^{n-1} \binom{n}{s} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \sum_{l=s}^{n-1} \binom{n-s}{l-s} (2i)^{l-n}\, \widetilde{B}_{l-s} \sum_{r=1}^{n-l} \binom{n-l}{r} 2^r\\
&\quad \times (\mathrm{D}^r \boldsymbol{G}\boldsymbol{y})(0)\, (\mathrm{D}^{n-l-r}\boldsymbol{P})(\omega)
\end{aligned}$$

$$= (-i)\sum_{s=0}^{n-1}\binom{n}{s}\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\sum_{l=s}^{n-1}\binom{n-s}{l-s}(2i)^{l-n}\,\widetilde{B}_{l-s}\sum_{r'=0}^{n-l-1}\binom{n-l}{r'}(-2i)^{n-l-r'}$$
$$\times(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0)\,(\mathrm{D}^{r'}\boldsymbol{P})(\omega),$$

where we have used that $(\mathrm{D}^r\boldsymbol{G}\boldsymbol{y})(0) = (-i)^{r-1}\,(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0) = (-i)^{n-l-r'+1}\,(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0)$ (see (4.3)). Thus,

$$\begin{aligned}\boldsymbol{B}_n^0(\omega) &= (-i)\sum_{s=0}^{n-1}\binom{n}{s}\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\sum_{l=0}^{n-s-1}\binom{n-s}{l}(2i)^{l+s-n}\,\widetilde{B}_l\sum_{r=0}^{n-l-s-1}\binom{n-l-s}{r}\\ &\quad\times(-2i)^{n-l-s-r}\,(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0)\,(\mathrm{D}^r\boldsymbol{P})(\omega)\\ &= (-i)\sum_{s=0}^{n-1}\binom{n}{s}\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\sum_{r=0}^{n-s-1}\binom{n-s}{r}(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0)\,(\mathrm{D}^r\boldsymbol{P})(\omega)(2i)^{-r}(-1)^{n-s-r}\\ &\quad\times\sum_{l=0}^{n-r-s-1}\binom{n-r-s}{l}\widetilde{B}_l\,(-1)^l.\end{aligned}$$

Observe that, by (2.16),

$$\sum_{l=0}^{k-1}\binom{k}{l}\widetilde{B}_l\,(-1)^l = \begin{cases}0 & \text{for } k>1,\\ 1 & \text{for } k=1.\end{cases} \tag{4.6}$$

Hence, the last term in the last representation of $\boldsymbol{B}_n^0$ vanishes for $n-s-1\neq r$, and so

$$\boldsymbol{B}_n^0(\omega) = i\sum_{s=0}^{n-1}\binom{n}{s}(n-s)\,\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\,(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0)\,(\mathrm{D}^{n-s-1}\boldsymbol{P})(\omega)\,(2i)^{-n+s+1}.$$

Shifting the summation index, we find for $\boldsymbol{A}_n(\omega)$ $(\omega = 0,\,\pi)$

$$\begin{aligned}\boldsymbol{A}_n(\omega) &= (-i)\sum_{l=0}^{n-1}\binom{n}{l+1}(2i)^{l+1-n}\,(l+1)\,\widetilde{\boldsymbol{y}}_l^{\mathrm{T}}\,(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0)\,(\mathrm{D}^{n-l+1}\boldsymbol{P})(\omega)\\ &= (-i)\sum_{l=0}^{n-1}\binom{n}{l}(n-l)\widetilde{\boldsymbol{y}}_l^{\mathrm{T}}\,(\mathrm{D}\boldsymbol{E})(0)\,(\mathrm{D}^{n-l-1}\boldsymbol{P})(\omega)\,(2i)^{-n+l+1}.\end{aligned}$$

Hence, $\boldsymbol{B}_n^0(\omega)+\boldsymbol{A}_n(\omega) = \boldsymbol{0}^{\mathrm{T}}$ for $\omega = 0,\ \pi$.

3. Let us consider $\boldsymbol{B}_n^1(\omega)$. We easily observe that $\boldsymbol{B}_n^1(\pi) = \boldsymbol{0}^{\mathrm{T}}$. For $\omega = 0$, we find by changing the order of summations over l and s that

$$\begin{aligned}\boldsymbol{B}_n^1(0) &= \sum_{l=0}^{n-1}\binom{n}{l}(2i)^{l-n}\sum_{s=0}^{l}\binom{l}{s}\widetilde{B}_{l-s}\,\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\,(D^{n-l}\widetilde{\boldsymbol{P}})(0)\,\boldsymbol{G}\boldsymbol{y}(0)\\ &= \sum_{s=0}^{n-1}\binom{n}{s}\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\sum_{l=s}^{n-1}\binom{n-s}{l-s}(2i)^{l-n}\,\widetilde{B}_{l-s}\,(\mathrm{D}^{n-l}\widetilde{\boldsymbol{P}})(0)\,\boldsymbol{G}\boldsymbol{y}(0)\\ &= \sum_{s=0}^{n-1}\binom{n}{s}\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\sum_{l=0}^{n-s-1}\binom{n-s}{l}(2i)^{l+s-n}\,\widetilde{B}_l\,(\mathrm{D}^{n-s-l}\widetilde{\boldsymbol{P}})(0)\,\boldsymbol{G}\boldsymbol{y}(0)\\ &= \sum_{l=0}^{n-1}\binom{n}{l}\widetilde{B}_l\sum_{s=0}^{n-1-l}\binom{n-l}{s}(2i)^{-n+l+s}\,\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\,(\mathrm{D}^{n-l-s}\boldsymbol{P})(0)\,\boldsymbol{G}\boldsymbol{y}(0).\end{aligned}$$

On the other hand, for $l > 1$, the equations (2.1) for $\widetilde{\boldsymbol{P}}$ and $\widetilde{\boldsymbol{y}}_n$ $(n = 0, \dots, m-1)$ imply that

$$\begin{aligned}&\sum_{s=0}^{n-1-l} \binom{n-l}{s} (2i)^{-n+l+s}\, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, (\mathrm{D}^{n-l-s}\widetilde{\boldsymbol{P}})(0)\, \boldsymbol{G}\boldsymbol{y}(0)\\ &= 2^{-n+l}\, \widetilde{\boldsymbol{y}}_{n-l}^{\mathrm{T}}\, \boldsymbol{G}\boldsymbol{y}(0) - \widetilde{\boldsymbol{y}}_{n-l}^{\mathrm{T}}\, \widetilde{\boldsymbol{P}}(0)\, \boldsymbol{G}\boldsymbol{y}(0) = (2^{-n+l}-1)\widetilde{\boldsymbol{y}}_{n-l}^{\mathrm{T}}\, \boldsymbol{G}\boldsymbol{y}(0),\end{aligned}$$

where we have used (4.4). Hence, we can write

$$\begin{aligned}\boldsymbol{B}_n^1(0) = &\sum_{l=1}^{n-1} \binom{n}{l} \widetilde{B}_l\, (2^{-n+l}-1)\, \widetilde{\boldsymbol{y}}_{n-l}^{\mathrm{T}}\, \boldsymbol{G}\boldsymbol{y}(0)\\ &+ \sum_{s=0}^{n-1} \binom{n}{s} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, (2i)^{-n+s}\, (\mathrm{D}^{n-s}\widetilde{\boldsymbol{P}})(0)\, \boldsymbol{G}\boldsymbol{y}(0).\end{aligned} \tag{4.7}$$

4. Let us concentrate on $\boldsymbol{B}_n^2(\omega)$. Putting

$$\widetilde{\boldsymbol{E}}_r(\omega) := \frac{(-i)^{r-1}}{2}\, \left[[(2^r-1)e^{-i\omega}+1](\mathrm{D}\boldsymbol{G}\boldsymbol{y})(\omega) - ie^{-i\omega}\, \boldsymbol{G}\boldsymbol{y}(\omega)\right)$$

we obtain, for $\omega = 0, \pi$ by changing the order of summations and shifting the summation indices

$$\begin{aligned}\boldsymbol{B}_n^2(\omega) &= \sum_{l=0}^{n-1} \binom{n}{l} (2i)^{l-n} \sum_{s=0}^{l} \binom{l}{s} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, \widetilde{B}_{l-s} \sum_{r=1}^{n-l} \binom{n-l}{r} (\mathrm{D}^{n-l-r}\widetilde{\boldsymbol{P}})(\omega)\, \widetilde{\boldsymbol{E}}_r(\omega)\\
&= \sum_{s=0}^{n-1} \binom{n}{s} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \sum_{l=s}^{n-1} \binom{n-s}{l-s} (2i)^{l-n}\, \widetilde{B}_{l-s} \sum_{r=1}^{n-l} \binom{n-l}{r} (\mathrm{D}^{n-l-r}\widetilde{\boldsymbol{P}})(\omega)\, \widetilde{\boldsymbol{E}}_r(\omega)\\
&= \sum_{s=0}^{n-1} \binom{n}{s} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \sum_{l=0}^{n-s-1} \binom{n-s}{l} (2i)^{l+s-n}\, \widetilde{B}_l \sum_{r=1}^{n-l-s} \binom{n-l-s}{r}\\
&\quad \times (\mathrm{D}^{n-l-s-r}\widetilde{\boldsymbol{P}})(\omega)\, \widetilde{\boldsymbol{E}}_r(\omega)\\
&= \sum_{l=0}^{n-1} \binom{n}{l} \widetilde{B}_l \sum_{s=0}^{n-l-1} \binom{n-l}{s} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, (2i)^{l+s-n} \sum_{r=1}^{n-l-s} \binom{n-l-s}{r} (\mathrm{D}^{n-l-s-r}\widetilde{\boldsymbol{P}})(\omega)\, \widetilde{\boldsymbol{E}}_r(\omega)\\
&= \sum_{l=0}^{n-1} \binom{n}{l} \widetilde{B}_l \sum_{r=1}^{n-l} \binom{n-l}{r} \left(\sum_{s=0}^{n-l-r} \binom{n-l-r}{s} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\, (2i)^{-n+l+r+s}\, (\mathrm{D}^{n-l-r-s}\widetilde{\boldsymbol{P}})(\omega) \right)\\
&\quad \times (2i)^{-r}\, \widetilde{\boldsymbol{E}}_r(\omega).\end{aligned}$$

Application of (2.1)–(2.2) for $\widetilde{\boldsymbol{P}}$ in the sum over s implies that $\boldsymbol{B}_n^2(\pi) = 0$ and

$$\begin{aligned}\boldsymbol{B}_n^2(0) &= \sum_{l=0}^{n-1} \binom{n}{l} \widetilde{B}_l \sum_{r=1}^{n-l} \binom{n-l}{r} 2^{-n+l+r}\, \widetilde{\boldsymbol{y}}_{n-l-r}^{\mathrm{T}}\, (2i)^{-r}\, \widetilde{\boldsymbol{E}}_r(0)\\
&= \sum_{l=0}^{n-1} \binom{n}{l} \widetilde{B}_l \sum_{r=1}^{n-l} \binom{n-l}{r} 2^{-n+l-1}\, \widetilde{\boldsymbol{y}}_{n-l-r}^{\mathrm{T}}\, (-1)^r \left(i2^r(\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0) + \boldsymbol{G}\boldsymbol{y}(0)\right).\end{aligned}$$

Putting $r' := n-l-r$ and changing again the order of summation we get

$$\begin{aligned}\boldsymbol{B}_n^2(0) &= \sum_{l=0}^{n-1}\binom{n}{l}\widetilde{B}_l \sum_{r'=0}^{n-l-1}\binom{n-l}{r'} 2^{-n+l-1}\,\widetilde{\boldsymbol{y}}_{r'}^{\mathrm{T}}\,(-1)^{n-l-r'} \\ &\quad\times\Big(i2^{n-l-r'}(\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0)+\boldsymbol{G}_{\boldsymbol{y}}(0)\Big) \\ &= i\sum_{r=0}^{n-1}\binom{n}{r}\widetilde{\boldsymbol{y}}_r^{\mathrm{T}}\left(\sum_{l=0}^{n-r-1}\binom{n-r}{l}\widetilde{B}_l\,(-1)^l\right)2^{-r-1}\,(-1)^{n-r}\,(\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0) \\ &\quad+\sum_{r=0}^{n-1}\binom{n}{r}\widetilde{\boldsymbol{y}}_r^{\mathrm{T}}\left(\sum_{l=0}^{n-r-1}\binom{n-r}{l}\widetilde{B}_l\,(-2)^l\right)2^{-n-1}\,(-1)^{n-r}\,\boldsymbol{G}_{\boldsymbol{y}}(0).\end{aligned}$$

Using the identities (2.17) and (4.6) for Bernoulli numbers and observing that $(-1)^k\widetilde{B}_k=\widetilde{B}_k$ for $k>1$, it follows that

$$\begin{aligned}\boldsymbol{B}_n^2(0) &= \frac{-in}{2^n}\widetilde{\boldsymbol{y}}_{n-1}^{\mathrm{T}}\,(\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0)-\frac{n}{2^{n+1}}\,\widetilde{\boldsymbol{y}}_{n-1}^{\mathrm{T}}\,\boldsymbol{G}_{\boldsymbol{y}}(0) \\ &\quad+2^{-n}\sum_{r=0}^{n-2}\binom{n}{r}\widetilde{\boldsymbol{y}}_r^{\mathrm{T}}\,(-2^{n-r}+1)\widetilde{B}_{n-r}\,(-1)^{n-r}\boldsymbol{G}_{\boldsymbol{y}}(0) \\ &= \frac{-in}{2^n}\widetilde{\boldsymbol{y}}_{n-1}^{\mathrm{T}}\,(\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0)-\frac{n}{2^{n+1}}\,\widetilde{\boldsymbol{y}}_{n-1}^{\mathrm{T}}\,\boldsymbol{G}_{\boldsymbol{y}}(0) \\ &\quad+\sum_{r=0}^{n-2}\binom{n}{r}\widetilde{\boldsymbol{y}}_r^{\mathrm{T}}\,(2^{-n}-2^{-r})\widetilde{B}_{n-r}\,(-1)^{n-r}\boldsymbol{G}_{\boldsymbol{y}}(0) \\ &= \frac{-in}{2^n}\widetilde{\boldsymbol{y}}_{n-1}^{\mathrm{T}}\,(\mathrm{D}\boldsymbol{G}_{\boldsymbol{y}})(0)+\sum_{r=0}^{n-1}\binom{n}{r}\widetilde{\boldsymbol{y}}_r^{\mathrm{T}}\,(2^{-n}-2^{-r})\widetilde{B}_{n-r}\,\boldsymbol{G}_{\boldsymbol{y}}(0).\end{aligned}$$

5. Now let $\omega=\pi$. Recall that $\boldsymbol{B}_n^1(\pi)=\boldsymbol{B}_n^2(\pi)=\boldsymbol{A}_n(\pi)+\boldsymbol{B}_n^0(\pi)=\boldsymbol{0}^{\mathrm{T}}$. Further, by $\boldsymbol{G}_{\boldsymbol{y}}(0)\boldsymbol{P}(\pi)=\boldsymbol{0}$ we have $\boldsymbol{C}_n(\pi)=\boldsymbol{D}_n(\pi)=\boldsymbol{0}^{\mathrm{T}}$. Hence,

$$\boldsymbol{A}_n(\pi)+\boldsymbol{B}_n(\pi)+\boldsymbol{C}_n(\pi)+\boldsymbol{D}_n(\pi)=\boldsymbol{0}^{\mathrm{T}}.$$

6. Let $\omega=0$. By $\boldsymbol{G}_{\boldsymbol{y}}(0)\,\boldsymbol{P}(0)=\widetilde{\boldsymbol{P}}(0)\,\boldsymbol{G}_{\boldsymbol{y}}(0)=\boldsymbol{G}_{\boldsymbol{y}}(0)$ we obtain

$$\begin{aligned}\boldsymbol{C}_n(0) &= \sum_{s=0}^{n-1}\binom{n}{s}\,\widetilde{B}_{n-s}\,\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\,\boldsymbol{G}_{\boldsymbol{y}}(0), \\ \boldsymbol{D}_n(0) &= -\frac{2^n}{2^n-1}\sum_{s=0}^{n-1}\binom{n}{s}\,(2i)^{s-n}\,\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\,(\mathrm{D}^{n-s}\widetilde{\boldsymbol{P}})(0)\,\boldsymbol{G}_{\boldsymbol{y}}(0).\end{aligned}$$

Observing that by (4.7)

$$\begin{aligned}\boldsymbol{D}_n(0)+\boldsymbol{B}_n^1(0) &= \sum_{l=1}^{n-1}\binom{n}{l}\,\widetilde{B}_l\,(2^{-n+l}-1)\,\widetilde{\boldsymbol{y}}_{n-l}^{\mathrm{T}}\,\boldsymbol{G}_{\boldsymbol{y}}(0) \\ &\quad-\frac{1}{2^n-1}\sum_{s=0}^{n-1}\binom{n}{s}\,(2i)^{s-n}\widetilde{\boldsymbol{y}}_s^{\mathrm{T}}\,(\mathrm{D}^{n-s}\widetilde{\boldsymbol{P}})(0)\,\boldsymbol{G}_{\boldsymbol{y}}(0)\end{aligned}$$

and using the expression for $\boldsymbol{B}_n^2$ found in part 4, we obtain

$$\begin{aligned}
&\boldsymbol{D}_n(0) + \boldsymbol{B}_n^1(0) + \boldsymbol{B}_n^2(0) + \boldsymbol{C}_n(0) \\
&= -\frac{1}{2^n - 1} \sum_{s=0}^{n-1} \binom{n}{s} (2i)^{s-n} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \, (\mathrm{D}^{n-s} \widetilde{\boldsymbol{P}})(0) \, \boldsymbol{G}\boldsymbol{y}(0) \\
&\quad + \sum_{l=0}^{n-1} \binom{n}{l} \widetilde{B}_{n-l} \, (2^{-l} - 1) \, \widetilde{\boldsymbol{y}}_l^{\mathrm{T}} \, \boldsymbol{G}\boldsymbol{y}(0) \\
&\quad - \frac{in}{2^n} \, \widetilde{\boldsymbol{y}}_{n-1}^{\mathrm{T}} \, (\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0) + \sum_{r=0}^{n-1} \binom{n}{r} \widetilde{\boldsymbol{y}}_r^{\mathrm{T}} \, (2^{-n} - 2^{-r}) \, \widetilde{B}_{n-r} \, \boldsymbol{G}\boldsymbol{y}(0) \\
&\quad + \sum_{s=0}^{n-1} \binom{n}{s} \widetilde{B}_{n-s} \, \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \, \boldsymbol{G}\boldsymbol{y}(0) \\
&= \frac{-in}{2^n} \, \widetilde{\boldsymbol{y}}_{n-1}^{\mathrm{T}} \, (\mathrm{D}\boldsymbol{G}\boldsymbol{y})(0) - \frac{1}{2^n - 1} \sum_{s=0}^{n-1} \binom{n}{s} (2i)^{s-n} \widetilde{\boldsymbol{y}}_s^{\mathrm{T}} \, (\mathrm{D}^{n-s} \widetilde{\boldsymbol{P}})(0) \, \boldsymbol{G}\boldsymbol{y}(0) \\
&\quad + \sum_{l=0}^{n-1} \binom{n}{l} \widetilde{B}_{n-l} \, \widetilde{\boldsymbol{y}}_l^{\mathrm{T}} \, \boldsymbol{G}\boldsymbol{y}(0) \, \left(2^{-l} - 1 + 2^{-n} - 2^{-l} + 1\right) = 2^{-n} \, \boldsymbol{y}_n^{\mathrm{T}}.
\end{aligned}$$

Recalling that $\boldsymbol{A}_n(0) + \boldsymbol{B}_n^0(0) = \boldsymbol{0}^{\mathrm{T}}$, the proof is complete. □

Acknowledgments. We would like to thank Ingrid Daubechies and Gilbert Strang for very useful discussions, advice, and comments.

REFERENCES

[AS] M. Abramowitz and J. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[AR] B. K. Alpert and V. Rokhlin, *A fast algorithm for the evaluation of Legendre expansions*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 158–179.

[CDM] A. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc., 453 (1991), pp. 1–186.

[CDP] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.

[CL] C. K. Chui and J. A. Lian, *A Study of Orthonormal Multi–Wavelets*, CAT report 351, Texas A&M University, College Station, TX, 1995.

[DM] W. Dahmen and C. A. Micchelli, *Biorthogonal wavelet expansions*, Const. Approx., 13 (1997), pp. 293–328.

[D1] I. Daubechies, *Orthonormal bases of wavelets with compact support*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[D2] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[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] I. Daubechies and J. Lagarias, *Two-scale difference equations:* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[DGHM] G. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1158–1192.

[GHM] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Fractal functions and wavelet expansions based on several scaling functions*, J. Approx. Theory, 78 (1994), pp. 373–401.

[GL] T. N. T. Goodman and S. L. Lee, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.

[GLT] T. N. T. Goodman, S. L. Lee, and W. S. Tang, *Wavelets in wandering subspaces*, Trans. Amer. Math. Soc., 338 (1993), pp. 639–654.

[HC] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.

[HSS] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., 73 (1996), pp. 75–94.

[H] L. Hervé, *Multi–Resolution analysis of multiplicity d. Application to dyadic interpolation*, Appl. Comput. Harmonic Anal., 1 (1994), pp. 299–315.

[JL] R. Q. Jia and J. J. Lei, *Approximation by multiinteger translates of functions having global support*, J. Approx. Theory, 72 (1993), pp. 2–23.

[J] Q. Jiang, *On the regularity of matrix refinable functions*, SIAM, J. Math. Anal., to appear.

[L] J. A. Lian, *Characterization of the order of polynomial reproduction for multi–scaling functions*, in Approximation Theory VIII, Vol 2: Wavelets and Multilevel Approximation, C. K. Chui and L. L. Schumaker, eds., World Scientific Publishing, Singapore, 1995, pp. 251–258.

[MRV] P. R. Massopust, D. K. Ruch, and P. J. Van Fleet, *On the support properties of scaling vectors*, Appl. Comput. Harmonic Anal., 3 (1996), pp. 229–238.

[P1] G. Plonka, *Two-scale symbol and autocorrelation symbol for B-splines with multiple knots*, Adv. Comput. Math., 3 (1995), pp. 1–22.

[P2] G. Plonka, *Generalized spline wavelets*, Constr. Approx., 12 (1996), pp. 127–155.

[P3] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.

[P4] G. Plonka, *Factorization of refinement masks of function vectors*, in Approximation Theory VIII, Vol 2: Wavelets and Multilevel Approximation, C. K. Chui and L. L. Schumaker, eds., World Scientific Publishing, Singapore, 1995, pp. 317–324.

[Sh] Z. Shen, *Refinable Function Vectors*, preprint, Department of Mathematics, The National University of Singapore, 1995.

[SB] M. J. T. Smith and T. P. Barnwell, *Exact reconstruction techniques for tree–structured subband coders*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 434–441.

[St1] G. Strang, *Eigenvalues of* $(\downarrow 2)H$ *and convergence of the cascade algorithm*, IEEE Trans. Signal Process., 44 (1996), pp. 233–238.

[St2] G. Strang, *Introduction to Linear Algebra*, Wellesley–Cambridge Press, Wellesley, MA, 1993.

[SF] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Wellesley–Cambridge Press, Wellesley, MA, 1973.

[SN] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley–Cambridge Press, Wellesley, MA, 1996.

[SS1] G. Strang and V. Strela, *Orthogonal multiwavelets with vanishing moments*, J. Optical Eng., 33 (1994), pp. 2104–2107.

[SS2] G. Strang and V. Strela, *Short wavelets and matrix dilation equations*, IEEE Trans. Signal Process., 43 (1995), pp. 108–115.

[SS3] V. Strela and G. Strang, *Finite element multiwavelets*, in Approximation Theory, Wavelets and Applications, S. P. Singh, ed., Kluwer Academic Publ., Dordrecht, 1995, pp. 485–496.

[SS4] V. Strela and G. Strang, *Biorthogonal Multiwavelets and Finite Elements*, preprint, Department of Mathematics, MIT, Cambridge, MA, 1996.

[S1] V. Strela, *Multiwavelets: Regularity, orthogonality, and symmetry via two–scale similarity transform*, Stud. Appl. Math., 98 (1997), pp. 335–354.

[S2] V. Strela, *Multiwavelets: Theory and Applications*, PhD thesis, Department of Mathematics, MIT, Cambridge, MA, 1996.

[SHSTH] V. Strela, P. N. Heller, G. Strang, P. Topiwala, and C. Heil, *The application of multiwavelet filter banks to image processing*, IEEE Trans. Signal Process., to appear.

[XGHS] X.-G. Xia, J. S. Geronimo, D. Hardin, and B. Suter, *Design of prefilters for discrete multiwavelet transforms*, IEEE Trans. Signal Process., 44 (1996), pp. 25–35.

SIAM J. MATH. ANAL.
Vol. 29, No. 2, pp. 511–546, March 1998

011

THE LIFTING SCHEME: A CONSTRUCTION OF SECOND GENERATION WAVELETS*

WIM SWELDENS†

Abstract. We present the lifting scheme, a simple construction of second generation wavelets; these are wavelets that are not necessarily translates and dilates of one fixed function. Such wavelets can be adapted to intervals, domains, surfaces, weights, and irregular samples. We show how the lifting scheme leads to a faster, in-place calculation of the wavelet transform. Several examples are included.

Key words. wavelet, multiresolution, second generation wavelet, lifting scheme

AMS subject classification. 42C15

PII. S0036141095289051

1. Introduction.

Wavelets form a versatile tool for representing general functions or data sets. Essentially we can think of them as data building blocks. Their fundamental property is that they allow for representations which are *efficient* and which can be computed *fast*. In other words, wavelets are capable of quickly capturing the essence of a data set with only a small set of coefficients. This is based on the fact that most data sets have correlation both in time (or space) and frequency. Because of the time-frequency localization of wavelets, efficient representations can be obtained. Indeed, building blocks which already reflect the correlation present in the data lead to more compact representations. This is the key to applications. Over the last decade wavelets have found applications in numerous areas of mathematics, engineering, computer science, statistics, physics, etc.

Wavelet functions $\psi_{j,m}$ are traditionally defined as the dyadic translates and dilates of one particular $L_2(\mathbf{R})$ function, the *mother wavelet* ψ: $\psi_{j,m}(x) = \psi(2^j x - m)$. We refer to such wavelets as *first generation wavelets*. In this paper we introduce a more general setting where the wavelets are not necessarily translates and dilates of each other but still enjoy all the powerful properties of first generation wavelets. These wavelets are referred to as *second generation wavelets*. We present the *lifting scheme*, a simple, but quite powerful, tool to construct second generation wavelets.

Before we consider the generalization to the second generation case, let us review the properties of first generation wavelets which we would like to preserve.

P1: Wavelets form a Riesz basis for $L_2(\mathbf{R})$ and an unconditional basis for a wide variety of function spaces $\mathcal{F}$, such as Lebesgue, Lipschitz, Sobolev, and Besov spaces. If we denote the wavelet basis by $\{\psi_{j,m} \mid j, m\}$, we can represent a general function f in $\mathcal{F}$ as $f = \sum_{j,m} \gamma_{j,m} \psi_{j,m}$, with unconditional convergence in the norm of $\mathcal{F}$. Simple characterizations of the $\mathcal{F}$-norm of f in terms of the absolute value of its wavelet coefficients $\gamma_{j,m}$ exist.

*Received by the editors July 14, 1995; accepted for publication (in revised form) January 15, 1997. While writing this paper the author was employed at the Katholieke Universiteit Leuven, Belgium and the University of South Carolina, where he was partially supported by NSF EPSCoR grant EHR 9108772 and DARPA grant AFOSR F49620-93-1-0083. He is also on leave as Senior Research Assistant of the National Fund of Scientific Research Belgium (NFWO).

http://www.siam.org/journals/sima/29-2/28905.html

†Lucent Technologies, Bell Laboratories, Rm. 2C-175, 700 Mountain Avenue, Murray Hill, NJ 07974 (wim@bell-labs.com).

P2: One has explicit information concerning the coordinate functionals $\widetilde{\psi}_{j,m}$ where $\gamma_{j,m} = \widetilde{\psi}_{j,m}(f)$. The wavelets are either orthogonal or the dual (biorthogonal) wavelets are known.

P3: The wavelets and their duals are local in space and frequency. Some wavelets are even compactly supported. The frequency localization follows from the smoothness of the wavelets (decay towards high frequencies) and the fact that they have vanishing polynomial moments (decay towards low frequencies).

P4: Wavelets fit into the framework of multiresolution analysis. This leads to the *fast wavelet transform*, which allows us to pass between the function f and its wavelet coefficients $\gamma_{j,m}$ in linear time.

These properties result in the fact that, quoted from Donoho in [58], "*wavelets are optimal bases for compressing, estimating, and recovering functions in* $\mathcal{F}$." Roughly speaking, for a general class of functions, the essential information contained in a function is captured by a small fraction of the wavelet coefficients. Again this is the key to applications. Wavelets have proved to be useful in various application domains such as signal and image processing, data compression, data transmission, the numerical solution of differential and integral equations, and noise reduction.

Many first generation wavelet families have been constructed over the last ten years. We refer to the work of (in alphabetical order) Aldroubi and Unser [2, 3, 108, 107], Battle and Lemarié [13, 78], Chui and Wang [19, 25, 24, 23], Cohen and Daubechies [28], Cohen, Daubechies, and Feauveau [29], Daubechies [47, 49, 48], Donoho [57, 56], Frazier and Jawerth [65, 67, 66], Herley and Vetterli [73, 110], Kovačević and Vetterli [77, 111], Mallat [85, 84, 86], Meyer [87], and many more. Except for Donoho, they all rely on the Fourier transform as a basic construction tool. The reason is that translation and dilation become algebraic operations in the Fourier domain.

In fact, in the early 1980s, several years before the above developments, Strömberg discovered the first orthogonal wavelets with a technique based on spline interpolation which does not rely on the Fourier transform [103].

The construction as initiated by Daubechies and coworkers essentially consists of three stages. The *algebraic stage* involves constructing the filters that are used in the fast wavelet transform; more precisely, it consists of finding certain polynomials and assuring that the above property P4 is satisfied. In the *analytic stage*, one shows that wavelets associated with these filters exist, that they are localized (property P3), and that they form a basis for the proper function space (property P1). In the *geometrical stage*, one checks the smoothness of the basis functions (property P3). In this context, we mention the work of Collela and Heil [37, 38], Daubechies and Lagarias [50, 51], Eirola [63], Rioul [94], and Villemoes [113, 112].

Let us next consider applications which illustrate the need for generalizations of first generation wavelets.

G1: While first generation wavelets provided bases for functions defined on $\mathbf{R}^n$, applications such as data segmentation and the solution of partial differential and integral equations on general domains require wavelets that are defined on arbitrary, possibly nonsmooth, domains of $\mathbf{R}^n$, as well as wavelets adapted to "life" on curves, surfaces, or manifolds.

G2: Diagonalization of differential forms, analysis on curves and surfaces, and weighted approximation require a basis adapted to weighted measures; however, first generation wavelets typically provide bases only for spaces with translation invariant (Haar–Lebesgue) measures.

G3: Many real life problems require algorithms adapted to irregular sampled data, while first generation wavelets imply a regular sampling of the data.

A generalization of first generation wavelets to the settings G1–G3, while preserving the properties P1–P4, is needed. We refer to such wavelets as *second generation wavelets*. The key lies in the observations (A) that translation and dilation cannot be maintained in the settings G1–G3, and (B) that translation and dilation are not essential in obtaining the properties P1–P4. Giving up translation and dilation, however, implies that the Fourier transform can no longer be used as a construction tool. A proper substitute is needed.

Several results concerning the construction of wavelets adapted to some of the cases in G1–G3 already exist. For example, we have wavelets on an interval [8, 10, 18, 30, 31, 88], wavelets on bounded domains [27, 74], spline wavelets for irregular samples, [15, 7, 45], and weighted wavelets [11, 12, 104]. These constructions are tailored toward one specific setting. Other instances of second generation wavelets have been reported in the literature, e.g., the construction of scaling functions through subdivision [41], basis constructions [43], as well as the development of stability criteria [41, 42].

In this paper, we present the *lifting scheme*, a simple, general construction of second generation wavelets. The basic idea, which inspired the name, is to start with a very simple or trivial multiresolution analysis and gradually work one's way up to a multiresolution analysis with particular properties. The lifting scheme allows one to custom design the filters, needed in the transform algorithms, to the situation at hand. In this sense it provides an answer to the algebraic stage of a wavelet construction. Whether these filters actually generate functions which form a stable basis (analytic stage) or have smoothness (geometric stage) remains to be checked in each particular case. The lifting scheme also leads to a fast in-place calculation of the wavelet transform, i.e., an implementation that does not require auxiliary memory.

The paper is organized as follows. We start out by discussing related work in section 2. In sections 3, 4, 5, and 6 we generalize, respectively, multiresolution analysis, cascade algorithm, wavelets, and the fast wavelet transform to the second generation setting. With the notation introduced in section 7 we are able to state and prove the lifting scheme in section 8. Section 9 discusses the lifted fast wavelet transform, while section 10 covers the cakewalk construction, an enhanced version of the lifting scheme. Sections 11, 12, and 13 introduce three possible examples of an initial multiresolution analysis to start lifting: respectively, generalized Haar wavelets, the Lazy wavelet, and biorthogonal Haar wavelets. Finally, section 14 contains a discussion of applications and future research.

2. Related work. The idea of second generation wavelets and abandoning the Fourier transform as a construction tool for wavelets is not entirely new and, over the last few years, has been researched by several independent groups. In this section we discuss these developments and their relationship with lifting.

The lifting scheme was originally inspired by the work of Donoho on one side and Lounsbery, De Rose, and Warren on the other. In [56, 57], Donoho presents the idea of interpolating and average-interpolating wavelets, a construction of first generation wavelets which relies on polynomial interpolation and subdivision as construction tools rather than the Fourier transform. It thus can be generalized to interval constructions [59] or weighed wavelets [104]. Lounsbery, De Rose, and Warren [79, 80] construct wavelets for the approximation of polyhedral surfaces of arbitrary genus. The wavelets are constructed by orthogonalizing scaling functions in a local neighborhood. We will

show later how this can be seen as a special case of lifting.

The lifting scheme can also be used to construct first generation wavelets; see [105, 52]. Although in this setting, the lifting will never come up with wavelets which could not have been found using the Cohen–Daubechies–Feauveau machinery in [29], it leads to two new insights: a custom-design construction of wavelets and a faster, in-place implementation of existing wavelet transforms [52]. In the first generation setting, lifting has many contacts with certain filter design algorithms used in signal processing. Those connections are pointed out in [105, 52].

Over the last few years Donovan, Hardin, Geronimo, and Massopust have developed techniques to construct wavelets based on fractal interpolation functions [60, 61, 62, 70]. They also introduced the concept of several generating functions (multiwavelets). As this technique does not rely on the Fourier transform either, it too potentially can be used to construct second generation wavelets.

Several spatial constructions of spline wavelets on irregular grids have been proposed [15, 7]. In [45], Dahmen and Micchelli propose a spatial construction of compactly supported wavelets that generate complementary spaces in a multiresolution analysis of univariate irregular knot splines.

Dahmen already made use of a technique related to lifting in the first generation setting [40] and later introduced a multiscale framework related to second generation wavelets [41].

Finally, after finishing this work, the author learned of two other very similar techniques developed independent of each other and of lifting. Harten and Abgrall developed a general multiresolution approximation framework based on prediction [71, 1], while Dahmen and co-workers [17, 46] developed a mechanism to characterize *all* stable biorthogonal decomposition. We will come back to this toward the end of the paper.

3. Multiresolution analysis. In this section we present the second generation version of multiresolution analysis. We keep most of the terminology and symbols of the first generation case, although their meaning can be quite different. For example, we maintain the name scaling function although it can be a little misleading since the scaling function can no longer be written as linear combinations of scaled versions of itself.

Consider a general function space $\mathrm{L}_2 = \mathrm{L}_2(X, \Sigma, \mu)$, with $X \subset \mathbf{R}^n$ being the spatial domain, Σ a σ-algebra, and μ a nonatomic measure on Σ. We do not require the measure to be translation invariant, so weighted measures are allowed. We assume (X, d) is a metric space.

DEFINITION 3.1. *A* multiresolution analysis $\mathbf{M}$ *of* L_2 *is a sequence of closed subspaces* $\mathbf{M} = \{V_j \subset \mathrm{L}_2 \mid j \in \mathcal{J} \subset \mathbf{Z}\}$ *so that*

1. $V_j \subset V_{j+1}$,
2. $\bigcup_{j\in\mathcal{J}} V_j$ *is dense in* L_2,
3. *for each* $j \in \mathcal{J}$, V_j *has a Riesz basis given by* scaling functions $\{\varphi_{j,k} \mid k \in \mathcal{K}(j)\}$.

One can think of $\mathcal{K}(j)$ as a general index set. We assume that $\mathcal{K}(j) \subset \mathcal{K}(j+1)$. We consider two cases:

I: $\mathcal{J} = \mathbf{N}$: This means there is one coarsest level V_0. This is the case if $\mu(X) < \infty$.

II: $\mathcal{J} = \mathbf{Z}$: We have a fully bi-infinite setting. This is typical when $\mu(X) = \infty$.

We then add the condition that

$$\bigcap_{j\in\mathcal{J}} V_j = \{\mathbf{0}\}\,.$$

A *dual multiresolution analysis* $\widetilde{\mathbf{M}} = \{\widetilde{V}_j \mid j \in \mathcal{J}\}$ consists of spaces $\widetilde{V}_j$ with Riesz bases given by *dual scaling functions* $\widetilde{\varphi}_{j,k}$. These dual scaling functions are biorthogonal with the scaling functions in the sense that

$$\langle\, \varphi_{j,k}, \widetilde{\varphi}_{j,k'} \,\rangle \;=\; \delta_{k,k'} \quad \text{for} \quad k,\, k' \in \mathcal{K}(j)\,. \tag{3.1}$$

For $f \in \mathrm{L}_2$, define the coefficients $\lambda_{j,k} \;=\; \langle\, f, \widetilde{\varphi}_{j,k} \,\rangle$ and consider the projections

$$P_j\, f \;=\; \sum_{k\in\mathcal{K}(j)} \lambda_{j,k}\, \varphi_{j,k}.$$

If the projection operators P_j are uniformly bounded in L_2, then

$$\lim_{j\to\infty} \|f - P_j\, f\| = 0.$$

First generation scaling functions reproduce polynomials up to a certain degree. To generalize this, consider a set of $\mathcal{C}^\infty$ functions on X, $\{P_p \mid p = 0, 1, 2, \ldots\}$, with $P_0 \equiv 1$ and so that the restrictions of a finite number of these functions to any ϵ-ball are linearly independent. We then say that the *order of the multiresolution analysis* is N if for all $j \in \mathcal{J}$, each P_p with $0 \le p < N$ can be represented pointwise as a linear combination of the $\{\varphi_{j,k} \mid k \in \mathcal{K}(j)\}$,

$$P_p(x) \;=\; \sum_{k\in\mathcal{K}(j)} c^p_{j,k}\, \varphi_{j,k}(x)\,.$$

We let $\widetilde{N}$ be the order of the dual multiresolution analysis, where we use a similar set of functions $\widetilde{P}_p$. In case X is a domain in $\mathbf{R}^n$, the functions P_p typically will be polynomials; in case X is a manifold, the functions P_p can, e.g., be parametric images of polynomials. However, in a practical situation one often has no explicit knowledge of the parameterization. This is why we use a very general definition of the order. Our definition obviously depends on the choice of P_p, but we do not include this dependency in the notation to avoid overloading. Most of the examples only have $N = 1$ in which case there is no dependency as $P_0 = 1$.

We assume that the dual functions are integrable and normalize them as

$$\int_X \widetilde{\varphi}_{j,k}\, d\mu \;=\; 1\,. \tag{3.2}$$

This implies that if $N > 0$,

$$\sum_{k\in\mathcal{K}(j)} \varphi_{j,k}(x) \;=\; 1\,. \tag{3.3}$$

4. Cascade algorithm. A question which immediately arises is how to construct scaling functions and dual scaling functions. As in the first generation case, there is often no analytic expression for them, and they are only defined through an iterative procedure, the *cascade algorithm*. In this section we present the second

generation version of the cascade algorithm. To do so we need two things: a *set of partitionings* and a *filter*.

Let us start by defining a filter. The definition of multiresolution analysis implies that for every scaling function $\varphi_{j,k}$ ($j \in \mathcal{J}$, $k \in \mathcal{K}(j)$), coefficients $\{h_{j,k,l} \mid l \in \mathcal{K}(j+1)\}$ exist so that formally

$$\varphi_{j,k} = \sum_{l \in \mathcal{K}(j+1)} h_{j,k,l}\, \varphi_{j+1,l}\,. \tag{4.1}$$

We refer to this equation as a *refinement relation*. Each scaling function can be written as a linear combination of scaling functions on the next finer level. To ensure that the summation in (4.1) is well defined we need to clearly state the definition of a *filter*. In this paper we only consider finite filters.

DEFINITION 4.1. *A set of real numbers* $\{h_{j,k,l} \mid j \in \mathcal{J}, k \in \mathcal{K}(j), l \in \mathcal{K}(j+1)\}$ *is called a* finite filter *if*

1. *For each* j *and* k *only a finite number of coefficients* $h_{j,k,l}$ *are nonzero, and thus the set*

$$\mathcal{L}(j,k) = \{l \in \mathcal{K}(j+1) \mid h_{j,k,l} \neq 0\}$$

   is finite.
2. *For each* j *and* l *only a finite number of coefficients* $h_{j,k,l}$ *are nonzero, and thus the set*

$$\mathcal{K}(j,l) = \{k \in \mathcal{K}(j) \mid h_{j,k,l} \neq 0\}$$

   is finite.
3. *The size of sets* $\mathcal{L}(j,k)$ *and* $\mathcal{K}(j,l)$ *is uniformly bounded for all* j, k, *and* l.

Note that in the first generation case $h_{j,k.l} = h_{l-2k}$, so if $\{h_k \mid k\}$ is a finite sequence, the filter is finite according to the above definition. We will always choose our indices consistently so that $j \in \mathcal{J}$, $k \in \mathcal{K}(j)$, and $l \in \mathcal{K}(j+1)$, even though it will not always be explicitly mentioned. The above defined index sets indicate which elements are nonzero on each row (respectively, column) of the (possibly infinite) matrix $\{h_{j,k,l} \mid k \in \mathcal{K}(j), l \in \mathcal{K}(j+1)\}$. We can think of them as adjoints of each other as

$$\mathcal{K}(j,l) = \{k \in \mathcal{K}(j) \mid l \in \mathcal{L}(j,k)\}\,.$$

The dual scaling functions satisfy refinement relations with coefficients $\{\widetilde{h}_{j,k,l}\}$. We can define similar index sets (denoted with a tilde).

A set of partitionings $\{S_{j,k}\}$ can be thought of as the replacement for the dyadic intervals on the real line in the first generation case. Again each scaling function $\varphi_{j,k}$ is associated with exactly one set $S_{j,k}$. We use the following definition.

DEFINITION 4.2. *A set of measurable subsets* $\{S_{j,k} \in \Sigma \mid j \in \mathcal{J}, k \in \mathcal{K}(j)\}$ *is called a* set of partitionings *if*

1. $\forall j \in \mathcal{J}$: $\operatorname{clos} \bigcup_{k \in \mathcal{K}(j)} S_{j,k} = X$ *and the union is disjoint,*
2. $\mathcal{K}(j) \subset \mathcal{K}(j+1)$,
3. $S_{j+1,k} \subset S_{j,k}$,
4. *For a fixed* $k \in \mathcal{K}(j_0)$, $\bigcap_{j > j_0} S_{j,k}$ *is a set which contains one point. We denote this point with* x_k.

The purpose now is to use a filter and a set of partitionings to construct scaling functions that satisfy (4.1). Assume we want to synthesize φ_{j_0,k_0}. First define a Kronecker sequence $\{\lambda_{j_0,k} = \delta_{k,k_0} \mid k \in \mathcal{K}(j_0)\}$. Then, generate sequences $\{\lambda_{j,k} \mid k \in \mathcal{K}(j)\}$ for $j > j_0$ by recursively applying the formula

$$\lambda_{j+1,l} \;=\; \sum_{k \in \mathcal{K}(j,l)} h_{j,k,l}\, \lambda_{j,k}\,.$$

Next we construct the functions

$$f^{(j)}_{j_0,k_0} \;=\; \sum_{k \in \mathcal{K}(j)} \lambda_{j,k}\, \chi_{S_{j,k}}, \qquad j \geq j_0\,. \tag{4.2}$$

These functions satisfy, for $j > j_0$,

$$f^{(j)}_{j_0,k_0} \;=\; \sum_{l} h_{j_0,k_0,l}\, f^{(j)}_{j_0+1,l}\,. \tag{4.3}$$

If $\lim_{j\to\infty} f^{(j)}_{j_0,k_0}$ converges to a function in L_2, we define this function to be φ_{j_0,k_0}. This procedure is called the *cascade algorithm*. The limit functions satisfy

$$\lim_{j\to\infty} \lambda_{j,k} \;=\; \varphi_{j_0,k_0}(x_k) \quad \text{a.e.}$$

If the cascade algorithm converges for all j_0 and k_0, we get a set of scaling functions that satisfies the refinement equation (4.1). This can be seen by letting j go to infinity in (4.3). Note how the resulting functions depend both on the filter and the set of partitionings. If the scaling functions generate a multiresolution analysis, the cascade algorithm started with a sequence $\{\lambda_{j_0,k} \mid k \in \mathcal{K}(j)\}$ that belongs to $\ell^2(\mathcal{K}(j))$ converges to

$$\sum_k \lambda_{j_0,k}\, \varphi_{j_0,k}\,.$$

The dual scaling functions are constructed similarly starting from a finite filter $\widetilde{h}$, the same set of partitionings, and an initial Kronecker sequence $\{\lambda_{j_0,k} = \delta_{k,k_0}/\mu(S_{j_0,k_0}) \mid k \in \mathcal{K}(j_0)\}$. The normalization of the initial sequences assures that $\langle\, \widetilde{\varphi}_{j,k}, \varphi_{j,k}\,\rangle = 1$.

An interesting question is now whether the biorthogonality condition (3.1) can be related back to the filters h and $\widetilde{h}$. By writing out the refinement relations we see that the biorthogonality (3.1) implies that

$$\sum_l h_{j,k,l}\, \widetilde{h}_{j,k',l} \;=\; \delta_{k,k'} \quad \text{for} \quad j \in \mathcal{J},\, k, k' \in \mathcal{K}(j)\,, \tag{4.4}$$

but the converse is not immediately true. More precisely, if the filter coefficients satisfy (4.4) and the cascade algorithm for the primal and dual scaling functions converges, then the resulting scaling functions are biorthogonal. This follows from the fact that (4.4) assures that the intermediate functions of the form $f^{(j)}_{j_0,k_0}$ in (4.2) (which converge to the scaling functions) are biorthogonal at each stage j.

It is important to note that not every filter corresponds to a set of scaling functions; i.e., the convergence of the cascade algorithm is not guaranteed. We would like to have a condition which relates convergence of the cascade algorithm and the Riesz basis property back to the filter coefficients, similar to the Cohen criterion in the first generation case [26] or the Cohen–Daubechies–Feauveau theorem [29] or [48, Theorem 8.3.1]. This result is part of the analysis phase of the construction. As we mentioned earlier, this paper is mostly concerned with the algebraic phase and the generation of the filter coefficients.

5. Wavelets. First generation wavelets are defined as basis functions for spaces complementing V_j in V_{j+1}. The same idea remains in the second generation case. This leads to the following definition.

DEFINITION 5.1. *A set of functions* $\{\psi_{j,m} \mid j \in \mathcal{J},\, m \in \mathcal{M}(j)\}$, *where* $\mathcal{M}(j) = \mathcal{K}(j+1) \setminus \mathcal{K}(j)$, *is a* set of wavelet functions *if*

1. *The space* $W_j = \operatorname{clos}\operatorname{span}\{\psi_{j,m} \mid m \in \mathcal{M}(j)\}$ *is a complement of* V_j *in* V_{j+1} *and* $W_j \perp \widetilde{V}_j$.
2. *If* $\mathcal{J} = \mathbf{Z}$, *the set* $\{\psi_{j,m}/\|\psi_{j,m}\| \mid j \in \mathcal{J}, m \in \mathcal{M}(j)\}$ *is a Riesz basis for* L_2. *If* $\mathcal{J} = \mathbf{N}$, *the set* $\{\psi_{j,m}/\|\psi_{j,m}\| \mid j \in \mathcal{J}, m \in \mathcal{M}(j)\} \cup \{\varphi_{0,k}/\|\varphi_{0,k}\| \mid k \in \mathcal{K}(0)\}$ *is a Riesz basis for* L_2.

We always assume that the index m belongs to the set $\mathcal{M}(j)$. The dual basis is given by *dual wavelets* $\widetilde{\psi}_{j,m}$, which are biorthogonal to the wavelets

$$\langle\, \psi_{j,m}, \widetilde{\psi}_{j',m'} \,\rangle \;=\; \delta_{m,m'}\, \delta_{j,j'}\,. \tag{5.1}$$

The dual wavelets span spaces $\widetilde{W}_j$ which complement $\widetilde{V}_j$ in $\widetilde{V}_{j+1}$ and $\widetilde{W}_j \perp V_j$. For $f \in \mathrm{L}_2$, define the coefficients $\gamma_{j,m} \;=\; \langle\, f, \widetilde{\psi}_{j,m} \,\rangle$. Then

$$f \;=\; \sum_{j,m} \gamma_{j,m}\, \psi_{j,m}\,.$$

Their definition implies that the wavelets satisfy refinement relations of the form

$$\psi_{j,m} \;=\; \sum_{l} g_{j,m,l}\, \varphi_{j+1,l}\,. \tag{5.2}$$

We assume that $g = \{g_{j,m,l} \mid j \in \mathcal{J}, m \in \mathcal{M}(j), l \in \mathcal{K}(j+1)\}$ is a finite filter according to Definition 4.1 with k substituted by m. This leads to the definition of the uniformly bounded finite sets

$$\mathcal{M}(j,l) \;=\; \{m \in \mathcal{M}(j) \mid g_{j,m,l} \neq 0\} \quad \text{and} \quad \mathcal{L}(j,m) \;=\; \{l \in \mathcal{K}(j+1) \mid m \in \mathcal{M}(j,l)\}\,.$$

The dual wavelets satisfy refinement relations with a finite filter $\widetilde{g}$.

Also, since $\varphi_{j+1,l} \in V_j \oplus W_j$, it holds that

$$\varphi_{j+1,l} \;=\; \sum_{k} \widetilde{h}_{j,k,l}\, \varphi_{j,k} + \sum_{m} \widetilde{g}_{j,m,l}\, \psi_{j,m}\,.$$

The biorthogonality (5.1) combined with (3.1) implies the following relations between the filters:

$$\begin{aligned} \sum_l g_{j,m,l}\, \widetilde{g}_{j,m',l} &= \delta_{m,m'}\,, & \sum_l h_{j,k,l}\, \widetilde{g}_{j,m,l} &= 0\,, \\ \sum_l h_{j,k,l}\, \widetilde{h}_{j,k',l} &= \delta_{k,k'}\,, & \sum_l g_{j,m,l}\, \widetilde{h}_{j,k,l} &= 0\,. \end{aligned} \tag{5.3}$$

DEFINITION 5.2. *A set of filters* $\{h, \widetilde{h}, g, \widetilde{g}\}$ *is a* set of biorthogonal *filters if condition* (5.3) *is satisfied.*

Now given a set of biorthogonal filters and a set of partitionings and assuming that the cascade algorithm converges, the resulting scaling functions, wavelets, dual

scaling functions, and dual wavelets are biorthogonal in the sense that

$$\begin{aligned} \langle\, \widetilde{\varphi}_{j,k}, \varphi_{j,k'} \,\rangle &= \delta_{k,k'}\,, \\ \langle\, \widetilde{\psi}_{j,m}, \psi_{j,m'} \,\rangle &= \delta_{m,m'}\,, \\ \langle\, \widetilde{\varphi}_{j,k}, \psi_{j,m} \,\rangle &= 0\,, \\ \langle\, \widetilde{\psi}_{j,m}, \varphi_{j,k} \,\rangle &= 0\,. \end{aligned}$$

Next we need to generalize the notion of vanishing polynomial moments. We therefore use the (nonpolynomial) functions P_p defined in section 3. If the scaling functions $\varphi_{j,k}$ with $k \in \mathcal{K}(j)$ reproduce P_p, then

$$\int_X P_p\, \widetilde{\psi}_{j,m}\, d\mu \;=\; 0 \quad \text{for} \quad 0 \le p < N,\ j \in \mathcal{J},\ m \in \mathcal{M}(j)\,.$$

We say that the dual wavelets have N vanishing moments. Similarly, the wavelets have $\widetilde{N}$ vanishing moments.

6. Fast wavelet transform. The basic idea of a wavelet transform is the same as in the first generation case. Given the set of coefficients $\{\lambda_{n,k} \mid k \in \mathcal{K}(n)\}$, calculate the $\{\gamma_{j,m} \mid n_0 \le j < n,\, m \in \mathcal{M}(j)\}$ and $\{\lambda_{n_0,k} \mid k \in \mathcal{K}(n_0)\}$. From the refinement relation of the dual scaling functions and wavelets, we see that a fast forward wavelet transform is given by recursive application of

$$\lambda_{j,k} = \sum_{l \in \widetilde{\mathcal{L}}(j,k)} \widetilde{h}_{j,k,l}\, \lambda_{j+1,l} \quad \text{and} \quad \gamma_{j,m} = \sum_{l \in \widetilde{\mathcal{L}}(j,m)} \widetilde{g}_{j,m,l}\, \lambda_{j+1,l}\,.$$

Similarly, the inverse transform follows from the recursive application of

$$\lambda_{j+1,l} \;=\; \sum_{k \in \mathcal{K}(j,l)} h_{j,k,l}\, \lambda_{j,k} \;+\; \sum_{m \in \mathcal{M}(j,l)} g_{j,m,l}\, \gamma_{j,m}\,.$$

The major difference with the first generation fast wavelet transform, and thus with traditional subband transforms, is that the filter coefficients are different for every coefficient. One has to be careful analyzing the complexity of the second generation fast wavelet transform. For general filters the complexity need not be linear as the number of terms in the above summation, albeit finite, can grow from level to level. This is precisely why Definition 4.1 of a finite filter requires the sizes of the index sets $\mathcal{L}$, $\mathcal{K}$, and $\mathcal{M}$ to be *uniformly* bounded. This leads to the following corollary.

COROLLARY 6.1. *In case the filters h, g, $\widetilde{h}$, and $\widetilde{g}$ are finite, the second generation fast wavelet transform is a linear time algorithm.*

Note that in a computer implementation the data structure for the filters can become much more complex than in the first generation case and therefore has to be designed carefully.

In case the wavelets form an unconditional basis, the condition number of the wavelet transform is bounded independent of the number of levels. Consequently the propagation of numerical round-off error in floating point calculations will be bounded. As we mentioned before, lifting does not guarantee stability and bounded condition numbers. However, in a practical situation involving spherical wavelets [99] we numerically estimated the condition number and found it to vary little with the number of levels. For a spherical wavelet transform involving roughly $650,000$ coefficients we found the condition number to be approximately 8.

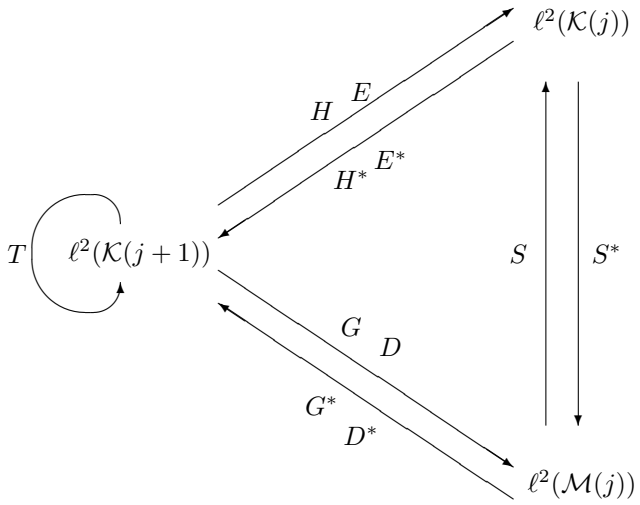

FIG. 7.1. *Schematic representation of operators, their domain, and range. This scheme can be used to verify that the order in which operators are applied is correct and which operators can be added.*

7. A word on notation. So far we have been using a notation involving the filter coefficients $h_{j,k,l}$ and $g_{j,m,l}$. As one can see this leads to expressions involving many indices. We will refer to it as the *index notation*. In this section we introduce a new notation, which we refer to as the *operator notation*. The advantage is that both the statement and the proof of some results become more elegant. Statements in the operator notation will also formally look the same as in the first generation case. In this way it helps to shed light on why things work. The disadvantage is that the operator notation is not practical and that it obscures implementation. Therefore we always state results in the index notation as well.

First consider the spaces $\ell^2(\mathcal{K}(j+1))$, $\ell^2(\mathcal{K}(j))$, and $\ell^2(\mathcal{M}(j))$, with their usual norm and inner product. We denote elements of these spaces by, respectively, a, b, and c so that

$$a \;=\; \{a_l \mid l \in \mathcal{K}(j+1)\} \in \ell^2(\mathcal{K}(j+1))\,,$$

and, similarly, mutatis mutandis, for $b \in \ell^2(\mathcal{K}(j))$ and $c \in \ell^2(\mathcal{M}(j))$. We always denote the identity operator on these spaces with 1. It should be clear from the context which one is meant. Next we introduce two operators (see also Figure 7.1):

1. $H_j : \ell^2(\mathcal{K}(j+1)) \to \ell^2(\mathcal{K}(j))$, where $b = H_j\, a$ means that

$$b_k \;=\; \sum_{l \in \mathcal{K}(j+1)} h_{j,k,l}\, a_l\,.$$

2. $G_j : \ell^2(\mathcal{K}(j+1)) \to \ell^2(\mathcal{M}(j))$, where $c = G_j\, a$ means that

$$c_m \;=\; \sum_{l \in \mathcal{K}(j+1)} g_{j,m,l}\, a_l\,.$$

The operators $\widetilde{H}_j$ and $\widetilde{G}_j$ are defined similarly. We refer to these operators as *filter operators* or sometimes simply as filters.

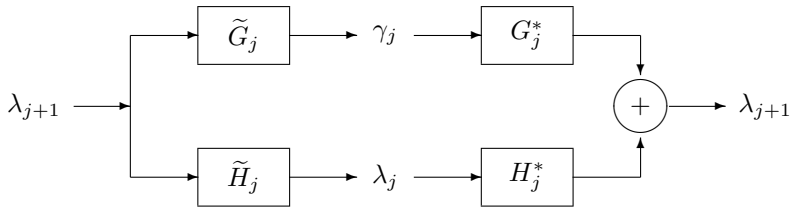

FIG. 7.2. *The fast wavelet transform. The major difference with the first generation fast wavelet transform is that the filters potentially are different for each coefficient. Observe that the subsampling is absorbed into the filters.*

We can now write the fast wavelet transform in operator notation. Define the sequences $\lambda_j = \{\lambda_{j,k} \mid k\}$ and $\gamma_j = \{\gamma_{j,k} \mid m\}$. Then one step in the forward transform is given by

$$\lambda_j \;=\; \widetilde{H}_j\, \lambda_{j+1} \quad \text{and} \quad \gamma_j \;=\; \widetilde{G}_j\, \lambda_{j+1}\,,$$

and one step in the inverse transform is given by

$$\lambda_{j+1} \;=\; H_j^*\, \lambda_j + G_j^*\, \gamma_j\,.$$

One step of the transform is depicted as a block diagram in Figure 7.2. We use here a scheme similar to a subband transform. Note how the traditional subsampling is absorbed into the filter operators.

The conditions on the filter operators for exact reconstruction now readily follow:

$$\widetilde{H}_j\, H_j^* \;=\; \widetilde{G}_j\, G_j^* \;=\; 1\,, \qquad\qquad \widetilde{G}_j\, H_j^* \;=\; \widetilde{H}_j\, G_j^* \;=\; 0\,,$$

and

$$H_j^*\, \widetilde{H}_j + G_j^*\, \widetilde{G}_j \;=\; 1\,.$$

These we can write in matrix form as

$$\text{(7.1)} \quad \begin{bmatrix} \widetilde{H}_j \\ \widetilde{G}_j \end{bmatrix} \begin{bmatrix} H_j^* & G_j^* \end{bmatrix} \;=\; \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} H_j^* & G_j^* \end{bmatrix} \begin{bmatrix} \widetilde{H}_j \\ \widetilde{G}_j \end{bmatrix} \;=\; 1\,.$$

DEFINITION 7.1. *The set of filter operators* $\{H_j, \widetilde{H}_j, G_j, \widetilde{G}_j\}$ *is a* set of biorthogonal filter operators *if condition* (7.1) *is satisfied.*

With slight abuse of notation, i.e., by letting the operators work on sequences of functions, we can write the refinement relations. Define $\varphi_j = \{\varphi_{j,k} \mid k \in \mathcal{K}(j)\}$ and $\psi_j = \{\psi_{j,m} \mid m \in \mathcal{M}(j)\}$. Then

$$\varphi_j \;=\; H_j\, \varphi_{j+1} \quad \text{and} \quad \psi_j \;=\; G_j\, \varphi_{j+1}\,.$$

In the other direction we have

$$\varphi_{j+1} \;=\; \widetilde{H}_j^*\, \varphi_j + \widetilde{G}_j^*\, \psi_j\,.$$

Armed with this operator notation, we now can state the lifting scheme.

8. The lifting scheme. In this section we state and prove the lifting scheme and show how it can be used to construct second generation wavelets.

THEOREM 8.1 (lifting). *Take an initial set of biorthogonal filter operators* $\{H_j^{\text{old}}, \widetilde{H}_j^{\text{old}}, G_j^{\text{old}}, \widetilde{G}_j^{\text{old}}\}$. *Then a new set of biorthogonal filter operators* $\{H_j, \widetilde{H}_j, G_j, \widetilde{G}_j\}$ *can be found as*

$$\begin{aligned} H_j &= H_j^{\text{old}}\,, \\ \widetilde{H}_j &= \widetilde{H}_j^{\text{old}} + S_j\,\widetilde{G}_j^{\text{old}}\,, \\ G_j &= G_j^{\text{old}} - S_j^*\,H_j^{\text{old}}\,, \\ \widetilde{G}_j &= \widetilde{G}_j^{\text{old}}\,, \end{aligned}$$

where S_j *is an operator from* $\ell^2(\mathcal{M}(j))$ *to* $\ell^2(\mathcal{K}(j))$.

Proof. We write the lifting scheme in matrix notation:

$$\begin{bmatrix} \widetilde{H}_j \\ \widetilde{G}_j \end{bmatrix} = \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \widetilde{H}_j^{\text{old}} \\ \widetilde{G}_j^{\text{old}} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} H_j \\ G_j \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -S^* & 1 \end{bmatrix} \begin{bmatrix} H_j^{\text{old}} \\ G_j^{\text{old}} \end{bmatrix}.$$

If we think of the biorthogonality conditions (7.1) in the matrix notation, the proof simply follows from the fact that

$$\begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -S \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \qquad \square$$

One can use Figure 7.1 to assert that the order of the operators H, G, and S is correct. The theorem in the index notation reads as follows.

THEOREM 8.2 (lifting in index notation). *Take an initial set of biorthogonal filters* $\{h^{\text{old}}, \widetilde{h}^{\text{old}}, g^{\text{old}}, \widetilde{g}^{\text{old}}\}$; *then a new set of biorthogonal filters* $\{h, \widetilde{h}, g, \widetilde{g}\}$ *can be constructed as*

$$\begin{aligned} h_{j,k,l} &= h_{j,k,l}^{\text{old}}\,, \\ \widetilde{h}_{j,k,l} &= \widetilde{h}_{j,k,l}^{\text{old}} + \sum_m s_{j,k,m}\,\widetilde{g}_{j,m,l}^{\text{old}}\,, \\ g_{j,m,l} &= g_{j,m,l}^{\text{old}} - \sum_k s_{j,k,m}\,h_{j,k,l}^{\text{old}}\,, \\ \widetilde{g}_{j,m,l} &= \widetilde{g}_{j,m,l}^{\text{old}}\,. \end{aligned}$$

After lifting, the filters h and $\widetilde{g}$ remain the same, while the filters $\widetilde{h}$ and g change. As h remains the same, so do the primal scaling functions. The dual scaling functions and primal wavelets change since $\widetilde{h}$ and g change. The dual wavelets also change because the dual scaling functions, from which they are built, change. However, the coefficients $\widetilde{g}$ of the refinement equation of the dual wavelet remain the same. More precisely, we have

$$\begin{aligned} \varphi_j &= \varphi_j^{\text{old}}\,, \\ \widetilde{\varphi}_j &= \widetilde{H}_j^{\text{old}}\widetilde{\varphi}_{j+1} + S_j\,\widetilde{G}_j^{\text{old}}\,\widetilde{\varphi}_{j+1} \;=\; \widetilde{H}_j^{\text{old}}\widetilde{\varphi}_{j+1} + S_j\,\widetilde{\psi}_j\,, \\ \psi_j &= G_j^{\text{old}}\,\varphi_{j+1} - S_j^*\,H_j^{\text{old}}\varphi_{j+1} \;=\; \psi_j^{\text{old}} - S_j\,\varphi_j^{\text{old}}\,, \\ \widetilde{\psi}_j &= \widetilde{G}_j^{\text{old}}\,\widetilde{\varphi}_j\,, \end{aligned}$$

or

$$\varphi_{j,k} = \varphi^{\text{old}}_{j,k}\,,$$

$$\widetilde{\varphi}_{j,k} = \sum_l \widetilde{h}^{\text{old}}_{j,k,l}\,\widetilde{\varphi}_{j+1,l} + \sum_m s_{j,k,m}\,\widetilde{\psi}_{j,m}\,, \tag{8.1}$$

$$\psi_{j,m} = \psi^{\text{old}}_{j,m} - \sum_k s_{j,k,m}\,\varphi^{\text{old}}_{j,k}\,, \tag{8.2}$$

$$\widetilde{\psi}_{j,m} = \sum_l \widetilde{g}^{\text{old}}_{j,k,m}\,\widetilde{\varphi}_{j+1,l}\,. \tag{8.3}$$

Although formally similar, the expressions in (8.2) and (8.1) are quite different. The difference lies in the fact that in (8.2) the scaling functions on the right-hand side did not change after lifting, while in (8.1) the functions on the right-hand side did change after lifting. Indeed, the dual wavelets on the right-hand side of (8.1) already are the new ones.

The power behind the lifting scheme is that through the operator S we have full control over all wavelets and dual functions that can be built from a particular set of scaling functions. This means we can start from a simple or trivial multiresolution analysis and use (8.2) to choose S so that the wavelets after lifting have particular properties. This allows custom design of the wavelet, and it is the motivation behind the name "lifting scheme."

The fundamental idea behind the lifting scheme is that instead of using scaling functions on the *finer* level to build a wavelet, as in (5.2), we use an old, simple wavelet and scaling functions on the *same* level to synthesize a new wavelet; see (8.2). Thus instead of using "sister" scaling functions, we use "aunt" scaling functions of the family tree to build wavelets. As we will point out later, the "aunt" property is fundamental when building adaptive wavelets. The advantage of using (8.2) as opposed to (5.2) for the construction of $\psi_{j,m}$ is that in the former we have total freedom in the choice of S. Once S is fixed, the lifting scheme assures that all filters are biorthogonal. If we use (5.2) to construct ψ, we would have to check the biorthogonality separately.

Equation (8.2) is also the key to finding the S operator, since functions on the right-hand side do not change. Conditions on $\psi_{j,m}$ thus immediately translate into conditions on S. For example, we can choose S to increase the number of vanishing moments of the wavelet or choose S so that $\psi_{j,m}$ resembles a particular shape.

If the original filters and S are finite filters, then the new filters will be finite as well. In such case define the (adjoint) sets

$$\mathcal{K}(j,m) \;=\; \{k \mid s_{j,k,m} \neq 0\} \quad \text{and} \quad \mathcal{M}(j,k) \;=\; \{m \mid k \in \mathcal{K}(j,m)\}\,.$$

If we want the wavelet to have vanishing moments, the condition that the integral of a wavelet multiplied with a certain function P_p is zero leads to

$$\int_X P_p\,\psi_{j,m}\,d\mu \;=\; 0 \;\Rightarrow\; \int_X P_p\,\psi^{\text{old}}_{j,m}\,d\mu \;=\; \sum_{k\in\mathcal{K}(j,m)} s_{j,k,m} \int_X P_p\,\varphi^{\text{old}}_{j,k}\,d\mu\,.$$

For fixed indices j and m, the latter is a linear equation in the unknowns $\{s_{j,k,m} \mid k \in \mathcal{K}(j,m)\}$. All coefficients only depend on the old multiresolution analysis. If we choose the number of unknown coefficients $s_{j,k,m}$ equal to the number of equations, we simply need to solve a linear system for each j and m. Remember that the functions

P_p had to be independent, so if the functions $\varphi_{j,k}^{\text{old}}$ are independent as well, the system will be full rank.

Notes.

1. Other constraints than vanishing moments can be used for the choice of S. For example one can custom design the shape of the wavelet for use in feature recognition. Given the scaling functions, choose S so that $\psi_{j,m}$ resembles the particular feature we want to recognize. The magnitude of the wavelet coefficients is now proportional to how much the original signal at the particular scale and place resembles the feature. This has important applications in automated target recognition and medical imaging. Other ideas are fixing the value of the wavelet or the value of the derivative of the wavelet at a certain location. This is useful to accommodate boundary conditions.
2. In general it is not possible to use lifting to build orthogonal or semi-orthogonal wavelets using only finite lifting filters. In the semi-orthogonal case, the condition that a new wavelet $\psi_{j,m}$ is orthogonal to the V_j typically will require using all $\varphi_{j,k}$ scaling functions of level j in the lifting (8.2). In [80] this was bypassed by *pseudo-orthogonalization,* a scheme where $\psi_{j,m}$ is only required to be orthogonal to the (interpolating) scaling functions in a certain neighborhood. As mentioned in the introduction, part of the inspiration of the lifting scheme came from generalizing this idea to a fully biorthogonal setting.
3. In [17] several examples, including splines with nonuniform knot sequences, are given where semi-orthogonal wavelets are constructed. This construction uses all possible degrees of freedom for the construction of the wavelet, which is more than what lifting allows, but does not lead to finite primal and dual filters.
4. One of the appealing features of using the lifting scheme in the construction of second generation wavelets is that one gets the filters for the scaling functions *and* the wavelets together. Other constructions, such as nonstationary subdivision, only give the filters for the scaling functions; see for example [104, Chapter 5]. One then needs to use a technical trick to find the wavelet filters with the right biorthogonality properties. There is no guarantee that this is always possible.

9. Fast lifted wavelet transform. In this section we show how the lifting scheme can be used to facilitate and accelerate the implementation of the fast wavelet transform. The basic idea is to never explicitly form the new filters but only work with the old filter, which can be trivial, and the S filter.

For the forward transform we get

$$\lambda_j \;=\; \widetilde{H}_j\,\lambda_{j+1} \;=\; \widetilde{H}_j^{\text{old}}\lambda_{j+1} + S_j\,\gamma_j\,.$$

In index notation this becomes

$$\lambda_{j,k} \;=\; \sum_l \widetilde{h}_{j,k,l}^{\text{old}}\,\lambda_{j+1,l} + \sum_m s_{j,k,m}\,\gamma_{j,m}\,.$$

This implies that if we first calculate the wavelet coefficients γ_j as $\widetilde{G}_j^{\text{old}}\lambda_{j+1}$, we can later *reuse* them in the calculation of the λ_j coefficients. The λ_j are first calculated as $\widetilde{H}_j^{\text{old}}\,\lambda_{j+1}$ and later updated (lifted) with the γ_j coefficients. This way we never have to form the (potentially large) filter $\widetilde{H}_j$. In other words, the lifting scheme makes

optimal use of the similarities between the $\widetilde{H}$ and $\widetilde{G}$ filter. This both facilitates and accelerates the implementation.

For the inverse transform we find that

$$\lambda_{j+1} \;=\; H_j^*\,\lambda_j + G_j^*\,\gamma_j \;=\; H_j^{\mathrm{old}\,*}(\lambda_j - S_j\,\gamma_j) + G_j^{\mathrm{old}\,*}\gamma_j\,.$$

In index notation this becomes

$$\lambda_{j+1,l} \;=\; \sum_k h^{\mathrm{old}}_{j,k,l}\left(\lambda_{j,k} - \sum_m s_{j,k,m}\gamma_{j,m}\right) + \sum_m g^{\mathrm{old}}_{j,m,l}\gamma_{j,m}\,.$$

The inverse transform thus first undoes the lifting (between the parentheses) and then does an inverse transform with the old filters.

This leads to the following algorithm for the *fast lifted wavelet transform* depicted in Figure 9.1. On each level the forward transform consists of two stages. Stage I is simply the forward transform with the old filters while stage II is the lifting. In the inverse transform, stage I simply undoes the lifting and stage II is an inverse transform with the old filters. In pseudo code this becomes

Forward wavelet transform	Inverse wavelet transform

`For j = n-1 downto 0`
`  Forward I(j)`
`  Forward II(j)`
	`For level = 0 to n-1`
	`  Inverse I(j)`
	`  Inverse II(j)`

`Forward I(j)`: Calculate the $\gamma_{j,m}$ and first stage of $\lambda_{j,k}$

$$\forall k \in \mathcal{K}(j)\,:\,\lambda_{j,k} := \sum_{l\in\widetilde{\mathcal{L}}(j,k)} \widetilde{h}^{\mathrm{old}}_{j,k,l}\,\lambda_{j+1,l}$$

$$\forall m \in \mathcal{M}(j)\,:\,\gamma_{j,m} := \sum_{l\in\widetilde{\mathcal{L}}(j,m)} \widetilde{g}^{\mathrm{old}}_{j,m,l}\,\lambda_{j+1,l}$$

`Forward II(j)`: Lift the $\lambda_{j,k}$ using the $\gamma_{j,m}$ calculated in Stage I

$$\forall k \in \mathcal{K}(j)\,:\,\lambda_{j,k} \mathrel{+}= \sum_{m\in\mathcal{M}(j,k)} s_{j,k,m}\,\gamma_{j,m}$$

`Inverse I(j)`: Undo the lifting

$$\forall k \in \mathcal{K}(j)\,:\,\lambda_{j,k} \mathrel{-}= \sum_{m\in\mathcal{M}(j,k)} s_{j,k,m}\,\gamma_{j,m}$$

`Inverse II(j)`: Calculate the $\lambda_{j+1,l}$ using the $\lambda_{j,k}$ from Stage I:

$$\forall l \in \mathcal{K}(j+1)\,:\,\lambda_{j+1,l} := \sum_{k\in\mathcal{K}(j,l)} h^{\mathrm{old}}_{j,k,l}\,\lambda_{j,k} + \sum_{m\in\mathcal{M}(j,l)} g^{\mathrm{old}}_{j,m,l}\,\gamma_{j,m}$$

As noted in [100], there are always two possibilities to implement these sums. For example, take the sum in the `Forward I` routine. We can either implement this as (after assigning 0 to $\gamma_{j,m}$)

$$\forall m \in \mathcal{M}(j)\,:\,\forall l \in \mathcal{L}(j,m)\,:\,\gamma_{j,m} \mathrel{+}= \widetilde{g}^{\mathrm{old}}_{j,m,l}\,\lambda_{j+1,l}$$

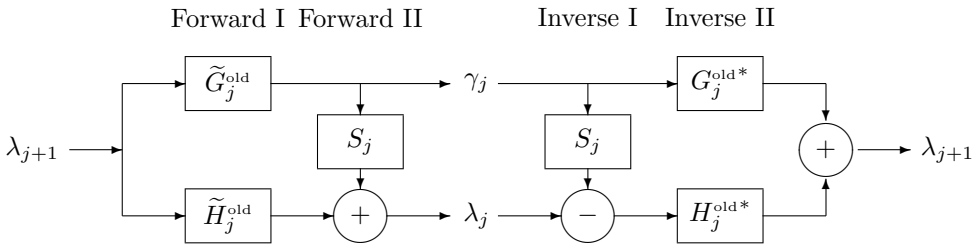

FIG. 9.1. *The fast lifted wavelet transform: the basic idea is to first perform a transform with the old, simple filters and later "lift" the scaling function coefficients with the help of wavelet coefficients. The inverse transform first undoes the lifting and then performs an inverse transform with the old filters.*

or as

$$\forall l \in \mathcal{K}(j+1) \,:\, \forall m \in \mathcal{M}(j,l) \,:\, \gamma_{j,m} \mathrel{+}= \widetilde{g}^{\text{old}}_{j,m,l}\, \lambda_{j+1,l}\,.$$

The first option loops over all m, for each $\gamma_{j,m}$ identifies the $\lambda_{j+1,l}$ that determine its value, then calculates the linear combination and assigns it into $\gamma_{j,m}$. The second option loops over all l, identifies the $\gamma_{j,m}$ which are influenced by $\lambda_{j+1,l}$, and then adds on the right amount to each $\gamma_{j,m}$. Both options are theoretically equivalent, but often one of the two is much easier to implement than the other; see for example [100]. There one of the index sets always contains the same number of elements, while the cardinality of the other can vary depending on the mesh.

10. Cakewalk construction. In this section we discuss how one can iterate the lifting scheme to bootstrap one's way up to a multiresolution analysis with desired properties.

We first introduce the *dual lifting scheme*. The basic idea is the same as for the lifting scheme except that we now leave the dual scaling function and the $\widetilde{H}$ and G filters untouched. The H and $\widetilde{G}$ filters and the dual wavelet, scaling function, and wavelet (by refinement) change. We can use the dual lifting scheme to custom design the dual wavelet. If we denote the operator involved with $\widetilde{S}_j$, the new set of biorthogonal filter operators is given by

$$\begin{aligned}
H_j &= H_j^{\text{old}} + \widetilde{S}_j\, G_j^{\text{old}}\,,\\
\widetilde{H}_j &= \widetilde{H}_j^{\text{old}}\,,\\
G_j &= G_j^{\text{old}}\,,\\
\widetilde{G}_j &= \widetilde{G}_j^{\text{old}} - \widetilde{S}_j^*\, \widetilde{H}_j^{\text{old}}\,,
\end{aligned}$$

where $\widetilde{S}_j$ is an operator from $\ell^2(\mathcal{M}(j))$ to $\ell^2(\mathcal{K}(j))$. Relationships like (8.2) and (8.1) can be obtained by simply toggling the tildes. In the second stage of the fast wavelet transform, the γ_j coefficients are now lifted with the help of the λ_j coefficients calculated in the first stage.

We now can alternate lifting and dual lifting. For example, after increasing the number of vanishing moments of the wavelet with the lifting scheme, one can use the dual lifting scheme to increase the number of vanishing moments of the dual wavelet. By iterating lifting and dual lifting, one can bootstrap one's way up to a multiresolution analysis with desired properties on primal and dual wavelets. This is the basic idea behind the *cakewalk* construction.

There is one issue that remains to be checked to allow cakewalk constructions. Suppose we first use dual lifting to increase the number of vanishing moments of the dual wavelet. How do we know that this will not be ruined by later lifting? Remember that lifting changes the dual scaling function and thus, by refinement, the dual wavelet. The answer is given by the following theorem.

THEOREM 10.1. *Consider a multiresolution analysis with order N. After lifting, the first N moments of the dual scaling function and dual wavelet do not change.*

Proof. The primal scaling functions do not change after lifting. This means that

$$P_p \;=\; \sum_k \langle\, P_p, \widetilde{\varphi}^{\,\mathrm{old}}_{j,k} \,\rangle\, \varphi_{j,k} \;=\; \sum_k \langle\, P_p, \widetilde{\varphi}_{j,k} \,\rangle\, \varphi_{j,k} \quad \text{for} \quad 0 \le p < N\,.$$

This implies that the first N moments of the dual scaling functions do not change after lifting. Since the coefficients of the refinement relations of the dual wavelets do not change (8.3), neither do their moments. ∎

Thus lifting does not alter the number of vanishing moments of the dual wavelet obtained by prior lifting.

Suppose we use dual lifting to increase the number of dual vanishing moments from N^{old} to N. This involves solving a linear system of size N, independent of how many vanishing moments the old dual wavelets already had. This means that if we use a cakewalk construction the linear systems to be solved become larger and larger, and so do the S filters. Therefore we present a scheme which allows us to exploit the fact that the dual wavelets already have N^{old} moments and thus only solve a system of size $N - N^{\mathrm{old}}$. The basic idea is to lift an old dual wavelet ($\widetilde{\psi}^{\,\mathrm{old}}_{j,m}$) not with old dual *scaling functions* on the same level ($\widetilde{\varphi}^{\,\mathrm{old}}_{j,k}$) but with old dual *wavelets* on the coarser level ($\widetilde{\psi}^{\,\mathrm{old}}_{j-1,n}$). This leads to a new dual wavelet of the form

$$\widetilde{\psi}_{j,m} \;=\; \widetilde{\psi}^{\,\mathrm{old}}_{j,m} - \sum_{n \in \mathcal{M}(j-1)} \widetilde{t}_{j,n,m}\, \widetilde{\psi}^{\,\mathrm{old}}_{j-1,n}\,.$$

Here the $\widetilde{t}_{j,n,m}$ are the coefficients of a filter operator $\widetilde{T}_j : \ell^2(\mathcal{M}(j)) \to \ell_2(\mathcal{M}(j-1))$. We always assume that the index n belongs to $\mathcal{M}(j-1)$. Note that the new dual wavelets independent of $\widetilde{T}$ immediately have at least as many vanishing moments as the old ones (N^{old}). Expressing that the new dual wavelets have N vanishing moments leads to only $N - N^{\mathrm{old}}$ equations in the unknowns $\{\widetilde{t}_{j,n,m} \mid n\}$.

Let us try to find a fast wavelet transform associated with this. In operator notation we have

$$\widetilde{\psi}_j \;=\; \widetilde{\psi}^{\,\mathrm{old}}_j - \widetilde{T}^*_j\, \widetilde{\psi}^{\,\mathrm{old}}_{j-1} \;=\; \widetilde{\psi}^{\,\mathrm{old}}_j - \widetilde{T}^*_j\, \widetilde{G}^{\,\mathrm{old}}_{j-1}\, \widetilde{\varphi}^{\,\mathrm{old}}_j\,.$$

This construction thus corresponds to letting $\widetilde{S}^*_j = \widetilde{T}^*_j\, \widetilde{G}^{\,\mathrm{old}}_{j-1}$. The basic idea is to use only the old filters and the filter $\widetilde{T}$ and never construct the $\widetilde{S}$ filter or any of the new filters explicitly. The forward transform takes three stages:

- I: Given the sequence λ_{j+1} calculate the forward transform with the old filters: $\lambda_j := \widetilde{H}^{\,\mathrm{old}}_j\, \lambda_{j+1}$ and $\gamma_j := \widetilde{G}^{\,\mathrm{old}}_j\, \lambda_{j+1}$.
- II: Calculate another level with the old filters: $\lambda_{j-1} := \widetilde{H}^{\,\mathrm{old}}_{j-1}\, \lambda_j$ and $\gamma_{j-1} := \widetilde{G}^{\,\mathrm{old}}_{j-1}\, \lambda_j$.
- III: Lift the γ_j with the γ_{j-1}: $\gamma_j - = \widetilde{T}^*_j\, \gamma_{j-1}$.

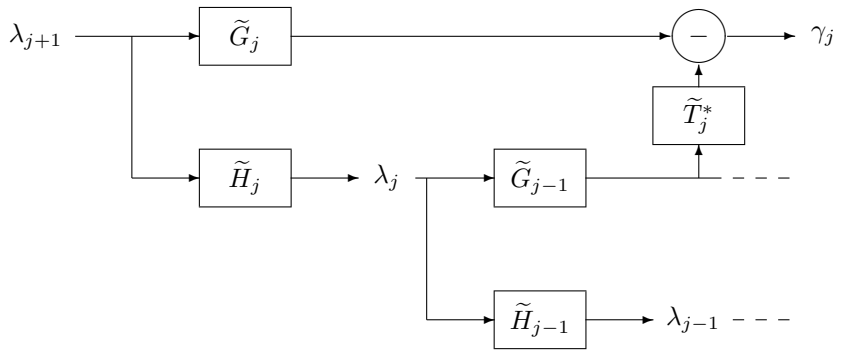

FIG. 10.1. *Part of a cakewalk construction. The basic idea is to lift the wavelet coefficients with wavelet coefficients on the coarser level. This way the fact that the old dual wavelets already have N^{old} vanishing moments can be exploited.*

Note that the second stage on level j coincides with the first stage on level $j-1$; see Figure 10.1 for a block diagram. The inverse transform in a first stage undoes the lifting and then applies an inverse transform with the old filters.

We have seen how the lifting scheme can pass between an old and a new multiresolution analysis. To start the construction of second generation wavelets we therefore need an initial multiresolution analysis. In the following sections we will give three examples of an initial multiresolution analysis to start the lifting scheme.

11. Orthogonal Haar wavelets. In this section we present the generalized orthogonal Haar wavelets, which form a first example of an initial multiresolution analysis to start the lifting scheme. The idea was first introduced by Coifman, Jones, and Semmes for dyadic cubes in [33], generalized for Clifford-valued measures in [9, 91], and later generalized for arbitrary partitionings in [68].

We first introduce the notion of a *nested set of partitionings*.

DEFINITION 11.1. *A set of measurable subsets $\{X_{j,k} \mid j,k\}$ is a* nested set of partitionings *if it is a set of partitionings and if, for every j and k, $X_{j,k}$ can be written as a finite disjoint union of at least two sets $X_{j+1,l}$:*

$$X_{j,k} = \bigcup_{l \in \mathcal{L}(j,k)} X_{j+1,l}\,.$$

Note that because of the partition property ($X_{j,k} \subset X_{j+1,k}$) we have that $k \in \mathcal{L}(j,k)$. Let $\varphi_{j,k} = \chi_{X_{j,k}}$ and $\widetilde{\varphi}_{j,k} = \chi_{X_{j,k}}/\mu(X_{j,k})$ according to our normalization (3.2). Define the $V_j \subset \mathrm{L}_2$ as

$$V_j \;=\; \operatorname{clos}\operatorname{span}\{\varphi_{j,k} \mid k \in \mathcal{K}(j)\}\,.$$

The spaces V_j generate a multiresolution analysis of L_2; see, e.g., [68] for a proof. As the scaling functions are orthogonal, we let W_j be the orthogonal complement of V_j in V_{j+1} so that $\widetilde{V}_j = V_j$.

Now fix a scaling function $\varphi_{j,k}$. For the construction of the Haar wavelets, we only need to consider the set $X_{j,k}$. First we assume without loss of generality that $\mathcal{L}(j,k)$ contains either two or three elements. Indeed, if $\mathcal{L}(j,k)$ contains more elements, we can split them into two groups whose numbers of elements differ by at most one. For each group we can introduce (implicitly) a new corresponding $X_{j',k'}$. We can continue

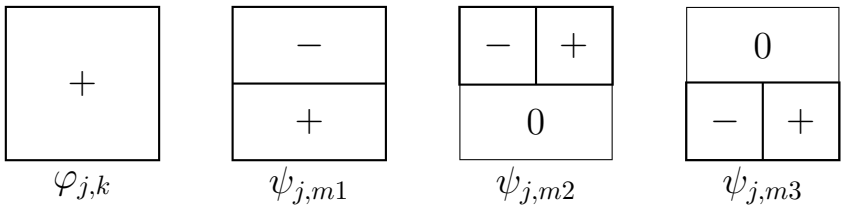

FIG. 11.1. *The generalized orthogonal Haar wavelets for square partitionings. The wavelets are piecewise constant and have a vanishing integral. The sign is indicated in the support. The orthogonality follows immediately from the support and the vanishing integral of the wavelets. Similar constructions apply to arbitrary partitionings.*

to do this until the number of elements is either 2 or 3. In case $\mathcal{L}(j,k) = \{k, m\}$ we let the wavelet $\psi_{j,m}$ be

$$\psi_{j,m} \;=\; \frac{\varphi_{j+1,k}}{2\,\mu(X_{j+1,k})} - \frac{\varphi_{j+1,m}}{2\,\mu(X_{j+1,m})}\,. \tag{11.1}$$

In case $\mathcal{L}(j,k) = \{k, m, m'\}$ we keep $\psi_{j,m}$ as above and let

$$\psi_{j,m'} \;=\; \frac{\varphi_{j+1,k} + \varphi_{j+1,m}}{2\,\mu(X_{j+1,k}) + 2\,\mu(X_{j+1,m})} - \frac{\varphi_{j+1,m'}}{2\,\mu(X_{j+1,m'})}\,.$$

In case $\mathcal{L}(j,k) = \{k, m1, m2, m3\}$, we need two stages. Each stage involves two sets and a wavelet of the form (11.1); see Figure 11.1. The Haar wavelets are constructed so that

$$\int_X \psi_{j,m}\,d\mu \;=\; 0 \quad \text{and} \quad \int_X |\psi_{j,m}|\,d\mu \;=\; 1\,.$$

They are orthogonal to $\varphi_{j,k}$ because they have a vanishing integral. Two different wavelets are orthogonal, since either their supports are disjoint or one is constant on the support of the other.

These wavelets form an orthogonal basis for L_2. In fact, they also form an unconditional basis for L_p.

THEOREM 11.2 (see [68]). *The generalized orthogonal Haar wavelets* $\{\psi_{j,m} \mid j, m\}$ *form an unconditional basis for* L_p *with* $1 < p < \infty$*, with unconditional basis constant* $p^* - 1$*, where* $1/p + 1/p^* = 1$.

This construction allows for Haar wavelets adapted to the settings G1–G3 mentioned in the introduction. Their advantage is their generality. Their disadvantages are that they are nonsmooth and that they have only one vanishing moment. However, they form a perfect example of an initial multiresolution analysis to start the lifting scheme with. With the lifting scheme we can build wavelets with more vanishing moments and/or more smoothness.

In the case of the real line and the classical Haar wavelet, the dual lifting scheme corresponds to a technique called *average interpolation* introduced by Donoho in [56]. Here $\widetilde{\varphi}$ is the indicator function on $[0,1]$, while φ is constructed through a subdivision scheme which ensures that polynomials up to a certain order can be reproduced with the scaling functions. This condition is precisely the same as the vanishing moment condition of the dual wavelet as used in dual lifting. The average interpolating technique can be generalized to a second generation setting; see, e.g., [104] for the construction of weighted wavelets. It generates primal and dual scaling functions which

are biorthogonal. However, it is not immediately clear what the wavelets and dual wavelets are. In other words there is no immediate generalization for the quadrature mirror filter construction where one takes $\widetilde{g}_k = (-1)^k h_{1-k}$. The dual lifting scheme provides a very simple solution to this problem. Again, the idea is to *first* construct the dual wavelets and later check what happens to the scaling functions using the cascade algorithm.

12. Interpolating scaling functions and wavelets. In this section we introduce the Lazy wavelet, another candidate to start the lifting scheme with, which is even simpler than the Haar wavelets. We show how it is connected with interpolating scaling functions.

12.1. The Lazy wavelet. One way to look at the general index sets $\mathcal{K}(j)$ and $\mathcal{M}(j)$ is to think of $\mathcal{K}(j)$ (respectively, $\mathcal{M}(j)$) as the generalization of the even (respectively, odd) indices. This inspires us to define two *subsampling operators* E (**e**ven) and D (o**d**d) as follows:

$E : \ell^2(\mathcal{K}(j+1)) \to \ell^2(\mathcal{K}(j))$, where $b = E\,a$ means $b_k = a_k$ for $k \in \mathcal{K}(j)$.

$D : \ell^2(\mathcal{K}(j+1)) \to \ell^2(\mathcal{M}(j))$, where $c = D\,a$ means $c_m = a_m$ for $m \in \mathcal{M}(j)$.

Although these operators depend on the level j we will not supply them with an extra subscript, since no confusion is possible. These operators provide a trivial orthogonal splitting, as

$$\begin{bmatrix} E \\ D \end{bmatrix} \begin{bmatrix} E^* & D^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} E^* & D^* \end{bmatrix} \begin{bmatrix} E \\ D \end{bmatrix} = 1\,.$$

We can now decompose any operator $W : \ell^2(\mathcal{K}(j)) \to \ell^2(\mathcal{K}(j))$ as

$$W = W_e\,E + W_d\,D\,, \quad \text{with} \quad W_e = W\,E^* \quad \text{and} \quad W_d = W\,D^*\,. \tag{12.1}$$

The filter operators of the *Lazy wavelet* are precisely these subsampling operators

$$H_j^{\text{Lazy}} = \widetilde{H}_j^{\text{Lazy}} = E \quad \text{and} \quad G_j^{\text{Lazy}} = \widetilde{G}_j^{\text{Lazy}} = D\,.$$

The Lazy wavelet transform thus is an orthogonal transform that essentially does nothing. It only resamples the coefficients into two groups for each step and thus can be seen as the generalization of the polyphase transform to the second generation setting. However, it is important to consider since it is connected with interpolating scaling functions. The operators E and D are crucial when implementing the lifting scheme. Although they are mathematically trivial, the data structure in the program has to be designed carefully to make them easy to implement. With such a data structure, the implementation of the lifting scheme is straightforward.

Given a set of partitionings, one can formally associate scaling functions and dual scaling functions with the Lazy wavelet. By using the cascade algorithm pointwise and respecting the normalization, one can see that $\widetilde{\varphi}_{j,k} = \delta(\cdot - x_k)$ and that $\varphi_{j,k}$ is zero everywhere except at x_k where it is one. Formally they are biorthogonal, but in the L_2 setting, $\widetilde{\varphi}_{j,k}$ does not belong to the space while $\varphi_{j,k}$ is zero. The wavelets and dual wavelets are given by $\psi_{j,m} = \varphi_{j+1,m}$ and $\widetilde{\psi}_{j,m} = \varphi_{j+1,m}$, and $N = \widetilde{N} = 0$.

12.2. Interpolating scaling functions. Next, we generalize the notion of an interpolating scaling function. We first need a *set of interpolation points* $\{x_k \mid j \in \mathcal{J}, k \in \mathcal{K}(j)\}$. Remember that such a set can be defined by a set of partitionings. In

the other direction, we can associate a set of partitionings with a set of interpolating points as follows. Assume that

$$\forall k \; : \; \inf_{j \in \mathcal{J},\, k' \in \mathcal{K}(j)} \; d(x_k, x_{k'}) = 0$$

for all k. Then let

$$S_{j,k} = \{x \in X \mid d(x, x_k) < d(x, x_{k'}) \text{ for } k' \in \mathcal{K}(j), k \neq k'\}\,.$$

The sets $S_{j,k}$ are the *Voronoi cells* of the set of points $\{x_k \mid k \in \mathcal{K}(j)\}$.

DEFINITION 12.1. *A set of scaling functions* $\{\varphi_{j,k} \mid j,k\}$ *is* interpolating *if a set of interpolation points* x_k *exists so that* $\varphi_{j,k}(x_{k'}) = \delta_{k,k'}$ *for* $k, k' \in \mathcal{K}(j)$.

As in the first generation case, the interpolating property can be characterized by means of the coefficients of the refinement relation. We state and prove the result in the index notation.

LEMMA 12.2. *If a set of second generation scaling functions is interpolating, then*

$$\forall\, k, k' \in \mathcal{K}(j) : h_{j,k,k'} \;=\; \delta_{k,k'}\,. \tag{12.2}$$

Proof.

$$\delta_{k,k'} = \varphi_{j,k}(x_{k'}) = \sum_m h_{j,k,l}\, \varphi_{j+1,l}(x_{k'}) = \sum_m h_{j,k,l}\, \delta_{l,k'} \;=\; h_{j,k,k'}\,. \qquad \square$$

Note that this lemma can be seen as a special case of Remark 4.2 in [41]. A filter h is called an *interpolating filter* if condition (12.2) holds. This condition can be written in operator notation as

$$H_j^{\text{int}}\, E^* \;=\; 1\,.$$

Note that this is the generalization of an *à trous* filter in the first generation case.

If we have an interpolating scaling function, we can always take Dirac functions as a formal dual

$$\widetilde{\varphi}_{j,k}^{\text{int}} \;=\; \delta(\cdot - x_k)\,.$$

The biorthogonality follows immediately from the interpolation property. The filter corresponding to the dual scaling function is

$$\widetilde{H}^{\text{int}} \;=\; E\,.$$

Now define $\widetilde{S}_j$ as $H_j^{\text{int}}\, D^*$. Then it follows from (12.1) that any interpolating filter can be written as $H_j^{\text{int}} = E + \widetilde{S}_j\, D$. But this expression can be seen as the result of applying the dual lifting scheme to the Lazy wavelet. We can then write a set of biorthogonal filters as

$$\begin{aligned}
H_j^{\text{int}} &= E + \widetilde{S}_j\, D\,,\\
\widetilde{H}_j^{\text{int}} &= E\,,\\
G_j^{\text{int}} &= D\,,\\
\widetilde{G}_j^{\text{int}} &= D - \widetilde{S}_j^*\, E\,.
\end{aligned}$$

We have thus shown the following theorem.

THEOREM 12.3. *The set of filters resulting from interpolating scaling functions, and Diracs as their formal dual, can be seen as a dual lifting of the Lazy wavelet.*

In index notation the filters become

$$\begin{aligned}
h^{\text{int}}_{j,k,l} &= \begin{cases} \delta_{k,l} & \text{if } l \in \mathcal{K}(j)\,, \\ \widetilde{s}_{j,k,l} & \text{if } l \in \mathcal{M}(j)\,, \end{cases} \\
\widetilde{h}^{\text{int}}_{j,k,l} &= \delta_{k,l}\,, \\
g^{\text{int}}_{j,m,l} &= \delta_{m,l}\,, \\
\widetilde{g}^{\text{int}}_{j,m,l} &= \begin{cases} -\widetilde{s}_{j,l,m} & \text{if } l \in \mathcal{K}(j)\,, \\ \delta_{m,l} & \text{if } l \in \mathcal{M}(j)\,. \end{cases}
\end{aligned}$$

Formally the dual wavelets are given by

$$\widetilde{\psi}_{j,m} \;=\; \delta(\cdot - x_m) - \sum_k h_{j,k,m}\delta(\cdot - x_k)\,.$$

The primal wavelets are $\psi_{j,m} \;=\; \varphi_{j+1,m}$. We have $\widetilde{N} = 0$ and N possibly > 0. These filters do not correspond to a multiresolution analysis of L_2, as the dual functions are Dirac distributions which do not even belong to L_2. In the case of linear interpolation, this example corresponds to what is known in finite elements as "hierarchical basis functions" [116].

We next apply the lifting scheme to find wavelets which have $\widetilde{N} > 0$. This leads to new filters of the form

$$\begin{aligned}
H_j = H^{\text{int}}_j \;&=\; E + \widetilde{S}_j\, D\,, \\
\widetilde{H}_j = \widetilde{H}^{\text{int}}_j + S_j\, \widetilde{G}^{\text{int}}_j \;&=\; (1 - S_j\, \widetilde{S}^*_j)\, E + S_j\, D\,, \\
G_j = G^{\text{int}}_j - S^*_j\, H^{\text{int}}_j \;&=\; -S^*_j\, E + (1 - S^*_j\, \widetilde{S}_j)\, D\,, \\
\widetilde{G}_j = \widetilde{G}^{\text{int}}_j \;&=\; -\widetilde{S}^*_j\, E + D\,.
\end{aligned}$$

This can be verified using Figure 12.1. For example, to find $\widetilde{H}_j$, follow the paths from λ_{j+1} to λ_j. There are three: one direct through E, one through D and then down through S_j, and one through E then up through $\widetilde{S}^*_j$ and down through S_j. Consequently $\widetilde{H}_j = (1 - S_j\, \widetilde{S}^*_j)\, E + S_j\, D$. In index notation this becomes

$$\widetilde{h}_{j,k,l} = \delta_{k,l} + \sum_m s_{j,k,m}\, \widetilde{g}_{j,m,l},$$

$$g_{j,m,l} = \delta_{m,l} - \sum_k s_{j,k,m}\, h_{j,k,l}\,.$$

The new wavelet can be written as

$$\psi_{j,m} \;=\; \varphi_{j+1,m} - \sum_{k \in \mathcal{K}(j,m)} s_{j,m,k}\, \varphi_{j,k}. \tag{12.3}$$

One can find the $s_{j,k,m}$ in the same way as described above.

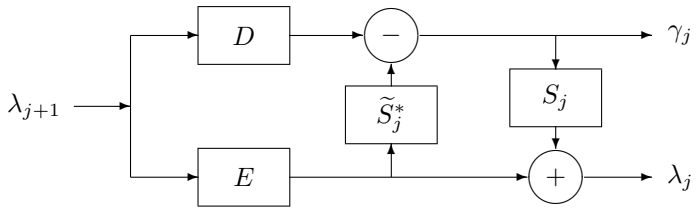

FIG. 12.1. *The fast wavelet transform for wavelets built from interpolating scaling functions. First apply a Lazy wavelet transform, then a dual lifting, and finally a regular lifting.*

12.3. Algorithm. The algorithm for the wavelet transform associated with the wavelets constructed in the previous section consists of three stages. First a Lazy wavelet transform, then a dual lifting, and finally a primal lifting; see Figure 12.1. The inverse transform can be derived immediately by simply inverting each step of the forward transform.

```
Forward(j):
```

$$\begin{aligned}
\forall k \in \mathcal{K}(j): \quad & \lambda_{j,k} \;:=\; \lambda_{j+1,k} \\
\forall m \in \mathcal{M}(j): \quad & \gamma_{j,m} \;:=\; \lambda_{j+1,m} \\
\forall m \in \mathcal{M}(j): \quad & \gamma_{j,m} \;-\!=\; \sum_{k \in \widetilde{\mathcal{K}}(j,m)} \widetilde{s}_{j,k,m}\, \lambda_{j,k} \\
\forall k \in \mathcal{K}(j): \quad & \lambda_{j,k} \;+\!=\; \sum_{m \in \mathcal{M}(j,k)} s_{j,k,m}\, \gamma_{j,m}
\end{aligned}$$

```
Inverse(j):
```

$$\begin{aligned}
\forall k \in \mathcal{K}(j): \quad & \lambda_{j,k} \;-\!=\; \sum_{m \in \mathcal{M}(j,k)} s_{j,k,m}\, \gamma_{j,m} \\
\forall m \in \mathcal{M}(j): \quad & \gamma_{j,m} \;+\!=\; \sum_{k \in \widetilde{\mathcal{K}}(j,m)} \widetilde{s}_{j,k,m}\, \lambda_{j,k} \\
\forall m \in \mathcal{M}(j): \quad & \lambda_{j+1,m} \;:=\; \gamma_{j,m} \\
\forall k \in \mathcal{K}(j): \quad & \lambda_{j+1,k} \;:=\; \lambda_{j,k}
\end{aligned}$$

One of the nice properties of the fast lifted wavelet transform is that all calculations can be done in place, i.e., without auxiliary memory. It is sufficient to provide storage locations only for the coefficients $\lambda_{n,k}$ of the finest levels. No additional auxiliary memory is needed. A coefficient $\lambda_{j,k}$ with $j < n$ can be stored in the same location as $\lambda_{n,k}$, while a wavelet coefficient $\gamma_{j,m}$ with $j < n$ can be stored in the same location as $\lambda_{n,m}$. The Lazy wavelet transform now simply requires blinking your eyes. Lifting will only require updates with local neighboring coefficients (typically $+=$ or $-=$ operators in the implementation) and thus does not need extra storage.

13. Biorthogonal Haar wavelets. In this section we introduce a third example of an initial multiresolution analysis: the biorthogonal Haar wavelets. They were first used in triangular subdivision in [99]. On triangles, biorthogonal Haar wavelets have more symmetry than orthogonal Haar wavelets. We here show how the biorthogonal Haar wavelets themselves can be seen as a result of lifting from the Lazy wavelet.

Take a set of nested partitionings $X_{j,k}$. Note that this defines the index sets $\mathcal{L}(j,k)$. Consider the Lazy scaling function and wavelet

$$\widetilde{\varphi}_{j,k}^{\text{Lazy}} \;=\; \delta(\cdot - x_k) \quad \text{and} \quad \widetilde{\psi}_{j,m}^{\text{Lazy}} \;=\; \delta(\cdot - x_m)\,.$$

Let us first apply dual lifting and denote the resulting functions with a superscript (1). Fix a $k^* \in \mathcal{K}(j)$ and let $\mathcal{M}(j,k^*) = \mathcal{L}(j,k^*)\backslash\{k^*\}$. In order for the new wavelet $\widetilde{\psi}_{j,m}^{(1)}$ with $m \in \mathcal{M}(j,k^*)$ to have one vanishing moment, we let

$$\widetilde{\psi}_{j,m}^{(1)} \;=\; \delta(\cdot - x_m) - \delta(\cdot - x_{k^*}) \;=\; \widetilde{\varphi}_{j+1,m}^{(1)} - \widetilde{\varphi}_{j+1,k^*}^{(1)}$$

so that $\mathcal{K}(j,m) = \{k^*\}$ and $\widetilde{s}_{j,k,m} = \delta_{k,k^*}$. This implies that the scaling function satisfies

$$\varphi_{j,k^*}^{(1)} = \varphi_{j+1,k^*}^{(1)} + \sum_{m\in\mathcal{M}(j,k^*)} \widetilde{s}_{j,k,m}\,\varphi_{j+1,m}^{(1)} = \sum_{l\in\mathcal{L}(j,k^*)} \varphi_{j+1,l}^{(1)} \tag{13.1}$$

which yields that $\varphi_{j,k}^{(1)} = \chi_{X_{j,k}}$ and thus $\psi_{j,m}^{(1)} = \chi_{X_{j+1,m}}$. We now have $N = 1$ and $\widetilde{N} = 0$ and could call this a *half-Haar basis*. Note that this half-Haar wavelet is used in the interlaced GIF format which is currently quite popular on the World Wide Web.

Next we use lifting to obtain a primal wavelet with a vanishing moment. We choose

$$\psi_{j,m} = \psi_{j,m}^{(1)} - s_{j,k^*,m}\,\varphi_{j,k*}^{(1)} \quad \text{with} \quad k^* \in \mathcal{K}(j,m)\,,$$

where $s_{j,k,m} = \mu(X_{j+1,m})/\mu(X_{j,k})$ if $m \in \mathcal{M}(j,k)$ and zero otherwise. In this way $\psi_{j,m}$ has one vanishing moment. The new dual scaling function becomes

$$\begin{aligned}
\widetilde{\varphi}_{j,k} &= \widetilde{\varphi}_{j+1,k} + \sum_{m\in\mathcal{M}(j,k)} \mu(X_{j+1,m})/\mu(X_{j,k})\,\widetilde{\psi}_{j,m} \\
&= \widetilde{\varphi}_{j+1,k} + \sum_{m\in\mathcal{M}(j,k)} \mu(X_{j+1,m})/\mu(X_{j,k})\,(\widetilde{\varphi}_{j+1,m} - \widetilde{\varphi}_{j+1,k}) \quad \text{(see (13.1))} \\
&= \sum_{m\in\mathcal{M}(j,k)} \mu(X_{j+1,m})/\mu(X_{j,k})\,\widetilde{\varphi}_{j+1,m} \\
&\qquad + \left(1 - \sum_{m\in\mathcal{M}(j,k)} \mu(X_{j+1,m})/\mu(X_{j,k})\right)\,\widetilde{\varphi}_{j+1,k} \\
&= \sum_{l\in\mathcal{L}(j,k)} \mu(X_{j+1,l})/\mu(X_{j,k})\,\widetilde{\varphi}_{j+1,l} \\
&= \chi_{X_{j,k}}/\mu(X_{j,k})\,.
\end{aligned}$$

Summarizing, we have the following basis functions, which generate the *biorthogonal Haar* multiresolution analysis:

$$\begin{aligned}
&\varphi_{j,k} = \chi_{X_{j,k}}\,, \\
&\widetilde{\varphi}_{j,k} = \chi_{X_{j,k}}/\mu(X_{j,k})\,, \\
&\psi_{j,m} = \varphi_{j+1,m} - \mu(X_{j+1,m})/\mu(X_{j,k^*})\,\varphi_{j,k^*} \quad \text{with} \quad \{k^*\} = \mathcal{K}(j,m)\,, \\
&\widetilde{\psi}_{j,m} = \widetilde{\varphi}_{j+1,m} - \widetilde{\varphi}_{j+1,k^*}\,.
\end{aligned}$$

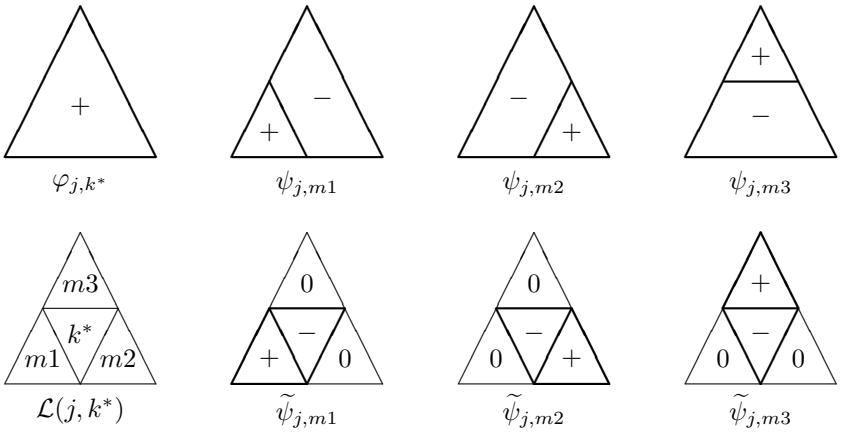

FIG. 13.1. *The biorthogonal Haar wavelets on triangles. Biorthogonality follows from the support and the vanishing integral of wavelets and dual wavelets. On triangles the biorthogonal Haar wavelets are more symmetric than the orthogonal Haar. This is another example to start the lifting scheme with.*

Figure 13.1 shows the biorthogonal Haar wavelets for a triangular partitioning. Given that the scaling function and dual scaling function are multiples of each other, we actually have a semi-orthogonal setting. This means that the V_j and $\widetilde{V}_j$ spaces coincide. Consequently wavelets on different levels are orthogonal, but within one level the wavelets are not orthogonal. The biorthogonal Haar multiresolution analysis is another example of an initial multiresolution analysis with $N = \widetilde{N} = 1$. Here we actually showed how it can be constructed by twice lifting the Lazy wavelet.

The algorithm for the biorthogonal Haar transform is given below. Again all calculations can be done in place.

`Forward(j):`

$$\forall m \in \mathcal{M}(j): \quad \gamma_{j,m} := \lambda_{j+1,m} - \lambda_{j+1,k} \qquad (\{k^*\} = \mathcal{K}(j,m))$$
$$\forall k \in \mathcal{K}(j): \quad \lambda_{j,k} := \lambda_{j+1,k} + \sum_{m \in \mathcal{M}(j,k)} s_{j,k,m}\, \gamma_{j,m}$$

`Inverse(j):`

$$\forall k \in \mathcal{K}(j): \quad \lambda_{j+1,k} := \lambda_{j,k} - \sum_{m \in \mathcal{M}(j,k)} s_{j,k,m}\, \gamma_{j,m}$$
$$\forall m \in \mathcal{M}(j): \quad \lambda_{j+1,m} := \gamma_{j,m} + \lambda_{j+1,k} \qquad (\{k^*\} = \mathcal{K}(j,m))$$

14. Applications and future research. Now that we understand the machinery of the lifting scheme, we can start to apply it in the settings described in the introduction. We discuss a few cases in more detail.

14.1. Wavelets on an interval. As pointed out in the introduction, many wavelet constructions on the interval already exist. They all involve modifying the wavelets and scaling functions close to the end point of the interval, which leads to special boundary filters. The derivation of the boundary filters is actually quite technical and it is not immediately clear to the user why they work. With the aid of the lifting scheme, the construction of interval wavelets and the implementation of

the associated transform become much more transparent. The Haar and Lazy wavelet can be trivially defined on the interval. Lifting then only requires pulling in the right aunts (scaling functions on the coarser level) at the boundary of the interval. All calculations can be done in place. For details we refer to [106].

A software package, `LIFTPACK`, to calculate the wavelet transformation of images is currently available [64]. Its properties are in-place calculation, correct treatment of boundaries, arbitrary size images (not only powers of two), and a faster implementation of existing biorthogonal wavelet filters (speedup can be a factor of two).

14.2. Weighted wavelets. Let X be $\mathbf{R}$ and consider the weight function $w(x) = d\mu/dx$, where dx stands for the Lebesgue measure. The wavelets constructed with the lifting scheme are orthogonal with respect to a weighted inner product, where $w(x)$ is the weight function. We refer to them as weighted wavelets. They are useful for the approximation of functions with singularities. If a function f contains a singularity, then the approximation with first generation wavelets will be slow, independent of the number of dual vanishing moments N. If we can now choose a weight function w so that $w \cdot f$ is a smooth function, then the approximation with weighted wavelets will again be of the order of the number of vanishing moments. An example of this behavior is given in [106].

Weighted wavelets are also useful in the solution of boundary value ODEs; see [75, 104]. If the operator is of the form $-DaD$, then operator wavelets defined as the antiderivative of weighted wavelets with weight function $w(x) = \sqrt{a(x)}$ diagonalize the operator. The solution algorithm is thus simply a forward and inverse wavelet transform. Future research involves the incorporation of the operator wavelets construction directly into the lifting scheme.

14.3. Wavelets on curves, surfaces, and manifolds. The only thing needed to construct wavelets on manifolds is either a set of interpolating points to define a Lazy wavelet or a set of nested partitionings to define Haar wavelets. Lifting will take care of the rest. The resulting wavelets are defined intrinsically on the manifold and do not depend on any parameterization or atlas.

In [99] the lifting scheme is used to construct wavelets on a sphere. Partitionings of the sphere were obtained by starting from a Platonic solid and alternating triangular subdivision and projection out to the sphere. This is known as a geodesic sphere construction. The Lazy wavelet is the starting point for a family of vertex-based wavelets, while the biorthogonal Haar wavelets lead to a family of face-based wavelets. In [100] these wavelets were used for the processing of spherical images. Current research involves the generalization of the construction and the applications to arbitrary surfaces.

14.4. Adaptive wavelets. The idea of adaptive wavelets was introduced in [69, 97, 98] in the context of the numerical solution of integral equations for illumination computations. The idea is the following. Assume the solution can be approximated with sufficient accuracy in a linear space V_n of dimension M. We know that out of the M^2 matrix entries representing the integral operator in the wavelet basis, only a fraction $\mathcal{O}(M)$ is relevant. If we have these entries, solving the matrix equation can be done in linear ($\mathcal{O}(M)$) time.

However, calculating all wavelet coefficients of the kernel from the finest level n to the coarsest level 0 with the fast wavelet transform requires $\mathcal{O}(M^2)$ operations and is thus a waste of CPU time and memory. Indeed the majority of all computations and memory use will be in vain. If we want an algorithm with linear complexity we can

only afford to calculate the wavelet coefficients which we actually need or a slightly larger set.

Gortler, Schröder, et al. [69] achieve this with the use of an *oracle* function. This function predicts, in a conservative fashion and based upon knowledge of the kernel of the integral equation, which wavelet coefficients need to be calculated. They were able to implement this with the use of what they call *tree wavelets*. Tree wavelets have the property that each wavelet of level j is supported within the support of only one scaling function of level j. Haar wavelets and Alpert wavelets [4, 5, 6] have this property. The advantage is that subdividing the support of a scaling function on level j, and thus constructing the wavelets of level j associated with it, does not imply subdividing any other support sets on level j. This way they can calculate the wavelet coefficients from the *coarsest* level to the *finest* level, thereby only subdividing (adding wavelets) where the oracle tells them to.

With traditional (nontree) wavelets, subdividing a support set S on level j and constructing the wavelets associated with it (whose support may reach outside of S) will imply subdividing a neighboring set of S and dragging in the wavelets associated with that set. This process cascades out and would imply subdividing the *whole* level j and thus makes adaptive constructions awkward.

Tree wavelets are discontinuous and this is a drawback in many applications. As shown in [99], lifting provides a solution here. Indeed, because of the "aunt" property, subdividing a set S on level j only requires its neighbors to exist; it does not necessarily require them to be subdivided as well. The mesh only needs to satisfy a restriction criterion in the sense that neighboring sets only differ by at most one level. This does not cascade out. Lifting thus opens the door to smooth adaptive wavelets. Current research involves the incorporation of these wavelets in illumination computations.

A word of caution is needed here. In [46, 96] it is shown that adaptive wavelet algorithms require wavelets on manifolds satisfying specific conditions concerning stability, regularity, and norm equivalence. As pointed out earlier, lifting does not guarantee these conditions and they have to be verified in each particular case.

14.5. Recursive wavelets. The principle of *recursive wavelets* is explained in [8, 56]. The basic idea is not to use the cascade algorithm ad infinitum to construct the scaling functions, but instead fix the scaling functions on an arbitrary finest level n. This can be generalized easily to the second generation case. Consider a set of partitionings and let

$$\varphi_{n,k} \;=\; \chi_{S_{n,k}} \quad \text{for} \quad k \in \mathcal{K}(n)\,.$$

Next consider a general filter h (not necessarily a Haar filter) and define the scaling functions on the coarser level ($\varphi_{j,k}$ with $j < n$) through recursive applications of the refinement relation (4.1). By definition all scaling functions $\varphi_{j,k}$ are piecewise constant on the sets $\{S_{n,k} \mid k \in \mathcal{K}(n)\}$. This is precisely the advantage of recursive wavelets; no need to go through an infinite limit process to find the scaling functions, instead apply the refinement relation a finite number of times.

One can choose other functions than indicator functions as scaling functions on the finest level. The advantage of indicator functions is their generality; the disadvantage is that they are not smooth. If the topology admits it, smoother choices are piecewise linear (hat) functions or B/box-splines.

In the setting of recursive wavelets, there are L_2 functions associated with the Lazy wavelet. Indeed

$$\psi^{\text{Lazy}}_{j,m} \;=\; \varphi^{\text{Lazy}}_{j+1,m} \;=\; \varphi^{\text{Lazy}}_{n,m} \;=\; \chi_{S_{n,m}}\,.$$

In this paper, we have always assumed that the measure is nonatomic. This restriction, however, is not fundamental. Recursive wavelets allow for atomic measures. Indeed, there is no reason why any of the subsets of $S_{n,k}$ should be measurable. As is shown in [68], it is possible to build Haar wavelets on fully discrete sets such as the integers, or on sets which are of mixed continuous/discrete nature. Also the lifting scheme remains valid.

The idea of recursive wavelets can also be combined with the idea of adaptive wavelets of the previous section. Instead of fixing the scaling function on one finest level n, one can let the notion of finest level depend on the location. Indeed, the oracle of the previous section typically leads to finer subdivisions in certain locations and coarser subdivisions in other. The finest level $n(k)$ thus depends on the location k. We then fix the scaling functions

$$\varphi_{n(k),k} \;=\; \chi_{S_{n(k),k}}\,,$$

where

$$X = \bigcup_k S_{n(k),k}\,.$$

In other words, we are using an adaptive mesh.

It is important to note that even though recursive wavelets only use a finite number of levels, the stability issue does not go away but rather manifests itself as a problem concerning ill-conditioning.

14.6. Wavelet packets. *Wavelet packets* were introduced by Coifman, Meyer, and Wickerhauser [34, 35, 114, 115]. The idea is to also further split the W_j spaces with the help of the h and g filters. This way one obtains a better frequency localization. The splitting leads to a full binary tree of wavelet packets, which form a redundant set. For a given function, one can choose the *best basis* with respect to a criterion such as the entropy of the basis coefficients. A fast tree algorithm to find the best basis was introduced in [36]; see also [115].

This idea again carries over into the second generation setting and can be combined with lifting. The conditions for exact reconstruction have exactly the same algebraic structure as in the wavelet case. One can start with defining a Lazy wavelet packet or a generalized Haar wavelet packet (which could be called generalized Walsh functions). A new wavelet packet is now defined as an old wavelet packet plus a linear combination of wavelet packets that live on a coarser level. From a practical point of view, one extra index comes in, and proper data structures need to be designed to incorporate the new filters.

14.7. M-band wavelets. The idea of *M-band wavelets*, or *p-adic wavelets*, a name more common in the mathematical literature, is to split a space V_{j+1} into M (as opposed to 2) subspaces: $V_j \oplus W_j^1 \oplus \cdots \oplus W_j^{M-1}$. For each subspace a different filter is used. Several constructions were introduced in [72, 81, 102, 109]. In some sense the second generation wavelet setting already incorporates this. Indeed, it even allows for different filters for each individual wavelet. However, thinking of lifting combined with M-band wavelets can lead to new constructions. Let us start with the Lazy wavelet. An M-band Lazy wavelet is easily defined and again is the standard polyphase transform [109]. Now one can define a new wavelet as an old wavelet plus a linear combination of scaling functions on the coarser level. This would be ordinary lifting. But we could also define a new wavelet as an old wavelet plus a

linear combination of scaling functions on a coarser level plus *wavelets belonging to another (lower index) subband*. This allows more flexibility in the construction. It requires that the M subband be calculated in ascending index order in the transform. In the extreme case one can let each subband contain precisely one wavelet. A new wavelet is now an old wavelet plus previously constructed new wavelets. This requires an ordering within the wavelets of one level. This way we can construct noncompactly supported wavelets with only finite filters.

14.8. Nonseparable wavelets in $\mathbf{R}^n$. As mentioned earlier, the lifting scheme also leads to new insights in the construction of first generation wavelets. This was shown in the one-dimensional case in [105]. Higher dimensional wavelets can always be constructed using tensor products, but this leads to severe axial directional dependencies. Instead one prefers to work with nonseparable wavelets which have more axial symmetry and which do not necessarily use a product lattice; see, for example, [22, 28, 32, 76, 77, 89, 90, 92, 93].

Here too lifting can help. Each lattice allows for the immediate definition of a Lazy wavelet or a Haar wavelet, either 2-band or M-band. Polynomial cancelation then leads to the filter coefficients. Whether lifting will actually lead to new wavelets in this context or rather provide faster implementations of already existing filters as in the one-dimensional case remains a topic for further study.

14.9. Wavelets on bounded domains and wavelet probing. One of the important application domains of wavelets is the solution of partial differential equations. In [44] it is shown that one can use wavelets to build multilevel preconditioners which result in stiffness matrices with uniformly bounded condition numbers. This leads to linear solution algorithms. To solve real life problems one needs wavelets constructed on nonsmooth (Lipschitz) domains in $\mathbf{R}^n$. In [27, 74] such a construction is presented. In both cases, tensor product wavelets are used in the interior of the domain while at the boundary specially adapted wavelets are constructed. In this sense these constructions are the natural generalization of the interval constructions to higher dimensions. With the lifting scheme one can build nonseparable wavelets adapted to general domains. Again the only thing needed is a set of partitionings. One nice property here is that lifting allows for adaptive meshes.

Already in the simple case of the Laplace equation and a nonsmooth domain it was recently shown that one cannot obtain an $\mathcal{O}(M^{-2})$ accuracy (where M is the number of elements) unless one uses nonlinear approximation [39]. The underlying reason is that the solution does not belong to the second-order Sobolev space but rather to a second-order Besov space. In other words, one has to use adaptive grids to obtain the correct convergence order.

There is another important application of wavelets on domains. It is a technique called *wavelet probing* introduced independently in [59] and [8, 53]. Let us discuss the idea first on the real line. Consider a function which is smooth except for jump discontinuities at isolated points. We know that the decay of the wavelet coefficients is fast away from the jumps, and slow in the neighborhood of the jumps. Increasing the number of dual vanishing moments leads to faster decay away from the jumps but also to a larger set of coefficients affected by the jumps because of the larger support. It is thus not possible to obtain more efficient approximations and thus better compression by simply increasing the number of dual vanishing moments.

Suppose now we know the location of the jumps. If we use interval wavelets on each interval between two jumps and thus segment the signal accordingly, we would get fast convergence *everywhere*. Wavelet probing is a technique which allows us to

locate the jumps. It simply tries every location between two samples and checks whether it would pay off to segment at this location. The payoff can be measured, e.g., with the entropy of the wavelet coefficients. Probing one location only requires altering $\log M$ coefficients where M is the number of samples. The whole algorithm thus has a complexity of $M \log M$.

Wavelet probing has important applications in image compression. Indeed it allows quick localization of the edges in an image, and then builds wavelets on the segments defined by those edges. Considerably higher compression ratios can be obtained this way. Wavelet probing thus provides an alternative for the zero crossing representation introduced by Mallat and Zhong [82, 83]. The advantage is that no iterative reconstruction is needed. Wavelet probing interacts easily with lifting and the adaptive wavelets mentioned above.

14.10. Wavelets adapted to irregular samples. As we mentioned in the introduction, one of the motivations for the generalization to second generation wavelets was the processing of irregularly sampled data. Let us discuss this first on the real line. Assume we are given irregular samples of a function $f(x)$: $\lambda_{n,k} = f(x_k)$ with $k \in \mathcal{K}(n)$. We first need to define a Lazy wavelet; i.e., we need the sets $\mathcal{K}(j)$ with $j < n$ (the coarser levels) and the sets $\mathcal{M}(j)$ with $j \geq n$, together with the x_m for $m \in \mathcal{M}(j)$ (the finer levels). For the coarser levels we only need to decide *which* sample locations to retain, while for the finer levels we also have to decide *where* to put the new locations. Coarser levels are needed in the wavelet transform, finer in the cascade algorithm. A simple strategy, which was used in [106], is to retain every other sample on the coarser level and put a new sample in the middle of two old samples on the finer levels. Once a Lazy wavelet is defined, dual lifting provides interpolating scaling functions, and lifting yields wavelets with vanishing moments. The dual lifting can be seen as an instance of irregular Deslauriers–Dubuc subdivision [54, 55]. Current research involves the study of more advanced choices for the Lazy wavelet. One of the strategies is to choose the sample locations so that the ratio of the largest versus the smallest interval of a level becomes closer and closer to one on both the finer ($j > n$) and the coarser ($j < n$) levels. Future research also involves the study of these schemes in higher dimensions.

14.11. Integer to integer wavelet transforms. In [16] lifting is used to build reversible wavelets which map integers to integers for applications to lossless image coding. The idea is to introduce a nonlinear round-off in each lifting step. This way the result is guaranteed to be integer while the lifting assures that the transform is invertible. The exact same idea works in the second generation setting, and using this in second generation compression applications is another line of future research.

14.12. Conclusion. In this paper we presented the lifting scheme, a construction tool for wavelets adapted to general settings. We showed how one can start from a trivial multiresolution analysis and use lifting to work one's way up to a multiresolution analysis with particular properties. As we mentioned in the introduction, the lifting scheme provides an answer to the algebraic phase of a wavelet construction. For each of the applications mentioned in this section, one still has to verify whether the cascade algorithm converges, whether the resulting wavelets form a Riesz basis (analytic phase), and what their smoothness is (geometric phase).

Note. As mentioned in the introduction, we learned after finishing this work that Dahmen and collaborators independently obtained a construction of multiscale bases with a technique very similar to lifting [17]. Here we go into more detail comparing

the two approaches. Recall from Theorem 8.1 that the lifting involved an operator which can be written in matrix location as

$$\begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix}.$$

The main advantage of this operator is that, independent of the choice of S, it is guaranteed to be invertible and the inverse can be found by flipping the sign on S. In the setting of [17] a more general operator of the form

$$\begin{bmatrix} 1 & S \\ 0 & K \end{bmatrix}$$

is used (which for $K = I$ becomes lifting). This operator is invertible if and only if K is invertible. The inverse is then given by

$$\begin{bmatrix} 1 & -SK^{-1} \\ 0 & K^{-1} \end{bmatrix}.$$

This setting is more general and allows us to explore *all* the degrees of freedom one has to generate new biorthogonal filters. In section 3.3 of [17] a $K \neq I$ is constructed which generates orthogonal decompositions. Section 4.1 of [17] discusses the case of compactly supported semi-orthogonal spline wavelets on irregular knot sequences. However, these settings do not allow both K and K^{-1} to be sparse.

There are certain advantages to the lifting ($K = I$) approach. One problem with the more general approach ($K \neq I$) is that it involves taking the inverse of K and does not guarantee that all primal and dual filters are finite. Infinitely supported filters are less useful practically and do not necessarily lead to fast transforms. Moreover, K^{-1} might be difficult to compute numerically. Many of the attractive features of lifting such as in-place computation, no need for inverting operators, and adaptive transforms using "aunt" functions disappear when allowing a general $K \neq I$.

In the first generation setting it was shown recently [52] that in case of finite filters no generality is lost when restricting oneself to the lifting setting: all finite filters can be obtained using multiple alternate primal and dual lifting steps.

On several occasions in this paper we mentioned that lifting does not guarantee stable bases or convergence of the associated subdivision scheme. In fact, many of these issues have been addressed carefully in [17, 41, 42] and we refer to those papers for details.

Acknowledgments. While this paper was being written (spring 1995), the author worked in close collaboration with Peter Schröder on the application of the lifting scheme to the construction of spherical wavelets [99, 100] and on a tutorial paper [106]. The feedback from the numerous and stimulating discussions with Peter was a tremendous help in writing this paper and has led to several new insights. Peter also helped by improving the exposition in this paper considerably. Also special thanks to Jelena Kovačević for pointing out the connection between the Lazy wavelet and the polyphase transform and to Thomas Barregren, Maurits Malfait, and Geert Uytterhoeven for carefully proofreading the text. Finally, the author would like to thank the two referees for their thorough reports and for pointing out oversights in the original submission. Their constructive comments have led to a much improved paper.

REFERENCES

[1] R. Abgrall and A. Harten, *Multiresolution representation in unstructured meshes*, SIAM J. Numer. Anal., submitted.

[2] A. Aldroubi, M. Eden, and M. Unser, *Discrete spline filters for multiresolutions and wavelets of ℓ_2*, SIAM J. Math. Anal., 25 (1994), pp. 1412–1432.

[3] A. Aldroubi and M. Unser, *Families of multiresolution and wavelet spaces with optimal properties*, Numer. Funct. Anal. Optim., 14 (1993), pp. 417–446.

[4] B. Alpert, *A class of bases in L_2 for the sparse representation of integral operators*, SIAM J. Math. Anal., 24 (1993), pp. 246–262.

[5] B. Alpert, G. Beylkin, R. Coifman, and V. Rokhlin, *Wavelet-like bases for the fast solution of second-kind integral equations*, SIAM J. Sci. Comput., 14 (1993), pp. 159–184.

[6] B. K. Alpert, *Wavelets and other bases for fast numerical linear algebra*, in [20], pp. 181–216.

[7] A. Amellaoui and P. G. Lemarié-Rieusset, *Ondelettes splines sur grilles irregulieres*, manuscript.

[8] L. Andersson, N. Hall, B. Jawerth, and G. Peters, *Wavelets on closed subsets of the real line*, in [101], pp. 1–61.

[9] L. Andersson, B. Jawerth, and M. Mitrea, *The Cauchy singular integral operator and Clifford wavelets*, in [14], pp. 525–546.

[10] P. Auscher, *Wavelets with boundary conditions on the interval*, in [20], pp. 217–236.

[11] P. Auscher and Ph. Tchamitchian, *Bases d'ondelettes sur les courbes corde-arc, noyau de Cauchy et espaces de Hardy associés*, Rev. Mat. Iberoamericana, 5 (1989), pp. 139–170.

[12] P. Auscher and Ph. Tchamitchian, *Ondelettes et conjecture de Kato*, C. R. Acad. Sci. Paris Sér. I Math. I, 313 (1991), pp. 63–66.

[13] G. Battle, *A block spin construction of ondelettes*, Comm. Math. Phys., 110 (1987), pp. 601–615.

[14] J. Benedetto and M. Frazier, eds., *Wavelets: Mathematics and Applications*, CRC Press, Boca Raton, FL, 1993.

[15] M. D. Buhmann and C. A. Micchelli, *Spline prewavelets for non-uniform knots*, Numer. Math., 61 (1992), pp. 455–474.

[16] R. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, *Wavelet transforms that map integers to integers*, Appl. Comput. Harm. Anal., to appear.

[17] J. M. Carnicer, W. Dahmen, and J. M. Peña, *Local decompositions of refinable spaces*, J. Appl. Comput. Harmonic Anal., 3 (1996), pp. 127–153.

[18] C. Chui and E. Quak, *Wavelets on a bounded interval*, in Numerical Methods of Approximation Theory, D. Braess and L. L. Schumaker, eds., Birkhäuser-Verlag, Basel, 1992, pp. 1–24.

[19] C. K. Chui, *An Introduction to Wavelets*, Academic Press, San Diego, CA, 1992.

[20] C. K. Chui, ed., *Wavelets: A Tutorial in Theory and Applications*. Academic Press, San Diego, CA, 1992.

[21] C. K. Chui, L. Montefusco, and L. Puccio, eds., *Conference on Wavelets: Theory, Algorithms, and Applications*. Academic Press, San Diego, CA, 1994.

[22] C. K. Chui, J. Stöckler, and J. D. Ward, *Compactly Supported Box-Spline Wavelets*, Technical report CAT 230, Center for Approximation Theory, Department of Mathematics, Texas A&M University, College Station, TX, 1991.

[23] C. K. Chui and J. Z. Wang, *A cardinal spline approach to wavelets*, Proc. Amer. Math. Soc., 113 (1991), pp. 785–793.

[24] C. K. Chui and J. Z. Wang, *A general framework of compactly supported splines and wavelets*, J. Approx. Theory, 71 (1992), pp. 263–304.

[25] C. K. Chui and J. Z. Wang, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–915.

[26] A. Cohen, *Ondelettes, analyses multiresolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 439–459.

[27] A. Cohen, W. Dahmen, and R. DeVore, *Multiscale decompositions on bounded domains*, Trans. Amer. Math. Soc., to appear.

[28] A. Cohen and I. Daubechies, *Non-separable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–137.

[29] A. Cohen, I. Daubechies, and J. Feauveau, *Bi-orthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[30] A. Cohen, I. Daubechies, B. Jawerth, and P. Vial, *Multiresolution analysis, wavelets and fast algorithms on an interval*, C. R. Acad. Sci. Paris Sér. I Math. I, 316 (1993), pp. 417–421.

[31] A. Cohen, I. Daubechies, and P. Vial, *Multiresolution analysis, wavelets and fast algorithms on an interval*, J. Appl. Comput. Harmonic Anal., 1 (1993), pp. 54–81.
[32] A. Cohen and J.-M. Schlenker, *Compactly supported bidimensional wavelet bases with hexagonal symmetry*, Constr. Approx., 9 (1993), pp. 209–236.
[33] R. R. Coifman, P. W. Jones, and S. Semmes, *Two elementary proofs of the* L_2 *boundedness of Cauchy integrals on Lipschitz curves*, J. Amer. Math. Soc., 2 (1989), pp. 553–564.
[34] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, *Signal processing and compression with wave packets*, in Proc. International Conference on Wavelets, Y. Meyer, ed., Marseille, 1989, Masson, Paris, 1992.
[35] R. R. Coifman, Y. Meyer, and V. Wickerhauser, *Size properties of wavelet packets*, in [95], pp. 453–470.
[36] R. R. Coifman and M. L. Wickerhauser, *Entropy based algorithms for best basis selection*, IEEE Trans. Inform. Theory, 38 (1992), pp. 713–718.
[37] D. Colella and C. Heil, *The characterization of continuous four-coefficient scaling functions and wavelets*, IEEE Trans. Inform. Theory, 38 (1992), pp. 876–881.
[38] D. Colella and C. Heil, *Characterizations of scaling functions: Continuous solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 496–518.
[39] S. Dahlke and R. DeVore, *Elliptic Boundary Value Problems and Besov Spaces*, Preprint, Dept. of Mathematics, University of South Carolina, Columbia, SC, 1996.
[40] W. Dahmen, *Decomposition of refinable spaces and applications to operator equations*, Numer. Algorithms, 5 (1993), pp. 229–245.
[41] W. Dahmen, *Some remarks on multiscale transformations, stability, and biorthogonality*, in Wavelets, Images, and Surface Fitting, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., A. K. Peters, Wellesley, MA, 1994, pp. 157–188.
[42] W. Dahmen, *Stability of Multiscale Transformations*, Technical report, Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 1994.
[43] W. Dahmen, B. Kleemann, S. Prössdorf, and R. Schneider, *Multiscale methods for the double layer potential equation on a polyhedron*, in Advances in Computational Mathematics, H. P. Dikshit and C. A. Micchelli, eds., World Scientific, Singapore, 1994, pp. 15–57.
[44] W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math., 63 (1992), pp. 315–344.
[45] W. Dahmen and C. A. Micchelli, *Banded matrices with banded inverses* II: *Locally finite decompositions of spline spaces*, Constr. Approx., 9 (1993), pp. 263–281.
[46] W. Dahmen, S. Prössdorf, and R. Schneider, *Multiscale methods for pseudo-differential equations on smooth manifolds*, in [21], 1994, pp. 385–424.
[47] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.
[48] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Series in Appl. Math. 61, SIAM, Philadelphia, PA, 1992.
[49] I. Daubechies, *Orthonormal bases of compactly supported wavelets* II: *Variations on a theme*, SIAM J. Math. Anal., 24 (1993), pp. 499–519.
[50] 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.
[51] I. Daubechies and J. C. Lagarias, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.
[52] I. Daubechies and W. Sweldens, *Factoring wavelet transforms into lifting steps*, J. Fourier Anal. Appl., to appear.
[53] B. Deng, B. Jawerth, G. Peters, and W. Sweldens, *Wavelet probing for compression based segmentation*, in Wavelet Applications in Signal and Image Processing, A. F. Laine, ed., Proc. SPIE 2034, International Society for Optical Engineering, 1993, pp. 266–276.
[54] G. Deslauriers and S. Dubuc, *Interpolation dyadique*, in Fractals, dimensions non entières et applications, Masson, Paris, 1987, pp. 44–55.
[55] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.
[56] D. L. Donoho, *Smooth wavelet decompositions with blocky coefficient kernels*, in [101], pp. 259–308.
[57] D. L. Donoho, *Interpolating Wavelet Transforms*, Department of Statistics, Preprint, Stanford University, Stanford, CA, 1992.
[58] D. L. Donoho, *Unconditional bases are optimal bases for data compression and for statistical estimation*, J. Appl. Comput. Harmonic Analysis, 1 (1993), pp. 100–115.
[59] D. L. Donoho, *On minimum entropy segmentation*, Preprint, Department of Statistics, Stanford University, Stanford, CA, 1994.

[60] G. C. Donovan, J. S. Geronimo, and D. P. Hardin, *A class of orthogonal multiresolution analyses in* 2*D*, in Mathematical Methods for Curves and Surfaces, Daehlen et al., eds., Vanderbilt Univ. Press, Nashville, TN, 1995, pp. 99–110.

[61] G. C. Donovan, J. S. Geronimo, and D. P. Hardin, *Intertwining multiresolution analyses and the construction of piecewise polynomial wavelets*, SIAM J. Math. Anal., 27 (1996), pp. 1791–1815.

[62] G. C. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1158–1192.

[63] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[64] G. Fernández, S. Periaswamy, and W. Sweldens, *LIFTPACK: A software package for wavelet transforms using lifting*, in Wavelet Applications in Signal and Image Processing IV, M. Unser, A. Aldroubi, and A. F. Laine, eds., Proc. SPIE 2825, International Society of Optical Engineering, 1996.

[65] M. Frazier and B. Jawerth, *Decomposition of Besov spaces*, Indiana Univ. Math. J., 34 (1995), pp. 777–799.

[66] M. Frazier and B. Jawerth, *The φ-Transform and Applications to Distribution Spaces*, in Function Spaces and Applications, M. Cwikel et al., eds., Lecture Notes in Math. 1302, Springer, New York, 1988, pp. 223–246.

[67] M. Frazier and B. Jawerth, *A discrete transform and decompositions of distribution spaces*, J. Funct. Anal., 93 (1990), pp. 34–170.

[68] M. Girardi and W. Sweldens, *A new class of unbalanced Haar wavelets that form an unconditional basis for* L_p *on general measure spaces*, J. Fourier Anal. Appl., 3 (1997), pp. 457–474.

[69] S. J. Gortler, P. Schröder, M. F. Cohen, and P. Hanrahan, *Wavelet radiosity*, in Computer Graphics (SIGGRAPH '93 Proceedings), ACM SIGGRAPH, New York, 1993, pp. 221–230.

[70] D. P. Hardin, B. Kessler, and P. R. Massopust, *Multiresolution analyses based on fractal functions*, J. Approx. Theory, 71 (1992), pp. 104–120.

[71] A. Harten, *Multiresolution representation of data: A general framework*, SIAM J. Numer. Anal., 33 (1996), pp. 1205–1256.

[72] P. N. Heller, *Rank M wavelets with N vanishing moments*, SIAM J. Matrix. Anal. Appl., 16 (1995), pp. 502–519.

[73] C. Herley and M. Vetterli, *Wavelets and recursive filter banks*, IEEE Trans. Signal Process, 41 (1993), pp. 2536–2556.

[74] B. Jawerth and G. Peters, *Wavelets on Non Smooth Sets of* $\mathbf{R}^n$, manuscript.

[75] B. Jawerth and W. Sweldens, *Wavelet multiresolution analyses adapted for the fast solution of boundary value ordinary differential equations*, in Sixth Copper Mountain Conference on Multigrid Methods, N. D. Melson, T. A. Manteuffel, and S. F. McCormick, eds., NASA Conference Publication 3224, 1993, pp. 259–273.

[76] R.-Q. Jia and C. A. Micchelli, *Using the refinement equations for the construction of pre-wavelets* II: *Powers of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, New York, 1991.

[77] J. Kovačević and M. Vetterli, *Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for* $\mathbf{R}^n$, IEEE Trans. Inform. Theory, 38 (1992), pp. 533–555.

[78] P.-G. Lemarié, *Ondelettes a localisation exponentielle*, J. Math. Pures Appl., 67 (1988), pp. 227–236.

[79] M. Lounsbery, *Multiresolution Analysis for Surfaces of Arbitrary Topological Type*, Ph.D. thesis, Department of Computer Science, University of Washington, Seattle, WA, 1994.

[80] M. Lounsbery, T. D. DeRose, and J. Warren, *Multiresolution surfaces of arbitrary topological type*, ACM Trans. Graphics, 16 (1997), pp. 34–73.

[81] M. Lundberg and G. Welland, *Construction of compact p-wavelets*, Constr. Approx., 9 (1993), pp. 347–370.

[82] S. Mallat and S. Zhong, *Wavelet transform maxima and multiscale edges*, in [95], pp. 67–104.

[83] S. Mallat and S. Zhong, *Characterization of signals from multiscale edges*, IEEE Trans. Patt. Anal. Mach. Intell., 14 (1992), pp. 710–732.

[84] S. G. Mallat, *Multifrequency channel decompositions of images and wavelet models*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 2091–2110.

[85] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of* $l^2(\mathbf{R})$, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[86] S. G. Mallat, *A theory for multiresolution signal decomposition: The wavelet representation*, IEEE Trans. Patt. Anal. Mach. Intell., 11 (1989), pp. 674–693.
[87] Y. Meyer, *Ondelettes et Opérateurs,* I: *Ondelettes,* II: *Opérateurs de Calderón-Zygmund,* III: (with R. Coifman)*, Opérateurs multilinéaires*, Hermann, Paris, 1990. English translation of Vol. 1, *Wavelets and Operators*, is published by Cambridge University Press, 1993.
[88] Y. Meyer, *Ondelettes sur l'intervalle*, Rev. Mat. Iberoamericana, 7 (1992), pp. 115–133.
[89] C. A. Micchelli, *Using the refinement equations for the construction of pre-wavelets*, Numer. Algorithms, 1 (1991), pp. 75–116.
[90] C. A. Micchelli, C. Rabut, and F. I. Utretas, *Using the refinement equations for the construction of pre-wavelets* III: *Elliptic splines*, Numer. Algorithms, 1 (1991), pp. 331–352.
[91] M. Mitrea, *Singular integrals, Hardy spaces and Clifford wavelets*, Lecture Notes in Math 1575, Springer, New York, 1994.
[92] S. D. Riemenschneider and Z. Shen, *Wavelets and pre-wavelets in low dimensions*, J. Approx. Theory, 71 (1992), pp. 18–38.
[93] S. D. Riemenschneider and Z. W. Shen, *Box splines, cardinal series and wavelets*, in Approximation Theory and Functional Analysis, C. K. Chui, ed., Academic Press, New York, 1991.
[94] O. Rioul, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.
[95] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., *Wavelets and Their Applications*, Jones and Bartlett, Boston, MA, 1992.
[96] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösing großer vollbesetzter Gleichungssysteme*, Habilitationsschrift, TH Darmstadt, 1995.
[97] P. Schröder, *Wavelet Algorithms for Illumination Computations*, Ph.D. thesis, Department of Computer Science, Princeton University, Princeton, NJ, 1994.
[98] P. Schröder, S. J. Gortler, M. F. Cohen, and P. Hanrahan, *Wavelet projections for radiosity*, Computer Graphics Forum, 13 (1994), pp 141–152.
[99] P. Schröder and W. Sweldens, *Spherical wavelets: Efficiently representing functions on the sphere*, Computer Graphics Proceedings (SIGGRAPH 95), ACM SIGGRAPH, New York, 1995, pp. 161–172.
[100] P. Schröder and W. Sweldens, *Spherical wavelets: Texture processing*, in Rendering Techniques '95, P. Hanrahan and W. Purgathofer, eds., Springer-Verlag, Wien, New York, 1995.
[101] L. L. Schumaker and G. Webb, eds., *Recent Advances in Wavelet Analysis*, Academic Press, New York, 1993.
[102] P. Steffen, P. Heller, R. A. Gopinath, and C. S. Burrus, *Theory of regular m-band wavelets*, IEEE Trans. Signal Process., 41 (1993), pp. 3497–3511.
[103] J. O. Strömberg, *A modified Franklin system and higher order spline systems on* $\mathbf{R}^n$ *as unconditional bases for Hardy spaces*, in Conference on Harmonic Analysis in Honor of Antoni Zygmund, Vol. II, Beckner et al., ed., University of Chicago Press, Chicago, IL, 1981, pp. 475–494.
[104] W. Sweldens, *Construction and Applications of Wavelets in Numerical Analysis*, Ph.D. thesis, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, 1994.
[105] W. Sweldens, *The lifting scheme: A custom-design construction of biorthogonal wavelets*, J. Appl. Comput. Harmonic Analysis, 3 (1996), pp. 186–200.
[106] W. Sweldens and P. Schröder, *Building your own wavelets at home*, in Wavelets in Computer Graphics, ACM SIGGRAPH Course notes, 1996, pp. 15–87; also available online from http://cm.bell-labs.com/who/wim/papers/papers.html#athome.
[107] M. Unser, A. Aldroubi, and M. Eden, *On the asymptotic convergence of B-spline wavelets to Gabor functions*, IEEE Trans. Inform. Theory, 38 (1992), pp. 864–872.
[108] M. Unser, A. Aldroubi, and M. Eden, *A family of polynomial spline wavelet transforms*, Signal Process., 30 (1993), pp. 141–162.
[109] P. P. Vaidyanathan, T. Q. Nguyen, Z. Doğanata, and T. Saramäki, *Improved technique for design of perfect reconstruction fir QMF banks with lossless polyphase matrices*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 1042–1055.
[110] M. Vetterli and C. Herley, *Wavelets and filter banks: Theory and design*, IEEE Trans. Acoust. Speech Signal Process., 40 (1992), pp. 2207–2232.
[111] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

[112] L. F. Villemoes, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1119–1543.
[113] L. F. Villemoes, *Wavelet analysis of refinement equations*, SIAM J. Math. Anal., 25 (1994), pp. 1433–1460.
[114] M. V. Wickerhauser, *Acoustic signal compression with wavelet packets*, in [20], pp. 679–700.
[115] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A. K. Peters, Wellesley, MA, 1994.
[116] H. Yserentant, *On the multi-level splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 547–554, May 1998

001

THE SHAPE OF THE TALLEST COLUMN*

STEVEN J. COX† AND C. MAEVE MCCARTHY†

Abstract. The height at which an unloaded column will buckle under its own weight is the fourth root of the least eigenvalue of a certain Sturm–Liouville operator. We show that the operator associated with the column proposed by Keller and Niordson [*J. Math. Mech.*, 16 (1966), pp. 433–446] *does not* possess a discrete spectrum. This calls into question their formal use of perturbation theory, so we consider a class of designs that permits a tapered summit yet still guarantees a discrete spectrum. Within this class we prove that the least eigenvalue increases when one replaces a design with its decreasing rearrangement. This leads to a very simple proof of the existence of a tallest column.

Key words. buckling load, self-weight, continuous spectrum, rearrangement

AMS subject classifications. 34L15, 49J99, 73H05

PII. S0036141097314537

1. Introduction.

Euler [4], [5] posed and solved the problem of buckling of prismatic columns under self-weight. He found that a column, clamped at its base and free at its summit, could be built to a height of

$$H_c = \left(\frac{9EI}{4\rho A} j_{-1/3}^2\right)^{1/3}$$

before buckling under its own weight. Here E denotes Young's modulus, ρ denotes weight density, A and I denote cross-sectional area and its second moment, and $j_{-1/3} \approx 1.8663$ is the least positive root of the Bessel function of order $-1/3$. To take a particular instance, the critical height of a circular cylinder of volume V is

$$H_c = \left(\frac{9EV}{16\pi\rho} j_{-1/3}^2\right)^{1/4}.$$

Almost two hundred years elapsed before Keller and Niordson [9] asked what height one could reach if, while fixing V, the overall volume, one was permitted to taper the column by varying A, and hence I, from point to point. Keller and Niordson formulated the Euler problem of critical height in terms of an eigenvalue problem for an ordinary differential operator and proceeded to maximize its least eigenvalue over a large class of shapes. In the spirit of Keller's previous attacks on buckling problems, [8] and [14], it was supposed that this least eigenvalue varied smoothly over the admissible class of shapes. This approach led Keller and Niordson to propose a shape that is so severely tapered at its summit that we are able to show that the associated differential operator *does not* possess a discrete spectrum. As this calls into doubt their method, if not their result, we argue that the problem merits reconsideration. Indeed, though the literature on column buckling is vast, see, e.g., Gajewski and Zyczkowski [7], the tallest column problem appears to have started and ended with the work of Keller and Niordson!

*Received by the editors January 6, 1997; accepted for publication (in revised form) June 3, 1997. This research was supported by NSF grant DMS-9258312.
http://www.siam.org/journals/sima/29-3/31453.html
†Department of Computational and Applied Mathematics, Rice University, 6100 Main St., Houston, TX 77005 (cox@rice.edu, maeve@rice.edu).

In section 2 we recall Keller and Niordson's formulation of the problem, their proposed tallest column, and present a proof that the associated spectrum is not discrete. In section 3 we isolate an admissible class of designs that permits tapered ends and guarantees a discrete spectrum. Within this class we prove, in section 4, the intuitively obvious, though technically elusive, fact that the least eigenvalue increases when one replaces a shape with its decreasing rearrangement. This leads to a very simple proof, in section 5, of the existence of a tallest column.

2. The work of Keller and Niordson. Tapered cross sections appear through the dependence of A on z, the distance from the clamped base. Assuming simply connected, geometrically similar cross sections, we find $I(z) = \alpha A^2(z)$, where α is a geometric constant. If $y(z)$ is the lateral deflection, from vertical, of the cross section at z and $EI(z)y''(z)$ is the associated bending moment then a balance of moments brings

$$E\alpha A^2(z)y''(z) = \int_z^H \rho A(\tilde{z})[y(\tilde{z}) - y(z)]\,d\tilde{z}, \quad 0 < z < H, \tag{2.1}$$

where H is the height of the column. Clamping the column at its base is synonymous with $y(0) = y'(0) = 0$. With V denoting the column's volume, Keller and Niordson introduce the dimensionless variables

$$x = z/H, \quad a(x) = HA(xH)/V, \quad \eta(x) = y(xH)/H, \quad \lambda = \rho H^4/\alpha EV$$

and so arrive at

$$a^2(x)\eta'' = \lambda \int_x^1 a(\tilde{x})[\eta(\tilde{x}) - \eta(x)]\,d\tilde{x}, \quad \eta(0) = \eta'(0) = 0, \tag{2.2}$$

and the area normalization

$$\int_0^1 a\,dx = 1. \tag{2.3}$$

Differentiating (2.2) with respect to x and calling $u(x) = \eta'(x)$, they finally obtain

$$-(a^2(x)u'(x))' = \lambda \left(\int_x^1 a(t)\,dt \right) u(x), \quad 0 < x < 1, \quad u(0) = a^2(1)u'(1) = 0. \tag{2.4}$$

We shall refer to this eigenvalue problem as the Euler problem and denote by $\lambda_1(a)$ its least eigenvalue. Keller and Niordson took up the problem of maximizing $\lambda_1(a)$ over those nonnegative a satisfying the volume constraint, (2.3). Supposing the existence of an optimal design, formal perturbation theory led them to a candidate $\tilde{a}$ for which

$$\tilde{a}(x) = \begin{cases} O(1), & \text{as } x \to 0, \\ c(1-x)^3 + O((1-x)^4), & \text{as } x \to 1, \end{cases} \tag{2.5}$$

where c is a positive constant. We shall now argue that the spectrum of the Euler problem is not discrete for such a design.

Following Friedrichs [6], we consider

$$\frac{1}{Z(x)} \equiv 4\tilde{a}^2(x) \left(\int_x^1 \tilde{a}(t)\,dt \right) \left(\int_0^x \tilde{a}(t)^{-2}\,dt \right)^2 \tag{2.6}$$

and recall [6, Criteria II & III] that if

$$\lambda_* \equiv \lim_{x\to 1} Z(x)$$

exists and

$$\lambda_* \le \liminf_{x\to 0} Z(x)$$

then the spectrum of (2.4) is discrete below λ_* and nondiscrete above λ_*. On substituting (2.5) into (2.6) we find, near $x = 1$, that

$$\begin{aligned}\frac{1}{Z(x)} &= 4\left(c^2(1-x)^6 + O((1-x)^7)\right)\left(\frac{c(1-x)^4}{4} + O((1-x)^5)\right)\\ &\qquad\times\left(\frac{1}{5c^2(1-x)^5} + \frac{1}{O((1-x)^4)}\right)^2\\ &= \frac{1}{25c} + O(1-x)\end{aligned}$$

and so $\lambda_* = 25c$. As $\tilde{a}$ is well behaved near $x = 0$, it follows easily that

$$\liminf_{x\to 0} Z(x) = +\infty.$$

We have just established the following proposition.

PROPOSITION 2.1. *The spectrum of the Euler problem,* (2.4), *with the Keller–Niordson design,* $\tilde{a}$, *is discrete below* $25c$ *and nondiscrete above* $25c$.

We note that the result states that *if* (2.4) has spectrum below $25c$ then this spectrum is discrete. It remains to see whether any such design produces eigenvalues below $25c$. Keller and Niordson's calculation of $c = \lambda_1/24$ suggests that their design indeed gives rise to an isolated eigenvalue, $\lambda_1 = 24c$, just below the continuous spectrum. Their result however was predicated on the false assumption that (2.4) possessed a purely discrete spectra. Nevertheless, we now construct a concrete design of unit volume that satisfies (2.5) and possesses at least one eigenvalue below $25c$. Again, the main idea lies with Friedrichs [6]; if there exists a function u for which $u(0) = 0$ and

$$\int_0^1 \tilde{a}^2|u'|^2\,dx < \lambda_* \int_0^1 \left(\int_x^1 \tilde{a}(t)\,dt\right) u^2\,dx < \infty$$

then (2.4) possesses at least one eigenvalue below λ_*. We shall apply this to

$$\tilde{a}(x) = 5(1-x)^3 - (5/4)(1-x)^4 \quad \text{and} \quad u(x) = \frac{x}{(1-x)^2}.$$

This $\tilde{a}$ is clearly positive away from $x = 1$ and satisfies the volume constraint (2.3). These choices produce $\lambda_* = 125$,

$$\int_0^1 \tilde{a}^2|u'|^2\,dx = \frac{1145}{24}, \quad \text{and} \quad \int_0^1 \left(\int_x^1 \tilde{a}(t)\,dt\right) u^2\,dx = \frac{19}{48},$$

and so this design has an eigenvalue below

$$\frac{1145/24}{19/48} \approx 120.5263.$$

Although one may easily generalize this example we have not been able to produce an exact characterization of those $\tilde{a}$ that satisfy (2.5) *and* possess at least one isolated eigenvalue. In other words, we have not been able to formulate a (large) class of admissible designs that accommodate Keller and Niordson's belief in cubic taper and presence of discrete spectra. In the interest of rigorously establishing the existence of a tallest column we have been compelled to exclude the possibility of cubic taper.

3. The Green's function. When the Green's function associated with the Euler problem is square integrable, the associated Green's operator is compact on $L^2(0,1)$ and therefore in possession of a discrete spectrum.

Following the standard recipe, see, e.g., Porter and Stirling [12, Example 6.13], the Green's function associated with the Euler problem (2.4) is

$$g(x,y;a) = \sqrt{w(x)w(y)} \int_0^{x\wedge y} \frac{dt}{a^2(t)},$$

where $x \wedge y = \min\{x, y\}$ and

$$w(x) \equiv \int_x^1 a(t)\, dt.$$

We now isolate a class of a that is rich enough to describe a large number of columns yet narrow enough to guarantee purely discrete spectra. On physical grounds it seems apparent that the tallest columns will be those that taper at their summit. We control the degree of taper by asking the cross-sectional area to lie in

$$ad_p \equiv \{a : k_1(1-x)^p \le a(x) \le k_2(1-x)^p\}$$

for some positive values of $k_1,\ k_2$, and p. It follows immediately that for such a

$$g^2(x,y;a) \le \frac{k_2^2}{k_1^2(1+p)^2(1-2p)^2}(1-x)^{1+p}(1-y)^{1+p}(1+(1-x\wedge y)^{2-4p}).$$

This being integrable over $(0,1)\times(0,1)$ so long as $0 \le p < 3$, we find the following proposition.

PROPOSITION 3.1. *If $a \in ad_p$ and $0 \le p < 3$ then the spectrum of the Euler problem is discrete.*

Proof. As $g(\cdot,\cdot;a) \in L^2((0,1)\times(0,1))$, it follows from [12, Theorem 3.4] that the Green's operator

$$G(a)\phi(x) \equiv \int_0^1 g(x,y;a)\phi(y)\, dy$$

is a compact operator on $L^2(0,1)$. This operator is also self-adjoint and positive and so, see, e.g., [12, Theorem 4.15], its spectrum is composed solely of a discrete sequence of nonnegative real numbers. □

We remark that $\tilde{a}$, the design of Keller and Niordson, remains just out of our reach. We next invoke [12, Lemma 5.1] in the variational characterization

$$\frac{1}{\lambda_1(a)} = \max_{\|\phi\|=1} \langle G(a)\phi, \phi\rangle, \tag{3.1}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard $L^2(0,1)$ inner product and $\|\cdot\|$ denotes the associated norm. The maximum is attained at $\phi_1 = \sqrt{w}u_1$, where u_1 is the first eigenfunction of

(2.4). As the boundary conditions on (2.4) are separated, the standard oscillation theory implies that $\lambda_1(a)$ is simple and that u_1 may be assumed everywhere nonnegative. Our first application of (3.1) is the following proposition.

PROPOSITION 3.2. *If $a \in ad_p$ and $0 \le p < 3$ and a satisfies the volume constraint* (2.3) *then*

$$\lambda_1(a) \le \frac{720}{k_1}.$$

Proof. Choosing $\phi \equiv 1$ in (3.1) brings

$$\frac{1}{\lambda_1(a)} \ge \int_0^1 \int_0^1 g(x,y;a)\,dx\,dy,$$

so we proceed to establish a pointwise lower bound for g.

$$\begin{aligned} g(x,y;a) &\ge \frac{k_1}{4}(1-x)^2(1-y)^2(x\wedge y)\left(\frac{1}{x\wedge y}\int_0^{x\wedge y} a^{-2}(t)\,dt\right) \\ &\ge \frac{k_1}{4}(1-x)^2(1-y)^2(x\wedge y)\left(\frac{1}{x\wedge y}\int_0^{x\wedge y} a(t)\,dt\right)^{-2} \\ &\ge \frac{k_1}{4}(1-x)^2(1-y)^2(x\wedge y)^3 \\ &\equiv g_0(x,y). \end{aligned}$$

The first inequality stems from $a(x) \ge k_1(1-x)^3$, the second is Jensen's inequality, while the third uses the nonnegativity of a and the volume constraint (2.3). As

$$\int_0^1 \int_0^1 g_0(x,y)\,dx\,dy = \frac{k_1}{720},$$

our result follows. □

Our second application of (3.1) states that the eigenvalues depend continuously on the Green's function.

PROPOSITION 3.3. *If a_1 and a_2 each lie in ad_p for $p < 3$ then*

$$\left|\frac{1}{\lambda_1(a_1)} - \frac{1}{\lambda_1(a_2)}\right| \le \|g(\cdot,\cdot;a_1) - g(\cdot,\cdot;a_2)\|.$$

Proof. Set $d(a_1,a_2) \equiv \|g(\cdot,\cdot;a_1) - g(\cdot,\cdot;a_2)\|$. Hölder's inequality provides

$$\langle G(a_2)\phi,\phi\rangle - d(a_1,a_2) \le \langle G(a_1)\phi,\phi\rangle \le \langle G(a_2)\phi,\phi\rangle + d(a_1,a_2),$$

when $\|\phi\| = 1$. Applying (3.1) throughout now gives

$$\frac{1}{\lambda_1(a_2)} - d(a_1,a_2) \le \frac{1}{\lambda_1(a_1)} \le \frac{1}{\lambda_1(a_2)} + d(a_1,a_2). \qquad \square$$

As preparation for our result on rearrangements we express (3.1) in a form reminiscent of that invoked by Alvino & Trombetti [1].

PROPOSITION 3.4. *If $a \in ad_p$ with $p < 3$ then*

$$\frac{1}{\lambda_1(a)} = \max_{\|\phi\|=1} \int_0^1 \frac{1}{a^2(x)}\left(\int_x^1 \sqrt{w(y)}\phi(y)\,dy\right)^2 dx.$$

Proof. This follows directly from

$$\langle G(a)\phi,\phi\rangle = \int_0^1\int_0^1 \sqrt{w(x)w(y)}\int_0^{x\wedge y}\frac{dt}{a^2(t)}\phi(x)\phi(y)\,dy\,dx$$
$$= \int_0^1 \sqrt{w(x)}\phi(x)\left(\int_0^x \sqrt{w(y)}\phi(y)\int_0^y \frac{dt}{a^2(t)}\,dy + \int_x^1 \sqrt{w(y)}\phi(y)\int_0^x \frac{dt}{a^2(t)}\,dy\right)dx.$$

Integrating by parts brings

$$\int_0^x \sqrt{w(y)}\phi(y)\int_0^y \frac{dt}{a^2(t)}\,dy = -\int_0^x \frac{dt}{a^2(t)}\int_x^1 \sqrt{w(t)}\phi(t)\,dt + \int_0^x \frac{1}{a^2(y)}\int_y^1 \sqrt{w(t)}\phi(t)\,dt\,dy,$$

so

$$\langle G(a)\phi,\phi\rangle = \int_0^1 \sqrt{w(x)}\phi(x)\int_0^x\int_y^1 \sqrt{w(t)}\phi(t)\,dt\frac{dy}{a^2(y)}\,dx.$$

Integrating this by parts, we arrive at the final form

$$\langle G(a)\phi,\phi\rangle = \int_0^1 \frac{1}{a^2(x)}\left(\int_x^1 \sqrt{w(t)}\phi(t)\,dt\right)^2 dx. \qquad \square$$

4. Increasing height via decreasing rearrangement. Expecting that the most efficient use of material will start from a large base and suffer a gradual diminution, we here show that replacing a by its decreasing rearrangement can but increase $\lambda_1(a)$. We follow a line of reasoning which, in our context, goes back to Krein [10] and Beesack & Schwarz [2]. The former considered the effect of the rearrangement of mass density while the latter addressed the rearrangement of a potential term. Our problem, with the design variable appearing in a nonlinear fashion in the highest order term and in a nonlocal fashion in the lowest order term, is considerably more cumbersome. Our contribution amounts to striking upon a variational characterization of $\lambda_1(a)$ which permits the application of the methods of [10] and [2].

Recall that the decreasing rearrangement of a nonnegative function, f, on $(0,1)$ is simply

$$f^*(x) \equiv \sup\{t>0 : \mu_f(t) > x\},$$

where

$$\mu_f(t) = |\{x\in(0,1) : f(x) > t\}|$$

is the measure of the set on which f exceeds t. The increasing rearrangement of f is simply $f_*(x)\equiv f^*(1-x)$. It is not difficult to show that

$$\int_0^1 f\,dx = \int_0^1 f^*\,dx = \int_0^1 f_*\,dx. \tag{4.1}$$

Regarding integrals of products, we recall the following propposition.

PROPOSITION 4.1. *Let f,ξ, and η be nonnegative functions, with ξ increasing and η decreasing. Then*

$$\int_0^1 f^*\xi\,dx \le \int_0^1 f\xi\,dx \quad \textit{and} \quad \int_0^1 f_*\eta\,dx \le \int_0^1 f\eta\,dx. \tag{4.2}$$

Proof. These are both special cases of inequalities established in Pólya and Szegö [11, p. 153]. □

As final preparation we recall the increasing rearrangement of a certain composition.

PROPOSITION 4.2. *If ψ is decreasing on the range of f then $(\psi \circ f)_* = \psi \circ f^*$.*

Proof. This is a special case of Cox [3, Theorem 1]. □

PROPOSITION 4.3. *If $a \in ad_p$ and $p < 3$ then $\lambda_1(a) \le \lambda_1(a^*)$.*

Proof. Denote by v the eigenfunction of $G(a^*)$ corresponding to $\lambda_1(a^*)$. As previously remarked, v is nonnegative. Now

$$\begin{aligned}
\frac{1}{\lambda_1(a)} &\ge \int_0^1 \frac{1}{a^2(x)} \left(\int_x^1 \left(\int_y^1 a(t)\,dt \right)^{1/2} v(y)\,dy \right)^2 dx \\
&\ge \int_0^1 \frac{1}{a^2(x)} \left(\int_x^1 \left(\int_y^1 a^*(t)\,dt \right)^{1/2} v(y)\,dy \right)^2 dx \\
&\ge \int_0^1 \left(\frac{1}{a^2(x)} \right)_* \left(\int_x^1 \left(\int_y^1 a^*(t)\,dt \right)^{1/2} v(y)\,dy \right)^2 dx \\
&= \int_0^1 \frac{1}{(a^*)^2(x)} \left(\int_x^1 \left(\int_y^1 a^*(t)\,dt \right)^{1/2} v(y)\,dy \right)^2 dx \\
&= \frac{1}{\lambda_1(a^*)}.
\end{aligned}$$

The first inequality is a direct consequence of Proposition 3.4. The second inequality comes from the first in (4.2) with ξ being the characteristic function of $(x, 1)$. The third inequality follows from the second in (4.2) with

$$\eta(x) = \left(\int_x^1 \left(\int_y^1 a^*(t)\,dt \right)^{1/2} v(y)\,dy \right)^2.$$

We remark that the nonnegativity of v leads to the nonincreasing of η. The first equality is a consequence of Proposition 4.2 with $\psi(t) = t^{-2}$. The final equality follows directly from the definition of v. □

5. Existence of a tallest column. Let us denote by ad_p^1 the collection of $a \in ad_p$ obeying the integral constraint (2.3). For $p < 3$ it follows from Proposition 3.2 that λ_1 is bounded on ad_p^1 and so

$$\lambda_1^{(p)} \equiv \sup_{a \in ad_p^1} \lambda_1(a)$$

is finite. That this sup is attained will follow directly from the following proposition.

PROPOSITION 5.1 (Helly's selection theorem, [13, p. 167]). *If $\{f_n\}$ is a sequence of nonnegative nonincreasing functions on $[0, 1]$ then there exists a subsequence $\{f_{n_k}\}$ and a function f such that $f_{n_k}(x) \to f(x)$ as $k \to \infty$ for each $x \in [0, 1]$.*

PROPOSITION 5.2. *If $p < 3$ then $a \mapsto \lambda_1(a)$ attains its maximum on $a \in ad_p^1$.*

Proof. As $\lambda_1^{(p)} < \infty$ there exists a maximizing sequence $\{a_n\} \subset ad_p^1$ for which $\lambda_1(a_n) \to \lambda_1^{(p)}$. By (4.1) and Proposition 4.3 we may assume that each a_n is nonincreasing and hence, by Helly's selection theorem, that there exists an $\hat{a}$ and a subsequence (that we neglect to relabel) such that $a_n \to \hat{a}$ pointwise. It follows by the

dominated convergence theorem (Rudin [13, Theorem 11.32]) that

$$\int_x^1 a_n(t)\,dt \to \int_x^1 \hat{a}(t)\,dt \quad \text{and} \quad \int_0^{x\wedge y} \frac{dt}{a_n^2(t)} \to \int_0^{x\wedge y} \frac{dt}{\hat{a}^2(t)}$$

for each x and y. In particular,

$$\int_0^1 a_n\,dx \to \int_0^1 \hat{a}\,dx \quad \text{and} \quad g(x,y;a_n) \to g(x,y;\hat{a}).$$

By the dominated convergence theorem it follows that $g(\cdot,\cdot,a_n) \to g(\cdot,\cdot,\hat{a})$ in $L^2((0,1)\times(0,1))$. This implies, via Proposition 3.3, that $\lambda_1(a_n) \to \lambda_1(\hat{a})$. But, by construction, $\lambda_1(a_n) \to \lambda_1^{(p)}$, and so $\lambda_1(\hat{a}) = \lambda_1^{(p)}$. □

REFERENCES

[1] A. Alvino and G. Trombetti, *A lower bound for the first eigenvalue of an elliptic operator*, J. Math. Anal. Appl., 94 (1983), pp. 328–337.

[2] P. R. Beesack and B. Schwarz, *On the zeros of solutions of second-order linear differential equations*, Canadian J. Math., 8 (1956), pp. 504–515.

[3] S. J. Cox, *The two phase drum with the deepest bass note*, Japan J. Indust. Appl. Math., 8 (1991), pp. 345–355.

[4] L. Euler, *Determinatio onerum, quae columnae gestare valent*, Leonhardi Euleri Opera Omnia 2, Vol. 17, C. Blanc and P. de Haller, eds., Orell Füssli Turici, Switzerland, 1982, pp. 232–251.

[5] L. Euler, *Examen insignis paradoxi in theoria columnarum occurentis*, Leonhardi Euleri Opera Omnia 2, Vol. 17, C. Blanc and P. de Haller, eds., Orell Füssli Turici, Switzerland, 1982, pp. 252–265.

[6] K. O. Friedrichs, *Criteria for discrete spectra*, Comm. Pure Appl. Math., 3 (1950), pp. 439–449.

[7] A. Gajewski and M. Zyczkowski, *Optimal Structural Design under Stability Constraints*, Kluwer, Boston, 1988.

[8] J. B. Keller, *The shape of the strongest column*, Arch. Rational Mech. Anal., 5 (1960), pp. 275–285.

[9] J. B. Keller and F. I. Niordson, *The Tallest Column*, J. Math. Mech., 16 (1966), pp. 433–446.

[10] M. G. Krein, *On certain problems on the maximum and minimum of characteristic values and on the Lyapunov zones of stability*, Transl. Amer. Math. Ser., 2 (1955), pp. 163–187.

[11] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Ann. of Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.

[12] D. Porter and D. Stirling, *Integral Equations*, Cambridge University Press, Cambridge, UK, 1990.

[13] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., McGraw–Hill, New York, 1976.

[14] I. Tadjbakhsh and J. B. Keller, *Strongest columns and isoperimetric inequalities for eigenvalues*, J. Appl. Mech., 29 (1962), pp. 159–164.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 555–595, May 1998

002

SLOW DYNAMICS OF INTERFACES IN THE ALLEN–CAHN EQUATION ON A STRIP-LIKE DOMAIN*

SHIN-ICHIRO EI† AND EIJI YANAGIDA ‡

Abstract. The dynamics of interfaces in the Allen–Cahn equation is studied. If a domain in $\boldsymbol{R^2}$ has constant width along a smooth curve, it is called a strip-like domain. We derive an equation which describes the motion of a straight interface intersecting the boundary of the strip-like domain. The equation shows that the motion is slower than the mean curvature flow, but faster than the very slow dynamics.

Key words. dynamics of interfaces, Allen–Cahn equation, super- and subsolutions

AMS subject classifications. 35B25, 35K57

PII. S0036141096307205

1. Introduction.

Recently, the dynamics of interfaces appearing in various reaction diffusion equations has received much attention. Among them, the motion of interfaces in the Allen–Cahn equation

$$\begin{cases} u_t = \varepsilon^2 \Delta u + f(u), & x \in \Omega, \\ \dfrac{\partial}{\partial n} u = 0, & x \in \partial\Omega, \end{cases} \tag{1.1}$$

has been most extensively studied, where $\varepsilon > 0$ is a sufficiently small parameter, Ω a (bounded or unbounded) domain in $\boldsymbol{R}^N$ with boundary $\partial\Omega$, Δ the Laplace operator, $\partial/\partial n$ the outer normal derivative. The nonlinearity f is given by $f = u(1-u^2)$, although we can extend our results to more general nonlinearities. In fact, we may assume that f is a sufficiently smooth cubic-like odd function which is derived from double-well potential with equal depth.

Given initial data, the solution u of (1.1) behaves in the following way. Since $\varepsilon > 0$ is small, the diffusion term is negligible in the first stage so that the solution approaches $+1$ or -1 depending only on the sign of the initial value. Thus, after some time, there must appear an inner layer (or an interface) in which the value of u rapidly changes from -1 to $+1$. At this stage, the diffusion term is not negligible near the interface, and the interface begins to move slowly.

In the case where $N = 1$ and Ω is a compact interval, the dynamics of interfaces was precisely investigated by Carr and Pego [3] and Fusco and Hale [7]. They showed that the motion of interfaces is governed by the *very slow dynamics*. More precisely, it moves with the speed $O(e^{-A/\varepsilon})$ for some positive constant A.

In higher dimension, it is known (see, e.g., [2, 4, 5, 6, 12]) that the motion of interfaces is approximately described by the *mean curvature flow*

$$V = -\varepsilon^2 (N-1)k, \quad x \in \Gamma. \tag{1.2}$$

Here $\Gamma = \Gamma(t)$ is a hypersurface in $\boldsymbol{R}^N$ which represents the location of an interface,

*Received by the editors July 24, 1996; accepted for publication (in revised form) May 27, 1997. http://www.siam.org/journals/sima/29-3/30720.html

†Graduate School of Integrated Science, Yokohama City University, Yokohama 236-0027, Japan (ei2s@yokohama-cu.ac.jp).

‡Graduate School of Mathematical Sciences, University of Tokyo, Meguro-ku, Tokyo 153-0041, Japan (yanagida@ms.u-tokyo.ac.jp).

and V and k stand for the normal velocity and the mean curvature of Γ, respectively. When Γ intersects $\partial\Omega$, it must be orthogonal to $\partial\Omega$ (cf. [8, 11]).

If Γ is a minimal surface (i.e., $k = 0$), then it is a steady state of (1.2) so that the interface does not move under the dynamics (1.2). However, if Ω is a cylindrical domain and Γ is a plane, the interface may move with a slower time scale. In fact, if a solution is constant in the transectional direction, then the solution behaves exactly in the same manner as in the one-dimensional case so that the motion of the interface is governed by the very slow dynamics.

Recently, Alikakos, Fusco, and Kowalczyk [1] dealt with (1.1) on a domain in $\boldsymbol{R}^2$ which consists of a rectangle and other parts. When Γ is a line segment inside the rectangle, they showed that the motion of Γ depends on the shape of Ω and the speed is $O(e^{-A/\varepsilon})$. Namely, the motion of Γ is governed by the very slow dynamics. It seems that this result can be extended to Ω which consists of a part of an annulus and other parts [9].

In this paper, we show that in a certain kind of domain, there may appear a motion with the speed $O(\varepsilon^4)$, which we will call the *slow dynamics*. Let $C(s), 0 < s < L$, be a smooth curve in $\boldsymbol{R}^2$ with its length L, where s is the arclength parameter. We define a strip-like domain Ω in $\boldsymbol{R}^2$ along $C(s)$ with its width $2d$ by

$$\Omega = \{C(s) + z\nu(s) \mid 0 < s < L,\ -d < z < d\},$$

where $\nu(s)$ is a unit normal vector of $C(s)$ (see Fig. 1.1). We assume that the curvature $\kappa(s)$ of $C(s)$ satisfies

$$d \sup_{0<s<L} |\kappa(s)| \le \delta \tag{1.3}$$

for some $0 < \delta < 1$. We also assume that Ω is simply connected and any point in Ω corresponds one-to-one to (s, z) by the relation

$$(x, y) = C(s) + z\nu(s).$$

Let Γ be a line segment intersecting $\partial\Omega$ orthogonally. Then Γ stands still under the dynamics (1.2). However, the interface may move in a slower time scale by keeping its shape straight, because the line segment can be translated freely along $C(s)$.

Our results assert that any straight interface moves with the speed $O(\varepsilon^4)$ toward the direction where C is less curved. More precisely, let $s = S(t)$ represent the s coordinate of the position of Γ. Then $S(t)$ is approximately governed by

$$\frac{d}{dt}S(t) = \varepsilon^4 H(S(t)), \tag{1.4}$$

where

$$H(s) = -\frac{M}{\{1 - d^2\kappa(s)^2\}^2\ \left\{1 + \frac{1}{3}d^2\kappa(s)^2\right\}}\ \kappa(s)\kappa_s(s),$$

$$M = \frac{\displaystyle\int_{-\infty}^{+\infty} \eta^2\varphi_\eta(\eta)^2 d\eta}{\displaystyle\int_{-\infty}^{+\infty} \varphi_\eta(\eta)^2 d\eta} > 0,$$

and φ is a unique solution of

$$\begin{cases} \varphi_{\eta\eta} + f(\varphi) = 0, \quad -\infty < \eta < \infty, \\ \varphi(0) = 0, \quad \varphi(\pm\infty) = \pm 1. \end{cases}$$

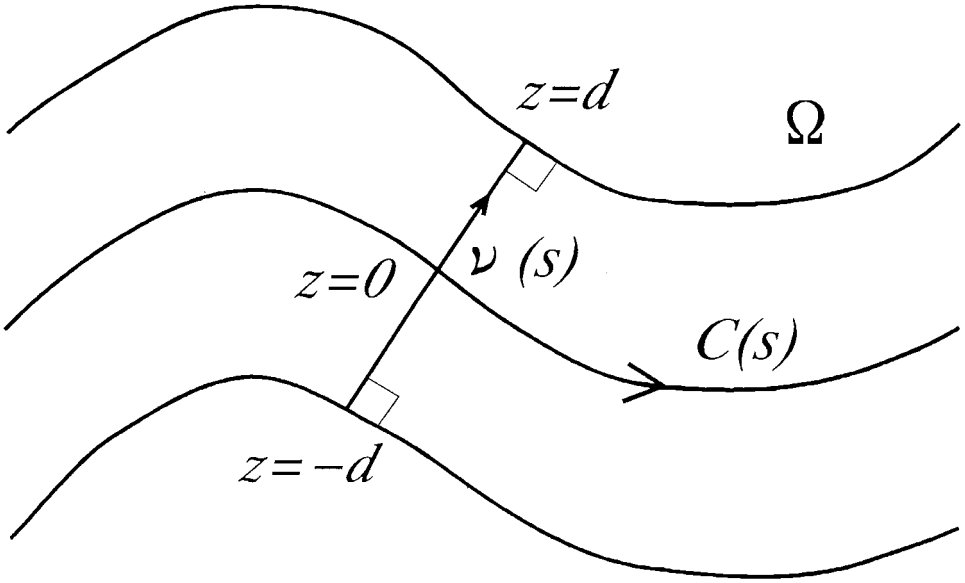

FIG. 1.1. *A strip-like domain* Ω.

Under our assumption on f, the solution φ is a sufficiently smooth function which is monotone increasing and odd symmetric with respect to η. Since $H(s)$ can be represented as

$$H(s) = -A(s)\{\kappa(s)^2\}_s$$

for some positive function $A(s)$, (1.4) implies that the interface moves toward a minimal point of $|k(s)|$. In Fig. 1.2, we present some numerical results which demonstrate this slow dynamics of interfaces.

When $\kappa(s)$ is constant, then $H(s) \equiv 0$, which means that the interface moves in slower dynamics. On the other hand, if $\kappa(s)$ is constant, then the domain must be a rectangle or a part of annulus. In this case, the motion of interface is governed by the very slow dynamics as mentioned previously. Thus, for a domain in $\boldsymbol{R}^2$, there are three kinds of dynamics, that is, the mean curvature flow, the slow dynamics, and the very slow dynamics.

This paper is organized as follows. In section 2, we give some definitions and our main results. In section 3, we state several important propositions. In section 4, we give proofs of the main results by using these propositions. Sections 5 and 6 are devoted to proofs of the propositions.

2. Definitions and results. Before stating our main results, we list our notation.

$\kappa(s)$: the curvature of $C(s)$ measured in the direction of $\nu(s)$.

$l(s)$: the straight line segment through $C(s)$ intersecting $\partial\Omega$ orthogonally, that is,

$$l(s) = \{C(s) + z\nu(s)\ ;\ -d < z < d\}.$$

C^z : the curve defined by

$$C^z = \{C(s) + z\nu(s)\ ;\ 0 < s < L\}.$$

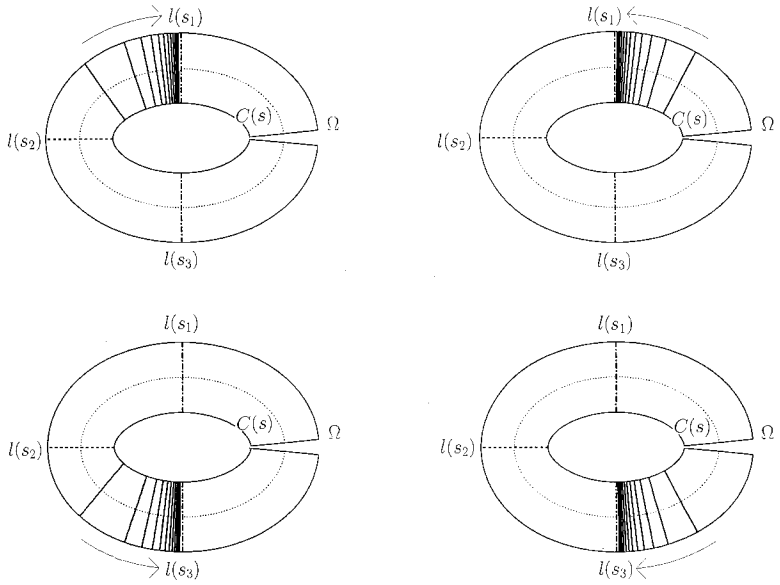

FIG. 1.2. *The dynamics of interfaces in a strip-like domain Ω whose center curve (a dotted curve) is a part of an ellipse. An arrow indicates the direction of movement of an interface. Broken lines denote the line segments $l(s_1), l(s_2), l(s_3)$, where s_1, s_3 are the values of s corresponding to the minima of $|\kappa(s)|$, while s_2 is the value of maximum of $|\kappa(s)|$.*

$\mathrm{Dist}\{(x,y), l(s)\}$: the signed distance between the point $(x,y) \in \Omega$ and the line segment $l(s)$ measured along C^z, that is,

$$\mathrm{Dist}\{(x,y), l(s)\} = (1 - z\kappa(s))(r - s) \quad \text{for } (x,y) = C(r) + z\nu(r).$$

$\Pi(P, Q, s)$: the set of functions satisfying the conditions

$$\begin{array}{rcccccccccc} -1-Q & \le & u(x,y) & \le & -1+Q & \text{if} & -\infty & \le & \mathrm{Dist}\{(x,y),l(s)\} & \le & -P, \\ -1-Q & \le & u(x,y) & \le & +1+Q & \text{if} & -P & \le & \mathrm{Dist}\{(x,y),l(s)\} & \le & +P, \\ +1-Q & \le & u(x,y) & \le & +1+Q & \text{if} & +P & \le & \mathrm{Dist}\{(x,y),l(s)\} & \le & +\infty, \end{array}$$

where $P > 0$ and $Q > 0$.

T : the time scale $\varepsilon^4 t$.

$\langle \cdot, \cdot \rangle_\eta$: the inner product of $L^2(\boldsymbol{R})$ with respect to η.

In the time scale $T = \varepsilon^4 t$, (1.4) can be written as

$$\frac{d}{dT} S(T) = H(S(T)), \quad T > 0. \tag{2.1}$$

In the following theorems, we denote a solution of (1.1) by $u(t,x,y)$.

THEOREM 2.1. *Let $S(T)$ be a solution of* (2.1) *defined for $T \in [0, T^*]$ for some $T^* < \infty$. Then for any $m \in (0,1)$, there exist positive constants P_1, Q_1, P_2, Q_2, and ε_0 such that if $0 < \varepsilon < \varepsilon_0$ and*

$$u(0,x,y) \in \Pi(P_1\varepsilon^m, Q_1\varepsilon^{m+3}, S(0)),$$

then

$$u(t,x,y) \in \Pi(P_2\varepsilon^m, Q_2\varepsilon^{m+3}, S(\varepsilon^4 t))$$

for $t \in [0, T^*/\varepsilon^4]$.

Let S^* be an equilibrium of (2.1). We say that S^* is exponentially stable if S^* satisfies $H'(S^*) < 0$.

THEOREM 2.2. *Suppose that a solution $S(T)$ of (2.1) converges to an exponentially stable equilibrium S^*. Then for any $m \in (0,1)$, there exist positive constants P_1, Q_1, P_2, Q_2, and ε_0 such that if $0 < \varepsilon < \varepsilon_0$ and*

$$u(0,x,y) \in \Pi(P_1\varepsilon^m, Q_1\varepsilon^{m+3}, S(0)),$$

then

$$u(t,x,y) \in \Pi(P_2\varepsilon^m, Q_2\varepsilon^{m+3}, S(\varepsilon^4 t))$$

for all $t \in [0,\infty)$.

THEOREM 2.3. *Let S^* be an exponentially stable equilibrium of (2.1). Then for any $m \in (0,1)$, there exist positive constants P, Q, and ε_0 such that if $0 < \varepsilon < \varepsilon_0$, (1.1) possesses a stable stationary solution $u^*(x,y)$ satisfying*

$$u^*(x,y) \in \Pi(P\varepsilon^m, Q\varepsilon^{m+3}, S^*).$$

Remark. It is easily seen from proofs of the above theorems that if Ω consists of a strip-like domain and other parts, similar results hold as long as a straight interface remains in the strip-like domain.

In (2.1), $H(s)$ is expressed as

$$H(s) = -A(s)\{\kappa(s)^2\}_s$$

for some positive function $A(s)$ as mentioned in section 1. This implies that the exponentially stable equilibrium S^* of (2.1) corresponds to a minimal point of $|\kappa(s)|$. Also Theorems 2.2 and 2.3 imply that the straight interface moves toward a minimal point of $|\kappa(s)|$ and converges to a stable stationary solution u^* with its inner layer near the line segment $l(S^*)$.

3. Preliminaries. In this section, we give several propositions as preliminaries for the subsequent sections. Throughout this paper, we let m be an arbitrarily fixed constant satisfying $0 < m < 1$. Also, we denote by A_j $(j = 0,1,2,\ldots)$ positive constants independent of t, $(x,y) \in \Omega$, sufficiently small $\varepsilon > 0$, and u.

In order to prove the theorems, we will employ the comparison method. We say that $u^+(t,x,y)$ $(t \in [0,t_0], (x,y) \in \Omega)$ is a supersolution of (1.1) if u^+ satisfies

$$u_t^+ \geq \Delta u^+ + f(u^+), \quad t \in [0,t_0],\ (x,y) \in \Omega,$$

and the Neumann boundary condition. Similarly, we say that $u^-(t,x,y)$, $(t \in [0,t_0], (x,y) \in \Omega)$ is a subsolution of (1.1) if u^- satisfies

$$u_t^- \leq \Delta u^- + f(u^-), \quad t \in [0,t_0],\ (x,y) \in \Omega,$$

and the Neumann boundary condition.

The following two propositions are well known (see, e.g., [10]).

PROPOSITION 3.1. *Let $u(t,x,y)$ be a solution of (1.1). If $u^+(t,x,y)$, $t \in [0,t_0]$, is a supersolution satisfying $u^+(0,x,y) \geq u(0,x,y)$ for $(x,y) \in \Omega$, then $u^+(t,x,y) \geq u(t,x,y)$ for $t \in [0,t_0], (x,y) \in \Omega$. Similarly, if $u^-(t,x,y)$, $t \in [0,t_0]$, is a subsolution satisfying $u^-(0,x,y) \leq u_0(x,y)$ for $(x,y) \in \Omega$, then $u^-(t,x,y) \leq u(t,x,y)$ for $t \in [0,t_0], (x,y) \in \Omega$.*

PROPOSITION 3.2. *If $u^+(x,y)$ and $u^-(x,y)$ are, respectively, super- and subsolutions independent of t satisfying*

$$u^-(x,y) \le u^+(x,y), \quad (x,y) \in \Omega,$$

then there exists a stable stationary solution $u^*(x,y)$ of (1.1) satisfying

$$u^-(x,y) \le u^*(x,y) \le u^+(x,y), \quad (x,y) \in \Omega.$$

The following two propositions are easily shown.

PROPOSITION 3.3. *For small $|q|$, there exist smooth functions $\alpha^\pm(q)$ such that $f(\alpha^\pm(q))+q=0$ and $\alpha^\pm(q) = \pm 1 + M_0 q + O(q^2)$, where $M_0 = -\frac{1}{f'(1)} = -\frac{1}{f'(-1)} > 0$.*

Here we define

$$|v(\eta)|_{C^h(\eta)} = \sum_{j=0}^{h} \left| \frac{d^j}{d\eta^j} v(\eta) \right|,$$
$$|v(\eta)|_{D^h(\eta)} = |v(\eta)|_{C^h(\eta)} - |v(\eta)|$$

for $v \in C^h$, the set of h times continuously differentiable functions.

PROPOSITION 3.4. *There exist positive constants a_0, A_0, and β_0 such that*

$$a_0 e^{-\beta_0|\eta|} \le |\varphi(\eta)+1| \le A_0 e^{-\beta_0|\eta|} \quad \text{for } \eta < 0,$$
$$a_0 e^{-\beta_0|\eta|} \le |\varphi(\eta)-1| \le A_0 e^{-\beta_0|\eta|} \quad \text{for } \eta > 0,$$

and for any positive integer h,

$$|\varphi(\eta)|_{D^h(\eta)} = O(e^{-\beta_0|\eta|}) \text{ as } |\eta| \to \infty.$$

Let $S^\pm(T)$ be solutions of

$$S_T^\pm = H^\pm(S^\pm) \tag{3.1}$$

with $S^\pm(0) = S(0) \mp \varepsilon^m$, where

$$H^\pm(s) = H(s) \mp \frac{2 - d\kappa(s)}{M_1\{1 + \frac{1}{3}d^2\kappa^2(s)\}}\varepsilon^m$$

and $M_1 = \langle\, \varphi, \varphi \,\rangle_\eta$. We assume that a solution $S(T)$ of (2.1) exists and is bounded for $T \in [0, T^*]$ for some $T^* \in (0, \infty]$. Then, we can construct super- and subsolutions as stated in the following proposition.

PROPOSITION 3.5. *Let $\gamma^\pm(T)$ be arbitrary smooth functions of $T \in [0, T^*]$, and set*

$$\eta^\pm = \{1 - z\kappa(S^\pm(T))\}\left(\frac{s - S^\pm(T)}{\varepsilon} - \gamma^\pm(T)\right)$$

with $(x,y) = C(s) + z\nu(s)$. Then for small $\varepsilon > 0$, there exist super- and subsolutions $u^\pm$ of (1.1) of the form

$$u^\pm(t,x,y) = \varphi(\eta^\pm) + \varepsilon\Phi^\pm(t,z,\eta^\pm;\varepsilon) \pm M_0\varepsilon^{m+3} + \varepsilon^4 R^\pm(t,x,y;\varepsilon), \tag{3.2}$$

where $\Phi^\pm$ satisfy

$$|\Phi^\pm(t,z,\eta;\varepsilon)|_{C^2(\eta)} \le A_1 e^{-\beta|\eta|}, \quad (t,z,\eta) \in [0, T^*/\varepsilon^4] \times (-d,d) \times (-\infty,\infty) \tag{3.3}$$

with some positive constants A_1 and $\beta \in (\frac{3}{4}\beta_0, \beta_0)$, and $R^\pm(t,x,y;\varepsilon)$ are uniformly bounded in $t \in [0, T^*/\varepsilon^4]$ and $(x,y) \in \Omega$.

A proof of this proposition will be given in section 5. All theorems in section 2 are derived from this proposition.

4. Proofs of theorems. In this and the following sections, B denotes a generic positive constant independent of sufficiently small ε and the solution u which may vary from line to line.

Let $S^{\pm}(T)$ be solutions of (3.1) with $S^{\pm}(0) = S(0) \mp \varepsilon^m$, respectively. Then there exists $B > 0$ such that

$$S^+(T) + B\varepsilon^m \le S(T) \le S^-(T) - B\varepsilon^m \tag{4.1}$$

uniformly in $T \in [0, T^*]$. For such $S^{\pm}(T)$, let $u^{\pm}(t, x, y)$ be, respectively, super- and subsolutions given in Proposition 3.5, and set

$$v^{\pm}(t, z, s) = u^{\pm}(t, C(s) + z\nu(s)).$$

LEMMA 4.1. *If $\varepsilon > 0$ is sufficiently small, then*

$$v^+(t, z, s) > v^-(t, z, s) + \frac{5}{3} M_0 \varepsilon^{m+3}, \quad (t, z, s) \in [0, T^*/\varepsilon^4] \times (-d, d) \times (0, L). \tag{4.2}$$

Proof. Let $\mu^{\pm} = \frac{s - S^{\pm}(T)}{\varepsilon}$ and let $\gamma^{\pm}(T)$ be a function as in Proposition 3.5. Then $v^{\pm}$ satisfy (3.2). Therefore it suffices to show (4.2) for $s > S^+(T) + \varepsilon\gamma^+(T)$.

First we consider the case where $S^+(T) + \varepsilon\gamma^+(T) < s < S^-(T) + \varepsilon\gamma^-(T)$. In this case, we have $\eta^+ > 0$, $\eta^- < 0$, and

$$\begin{aligned}
\eta^+ - \eta^- &= (\mu^+ - \mu^-) - (\gamma^+ - \gamma^-) - z\kappa^+(\mu^+ - \gamma^+) + z\kappa^-(\mu^- - \gamma^-) \\
&\ge (\mu^+ - \mu^-) - (\gamma^+ - \gamma^-) - \delta(\mu^+ - \mu^- - \gamma^+ + \gamma^-) \\
&= (1 - \delta)(\mu^+ - \mu^- - \gamma^+ + \gamma^-) \\
&= (1 - \delta)\left(\frac{S^- - S^+}{\varepsilon} - \gamma^+ + \gamma^-\right) \\
&\ge (1 - \delta)\left(\frac{2B}{\varepsilon^{1-m}} - \gamma^+ + \gamma^-\right) \\
&\ge \frac{B}{\varepsilon^{1-m}}
\end{aligned}$$

by noting $\mu^+ - \gamma^+ \ge 0$, $\mu^- - \gamma^- \le 0$, and (4.1). Hence the monotonicity of φ yields

$$\varphi(\eta^+) - \varphi(\eta^-) \ge \varphi(\eta^+) - \varphi\left(\eta^+ - \frac{B}{\varepsilon^{1-m}}\right).$$

Here, if $0 \le \eta^+ \le \frac{B}{2\varepsilon^{1-m}}$, we have

$$\begin{aligned}
\varphi(\eta^+) - \varphi\left(\eta^+ - \frac{B}{2\varepsilon^{1-m}}\right) &\ge \varphi(\eta^+) - \varphi\left(-\frac{B}{2\varepsilon^{1-m}}\right) \\
&\ge \varphi(\eta^+) + 1 - \exp\left(-\frac{\beta_0 B}{2\varepsilon^{1-m}}\right) \\
&\ge 1 - \exp\left(-\frac{\beta_0 B}{2\varepsilon^{1-m}}\right) \\
&\ge \frac{1}{2},
\end{aligned}$$

which means (4.2) by (3.2) and (3.3). Also, for $\eta^+ \ge \frac{B}{2\varepsilon^{1-m}}$, we can show

$$\varphi(\eta^+) - \varphi(\eta^-) \ge \frac{1}{2}$$

in quite a similar manner. Thus, (4.2) holds when $S^+(T)+\varepsilon\gamma^+(T) < s < S^-(T)+\varepsilon\gamma^-(T)$.

Next we consider the case where $s > S^-(T)+\varepsilon\gamma^-(T)$. Since $\eta^\pm \geq 0$ in this case, we see from Proposition 3.4, (3.2), and (3.3) that

$$v^+(t,z,s) \geq 1 - A_0 e^{-\beta_0|\eta^+|} - \varepsilon A_1 e^{-\beta|\eta^+|} + M_0\varepsilon^{m+3} - B\varepsilon^4, \tag{4.3}$$

$$v^-(t,z,s) \leq 1 - a_0 e^{-\beta_0|\eta^-|} + \varepsilon A_1 e^{-\beta|\eta^-|} - M_0\varepsilon^{m+3} + B\varepsilon^4. \tag{4.4}$$

Here $\eta^+ \geq \frac{B}{\varepsilon^{1-m}}$ holds. Hence, in view of (4.3) and (4.4), it suffices to show that

$$1 + M_0\varepsilon^{m+3} - B\varepsilon^4 \geq 1 - a_0 e^{-\beta_0|\eta^-|} + \varepsilon A_1 e^{-\beta|\eta^-|} - M_0\varepsilon^{m+3} + B\varepsilon^4 + \frac{5}{3}M_0\varepsilon^{m+3},$$

that is,

$$a_0 e^{-\beta_0|\eta^-|} - \varepsilon A_1 e^{-\beta|\eta^-|} \geq -\frac{1}{3}M_0\varepsilon^{m+3} + B\varepsilon^4. \tag{4.5}$$

Noting $\eta^- \geq 0$, we let

$$\sigma(\eta) = a_0 e^{-\beta_0\eta} - \varepsilon A_1 e^{-\beta\eta}.$$

By

$$\sigma'(\eta) = e^{-\beta_0\eta}\left(-a_0\beta_0 + \varepsilon A_1\beta e^{(\beta_0-\beta)\eta}\right),$$

$\sigma(\eta)$ has a minimum at $\eta = \eta^*$, where

$$\eta^* = \frac{1}{\beta_0-\beta}\log\left(\frac{a_0\beta_0}{\varepsilon A_1\beta}\right)$$

satisfies $\sigma'(\eta^*) = 0$. Substituting η^*, we have

$$\sigma(\eta^*) = A_1\left(\frac{A_1\beta}{a_0\beta_0}\right)^{\frac{\beta}{\beta_0-\beta}}\left(1-\frac{\beta_0}{\beta}\right)\varepsilon^{\frac{\beta_0}{\beta_0-\beta}} < 0.$$

Since $\frac{3}{4}\beta_0 < \beta < \beta_0$ and $0 < m < 1$, the inequality $\frac{\beta_0}{\beta_0-\beta} > m+3$ holds, which implies (4.5) for sufficiently small $\varepsilon > 0$. □

LEMMA 4.2. *There exist positive constants P_1 and Q_1 such that*

$$\Pi(P_1\varepsilon^m, Q_1\varepsilon^{m+3}, s_0) \subset \Pi^* \tag{4.6}$$

for sufficiently small $\varepsilon > 0$, where $s_0 = S(0)$ and

$$\Pi^* = \{w(z,s);\ v^-(0,z,s) \leq w(z,s) \leq v^+(0,z,s)\}.$$

Proof. Suppose $|\mathrm{Dist}\{(x,y), l(s_0)\}| \leq P_1\varepsilon^m$. This is equivalent to

$$|(1-z\kappa(s_0))(s-s_0)| \leq P_1\varepsilon^m,$$

where $(x,y) = C(s) + z\nu(s)$. Hence, we have

$$\left|\frac{s-s_0}{\varepsilon}\right| \leq \frac{P_1}{(1-z\kappa(s_0))\varepsilon^{1-m}} \leq \frac{P_1}{(1-\delta)\varepsilon^{1-m}}$$

and

$$\begin{aligned}\frac{s-S^+(0)}{\varepsilon} &= \frac{s-s_0}{\varepsilon}+\frac{s_0-S^+(0)}{\varepsilon}\\ &\geq -\frac{P_1}{(1-\delta)\varepsilon^{1-m}}+\frac{1}{\varepsilon^{1-m}}\\ &\geq \frac{B}{\varepsilon^{1-m}}\end{aligned}$$

for appropriately small constant $P_1>0$, and similarly

$$\frac{s-S^-(0)}{\varepsilon}\leq -\frac{B}{\varepsilon^{1-m}}.$$

Therefore, it follows that

$$\begin{aligned}\eta^+ &= (1-z\kappa^+)\left(\frac{s-S^+(0)}{\varepsilon}-\gamma^+(0)\right)\\ &\geq (1-\delta)\left(\frac{B}{\varepsilon^{1-m}}-\gamma^+(0)\right)\\ &\geq \frac{B}{\varepsilon^{1-m}},\end{aligned}\tag{4.7}$$

and similarly

$$\begin{aligned}\eta^- &= (1-z\kappa^-)\left(\frac{s-S^-(0)}{\varepsilon}-\gamma^-(0)\right)\\ &\leq -\frac{B}{\varepsilon^{1-m}}.\end{aligned}\tag{4.8}$$

By the inequalities (3.2), (3.3), (4.3) and (4.7), (4.8), we get

$$v^+(0,z,s)\geq 1+B\varepsilon^{m+3},\tag{4.9}$$

and similarly

$$v^-(0,z,s)\leq -1-B\varepsilon^{m+3}\tag{4.10}$$

for sufficiently small $\varepsilon>0$.

Next, suppose $\mathrm{Dist}\{(x,y),l(s_0)\}\geq P_1\varepsilon^m$. In this range, (4.7) and (4.9) also hold. On the other hand, (3.2) and (4.7) imply

$$v^+(0,z,s)\leq 1+\frac{4}{3}M_0\varepsilon^{m+3}$$

for sufficiently small $\varepsilon>0$. Hence, by Lemma 4.1, we obtain

$$v^-(0,z,s)\leq 1-\frac{1}{3}M_0\varepsilon^{m+3}.$$

Thus,

$$v^-(0,z,s)\leq 1-\frac{1}{3}M_0\varepsilon^{m+3}\leq 1+B\varepsilon^{m+3}\leq v^+(0,z,s)\tag{4.11}$$

holds for $\mathrm{Dist}\{(x,y),l(s_0)\}\geq P_1\varepsilon^m$.

The case that $\mathrm{Dist}\{(x,y), l(s_0)\} \leq -P_1\varepsilon^m$ can be treated in quite a similar manner.

Taking $0 < Q_1 \leq \min\{B, \frac{1}{3}M_0\}$, the proof is complete. □

Now let us complete the proofs of Theorems 2.1–2.3.

Proof of Theorem 2.1. Let $u(t,x,y)$ be the solution of (1.1) with

$$u(0,\cdot) \in \Pi(P_1\varepsilon^m, Q_1\varepsilon^{m+3}, s_0).$$

Then the inequalities

$$v^-(t,z,s) \leq u(t,x,y) \leq v^+(t,z,s)$$

hold for $(t,z,s) \in [0,T^*/\varepsilon^4] \times (-d,d) \times (0,L)$ by Lemma 4.2, where $(x,y) = C(s) + z\nu(s)$. On the other hand, we have

$$S(T) - B\varepsilon^m < S^+(T) < S(T) < S^-(T) < S(T) + B\varepsilon^m$$

for some B uniformly in $T \in [0,T^*]$. Therefore, we have

$$v^\pm(t,z,s) \in \Pi(P_2\varepsilon^m, Q_2\varepsilon^{m+3}, S(\varepsilon^4 t))$$

for certain positive constants P_2, Q_2 by a similar argument to the above. Thus the proof is complete. □

Proof of Theorem 2.2. Since S^* is an exponentially stable equilibrium of (2.1), the equation (3.1) also has exponentially stable equilibria $S^*_\pm$, respectively, satisfying

$$S^* - B\varepsilon^m < S^*_+ < S^* < S^*_- < S^* + B\varepsilon^m \tag{4.12}$$

for some B. Let $S^\pm(T)$ be solutions of (3.1) with $S^\pm(0) = S(0) \mp \varepsilon^m$. Then, for sufficiently small $\varepsilon > 0$, $S^\pm(T)$ converge to $S^*_\pm$ as $T \to \infty$, respectively, and the inequalities

$$S(T) - B\varepsilon^m < S^+(T) < S(T) < S^-(T) < S(T) + B\varepsilon^m$$

hold uniformly in $T \geq 0$. Therefore, we can prove Theorem 2.2 by quite a similar manner to Theorem 2.1. □

Proof of Theorem 2.3. Let $S^\pm(T)$ be solutions of (3.1) with $S^\pm(0) = S^*_\pm$, respectively. Then we see that $S^\pm(T) \equiv S^*_\pm$ and (4.12) hold. Since $v^\pm(t,z,s)$ corresponding to $S^\pm(T)$ are independent of t, the proof of Theorem 2.3 is complete by Proposition 3.2. □

5. Proof of Proposition 3.5.

5.1. Preliminaries. Now let us explain the way to construct a suitable supersolution $u^+(t,x,y)$. We consider the equation

$$u_t = \varepsilon^2\Delta u + f(u) + \varepsilon^{m+3}, \quad t > 0,\ (x,y) \in \Omega. \tag{5.1}$$

Suppose that a function u^+ approximates (5.1) up to $O(\varepsilon^n)$ for some $n > m+3$. More precisely, suppose that u^+ satisfies

$$\varepsilon^2\Delta u^+ + f(u^+) + \varepsilon^{m+3} - u^+_t = O(\varepsilon^n)$$

uniformly in $t \in [0,T^*/\varepsilon^4]$ and $(x,y) \in \Omega$. Then u^+ must be a supersolution for sufficiently small $\varepsilon > 0$, because

$$\varepsilon^2\Delta u^+ + f(u^+) - u^+_t = -\varepsilon^{m+3} + O(\varepsilon^n) < 0.$$

Similarly, we consider the equation

$$u_t = \varepsilon^2 \Delta u + f(u) - \varepsilon^{m+3}, \quad t > 0,\ (x,y) \in \Omega. \tag{5.2}$$

If a function u^- approximates (5.2) up to $O(\varepsilon^n)$ for some $n > m+3$, then u^- becomes a subsolution for sufficiently small $\varepsilon > 0$. Thus, it is sufficient to find approximate functions u^+ and u^-.

Along the above argument, the following proposition gives the direct proof of Proposition 3.5.

PROPOSITION 5.1. *Let $\gamma^{\pm}(T)$ be arbitrary smooth functions of $T \in [0, T^*]$, and set*

$$\eta^{\pm} = \{1 - z\kappa(S^{\pm}(T))\} \left(\frac{s - S^{\pm}(T)}{\varepsilon} - \gamma^{\pm}(T) \right)$$

with $(x,y) = C(s) + z\nu(s)$. Then for small $\varepsilon > 0$, there exist a number $n \in (m+3, 4)$ and functions $u^{\pm}(t,x,y)$ of the form

$$u^{\pm}(t,x,y) = \varphi(\eta^{\pm}) + \varepsilon \Phi^{\pm}(t,z,\eta^{\pm};\varepsilon) \pm M_0 \varepsilon^{m+3} + \varepsilon^4 R^{\pm}(t,x,y;\varepsilon) \tag{5.3}$$

which satisfy (5.1) and (5.2) up to $O(\varepsilon^n)$ uniformly in $t \in [0, T^*/\varepsilon^4]$ and $(x,y) \in \Omega$ together with the Neumann boundary condition, where $\Phi^{\pm}$ satisfy

$$|\Phi^{\pm}(t,z,\eta;\varepsilon)|_{C^2(\eta)} \le A_1 e^{-\beta|\eta|}, \quad (t,z,\eta) \in [0, T^*/\varepsilon^4] \times (-d,d) \times (-\infty, \infty) \tag{5.4}$$

with some positive constants A_1 and $\beta \in (\frac{3}{4}\beta_0, \beta_0)$, and $R^{\pm}(t,x,y;\varepsilon)$ are uniformly bounded in $t \in [0, T^*/\varepsilon^4]$ and $(x,y) \in \Omega$.

In order to prove this proposition, we need some preliminary results.

PROPOSITION 5.2. *For any $\beta \in (0, \beta_0)$ and any nonnegative integer h, the equation*

$$v_{\eta\eta} + f'(\varphi(\eta))v = g(\eta), \quad -\infty < \eta < \infty, \tag{5.5}$$

has a unique solution v^* satisfying $|v^*(\eta)|_{C^{h+2}(\eta)} = O\left(e^{-\beta|\eta|}\right)$ as $|\eta| \to \infty$ and $\langle\, v^*, \varphi_\eta \,\rangle_\eta = 0$ if and only if $g \in C^h$ satisfies $|g(\eta)|_{C^h(\eta)} = O\left(e^{-\beta|\eta|}\right)$ as $|\eta| \to \infty$ and the solvability condition

$$\langle\, g, \varphi_\eta \,\rangle_\eta = 0. \tag{5.6}$$

Moreover, any bounded solution of (5.5) is written as

$$v(\eta) = v^*(\eta) + \gamma \varphi_\eta \tag{5.7}$$

for some constant $\gamma \in \boldsymbol{R}$.

Here we define

$$|v(r,\eta)|_{C^{h,k}(r,\eta)} = \sum_{\substack{0 \le i \le h,\ 0 \le j \le k \\ 0 \le i+j \le \max\{h,k\}}} \left| \frac{\partial^{i+j}}{\partial r^i \eta^j} v(r,\eta) \right|,$$

$$|v(r,\eta)|_{D^{h,k}(r,\eta)} = |v(r,\eta)|_{C^{h,k}(r,\eta)} - |v(r,\eta)|$$

for $v \in C^{h,k}$, the set of functions such that the partial derivatives in the right-hand side are defined and continuous.

PROPOSITION 5.3. *There exists a positive constant b_0 such that for any $b \in (0, b_0)$ and $\beta \in (\frac{3}{4}\beta_0, \beta_0)$, if $g \in C^{0,0}$ satisfies $|g(r,\eta)| \le A_2 \min\{e^{-br}, e^{-\beta|\eta|}\}$ for some A_2 and the solvability condition*

$$\int_0^\infty \langle\, g, \varphi_\eta \,\rangle_\eta \, dr = 0, \tag{5.8}$$

then the equation

$$v_{rr} + v_{\eta\eta} + f'(\varphi(\eta))v = g(r,\eta), \quad 0 < r < \infty,\ -\infty < \eta < \infty, \tag{5.9}$$

has a solution v satisfying $v_r(0,\eta) = 0$ and $|v(r,\eta)|_{C^{2,2}(r,\eta)} \le A_3 \min\{e^{-b'r}, e^{\beta'|\eta|}\}$ for some $b' \in (0,b)$, $\beta' \in (\frac{3}{4}\beta_0, \beta)$ and A_3.

Proposition 5.2 is a well-known fact. A proof of Proposition 5.3 will be given in section 6. By using these propositions, we will prove Proposition 5.1 in the following.

Let us construct u^+. By changing the variables from (x,y) to (z,s), (5.1) is rewritten as

$$v_t = \varepsilon^2 \left\{ v_{zz} - \frac{\kappa}{1-\kappa z} v_z + \frac{1}{1-\kappa z}\left(\frac{1}{1-\kappa z} v_s\right)_s \right\} + f(v) + \varepsilon^{m+3}, \tag{5.10}$$

$$t > 0,\ -d < z < d,\ 0 < s < L,$$

with the boundary conditions

$$v_z(t, \pm d, s) = 0, \tag{5.11}$$

where $\kappa = \kappa(s)$ and $v(t,z,s) = u^+(t, C(s) + z\nu(s))$.

First, we pay our attention to a neighborhood of the inner layer of v. Let $s = S^+$ be the position of the inner layer of u^+ and transform s into μ by $s = S^+ + \varepsilon\mu$. Then (5.10) is rewritten as

$$v_t - \frac{S_t^+}{\varepsilon} v_\mu = \varepsilon^2 \left(v_{zz} - \frac{\kappa}{1-\kappa z} v_z \right) + \frac{1}{1-\kappa z}\left(\frac{1}{1-\kappa z} v_\mu\right)_\mu + f(v) + \varepsilon^{m+3}. \tag{5.12}$$

We solve this equation under the conditions

$$v(t, z, \pm\infty) = \alpha^\pm(\varepsilon^{m+3}). \tag{5.13}$$

Here we assume that v and S^+ are functions of T, where $T = \varepsilon^4 t$, that is, $v = v(T,z,\mu)$ and $S^+ = S^+(T)$. Then (5.12) takes the form

$$\varepsilon^4 v_T = F^\varepsilon(v), \tag{5.14}$$

where

$$F^\varepsilon(v) = \varepsilon^2 \left(v_{zz} - \frac{\kappa}{1-\kappa z} v_z \right) + \frac{1}{1-\kappa z}\left(\frac{1}{1-\kappa z} v_\mu\right)_\mu + f(v) + \varepsilon^{m+3} + \varepsilon^3 S_T^+ v_\mu.$$

We expand $v(T,z,\mu)$ and $F^\varepsilon(v)$ as

$$v(T,z,\mu) = v^0(T,z,\mu) + \varepsilon v^1(T,z,\mu) + \cdots, \tag{5.15}$$

$$F^\varepsilon(v) = F^0(v) + \varepsilon F^1(v) + \cdots. \tag{5.16}$$

Noting that

$$\begin{aligned}\kappa &= \kappa(S^+ + \varepsilon\mu)\\ &= \kappa^+ + \varepsilon\mu\kappa_s^+ + \frac{1}{2}\varepsilon^2\mu^2\kappa_{ss}^+ + \frac{1}{6}\varepsilon^3\mu^3\kappa_{sss}^+ + O((\varepsilon\mu)^4),\end{aligned}$$

where $\kappa^+ = \kappa(S^+(T))$, we have

$$F^0(v) = \frac{1}{(1-z\kappa^+)^2}v_{\mu\mu} + f(v), \tag{5.17}$$

$$F^1(v) = F^1(z,\mu)v = \frac{\kappa_s^+ z}{(1-z\kappa^+)^2}(v_\mu + 2\mu v_{\mu\mu}), \tag{5.18}$$

$$\begin{aligned}F^2(v) &= F^2(z,\mu)v\\ &= v_{zz} - \frac{\kappa^+}{1-z\kappa^+}v_z + \frac{z}{(1-z\kappa^+)^3}\left(\kappa_{ss}^+ + \frac{3z{\kappa_s^+}^2}{1-z\kappa^+}\right)(\mu v_\mu + \mu^2 v_{\mu\mu}),\end{aligned} \tag{5.19}$$

$$F^3(v) = F^3(z,\mu)v + S_T^+ v_\mu + \varepsilon^m, \tag{5.20}$$

where

$$\begin{aligned}F^3(z,\mu)v = &-\frac{\kappa_s^+}{(1-z\kappa^+)^2}\mu v_z\\ &+\frac{z}{(1-z\kappa^+)^3}\left\{\frac{1}{3}\kappa_{sss}^+ + \frac{3z\kappa_s^+\kappa_{ss}^+}{1-z\kappa^+} + \frac{4z^2{\kappa_s^+}^3}{(1-z\kappa^+)^2}\right\}\left(\frac{3}{2}\mu^2 v_\mu + \mu^3 v_{\mu\mu}\right).\end{aligned}$$

Define

$$|\mu|_h = \sum_{j=0}^{h}|\mu|^j.$$

Then it is obvious that

$$\left|F^\varepsilon(v) - \sum_{j=0}^{3}\varepsilon^j F^j(v)\right| \le B\varepsilon^4|\mu|_4|v(z,\mu)|_{D^{1,2}(z,\mu)}. \tag{5.21}$$

Substituting (5.15) into (5.14) and equating coefficients of the same power of ε, we have

$$0 = F^0(v^0), \tag{5.22}$$

$$0 = F_v^0(v^0)v^1 + F^1(z,\mu)v^0, \tag{5.23}$$

$$0 = F_v^0(v^0)v^2 + \frac{1}{2}F_{vv}^0(v^0)v^1\cdot v^1 + F^1(z,\mu)v^1 + F^2(z,\mu)v^0, \tag{5.24}$$

$$\begin{aligned}0 = &F_v^0(v^0)v^3 + F_{vv}^0(v^0)v^1\cdot v^2 + \frac{1}{6}F_{vvv}^0(v^0)(v^1)^3\\ &+F^1(z,\mu)v^2 + F^2(z,\mu)v^1 + F^3(z,\mu)v^0 + S_T^+ v_\mu^0 + \varepsilon^m.\end{aligned} \tag{5.25}$$

We solve these equations to find $v^0,\dots,v^3$. We note that by (5.13) and Proposition 3.3, v^j must satisfy

$$v^0(t,z,\pm\infty) = \pm 1, \tag{5.26}$$

$$v^j(t,z,\pm\infty) = 0 \quad (j=1,2), \tag{5.27}$$

$$v^3(t,z,\pm\infty) = M_0\varepsilon^m. \tag{5.28}$$

First we can obtain v^0 from (5.22), (5.17), and (5.26) as

$$v^0(T,z,\mu)=\varphi((1-z\kappa^+)(\mu-\gamma^+)) \tag{5.29}$$

for a certain function $\gamma^+=\gamma^+(T,z)$. However, this does not satisfy (5.11). This implies that we have to consider boundary layers. We will consider only the boundary layer in the neighborhood of $z=-d$ here, because the neighborhood of $z=d$ is treated in the same manner.

We set $z=-d+\varepsilon r$ near $z=-d$. Then (5.14) is written as

$$\varepsilon^4\tilde{v}_T=\widetilde{F}^\varepsilon(\tilde{v}) \tag{5.30}$$

with the boundary conditions

$$\tilde{v}(T,r,\pm\infty)=\alpha^\pm(\varepsilon^{m+3}),\quad \tilde{v}_r(T,0,\mu)=0, \tag{5.31}$$

where

$$\tilde{v}(T,r,\mu)=v(T,-d+\varepsilon r,\mu) \tag{5.32}$$

and

$$\widetilde{F}^\varepsilon(\tilde{v})=\tilde{v}_{rr}-\varepsilon\frac{\kappa}{1-\kappa z}\tilde{v}_r+\frac{1}{1-\kappa z}\left(\frac{1}{1-\kappa z}\tilde{v}_\mu\right)_\mu+f(\tilde{v})+\varepsilon^{m+3}+\varepsilon^3S_T^+\tilde{v}_\mu.$$

We expand $\tilde{v}(T,r,\mu)$ and $\widetilde{F}^\varepsilon(\tilde{v})$ as

$$\begin{aligned}\tilde{v}(T,r,\mu)&=\tilde{v}^0(T,r,\mu)+\varepsilon\tilde{v}^1(T,r,\mu)+\cdots,\\ \widetilde{F}^\varepsilon(\tilde{v})&=\widetilde{F}^0(\tilde{v})+\varepsilon\widetilde{F}^1(\tilde{v})+\cdots.\end{aligned}$$

Then, by the same way as in the case of F^ε, we have

$$\widetilde{F}^0(\tilde{v})=\tilde{v}_{rr}+\frac{1}{(1+d\kappa^+)^2}\tilde{v}_{\mu\mu}+f(\tilde{v}), \tag{5.33}$$

$$\begin{aligned}\widetilde{F}^1(\tilde{v})&=\widetilde{F}^1(r,\mu)\tilde{v}\\ &=-\frac{\kappa^+}{1+d\kappa^+}\tilde{v}_r+\frac{2\kappa^+}{(1+d\kappa^+)^3}r\tilde{v}_{\mu\mu}-\frac{d\kappa_s^+}{(1+d\kappa^+)^3}(\tilde{v}_\mu+2\mu\tilde{v}_{\mu\mu}),\end{aligned} \tag{5.34}$$

$$\begin{aligned}\widetilde{F}^2(\tilde{v})&=\widetilde{F}^2(r,\mu)\tilde{v}\\ &=-\frac{\kappa_s^+\mu+(\kappa^+)^2r}{(1+d\kappa^+)^2}\tilde{v}_r+\frac{r}{(1+d\kappa^+)^3}\left(\kappa_s^+-\frac{3d\kappa^+\kappa_s^+}{1+d\kappa^+}\right)(2\mu\tilde{v}_{\mu\mu}+\tilde{v}_\mu)\\ &\quad+\frac{3(\kappa^+)^2r^2}{(1+d\kappa^+)^4}\tilde{v}_{\mu\mu}-\frac{d}{(1+d\kappa^+)^3}\left(\kappa_{ss}^+\mu-\frac{3d(\kappa_s^+)^2}{1+d\kappa^+}\right)(\mu^2\tilde{v}_{\mu\mu}+\mu\tilde{v}_\mu).\end{aligned} \tag{5.35}$$

On the other hand, $v(T,-d+\varepsilon r,\mu)$ is expanded as

$$\begin{aligned}&v(T,-d+\varepsilon r,\mu)\\ &=v^0(T,-d+\varepsilon r,\mu)+\varepsilon v^1(T,-d+\varepsilon r,\mu)+\varepsilon^2v^2(T,-d+\varepsilon r,\mu)+\cdots\\ &=v^0(T,-d,\mu)+\varepsilon\left\{v_z^0(T,-d,\mu)r+v^1(T,-d,\mu)\right\}\\ &\qquad+\varepsilon^2\left\{\frac{1}{2}v_{zz}^0(T,-d,\mu)r^2+v_z^1(T,-d,\mu)r+v^2(T,-d,\mu)\right\}\\ &\qquad+\varepsilon^3\left\{\frac{1}{6}v_{zzz}^0(T,-d,\mu)r^3+\frac{1}{2}v_{zz}^1(T,-d,\mu)r^2\right.\\ &\qquad\qquad\left.+v_z^2(T,-d,\mu)r+v^3(T,-d,\mu)\right\}+\cdots,\end{aligned}$$

which must be equal to $\tilde{v}(T,r,\mu)$ by (5.32). Hence we have as $r \to \infty$

$$\tilde{v}^0(T,r,\mu) = \widetilde{W}^0(T,\mu) + o(1), \tag{5.36}$$

$$\tilde{v}^1(T,r,\mu) = \widetilde{W}^1(T,r,\mu) + o(1), \tag{5.37}$$

$$\tilde{v}^2(T,r,\mu) = \widetilde{W}^2(T,r,\mu) + o(1), \tag{5.38}$$

$$\tilde{v}^3(T,r,\mu) = \widetilde{W}^3(T,r,\mu) + o(1), \tag{5.39}$$

where

$$\begin{aligned}
\widetilde{W}^0(T,\mu) &= v^0(T,-d,\mu),\\
\widetilde{W}^1(T,r,\mu) &= v^0_z(T,-d,\mu)r + v^1(T,-d,\mu),\\
\widetilde{W}^2(T,r,\mu) &= \frac{1}{2}v^0_{zz}(T,-d,\mu)r^2 + v^1_z(T,-d,\mu)r + v^2(T,-d,\mu),\\
\widetilde{W}^3(T,r,\mu) &= \frac{1}{6}v^0_{zzz}(T,-d,\mu)r^3 + \frac{1}{2}v^1_{zz}(T,-d,\mu)r^2\\
&\qquad + v^2_z(T,-d,\mu)r + v^3(T,-d,\mu).
\end{aligned}$$

We will find $\tilde{v}^j$ $(j = 0,\dots,3)$ which satisfy (5.36)–(5.39) up to the second derivative with respect to r.

Similar to (5.22)–(5.28), the following equalities must hold:

$$0 = \widetilde{F}^0(\tilde{v}^0), \tag{5.40}$$

$$0 = \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v}^1 + \widetilde{F}^1(z,\mu)\tilde{v}^0, \tag{5.41}$$

$$0 = \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v}^2 + \frac{1}{2}\widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\tilde{v}^1\cdot\tilde{v}^1 + \widetilde{F}^1(z,\mu)\tilde{v}^1 + \widetilde{F}^2(z,\mu)\tilde{v}^0, \tag{5.42}$$

$$\begin{aligned}
0 = {}& \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v}^3 + \widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\tilde{v}^1\cdot\tilde{v}^2 + \frac{1}{6}\widetilde{F}^0_{\tilde{v}\tilde{v}\tilde{v}}(\tilde{v}^0)(\tilde{v}^1)^3\\
&+\widetilde{F}^1(z,\mu)\tilde{v}^2 + \widetilde{F}^2(z,\mu)\tilde{v}^1 + \widetilde{F}^3(z,\mu)\tilde{v}^0 + S^+_T\tilde{v}^0_\mu + \varepsilon^m,
\end{aligned} \tag{5.43}$$

and

$$\begin{aligned}
\tilde{v}^0(t,z,\pm\infty) &= \pm 1,\\
\tilde{v}^j(t,z,\pm\infty) &= 0 \ (j=1,2),\\
\tilde{v}^3(t,z,\pm\infty) &= M_0\varepsilon^m.
\end{aligned} \tag{5.44}$$

We remark here that (5.40)–(5.43) hold when $\tilde{v}^j = \widetilde{W}^j$ $(j = 0,1,2,3)$.

By (5.29), (5.33) and (5.36), (5.40), (5.44), we immediately have

$$\begin{aligned}
\tilde{v}^0(T,r,\mu) &= \widetilde{W}^0(T,\mu)\\
&= \varphi((1+d\kappa^+)(\mu-\gamma^+(T,-d))),
\end{aligned} \tag{5.45}$$

which is independent of r.

5.2. Constructions of v^j and $\tilde{v}^j$ $(j = 1,2,3)$. First, let us consider v^1 and $\tilde{v}^1$.

LEMMA 5.4. *There exists a function v^1 satisfying* (5.23) *and* (5.27).

Proof. By (5.29) and the transformation $\eta = (1-z\kappa^+)(\mu-\gamma^+)$, (5.23) is equivalent to

$$0 = v^1_{\eta\eta} + f'(\varphi(\eta))v^1 + g^1, \tag{5.46}$$

where

$$g^1 = \frac{z\kappa_s^+}{(1-z\kappa^+)^2}\left\{2\eta\varphi_{\eta\eta} + \varphi_\eta + 2\gamma^+(1-z\kappa^+)\varphi_{\eta\eta}\right\}.$$

Since we can easily see that $|g^1(\eta)|_{C^3(\eta)} \le Be^{-\beta_1|\eta|}$ for constants $\beta_1 \in (\frac{3}{4}\beta_0, \beta_0)$ and B and that the solvability condition

$$\langle\, g^1, \varphi_\eta \,\rangle_\eta = 0$$

holds, there exists a unique function $V^1(T,z,\eta)$ satisfying (5.46),

$$\langle\, V^1, \varphi_\eta \,\rangle_\eta = 0, \tag{5.47}$$

and

$$|V^1(T,z,\eta)|_{C^5(\eta)} \le Be^{-\beta_1|\eta|} \tag{5.48}$$

by Proposition 5.2 and the smoothness of φ. Here we used the oddness of φ and

$$\langle\, \eta\varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta = -\frac{1}{2}M_1. \tag{5.49}$$

Hence, from (5.7), v^1 is given by

$$\begin{aligned} v^1(T,z,\mu) = V^1(T,z,(1-z\kappa^+)(\mu-\gamma^+(T,z)) \\ +\gamma^1(T,z)\varphi_\eta((1-z\kappa^+)(\mu-\gamma^+(T,z))) \end{aligned} \tag{5.50}$$

for some function γ^1. □

In general, we may assume

$$|V^1(T,z,\eta)|_{C^{5,5}(z,\eta)} \le Be^{-\beta_1|\eta|}. \tag{5.51}$$

LEMMA 5.5. *There exists a function $\tilde{v}^1$ satisfying* (5.31), (5.37) *and* (5.41), (5.44) *if and only if*

$$\gamma_z^+(T,-d) = 0. \tag{5.52}$$

Proof. In this proof, we simply express $V^1(T,-d,\eta)$, $\gamma^+(T,-d)$, and $\gamma_z^+(T,-d)$ as $V^1(\eta)$, γ^+, and γ_z^+, respectively.

By (5.34), (5.41), and (5.45), the equality (5.41) is represented as

$$0 = \tilde{v}^1_{rr} + \tilde{v}^1_{\eta\eta} + f'(\varphi)\tilde{v}^1 + \tilde{h}^1, \tag{5.53}$$

where $\eta = (1+d\kappa^+)(\mu-\gamma^+)$ and

$$\tilde{h}^1 = \frac{2\kappa^+}{(1+d\kappa^+)}r\varphi_{\eta\eta} - \frac{d\kappa_s^+}{(1+d\kappa^+)^2}\left\{\varphi_\eta + 2\eta\varphi_{\eta\eta} + 2\gamma^+(1+d\kappa^+)\varphi_{\eta\eta}\right\}.$$

Now it follows from (5.29) and (5.50) that

$$\widetilde{W}^1(T,r,\mu) = -\left\{\frac{\kappa^+}{1+d\kappa^+}\eta + (1+d\kappa^+)\gamma_z^+\right\}\varphi_\eta r + V^1(\eta) + \gamma^1\varphi_\eta,$$

where $\gamma^1 = \gamma^1(T, -d)$. Denote the right-hand side of this equation by $W^1 = W^1(T, r, \eta)$ and set $w^1 = \tilde{v}^1 - W^1$. Then w^1 satisfies

$$\begin{cases} w^1_{rr} + w^1_{\eta\eta} + f'(\varphi)w^1 = 0, \\ w^1_r(T, 0, \eta) = \left(\dfrac{\kappa^+}{1 + d\kappa^+}\eta + (1 + d\kappa^+)\gamma^+_z \right) \varphi_\eta, \\ w^1(T, \infty, \eta) = 0, \end{cases} \tag{5.54}$$

because $V^1(\eta)$, φ_η, respectively, satisfy (5.46) at $z = -d$ and the equation

$$0 = (\varphi_\eta)_{\eta\eta} + f'(\varphi)\varphi_\eta$$

and $U^{1,1} = -\eta\varphi_\eta$ satisfies

$$0 = U^{1,1}_{\eta\eta} + f'(\varphi)U^{1,1} + 2\varphi_{\eta\eta}. \tag{5.55}$$

Let $\chi(r)$ be a smooth function satisfying

$$\chi_r(0) = 1, \quad |\chi(r)|_{C^2(r)} \le Be^{-b_0 r},$$

and let

$$\tilde{w}^1 = w^1 - \chi(r) \left(\frac{\kappa^+}{1 + d\kappa^+}\eta + (1 + d\kappa^+)\gamma^+_z(T, -d) \right) \varphi_\eta.$$

Then, by (5.55), the equation for $\tilde{w}^1$ becomes

$$\begin{aligned} &\tilde{w}^1_{rr} + \tilde{w}^1_{\eta\eta} + f'(\varphi)\tilde{w}^1 + \tilde{g}^1 = 0, \\ &\tilde{w}^1_r(T, 0, \eta) = 0, \end{aligned} \tag{5.56}$$

where

$$\tilde{g}^1 = \frac{\kappa^+}{1 + d\kappa^+} \left\{ -\chi_{rr}U^{1,1} + 2\chi\varphi_{\eta\eta} \right\} + \chi_{rr}(1 + d\kappa^+)\gamma^+_z\varphi_\eta.$$

Since $|\tilde{g}^1(T, r, \eta)| \le O\left(e^{-(b_0 r + \beta_1|\eta|)}\right)$, Proposition 5.3 implies that there exists a function $\tilde{w}^1$ satisfying (5.56) and $|\tilde{w}^1|_{C^{2,2}(r,\eta)} \le O(\min\{e^{-b_1 r}, e^{-\beta_2|\eta|}\})$ for $b_1 \in (0, b_0)$ and $\beta_2 \in (\frac{3}{4}\beta_0, \beta_1)$ if and only if

$$\begin{aligned} 0 &= \int_0^\infty \langle\, \tilde{g}^1, \varphi_\eta \,\rangle_\eta \, dr \\ &= \int_0^\infty \chi_{rr}(1 + d\kappa^+)\gamma^+_z \langle\, \varphi_\eta, \varphi_\eta \,\rangle_\eta \, dr \\ &= M_1(1 + d\kappa^+)\gamma^+_z[\chi_r]_0^\infty \\ &= -M_1(1 + d\kappa^+)\gamma^+_z. \end{aligned}$$

This is equivalent to (5.52). Now, we set

$$\widetilde{V}^1 = \tilde{w}^1 + \chi(r) \left(\frac{\kappa^+}{1 + d\kappa^+}\eta + (1 + d\kappa^+)\gamma^+_z \right) \varphi_\eta + W^1. \tag{5.57}$$

Then, by (5.52) and the definition of W^1, we have

$$\begin{aligned} \widetilde{V}^1 &= \tilde{w}^1 + \chi(r)\frac{\kappa^+}{1 + d\kappa^+}\eta\varphi_\eta + W^1 \\ &= \tilde{w}^1 + \frac{(\chi - r)\kappa^+}{1 + d\kappa^+}\eta\varphi_\eta + V^1(\eta) + \gamma^1\varphi_\eta. \end{aligned}$$

Thus the function $\widetilde{V}^1$ satisfies (5.53) and

$$|\widetilde{V}^1(T,r,\eta) - W^1(T,r,\eta)|_{C^{2,2}(r,\eta)} \le B\min\{e^{-b_1 r}, e^{-\beta_2|\eta|}\}, \tag{5.58}$$

and $\tilde{v}^1$ is given by

$$\tilde{v}^1(T,r,\mu) = \widetilde{V}^1(T,r,(1+d\kappa^+)(\mu-\gamma^+)). \qquad \square \tag{5.59}$$

In quite a similar way, we can show that

$$\gamma_z^+(T,d) = 0 \tag{5.60}$$

is a necessary and sufficient condition for the existence of $\tilde{v}^1$ in a neighborhood of $z = d$.

Next we consider (5.24) in order to obtain v^2. A condition so that v^2 exists in (5.24) is given by

$$0 = \left\langle \frac{1}{2}F^0_{vv}(v^0)v^1\cdot v^1 + F^1(z,\mu)v^1 + F^2(z,\mu)v^0, v^0_\mu \right\rangle_\mu, \tag{5.61}$$

which is easily found by Proposition 5.2 and the transformation of μ into $\eta = (1 - z\kappa^+)(\mu - \gamma^+(T,z))$.

LEMMA 5.6. *The equality*

$$0 = \left\langle \frac{1}{2}F^0_{vv}(v^0)v^1\cdot v^1 + F^1(z,\mu)v^1, v^0_\mu \right\rangle_\mu \tag{5.62}$$

holds.

Proof. By differentiating (5.23) with respect to μ, we have

$$0 = F^0_{vv}(v^0)v^0_\mu\cdot v^1 + F^0_v(v^0)v^1_\mu + \frac{\partial}{\partial\mu}F^1v^0. \tag{5.63}$$

Hence, noting that $F^0_v(v^0)$ is self-adjoint, we see

$$\begin{aligned} 0 &= \left\langle F^0_{vv}(v^0)v^0_\mu\cdot v^1 + F^0_v(v^0)v^1_\mu + \frac{\partial}{\partial\mu}F^1v^0, v^1 \right\rangle_\mu \\ &= \langle\, F^0_v(v^0)v^1 - F^1v^0, v^1_\mu \,\rangle_\mu + \langle\, F^0_{vv}(v^0)v^1\cdot v^1, v^0_\mu \,\rangle_\mu \\ &= -2\langle\, F^1v^0, v^1_\mu \,\rangle_\mu + \langle\, F^0_{vv}(v^0)v^1\cdot v^1, v^0_\mu \,\rangle_\mu . \end{aligned} \tag{5.64}$$

Substituting (5.64) into the right-hand side of (5.62), we have

$$\left\langle \frac{1}{2}F^0_{vv}(v^0)v^1\cdot v^1 + F^1v^1, v^0_\mu \right\rangle_\mu = \langle\, F^1v^0, v^1_\mu \,\rangle_\mu + \langle\, F^1v^1, v^0_\mu \,\rangle_\mu . \tag{5.65}$$

Here it follows in general that, by a simple calculation,

$$\langle\, F^1v, w_\mu \,\rangle_\mu = -\langle\, F^1w, v_\mu \,\rangle_\mu . \tag{5.66}$$

From (5.65) and (5.66), the proof of this lemma is complete. $\square$

Lemma 5.6 implies that (5.61) is equivalent to

$$0 = \langle\, F^2(z,\mu)v^0, v^0_\mu \,\rangle_\mu . \tag{5.67}$$

LEMMA 5.7. *The equality* (5.67) *holds if and only if*

$$\gamma^+ = \gamma^+(T) \tag{5.68}$$

is independent of z.

Proof. From (5.29), we have

$$v_z^0 = -\left\{\frac{\kappa^+}{1-z\kappa^+}\eta + (1-z\kappa^+)\gamma_z^+\right\}\varphi_\eta,$$
$$v_{zz}^0 = \left\{\frac{\kappa^+}{1-z\kappa^+}\eta + (1-z\kappa^+)\gamma_z^+\right\}^2 \varphi_{\eta\eta} + \{2\kappa^+\gamma_z^+ - (1-z\kappa^+)\gamma_{zz}^+\}\varphi_\eta.$$

Noting

$$\langle\, \eta^2\varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta = 0,$$
$$\langle\, \varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta = 0,$$
$$\langle\, \eta\varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta = -\frac{1}{2}M_1,$$

we see

$$\begin{aligned}
&\left\langle v_{zz}^0 - \frac{\kappa^+}{1-z\kappa^+}v_z^0, v_\mu^0\right\rangle_\mu \qquad (5.69)\\
&= \langle\, v_{zz}^0, v_\mu^0 \,\rangle_\mu - \frac{\kappa^+}{1-z\kappa^+}\langle\, v_z^0, v_\mu^0 \,\rangle_\mu\\
&= \left(\frac{\kappa^+}{1-z\kappa^+}\right)^2 \langle\, \eta^2\varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta + 2\kappa^+\gamma_z^+ \langle\, \eta\varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta\\
&\qquad +(1-z\kappa^+)^2(\gamma_z^+)^2 \langle\, \varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta\\
&\qquad +\{2\kappa^+\gamma_z^+ - (1-z\kappa^+)\gamma_{zz}^+\}\langle\, \varphi_\eta, \varphi_\eta \,\rangle_\eta\\
&\qquad -\frac{\kappa^+}{1-z\kappa^+}\left(-\frac{\kappa^+}{1-z\kappa^+}\langle\, \eta\varphi_\eta, \varphi_\eta \,\rangle_\eta - (1-z\kappa^+)\gamma_z^+\langle\, \varphi_\eta, \varphi_\eta \,\rangle_\eta\right)\\
&= -M_1\kappa^+\gamma_z^+ + M_1\{2\kappa^+\gamma_z^+ - (1-z\kappa^+)\gamma_{zz}^+\}\\
&\qquad -\frac{\kappa^+}{1-z\kappa^+}\cdot\{-M_1(1-z\kappa^+)\gamma_z^+\}\\
&= 2M_1\kappa^+\gamma_z^+ - M_1(1-z\kappa^+)\gamma_{zz}^+.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&\langle\, \mu v_\mu^0 + \mu^2 v_{\mu\mu}^0, v_\mu^0 \,\rangle_\mu \qquad (5.70)\\
&= \left\langle\left(\frac{\eta}{1-z\kappa^+} + \gamma^+\right)(1-z\kappa^+)\varphi_\eta, \varphi_\eta\right\rangle_\eta\\
&\qquad +\left\langle\left(\frac{\eta}{1-z\kappa^+} + \gamma^+\right)^2(1-z\kappa^+)^2\varphi_{\eta\eta}, \varphi_\eta\right\rangle_\eta\\
&= \langle\, \eta\varphi_\eta, \varphi_\eta \,\rangle_\eta + (1-z\kappa^+)\gamma^+M_1 + \langle\, \eta^2\varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta\\
&\qquad +2\gamma^+(1-z\kappa^+)\langle\, \eta\varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta + (\gamma^+)^2(1-z\kappa^+)^2\langle\, \varphi_{\eta\eta}, \varphi_\eta \,\rangle_\eta\\
&= (1-z\kappa^+)\gamma^+M_1 - (1-z\kappa^+)\gamma^+M_1\\
&= 0.
\end{aligned}$$

By (5.69), (5.70), and (5.67), $\gamma^+(T,z)$ must satisfy

$$\gamma^+_{zz} - \frac{2\kappa^+}{1-z\kappa^+}\gamma^+_z = 0, \quad -d < z < d, \tag{5.71}$$

in order that (5.67) holds. Thus, by (5.71) with (5.52), (5.60), we obtain (5.68). □

By Lemma 5.7, we can show the existence of v^2 satisfying (5.24) if and only if (5.68) holds.

Next, let us consider $\tilde{v}^2$ of (5.42).

LEMMA 5.8. *The equality*

$$\left\langle \frac{1}{2}\widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\tilde{v}^1\cdot\tilde{v}^1 + \widetilde{F}^1\tilde{v}^1, \tilde{v}^0_\mu \right\rangle_\mu = -\frac{\kappa^+}{1+d\kappa^+}\langle\, \tilde{v}^1_r, \tilde{v}^0_\mu \,\rangle_\mu + \langle\, \tilde{v}^1_{rr}, \tilde{v}^1_\mu \,\rangle_\mu \tag{5.72}$$

holds.

Proof. By differentiating (5.41) with respect to μ, we have

$$\widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v}^1_\mu + \widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\tilde{v}^0_\mu\cdot\tilde{v}^1 + \frac{\partial}{\partial\mu}(\widetilde{F}^1\tilde{v}^0) = 0,$$

and taking an inner product with $\tilde{v}^1$, we also have

$$\langle\, \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v}^1_\mu, \tilde{v}^1 \,\rangle_\mu + \langle\, \widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\tilde{v}^1\cdot\tilde{v}^1, \tilde{v}^0_\mu \,\rangle_\mu - \langle\, \widetilde{F}^1\tilde{v}^0, \tilde{v}^1_\mu \,\rangle_\mu = 0. \tag{5.73}$$

Since

$$\widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v} = \tilde{v}_{rr} + F^0_v(\tilde{v}^0)\tilde{v} \tag{5.74}$$

and $F^0_v(\tilde{v}^0)$ is self-adjoint in $L^2(\boldsymbol{R}^1)$, it follows that

$$\begin{aligned}\langle\, \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v}^1_\mu, \tilde{v}^1 \,\rangle_\mu &= \langle\, (\tilde{v}^1_\mu)_{rr}, \tilde{v}^1 \,\rangle_\mu + \langle\, F^0_v(\tilde{v}^0)\tilde{v}^1_\mu, \tilde{v}^1 \,\rangle_\mu \\ &= -\langle\, \tilde{v}^1_{rr}, \tilde{v}^1_\mu \,\rangle_\mu + \langle\, \tilde{v}^1_\mu, F^0_v(\tilde{v}^0)\tilde{v}^1 \,\rangle_\mu \\ &= \langle\, \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{v}^1, \tilde{v}^1_\mu \,\rangle_\mu - 2\langle\, \tilde{v}^1_{rr}, \tilde{v}^1_\mu \,\rangle_\mu .\end{aligned} \tag{5.75}$$

On the other hand, we have

$$\begin{aligned}&\langle\, \widetilde{F}^1\tilde{v}^0, \tilde{v}^1_\mu \,\rangle_\mu \\ &= -\frac{2\kappa^+ r}{(1+d\kappa^+)^3}\langle\, \tilde{v}^1_{\mu\mu}, \tilde{v}^0_\mu \,\rangle_\mu + \frac{d\kappa^+_s}{(1+d\kappa^+)^3}\langle\, \tilde{v}^1_\mu + 2\mu\tilde{v}^1_{\mu\mu}, \tilde{v}^0_\mu \,\rangle_\mu \\ &= -\langle\, \widetilde{F}^1\tilde{v}^1, \tilde{v}^0_\mu \,\rangle_\mu - \frac{\kappa^+}{1+d\kappa^+}\langle\, \tilde{v}^1_r, \tilde{v}^0_\mu \,\rangle_\mu\end{aligned} \tag{5.76}$$

by (5.34), (5.45), and the same equality as (5.66). Substituting (5.75) and (5.76) into (5.73) and using (5.41), we complete the proof. □

LEMMA 5.9. *The unique solution V^1 of* (5.46) *with* (5.47) *is explicitly obtained as*

$$V^1(T,z,\eta) = \zeta^{1,1}(T,z)U^{1,1}(\eta) + \zeta^{1,2}(T,z)U^{1,2}(\eta), \tag{5.77}$$

where

$$\begin{aligned}\zeta^{1,1}(T,z) &= \frac{\gamma^+(T)\kappa^+_s z}{1-z\kappa^+}, \\ U^{1,1}(\eta) &= -\eta\varphi_\eta(\eta), \\ \zeta^{1,2}(T,z) &= \frac{z\kappa^+_s}{(1-z\kappa^+)^2}, \\ U^{1,2}(\eta) &= -\frac{1}{2}\eta^2\varphi_\eta(\eta) + \frac{M_2}{2M_1}\varphi_\eta(\eta)\end{aligned}$$

and $M_2 = \langle\, \eta\varphi_\eta, \eta\varphi_\eta \,\rangle_\eta$.

Proof. We have equalities

$$\langle\, U^{1,1}, \varphi_\eta \,\rangle_\eta = \langle\, U^{1,2}, \varphi_\eta \,\rangle_\eta = 0.$$

Moreover, $U^{1,1}$ and $U^{1,2}$ satisfy (5.55) and

$$U^{1,2}_{\eta\eta} + F'(\varphi)U^{1,2} + \varphi_\eta + 2\eta\varphi_\eta = 0.$$

Thus the proof is complete. □

LEMMA 5.10. *There exists a function $\tilde{v}^2$ satisfying* (5.38), (5.42), *and* (5.44) *if and only if*

$$\gamma^1_z(T,-d) + \frac{\kappa^+}{1+d\kappa^+}\gamma^1(T,-d) + \frac{M_2\kappa^+}{M_1(1+d\kappa^+)}\zeta^{1,2}(T,-d) = 0. \tag{5.78}$$

Proof. Let

$$\tilde{h}^2 = \frac{1}{2}\widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\tilde{v}^1\cdot\tilde{v}^1 + \widetilde{F}^1\tilde{v}^1 + \widetilde{F}^2\tilde{v}^0$$

and $w^2 = \tilde{v}^2 - \widetilde{W}^2$. Then w^2 satisfies

$$\begin{aligned} &0 = \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)w^2 + \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\widetilde{W}^2 + \tilde{h}^2,\\ &w^2_r(T,0,\mu) = -v^1_z(T,-d,\mu),\\ &w^2(T,r,\mu)\to 0 \ \text{ as } \ |\mu|,\ r\to\infty, \end{aligned} \tag{5.79}$$

because $\widetilde{W}^2_r(T,0,\mu) = v^1_z(T,-d,\mu)$. Putting

$$\tilde{w}^2 = w^2 + \chi(r)v^1_z$$

and

$$\tilde{g}^2 = -\widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\chi v^1_z + \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\widetilde{W}^2 + \tilde{h}^2,$$

where $v^1_z = v^1_z(T,-d,\mu)$, we get

$$\begin{aligned} &0 = \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\tilde{w}^2 + \tilde{g}^2,\\ &\tilde{w}^2_r(T,0,\mu) = 0,\\ &w^2(T,r,\mu)\to 0 \ \text{ as } \ |\mu|,\ r\to\infty. \end{aligned} \tag{5.80}$$

Here we have

$$\widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\widetilde{W}^2 + \frac{1}{2}\widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\widetilde{W}^1\cdot\widetilde{W}^1 + \widetilde{F}^1\widetilde{W}^1 + \widetilde{F}^2\widetilde{W}^0 = 0. \tag{5.81}$$

Moreover (5.36) and (5.37) hold with the exponential order $O(e^{-b_1 r})$ from (5.45), (5.58), and (5.59). Hence

$$|\tilde{g}^2| \le O(\min\{e^{-b_1 r}, e^{-\beta_2|\eta|}\}),$$

where $\eta = (1+d\kappa^+)(\mu-\gamma^+)$. Therefore, by Proposition 5.3, there exists $\tilde{w}^2$ of (5.80) if and only if

$$\int_0^\infty \langle\, \tilde{g}^2, \tilde{v}^0_\mu \,\rangle_\mu \, dr = 0. \tag{5.82}$$

Here we note that $\langle\, \tilde{g}^2, \varphi_\eta \,\rangle_\eta = \langle\, \tilde{g}^2, \tilde{v}^0_\mu \,\rangle_\mu$ with $\eta = (1+d\kappa^+)(\mu-\gamma^+)$.

We will calculate (5.82). From (5.74), we have

$$\tilde{g}^2 = -\chi_{rr} v_z^1 - \chi F_v^0(\tilde{v}^0) v_z^1 + \widetilde{F}_{\tilde{v}}^0(\tilde{v}^0)\widetilde{W}^2 + \tilde{h}^2. \tag{5.83}$$

Since

$$\begin{aligned} \langle\, F_v^0(\tilde{v}^0) v_z^1, \tilde{v}_\mu^0 \,\rangle_\mu &= \langle\, F_v^0(\widetilde{W}^0) v_z^1, \widetilde{W}_\mu^0 \,\rangle_\mu \\ &= \langle\, v_z^1, F_v^0(\widetilde{W}^0)\widetilde{W}_\mu^0 \,\rangle_\mu \\ &= 0, \end{aligned}$$

it suffices to calculate

$$-\int_0^\infty \chi_{rr} \langle\, v_z^1, \tilde{v}_\mu^0 \,\rangle_\mu \, dr \tag{5.84}$$

and

$$\int_0^\infty \langle\, \widetilde{F}_{\tilde{v}}^0(\tilde{v}^0)\widetilde{W}^2 + \tilde{h}^2, \tilde{v}_\mu^0 \,\rangle_\mu \, dr. \tag{5.85}$$

In (5.84), $v_z^1 = v_z^1(T, -d, \mu)$ and $\tilde{v}_\mu^0$ are independent of r. Hence we have

$$\begin{aligned} -\int_0^\infty \chi_{rr} \langle\, v_z^1, \tilde{v}_\mu^0 \,\rangle_\mu \, dr &= -[\chi_r]_0^\infty \langle\, v_z^1, \tilde{v}_\mu^0 \,\rangle_\mu \\ &= \langle\, v_z^1, \tilde{v}_\mu^0 \,\rangle_\mu . \end{aligned} \tag{5.86}$$

On the other hand, (5.49) and (5.50) imply

$$\begin{aligned} &\langle\, v_z^1, \tilde{v}_\mu^0 \,\rangle_\mu \\ &= \left\langle V_z^1 - \frac{\kappa^+}{1+d\kappa^+}\eta V_\eta^1 + \gamma_z^1 \varphi_\eta - \gamma^1 \frac{\kappa^+}{1+d\kappa^+}\eta\varphi_{\eta\eta}, \varphi_\eta \right\rangle_\eta \\ &= M_1\gamma_z^1 + \langle\, V_z^1, \varphi_\eta \,\rangle_\eta - \frac{\kappa^+}{1+d\kappa^+}\left(\langle\, \eta V_\eta^1, \varphi_\eta \,\rangle_\eta - \frac{1}{2} M_1 \gamma^1 \right) \end{aligned} \tag{5.87}$$

at $z = -d$. Here, by Lemma 5.9, we have

$$\begin{aligned} \langle\, \eta V_\eta^1, \varphi_\eta \,\rangle_\eta &= \zeta^{1,2} \langle\, \eta U_\eta^{1,2}, \varphi_\eta \,\rangle_\eta \\ &= -\frac{1}{2} M_2 \zeta^{1,2}, \end{aligned}$$

because $U^{1,1}(\eta)$ and $U^{1,2}(\eta)$ are, respectively, odd and even. Hence

$$\begin{aligned} &\langle\, v_z^1, \tilde{v}_\mu^0 \,\rangle_\mu \\ &= M_1\gamma_z^1(T,-d) - \frac{\kappa^+}{1+d\kappa^+}\left(-\frac{1}{2} M_2\zeta^{1,2}(T,-d) - \frac{1}{2} M_1\gamma^1(T,-d) \right) \\ &= M_1\gamma_z^1(T,-d) + \frac{\kappa^+}{2(1+d\kappa^+)}\left(M_2\zeta^{1,2}(T,-d) + M_1\gamma^1(T,-d) \right). \end{aligned} \tag{5.88}$$

Next we calculate (5.85). From Lemma 5.8, we have

$$\begin{aligned} &\left\langle \frac{1}{2}\widetilde{F}_{\tilde{v}\tilde{v}}^0(\widetilde{W}^0)\widetilde{W}^1 \cdot \widetilde{W}^1 + \widetilde{F}^1\widetilde{W}^1, \widetilde{W}_\mu^0 \right\rangle_\mu \\ &= -\frac{\kappa^+}{1+d\kappa^+} \langle\, \widetilde{W}_r^1, \widetilde{W}_\mu^0 \,\rangle_\mu + \langle\, \widetilde{W}_{rr}^1, \widetilde{W}_\mu^1 \,\rangle_\mu , \end{aligned} \tag{5.89}$$

since $\widetilde{W}^j$ $(j=0,1,2)$ satisfy the same equations as (5.40) $\sim$ (5.42) and $\tilde{v}^0=\widetilde{W}^0$. In (5.89), it is obvious that

$$\begin{aligned}\langle\, \widetilde{W}^1_r, \widetilde{W}^0_\mu \,\rangle_\mu &= \langle\, v^0_z(T,-d,\mu), v^0_\mu(T,-d,\mu)\,\rangle_\mu \\ &= -\frac{\kappa^+}{1+d\kappa^+}\,\langle\, \eta\varphi_\eta, \varphi_\eta\,\rangle_\eta \\ &= 0,\end{aligned} \tag{5.90}$$

$$\langle\, \widetilde{W}^1_{rr}, \widetilde{W}^1_\mu \,\rangle_\mu = 0. \tag{5.91}$$

Hence, by (5.89), (5.90), and (5.91), we have

$$\left\langle \frac{1}{2}\widetilde{F}^0_{\tilde{v}\tilde{v}}(\widetilde{W}^0)\widetilde{W}^1\cdot\widetilde{W}^1 + \widetilde{F}^1\widetilde{W}^1, \widetilde{W}^0_\mu \right\rangle_\mu = 0. \tag{5.92}$$

In order to use (5.72), we calculate the right-hand side of (5.72). By (5.59), we have

$$\langle\, \tilde{v}^1_r, \tilde{v}^0_\mu \,\rangle_\mu = \langle\, \widetilde{V}^1_r, \varphi_\eta\,\rangle_\eta\,, \tag{5.93}$$

$$\langle\, \tilde{v}^1_{rr}, \tilde{v}^1_\mu \,\rangle_\mu = \langle\, \widetilde{V}^1_{rr}, \widetilde{V}^1_\eta\,\rangle_\eta\,. \tag{5.94}$$

Since $\tilde{w}^1$ is odd with respect to η due to (5.52) and (5.56), it follows from (5.57) that

$$\begin{aligned}\langle\, \widetilde{V}^1_r, \varphi_\eta\,\rangle_\eta &= \langle\, \tilde{w}^1_r, \varphi_\eta\,\rangle_\eta - (\chi_r+1)\frac{\kappa^+}{1+d\kappa^+}\,\langle\, \eta\varphi_\eta, \varphi_\eta\,\rangle_\eta \\ &= 0.\end{aligned} \tag{5.95}$$

Therefore, (5.93) and (5.95) imply

$$\langle\, \tilde{v}^1_r, \tilde{v}^0_\mu \,\rangle_\mu = 0. \tag{5.96}$$

Similarly, (5.94) becomes

$$\begin{aligned}&\langle\, \widetilde{V}^1_{rr}, \widetilde{V}^1_\eta\,\rangle_\eta \\ &= \left\langle \tilde{w}^1_{rr} + \frac{\kappa^+}{1+d\kappa^+}\chi_{rr}\eta\varphi_\eta\ ,\ \tilde{w}^1_\eta + \frac{\kappa^+}{1+d\kappa^+}(\chi-r)(\eta\varphi_\eta)_\eta + V^1_\eta + \gamma^1\varphi_{\eta\eta}\right\rangle_\eta \\ &= \langle\, \tilde{w}^1_{rr}, V^1_\eta\,\rangle_\eta + \frac{\kappa^+}{1+d\kappa^+}\chi_{rr}\,\langle\, \eta\varphi_\eta, V^1_\eta + \gamma^1\varphi_{\eta\eta}\,\rangle_\eta \\ &= \langle\, \tilde{w}^1_{rr}, V^1_\eta\,\rangle_\eta - \frac{\kappa^+}{2(1+d\kappa^+)}\chi_{rr}(M_2\zeta^{1,2}+M_1\gamma^1),\end{aligned} \tag{5.97}$$

where $V^1=V^1(T,-d,\eta)$, $\zeta^{1,2}=\zeta^{1,2}(T,-d)$, and $\gamma^1=\gamma^1(T,-d)$. Since

$$\langle\, \widetilde{F}^2\tilde{v}^0, \tilde{v}^0_\mu\,\rangle_\mu = 0$$

by (5.35), (5.45), (5.49), and the oddness of φ, it follows from (5.72), (5.92) and (5.96), (5.97) that

$$\begin{aligned}&\langle\, \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\widetilde{W}^2 + \tilde{h}^2, \tilde{v}^0_\mu\,\rangle_\mu \\ &= \langle\, \widetilde{F}^0_{\tilde{v}}(\tilde{v}^0)\widetilde{W}^2\ ,\ \tilde{v}^0_\mu\,\rangle_\mu + \left\langle \frac{1}{2}\widetilde{F}^0_{\tilde{v}\tilde{v}}(\tilde{v}^0)\tilde{v}^1\cdot\tilde{v}^1 + \widetilde{F}^1\tilde{v}^1\ ,\ \tilde{v}^0_\mu\right\rangle_\mu \\ &= -\left\langle \frac{1}{2}\widetilde{F}^0_{\tilde{v}\tilde{v}}(\widetilde{W}^0)\widetilde{W}^1\cdot\widetilde{W}^1 + \widetilde{F}^1\widetilde{W}^1\ ,\ \tilde{v}^0_\mu\right\rangle_\mu \\ &\qquad + \langle\, \tilde{w}^1_{rr}, V^1_\eta\,\rangle_\eta - \frac{\kappa^+}{2(1+d\kappa^+)}\chi_{rr}(M_2\zeta^{1,2}+M_1\gamma^1).\end{aligned}$$

Therefore, (5.85) is equal to

$$\begin{aligned}(5.98)\qquad &\int_0^\infty \left(\langle\, \tilde{w}^1_{rr}, V^1_\eta \,\rangle_\eta - \frac{\kappa^+}{2(1+d\kappa^+)}\chi_{rr}(M_2\zeta^{1,2}+M_1\gamma^1)\right) dr\\ &= \langle\, \tilde{w}^1_r(T,\infty,\eta) - \tilde{w}^1_r(T,0,\eta), V^1_\eta \,\rangle_\eta\\ &\qquad -\frac{\kappa^+}{2(1+d\kappa^+)}(M_2\zeta^{1,2}+M_1\gamma^1)(\chi_r(\infty)-\chi_r(0))\\ &= \frac{\kappa^+}{2(1+d\kappa^+)}(M_2\zeta^{1,2}+M_1\gamma^1).\end{aligned}$$

It follows from (5.82), (5.88), and (5.98) that

$$\begin{aligned}(5.99)\qquad 0 &= \int_0^\infty \langle\, \tilde{g}^2, \tilde{v}^0_\mu \,\rangle_\mu\, dr\\ &= \left(M_1\gamma^1_z + \frac{\kappa^+}{2(1+d\kappa^+)}(M_2\zeta^{1,2}+M_1\gamma^1)\right)\\ &\qquad +\frac{\kappa^+}{2(1+d\kappa^+)}(M_2\zeta^{1,2}+M_1\gamma^1)\\ &= M_1\gamma^1_z + \frac{M_1\kappa^+}{1+d\kappa^+}\gamma^1 + \frac{M_2\kappa^+}{1+d\kappa^+}\zeta^{1,2}\end{aligned}$$

at $z=-d$. This shows (5.78). □

Similarly, $\tilde{v}^2$ exists in a neighborhood of $z=d$ if and only if

$$(5.100)\qquad \gamma^1_z(T,d) + \frac{\kappa^+}{1-d\kappa^+}\gamma^1(T,d) + \frac{M_2\kappa^+}{M_1(1-d\kappa^+)}\zeta^{1,2}(T,d) = 0.$$

Finally, we consider v^3 of (5.25). Put $w = v^3 - M_0\varepsilon^m$. Then w satisfies

$$\begin{aligned}(5.101)\qquad 0 &= F^0_v(v^0)w + F^0_{vv}(v^0)v^1\cdot v^2 + \frac{1}{6}F^0_{vvv}(v^0)(v^1)^3 + F^1(z,\mu)v^2\\ &\quad +F^2(z,\mu)v^1 + F^3(z,\mu)v^0 + S^+_T v^0_\mu + (M_0f'(v^0)+1)\varepsilon^m\end{aligned}$$

with

$$w(T,z,\mu)\to 0 \ as\ |\mu|\to\infty.$$

Hence the condition for the existence of w in (5.101) is

$$\begin{aligned}(5.102)\qquad 0 = \Big\langle & F^0_{vv}(v^0)v^1\cdot v^2 + \frac{1}{6}F^0_{vvv}(v^0)(v^1)^3 + F^1(z,\mu)v^2\\ &+F^2(z,\mu)v^1 + F^3(z,\mu)v^0 + S^+_T v^0_\mu + (M_0f'(v^0)+1)\varepsilon^m, v^0_\mu \Big\rangle_\mu.\end{aligned}$$

LEMMA 5.11. *The condition* (5.102) *is equivalent to*

$$(5.103)\qquad 0 = \langle\, F^2v^0, v^1_\mu \,\rangle_\mu + \langle\, F^2v^1, v^0_\mu \,\rangle_\mu + \langle\, F^3v^0, v^0_\mu \,\rangle_\mu + M_1(1-z\kappa^+)S^+_T + 2\varepsilon^m.$$

Proof. It is easily shown that

$$\begin{aligned}(5.104)\qquad &\langle\, S^+_T v^0_\mu + (M_0f'(v^0)+1)\varepsilon^m, \tilde{v}^0_\mu \,\rangle_\mu\\ &= \langle\, (1-z\kappa^+)S^+_T\varphi_\eta + (M_0f'(\varphi)+1)\varepsilon^m, \varphi_\eta \,\rangle_\eta\\ &= M_1(1-z\kappa^+)S^+_T + 2\varepsilon^m.\end{aligned}$$

So we consider the remaining terms of (5.103).

By (5.63), we have

$$0 = \left\langle F^0_{vv}(v^0)v^0_\mu \cdot v^1 + F^0_v(v^0)v^1_\mu + \frac{\partial}{\partial\mu}F^1v^0, v^2 \right\rangle_\mu \tag{5.105}$$
$$= \langle\, F^0_{vv}(v^0)v^1 \cdot v^2, v^0_\mu \,\rangle_\mu + \langle\, F^0_v(v^0)v^2, v^1_\mu \,\rangle_\mu - \langle\, F^1v^0, v^2_\mu \,\rangle_\mu .$$

Here

$$-\langle\, F^1v^0, v^2_\mu \,\rangle_\mu = \langle\, F^1v^2, v^0_\mu \,\rangle_\mu$$

holds by (5.66). Hence (5.105) is written as

$$0 = \langle\, F^0_{vv}(v^0)v^1 \cdot v^2 + F^1v^2, v^0_\mu \,\rangle_\mu + \langle\, F^0_v(v^0)v^2, v^1_\mu \,\rangle_\mu . \tag{5.106}$$

On the other hand, it follows from (5.24) that

$$0 = -\langle\, F^0_v(v^0)v^2, v^1_\mu \,\rangle_\mu + \frac{1}{2}\langle\, F^0_{vv}(v^0)v^1 \cdot v^1, v^1_\mu \,\rangle_\mu \tag{5.107}$$
$$+ \langle\, F^1v^1, v^1_\mu \,\rangle_\mu + \langle\, F^2v^0, v^1_\mu \,\rangle_\mu .$$

Now we have

$$\begin{aligned}\frac{1}{2}\langle\, F^0_{vv}(v^0)v^1 \cdot v^1, v^1_\mu \,\rangle_\mu &= -\frac{1}{2}\langle\, (F^0_{vv}(v^0)v^1 \cdot v^1)_\mu, v^1 \,\rangle_\mu \\ &= -\frac{1}{2}\langle\, F^0_{vvv}(v^0)v^1 \cdot v^1 \cdot v^0_\mu + 2F^0_{vv}(v^0)v^1 \cdot v^1_\mu, v^1 \,\rangle_\mu \\ &= -\frac{1}{2}\langle\, F^0_{vvv}(v^0)(v^1)^3, v^0_\mu \,\rangle_\mu - \langle\, F^0_{vv}(v^0)(v^1)^2, v^1_\mu \,\rangle_\mu ,\end{aligned}$$

so

$$\langle\, F^0_{vv}(v^0)(v^1)^2, v^1_\mu \,\rangle_\mu = -\frac{1}{3}\langle\, F^0_{vvv}(v^0)(v^1)^3, v^0_\mu \,\rangle_\mu . \tag{5.108}$$

Substituting (5.108) into (5.107), we obtain

$$0 = \langle\, F^0_v(v^0)v^2, v^1_\mu \,\rangle_\mu - \frac{1}{6}\langle\, F^0_{vvv}(v^0)(v^1)^3, v^0_\mu \,\rangle_\mu \tag{5.109}$$
$$+ \langle\, F^1v^1, v^1_\mu \,\rangle_\mu + \langle\, F^2v^0, v^1_\mu \,\rangle_\mu .$$

Hence, by (5.106), (5.109), and the equality $\langle\, F^1v^1, v^1_\mu \,\rangle_\mu = 0$, we have

$$\begin{aligned}&\left\langle F^0_{vv}(v^0)v^1 \cdot v^2 + \frac{1}{6}F^0_{vvv}(v^0)(v^1)^3 + F^1(z,\mu)v^2 + F^2(z,\mu)v^1 + F^3(z,\mu)v^0, v^0_\mu \right\rangle_\mu \\ &= \langle\, F^1v^1, v^1_\mu \,\rangle_\mu + \langle\, F^2v^0, v^1_\mu \,\rangle_\mu + \langle\, F^2v^1, v^0_\mu \,\rangle_\mu + \langle\, F^3v^0, v^0_\mu \,\rangle_\mu \\ &= \langle\, F^2v^0, v^1_\mu \,\rangle_\mu + \langle\, F^2v^1, v^0_\mu \,\rangle_\mu + \langle\, F^3v^0, v^0_\mu \,\rangle_\mu ,\end{aligned}$$

which completes the proof. □

LEMMA 5.12. *The equality*

$$\langle\, F^2v^0, v^1_\mu \,\rangle_\mu + \langle\, F^2v^1, v^0_\mu \,\rangle_\mu = M_1\gamma^1_{zz} + \frac{M_2(1+z\kappa^+)}{(1-z\kappa^+)^4}\kappa^+\kappa^+_s \tag{5.110}$$

holds.

Proof. Put $F^2v = F^{2,1}v + F^{2,2}v$, where

$$F^{2,1}v = v_{zz} - \frac{\kappa^+}{1-z\kappa^+}v_z,$$
$$F^{2,2}v = \frac{z}{(1-z\kappa^+)^3}\left(\kappa^+_{ss} + \frac{3z\kappa^{+2}_s}{1-z\kappa^+}\right)(\mu v_\mu + \mu^2 v_{\mu\mu}).$$

First we note that

$$\langle\, F^{2,2}v, w_\mu \,\rangle_\mu = -\langle\, v_\mu, F^{2,2}w \,\rangle_\mu$$

holds in general. Therefore, we have

$$\langle\, F^2v^0, v^1_\mu \,\rangle_\mu + \langle\, F^2v^1, v^0_\mu \,\rangle_\mu = \langle\, F^{2,1}v^0, v^1_\mu \,\rangle_\mu + \langle\, F^{2,1}v^1, v^0_\mu \,\rangle_\mu . \tag{5.111}$$

Let $\eta = (1-z\kappa^+)(\mu - \gamma^+)$. Then $F^{2,1}$ is expressed as

$$\begin{aligned} F^{2,1}v &= v_{zz} - \frac{2\kappa^+}{1-z\kappa^+}\eta v_{z\eta} + \left(\frac{\kappa^+}{1-z\kappa^+}\right)^2 \eta^2 v_{\eta\eta} \\ &\quad - \frac{\kappa^+}{1-z\kappa^+}\left(v_z - \frac{\kappa^+}{1-z\kappa^+}\eta v_\eta\right) \\ &= v_{zz} - \frac{2\kappa^+}{1-z\kappa^+}\eta v_{z\eta} - \frac{\kappa^+}{1-z\kappa^+}v_z \\ &\quad + \left(\frac{\kappa^+}{1-z\kappa^+}\right)^2 (\eta v_\eta + \eta^2 v_{\eta\eta}), \end{aligned} \tag{5.112}$$

and v^0 and v^1 are already obtained as

$$v^0 = \varphi(\eta), \tag{5.113}$$
$$v^1 = V^1(T,z,\eta) + \gamma^1(T,z)\varphi_\eta(\eta) \tag{5.114}$$

by (5.29) and (5.50). Since $d\eta = (1-z\kappa^+)d\mu$ and

$$v^0_\mu = (1-z\kappa^+)\varphi_\eta,$$
$$v^1_\mu = (1-z\kappa^+)\{V^1 + \gamma^1\varphi_\eta\}_\eta$$

by (5.113) and (5.114), the right-hand side of (5.111) is written as

$$\begin{aligned} &\langle\, F^{2,1}v^0, v^1_\mu \,\rangle_\mu + \langle\, F^{2,1}v^1, v^0_\mu \,\rangle_\mu \\ &= \langle\, F^{2,1}\varphi, (V^1 + \gamma^1\varphi_\eta)_\eta \,\rangle_\eta + \langle\, F^{2,1}(V^1 + \gamma^1\varphi_\eta), \varphi_\eta \,\rangle_\eta . \end{aligned} \tag{5.115}$$

Let $F^{2,1}v = F^*v + F^{**}v$ with

$$F^*v = v_{zz} - \frac{2\kappa^+}{1-z\kappa^+}\eta v_{z\eta} - \frac{\kappa^+}{1-z\kappa^+}v_z,$$
$$F^{**}v = \left(\frac{\kappa^+}{1-z\kappa^+}\right)^2 (\eta v_\eta + \eta^2 v_{\eta\eta}).$$

Since F^{**} also satisfies $\langle\, F^{**}v, w_\eta \,\rangle_\eta = -\langle\, v_\eta, F^{**}w \,\rangle_\eta$ in general, we have

$$\begin{aligned} &\langle\, F^{2,1}\varphi, (V^1 + \gamma^1\varphi_\eta)_\eta \,\rangle_\eta + \langle\, F^{2,1}(V^1 + \gamma^1\varphi_\eta), \varphi_\eta \,\rangle_\eta \\ &= \langle\, F^*\varphi, (V^1 + \gamma^1\varphi_\eta)_\eta \,\rangle_\eta + \langle\, F^*(V^1 + \gamma^1\varphi_\eta), \varphi_\eta \,\rangle_\eta . \end{aligned} \tag{5.116}$$

Since $F^*\varphi = 0$, it follows from (5.111), (5.115), and (5.116) that

$$\langle F^2 v^0, v_\mu^1 \rangle_\mu + \langle F^2 v^1, v_\mu^0 \rangle_\mu = \langle F^*(V^1 + \gamma^1 \varphi_\eta), \varphi_\eta \rangle_\eta . \tag{5.117}$$

We will show

$$\langle F^*(\gamma^1 \varphi_\eta), \varphi_\eta \rangle_\eta = M_1 \gamma_{zz}^1. \tag{5.118}$$

Since $\langle \eta\varphi_\eta, \varphi_{\eta\eta} \rangle_\eta = -\frac{1}{2} M_1$ and

$$F^*(\gamma^1 \varphi_\eta) = \gamma_{zz}^1 \varphi_\eta - \frac{2\kappa^+}{1 - z\kappa^+} \eta \gamma_z^1 \varphi_{\eta\eta} - \frac{\kappa^+}{1 - z\kappa^+} \gamma_z^1 \varphi_\eta,$$

we have

$$\begin{aligned} \langle F^*(\gamma^1 \varphi_\eta), \varphi_\eta \rangle_\eta &= M_1 \gamma_{zz}^1 + \frac{M_1 \kappa^+}{1 - z\kappa^+} \gamma_z^1 - \frac{M_1 \kappa^+}{1 - z\kappa^+} \gamma_z^1 \\ &= M_1 \gamma_{zz}^1. \end{aligned}$$

This shows (5.118).

Next, we will show

$$\langle F^* V^1, \varphi_\eta \rangle_\eta = \frac{M_2 (1 + z\kappa^+)}{(1 - z\kappa^+)^4} \kappa^+ \kappa_s^+. \tag{5.119}$$

Substituting (5.77), we have

$$F^* V^1 = \sum_{j=1}^{2} \left(\zeta_{zz}^{1,j} U^{1,j} - \frac{2\kappa^+}{1 - z\kappa^+} \eta \zeta_z^{1,j} U_\eta^{1,j} - \frac{\kappa^+}{1 - z\kappa^+} \zeta_z^{1,j} U^{1,j} \right).$$

Since $\langle U^{1,j}, \varphi_\eta \rangle_\eta = 0$ $(j = 1, 2)$, the left-hand side of (5.119) satisfies the equality

$$\langle F^* V^1, \varphi_\eta \rangle_\eta = -\frac{2\kappa^+}{1 - z\kappa^+} \sum_{j=1}^{2} \zeta_z^{1,j} \langle \eta U_\eta^{1,j}, \varphi_\eta \rangle_\eta . \tag{5.120}$$

Since both $U_\eta^{1,1}$ and φ_η are even functions, it follows from (5.120) that

$$\begin{aligned} \langle F^* V^1, \varphi_\eta \rangle_\eta &= -\frac{2\kappa^+}{1 - z\kappa^+} \zeta_z^{1,2} \langle \eta U_\eta^{1,2}, \varphi_\eta \rangle_\eta \\ &= -\frac{2\kappa^+}{1 - z\kappa^+} \left(\frac{z\kappa_s^+}{(1 - z\kappa^+)^2} \right)_z \langle \eta U_\eta^{1,2}, \varphi_\eta \rangle_\eta \\ &= -\frac{2(1 + z\kappa^+)}{(1 - z\kappa^+)^4} \kappa^+ \kappa_s^+ \langle \eta U_\eta^{1,2}, \varphi_\eta \rangle_\eta . \end{aligned} \tag{5.121}$$

Here, by Lemma 5.9, we have

$$\begin{aligned} \langle \eta U_\eta^{1,2}, \varphi_\eta \rangle_\eta &= -\left\langle \eta\varphi_\eta + \frac{1}{2}\eta^2 \varphi_{\eta\eta} - \frac{M_2}{2M_1} \varphi_{\eta\eta}, \eta\varphi_\eta \right\rangle_\eta \\ &= -M_2 - \frac{1}{2} \langle \eta^3 \varphi_{\eta\eta}, \varphi_\eta \rangle_\eta + \frac{M_2}{2M_1} \langle \varphi_{\eta\eta}, \eta\varphi_\eta \rangle_\eta \\ &= -M_2 - \frac{1}{4} \langle \eta^3, (\varphi_\eta{}^2)_\eta \rangle_\eta + \frac{M_2}{2M_1} \left(-\frac{1}{2} M_1 \right) \\ &= -M_2 + \frac{3}{4} M_2 - \frac{1}{4} M_2 \\ &= -\frac{1}{2} M_2. \end{aligned} \tag{5.122}$$

Thus, (5.121) and (5.122) show (5.119).

By (5.117), (5.118), and (5.119), the proof is complete. □

LEMMA 5.13. *The equality*

$$\langle\, F^3 v^0, v^0_\mu \,\rangle_\mu = \frac{M_2}{(1-z\kappa^+)^4}\kappa^+\kappa^+_s \tag{5.123}$$

holds.

Proof. First it is easily shown that

$$\left\langle \frac{3}{2}\mu^2 v^0_\mu + \mu^3 v^0_{\mu\mu}, v^0_\mu \right\rangle_\mu = 0.$$

Hence

$$\langle\, F^3 v^0, v^0_\mu \,\rangle_\mu = -\frac{\kappa^+_s}{(1-z\kappa^+)^2}\,\langle\, \mu v^0_z, v^0_\mu \,\rangle_\mu\,. \tag{5.124}$$

Since

$$\mu v^0_z = \left(\frac{\eta}{1-z\kappa^+} + \gamma^0\right)\left(-\frac{\kappa^+}{1-z\kappa^+}\right)\eta\varphi_\eta$$

by (5.29) and (5.68), we have

$$\begin{aligned}
\langle\, \mu v^0_z, v^0_\mu \,\rangle_\mu &= \langle\, \mu v^0_z, \varphi_\eta \,\rangle_\eta \\
&= -\frac{\kappa^+}{1-z\kappa^+}\left(\frac{1}{1-z\kappa^+}\,\langle\, \eta^2\varphi_\eta, \varphi_\eta \,\rangle_\eta + \gamma^0\,\langle\, \eta\varphi_\eta, \varphi_\eta \,\rangle_\eta\right) \\
&= -\frac{\kappa^+}{1-z\kappa^+}\times\frac{1}{1-z\kappa^+}M_2 \\
&= -\frac{\kappa^+}{(1-z\kappa^+)^2}.
\end{aligned}$$

Substituting this into (5.124), we complete the proof. □

By (5.102) and Lemmas 5.11–5.13, there exists v^3 satisfying (5.25) and (5.28) if and only if

$$\begin{aligned}
0 &= M_1\gamma^1_{zz} + \frac{M_2(1+z\kappa^+)}{(1-z\kappa^+)^4}\kappa^+\kappa^+_s + \frac{M_2}{(1-z\kappa^+)^4}\kappa^+\kappa^+_s \\
&\qquad + M_1(1-z\kappa^+)S^+_T + 2\varepsilon^m \\
&= M_1\gamma^1_{zz} + \frac{M_2(2+z\kappa^+)}{(1-z\kappa^+)^4}\kappa^+\kappa^+_s + M_1(1-z\kappa^+)S^+_T + 2\varepsilon^m.
\end{aligned} \tag{5.125}$$

LEMMA 5.14. *There exists* γ^1 *satisfying* (5.125) *with* (5.78) *and* (5.100) *if and only if*

$$S^+_T = H^+(S^+). \tag{5.126}$$

Proof. Set

$$\begin{aligned}
K(z) &= \frac{1}{M_1}\left(\frac{M_2(2+z\kappa^+)}{(1-z\kappa^+)^4}\kappa^+\kappa^+_s + M_1(1-z\kappa^+)S^+_T + 2\varepsilon^m\right), \\
\omega^\pm &= \frac{\kappa^+}{1\mp d\kappa^+}, \\
\lambda^\pm &= \frac{M_2\kappa^+}{M_1(1\mp d\kappa^+)}\zeta^{1,2}(T, \pm d) \\
&= \pm\frac{M_2 d}{M_1(1\mp d\kappa^+)^3}\kappa^+\kappa^+_s.
\end{aligned}$$

Then (5.125) with (5.100) and (5.78) is expressed as

$$\left\{ \begin{array}{l} \gamma^1_{zz} + K = 0, \quad -d < z < d, \\ \gamma^1_z + \omega^{\pm}\gamma^1 + \lambda^{\pm} = 0, \quad z = \pm d. \end{array} \right. \tag{5.127}$$

Let $\gamma^{1,\pm} = \gamma^1(\pm d)$. Then by integrating (5.127) over $(-d, d)$, we have

$$-\omega^+\gamma^{1,+} - \lambda^+ + \omega^-\gamma^{1,-} + \lambda^- + K_1 = 0, \tag{5.128}$$

where

$$K_1 = \int_{-d}^{d} K(z)dz.$$

On the other hand, by integrating (5.127) over $(-d, d)$ twice, we have

$$\gamma^{1,+} - \gamma^{1,-} + 2(\omega^-\gamma^{1,-} + \lambda^-)d + K_2 = 0, \tag{5.129}$$

where

$$K_2 = \int_{-d}^{d}\int_{-d}^{z} K(t)dtdz.$$

Set

$$D = \begin{pmatrix} -\omega^+ & \omega^- \\ 1 & 2d\omega^- - 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} K_1 - \lambda^+ + \lambda^- \\ K_2 + 2d\lambda^- \end{pmatrix}.$$

Then the equalities (5.128) and (5.129) are expressed as

$$D\begin{pmatrix} \gamma^{1,+} \\ \gamma^{1,-} \end{pmatrix} + \boldsymbol{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{5.130}$$

The determinant of D is computed as

$$\begin{aligned} \det\ D &= -\omega^+(2d\omega^- - 1) - \omega^- \\ &= -2d\omega^+\omega^- + \omega^+ - \omega^- \\ &= -2d\frac{\kappa^+}{1 - d\kappa^+} \times \frac{\kappa^+}{1 + d\kappa^+} + \frac{\kappa^+}{1 - d\kappa^+} - \frac{\kappa^+}{1 + d\kappa^+} \\ &= 0. \end{aligned}$$

Hence $\boldsymbol{b}$ must be orthogonal to the eigenfunction

$$\boldsymbol{x} = \begin{pmatrix} 2d\omega^- - 1 \\ -\omega^- \end{pmatrix}$$

of ${}^tD$ (the transposed matrix of D) in order that $(\gamma^{1,+}, \gamma^{1,-})$ exists. Namely

$$(2d\omega^- - 1)(K_1 - \lambda^+ + \lambda^-) - \omega^-(K_2 + 2d\lambda^-) = 0 \tag{5.131}$$

must be satisfied. Here we have

$$K_1 = \frac{4M_2 d(1 + d^2(\kappa^+)^2)}{M_1(1 - d\kappa^+)^3(1 + d\kappa^+)^3}\kappa^+\kappa_s^+ + 2dS_T^+ + \frac{4dp}{M_1}\varepsilon^m,$$

$$K_2 = \frac{M_2 d}{M_1}\left(\frac{1+d^2(\kappa^+)^2}{\kappa^+(1-d\kappa^+)^2(1+d\kappa^+)^2} - \frac{1-d\kappa^+}{\kappa^+(1+d\kappa^+)^3}\right)\kappa^+\kappa_s^+ + 2d^2\left(\frac{1}{3}d\kappa^+ + 1\right)S_T^+ + \frac{2d^2 p}{M_1}\varepsilon^m.$$

Substituting these into (5.131), we complete the proof. □

Let $S^+(T)$ be the solution of (5.126) defined for $T \in [0, T^*]$. Then we can construct $v^0, \dots, v^3$ and $\tilde{v}^0, \dots, \tilde{v}^2$ as we have seen so far. In quite a similar manner, we can show the existence of $\tilde{v}^3$ satisfying (5.39), (5.31) and (5.43), (5.44). We will roughly explain it and omit the details.

By Proposition 5.2 and Lemma 5.7, v^2 is expressed as

$$v^2 = V^2 + \gamma^2\varphi_\eta$$

for certain functions $V^2 = V^2(T, z, \eta)$ and $\gamma^2 = \gamma^2(T, z)$, where $\eta = (1 - z\kappa^+)(\mu - \gamma^+)$. Then the solvability condition for the existence of $\tilde{v}^3$ gives boundary conditions of γ^2 at $z = -d$ as (5.41) and (5.42) give (5.52) and (5.78), respectively. On the other hand, the equation of γ^2 for $z \in (-d, d)$ is derived from the solvability condition of v^4 as (5.71) and (5.125) are derived from (5.24) and (5.25), respectively. However we need not consider v^4 here, because the complete properties of γ^2, except that v^2 satisfies (5.24), and (5.27), are not necessary. Therefore, we can give an adequate boundary value of γ^2 such that $\tilde{v}^3$ exists.

5.3. Construction of the approximate function v^+. We will show the existence of a positive constant $n \in (m+3, 4)$ such that there exists a function $v^+(t, z, s)$ approximating (5.10) up to $O(\varepsilon^n)$ uniformly in $t \in [0, \leq T^*/\varepsilon^4]$ and $(z, s) \in (-d, d) \times (0, L)$. We note that v^j and $\tilde{v}^j$ $(j = 0, \dots, 3)$ are represented as functions of (T, z, η_z) and (T, r, η_{-d}), respectively, with $\eta_z = (1 - z\kappa^+)(\mu - \gamma^+(T))$ and $\eta_{-d} = (1 + d\kappa^+)(\mu - \gamma^+(T))$. Hence we can write v^j and $\tilde{v}^j$ $(j = 0, \dots, 3)$ as $v^j(T, z, \eta_z)$ and $\tilde{v}^j(T, r, \eta_{-d})$ $(j = 0, \dots, 3)$, respectively, and expand

$$v^\varepsilon(T, z, \eta) = \sum_{j=0}^{3} \varepsilon^j v^j(T, z, \eta),$$

$$\tilde{v}^\varepsilon(T, r, \eta) = \sum_{j=0}^{3} \varepsilon^j \tilde{v}^j(T, r, \eta).$$

Also $\widetilde{W}^j$ $(j = 0, \dots, 3)$ are represented as functions of (T, r, η_{-d}). Hence we write those as $\widetilde{W}^j(T, r, \eta_{-d})$ $(j = 0, \dots, 3)$ and expand

$$\widetilde{W}^\varepsilon(T, r, \eta) = \sum_{j=0}^{3} \varepsilon^j \widetilde{W}^j(T, r, \eta).$$

Then v^ε and $\tilde{v}^\varepsilon$ satisfy

$$|v^\varepsilon(T, z, \eta) - (\pm 1 + M_0\varepsilon^{m+3})|_{C^{5,5}(z,\eta)} \leq Be^{-\beta|\eta|}, \tag{5.132}$$

$$|\tilde{v}^\varepsilon(T, r, \eta) - \widetilde{W}^\varepsilon(T, r, \eta)|_{C^{2,2}(r,\eta)} \leq B\min\{e^{-br}, e^{-\beta|\eta|}\}, \tag{5.133}$$

respectively, for $b \in (0, b_1)$, $\beta \in (\frac{3}{4}\beta_0, \beta_2)$ by (5.51) and so on. Furthermore, it follows from the definition of $\widetilde{W}^j$ and (5.132) that

$$\begin{aligned} |\widetilde{W}^\varepsilon(T, r, \eta_{-d})|_{C^{2,2}(r,\mu)} &\leq B|\varepsilon r|_3(1 + |\eta_{-d}|^3)|v^\varepsilon(T, z, \eta_{-d})|_{C^{3,5}(z,\eta)} \\ &\leq B|\varepsilon r|_3 e^{-\beta'|\eta_{-d}|} \end{aligned} \tag{5.134}$$

for $r > 0$ and some $\beta' \in (\frac{3}{4}\beta_0, \beta_0)$.

Let $\rho(\eta)$ be a smooth function satisfying $0 \le \rho(\eta) \le 1$, $\rho(\eta) = 0$ for $\eta \le -1$, and $\rho(\eta) = 1$ for $\eta \ge 1$, and let

$$\hat{v}(t,z,\mu) = \left\{1 - \rho\left(\frac{r - 2q|\log\varepsilon|}{q|\log\varepsilon|}\right)\right\}\tilde{v}^\varepsilon(T,r,\eta_{-d}) + \rho\left(\frac{r-2q|\log\varepsilon|}{q|\log\varepsilon|}\right)v^\varepsilon(T,z,\eta_z),$$

where $T = \varepsilon^4 t$, $r = \frac{z+d}{\varepsilon}$ and q is a positive constant with $bq \ge 4$.

LEMMA 5.15. *For any fixed $n \in (m+3, 4)$, $\hat{v}$ satisfies*

$$\hat{v}_z(t,-d,\mu) = 0 \tag{5.135}$$

and

$$F^\varepsilon(\hat{v}) - \hat{v}_t = O(\varepsilon^n) \tag{5.136}$$

uniformly in $(t,z,\mu) \in [0, T^*/\varepsilon^4] \times (-d,d) \times (-\infty,\infty)$.

Proof. From the definition of $\hat{v}$, we may write it by $\hat{v}(T,z,\mu)$, and (5.135) obviously holds. Therefore, it suffices to show that

$$F^\varepsilon(\hat{v}) = O(\varepsilon^n) \tag{5.137}$$

uniformly in $(t,z,\mu) \in J$, where $J = [0,T^*] \times (-d,d) \times (-\infty,\infty)$.

First, let us consider (5.137) in the range of $0 \le r \le q|\log\varepsilon|$, that is, $-d \le z \le -d + q\varepsilon|\log\varepsilon|$. In this range, $\hat{v}(T,z,\mu) = \tilde{v}^\varepsilon(T,r,\eta_{-d})$ holds. Now it follows from (5.21) that

$$\left|\widetilde{F}^\varepsilon(\tilde{v}) - \sum_{j=0}^{3}\varepsilon^j\widetilde{F}^j(\tilde{v})\right| \le B\varepsilon^4 r^4 |\mu|_4 |\tilde{v}(r,\mu)|_{D^{1,2}(r,\mu)} \tag{5.138}$$

for $r > 1$. Hence, by (5.133) and inequalities $m+3 < n < 4$, $0 < \beta' < \beta$, we have

$$\begin{aligned}
&|\widetilde{F}^\varepsilon(\tilde{v}^\varepsilon)| \\
&\le \left|\sum_{j=0}^{3}\varepsilon^j\widetilde{F}^j(\tilde{v}^\varepsilon)\right| + B\varepsilon^4 r^4|\mu|_4|\tilde{v}^\varepsilon(T,r,\mu)|_{D^{1,2}(r,\mu)} \\
&\le B\varepsilon^4 r^4\left|\frac{\eta_{-d}}{1+d\kappa^+} + \gamma^+\right|_4 |\tilde{v}^\varepsilon(T,r,\eta_{-d})|_{C^{1,2}(r,\eta)} \\
&\le B\varepsilon^4 r^4\left|\frac{\eta_{-d}}{1+d\kappa^+} + \gamma^+\right|_4 \left\{|\widetilde{W}^\varepsilon(T,r,\eta_{-d})|_{C^{1,2}(r,\eta)} + B\min\{e^{-br}, e^{\beta|\eta_{-d}|}\}\right\} \\
&\le B\varepsilon^4 r^4\left|\frac{\eta_{-d}}{1+d\kappa^+} + \gamma^+\right|_4 \left\{B|\varepsilon r|_3 e^{-\beta|\eta_{-d}|} + B\min\{e^{-br}, e^{-\beta|\eta_{-d}|}\}\right\} \\
&\le B\varepsilon^4 r^4(|\varepsilon r|_3 + 1)\left|\frac{\eta_{-d}}{1+d\kappa^+} + \gamma^+\right|_4 e^{\beta|\eta_{-d}|} \\
&\le B\varepsilon^4 r^4 \\
&\le B(q\varepsilon|\log\varepsilon|)^4 \\
&\le B\varepsilon^n
\end{aligned}$$

for $1 < r \le q|\log\varepsilon|$, where $\tilde{v}^\varepsilon(T,r,\mu) = \tilde{v}^\varepsilon(T,r,\eta_{-d})$. For $0 \le r \le 1$, the inequality $|\widetilde{F}^\varepsilon(\tilde{v}^\varepsilon)| \le B\varepsilon^n$ obviously holds. Thus, we have

$$|\widetilde{F}^\varepsilon(\tilde{v}^\varepsilon)| \le B\varepsilon^n \tag{5.139}$$

uniformly in $(T,r,\mu) \in [0,T^*] \times [0, q|\log\varepsilon|] \times (-\infty,\infty)$.

Next, consider (5.137) in the range of $r \geq 3q|\log\varepsilon|$, that is, $z \geq -d + 3q\varepsilon|\log\varepsilon|$. In this range of r, $\hat{v}(T,z,\mu) = v^\varepsilon(T,z,\eta_z)$ holds. By

$$\frac{\partial}{\partial z} v(z,\eta_z) = v_z - \frac{\eta_z}{1-z\kappa^+} v_\eta,$$

we have

$$|v(z,\eta_z)|_{D^{1,2}(z,\mu)} \leq B(1+|\eta_z|)|v(z,\eta_z)|_{D^{1,2}(z,\eta)}.$$

Hence, by (5.21) and (5.132), we have

$$\begin{aligned}(5.140)\quad |F^\varepsilon(v^\varepsilon)| &\leq \left|\sum_{j=0}^{3} \varepsilon^j F^j(v^\varepsilon)\right| \\ &\qquad + B\varepsilon^4 \left|\frac{\eta_z}{1-z\kappa^+} + \gamma^+\right|_4 (1+|\eta_z|)|v^\varepsilon(T,z,\eta_z)|_{D^{1,2}(z,\eta)} \\ &\leq B\varepsilon^4(1+|\eta_z|)^5 e^{-\beta|\eta_z|} \\ &\leq B\varepsilon^4\end{aligned}$$

uniformly in $(T,z,\mu) \in J$.

Finally, consider (5.137) in the range of $q|\log\varepsilon| \leq r \leq 3q|\log\varepsilon|$, that is, $-d + q\varepsilon|\log\varepsilon| \leq z \leq -d + 3q\varepsilon|\log\varepsilon|$. In this range of r, we have

$$\begin{aligned}(5.141)\qquad \left|\sum_{j=0}^{3} \varepsilon^j F^\varepsilon(\widetilde{W}^\varepsilon)\right| &\leq B|\varepsilon r|^4 e^{-\beta|\eta|} \\ &\leq B(\varepsilon|\log\varepsilon|)^4 e^{-\beta|\eta|}\end{aligned}$$

uniformly in $(T,r,\eta) \in [0,T^*] \times [q|\log\varepsilon|, 3q|\log\varepsilon|] \times (-\infty,\infty)$. Furthermore, we have

$$\begin{aligned}(5.142)\qquad |\tilde{v}^\varepsilon(T,r,\eta) - \widetilde{W}^\varepsilon(T,r,\eta)|_{C^{2,2}(r,\eta)} &\leq B\min\{e^{-br}, e^{-\beta|\eta|}\} \\ &\leq B\min\{\varepsilon^{bq}, e^{-\beta|\eta|}\} \\ &\leq B\min\{\varepsilon^4, e^{-\beta|\eta|}\}\end{aligned}$$

by (5.133) and the inequalities $bq \geq 4$, $q|\log\varepsilon| \leq r \leq 3q|\log\varepsilon|$. Then the inequalities

$$(5.143)\qquad \begin{cases} |v^\varepsilon(T,z,\eta_z) - \sum_{j=0}^{3} \varepsilon^j \widetilde{W}^j(T,r,\eta_{-d})|_{C^{0,2}(z,\mu)} \\ \qquad \leq B\,|z+d|^4(e^{-\beta|\eta_z|} + e^{-\beta|\eta_{-d}|}), \\ \left|v^\varepsilon(T,z,\eta_z) - \displaystyle\sum_{j=0}^{3} \varepsilon^j \widetilde{W}^j(T,r,\eta_{-d})\right|_{C^{2,2}(z,\mu)} \\ \qquad \leq B\,|z+d|^2\left((1+|\eta_z|^2)e^{-\beta|\eta_z|} + e^{-\beta|\eta_{-d}|}\right) \\ \qquad \leq B\,|z+d|^2(e^{-\beta'|\eta_z|} + e^{-\beta|\eta_{-d}|}) \end{cases}$$

hold for $(T,z,\mu) \in J$ and $\beta' \in (\frac{3}{4}\beta_0, \beta)$. Therefore, by (5.142) and (5.143), $\hat{v}(T,z,\mu)$ is expressed as

$$\begin{aligned}(5.144)\qquad &\hat{v}(T,z,\mu) \\ &= (1-\rho)\cdot(\widetilde{W}^\varepsilon + a(T,r,\eta_{-d})) + \rho\cdot(\widetilde{W}^\varepsilon + b(T,z,\eta_z,\eta_{-d})) \\ &= \widetilde{W}^\varepsilon + c(T,z,\eta_z,\eta_{-d}),\end{aligned}$$

where

$$\widetilde{W}^\varepsilon = \widetilde{W}^\varepsilon(T, r, \eta_{-d}),$$
$$\rho = \rho\left(\frac{r - 2q|\log\varepsilon|}{q|\log\varepsilon|}\right),$$
$$c(T, z, \eta_z, \eta_{-d}) = (1-\rho)a\left(T, \frac{z+d}{\varepsilon}, \eta_{-d}\right) + \rho \cdot b(T, z, \eta_z, \eta_{-d}),$$

and $a(T, r, \eta_{-d}), b(T, z, \eta_z, \eta_{-d})$ satisfy

$$|a(T, r, \eta_{-d})|_{C^{2,2}(r,\mu)}, |b(T, z, \eta_z, \eta_{-d})|_{C^{0,2}(z,\mu)} \le B(|z+d|^4 + \varepsilon^4)(e^{-\beta'|\eta_z|} + e^{-\beta|\eta_{-d}|}),$$

$$|b(T, z, \eta_z, \eta_{-d})|_{C^{2,2}(z,\mu)} \le B(|z+d|^2 + \varepsilon^4)(e^{-\beta'|\eta_z|} + e^{-\beta|\eta_{-d}|}).$$

Since $q\varepsilon|\log\varepsilon| \le z + d \le 3q\varepsilon|\log\varepsilon|$,

$$\begin{aligned} |c(T, z, \eta_z, \eta_{-d})|_{C^{2,2}(z,\mu)} &\le B(\varepsilon|\log\varepsilon|)^2(e^{-\beta|\eta_z|} + e^{-\beta|\eta_{-d}|}), \\ |c(T, z, \eta_z, \eta_{-d})|_{C^{0,2}(z,\mu)} &\le B(\varepsilon|\log\varepsilon|)^4(e^{-\beta|\eta_z|} + e^{-\beta|\eta_{-d}|}) \end{aligned} \tag{5.145}$$

hold. Therefore, by (5.21), (5.134), (5.141) and (5.142), (5.144), (5.145), we have

$$\begin{aligned} |F^\varepsilon(\hat{v})| &\le \left|\sum_{j=0}^{3} \varepsilon^j F^j(\hat{v})\right| + B\varepsilon^4|\mu|_4|\hat{v}|_{D^{1,2}(z,\mu)} \\ &\le \left|\sum_{j=0}^{3} \varepsilon^j F^j(\widetilde{W}^\varepsilon + c)\right| \\ &\qquad + B\varepsilon^4|\mu|_4(|\widetilde{W}^\varepsilon|_{D^{1,2}(z,\mu)} + |c(T, z, \eta_z, \eta_{-d})|_{D^{1,2}(z,\mu)}) \\ &\le \left|\sum_{j=0}^{3} \varepsilon^j F^j(\widetilde{W}^\varepsilon)\right| + B|c(T, z, \eta_z, \eta_{-d})| + \left|\sum_{j=1}^{3} \varepsilon^j F^j c\right| \\ &\qquad + B\varepsilon^4|\mu|_4(|\varepsilon r|_3 + (\varepsilon|\log\varepsilon|)^2)(e^{-\beta'|\eta_z|} + e^{-\beta'|\eta_{-d}|}) \\ &\le B(\varepsilon|\log\varepsilon|)^4 e^{-\beta'|\eta_{-d}|} + B(\varepsilon|\log\varepsilon|)^4(e^{-\beta'|\eta_z|} + e^{-\beta'|\eta_{-d}|}) \\ &\qquad + \varepsilon|F^1 c| + \left|\sum_{j=2}^{3} \varepsilon^j F^j c\right| + B\varepsilon^4|\mu|_4(e^{-\beta'|\eta_z|} + e^{-\beta'|\eta_{-d}|}) \\ &\le B\varepsilon^n|\mu|_4(e^{-\beta'|\eta_z|} + e^{-\beta'|\eta_{-d}|}) + B\varepsilon|c(T, z, \eta_z, \eta_{-d})|_{C^{0,2}(z,\mu)} \\ &\qquad + B\varepsilon^2|c(T, z, \eta_z, \eta_{-d})|_{C^{2,2}(z,\mu)} \\ &\le B\varepsilon^n + B\varepsilon(\varepsilon|\log\varepsilon|)^4(e^{-\beta'|\eta_z|} + e^{-\beta'|\eta_{-d}|}) \\ &\qquad + B\varepsilon^2(\varepsilon|\log\varepsilon|)^2(e^{-\beta'|\eta_z|} + e^{-\beta'|\eta_{-d}|}) \\ &\le B\varepsilon^n \end{aligned} \tag{5.146}$$

uniformly in $(T, z, \mu) \in [0, T^*] \times [q|\log\varepsilon|, 3q|\log\varepsilon|] \times (-\infty, \infty)$. Here we note that F^1 does not contain any terms differentiated with respect to z.

Thus, it follows from (5.139), (5.140), and (5.146) that

$$F^\varepsilon(\hat{v}) \le B\varepsilon^n \tag{5.147}$$

uniformly in $(T, z, \mu) \in J$. □

In the vicinity of $z = d$, we can treat it by quite a similar way, and consequently we can construct $\hat{v}(t, z, \mu)$ satisfying (5.136), (5.135),

$$\hat{v}_z(t, d, \mu) = 0, \tag{5.148}$$

and the inequality

$$|\hat{v}(T, z, \mu) - (1 + M_0\varepsilon^m)|_{C^{2,2}(z,\mu)} \le Be^{-\beta|\eta_z|}$$

for some $\beta > 0$. Now, we may assume q is sufficiently large so that $bq, \beta q \ge 4$.

Let $\bar{\gamma} = \max_{0\le T\le T^*} |\gamma^+(T)|$ and $\bar{q} = \frac{q}{1-\delta} + \bar{\gamma}$. Define

$$v^+(t, z, s) = \begin{cases} (1-\rho)\hat{v}(t, z, \mu) + (1 + M_0\varepsilon^{m+3})\rho, & \mu > \bar{q}|\log\varepsilon|, \\ \hat{v}(t, z, \mu), & -\bar{q}|\log\varepsilon| \le \mu \le \bar{q}|\log\varepsilon|, \\ \rho\hat{v}(t, z, \mu) + (1-\rho)(-1 + M_0\varepsilon^{m+3}), & \mu < -\bar{q}|\log\varepsilon|, \end{cases}$$

where

$$\rho = \rho\left(\frac{\mu - 2\bar{q}|\log\varepsilon|}{\bar{q}|\log\varepsilon|}\right), \quad \mu = \frac{s - S^+(T)}{\varepsilon}.$$

It is obvious from the manner of construction of v^+ that $v^+(t, z, s) = u^+(t, C(s) + z\nu(s))$ satisfies (5.3) and (5.4).

LEMMA 5.16. *$v^+(t, z, s)$ satisfies* (5.11) *and*

$$\varepsilon^2\left\{v_{zz} - \frac{\kappa}{1-\kappa z}v_z + \frac{1}{1-\kappa z}\left(\frac{1}{1-\kappa z}v_s\right)_s\right\} + f(v) + \varepsilon^{m+3} - v_t = O(\varepsilon^n) \tag{5.149}$$

uniformly in $(t, z, s) \in [0, T^*/\varepsilon^4] \times (-d, d) \times (0, L)$.

Proof. It suffices to consider only the case $\mu > \bar{q}|\log\varepsilon|$.

Since $\mu > \bar{q}|\log\varepsilon|$ implies

$$\begin{aligned} \eta_z &\ge (1-\delta)(\bar{q}|\log\varepsilon| - \bar{\gamma}) \\ &\ge (1-\delta)(\bar{q} - \bar{\gamma})|\log\varepsilon| \\ &= q|\log\varepsilon|, \end{aligned}$$

$\hat{v}$ satisfies

$$\begin{aligned} |\hat{v}(T, z, \mu) - (1 + M_0\varepsilon^{m+3})|_{C^{2,2}(z,\mu)} &\le Be^{-\beta|\eta_z|} \\ &\le Be^{-\beta q|\log\varepsilon|} \\ &\le B\varepsilon^4 \end{aligned} \tag{5.150}$$

by (5.132). Hence v^+ is represented as

$$\begin{aligned} v^+(t, z, s) &= (1-\rho)(1 + M_0\varepsilon^{m+3} + d(T, z, \mu)) + \rho(1 + M_0\varepsilon^{m+3}) \\ &= (1 + M_0\varepsilon^{m+3}) + (1-\rho)d(T, z, \mu). \end{aligned}$$

Here (5.150) implies

$$|d(T, z, \mu)|_{C^{2,2}(z,\mu)} \le B\varepsilon^4, \tag{5.151}$$

$$|(1-\rho)d(T, z, \mu)|_{C^{2,2}(z,\mu)} \le B\varepsilon^4. \tag{5.152}$$

Hence, by (5.151), (5.152), and Proposition 3.3, we obtain

$$\begin{aligned}|F^\varepsilon(v^+) - v_t^+| &\le B\varepsilon^4 + |f(1+M_0\varepsilon^{m+3}) + \varepsilon^{m+3}| + \varepsilon^4|v_T^+| \\ &\le B\varepsilon^4 + O(\varepsilon^{2(m+3)}) \\ &\le B\varepsilon^4.\end{aligned}$$

This completes the proof. □

Quite similarly, we can construct v^- satisfying (5.3), (5.4), and (5.11), and

$$\begin{aligned}\varepsilon^2\left\{v_{zz} - \frac{\kappa}{1-\kappa z}v_z + \frac{1}{1-\kappa z}\left(\frac{1}{1-\kappa z}v_s\right)_s\right\} \\ +f(v) - \varepsilon^{m+3} - v_t = O(\varepsilon^n)\end{aligned} \tag{5.153}$$

uniformly in $(t,z,s) \in [0, T^*/\varepsilon^4] \times (-d,d) \times (0,L)$ if $S^-(T)$ is a solution of

$$S_T^- = H^-(S^-). \tag{5.154}$$

6. Proof of Proposition 5.3. Suppose $|g(r,\eta)| \le \min\{e^{-br}, e^{-\beta|\eta|}\}$ for certain $b > 0$ and $\beta \in (\frac{3}{4}\beta_0, \beta_0)$. Let

$$\begin{aligned}\theta(r) &= \langle v, \varphi_\eta\rangle_\eta, \\ \psi(r,\eta) &= v - \langle v, \varphi_\eta\rangle_\eta\varphi_\eta,\end{aligned}$$

and

$$\begin{aligned}g^\theta(r) &= \langle g, \varphi_\eta\rangle_\eta, \\ g^\psi(r,\eta) &= g - \langle g, \varphi_\eta\rangle_\eta\varphi_\eta.\end{aligned}$$

Then (5.9) is equivalent to the equations

$$\theta_{rr} = g^\theta, \tag{6.1}$$

$$\psi_{rr} + E\psi = g^\psi, \tag{6.2}$$

with $|\theta(r)|, |\psi(r,\eta)| \to 0$ as $r, |\eta| \to +\infty$, where

$$E\psi = \psi_{\eta\eta} + f'(\varphi)\psi$$

for

$$\psi \in D(E) = \{\psi \in H^2(\boldsymbol{R}^1);\ \langle\psi, \varphi_\eta\rangle_\eta = 0\}.$$

Let $\|v\|_\eta = \sqrt{\langle v, v\rangle_\eta}$. First, we obtain θ from (6.1) as

$$\theta(r) = \int_0^r\int_0^q g^\theta(s)dsdq - \int_0^\infty\int_0^r g^\theta(q)dqdr.$$

LEMMA 6.1. *The inequality*

$$|\theta(r)|_{C^2(r)} \le Be^{-b'r} \tag{6.3}$$

holds for $b' \in (0,b)$.

Proof. Since $\|\varphi_\eta\|_\eta^2 = M_1$ and

$$
\begin{aligned}
|g^\theta(r)| &\le \| g(r,\cdot) \|_\eta \cdot \| \varphi_\eta \|_\eta \\
&\le \sqrt{M_1} \left(\int_{-\infty}^{\infty} (\min\{e^{-br}, e^{-\beta|\eta|}\})^2 d\eta \right)^{\frac{1}{2}} \\
&= \sqrt{M_1} \left(\frac{2br+1}{\beta} \right)^{\frac{1}{2}} e^{-br} \\
&\le B e^{-b'r}
\end{aligned} \tag{6.4}
$$

for $b' \in (0, b)$, we have the inequality

$$
|\theta_{rr}(r)| \le B e^{-b'r} \tag{6.5}
$$

directly from (6.1) and (6.4).

On the other hand, θ_r is estimated as

$$
\begin{aligned}
|\theta_r(r)| &= \left| \int_0^r g^\theta(r) dr \right| \\
&= \left| \int_r^\infty g^\theta(r) dr \right| \\
&\le \int_r^\infty |g^\theta(r)| dr \\
&\le B e^{-b'r}
\end{aligned} \tag{6.6}
$$

by using (5.8) and (6.4). Therefore, by (6.6), we have

$$
\begin{aligned}
|\theta(r)| &= \left| \int_0^r \theta_r(r) dr + \theta(0) \right| \\
&= \left| \int_0^r \theta_r(r) dr - \int_0^\infty \theta_r(r) dr \right| \\
&\le \int_r^\infty |\theta_r(r)| dr \\
&\le B e^{-b'r}.
\end{aligned} \tag{6.7}
$$

Inequalities (6.5), (6.6), and (6.7) complete the proof. □

Next, consider the equation (6.2). Let $\beta' \in (\frac{3}{4}\beta_0, \beta)$, and let $\omega(\eta)$ be a positive smooth function satisfying

$$
\begin{aligned}
&\omega(\eta) = e^{\beta'|\eta|}, \ |\eta| \ge A_4, \\
&\omega(\eta) \ge A_5, \ \eta \in \boldsymbol{R}^1
\end{aligned}
$$

for some positive constants A_4 and A_5. Define the Banach space X with a weighted sup-norm by

$$
X = \left\{ \psi \in C^0(\boldsymbol{R}^1) \cap D(E) \ ; \ \| \psi \|_\infty = \sup_\eta |\psi(\eta)\omega(\eta)| < \infty \right\}.
$$

Let $\mathcal{F}$ be the Fourier transformation on $(0, \infty)$ defined by

$$
(\mathcal{F}\psi)(\xi) = \sqrt{\frac{2}{\pi}} \int_0^\infty \cos \xi r \psi(r) dr.
$$

LEMMA 6.2. *The equation* (6.2) *has a solution* $\psi(r,\eta)$ *satisfying*

$$|\psi(r,\eta)|_{C^{0,2}(r,\eta)} \le Be^{-\beta'|\eta|}. \tag{6.8}$$

Proof. Operating $\mathcal{F}$ on both sides of (6.2), we have

$$(E-\xi^2)(\mathcal{F}\psi) = \mathcal{F}g^{\psi}. \tag{6.9}$$

Since

$$\begin{aligned} |g^{\psi}(r,\eta)| &\le |g(r,\eta)| + |g^{\theta}(r)||\varphi_{\eta}(\eta)| \\ &\le \min\{e^{-br}, e^{-\beta|\eta|}\} + Be^{-b'r}e^{-\beta_0|\eta|} \end{aligned} \tag{6.10}$$

by (6.4), we have

$$\begin{aligned} |(\mathcal{F}g^{\psi}(\xi,\cdot))(\eta)| &\le \int_0^{\infty} |g^{\psi}(r,\eta)|dr \\ &\le \int_0^{\infty} \left(\min\{e^{-br}, e^{-\beta|\eta|}\} + Be^{-b'r}e^{-\beta_0|\eta|}\right) dr \\ &\le \frac{\beta|\eta|+1}{b}e^{-\beta|\eta|} + Be^{-\beta_0|\eta|} \\ &\le Be^{-\beta'|\eta|}. \end{aligned}$$

Hence $\mathcal{F}g^{\psi}(\xi,\cdot) \in X$ and the equation (6.9) is solvable with respect to $\mathcal{F}\psi(\xi,\cdot)$ in X with $\| \frac{\partial^j}{\partial\eta^j}\mathcal{F}\psi(\xi,\cdot) \|_{\infty} \le \frac{B}{1+\xi^2}$ $(j=0,1,2)$. This implies the existence of ψ satisfying

$$|\psi(r,\eta)|_{C^{0,2}(r,\eta)} \le Be^{-\beta'|\eta|}, \tag{6.11}$$

because $\mathcal{F}(\mathcal{F}\psi) = \psi$. □

Let

$$L^2_{\varphi_\eta} = \left\{\psi \in L^2(\boldsymbol{R}^1)\ ;\ \langle\ \psi, \varphi_\eta\ \rangle_\eta = 0\right\}$$

with the norm $\|\psi\|_\eta = \sqrt{\langle\psi,\psi\rangle_\eta}$. Then E is self-adjoint in $L^2_{\varphi_\eta}$ and there exists $\underline{\lambda} > 0$ such that the spectrum of $-E$ is contained in $[\underline{\lambda}, \infty)$. Therefore, there exists a partition of unity $\{E(\lambda)\}$ with respect to $-E$ such that E is represented as

$$E = -\int_{\underline{\lambda}}^{\infty} \lambda dE(\lambda).$$

Let

$$g(\lambda; r)(\cdot) = E(\lambda)g^{\psi}(r,\cdot),$$

and let

$$\begin{aligned} \psi(\lambda; r) = \frac{1}{2}e^{-\Lambda r}\psi^0(\lambda) &- \frac{1}{2\Lambda}e^{\Lambda r}\int_r^{\infty} e^{-\Lambda s}g(\lambda; s)ds \\ &- \frac{1}{2\Lambda}e^{-\Lambda r}\int_0^{r} e^{\Lambda s}g(\lambda; s)ds, \end{aligned} \tag{6.12}$$

$$\psi(r,\cdot) = \int_{\underline{\lambda}}^{\infty} d\psi(\lambda; r) \tag{6.13}$$

for $\lambda \geq \underline{\lambda}$, where

$$\psi^0(\lambda) = -\frac{1}{\Lambda}\int_0^\infty e^{-\Lambda s} g(\lambda; s) ds$$

and $\Lambda = \sqrt{\lambda}$. Then ψ defined in (6.13) satisfies (6.2) because $g^\psi(r)(\cdot) = \int_{\underline{\lambda}}^\infty dg(\lambda; r)$ and the equation

$$\frac{d^2}{dr^2}\psi(\lambda; r) - \lambda\psi(\lambda; r) = g(\lambda; r) \quad (\lambda \geq \underline{\lambda}) \tag{6.14}$$

are satisfied.

In the following, we will estimate $\psi(r, \eta)$ given in (6.13). Note that

$$|g^\psi(r, \eta)|^2 \leq B(1 + r)e^{-2br} \tag{6.15}$$

by (6.4) and (6.10). Since the inequality $b' < b$ holds, it follows from (6.15) that

$$\begin{aligned} \langle\, e^{b'r} g^\psi(r, \cdot), e^{b'r} g^\psi(r, \cdot)\,\rangle_\eta &= e^{2b'r}\,\langle\, g^\psi(r, \cdot), g^\psi(r, \cdot)\,\rangle_\eta \\ &= e^{2b'r} B(1 + r)e^{-2br} \\ &\leq Be^{-2(b''-b')r} \end{aligned}$$

for some $b'' \in (b', b)$. Hence

$$\begin{aligned} \langle\, e^{b'r} g^\psi(r, \cdot), e^{b'r} g^\psi(r, \cdot)\,\rangle_\eta &= \|\bar{g}(r)\|_\eta^2 \\ &= \int_{\underline{\lambda}}^\infty d\,\langle\, E(\lambda)\bar{g}, \bar{g}\,\rangle_\eta \\ &= \int_{\underline{\lambda}}^\infty d\|\bar{g}(\lambda; r)\|_\eta^2 \\ &\leq Be^{-2(b''-b')r}, \end{aligned} \tag{6.16}$$

where $\bar{g}(r)(\cdot) = e^{b'r} g^\psi(r, \cdot)$ and $\bar{g}(\lambda; r) = E(\lambda)\bar{g}(r)(\cdot)$. Here we fix a constant b_0 with

$$0 < b_0 < \underline{\Lambda},$$

where $\underline{\Lambda} = \sqrt{\underline{\lambda}}$, and suppose the inequalities $0 < b' < b < b_0$.

LEMMA 6.3. *The inequality*

$$\|\,\psi(r, \cdot)\,\|_{H^1(\boldsymbol{R}^1)} \leq Be^{-b'r} \tag{6.17}$$

holds.

Proof. By (6.12) and (6.16) $d\psi(\lambda; r)$ is estimated as

$$\begin{aligned} &d\,\langle\, E(\lambda)\psi, \psi\,\rangle_\eta = d\|\psi(\lambda; r)\|_\eta^2 \\ &\leq B\left\{e^{-2\Lambda r} d\|\psi^0(\lambda; r)\|_\eta^2 + \frac{1}{\Lambda^2} e^{2\Lambda r}\left(\int_r^\infty e^{-\Lambda s} d\|g(\lambda; s)\|_\eta ds\right)^2\right. \\ &\qquad\left. + \frac{1}{\Lambda^2} e^{-2\Lambda r}\left(\int_0^r e^{\Lambda s} d\|g(\lambda; s)\|_\eta ds\right)^2\right\} \end{aligned}$$

$$
\begin{aligned}
&\le B\left\{e^{-2\Lambda r}\cdot\frac{1}{\Lambda^2}\int_0^\infty e^{-2\Lambda s}ds\int_0^\infty d\|g(\lambda;s)\|_\eta^2 ds\right.\\
&\qquad\qquad +\frac{1}{\Lambda^2}e^{2\Lambda r}\int_r^\infty e^{-2\Lambda s}ds\int_r^\infty d\|g(\lambda;s)\|_\eta^2 ds\\
&\qquad\qquad \left.+\frac{1}{\Lambda^2}e^{-2\Lambda r}\int_0^r e^{2(\Lambda-b')s}ds\int_0^r d\|\bar g(\lambda;s)\|_\eta^2 ds\right\}\\
&=\frac{B}{\Lambda^2}\left\{\frac{1}{2\Lambda}e^{-2\Lambda r}\int_0^\infty d\|g(\lambda;s)\|_\eta^2 dr+\frac{1}{2\Lambda}\int_r^\infty d\|g(\lambda;s)\|_\eta^2 ds\right.\\
&\qquad\qquad \left.+\frac{1}{2(\Lambda-b')}\left(e^{-2b'r}-e^{-2\Lambda r}\right)\int_0^r d\|\bar g(\lambda;s)\|_\eta^2 ds\right\}\\
&\le\frac{B}{\lambda}\left\{\frac{1}{2\underline{\Lambda}}e^{-2\underline{\Lambda}r}\int_0^\infty d\|g(\lambda;s)\|_\eta^2 ds+\frac{1}{2\underline{\Lambda}}\int_r^\infty d\|g(\lambda;s)\|_\eta^2 ds\right.\\
&\qquad\qquad \left.+\frac{1}{2(\underline{\Lambda}-b')}e^{-2b'r}\int_0^r d\|\bar g(\lambda;s)\|_\eta^2 ds\right\}
\end{aligned}
$$

for some B, where we used the relation $g^\psi(r)=e^{-b'r}\bar g(r)$. Thus, by (6.16), we have

$$
\begin{aligned}
&\langle -E\psi(r,\cdot),\psi(r,\cdot)\rangle_\eta\\
&=\int_{\underline{\lambda}}^\infty \lambda d\langle E(\lambda)\psi,\psi\rangle_\eta\\
&=\int_{\underline{\lambda}}^\infty \lambda d\|\psi(\lambda;r)\|_\eta^2\\
&\le B\left\{e^{-2\underline{\Lambda}r}\int_0^\infty\int_{\underline{\lambda}}^\infty d\|g(\lambda;s)\|_\eta^2 ds+\int_r^\infty\int_{\underline{\lambda}}^\infty d\|g(\lambda;s)\|_\eta^2 ds\right.\\
&\qquad\qquad \left.+e^{-2b'r}\int_0^r\int_{\underline{\lambda}}^\infty d\|\bar g(\lambda;s)\|_\eta^2 ds\right\}\\
&\le B\left(e^{-2\underline{\Lambda}r}\int_0^\infty\|g^\psi(s)\|_\eta^2 ds+\int_r^\infty\|g^\psi(s)\|_\eta^2 ds+e^{-2b'r}\int_0^r\|\bar g(s)\|_\eta^2 ds\right)\\
&\le B\left(e^{-2\underline{\Lambda}r}\int_0^\infty e^{-2b''s}ds+\int_r^\infty e^{-2b''s}ds+e^{-2b'r}\int_0^r e^{-2(b''-b')s}ds\right)\\
&\le B\left(e^{-2\underline{\Lambda}r}+e^{-2b''r}+e^{-2b'r}\right)\\
&\le Be^{-2b'r}.
\end{aligned}
$$

Since

$$\langle -E\psi(r,\cdot),\psi(r,\cdot)\rangle_\eta\ge\underline{\lambda}\,\|\,\psi(r,\cdot)\,\|_\eta^2,$$

the above inequality yields (6.17). □

It follows from Lemma 6.3 that

$$|\psi(r,\eta)|\le Be^{-b'r}.$$

Hence, by Lemma 6.2, we get

$$|\psi(r,\eta)|\le B\min\{e^{-b'r},e^{-\beta'|\eta|}\}$$

for some B. Thus the inequalities (6.3) and (6.8) imply that a solution $v(r,\eta) = \theta\varphi_\eta + \psi$ of (5.9) satisfies

$$|v(r,\eta)| \le B\min\{e^{-b'r}, e^{-\beta'|\eta|}\}. \tag{6.18}$$

By using the inequalities (6.18) and (6.8) again, we get

$$|v_{rr}(r,\eta)| \le Be^{-\beta'|\eta|} \tag{6.19}$$

by using (5.9).

On the other hand, equation (5.9) is rewritten as

$$v_{rr} + v_{\eta\eta} + f'(\varphi(r))v = h(r,\eta), \tag{6.20}$$

where

$$h(r,\eta) = f'(\varphi(r))v(r,\eta) - f'(\varphi(\eta))v(r,\eta) + g(r,\eta).$$

We can estimate h as

$$|h(r,\eta)| \le B\min\{e^{-b'r}, e^{-\beta'|\eta|}\} \tag{6.21}$$

by (6.18). Hence we can apply the same argument as above to (6.20) by exchanging the role of r and η. Consequently, we obtain

$$\begin{array}{rcl} |v(r,\eta)|_{C^{2,0}(r,\eta)} & \le & Be^{-b'''r}, \\ |v_{\eta\eta}(r,\eta)| & \le & Be^{-b'''r} \end{array} \tag{6.22}$$

for some $b''' \in (0,b')$.

Thus, we have

$$\begin{array}{rcl} |v(r,\eta)|_{C^{2,0}(r,\eta)} & \le & B\min\{e^{-b'''r}, e^{-\beta''|\eta|}\}, \\ |v(r,\eta)|_{C^{0,2}(r,\eta)} & \le & B\min\{e^{-b'''r}, e^{-\beta''|\eta|}\} \end{array} \tag{6.23}$$

for some $\beta'' \in (\frac{3}{4}\beta_0, \beta')$ by using (6.3), (6.8), (6.18) and (6.19), (6.22) and the same argument as (6.6). The rest of the proof is easily obtained by using the Poisson Kernel on the upper half-plane. □

Acknowledgments. Both of the authors thank undergraduate students of the first author K. Aoyama, Y. Kuroda, E. Okutu, and S. Tanaka, who carried out numerical computations in Fig. 1.2.

REFERENCES

[1] N. D. Alikakos, G. Fusco, and M. Kowalczyk, *Finite Dimensional Dynamics and Interfaces Intersecting the Boundary* I, preprint.

[2] L. Bronsard and R. V. Kohn, *Motion by the mean curvature as the singular limit of Ginzburg-Landau dynamics*, J. Differential Equations, 90 (1992), pp. 211–237.

[3] J. Carr and R. L. Pego, *Metastable patterns in solutions of $u_t = \varepsilon^2 u_{xx} - f(u)$*, Comm. Pure Appl. Math., XLII (1989), pp. 523–576.

[4] X. Chen, *Generation and propagation of interfaces for reaction-diffusion equations*, J. Differential Equations, 33 (1991), pp. 749–786.

[5] P. DeMottoni and M. Schatzman, *Geometrical evolution of developed interfaces*, Trans. Amer. Math. Soc., 347 (1995), pp. 1533–1589.

[6] L. C. Evans, H. M. Soner, and P. E. Souganidis, *Phase transition and generalized motion by mean curvature*, Comm. Pure Appl. Math., 45 (1992), pp. 1097–1123.
[7] G. Fusco and J. K. Hale, *Slow-motion manifolds, dormant instability, and singular perturbations*, J. Dynam. Differential Equations, 1 (1989), pp. 75–94.
[8] R. V. Kohn and P. Sternberg, *Local minimizers and singular perturbations*, Proc. Roy. Soc. Edinburgh, 111A (1989), pp. 69–84.
[9] M. Kowalczyk, private communication, 1996.
[10] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. Res. Inst. Math. Sci. Kyoto Univ., 15 (1979), pp. 401–454.
[11] N. C. Owen and P. Sternberg, *Gradient flow and front propagation with boundary contact energy*, Proc. Roy. Soc. London Ser. A, 437 (1992), pp. 715–728.
[12] J. Rubinstein, P. Sternberg, and J. B. Keller, *Fast reaction, slow diffusion and curve shortening*, SIAM J. Appl. Math., 49 (1989), pp. 116–133.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 596–599, May 1998

003

AN EXTENSION OF MARCHIORO'S BOUND ON THE GROWTH OF A VORTEX PATCH TO FLOWS WITH L^p VORTICITY*

M. C. LOPES FILHO† AND H. J. NUSSENZVEIG LOPES†

Abstract. We observe that C. Marchioro's cubic-root bound in time on the growth of the diameter of a patch of vorticity [*Comm. Math. Phys*, 164 (1994), pp. 507–524] can be extended to incompressible two-dimensional Euler flows with compactly supported initial vorticity in L^p, $p > 2$, and with a distinguished sign.

Key words. L^p vorticity, two-dimensional incompressible flow, support of vorticity

AMS subject classifications. 35Q35, 76C05

PII. S0036141096310910

Let $\omega_0 \in L^p_c(\mathbb{R}^2)$, $p \geq 1$, be a compactly supported function. It was first shown by A. Majda (see [2] and references therein) that a weak solution to the two-dimensional inviscid, incompressible vorticity equation with ω_0 as initial data exists if $p > 4/3$. An extension of this result to all $p \geq 1$ is a consequence of the work of J.-M. Delort in [1], as was observed by S. Schochet in [4]. There are, however, very few results on weak solutions to two-dimensional Euler beyond existence.

The purpose of this note is to extend to a nonnegative initial vorticity ω_0 in L^p_c, $p > 2$, the $\mathcal{O}(t^{1/3})$ bound on the diameter of the support of $\omega(\cdot, t)$ obtained previously by C. Marchioro [3] for $\omega_0 \in L^\infty$. The exponent $p = 2$ is precisely the exponent for which velocity is no longer a priori bounded.

Our strategy is as follows. Fix $p > 2$ and $\omega_0 \in L^p_c$ a nonnegative, compactly supported function. Assume that $\operatorname{supp}(\omega_0) \subset\subset B_{R_0}$, the ball of radius R_0, centered at the origin. Let $\omega_0^\varepsilon \in C^\infty_c$ be a sequence of nonnegative functions, obtained by regularizing ω_0, so that $\operatorname{supp}(\omega_0^\varepsilon) \subset\subset B_{R_0}$. Let $\omega^\varepsilon = \omega^\varepsilon(x,t)$ be the sequence of smooth solutions of the two-dimensional vorticity equation, with initial data ω_0^ε. By the results proven in [1], [2], [4] there exists a subsequence of ω^ε converging weakly to a weak solution. Let $\omega = \omega(x,t)$ be a global weak solution obtained as the weak limit of such a subsequence. We will show that the support of ω^ε is contained in the disk of radius $r_t = (R_0^3 + b_1 t)^{1/3}$ for some positive constant b_1, independent of ε, thereby implying that the support of ω is contained in the same disk. Our result has the nature of an a priori estimate on any weak solution obtained by the process of regularization of initial data as described above. Hereafter we will omit the superscript ε.

Let $u = u(x,t)$ be the incompressible velocity field induced by the vorticity ω, given by $u = K * \omega$, the Biot–Savart law. We show below that, although the velocity field is only locally $W^{1,p}$, a simple estimate gives a global L^∞ bound. We will denote $p' = p/(p-1)$, the conjugate Lebesgue exponent, throughout.

LEMMA 1. *We have* $\|u\|_{L^\infty} \leq C_p \|\omega_0\|_{L^p} + (2\pi)^{-1} \int \omega_0$.

Proof. We estimate directly

$$|u(x,t)| \leq \int_{|x-y|\leq 1} (2\pi)^{-1} |x-y|^{-1} \omega(y,t)dy + \int_{|x-y|>1} (2\pi)^{-1} |x-y|^{-1} \omega(y,t)dy$$

*Received by the editors October 22, 1996; accepted for publication (in revised form) April 9, 1997. The research of the first author was supported in part by CNPq grant 300962/91-6, and the research of the second author was supported in part by CNPq grant 300158/93-9.

http://www.siam.org/journals/sima/29-3/31091.html

†Departamento de Matematica, IMECC-UNICAMP, Caixa Postal 6065, Campinas, SP 13081-970, Brasil (mlopes@ime.unicamp.br, hlopes@ime.unicamp.br).

$$\leq \|\omega(\cdot,t)\|_{L^p(B_1(x))} \, \| \, |y|^{-1} \, \|_{L^{p'}(B_1(0))} + (2\pi)^{-1} \int \omega_0.$$

Take $C_p = \| \, |y|^{-1} \, \|_{L^{p'}(B_1(0))} < \infty$, since $p > 2$. The estimate follows, since the L^p-norm of vorticity is conserved. □

Before we state and prove our theorem, some comments on Marchioro's proof of the L^∞ result are in order. The basic issue in the proof of Theorem 2.1 in [3] is to carefully estimate the radial velocity field at a point far from the center of motion. This is performed by decomposing velocity into the portions generated by neighboring vorticity, referred to as near-field velocity, and by vorticity remaining near the center of motion, the far-field velocity. The difficult part of this problem is to estimate the near-field.

The heart of Marchioro's argument is to obtain exponential decay of the mass of vorticity relevant for the near-field estimate; this is encoded in [3, eq. (2.64)]. Nevertheless, this estimate is still a far-field calculation, which means that it is insensitive to the unboundedness of vorticity. Finally, the near-field estimate is performed using the technique in [3, eq. (2.29)]. It is this final step that needs modification in order to extend Marchioro's result to unbounded vorticities. We will use Lemma 2 to estimate the near-field. As in [3], $m_t(r)$ denotes the mass of vorticity outside the disk of radius r at time t.

LEMMA 2. *Let $\beta \in (p'/2, 1)$. Then there exists a constant C such that*

$$\left| \int_{|y| \geq R} K(x-y)\, \omega(y,t) dy \right| \leq C (m_t(R))^{1-\beta} \|\omega_0\|_{L^p}^{\beta}.$$

Moreover, C depends only on β, p, and the Lebesgue measure of the support of the initial vorticity, $|supp\ (\omega_0)|$. In addition, $C = \mathcal{O}(1/(p-2))$ as p approaches 2.

Proof. Let B_R^c denote the set $\{|y| \geq R\}$. Write $\omega = \omega^{1-\beta}\omega^{\beta}$ and estimate directly

$$\left| \int_{B_R^c} K(x-y)\, \omega(y,t) dy \right| \leq (m_t(R))^{1-\beta} \left(\int_{B_R^c} |K(x-y)|^{1/\beta}\, \omega(y,t) dy \right)^{\beta}$$

$$\leq (m_t(R))^{1-\beta} \|\omega_0\|_{L^p}^{\beta} \left(\int_{\text{supp } (\omega(\cdot,t))} |K(x-y)|^{p'/\beta} dy \right)^{\beta/p'}.$$

Above, we have used Hölder's inequality twice and the conservation of the L^p-norm of vorticity.

The proof is concluded once we find an upper bound for

$$\left(\int_{\text{supp } (\omega(\cdot,t))} |K(x-y)|^{p'/\beta} dy \right)^{\beta/p'}. \tag{1}$$

The condition $p'/2 < \beta < 1$ is used to guarantee that (1) is finite.

By incompressibility, $|\text{supp } (\omega(\cdot,t))|$ is constant and equal to $|\text{supp } (\omega_0)|$. We adapt the idea in [3, eqs. (2.29),(2.30)] to obtain

$$(1) \leq \left(\frac{1}{2\pi} \int_{B_\eta} \left(\frac{1}{|z|} \right)^{p'/\beta} dz \right)^{\beta/p'}.$$

The radius η is chosen so that $\pi\eta^2 = |\text{supp }(\omega_0)|$. Denote $q = 2\beta(p-1) - p$.

Hence, (1) is bounded above by

$$\eta^{q/p}\left(\frac{\beta(p-1)}{q}\right)^{\beta/p'} \equiv C(\beta, p, |\text{supp }(\omega_0)|),$$

and, clearly, this constant C is $\mathcal{O}(1/(p-2))$ as $p \to 2$, as we wanted. □

THEOREM 3. *There exists a constant $b_1 = b_1(p, R_0, \|\omega_0\|_{L^p}) > 0$ such that the diameter of the support of $\omega(\cdot, t)$ is at most $2(R_0^3 + b_1 t)^{1/3}$ for $t \geq 0$.*

Proof. In this proof we will mention only those portions of the proof of Theorem 2.1 in [3] which need to be changed.

We begin by using Lemma 1 to ensure the existence of $t^* > 0$ so that the support of vorticity is contained in the disk of radius $r_t = (R_0^3 + b_1 t)^{1/3}$, for $0 \leq t \leq t^*$, for some positive b_1.

All subsequent arguments and estimates in Marchioro's proof are far-field estimates up to [3, eq. (2.64)] and can be adapted to the L^p case in a straightforward manner: simply substitute $K(x-y)$ by $K(x-y)\omega(y,t)$ whenever it appears.

Marchioro's argument consists of estimating the radial velocity at a point x, with $|x| = r_t$. This is done by decomposing the disk of radius r_t into a union of annuli: $\{a_{k-1} \leq |y| < a_k\}$, $1 \leq k \leq k^*$, and $\{a_{k^*} \leq |y| < r_t\}$. Here, $a_0 = 0$, $a_1 = R_0$, $a_k = 2a_{k-1}$, and k^* is chosen so that $a_{k^*+1} \leq r_t < a_{k^*+2}$. We restrict our attention to the near-field velocity, generated by vorticity outside the disk of radius a_{k^*}. Recall that $n = 2^{k^*-1} - 1$ and fix $\beta \in (p'/2, 1)$.

Substitute estimate [3, eq. (2.65)] by the following:

$$m_t(a_{k^*}) < C^n b_1^{-n} < \bar{C} n^{-2M}, \tag{2}$$

where $M = (1-\beta)^{-1}$. This is possible by choosing b_1 sufficiently large. Observe that b_1 blows up exponentially as $p \to 2$. Thus,

$$m_t(a_{k^*}) \leq C a_{k^*}^{-2M} \leq C r_t^{-2M}. \tag{3}$$

Finally, consider estimate [3, eq. (2.66)]. This is Marchioro's near-field estimate, which we substitute by Lemma 2, at $R = a_{k^*}$:

$$\left|\int_{|y|\geq a_{k^*}} K(x-y)\,\omega(y,t)dy\right| \leq C(m_t(a_{k^*}))^{1-\beta}\|\omega_0\|_{L^p}^{\beta} \leq \tilde{C} r_t^{-2M(1-\beta)} = \tilde{C} r_t^{-2}.$$

This concludes the proof. □

This result raises the question of what happens with more singular vorticity, such as L^p-vorticity, $1 \leq p \leq 2$ or even vortex sheets, keeping the distinguished sign restriction. Since velocity is no longer bounded, it could happen that the support of vorticity escapes to infinity instantly. This will be the object of forthcoming work.

Acknowledgments. The authors are grateful to S. Schochet for helpful comments. The authors also thank the referee for pointing out a substantial simplification of our original argument.

REFERENCES

[1] J.-M. Delort, *Existence de nappes de tourbillons en dimension deux*, J. Amer. Math. Soc., 4 (1991), pp. 553–586.
[2] A. Majda, *Vorticity and the mathematical theory of incompressible fluid flow*, Comm. Pure Appl. Math., 39 (1986), pp. S187–S220.
[3] C. Marchioro, *Bounds on the growth of the support of a vortex patch*, Comm. Math. Phys., 164 (1994), pp. 507–524.
[4] S. Schochet, *The weak vorticity formulation of the* 2*D Euler equations and concentration-cancellation*, Comm. Partial Differential Equations, 20 (1995), pp. 1077–1104.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 600–618, May 1998

004

NONLINEAR INSTABILITY OF A PRECESSING BODY WITH A CAVITY FILLED BY AN IDEAL FLUID*

ANDREI A. LYASHENKO† AND SUSAN J. FRIEDLANDER‡

Abstract. A sufficient condition is derived for the nonlinear instability of a rotating body with a fluid-filled cavity. This condition is given in terms of quantities completely determined by the shape of the body/cavity and the relative density of the fluid and the body. It is shown, for example, that such a system with a prolate ellipsoidal cavity of appropriate ellipticity is nonlinearly unstable.

Key words. nonlinear instability, precession, rotating ideal fluids

AMS subject classifications. 76EXX, 35

PII. S0036141096302160

Introduction. We consider the stability of a rigid body with a cavity completely filled with an inviscid, incompressible fluid. The entire system initially rotates with constant angular velocity about an axis through its center of mass. The angular velocity of the rigid body is free to vary with time (i.e., the body may precess) and the fluid moves under the constraints of incompressibility and is subject to the condition that there are no velocity components normal to the boundary of the cavity. Such a system gives a model for an astronomical body with a rigid crust surrounding a liquid core, and there is a long history of examining the mathematics of such fluid-body systems (eg., [6, 5, 12]). More recent studies have been inspired by the question of stability of projectiles with fuel-filled cavities and the mathematical theory of rotating fluids (eg., [14, 15, 4]). In this present paper we give a sufficient condition for nonlinear instability of such a fluid-body system.

Recently Friedlander, Strauss, and Vishik [3] proved the following theorem for rather general nonlinear evolution partial differential equations (PDEs). It is shown that, under appropriate conditions, instability of the linearized operator implies nonlinear instability. The crucial idea underlying this theorem is to use two Banach spaces: a large space Z where the spectrum of the linearized operator is studied and a "small" space $X \hookrightarrow Z$ where a local existence theorem for the nonlinear equation can be proved. The method of proof of this theorem utilizes a projection onto the subspaces of growing and decaying modes. Such a decomposition requires the existence of a suitable "gap" in the spectrum of the linearized operator. This spectral gap condition may be hard to verify for a PDE where the unstable spectrum of the linearized operator has both continuous and discrete parts (eg., the Euler equations for an ideal fluid perturbed about an arbitrary steady state). However, any problem for which the unstable spectrum is nonempty and purely discrete automatically satisfies the spectral gap condition.

For the equations governing the motion of a precessing body with a fluid-filled cavity Ω, we verify the conditions of the nonlinear instability theorem of [3]. We prove

*Received by the editors April 15, 1996; accepted for publication (in revised form) December 12, 1996. The second author was supported by NSF grants DMS 95-00466 and DMS 93-00752.
http://www.siam.org/journals/sima/29-3/30216.html

†Department of Mathematics (M/C 249), 851 S. Morgan Street, University of Illinois-Chicago, Chicago, IL 60607 and Institute of Mathematics, Novosibirsk 630090, Russia (lyashenk@math.uic.edu).

‡Department of Mathematics (M/C 249), 851 S. Morgan Street, University of Illinois-Chicago, Chicago, IL 60607 (susan@math.nwu.edu).

a local existence theorem with the functional space X being $X_s \times \mathbb{R}^3$, where X_s is the space of divergence-free vectors, tangential to the boundary of Ω with components in the Sobolev space H^s, $s > 5/2$. We note that the "infinite dimensional" contribution to the nonlinearity comes from the nonlinearity in the fluid equation and the method of proof of local existence follows the lines of well-known proofs of local existence for the Euler equation (see, for example, [16]). We consider the spectrum in a space Z of the operator obtained from the equations linearized about a state of uniform rotation. The space Z is taken to be $J_0(\Omega) \times \mathbb{R}^3$, where $J_0(\Omega)$ is the space of divergence-free square integrable vectors tangential to the boundary of Ω. We study the unstable spectrum under the simplifying assumption that the body and the cavity have 4-fold symmetry about the axis of rotation (see [14, 10]). We obtain an explicit formula for the unstable eigenvalues of the linearized operator in terms of the zeros of an analytic function. It follows from this formula that

(1) all the unstable eigenvalues are discrete and for any $\epsilon > 0$ there exist at most a finite number of spectral points λ satisfying $|\mathrm{Re}\lambda| \geq \epsilon$;
(2) we have a necessary and sufficient condition for linear instability and a bound on $|\lambda|$ explicitly given in terms of the geometry of the configuration and the relative densities of the liquid and the body.

Since the unstable spectrum is purely discrete, it is straightforward to apply the nonlinear instability theorem. Hence we obtain the result that any fluid-body configuration for which the sufficient condition for linear instability holds is nonlinearly unstable in X. To our knowledge this is the first nonlinear instability result for this system.

We note that the property of discreteness of the unstable spectrum follows from the nature of the linearized operator and is valid independent of any geometrical symmetry constraints. Hence the nonlinear instability theorem of [3] can be invoked to show that any configuration that is linearly unstable is also nonlinearly unstable. The constraint of 4-fold symmetry is used to obtain the explicit formulas for the unstable eigenvalues. As a particular example, in the final section we use these formulas to compute necessary and sufficient conditions for linear instability for the case of an ellipsoidal cavity. We show, for example, that a spinning body with a prolate ellipsoidal cavity of appropriate ellipticity is nonlinearly unstable.

1. Equations of motion. We consider a rigid body G with a cavity Ω entirely filled with a homogeneous incompressible inviscid fluid. It is assumed that the boundary $\partial\Omega$ is smooth and that G and Ω satisfy a condition of 4-fold symmetry (following [14], a domain is said to satisfy a condition of k-fold symmetry if it is symmetric with respect to turning through an angle of $2\pi/k$ about the axis of symmetry). Let O be the center of mass of the entire "body+fluid" system and let Ox_1, x_2, x_3 be an orthogonal system rigidly connected with the body (Ox_3 is the axis of 4-fold symmetry). With respect to the center of mass O, the motion of the system is completely described by seven scalar functions: the fluid velocity,

$$\mathbf{u}(\mathbf{r}, t) = (u_1(\mathbf{r}, t), u_2(\mathbf{r}, t), u_3(\mathbf{r}, t)), \quad \mathbf{r} \in \Omega,\ t \geq 0,$$

the fluid "pressure" term,

$$p(\mathbf{r}, t), \quad \mathbf{r} \in \Omega,\ t \geq 0,$$

and the angular velocity of the body,

$$\mathbf{W}(t) = (W_1(t), W_2(t), W_3(t)), \quad t \geq 0.$$

The position vector $\mathbf{r}$ of a fluid particle in Ω has components (x_1, x_2, x_3). By the "pressure" we consider the scalar potential for the conservative forces acting on the fluid.

The motion is governed by the coupled system of equations: the Euler equations for the motion of the fluid in Ω and the equations for conservation of angular momentum of the entire "body+fluid" system. Following [7, 10], the equations of motion are written in the form

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} + 2\mathbf{W}\times\mathbf{u} + \frac{d\mathbf{W}}{dt}\times\mathbf{r} = -\nabla p, \tag{1.1}$$

$$\frac{d}{dt}\left(J\mathbf{W} + \int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right) + \mathbf{W}\times\left(J\mathbf{W} + \int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right) = 0, \tag{1.2}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{1.3}$$

$$\mathbf{u}\cdot\mathbf{n}|_{\partial\Omega} = 0, \tag{1.4}$$

where

$$\begin{cases} J = \mathrm{Diag}(a, a, b), \\ a = \displaystyle\int_\Omega (x_1^2 + x_3^2)\,d\Omega + \mu\int_G (x_1^2 + x_3^2)\,dG \\ b = 2\displaystyle\int_\Omega x_1^2\,d\Omega + 2\mu\int_G x_1^2\,dG, \end{cases} \tag{1.5}$$

and μ is the ratio of the body density to the fluid density.

The solenoidality condition of (1.3) can be replaced by writing (1.1) in operator form:

$$\frac{\partial \mathbf{u}}{\partial t} + B(\mathbf{u},\mathbf{u}) + 2P_0(\mathbf{W}\times\mathbf{u}) + P_0\left(\frac{d\mathbf{W}}{dt}\times\mathbf{r}\right) = 0, \tag{1.6}$$

where $P_0 : L^2(\Omega) \to L^2(\Omega)$ is the Weyl orthogonal projection onto the subspace $J_0(\Omega)$ of solenoidal vectors satisfying $\mathbf{u}\cdot\mathbf{n}|_{\partial\Omega} = 0$ and $B(\mathbf{u},\mathbf{v}) = P_0(\mathbf{u}\cdot\nabla)\mathbf{v}$.

We consider perturbations of the steady state of rigid body rotation about the axis Ox_3 with angular velocity $\omega_0 > 0$:

$$\mathbf{u}_0 = \mathbf{0}, \quad \mathbf{W}_0 = \omega_0\mathbf{e}_3; \quad \mathbf{W} = \mathbf{w} + \omega_0\mathbf{e}_3,$$

where $\mathbf{e}_3 = (0, 0, 1)$. The system (1.2), (1.6) is then written as follows:

$$\frac{\partial}{\partial t}\big(\mathbf{u} + P_0(\mathbf{w}\times\mathbf{r})\big) + 2i\omega_0 T\mathbf{u} + B(\mathbf{u},\mathbf{u}) + 2P_0(\mathbf{w}\times\mathbf{u}) = 0, \tag{1.7}$$

$$\begin{aligned} &\frac{d}{dt}\left(J\mathbf{w} + \int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right) + \omega_0\mathbf{e}_3\times J\mathbf{w} + \omega_0 b\mathbf{w}\times\mathbf{e}_3 + \omega_0\mathbf{e}_3\times\int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega \\ &\qquad + \mathbf{w}\times J\mathbf{w} + \mathbf{w}\times\int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega = 0, \end{aligned} \tag{1.8}$$

where $T\mathbf{u} = iP_0(\mathbf{u} \times \mathbf{e}_3)$ is a self-adjoint operator with spectrum $\sigma(T) = [-1, 1]$. Equations (1.7), (1.8) can be written in the form

$$\frac{\partial}{\partial t} K\mathbf{V} + \omega_0 L\mathbf{V} + N(\mathbf{V}) = 0, \tag{1.9}$$

where $\mathbf{V} = \binom{\mathbf{u}}{\mathbf{w}}$,

$$K\mathbf{V} = \begin{pmatrix} \mathbf{u} + P_0(\mathbf{w} \times \mathbf{r}) \\ J\mathbf{w} + \displaystyle\int_\Omega \mathbf{r} \times \mathbf{u}\, d\Omega \end{pmatrix}, \tag{1.10}$$

$$L\mathbf{V} = \begin{pmatrix} 2iT\mathbf{u} \\ \mathbf{e}_3 \times J\mathbf{w} + b\mathbf{w} \times \mathbf{e}_3 + \mathbf{e}_3 \times \displaystyle\int_\Omega \mathbf{r} \times \mathbf{u}\, d\Omega \end{pmatrix}, \tag{1.11}$$

$$N(\mathbf{V}) = \begin{pmatrix} B(\mathbf{u}, \mathbf{u}) + 2P_0(\mathbf{w} \times \mathbf{u}) \\ \mathbf{w} \times J\mathbf{w} + \mathbf{w} \times \displaystyle\int_\Omega \mathbf{r} \times \mathbf{u}\, d\Omega \end{pmatrix}. \tag{1.12}$$

In the Hilbert space $Z = J_0(\Omega) \times \mathbb{R}^3$ with the inner product

$$(\mathbf{V}_1, \mathbf{V}_2)_Z = \int_\Omega \mathbf{u}_1 \cdot \mathbf{u}_2\, d\Omega + \mathbf{w}_1 \cdot \mathbf{w}_2, \tag{1.13}$$

the operator K is self-adjoint and satisfies

$$c_1 \|\mathbf{V}\|_Z^2 \le (K\mathbf{V}, \mathbf{V})_Z \le c_2 \|\mathbf{V}\|_Z^2 \tag{1.14}$$

for some positive constants $c_1, c_2 > 0$ (see [7]). Hence the operator K is bounded and K^{-1} exists. Therefore (1.9) can be written in the form

$$\frac{\partial}{\partial t}\mathbf{V} + \omega_0 K^{-1} L\mathbf{V} + K^{-1} N(\mathbf{V}) = 0. \tag{1.15}$$

We will prove that, under appropriate conditions on Ω and μ, the steady state $\mathbf{V}_0 = \mathbf{0}$ (i.e., rigid body rotation $\mathbf{u}_0 = 0$, $\mathbf{W}_0 = \omega_0 \mathbf{e}_3$) is nonlinearly unstable.

2. General nonlinear instability theorem. Let us fix a pair of Banach spaces $X \hookrightarrow Z$ with a dense embedding. Consider a nonlinear evolution equation of the form

$$\frac{d}{dt}\mathbf{v} = L\mathbf{v} + N(\mathbf{v}), \quad \mathbf{v}(0) = \mathbf{v}_0, \tag{2.1}$$

where L is the generator of a C_0-group of operators $\mathcal{L}(Z)$, e^{Lt} leaves X invariant for $t \in \mathbb{R}$, $X \subset D(L)$, and N is a nonlinear operator

$$N : X \to Z. \tag{2.2}$$

We assume that the nonlinear term N satisfies the inequality

$$\begin{aligned} &\|N(\mathbf{v})\|_Z \le c_0 \|\mathbf{v}\|_X \|\mathbf{v}\|_Z \quad \text{for } \mathbf{v} \in X \text{ with} \\ &\|\mathbf{v}\|_X < \rho \text{ for some } \rho > 0. \end{aligned} \tag{2.3}$$

We assume that a local existence theorem holds for the nonlinear equation (2.1). This means that for any $\mathbf{v}_0 \in X$ there exists $T > 0$ and a unique

$$\mathbf{v}(t) \in L^\infty((0,T);X) \cap C([0,T];Z), \tag{2.4}$$

which is a solution to (2.1) in the following sense: for any $\phi \in D(0,T)$

$$\int_0^T \{\mathbf{v}(\tau)\phi'(\tau) + [L\mathbf{v}(\tau) + N(\mathbf{v}(\tau))]\phi(\tau)\}\, d\tau = 0. \tag{2.5}$$

The initial condition is assumed in the sense of strong convergence in Z:

$$\lim_{t\to 0^+} \|\mathbf{v}(t) - \mathbf{v}_0\|_Z = 0.$$

We consider the following definition of nonlinear stability/instability. The trivial solution $\mathbf{v}_0 = 0$ of equation (1.1) is called nonlinearly stable in X (i.e., Lyapunov stable) if no matter how small $\epsilon > 0$ is, there exists a $\delta > 0$ such that $\|\mathbf{v}(0)\|_X < \delta$ implies

(a) we can choose $T = \infty$ in (2.4) and

(b) $\|\mathbf{v}(t)\|_X < \epsilon$ for a.e. $t \in [0,\infty)$.

The trivial solution $\mathbf{v}_0 = 0$ is called nonlinearly unstable if it is not nonlinearly stable.

Remark. By this definition we regard a "blowing up" solution (i.e., there exists a maximal finite $T > 0$ in (2.4)) as a particular case of nonlinear instability.

The following theorem is proved by Friedlander, Strauss, and Vishik [3].

THEOREM 2.1. *Let the nonlinear equation* (2.1) *admit a local existence theorem as described above with N satisfying* (2.3). *Let the spectrum σ of $e^{Lt} \in \mathcal{L}(Z)$ be of the following structure:*

$$\sigma = \sigma(e^{Lt}) = \sigma_+ \cup \sigma_-, \quad \sigma_+ \neq \emptyset,$$

where

$$\sigma_+ \subset \{z \in \mathbb{C} \mid e^{Mt} < |z| < e^{\Lambda t}\},$$
$$\sigma_- \subset \{z \in \mathbb{C} \mid e^{\lambda t} < |z| < e^{\mu t}\}$$

with $-\infty < \lambda < \mu < M < \Lambda < \infty$ and $M > 0$. Then the trivial solution $\mathbf{v}_0 = 0$ to equation (2.1) *is nonlinearly unstable in X.*

We remark that any operator L for which the unstable spectrum is nonempty and purely discrete automatically satisfies the "spectral gap" condition stated above which is utilized in the proof of Theorem 2.1.

We will apply Theorem 2.1 to the nonlinear equation for the coupled fluid-body system given by (1.15). We choose the following spaces for X and Z:

$$X = X_s \times \mathbb{R}^3, \text{ where } X_s = \{\mathbf{u} \in (H^s(\Omega))^3 \mid \operatorname{div}\mathbf{u} = 0 \text{ in } \Omega;\ \mathbf{u}\cdot\mathbf{n}|_{\partial\Omega} = 0\},$$
$$Z = J_0(\Omega) \times \mathbb{R}^3, \text{ where } J_0(\Omega) = \{\mathbf{u} \in (L^2(\Omega))^3 \mid \operatorname{div}\mathbf{u} = 0 \text{ in } \Omega;\ \mathbf{u}\cdot\mathbf{n}|_{\partial\Omega} = 0\}.$$

In section 3 we will show that the nonlinear equation (1.15) admits a local existence theorem in X and that the operator $(K^{-1}N)$ given by (1.10), (1.12) satisfies the inequality (2.3). In section 4 we study the spectrum of the linear operator $(K^{-1}L)$ given by (1.10), (1.11). We prove that the unstable spectrum is purely discrete and we obtain a necessary and sufficient condition for the existence of an unstable eigenvalue. It therefore follows from Theorem 2.1 that geometries for which this condition holds are nonlinearly unstable.

3. General conditions. Let s be an integer with $s > 5/2$ and consider the pair of Hilbert spaces

$$Z = J_0(\Omega) \times \mathbb{R}^3, \quad (\mathbf{V}_1, \mathbf{V}_2)_Z = (\mathbf{u}_1, \mathbf{u}_2)_{\mathbf{L}^2} + \mathbf{w}_1 \cdot \mathbf{w}_2,$$
$$X = X_s \times \mathbb{R}^3, \quad (\mathbf{V}_1, \mathbf{V}_2)_X = (\mathbf{u}_1, \mathbf{u}_2)_{\mathbf{H}^s} + \mathbf{w}_1 \cdot \mathbf{w}_2,$$

where from now on we use the notations $\mathbf{L}^2 = (L^2(\Omega))^3$ and $\mathbf{H}^s = (H^s(\Omega))^3$.

PROPOSITION 3.1. *The operator $K^{-1}N$ given by* (1.10) *and* (1.12) *satisfies*

$$\|K^{-1}N(\mathbf{V})\|_Z \le \text{const}\, \|\mathbf{V}\|_X \|\mathbf{V}\|_Z \qquad \textit{for all } \mathbf{V} \in X.$$

Proof. Because of (1.10), (1.12), and (1.14),

$$\|K^{-1}N(\mathbf{V})\|_Z^2 \le \text{const}\, \|N(\mathbf{V})\|_Z^2$$
$$= \text{const} \left(\|P_0(\mathbf{u} \cdot \nabla)\mathbf{u} + 2P_0(\mathbf{w} \times \mathbf{u})\|_{\mathbf{L}^2}^2 + \left| \mathbf{w} \times J\mathbf{w} + \mathbf{w} \times \int_\Omega \mathbf{r} \times \mathbf{u}\, d\Omega \right|^2 \right)$$
$$\le \text{const} \left(\|(\mathbf{u} \cdot \nabla)\mathbf{u}\|_{\mathbf{L}^2}^2 + \|\mathbf{V}\|_Z^4 \right) \le \text{const}\, \|\mathbf{V}\|_X^2 \|\mathbf{V}\|_Z^2$$

due to the Sobolev embedding theorem ($s > 1 + 3/2$). ∎

In the proof of the following local existence theorem we follow the lines of Temam (see [16]).

THEOREM 3.1. *For any $\mathbf{V}_0 \in X$ there exist $T > 0$ and unique $\mathbf{V}(t) \in L^\infty((0,T); X) \cap C([0,T]; Z)$ satisfying* (1.15) *and $V(0) = V_0$.*

Proof. It suffices to prove this local existence result for the original problem (1.1)–(1.4). Following Temam [16] we first derive a priori estimates. Assume that $\mathbf{u}$, p, and $\mathbf{W}$ are sufficiently regular real-valued solutions to (1.1)–(1.4). Taking the $\mathbf{L}^2$-scalar product of (1.1) with $\mathbf{u}$ and adding it to the $\mathbb{R}^3$-scalar product of (1.2) with $\mathbf{W}$ we obtain

$$\frac{d}{dt} \left(\frac{1}{2} \|\mathbf{u}(t)\|_{\mathbf{L}^2}^2 + \frac{1}{2} J\mathbf{W}(t) \times \mathbf{W}(t) + \left(\int_\Omega \mathbf{r} \times \mathbf{u}(t)\, d\Omega \right) \cdot \mathbf{W}(t) \right) = 0. \tag{3.1}$$

It is easy to see that

$$\|\mathbf{u}\|_{\mathbf{L}^2}^2 + J\mathbf{W} \times \mathbf{W} + 2 \left(\int_\Omega \mathbf{r} \times \mathbf{u}\, d\Omega \right) \cdot \mathbf{W} = (K\mathbf{V}, \mathbf{V})_Z, \quad \mathbf{V} = \begin{pmatrix} \mathbf{u} \\ \mathbf{W} \end{pmatrix} \in Z. \tag{3.2}$$

Thus (3.1) and (3.2) imply

$$\frac{d}{dt} (K\mathbf{V}(t), \mathbf{V}(t))_Z = 0. \tag{3.3}$$

It follows from (1.14) and (3.3) that

$$\|\mathbf{V}(t)\|_Z^2 \le \frac{1}{c_1} (K\mathbf{V}(t), \mathbf{V}(t))_Z = \frac{1}{c_1} (K\mathbf{V}_0, \mathbf{V}_0)_Z \le \frac{c_2}{c_1} \|\mathbf{V}_0\|_Z^2. \tag{3.4}$$

We apply $D^\alpha_{\mathbf{r}}$ to (1.1), take the $\mathbf{L}^2$-scalar product with $D^\alpha_{\mathbf{r}}\mathbf{u}$, and sum for $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3 = s$. Since s is an integer satisfying $s > 5/2$, we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\mathbf{u}(t)\|_s^2 &= -[\nabla p, \mathbf{u}]_s - \sum_{|\alpha|=s} \left(D^\alpha_{\mathbf{r}}((\mathbf{u}\cdot\nabla)\mathbf{u}), D^\alpha_{\mathbf{r}}\mathbf{u}\right)_{\mathbf{L}^2} \\ &= -[\nabla p, \mathbf{u}]_s - \sum_{|\alpha|=s} \left(D^\alpha_{\mathbf{r}}((\mathbf{u}\cdot\nabla)\mathbf{u}) - (\mathbf{u}\cdot\nabla)D^\alpha_{\mathbf{r}}\mathbf{u}, D^\alpha_{\mathbf{r}}\mathbf{u}\right)_{\mathbf{L}^2},\end{aligned}$$

where

$$\|\mathbf{u}\|_s^2 = \sum_{|\alpha|=s} \|D^\alpha_{\mathbf{r}}\mathbf{u}\|^2_{\mathbf{L}^2}, \quad [\mathbf{u},\mathbf{v}]_s = \sum_{|\alpha|=s} (D^\alpha_{\mathbf{r}}\mathbf{u}, D^\alpha_{\mathbf{r}}\mathbf{v})_{\mathbf{L}^2}.$$

Therefore, by (1.14) of [16],

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{u}(t)\|_s^2 \le \|\nabla p(t)\|_{\mathbf{H}^s}\|\mathbf{u}(t)\|_s + \text{const}\,\|\mathbf{u}(t)\|_s^3. \tag{3.5}$$

In order to estimate $\|\nabla p(t)\|_{\mathbf{H}^s}$ we use (1.1). Applying div to (1.1), we obtain

$$\Delta p = -\sum_{i,j=1}^3 D_{x_i}u_j D_{x_j}u_i + 2(\nabla\times\mathbf{u})\cdot\mathbf{W}.$$

Taking the $\mathbb{R}^3$-inner product of the trace of (1.1) with the normal vector $\mathbf{n}$ we obtain

$$\begin{aligned}\frac{\partial p}{\partial \mathbf{n}} &= -2\mathbf{W}\cdot(\mathbf{u}\times\mathbf{n}) - \frac{d\mathbf{W}}{dt}\cdot(\mathbf{r}\times\mathbf{n}) - ((\mathbf{u}\cdot\nabla)\mathbf{u})\cdot\mathbf{n} \\ &= -2\mathbf{W}\cdot(\mathbf{u}\times\mathbf{n}) - \frac{d\mathbf{W}}{dt}\cdot(\mathbf{r}\times\mathbf{n}) + \sum_{i,j=1}^3 u_i u_j \frac{D_{x_i x_j}\phi}{|\nabla\phi|},\end{aligned}$$

where locally $\partial\Omega = \{\phi(x_1,x_2,x_3) = 0\}$. By the results of Agmon, Douglis, and Nirenberg [1],

$$\begin{aligned}&\|\nabla p\|_{\mathbf{H}^s} \le \|p\|_{H^{s+1}(\Omega)} \le \text{const}\left(\|\Delta p\|_{H^{s-1}(\Omega)} + \left\|\frac{\partial p}{\partial\mathbf{n}}\right\|_{H^{s-1/2}(\partial\Omega)}\right) \\ &\le \text{const}\left[\sum_{i,j=1}^3 \left(\|D_{x_i}u_jD_{x_j}u_i\|_{H^{s-1}} + \|u_ju_i\|_{H^s}\right) + |\mathbf{W}|\,\|\mathbf{u}\|_{\mathbf{H}^s} + \left|\frac{d\mathbf{W}}{dt}\right|\right] \\ &\le \text{(by (1.13) of [16])} \\ &\le \text{const}\left(\|\mathbf{u}\|^2_{\mathbf{H}^s} + |\mathbf{W}|^2 + \left|\frac{d\mathbf{W}}{dt}\right|\right).\end{aligned}$$

Therefore

$$\|\nabla p(t)\|_{\mathbf{H}^s} \le \text{const}\left(\|\mathbf{V}(t)\|_X^2 + \left|\frac{d\mathbf{W}}{dt}(t)\right|\right). \tag{3.6}$$

Denote

$$\mathbf{a}_j = P_0(\mathbf{e}_j\times\mathbf{r}), \quad j = 1,2,3.$$

It is shown in Lyashenko [11] that

$$\begin{cases} (\mathbf{a}_j, \mathbf{a}_i)_{\mathbf{L}^2} = 0, \quad i \neq j; \quad \|\mathbf{a}_1\|_{\mathbf{L}^2} = \|\mathbf{a}_2\|_{\mathbf{L}^2}, \\ \displaystyle\int_\Omega \mathbf{r} \times \mathbf{u}\, d\Omega = \big((\mathbf{u}, \mathbf{a}_1)_{\mathbf{L}^2}, (\mathbf{u}, \mathbf{a}_2)_{\mathbf{L}^2}, (\mathbf{u}, \mathbf{a}_3)_{\mathbf{L}^2}\big), \quad \mathbf{u} \in J_0(\Omega). \end{cases} \tag{3.7}$$

We have

$$\|\mathbf{a}_1\|^2_{\mathbf{L}^2} = \|\mathbf{a}_2\|^2_{\mathbf{L}^2} \le \|\mathbf{e}_1 \times \mathbf{r}\|^2_{\mathbf{L}^2} = \int_\Omega x_1^2 + x_3^2\, d\Omega < \int_\Omega x_1^2 + x_3^2\, d\Omega + \mu \int_G x_1^2 + x_3^2\, dG = a,$$

$$\|\mathbf{a}_3\|^2_{\mathbf{L}^2} \le \|\mathbf{e}_3 \times \mathbf{r}\|^2_{\mathbf{L}^2} = 2\int_\Omega x_1^2\, d\Omega < 2\int_\Omega x_1^2\, d\Omega + 2\mu \int_G x_1^2\, dG = b.$$

Therefore

$$\|\mathbf{a}_1\|^2_{\mathbf{L}^2} = \|\mathbf{a}_2\|^2_{\mathbf{L}^2} < a, \quad \|\mathbf{a}_3\|^2_{\mathbf{L}^2} < b. \tag{3.8}$$

Denote

$$\mathbf{U}(t) = (U_1(t), U_2(t), U_3(t)) = \big((\mathbf{u}(t), \mathbf{a}_1)_{\mathbf{L}^2}, (\mathbf{u}(t), \mathbf{a}_2)_{\mathbf{L}^2}, (\mathbf{u}(t), \mathbf{a}_3)_{\mathbf{L}^2}\big).$$

Because of (1.5) and (3.7), equation (1.2) can be written as follows:

$$\begin{cases} \dfrac{d}{dt}(aW_1 + U_1) + (b-a)W_3W_2 + W_2U_3 - W_3U_2 = 0, \\ \dfrac{d}{dt}(aW_2 + U_2) - (b-a)W_3W_1 + W_3U_1 - W_1U_3 = 0, \\ \dfrac{d}{dt}(bW_3 + U_3) + W_1U_2 - W_2U_1 = 0. \end{cases} \tag{3.9}$$

Taking the $\mathbf{L}^2$-inner product of (1.1) with $\mathbf{a}_j$, $j = 1, 2, 3$ and using

$$\left(\frac{d\mathbf{W}}{dt} \times \mathbf{r}, \mathbf{a}_j\right)_{\mathbf{L}^2} = \sum_{k=1}^3 \frac{dW_k}{dt}(\mathbf{e}_k \times \mathbf{r}, \mathbf{a}_j)_{\mathbf{L}^2} = \sum_{k=1}^3 \frac{dW_k}{dt}(\mathbf{a}_k, \mathbf{a}_j)_{\mathbf{L}^2} = \frac{dW_j}{dt}\|\mathbf{a}_j\|^2_{\mathbf{L}^2},$$

we obtain

$$\frac{d}{dt}\big(\|\mathbf{a}_j\|^2_{\mathbf{L}^2} W_j + U_j\big) = (\mathbf{u}, (\mathbf{u} \cdot \nabla)\mathbf{a}_j)_{\mathbf{L}^2} - 2(\mathbf{W} \times \mathbf{u}, \mathbf{a}_j)_{\mathbf{L}^2}, \quad j = 1, 2, 3. \tag{3.10}$$

Because of (3.8), equations (3.9) and (3.10) imply

$$\left|\frac{d\mathbf{W}}{dt}(t)\right| \le \text{const}\,(\|\mathbf{u}(t)\|^2_{\mathbf{L}^2} + |\mathbf{W}(t)|^2) = \text{const}\,\|\mathbf{V}(t)\|^2_Z. \tag{3.11}$$

Estimates (3.5), (3.6), and (3.11) yield

$$\frac{d}{dt}\|\mathbf{u}(t)\|_s \le \text{const}\,\|\mathbf{V}(t)\|^2_X. \tag{3.12}$$

Taking the $\mathbf{L}^2$-inner product of (1.1) with $\mathbf{u}$ we obtain

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{u}(t)\|^2_{\mathbf{L}^2} = -\left(\frac{d\mathbf{W}}{dt}(t) \times \mathbf{r}, \mathbf{u}(t)\right)_{\mathbf{L}^2},$$

which implies

$$\frac{d}{dt}\|\mathbf{u}(t)\|_{\mathbf{L}^2} \le \text{const}\,\|\mathbf{V}(t)\|_Z^2. \tag{3.13}$$

Thus estimates (3.11)–(3.13) yield

$$\frac{d}{dt}\|\mathbf{V}(t)\|_X \le \text{const}\,\|\mathbf{V}(t)\|_X^2, \tag{3.14}$$

where the constant does not depend on t.

Following Temam [16], we consider the following singular perturbation to the system (1.1)–(1.4):

$$\frac{\partial \mathbf{u}}{\partial t} + \epsilon[(-\Delta)^s\mathbf{u} + \mathbf{u}] + (\mathbf{u}\cdot\nabla)\mathbf{u} + 2\mathbf{W}\times\mathbf{u} + \frac{d\mathbf{W}}{dt}\times\mathbf{r} = -\nabla p, \tag{3.15}$$

$$\frac{d}{dt}\left(J\mathbf{W} + \int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right) + \mathbf{W}\times\left(J\mathbf{W} + \int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right) = 0, \tag{3.16}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{3.17}$$

$$\mathbf{u}\cdot\mathbf{n}|_{\partial\Omega} = 0, \tag{3.18}$$

$$\Delta^j\mathbf{u}|_{\partial\Omega} = 0, \quad \frac{s}{2} \le j \le s-1, \quad s \text{ even}; \quad \frac{s+1}{2} \le j \le s-1, \quad s \text{ odd}, \tag{3.19}$$

$$\frac{\partial\Delta^j\mathbf{u}}{\partial\mathbf{n}}|_{\partial\Omega} = 0, \quad \frac{s}{2} \le j \le s-2, \quad s \text{ even}; \quad \frac{s-1}{2} \le j \le s-2, \quad s \text{ odd}, \tag{3.20}$$

$$\frac{\partial\Delta^{s-1}\mathbf{u}}{\partial\mathbf{n}} = \left(\frac{\partial\Delta^{s-1}\mathbf{u}}{\partial\mathbf{n}}\cdot\mathbf{n}\right)\mathbf{n}. \tag{3.21}$$

It is easy to see (cf. [9, 17]) that for any $\epsilon > 0$ fixed there exists $T > 0$ and unique $\mathbf{u}_\epsilon(t) \in L^2((0,T);X_s)\cap L^\infty((0,T);J_0(\Omega))$, $\mathbf{W}_\epsilon(t) \in (L^\infty(0,T))^3$ satisfying (3.15)–(3.21) and $\mathbf{u}_\epsilon(0) = \mathbf{u}_0$, $\mathbf{W}(0) = \mathbf{W}_0$. Applying the same arguments as in deriving (3.14) and using the additional boundary conditions (3.19)–(3.21), we obtain the following a priori estimates:

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{u}_\epsilon(t)\|_{\mathbf{H}^s}^2 + \epsilon\|(-\Delta)^s\mathbf{u}_\epsilon(t) + \mathbf{u}_\epsilon(t)\|_{\mathbf{L}^2}^2 \le \text{const}\,\|\mathbf{u}_\epsilon(t)\|_{\mathbf{H}^s}\|\mathbf{V}_\epsilon(t)\|_X^2, \tag{3.22}$$

$$\left|\frac{d\mathbf{W}}{dt}(t)\right| \le \text{const}\,(\|\mathbf{V}_\epsilon(t)\|_Z^2 + \epsilon\|\mathbf{u}_\epsilon(t)\|_{\mathbf{H}^s}) \tag{3.23}$$

which imply

$$\frac{d}{dt}\|\mathbf{V}_\epsilon(t)\|_X \le \text{const}\,\|\mathbf{V}_\epsilon(t)\|_X^2. \tag{3.24}$$

Standard passage to the limit $\epsilon \to 0$ (cf. [16]) proves existence of a solution to the system (1.1)–(1.4). Uniqueness follows from (3.4). The inclusion $\mathbf{V}(t) \in C([0,T];Z)$ follows from (3.11) and (3.13). □

4. The unstable spectrum of the linearized operator. In the present section we study the spectrum of the linear part $\omega_0 K^{-1}L$ of equation (1.15). We show that the unstable spectrum is purely discrete and we obtain a necessary and sufficient condition for the existence of an unstable eigenvalue. Thus, for configurations where this condition is satisfied, the general theorem described in section 2 implies nonlinear instability of the uniform rotation of the fluid+body system. Since ω_0 is a positive constant, it is sufficient to discuss the spectral properties of the operator $K^{-1}L$.

PROPOSITION 4.1. *If $\lambda \in \mathbb{C}$ is a spectral point of $K^{-1}L$ satisfying $\operatorname{Re}\lambda \neq 0$ then λ is an eigenvalue of finite multiplicity.*

Proof. Let $\operatorname{Re}\lambda \neq 0$. Using (1.14) and the self-adjointness of K we obtain

$$\begin{aligned}(K^{-1}L - \lambda I) &= K^{-1/2}\left(iK^{-1/2}L_1K^{-1/2} + K^{-1/2}L_2K^{-1/2} - \lambda I\right)K^{1/2}\\ &= K^{-1/2}\left(iK^{-1/2}L_1K^{-1/2} - \lambda I\right)\\ &\qquad\cdot\left(I + (iK^{-1/2}L_1K^{-1/2} - \lambda I)^{-1}K^{-1/2}L_2K^{-1/2}\right)K^{1/2},\end{aligned}$$

where

$$L_1\mathbf{V} = \begin{pmatrix}2T\mathbf{u}\\ 0\end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix}\mathbf{u}\\ \mathbf{w}\end{pmatrix},$$

and

$$L_2 = L - iL_1$$

is a finite-dimensional operator. Since $K^{-1/2}L_1K^{-1/2}$ is a self-adjoint operator then λ is a spectral point of $K^{-1}L$ if and only if (-1) is an eigenvalue of the finite-dimensional operator

$$(iK^{-1/2}L_1K^{-1/2} - \lambda I)^{-1}K^{-1/2}L_2K^{-1/2}.$$

Therefore any spectral point λ of $K^{-1}L$ satisfying $\operatorname{Re}\lambda \neq 0$ is an eigenvalue of finite multiplicity. □

Remark. We note that the above arguments do not make use of the condition of 4-fold symmetry. Therefore Proposition 4.1 is valid independent of any geometrical symmetry constraints.

Consider the spectral problem

$$K^{-1}L\mathbf{V} = \lambda\mathbf{V}, \quad \mathbf{V} \in Z. \tag{4.1}$$

Because of (1.5), (1.10), and (1.11) it can be written as follows:

$$2iT\mathbf{u} = \lambda(\mathbf{u} + P_0(\mathbf{w}\times\mathbf{r})), \tag{4.2}$$

$$\mathbf{e}_3 \times \left((a-b)\mathbf{w} + \int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right) = \lambda\left(J\mathbf{w} + \int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right). \tag{4.3}$$

It is easy to see that $\lambda = 0$ is always an eigenvalue of (4.2), (4.3). Henceforth we assume that $\lambda \neq 0$. Then (4.3) can be solved for $\mathbf{w}$ (cf. [11]):

$$\mathbf{w} = C(\lambda)\int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega, \tag{4.4}$$

where

$$C(\lambda) = \begin{pmatrix} \dfrac{(b-a)-\lambda^2 a}{(b-a)^2+\lambda^2 a^2} & -\lambda b & 0 \\ \lambda b & \dfrac{(b-a)-\lambda^2 a}{(b-a)^2+\lambda^2 a^2} & 0 \\ 0 & 0 & -\dfrac{1}{b} \end{pmatrix}.$$

Using the notation of section 3

$$\mathbf{a}_j = P_0(\mathbf{e}_j \times \mathbf{r}), \quad j = 1, 2, 3,$$

and formulas (3.7) one can verify that (cf. [11])

$$P_0\left(\left(C(\lambda)\int_\Omega \mathbf{r}\times\mathbf{u}\,d\Omega\right)\times\mathbf{r}\right) = -m\left[\frac{\lambda+i}{\lambda+ik}P_{\mathbf{c}_1}\mathbf{u} + \frac{\lambda-i}{\lambda-ik}P_{\mathbf{c}_2}\mathbf{u} + dP_{\mathbf{c}_3}\mathbf{u}\right], \tag{4.5}$$

where

$$\begin{cases} m = \dfrac{\|\mathbf{a}_1\|^2_{\mathbf{L}^2}}{a}, \quad k = \dfrac{a-b}{a}, \quad d = \dfrac{a}{b}\dfrac{\|\mathbf{a}_3\|^2_{\mathbf{L}^2}}{\|\mathbf{a}_1\|^2_{\mathbf{L}^2}}, \\ P_{\mathbf{c}_j}\mathbf{u} = \dfrac{(\mathbf{u},\mathbf{c}_j)_{\mathbf{L}^2}}{\|\mathbf{c}_j\|^2_{\mathbf{L}^2}}\mathbf{c}_j, \quad j = 1,2,3, \\ \mathbf{c}_1 = \mathbf{a}_1 + i\mathbf{a}_2, \quad \mathbf{c}_2 = \mathbf{a}_1 - i\mathbf{a}_2, \quad \mathbf{c}_3 = \mathbf{a}_3. \end{cases} \tag{4.6}$$

Thus (4.4), (4.5) imply that for $\lambda \neq 0$ system (4.2), (4.3) is equivalent to the following spectral problem:

$$2iT\mathbf{u} + m\lambda D(\lambda, k, d)\,\mathbf{u} = \lambda\mathbf{u}, \tag{4.7}$$

where

$$D(\lambda,k,d) = \frac{\lambda+i}{\lambda+ik}P_{\mathbf{c}_1} + \frac{\lambda-i}{\lambda-ik}P_{\mathbf{c}_2} + dP_{\mathbf{c}_3}$$

is a three-dimensional operator in $J_0(\Omega)$. It follows from (3.7), (4.6) that

$$(\mathbf{c}_j, \mathbf{c}_n)_{\mathbf{L}^2} = 0\,, \quad j \neq n; \quad \|\mathbf{c}_1\|_{\mathbf{L}^2} = \|\mathbf{c}_2\|_{\mathbf{L}^2}. \tag{4.8}$$

Therefore $P_{\mathbf{c}_j}$, $j = 1, 2, 3$ are pair-wise orthogonal one-dimensional projections. It is easy to see that (1.5), (4.6), and (3.8) imply

$$0 < m < 1, \quad 0 < md < 1, \quad -1 < k < 1. \tag{4.9}$$

Since T is a self-adjoint operator then for any λ with $\operatorname{Re}\lambda \neq 0$ spectral problem (4.1) is equivalent to

$$\mathbf{u} + m\lambda\,(2iT - \lambda I)^{-1}\,D\,(\lambda,k,d)\,\mathbf{u} = 0. \tag{4.10}$$

PROPOSITION 4.2. *If there exists an eigenvalue λ of $K^{-1}L$ satisfying $\operatorname{Re}\lambda \neq 0$ then $k = (a-b)/a > 0$.*

Proof. Let $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \neq 0$ be an eigenvalue of $K^{-1}L$ and $\mathbf{u}$ be a corresponding eigenfunction. Taking the real part of the $\mathbf{L}^2$-inner product of (4.7) with $\mathbf{u}$ we obtain

$$\begin{aligned}\operatorname{Re}\lambda\, k(1-k)\, m \left[\frac{\|P_{\mathbf{c}_1}\mathbf{u}\|^2_{\mathbf{L}^2}}{|\lambda+ik|^2} + \frac{\|P_{\mathbf{c}_2}\mathbf{u}\|^2_{\mathbf{L}^2}}{|\lambda-ik|^2}\right] \\ = \operatorname{Re}\lambda \left(\|\mathbf{u}\|^2_{\mathbf{L}^2} - m\|P_{\mathbf{c}_1}\mathbf{u}\|^2_{\mathbf{L}^2} - m\|P_{\mathbf{c}_2}\mathbf{u}\|^2_{\mathbf{L}^2} - md\|P_{\mathbf{c}_3}\mathbf{u}\|^2_{\mathbf{L}^2}\right).\end{aligned} \tag{4.11}$$

Because of (4.8),(4.9) and $\operatorname{Re}\lambda \neq 0$, equality (4.11) implies $k > 0$. □

Corollary. *Condition*

$$a - b = \int_\Omega x_3^2 - x_1^2\, d\Omega + \mu \int_G x_3^2 - x_1^2\, dG > 0 \tag{4.12}$$

is a necessary condition for $K^{-1}L$ to have an eigenvalue λ satisfying $\operatorname{Re}\lambda < 0$.

Remark. The physical meaning of (4.12) is that the axis of rotation is the axis of the least moment of inertia of the entire system body+liquid. This necessary condition for the linear instability is known in the literature (cf. [14]).

Denote

$$P_3 = P_{\mathbf{c}_1} + P_{\mathbf{c}_2} + P_{\mathbf{c}_3}$$

an orthogonal projection onto $span\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ by virtue of (4.8). Since $D(\lambda,k,d)P_3 = D(\lambda,k,d)$ then (4.10) is equivalent to the following system:

$$P_3\mathbf{u} + m\lambda P_3 \left(2iT - \lambda I\right)^{-1} D\left(\lambda,k,d\right) P_3\mathbf{u} = 0, \tag{4.13}$$

$$(I - P_3)\mathbf{u} + m\lambda(I - P_3) \left(2iT - \lambda\right)^{-1} D\left(\lambda,k,d\right) P_3\mathbf{u} = 0. \tag{4.14}$$

Thus $(I-P_3)\mathbf{u}$ is uniquely determined by $P_3\mathbf{u}$. Hence $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \neq 0$ is an eigenvalue of $K^{-1}L$ if and only if (4.13) admits a nontrivial solution. Since

$$P_3 + m\lambda P_3(2iT - \lambda I)^{-1} D(\lambda,k,d) P_3$$

is a three-dimensional linear operator in $P_3 J_0(\Omega) = span\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ then (4.13) admits a nontrivial solution if and only if

$$\det\left[(\mathbf{c}_j, \mathbf{c}_n)_{\mathbf{L}^2} + m\lambda \left(\left(2iT - \lambda I\right)^{-1} D\left(\lambda,k,d\right)\mathbf{c}_j, \mathbf{c}_n\right)_{\mathbf{L}^2}\right] = 0. \tag{4.15}$$

It was shown in [11] that

$$((2iT - \lambda I)^{-1}\mathbf{c}_j, \mathbf{c}_n)_{\mathbf{L}^2} = 0, \quad j \neq n,\ \operatorname{Re}\lambda \neq 0. \tag{4.16}$$

Because of (4.16), equation (4.15) can be written as follows:

$$f_1(\lambda,m,k) \cdot f_2(\lambda,m,k) \cdot f_3(\lambda,md,k) = 0, \tag{4.17}$$

where

$$\begin{aligned} f_1(\lambda,m,k) &= \|\mathbf{c}_1\|^2_{\mathbf{L}^2} + m\lambda\frac{\lambda+i}{\lambda+ik}((2iT - \lambda I)^{-1}\mathbf{c}_1, \mathbf{c}_1)_{\mathbf{L}^2}, \\ f_2(\lambda,m,k) &= \|\mathbf{c}_2\|^2_{\mathbf{L}^2} + m\lambda\frac{\lambda-i}{\lambda-ik}((2iT - \lambda I)^{-1}\mathbf{c}_2, \mathbf{c}_2)_{\mathbf{L}^2}, \\ f_3(\lambda,m,k) &= \|\mathbf{c}_3\|^2_{\mathbf{L}^2} + m\lambda((2iT - \lambda I)^{-1}\mathbf{c}_3, \mathbf{c}_3)_{\mathbf{L}^2} \end{aligned}$$

are functions that are analytic in $\lambda \in \mathbb{C} \setminus \{\alpha i \mid \alpha \in [-2,2]\}$. Thus we have verified the following criterion.

THEOREM 4.1. *$\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \neq 0$ is an eigenvalue of $K^{-1}L$ if and only if (4.17) holds.*

COROLLARY. *For any $\epsilon > 0$ there exist at most a finite number of spectral points λ of $K^{-1}L$ satisfying $|\operatorname{Re}\lambda| \geq \epsilon$.*

Proof. It is easy to see that

$$\lim_{\lambda\to\infty} f_j(\lambda, m, k) = \|\mathbf{c}_j\|^2_{\mathbf{L}^2}(1-m) \neq 0, \quad j = 1,2,3.$$

Therefore there exist $R_j(m,k) > 0$, $j = 1,2,3$ such that

$$f_j(\lambda, m, k) \neq 0, \quad |\lambda| > R_j(m,k)\,,\ j = 1,2,3.$$

Then the statement of the corollary follows from Proposition 4.1, Theorem 4.1, and the analyticity of $f_j(\lambda, m, k)$, $j = 1,2,3$. □

Now we are going to simplify the criterion presented in Theorem 4.1 by showing that $f_3(\lambda, md, k) \neq 0$, $\operatorname{Re}\lambda \neq 0$, and $\overline{f_1(\lambda, m, k)} = f_2(\bar\lambda, m, k)$.

LEMMA 4.1. *For any $\mathbf{a} \in J_0(\Omega)$, $\mathbf{a} \neq 0$, $m \in (0,1)$, $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \neq 0$ we have*

$$\|\mathbf{a}\|^2_{\mathbf{L}^2} + m\lambda\left((2iT - \lambda I)^{-1}\mathbf{a}\,,\,\mathbf{a}\right)_{\mathbf{L}^2} \neq 0.$$

Proof. Let E_γ be the resolution of identity corresponding to the self-adjoint operator T. Since $\sigma(T) = [-1,1]$, then for any $\epsilon > 0$

$$\begin{aligned}\|\mathbf{a}\|^2_{\mathbf{L}^2} + m\lambda\left((2iT - \lambda I)^{-1}\mathbf{a}\,,\,\mathbf{a}\right)_{\mathbf{L}^2} &= \int_{-1}^{1+\epsilon}\left(1 + \frac{m\lambda}{2i\gamma - \lambda}\right) d(E_\gamma\mathbf{a},\mathbf{a})_{\mathbf{L}^2} \\ &= \int_{-1}^{1+\epsilon}\frac{|2i\gamma-\lambda|^2 - m|\lambda|^2}{|2i\gamma-\lambda|^2}\, d(E_\gamma\mathbf{a},\mathbf{a})_{\mathbf{L}^2} - 2m\lambda i\int_{-1}^{1+\epsilon}\frac{\gamma}{|2i\gamma-\lambda|^2}\, d(E_\gamma\mathbf{a},\mathbf{a})_{\mathbf{L}^2}.\end{aligned}$$

Assume there exists $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \neq 0$ such that $\|\mathbf{a}\|^2_{\mathbf{L}^2} + m\lambda\left((2iT - \lambda I)^{-1}\mathbf{a},\mathbf{a}\right)_{\mathbf{L}^2} = 0$. Then from the formula above it follows that

$$\int_{-1}^{1+\epsilon}\frac{\gamma}{|2i\gamma-\lambda|^2}\, d(E_\gamma\mathbf{a},\mathbf{a})_{\mathbf{L}^2} = 0,$$

$$\begin{aligned}0 = \int_{-1}^{1+\epsilon}\frac{|2i\gamma-\lambda|^2 - m|\lambda|^2}{|2i\gamma-\lambda|^2}\, d(E_\gamma\mathbf{a},\mathbf{a})_{\mathbf{L}^2} &= \int_{-1}^{1+\epsilon}\frac{4\gamma^2 + (1-m)|\lambda|^2}{|2i\gamma-\lambda|^2}\, d(E_\gamma\mathbf{a},\mathbf{a})_{\mathbf{L}^2} \\ &> \frac{(1-m)|\lambda|^2}{(\operatorname{Re}\lambda)^2}\int_{-1}^{1+\epsilon} d(E_\gamma\mathbf{a},\mathbf{a})_{\mathbf{L}^2} = \frac{(1-m)|\lambda|^2}{(\operatorname{Re}\lambda)^2}\|\mathbf{a}\|^2_{\mathbf{L}^2} > 0.\end{aligned}$$

The obtained contradiction proves the lemma. □

COROLLARY. *From Lemma 4.1 and $\mathbf{c}_3 = P_0(-x_2, x_1, 0) \neq 0$ it follows that*

$$f_3(\lambda, m, k) \neq 0, \quad \operatorname{Re}\lambda \neq 0\,,\ m \in (0,1).$$

Let $\mathbf{v} \in J_0(\Omega) \cap \mathbf{H}^1$, $\lambda \in \mathbb{C}$, $\mathrm{Re}\lambda \neq 0$, and

$$\mathbf{u} = (2iT - \lambda I)^{-1}\mathbf{v}. \tag{4.18}$$

Then

$$\mathbf{v} = (2iT - \lambda I)\mathbf{u} = 2P_0(\mathbf{e}_3 \times \mathbf{u}) - \lambda\mathbf{u} = 2(-u_2, u_1, 0) - \lambda(u_1, u_2, u_3) - \nabla p \tag{4.19}$$

for some $p(x) \in H^1(\Omega)$. Therefore

$$\begin{cases} u_1 = \dfrac{2v_2 - \lambda v_1 + 2p_{x_2} - \lambda p_{x_1}}{\lambda^2 + 4}, \\ u_2 = -\dfrac{2v_1 + \lambda v_2 + 2p_{x_1} + \lambda p_{x_2}}{\lambda^2 + 4}, \\ u_3 = -\dfrac{v_3 + p_{x_3}}{\lambda}. \end{cases} \tag{4.20}$$

Since $\mathbf{u}, \mathbf{v} \in J_0(\Omega)$ then

$$0 = \operatorname{div}\mathbf{u} = \frac{1}{\lambda(\lambda^2+4)}\Big(-\lambda^2\Delta p - 4p_{x_3x_3} + 2\lambda(v_{2x_1} - v_{1x_2}) - 4v_{3x_3}\Big),$$

$$0 = \mathbf{u}\cdot\mathbf{n} = \frac{-\lambda^2\nabla p\cdot\mathbf{n} + 2\lambda(p_{x_2}n_1 - p_{x_1}n_2) - 4p_{x_3}n_3 + 2\lambda(v_2n_1 - v_1n_2) - 4v_3n_3}{\lambda(\lambda^2+4)}.$$

Hence $p(x)$ satisfies

$$\left(-\frac{\lambda}{2}\right)^2 \Delta p + p_{x_3x_3} = -\left(-\frac{\lambda}{2}\right)(v_{2x_1} - v_{1x_2}) - v_{3x_3}, \tag{4.21}$$

$$\begin{aligned} \left(-\frac{\lambda}{2}\right)^2 \nabla p\cdot\mathbf{n} + \left(-\frac{\lambda}{2}\right)(p_{x_2}n_1 - p_{x_1}n_2) + p_{x_3}n_3\big|_{\partial\Omega} \\ = -\left(-\frac{\lambda}{2}\right)(v_2n_1 - v_1n_2) - v_3n_3\big|_{\partial\Omega}. \end{aligned} \tag{4.22}$$

Remark. If $\mathbf{v} = 0$ then (4.21), (4.22) is the spectral problem corresponding to the second mixed problem for the Poincaré–Sobolev equation (cf. [18]).

Applying the standard arguments (cf. [13, 2]), one obtains the following result.

LEMMA 4.2 (see [10]). *Let $\lambda \in \mathbb{C}$, $\mathrm{Re}\,\lambda \neq 0$, $s \in \mathbb{N}\cup\{0\}$, $\mathbf{v} \in X_{s+1}$, $\partial\Omega \in C^{4+s}$. Then there exists unique $p \in H^{s+2}(\Omega)$ satisfying* (4.21), (4.22), *and $\int_\Omega p\,d\Omega = 0$.*

Denote by $p(x, \lambda, \mathbf{v})$ the unique solution of (4.21), (4.22) satisfying $\int_\Omega p\,d\Omega = 0$. From (4.21), (4.22) it follows that

$$\overline{p(x, \lambda, \mathbf{v})} = p(x, \bar\lambda, \bar{\mathbf{v}}). \tag{4.23}$$

Let $\mathrm{Re}\lambda \neq 0$, $\mathbf{v} = (v_1, v_2, v_3) \in J_0(\Omega) \cap \mathbf{H}^1$. Then using (4.18), (4.19) we obtain

(4.24)

$$\begin{aligned} &g(\mathbf{v}, \lambda) \stackrel{\text{def}}{=} \big((2iT - \lambda I)^{-1}\mathbf{v}, \mathbf{v}\big)_{\mathbf{L}^2} \\ &= \frac{1}{\lambda(\lambda^2+4)}\int_\Omega \big(2\lambda(v_2\bar v_1 - v_1\bar v_2) - \lambda^2|\mathbf{v}|^2 - 4|v_3|^2 + 2\lambda(p_{x_2}\bar v_1 - p_{x_1}\bar v_2) - 4p_{x_3}\bar v_3\big)\,d\Omega, \end{aligned}$$

where $p = p(x, \lambda, \mathbf{v})$. Since $\overline{\mathbf{c}_1} = \mathbf{c}_2$ then (4.23), (4.24) imply

$$\overline{\left((2iT - \lambda I)^{-1}\mathbf{c}_1\,,\,\mathbf{c}_1\right)_{\mathbf{L}^2}} = \left((2iT - \bar{\lambda} I)^{-1}\mathbf{c}_2\,,\,\mathbf{c}_2\right)_{\mathbf{L}^2}.$$

Hence

$$\overline{f_1(\lambda, m, k)} = f_2(\bar{\lambda}, m, k).$$

Thus, Theorem 4.1 can be rewritten as follows.

THEOREM 4.2. *$\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \neq 0$ is an eigenvalue of $K^{-1}L$ if and only if*

$$\left[\|\mathbf{c}_1\|^2_{\mathbf{L}^2} + m\lambda\frac{\lambda + i}{\lambda + ik}g(\mathbf{c}_1, \lambda)\right] \cdot \left[\|\mathbf{c}_1\|^2_{\mathbf{L}^2} + m\bar{\lambda}\frac{\bar{\lambda} + i}{\bar{\lambda} + ik}g(\mathbf{c}_1, \bar{\lambda})\right] = 0, \tag{4.25}$$

where $g(\mathbf{v}, \lambda)$ is defined by (4.24).

To conclude this section we derive a representation of $p(x, \lambda, \mathbf{v})$ in the form of a power series in $1/\lambda$. It follows from (4.18), (4.19) that

$$\nabla p(x, \lambda, \mathbf{v}) = (I - P_0)2(-u_2, u_1, 0) = 2(I - P_0)(\mathbf{e}_3 \times (2iT - \lambda I)^{-1}\mathbf{v}).$$

Since $\|T\| = 1$ then for any $\lambda \in \mathbb{C}$, $|\lambda| > 2$ we have

$$(2iT - \lambda I)^{-1}\mathbf{v} = -\frac{1}{\lambda}\sum_{j=0}^{\infty}\left(\frac{2}{\lambda}\right)^j (iT)^j\mathbf{v}.$$

Therefore

$$\nabla p(x, \lambda, \mathbf{v}) = \sum_{j=1}^{\infty}\frac{1}{\lambda^j}\left[-2^j(I - P_0)(\mathbf{e}_3 \times (iT)^{j-1}\mathbf{v})\right].$$

Due to the Weyl decomposition of $\mathbf{L}^2$, for any $j \in \mathbb{N}$ there exists a unique $p_j(x, \mathbf{v}) \in H^1(\Omega)$ such that $\int_\Omega p_j\, d\Omega = 0$ and

$$\nabla p_j(x, \mathbf{v}) = -2^j(I - P_0)(\mathbf{e}_3 \times (iT)^{j-1}\mathbf{v}).$$

Hence we have

$$p(x, \lambda, \mathbf{v}) = \sum_{j=1}^{\infty}\frac{1}{\lambda^j}p_j(x, \mathbf{v})\,, \quad \lambda \in \mathbb{C} \setminus \{i\alpha \mid \alpha \in [-2, 2]\}. \tag{4.26}$$

Substituting (4.26) into (4.21), (4.22) we obtain that $p_j(x, \mathbf{v})$, $j \in \mathbb{N}$ satisfy the following system:

(4.27)
$$\begin{cases} \Delta p_1 = 2(v_{2x_1} - v_{1x_2}), & \nabla p_1 \cdot \mathbf{n}\big|_{\partial\Omega} = 2(v_2 n_1 - v_1 n_2)\big|_{\partial\Omega}, \\ \Delta p_2 = -4v_{3x_3}, & \nabla p_2 \cdot \mathbf{n}\big|_{\partial\Omega} = 2(p_{1x_2}n_1 - p_{1x_1}n_2) - 4v_3 n_3\big|_{\partial\Omega}, \\ \Delta p_{j+2} = -4p_{jx_3x_3}, & \nabla p_{j+2} \cdot \mathbf{n}\big|_{\partial\Omega} = 2(p_{j+1x_2}n_1 - p_{j+1x_1}n_2) - 4p_{jx_3}n_3\big|_{\partial\Omega}. \end{cases}$$

It is easy to see that the system of Neumann problems (4.27) is compatible and the functions $p_j(x, \mathbf{v})$, $j \in \mathbb{N}$ are uniquely determined by (4.27) and the condition $\int_\Omega p_j\, d\Omega = 0$.

The following theorem summarizes the obtained results regarding the spectrum of the operator $K^{-1}L$.

THEOREM 4.3.

(1) *All the spectral points* λ *of* $K^{-1}L$ *satisfying* $\operatorname{Re}\lambda \neq 0$ *are eigenvalues of finite multiplicity;*
(2) *For any* $\epsilon > 0$ *there exist at most a finite number of spectral points* λ *of* $K^{-1}L$ *satisfying* $|\operatorname{Re}\lambda| \geq \epsilon$*;*
(3) $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \neq 0$ *is an eigenvalue of* $K^{-1}L$ *if and only if* (4.25) *holds, where* $g(\mathbf{v}, \lambda)$ *is defined by* (4.24) *and* $p = p(x, \lambda, \mathbf{v})$ *is given by* (4.26), (4.27).

Thus the spectral gap condition of Theorem 2.1 holds whenever there exists a $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda < 0$ satisfying (4.25) and the linear instability of uniform rotation of the entire body+liquid system implies nonlinear instability.

5. Ellipsoidal cavity. To demonstrate that Theorem 4.3 gives an effective condition for nonlinear instability, we consider the example of a body with an ellipsoidal cavity Ω. We calculate the geometric quantities occuring in equation (4.25). We obtain an explicit quadratic equation for the spectral parameter λ and demonstrate that for cavities of appropriate ellipticity there exists an eigenvalue with $\operatorname{Re}\lambda < 0$. In fact either the unstable spectrum is empty or there exist exactly two pairs of eigenvalues $\pm\operatorname{Re}\lambda + i\operatorname{Im}\lambda$ and $\pm\operatorname{Re}\lambda - i\operatorname{Im}\lambda$ (the volume preserving nature of the fluid motion dictates that a growing mode is matched by a decaying mode). The growth rate $|\operatorname{Re}\lambda|$ of the instability is given by an explicit function of the geometry and the relative fluid/body density.

For any $r, R \in (0, \infty)$ we consider the following ellipsoid of revolution:

$$\Omega = \Omega_{r,R} = \{x \in \mathbb{R}^3 \mid r^2(x_1^2 + x_2^2) + x_3^2 < R^2\}.$$

Then

$$\mathbf{n}(x) = \frac{(r^2 x_1, r^2 x_2, x_3)}{\sqrt{r^2(R^2 - x_3^2) + x_3^2}}$$

is the outer normal to $\Omega_{r,R}$. Denote

$$\tilde{\mathbf{n}}(x) = (r^2 x_1, r^2 x_2, x_3).$$

Consider $\mathbf{a}_k = P_0(\mathbf{e}_k \times \mathbf{r})$, $k = 1, 2, 3$, $\mathbf{r} = (x_1, x_2, x_3)$. We have

$$\begin{aligned}
&\mathbf{a}_1 = (0, -x_3, x_2) - \nabla p^1, \quad \nabla p^1 \cdot \tilde{\mathbf{n}}\big|_{\partial\Omega} = -x_3\tilde{n}_2 + x_2\tilde{n}_3\big|_{\partial\Omega} = (1 - r^2)x_2x_3,\\
&\mathbf{a}_2 = (x_3, 0, -x_1) - \nabla p^2, \quad \nabla p^2 \cdot \tilde{\mathbf{n}}\big|_{\partial\Omega} = x_3\tilde{n}_1 - x_1\tilde{n}_3\big|_{\partial\Omega} = (r^2 - 1)x_1x_3,\\
&\mathbf{a}_3 = (-x_2, x_1, 0) - \nabla p^3, \quad \nabla p^3 \cdot \tilde{\mathbf{n}}\big|_{\partial\Omega} = -x_2\tilde{n}_1 + x_1\tilde{n}_2\big|_{\partial\Omega} = 0,
\end{aligned}$$

$$\Delta p^j(x) = 0, \qquad j = 1, 2, 3.$$

Assuming $\int_\Omega p^j \, d\Omega = 0$ we obtain

$$p^1(x) = \frac{1 - r^2}{1 + r^2}x_2x_3, \quad p^2(x) = \frac{r^2 - 1}{1 + r^2}x_1x_3, \quad p^3(x) = 0.$$

Therefore

$$\mathbf{a}_1 = \frac{2}{1 + r^2}(0, -x_3, r^2x_2), \quad \mathbf{a}_2 = \frac{2}{1 + r^2}(x_3, 0, -r^2x_1), \quad \mathbf{a}_3 = (-x_2, x_1, 0).$$

According to (4.6),

$$\mathbf{c}_1 = \mathbf{a}_1 + i\mathbf{a}_2 = \frac{2}{1+r^2}(ix_3, -x_3, r^2(x_2 - ix_1)),$$
$$\mathbf{c}_2 = \mathbf{a}_1 - i\mathbf{a}_2 = \frac{2}{1+r^2}(-ix_3, -x_3, r^2(x_2 + ix_1)),$$
$$\mathbf{c}_3 = \mathbf{a}_3 = (-x_2, x_1, 0).$$

By (4.26), (4.27),

$$p(x, \lambda, \mathbf{c}_1) = \sum_{j=1}^{\infty} \frac{1}{\lambda^j} p_j(x, \mathbf{c}_1),$$

where $p_j(x, \mathbf{c}_1)$, $j \in \mathbb{N}$ satisfy

$$\Delta p_1 = 0, \qquad \nabla p_1 \cdot \tilde{\mathbf{n}}\big|_{\partial\Omega} = -4i\frac{r^2}{1+r^2} x_3(x_2 - ix_1),$$
$$\Delta p_2 = 0, \qquad \nabla p_2 \cdot \tilde{\mathbf{n}}\big|_{\partial\Omega} = 2(p_{1_{x_2}}\tilde{n}_1 - p_{1_{x_1}}\tilde{n}_2) - 8\frac{r^2}{1+r^2} x_3(x_2 - ix_1),$$
$$\Delta p_{j+2} = -4p_{j_{x_3x_3}}, \quad \nabla p_{j+2} \cdot \tilde{\mathbf{n}}\big|_{\partial\Omega} = 2(p_{j+1_{x_2}}\tilde{n}_1 - p_{j+1_{x_1}}\tilde{n}_2) - 4p_{j_{x_3}}\tilde{n}_3.$$

Since $\nabla(x_3(x_2 - ix_1)) \cdot \tilde{\mathbf{n}}(x) = (r^2+1)x_3(x_2 - ix_1)$ then one can easily verify that

$$p_1(x, \mathbf{c}_1) = -4i\frac{r^2}{(1+r^2)^2} x_3(x_2 - ix_1),$$

$$p_{j+1}(x, \mathbf{c}_1) = -\frac{2i}{1+r^2} p_j(x, \mathbf{c}_1), \quad j \in \mathbb{N}.$$

Therefore

$$p(x, \lambda, \mathbf{c}_1) = \sum_{j=1}^{\infty} \frac{1}{\lambda^j} p_j(x, \mathbf{c}_1) = -4i\frac{r^2}{1+r^2} \cdot \frac{x_3(x_2 - ix_1)}{\lambda(1+r^2) + 2i}.$$

According to (4.24) we have

$$g(\mathbf{c}_1, \lambda) = \frac{1}{\lambda(\lambda^2+4)(1+r^2)^2} \left[16i\lambda \int_\Omega x_3^2 \, d\Omega - 8\lambda^2 \int_\Omega (x_3^2 + r^4 x_1^2) \, d\Omega \right.$$
$$\left. -32r^4 \int_\Omega x_1^2 \, d\Omega + \frac{32ir^2}{\lambda(1+r^2) + 2i} \left(i\lambda \int_\Omega x_3^2 \, d\Omega + 2r^2 \int_\Omega x_1^2 \, d\Omega \right) \right].$$

Since

$$\int_{\Omega_{r,R}} x_3^2 \, d\Omega = \frac{4\pi}{15} R^5 \frac{1}{r^2}, \quad \int_{\Omega_{r,R}} x_1^2 \, d\Omega = \frac{4\pi}{15} R^5 \frac{1}{r^4},$$

we obtain

$$g(\mathbf{c}_1, \lambda) = -\frac{32\pi}{15} R^5 \frac{1}{r^2(\lambda(1+r^2) + 2i)}.$$

Therefore

$$\begin{aligned} f_1(\lambda,m,k) &= \|\mathbf{c}_1\|^2_{\mathbf{L}^2} + m\lambda\frac{\lambda+i}{\lambda+ik}g(\mathbf{c}_1,\lambda) \\ &= 2\left(\frac{2}{1+r^2}\right)^2\int_\Omega (x_3^2+r^4x_1^2)\,d\Omega - \frac{32\pi}{15}R^5 m\lambda\frac{\lambda+i}{\lambda+ik}\cdot\frac{1}{r^2(\lambda(1+r^2)+2i)} \\ &= \frac{32\pi}{15}R^5\frac{(1+r^2)(1-m)\lambda^2+((1+r^2)(k-m)+2)i\lambda-2k}{r^2(r^2+1)(\lambda+ik)(\lambda(1+r^2)+2i)}. \end{aligned}$$

Thus, from Theorem 4.2 it follows that $K^{-1}L$ has an unstable eigenvalue λ if and only if the roots of the numerator have the form

$$\lambda = \frac{\pm\sqrt{\Delta} - i((1+r^2)(k-m)+2)}{2(1+r^2)(1-m)}$$

with

$$\Delta(r,m,k) = 8(1+r^2)(1-m)k - ((1+r^2)(k-m)+2)^2 > 0. \tag{5.1}$$

The quantities m and k occurring in the instability condition (5.1) are defined in (4.6) and (1.5) and involve only the shape of the body/cavity and the relative density μ of the body/fluid. For configurations such that $\Delta > 0$ the linearized operator has an eigenfunction whose growth rate is given by $\sqrt{\Delta}$ and such a configuration is nonlinearly unstable.

Remark. Condition (5.1) reproduces the linear instability condition obtained by Sobolev [14] using a different approach.

Consider the limit case of the weightless body. Then the ratio μ of the body density to the fluid density is zero and we have

$$m = \frac{4r^2}{(1+r^2)^2}, \quad k = \frac{r^2-1}{r^2+1}.$$

Therefore, in this case

$$\Delta = \frac{(r^2-1)^3}{(1+r^2)^2}(9-r^2).$$

Thus we have verified the following result: in the case of weightless body with ellipsoidal cavity $\Omega_{r,R}$ the uniform rotation is linearly unstable if and only if $r \in (1,3)$. This is a classical result known to Kelvin (see, for example, [8]).

Since m, k depend continuously on $\mu \in [0,\infty)$ and $\Delta(r,m,k)$ depends continuously on m, k then from the above arguments it follows that for any $r \in (1,3)$, $R \in (0,\infty)$ there exists $\mu_0 > 0$ such that for any $\mu \in (0,\mu_0)$ uniform rotation of the entire system body+liquid is unstable in the case $\Omega = \Omega_{r,R}$.

Remark. For general geometries the two separate problems,

(i) free oscillations of a uniformly rotating fluid in a bounded domain (the so-called inertial modes of the Poincaré problem for rotating fluids),

(ii) free oscillations of a solid spinning body with no cavity,

both give rise to purely stable spectrum. (This is easily seen by setting (i) $\mathbf{W} = \omega_0\mathbf{e}_3$ or (ii) $\mathbf{u}(\mathbf{r},t) = 0$). However, as we have proved, under certain conditions on shape and relative density the coupled fluid-body system is (nonlinearly) unstable.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I, Comm. Pure Appl. Math., 17 (1959), pp. 623–727.
[2] Yu. M. Berezanskii, *Expansion in Eigenfunctions of Selfadjoint Operators*, Trans. Math. Monographs 17, Amer. Math. Soc., Providence, RI, 1968.
[3] S. J. Friedlander, W. A. Strauss, and M. M. Vishik, *Nonlinear instability in an ideal fluid*, Ann. Inst. H. Poincaré Anal. Non Lineaire, 14 (1997), pp. 187–209.
[4] H. P. Greenspan and L. N. Howard, *On a time-dependent motion of a rotating fluid*, J. Fluid Mech., 17 (1963), pp. 385–404.
[5] S. S. Hough, *The oscillations of a rotating ellipsoidal shell containing fluid*, Philos. Trans. Roy. Soc. London Ser. A, 186 (1895), pp. 469–506.
[6] L. Kelvin, *Mathematical and Physical Papers*, Vol. 3, Cambridge University Press, Cambridge, UK, 1876, p. 322.
[7] N. D. Kopachevsky, S. G. Krein, and Ngo Zui Kan, *Operator Methods in Linear Hydrodynamics: Evolution and Spectral Problems*, Nauka, Moscow, 1989 (in Russian).
[8] H. Lamb, *Hydrodynamics*, 6th ed., Cambridge University Press, Cambridge, UK, 1932.
[9] J. L. Lions, *Quelques méthodes de résolution des prolbèmes aux limites non linéaire*, Dunod-Gauthier-Villars, Paris, 1969.
[10] A. A. Lyashenko, *Instability criterion for a rotating body with an ideal fluid inside*, Stability Appl. Anal. Continuous Media, 2 (1992), pp. 501–528.
[11] A. A. Lyashenko, *On the instability of a rotating body with a cavity filled with viscous liquid*, Japan J. Industrial Appl. Math., 10 (1993), pp. 451–469.
[12] H. Poincaré, *Sur la precession des corps deformables*, Bull. Astronomique, 27 (1910), pp. 321–356.
[13] M. Schechter, *General boundary value problems for elliptic partial differential equations*, Comm. Pure Appl. Math., 12 (1959), pp. 457–486.
[14] S. L. Sobolev, *Motion of a symmetric top with a cavity filled with fluid*, Ž. Prikl. Meh. i Tehn. Fiz., 1 (1960), pp. 20–55 (in Russian).
[15] K. Stewartson and P. H. Roberts, *On the motion of a liquid in a spheroidal cavity of a precessing rigid body*, J. Fluid Mech., 17 (1963), pp. 1–20.
[16] R. Temam, *Local Existence of C^∞ Solutions of the Euler Equations of Incompressible Perfect Fluids*, Lecture Notes in Math 565, Springer, New York, 1976, pp. 184–194.
[17] V.I. Yudovich, *The Linearization Method in Hydrodynamical Stability Theory*, Trans. Math. Monographs 74, Amer. Math. Soc., Providence, RI, 1989.
[18] T.I. Zelenyak and V.P. Mikhaĭlov, *Asymptotic behavior of the solutions of certain boundary value problems of mathematical physics for $t \to \infty$*, Partial Differential Equations, (1970), pp. 96–118 (English translation in Amer. Math. Soc. Transl. Ser 2, 105 (1976), pp. 139–171.).

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 619–636, May 1998

005

ON THE CHAPMAN–JOUGUET LIMIT FOR A COMBUSTION MODEL*

BERNARD HANOUZET†, ROBERTO NATALINI‡, AND ALBERTO TESEI§

Abstract. We study the limiting behavior of solutions to a simple model for combustion waves when the reaction rate tends to infinity. First we establish strong convergence for locally uniformly bounded sequences of solutions. Next we show the uniform boundedness with respect to the reaction rate of piecewise smooth solutions, both for detonation and for deflagration waves.

Key words. small combustion, detonation, deflagration, Chapman–Jouguet limit, conservation laws with source, shock waves

AMS subject classifications. Primary, 35L60, 35L67; Secondary, 76L05, 80A25

PII. S0036141096299351

1. Introduction.

In this paper we investigate the qualitative behavior of solutions to the following simple model for combustion waves:

$$\partial_t(u+v) + \partial_x f(u) = 0 \ , \tag{1.1}$$

$$\partial_t v = -k\varphi(u)v \qquad (k>0) \tag{1.2}$$

for $(x,t) \in \mathbb{R} \times (0,\infty)$, with initial conditions

$$u(x,0) = u_0(x) \ , \quad v(x,0) = v_0(x) \qquad (x \in \mathbb{R}). \tag{1.3}$$

Throughout the paper the functions f and φ are smooth given functions. We also assume that

$$\varphi \text{ is monotone increasing, } \ \varphi(u) \equiv 0 \text{ for } u \le 0. \tag{1.4}$$

The model (1.1)–(1.2) has been proposed independently by Fickett [Fi] and Majda [Ma1] as a simple mathematical analogous for the equations describing one-dimensional compressible flows in a chemically reacting fluid. Existence, uniqueness, and stability of the (entropy) solutions to the Cauchy problem (1.1)–(1.3) have been first established in [TY1]; further results can be found in [TY2], [Le], [HN]. A slightly different model, yet with the same viscous travelling waves of model (1.1)–(1.2), was considered in [MR] and [L1], [L2].

The main purpose of this paper is to investigate the limit of solutions to (1.1)–(1.3) as the reaction rate k diverges. Our approach can be described as follows. Denote by (u^k, v^k) the solution of (1.1)–(1.3) for any fixed $k > 0$. Consider the following Chapman–Enskog-type expansion with respect to the small parameter $1/k$:

$$u^k = u^\infty + \frac{1}{k}\tilde{u} + 0\left(\frac{1}{k}\right), \ v^k = v^\infty + \frac{1}{k}\tilde{v} + 0\left(\frac{1}{k}\right) \ . \tag{1.5}$$

*Received by the editors February 23, 1996; accepted for publication (in revised form) May 27, 1997.

http://www.siam.org/journals/sima/29-3/29935.html

†Département de Mathématiques Appliquées, Université de Bordeaux I, F-33405 Talence cedex, France.

‡Istituto per le Applicazioni del Calcolo "M. Picone," CNR, Viale del Policlinico 137, I-00161 Roma, Italy (natalini@vaxiac.iac.rm.cnr.it).

§Dipartimento di Matematica, Università degli Studi di Roma "La Sapienza," P. le A. Moro 5, I-00185 Roma, Italy (tesei@mat.uniroma1.it).

Introducing (1.5) into the system (1.1)–(1.2) gives, up to terms of order $\frac{1}{k}$,

$$\partial_t(u^\infty + v^\infty) + \partial_x f(u^\infty) = 0 \ , \tag{1.6}$$

$$\partial_t v^\infty \le 0 \ , \qquad v^\infty \ge 0 \ , \tag{1.7}$$

$$\varphi(u^\infty)v^\infty = 0 \ . \tag{1.8}$$

It is natural to try to give a rigorous justification of the above formal argument. This is made in section 3, assuming that the sequence $\{u^k\}$ is locally uniformly bounded in the supremum norm. Under this assumption, by compensated compactness arguments we prove that any sequence $\{(u^k, v^k)\}$ of entropy solutions of (1.1)–(1.3) contains a subsequence that converges to a weak solution (u^∞, v^∞) of problem (1.6)–(1.8) (see Theorem 3.1). Let us observe that this convergence result is far from obvious, since no stability condition—as, for example, the subcharacteristic condition in the case of relaxation problems; see [CLL], [Na]—is satisfied in the present case. In fact, the technical tools of [CLL] have partially inspired our convergence proof; however, an important difference is given by the intrinsic instability of the present problem (see [LZ]).

The remainder of the paper is devoted to investigating the uniform boundedness of the sequence $\{u^k\}$, at least for some classes of initial data.

In section 4 we deal with simple detonation waves; these waves are in general nonmonotone, so usual comparison arguments (like those used, e.g., for relaxation phenomena) are not expedient in this case. On the other hand, using energy methods seems cumbersome for the present problem, due to the lack of regularity of solutions. In our approach, the crucial remark is that the dissipative character of the shock wave affects the reaction mechanism in equation (1.1); hence the absorption effect of equation (1.2) is able to make the solution uniformly bounded. In this connection, let us observe that the uniform boundedness depends in an essential way on the nonlinear character of the flux function $f = f(u)$ (see section 4 for a counterexample when the flux function is linear). A careful study of propagation along characteristics of the above mechanism allows us to extend elementary arguments of the space clamp situation to the general detonation case.

Finally, in section 5 we study deflagration waves using a fixed-point approach. Then we generalize previous results in [TY1], both for more general data and for unbounded reaction rates.

Before proceeding, it is important to motivate our interest in the present problem. Let us recall that two different mathematical theories are widely used to describe the propagation of combustion waves in reacting flows: the Chapman–Jouguet (CJ) theory and the Zeldovich-Von Neumann-Döring (ZND) theory (see [CF], [Wi], [Ma2]). The CJ theory assumes that the reaction rate is *infinitely large* or, equivalently, that the reaction region is *infinitely thin*; on the contrary, the ZND theory makes the assumption that the reaction rate is *finite*. Both theories disregard the effects of viscosity and heat conduction; in fact, they are formulated starting from the classical Euler equations for gas dynamics, which in Lagrangian coordinates for the one-dimensional case read

$$\begin{cases} \partial_t \tau - \partial_x w = 0, \\ \partial_t w + \partial_x P = 0, \\ \partial_t(E) + \partial_x(wP) = 0 \ . \end{cases} \tag{1.9}$$

Here τ is the specific volume, w is the fluid velocity, P is the pressure, and E is the total specific energy—namely, $E = e + q_0 z + \frac{1}{2}w^2$, the quantity e being the specific internal energy, while q_0 denotes the amount of heat released by the chemical reaction and z is the mass fraction of the reactant. The internal energy e and the temperature T are given through state equations $e = e(\tau, P),\ T = T(\tau, P)$, which depend on the gas mixture under consideration.

To make the above system well defined, an equation for the mass fraction z must be provided. This is where the difference between the theories occurs. To specify the variable z, the CJ theory introduces the following relation:

$$z(x,t) = \begin{cases} 0 & \text{if } \sup\limits_{0\le s\le t} T(x,s) > T_i, \\ z(x,0) & \text{if } \sup\limits_{0\le s\le t} T(x,s) \le T_i\ , \end{cases} \tag{1.10}$$

where T_i is a given ignition temperature. On the contrary, in the ZND theory the Ansatz (1.10) is replaced by the following:

$$\partial_t z = -k\varphi(T)z\ , \tag{1.11}$$

where the rate function $\varphi(T)$ has the form

$$\varphi(T) = \begin{cases} 0 & \text{if } T < T_i, \\ T^\alpha \exp\left\{-\frac{A}{T-T_i}\right\} & \text{if } T \ge T_i\ . \end{cases} \tag{1.12}$$

Here T_i is again the ignition temperature, A is the activation energy, and k is the reaction rate ($T_i, A, k > 0$); as for α, this is a dimensionless parameter in the range $(-1, 2]$.

It is natural to ask whether the CJ system (1.9)–(1.10) is in some sense the limit as $k \to \infty$ of the ZND system (1.9), (1.11) (with $\varphi(T)$ as in (1.12)). Observe that this is formally the case, as it is apparent from the equality

$$z(x,t) = z(x,0)e^{-k\int_0^t \varphi(T(x,s))ds}\ .$$

This remark led some authors (see [LZ]) to investigate the equation (1.1) supplemented by the equalities

$$v(x,t) = \begin{cases} 0 & \text{if } \sup\limits_{0\le s\le t} u(x,s) > 0, \\ v(x,0) & \text{if } \sup\limits_{0\le s\le t} u(x,s) \le 0\ . \end{cases} \tag{1.13}$$

Then the system (1.1), (1.13) can be regarded as a simplified version of the CJ model (1.9)–(1.10), while the system (1.1)–(1.2) is the counterpart of the ZND system (1.9), (1.11). The unknown u plays a role similar to that of density, velocity, and temperature in (1.9), while v represents the mass fraction of the unburnt gas. The function φ can be chosen as in (1.12), with $T = u$ and $T_i = 0$.

The reason to consider these simpler models is apparent, due to the complexity, e.g., of (1.9), (1.11). Let us mention that local existence of solutions to the Cauchy problem for this system was proved in [DH] for initial data of small bounded variation (see also [YW]); however, global existence—even for the Riemann problem—seems to be still an open problem.

The Riemann problem for system (1.1), (1.13) was investigated in [LZ] (see also [Zh], where an interesting survey on the whole theory is given). In particular, in [LZ] the authors found up to four distinct configurations for some given initial data to exist. Therefore they proposed to consider as physically eligible those (piecewise smooth, self-similar) solutions for which the number of combustion waves is as small as possible. It was proved in [TY1] that these admissible solutions are the limit of solutions to the Riemann problem (1.1)–(1.3) as the reaction rate k diverges, the limit as $k \to \infty$ was taken by using the explicit structure of the combustion waves (both for detonation and for deflagration) of the Riemann problem, under suitable assumptions on the function φ.

Although the above results are very interesting in themselves, it can be observed that the very mathematical formulation of the limiting problem (1.1), (1.13) is meaningless for generic discontinuous solutions. This seems a strong argument to advocate the use of (1.6)–(1.8) instead of (1.1), (1.13) as the limit of (1.1)–(1.2) when $k \to \infty$. Observe that the solutions of (1.1), (1.13) considered in [LZ] satisfy almost everywhere equation (1.8) (with φ as in (1.12)).

2. Preliminaries. In this section we shall recall some basic definitions and results concerning entropy solutions of the Cauchy problem (1.1)–(1.3). First let us fix some notation. Throughout the paper Σ_T denotes the strip $\mathbb{R} \times (0,T)$ $(T > 0)$, $\Sigma = \Sigma_\infty$. Let us recast the problem (1.1)–(1.2) in a more general way. Consider the weakly coupled first-order quasilinear hyperbolic 2×2 system

$$\partial_t u_i + \partial_x f_i(u_i) = g_i(U) \ , \qquad i = 1,2 \ , \tag{2.1}$$

$$U(x,0) = U^0(x) \ , \tag{2.2}$$

where $U = (u_1, u_2)$ and $U^0 = (u_1^0, u_2^0)$. Here $F(U) = (f_1(u_1), f_2(u_2))$, $G(U) = (g_1(U), g_2(U))$ are given smooth functions.

Let us define the entropy solutions of (2.1)–(2.2) as follows (see [HN]; see also [Kr] for the scalar case).

DEFINITION 2.1. *A vector-valued function $U \in L^\infty(\Sigma_T)^2$ is said to be an entropy solution of* (2.1)–(2.2) *in Σ_T if, for any $\Psi \in C_0^\infty(\Sigma_T)^2$, $\Psi = (\psi_1, \psi_2)$, $\psi_i \geq 0$ $(i = 1,2)$, and any l_1, $l_2 \in \mathbb{R}$, we have*

$$\sum_{i=1}^{2} \iint_{\Sigma_T} \{|u_i - l_i| \, \partial_t \psi_i + \operatorname{sgn}(u_i - l_i)(f(u_i) - f(l_i))\partial_x \psi_i + g_i(U)\psi\} \, dx \, dt \geq 0 \tag{2.3}$$

and, for any interval I,

$$\lim_{T \to 0^+} \frac{1}{T} \int_0^T \int_I |u_i(x,t) - u^0(x)| dx = 0 \ . \tag{2.4}$$

Remark 2.2. Observe that U verifies the condition (2.3) if and only if for any convex function η, setting $q_i := \int^u \eta'(s) f_i'(s) ds$, we have

$$\partial_t \eta(u^i) + \partial_x q_i(u^i) \leq \eta'(u^i) g_i(U)$$

in $\mathcal{D}'$. The pair (η, q_i) is called an entropy pair for the ith equation in (2.1) $(i = 1,2)$.

The following results were proved in [HN].

THEOREM 2.3. (a) *For any $U^\circ \in L^\infty(\mathbb{R})^2$ and $T > 0$ there exists at most one entropy solution U to the Cauchy problem* (2.1)–(2.2) *in* Σ_T.

(b) *For any $U^0 \in L^\infty(\mathbb{R})^2$ there is $T > 0$ (depending only on $\|U_0\|_\infty$) such that in Σ_T there exists an entropy solution U of* (2.1)–(2.2) *and $U \in C([0,T], L^1_{loc}(\mathbb{R})^2)$. Moreover, only two possibilities occur: either $U \in L^\infty(\Sigma_T)^2$ for all $T > 0$, or there exists $T^* < +\infty$ such that U is defined for all $T < T^*$ and*

$$\lim_{T \to T^*_-} \|U\|_{L^\infty(\Sigma_T)^2} = +\infty \ .$$

Concerning our problem (1.1)–(1.2), the following statement can be proved by using the comparison tools in [HN] (see also [TY1], [TY2], [Le]).

THEOREM 2.4. *Let φ verify* (1.4); *assume that φ is Lipschitz continuous and sublinear*; *namely, there is a constant $C > 0$ such that*

$$0 \le \varphi(u) \le Cu \qquad \text{for all } u \ge 0 \ . \tag{2.5}$$

Then for any u_0, $v_0 \in L^\infty(\mathbb{R})$, $v_0 \ge 0$, there is a unique entropy solution (u, v) to the Cauchy problem (1.1)–(1.2). *Moreover, the following global estimates hold true for almost every $(x,t) \in \Sigma$:*

$$0 \le v(x,t) \le v_0(x) \ , \tag{2.6}$$

$$-\|u_0\|_\infty \le u(x,t) \le \|u_0\|_\infty e^{Ck\|v_0\|_\infty t} \ . \tag{2.7}$$

3. The Chapman–Jouguet limit. In this section we are dealing with the limit as $k \to \infty$ of globally defined entropy solutions (u^k, v^k) to problem (1.1)–(1.2) (see Theorem 2.4).

We assume the sequence $\{u^k\}$ to be uniformly locally bounded in L^∞; namely,

$$\begin{cases} \text{for any compact set } B \subseteq \overline{\Sigma} \text{ there exists} \\ \text{a constant } C_B > 0 \text{ such that, for any } k > 0 \ , \\ \qquad \|u^k\|_{L^\infty(B)} \le C_B \ . \end{cases} \tag{3.1}$$

Then we have the following result.

THEOREM 3.1. *Assume that the function f is not affine on any interval. Let the hypotheses of Theorem* 2.4 *and* (3.1) *hold. Then there exist a subsequence (also denoted $\{(u^k, v^k)\}$) of solutions to* (1.1)–(1.2) *and a couple $(\overline{u}, \overline{v}) \in L^\infty_{loc}(\Sigma)$ such that*

$$u^k \longrightarrow \overline{u} \qquad \text{in } L^p_{loc}(\Sigma) \quad (1 \le p < +\infty) \ , \tag{3.2}$$

$$v^k \rightharpoonup \overline{v} \qquad \text{in } L^\infty\text{-weak}^* \ . \tag{3.3}$$

Moreover, for any $\psi \in C^\infty_0(\overline{\Sigma})$ the following holds:

$$\iint (\overline{u} + \overline{v})\partial_t \psi + f(\overline{u})\partial_x \psi \ dx\, dt + \int (u_0(x) + v_0(x))\psi(x,0)dx = 0 \ ; \tag{3.4}$$

$$\overline{v} \ge 0 \ ; \tag{3.5}$$

$$\varphi(\overline{u})\overline{v} = 0 \quad \text{for almost every } (x,t) \in \Sigma; \tag{3.6}$$

$$\partial_t \overline{v} \le 0 \qquad \text{in } \mathcal{D}' \ . \tag{3.7}$$

To prove the above theorem we shall use the following results.

PROPOSITION 3.2 (see [Ta]). *Let $f \in C^1$ and $\Omega \subseteq \mathbb{R}^n$ be a bounded open set. Let $\{u^k\}$ be a sequence of functions satisfying $u^k \rightharpoonup u$ in $L^\infty(\Omega)$–weak*. Suppose that for any convex function η*

$$\partial_t \eta(u^k) + \partial_x q(u^k) \in (\text{compact set of } H^{-1}(\Omega)), \tag{3.8}$$

where $q' = \eta' f'$. Then, if no interval exists on which f is affine,

$$u^k \longrightarrow u \qquad \text{in } L^p(\Omega) \tag{3.9}$$

for any $p \in [1,\infty)$.

LEMMA 3.3 (see [Mu]). *Let Ω be a bounded open set of $\mathbb{R}^n$ and $p \in (1,\infty)$. Let $\{u^k\}$ be any sequence such that $\{u^k\} \in$ (bounded set of $W^{-1,p}(\Omega)$) and $u^k \ge 0$. Then $\{u^k\} \in$ (compact set of $W^{-1,q}(\Omega)$) for all $1 < q < p$.*

LEMMA 3.4 (see [DCL]). *Let Ω be a bounded open set of $\mathbb{R}^n$ and $1 < p \le q < r < +\infty$. Let $\{u^k\}$ be a sequence such that $\{u^k\} \in$ (compact set of $W^{-1,p}(\Omega) \cap$ (bounded set of $W^{-1,r}(\Omega)$. Then $\{u^k\} \in$ (compact set of $W^{-1,q}(\Omega)$).*

Proof of Theorem 3.1. Let $\psi \in C_0^\infty(\mathbb{R})$, $\psi \ge 0$. From the second equation in (1.1) we have, for all $t \ge 0$ and $k > 0$,

$$\partial_t \int v^k(x,t)\psi(x)dx + k \int \varphi(u^k(x,t))v^k(x,t)\psi(x)dx = 0 \ .$$

Then, for any $T > 0$,

$$0 \le \int_0^T \int \varphi(u^k(x,t))v^k(x,t)\psi(x)dx\,dt \le \frac{\|v_0\|_\infty \|\psi\|_{L^1}}{k} \ . \tag{3.10}$$

Now let (η, q) be an entropy pair; namely, η is convex and $q' = \eta' f'$. Then, in view of Remark 2.2, we have, for all $k > 0$,

$$\partial_t \eta(u^k) + \partial_x q(u^k) \le k\eta'(u^k)\varphi(u^k)v^k \qquad \text{in } \mathcal{D}' \ .$$

Setting

$$I^k := k\eta'(u^k)\varphi(u^k)v^k,$$

we have, by (3.1) and (3.10),

$$\|I^k\psi\|_{L^1(\Sigma_T)} \le C \sup_{(0,T)\times(\text{supp}\psi)} |\eta'(u^k)| \le \overline{C} \ ,$$

where $\overline{C}$ depends only on $\|v_0\|_\infty$, η, ψ, and T.

Then the function I^k is in a bounded set of L^1_{loc} for any $k > 0$. Hence by the Sobolev embedding lemma

$$I^k \in \ (\text{compact set in } W^{-1,p_1}_{loc}) \tag{3.11}$$

with $p_1 < 2$. Moreover,

$$\Lambda^k := \partial_t \eta(u^k) + \partial_x q(u^k) - I^k \leq 0,$$

Λ^k being in a bounded subset in W^{-1,p_1}_{loc}. In view of Lemma 3.3 we have

$$\Lambda^k \in (\text{compact set in } W^{-1,p}_{loc}) \tag{3.12}$$

for all $1 \leq p < p_1$. Hence by (3.11)–(3.12), there holds

$$\partial_t \eta(u^k) + \partial_x q(u^k) \in \ (\text{compact set of } W^{-1,p}_{loc}) \ , \tag{3.13}$$

which by assumption (3.1) implies

$$\partial_t \eta(u^k) + \partial_x q(u^k) \in \ (\text{bounded set of } W^{-1,\infty}_{loc}) \ .$$

Then, using Lemma 3.4, we get

$$\partial_t \eta(u^k) + \partial_x q(u^k) \in \ (\text{compact set of } H^{-1}_{loc}) \ .$$

By Proposition 3.2 there exists $\overline{u} \in L^\infty_{loc}(\Sigma)$ and a subsequence (also denoted u^k) such that

$$u^k \longrightarrow \overline{u} \qquad \text{in } L^p_{loc}(\Sigma) \ .$$

Moreover, there exists $\overline{v} \in L^\infty(\Omega)$ and a further subsequence v^k such that

$$v^k \rightharpoonup \overline{v} \ , \ \varphi(u^k) v^k \rightharpoonup \varphi(\overline{u})\overline{v} \ ,$$

both convergences being in the L^∞-weak* sense. On the other hand, using (3.10), we obtain

$$\varphi(u^k) v^k \longrightarrow 0 \qquad \text{in } L^1_{loc} \ ,$$

which implies (3.6). The proof of (3.4), (3.5), and (3.7) is now easy. Hence the conclusion follows. □

4. Simple detonations. In this section we establish a uniform bound in the supremum norm for the solutions of (1.1)–(1.2) in the case where the flow is piecewise smooth with finitely many noninteracting detonation waves. According to the results of section 3, this estimate is of fundamental importance to prove the convergence of the sequence $\{u^k, v^k\}$ as $k \to \infty$.

For simplicity we concentrate on the case where the solution contains only a single detonation; the general case easily follows by similar arguments.

Let us make the following definition.

DEFINITION 4.1. *An entropy solution* (u, v) *of problem* (1.1)–(1.2) *is a* simple detonation *if:*

(i) *the initial data* $(u_0, v_0) \in L^\infty(\mathbb{R}) \cap C^1(\mathbb{R}\backslash\{0\})$;

(ii) $v \in C^1(\mathbb{R}\backslash\{0\})$;

(iii) u *is a* C^1 *function away from a shock curve* $t = \theta(x)$*, which has the following properties:* θ *is defined for* $x > 0$ *and is monotonically increasing;* $\theta(0) = 0$, $\lim_{x\to\infty} \theta(x) = +\infty$;

(iv) *for $x < 0$ and any $t > 0$, u is bounded, uniformly with respect to k. For $x \geq 0$, $t \geq \theta(x)$, u is decreasing in time and is nonnegative. For $x > 0$, $0 < t < \theta(x)$, u is nonpositive.*

Notice that, due to the entropy conditions, the function θ satisfies both the Rankine–Hugoniot condition and the Oleinik entropy condition.

Examples of simple detonations which exhibit the above structure are given in [TY1, Theorem 3] (see also [Le]); there the function φ is increasing such that $\varphi(u) \equiv 1$ if $u \geq \beta > 0$.

Let us assume the following:

(H_1) the function f is C^2 and convex, with $f(0) = f'(0) = 0$. Moreover, there exist $\delta \in (1/2, 1)$ and $M > 0$ such that

$$\frac{uf'(\delta u)}{f(u)} \geq M + 1 \qquad \text{for all } u > 0 \ . \tag{4.1}$$

If $f(u) = Au^\beta$ for $u > 0$ ($A > 0$, $\beta > 1$), assumption (H_1) is satisfied with $\delta^{\beta-1} \geq \frac{M+1}{\beta}$;

(H_2) the function φ is Lipschitz continuous, satisfies (1.4), and there holds $\varphi(u) > 0$ for $u > 0$. Moreover, there exists $\alpha \geq 0$ such that

$$\varphi(u) = 0(u^\alpha) \quad \text{as } u \to +\infty \ .$$

Among the above assumptions the convexity of f is important, as the following example shows.

Example 4.2. Consider the following semilinear problem ($f(u) \equiv u$ in (1.1)):

$$\begin{cases} \partial_t u + \partial_x u = k\varphi(u)v \ , \\ \partial_t v = -k\varphi(u)v \ , \end{cases}$$

with $u_0(x) = H(-x)$ and $v_0(x) = H(x)$, where H is the Heaviside function. It is easily seen that

$$v(x,t) = v_0(x)e^{-k\int_0^t \varphi(u(x,\tau))d\tau} \ ;$$

in particular,

$$v(x,t) \equiv 1 \qquad \text{for } 0 \leq t \leq x.$$

On the other hand, on the characteristic $x = t$ the function u verifies

$$u(t^-,t) = k\int_0^t \varphi(u(\tau^-,\tau))v\,d\tau + 1 \geq k\varphi(1)t + 1 \ .$$

It is clear from the above inequality that condition (3.1) is not satisfied in the present case.

Let us prove the following result.

THEOREM 4.3. *Let (u^k, v^k) be a simple detonation; let assumptions (H_1)–(H_2) be satisfied. Then there exists a constant $C > 0$ (only depending on f, φ, $\|u_0\|_\infty$, $\|v_0\|_\infty$) such that*

$$u^k(x,t) \leq C$$

for all $x > 0$, $t > \theta^k(x)$.

Proof. By the change of variable $\tilde{x} = \frac{x}{k}$, $\tilde{t} = \frac{t}{k}$ it suffices to consider the case $k = 1$, since all the involved quantities do not depend on the scaling; hence we omit the index k in the following. Assume $x \geq 0$, $t \geq \theta(x)$. Fix $x_0 > 0$ and $t_0 = \theta(x_0)$; also set

$$\overline{u} = \max_{x \in [0, x_0]} u(x^-, \theta(x)) \ .$$

Let $\overline{x}$ be such that $\overline{u} = u(\overline{x}^-, \theta(\overline{x}))$; set $\overline{t} = \theta(\overline{x})$.

Observe that by Definition 4.1 (iv) we have

$$\overline{u} = \max_{\substack{\theta(x) \leq t \leq t_0 \\ 0 \leq x \leq x_0}} u(x,t) \ .$$

Consider now the backward characteristic $(X(t), U(t))$ issued from the point $(\overline{x}, \overline{t})$. There holds

$$\begin{cases} \dot{X} = f'(U) \ , \\ \dot{U} = v_0(X(t))\varphi(U) e^{-\int_{\theta(X(t))}^{t} \varphi(u(X(t),\tau))d\tau} \ , \end{cases} \tag{4.2}$$

with $x(\overline{t}) = \overline{x}$, $U(\overline{t}) = \overline{u}$. Let $\overline{\gamma} \in [0, \overline{t}]$ be such that $x(\overline{\gamma}) = 0$, $\overline{\alpha} := u(0, \overline{\gamma})$; set

$$I(t) := \int_t^{\overline{t}} v_0(x(s))\varphi(u(s)) e^{-\int_{\theta(x(s))}^{s} \varphi(u(x(s),\tau))d\tau} ds \tag{4.3}$$

for all $t \in [\overline{\gamma}, \overline{t}]$. Then

$$U(t) + I(t) = \overline{u} \tag{4.4}$$

for all $t \in [\overline{\gamma}, \overline{t}]$. In particular,

$$\overline{\alpha} + I(\overline{\gamma}) = \overline{u} \ . \tag{4.5}$$

If $I(\overline{\gamma}) \leq \frac{1}{2}\overline{u}$, then $\overline{u} \leq 2\overline{\alpha}$. Otherwise we have $\overline{u} > 2u(0, \overline{\gamma})$. For any fixed $\delta \in \left(\frac{1}{2}, 1\right)$ there exists $t_\delta \in (\overline{\gamma}, \overline{t})$ such that $U(t_\delta) = \delta\overline{u}$, from which $I(t_\delta) = (1-\delta)\overline{u}$. Therefore the following estimate holds:

$$\overline{u} \leq \frac{1}{1-\delta} \|v_0\|_\infty \varphi(\overline{u}) \int_{t_\delta}^{\overline{t}} e^{-\int_{\theta(x(s))}^{s} \varphi(u(x(s),\tau))d\tau} ds \ . \tag{4.6}$$

To estimate the integral on the right-hand side we observe that, since $u_t \leq 0$ for $t \geq \theta(x)$, there holds

$$\varphi(u(x(s),\tau)) \geq \varphi(U(s)) \geq \varphi(U(t_\delta)) = \varphi(\delta\overline{u}) \ .$$

Then

$$-\int_{\theta(x(s))}^{s} \varphi(u(x(s),\tau))d\tau \leq -\varphi(\delta\overline{u})(s - \theta(x(s))) \tag{4.7}$$

for $s \in [t_\delta, \overline{t}]$.

Due to the mean value theorem, there exists $\xi \in (x(s), \overline{x})$ and $\tau \in (s, \overline{t})$ such that

$$s - \theta(x(s)) = (s - \overline{t}) \left(1 - \frac{d\theta}{dx}(\xi) f'(U(\tau))\right) \ . \tag{4.8}$$

On the other hand, by the Rankine–Hugoniot condition we have

$$\frac{d\theta}{dx}(\xi) = \frac{u(\xi^-, \theta(\xi)) - u(\xi^+, \theta(\xi))}{f(u(\xi^-, \theta(\xi))) - f(u(\xi^+, \theta(\xi)))} \ ,$$

where $u(\xi^-, \theta(\xi))$ and $u(\xi^+, \theta(\xi))$ are the left, respectively, the right, limit of u on the shock curve. Since θ is monotone increasing, we have $\frac{d\theta}{dx} \geq 0$. Moreover, the function f is convex, $f(0) = f'(0) = 0$, and

$$u(\xi^+, \theta(\xi)) \leq 0 \leq u(\xi^-, \theta(\xi)) \leq \overline{u} \ .$$

It follows that

$$\frac{d\theta}{dx}(\xi) \geq \frac{\overline{u}}{f(\overline{u})}$$

for all $\xi \geq 0$.

On the other hand $U(\tau) \geq \delta\overline{u}$. Combined with (4.8), this gives

$$\begin{aligned} s - \theta(x(s)) &\geq (\overline{t} - s)\left(\frac{\overline{u}}{f(\overline{u})} f'(\delta\overline{u}) - 1\right) \\ &\geq (\overline{t} - s)M \end{aligned} \tag{4.9}$$

by assumption (H_1). Now the inequality (4.6) yields

$$\begin{aligned} \overline{u} &\leq \frac{1}{1-\delta}\|v_0\|_\infty \varphi(\overline{u}) \int_{t_\delta}^{\overline{t}} e^{-\varphi(\delta\overline{u})M(\overline{t}-s)} ds \\ &\leq \frac{1}{1-\delta}\|v_0\|_\infty \frac{\varphi(\overline{u})}{\varphi(\delta\overline{u})M} \ . \end{aligned} \tag{4.10}$$

Then by assumption (H_2) there exist $\alpha \geq 0$, $\tilde{C} > 0$ such that

$$\overline{u} \leq \tilde{C}\frac{\delta^{-\alpha}}{1-\delta}\|v_0\|_\infty \ ,$$

where $\tilde{C}$ depends only on φ, f. Hence the conclusion follows choosing

$$C \geq \max\left(2\overline{\alpha}, \tilde{C}\frac{\delta^{-\alpha}}{1-\delta}\|v_0\|_\infty\right) \ . \qquad \square$$

5. Deflagrations. Deflagration waves are analogous to the rarefaction waves of the nonreactive case. This section is devoted to a detailed investigation of their structure.

Let (u, v) be a solution of problem (1.1)–(1.3) with initial data

$$(u_0(x), v_0(x)) = \begin{cases} (\alpha_1, 0) & \text{if } x < 0 \ , \\ (\alpha_2, \beta_2) & \text{if } x > 0 \ . \end{cases} \tag{5.1}$$

As usual the function φ satisfies condition (1.4), while concerning the function f we make the following assumption:

(H_3) $f(0) = f'(0) = 0$; there exists a constant $\nu > 0$ such that $f''(u) \geq \nu > 0$ for all $u \geq 0$.

Due to the convexity of f, we say that a deflagration wave is generated if

$$0 < \alpha_1 < \alpha_2 \ , \qquad \beta_2 > 0 \ . \tag{5.2}$$

Even in this case it is enough to consider the case $k = 1$. The main result of this section is given in the following theorem, which provides a (global) uniform L^∞ bound for the solution (u, v), as required by Theorem 3.1.

THEOREM 5.1. *Let the assumption* (H_3) *hold and the initial data satisfy* (5.1)–(5.2). *Then the problem* (1.1)–(1.3) *has a unique global entropy solution* (u, v) *such that*

$$\alpha_1 \le u(x,t) \le \alpha_2 + \beta_2 \ , \tag{5.3}$$

$$0 \le v(x,t) \le \beta_2 \tag{5.4}$$

for almost every $(x, t) \in \mathbb{R} \times (0, \infty)$.

Observe that, by contrast with Theorem 2.4, no growth assumption is made here on the function φ. In fact, we shall assume, without loss of generality,

$$\varphi(u) = \varphi(\alpha_2 + \beta_2) \quad \text{for } u \ge \alpha_2 + \beta_2 \ . \tag{H_4}$$

Let us introduce for any $\alpha, \beta \ge 0$ the solution $(\tilde u, \tilde v) = (\tilde u, \tilde v)(\alpha, \beta)$ of the ordinary differential system

$$\begin{cases} \dot{\tilde u} = \varphi(\tilde u)\tilde v, \\ \dot{\tilde v} = -\varphi(\tilde u)\tilde v, \end{cases} \tag{5.5}$$

with Cauchy data $\tilde u(0) = \alpha,\ \tilde v(0) = \beta$. It is easily seen that

$$\tilde u + \tilde v = \alpha + \beta, \quad 0 \le \tilde v \le \beta, \quad \alpha \le \tilde u \le \alpha + \beta \ .$$

Set

$$\tilde x(t) = \int_0^t f'(\tilde u(\tau))d\tau \qquad (t > 0) \ ; \tag{5.6}$$

then the half-plane Σ is the union of the following regions:

$$\begin{aligned} I &:= \{(x,t) \,|\, x \le 0 \ ,\ t > 0\} \ ; \\ II &:= \{(x,t) \,|\, 0 < x \le \tilde x(t) \ ,\ t > 0\} \ ; \\ III &:= \{(x,t) \,|\, x > \tilde x(t) \ ,\ t > 0\} \ . \end{aligned}$$

Since $\dot{\tilde x} = f'(\tilde u) \ge 0$ and $\ddot{\tilde x} = f''(\tilde u)\dot{\tilde u} = f''(\tilde u)\varphi(\tilde u)\tilde v \ge 0$, the curve $x = \tilde x(t)$ is strictly increasing and convex for $t > 0$.

Now we can state the following result, which completely describes the structure of deflagration waves.

THEOREM 5.2. *Let the assumption* (H_3)–(H_4) *hold and the initial data satisfy* (5.1)–(5.2). *Then there exists a unique global entropy solution* (u, v) *to problem* (1.1)–(1.3) *such that* $u \in \mathrm{Lip}_{loc}(\overline{\Sigma} \setminus (0,0))$. *Moreover,*

$$(u, v) = (\alpha_1, 0) \quad \text{in region } I \ , \tag{5.7}$$

$$(u, v) = (\tilde u, \tilde v)(\alpha_2, \beta_2) \quad \text{in region } III \tag{5.8}$$

and there exists $\overline{M} > 0$ *such that*

$$0 \le \partial_x u \le \overline{M} \max\left(\frac{1}{t}, 1\right) \ , \tag{5.9}$$

$$-\frac{\overline{M}}{t} \le \partial_t u \le 0 \quad \text{in region } II \ . \tag{5.10}$$

It is easily seen that Theorem 5.1 follows from the above result; hence the remainder of the paper is devoted to the proof of the latter.

For this purpose, observe that (1.2)–(1.3) (with $k = 1$) entail

$$v(x,t) = v_0(x) e^{-\int_0^t \varphi(u(x,s))ds} \ . \tag{5.11}$$

Then it is natural to investigate the integro-differential problem

$$\partial_t u + \partial_x f(u) = \varphi(u) v_0(x) e^{-\int_0^t \varphi(u(x,s))ds} \ , \tag{5.12}$$

$$u(x,0) = u_0(x) \ . \tag{5.13}$$

Let us introduce the space $\mathcal{A}^M(\Sigma)$, as the space of $w \in \mathrm{Lip}_{loc}(\overline{\Sigma} \setminus (0,0))$ such that

$$\alpha_1 \le w \le \alpha_2 + \beta_2; \tag{P_1}$$

$$w = \alpha_1 \quad \text{in region } I; \tag{P_2}$$

$$w = \tilde{u}(\alpha_2, \beta_2) \quad \text{in region } III; \tag{P_3}$$

(P_4) there exists $M > 0$ such that

$$0 \le \partial_x w \le M \max\left(\frac{1}{t}, 1\right) \quad \text{in } \Sigma;$$

(P_5) there exists $M > 0$ such that

$$-\frac{M}{t} \le \partial_t w \le 0 \quad \text{in region } II \ .$$

For any $w \in \mathcal{A}^M(\Sigma)$ we define a map $u = \Lambda(w)$ by considering the (unique) entropy solution of the problem

$$\partial_t u + \partial_x f(u) = \varphi(u) v_0(x) e^{-\int_0^t \varphi(w(x,s))ds} \ , \tag{5.14}$$

$$u(x,0) = u_0(x) \ . \tag{5.15}$$

This solution exists by classical results (see section 2), in view of the assumption (H_4).

The first step in the proof of Theorem 5.2 is given in the following proposition.

PROPOSITION 5.3. *There exists* $M > 0$ *such that, for any* $w \in \mathcal{A}^M(\Sigma)$*, there holds* $\Lambda(w) \in \mathcal{A}^M(\Sigma)$.

The proof of Proposition 5.3 will be given after two preliminary lemmas (see Lemmas 5.4–5.5 below). Set

$$v(x,t) = v_0(x) e^{-\int_0^t \varphi(w(x,s))ds} \ , \tag{5.16}$$

where $v_0 \equiv 0$ for $x \le 0$, $v_0 \equiv \beta_2$ for $x > 0$. Then $u \equiv \alpha_1$ for $x \le 0$. To deal with the case $x > 0$ we solve the system for the characteristics of equation (5.14), namely,

$$\dot{X} = f'(U),\ \dot{U} = \varphi(U)V\ , \tag{5.17}$$

where $X = X(t),\ U = U(t),\ V(t) = v(X(t), t)$. Since $\dot{X} > 0$, the function X is invertible; denote by $T = T(x)$ its inverse function for $x > 0$. Then the system (5.17) is equivalent to

$$\frac{dT}{dx} = \frac{1}{f'(U)}, \quad \frac{dU}{dx} = \frac{\varphi(U)}{f'(U)}V\ , \tag{5.18}$$

where $U(x) = U(T(x))$ and $V(x) = V(T(x))$.

We supplement (5.18) with two different sets of initial data, namely,

$$(T(0), U(0)) = (\gamma, \alpha_1) \qquad \text{for } \gamma \ge 0, \tag{5.19}$$

or, respectively,

$$(T(0), U(0)) = (0, \alpha) \qquad \text{for } \alpha \in [\alpha_1, \alpha_2]\ . \tag{5.20}$$

Due to the above assumptions (in particular (H_4)), it is easily seen that the solutions of system (5.18) with initial conditions (5.19) or (5.20) are globally defined for $x > 0$. The following holds.

LEMMA 5.4. *Characteristic curves (T, U) corresponding to different values of the parameter γ or α do not intersect themselves for $x > 0$. The following inequalities hold:*

(i) $\frac{\partial U}{\partial \gamma}(x, \gamma) < 0,\ \frac{\partial T}{\partial \gamma}(x, \gamma) \ge 1\ (x > 0, \gamma \ge 0)$;

(ii) $\frac{\partial U}{\partial \alpha}(x, \alpha) > 0,\ \frac{\partial T}{\partial \alpha}(x, \alpha) < 0\ (x > 0, \alpha \in [\alpha_1, \alpha_2])$.

Moreover, the families $\{T(x, \gamma), \gamma \ge 0\}$ and $\{T(x, \alpha), \alpha \in [\alpha_1, \alpha_2]\}$ cover the entire region II.

Proof. Let us only consider (5.18)–(5.19), since problem (5.18), (5.20) can be dealt with by similar arguments. Let $0 \le \gamma_1 < \gamma_2$; set

$$\tilde{T}(x) = T_2(x) - T_1(x)\ ,\ \tilde{U}(x) = U_2(x) - U_1(x)\ ,$$

with obvious meaning of the symbols. An elementary calculation shows that

$$\tilde{T}(0) > 0\ , \qquad \frac{d\tilde{T}}{dx}(0) = 0\ , \qquad \frac{d^2\tilde{T}}{dx^2}(0) > 0\ , \tag{5.21}$$

$$\frac{d\tilde{T}}{dx} = -\frac{(\int_0^1 f''(\theta U_1 + (1-\theta)U_2)d\theta)\tilde{U}}{f'(U_1)f'(U_2)} \qquad (x > 0)\ , \tag{5.22}$$

$$\frac{d\tilde{U}}{dx} = \frac{\varphi(U_2)V(x, \gamma_2)}{f'(U_2)} - \frac{\varphi(U_1)V(x, \gamma_1)}{f'(U_1)} \qquad (x > 0). \tag{5.23}$$

By (5.22) there is a right neighborhood of $x = 0$ where $\tilde{T} > 0$, $\frac{d\tilde{T}}{dx} > 0$. Let $x_1 > 0$ exist such that $\tilde{T}(x) > 0$ for $x \in (0, x_1)$, $\tilde{T}(x_1) = 0$. Then there is $x_0 \in (0, x_1)$ such that $\frac{d\tilde{T}}{dx} > 0$ in $(0, x_0)$ and $\frac{d\tilde{T}}{dx}(x_0) = 0$. Hence by (5.22), $\tilde{U} < 0$ in $(0, x_0)$, $\tilde{U}(x_0) = 0$, which gives $\frac{d\tilde{U}}{dx}(x_0) \ge 0$. On the other hand, since $\tilde{T}(x_0) > 0$ we have

$$\begin{aligned} &V(x_0, \gamma_2) - V(x_0, \gamma_1) \\ &\quad = \beta_2 \left\{ e^{-\int_0^{T(x,\gamma_2)} \varphi(w(x_0,s))ds} - e^{-\int_0^{T(x,\gamma_1)} \varphi(w(x_0,s))ds} \right\} < 0\ . \end{aligned}$$

It follows from (5.23) that $\frac{d\tilde{U}}{dx}(x_0) < 0$; the contradiction proves the claim.

Inequalities (i) and (ii) are proved by similar arguments. To prove the last claim it suffices to show that $T(x, \alpha_2) = \tilde{t}(x)$ $(x \geq 0)$, where $\tilde{t}(\cdot)$ denotes the inverse function of $\tilde{x}(\cdot)$. Observe first that $(X(t, \alpha_2), U(t, \alpha_2))$ satisfies the equations

$$\dot{X} = U,\ \dot{U} = \varphi(U)V\ ,$$

with $X(0) = 0$, $U(0) = \alpha_2$, and $V(t) = \beta_2\, e^{-\int_0^t \varphi(w(X(t),s))ds}$. On the other hand, setting $X(t) = \tilde{x}(t)$ in the expression for $V(t)$ gives $V(t) = \tilde{v}(t)$. Since the couple $(\tilde{u}, \tilde{v}) = (\tilde{u}, \tilde{v})(\alpha_2, \beta_2)$ verifies

$$\dot{\tilde{x}} = \tilde{u},\ \dot{\tilde{u}} = \varphi(\tilde{u})\tilde{v} = \varphi(\tilde{u})V\ ,$$

and $\tilde{x}(0) = 0$, $\tilde{u}(0) = \alpha_2$, the conclusion follows by the uniqueness theorem for ordinary differential equations. □

Set

$$II_a = \{(x,t) \in II \,|\, t > T(x, \alpha_1)\}\ ;$$
$$II_b = \{(x,t) \in II \,|\, T(x, \alpha_2) < t \leq T(x, \alpha_1)\}\ .$$

For any $(x,t) \in II_a$, there exists a unique value $\gamma = \gamma(x,t) > 0$ such that $t = T(x, \gamma)$. In the same way, for any $(x,t) \in II_b$, there exists a unique $\alpha = \alpha(x,t) \in [\alpha_1, \alpha_2)$ such that $t = T(x, \alpha)$. Hence we can define, in region II,

$$\lambda(x,t) = \begin{cases} \gamma(x,t) & \text{if } (x,t) \in II_a, \\ \alpha(x,t) & \text{if } (x,t) \in II_b\ . \end{cases} \tag{5.24}$$

The function λ is invertible and differentiable. Define a function $u : \Sigma \to [\alpha_1, \infty)$ as follows:

$$u(x,t) = \begin{cases} \alpha_1 & \text{in region } I \setminus (0,0), \\ U(x, \lambda, (x,t)) & \text{in region } II, \\ \tilde{u}(t) & \text{in region } III\ . \end{cases} \tag{5.25}$$

The elementary proof of the following result is omitted.

LEMMA 5.5. *The function u given by* (5.25) *is in $C^1(\overline{\Sigma} \setminus \{(0,0)\})$ and solves problem* (5.12)–(5.13).

Proof of Proposition 5.3. Let $u = \Lambda w$ as defined by (5.25). First let us show that

$$\partial_x u \geq 0 \quad \text{in } \Sigma,\ \partial_t u < 0 \quad \text{in region } II\ .$$

In region I or region III, $\partial_x u \equiv 0$. On the other hand $u(x,t) = U(x, \lambda(x,t))$ in region II; hence $\partial_t u < 0$ by Lemma 5.4. Also, $f'(u)\partial_x u = -\partial_t u + \varphi(u)v > 0$. Since $\tilde{u}(t) \leq \alpha_2 + \beta_2$ in region III and $\partial_t u < 0$ in region II, we have

$$u(x,t) \leq \alpha_2 + \beta_2 \qquad \text{in region } II\ .$$

Hence properties (P_1), (P_2), (P_3) of the space $\mathcal{A}^M(\Sigma)$ are satisfied. We must check that there exists $M > 0$ such that, if $w \in \mathcal{A}^M(\Sigma)$, (P_4) and (P_5) are verified for u. Set $p = \partial_x u$ and differentiate equation (5.12). We have

$$\partial_t p + f'(u)\partial_x p + p^2 = \varphi'(u)vp + \varphi(u)\partial_x v\ . \tag{5.26}$$

For $x = 0$, it follows from the equation that

$$p(0,\gamma) = (\partial_x u)(0,\gamma) = \frac{\varphi(\alpha_1)}{\alpha_1}\beta_2 e^{-\varphi(\alpha_1)\gamma} \ . \tag{5.27}$$

Let $M_1 = \beta_2 \max(\sup_{u\in(\alpha_1,\alpha_2+\beta_2)} \varphi'(u), \frac{\varphi(\alpha_1)}{\alpha_1})$. Then $\overline{p} := M_1$ is a supersolution for problem (5.26)–(5.27) in region II_a. By standard comparison arguments on the characteristic curves we have

$$0 \le \partial_x u \le M_1 \tag{5.28}$$

and

$$\begin{aligned} 0 \le -\partial_t u &= u\partial_x u - \varphi(u)v \\ &\le (\alpha_2+\beta_2)M_1 \quad \text{for all } (x,t) \in II_a \ . \end{aligned} \tag{5.29}$$

To estimate the derivatives of u in region II_b we set

$$P(t,\alpha) = p(X(t,\alpha),t) = (\partial_x u)(X(t,\alpha),t) \ .$$

Using (5.18), (5.20), this yields

$$P(t,\alpha) = \frac{\varphi(U)}{f'(U)}V - \frac{1}{f'(U)}\ \frac{\partial U/\partial\alpha}{\partial T/\partial\alpha} \ .$$

Then

$$\lim_{t\to 0^+} t\ P(t,\alpha) = -\lim_{t\to 0^+} \frac{1}{f'(U)}\ \frac{\partial U}{\partial\alpha}\ \frac{t}{\frac{\partial T}{\partial\alpha}(X(t,\alpha),\alpha)} \ .$$

By elementary arguments it easy to show that

$$\lim_{t\to 0^+} t\ P(t,\alpha) = \frac{1}{f''(\alpha)} \ . \tag{5.30}$$

Observe now that P is a solution of the ordinary differential equation

$$\dot{P} = P^2 + \varphi'(U)VP + \varphi(U)v_x(X(t,\alpha),t). \tag{5.31}$$

Then

$$P(t,\alpha) \le \frac{M_2}{t} \ ,$$

where

$$M_2 = \max\left(\frac{1}{f''(\alpha)}, 1 + \left[\sup_{\alpha_1\le u\le\alpha_2+\beta_2} \varphi'(U)\right]\frac{\beta_2}{\varphi(\alpha_1)}\right) \ ,$$

since $\overline{P} = \frac{M_2}{t}$ is a supersolution of (5.31)–(5.30). Choosing $M = \max(M_1, M_2)$, the conclusion follows easily. □

The second step in the proof of Theorem 5.2 is given by the following fixed point result.

PROPOSITION 5.6. *Let M be given by Proposition* 5.3. *Then the map* $\Lambda : \mathcal{A}^M(\Sigma) \to \mathcal{A}^M(\Sigma)$ *has a fixed point.*

Proof. We shall use the Schauder theorem in the space

$$\mathcal{A}(Q) = \mathcal{A}^M(\Sigma)|_Q \ ,$$

where $Q = [0,T] \times [-R,R]$ $(T > 0, R > 0)$. Set also

$$B(\Sigma) = t\mathcal{A}^M(\Sigma), \ B(Q) = t\mathcal{A}(Q) \ .$$

Step 1. For any fixed R, T, the set $B(Q)$ is obviously a convex, bounded subset of $L^1(Q)$. Moreover, due to the properties (P_4)–(P_5) of the space $\mathcal{A}^M$, for any $z \in B(Q)$, $\partial_t z, \partial_x z \in L^\infty(Q)$. Next we claim that $B(Q)$ is a closed subset of $L^1(Q)$. In fact, let $\{z_n\} \subset B(Q)$ such that

$$z_n \longrightarrow z \qquad \text{in } L^1(Q) \ .$$

Then for any multi-index $\alpha = (\alpha_1, \alpha_2)$,

$$D^\alpha z_n \rightharpoonup D^\alpha x \qquad \text{in } \mathcal{D}(Q) \ ,$$

where D^α denotes any partial derivative with respect to the (x,t)-variables. Since $D^\alpha z_n$ is in a bounded set of $L^\infty(Q)$ for any α with $|\alpha| = 1$, there exist $\xi^\alpha \in L^\infty(Q)$ and a subsequence of $\{z_n\}$, say $\{z_{n_k}\}$, such that

$$D^\alpha z_{n_k} \rightharpoonup \xi^\alpha \qquad \text{in } L^\infty - \text{weak}^* \ .$$

Hence $D^\alpha z = \xi^\alpha \in L^\infty(Q)$ and the claim follows. Then the compact embedding of $B(Q)$ in $L^1(Q)$ follows, since $B(Q) \subseteq W^{1,\infty}(Q)$.

Step 2. To define a continuous map from $B(Q)$ in itself, fix $t \le T$ and $R > \tilde{x}(T)$. For $w_1, w_2 \in \mathcal{A}(Q)$, let $u_1 = \Lambda w_1$ and $u_2 = \Lambda w_2$. By the Gauss–Green formula applied to equation (5.12) we have

$$\begin{aligned}
&\int_{-R}^{R} |u_2(x,t) - u_1(x,t)|dx \\
&\quad = -\int_0^t [f(u_2) - f(u_1)](\tau, R)d\tau + \int_0^t [f(u_2) - f(u_1)](\tau, -R)d\tau \\
&\quad + \int_0^t \int_{-R}^{R} |\varphi(u_2)v_2 - \varphi(u_1)v_1|d\tau\, dx \\
&\quad \le \int_0^t \int_{-R}^{R} |\varphi(u_2)v_2 - \varphi(u_1)v_1|d\tau\, dx.
\end{aligned} \tag{5.32}$$

Since v_i is bounded, φ is Lipschitz continuous and by (5.16)

$$\begin{aligned}
&\int_0^t |v_2(x,\tau) - v_1(x,\tau)|d\tau \\
&\quad = \beta_2 \int_0^t \left| e^{-\int_0^\tau \varphi(w_2(x,s))ds} - e^{-\int_0^\tau \varphi(w_1(x,s))ds} \right| d\tau \\
&\quad \le \overline{C}_1 \int_0^t |w_2(x,\tau) - w_1(x,\tau)|d\tau \ ,
\end{aligned}$$

there exists a constant $\overline{C} > 0$ (depending only on T, φ, α_1, α_2, β_2, M) such that

$$\begin{aligned}\int_0^t \int_{-R}^R |\varphi(u_1)v_1 - \varphi(u_2)v_2| dx\, d\tau \\ \leq \overline{C} \int_0^t \int_{-R}^R [\,|u_1(x,\tau) - u_2(x,\tau)| \\ + |w_1(x,\tau) - w_2(x,\tau)|\,] dx\, d\tau\ .\end{aligned} \tag{5.33}$$

Due to (5.32)–(5.33), by the Gronwall inequality there exists a constant $C > 0$ (which depends on T, φ, α_1, α_2, β_2, M) such that

$$\int_{-R}^R |u_2(x,t) - u_1(x,t)| dx \leq C \int_0^t \int_{-R}^R |w_2(x,\tau) - w_1(x,\tau)| dx\, d\tau\ . \tag{5.34}$$

Now define a map $\mathcal{T} : B(Q) \to B(Q)$ setting

$$\mathcal{T}(z) = t\Lambda\left(\frac{1}{t}z\right) \quad \text{for all } z \in B(Q)\ . \tag{5.35}$$

Step 3. Let us prove that the map $\mathcal{T}$ defined in (5.35) is continuous in the L^1 topology. Set $z_i = tw_i$ $(i = 1, 2)$. Then (5.34) yields

$$\begin{aligned}\|\mathcal{T}z_1 - \mathcal{T}z_2\|_{L^1(Q)} &= \int_0^T \int_{-R}^R t|u_1(x,\tau) - u_2(x,\tau) dx\, d\tau \\ &\leq T \int_0^T \int_{-R}^R |u_1(x,\tau) - u_2(x,\tau)| dx\, d\tau \\ &\leq CT \int_0^T dt \int_0^t \int_{-R}^R \frac{1}{\tau}|z_1(x,\tau) - z_2(x,\tau)| dx\, d\tau \\ &\leq C' \int_0^T \int_{-R}^R \frac{1}{\tau}|z_1(x,\tau) - z_1(x,\tau)| dx\, d\tau\ .\end{aligned}$$

Fix $\varepsilon > 0$. Since $z_i = tw_i$ and $\alpha_1 \leq w_i \leq \alpha_2 + \beta_2$ $(1 = 1, 2)$, there exists $t_\varepsilon > 0$ such that

$$\int_0^{t_\varepsilon} \int_{-R}^R \frac{1}{\tau}|z_1(x,\tau) - z_2(x,\tau)| dx\, d\tau \leq \varepsilon\ .$$

On the other hand,

$$\int_{t_\varepsilon}^T \int_{-R}^R \frac{1}{\tau}|z_1(x,\tau) - z_2(x,\tau)| dx\, d\tau \leq \frac{1}{t_\varepsilon}\|z_1 - z_2\|_{L^1(Q)}\ .$$

Therefore, for any $\varepsilon > 0$ there is $\delta > 0$ such that if $\|z_1 - z_2\|_{L^1(Q)} < \delta$, then $\|\mathcal{T}z_1 - \mathcal{T}z_2\|_{L^1(Q)} < \varepsilon$. Hence the continuity of the map $\mathcal{T}$ is proved.

Due to Steps 1–3 above, for any fixed rectangle Q there is a fixed point of the map $\mathcal{T}$ in $B(Q)$. Let $\overline{z}$ be this fixed point. Then, the function $\overline{u} = \frac{1}{z}\overline{z}$ is a fixed point of Λ in $B(Q)$ and the conclusion follows. □

Proof of Theorem 5.2. It follows by Propositions 5.3 and 5.6. □

Acknowledgments. The authors are grateful to Dr. Dechun Tan for helpful discussions.

REFERENCES

[CLL] G.-Q. Chen, C. Levermore, and T.P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[CF] R. Courant and K.O. Friedrichs, *Supersonic Flows and Shock Waves*, Springer-Verlag, New York, 1949.

[DCL] X. Ding, G.-Q. Chen, and P. Luo, *Convergence of the Lax-Friedrichs scheme for isentropic gas dynamics* (I), Acta Math. Sci., 5 (1985), pp. 415–432.

[DH] C.M. Dafermos and L. Hsiao, *Hyperbolic systems of balance laws with inhomogenity and dissipation*, Indiana U. Math. J., 31 (1982), pp. 471–491.

[Fi] W. Fickett, *Detonation in miniature*, Amer. J. Phys., 47 (1979), pp. 1050–1059.

[HN] B. Hanouzet and R. Natalini, *Weakly coupled systems of quasilinear hyperbolic equations*, Differential Integral Equations, 9 (1996), pp. 1279–1292.

[Kr] S.N. Kruzhkov, *First order quasilinear equations in several independent variables*, Mat. Sb., 81 (1970) (in Russian); Math. USSR Sb., 10 (1970), pp. 217–243 (in English).

[Le] A. Levy, *On Majda's model for dynamic combustion*, Comm. Partial Differential Equations, 17 (1992), pp. 657–698.

[L1] T. Li, *On the Riemann problem for a combustion model*, SIAM J. Math. Anal., 24 (1993), pp. 59–75.

[L2] T. Li, *On the initiation problem for a combustion model*, J. Differential Equations, 112 (1994), pp. 351–373.

[LZ] T.P. Liu and T. Zhang, *A scalar combustion model*, Arch. Rational Mech. Anal., 114 (1991), pp. 297–312.

[Ma1] A. Majda, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math., 41 (1981), pp. 70–93.

[Ma2] A. Majda, *High Mach Number Combustion*, Lectures in Applied Mathematics 24, American Mathematical Society, Providence, RI, 1986.

[MR] A. Majda and R. Rosales, *Weakly nonlinear detonation waves*, SIAM J. Appl. Math., 43 (1983), pp. 1086–1118.

[Mu] F. Murat, *L'injection du cone positif de H^{-1} dans $W^{-1,q}$ est compacte pour tout $q < 2$*, J. Math. Pures Appl., 60 (1981), pp. 309–322.

[Na] R. Natalini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 795–823.

[Ta] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Research Notes in Mathematics, Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. 4, R.J. Knops, ed., Pitman Press, New York, 1979, pp. 136–212.

[TY1] Z. Teng and L. Ying, *Riemann problem for a reacting and convection hyperbolic system*, Approx. Theory Appl., 1 (1984), pp. 95–122.

[TY2] Z. Teng and L. Ying, *Existence, uniqueness and convergence as vanishing viscosity for a reaction-diffusion-convection system*, Acta Math. Sinica (N.S.), 5 (1989), pp. 114–135.

[YW] L. Ying and C. Wang, *The discontinuous initial value problem of a reacting gas flow system*, Trans. Amer. Math. Soc., 266 (1981), pp. 361–387.

[Wi] F.A. Williams, *Combustion Theory*, Benjamin/Cummings, Menlo Park, CA, 1985.

[Zh] T. Zhang, *The Riemann problem for combustion*, Contemp. Math., 100 (1989), pp. 111–124.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 637–651, May 1998

006

ON MAXWELL'S EQUATIONS IN AN ELECTROMAGNETIC FIELD WITH THE TEMPERATURE EFFECT*

HONG-MING YIN†

Abstract. This paper deals with Maxwell's equations coupled with a nonlinear heat equation. The system models an induction heating process for a conductive material in which the electrical conductivity strongly depends on the temperature. It is shown that the evolution system has a global weak solution if the electrical conductivity is bounded. For the case of one space dimension, the existence of a global classical solution is established. Moreover, for a quasi-stationary state field it is proved that the temperature will blow up in finite time if the electric conductivity satisfies certain growth conditions.

Key words. macrowave heating, Maxwell's system with thermal effect

AMS subject classifications. 35K55, 35L40, 35Q20

PII. S0036141097316159

1. Introduction.

Induction heating is commonly used in industrial operations such as metal hardening and preheating for forging operations (see [5, 14]). The investigation of an induction heating system usually relies upon a series of expensive, long, and complicated experiments. The mathematical analysis and numerical simulation for induction heating play an important role in the designing process. In this paper we shall investigate how a conductive material is heated up by using electromagnetic waves.

The mathematical model of induction heating consists of Maxwell's equations coupled with a nonlinear heat equation (see section 2). The analysis of this full system is quite complicated. The common method in the previous study of induction heating is to assume that the electric field $\mathbf{E}$ is given by certain special time-harmonic form. With this assumption for $\mathbf{E}$, one can decouple Maxwell's equations from the nonlinear heat equation and then study the heat equation alone (see [4, 5, 8, 9, 10, 14, 16] for example). This method is a good approximation of the full system when the electrical conductivity of the targeted material is not sensitive with respect to the change of the temperature. The microwave cooking is partly based on this approximation (see [14]), since the electric conductivity of most water-like foodstuffs has only a small change with respect to the change of temperature. However, the electric field $\mathbf{E}$ is much more complicated than the guessed form if the electric conductivity strongly depends on the temperature. One must take into account the effect of the temperature and investigate the full Maxwell equations along with the nonlinear heat equation.

This paper is devoted to the investigation of the full Maxwell equations coupled with a nonlinear heat equation. We show that the full evolution system has a global solution in some weak sense. The existence proof is based on a standard fixed point argument. The technical difficulty is to show that the mapping is continuous in applying Schauder's fixed point theorem. This difficulty is resolved from the help of the boundedness assumption on the electric conductivity. For a special case where an electrical field depends only one space variable, Maxwell's equation becomes a wave equation with a nonlinear damping which depends on the temperature. For the

*Received by the editors February 4, 1997; accepted for publication March 21, 1997.
http://www.siam.org/journals/sima/29-3/31615.html
†Department of Mathematics, University of Notre Dame, Notre Dame, IN 46656 (hong-ming.yin.3@nd.edu).

damped wave equation, we are able to derive the a priori L^∞-bound of the electric field. With this a priori bound, we establish the existence of a global solution in the classical sense.

The paper is organized as follows. In section 2, we formulate the mathematical model for completeness. The existence of a global weak solution is established in section 3. In section 4, we consider a special case where the electrical field depends only on one space variable and show that the full system has a unique global solution in the classical sense. In section 5, we present a blowup result for temperature in a quasi-stationary field.

2. The mathematical model. For completeness, we shall formulate the mathematical model of an induction heating process. Let a certain conductive material occupy a bounded domain $\Omega \subset R^3$. Let $\mathbf{E}(x,t)$ and $\mathbf{H}(x,t)$ denote the electric and magnetic fields in Ω (hereafter, a bold letter means a vector in R^3). Let $\mathbf{D}(x,t)$ and $\mathbf{B}(x,t)$ be the electric displacement and magnetic induction in Ω, respectively.

Then the classical Maxwell equations hold in Ω (see [11]):

$$\nabla \times \mathbf{E} + \mathbf{B}_t = 0 \qquad \text{(Faraday's law)},$$
$$\nabla \times \mathbf{H} = \mathbf{J} + \mathbf{D}_t \qquad \text{(generalized Ampere's law)},$$

where $\mathbf{J}$ represents the electric current and $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$.

Moreover, we will assume the following constitutive relations [11]:

$$\mathbf{D} = \varepsilon \mathbf{E}, \qquad \mathbf{B} = \mu \mathbf{H}, \qquad \mathbf{J} = \sigma \mathbf{E} \qquad \text{(Omh's law)},$$

where ε, μ, and σ are electric permittivity, the magnetic permeability, and the electric conductivity of the medium, respectively.

Let $u(x,t)$ denote the temperature in Ω. Due to the complexity of the dynamical process, we shall only study how the temperature through the electric conductivity, $\sigma = \sigma(u)$, affects the electromagnetic field in the conductive material. Therefore, we shall normalize the physical constants ε and μ and set $\varepsilon = \mu = 1$.

Note that the local Joule heat generated by the current equals

$$\mathbf{E} \cdot \mathbf{J} = \mathbf{E} \cdot [\sigma(u)\mathbf{E}] = \sigma(u)|\mathbf{E}|^2.$$

Eliminating $\mathbf{D}$ and $\mathbf{B}$ and applying Fourier's law and the conservation of energy for the temperature $u(x,t)$, we obtain the following evolution system for $\mathbf{E}$(x,t), $\mathbf{H}(x,t)$, and $u(x,t)$:

$$\mathbf{E}_t + \sigma(u)\mathbf{E} = \nabla \times \mathbf{H}, \qquad (x,t) \in Q_T, \tag{2.1}$$
$$\mathbf{H}_t + \nabla \times \mathbf{E} = 0, \qquad (x,t) \in Q_T, \tag{2.2}$$
$$u_t - \nabla[k(u)\nabla u] = \sigma(u)|\mathbf{E}|^2, \qquad (x,t) \in Q_T, \tag{2.3}$$

where $Q_T = \Omega \times (0,T]$ with $T > 0$, $\sigma(u)$ and $k(u)$ denote the electrical and the thermal conductivity, respectively, and other physical parameters have been normalized.

When a conductive material is heated up by using induction heating, the mechanism in the material for the electric field $\mathbf{E}$, the magnetic field $\mathbf{H}$, and the temperature u will be determined by the coupled system (2.1)–(2.3) subject to appropriate initial and boundary values.

The system (2.1)–(2.3) has certain similarity to the well-known Maxwell–Vlasov system which has been studied by many authors (see [6, 7] and the references therein). However, to the best of the author's knowledge, little is known about the evolution

system (2.1)–(2.3). This paper is the first attempt to answer some mathematical questions such as the well-posedness of the system (2.1)–(2.3) subject to appropriate initial and boundary conditions. We shall prescribe the following initial-boundary conditions:

$$\nu \times \mathbf{E} = 0,\ u = 0, \qquad (x,t) \in \partial\Omega, \quad t > 0, \tag{2.4}$$

$$\mathbf{E}(x,0) = \mathbf{E}_0(x), \qquad \mathbf{H}(x,0) = \mathbf{H}_0(x), \qquad u(x,0) = u_0(x), \qquad x \in \Omega. \tag{2.5}$$

In some industrial applications (see [3] for example), the electrical displacement $\mathbf{D}$ is negligible. This field is often referred to as quasi-stationary in physics. Since $\mathbf{D} = \varepsilon\mathbf{E}$, we see that $\mathbf{E}_t = 0$. In this case, Maxwell's equations (2.1)–(2.2) reduce to

$$\mathbf{H}_t + \nabla \times [r(u)\nabla \times \mathbf{H}] = 0,$$

where $r(u) = \frac{1}{\sigma(u)}$ is the resistivity of the material. This system was studied in [19]. A global weak solution was obtained in [19], and regularity of the weak solution was studied in [20]. In this paper we shall show that the certain growth condition $r(u)$ is necessary in order to avoid the thermal runaway phenomenon.

3. Global existence of weak solutions. For the reader's convenience, we recall some standard notation.

Let B be a Banach space and $1 \le p < \infty$, and let

$$L^p(0,T;B) = \left\{ f | f : [0,T] \to B \text{ with the norm } ||f||_{L^p(0,T;B)} < \infty \right\},$$

where

$$||f||_{L^p(0,T;B)} = \left[\int_0^T ||f||_B^p dt \right]^{\frac{1}{p}}.$$

The spaces $W^{m,p}(\Omega), H_0^1(\Omega), W_p^{2,1}(Q_T)$, and $C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T), C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$, etc. are the usual Sobolev and classical spaces.

Let

$$C^\infty_{curl}(\bar{\Omega}, R^3) = \left\{ \mathbf{F}(x) : \Omega \to R^3 \text{ is smooth, } \nu \times \mathbf{F} = 0 \text{ on } \partial\Omega \right\},$$

$$H_*(\Omega) = \left\{ \mathbf{F}(x) \in H^1(\Omega; R^3) : \nabla \times \mathbf{F} \in L^2(\Omega; R^3), \nu \times \mathbf{F} = 0 \text{ a.e. on } \partial\Omega \right\}.$$

DEFINITION 3.1. *A bounded domain $\Omega \subset R^3$ is said to be cylindrical if $\Omega = (l_1, l_2) \times \Omega_1$ for some l_1, l_2 and a bounded domain Ω_1 in R^2.*

PROPOSITION 3.2 (see [13]). *Let Ω be a cylinder-type domain in R^3; then $H_*(\Omega)$ is a separable Hilbert space. Moreover, $C^\infty_{curl}(\bar{\Omega}; R^3)$ is dense in $H_*(\Omega)$. Hence, the trace operator*

$$Tr : \mathbf{F}(x) \in H_*(\Omega) \to H^{1/2}(\partial\Omega)$$

is well defined.

From now on, we shall always assume that the domain Ω is a cylinder-type one in R^3.

Consider the following linear Maxwell's equations:

$$\mathbf{E}_t + \sigma(x,t)\mathbf{E} = \nabla \times \mathbf{H}, \qquad (x,t) \in Q_T, \tag{3.1}$$

$$\mathbf{H}_t + \nabla \times \mathbf{E} = 0, \qquad (x,t) \in Q_T, \tag{3.2}$$

$$\nu \times \mathbf{E} = 0, \qquad (x,t) \in \partial\Omega \times (0,T], \tag{3.3}$$

$$\mathbf{E}(x,0) = \mathbf{E}_0(x), \qquad \mathbf{H}(x,0) = \mathbf{H}_0(x), \qquad x \in \Omega. \tag{3.4}$$

H(3.1). Assume that $\mathbf{E}_0(x)$ and $\mathbf{H}_0(x)$ are in $L^2(\Omega; R^3)$.

LEMMA 3.3. *Let the assumption* (3.1) *hold and let the function* $\sigma(x,t)$ *be measurable, nonnegative, and bounded. Then the problem* (3.1)–(3.4) *has a unique weak solution* $(\mathbf{E}, \mathbf{H}) \in L^\infty(0,T; L^2(\Omega; R^3))^2$ *in the following sense:*

$$\int_0^T \int_\Omega \left[-\mathbf{E}\cdot\mathbf{\Phi}_t + \sigma\mathbf{E}\cdot\mathbf{\Phi} - \mathbf{H}\cdot(\nabla\times\mathbf{\Phi})\right] dxdt = \int_\Omega E_0(x)\cdot\mathbf{\Phi}(x,0)dx; \tag{3.5}$$

$$\int_0^T \int_\Omega \left[-\mathbf{H}\cdot\mathbf{\Psi}_t + \mathbf{H}\cdot(\nabla\times\mathbf{\Psi})\right] dxdt = \int_\Omega \mathbf{H}_0(x)\cdot\mathbf{\Psi}(x,0)dx \tag{3.6}$$

for any vector test functions $\mathbf{\Phi}, \mathbf{\Psi} \in H^1(0,T; H^1_*(\Omega; R^3))$.

Moreover, the following energy estimate holds:

$$\int_\Omega \left[|\mathbf{E}|^2 + |\mathbf{H}|^2\right] dx + \int_0^t \int_\Omega \sigma|\mathbf{E}|^2 dxdt = \int_\Omega \left[|\mathbf{E}_0|^2 + |\mathbf{H}_0|^2\right] dx. \tag{3.7}$$

Proof. Since the domain Ω is a cylinder-type one in R^3, the Hilbert space $H_*(\Omega)$ is separable. We can use the finite element method to prove the existence when $\sigma(x,t)$ is only a function of the space variables and belongs to $W^{1,\infty}(\Omega)$; this was done in [13, Theorem 3.1]. For the current case, one can follow the same argument to establish the existence. However, we must clarify how a weak solution takes the initial and boundary value since $(\mathbf{E}, \mathbf{H})$ only belongs to $L^\infty(0,T; L^2(\Omega; R^3))^2$. Introduce a new function

$$\mathbf{W}(x,t) = \int_0^t \mathbf{E}(x,\tau)d\tau.$$

Then the system (3.1)–(3.2) becomes

$$\mathbf{W}_{tt} + \sigma(x,t)\mathbf{W}_t = \nabla\times[\mathbf{H}_0 - \nabla\times\mathbf{W}], \qquad (x,t) \in Q_T, \tag{3.8}$$

$$\nu\times\mathbf{W}_\mathbf{t} = 0, \qquad (x,t) \in \partial\Omega\times(0,T], \tag{3.9}$$

$$\mathbf{W}(x,0) = 0,\ \mathbf{W}_t(x,0) = \mathbf{E}_0(x), \qquad x \in \Omega. \tag{3.10}$$

Now we can easily show that $\mathbf{W} \in H^1(0,T; L^2(\Omega; R^3)) \bigcap L^2(0; T; H_*(\Omega))$ by using the energy method. Moreover, by Lemma 4.1 in Chapter 3 of [12] $\mathbf{W}_t$ is strongly continuous in t in the norm of $L^2(\Omega; R^3)$. It follows that the initial condition (3.10) for $\mathbf{W}$ makes sense. On the other hand, by Proposition 2.1 of [12], the trace of $\mathbf{W}$ on $\partial\Omega$ makes sense. Consequently we can uniquely define the value of $\mathbf{W}_t$ on the boundary $\partial\Omega$ as follows:

$$\nu\times\mathbf{W}_t = 0 \text{ if and only if } \quad \nu\times\mathbf{W} = 0 \qquad \text{on } \partial\Omega\times(0,T].$$

To derive the energy estimate, we note that, for any smooth vector fields $\mathbf{E}(x,t)$ and $\mathbf{H}(x,t)$ with $\nu\times\mathbf{E} = 0$ on $\partial\Omega\times(0,T]$,

$$\int_\Omega (\nabla\times\mathbf{H})\cdot\mathbf{E}dx = \int_\Omega \mathbf{H}\cdot(\nabla\times\mathbf{E})dx.$$

By taking the inner production to (2.1) with $\mathbf{E}$ and to (2.2) with $\mathbf{H}$, respectively, we obtain

$$\frac{d}{dt}\int_\Omega \frac{1}{2}\left[|\mathbf{E}|^2 + |\mathbf{H}|^2\right] dx + \int_\Omega \sigma|\mathbf{E}|^2 dx = 0.$$

Integration over $[0,t]$ yields the desired identity (3.7). The uniqueness of the solution is obvious from the identity (3.7).

The following lemma is elementary, but important, in the proof of the main result.

LEMMA 3.4. *Assume that a uniformly bounded sequence* $\{\sigma_n(x,t)\}$ *converges to* $\sigma(x,t)$ *in* $L^2(Q_T)$. *Let* $(\mathbf{E}_n,\mathbf{H}_n)$ *and* $(\mathbf{E},\mathbf{H})$ *be the corresponding solutions to the system* (3.1)–(3.4). *Then* $\mathbf{E}_n$ *and* $\mathbf{H}_n$ *converge to* $\mathbf{E}$ *and* $\mathbf{H}$, *respectively, in* $L^2(Q_T)$ *as* $n\to\infty$.

Proof. To prove the convergence, we define

$$\hat{\mathbf{E}}_n=\mathbf{E}_n-\mathbf{E},\qquad \hat{\mathbf{H}}_n=\mathbf{H}_n-\mathbf{H}.$$

Then, we have

$$\int_0^T\int_\Omega\Big[-\hat{\mathbf{E}}_n\cdot\mathbf{\Phi}_t-\hat{\mathbf{H}}_n\cdot(\nabla\times\mathbf{\Phi})\Big]\,dxdt=\int\int_{Q_T}[\sigma_n\mathbf{E}_n-\sigma\mathbf{E}]\cdot\mathbf{\Phi}dxdt,$$
$$\int_0^T\int_\Omega\Big[-\hat{\mathbf{H}}_n\cdot\mathbf{\Psi}_t+\hat{\mathbf{E}}_n\cdot(\nabla\times\mathbf{\Psi})\Big]\,dxdt=0.$$

By using Steklov averaging and standard approximation if necessary, we may choose $\mathbf{\Phi}=\hat{\mathbf{E}}_n(x,t)$, $\mathbf{\Psi}=\hat{\mathbf{H}}_n(x,t)$ as test functions to obtain

$$\begin{aligned}&\sup_{0\le t\le T}\int_\Omega[|\hat{\mathbf{E}}_n|^2+|\hat{\mathbf{H}}_n|^2]dx\\&\le\int\int_{Q_T}\Big[\sigma_n|\hat{\mathbf{E}}_n|^2+|\sigma_n-\sigma|\cdot|\mathbf{E}|\cdot|\hat{\mathbf{E}}_n|\Big]\,dxdt\\&\le C\int\int_{Q_T}|\hat{\mathbf{E}}_n|^2dxdt+\int\int_{Q_T}|\sigma_n-\sigma|^2\cdot|\mathbf{E}|^2dxdt\end{aligned}$$

since $0\le\sigma(s)\le\sigma_0$.

Gronwall's inequality implies

$$\sup_{0\le t\le T}\int_\Omega[|\hat{\mathbf{E}}_n|^2+|\hat{\mathbf{H}}_n|^2]dx\le C\int\int_{Q_T}|\sigma_n-\sigma|^2\cdot|\mathbf{E}|^2dxdt.$$

Since σ_n converges to σ in $L^2(Q_T)$, there exists a subsequence, say, σ_{n_k} which converges to σ a.e. on Q_T. By Lebesgue's dominated convergence theorem, it follows that

$$\int\int_{Q_T}|\sigma_{n_k}-\sigma|^2\cdot|\mathbf{E}|^2dxdt\to 0$$

as $n_k\to\infty$. Consequently, we have

$$\sup_{0\le t\le T}\int_\Omega[|\hat{\mathbf{E}}_{n_k}|^2+|\hat{\mathbf{H}}_{n_k}|^2]dx\to 0$$

as $n_k\to\infty$.

On the other hand, since the solution to the problem (3.1)–(3.4) is unique, the whole sequence $(\mathbf{E}_n,\mathbf{H}_n)$ must converge to $(\mathbf{E},\mathbf{H})$ in $L^2(Q_T)$.

Without essential difference, we shall assume that $k(u)=1$ in (2.3) for simplicity.

DEFINITION 3.5. *We say a triple of functions* $\mathbf{E}(x,t),\mathbf{H}(x,t)$, *and* $u(x,t)$ *is a weak solution of the system* (2.1)–(2.5) *if*

$$\mathbf{E},\mathbf{H}\in L^\infty(0,T;L^2(Q_T))\ \textit{and}\ u(x,t)\in L^q(0,T;W_0^{1,q}(\Omega))\ \textit{for some}\ q\in(1,\tfrac{5}{4}),$$

which satisfy

$$\int_0^T \int_\Omega \left[-\mathbf{E}\cdot\mathbf{\Phi}_t + \sigma\mathbf{E}\cdot\mathbf{\Phi} - \mathbf{H}\cdot(\nabla\times\mathbf{\Phi})\right] dxdt = \int_\Omega \mathbf{E}_0(x)\cdot\mathbf{\Phi}(x,0)dx,$$
$$\int_0^T \int_\Omega \left[-\mathbf{H}\cdot\mathbf{\Psi}_t + \mathbf{H}\cdot(\nabla\times\mathbf{\Psi})\right] dxdt = \int_\Omega \mathbf{H}_0(x)\cdot\mathbf{\Psi}(x,0)dx,$$
$$\int_0^T \int_\Omega u[\phi_t + \Delta\phi]dxdt = \int_0^T \int_\Omega \sigma(u)|\mathbf{E}|^2\phi dxdt$$

for any vector functions $\mathbf{\Phi}, \mathbf{\Psi} \in H^1(0,T;H_*(\Omega,R^3))$ *and for any* $\phi(x,t) \in C_0^\infty(Q_T)$. *Moreover, the limit of* $u(x,t)$ *as* t *tends to 0 is* $u_0(x)$ *in the sense of Lebesgue measure.*

Now we study the full evolution system (2.1)–(2.5).

H(3.2). (a) Let $\sigma(s)$ be bounded and uniformly Lipschitz continuous in $[0,\infty)$.

(b) $u_0(x) \in L^2(\Omega)$ with $u_0(x) \geq 0$.

THEOREM 3.6. *Under the assumptions* **H**(3.1)–**H**(3.2), *the problem* (2.1)–(2.5) *has a global weak solution.*

Proof. We shall use Schauder's fixed point theorem to prove the desired result. Let T be an arbitrary fixed number. Let

$$K = \{u(x,t) \in L^{1+\varepsilon_0}(Q_T) : ||u||_{L^{1+\varepsilon_0}(Q_T)} \leq K_0\},$$

where $\varepsilon_0 \in (0, \frac{1}{4})$ is a fixed number and K_0 will be determined later.

For any $v(x,t) \in K$, we solve Maxwell's equations:

$$\mathbf{E}_t + \sigma(v)\mathbf{E} = \nabla\times\mathbf{H}, \qquad (x,t) \in Q_T, \tag{3.11}$$
$$\mathbf{H}_t + \nabla\times\mathbf{E} = 0, \qquad (x,t) \in Q_T, \tag{3.12}$$
$$\nu\times\mathbf{E} = 0, \qquad (x,t) \in \partial\Omega\times(0,T], \tag{3.13}$$
$$\mathbf{E}(x,0) = \mathbf{E}_0(x), \mathbf{H}(x,0) = \mathbf{H}_0(x), \qquad x \in \Omega. \tag{3.14}$$

By Lemma 3.3, the system (3.11)–(3.14) has a unique weak solution $(\mathbf{E},\mathbf{H}) \in L^\infty(0,T;L^2(\Omega;R^3))^2$ with

$$\sup_{0\leq t\leq T}\int_\Omega \left[|\mathbf{E}|^2 + |\mathbf{H}|^2\right] dx + \int_0^T\int_\Omega \sigma(v)|\mathbf{E}|^2 dxdt \leq C_0,$$

where C_0 depends only on known data and the bound of $\sigma(s)$, but not on K_0.

Now we define a mapping $\mathcal{M} : v \in K \to u(x,t) = M[v] \in L^{1+\varepsilon_0}(Q_T)$ as follows. For any $v(x,t) \in K$, we define $\mathcal{M}[v] = u(x,t)$ to be the weak solution of the parabolic problem

$$u_t - \Delta u = \sigma(v)|\mathbf{E}|^2, \qquad (x,t) \in Q_T, \tag{3.15}$$
$$u(x,t) = 0, \qquad (x,t) \in \partial\Omega\times(0,T], \tag{3.16}$$
$$u(x,0) = u_0(x), \qquad x \in \Omega, \tag{3.17}$$

where $\mathbf{E}(x,t)$ is the solution of the system (3.11)–(3.14).

Since $\sigma(v)|\mathbf{E}|^2 \in L^1(Q_T)$, the result of Theorem 4 in [2] implies that the parabolic problem (3.15)–(3.17) has a weak solution:

$$u(x,t) \in L^\infty(0,T;L^1(\Omega))\bigcap L^q(0,T;W_0^{1,q}(\Omega))$$

for any $q \in (0, \frac{5}{4})$ (recall the space dimension of Ω is 3). Moreover, there exists a constant C_1 such that for any $q \in (1, \frac{5}{4})$

$$\sup_{0\le t\le T}\int_\Omega |u|dxdt + \int\int_{Q_T} |\nabla u|^q dxdt \le C_1,$$

where C_1 depends only on known data and the bound of $\sigma(s)$, but not on K_0.

Hence, the mapping $\mathcal{M}$ is well defined if we choose $q \in (1+\varepsilon_0, \frac{5}{4})$ and from K to itself if we choose $K_0 = C_1$.

Next we show that $\mathcal{M}$ is continuous. Let $\{v_n(x,t)\} \subset K \subset L^{1+\varepsilon_0}(Q_T)$ be a sequence which converges to $v(x,t)$ in $L^{1+\varepsilon_0}(Q_T)$. Let $(\mathbf{E}_n, \mathbf{H}_n)$ and $(\mathbf{E}, \mathbf{H})$ be the weak solutions of the system (3.11)–(3.14) corresponding to $v_n(x,t)$ and $v(x,t)$, respectively. It is clear by using the same argument that the following energy estimates hold:

$$\sup_{0\le t\le T}\int_\Omega \left[|\mathbf{E}_n|^2 + |\mathbf{H}_n|^2\right] dx + \int_0^T\int_\Omega \sigma(v_n)|\mathbf{E}_n|^2 dxdt \le C_0,$$

$$\sup_{0\le t\le T}\int_\Omega \left[|\mathbf{E}|^2 + |\mathbf{H}|^2\right] dx + \int_0^T\int_\Omega \sigma(v)|\mathbf{E}|^2 dxdt \le C_0.$$

We claim that $\mathbf{E}_n \to \mathbf{E}$ and $\mathbf{H}_n(x,t) \to \mathbf{H}$ in $L^2(Q_T)$ as $n \to \infty$. Indeed, since $v_n \to v$ in $L^{1+\varepsilon_0}(Q_T)$ and $\sigma(s)$ is bounded and uniformly Lipschitz continuous, then

$$\int_{Q_T} |\sigma(v_n) - \sigma(v)|^2 dx \le C\int_{Q_T} |\sigma(v_n) - \sigma(v)|^{1+\varepsilon_0} dxdt \le C\int_{Q_T} |v_n - v|^{1+\varepsilon_0} dxdt \to 0,$$

as $n \to \infty$. The claim follows from Lemma 3.3.

Since the L^1-bound of $\sigma(u_n)|\mathbf{E}|^2$ is independent of n, we have by [2] that, for any $q \in (1, \frac{5}{4})$,

$$||u_n||_{L^q(0,T;W_0^{1,q}(\Omega))} \le C_1,$$

where C_1 depends only on C_0 and known data, but not on n.

From a compactness result of [15], we see that there exists a subsequence of u_n (still denoted by $u_n(x,t)$) such that, for any $q \in (1+\varepsilon_0, \frac{5}{4})$,

$$\begin{aligned} &u_n \to u(x,t), && \text{weakly in } L^q(0,T;W_0^{1,q}(\Omega)) \ ; \\ &u_n \to u(x,t), && \text{strongly in } L^q(Q_T). \end{aligned}$$

Now we show that $u(x,t)$ is the only limit point of the sequence $\{u_n(x,t)\}$ in $L^q(Q_T)$. Assume that $u^*(x,t)$ is a limit point of $\{u_n(x,t)\}$ in $L^q(Q_T)$. Then we can extract a subsequence, denoted by $v_{n_k}(x,t)$, of $\{v_n(x,t)\}$ such that $v_{n_k}(x,t) \to v(x,t)$ a.e. in Q_T.

Now, by Green's representation,

$$u_{n_k}(x,t) = \int_\Omega G(x,y,t,0)u_0(y)dy + \int_0^t\int_\Omega G(x,y;t,\tau)f_{n_k}(y,\tau)dyd\tau,$$

where

$$f_{n_k}(x,t) = \sigma(v_{n_k})|\mathbf{E}_{n_k}|^2$$

and $G(x,y;t,\tau)$ is Green's function of the heat operator with homogeneous Dirichlet boundary conditions. It follows (see [19]) that

$$\sup_{0\le t\le T}\int_\Omega |u_{n_k} - u|dx + \int_0^T\int_\Omega |\nabla u_{n_k} - \nabla u|dxdt \le C||f_{n_k} - f||_{L^1(Q_T)}. \tag{3.18}$$

Since $\mathbf{E}_{n_k}$ converges to $\mathbf{E}$ in $L^2(Q_T)$, it follows that $|\mathbf{E}_{n_k}|^2$ converges to $|\mathbf{E}|^2$ in $L^1(Q_T)$. On the other hand, since $\sigma(s)$ is bounded and v_{n_k} converges to v a.e. in Q_T, it follows that $f_{n_k}(x,t)$ converges to $f(x,t)$ in $L^1(Q_T)$.

Consequently, taking the limit in (3.18), we obtain

$$\sup_{0\le t\le T}\int_\Omega |u^*-u|dx+\int_0^T\int_\Omega |\nabla u^*-\nabla u|dxdt\le 0,$$

i.e., $u^*(x,t)=u(x,t)$ a.e. in Q_T. Therefore, the mapping $\mathcal{M}$ is continuous.

In order to apply Schauder's fixed point theorem, we need to show that the mapping $\mathcal{M}$ is compact. It is clear that $W^{1,1}(\Omega)\subset BV(\Omega)$, where $BV(\Omega)$ denotes the space of functions with bounded total variations in Ω. On the other hand, the embedding from $BV(\Omega)$ to $L^q(\Omega)$ with $q\in(1,\frac{3}{2})$ (recall the space dimension of Ω is 3) is compact. Moreover, since $\nabla u\in L^q(Q_T)$ for any $q\in(1,5/4)$ and $\sigma(v)|\mathbf{E}|^2\in L^1(Q_T)$, it follows that from the equation (3.15),

$$u_t\in L^1(Q_T)+L^1(0,T;W^{-1,q'}(\Omega)),$$

where $q'=\frac{q}{q-1}$.

It follows by [15] that the mapping $\mathcal{M}$ is also compact. Finally, by Schauder's fixed point theorem, we know that the mapping $\mathcal{M}$ has a fixed point, denoted by $u(x,t)$. This fixed point $u(x,t)$ along with the solution $(\mathbf{E},\mathbf{H})$ of (3.11)–(3.14) consists of a weak solution of (2.1)–(2.5). □

The temperature $u(x,t)$ satisfied (2.3) only in the sense of distribution. However, we can improve the regularity of $u(x,t)$ if some additional condition is imposed on $\sigma(u)$.

COROLLARY 3.7. *In addition to the assumptions* **H**(3.1)–**H**(3.2), *assume that there exists a constant* σ_1 *such that*

$$s^p\sigma(s)\le\sigma_1\qquad\text{for } s\ge 0 \text{ and some } p>0;$$

then $v(x,t)=u(x,t)^{\frac{p}{2}}\in L^\infty(0,T;L^2(\Omega))\bigcap L^2(0,T;H_0^1(\Omega))$.

Proof. By choosing $\psi=u^p$ as a test function,

$$\int\int_{Q_T}u^p\sigma(u)|\mathbf{E}|^2dxdt\le C_0\int\int_{Q_T}|\mathbf{E}|^2dxdt\le C,$$

where C is a constant depending only upon the known constants.

Set $v(x,t)=u^{p/2}(x,t)$. Then from (2.3) we easily derive

$$\int_\Omega v^2dx+\int\int_{Q_T}|\nabla v|^2dxdt\le C,$$

where C depends only on the known data and p. We can choose the set K as a subset of the space such that

$$u(x,t)^{\frac{p}{2}}\in L^\infty(0,T;L^2(\Omega))\bigcap L^2(0,T;H_0^1(\Omega)).$$

By the same argument as in the proof of Theorem 3.6, we can obtain the desired result. □

4. One-dimension problem. In this section we consider a special case where the electrical field is parallel to the y-axis; that is, $\mathbf{E}(x,t) = \{0, g(x,t), 0\}$. Then $\mathbf{H}$ must have the form $\mathbf{H} = \{0, 0, h(x,t)\}$. For this special case, the system (2.1)–(2.5) becomes

$$g_t + \sigma(u)g = -h_x, \qquad 0 < x < 1, \quad t > 0, \tag{4.1}$$

$$h_t + g_x = 0, \qquad 0 < x < 1, \quad t > 0, \tag{4.2}$$

$$u_t - (k(u)u_x)_x = \sigma(u)g^2, \qquad 0 < x < 1, \quad t > 0, \tag{4.3}$$

$$g(i,t) = u(i,t) = 0, \qquad t > 0, \quad i = 0, 1, \tag{4.4}$$

$$g(x,0) = g_0(x), h(x,0) = h_0(x), u(x,0) = u_0(x) \geq 0, \qquad 0 \leq x \leq 1. \tag{4.5}$$

H(4.1). (a) Let $\sigma(s)$ and $k(s)$ be C^2 functions and

$$0 \leq \sigma(s) \leq \sigma_0, \quad 0 < k_0 \leq k(s) \leq k_1 < \infty.$$

(b) Let $g_0(x)$ and $h_0(x)$ be $C^3[0,1]$. Moreover, $g_0(x)$ and $h_{0x}(x)$ can be extended as odd functions about $x = 0$ and $x = 1$, and the extended functions are in $C^2(R)$. Furthermore, the following consistency conditions hold:

$$\begin{aligned}
&\sigma(0)g_0(0) = h_0^{'}(0),\ \sigma(0)g_0(1) = h_0^{'}(1),\ g_0^{''}(0) = g_0^{''}(1) = 0,\\
&u_0(0) = u_0(1) = 0,\\
&-[k(0)u_0^{''}(0) + k'(0)u_0^{'}(0)^2] = \sigma(0)g_0(0)^2,\\
&-[k(0)u_0^{''}(1) + k'(0)u_0^{'}(1)^2] = \sigma(0)g_0(1)^2.
\end{aligned}$$

THEOREM 4.1. *Under the assumption* **H**(4.1), *the problem* (4.1)–(4.5) *has a unique global solution in the classical sense:* $g, h \in C^{1+1,1+1}(\mathbf{Q}_T)$ *and* $u(x,t) \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$.

Proof. Introduce

$$w(x,t) = \int_0^t g(x,\tau)d\tau, \qquad (x,t) \in Q_T.$$

Then it is easy to see that $w(x,t)$ satisfies

$$w_{tt} - w_{xx} + \sigma(u)w_t = h_{0x}, \qquad (x,t) \in Q_T, \tag{4.6}$$

$$u_t - (k(u)u_x)_x = \sigma(u)w_t^2, \qquad 0 < x < 1, t > 0, \tag{4.7}$$

$$w(i,t) = u(i,t) = 0, \qquad 0 \leq t \leq T, i = 0, 1, \tag{4.8}$$

$$w(x,0) = 0, w_t(x,0) = g_0(x), u(x,0) = u_0(x), \qquad 0 \leq x \leq 1. \tag{4.9}$$

By the assumption **H**(4.1), we see that a classical solution $g(x,t), h(x,t)$, and $u(x,t)$ exists in Q_{T_0} for a small $T_0 > 0$, which can be done by a standard argument (see [17] for example). To extend the local solution to an arbitrary interval $[0,T]$, we only need to derive an a priori estimate in the classical space.

Since we are deriving a priori estimates, we can always assume that h, g, and u are smooth. We will use C to denote a generic constant which depends only on known data.

By the assumption, we extend the function $g_0(x)$ as an odd function about $x = 0$ and $x = 1$ and the extended function $g_0(x) \in C^2(R)$. By considering $F(x,t) = h_{0x}(x) - \sigma(u)w_t$ as an inhomogeneous term in (4.6), the solution $w(x,t)$ will be an odd function as long as the inhomogeneous term $h_{0x}(x) - \sigma(u)w_t$ is an odd function.

Note that $\sigma(u(x,t)) \geq 0$ is an even function if $u(x,t)$ is an even function about $x=0$ and $x=1$. If we extend $u_0(x)$ as an even function about $x=0$ or $x=1$, then $u(x,t)$ will be an even function since $\sigma(u)w_t^2 \geq 0$. Now it is clear that $w(x,t)$ and $w_t(x,t)$ have the same sign when the space variable x is replaced by $-x$ or $1-x$; it follows that the solution $w(x,t)$ can be represented by the following formula (see [18]):

$$\begin{aligned} w(x,t) &= \frac{1}{2}\int_{x-t}^{x+t} g_0(\xi)d\xi + \frac{1}{2}\int_0^t \int_{x-(t-\tau)}^{x+(t-\tau)} [h_{0y}(y) - \sigma(u(y,\tau))w_\tau(y,\tau)]dyd\tau \\ &= \frac{1}{2}\int_{x-t}^{x+t} g_0(\xi)d\xi + \frac{1}{2}\int_0^t [h_0(x+t-\tau) - h_0(x-t+\tau)]d\tau \\ &\quad -\frac{1}{2}\int_0^t \int_{x-(t-\tau)}^{x+t-\tau} \sigma(u(y,\tau))w_\tau(y,\tau)dyd\tau. \end{aligned} \tag{4.10}$$

Then

$$\begin{aligned} w_t(x,t) &= \frac{1}{2}[g_0(x+t) - g_0(x-t)] + \frac{1}{2}\int_0^t h_0^{'}(x+t-\tau) + h_0^{'}(x-t+\tau)]d\tau \\ &\quad -\frac{1}{2}\int_0^t \sigma(u(y,\tau))w_\tau(y,\tau)\,|_{y=x-t+\tau}^{y=x+t-\tau}\, d\tau. \end{aligned} \tag{4.11}$$

Since $\sigma(s)$ is bounded by σ_0, we have

$$||w_t(\cdot,t)||_{L^\infty(0,1)} \leq ||g_0||_{L^\infty(0,1)} + T||h_0^{'}||_{L^\infty(0,1)} + \sigma_0 \int_0^t ||w_\tau||_{L^\infty(0,1)}d\tau.$$

Gronwall's inequality yields

$$||w_t||_{L^\infty(0,1)} \leq C,$$

where C depends only on σ_0, T, and $||g_0||_{L^\infty(0,1)} + ||h_0||_{L^\infty(0,1)}$.

From the classical theory of parabolic equations, we know from (4.7) that for any $p > 1$,

$$||u||_{W_p^{2,1}(Q_T)} \leq C||\sigma(u)w_t^2||_{L^p(Q_T)} \leq C,$$

where C depends only on p, σ_0, and known data.

Sobolev's embedding implies that for any $\alpha \in (0,1)$,

$$||u||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} \leq C.$$

Note that

$$\begin{aligned} &\frac{d}{dx}\int_0^t \sigma(u(x+t-\tau,\tau)w_\tau(x+t-\tau,\tau))d\tau \\ &= \int_0^t [\sigma'(u(y,\tau))u_x(y,\tau)w_\tau(y,\tau) + \sigma(u(y,\tau))w_{\tau y}(y,\tau)]_{y=x+t-\tau}\, d\tau. \end{aligned}$$

We can differentiate (4.11) with respect to x and use the boundedness of u_x and w_t to obtain

$$||w_{tx}||_{L^\infty(0,1)} \leq C + C\int_0^t ||w_{\tau x}||_{L^\infty(0,1)}d\tau,$$

which gives the $L^\infty(0,1)$-bound of w_{tx} by Gronwall's inequality. Moreover, the $L^\infty(0,1)$-bound of w_{tx} depends only on known data.

Next, we differentiate (4.11) with respect to t again, after some routine calculations, to obtain

$$||w_{tt}||_{L^\infty(0,1)} \le C + C\int_0^t ||w_{tx}||_{L^\infty(0,1)} d\tau \le C.$$

Set $v(x,t) = u_t(x,t)$. Then it is easy to see that $v(x,t)$ solves the following initial-boundary value problem:

$$\begin{aligned} &v_t - k(u)v_{xx} - k''(u)u_x^2 v - 2k'(u)v_x u_x - k'(u)u_{xx}v \\ &= \sigma'(u)vw_t^2 + 2\sigma(u)w_t w_{tt}, \qquad (x,t) \in Q_T, \end{aligned} \tag{4.12}$$

$$v(0,t) = v(1,t) = 0, \qquad 0 \le t \le T, \tag{4.13}$$

$$v(x,0) = v_0(x), \qquad 0 \le x \le 1, \tag{4.14}$$

where

$$v_0(x) = k(u_0)u_0''(x) + k'(u_0)u_0'(x)^2 + \sigma(u_0)g_0(x)^2.$$

By the $W_p^{2,1}(Q_T)$-estimate for parabolic equations, we see

$$||v||_{W_p^{2,1}(Q_T)} \le C + ||v||_{L^{2p}(Q_T)} + ||u_{xx}||_{L^{2p}(Q_T)} + C||w_{tt}||_{L^p(Q_T)} \le C,$$

where C depends only on known data.

Again, Sobolev's embedding theorem implies that for any $\alpha \in (0,1)$

$$||u_t||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} \le C.$$

Similarly, by differentiating the equation (4.7) with respect to x we can easily derive that

$$||u_x||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} \le C.$$

Now we take the derivative with respect to x twice in (4.11) and then use the boundedness of u_{xx} and w_t to deduce

$$||w_{txx}||_{L^\infty(0,1)} \le C + C\int_0^t ||w_{\tau xx}||_{L^\infty(0,1)} d\tau,$$

which yields an a priori bound of $||w_{txx}||_{L^\infty(0,1)}$. Finally, from the definition of w and the system (4.1)–(4.2) we see

$$g(x,t), h(x,t) \in C^{1+1,1+1}(\bar{Q}_T), \qquad u(x,t) \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T),$$

and

$$||g||_{C^{1+1,1+1}(\bar{Q}_T)} + ||h||_{C^{1+1,1+1}(\bar{Q}_T)} + ||u||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} \le C,$$

where $\alpha \in (0,1)$ is arbitrary and C depends only on known data.

With the above a priori estimate, we can use the method of continuity to extend a local solution to an arbitrary interval $[0,T]$. □

5. Maxwell's equations in quasi-stationary fields. In this section we study an electromagnetic field with a negligible electric displacement $\mathbf{D}$. When $\mathbf{D}_t = 0$, then $\mathbf{E}_t = 0$. It follows that the system (2.1)–(2.3) becomes

$$\mathbf{H}_t + \nabla \times [r(u)\nabla \times \mathbf{H}] = 0, \qquad (x,t) \in Q_T, \tag{5.1}$$

$$u_t - \nabla[k(u)\nabla u] = r(u)|\nabla \times \mathbf{H}|^2, \qquad (x,t) \in Q_T, \tag{5.2}$$

where $r(u) = \frac{1}{\sigma(u)}$.

We shall prescribe the following initial-boundary conditions for $\mathbf{H}$ and $u(x,t)$:

$$\mathbf{H}(x,t) = \mathbf{F}(x,t),\ u_\nu(x,t) = 0, \qquad (x,t) \in S_T = \partial\Omega \times (0,T], \tag{5.3}$$

$$\mathbf{H}(x,0) = \mathbf{H}_0(x), u(x,0) = u_0(x), \qquad x \in \Omega, \tag{5.4}$$

where u_ν denotes the outward normal derivative to $\partial\Omega$.

$\mathbf{H}(5.1)$. (a) Let $r(u)$ and $k(u)$ be $C^{1+\alpha}(R)$ functions. There exists a constant $a_0 > 0$ such that

$$r(u) \geq a_0, k(u) \geq a_0.$$

(b) Let $u_0(x) \geq 0, u_0(x) \in C^{2+\alpha}(\bar{\Omega})$, and $\mathbf{H}_0 \in C^{2+\alpha}(\bar{\Omega})^3, \mathbf{F}(x,t) \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)^3$. The following consistency conditions hold:

$$\begin{aligned}
&\nabla \cdot \mathbf{H}_0(x) = 0, && x \in \Omega,\\
&u_{0\nu}(x) = 0, \mathbf{F}(x,0) = \mathbf{H}_0(x), && \text{on } \partial\Omega,\\
&\mathbf{F}_t(x,0) + \nabla \times [r(u_0(x))\nabla \times \mathbf{H}_0] = 0, && \text{on } \partial\Omega.
\end{aligned}$$

LEMMA 5.1. *Under the assumption* $\mathbf{H}(5.1)$ *the evolution system* (5.1)–(5.4) *has a classical solution* $\mathbf{H}(x,t) \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)^3, u(x,t) \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$ *for some* $T > 0$.

Proof. The proof is quite standard. We shall only outline the main steps. For simplicity, we take $k(u) = 1$. It will be seen that the general case can be handled similarly. We also assume that $T \leq 1$ since we are only concerned with the local existence.

Let

$$K = \{u(x,t) \in C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T) : u(x,t) \geq 0, ||u||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} \leq K_0\},$$

where K_0 is a constant to be specified later. It is clear that K is a bounded convex subset of $C^{2,1}(\bar{Q}_T)$. Given $u^*(x,t) \in K$, we consider the following evolution system:

$$\mathbf{H}_t + \nabla \times [r(u^*)\nabla \times \mathbf{H}] = 0, \qquad (x,t) \in Q_T, \tag{5.5}$$

$$\mathbf{H}(x,t) = \mathbf{F}(x,t), \qquad (x,t) \in \partial\Omega \times (0,T]; \tag{5.6}$$

$$\mathbf{H}(x,0) = \mathbf{H}_0(x), \qquad x \in \Omega. \tag{5.7}$$

Observe that

$$\nabla \times [r(u^*)\nabla \times \mathbf{H}] = -r(u^*)\Delta\mathbf{H} + r'(u^*)\nabla u^* \times (\nabla \times \mathbf{H}).$$

By the results of parabolic systems (see [12]), there exists a unique classical solution $\mathbf{H}(x,t)$ for some $T > 0$. Moreover, there exists a constant $C(K_0)$ such that

$$||\mathbf{H}||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} \leq C(K_0),$$

where $C(K_0)$ depends only on K_0 and the known data but is independent of the lower bound of T.

Now we define a mapping as follows:

$$M : u^*(x,t) \in K \to u(x,t) = M[u^*],$$

where $u(x,t)$ is the solution to the following problem:

$$u_t - \Delta u = r(u^*)|\nabla \times \mathbf{H}|^2, \qquad (x,t) \in Q_T, \tag{5.8}$$

$$u_\nu = 0, \qquad (x,t) \in \partial\Omega \times (0,T], \tag{5.9}$$

$$u(x,0) = u_0(x), \qquad x \in \Omega, \tag{5.10}$$

where $\mathbf{H}$ is the solution to the system (5.5)–(5.7).

The classical theory of parabolic equations ensures that the above parabolic problem has a unique classical solution on $[0,T]$. Moreover,

$$||u||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} \le C(K_0),$$

where $C(K_0)$ depends on known data and K_0, but not on the lower bound of T. We know that the mapping M is well defined. The continuity of the mapping M is quite standard since $u^* \in C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)$. The compactness of the mapping M is clear since $u(x,t) = M[u^*] \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$ and the embedding operator from $C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$ to $C^{2,1}(\bar{Q}_T)$ is compact.

Now we show that the mapping M is from K into K. Indeed,

$$\begin{aligned} ||u||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} &\le \sum_{|k|\le 2} \sup_{Q_T} |D^k u| + \sup_{Q_T} |u_t| \\ &\le [T^\alpha + T^{\alpha/2}]C(K_0) + ||u_0||_{C^2(\bar{\Omega})} \\ &\le 1 + ||u_0||_{C^2(\bar{\Omega})}, \end{aligned}$$

provided that we restrict T such that $TC(K_0) \le 1$.

By Schauder's fixed point theorem, the mapping M has a fixed point. This fixed point $u(x,t)$ along $\mathbf{H}$ obtained from (5.5)–(5.8) consists of a solution to the system (5.1)–(5.4). The uniqueness is clear since the solution is classical.

H(5.2). (a) The function $r(s)$ satisfies

$$\int_{a_0}^{\infty} \frac{1}{r(s)} ds < \infty$$

for some constant $a_0 > 0$.

(b) The initial and boundary data $u_0(x)$ and $\mathbf{F} = (F_1, F_2, F_3)$ satisfy the following inequality:

$$u_0(x) \ge a_0, \int_\Omega \int_{u_0(x)}^{\infty} \frac{1}{r(s)} ds dx < \frac{1}{C_0} \sum_{i=1}^{3} \int_0^\infty \int_S |F_i - (F_i)_s|^2 ds dt,$$

where C_0 is the best constant from the trace inequality

$$\sum_{i=1}^{3} \int_S |F_i - (F_i)_s|^2 ds \le C_0 \sum_{i,j=1}^{3} \int_\Omega |H_{ix_j}|^2 dx,$$

$$\mathbf{H} = (H_1, H_2, H_3),\ (F_i)_s = \frac{1}{|S|} \int_S F_i ds, i = 1,2,3,$$

while ds represents the surface element and $|S|$ denotes the $(n-1)$-dimension Lebesgue measure.

Now we can state the following blowup result.

THEOREM 5.2. *Let the assumptions* $\mathbf{H}(5.1)$–$\mathbf{H}(5.2)$ *hold. Then* $u(x,t)$ *will blow up in finite time.*

Proof. The argument below follows the idea from [1]. It is clear from the assumption and the maximum principle that $u(x,t) \geq a_0$ whenever the solution exists. Suppose the problem (5.1)–(5.4) has a smooth solution for any $T > 0$. Define

$$A(t) = \int_\Omega \int_{u(x,t)}^{\infty} \frac{1}{r(s)} ds dx.$$

Then $A(t)$ is well defined for all $t < \infty$. By using (5.2) and performing the integration by parts, we see

$$\begin{aligned} A'(t) &= -\int_\Omega \frac{u_t}{r(u)} dx \\ &= -\int_\Omega \frac{\nabla[k(u)\nabla u]}{r(u)} dx - \int_\Omega |\nabla \times \mathbf{H}|^2 dx \\ &= -\int_\Omega \frac{r'(u)k(u)|\nabla u|^2}{r(u)^2} dx - \int_\Omega |\nabla \times \mathbf{H}|^2 dx \\ &\leq -\int_\Omega |\nabla \times \mathbf{H}|^2 dx. \end{aligned}$$

It follows that

$$0 \leq A(t) \leq A(0) - \int_0^t \int_\Omega |\nabla \times \mathbf{H}|^2 dx dt.$$

Since $\nabla \cdot \mathbf{H}_t(x,t) = 0$ from (5.1), it follows that

$$\nabla \cdot \mathbf{H}(x,t) = \nabla \cdot \mathbf{H}_0(x) = 0.$$

Performing integration by parts, we obtain

$$\begin{aligned} &\int_\Omega |\nabla \times \mathbf{H}|^2 dx \equiv \int_\Omega [|\nabla \cdot \mathbf{H}|^2 + |\nabla \times \mathbf{H}|^2] dx \\ &= \sum_{i,j=1}^{3} \int_\Omega |H_{ix_j}|^2 dx. \end{aligned}$$

It follows that

$$0 \leq A(t) \leq A(0) - \sum_{i,j=1}^{3} \int_0^t \int_\Omega |H_{ix_j}|^2 dx dt.$$

Now the trace inequality implies

$$\sum_{i=1}^{3} \int_S |F_i - (F_i)_s|^2 ds \leq C_0 \sum_{i,j=1}^{3} \int_\Omega |H_{ix_j}|^2 dx,$$

where C_0 is a constant depending only upon the space dimension n and the boundary $S = \partial\Omega$. In particular, C_0 does not depend on t. Consequently,

$$0 \leq A(t) \leq A(0) - \frac{1}{C_0} \int_0^t \int_S |F - (F)_s|^2 ds$$

will become negative for large t by the assumption $\mathbf{H}(5.2)$, a contradiction.

It follows that $u(x,t)$ will become unbounded in finite time. ∎

Acknowledgment. The author would like to thank Professor Bei Hu for many helpful discussions.

REFERENCES

[1] S. N. Antontsev and M. Chipot, *The thermistor problem: Existence, smoothness, uniqueness, blowup*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.
[2] L. Boccardo and T. Gallouet, *Nonlinear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 87 (1989), pp. 149–169.
[3] A. Bossavit, *Free boundaries in induction heating*, Control Cybernet., 14 (1985), pp. 69–96.
[4] C. J. Coleman, *On the microwave hot spot problem*, J. Austral. Math. Soc. Ser. B., 33 (1991), pp. 1–8.
[5] E. J. Davies, *Conduction and Induction Heating*, Per Peregrimus Ltd., London, 1990.
[6] R. DiPerna and P. L. Lions, *Global weak solutions of Vlasov–Maxwell system*, Comm. Pure Appl. Math., 42 (1989), pp. 729–757.
[7] R. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, PA, 1996.
[8] J. M. Hill and A. H. Pincombe, *Some similarity temperature profiles for the microwave heating of a half-space*, J. Austral. Math. Ser. B., 33 (1992), pp. 290–320.
[9] G. A. Kriegsmann, *Microwave heating of dispersive media*, SIAM J. Appl. Math., 53 (1993), pp. 655–669.
[10] G. A. Kriegsmann, *Hot Spot Formation in Microwave Heated Ceramic Fibers*, CAMS Research Report 96-17, New Jersey Institute of Technology, 1996.
[11] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Pergamon Press, New York, 1960.
[12] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, Amer. Math. Soc., Providence, RI, 1968.
[13] A. Lorenzi, *Direct and inverse integrodifferential Maxwell's problems for dispersive media related to cylindrical domains*, in Inverse Problems in Diffusion Processes, H. W. Engl and W. Rundell, eds., SIAM, Philadelphia, PA, 1995.
[14] A. C. Metaxas and R. J. Meredith, *Industrial Microwave Heating*, IEE Power Engineering Series, Vol. 4, Per Peregrimus Ltd., London, 1983.
[15] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pure Appl., 146 (1987), pp. 65–96.
[16] N. F. Smyth, *Microwave heating of bodies with temperature dependent properties*, Wave Motion, 12 (1990), pp. 171–186.
[17] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1982.
[18] H. F. Weiberger, *A first course in Partial Differential Equations*, Blaisdell Publishing Company, Waltham, MA, 1965.
[19] H. M. Yin, *Global solutions of Maxwell's equations in an electromagnetic field with the temperature-dependent electrical conductivity*, European J. Appl. Math., 5 (1994), pp. 57–64.
[20] H. M. Yin, *Regularity of Solutions to Maxwell's System in Quasi-Stationary Electromagnetic Fields and Applications*, Comm. Partial Differential Equations, 22 (1997), pp. 1029–1053.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 652–672, May 1998

007

ALL-TIME EXISTENCE OF CLASSICAL SOLUTIONS FOR SLIGHTLY COMPRESSIBLE FLOWS*

THOMAS HAGSTROM† AND JENS LORENZ†

Abstract. In two space dimensions and under periodic boundary conditions, the solution of the incompressible Navier–Stokes equations is known to remain smooth for all time. In this paper we consider the system of equations describing isentropic, compressible flow and show a similar result if the Mach number is sufficiently small and the initial data are almost incompressible. It is not assumed that the initial data are small. To the leading order, the solution consists of the corresponding incompressible flow plus a highly oscillatory part describing sound waves.

Key words. Navier–Stokes equations, compressible flows, all-time existence, multiple scales

AMS subject classifications. 76N99, 35M20

PII. S0036141097315312

1. Introduction. In this paper we consider the Navier–Stokes equations for isentropic, compressible flows of a polytropic gas:

$$\begin{aligned} &u_t + (u\cdot\nabla)u + (1+\varepsilon^2\rho)^{\gamma-2}\nabla\rho \\ &\qquad = (1+\varepsilon^2\rho)^{-1}\left(\nu\Delta u + \eta\nabla\nabla\cdot u\right), \quad \nu>0, \quad \eta+\nu>0, \end{aligned} \tag{1.1}$$

$$\varepsilon^2\left(\rho_t + (u\cdot\nabla)\rho\right) + (1+\varepsilon^2\rho)\nabla\cdot u = 0\ . \tag{1.2}$$

The scalings and assumptions leading to these equations are given in the Appendix, where it also is explained that ρ describes the scaled fluctuations of the density. If the parameter ε, which represents the Mach number, is formally set to zero, we obtain the equations for incompressible flow,

$$U_t + (U\cdot\nabla)U + \nabla P = \nu\Delta U, \quad \nu>0, \tag{1.3}$$

$$\nabla\cdot U = 0\ . \tag{1.4}$$

We restrict the discussion to two space variables

$$(x,y)\in(\mathbb{R}\mathrm{mod}2\pi)^2 =: T^2;$$

i.e., we assume that the functions

$$u(x,y,t),\ \rho(x,y,t),\ U(x,y,t),\ P(x,y,t)$$

are 2π-periodic in x and in y. (We comment below on the three-dimensional case.)

*Received by the editors January 22, 1997; accepted for publication June 4, 1997. The first author was supported in part by NSF grants DMS-9304406 and DMS-9600146, DOE grant DE-FG03-92ER25128 and ICOMP, NASA-Lewis Res. Ctr. The second author was supported in part by NSF grant DMS-9404124 and DOE grant DE-FG03-92ER25128.

http://www.siam.org/journals/sima/29-3/31531.html

†Department of Mathematics and Statistics, The University of New Mexico, Albuquerque, NM 87131 (hagstrom@math.unm.edu, lorenz@math.unm.edu).

For $\varepsilon > 0$, the equations (1.1), (1.2) form a coupled parabolic–hyperbolic system, for which we give initial conditions

$$u = u_0(x,y), \quad \rho = \rho_0(x,y) \qquad \text{at} \quad t = 0\ . \tag{1.5}$$

For simplicity,[1] we assume that $u_0, \rho_0 \in C^\infty(T^2)$ and that they have mean zero:

$$\int_{T^2} u_0 dxdy\ , \quad \int_{T^2} \rho_0 dxdy = 0. \tag{1.6}$$

(Note that these assumptions on the mean are without essential loss of generality, as they may be imposed by a Galilean transformation and by choosing the reference density to be the mean initial density.) For the incompressible equations (1.3), (1.4), one can only prescribe an initial velocity field $U_0(x,y)$, which must be divergence-free, to obtain a classical solution,

$$U = U_0(x,y) \qquad \text{at} \quad t = 0\ . \tag{1.7}$$

Here we assume $U_0 \in C^\infty$, $\int_{T^2} U_0 = 0$, $\nabla \cdot U_0 = 0$. To eliminate the free (time-dependent) constant in the incompressible pressure, we impose the side condition

$$\int_{T^2} P(x,y,t)dxdy = 0, \quad t \geq 0\ . \tag{1.8}$$

It is well known that the incompressible problem (1.3), (1.4), (1.7), (1.8) has a unique classical solution $(U,P) \in C^\infty(T^2 \times [0,\infty))$, and this solution—together with all its derivatives—tends to zero at an exponential rate as $t \to \infty$. (See, e.g., [5, Chap. 9] and extensions of the arguments therein.)

The aim of this paper is to show all-time existence also for the compressible problem (1.1), (1.2), (1.5) provided that $\varepsilon > 0$ is sufficiently small and the initial data u_0, ρ_0 are *almost incompressible.* A precise statement is given in the next section.

Under the assumptions of our theorem, the compressible solution $u = u^\varepsilon, \rho = \rho^\varepsilon$ consists of leading order of the incompressible flow U, P, and it is essential for our proof that U, P decays to zero as $t \to \infty$. Thus, at present it is not clear if our result generalizes to Euler flow ($\nu = \eta = 0$) or to the forced equation

$$u_t + (u \cdot \nabla)u + (1+\varepsilon^2\rho)^{\gamma-2}\nabla\rho = (1+\varepsilon^2\rho)^{-1}\left(\nu\Delta u + \eta\nabla\nabla\cdot u + F(x,y,t)\right),$$

$$\varepsilon^2\left(\rho_t + (u\cdot\nabla)\rho\right) + (1+\varepsilon^2\rho)\nabla\cdot u = 0$$

with $F \in C^\infty$. In both cases, the solution U, P is C^∞ for $0 \leq t < \infty$, but generally does not tend to zero.

A generalization of our result to three-dimensional flow is straightforward, however, if we assume that the initial incompressible velocity field $U_0(x,y,z)$ leads to a classical solution U, P for $0 \leq t < \infty$, which—together with all its derivatives—tends to zero as $t \to \infty$.

Our approach can also be used to establish an asymptotic expansion result for $u^\varepsilon, \rho^\varepsilon$. Leading order corrections to U, P consist of fast oscillations of the pressure and the dilatation, $\nabla \cdot u^\varepsilon$. These corrections are solutions of the damped wave equation

$$w_{tt} = \frac{1}{\varepsilon^2}\Delta w + (\nu+\eta)\Delta w_t\ .$$

[1] Only a finite number of derivatives will be used. We work with C^∞-functions so that we do not have to keep track of the exact number of derivatives required at every step.

We will not elaborate on this here, however, since corresponding asymptotic expansions in finite time intervals have already been derived in [3, 4, 6].

There has been much recent work on weak solutions of the equations of compressible flow. Hoff (see [2] and references therein) has considered discontinuous initial data in one space dimension. Lions [7, 8, 9] has developed a multidimensional theory of weak solutions for the isentropic system we consider and has established connections with the theory of weak solutions for the incompressible case. Results on classical solutions for small data are also known. See [1] for a proof using the techniques of this paper and for references to earlier work.

2. Notations and statement of main theorem. The Euclidean inner product and norm on $\mathbb{C}^n$ (and $\mathbb{R}^n$) are denoted by

$$\langle u, v\rangle = \sum_{j=1}^{n} \bar{u}_j v_j, \quad |u| = \langle u, u\rangle^{1/2} .$$

Also, if $A \in \mathbb{C}^{n\times n}$ is an $n\times n$ matrix, then $|A|$ denotes the corresponding matrix norm. If $H_1, H_2 \in \mathbb{C}^{n\times n}$ are Hermitian matrices, then we write $H_1 \le H_2$ iff $u^* H_1 u \le u^* H_2 u$ for all $u \in \mathbb{C}^n$. For functions $u, v \in L_2 = L_2(T^2, \mathbb{R}^n)$, the L_2-inner product and norm are defined by

$$(u, v) = \int_{T^2} \langle u(x, y), v(x, y)\rangle dxdy, \quad \|u\| = (u, u)^{1/2} .$$

We recall Parseval's relation,

$$(u, v) = \sum_{k\in\mathbb{Z}^2} \langle \hat{u}(k), \hat{v}(k)\rangle,$$

where

$$\hat{u}(k) = \frac{1}{2\pi} \int_{T^2} e^{-i(k_1 x + k_2 y)} u(x, y) dxdy .$$

For spatial differential expressions we use multi-index notation,

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x^{\alpha_1} \partial y^{\alpha_2}}, \quad |\alpha| = \alpha_1 + \alpha_2 .$$

We partially order multi-indices by

$$\beta < \alpha \iff \beta_1 \le \alpha_1, \quad \beta_2 \le \alpha_2, \quad \beta \ne \alpha.$$

The Sobolev inner products and norms, based on L_2, are

$$(u, v)_j = \sum_{|\alpha|\le j} (D^\alpha u, D^\alpha v) , \quad \|u\|_j = (u, u)_j^{1/2} , \quad j = 1, 2, \ldots .$$

We write $H^j = H^j(T^2, \mathbb{R}^n)$ for the corresponding Sobolev space. The sup-norm of any bounded function u is denoted by $|u|_\infty$.

In our main theorem formulated next, let $U_0 \in C^\infty(T^2)$ denote an incompressible velocity field and let $P_0 = P_0(x, y)$ denote the corresponding pressure at $t = 0$; i.e., if U, P is the solution of (1.3), (1.4), (1.7), (1.8), then $P_0(x, y) = P(x, y, 0)$.

THEOREM 2.1. *Consider the compressible problem,* (1.1), (1.2), (1.5), (1.6), *with* $u_0, \rho_0 \in C^\infty(T^2)$, $\nu > 0, \varepsilon > 0$, $\eta + \nu > 0$. *There is* $\varepsilon_0 = \varepsilon_0(U_0, \nu, \eta) > 0$ *and* $\delta_0 = \delta_0(U_0, \nu, \eta) > 0$ *so that the following holds. If* $0 < \varepsilon \le \varepsilon_0$ *and*

$$\|u_0 - U_0\|_3^2 + \varepsilon^2 \|\rho_0 - P_0\|_3^2 \le \delta_0^2 \ , \tag{2.1}$$

then the solution u, ρ *is in* $C^\infty(T^2 \times [0, \infty))$ *and approaches a uniform state* $(\bar{u}'_\infty, 0)$ *as* $t \to \infty$.

A main point of the theorem is that the radius, given by δ_0, of the ball about incompressible data can be chosen independently of the singular perturbation parameter $0 < \varepsilon < \varepsilon_0$. We note that the bounded derivative principle (in finite time intervals) is valid for the system (1.1), (1.2); see [6]. Thus, a highly oscillatory part of the solution could be suppressed by proper initialization. We emphasize that our theorem does *not* require such an initialization; i.e., a highly oscillatory part (sound waves) is generally present in the solution. The amplitude of the sound waves is initially controlled by δ_0.

An outline of the proof follows. If we subtract the incompressible solution U, P from u, ρ, then we obtain equations for the differences $u' = u - U, \rho' = \rho - P$ with a forcing term of order $\mathcal{O}(\varepsilon^2)$. The details are carried out in section 3. The 3-vector

$$w = (u', \varepsilon\rho')^T$$

satisfies a coupled parabolic–hyperbolic system, where the large hyperbolic part is symmetrized. This symmetrization is the reason for multiplying ρ' by ε. The system satisfied by w is central for our discussion.

General results on coupled parabolic–hyperbolic systems (see, for example, [1]) imply that w is C^∞ in some maximal interval $0 \le t < t_\varepsilon$, and if t_ε is finite, then

$$\sup_{0 \le t < t_\varepsilon} \|w(\cdot, t)\|_3 = \infty \ .$$

(In N space dimensions, the crucial degree of smoothness is given by the smallest integer $s > \frac{N}{2} + 1$. For $N = 2$ we have $s = 3$.) Thus it suffices to show that $\|w(\cdot, t)\|_3$ remains bounded in order to obtain all-time existence and C^∞ smoothness.

Standard L_2-estimates for w and its space derivatives are not good enough to imply all-time existence. A crucial step is the construction of a new norm $\|\cdot\|_H$, in which the solution of the linearized constant coefficient problem decays exponentially. The construction of this new norm is carried out in section 4. A main technical difficulty is to obtain ε-independent bounds for the related symmetrizer.

3. Equations for the perturbed variables. The solution u, ρ of (1.1), (1.2), (1.5) consists of a slow part and a part which is highly oscillatory in time. Following [6], we first subtract the slow part of the solution in order to obtain an equation with small initial data for the remainder.

If U, P denotes the solution of the incompressible problem (1.3), (1.4), (1.7), (1.8), then we define new variables u', ρ' by

$$u = U + u', \quad \rho = P + \rho' \ .$$

A lengthy but straightforward computation yields

$$\begin{aligned} &u'_t + (U \cdot \nabla)u' + (u' \cdot \nabla)U + (u' \cdot \nabla)u' + \nabla\rho' + \varepsilon^2(P + \rho')\Pi(0, \varepsilon^2(P + \rho'))\nabla\rho' \\ &\quad + \varepsilon^2\rho' F(P, U, \rho', \varepsilon) = (1 + \varepsilon^2 P + \varepsilon^2\rho')^{-1} (\nu\Delta u' + \eta\nabla\nabla \cdot u') + \varepsilon^2 g_1 \ , \end{aligned} \tag{3.1}$$

$$\varepsilon^2 (\rho'_t + (U \cdot \nabla)\rho' + (u' \cdot \nabla)P + (u' \cdot \nabla)\rho') + (1 + \varepsilon^2 P + \varepsilon^2\rho')\nabla \cdot u' = \varepsilon^2 g_2 \tag{3.2}$$

with

$$\Pi(a,b) = (\gamma-2)\int_0^1 (1+a+bz)^{\gamma-3}dz,$$
$$F = \Pi(\varepsilon^2 P, \varepsilon^2\rho')\nabla P + \nu(1+\varepsilon^2 P)^{-1}(1+\varepsilon^2 P+\varepsilon^2\rho')^{-1}\Delta U,$$
$$g_1 = -P\Pi(0,\varepsilon^2 P)\nabla P - \nu P(1+\varepsilon^2 P)^{-1}\Delta U,$$
$$g_2 = -(P_t + (U\cdot\nabla)P).$$

The variables u', ρ' satisfy initial conditions

$$u' = u_0'(x,y), \quad \rho' = \rho_0'(x,y) \qquad \text{at} \quad t=0\ , \tag{3.3}$$

where $u_0' = u_0 - U_0, \rho_0' = \rho_0 - P_0$, and therefore

$$\|u_0'\|_3^2 + \varepsilon^2\|\rho_0'\|_3^2 \le \delta_0^2 \tag{3.4}$$

by assumption (2.1). We note that the incompressible solution satisfies estimates

$$\|U(\cdot,t)\|_j + \|P(\cdot,t)\|_j \le C_j e^{-c_j t}, \quad t\ge 0, \quad j=0,1,\dots \tag{3.5}$$

with $C_j > 0$ and $c_j > 0$ independent of t. Corresponding estimates for time derivatives follow from the differential equation. (We will use the following convention: C, C_j, etc. denote sufficiently large constants which may depend on U_0 and the viscosities ν, η. Similarly, c, c_j, etc. denote sufficiently small positive constants which may depend on U_0 and ν, η. All constants are independent of $u_0, \rho_0, t, \varepsilon$, and the wave vector k used below. At different appearances, the constants may have different meanings.)

The inhomogeneous terms $\varepsilon^2 g_1$, $\varepsilon^2 g_2$ in (3.1), (3.2) can be reduced to higher order by constructing additional terms in the expansion of the "slow" part of the solution [6], but this is unnecessary for our purposes.

We also note that the nonlinearities in (3.1), (3.2) are smooth only under some restrictions on the arguments. To be specific, we must have $(1+\varepsilon^2 P), (1+\varepsilon^2\rho'), (1+\varepsilon^2 P+\varepsilon^2\rho') > 0$. As we derive time-uniform bounds on the solution valid for all ε sufficiently small, we can guarantee that the arguments remain in the appropriate domains by further restricting ε_0 if necessary.

To motivate a scaling that we use below, we first consider the linear constant coefficient system obtained from (3.1), (3.2) by setting the incompressible solution, (U,P) to zero and ignoring nonlinearities:

$$u_t' + \nabla\rho' = \nu\Delta u' + \eta\nabla\nabla\cdot u', \tag{3.6}$$

$$\varepsilon^2\rho_t' + \nabla\cdot u' = 0\ . \tag{3.7}$$

For $\nu = \eta = 0$ this system is hyperbolic, but unsymmetric, and in fact strongly unbalanced for $0 < \varepsilon \ll 1$. To symmetrize this underlying hyperbolic system, we use the variable $\varepsilon\rho'$ instead of ρ'. Thus we introduce the 3-vector

$$w = \begin{pmatrix} u' \\ \varepsilon\rho' \end{pmatrix}.$$

Returning to the full system (3.1), (3.2) and dividing (3.2) by ε, we obtain the following system for w:

$$w_t + ((U+u')\cdot\nabla)w = A_\varepsilon w + \varepsilon G + Q. \tag{3.8}$$

Here A_ε is the constant coefficient operator (compare with (3.6), (3.7))

$$A_\varepsilon = -\frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 & \partial_x \\ 0 & 0 & \partial_y \\ \partial_x & \partial_y & 0 \end{pmatrix} + \begin{pmatrix} \nu\Delta + \eta\frac{\partial^2}{\partial x^2} & \eta\frac{\partial^2}{\partial x \partial y} & 0 \\ \eta\frac{\partial^2}{\partial x \partial y} & \nu\Delta + \eta\frac{\partial^2}{\partial y^2} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{3.9}$$

G is the inhomogeneous term

$$G = \begin{pmatrix} \varepsilon g_1 \\ g_2 \end{pmatrix}, \tag{3.10}$$

and the remaining terms Q are given by

$$Q = Q_1 + Q_2, \tag{3.11}$$

$$Q_1 = \begin{pmatrix} \frac{(\varepsilon w^{(3)} + \varepsilon^2 P)}{(1+\varepsilon w^{(3)} + \varepsilon^2 P)}\,(\nu\Delta u' + \eta\nabla\nabla\cdot u') - (w^{(3)} + \varepsilon P)\Pi(0, \varepsilon w^{(3)} + \varepsilon^2 P)\nabla w^{(3)} \\ 0 \end{pmatrix}, \tag{3.12}$$

$$Q_2 = \begin{pmatrix} -(u'\cdot\nabla)U - \varepsilon w^{(3)} F \\ -\varepsilon(u'\cdot\nabla)P - (w^{(3)} + \varepsilon P)\nabla\cdot u' \end{pmatrix}. \tag{3.13}$$

The initial condition for w is

$$w = w_0(x,y) := \begin{pmatrix} u_0' \\ \varepsilon\rho_0' \end{pmatrix} \quad \text{at} \quad t = 0;$$

thus

$$\|w_0\|_3 \le \delta_0. \tag{3.14}$$

In the next section we consider the constant coefficient system $w_t = A_\varepsilon w$.

4. Decay estimates for $w_t = A_\varepsilon w$. Let $w = w(x,y,t)$ denote a solution of the linear constant coefficient system $w_t = A_\varepsilon w$ with A_ε given by (3.9). The standard L_2-estimate reads

$$\begin{aligned} \frac{d}{dt}\|w\|^2 &= 2(w, w_t) \\ &= -2\nu\sum_{j=1}^{2}\|Dw^{(j)}\|^2 - 2\eta\|\nabla\cdot u'\|^2 \ \le\ 0\ . \end{aligned}$$

Here, $\|Dv\|^2 = \|v_x\|^2 + \|v_y\|^2$. Clearly, the third component of w does not appear on the right side (since the third component of (3.9) does not contain the viscosity terms), and therefore we cannot obtain exponential decay of w in the L_2-norm. This is the fundamental reason for the difficulty in proving our main theorem. We will, however, construct a new norm, which is equivalent to the L_2 norm, in which we have exponential decay, except for the spatial averages of w. (The spatial averages of the solutions of $w_t = A_\varepsilon w$ remain constant in time. For the full system (3.8), they will require a separate consideration, with bounds following from the conservation of mass and momentum.)

To construct the new norm, we use Fourier expansion in space. The Fourier representation of w is

$$w(x,y,t) = \frac{1}{2\pi}\sum_{k\in\mathbb{Z}^2} e^{ik_1x+ik_2y}\hat{w}(k,t)$$

with

$$\hat{w}(k,t) = \frac{1}{2\pi}\int_{T^2} e^{-ik_1x-ik_2y}w(x,y,t)dxdy\ ,$$

and the system $w_t = A_\varepsilon w$ transforms to

$$\hat{w}_t(k,t) = \hat{A}_\varepsilon(k)\hat{w}(k,t), \quad k\in\mathbb{Z}^2\ . \tag{4.1}$$

Here

$$\hat{A}_\varepsilon(k) = -\frac{i}{\varepsilon}\begin{pmatrix} 0 & 0 & k_1\\ 0 & 0 & k_2\\ k_1 & k_2 & 0\end{pmatrix} - \begin{pmatrix} \nu|k|^2+\eta k_1^2 & \eta k_1k_2 & 0\\ \eta k_1k_2 & \nu|k|^2+\eta k_2^2 & 0\\ 0 & 0 & 0\end{pmatrix} \tag{4.2}$$

is the symbol of A_ε; see (3.9). We first show that the eigenvalues of $\hat{A}_\varepsilon(k)$ have negative real parts for all $k\neq 0$.

LEMMA 4.1. *Let $\lambda_j = \lambda_j(\nu,\eta,\varepsilon,k)$ denote the eigenvalues of $\hat{A}_\varepsilon(k)$. Then we have*

$$\operatorname{Re}\lambda_j \le \max\left\{-\nu, -\frac{\nu+\eta}{2}, -\frac{1}{(\nu+\eta)\varepsilon^2}\right\} < 0, \quad j=1,2,3$$

for all $\nu>0, \nu+\eta>0, \varepsilon>0, k\in\mathbb{Z}^2, k\neq 0$.

Proof. For

$$\phi_1 = \frac{1}{|k|}\begin{pmatrix} -k_2\\ k_1\\ 0\end{pmatrix}$$

we have

$$\hat{A}_\varepsilon(k)\phi_1 = -\nu|k|^2\phi_1;$$

thus $\lambda_1 = -\nu|k|^2$. Next let ϕ_2,ϕ_3, which span the subspace of $\mathbb{R}^3$ orthogonal to ϕ_1, be given by

$$\phi_2 = \frac{1}{|k|}\begin{pmatrix} k_1\\ k_2\\ 0\end{pmatrix}\ , \qquad \phi_3 = \begin{pmatrix} 0\\ 0\\ 1\end{pmatrix}\ .$$

Then we have

$$\hat{A}_\varepsilon(k)\phi_2 = -(\nu+\eta)|k|^2\phi_2 - \frac{i}{\varepsilon}|k|\phi_3,$$
$$\hat{A}_\varepsilon(k)\phi_3 = -\frac{i}{\varepsilon}|k|\phi_2.$$

This implies that the eigenvalues $\lambda_{2,3}$ of $\hat{A}_\varepsilon(k)$ are the eigenvalues of

$$B = \begin{pmatrix} -(\nu+\eta)|k|^2 & -\frac{i}{\varepsilon}|k|\\ -\frac{i}{\varepsilon}|k| & 0\end{pmatrix}\ , \tag{4.3}$$

and we obtain that

$$\lambda_{2,3} = -\frac{(\nu+\eta)|k|^2}{2}\left\{1 \pm \sqrt{1 - \frac{4}{(\nu+\eta)^2\varepsilon^2|k|^2}}\right\}. \tag{4.4}$$

If $4 > (\nu+\eta)^2\varepsilon^2|k|^2$, then the root is purely imaginary; thus

$$\operatorname{Re}\lambda_{2,3} = -\frac{(\nu+\eta)|k|^2}{2} \le -\frac{\nu+\eta}{2}\ .$$

If $(\nu+\eta)^2\varepsilon^2|k|^2 \ge 4$, then $\lambda_{2,3}$ are real, and we have

$$\lambda_2 \le \lambda_3 = -\frac{r}{2} + \frac{r}{2}\sqrt{1 - \frac{4}{(\nu+\eta)\varepsilon^2 r}} \tag{4.5}$$

with $r = (\nu+\eta)|k|^2$. Clearly, $\lambda_3 \to -\frac{1}{(\nu+\eta)\varepsilon^2}$ as $r \to \infty$. Furthermore, the function of r on the right side of (4.5) is monotonically increasing. Therefore,

$$\lambda_2 \le \lambda_3 \le -\frac{1}{(\nu+\eta)\varepsilon^2}\ ,$$

and the lemma is proved. □

Remarks. 1. Let S denote the orthogonal matrix containing as columns the vectors ϕ_j defined in the proof of the previous lemma,

$$S = \frac{1}{|k|}\begin{pmatrix} -k_2 & k_1 & 0 \\ k_1 & k_2 & 0 \\ 0 & 0 & |k| \end{pmatrix}\ . \tag{4.6}$$

Then the proof shows that

$$\hat{A}_\varepsilon(k)S = S\begin{pmatrix} -\nu|k|^2 & 0 & 0 \\ 0 & b_{11} & b_{12} \\ 0 & b_{21} & b_{22} \end{pmatrix}, \tag{4.7}$$

where B is given in (4.3).

2. In the three-dimensional case, let $\psi_1, \psi_2 \in \mathbb{R}^3$ denote vectors so that ψ_1, ψ_2, $k/|k|$ form an orthonormal system of $\mathbb{R}^3$. Setting $\phi_j = (\psi_j, 0)^T$, we obtain

$$\hat{A}_\varepsilon(k)\phi_j = -\nu|k|^2\phi_j, \quad j = 1, 2\ .$$

Furthermore, let

$$\phi_3 = \frac{1}{|k|}(k_1, k_2, k_3, 0)^T, \quad \phi_4 = (0, 0, 0, 1)^T\ .$$

Computation of $\hat{A}_\varepsilon(k)\phi_j, j = 3, 4$, shows that the same matrix B as given in (4.3) determines the eigenvalues $\lambda_{3,4}$ of $\hat{A}_\varepsilon(k)$. Thus, Lemma 4.1 generalizes directly to the three-dimensional case.

Since the matrices $\hat{A}_\varepsilon(k)$ in (4.1) all have their eigenvalues in the left half-plane, all solutions $\hat{w}(k,t), k \ne 0$, tend to zero as $t \to \infty$. We need to quantify this decay. For clarity, consider first any ODE system

$$\frac{dy}{dt} = Ay, \quad y(t) \in \mathbb{C}^n, \tag{4.8}$$

where A is a constant $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n, \operatorname{Re}\lambda_j \le -c_0 < 0$. One can construct a norm, based on an inner product, in which all solutions of (4.8) decay exponentially. In particular, we can construct a positive definite Hermitian matrix H satisfying

$$HA + A^*H \le -c_0 H. \tag{4.9}$$

The positive definite Hermitian matrix H defines an inner product and a norm on $\mathbb{C}^n$,

$$\langle u, v\rangle_H = u^*Hv, \quad |u|_H = \langle u, u\rangle_H^{1/2} .$$

For any solution $y(t)$ of (4.8) we have

$$\begin{aligned}\frac{d}{dt}|y|_H^2 &= \langle y, y'\rangle_H + \langle y', y\rangle_H \\ &= \langle y, Ay\rangle_H + \langle Ay, y\rangle_H \\ &= \langle y, HAy\rangle + \langle y, A^*Hy\rangle \\ &\le -c_0|y|_H^2.\end{aligned}$$

The last inequality follows from the crucial property $HA + A^*H \le -c_0 H$ of the matrix H. Therefore, for all solutions $y(t)$ of (4.8), the expression $|y(t)|_H$ decays exponentially. The matrix $H = H^* > 0$ is called a symmetrizer for the system (4.8) since—in $|\cdot|_H$—A acts like a matrix with negative definite symmetric part.

For the family of systems (4.1) we need additional properties of the symmetrizer $H = H(\nu, \eta, \varepsilon, k)$, namely some uniform estimates. As we will see, it suffices to consider the matrices B given in (4.3). Recall that, by Lemma 4.1, the eigenvalues $\lambda_{2,3}$ of B satisfy

$$\operatorname{Re}\lambda_{2,3} \le -c_0 < 0 , \quad c_0 = c_0(\nu + \eta) \tag{4.10}$$

for all $0 < \varepsilon \le 1, k \in \mathbb{Z}^2, k \ne 0$, and any fixed $\nu + \eta > 0$.

LEMMA 4.2. *For fixed $\nu + \eta > 0$, consider the family of matrices B given in* (4.3) *for $k \in \mathbb{Z}^2, k \ne 0$. There are positive constants $c_0, c_1, C_1, C_2, \varepsilon_0$, depending only on $\nu + \eta$, and there are Hermitian matrices $H = H(\nu + \eta, \varepsilon, k) \in \mathbb{C}^{2\times 2}$ for $0 < \varepsilon \le \varepsilon_0$ with the following properties:*

$$0 < (1 - C_1\varepsilon)I \le H \le (1 + C_1\varepsilon)I \tag{4.11}$$

$$q^*(HB + B^*H)q \le -c_0 q^*Hq - c_1|k|^2|q_1|^2 \quad \forall q \in \mathbb{C}^2, \tag{4.12}$$

$$|H - I| \le C_2\frac{\varepsilon}{|k|}. \tag{4.13}$$

Proof. We seek H in the form

$$H = \begin{pmatrix} 1 & i\frac{\varepsilon\mu}{|k|} \\ -i\frac{\varepsilon\mu}{|k|} & 1 \end{pmatrix}, \tag{4.14}$$

which, for fixed $\mu > 0$, clearly satisfies (4.11) and (4.13). To check (4.12) we directly compute $HB + B^*H$ and obtain

$$\begin{aligned} HB + B^*H &= \begin{pmatrix} -2(\nu+\eta)|k|^2 + 2\mu & -i\mu\varepsilon(\nu+\eta)|k| \\ i\mu\varepsilon(\nu+\eta)|k| & -2\mu \end{pmatrix} \\ &= \begin{pmatrix} -c_1|k|^2 & 0 \\ 0 & 0 \end{pmatrix} + R, \end{aligned} \tag{4.15}$$

where we choose $c_1 = \nu + \eta$. We must only show that R is negative definite and that its largest eigenvalue is bounded above by $-\bar{c}_0 < 0$ independent of $|k|$ and ε sufficiently small. Then (4.15) combined with (4.11) will imply (4.12). To that end we compute

$$\det(R) = \mu(\nu+\eta)|k|^2 \left(2 - \mu(\nu+\eta)\varepsilon^2 - 4\frac{\mu}{(\nu+\eta)|k|^2}\right), \tag{4.16}$$

$$\operatorname{tr}(R) = -(\nu+\eta)|k|^2. \tag{4.17}$$

Clearly, the trace is negative and the determinant positive for ε, μ sufficiently small. For 2×2 Hermitian matrices this implies negative definiteness. Choosing, for example,

$$\mu = \frac{\nu+\eta}{8}, \quad \varepsilon_0 < \frac{2}{\nu+\eta}, \tag{4.18}$$

we also have the bound

$$\bar{c}_0 \geq -\frac{\det(R)}{\operatorname{tr}(R)} > \frac{\nu+\eta}{8}. \tag{4.19}$$

This completes the proof. □

Combining this lemma with Remark 1 above we obtain the following result for the systems (4.1).

LEMMA 4.3. *For fixed $\nu > 0$, $\nu + \eta > 0$, and ε_0 sufficiently small consider the family of matrices $\hat{A}_\varepsilon(k)$ given by (4.2) for $0 < \varepsilon \leq \varepsilon_0, k \in \mathbb{Z}^2, k \neq 0$. There are positive constants c_0, c_1, C_1, C_2, depending only on ν, η, and there are Hermitian matrices $H = H(\nu, \eta, \varepsilon, k) \in \mathbb{C}^{3\times 3}$ with the following properties:*

$$0 < (1 - C_1\varepsilon)I \leq H \leq (1 + C_1\varepsilon)I, \tag{4.20}$$

$$q^*(H\hat{A}_\varepsilon + \hat{A}_\varepsilon^* H)q \leq -c_0 q^* H q - c_1|k|^2(|q_1|^2 + |q_2|^2) \quad \forall q \in \mathbb{C}^3, \tag{4.21}$$

$$|H - I| \leq \frac{C_2\varepsilon}{|k|}. \tag{4.22}$$

Proof. Let $\tilde{H} \in \mathbb{C}^{2\times 2}$ denote the symmetrizer constructed in the proof of Lemma 4.2 for the matrices B. We set

$$H = S \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H} \end{pmatrix} S^*,$$

where the orthogonal matrix S is given in (4.6). Then H clearly satisfies (4.20) and (4.22). Setting $p = S^*q$ we find

$$q^*(H\hat{A}_\varepsilon + \hat{A}_\varepsilon^* H)q = p^* \begin{pmatrix} -\nu|k|^2 & 0 \\ 0 & \tilde{H}B + B^*\tilde{H} \end{pmatrix} p \leq -c_0 p^* H p - c_1|k|^2(|p_1|^2 + |p_2|^2).$$

Inequality (4.21) follows from the orthogonality of S and its block structure, completing the proof. □

By the previous lemma, the symmetrizer $H = H(\nu, \eta, \varepsilon, k)$ of (4.1) is constructed for $\nu > 0$, $\nu + \eta > 0$, $0 < \varepsilon \leq \varepsilon_0$, $k \in \mathbb{Z}^2$, $k \neq 0$. For $k = 0$ we set $H = I$. Then we define a new inner product on $L_2 = L_2(T^2, \mathbb{R}^3)$ by

$$(w_1, w_2)_H = \sum_{k\in\mathbb{Z}^2} \hat{w}_1(k)^* H(\nu, \eta, \varepsilon, k)\hat{w}_2(k) \ .$$

The corresponding norm is denoted by $\|\cdot\|_H$, where the dependency on ν, η, and ε is suppressed in our notation. The following result is an easy consequence of Lemma 4.3 and Parseval's relation.

LEMMA 4.4.

(a) *For all* $w = (u', w^{(3)})^T \in L_2$,

$$(1 - C_1\varepsilon)\|w\|^2 \le \|w\|_H^2 \le (1 + C_1\varepsilon)\|w\|^2 \ ,$$

i.e., the norms $\|\cdot\|_H$ *are equivalent to the* L_2*-norm, and in fact are* ε*-close to it.*

(b) *If* $w = w(x, y)$ *is sufficiently regular (e.g.,* $w \in H^2$*) and* $\hat{w}(0) = 0$ *(i.e., the spatial averages of the components of* w *are zero), then*

$$(w, A_\varepsilon w)_H + (A_\varepsilon w, w)_H \le -c_0\|w\|_H^2 - c_1\|Du'\|^2 \ .$$

In this sense, A_ε *is negative definite, uniformly in* ε*, with "parabolic" estimates on the velocities.*

(c) *If* $w_1 \in L_2, w_2 \in H^1$*, then*

$$|(w_1, Dw_2)_H - (w_1, Dw_2)| \le \varepsilon C_2\|w_1\|\|w_2\| \ .$$

Here $D = \partial_x$ *or* $D = \partial_y$*. In this sense, we "gain" one derivative when comparing the* H*-inner product and the* L_2*-inner product, and the constant in the estimate is proportional to* ε*.*

(d) *If* $w_1, w_2 \in H^1$ *then*

$$(w_1, Dw_2)_H = -(Dw_1, w_2)_H \quad \textit{for} \quad D = \partial_x \textit{ or } D = \partial_y \ ,$$

i.e., the rule of integration by parts, $(w_1, Dw_2) = -(Dw_1, w_2)$*, is valid in the* H*-inner product.*

As a consequence of (b) we note the following: if $w = w(x, y, t)$ solves

$$w_t = A_\varepsilon w, \quad w = w_0(x, y) \quad \text{at} \quad t = 0,$$

and $\hat{w}_0(0) = 0$, then

$$\begin{aligned}\frac{d}{dt}\|w\|_H^2 &= (w, A_\varepsilon w)_H + (A_\varepsilon w, w)_H \\ &\le -c_0\|w\|_H^2 - c_1\|Du'\|_H^2 \ .\end{aligned}$$

Therefore, $\|w(\cdot, t)\|_H$ decays exponentially in time. The rate of decay is independent of $0 < \varepsilon \le \varepsilon_0$.

The following result will be useful when we estimate derivative terms in the sections below. We first show the result for the L_2-inner product, then for the H-inner product.

LEMMA 4.5. *Let* $w \in H^1(T^2, \mathbb{R}^3)$ *and let* $v \in C^1(T^2)$ *be a real valued function. Then we have*

$$|(w, vDw)| \le \frac{1}{2}|Dv|_\infty\|w\|^2 \tag{4.23}$$

for $D = \partial_x$ *or* $D = \partial_y$*.* The crucial point is that no derivative of w appears on the right side.

Proof. Integration by parts yields

$$(w, vDw) = (vw, Dw) = -((Dv)w, w) - (vDw, w) \ . \tag{4.24}$$

Therefore,

$$(w, vDw) = -\frac{1}{2}((Dv)w, w)$$

and (4.23) follows. □

Equation (4.24) is valid if v takes Hermitian matrices as values; the estimate corresponding to (4.23) is a standard tool for symmetric hyperbolic systems; see, e.g., [5]. Equation (4.24) is not valid, however, if the L_2-inner product is replaced by the H-inner product. For this case, we obtain the following slightly weaker result.

LEMMA 4.6. *Under the assumptions of Lemma* 4.5 *we have*

$$|(w, vDw)_H| \le C(\varepsilon|v|_\infty + |Dv|_\infty)\|w\|^2 \ . \tag{4.25}$$

Proof. We have

$$(w, vDw)_H = (w, D(vw))_H - T_1$$

with $T_1 = (w, (Dv)w)_H$; thus $|T_1| \le C|Dv|_\infty\|w\|^2$. By Lemma 4.4 it follows that

$$(w, D(vw))_H = (w, D(vw)) + T_2$$

with

$$|T_2| \le \varepsilon C\|w\|\|vw\| \le \varepsilon C|v|_\infty\|w\|^2.$$

Finally,

$$(w, D(vw)) = (w, vDw) + T_3$$

with $T_3 = (w, (Dv)w)$; thus $|T_3| \le |Dv|_\infty\|w\|^2$. Using the estimate of the previous lemma for (w, vDw), estimate (4.25) follows. □

5. Conservation laws and estimates of the mean. The results of the preceding section cannot be used directly to estimate the spatial averages of the perturbed quantities, $\hat{w}(0,t)$, as these are invariant for the linearized evolution. However, we can bound them using the basic conservation laws of mass and linear momentum.

First recall that $u' = u - U$, $\rho' = \rho - P$, $w = (u', \varepsilon\rho')^T$, and that total mass

$$s = \int_{T^2}(1 + \varepsilon^2\rho)dxdy$$

and total momentum

$$m = \int_{T^2}(1 + \varepsilon^2\rho)udxdy$$

are constant in time, as follows directly from (1.2) and (1.1). Moreover, we have assumed (see (1.6), (1.8)) that

$$\int_{T^2} u_0 dxdy = \int_{T^2} U_0 dxdy = 0, \quad \int_{T^2} \rho_0 dxdy = \int_{T^2} P dxdy = 0, \tag{5.1}$$

where u_0, U_0, and ρ_0 are the initial data for u, U, and ρ, respectively. Combining these we have the following lemma.

LEMMA 5.1. *If the initial data satisfy* (5.1) *and* w *is a solution of* (3.8), *we have,* $\forall t \geq 0$,

$$\hat{w}^{(3)}(0,t) = \varepsilon \hat{\rho}'(0,t) = 0. \tag{5.2}$$

Proof. By conservation of mass and (5.1),

$$0 = \int_{T^2} (\rho' + P) dxdy,$$

which implies, using (1.8),

$$\int_{T^2} \rho' dxdy = -\int_{T^2} P dxdy = 0,$$

completing the proof. ∎

We now consider conservation of momentum. Note that by (1.6) the momentum satisfies

$$m = \int_{T^2} (1 + \varepsilon^2 \rho_0) u_0 dxdy = \varepsilon^2 \int_{T^2} \rho_0 u_0 dxdy. \tag{5.3}$$

We then have the following lemma.

LEMMA 5.2. *Suppose the initial data satisfy* (5.1) *and that* w *is a solution of* (3.8). *Let*

$$\bar{u}' = (2\pi)^{-2} \int_{T^2} u' dxdy,$$

and let[2]

$$w^c = w - (\bar{u}', 0)^T.$$

Then, for some constants C, c *depending on* U_0 *and* ν *but independent of* u_0, ρ_0, ε, *and* $\forall t \geq 0$ *we have*

$$|\bar{u}'| \leq \varepsilon C \left(\delta_0 + \varepsilon + \|w^c\|_H (e^{-ct} + \|w^c\|_H) \right). \tag{5.4}$$

Proof. We have

$$m = \int_{T^2} (1 + \varepsilon^2 \rho' + \varepsilon^2 P)(u' + U) dxdy = \varepsilon^2 \int_{T^2} \rho_0 u_0 dxdy.$$

Recalling that ρ', U, and P all have mean zero and solving for $\bar{u}'$, we obtain

$$\bar{u}' = \varepsilon^2 (2\pi)^{-2} \int_{T^2} (\rho_0 u_0 - PU) dxdy - \varepsilon (2\pi)^{-2} \int_{T^2} \left(\varepsilon (\rho' + P)(u' - \bar{u}') + \varepsilon \rho' U \right) dxdy.$$

From (3.14) we conclude

$$\varepsilon \int_{T^2} |\rho_0 u_0 - PU| dxdy \leq \varepsilon \int_{T^2} |\rho_0 u_0 - P_0 U_0| dxdy + \varepsilon \int_{T^2} |P_0 U_0 - PU| dxdy \leq C\delta_0 + C\varepsilon.$$

[2]Here $\bar{u}'$ is the spatial average of u'. Also w^c is centered in the sense that its spatial mean is zero.

Using the Cauchy–Schwarz inequality on the remaining terms, the estimate readily follows. ∎

Noting that the decomposition into spatially constant and spatially centered components is orthogonal with respect to the H inner product we have, as an immediate corollary of these lemmas,

$$\|w^c\|_H^2 \geq \|w\|_H^2 - \varepsilon^2 C(\varepsilon^2 + \delta_0^2 + \|w^c\|_H^2(e^{-ct} + \|w^c\|_H^2)). \tag{5.5}$$

6. Proof of Theorem 2.1. We begin by estimating time derivatives of the H-norm of the solution, w, to (3.8) along with its first three space derivatives. Set

$$\phi^2(t) = \frac{1}{2}\sum_{|\alpha|\leq 3} \|D^\alpha w(\cdot,t)\|_H^2, \quad h^2(t) = \sum_{|\alpha|=4} \|D^\alpha u'(\cdot,t)\|^2. \tag{6.1}$$

Note that short-time existence of a solution is guaranteed by the standard theory of parabolic–hyperbolic systems [5]. Moreover, we assume a bound on ϕ, $\phi(t) \leq 1$, so that the various nonlinear functions appearing in the equations remain smooth and so that higher powers of ϕ can be bounded by lower ones. We then have the fundamental estimate.

LEMMA 6.1. *Under the assumptions of Theorem* 2.1 *and in an interval* $0 \leq t < t_\varepsilon$ *of existence of a smooth solution,* w*, with* $\phi(t) \leq 1$*, the function* $\phi^2(t)$ *satisfies*

$$\begin{aligned}\frac{d}{dt}\phi^2(t) \leq \left(Ce^{-ct} - 2\tilde{c}_0\right)\phi^2(t) + \left(\varepsilon^2 Ce^{-ct} - \frac{1}{2}c_1\right)h^2(t)\\ +\varepsilon^2 Ce^{-ct} + \varepsilon^4 C + \varepsilon^2\delta_0^2 C + C\phi^3(t) + \varepsilon C\phi(t)h^2(t),\end{aligned} \tag{6.2}$$

$$\phi^2(0) \leq \frac{1}{2}(1 + C_1\varepsilon)\delta_0^2. \tag{6.3}$$

(Here and throughout constants depend on ν*,* η*,* U_0 *but are independent of* ε *for* $0 < \varepsilon \leq \varepsilon_0(\nu,\eta,U_0)$*.)*

Proof. We have

$$\frac{d}{dt}\phi^2(t) = \operatorname{Re}\sum_{|\alpha|\leq 3}(D^\alpha w, D^\alpha w_t)_H, \tag{6.4}$$

where

$$D^\alpha w_t = -D^\alpha(((U+u')\cdot\nabla)w) + A_\varepsilon D^\alpha w + \varepsilon D^\alpha G + D^\alpha Q_1 + D^\alpha Q_2. \tag{6.5}$$

We bound the right-hand side of (6.4) term-by-term in a sequence of lemmas. We recall first some standard general results, specialized here to two space dimensions. Proofs of some can be found in the Appendix of [1]. See also [3]. We assume throughout that f, g are (possibly vector-valued) functions in the Sobolev space indicated, and for the chain rule we assume that Φ is in C^3. All constants are independent of f, g but may depend on Φ. The multi-index, α, satisfies $1 \leq |\alpha| \leq 3$.

Sobolev's inequality:

$$|f|_\infty \leq C\|f\|_3, \quad |Df|_\infty \leq C\|f\|_3, \quad |D^2 f|_\infty \leq C\|f\|_4. \tag{6.6}$$

Estimate based on the chain rule:

$$\|D^\alpha(\Phi\circ f)\| \leq C(1 + |f|_\infty)^{|\alpha|-1}\|f\|_{|\alpha|}. \tag{6.7}$$

Estimates based on Leibniz' rule:

$$\|D^\alpha(fg)\| \le C(|f|_\infty \|g\|_{|\alpha|} + |g|_\infty \|f\|_{|\alpha|}), \tag{6.8}$$

$$\|D^\alpha(fg) - fD^\alpha g\| \le C(|Df|_\infty \|g\|_{|\alpha|-1} + |g|_\infty \|f\|_{|\alpha|}). \tag{6.9}$$

We remark that by Lemma 4.4(a) the L_2-norms in the inequalities above can be replaced by H-norms.

LEMMA 6.2. *Under the assumptions of Lemma* 6.1 *and for* $|\alpha| \le 3$ *we have*

$$|(D^\alpha w, D^\alpha(((U+u')\cdot\nabla)w))_H| \le Ce^{-ct}\phi^2(t) + C\phi^3(t). \tag{6.10}$$

Proof. We have

$$(D^\alpha w, D^\alpha(((U+u')\cdot\nabla)w))_H = (D^\alpha w, ((U+u')\cdot\nabla)D^\alpha w)_H + T,$$

$$T = \sum_{\beta<\alpha} c_{\alpha\beta}(D^\alpha w, (((D^{\alpha-\beta}(U+u'))\cdot\nabla)D^\beta w))_H.$$

We estimate the first term using (4.25), (3.5), and Sobolev's inequality to conclude

$$|(D^\alpha w, ((U+u')\cdot\nabla)D^\alpha w)_H| \le C(\varepsilon|U+u'|_\infty + |D(U+u')|_\infty)\phi^2 \le Ce^{-ct}\phi^2 + C\phi^3.$$

To estimate T, we use the Cauchy–Schwarz inequality, (3.5), and (6.9) to conclude

$$\begin{aligned} |T| &\le C\phi\,(|D(U+u')|_\infty \|w\|_3 + |Dw|_\infty \|U+u'\|_3) \\ &\le Ce^{-ct}\phi^2 + C\phi^3, \end{aligned}$$

completing the proof. ∎

LEMMA 6.3. *Under the assumptions of Lemma* 6.1 *we have*

$$\begin{aligned} &\operatorname{Re}\sum_{|\alpha|\le 3}(D^\alpha w, A_\varepsilon D^\alpha w)_H \\ &\quad\le -2\tilde{c}_0\phi^2(t) - c_1h^2(t) + \varepsilon^4 C + \varepsilon^2\delta_0^2 C + \varepsilon^2 Ce^{-ct}\phi^2(t) + C\varepsilon^2\phi^4(t). \end{aligned} \tag{6.11}$$

Proof. Since $\hat{A}_\varepsilon(0) = 0$, Lemma 4.4(b) implies

$$\operatorname{Re}\sum_{|\alpha|\le 3}(D^\alpha w, A_\varepsilon D^\alpha w)_H \le -2\tilde{c}_0\left(\|w^c\|_H^2 + \sum_{1\le|\alpha|\le 3}\|D^\alpha w\|_H^2\right) - c_1h^2.$$

Using (5.5) to replace $\|w^c\|_H^2$ by $\|w\|_H^2$, we obtain the desired estimate. ∎

LEMMA 6.4. *Under the assumptions of Lemma* 6.1 *we have*

$$|(D^\alpha w, \varepsilon D^\alpha G)_H| \le Ce^{-ct}\phi^2(t) + \varepsilon^2 Ce^{-ct}. \tag{6.12}$$

Proof. The lemma follows directly from the Cauchy–Schwarz and Cauchy inequalities combined with (3.5). ∎

LEMMA 6.5. *Under the assumptions of Lemma* 6.1 *we have*

$$\sum_{|\alpha|\le 3}|(D^\alpha w, D^\alpha Q_1)_H| \le Ce^{-ct}(\phi^2(t)+\varepsilon^2h^2(t)) + C\phi^3(t) + C\phi^2(t)h(t) + \varepsilon C\phi(t)h^2(t). \tag{6.13}$$

Proof. Recall that

$$Q_1 = \begin{pmatrix} \tilde{Q}_1 \\ 0 \end{pmatrix},$$

$$\tilde{Q}_1 = \frac{\varepsilon w^{(3)} + \varepsilon^2 P}{1 + \varepsilon w^{(3)} + \varepsilon^2 P}(\nu \Delta u' + \eta \nabla \nabla \cdot u') - (w^{(3)} + \varepsilon P)\Pi(0, \varepsilon w^{(3)} + \varepsilon^2 P)\nabla w^{(3)}.$$

We split the quantity to be estimated into three parts:

$$\sum_{|\alpha| \le 3} (D^\alpha w, D^\alpha Q_1)_H = T_1 + T_2 + T_3,$$
$$T_1 = \sum_{|\alpha|=3} (D^\alpha u', D^\alpha \tilde{Q}_1),$$
$$T_2 = \sum_{|\alpha| \le 2} (D^\alpha u', D^\alpha \tilde{Q}_1),$$
$$T_3 = \sum_{|\alpha| \le 3} (D^\alpha w, D^\alpha Q_1)_H - (D^\alpha w, D^\alpha Q_1).$$

Integrating once by parts and applying the Cauchy–Schwarz inequality, we find

$$|T_1| \le Ch \|\tilde{Q}_1\|_2.$$

Estimating directly by Cauchy–Schwarz, we obtain

$$|T_2| \le C\phi \|\tilde{Q}_1\|_2.$$

Finally, from Lemma 4.4(c) we conclude

$$|T_3| \le \varepsilon C\phi \|\tilde{Q}_1\|_2.$$

It remains then to estimate $\|\tilde{Q}_1\|_2$. By first applying (6.8) we have

$$\|\tilde{Q}_1\|_2 \le C|\varepsilon w^{(3)} + \varepsilon^2 P|_\infty h + C|D^2 u'|_\infty \left\| \frac{\varepsilon w^{(3)} + \varepsilon^2 P}{1 + \varepsilon w^{(3)} + \varepsilon^2 P} \right\|_2$$
$$+ C|(w^{(3)} + \varepsilon P)\Pi|_\infty \phi + C|Dw^{(3)}|_\infty \|(w^{(3)} + \varepsilon P)\Pi\|_2.$$

By Sobolev's inequality and (3.5) we may bound the maximum norm terms:

$$|\varepsilon w^{(3)} + \varepsilon^2 P|_\infty \le \varepsilon C\phi + \varepsilon^2 C e^{-ct},$$

$$|(w^{(3)} + \varepsilon P)\Pi|_\infty \le C\phi + \varepsilon C e^{-ct},$$

$$|D^2 u'|_\infty \le Ch, \quad |Dw^{(3)}|_\infty \le \phi.$$

Using (6.7) in combination with Sobolev's inequality and (3.5) we obtain

$$\left\| \frac{\varepsilon w^{(3)} + \varepsilon^2 P}{1 + \varepsilon w^{(3)} + \varepsilon^2 P} \right\|_2 \le \varepsilon C(\phi + \varepsilon e^{-ct}),$$

$$\|(w^{(3)} + \varepsilon P)\Pi(0, \varepsilon w^{(3)} + \varepsilon^2 P)\|_2 \le C(\phi + \varepsilon e^{-ct}).$$

Combining these, we have

$$\|\tilde{Q}_1\|_2 \le C(\phi + \varepsilon h)(\phi + \varepsilon e^{-ct}).$$

According to the bounds on the T_j, the final estimate follows from multiplying the right-hand side by $C(\phi + h)$, yielding

$$|T_j| \le C(\phi + h)(\phi + \varepsilon h)(\phi + \varepsilon e^{-ct}).$$

Application of Cauchy's inequality completes the proof. ∎

LEMMA 6.6. *Under the assumptions of Lemma* 6.1 *we have*

$$\sum_{|\alpha|\le 3} |(D^\alpha w, D^\alpha Q_2)_H| \le Ce^{-ct}\phi^2(t) + \varepsilon^2 Ce^{-ct}h^2(t) + C\phi^2(t)h(t). \tag{6.14}$$

Proof. We directly estimate $\|Q_2\|_3$ and then use the equivalence of the H-norm to the L_2-norm in conjunction with the Cauchy–Schwarz inequality to obtain the estimate. We recall

$$Q_2 = \begin{pmatrix} -(u'\cdot\nabla)U - \varepsilon w^{(3)}F \\ -\varepsilon(u'\cdot\nabla)P - (w^{(3)} + \varepsilon P)\nabla\cdot u' \end{pmatrix}.$$

Using (6.8), (6.6), and (3.5) we obtain

$$\|Q_2\|_3 \le Ce^{-ct}\phi + (\varepsilon Ce^{-ct} + C\phi)(\phi + h) + \varepsilon C\phi\|F\|_3.$$

Now F is given by

$$F = \Pi(\varepsilon^2 P, \varepsilon w^{(3)})\nabla P + \nu(1+\varepsilon^2 P)^{-1}(1 + \varepsilon w^{(3)} + \varepsilon^2 P)^{-1}\Delta U.$$

Therefore, by (6.8), the chain rule estimate, (6.7), (6.6), and (3.5) we have

$$\begin{aligned}\|F\|_3 &\le Ce^{-ct}(\|\Pi(\varepsilon^2 P, \varepsilon w^{(3)})\|_3 + \|\nu(1+\varepsilon^2 P)^{-1}(1+\varepsilon w^{(3)} + \varepsilon^2 P)^{-1}\|_3) \\ &\le Ce^{-ct}(1 + \varepsilon\phi + \varepsilon^2 Ce^{-ct})^3 \le Ce^{-ct}.\end{aligned}$$

Combining these terms and multiplying by ϕ (by applying Cauchy–Schwarz) the bound becomes

$$Ce^{-ct}\phi^2 + \varepsilon Ce^{-ct}\phi h + C\phi^2 h.$$

Applying the Cauchy inequality to the second term yields the desired result. ∎

Substituting the estimates of Lemmas 6.2–6.6 into the right-hand side of (6.4) almost yields (6.2), with $-c_1$ multiplying h^2 (rather than $-(1/2)c_1$), and an additional term $C\phi^2 h$. The latter is approximated using Cauchy's inequality and the assumed bound on ϕ:

$$C\phi^2 h \le C\phi^3 + \frac{1}{2}c_1 h^2,$$

finally producing (6.2). Equation (6.3) follows from (3.14) and Lemma 4.4(a), thus completing the proof of Lemma 6.1. ∎

We are now in a position to prove Theorem 2.1. Assume that δ_0 and ε_0 are chosen so that $\phi(0) < 1$. Recall that the short-time existence theory of hyperbolic–parabolic

systems guarantees a maximal interval of existence, $0 \le t < T$, with $\phi(t) < 1$ and, if $T < \infty$,

$$\limsup_{t \to T^-} \phi^2(t) = 1. \tag{6.15}$$

Therefore, all-time existence is proved if we can bound $\phi^2(t) < 1$ in arbitrary intervals of existence. To that end, consider the scalar ordinary differential inequality

$$\frac{dy}{dt} \le \left(Ce^{-ct} - \tilde{c}_0\right) y + \varepsilon^2 C \left(e^{-ct} + \varepsilon^2 + \delta_0^2\right), \tag{6.16}$$

$$y(0) \le \frac{1}{2}(1 + C_1\varepsilon)\delta_0^2. \tag{6.17}$$

(For clarity, we now fix the constants C, $\tilde{c}_0$, c, C_1 to values appearing in Lemma 6.1.) We then have the following lemma.

LEMMA 6.7. *There exists K depending only on C, $\tilde{c}_0$, c, C_1 such that any solution, $y(t)$, of* (6.16)–(6.17) *satisfies*

$$y(t) < K^2(\varepsilon^2 + \delta_0^2), \quad 0 \le t < \infty, \tag{6.18}$$

$$\limsup_{t \to \infty} y(t) \le \varepsilon^2 K^2 (1 + \varepsilon^2 + \delta_0^2). \tag{6.19}$$

Proof. Set

$$\psi(t) = \int_0^t \left(Ce^{-cs} - \tilde{c}_0\right) ds = c^{-1}C(1 - e^{-ct}) - \tilde{c}_0 t \le c^{-1}C - \tilde{c}_0 t,$$

$$z(t) = e^{-\psi(t)} y(t), \quad z(0) = y(0).$$

Then

$$\frac{dz}{dt}(t) = e^{-\psi(t)} \left(\frac{dy}{dt} - (Ce^{-ct} - \tilde{c}_0) y\right) \le \varepsilon^2 C e^{-\psi(t)} \left(e^{-ct} + \varepsilon^2 + \delta_0^2\right).$$

Integrating, we obtain

$$y(t) = e^{\psi(t)} z(t) \le \frac{1}{2}(1 + C_1\varepsilon)\delta_0^2 e^{\psi(t)} + \varepsilon^2 C \int_0^t e^{\psi(t) - \psi(s)} \left(e^{-cs} + \varepsilon^2 + \delta_0^2\right) ds.$$

Note that $e^{\psi(t)}$ is uniformly bounded and decays exponentially to zero as $t \to \infty$. Moreover,

$$\int_0^t e^{\psi(t) - \psi(s)} ds \le \tilde{c}_0^{-1} e^{c^{-1}C}.$$

The inequalities (6.18) and (6.19) then follow, completing the proof. ∎

Now choose $\varepsilon_0(\nu, \eta, U_0)$ and $\delta_0(\nu, \eta, U_0)$ sufficiently small such that all previous lemmas hold and that

$$K\sqrt{\varepsilon_0^2 + \delta_0^2} \le 1, \tag{6.20}$$

$$2\left(\varepsilon_0^2 C + \varepsilon_0 C K \sqrt{\varepsilon_0^2 + \delta_0^2}\right) \le c_1, \tag{6.21}$$

$$CK\sqrt{\varepsilon_0^2 + \delta_0^2} \le \tilde{c}_0. \tag{6.22}$$

Our claim is that, given (6.20), (6.21), (6.22),

$$\phi^2(t) < K^2(\varepsilon^2 + \delta_0^2) \tag{6.23}$$

holds in any interval of existence. Suppose the contrary. Then for some time T we have (6.23) for $0 \le t < T$ and

$$\phi^2(T) = K^2(\varepsilon_0^2 + \delta_0^2).$$

However, inequalities (6.21), (6.22) combined with (6.2) imply, for $0 \le t \le T$,

$$\frac{d}{dt}\phi^2(t) \le \left(Ce^{-ct} - \tilde{c}_0\right)\phi^2(t) + \varepsilon^2 C\left(e^{-ct} + \varepsilon^2 + \delta_0^2\right).$$

Therefore (see (6.3)), ϕ^2 satisfies (6.16) and (6.17), so by (6.18)

$$\phi^2(T) < K^2(\varepsilon^2 + \delta_0^2).$$

We have thus reached a contradiction. Hence, (6.23) must hold and, by the remarks at the start of the proof, all-time existence is proved.

From (6.19) we may now conclude

$$\limsup_{t\to\infty} \phi^2(t) \le \varepsilon^2 K^2(1 + \varepsilon^2 + \delta_0^2). \tag{6.24}$$

To prove that a uniform state is attained, we derive an inequality analogous to (6.2) for

$$\frac{d}{dt}\sum_{|\alpha|\le 3} \|D^\alpha w^c\|_H^2 \equiv \phi_c^2$$

by essentially repeating the estimates in Lemmas 6.2–6.6, while being careful to keep track of terms which are bounded independent of the mean. (See, e.g., [1].) We thus obtain

$$\frac{d}{dt}\phi_c^2(t) \le \left(Ce^{-ct} - 2\bar{c}_0\right)\phi_c^2(t) + C\phi(t)\phi_c(t)\left(\phi_c(t) + e^{-ct}\right) + \varepsilon^2 Ce^{-ct}. \tag{6.25}$$

Using (5.5), we find that

$$\phi \le \phi_c + \varepsilon C(\varepsilon + \delta_0 + \phi_c^2 + \phi_c e^{-ct}). \tag{6.26}$$

Applying Cauchy's inequality and further restricting ε_0 and δ_0 if necessary, we find

$$\frac{d}{dt}\phi_c^2(t) \le \left(Ce^{-ct} - \bar{c}_0\right)\phi_c^2(t) + C\phi_c^3(t) + \varepsilon^2 Ce^{-ct}. \tag{6.27}$$

Choosing T sufficiently large and using (6.24), we finally conclude

$$\frac{d}{dt}\phi_c^2(t) \le -\frac{1}{2}\bar{c}_0\phi_c^2(t) + \varepsilon^2 Ce^{-ct}, \quad t \ge T. \tag{6.28}$$

Integrating (6.28), we find

$$\lim_{t\to\infty} \phi_c = 0,$$

at an exponential rate. This, in combination with the conservation of mass and momentum, implies

$$\lim_{t\to\infty} w = (\bar{u}'_\infty, 0)^T, \tag{6.29}$$

where

$$\bar{u}'_\infty = (2\pi)^{-2}\varepsilon^2 \int_{T^2} \rho_0 u_0 dxdy = (2\pi)^{-2} m. \tag{6.30}$$

That is, we approach the unique uniform flow field with mass and momentum equal to the initial mass and momentum. This completes the proof. □

Appendix. If entropy variations are neglected and the viscosity coefficients are assumed constant, the equations for compressible flow read

$$\begin{aligned} \rho\,(u_t + (u\cdot\nabla)u) + \nabla p &= \mu\Delta u + \xi\nabla(\nabla\cdot u), \\ \rho_t + \nabla\cdot(\rho u) &= 0, \\ p/p_* &= (\rho/\rho_*)^\gamma, \quad \gamma \geq 1\ . \end{aligned}$$

Here we have assumed a γ-gas law as equation of state for simplicity.

Let $x_*, t_*, u_*, p_*, \rho_*$ denote units of length, time, velocity, pressure, and density, respectively. We assume

$$u_* = \frac{x_*}{t_*}\ .$$

The sound speed a_* corresponding to the state ρ_*, p_* satisfies

$$\frac{dp}{d\rho}(\rho_*) = \frac{\gamma p_*}{\rho_*} = a_*^2\ ,$$

and

$$\varepsilon = \frac{u_*}{a_*}$$

is the Mach number. If we introduce dimensionless variables $\tilde{x}, \tilde{t}, \tilde{u}, \tilde{p}, \tilde{\rho}$ by $x = x_*\tilde{x}$, etc., the equations become

$$\begin{aligned} \rho\,(u_t + (u\cdot\nabla)u) + \frac{1}{\gamma\varepsilon^2}\nabla p &= \nu\Delta u + \eta\nabla(\nabla\cdot u), \\ \rho_t + \nabla\cdot(\rho u) &= 0, \\ p &= \rho^\gamma, \end{aligned}$$

where $\tilde{}$ is dropped in the notation. Here

$$\nu = \frac{\mu}{x_*\rho_* u_*}, \quad \eta = \frac{\xi}{x_*\rho_* u_*}\ .$$

Eliminating p and dividing the momentum equations by ρ, the system reads

$$\begin{aligned} u_t + (u\cdot\nabla)u + \frac{1}{\varepsilon^2}\rho^{\gamma-2}\nabla\rho &= \frac{\nu}{\rho}\Delta u + \frac{\eta}{\rho}\nabla(\nabla\cdot u), \\ \rho_t + (u\cdot\nabla)\rho + \rho\nabla\cdot u &= 0. \end{aligned}$$

Introduce a new variable $r = r(x,t)$ by

$$\rho = 1 + \varepsilon^2 r \ .$$

Then we obtain

$$u_t + (u\cdot\nabla)u + (1+\varepsilon^2 r)^{\gamma-2}\nabla r = \frac{\nu}{1+\varepsilon^2 r}\Delta u + \frac{\eta}{1+\varepsilon^2 r}\nabla(\nabla\cdot u),$$
$$\varepsilon^2(r_t + (u\cdot\nabla)r) + (1+\varepsilon^2 r)\nabla\cdot u = 0.$$

We obtain the equations (1.1), (1.2) if we write ρ instead of r.

The introduction of r can be motivated as follows. Assuming that u, ρ, and their derivatives are $\mathcal{O}(1)$, the first equation implies that ρ is constant in space, to leading order in ε. Then, if we assume $\rho = \rho_0(t) + \mathcal{O}(\varepsilon^2)$, the second equation implies $\rho_0(t) =$ constant. We take $\rho_0 = 1$. This is an assumption on the initial data and our choice of ρ_*.

REFERENCES

[1] T. Hagstrom and J. Lorenz, *All-time existence of smooth solutions to PDEs of mixed type and the invariant subspace of uniform states*, Adv. Appl. Math., 16 (1995), pp. 219–257.

[2] D. Hoff, *Continuous dependence on initial data for discontinuous solutions of the Navier–Stokes equations for one-dimensional, compressible flow*, SIAM J. Math. Anal., 27 (1996), pp. 1193–1211.

[3] S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 34 (1981), pp. 481–525.

[4] S. Klainerman and A. Majda, *Compressible and incompressible fluids*, Comm. Pure Appl. Math., 35 (1982), pp. 629–651.

[5] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, New York, 1989.

[6] H.-O. Kreiss, J. Lorenz, and M. J. Naughton, *Convergence of the solutions of the compressible to the solutions of the incompressible Navier-Stokes equations*, Adv. Appl. Math., 12 (1991), pp. 187–214.

[7] P.-L. Lions, *Existence globale de solutions pour les equations de Navier-Stokes compressibles isentropiques*, C. R. Acad. Sci. Paris, 316 (1993), pp. 1335–1340.

[8] P.-L. Lions, *Compacite des solutions des equations de Navier-Stokes compressibles isentropiques*, C. R. Acad. Sci. Paris, 317 (1993), pp. 115–120.

[9] P.-L. Lions, *Limites incompressible et acoustique pour des fluides visqueux, compressibles et isentropique*, C. R. Acad. Sci. Paris, 317 (1993), pp. 1197–1202.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 673–680, May 1998

008

THE SECOND STEKLOFF EIGENVALUE AND ENERGY DISSIPATION INEQUALITIES FOR FUNCTIONALS WITH SURFACE ENERGY*

ROBERT LIPTON†

Abstract. A functional with both bulk and interfacial surface energy is considered. It corresponds to the energy dissipated inside a two-phase electrical conductor in the presence of an electrical contact resistance at the two-phase interface. The effect of embedding a highly conducting particle into a matrix of lesser conductivity is investigated. We find the criterion that determines when the increase in surface energy matches or exceeds the reduction in bulk energy associated with the particle. This criterion is general and applies to any particle with Lipschitz continuous boundary. It is given in terms of the of the second Stekloff eigenvalue of the particle. This result provides the means for selecting energy-minimizing configurations.

Key words. Stekloff eigenvalue, heat conduction, size effects, isoperimetric inequalities

AMS subject classifications. 31B20, 35A15, 73B27

PII. S0036141096310144

1. Introduction.

We consider a suspension of electrically conducting particles embedded in a matrix with a lower electrical conductivity. The two-phase conductor fills out a domain $\Omega \subset R^3$ with Lipschitz continuous boundary $\partial\Omega$. The electric conductivity tensor associated with the particle is denoted by σ_r and that of the matrix by σ_m. Here, both conductors are assumed anisotropic, and σ_r, σ_m are given by 3×3 symmetric, positive definite matrices. The tensors satisfy the inequality $\sigma_r > \sigma_m$ in the sense of quadratic forms. We suppose that there is an interfacial contact resistance between the two phases. The contact resistance is characterized by a scalar β with dimensions of conductivity per unit length.

The region occupied by the better conductor is denoted by A_r, and the region occupied by the matrix is denoted by A_m. The interface separating them is assumed Lipschitz continuous and is denoted by Γ and $\Omega = A_r \cup A_m \cup \Gamma$. The resistivity tensor inside the composite is described by $\sigma^{-1}(x) = \sigma_r^{-1}\chi_{A_r} + \sigma_m^{-1}(1 - \chi_{A_r})$, where χ_{A_r} equals one in A_r and zero otherwise. For a prescribed current $g \in H^{-1/2}(\partial\Omega)$, such that $\int_{\partial\Omega} g ds = 0$, the thermal energy dissipated inside the composite is given by $E(A_r, g)$, where

$$E(A_r, g) = \min\{C(A_r, j) : j \in L^2(\Omega)^3, \operatorname{div} j = 0,\ j \cdot n = g \text{ on } \partial\Omega\} \tag{1.1}$$

and

$$C(A_r, j) = \int_\Omega \sigma^{-1}(x)\ j \cdot j dx + \beta^{-1} \int_\Gamma (j \cdot n)^2 ds. \tag{1.2}$$

*Received by the editors October 2, 1996; accepted for publication February 19, 1997. This research was supported by NSF grant DMS-9403866 and by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant F49620-96-1-0055. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

http://www.siam.org/journals/sima/29-3/31014.html

†Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Rd., Worcester, MA 01609 (lipton@wpi.edu).

Here div $j = 0$ holds in the sense of distributions, ds is the element of surface area, and the vector n is the unit normal pointing into the matrix phase. The first term of the functional $C(A_r, j)$ is associated with bulk energy dissipation, while the second term gives the energy dissipation at the two-phase interface. The minimizer j_{A_r} is precisely the current in the composite and is related to the potential u_{A_r} by the constituitive law: $j_{A_r} = \sigma(x)\nabla u_{A_r}$ and

$$\operatorname{div}(\sigma(x)\nabla u_{A_r}) = 0 \text{ in } A_r \cup A_m. \tag{1.3}$$

Across the interface one has

$$[j_{A_r} \cdot n] = 0 \text{ on } \Gamma, \tag{1.4}$$

and

$$j_{A_r} \cdot n_{|_2} = -\beta[u_{A_r}] \text{ on } \Gamma,\ \sigma_m \nabla u_{A_r} \cdot n = g \text{ on } \partial\Omega. \tag{1.5}$$

Here $u_{A_r} \in H^1(\Omega\backslash\Gamma)$ and $[u_{A_r}] = u_{A_r|_2} - u_{A_r|_1}$, where the subscripts indicate the side of the interface where the trace is taken. The requirement $\int_{\partial\Omega} g ds = 0$ is the solvability condition for the equation of state, and the potential u_{A_r} is determined uniquely up to a constant. To expedite the presentation we denote the subspace of all elements $g \in H^{-1/2}(\partial\Omega)$ such that $\int_{\partial\Omega} g ds = 0$ by $H^{-1/2}(\partial\Omega) \setminus R$.

The replacement of a region of matrix denoted by "Σ" with material of better conductivity amounts to a nonlocal perturbation of the functional $C(A_r, j)$. The region Σ is assumed to be compactly contained within the matrix (i.e., $\Sigma \subset A_m$ and $\partial\Sigma \cap \partial A_m = \emptyset$). The perturbed functional is written as

$$C(A_r \cup \Sigma, j) = \int_\Omega \tilde{\sigma}^{-1}(x)\ j \cdot j dx + \beta^{-1} \int_{\Gamma\cup\partial\Sigma} (j \cdot n)^2 ds, \tag{1.6}$$

where $\partial\Sigma$ is the reinforcement (or particle) boundary and

$$\tilde{\sigma}^{-1}(x) = \sigma_r^{-1}\chi_{A_r\cup\Sigma} + \sigma_m^{-1}(1 - \chi_{A_r\cup\Sigma}). \tag{1.7}$$

In this article we present the geometric criterion that determines when effects due to surface energy overcome the benefits of a highly conducting particle. This criterion is general and applies to any particle with Lipschitz continuous boundary. In order to give the criterion, we introduce the 3×3 symmetric matrix $\mathbf{R_{cr}}$ given by

$$\mathbf{R_{cr}} = \beta^{-\mathbf{1}}(\sigma_{\mathbf{m}}^{-\mathbf{1}} - \sigma_{\mathbf{r}}^{-\mathbf{1}})^{-\mathbf{1}}. \tag{1.8}$$

Here each element of $\mathbf{R_{cr}}$ has dimensions of length. This tensor provides a measure of the relative magnitude of the interfacial barrier resistance with respect to the mismatch between the resistivity tensors of the matrix and particle. For a given particle occupying the set "Σ," the geometric parameter of interest is its second Stekloff eigenvalue ρ_2. The second Stekloff eigenvalue has dimensions of conductivity per unit length and we write $\rho_2(\Sigma, \sigma_r)$ to indicate its dependence on the conductivity and geometry of the particle. When Σ has Lipschitz continuous boundary the variational formulation for the second Stekloff eigenvalue is given by

$$\rho_2(\Sigma, \sigma_r) = \min_{\operatorname{div}(\sigma_r\nabla\varphi)=0} \frac{\int_{\partial\Sigma}(\sigma_r\nabla\varphi \cdot n)^2 ds}{\int_\Sigma \sigma_r \nabla\varphi \cdot \nabla\varphi dx}; \tag{1.9}$$

cf. Kuttler and Sigillito [9] and Alessandrini and Magnanini [1]. Equality in (1.9) holds for the second Stekloff eigenfunction φ_2, where div $(\sigma_r \nabla \varphi_2) = 0$ in Σ, $\int_{\partial\Sigma} \varphi_2 ds = 0$, and

$$\sigma_r \nabla \varphi_2 \cdot n = \rho_2(\Sigma, \sigma_r)\varphi_2 \text{ on } \partial\Sigma. \tag{1.10}$$

The study of this eigenvalue problem was initiated in the work of Stekloff [17]. It is evident that the second Stekloff eigenvalue and boundary traces of the Stekloff eigenfunction correspond to the first nonzero eigenvalue and eigenfunction of the Dirichlet to Neumann map on $\partial\Sigma$.

Let $\mathrm{E}(\mathrm{A_r} \cup \Sigma, \mathrm{g})$ denote the associated energy dissipation obtained by replacing a region Σ compactly contained inside A_m with the better conductor. It is given by

$$\mathrm{E}(\mathrm{A_r} \cup \Sigma, \mathrm{g}) = \min\{C(A_r \cup \Sigma, j) : j \in L^2(\Omega)^3, \text{div } j = 0,\ j \cdot n = g \text{ on } \partial\Omega\} \tag{1.11}$$

We state the following theorem.

THEOREM 1.1 (energy dissipation inequality). *Let Σ be a set with Lipschitz continuous boundary that is compactly contained in A_m. If $\rho_2(\Sigma, \sigma_r)$ satisfies*

$$\mathbf{R_{cr}^{-1}} \leq \sigma_{\mathbf{r}}^{-\mathbf{1}} \rho_{\mathbf{2}}(\mathbf{\Sigma}, \sigma_{\mathbf{r}}), \tag{1.12}$$

then

$$E(A_r \cup \Sigma, g) \geq E(A_r, g) \tag{1.13}$$

for all $g \in H^{-1/2}(\partial\Omega) \setminus R$.

Here (1.12) holds in the sense of quadratic forms. No assumptions on the topological nature of the particle domain Σ is made. Indeed it can be a disjoint union of multiply-connected components. The proof of this theorem is provided in section 2. We emphasize that (1.13) holds for every current $g \in H^{-1/2}(\partial\Omega) \setminus R$.

When the particle is made from an isotropic conductor, one can readily compute ρ_2 for spheres and rectangular fibers; cf. Kuttler and Sigillito [9]. For starlike domains and domains with smooth boundary, isoperimetric inequalities bounding ρ_2 from below have been obtained in the work of Payne [15], Bramble and Payne [2]: see also the review article of Payne [16]. These observations are applied in section 3, where heat dissipation inequalities are given in terms of the physical dimensions of the reinforcement. Such *size effect* inequalities predict the existence of a critical particle dimension below which the particle will no longer reduce the total heat dissipated inside the composite. These results show that the size of the domain Ω must be taken into consideration. Indeed, if the domain is "too thin," then the particle will have to have dimensions below the critical value in order to fit inside it. For such domains, the addition of highly conducting particles will not reduce the energy.

Theorem 1.1 can be applied to problems of energy minimization over various classes of configurations. We consider mixtures of two isotropically conducting materials. For this case, the particle and matrix phases have scalar conductivities and we continue to denote them as σ_r and σ_m, respectively, where $\sigma_r > \sigma_m$. The admissible class is chosen to be all suspensions of spheres of conductivity σ_r suspended in a matrix of σ_m. Here we allow the suspension to contain spheres of different radii. This class of suspensions is referred to as the class of *polydisperse* suspensions of spheres. We assume that each suspension consists of a finite number of spheres and that the spheres do not intersect. It is emphasized that no lower bound is placed on the size

of the spheres appearing in the suspension. We suppose that the total amount of good conductor occupies no more than a prescribed volume fraction θ_r of the domain denoted by Ω. Theorem 4.1 shows that one needs only to consider suspensions of spheres with radii greater than or equal to $R_{cr} = \beta^{-1}(\sigma_m^{-1} - \sigma_r^{-1})^{-1}$ when looking for energy-minimizing configurations. This result rules out the appearance of fine scale mixtures of spheres (i.e., minimizing sequences of suspensions made with progressively smaller spheres). An existence proof of optimal designs within this class follows from a suitable Poincaré inequality together with the theory of Chenais [3], [4] for shape optimization problems over a restricted class of Lipschitz domains. This topic is pursued elsewhere and will appear in [10]. These results are in striking contrast to what is seen when there is perfect bonding between the two conductors. For this situation it is often the case that no optimal design exists. Instead, minimizing sequences of designs exhibit regions consisting of progressively finer mixtures of the two conductors; see Lurie and Cherkaev [13] and Murat and Tartar [14].

More generally, we consider Lipschitz domains A_r of good conductor compactly contained within the design domain Ω. As before, we place no constraints on the topological nature of the reinforcing set A_r. We show, subject to the resource constraint $\text{meas}(A_r) \leq \theta_r \text{ meas}(\Omega)$, that all energy minimizing configurations lie within a subclass of domains determined by bounds on $\rho_2(A_r, \sigma_r)$: see Theorem 4.2.

2. Energy dissipation inequalities. In this section we establish Theorem 1.1.

For any $g \in H^{-1/2}(\partial\Omega) \setminus R$ we write the difference $\Delta E = E(A_r \cup \Sigma, g) - E(A_r, g)$ as

$$\Delta E = C(A_r, \tilde{j}) - C(A_r, \hat{j}) + D(\Sigma, \tilde{j}), \tag{2.14}$$

where $\tilde{j} = \text{argmin}\{C(A_r \cup \Sigma, j)\}$, $\hat{j} = \text{argmin}\{C(A_r, j)\}$, and $D(\Sigma, \tilde{j})$ is given by

$$D(\Sigma, \tilde{j}) = \beta^{-1}\left\{\int_{\partial\Sigma} (\tilde{j} \cdot n)^2 ds - \int_\Sigma \beta(\sigma_m^{-1} - \sigma_r^{-1})\tilde{j} \cdot \tilde{j} dx\right\}. \tag{2.15}$$

Noting that the field $\tilde{j}$ is an admissible trial for the variational principle (1.1), we have

$$C(A_r, \tilde{j}) - C(A_r, \hat{j}) \geq 0. \tag{2.16}$$

Thus

$$\Delta E \geq D(\Sigma, \tilde{j}). \tag{2.17}$$

Now, the equations of state for the potential $\tilde{u} \in H^1(\Omega \backslash (\Gamma \cup \partial\Sigma))$ imply that $\tilde{j} = \sigma_r \nabla \tilde{u}$ in Σ, $[\sigma \nabla \tilde{u} \cdot n] = 0$ on $\partial\Sigma$, and $\tilde{j} \cdot n_{|_2} = \sigma_r \nabla \tilde{u} \cdot n_{|_2}$ on $\partial\Sigma$. Thus from (2.15) and (2.17) we obtain

$$\Delta \text{E} \geq \beta^{-1}\left\{\int_{\partial\Sigma} (\sigma_r \nabla \tilde{u} \cdot n)^2 ds - \int_\Sigma \beta(\sigma_m^{-1} - \sigma_r^{-1})\sigma_r \nabla \tilde{u} \cdot \sigma_r \nabla \tilde{u} dx\right\}. \tag{2.18}$$

From (1.9), it follows that

$$\int_{\partial\Sigma} (\sigma_r \nabla \varphi \cdot n)^2 ds - \rho_2(\Sigma, \sigma_r) \int_\Sigma \sigma_r \nabla \varphi \cdot \nabla \varphi dx \geq 0 \tag{2.19}$$

for all $\varphi \in H^{3/2}(\Sigma)$ such that $\text{div}(\sigma_r \nabla \varphi) = 0$ in Σ.

Comparing the right-hand side of (2.18) with (2.19), we discover that

$$\Delta E \geq 0 \tag{2.20}$$

for

$$\sigma_r \beta(\sigma_m^{-1} - \sigma_r^{-1})\sigma_r \leq \sigma_r \rho_2(\Sigma, \sigma_r), \tag{2.21}$$

and the theorem follows.

We observe that strict inequality in (2.20) follows from strict inequality in (2.21), provided that $\nabla \tilde{u}$ is not identically equal to zero on Σ.

3. The second Stekloff eigenvalue for simple shapes and size effects. The second Stekloff eigenvalue for a sphere of radius a filled with isotropic conductor σ_r is given by $\rho_2 = \sigma_r/a$. It follows immediately from Theorem 1.1 that if both conducting phases are isotropic and if Σ is a sphere of radius a, then we have the following theorem.

THEOREM 3.1 (size effect for spheres). *For any current flux $g \in H^{-1/2}(\partial\Omega) \setminus R$,*

$$E(A_r \cup \Sigma, g) \geq E(A_r, g) \tag{3.22}$$

if

$$a \leq R_{cr} = \beta^{-1}(\sigma_m^{-1} - \sigma_r^{-1})^{-1}. \tag{3.23}$$

Other size-effect theorems have been obtained in the context of effective properties for isotropic suspensions of isotropically conducting spheres in an isotropic matrix. In that context the results have focused on critical radii for monodisperse suspensions of spheres; see Lipton and Vernescu [11]. Here the critical radius is precisely R_{cr} and is that for which the conductivity of the composite equals that of the matrix.

Results involving various averages of sphere radii have been found in the context of isotropic polydisperse suspensions of spheres; see Lipton and Vernescu [12]. There it is shown that if the harmonic mean of the sphere radii lies above R_{cr}, then the effective conductivity is greater than the matrix conductivity. Moreover, the effective conductivity lies below that of the matrix when the arithmetic mean of the radii lies below R_{cr}.

For size effects in the context of isotropic dilute suspensions of spheres, see Chiew and Glandt [5]. Prediction of size effects for isotropic monodisperse suspensions of spheres, by way of micromodels such as the effective medium theory and differential effective medium theory, can be found in the work of Every, Tzou, Hasselman, and Raj [7], Hasselman and Johnson [8], and Davis and Artz [6].

More generally, we consider starlike inclusions Σ filled with isotropic conductor σ_r embedded in an isotropic matrix with conductivity σ_m. Fixing the origin inside Σ, we denote by h_m the minimum distance from the origin to a tangent plane on $\partial\Sigma$. The maximum and minimum distance from the origin to $\partial\Sigma$ are denoted by r_M and r_m, respectively. For such shapes, Bramble and Payne [2] show

$$\sigma_r^{-1}\rho_2(\Sigma, \sigma_r) \geq \frac{1}{r_M}\left[\left(\frac{r_m}{r_M}\right)^2 \frac{h_m}{r_M}\right]. \tag{3.24}$$

It is evident from (3.24) and Theorem 1.1 that we have the following size effect theorem for starlike reinforcements.

THEOREM 3.2 (size effect theorem for starlike particles). *If the reinforcement Σ is starlike with geometric parameters r_m, r_M, and h_m, then for any $g \in H^{-1/2}(\partial\Omega) \setminus R$, we have*

$$E(A_r \cup \Sigma, g) \geq E(A_r, g) \tag{3.25}$$

if

$$\left(\frac{1}{r_M} \left[\left(\frac{r_m}{r_M} \right)^2 \frac{h_m}{r_M} \right] \right)^{-1} \leq R_{cr}. \tag{3.26}$$

To fix ideas we apply this theorem to an ellipsoidal particle. Here we suppose that the half-lengths of the major and minor axes are specified by a and c, respectively. For this case Theorem 3.2 implies the following corollary.

COROLLARY 3.3 (size effect theorem for ellipsoidal particles). *Given an ellipsoidal particle Σ with major and minor axes specified by a and c, respectively, then for any current flux $g \in H^{-1/2}(\partial\Omega) \setminus R$*

$$E(A_r \cup \Sigma, g) \geq E(A_r, g) \tag{3.27}$$

if

$$a \left(\frac{a}{c} \right)^3 \leq R_{cr}. \tag{3.28}$$

We consider an ellipsoidal inclusion such that $c = a(1-\lambda)$ for $0 < \lambda < 1$. It follows from the corollary that the introduction of an ellipsoidal inclusion will not lower the energy dissipated inside the composite when a lies below $R_{cr}(1-\lambda)^3$.

4. Energy minimizing configurations. We consider the problem of minimizing the thermal energy dissipation over the class of polydisperse suspensions of spheres of good conductor immersed in a matrix of lesser conductivity. The matrix and spheres are made from isotropically conducting material with conductivities specified by σ_m and σ_r, respectively. Here the suspensions consist of a finite number of nonintersecting spheres and we assume no lower bound on the sphere radii. Denoting the ith sphere by B_i, we write $A_r = \cup B_i$. We suppose that the suspension takes up no more than a prescribed volume fraction θ_r of the total composite; i.e., $\text{meas}(A_r) \leq \theta_r$ meas(Ω). We denote this class of suspensions by $\mathcal{C}_{\theta_r}$. We consider the subclass $\mathcal{SC}_{\theta_r}$ of $\mathcal{C}_{\theta_r}$, defined to be all suspensions with minimum sphere radii greater than or equal to R_{cr}. For a prescribed heat flux $g \in H^{-1/2}(\partial\Omega) \setminus R$ on the boundary, we consider the problem

$$\min\{E(A_r, g) : A_r \in \mathcal{C}_{\theta_r}\}. \tag{4.29}$$

Theorem 4.1 follows from Theorem 3.1.

THEOREM 4.1. *If a minimizer of problem* (4.1) *exists, then it can be found in the class $\mathcal{SC}_{\theta_r}$ or $A_r = \emptyset$. Moreover, if Ω has dimensions for which $\mathcal{SC}_{\theta_r}$ is empty, then the minimum energy dissipation is given by $E(\emptyset, g)$.*

Proof. We consider any suspension in the class $\mathcal{C}_{\theta_r}$. If there exist spheres of radius less than R_{cr}, then Theorem 3.1 shows that there is no advantage to keeping them

in the suspension. When $\mathcal{SC}_{\theta_r}$ is empty, we see that no reinforcement is needed, and the minimum is attained for $A_r = \emptyset$. □

Next we consider energy minimization over a wide class of particle configurations. We suppose that σ_m and σ_r are anisotropic and let $\mathcal{CL}_{\theta_r}$ be the class of Lipschitz continuous sets A_r compactly contained inside Ω for which $\text{meas}(A_r) \leq \theta_r \text{ meas}(\Omega)$. Here we assume that A_r is the union of one or more components and we make no assumption on the topological nature of each component. For a given reinforcement set A_r, we denote its ith component by A_r^i. The subclass $\mathcal{SCL}_{\theta_r}$ of $\mathcal{CL}_{\theta_r}$ is defined to be all $A_r \in \mathcal{CL}_{\theta_r}$ for which every component A_r^i satisfies

$$\sigma_r^{-1}\rho_2(A_r^i, \sigma_r) \leq \mathbf{R_{cr}^{-1}}. \tag{4.30}$$

For $g \in H^{-1/2}(\partial\Omega) \setminus R$ we consider the problem

$$\min\{E(A_r, g) : A_r \in \mathcal{CL}_{\theta_r}\}. \tag{4.31}$$

Theorem 4.2 follows immediately from Theorem 1.1.

THEOREM 4.2. *If a minimizer of problem* (4.3) *exists, then it can be found in* $\mathcal{SCL}_{\theta_r}$ *or* $A_r = \emptyset$. *Moreover, if* Ω *has dimensions for which* $\mathcal{SCL}_{\theta_r}$ *is empty, then the minimum energy dissipation is given by* $E(\emptyset, g)$.

5. Conclusions. The second Stekloff eigenvalue associated with the reinforcement phase is shown to be a basic tool for the study of nonlocal perturbations of functionals with bulk and surface energies associated with imperfectly bonded composite conductors. The associated energy dissipation inequalities establish a means for selecting energy minimizing configurations. For the problem treated in section 4, it is found that fine scale oscillations are rendered superfluous due to the electrical contact resistance associated with the interface.

REFERENCES

[1] G. ALESSANDRINI AND R. MAGNANINI, *Elliptic equations in divergence form, geometric critical points of solutions, and Stekloff eigenfunctions*, SIAM J. Math. Anal., 25 (1994), pp. 1259–1268.

[2] J. H. BRAMBLE AND L. E. PAYNE, *Bounds in the Neumann problem for second order uniformly elliptic operators*, Pacific J. Math., 12 (1962), pp. 823–833.

[3] D. CHENAIS, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189–219.

[4] D. CHENAIS, *Homéomorphisme entre ouverts lipschitziens*, Ann. Mat. Pura Appl., 118 (1980), pp. 343–398.

[5] Y. C. CHIEW AND E. D. GLANDT, *Effective conductivity of dispersions: The effect of resistance at particle interfaces*, Chem. Engrg. Sci., 42 (1987), pp. 2677–2685.

[6] L. C. DAVIS AND B. E. ARTS, *Thermal conductivity of metal-matrix composites*, J. Appl. Phys., 77 (1995), pp. 4954–4960.

[7] A. G. EVERY, Y. TZOU, D. P. H. HASSELMAN, AND R. RAJ, *The effect of particle size on the thermal conductivity of ZnS/diamond composites*, Acta Metall. Matter, 40 (1992), pp. 123–129.

[8] D. P. H. HASSELMAN AND L. F. JOHNSON, *Effective thermal conductivity of composites with interfacial thermal barrier resistance*, J. Composite Materials, 21 (1987), pp. 508–515.

[9] J. R. KUTTLER AND V. G. SIGILLITO, *Inequalities for membrane and Stekloff eigenvalues*, J. Math. Anal. Appl., 23 (1968), pp. 148–160.

[10] R. LIPTON, *Energy minimizing configurations for mixtures of two imperfectly bonded conductors*, J. Control Cybernet., to appear.

[11] R. LIPTON AND B. VERNESCU, *Composites with imperfect interface*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 329–358.

[12] R. Lipton and B. Vernescu, *Critical radius, size effects, and inverse problems for composites with imperfect interface*, J. Appl. Phys., 79 (1996), pp. 8964–8966.

[13] K. A. Lurie and A. V. Cherkaev, *Effective characteristics of composite materials and the optimal design of structural elemets*, Uspekhi Mekaniki, Advances in Mechanics, 9 (1986), pp. 3–81; in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev and R. Kohn, eds., Birkhäuser, Basel, 1997, pp. 175–258 (in English).

[14] F. Murat and L. Tartar, *Calcul des variations et homogénéisation*, in Les Methodes de l'Homogénéisation: Théorie et Applications en Physique Coll. de la Dir. des Etudes et Recherches de Electr. del. France, Eyrolles, Paris, 1985, pp. 319–370; in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev and R. Kohn, eds., Birkhäuser, Basel, 1997, pp. 139–174 (in English).

[15] L. E. Payne, *Some isoperimetric inequalities for harmonic functions*, SIAM J. Math. Anal., 3 (1970), pp. 354–359.

[16] L. E. Payne, *Isoperimetric inequalities and their applications*, SIAM Rev., 9 (1967), pp. 453–488.

[17] A. V. Stekloff, *Sur les problémes fondamentaux en physique mathématique*, Ann. Sci. Ecole Norm. Sup., 19 (1902), pp. 455–490.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 681–696, May 1998

009

GLOBAL STABILITY IN CHEMOSTAT-TYPE EQUATIONS WITH DISTRIBUTED DELAYS*

XUE-ZHONG HE†, SHIGUI RUAN‡, AND HUAXING XIA§

Abstract. We consider a chemostat-type model in which a single species feeds on a limiting nutrient supplied at a constant rate. The model incorporates a general nutrient uptake function and two distributed (infinite) delays. The first delay models the fact that the nutrient is partially recycled after the death of the biomass by bacterial decomposition, and the second delay indicates that the growth of the species depends on the past concentration of the nutrient. By constructing appropriate Liapunov-like functionals, we obtain sufficient conditions for local and global stability of the positive equilibrium of the model. Quantitative estimates on the size of the delays for local and global stability are also obtained with the help of the Liapunov-like functionals. The technique we use in this paper may be used as well to study global stability of other types of physical models with distributed delays.

Key words. chemostat-type equations, distributed delay, Liapunov functionals, local and global stability, nutrient recycling

AMS subject classifications. 34K15, 34K20, 45J05, 92A15

PII. S0036141096311101

1. Introduction.

The effect of material (nutrient) recycling on ecosystem stability has been previously studied for closed systems (see Nisbet and Gurney [16], Nisbet, McKinstry, and Gurney [17], and Ulanowicz [22]). Powell and Richerson [18] and Nisbet and Gurney [16] regarded nutrient recycling as an instantaneous process, thus neglecting the time required to regenerate the nutrient from the dead biomass by bacterial decomposition. However, as pointed out in Whittaker [23], a delay in nutrient recycling is always present in a natural system and it increases when temperature decreases. To simulate the growth of planktonic communities of unicellular algae in the lakes, Beretta, Bischi, and Solimano [1] proposed an open system in which a single species feeds on a limiting nutrient supplied at a constant rate. They assumed that the nutrient is partially recycled after the death of the organisms and used a distributed delay to model nutrient recycling. Bischi [5] observed that the delay involved in nutrient recycling alone does not have a destabilizing effect on the equilibrium.

Evidence of delayed growth response has also been observed from chemostat experiments with microalgae *Chlamidomonas Reinhardii* even when the limiting nutrient is at undetectable small concentration (see Caperon [7]). Following Caperon [7], Ruan [19] introduced a discrete delay to the model of Beretta, Bischi, and Solimano [1] to describe the delayed growth response of the species to nutrient uptake. It is shown (see He and Ruan [11]) that the positive equilibrium is globally stable if the delays

*Received by the editors October 23, 1996; accepted for publication (in revised form) April 25, 1997.

http://www.siam.org/journals/sima/29-3/31110.html

†School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia (he_t@maths.su.oz.au).

‡Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5 (ruan@mscs.dal.ca). The research of this author was supported by the NSERC of Canada and the Petro-Canada Young Innovator Award.

§Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1 (xia@icarus.math.mcmaster.ca).

are sufficiently small. However, there is a threshold value of the discrete delay involved in the growth response; when the discrete delay is increased and passes this critical value, the equilibrium may lose its stability and a Hopf bifurcation may occur (see Ruan [19]). Recently, Beretta and Takeuchi [2–4] used an additional distributed delay to model the delayed growth response. By assuming that the response function is either a Lotka–Volterra function or a Michaelis–Menten function, they studied the global stability of the positive equilibrium.

In this paper, we consider a chemostat-type model with (distributedly) delayed growth response and (distributedly) delayed nutrient recycling, namely, a system of two retarded functional differential equations with two distributed delays. This model was first proposed and studied by Beretta and Takeuchi [2]. However, their stability results are only local. Although global stability was also considered in [2–4] for similar models with an additional instantaneous negative feedback or without delayed growth response at all, the problem is more difficult to study when delayed growth response is introduced. By constructing appropriate Liapunov-like functionals, we study both local and global stability of the positive equilibrium of the model. It turns out that the positive equilibrium can be globally asymptotically stable if the mean delays are sufficiently small, and quantitative estimates on the size of these delays can be obtained with the help of the Liapunov-like functionals. Moreover, our approach to the local stability problem is slightly different than that used by Beretta and Takeuchi [2], and we improve their local stability result.

We remark that distributed (infinite) delay equations have been used in biological modeling since the work of Volterra (see Scudo and Ziegler [20]) and they are regarded to be more realistic than discrete (finite) delay equations (see Caperon [7]). The fundamental theory and some properties such as stability, existence of periodic solutions, etc. of distributed delay equations are well understood now and are discussed in the books of Burton [6], Hale and Verduyn Lunel [10], and Hino, Murakami and Naito [12]. The monographs of Cushing [8] and MacDonald [15] give excellent descriptions of distributed delay models and study the local stability, bifurcation, and periodic solutions of these models. Although global stability of some biological models with distributed delays has been studied (see Gopalsamy [9], Kuang [14], Wolkowicz, Xia, and Ruan [24] and the references cited therein), in general global results for models involving distributed delays are hard to obtain. The reason probably is that there are few methods available in investigating the global stability of infinite delay equations. The most powerful and most important method is perhaps the Liapunov function(al) method. However, there is no general procedure to follow in constructing a desirable Liapunov function(al), and completely different forms of Liapunov function(al)s are used for different kinds of equations. In the present paper, we try to construct the Liapunov-like functionals step by step so that the idea and technique can be easily followed. We believe that our technique can be used as well to study global stability of some other types of physical models with infinite delays.

The paper is organized as follows. In section 2 we describe the model equations. Local stability is studied in section 3 and global stability is discussed in section 4. Finally, a brief discussion is carried out in section 5.

2. The model. Let $N(t)$ denote the limiting nutrient concentration and $P(t)$ denote the plankton concentration at time t. Consider the following integrodifferential equations model of plankton–nutrient interaction with delayed growth response and delayed nutrient recycling:

$$\begin{aligned}\dot{N} =& D(N^0 - N) - aU(N)P + b\gamma \int_0^\infty f(s)P(t-s)\,ds \\ \dot{P} =& P\bigg[- (\gamma + D) + c\int_0^\infty g(s)U(N(t-s))\,ds\bigg],\end{aligned} \tag{2.1}$$

with initial value conditions

$$N(\theta) = \phi_1(\theta) \geq 0, \quad P(\theta) = \phi_2(\theta) \geq 0, \quad \theta \in (-\infty, 0], \tag{2.2}$$

where $\phi_1(\theta), \phi_2(\theta) \in BC(-\infty, 0]$, the Banach space of all continuous bounded functions, and all parameters are positive constants. N^0 is the input concentration of the limiting nutrient, a is the maximum uptake rate of nutrient, $c\,(\leq a)$ is the maximum specific growth rate of plankton, $b\,(0 < b < 1)$ is the fraction of the nutrient recycled by bacterial decomposition of the dead plankton, γ is the death rate of plankton, and D is the washout rate, so $\gamma + D$ represents the total loss rate of the plankton.

The function $U(N)$ describes the nutrient uptake rate of plankton. Throughout, we assume that $U(N)$ is nonnegative, increasing, and vanishes when there is no nutrient, and there is a saturation effect when the nutrient is very abundant. That is, we assume that $U(N)$ is a continuously differentiable function defined on $[0, \infty)$ and

$$U(0) = 0, \quad \frac{dU}{dN} > 0, \quad \lim_{N\to\infty} U(N) = 1. \tag{2.3}$$

These general hypotheses are satisfied by the Michaelis–Menten function (see [21])

$$U(N) = \frac{N}{L + N},$$

where $L > 0$ is the half-saturation constant or Michaelis–Menten constant.

The delay kernels $f(s)$ and $g(s)$ are nonnegative bounded functions defined on $[0, \infty)$. $f(s)$ describes the contribution of the plankton population dead in the past to the nutrient recycled and $g(s)$ describes the delayed growth response of plankton to nutrient uptake. The presence of the distributed time delays must not affect the equilibrium values, so we normalize the kernels such that

$$\int_0^\infty f(s)ds = \int_0^\infty g(s)ds = 1.$$

As in MacDonald [15], we define the average time delays as

$$T_f = \int_0^\infty sf(s)ds, \quad T_g = \int_0^\infty sg(s)ds.$$

Note that $E_0 = (N^0, 0)$ is always an equilibrium for system (2.1), and if

$$\gamma + D < c \quad \text{and} \quad U^{-1}\bigg(\frac{\gamma + D}{c}\bigg) < N^0, \tag{2.4}$$

system (2.1) has a positive interior equilibrium $E^* = (N^*, P^*)$ with

$$N^* = U^{-1}\bigg(\frac{\gamma + D}{c}\bigg), \quad P^* = \frac{D(N^0 - N^*)}{aU(N^*) - b\gamma}. \tag{2.5}$$

Throughout, we always assume that (2.4) is satisfied and T_f and T_g are finite.

Denote by $X(t,\phi) = (N(t,\phi), P(t,\phi))$ the solution of system (2.1) satisfying the initial value conditions (2.2), where $\phi = (\phi_1, \phi_2)$. We say that the positive equilibrium $E^* = (N^*, P^*)$ of (2.1) is *(locally) stable* if for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon)$ such that $|X(t,\phi) - E^*| < \varepsilon$ for $t \geq 0$ and $\phi \in B(E^*, \delta)$, where $B(E^*, \delta)$ is an open δ-ball of E^*. E^* is said to be *(locally) asymptotically stable* if it is (locally) stable and there is a $\delta_o > 0$ such that $\phi \in B(E^*, \delta_o)$ implies $X(t,\phi) \to E^*$ as $t \to \infty$. E^* is said to be *globally asymptotically stable* if it is (locally) asymptotically stable, and for any positive solution $X(t,\phi)$ of (2.1) and (2.2), we have $X(t,\phi) \to E^*$ as $t \to \infty$.

System (2.1) was introduced and studied by Beretta and Takeuchi [2, 4]. It was shown there that all solutions of system (2.1) are nonnegative if the initial data chosen from $BC(-\infty, 0]$ are nonnegative. They also discussed stability of the positive equilibrium E^*. However, their stability results about system (2.1) are only local, and global stability results hold only for systems (similar to (2.1)) with an additional instantaneous negative feedback or without delayed growth response at all. The object of this paper is to improve their local stability result and investigate as well the global stability of the positive equilibrium.

3. Local asymptotic stability. We first study the local stability of the positive equilibrium $E^* = (N^*, P^*)$. Let

$$x_1 = N - N^*, \quad x_2 = P - P^*,$$

where $-N^* \leq x_1 < \infty$, $-P^* \leq x_2 < \infty$, and define

$$\xi(x_1) = U(N) - U(N^*) \tag{3.1}$$

so that $-U(N^*) \leq \xi(x_1) < 1 - U(N^*)$ by assumption (2.3). Then the linearized equations about E^* are

$$\begin{aligned} \dot{x}_1 &= -\big(D + aP^*U'(N^*)\big)x_1 - aU(N^*)x_2 + b\gamma \int_0^\infty f(s)x_2(t-s)ds, \\ \dot{x}_2 &= cP^*U'(N^*) \int_0^\infty g(s)x_1(t-s)ds. \end{aligned} \tag{3.2}$$

Note that the asymptotic stability of the trivial equilibrium $x_1 = x_2 = 0$ of (3.2) implies the local asymptotic stability of the positive equilibrium E^* of (2.1). For convenience, we define

$$A = D + aP^*U'(N^*), \quad B = acP^*U(N^*)U'(N^*), \quad C = b\gamma cP^*U'(N^*). \tag{3.3}$$

Then $B > C$, and system (3.2) becomes

$$\begin{aligned} \dot{x}_1 &= -Ax_1 - \frac{B}{C}b\gamma x_2 + b\gamma \int_0^\infty f(s)x_2(t-s)ds, \\ \dot{x}_2 &= \frac{C}{b\gamma} \int_0^\infty g(s)x_1(t-s)ds. \end{aligned} \tag{3.4}$$

Let $(x_1(t), x_2(t))$ be an arbitrary solution of system (3.4). We first consider the function $V_{11}(t) = x_1^2(t)$. It follows from (3.4) that

$$\begin{aligned}
\dot V_{11}(t) &= 2x_1(t)\dot x_1(t)\\
&= -2Ax_1^2(t) - 2b\gamma\tfrac{B}{C}x_1(t)x_2(t) + 2b\gamma x_1(t)x_2(t)\\
&\quad - 2b\gamma x_1(t)\int_0^t f(s)\int_{t-s}^t \dot x_2(u)\,du\,ds + I(t)\\
&= -2Ax_1^2(t) - 2b\gamma\big(\tfrac{B}{C}-1\big)x_1(t)x_2(t) + I(t)\\
&\quad - 2Cx_1(t)\int_0^t f(s)\int_{t-s}^t\int_0^\infty g(v)x_1(u-v)\,dv\,du\,ds\\
&\le -2Ax_1^2(t) - 2b\gamma\big(\tfrac{B}{C}-1\big)x_1(t)x_2(t) + I(t)\\
&\quad + C\int_0^\infty f(s)\int_{t-s}^t\int_0^\infty g(v)[x_1^2(t)+x_1^2(u-v)]\,dv\,du\,ds\\
&= -2Ax_1^2(t) - 2b\gamma\big(\tfrac{B}{C}-1\big)x_1(t)x_2(t) + CT_f x_1^2(t)\\
&\quad + C\int_0^\infty f(s)\int_{t-s}^t\int_0^\infty g(v)x_1^2(u-v)\,dv\,du\,ds + I(t),
\end{aligned}\tag{3.5}$$

where

$$I(t) = -2b\gamma x_1(t)\int_t^\infty f(s)\big(x_2(t)-x_2(t-s)\big)\,ds.\tag{3.6}$$

For technical reasons, we assume that $\int_0^\infty s^2 f(s)\,ds < \infty$. Then the function

$$V_{12}(t) = C\int_0^\infty f(s)\int_{t-s}^t\int_r^t\int_0^\infty g(v)x_1^2(u-v)\,dv\,du\,dr\,ds$$

is well defined, and by (3.5), we have

$$\begin{aligned}
\dot V_{11}(t) + \dot V_{12}(t) &\le -2Ax_1^2(t) - 2b\gamma\big(\tfrac{B}{C}-1\big)x_1(t)x_2(t) + CT_f x_1^2(t)\\
&\quad + C\int_0^\infty f(s)\int_{t-s}^t\int_0^\infty g(v)x_1^2(t-v)\,dv\,du\,ds + I(t)\\
&= -2Ax_1^2(t) - 2b\gamma\big(\tfrac{B}{C}-1\big)x_1(t)x_2(t) + CT_f x_1^2(t)\\
&\quad + CT_f\int_0^\infty g(s)x_1^2(t-s)\,ds + I(t).
\end{aligned}\tag{3.7}$$

We now consider the function

$$V_1(t) = V_{11}(t) + V_{12}(t) + CT_f\int_0^\infty g(s)\int_{t-s}^t x_1^2(u)\,du\,ds.$$

It follows from (3.7) that

$$\dot V_1(t) \le -2\big(A - CT_f\big)x_1^2(t) - 2b\gamma\big(\tfrac{B}{C}-1\big)x_1(t)x_2(t) + I(t).\tag{3.8}$$

On the other hand, by the second equation of (3.4), we have

$$\frac{d}{dt}\left[x_2 + \frac{C}{b\gamma}\int_0^\infty g(s)\int_{t-s}^t x_1(u)\,du\,ds\right] = \frac{C}{b\gamma}x_1(t).$$

Assume $\int_0^\infty s^2 g(s)ds < \infty$ and define

$$V_2(t) = \left[x_2 + \frac{C}{b\gamma}\int_0^\infty \int_{t-s}^t g(s)x_1(u)du\right]^2 + \left(\frac{C}{b\gamma}\right)^2 \int_0^\infty g(s)\int_{t-s}^t \int_v^t x_1^2(u)dudvds.$$

We find that

$$\dot{V}_2(t) \le \frac{2C}{b\gamma}x_1(t)x_2(t) + 2\left(\frac{C}{b\gamma}\right)^2 T_g x_1^2(t). \tag{3.9}$$

Therefore, for the function

$$V(t) = V_1(t) + \frac{(b\gamma)^2}{C}\left(\int_t^\infty f(s)\,ds + \frac{B}{C} - 1\right)V_2(t),$$

we have from (3.6), (3.8), and (3.9) that

$$\begin{aligned}
\dot{V}(t) \le & -2\left[A - CT_f - \left(C\int_t^\infty f(s)\,ds + B - C\right)T_g\right]x_1^2(t) \\
& + 2b\gamma x_1(t)x_2(t)\int_t^\infty f(s)\,ds + I(t) \\
= & -2\left[A - CT_f - \left(C\int_t^\infty f(s)\,ds + B - C\right)T_g\right]x_1^2(t) \\
& + 2b\gamma x_1(t)\int_t^\infty f(s)x_2(t-s)\,ds \\
\le & -2\left[A - CT_f - \left(C\int_t^\infty f(s)\,ds + B - C\right)T_g\right]x_1^2(t) \\
& + 2b\gamma|x_1(t)|\|\phi_2\|\int_t^\infty f(s)\,ds \\
\le & -2\left[A - CT_f - \left(C\int_t^\infty f(s)\,ds + B - C\right)T_g\right]x_1^2(t) \\
& + b\gamma x_1^2(t)\int_t^\infty f(s)\,ds + b\gamma\|\phi_2\|^2\int_t^\infty f(s)\,ds \\
= & -2\big[A - CT_f - (B-C)T_g\big]x_1^2(t) \\
& + \big(2CT_g + b\gamma\big)x_1^2(t)\int_t^\infty f(s)\,ds + b\gamma\|\phi_2\|^2\int_t^\infty f(s)\,ds,
\end{aligned} \tag{3.10}$$

where $\phi_2 \in BC(-\infty, 0]$ is the initial data of $x_2(t)$. By using (3.10), we now can prove the following local stability result.

THEOREM 3.1. *Assume that* $\int_0^\infty s^2 f(s)\,ds < \infty$ *and* $\int_0^\infty s^2 g(s)ds < \infty$. *If*

$$CT_f + (B-C)T_g < A, \tag{3.11}$$

then the positive equilibrium E^* *of system* (2.1) *is locally asymptotically stable.*

Proof. Let $(x_1(t), x_2(t))$ be an arbitrary solution of (3.4) with $\phi_2 \in BC(-\infty, 0]$ being the initial data for $x_2(t)$. By (3.11), we can find $\varepsilon > 0$ such that

$$Q(\varepsilon) \triangleq CT_f + (B-C)T_g + \left(CT_g + \frac{1}{2}b\gamma\right)\varepsilon < A.$$

Let $T = T(\varepsilon) > 0$ be such that $\int_t^\infty f(s)\,ds < \varepsilon$ for all $t \geq T$. It then follows from (3.10) that for all $t \geq T$,

$$\dot{V}(t) \leq -2\big(A - Q(\varepsilon)\big)x_1^2(t) + b\gamma\|\phi\|^2 \int_t^\infty f(s)\,ds.$$

Integrating $\dot{V}(t)$ from T to $t \geq T$ gives

$$\begin{aligned} x_1^2(t) + \frac{(b\gamma)^2}{C^2}\big(B - C\big)V_2(t) + 2\big(A - Q(\varepsilon)\big)\int_T^t x_1^2(s)\,ds \\ \leq V(t) + 2\big(A - Q(\varepsilon)\big)\int_T^t x_1^2(s)\,ds \\ \leq V(T) + b\gamma\|\phi_2\|^2 \int_T^t \int_s^\infty f(u)\,du ds \\ \leq V(T) + b\gamma\|\phi_2\|^2 \int_0^\infty s f(s)\,ds \\ = V(T) + b\gamma\|\phi_2\|^2 T_f < \infty. \end{aligned}$$

Therefore, $x_1(t)$ and $x_2(t)$ are bounded, and $x_1^2(t) \in L_1[0,\infty)$. By the mean value theorem and the equations in (3.4), $x_1(t)$, $x_2(t)$, and their derivative functions are thus uniformly continuous on $[0,\infty)$. Applying the Barbălat lemma (see Lemmas 1.2.2 and 1.2.3 in Gopalsamy [9]), we conclude that $\big(x_1(t), \dot{x}_1(t)\big) \to 0$ as $t \to \infty$. Therefore, from the first equation of (3.4), we must have

$$\lim_{t\to\infty}\left[-\frac{B}{C}x_2(t) + \int_0^\infty f(s)x_2(t-s)\,ds\right] = 0. \tag{3.12}$$

Let $\alpha = \liminf_{t\to\infty} x_2(t)$, $\beta = \limsup_{t\to\infty} x_2(t)$, and $\{t_m\} \uparrow \infty$ be a sequence such that $x_2(t_m) \to \beta$ as $m \to \infty$. Then $\beta < \infty$, and from (3.12) we obtain

$$\frac{B}{C}\,\beta = \lim_{m\to\infty}\int_0^\infty f(s)x_2(t_m - s)\,ds \leq \beta.$$

Since $B > C$, this implies that $\beta \leq 0$. A similar argument shows that $\alpha \geq 0$. Therefore $\alpha = \beta = 0$, and $\big(x_1(t), x_2(t)\big) \to (0,0)$ as $t \to \infty$ for every solution $\big(x_1(t), x_2(t)\big)$ of system (3.4).

Note that the characteristic equation of (3.4) is

$$\Delta(\lambda) = \lambda^2 + A\lambda + G(\lambda)\big(B - CF(\lambda)\big) = 0,$$

where

$$F(\lambda) = \int_0^\infty f(s)e^{-\lambda s}ds, \quad G(\lambda) = \int_0^\infty g(s)e^{-\lambda s}ds.$$

Since $B > C$ and every solution of (3.4) approaches zero as $t \to \infty$, $\Delta(\lambda)$ has no roots with $\mathrm{Re}(\lambda) \geq 0$. Therefore, all roots of $\Delta(\lambda)$ have negative real parts and E^* is thus locally asymptotically stable. This completes the proof. □

Remark 3.2. Beretta and Takeuchi [2, 4] observed that system (3.4) has the same characteristic equation as the following system:

$$\begin{aligned} \dot{y}_1 &= y_2, \\ \dot{y}_2 &= -Ay_2 - \int_0^\infty Bg(s)y_1(t-s)ds + C\int_0^\infty\int_0^\infty g(s-v)f(v)y_1(t-s)dvds. \end{aligned} \tag{3.13}$$

They then constructed a Liapunov functional for (3.13) and showed that the sufficient condition for the local stability of E^* is (Theorem 2 in [4])

$$CT_f + (B + C)T_g < A.$$

Our condition (3.11) improves the above condition.

4. Global asymptotic stability. To study the global stability of E^*, we consider any positive solution $X(t,\phi) = (N(t,\phi), P(t,\phi))$ of (2.1) and (2.2). We make the change of variables

$$x_1 = N - N^*, \quad x_2 = \ln(P/P^*), \tag{4.1}$$

and define $\xi(x_1)$ as in (3.1). Then

$$N = x_1 + N^*, \quad P = P^* \exp(x_2), \tag{4.2}$$

and $x_1\xi(x_1) > 0$ for any $x_1 \in [-N^*, +\infty)$; $x_1\xi(x_1) = 0$ if and only if $x_1 = 0$. Using (4.1), we rewrite system (2.1) as follows:

$$\begin{aligned}
\dot{x}_1(t) =& - Dx_1(t) - aP^* \exp(x_2(t))\xi(x_1(t)) - P^*G\big(\exp(x_2(t)) - 1\big) \\
& - b\gamma P^* \int_0^t \int_{t-s}^t f(s) \exp(x_2(u))\dot{x}_2(u) du\, ds + J(t), \\
\dot{x}_2(t) =& c \int_0^\infty g(s)\xi(x_1(t-s))ds,
\end{aligned} \tag{4.3}$$

where $G = aU(N^*) - b\gamma > 0$ and

$$J(t) = -b\gamma P^* \int_t^\infty f(s)\big[\exp(x_2(t)) - \exp(x_2(t-s))\big]\, ds. \tag{4.4}$$

Let $(x_1(t), x_2(t))$ be an arbitrary solution of system (4.3). We consider the function

$$V_{11}(t) = \int_0^{x_1(t)} \xi(s)\, ds.$$

Then upon using (4.3), we obtain

$$\begin{aligned}
\dot{V}_{11}(t) =& -Dx_1(t)\xi(x_1(t)) - aP(t)\xi^2(x_1(t)) - P^*G\big(\exp(x_2(t)) - 1\big)\xi(x_1(t)) \\
& - b\gamma P^*\xi(x_1(t)) \int_0^t \int_{t-s}^t f(s) \exp(x_2(u))\dot{x}_2(u) du ds + \xi(x_1(t))J(t) \\
=& -Dx_1(t)\xi(x_1(t)) - aP(t)\xi^2(x_1(t)) \\
& - P^*G\big(\exp(x_2(t)) - 1\big)\xi(x_1(t)) + \xi(x_1(t))J(t) \\
& - bc\gamma P^*\xi(x_1(t)) \int_0^t \int_{t-s}^t f(s) \exp(x_2(u)) \int_0^\infty g(v)\xi(x_1(u-v)) dv du ds \\
\le& -Dx_1(t)\xi(x_1(t)) - aP(t)\xi^2(x_1(t)) - P^*G\big(\exp(x_2(t)) - 1\big)\xi(x_1(t)) \\
& + \frac{1}{2}bc\gamma\bigg(\int_0^\infty \int_{t-s}^t f(s)P(u) du ds\bigg)\xi^2(x_1(t)) + \xi(x_1(t))J(t) \\
& + \frac{1}{2}bc\gamma \int_0^\infty \int_{t-s}^t f(s)P(u) \int_0^\infty g(v)\xi^2(x_1(u-v)) dv du ds.
\end{aligned} \tag{4.5}$$

Let us now consider the following two functions:

$$V_{12}(t) = \frac{1}{2}bc\gamma \int_0^\infty f(s) \int_{t-s}^t \int_w^t P(u) \int_0^\infty g(v)\xi^2(x_1(u-v))dvdudwds,$$
$$V_{13}(t) = \frac{1}{2}bc\gamma T_f \int_0^\infty g(s) \int_{t-s}^t P(u+s)\xi^2(x_1(u))duds.$$

We now assume that $\int_0^\infty s^2 f(s)\,ds < \infty$ so that V_{12} is well defined. To see the existence of V_{13}, we note from (2.1) that

$$\dot{P}(t) \le P(t)[c - (\gamma + D)] = kP(t) \tag{4.6}$$

with $k = c - (\gamma + D) > 0$. This implies

$$P(s) \le P(t)e^{k(s-t)} \qquad \text{for} \qquad s \ge t \ge 0. \tag{4.7}$$

Since $|\xi(x)| \le 1$, using (4.7), we have from the definition of V_{13} that

$$\begin{aligned} V_{13}(t) &\le \frac{1}{2}bc\gamma T_f P(t) \int_0^\infty g(s) \int_{t-s}^t e^{k(u+s-t)}duds \\ &= \frac{1}{2k}bc\gamma T_f P(t) \int_0^\infty g(s)[e^{ks} - 1]ds. \end{aligned} \tag{4.8}$$

It follows that if $\int_0^\infty g(s)[e^{ks} - 1]\,ds < \infty$, then $V_{13}(t)$ exists. Thus, for $V_1(t) = V_{11}(t) + V_{12}(t) + V_{13}(t)$, we obtain from (4.5) that

$$\begin{aligned} \dot{V}_1(t) \le{}& -Dx_1(t)\xi(x_1(t)) - aP(t)\xi^2(x_1(t)) - P^*G\big(\exp(x_2(t)) - 1\big)\xi(x_1(t)) \\ &+ \frac{1}{2}bc\gamma\xi^2(x_1(t)) \int_0^\infty \int_{t-s}^t f(s)P(u)duds \\ &+ \frac{1}{2}bc\gamma T_f\xi^2(x_1(t)) \int_0^\infty g(s)P(t+s)\,ds + \xi(x_1(t))J(t). \end{aligned} \tag{4.9}$$

Notice that from the second equation of system (2.1), we have $\dot{P}(t) \ge -(\gamma + D)P(t)$ for all $t > 0$. Thus

$$P(s) \le P(t)\exp\big[(\gamma + D)(t - s)\big] \quad \text{for } t \ge s \ge 0. \tag{4.10}$$

This implies that

$$\begin{aligned} \int_0^\infty \int_{t-s}^t f(s)P(u)\,duds &= \int_0^t \int_{t-s}^t f(s)P(u)\,duds + \int_t^\infty \int_0^t f(s)P(u)\,duds + K(t) \\ &\le P(t) \int_0^t \int_{t-s}^t f(s)\,\exp\big[(\gamma + D)(t-u)\big]duds \\ &\quad + P(t) \int_t^\infty \int_0^t f(s)\exp\big[(\gamma + D)(t-u)\big]duds + K(t) \\ &\le \frac{P(t)}{\gamma + D} \int_0^\infty f(s)\big(\exp[(\gamma + D)s] - 1\big)\,ds + K(t) \\ &= T_f^* P(t) + K(t), \end{aligned} \tag{4.11}$$

where

$$K(t) = \int_t^\infty f(s) \int_{t-s}^0 P(u)\, duds, \tag{4.12}$$

$$T_f^* = \frac{1}{\gamma + D} \int_0^\infty f(s)\big(\exp\big[(\gamma + D)s\big] - 1\big)\, ds. \tag{4.13}$$

Similarly, using (4.10), we have

$$\int_0^\infty g(s)P(t+s)ds \le P(t) \int_0^\infty g(s) \exp\big[ks\big] ds = \big(1 + kT_g^*\big)P(t), \tag{4.14}$$

where

$$T_g^* = \frac{1}{k} \int_0^\infty g(s)[e^{ks} - 1]ds < \infty. \tag{4.15}$$

Then (4.9), together with (4.11) and (4.14), implies that

$$\begin{aligned} \dot{V}_1(t) \le & -Dx_1(t)\xi(x_1(t)) - aP(t)\xi^2(x_1(t)) \\ & - P^*G\big(\exp(x_2(t)) - 1\big)\xi(x_1(t)) \\ & + \frac{1}{2}bc\gamma\big(T_f^* + T_f + kT_fT_g^*\big)P(t)\xi^2(x_1(t)) \\ & + \xi(x_1(t))J(t) + \frac{1}{2}bc\gamma\xi^2(x_1(t))K(t). \end{aligned} \tag{4.16}$$

On the other hand, from the second equation of system (4.3), we have

$$\frac{d}{dt}\left[x_2(t) + c\int_0^\infty g(s) \int_{t-s}^t \xi(x_1(u))duds\right] = c\xi(x_1(t)).$$

Let

$$y(t) = x_2(t) + c\int_0^\infty g(s) \int_{t-s}^t \xi(x_1(u))duds.$$

We define

$$V_{21}(t) = \int_0^{y(t)} \big[\exp(s) - 1\big] ds.$$

Then it follows that

$$\begin{aligned} \dot{V}_{21}(t) &= c\xi(x_1(t))\left\{ \exp\left[x_2(t) + c\int_0^\infty g(s) \int_{t-s}^t \xi(x_1(u))duds\right] - 1\right\} \\ &= c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\ &\quad + c\xi(x_1(t))\left\{ \exp\left[x_2(t) + c\int_0^\infty g(s) \int_{t-s}^t \xi(x_1(u))duds\right] - \exp(x_2(t))\right\} \\ &= c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\ &\quad + c\exp(x_2(t))\xi(x_1(t))\left\{ \exp\left[c\int_0^\infty g(s) \int_{t-s}^t \xi(x_1(u))duds\right] - 1\right\} \\ &= c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\ &\quad + c^2\exp(x_2(t))\xi(x_1(t))\exp(\alpha(t)) \int_0^\infty g(s) \int_{t-s}^t \xi(x_1(u))duds \end{aligned}$$

for some function $\alpha(t)$ between 0 and $c\int_0^\infty g(s)\int_{t-s}^t \xi(x_1(u))duds$. Since $|\xi(x_1)| \le 1$, we have

$$|\alpha(t)| \le c\int_0^\infty g(s)\int_{t-s}^t |\xi(x_1(u))|duds \le cT_g.$$

Therefore,

$$\begin{aligned}
\dot V_{21}(t) &\le c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\
&\quad + \frac{1}{2}c^2\exp(x_2(t))\exp(\alpha(t))\int_0^\infty\int_{t-s}^t g(s)\big[\xi^2(x_1(t)) + \xi^2(x_1(u))\big]duds \\
&= c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\
&\quad + \frac{1}{2}c^2\exp(x_2(t))\exp(\alpha(t))\left[T_g\xi^2(x_1(t)) + \int_0^\infty\int_{t-s}^t g(s)\xi^2(x_1(u))duds\right] \\
&\le c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\
&\quad + \frac{1}{2}c^2T_g\exp(x_2(t))\exp(cT_g)\xi^2(x_1(t)) \\
&\quad + \frac{1}{2}c^2\exp(x_2(t))\exp(cT_g)\int_0^\infty\int_{t-s}^t g(s)\xi^2(x_1(u))duds. \qquad (4.17)
\end{aligned}$$

We now define

$$V_{22}(t) = \frac{1}{2}c^2\exp(cT_g)\int_0^\infty g(s)\int_{t-s}^t \exp(x_2(v+s))\int_v^t \xi^2(x_1(u)dudvds.$$

Then, by using (4.7) and

$$\begin{aligned}
&\int_0^\infty g(s)\int_{t-s}^t P(v+s)\int_v^t du\,dv\,ds \\
&\qquad \le \int_0^\infty sg(s)\int_{t-s}^t P(v+s)\,dv\,ds \\
&\qquad \le P(t)\int_0^\infty sg(s)\int_{t-s}^t e^{k(v+s-t)}\,dv\,ds = \frac{1}{k}P(t)\int_0^\infty sg(s)[e^{ks}-1]ds,
\end{aligned}$$

we can see that, under the assumption that $\int_0^\infty sg(s)[e^{ks}-1]ds < \infty$, $V_{22}(t)$ exists. Let $V_2(t) = V_{21}(t) + V_{22}(t)$; we have from (4.17) that

$$\begin{aligned}
\dot V_2(t) &\le c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\
&\quad + \frac{1}{2}c^2T_g\exp(cT_g)\,\exp(x_2(t))\xi^2(x_1(t)) \\
&\quad + \frac{1}{2}c^2\exp(cT_g)\xi^2(x_1(t))\int_0^\infty g(s)\int_{t-s}^t \exp(x_2(u+s))duds.
\end{aligned}$$

Note that from (4.7) we have

$$\int_{t-s}^t \exp(x_2(u+s))du \le \frac{1}{k}[e^{ks}-1]\exp(x_2(t)).$$

Therefore,

$$\begin{aligned}\dot{V}_2(t) \le{}& c\xi(x_1(t))\big[\exp(x_2(t)) - 1\big] \\ &+ \frac{1}{2}c^2\exp(cT_g)\big(T_g + T_g^*\big)\exp(x_2(t))\xi^2(x_1(t)).\end{aligned} \tag{4.18}$$

We finally define the following function:

$$V(t) = V_1(t) + \frac{P^*}{c}\bigg(b\gamma\int_t^\infty f(s)\,ds + G\bigg)V_2(t).$$

It then follows from (4.16) and (4.18) that

$$\begin{aligned}\dot{V}(t) \le{}& -Dx_1(t)\xi(x_1(t)) - \frac{1}{2}\big(2a - L\big)P(t)\xi^2(x_1(t)) \\ &+ \frac{1}{2}bc\gamma P^*\exp(cT_g)\big(T_g + T_g^*\big)\exp(x_2(t))\xi^2(x_1(t))\int_t^\infty f(s)\,ds \\ &+ b\gamma P^*\xi(x_1(t))\int_t^\infty f(s)\big[\exp(x_2(t-s)) - 1\big]\,ds \\ &+ \frac{1}{2}bc\gamma\xi^2(x_1(t))K(t),\end{aligned} \tag{4.19}$$

where

$$L = bc\gamma\big(T_f^* + T_f + kT_fT_g^*\big) + cG\exp(cT_g)\big(T_g + T_g^*\big).$$

Notice that from (4.2), $P(u) = P^*\exp(\phi_2(u))$ for $u \le 0$ and $\exp(x_2(t-s)) = \exp(\phi_2(t-s))$ for $t \le s$, where $\phi_2 \in BC(-\infty, 0]$ is the initial data for $x_2(t)$. By (4.12), we have

$$\begin{aligned}&|K(t)| \le P^*\exp\big(\|\phi_2\|\big)\int_t^\infty sf(s)\,ds, \\ &\left|\int_t^\infty f(s)\big[\exp(x_2(t-s)) - 1\big]ds\right| \le \big[\exp\big(\|\phi_2\|\big) - 1\big]\int_t^\infty f(s)\,ds.\end{aligned}$$

Since $|\xi(x_1)| \le 1$, using the above inequalities, we obtain from (4.19) that

$$\begin{aligned}\dot{V}(t) \le{}& -Dx_1(t)\xi(x_1(t)) - \frac{1}{2}\big(2a - L\big)P(t)\xi^2(x_1(t)) \\ &+ \frac{1}{2}bc\gamma\exp(cT_g)\big(T_g + T_g^*\big)P(t)\xi^2(x_1(t))\int_t^\infty f(s)\,ds \\ &+ b\gamma P^*\big[\exp\big(\|\phi_2\|\big) - 1\big]\int_t^\infty f(s)\,ds \\ &+ \frac{1}{2}bc\gamma P^*\exp\big(\|\phi_2\|\big)\int_t^\infty sf(s)\,ds.\end{aligned} \tag{4.20}$$

The above analysis now leads to the following global stability result.

THEOREM 4.1. *Let $G = aU(N^*) - b\gamma$ and T_f^* and T_g^* be the constants defined in* (4.13) *and* (4.15), *respectively. Assume that $\int_0^\infty sg(s)[e^{ks} - 1]ds < \infty$ and*

$$L \triangleq bc\gamma\big(T_f^* + T_f + kT_fT_g^*\big) + cG\exp(cT_g)\big(T_g + T_g^*\big) < 2a \tag{4.21}$$

with $k = c - (\gamma + D) > 0$. Then the positive equilibrium E^* is globally asymptotically stable.

Proof. Let $(x_1(t), x_2(t))$ be an arbitrary solution of system (4.3) with $\phi_2 \in BC(-\infty, 0]$ being the initial data of $x_2(t)$. We choose $\varepsilon > 0$ such that

$$L(\varepsilon) \triangleq L + \frac{1}{2}bc\gamma \exp(cT_g)\big(T_g + T_g^*\big)\varepsilon < 2a$$

and find $T = T(\varepsilon) > 1$ such that $\int_t^\infty f(s)\,ds < \varepsilon$ for all $t \geq T$. Notice that $L < 2a$ and $\int_0^\infty s^2 g(s)[e^{ks} - 1]ds < \infty$, thus $V(t)$ is well defined and (4.20) holds. Therefore, for all $t \geq T$, we have

$$\begin{aligned}\dot{V}(t) \leq & - Dx_1(t)\xi(x_1(t)) - \frac{1}{2}\big(2a - L(\varepsilon)\big)P(t)\xi^2(x_1(t)) \\ & + \frac{1}{2}b\gamma P^*\big[(c+2)\exp\big(\|\phi_1\|\big) - 2\big]\int_t^\infty sf(s)\,ds \\ \leq & - Dx_1(t)\xi(x_1(t)) + M\int_t^\infty sf(s)\,ds,\end{aligned}$$

where $M = \frac{1}{2}b\gamma P^*\big[(c+2)\exp\big(\|\phi_1\|\big) - 2\big] > 0$. Integrating $\dot{V}(t)$ from T to $t \geq T$ now gives

$$\begin{aligned}& V_{11}(t) + \frac{P^*G}{C}V_2(t) + D\int_T^t x_1(s)\xi(x_1(s))\,ds \\ & \quad \leq V(t) + D\int_T^t x_1(s)\xi(x_1(s))\,ds \\ & \quad \leq V(T) + M\int_T^t\int_s^\infty uf(u)\,duds \\ & \quad \leq V(T) + M\int_0^\infty s^2 f(s)\,ds \\ & \quad \leq V(T) + 2(\gamma + D)MT_f^* < \infty.\end{aligned}$$

This implies that $x_1(t)$ and $x_2(t)$ are bounded, and $x_1(t)\xi(x_1(t)) \in L_1[0,\infty)$. Since ξ is uniformly continuous on $[0,\infty)$, it follows from the equations in system (4.3) and the boundedness of $x_1(t)$ and $x_2(t)$ that $x_1(t)\xi(x_1(t))$ is also uniformly continuous. Thus, by the Barbălat lemma (see Lemma 1.2.2 in [9]), $x_1(t)\xi(x_1(t)) \to 0$ as $t \to \infty$. This leads to $\lim_{t\to\infty} x_1(t) = 0$, so $\lim_{t\to\infty} N(t) = N^*$ by (4.2). Note further that $\dot{x}_1(t)$ is also uniformly continuous on $[0,\infty)$ by the equations in (4.3). Applying the Barbălat lemma (see Lemma 1.2.3 in [9]) once again gives $\lim_{t\to\infty}\dot{x}_1(t) = \lim_{t\to\infty}\dot{N}(t) = 0$. Now taking the limit $(t \to \infty)$ on both sides of (2.1), we obtain

$$\lim_{t\to\infty}\left(aU(N^*)P(t) - b\gamma\int_0^\infty f(s)P(t-s)\,ds\right) = D\big(N^0 - N^*\big). \tag{4.22}$$

Let $\alpha = \liminf_{t\to\infty} P(t)$, $\beta = \limsup_{t\to\infty} P(t)$, and $\{t_m\} \uparrow \infty$ be a sequence such that $\lim_{m\to\infty} P(t_m) = \beta$. Since $\beta < \infty$, it follows from (4.22) that

$$\begin{aligned}aU(N^*)\beta &= D\big(N^0 - N^*\big) + \lim_{m\to\infty} b\gamma\int_0^\infty f(s)P(t_m - s)\,ds \\ &\leq D\big(N^0 - N^*\big) + b\gamma\beta.\end{aligned}$$

Thus,

$$\beta \le \frac{D\big(N^0 - N^*\big)}{aU(N^*) - b\gamma} = P^*.$$

Similarly, we can show that $\alpha \ge P^*$. Therefore, $\lim_{t\to\infty} P(t) = P^*$. This proves the global attractivity of E^*.

On the other hand, we claim that $L < 2a$ implies (3.11). In fact,

$$CT_f + (B - C)T_g = cP^*U'(N^*)[b\gamma T_f + GT_g] \tag{4.23}$$

with $G = aU(N^*) - b\gamma > 0$. It is noticed that $T_f \le T_f^*$ and $T_g \le T_g^*$. Then, from $L < 2a$, we have

$$2a > bc\gamma(T_f^* + T_f) + cG(T_g + T_g^*) \ge 2c[b\gamma T_f + GT_g]$$

and hence

$$CT_f + (B - C)T_g \le P^*U'(N^*)a < A,$$

which shows that (3.11) holds. By Theorem 3.1, E^* is locally asymptotically stable. This, together with global attractivity, implies that E^* is globally asymptotically stable. □

5. Discussion. In this paper, we have considered a chemostat-type plankton model with nutrient recycling. We assumed that there are a (distributed) delay in the growth response of plankton to nutrient uptake and a (distributed) delay in nutrient recycling. We have obtained some sufficient conditions for both local and global stability of the positive equilibrium by constructing appropriate Liapunov-like functionals. It is known that the delay in the growth response of the populations to nutrient uptake can cause oscillations in population density (see Caperon [7] and Ruan [19]); our results indicate that one can still have global stability of the positive equilibrium if the delays are sufficiently small, and explicit estimates (see (3.11), (4.18), and (4.19)) on the size of these delays for local and global stability can also be obtained.

We should point out that model (2.1) and other related models have been studied by Beretta and Takeuchi [2–4], He and Ruan [11], and Kolmanovskii, Torelli, and Vermiglio [13]. For local stability, following the arguments of Kolmanovskii, Torelli, and Vermiglio [13], Beretta and Takeuchi [2, 4] chose a Liapunov functional for the equivalent system (3.13) and obtained some sufficient conditions. Our local stability conditions improve their conditions. For global stability, Beretta and Takeuchi [2, 4] considered the special cases of model (2.1) when the second delay does not appear, that is, the system (2.1) with Dirac delta function $g(s) = \delta(s)$. In [2], they assumed that the interaction between nutrient and biotic species is described by (no-delayed) Lotka–Volterra coupling (i.e., $U(N) = N$). In [4], they adopted a (no-delayed) Michaelis–Menten law (i.e., $U(N) = N/(L+K)$), which is better than the former to describe the interaction from the biological point of view. With this choice, they proved (Theorem 7 in [4]) that the positive equilibrium E^* is attractive if

$$\gamma T_f < \min\left\{\frac{1}{b}, \frac{2}{b}\sqrt{\frac{aDN^*}{c^2U(N^*)K}}\right\} \le \frac{1}{b}, \tag{5.1}$$

where $K = \max\{\beta, N^o/(1 - b\gamma T_f)\}$, $\beta = (1 + b\gamma T_f)H$, and H is the bound for initial functions. Hence, the attractivity of N^* is indeed for all the solutions whose initial

functions are bounded by H. Obviously, $\gamma T_f \to 0$ as $H \to \infty$. If we consider Beretta and Takeuchi's model in [4] as a special case of our system (2.1), then we have $T_g = T_g^* = 0$ and our global asymptotic stability condition in Theorem 4.1 becomes

$$bc\gamma[T_f^* + T_f] < 2a. \tag{5.2}$$

Condition (5.2) was also obtained in He and Ruan [11, Theorem 2.1] for the system (2.1) with $g(s) = \delta(s)$ and the general growth response $U(N)$. Rewrite (5.2) as

$$\gamma T_f + \frac{1}{2}\gamma[T_f^* - T_f] < \frac{a}{bc}. \tag{5.3}$$

Note that $c \le a$, so $a/bc \ge 1/b$. Compared with (5.1), we can see that (5.3) is more restrictive on the delay kernel function f but less restrictive on the initial functions. Since condition (5.1) depends on the bound of the initial values, basically it is not a global stability condition. Thus, our condition (5.3) complements Beretta and Takeuchi's results in [4] by providing really global stability results with a little more restriction on the delay. Furthermore, our global stability results hold for the general case when both distributed delays are present and the growth response function is a general function satisfying (2.3).

Acknowledgments. Conditions in Theorem 4.1 were improved by following the referees' suggestions. We thank both referees for their comments. We are also grateful to Wanbiao Ma for his careful reading of the original version of the paper and for his helpful comments and suggestions.

REFERENCES

[1] E. Beretta, G. I. Bischi, and F. Solimano, *Stability in chemostat equations with delayed nutrient recycling*, J. Math. Biol., 28 (1990), pp. 99–111.
[2] E. Beretta and Y. Takeuchi, *Qualitative properties of chemostat equations with time delays: Boundedness, local and global asymptotic stability*, Differential Equations Dynam. Systems, 2 (1994), pp. 19–40.
[3] E. Beretta and Y. Takeuchi, *Qualitative properties of chemostat equations with time delays* II, Differential Equations Dynam. Systems, 2 (1994), pp. 263–288.
[4] E. Beretta and Y. Takeuchi, *Global stability for chemostat equations with delayed nutrient recycling*, Nonlinear World, 1 (1994), pp. 291–306.
[5] G. I. Bischi, *Effects of time lags on transient characteristics of a nutrient cycling model*, Math. Biosci., 109 (1992), pp. 151–175.
[6] T. A. Burton, *Volterra Integral and Differential Equations*, Academic Press, New York, 1983.
[7] J. Caperon, *Time lag in population growth response of isochrysis galbana to a variable nitrate environment*, Ecology, 50 (1969), pp. 188–192.
[8] J. M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*, Springer-Verlag, Heidelberg, 1977.
[9] K. Gopalsamy, *Stability and Oscillations in Delay Differential Equations of Population Dynamics*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.
[10] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
[11] X.-Z. He and S. Ruan, *Global Stability in Chemostat-Type Plankton Models with Delayed Nutrient Recycling*, Research report 96-8, The School of Math. & Stat., University of Sydney, 1996.
[12] Y. Hino, S. Murakami, and T. Naito, *Functional Differential Equations with Infinite Delay*, Lecture Notes in Math. 1473, Springer-Verlag, New York, 1991.
[13] V. B. Kolmanovskii, L. Torelli, and R. Vermiglio, *Stability of some test equations with delay*, SIAM J. Math. Anal., 25 (1994), pp. 948–961.
[14] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Academic Press, New York, 1993.

[15] N. MacDonald, *Time Lags in Biological Models*, Springer-Verlag, Heidelberg, 1978.
[16] R. M. Nisbet and W. S. C. Gurney, *Model of material cycling in a closed ecosystem*, Nature, 264 (1976), pp. 633–635.
[17] R. M. Nisbet, J. McKinstry, and W. S. C. Gurney, *A strategic model of material cycling in a closed ecosystem*, Math. Biosci., 64 (1983), pp. 99–113.
[18] T. Powell and P. J. Richerson, *Temporal variation, spatial heterogeneity and competition for resource in plankton system: A theoretical model*, Amer. Nat., 125 (1985), pp. 431–464.
[19] S. Ruan, *The effect of delays on stability and persistence in plankton models*, Nonlinear Analysis, 24 (1995), pp. 575–585.
[20] F. M. Scudo and J. R. Ziegler, *The Golden Age of Theoretical Ecology:* 1923-1940, Lecture Notes in Biomathematics 22, Springer-Verlag, Berlin, 1978.
[21] H. L. Smith and P. Waltman, *The Theory of the Chemostat*, Cambridge University Press, Cambridge, 1994.
[22] R. E. Ulanowicz, *Mass and energy flow in closed ecosystems*, J. Theor. Biol., 34 (1972), pp. 239–253.
[23] R. H. Whittaker, *Communities and Ecosystems*, Macmillan, New York, 1975.
[24] G. S. K. Wolkowicz, H. Xia, and S. Ruan, *Competition in the chemostat: A distributed delay model and its global asymptotic behavior*, SIAM J. Appl. Math., 57 (1997), pp. 1281–1310.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 697–711, May 1998

010

RECOVERY OF SINGULARITIES OF A MULTIDIMENSIONAL SCATTERING POTENTIAL*

LASSI PÄIVÄRINTA† AND VALERI SEROV‡

Abstract. We prove that in multidimensional potential scattering the leading order singularities of the unknown potential are obtained exactly from the scattering amplitude by the linearized inversion method. The proof is based on the appropriate mapping properties of the fundamental solution in weighted L^p-spaces and on a homogeneity argument concerning the bilinear term.

Key words. Born approximation, Schrödinger scattering, inverse problem

AMS subject classifications. 35P25, 35R30

PII. S0036141096305796

Introduction. A widely applied approximate method of estimating the potential from the scattering amplitude is to use the Born approximation for the scattering solution. The obvious advantage of this is that within the Born approximation, the scattering amplitude is simply the Fourier transform of the unknown potential. The weaker the potential, the better is this approximation. But even when the potential is not weak the Fourier transform of a scattering amplitude contains essential information of the potential as was shown in [PS] and [PSS] in two and three dimensions.

The purpose of this work is to generalize the results of the articles [PS], [PSS] to arbitrary dimensions. The assumptions on the potential for the Schrödinger operator are also reduced to allow stronger singularities. We recall that in the case of less singular potentials Sun and Uhlmann [SU1], [SU2] considered related problems in two dimensions with fixed energy data, while Greenleaf and Uhlmann [GU] considered related problems in $\mathbb{R}^n$ with backscattering data. The main result of this work is that the leading order singularity of the potential is obtained exactly from the scattering amplitude by the linearized inversion method (Born approximation). For a similar result in the one-dimensional case see [So], [No], and [ST].

1. Outline and results. Let $q(x)$ be a real valued potential in $\mathbb{R}^n (n \geq 3)$ appearing in the Schrödinger operator

$$H \equiv -\Delta + q(x). \tag{1.1}$$

Our basic assumption is that the potential $q(x)$ belongs to the weighted space $L^s_\sigma(\mathbb{R}^n)$ defined by the norm

$$\|q\|_{s,\sigma} = \left(\int_{\mathbb{R}^n} (1+|x|)^{\sigma s} |q(x)|^s dx \right)^{1/s} < \infty,$$

where

$$s > 2n, \quad \sigma > 1 + \frac{n}{s}. \tag{1.2}$$

*Received by the editors June 28, 1996; accepted for publication (in revised form) February 27, 1997.

http://www.siam.org/journals/sima/29-3/30579.html

†Department of Mathematical Sciences, University of Oulu, 90570 Oulu, Finland (lassi@rieska.oulu.fi).

‡Department of Computational Mathematics and Cybernetics, Moscow State University, 119899 Moscow, Russia.

Below we also use the following notation. The space $H^t,\ t \in \mathbb{R}$, denotes the usual L^2-based and W_s^t the L^s-based Sobolev space in $\mathbb{R}^n$.

Under the above assumptions on the potential $q(x)$, the operator H is a self-adjoint operator in $L^2(\mathbb{R}^n)$. The spectrum of this operator consists of an absolutely continuous spectrum, filling out the positive real axis, and a negative discrete spectrum of finite multiplicity with zero as the only possible accumulation point. In this case, for arbitrary $k \in \mathbb{R},\ k \neq 0$, we define the scattering solutions of the homogeneous Schrödinger equation

$$(H - k^2)u(x,k) = 0 \tag{1.3}$$

to be the unique solutions of the Lippmann–Schwinger equation

$$u(x,k,\theta) = e^{ik(x,\theta)} - \int\limits_{\mathbb{R}^n} G_k^+(|x-y|)q(y)u(y,k,\theta)dy, \tag{1.4}$$

where $\theta \in S^{n-1}$ and the outgoing fundamental solution of the Helmholtz equation G_k^+ is defined as

$$G_k^+(|x|) = \frac{i}{4}\left(\frac{|k|}{2\pi|x|}\right)^{\frac{n-2}{2}} H^{(1)}_{\frac{n-2}{2}}(|k||x|), \tag{1.5}$$

where $H^{(1)}_{\frac{n-2}{2}}$ is the Hankel function of the first kind and of order $\frac{n-2}{2}$. The function $G_k^+(x-y)$ is the kernel of the integral operator $(-\Delta - k^2)^{-1}$.

The solution $u(x,k,\theta)$ for $k > 0$ of the equation (1.3) admits asymptotically as $|x| \to +\infty$ the representation

$$u(x,k,\theta) = e^{ik(x,\theta)} + C_n \frac{e^{ik|x|}k^{\frac{n-3}{2}}}{|x|^{\frac{n-1}{2}}} A(k,\theta',\theta) + o\left(\frac{1}{|x|^{\frac{n-1}{2}}}\right). \tag{1.6}$$

Here $\theta \in S^{n-1},\ \theta' = \frac{x}{|x|} \in S^{n-1},\ C_n$ is a constant depending only on the dimension n, and the scattering amplitude $A(k,\theta',\theta)$ is defined by

$$A(k,\theta',\theta) = \int\limits_{\mathbb{R}^n} e^{-ik(\theta',y)}q(y)u(y,k,\theta)dy. \tag{1.7}$$

For $k < 0$ we set

$$(k,\theta',\theta) = \overline{A(-k,\theta',\theta)}$$

to obtain a well-defined scattering amplitude $A(k,\theta',\theta)$ for all $k \in \mathbb{R},\ k \neq 0,\ \theta',\theta \in S^{n-1}$. The inverse scattering problem is to recover the potential from the knowledge of $A(k,\theta',\theta)$.

To introduce the Born inversion scheme we proceed as in [PS] and define the manifolds M_0 and M by $M_0 = \mathbb{R} \times S^{n-1},\ M = M_0 \times S^{n-1}$ and the measures $d\mu_\theta(k,\theta')$ and $d\mu(k,\theta',\theta)$ over M_0 and M, correspondingly, as

$$d\mu_\theta(k,\theta') = \frac{1}{4}|k|^{n-1}dk|\theta - \theta'|^{n-1}d\theta',$$

$$d\mu(k,\theta',\theta) = \frac{1}{|S^{n-1}|}d\theta d\mu_\theta(k,\theta'), \tag{1.8}$$

where $|S^{n-1}| = 2\pi^{\frac{n}{2}}/\Gamma(n/2)$ is the area of the unit sphere S^{n-1} and $d\theta$ and $d\theta'$ denote the usual Lebesgue measures on S^{n-1}. We shall define the equivalent of the usual inverse Fourier transform on M_0 and M as

$$\left(\mathcal{F}_{M_0}^{-1}\phi_1\right)(x) = \frac{1}{(2\pi)^n}\int\limits_{M_0} e^{-ik(\theta-\theta',x)}\phi_1(k,\theta')d\mu_\theta(k,\theta'),$$

$$\left(\mathcal{F}_{M}^{-1}\phi_2\right)(x) = \frac{1}{(2\pi)^n}\int\limits_{M} e^{-ik(\theta-\theta',x)}\phi_2(k,\theta',\theta)d\mu(k,\theta',\theta). \tag{1.9}$$

If we write $\xi = k(\theta-\theta')$ then k and θ are obtained by

$$k = \frac{|\xi|}{2(\theta,\hat{\xi})},\ \theta' = \theta - 2(\theta,\hat{\xi})\hat{\xi},$$

$$\hat{\xi} = \frac{\xi}{|\xi|},\ (\theta,\hat{\xi}) \neq 0. \tag{1.10}$$

These formulas are found useful later on.

From (1.6), Proposition 1.4, and (1.7), it follows that

$$\left(\mathcal{F}q\right)(\xi) = \lim_{k\to\infty} A(k,\theta',\theta),\ \xi = k(\theta'-\theta), \tag{1.11}$$

where $\mathcal{F}$ is the usual Fourier transform in $\mathbb{R}^n$. This fact justifies the following definition.

DEFINITION. *The inverse Born approximation $q_B^\theta(x)$ and $q_B(x)$ of the potential $q(x)$ are defined as follows:*

$$q_B^\theta(x) = \left(\mathcal{F}_{M_0}^{-1}A\right)(x),\ q_B(x) = \left(\mathcal{F}_{M}^{-1}A\right)(x). \tag{1.12}$$

In this work we will prove the following three theorems.

THEOREM 1.1. *Assume that the potential $q(x)$ belongs to $L^s_\sigma(\mathbb{R}^n)$ $(n \geq 3)$ with conditions* (1.2) *fulfilled. Then the knowledge of $q_B^\theta(x)$ with θ restricted to an $(n-2)$-dimensional semisphere defines $q(x)$ uniquely.*

THEOREM 1.2. *Under the same assumptions for $q(x)$ as in Theorem* 1.1, *there exists $\varepsilon > 0$ such that*

$$q(x) - q_B(x) \in L^{s+\varepsilon}_{loc}(\mathbb{R}^n). \tag{1.13}$$

THEOREM 1.3. *Under the same assumptions for $q(x)$ as in Theorem* 1.1,

$$\lim_{k\to+\infty} k^{n-1}\int\limits_{S^{n-1}}\int\limits_{S^{n-1}} e^{-ik(\theta-\theta',x)}A(k,\theta',\theta)d\theta d\theta' = \frac{(2\pi)^n}{\pi}\int\limits_{\mathbb{R}^n}\frac{q(y)dy}{|x-y|^{n-1}}. \tag{1.14}$$

The following new estimates for the resolvent of the operator H (cf. [S]) on the continuous spectrum play the key role in the proofs of these theorems.

PROPOSITION 1.4. *Assume that the potential $q(x)$ is in $L^{\frac{p}{p-2}}_\sigma(\mathbb{R}^n)$ $(n \geq 3)$ for $2 \leq p < \frac{2n}{n-1/2}$, $\frac{1}{p}+\frac{1}{p'} = 1$, and $\sigma > 1 + n(1-\frac{2}{p})$. Then for all $k \in \mathbb{R}$, $k \neq 0$, the limit*

$$\mathbf{G}_q := \lim_{\varepsilon\to+0}(H - k^2 - i\varepsilon)^{-1} \tag{1.15}$$

exists in the uniform operator topology from $L^{p'}_{\sigma/2}(\mathbb{R}^n)$ to $L^p_{-\sigma/2}(\mathbb{R}^n)$.

Moreover, for large $|k|$

$$\|\mathbf{G}_q f\|_{L^p_{-\sigma/2}(\mathbb{R}^n)} \le \frac{C}{|k|^\alpha}\|f\|_{L^{p'}_{\sigma/2}(\mathbb{R}^n)}, \tag{1.16}$$

where $0 < \alpha < 1 - n(1 - \frac{2}{p})$.

The following well-known resolvent equation for the integral kernel $G_q(x, y, k)$ of the operator $\mathbf{G}_q$ in (1.15) will be useful in several points of the paper:

$$G_q(x, y, k) = G_k^+(|x - y|) - \int\limits_{\mathbb{R}^n} G_k^+(|x - z|)q(z)G_q(z, y, k)dz. \tag{1.17}$$

Finally, by $\mathbf{K}$ we denote the integral operator having the kernel

$$K(x, y) = |q(x)|^{1/2}G_k^+(|x - y|)q_{1/2}(y) \tag{1.18}$$

with $q_{1/2} = |q|^{1/2}\ \operatorname{sgn} q$. From Proposition 1.4 it follows that $\mathbf{K}$ is a bounded operator in $L^2(\mathbb{R}^n)$ with the norm estimate

$$\|\mathbf{K}\| \le \frac{C}{|k|^\alpha} \tag{1.19}$$

with α as in (1.16).

2. Reconstruction of L^p-singularities. This section is devoted to the proofs of Theorems 1.1 and 1.2.

Proof of Theorem 1.1. The beginning of the proof proceeds as in [PS]. For the readers' convenience we repeat the reasoning. The rest of the proof is concentrated on the right L^p-estimates.

The definition (1.9) of $q_B^\theta(x)$ together with formula (1.7) for the scattering amplitude $A(k, \theta', \theta)$ yields

$$q_B^\theta(x) = \frac{1}{(2\pi)^n}\int\limits_{M_0} e^{-ik(\theta-\theta',x)}A(k, \theta', \theta)d\mu_\theta(k, \theta')$$

$$= \frac{1}{(2\pi)^n}\int\limits_{M_0} d\mu_\theta(k, \theta')\int\limits_{\mathbb{R}^n} e^{-ik(\theta-\theta',x-y)}q(y)v(y, k, \theta)dy, \tag{2.1}$$

where $v(y, k, \theta) = e^{-ik(y,\theta)}u(y, k, \theta)$. After making the change of variables (1.10) in (2.1) we have

$$q_B^\theta(x) = \frac{1}{(2\pi)^n}\int\limits_{\mathbb{R}^n} d\xi \int\limits_{\mathbb{R}^n} e^{-i(\xi,x-y)}q(y)v\left(y, \frac{|\xi|}{2(\theta, \hat{\xi})}, \theta\right)dy.$$

Hence the usual Fourier transform of $q_B^\theta(x)$ is simply

$$\left(\mathcal{F}q_B^\theta\right)(\xi) = \left(\mathcal{F}(q)\right)(\xi) + \int\limits_{\mathbb{R}^n} e^{i(\xi,y)}q(y)\left[v\left(y, \frac{|\xi|}{2(\theta, \hat{\xi})}, \theta\right) - 1\right]dy.$$

This allows us to get the next inequality,

$$\left|\mathcal{F}(q_B^\theta - q)(\xi)\right| \le \int\limits_{\mathbb{R}^n} |q(y)| \left| v\left(y, \frac{|\xi|}{2(\theta,\hat{\xi})}, \theta\right) - 1 \right| dy. \tag{2.2}$$

Further, from the Lippmann–Schwinger equation (1.4) and the kernel equation (1.17) we have

$$v(y,k,\theta) - 1 = \tilde{\mathbf{G}}_q(q), \tag{2.3}$$

where $\tilde{\mathbf{G}}_q$ is the integral operator with the kernel $\tilde{G}_q(x,y,k) = e^{-ik(x-y,\theta)} G_q(x,y,k)$. Because the potential $q(x)$ satisfies the conditions (1.2), it is easy to check that $q(x)$ is in $L_{\sigma/2}^{\frac{2s}{s+1}}(\mathbb{R}^n)$, where s and σ satisfy (1.2). Thus, by (2.3) and Proposition 1.4 we get

$$\|v(y,k,\theta) - 1\|_{L_{-\sigma/2}^{\frac{2s}{s-1}}(\mathbb{R}^n)} \le \frac{C}{|k|^\alpha} \|q\|_{L_{\sigma/2}^{\frac{2s}{s+1}}(\mathbb{R}^n)}, \tag{2.4}$$

where $0 < \alpha < 1 - n/s$. Finally, from (2.2) and (2.4) we get

$$\left|\mathcal{F}(q_B^\theta - q)(\xi)\right| \le C \left(\frac{|(\hat{\xi},\theta)|}{|\xi|}\right)^\alpha \|q\|^2_{L_{\sigma/2}^{\frac{2s}{s+1}}(\mathbb{R}^n)}.$$

Clearly this implies

$$\mathcal{F}(q_B^\theta - q)(\xi) = 0 \text{ for } (\hat{\xi},\theta) = 0.$$

But the data $\{(\mathcal{F}f)(\xi) | (\hat{\xi},\theta) = 0\}$ when θ runs through an $(n-2)$-dimensional semisphere is enough for the complete recovery of f. □

For the proof of Theorem 1.2, we need some new definitions and lemmas.

For f in the Schwarz space S let v be the outgoing solution of the inhomogeneous Schrödinger equation

$$(H - k^2)v = f. \tag{2.5}$$

Then from (1.15) and (1.17) it follows that for v we have the following representation:

$$v(x) = \mathbf{G}_k^+(f - q \cdot \mathbf{G}_q(f))(x), \tag{2.6}$$

where $\mathbf{G}_k^+$ and $\mathbf{G}_q$ are the integral operators with kernels $G_k^+(|x-y|)$ and $G_q(x,y,k)$, correspondingly. This representation allows us to get the following asymptotic for $|x| \to +\infty$ behavior ($k > 0$):

$$v(x,k) = C_n \frac{e^{ik|x|} k^{\frac{n-3}{2}}}{|x|^{\frac{n-1}{2}}} A_f(k,\theta') + o(|x|^{\frac{1-n}{2}}), \tag{2.7}$$

where $\theta' = \frac{x}{|x|}$ and

$$A_f(k,\theta') = \int\limits_{\mathbb{R}^n} e^{-i(\theta',y)} (f(y) - q(y)\mathbf{G}_q(f)(y)) dy.$$

Below we will employ the following useful lemma.

LEMMA 2.1 (optical lemma). *For the function $A_f(k,\theta')$ the L^2-norm can be calculated by*

$$\int\limits_{S^{n-1}} |A_f(k,\theta')|^2 d\theta' = -\frac{1}{C_n^2 k^{n-2}} \operatorname{Im} \int\limits_{\mathbb{R}^n} f(x)\overline{v}(x,k)dx, \tag{2.8}$$

where C_n is the constant from (2.7).

The proof is the same as in [PS] with obvious modifications.

The function $A_f(k,\theta')$ is often called the "far field" of the equation (2.7). It satisfies the following equality:

$$A_f(k,\theta') = ((I - q\mathbf{G}_q)f, e^{ik(\theta',y)})_{L^2(\mathbb{R}^n)} = \int\limits_{\mathbb{R}^n} f(y)\overline{u(y,k,\theta')}dy, \tag{2.9}$$

where $u(y,k,\theta')$ is the solution of the Lippmann–Schwinger equation.

If a function $f : \mathbb{R}^n \to \mathbb{C}$ is decaying fast enough so that its Fourier transform has a trace on a unit sphere, then we can define the operator $A_0(k)$ as

$$(A_0(k)f)(\theta') = \int\limits_{\mathbb{R}^n} e^{-ik(\theta',y)} f(y)dy.$$

Similarly we define $A_q(k)$ as the trace of a generalized Fourier transform

$$(A_q(k)f)(\theta') = \int\limits_{\mathbb{R}^n} f(y)\overline{u(y,k,\theta')}dy, \tag{2.10}$$

whenever it exists.

The following lemma yields sufficient estimate for the existence of the above traces as well as k-dependent norm inequalities.

LEMMA 2.2. *Let the potential $q(x)$ satisfy the conditions* (1.2). *Then $A_q(k)$ and $A_0(k)$ are well defined bounded operators from $L^{\frac{2s}{s+1}}_{\sigma/2}(\mathbb{R}^n)$ to $L^2(S^{n-1})$ with the operator norm estimates*

$$\|A_0(k)\|,\ \|A_q(k)\| \le \frac{C}{|k|^{\frac{n-2}{2}+\frac{\alpha}{2}}}, \tag{2.11}$$

where s and σ are as in (1.2) *and $0 < \alpha < 1 - \frac{n}{s}$.*

Proof. By Lemma 2.1, we have

$$\|A_q(k)f\|^2_{L^2(S^{n-1})} = \int\limits_{S^{n-1}} |A_f(k,\theta')|^2 d\theta' \le \frac{C}{|k|^{n-2}} \|v\|_{L^{\frac{2s}{s-1}}_{-\sigma/2}(\mathbb{R}^n)} \|f\|_{L^{\frac{2s}{s+1}}_{\sigma/2}(\mathbb{R}^n)}.$$

Further, because $v = \mathbf{G}_q f$, from Proposition 1.4 and (1.16) we can get the estimate

$$\|A_q(k)f\|^2_{L^2(S^{n-1})} \le \frac{C}{|k|^{n-2+\alpha}} \|f\|_{L^{\frac{2s}{s+1}}_{\sigma/2}(\mathbb{R}^n)},$$

which proves the lemma. ∎

Now let $\Phi_0(k)$ and $\Phi(k)$ be the operators, defined for $f \in L^2(S^{n-1})$ as

$$\begin{aligned} \Phi_0(k)f(x) &= |q(x)|^{1/2} \int\limits_{S^{n-1}} e^{ik(\theta,x)} f(\theta)d\theta, \\ \Phi(k)f(x) &= |q(x)|^{1/2} \int\limits_{S^{n-1}} u(x,k,\theta)f(\theta)d\theta. \end{aligned} \tag{2.12}$$

It is readily seen that

$$\begin{aligned} \Phi_0(k)f(x) &= |q(x)|^{1/2}(A_0^*(k)f)(x), \\ \Phi(k)f(x) &= |q(x)|^{1/2}(A_q^*(k)f)(x), \end{aligned} \tag{2.13}$$

where A_q^* and A_0^* are the adjoint operators for A_q and A_0.

By Lemma 2.2 and (2.13) the operators $\Phi_0(k)$ and $\Phi(k)$ are bounded from $L^2(S^{n-1})$ to $L^2(\mathbb{R}^n)$ with the norm estimates

$$\|\Phi_0(k)\|,\ \|\Phi(k)\| \le \frac{C}{|k|^{\frac{n-2+\alpha}{2}}}. \tag{2.14}$$

A repeated use of the Lippmann–Schwinger equation yields the following representation for the scattering amplitude $A(k,\theta',\theta)$:

$$\begin{aligned} A(k,\theta',\theta) = \sum_{j=0}^{m} \int\limits_{\mathbb{R}^n} e^{-ik(\theta',y)} q_{1/2}(y)\mathbf{K}^j(|q|^{1/2}e^{ik(x,\theta)})(y)dy \\ + \int\limits_{\mathbb{R}^n} e^{-ik(\theta',y)} q_{1/2}(y)\mathbf{K}^{m+1}(|q|^{1/2}u(x,k,\theta))(y)dy, \end{aligned} \tag{2.15}$$

where $\mathbf{K}$ is the integral operator with kernel (1.18). This equality can be formulated in the sense of the integral operator in $L^2(S^{n-1})$:

$$\hat{A} = \sum_{j=0}^{m} \Phi_0^*(k)\ \operatorname{sgn} q\mathbf{K}^j\Phi_0(k) + \Phi_0^*(k)\ \operatorname{sgn} q\mathbf{K}^{m+1}\Phi(k), \tag{2.16}$$

where $\Phi_0^*(k)$ is the adjoint operator for $\Phi_0(k)$. Further, if we apply these formulas to the definition (1.12) of the Born approximation $q_B(x)$, we get

$$q_B(x) = \sum_{j=0}^{m} \mathcal{F}_M^{-1}[\Phi_0^*(k)\ \operatorname{sgn} q\mathbf{K}^j\Phi_0(k)] + \mathcal{F}_M^{-1}[\Phi_0^*(k)\ \operatorname{sgn} q\mathbf{K}^{m+1}\Phi(k)], \tag{2.17}$$

where the inverse Fourier transform is applied on the integral kernel of the corresponding integral operator. Let's rewrite the formula (2.17) as

$$q_B(x) = \sum_{j=0}^{m} q_j(x) + \tilde{q}_{m+1}(x). \tag{2.18}$$

By using this notation we have the following lemma.

LEMMA 2.3. *For any $j \ge 1$, $q_j(x)$ and $\tilde{q}_j x$ belong to the Sobolev space $H^t(\mathbb{R}^n)$ for any $t < \alpha(j+1) + \frac{n}{2} - 2$, where $0 < \alpha < 1 - \frac{n}{s}$.*

Proof. By the definition of the norm in the Sobolev space H^t and after the change of variables (1.10) we obtain

$$\|q_j\|^2_{H^t} = \|(1+|\xi|^2)^{t/2}\mathcal{F}(q_j)(\xi)\|^2_{L^2(\mathbb{R}^n)}$$

$$= \frac{1}{|S^{n-1}|}\int\limits_{\mathbb{R}^n}(1+|\xi|^2)^t d\xi \left| \int\limits_{S^{n-1}} [\Phi_0^*\ \operatorname{sgn} q\mathbf{K}^j\Phi_0] \left(\frac{|\xi|}{2(\theta,\hat{\xi})}, \theta - 2(\hat{\xi},\theta)\hat{\xi},\theta\right) d\theta \right|^2$$

$$\leq C\int\limits_0^\infty |k|^{n-1}(1+k^2)^t \|(\Phi_0^*\ \operatorname{sgn} q\mathbf{K}^j\Phi_0)(1)\|^2_{L^2(S^{n-1})}dk, \tag{2.19}$$

where $[\Phi_0^*\ \operatorname{sgn} q\mathbf{K}^j\Phi_0]$ denote the kernel of the integral operator inside and the integral operator applied to the function $f\equiv 1$ on S^{n-1} is denoted by $(\Phi_0^*\ \operatorname{sgn} q\mathbf{K}^j\Phi_0)(1)$. From estimates (1.18) and (2.14) we have, for $|k|>1$,

$$\|(\Phi_0^*\ \operatorname{sgn} q\mathbf{K}^j\Phi_0)(1)\|^2_{L^2(S^{n-1})} \leq C\|\Phi_0^*\|^2\|\Phi_0\|^2\|\mathbf{K}\|^{2j} \leq \frac{C}{|k|^{2\alpha(j+1)+2(n-2)}}.$$

Hence, for $q_j(x)$ we obtain the following estimate:

$$\|q_j\|_{H^t} \leq C\left(1+\int\limits_{|k|>1}\frac{|k|^{n-1+2t}d|k|}{|k|^{2\alpha(j+1)+2(n-2)}}\right)^{1/2}.$$

We leave it for the readers to check that the proof goes through for $\tilde{q}_j$ with obvious changes. Lemma 2.3 is thus proved. □

It is also easy to check that $q_0(x)$ in (2.18) is simply the potential $q(x)$ and hence we can rewrite (2.18) as

$$q_B(x) - q(x) = q_1(x) + \sum_{j=2}^m q_j(x) + \tilde{q}_{m+1}(x).$$

Here the first bilinear term $q_1(x)$ has the form

$$q_1(x) = \int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} F(y-x, z-x)q(y)q(z)dydz = (F*Q)(x,x),$$

where $Q = q\otimes q$ and F is a tempered distribution in $\mathbb{R}^{2n}$, given by

$$F(y,z) = \int\limits_{\mathbb{R}}\int\limits_{S^{n-1}}\int\limits_{S^{n-1}} e^{ik(\theta,y)-ik(\theta',z)}(|k||\theta-\theta'|)^{n-1}G_k^+(|y-z|)d\theta d\theta' dk. \tag{2.20}$$

LEMMA 2.4. *The Fourier transform $\hat{F}$ of the function F from* (2.20) *is homogeneous of degree -2 and the first nonlinear term $q_1(x)$ belongs to the Sobolev space W^1_s, where s satisfies the condition* (1.2).

Proof. See the Appendix.

Now we are ready to give the proof of Theorem 2.2.

Proof of Theorem 1.2. By Lemma 2.3 and (2.19), we obtain that

$$q_B(x) - q(x) - q_1(x) \in W_2^t(\mathbb{R}^n) = H^t \tag{2.21}$$

for any $t < 1 + \frac{n}{2} - \frac{3n}{s}$, where $s > 2n$. From the Sobolev embedding theorem we have

$$H^t \subset W_s^{t-n(1/2-1/s)}(\mathbb{R}^n).$$

But the number $t - n/2 + n/s$ can be chosen strictly positive since $1 + n/2 - 3n/s - n/2 + n/s = 1 - 2n/s > 0$ and $s > 2n$. This implies the existence of $\varepsilon > 0$ such that

$$q_B(x) - q(x) - q_1(x) \in L_{loc}^{s+\varepsilon}(\mathbb{R}^n). \tag{2.22}$$

By Lemma 2.4 we are through. □

3. Proof of Theorem 1.3. Theorem 1.3 generalizes the Saito formula to the case of singular potentials (cf. [Sa]). Again, Proposition 1.4 plays the main role for proving this theorem.

Proof of Theorem 1.3. By definition (1.7) of the scattering amplitude we have

$$\begin{aligned} & k^{n-1} \int\limits_{S^{n-1}} \int\limits_{S^{n-1}} A(k, \theta', \theta) e^{-ik(\theta-\theta', x)} d\theta d\theta' \\ & = k^{n-1} \int\limits_{\mathbb{R}^n} q(y) dy \int\limits_{S^{n-1}} \int\limits_{S^{n-1}} e^{ik(\theta-\theta', y-x)} d\theta d\theta' \\ & + k^{n-1} \int\limits_{\mathbb{R}^n} q(y) dy \int\limits_{S^{n-1}} \int\limits_{S^{n-1}} e^{-ik(\theta', y)} R(y, k, \theta) e^{-ik(\theta-\theta', x)} d\theta d\theta', \end{aligned} \tag{3.1}$$

where the function $R(y, k, \theta)$ is given by

$$R(y, k, \theta) = -\int\limits_{\mathbb{R}^n} G_k^+(|y - z|) q(z) u(z, k, \theta) dz. \tag{3.2}$$

Denote the right side of (3.1) as $I_1 + I_2$. Since

$$\int\limits_{S^{n-1}} \int\limits_{S^{n-1}} e^{ik(\theta-\theta', y-x)} d\theta d\theta' = \left| \int\limits_{S^{n-1}} e^{ik(\theta, y-x) d\theta} \right|^2$$

the integral I_1 from (3.1) rewrites as

$$I_1 = k^{n-1} \int\limits_{\mathbb{R}^n} q(y) dy \left| \int\limits_{S^{n-1}} e^{ik|y-x|(\theta, \omega)} d\theta \right|^2, \tag{3.3}$$

where $w = (y - x)/|y - x|$.

By the equality (cf. [W])

$$\int\limits_{S^{n-1}} e^{ik|y-x|(\theta, \omega)} d\theta = \frac{2\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} \int\limits_0^\pi e^{ik|y-x| \cos\phi} (\sin\phi)^{n-2} d\phi,$$

we obtain

$$I_1 = (2\pi)^n k \int\limits_{\mathbb{R}^n} \frac{q(y)}{|x-y|^{n-2}} J^2_{\frac{n-2}{2}}(k|x-y|)dy. \tag{3.4}$$

We consider the cases $k|x-y|<1$ and $k|x-y|>1$ separately. In the first case we see by using Hölder's inequality that the integral over $\{y:\ k|x-y|<1\}$ can be estimated by

$$C_n k \int\limits_{|x-y|<\frac{1}{k}} \frac{|q(y)|(k|x-y|)^{n-2}}{|x-y|^{n-2}}dy \le C_n k^{\frac{n}{s}-1}\left(\int\limits_{|x-y|<\frac{1}{k}} |q(y)|^s dy\right)^{\frac{1}{s}}, \tag{3.5}$$

where $s>2n$. This means that for every fixed x the integral over $\{y:\ k|x-y|<1\}$ approaches zero as $k\to+\infty$. Hence we need only to estimate the integral (3.3) over $\{y:\ k|x-y|>1\}$. Here we have

$$\begin{aligned}
&(2\pi)^n k \int\limits_{|x-y|>\frac{1}{k}} \frac{|q(y)|}{|x-y|^{n-2}}\left[\sqrt{\frac{2}{\pi k|x-y|}}\cos\left(k|x-y|-\frac{\pi n}{4}+\frac{\pi}{4}\right)\right.\\
&\left.+O\left(\frac{1}{(k|x-y|)^{\frac{3}{2}}}\right)\right]^2 dy\\
&= (2\pi)^n k \int\limits_{|x-y|>\frac{1}{k}} \frac{|q(y)|}{|x-y|^{n-2}}\left[\frac{2}{\pi k|x-y|}\cos^2(k|x-y|+\phi_n)+O\left(\frac{1}{(k|x-y|)^2}\right)\right]dy\\
&= 2^n\pi^{n-1}\int\limits_{|x-y|>\frac{1}{k}} \frac{|q(y)|}{|x-y|^{n-1}} + 2^n\pi^{n-1}\int\limits_{|x-y|>\frac{1}{k}} \frac{|q(y)|}{|x-y|^{n-1}}\cos(2k|x-y|+2\phi_n)dy\\
&+\frac{O(1)}{k^{1-\delta}}\int\limits_{|x-y|>\frac{1}{k}} \frac{|q(y)|}{|x-y|^{N-\delta}},
\end{aligned} \tag{3.6}$$

with $0<\delta<1$. By Sobolev's inequality we have that the L^1-norm of the function $q(y)|x-y|^{n-1}$ is uniformly bounded with respect to x. Hence, it follows from the Riemann–Lebesque lemma that the second summand in the right-hand side of (3.6) approaches zero uniformly with respect to x. The same is true for the third term and thus we have

$$\lim_{k\to+\infty} I_1 = 2^n\pi^{n-1}\int\limits_{\mathbb{R}^n}\frac{q(y)dy}{|x-y|^{n-1}}. \tag{3.7}$$

Further, for the function $R(y,k,\theta)$ defined by (3.2) we have

$$R(y,k,\theta) = \mathbf{G}_q\left(qe^{ik(\theta,z)}\right) \tag{3.8}$$

and hence for the integral I_2 we obtain the following presentation:

$$I_2 = -k^{n-1}\int\limits_{\mathbb{R}^n} q(y)dy \int\limits_{S^{n-1}} e^{ik(\theta',x-y)}d\theta \mathbf{G}_q\left(q(z)\int\limits_{S^{n-1}} e^{ik(\theta,z-x)}d\theta\right)$$

$$= -k^{n-1}(2\pi)^{n-1}\int\limits_{\mathbb{R}^n} q(y)\frac{J_{\frac{n-2}{2}}(k|x-y|)}{(k|x-y|)^{\frac{n-2}{2}}}\mathbf{G}_q\left(q(z)\frac{J_{\frac{n-2}{2}}(k|z-x|)}{(k|x-z|)^{\frac{n-2}{2}}}\right)dy$$

$$= (2\pi)^n k\int\limits_{\mathbb{R}^n} q_{1/2}(y)\frac{J_{\frac{n-2}{2}}(k|x-y|)}{(k|x-y|)^{\frac{n-2}{2}}}|q(y)|^{1/2}\mathbf{K}\left(|q(z)|^{1/2}\frac{J_{\frac{n-2}{2}}(k|z-x|)}{(|z-x|)^{\frac{n-2}{2}}}dy\right), \tag{3.9}$$

where $\mathbf{K}$ is the integral operator with the kernel (cf. (1.18)):

$$K(x,y) = -|q(x)|^{1/2}G_q(x,y,k)q_{1/2}(y).$$

From Proposition 1.4 and Hölder's inequality it follows that

$$|I_2| \le (2\pi)^n k\int\limits_{\mathbb{R}^n} |q(y)|\frac{J^2_{\frac{n-2}{2}}(k|x-y|)}{|x-y|^{n-2}}dy\|\mathbf{K}\|_{L^2\to L^2} \le \frac{C}{|k|^\alpha}.$$

This proves the theorem. □

Remark. It is easy to see that the result of Theorem 1.3 in the three-dimensional case can be applied to any potential $q(x)$ which satisfies the following conditions:

1. $q \in L^p_{loc}(\mathbb{R}^3)$ for some $p > 3/2$;
2. $|q(x)| \le \frac{C}{|x|^{2+\delta}}$ for $|x| \to +\infty$, where $\delta > 0$ and fixed.

COROLLARY 3.1. *Assume that the potential $q(x)$ satisfies the same assumptions as in Theorem* 1.3. *Then the following formula is true:*

$$q(x) = \lim_{k\to+\infty}\frac{\Gamma\left(\frac{n-1}{2}\right)k^n}{2^{n+1}\pi^{\frac{3n-1}{2}}}\int\limits_{S^{n-1}}\int\limits_{S^{n-1}} A(k,\theta',\theta)|\theta-\theta'|e^{-ik(\theta-\theta',x)}d\theta d\theta'. \tag{3.10}$$

Proof. Denote the left-hand side of (1.14) by $f(x)$. Then the Fourier transform of f is given by

$$\hat{f}(\xi) = 2^n\pi^{n-1}\hat{q}(\xi)\int_{\mathbb{R}^n}|x|^{1-n}e^{ix\xi}dx. \tag{3.11}$$

Thus

$$q(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{2^{n+1}\pi^{\frac{3n-1}{2}}}\mathcal{F}^{-1}(|\xi|\hat{f}(\xi))(x) \tag{3.12}$$

with $\mathcal{F}^{-1}$ the inverse Fourier transform. Further from (3.12) and (1.14) we get

$$q(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{2^{n+1}\pi^{\frac{3n-1}{2}}}\lim_{k\to+\infty}k^{n-1}\int\limits_{S^{n-1}}\int\limits_{S^{n-1}} A(k,\theta',\theta)k|\theta-\theta'|F^{-1}\left(\delta(\xi-k(\theta-\theta'))\right)(x)d\theta d\theta'$$

$$= \frac{\Gamma\left(\frac{n-1}{2}\right)}{2^{n+1}\pi^{\frac{3n-1}{2}}}\lim_{k\to+\infty}k^{n}\int\limits_{S^{n-1}}\int\limits_{S^{n-1}} A(k,\theta',\theta)|\theta-\theta'|e^{-ik(\theta-\theta',x)}d\theta d\theta'. \qquad \square$$

Remark. A related theorem for measure potentials is proved in [F].

4. Appendix.

Proof of Lemma 2.4. From (2.20) it follows that the function $F(y,z)$ has the representation

$$F(y,z) = C_n \int\limits_k \int\limits_{\theta'} \int\limits_{\theta} e^{ik(\theta,z)-ik(\theta',y)} G_k^+(|z-y|)\,(|k||\theta-\theta'|)^{n-1}\,dk d\theta' d\theta. \tag{4.1}$$

For the $2n$-dimensional Fourier transform of the function $F(y,z)$ we obtain

$$\hat{F}(\xi_1,\xi_2) \tag{4.2}$$

$$= \int\limits_{\mathbb{R}} \int\limits_{S^{n-1}} \int\limits_{S^{n-1}} (|k||\theta-\theta'|)^{n-1}\,dk d\theta' d\theta \int\limits_{\mathbb{R}^n} e^{-i(k(\theta'-\theta)+\xi_1+\xi_2,y)} dy \int\limits_{\mathbb{R}^n} G_k^+(|s|) e^{-i(\xi_2-k\theta,s)} ds.$$

We approximate $\hat{F}$ by $\hat{F}_\varepsilon$ given by

$$\hat{F}_\varepsilon(\xi_1,\xi_2) = \int\limits_k \int\limits_{\theta'} \int\limits_{\theta} (|k||\theta-\theta'|)^{n-1}\,dk d\theta' d\theta \int\limits_{\mathbb{R}^n} e^{-i(k(\theta'-\theta)+\xi_1+\xi_2,y)} \frac{1}{\xi_2^2 - 2k(\theta,\xi_2) - i\varepsilon} dy. \tag{4.3}$$

Here we used the fact that

$$\hat{G_k^+}(\eta) = \lim_{\varepsilon\to+0} \frac{1}{|\eta|^2 - k^2 - i\varepsilon}.$$

By (4.3) we have

$$\hat{F}_\varepsilon(\xi_1,\xi_2) = C_n \int\limits_{S^{n-1}} \frac{g\theta'}{f\theta' - i\varepsilon g\theta'} d\theta', \tag{4.4}$$

where $g = \xi_1+\xi_2$ and $f = \xi_1^2(\xi_1+\xi_2) - \xi_1|\xi_1+\xi_2|^2$. Suppose T is the two-dimensional plane spanned by f and g (for ξ_1 and ξ_2 fixed):

$$T = \mathrm{span}\{f,g\}.$$

We denote by S^1 the circle

$$S^1 = T \cap S^{n-1}.$$

Then we claim

$$\hat{F}_\varepsilon(\xi_1,\xi_2) = C_n \int\limits_{S^1} \frac{g\omega}{f\omega - i\varepsilon g\omega} d\omega, \tag{4.5}$$

where the constant C_n depends only on n and not on ξ_1 and ξ_2. To prove (4.5) we introduce the usual n-dimensional polar coordinates ($n \geq 3$) :

$$\begin{cases} x_1 = \rho\cos\phi_1, \\ x_2 = \rho\sin\phi_1\cos\phi_2, \\ \dots \\ x_n = \rho\sin\phi_1\sin\phi_2\cdots\sin\phi_{n-1}, \end{cases}$$

$$\left|\frac{\partial x}{\partial\phi}\right| = \rho^{n-1}\sin^{n-2}\phi_1\cdots\sin\phi_{n-2},$$

where we have chosen the coordinate system so that x_1 and x_2 span the plane T. Of course, this is not the original coordinate system but we use it only to prove (4.5) from (4.4).

Clearly, from (4.4) we get for any Φ

$$\int_{S^{n-1}} \Phi(x_1, x_2) d\theta = C \int_0^{\pi} \int_0^{2\pi} \Phi(\cos\phi_1, \sin\phi_1 \cos\phi_2) \sin^{n-2}\phi_1 \sin^{n-3}\phi_2 d\phi_2 d\phi_1, \tag{4.6}$$

where again $n \geq 3$. From $x_1 = \cos\phi_1,\ x_2 = \sin\phi_1 \cos\phi_2$ we get

$$\left|\frac{\partial x}{\partial \phi}\right| = \begin{vmatrix} -\sin\phi_1 & 0 \\ \cos\phi_1 \cos\phi_2 & -\sin\phi_1 \sin\phi_2 \end{vmatrix} = \sin^2\phi_1 \sin\phi_2.$$

Further, note that

$$x_1^2 = 1 - \sin^2\phi_1,\ x_2^2 = \sin^2\phi_1(1 - \sin^2\phi_2)$$

or

$$\sin^2\phi_1 \sin^2\phi_2 = -x_2^2 + \sin^2\phi_1 = 1 - x_1^2 - x_2^2 = 1 - |x|^2.$$

This together with (4.4) and (4.6) yield

$$\hat{F}_\varepsilon(\xi_1, \xi_2) = C \int_D \frac{gx}{fx - i\varepsilon gx} (1 - |x|^2)^{\frac{n-4}{2}} dx, \tag{4.7}$$

where D is the unit disc on T. Introducing polar coordinates on D, we get

$$\hat{F}_\varepsilon(\xi_1, \xi_2) = C \int_0^1 \int_{S^1} \frac{g\omega}{f\omega - i\varepsilon g\omega} r(1 - r^2)^{\frac{n-4}{2}} dr d\omega, \tag{4.8}$$

which proves (4.5).

As in the proof of Lemma 3.5 in [PSS] we obtain, after a straightforward calculation,

$$\hat{F}(\xi_1, \xi_2) = C_n \frac{\phi(\hat{\xi}_1, \hat{\xi}_2)}{|\xi_1||\xi_2|}, \tag{4.9}$$

where $\hat{\xi}_i = |\xi_i|/|\xi_i|$, $i = 1, 2$ and

$$\phi(\theta, \theta') = (\theta, \theta') + i\sqrt{1 - (\theta, \theta')} = e^{i\alpha}. \tag{4.10}$$

Here α is the angle between θ and θ'.

We introduce Riesz potential I^{-1} and Riesz transform R [St] defined as

$$I^{-1} f(x) = \mathcal{F}^{-1}\left(\frac{1}{|\xi|} \hat{f}(\xi)\right)(x)$$

and

$$Rf(x) = \mathcal{F}^{-1}(\hat{\xi} \hat{f}(\xi))(x).$$

Next we claim that

$$I^{-1} : L^s_\sigma(\mathbb{R}^n) \to W^1_s(\mathbb{R}^n) \tag{4.11}$$

for $\sigma > 1 + n/s$ and $s > 2n$. To see this we first observe that $D_j I^{-1} = R_j$ is bounded in L^s for every $1 < s < \infty$ [St]. Hence we only need to show

$$I^{-1} : L^s_\sigma(\mathbb{R}^n) \to L^s(\mathbb{R}^n) \tag{4.12}$$

for $2 \le s \le \infty$. But f belongs to L^2_σ if and only if $\hat f$ belongs to H^σ. Thus $\hat f(\xi)/|\xi|$ is integrable at the origin and (4.12) holds for $s = 2$. Since for $f \in L^\infty_\sigma, \sigma > 1$,

$$|I^{-1} f(x)| = \left| C_n \int \frac{1}{|x-y|^{n-1}} f(y) dy \right| \le C_n \int \frac{1}{|x-y|^{n-1}} (1+|y|)^{-\sigma} dy,$$

the claim holds for $s = \infty$, too. The case of a general $2 \le s \le \infty$ follows now by interpolation [T].

We write

$$q_1(x) = C_n \int e^{i(x,\xi_1+\xi_2)} \phi(\hat\xi_1, \hat\xi_2) \hat h(\xi_1) \hat h(\xi_2) d\xi_1 d\xi_2, \tag{4.13}$$

where $h = I^{-1} q$. By using binomial series for the square root in (4.10) we obtain

$$q_1(x) = C_n \left((Rh)^2 + i \sum_{k=0}^\infty \binom{1/2}{k} \int e^{i(x,\xi_1+\xi_2)} (\hat\xi_1, \hat\xi_2)^{2k} \hat h(\xi_1) \hat h(\xi_2) d\xi_1 d\xi_2 \right).$$

Since the components R_j of R are bounded also in W^1_s [St] and W^1_s is a multiplication algebra, i.e., $W^1_s \cdot W^1_s \subset W^1_s$ for $s > n$ [T], we obtain for $\|h\|_{W^1_s} < 1/2n$ that q_1 belongs to W^1_s and that

$$\|q_1(x)\|_{W^1_s} \le C_n \left(\|h\|^2_{W^1_s} + \sum_{k=0}^\infty \binom{1/2}{k} n^{2k} \|h\|^2_{W^1_s} \right) \le C_n \|h\|^2_{W^1_s}. \tag{4.14}$$

By the bilinearity of (4.13) the inequality (4.14) holds also without assuming that $\|h\|_{W^1_s}$ is small. The lemma is thus proved. □

As a corollary we obtain for less singular potentials that $q - q_B$ is, indeed, Hölder continuous.

COROLLARY 4.1. *Assume that $q(x)$ belongs to $L^s_\sigma(\mathbb{R}^n)$ for $s > 3n$ and $\sigma > 1 + \frac{n}{s}$. Then $q - q_B \in C^\alpha$ for $0 < \alpha < 1 - 3n/s$.*

Proof. By Sobolev embedding $W^1 \hookrightarrow C^\alpha$ for $1 - n/s > \alpha$. As shown in the proof of Theorem 1.2, $q - q_B - q_1 \in H^t, t < 1 + n/2 - 3n/s$. On the other hand, $q_1 \in W^1_s$ and $H^t \hookrightarrow C^\alpha$ for $t > \alpha + n/2$. □

Finally we remark that the statement in Theorem 2.2 in [PSS] is not correct and should be replaced by Lemma 2.4 of the present article.

REFERENCES

[F] R. FORD, *An inverse scattering result for measure potentials*, Inverse Problems, 11 (1995), pp. 939–948.

[No] R. NOVIKOV, *Inverse scattering for the Schrödinger equation in dimension* 1 *up to smooth functions*, Bull. Sci. Math., 120 (1996), pp. 473–491.

[PS] L. Päivärinta and E. Somersalo, *Inversion of discontinuities for the Schrödinger equation in three dimensions*, SIAM J. Math. Anal., 22 (1991), pp. 480–499.

[PSS] L. Päivärinta, V.S. Serov, and E. Somersalo, *Reconstruction of singularities of a scattering potential in two dimensions*, Adv. Appl. Maths., 15 (1994), pp. 97–113.

[S] V.S. Serov, *On estimates of the resolvent of the Laplace operator over the entire space*, Mat. Zametki, 52 (1992), pp. 109–118.

[So] E. Somersalo, *One-dimensional electromagnetic inverse reflection problem: Formulation as a Riemann–Hilbert problem and imaging of discontinuities*, SIAM J. Appl. Math., 49 (1989), pp. 944–951.

[Sa] Y. Saito, *Some properties of the scattering amplitude and inverse scattering problem*, Osaka J. Maths., 19 (1982), pp. 527–547.

[ST] V.S. Serov and D.S. Tkachenko, *On recovering the potential in the Schrödinger operator on the line using Born's approximation*, Differentsial'nye Uravneniya, 29 (1993), pp. 108–116.

[St] E.M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[GU] A. Greenleaf and G. Uhlmann, *Recovering singularities of a potential from singularities of scattering data*, Comm. Math. Phys., 157 (1993), pp. 549–572.

[SU1] Z. Sun and G. Uhlmann, *Inverse scattering for singular potentials in two dimensions*, Trans. Amer. Math. Soc., 338 (1993), pp. 363–374.

[SU2] Z. Sun and G. Uhlmann, *Recovery of singularities for formally determined inverse problems*, Comm. Math. Phys., 153 (1993), pp. 431–445.

[T] H. Triebel, *Theory of Function Spaces*, Akademische Verlagsgeschellchaft Geest and Portig K.-G., Leipzig, Germany, 1983.

[W] G.N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, UK, 1948.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 712–719, May 1998

011

RADIAL SYMMETRY AND DECAY RATE OF VARIATIONAL GROUND STATES IN THE ZERO MASS CASE*

M. FLUCHER† AND S. MÜLLER‡

Abstract. P.-L. Lions raised the question whether variational ground state solutions of the semilinear Dirichlet problem

$$-\Delta w = f(w) \text{ in } \mathbb{R}^n,$$
$$w(x) \to 0 \text{ as } |x| \to \infty$$

are radial with constant sign. We consider the zero mass case $f(0) = f'(0) = 0$ without regularity assumptions for the nonlinearity. The celebrated symmetry result of Gidas, Ni, and Nirenberg and its refinements do not apply. Nevertheless we give an affirmative answer to the question of Lions. We prove that every variational ground state is either strictly positive or strictly negative. For positive nonlinearities positive solutions are radially symmetric with respect to some point and strictly decreasing in radial direction. For general nonlinearities we show that the same is true outside a compact set. This is a consequence of our main result, the second-order decay estimate

$$w(r) = c\, r^{2-n}(1 + O(r^{-2})) \text{ in the } C^1\text{-sense.}$$

In addition we obtain an integral representation for the constant c.

Key words. elliptic boundary value problem, ground state, symmetry, decay rate

AMS subject classifications. 35J20, 35B40

PII. S0036141096314026

1. Introduction.

In this paper we derive exact decay estimates for variational ground state solutions of the semilinear Dirichlet problem

$$\begin{aligned} -\Delta w &= \lambda f(w) \ \text{ in } \ \mathbb{R}^n, \\ w(x) &\to 0 \ \text{ as } \ |x| \to \infty. \end{aligned} \tag{1.1}$$

We restrict our attention to the *zero mass case* $f(0) = 0$, $f'(0) = 0$. Variational ground state solutions are obtained by solving the variational problem for the *generalized Sobolev constant*

$$S^F := \sup \left\{ \int_{\mathbb{R}^n} F(u) \,:\, u \in D^{1,2}(\mathbb{R}^n),\ \|\nabla u\|_2 \le 1 \right\}$$

with $F' = f$. A *variational ground state* thus satisfies

$$\|\nabla w\|_2 = 1, \qquad \int_{\mathbb{R}^n} F(w) \ = \ S^F.$$

Lions [11, Remark II.6] raised the question whether for nonsymmetric integrands variational ground states are radial with constant sign. For $f \in C^{1,\alpha}$ with $f(0) = 0$ and $f'(0) < 0$, symmetry of positive solutions follows from the result of Gidas, Ni,

*Received by the editors December 23, 1996; accepted for publication (in revised form) June 3, 1997.

http://www.siam.org/journals/sima/29-3/31402.html

†Universität Basel, Mathematisches Institut, Rheinsprung 21, CH-4051 Basel, Switzerland (flucher@math.unibas.ch).

‡Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstr. 22-26, D-04103 Leipzig, Germany (stefan.mueller@mis.mpg.de).

and Nirenberg [7]. Symmetry for Lipschitz continuous nonlinearities is due to Kaper, Kwong, and Li [8]. In section 2 (Lemma 2) we answer Lions' question for the zero mass case. Our symmetry result is obtained as a consequence of the second-order decay estimates of section 4 in combination with a result of Brothers and Ziemer [1]. Due to our variational technique no regularity is needed for the nonlinearity besides a mild growth condition for its antiderivative. The precise hypotheses are as follows:

(F) The integrand F satisfies the growth condition $0 \le F(t) \le c\,|t|^{2^*}$ for some constant c. It is upper semicontinuous and $F \not\equiv 0$ in the L^1 sense.

For smooth integrands every variational ground state is a solution of the Euler–Lagrange equation (1.1). The value of the Lagrange multiplier can be obtained as follows. The scaling $u^s(x) := u(x/s)$ with $s := \|\nabla u\|_2^{-\frac{2}{n-2}}$ leads to $\|\nabla u^s\|_2 = 1$ and

$$\int_{\mathbb{R}^n} F(u^s) = \|\nabla u^s\|_2^{-2^*} \int_{\mathbb{R}^n} F(u). \tag{1.2}$$

Thus variational ground states are critical points of the functional

$$J(u) := \frac{\int_{\mathbb{R}^n} F(u)}{\|\nabla u\|_2^{2^*}} \tag{1.3}$$

defined on $D^{1,2}(\mathbb{R}^n) \setminus \{0\}$. The gradient of this functional is

$$\begin{aligned} J'(w)\phi &= \frac{d}{dt} J(w + t\phi)\bigg|_{t=0} \\ &= \frac{d}{dt}\left(\int_{\mathbb{R}^n} F(w + t\phi) - S^F \frac{2^*}{2} \int_{\mathbb{R}^n} |\nabla(w + t\phi)|^2 \right)\bigg|_{t=0} \\ &= \int_{\mathbb{R}^n} f(w)\phi - 2^* S^F \int_{\mathbb{R}^n} \nabla w \cdot \nabla \phi \end{aligned}$$

by normalization of w and $\int_{\mathbb{R}^n} F(w) = S^F$. Hence

$$\lambda = \frac{1}{2^* S^F}. \tag{1.4}$$

Existence of variational ground states has been discussed in [4]. Uniqueness of radial ground states is an unsolved problem. Most results available deal with the positive mass case $f'(0) < 0$ or require a similar condition leading to exponential decay at infinity. Results of this type have been obtained by many authors including Kwong [9], Kwong and Zhang [10], Chen and Lin [2], and Yanagida [13]. The most general uniqueness results (together with an up-to-date survey of the literature) are due to Franchi, Lanconelli, and Serrin [5]. They consider general divergence operators of the form $\operatorname{div}(A(|\nabla u|)\nabla u)$ including the p-Laplacian and the mean curvature operator.

2. Symmetry. In the following u^* denotes the Schwarz symmetrization or decreasing rearrangement of a positive function u. It is well known that $\|\nabla u^*\|_2 \le \|\nabla u\|_2$ and $\int_{\mathbb{R}^n} F(u^*) = \int_{\mathbb{R}^n} F(u)$. If w is a variational ground state then so is w^* as follows from (1.2). This reduces the computation of the generalized Sobolev constant to a one-dimensional problem. Our symmetry result is based on the following lemma of Brothers and Ziemer and the strict decay property of radial extremals (Theorem 5).

LEMMA 1 (Brothers and Ziemer [1]). *If a positive function $u \in D^{1,2}(\mathbb{R}^n)$ satisfies $\|\nabla u^*\|_2 = \|\nabla u\|_2$ then $u = u^*$ up to translation or u has a plateau of positive volume which is below the top level.*

LEMMA 2 (basic properties of variational ground states). *Assume* (F) *and let w be an extremal for S^F. Then:*

1. *Either $w \geq 0$ or $w \leq 0$.*
2. *There is a ball $B^{r_0}_{x_0}$ (we assume $x_0 = 0$) such that $w = w^*$ outside this ball. If we suppose for normalization that $w \geq 0$ then the function $r \mapsto w(r)$ is strictly decreasing on (r_0, ∞). In particular $w > 0$ in $\mathbb{R}^n$.*

Proof.

1. If w changes sign we split $w = w_+ + w_-$ and we have $\|\nabla w_\pm\|_2 < 1$, $S_\pm := \int_{\mathbb{R}^n} F(w_\pm) < S^F$. Both functions belong to $D^{1,2}(\mathbb{R}^n)$. Normalization by means of (1.2) leads to the contradiction

$$\begin{aligned} S^F = S_+ + S_- &= \|\nabla w_+\|_2^{2^*} \int_{\mathbb{R}^n} F(w_+^{s_+}) + \|\nabla w_-\|_2^{2^*} \int_{\mathbb{R}^n} F(w_-^{s_-}) \\ &\leq S^F \left(\left(\int_{\mathbb{R}^n} |\nabla w_+|^2 \right)^{\frac{n}{n-2}} + \left(\int_{\mathbb{R}^n} |\nabla w_-|^2 \right)^{\frac{n}{n-2}} \right) < S^F \end{aligned}$$

   by definition of the generalized Sobolev constant and strict convexity of the function $t \mapsto t^{\frac{n}{n-2}}$ on $\mathbb{R}^+$. Thus $w \geq 0$ or $w \leq 0$.
2. By Theorem 5 below the symmetrized function w^* is strictly decreasing outside some ball $B_0^{r_0}$. We show that after translation $w = w^*$ in $\mathbb{R}^n \setminus B_0^{r_0}$. Assume the contrary. Then by Lemma 1 the function w^* has a doubly connected plateau $B_0^R \setminus B_0^r$ with $r > r_0$ in contradiction to the fact that w^* is strictly decreasing outside $B_0^{r_0}$. ∎

3. Euler–Lagrange equation corresponding to variations of the independent variable. Our integrands may be discontinuous. Hence we cannot use (1.1). Instead we use the Euler–Lagrange equation corresponding to variations of the independent variable.

If w is a radial function we write $r := |x|$ and $w(r) := w(|x|)$. The jump $[g(r)]$ denotes the difference of right and left limit of a function g at the point r. Moreover, $w'(r)$ denotes an arbitrary element of the subdifferential of w at r. The tensor product of two $1 \times n$-matrices $a \otimes b = a^T b$ is an $n \times n$-matrix. The scalar product of $n \times n$-matrices is defined by $A : B = Tr(A^T B)$.

LEMMA 3. *Every extremal for S^F satisfies*

$$\int_{\mathbb{R}^n} \left(2\nabla w \otimes \nabla w - |\nabla w|^2 \,\mathrm{Id} + 2\lambda F(w)\mathrm{Id} \right) : D\eta = 0 \tag{3.1}$$

for every test function $\eta \in C_c^\infty(\mathbb{R}^n, \mathbb{R}^n)$ with λ as in (1.4). *If w is a radial extremal for S^F then*

$$\int_0^\infty r^n \left(|w'|^2 + 2\lambda F(w) \right) \psi' - r^{n-1} \left((n-2)\,|w'|^2 - 2n\lambda F(w) \right) \psi = 0 \tag{3.2}$$

for every $\psi \in C_c^\infty(\mathbb{R}^+)$. In particular the jump condition $[|w'|^2] = -2\lambda[F(w)]$ holds at the discontinuities of F.

Proof. Consider the variations

$$\phi_t(x) := x + t\eta(x), \qquad w_t := w \circ \phi_t^{-1}$$

of the independent variable. Then

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^n} F(w_t)\bigg|_{t=0} &= \frac{d}{dt}\int_{\mathbb{R}^n} F(w)\,\det D\phi_t\bigg|_{t=0}\\ &= \int_{\mathbb{R}^n} F(w)\,\mathrm{div}\,\eta \quad = \quad \int_{\mathbb{R}^n} F(w)\,\mathrm{Id} : D\eta.\end{aligned}$$

Similarly

$$\begin{aligned}\int_{\mathbb{R}^n} |\nabla w_t|^2 &= \int_{\mathbb{R}^n} \left|\nabla w (D\phi_t)^{-1}\right|^2 \det D\phi_t,\\ \frac{d}{dt}\int_{\mathbb{R}^n} |\nabla w_t|^2\bigg|_{t=0} &= \int_{\mathbb{R}^n} |\nabla w|^2\,\mathrm{Id} : D\eta \;- 2\int_{\mathbb{R}^n} \nabla w \otimes \nabla w : D\eta.\end{aligned}$$

Since w is a maximizer of the functional J defined in (1.3) we have

$$0 = \frac{d}{dt}J(w_t)\bigg|_{t=0} \quad = \quad \frac{d}{dt}\left(\int_{\mathbb{R}^n} F(w_t) - \frac{n}{n-2}S^F\int_{\mathbb{R}^n}|\nabla w_t|^2\right)\bigg|_{t=0}.$$

Inserting the above expressions for the variation of the individual terms yields (3.1). In the radial case set $\eta(x) = \psi(r)x$. □

For smooth integrands (3.2) is equivalent to

$$w'' + \frac{n-1}{r}w' + \lambda f(w) = 0,$$

i.e., to (1.1) at all points where w' is defined and nonzero. A similar identity leading to the same jump condition has been proposed by Franchi, Lanconelli, and Serrin [5, p. 179]. In our terms it assumes the form

$$\left(r\left(|w'|^2 + 2\lambda F(w)\right)^{\frac{1}{2(n-1)}}\right)' = 2\left(|w'|^2 + 2\lambda F(w)\right)^{-\frac{2n-3}{2(n-1)}} F(w).$$

Equation (3.2) is the radial component of

$$\mathrm{div}\left(|\nabla w|^2 x + 2\lambda F(w)x\right) + (n-2)\,|w'|^2 - 2n\lambda F(w) = 0$$

in the sense of distributions. Integration over $\mathbb{R}^n$ yields a special case of the Pohozaev identity, namely

$$\int_{\mathbb{R}^n} |\nabla w|^2 = 2^*\lambda \int_{\mathbb{R}^n} F(w).$$

This is the same as (1.4).

4. Decay estimates. Every radial $D^{1,2}$ function decays at least like $r^{-\frac{n-2}{2}}$ (Strauss [12]). We will see that variational ground states decay faster. In most applications F is nondecreasing on $\mathbb{R}^+$ and nonincreasing on $\mathbb{R}^-$ corresponding to $f \geq 0$ on $\mathbb{R}^+$ and $f \leq 0$ on $\mathbb{R}^-$, respectively. In this case every positive extremal is superharmonic in the sense of distributions. This can be seen as follows. For every positive test function ϕ and $t \geq 0$ we have $\int_{\mathbb{R}^n} F(w + t\phi) \geq \int_{\mathbb{R}^n} F(w)$ and therefore $\int_{\mathbb{R}^n} |\nabla(w + t\phi)|^2 \geq \int_{\mathbb{R}^n} |\nabla w|^2$ by definition of the generalized Sobolev constant. This

implies $\int_{\mathbb{R}^n} \nabla w \cdot \nabla \phi \geq 0$. On every subdomain a superharmonic function is pointwise bigger or equal to the harmonic function with the same boundary values. The properties listed below are consequences of this fact. Let

$$K(r) := \frac{1}{(n-2)\,|S^{n-1}|\,r^{n-2}}$$

denote the fundamental singularity of the Laplacian.

LEMMA 4 (monotone integrands). *Assume* (F) *and let w be a positive extremal for S^F. If F is nondecreasing on $\mathbb{R}^+$ then:*

1. *After translation $w = w^*$. We can assume that w is radial with respect to the origin.*
2. *The function $r \mapsto w(r)$ is strictly decreasing on $\{w < w(0)\}$.*
3. *The function $r \mapsto w(r)/K(r)$ is nondecreasing on $\mathbb{R}^+$.*
4. *The function $r \mapsto w'(r)/K'(r)$ is nondecreasing on $\mathbb{R}^+$. In particular every kink satisfies $[w'] \leq 0$.*
5. *The limit*

$$w_\infty := \lim_{r\to\infty} \frac{w(r)}{K(r)} \quad = \quad \lim_{r\to\infty} \frac{w'(r)}{K'(r)}$$

exists and $w_\infty > 0$. In particular $w(r) \leq w_\infty r^{2-n}$ and $F(w(r)) \leq c\, r^{-2n}$.

Proof. The symmetrized function w^* is also extremal. By weak superharmonicity of w^* outside the ball B_0^r we have

$$w^*(R) \geq \frac{w^*(r)}{K(r)} K(R)$$

for every $R \geq r$. In particular w^* is strictly decreasing below the top level and strictly positive. Lemma 1 implies $w = w^*$ up to translation. The above inequality also shows $w'/K' \geq w/K$. Weak superharmonicity of w implies that for every triple $r \leq s \leq R$ we have

$$w(s) \geq w(R) + \frac{K(s) - K(R)}{K(r) - K(R)} (w(r) - w(R)).$$

Therefore

$$w'(r) \geq K'(r) \frac{w(r) - w(R)}{K(r) - K(R)}, \qquad w'(R) \quad \leq \quad K'(R) \frac{w(r) - w(R)}{K(r) - K(R)},$$

which implies $w'(r)/K'(r) \leq w'(R)/K'(R)$. Thus the limit w_∞ exists in $(0, \infty]$. We are left to show that it is finite. This is trivial if $F = 0$ near 0. For general integrands we refer to Theorem 5. Also the last claim follows from weak superharmonicity of w. For every triple $r \leq s \leq R$ we have

$$w(s) \geq w(R) + \frac{K(s) - K(R)}{K(r) - K(R)} (w(r) - w(R)).$$

Therefore

$$w'(r) \geq K'(r) \frac{w(r) - w(R)}{K(r) - K(R)}, \qquad w'(R) \quad \leq \quad K'(R) \frac{w(r) - w(R)}{K(r) - K(R)}.$$

Thus $w'(r)/K'(r) \leq w'(R)/K'(R)$. ∎

The proof of our decay estimates involves an iteration technique that is common in elliptic regularity theory (see, e.g., Giaquinta [6, Chapter III, Lemma 2.1]).

THEOREM 5 (decay rate of variational ground states). *Assume* (F) *and let* $w = w^*$ *be a positive radial extremal for* S^F. *Then*

$$\begin{aligned} w_\infty K(r)\left(1 - O(r^{-2})\right) \le w(r) &\le w_\infty K(r), \\ w_\infty K'(r) \le w'(r) &\le w_\infty K'(r)\left(1 - O(r^{-2})\right) \end{aligned}$$

for $r \to \infty$, *where*

$$w_\infty^2 = \frac{2(n-1)}{nS^F} \int_{\mathbb{R}^n} \frac{F(w)}{K(|\cdot|)}.$$

In particular $w'(r) < 0$ *for* $r > r_0$ *with some* $r_0 \ge 0$, $w(r) \le c\, r^{2-n}$, $F(w(r)) \le c\, r^{-2n}$, *and*

$$\int_{\mathbb{R}^n \setminus B_0^R} |\nabla w|^2 \le c\, R^{2-n}, \quad \int_{\mathbb{R}^n \setminus B_0^R} F(w) \le c\, R^{-n}$$

for every $R > 0$. *The error terms depend only on the constant in the growth condition* (F).

Proof. With $\psi(r) = r^{n-2}\eta(r)$ equation (3.2) assumes the form

$$\int_0^\infty r^{2(n-1)} \left(|w'|^2 + 2\lambda F(w)\right) \eta' + 2(n-1) r^{2n-3} 2\lambda F(w)\eta = 0.$$

In terms of the auxiliary functions

$$\begin{aligned} a(r) &:= r^n \left(|w'(r)|^2 + 2\lambda F(w(r))\right), \\ b(r) &:= r^n 2\lambda F(w(r)), \end{aligned}$$

this can be written as

$$\left(r^{n-2} a(r)\right)' = 2(n-1) r^{n-3} b(r) \tag{4.1}$$

or

$$a'(r) = r^{n-1} \left(2n\lambda F(w) - (n-2)\,|w'|^2\right)$$

in the sense of distributions. The right-hand side is in $L^1(\mathbb{R}^+)$. Thus $\lim_{r\to\infty} a(r)$ exists. Together with

$$\int_0^\infty \frac{a(r)}{r} = \int_0^\infty r^{n-1} \left(|w'|^2 + 2\lambda F(w)\right) < \infty,$$

this shows that $\lim_{r\to\infty} a(R) = 0$. Also the decreasing envelope of a,

$$A(r) := \sup_{R \ge r} a(R),$$

tends to 0. In terms of this function we can estimate

$$w(r) = \int_\infty^r w' \le \int_r^\infty \left(\frac{a(R)}{R^n}\right)^{\frac{1}{2}} \le c\, A^{\frac{1}{2}}(r) r^{-\frac{n-2}{2}},$$

and therefore

$$b(r) \le c\, r^n w^{2^*} \quad \le \quad c\, A^{\frac{n}{n-2}}(r) \tag{4.2}$$

by (F). Integration of (4.1) over the interval (r, R) yields

$$a(R) - \left(\frac{r}{R}\right)^{n-2} a(r) \le c \left(1 - \left(\frac{r}{R}\right)^{n-2}\right) A^{\frac{n}{n-2}}(r) \quad \le \quad c\, A^{\frac{n}{n-2}}(r).$$

Since the larger radius appears in the denominator we can take the supremum over all $R' \ge R$ to obtain

$$A(R) \le \left(\left(\frac{r}{R}\right)^{n-2} + A^{\frac{2}{n-2}}(r)\right) A(r) \quad \text{for every} \quad R > r > 0. \tag{4.3}$$

Fix $\beta \in \left(\frac{n-2}{n}(n-2), n-2\right)$ and choose r such that

$$A(r)^{\frac{2}{n-2}} \le 2^{-\beta} - 2^{-(n-2)}.$$

This is possible because $A(r) \to 0$ as $r \to \infty$. Iterated application of (4.3) with $R = 2r$ and monotonicity of A yields

$$A(r) \le c r^{-\beta}, \quad b(r) \quad \le \quad c r^{-\frac{\beta n}{n-2}}$$

for every r by (4.2). Since $\frac{\beta n}{n-2} > n-2$, the right-hand side of (4.1) is integrable. In particular $a(r) \le c\, r^{2-n}$. Another application of (4.2) yields

$$b(r) \le c\, r^{-n}, \quad F(w(r)) \quad \le \quad c\, r^{-2n},$$

and $r^{n-3} b(r) \to 0$ as $r \to \infty$. Integration of (4.1) together with $a(0) = 0$ yields

$$\begin{aligned} r^{2(n-1)} \left(|w'|^2 + 2\lambda F(w)\right) \quad &\le \quad c_1^2 \quad \text{with} \\ c_1^2 \quad := \quad 2(n-1) \int_0^\infty r^{2n-3} 2\lambda F(w(r)) \quad &= \quad \frac{4(n-1)\lambda}{|S^{n-1}|} \int_{\mathbb{R}^n} |x|^{n-2} F(w) \end{aligned}$$

for every r with equality at infinity. Thus, in view of (1.4),

$$\begin{aligned} |w'(r)| \le c_1\, r^{1-n} \quad &= \quad w_\infty \left|K'(r)\right|, \\ w(r) \le \int_r^\infty |w'| \quad &\le \quad \frac{c_1}{(n-2) r^{n-2}} \quad = \quad w_\infty K(r) \end{aligned}$$

by integration. Again by (4.1)

$$\begin{aligned} R^{2(n-1)} \left(|w'|^2 + 2\lambda F(w)\right) &= c_1^2 - 2(n-1) \int_R^\infty r^{n-3} b(r) \\ &= c_1^2 - O(R^{-2}) \end{aligned}$$

since $b(r) \le c\, r^{-n}$. Together with $F(w(r)) \le c\, r^{-2n}$ we obtain

$$w'(r) = w_\infty K'(r) \left(1 + O(r^{-2})\right)$$

and the lower bound for w. □

The approximation formulas of Theorem 5 are second-order accurate. This is best possible as can be seen from the variational ground state

$$w(r) = \frac{c}{(1+r^2)^{\frac{n-2}{2}}} \quad = \quad w_\infty K(r)\left(1 + O(r^{-2})\right)$$

for the critical power $F(t) = |t|^{2^*}$. For $0 \le f(t) \le t^p$ with $p > \frac{n+2}{n-2}$, Egnell [3, Proposition C] found

$$w_\infty = \lambda \int_{\mathbb{R}^n} f(w).$$

Formally this follows by integration of (1.1) over a large ball.

REFERENCES

[1] J. E. Brothers and W. P. Ziemer, *Minimal rearrangements of Sobolev functions*, J. Reine Angew. Math., 384 (1988), pp. 153–179.

[2] C. C. Chen and C.-S. Lin, *Uniqueness of the ground state solutions of* $\Delta u + f(u) = 0$ *in* $\mathbb{R}^n$*;* $n \ge 3$, Comm. Partial Differential Equations, 16 (1991), pp. 1549–1572.

[3] H. Egnell, *Asymptotic results for finite energy solutions of semilinear elliptic equations*, J. Differential Equations, 98 (1992), pp. 34–56.

[4] M. Flucher and S. Müller, *Concentration of low energy extremals*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[5] B. Franchi, E. Lanconelli, and J. Serrin, *Existence and uniqueness of nonnegative solutions of quasilinear equations in* $\mathbb{R}^n$, Adv. Math., 118 (1996), pp. 177–243.

[6] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Annals of Mathematics Studies 105, Princeton University Press, Princeton, NJ, 1983.

[7] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[8] H. G. Kaper, M. K. Kwong, and Y. Li, *Symmetry results for reaction-diffusion equations*, Differential Integral Equations, 6 (1993), pp. 1045–1056.

[9] M. K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* $\mathbb{R}^n$, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[10] M. K. Kwong and L. Q. Zhang, *Uniqueness of the positive solution of* $\Delta u + f(u) = 0$ *in an annulus*, Differential Integral Equations, 4 (1991), pp. 583–599.

[11] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case*, II, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 223–283.

[12] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

[13] E. Yanagida, *Uniqueness of positive radial solutions of* $\Delta u + g(r)u + h(r)u^p = 0$ *in* $\mathbb{R}^n$, Arch. Rational Mech. Anal., 115 (1991), pp. 257–274.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 720–735, May 1998

012

L_∞ ESTIMATES ON THE SOLUTIONS OF NONSELFADJOINT ELLIPTIC AND PARABOLIC EQUATIONS IN BOUNDED DOMAINS*

ADRIAN T. HILL†

Abstract. This paper considers explicit upper bounds in L_∞ on the solution operator of a class of second-order parabolic Dirichlet problems defined in $(-1, 1)^N$. The elliptic part of the operator L is given by

$$Lu = -\sum_{i=1}^{N} a_i(x, t)\frac{\partial^2 u}{\partial x_i^2} + \sum_{i=1}^{N} b_i(x, t)\frac{\partial u}{\partial x_i},$$

where $a_i \geq d_i > 0$, $|b_i| \leq M_i$, $i = 1, \ldots, N$, uniformly across the domain.

Symmetry and the maximum principle are used to identify those coefficients, obeying these bounds, which result in the largest possible value for the norm of the solution operator in L_∞. The norm of this optimal case is found in terms of (d_i) and (M_i) and a family of constant coefficient problems in one space dimension. This representation is made quantitatively explicit by Laplace transform evaluation of the one-dimensional problems.

Similar sharp quantitative estimates on the resolvent $\|(\lambda I + L)^{-1}\|_\infty$, $\lambda \geq 0$, are obtained in $(-1, 1)^N$ as a corollary of the parabolic results. For comparison, a related, but direct, technique is used to derive optimal bounds on the resolvent of a slightly more general class of elliptic operators defined on the unit ball in $\mathbb{R}^N$.

Key words. heat kernels, nonselfadjoint operators

AMS subject classifications. 35B45, 35K05

PII. S0036141096310156

1. Introduction.

Our objective in this paper is to obtain explicit sharp L_∞ estimates for a class of elliptic and parabolic operators, defined on a bounded subdomain of $\mathbb{R}^N$ with Dirichlet boundary conditions. Elsewhere, we consider L_1 bounds in bounded domains and in $\mathbb{R}^N$ [8, 9]. Such quantitative estimates are of use in the applied analysis of numerical algorithms, control systems, and biological models. Thus, our motivation differs from that of Agmon, Douglis, and Nirenberg [2] and Stewart [13], whose far more general estimates are of a qualitative nature.

On the domain $\Omega = (-1, 1)^N$, we investigate the parabolic equation

$$\frac{\partial u}{\partial t} + Lu \equiv \frac{\partial u}{\partial t} - \sum_{i=1}^{N} a_i(x, t)\frac{\partial^2 u}{\partial x_i^2} + \sum_{i=1}^{N} b_i(x, t)\frac{\partial u}{\partial x_i}, \qquad (x, t) \in \Omega \times (0, \infty), \tag{1.1}$$

subject to the Dirichlet boundary conditions

$$u(x, t) = 0 \qquad \text{for } (x, t) \in \partial\Omega \times (0, \infty),$$

with initial data u_0 taken to be in $L_\infty(\Omega)$.

It is assumed that $a_i \in C(\overline{\Omega} \times [0, \infty))$ satisfies

$$\inf_{(x, t) \in \Omega \times [0, \infty)} a_i(x, t) \geq d_i > 0; \tag{1.2}$$

*Received by the editors October 7, 1996; accepted for publication March 25, 1997.
http://www.siam.org/journals/sima/29-3/31015.html
†School of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (ath@maths.bath.ac.uk).

also that $b(\cdot, t) = (b_i(\cdot, t)) \in L_\infty(\Omega)$ for $t \geq 0$, and

$$\text{ess. sup}_{(x,t)\in\overline{\Omega}\times[0,\infty)}|b_i(x, t)| \leq M_i, \qquad i = 1, \ldots, N. \tag{1.3}$$

Our objective is to derive sharp estimates for $\|u(t)\|_\infty/\|u_0\|_\infty$, for nonnull u_0, in terms of the bounds $d = (d_i)$ and $M = (M_i)$. This is rather similar to the work of Pucci [11] for elliptic equations, though we work with somewhat different assumptions on the coefficients. As in [11], we identify a maximizing operator, which here is given by

$$L_{d,M}u = -\sum_{i=1}^{N} d_i \frac{\partial^2 u}{\partial x_i^2} + \sum_{i=1}^{N} M_i \text{sign}[x_i] \frac{\partial u}{\partial x_i}. \tag{1.4}$$

$V(d, M; x, t)$, the solution of (1.1) when $L = L_{d,M}$ and $u_0(x) \equiv 1$, is symmetric about $x_i = 0$ and is therefore the solution of a constant coefficient equation in the octant $[0, 1]^N$ with Neumann boundary conditions at $x_i = 0$, $i = 1, \ldots, N$. As a result, we show that $V(d, M; x, t)$ may be written as the product of the solutions, $v(M_i/d_i; x_i, d_i t)$, of constant coefficient equations in one space dimension.

By analyzing the behavior of $v(M; x, t)$, for $M \geq 0$, $x \in [0, 1]$, we find that $V(d, M; x, t)$ also satisfies

$$u_t - \sum_{i=1}^{N} d_i \frac{\partial^2 u}{\partial x_i^2} - \sum_{i=1}^{N} M_i \left| \frac{\partial u}{\partial x_i} \right| = 0, \qquad (x, t) \in \Omega \times (0, \infty), \tag{1.5}$$

and is therefore an upper solution for any equation of the form (1.1), with initial data such that $\|u_0\|_\infty \leq 1$. Consequently, for nonnull $u_0 \in L_\infty$,

$$\frac{\|u(t)\|_\infty}{\|u_0\|_\infty} \leq \|V(d, M; \cdot, t)\|_\infty = V(d, M; 0, t) = \prod_{i=1}^{N} v(M_i/d_i; 0, d_i t). \tag{1.6}$$

The right-hand side is now explicitly computable via the Laplace transform.

In section 2, we establish the qualitative properties of v and V, leading to the proof of (1.6). As a corollary, we also deduce sharp upper bounds for the elliptic case in $(-1, 1)^N$, in terms of v. More quantitative properties of $v(M; 0, t)$ are obtained in section 3, following small and large time expansions of the Laplace transform. Explicitly computable estimates for both the parabolic and elliptic cases are derived as a consequence. In section 4, we consider elliptic boundary value problems defined on the unit ball in $\mathbb{R}^N$, under less restrictive assumptions on the form of the diffusion coefficients, and sharp estimates are obtained in L_∞ using a related, but direct, method.

We remark that our bounds are not confined exclusively to the special geometries we study here. Scaling easily extends all of our results on the cube to rectangular domains. More generally, it is known that the L_∞ norm of the parabolic solution operator and of the resolvent $(\lambda I + L)^{-1}$ are monotonically increasing with respect to domain inclusion; see Sattinger [12]. Thus, our bounds still hold, though less sharply, for operators defined on $\Omega' \subset \Omega$ whose coefficients can be extended to satisfy (1.2), (1.3) on Ω.

2. Qualitative bounds on the cube. In this section, we find bounds for the solutions of (1.1)–(1.3) in terms of the solution of a constant coefficient parabolic

problem in one space dimension. Subsequently, we extend our results to the elliptic case.

We begin with a detailed study of the properties of the solution of the following problem. For $M \in [0, \infty)$, we consider the initial boundary value problem

$$\begin{aligned} u_t + L_M u \equiv u_t - u_{xx} + M u_x = 0, &\qquad (x, t) \in (0, 1) \times (0, \infty), \\ u_x(0, t) = u(1, t) = 0, &\qquad t \in (0, \infty), \\ u(x, 0) = 1, &\qquad x \in (0, 1). \end{aligned} \tag{2.1}$$

Standard results imply that (2.1) possesses a solution $v(M)$, which is unique in $C[[0, \infty); L_2(0, 1)]$, and belongs to $C[[0, \infty) \times [0, 1)] \cap C^\infty[(0, \infty); C^\infty[0, 1]]$. There is a finite discontinuity of $v(M; x, t)$ at $(x, t) = (1, 0)$.

LEMMA 2.1. *For all $M \geq 0$, v satisfies*

$$v_t(M; x, t) < 0, \qquad (x, t) \in [0, 1) \times (0, \infty), \tag{2.2}$$

$$v_x(M; x, t),\ v_{xx}(M; x, t) < 0, \qquad (x, t) \in (0, 1] \times (0, \infty), \tag{2.3}$$

and $\|v(M; \cdot, t)\|_\infty = v(M; t)$ for $t \geq 0$, where

$$v(M; t) \equiv v(M; 0, t). \tag{2.4}$$

Proof. The initial data $u_0(x) \equiv 1$ is a strict upper solution for the elliptic boundary value problem corresponding to (2.1). Consequently, a monotonicity result of Sattinger [12, Lemma 3.5] now implies that (2.2) holds. Integrating e^{-Mx} times (2.1) with respect to x leads to the identity

$$v_x(M; x, t) = \int_0^x e^{M(x-y)} v_t(M; y, t)\, dy < 0, \qquad (x, t) \in (0, 1] \times (0, \infty). \tag{2.5}$$

This implies that, for all $t > 0$, $v(M; \cdot, t)$ has its maximum at $x = 0$. From (2.1),

$$v_{xx} = M v_x + v_t < 0, \qquad (x, t) \in (0, 1] \times (0, \infty).$$

(Here, $v_x(M; 1, t)$ and $v_{xx}(M; 1, t)$ are continuous limits as $x \to 1$.) ∎

The following theorem collects further results describing the qualitative behavior of $v(M; t)$, which is of use in studying the behavior of the upper bound (1.6). However, the remaining results of this section do not depend upon it.

THEOREM 2.2. *$v(M; t)$ increases strictly with M. For fixed $M \geq 0$, $v(M; t)$ satisfies*

$$-\lambda_1 < \frac{1}{v(M; t)} \frac{dv}{dt}(M; t) < 0, \qquad t > 0, \tag{2.6}$$

$$\frac{d}{dt}\left(\frac{1}{v(M; t)} \frac{dv}{dt}(M; t) \right) \leq 0, \qquad t > 0, \tag{2.7}$$

where λ_1 is the principal eigenvalue of the elliptic operator L_M, subject to the boundary conditions of (2.1).

Proof. For $M > 0$, (2.3) implies that, for any $\epsilon \in (0, 1]$, $v(M; x, t)$ is a strict upper solution for (2.1) when $M(1 - \epsilon)$ replaces M, and thus $v(M; t)$ is strictly increasing with M by parabolic monotonicity.

We now consider

$$p(M; x, t) = \frac{v_x(M; x, t)}{v(M; x, t)}, \qquad (x, t) \in [0, 1) \times (0, \infty).$$

Since v_x and v satisfy equation (2.1), p obeys

$$u_t - u_{xx} + (M - 2u)u_x = 0, \qquad (x, t) \in (0, 1) \times (0, \infty).$$

(2.3) states that $v_x(M; 1, t) < 0$ for $t > 0$. Since $v_t(M; 1, t) = 0$, $v_{xx}(M; 1, t) = Mv_x(M; 1, t)$ for $t > 0$. Hence, p satisfies the boundary conditions

$$u(0, t) = 0, \qquad \lim_{x \to 1} u(x, t) + \frac{1}{1 - x} = \frac{M}{2}.$$

We define $q^h(M; x, t) = (p(M; x, t + h) - p(M; x, t))/h$, $h > 0$. From the above considerations for p, we see that q^h satisfies

$$\begin{aligned} u_t - u_{xx} + (M - \overline{p})u_x - \overline{p}_x u = 0, &\qquad (x, t) \in (0, 1) \times (0, \infty), \\ u(0, t) = u(1, t) = 0, &\qquad t > 0, \end{aligned}$$

where $\overline{p}(x, t) = (p(M; x, t + h) + p(M; x, t))/2$. For $x \in [0, 1)$, (2.3) implies that $p(M; x, h) \le 0$, and so $q^h(M; x, 0) \le 0$. We note that $\overline{p}$ and $\overline{p}_x$ are unbounded as x tends to 1. However, one may argue by contradiction, using the usual maximum principle arguments and considering $e^{-Kt}q^h$ for sufficiently large positive values of K, to show that $q^h(M; x, t) \le 0$ for $(x, t) \in (0, 1) \times (0, \infty)$. Thus,

$$\frac{\partial}{\partial x}\left(\frac{v_t}{v}\right) = \frac{\partial^2 \log v}{\partial x \partial t} = \frac{\partial}{\partial t}\left(\frac{v_x}{v}\right) = \lim_{h \to 0} q^h \le 0, \qquad (x, t) \in (0, 1) \times (0, \infty). \tag{2.8}$$

Hence, for all $t > 0$,

$$-\frac{v_t(M; t)}{v(M; t)} = \inf_{x \in [0, 1)} -\frac{v_t}{v}(M; x, t) = \inf_{x \in [0, 1)} \frac{-v_{xx} + Mv_x}{v}(M; x, t) < \lambda_1,$$

where the last inequality is a consequence of a result of Protter and Weinberger [10], and hence (2.6) holds.

(2.8) also implies that for $(x, t) \in [0, 1) \times (0, \infty)$,

$$\frac{d}{dt}\left(\frac{v(M; x, t)}{v(M; t)}\right) = \left[\frac{v_t(M; x, t)}{v(M; x, t)} - \frac{v_t(M; t)}{v(M; t)}\right]\frac{v(M; x, t)}{v(M; t)} \le 0.$$

Thus for any $h > 0$,

$$0 \le \frac{v(M; t + h)}{v(M; t)} v(M; x, t) - v(M; x, t + h), \qquad (x, t) \in [0, 1) \times (0, \infty).$$

Considering the right-hand side as initial data for (2.1), over the time interval $[0, h]$, the maximum principle implies that

$$0 \le \frac{v(M; t + h)}{v(M; t)} v(M; x, t + h) - v(M; x, t + 2h), \qquad (x, t) \in [0, 1) \times (0, \infty).$$

Hence, letting $x = 0$, we conclude that

$$\begin{aligned} 0 &\ge \lim_{h \to 0} \frac{1}{h^2}\left[\frac{v(M; t + 2h) - v(M; t + h)}{v(M; t + h)} - \frac{v(M; t + h) - v(M; t)}{v(M; t)}\right] \\ &= \frac{d}{dt}\left(\frac{1}{v(M; t)}\frac{dv}{dt}(M; t)\right), \qquad t > 0, \end{aligned}$$

which implies (2.7). □

We now extend the definition of $v(M;\, x,\, t)$ to $x \in [-1,\, 0)$ by reflection:

$$v(M;\, x,\, t) \equiv v(M;\, -x,\, t), \qquad (x,\, t) \in [-1,\, 0) \times [0,\, \infty). \tag{2.9}$$

Since $v_x(M;\, 0,\, t) = 0$ and v is even in x, $v(M;\, \cdot,\, t) \in C^2[-1,\, 1]$, and v satisfies the equation

$$u_t - u_{xx} + M\text{sign}[x]u_x = 0, \qquad (x,\, t) \in (-1,\, 1) \times (0,\, \infty), \tag{2.10}$$

(with the convention that sign[0] = 0).

For $d = (d_i)$, $M = (M_i)$, and $\Omega = (-1,\, 1)^N$, we define

$$V(d,\, M;\, x,\, t) = \prod_{i=1}^{N} v(M_i/d_i;\, x_i,\, d_i t), \qquad (x,\, t) \in \overline{\Omega} \times [0,\, \infty). \tag{2.11}$$

From the foregoing analysis, we conclude that the following lemma is true.

LEMMA 2.3. *The function* $V(d,\, M;\, \cdot,\, \cdot) \in C^1[(0,\, \infty);\, C(\Omega)] \cap\, C\left[(0,\, \infty);\, C^2(\overline{\Omega})\right] \cap\, C[[0,\, \infty) \times \Omega]$ *is the solution of the problem*

$$\begin{aligned} u_t + L_{d,\, M} u &= 0, \qquad (x,\, t) \in \Omega \times (0,\, \infty), \\ u(x,\, t) &= 0, \qquad (x,\, t) \in \partial\Omega \times (0,\, \infty), \\ u(x,\, 0) &= 1, \qquad x \in \Omega. \end{aligned}$$

Furthermore,

$$\text{sign}\left[\frac{\partial V}{\partial x_i}\right] = -\,\text{sign}[x_i], \qquad \frac{\partial^2 V}{\partial x_i^2} \le 0, \qquad (x,\, t) \in [-1,\, 1] \times (0,\, \infty),$$

$V(d,\, M)$ *satisfies* (1.5), *and* $\|V(d,\, M;\, \cdot,\, t)\|_\infty = V(d,\, M;\, 0,\, t)$.

THEOREM 2.4. *Suppose that* $u(x,\, t)$ *is the solution of an initial boundary value problem of the form* (1.1)–(1.3), *for some* $d,\, M \in \mathbb{R}^N$, *and nonnull initial data* $u_0 \in L_\infty(\Omega)$. *Then,*

$$\frac{\|u(t)\|_\infty}{\|u_0\|_\infty} \le V(d,\, M;\, 0,\, t) = \prod_{i=1}^{N} v(M_i/d_i;\, d_i t). \tag{2.12}$$

This upper bound is attained when $a_i(x,\, t) = d_i$, $b_i = M_i\,\text{sign}[x_i]$, *and* $u_0(x)$ *is a nonzero constant.*

Proof. By linearity, we may assume without loss that $\|u_0\|_\infty = 1$. Now, Lemma 2.3 implies that for each $t > 0$, almost everywhere in Ω,

$$\begin{aligned} V_t + LV &= V_t - \sum_{i=1}^{N} a_i \frac{\partial^2 V}{\partial x_i^2} + \sum_{i=1}^{N} b_i \frac{\partial V}{\partial x_i} \\ &= -\sum_{i=1}^{N} (a_i - d_i) \frac{\partial^2 V}{\partial x_i^2} + \sum_{i=1}^{N} (b_i - M_i \text{sign}[x_i]) \frac{\partial V}{\partial x_i} \\ &\ge 0. \end{aligned}$$

Since V satisfies the boundary conditions and $V(d,\, M;\, x,\, 0) \ge u_0(x)$, V is an upper solution for the problem (1.1)–(1.3); similarly, $-V$ is a lower solution. The maximum principle now implies that

$$-V(d,\, M;\, x,\, t) \le u(x,\, t) \le V(d,\, M;\, x,\, t), \qquad (x,\, t) \in \overline{\Omega} \times [0,\, \infty).$$

Thus $\|u(t)\|_\infty \le \|V(d, M; \cdot, t)\|_\infty$. Lemma 2.3 implies that $\|V(d, M; \cdot, t)\|_\infty = V(d, M; 0, t)$. □

Since Laplace transform techniques will be used later to establish further properties of V, it is convenient to consider properties of the elliptic problem corresponding to the operator $L_{d,M}$. The boundary value problem

$$L_{d,M}u + \lambda u = 1, \quad x \in \Omega; \qquad u(x) = 0, \quad x \in \partial\Omega, \tag{2.13}$$

has a unique solution in $H_0^1(\Omega)$ for $\lambda \ge 0$ (see, e.g., Gilbarg and Trudinger [7, Chapter 8]), which we define as $W(d, M; x, \lambda)$.

LEMMA 2.5.

$$\begin{aligned} \overline{V}(d, M; x, \lambda) &\equiv \int_0^\infty e^{-\lambda t} V(d, M; x, t)\, dt \\ &= W(d, M; x, \lambda), \qquad (x, \lambda) \in [-1, 1]^N \times [0, \infty). \end{aligned} \tag{2.14}$$

Furthermore, $\operatorname{sign}[\partial W/\partial x_i] = -\operatorname{sign}[x_i]$ *and*

$$\|W(d, M; \lambda, \cdot)\|_\infty = W(d, M; \lambda, 0) = \int_0^\infty e^{-\lambda t} V(d, M; 0, t), dt. \tag{2.15}$$

Proof. Standard parabolic regularity implies that we may differentiate each term in (2.1) with respect to t for $t \ge 1$, say, so that $v_t(M; x, t)$ also satisfies (2.1), including the boundary conditions, for $t \ge 1$. We may now deduce from a result of Friedman [6, Chapter 6] that there are constants C, $\mu > 0$ such that

$$\|v_t(t)\|_\infty \le Ce^{-\mu t}, \qquad t \ge 1.$$

For fixed t, one may use (2.5) to express v, v_x, and v_{xx} in terms of v_t. Now, using the definition of $V(d, M; x, t)$ in terms of v, we conclude that, for each $x \in \Omega$, V and its first and second partial derivatives are absolutely integrable over the time interval $(0, \infty)$. Thus, by a standard result of calculus, x-differentiation commutes with Laplace transformation for $x \in \Omega$. Hence, $\overline{V}(d, M; \cdot, \lambda) \in H_0^1(\Omega)$ and satisfies (2.13) and so is equal to $W(d, M; \cdot, \lambda)$ by uniqueness.

Since $V(d, M; x, t)$ is nonnegative, with a maximum at $x = 0$ for each $t \ge 0$, $\overline{V}(d, M; x, \lambda)$ and therefore $W(d, M; x, \lambda)$ must also be nonnegative with a maximum at $x = 0$. Lastly, since the sign of $\partial V/\partial x_i$ is invariant for fixed x, $\partial W/\partial x_i$ takes the same sign, by the commutativity of x-differentiation and t-integration. □

3. Calculations of $v(M; t)$. In order to obtain practical bounds as a consequence of Theorem 2.4, it remains to explicitly compute the function $v(M; t)$ for $M, t \ge 0$. By a simple calculation one finds that the solution $w(M; x, \lambda)$ of the problem

$$(L_M + \lambda I)[u] = 1; \quad x \in (0, 1); \quad u_x(0) = u(1) = 0, \tag{3.1}$$

takes the following values at $x = 0$:

$$\left.\begin{aligned} &\frac{e^M - M - 1}{M^2}, && \lambda = 0, \\ &\frac{1}{\lambda} - \frac{2\gamma e^{-(M/2+\gamma)}}{\lambda(\gamma - M/2 + (\gamma + M/2)e^{-2\gamma})}, && \left[\gamma = (\lambda + M^2/4)^{1/2}\right], \quad \lambda > 0. \end{aligned}\right\} \tag{3.2}$$

Since Lemma 2.5 implies that $\overline{v}(M; 0, \lambda) = w(M; 0, \lambda)$, we calculate $v(M; t)$ by finding two expansions for the inverse Laplace transform of (3.2), one of which is rapidly convergent for large t and the other for small t.

3.1. Eigenmode expansion. Here, we apply the Mellin inversion and residue theorems directly to (3.2) to obtain an expansion suitable for small t, which is equivalent to an expansion in terms of the eigenfunctions of L_M. Extending the definition of $\overline{v}(M;\,\lambda)$ to complex λ, one sees that the function is single valued because it is even in γ. The singularity at $\lambda = 0$ is removable, and thus poles occur only when $\gamma = 0$ or $\tanh\gamma/\gamma = M/2$. The imaginary part of this last equation yields

$$b\sinh 2a = a\sin 2b \qquad \text{for } \gamma = a + bi, \quad a,\, b \in \mathbb{R},$$

but this can only be satisfied when either $a = 0$ or $b = 0$ since

$$\sinh x/x > 1 > \sin x/x \qquad \text{for } \ x \in \mathbb{R}\setminus\{0\}.$$

Hence poles only occur when γ is real or pure imaginary.

Negative γ is excluded, so when $b = 0$ we require solutions of

$$a\cosh a = M/2\sinh a, \qquad a \geq 0. \tag{3.3}$$

The solution $a = 0$ does not yield a pole, except when $M = 2$. For $M \in [0,\,2)$, (3.3) has no other solutions. When $M \in [2,\,\infty)$ there is exactly one solution, yielding a pole $\lambda = -\lambda_1$, where $\lambda_1 = M^2/4 - a^2$. λ_1 is the principal eigenvalue of the operator L_M, previously considered in Theorem 2.2.

The remaining poles occur when $\gamma = ib$, where b is a solution of

$$b\cos b = M/2\sin b, \qquad b > 0. \tag{3.4}$$

For all $M \geq 0$ there is exactly one solution b_n of (3.4) in every interval $((n-1)\pi,\,(n-1/2)\pi]$ for $n = 2,\,3,\,4,\,\ldots$. When $M \in [0,\,2)$ there is a further solution $b_1 \in (0,\,\pi/2]$. These solutions yield poles $\lambda = -\lambda_n$, where $\lambda_n = M^2/4 + b_n^2$. Hence, all poles are simple, real, and strictly negative.

An analysis similar to that of Carslaw and Jaeger [3, p. 96] demonstrates that $\overline{v}(M;\,\lambda)$ satisfies a version of Jordan's lemma on the sequence of contours $\Gamma_n = \partial\omega_n$, where

$$\omega_n = \left\{ z \in \mathbb{C} \;\middle|\; |z + M^2/4| < n^2\pi^2, \quad \mathrm{Re}[z] \leq 0 \right\}.$$

Thus,

$$v(M;\,t) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \overline{v}(M;\,\lambda)e^{\lambda t}\,d\lambda = \sum \mathrm{Res}[\overline{v}(M;\,\lambda)e^{\lambda t}].$$

A calculation shows that, for all $t \geq 0$,

$$v(M;\,t) = \sum_{n=1}^{\infty} c_n(M)\exp(-\lambda_n(M)t), \tag{3.5}$$

where

$$c_1(M) = \begin{cases} \frac{2e^{-M/2}(\lambda_1 - M^2/4)}{\lambda_1^{1/2}(\lambda_1 - M/2)} & M \in [0,\,\infty)\setminus\{2\}, \\ 3e^{-(1+t)} & M = 2, \end{cases}$$

$$c_n(M) = \frac{2(-1)^{n+1}e^{-M/2}(\lambda_n - M^2/4)}{\lambda_n^{1/2}(\lambda_n - M/2)}, \qquad n \geq 2.$$

We state some of the conclusions from the calculation as a theorem.

THEOREM 3.1. *The terms in the series* (3.5) *are alternating and decreasing,*

$$e^{-\lambda_1 t} \le v(M;\, t) \le \min\left[1,\, c_1(M)e^{-\lambda_1 t}\right], \tag{3.6}$$

and $v(M;\, t) \sim c_1(M)e^{-\lambda_1 t}$ *for large* t. $\lambda_1(M)$ *decreases with* M; $\lambda_1(0) = \pi^2/4$, $\lambda_1(2) = 1$, *and*

$$\lambda_1(M) = M^2 e^{-M} + 2M^2(M-1)e^{-2M} + o\left(e^{-(3-\epsilon)M}\right) \qquad \text{for any } \epsilon > 0, \tag{3.7}$$

for large M. $c_1(M)$ *also decreases with* M; $c_1(0) = 4/\pi$, $c_1(2) = 3/e$, *and, for large* M,

$$c_1(M) = 1 + (M-3)e^{-M} + o\left(e^{-(2-\epsilon)M}\right) \qquad \text{for any } \epsilon > 0. \tag{3.8}$$

Proof. Since $\lambda_n(M) \ge \max[M/2,\, M^2/4]$, except when $n = 1$ and $M > 2$, $(-1)^{n+1}c_n$ decreases with λ_n, and thus with n. Hence the sequence $(c_n e^{-\lambda_n t})$ is alternating and decreasing for $t \ge 0$. This implies $v(M;\, t) \le c_1(M)\exp(-\lambda_1(M)t)$. On the other hand, the maximum principle implies that $v(M;\, t) \le 1$. For the lower bound, we observe that if $\phi_1(x)$ is the first eigenmode of L_M, $u(x,\, t) = e^{-\lambda_1 t}\phi_1(x)$ is a solution of (2.1) such that $u_0 = \phi_1$. Thus, the last part of (3.6) follows from Theorem 2.4. Since $0 < \lambda_1 < \lambda_n$ for $n > 1$, (3.5) implies that $v(M;\, t) \sim c_1(M)e^{-\lambda_1 t}$. The values of $\lambda_1(M)$ and $c_1(M)$, and the asymptotic estimates for large M, are obtained by solving (3.3) and (3.4). ∎

A simple consequence of Theorems 2.4 and 3.1 is that the first eigenvalue of the operator $L_{d,\,M}$ attains the minimum possible value for an eigenvalue of an elliptic operator L (independent of t), of the form indicated by (1.1)–(1.3). As regards the behavior of $v(M;\, t)$ and the series (3.6), we postpone further comment until the end of the section.

3.2. Complementary error function expansion. We now consider expansions of $v(M;\, t)$ suitable for small t in terms of iterated complementary error functions. $\overline{v}(M;\, \lambda)$, given by (3.2), may be re-expressed as a convergent geometric series, provided λ is sufficiently small. Considering the nth partial sum of this series, we define

$$\overline{v}_n(M;\, \lambda) = \frac{1}{\lambda} - \frac{2\gamma e^{-(M/2+\gamma)}}{\lambda(\gamma - M/2)} \sum_{r=0}^{n} (-1)^r \left[\frac{(\gamma + M/2)e^{-2\gamma}}{\gamma - M/2}\right]^r \tag{3.9}$$

and set $v_n(M;\, t) \equiv \mathcal{L}^{-1}\{\overline{v}_n(M;\, \lambda)\}(t)$, where $\mathcal{L}$ is the Laplace transform operator. To obtain an explicit expression for $v_n(M;\, t)$, we apply the following theorem stated by Carslaw and Jaeger [3, Theorem XII, p. 259].

THEOREM 3.2. *If the transform of* $f(t)$ *is* $F(\lambda)$, $\lambda \ge \lambda_0$, *and that of* $K(t,\, u)$ *is* $\phi(\lambda)e^{-u\psi(\lambda)}$, *where* $\phi(\lambda)$ *and* $\psi(\lambda)$ *are independent of* u, *and* $\psi(\lambda) \ge \lambda_0$ *for* $\lambda \ge \lambda_1$, *then*

$$\mathcal{L}^{-1}\left\{\phi(\lambda)F(\psi(\lambda))\right\}(t) = \int_0^\infty K(t,\, u)f(u)\, du. \tag{3.10}$$

Here, we note that for $\phi(\lambda) = 1/\gamma(\lambda)$ and $\psi(\lambda) = \gamma(\lambda) - M/2$,

$$\frac{1}{\lambda} - \overline{v}_n(M;\, \lambda) = \phi(\lambda)F_n(\psi(\lambda)),$$

where

$$F_n(\lambda) = \frac{2(\lambda + M/2)^2 e^{-(M+\lambda)}}{\lambda^2(\lambda + M)} \sum_{r=0}^{n} (-1)^r \left(1 + \frac{M}{\lambda}\right)^r e^{-r(M+2\lambda)}. \tag{3.11}$$

Henceforth, for standard transform results we quote the tables of Erdélyi et al. [5], from where we deduce that

$$K(t,\, u) = \frac{1}{\sqrt{\pi t}} \exp\left(-\left(\frac{u}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right)^2\right).$$

On the other hand, expanding (3.11),

$$F_n(\lambda) = \left(\frac{2}{\lambda} + \frac{M^2}{2\lambda^2(\lambda + M)}\right) \sum_{r=0}^{n} (-1)^r \sum_{s=0}^{r} \binom{r}{s} \frac{M^s e^{-(r+1)M} e^{-(2r+1)\lambda}}{\lambda^s}.$$

Using the tables [5], $f_n(t)$, the inverse of $F_n(\lambda)$, is given by

$$\begin{aligned}
&\sum_{r=0}^{n} (-1)^r \sum_{s=0}^{r} 2 \binom{r}{s} M^s e^{-(r+1)M} \frac{(t - (2r+1))^s}{s!} H(t - (2r+1)) \\
&+ \frac{1}{2}(e^{-Mt} - e^{-M} + Me^{-M}(t-1)) H(t-1) \\
&+ \sum_{r=1}^{n} \frac{(-1)^r}{2} \sum_{s=0}^{r-1} \binom{r-1}{s} M^{s+2} e^{-(r+1)M} \frac{(t - (2r+1))^{s+2}}{(s+2)!} H(t - (2r+1)),
\end{aligned}$$

where $H(t)$ is the Heaviside function. Evaluating (3.10), we find that

$$\begin{aligned}
&v_n(M;\, t) \\
&= 1 - \sum_{r=0}^{n} (-1)^r \sum_{s=0}^{r} 2 \binom{r}{s} e^{-(r+1)M} (2M\sqrt{t})^s \mathrm{i}^s \mathrm{erfc}\left(\frac{2r+1}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right) \\
&- \frac{1}{2}\left(\mathrm{erfc}\left(\frac{1}{2\sqrt{t}} + \frac{M\sqrt{t}}{2}\right) - e^{-M} \mathrm{erfc}\left(\frac{1}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right)\right) \\
&- Me^{-M}\sqrt{t}\,\mathrm{i}^1 \mathrm{erfc}\left(\frac{1}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right) \\
&- \sum_{r=1}^{n} (-1)^r \sum_{s=0}^{r-1} \binom{r-1}{s} \frac{e^{-(r+1)M}}{2} (2M\sqrt{t})^{s+2} \mathrm{i}^{s+2} \mathrm{erfc}\left(\frac{2r+1}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right),
\end{aligned} \tag{3.12}$$

where (see [1, Chapter 7]),

$$i^n \mathrm{erfc}(x) = \int_x^\infty \frac{(t-x)^n}{n!} e^{-t^2}\, dt, \quad n = 0,\, 1,\, 2,\, \ldots; \quad \mathrm{erfc}(x) \equiv \mathrm{i}^0 \mathrm{erfc}(x), \quad x \in \mathbb{R}$$

and $i^n \mathrm{erfc}(x)$ satisfies the recurrence relation

$$\mathrm{i}^n \mathrm{erfc}(x) = -\frac{x}{n} \mathrm{i}^{n-1} \mathrm{erfc}(x) + \frac{1}{2n} \mathrm{i}^{n-2} \mathrm{erfc}(x), \quad n \geq 1, \tag{3.13}$$

where $\mathrm{i}^{-1} \mathrm{erfc}(x) = (2/\sqrt{\pi}) \exp(-x^2)$.

THEOREM 3.3. *The sequence* $(v_n(M;\, t))$ *is such that, for* $n \geq 0$,

$$v_{2n}(M;\, t) \leq v(M;\, t) \leq v_{2n+1}(M;\, t), \qquad (M,\, t) \in [0,\, \infty) \times (0,\, \infty), \tag{3.14}$$

and for fixed $(M,\, t) \in [0,\, \infty) \times (0,\, \infty)$, $\lim_{n\to\infty} v_n(M;\, t) = v(M;\, t)$. *Furthermore, for small* t,

$$v(M;\, t) = 1 - \frac{2e^{-M/2}\sqrt{t}}{\sqrt{\pi}} \exp\left(-\frac{1}{4t}\right)[1 + 2M\sqrt{t} + O(t)]. \tag{3.15}$$

Proof. Subtracting (3.9) from (3.2), we obtain

$$\overline{v}(M;\, \lambda) - \overline{v}_n(M;\, \lambda) = (-1)^n \left[\frac{(\gamma + M/2)e^{-2\gamma}}{\gamma - M/2}\right]^{(n+1)} \left(\frac{1}{\lambda} - \overline{v}(M;\, \lambda)\right).$$

Applying the iterated product formula for inverse Laplace transforms and using the positivity of $\mathcal{L}^{-1}\{1/\lambda - \overline{v}(M;\, \lambda)\}(t)$ and $\mathcal{L}^{-1}\{(\gamma + M/2)e^{-2\gamma}/(\gamma - M/2)\}(t)$, we conclude that $(-1)^n(v(M;\, t) - v_n(M;\, t))$ is nonnegative, and (3.14) follows.

Consequently,

$$\begin{aligned}
&|v(M;\, t) - v_n(M;\, t)| \\
&\quad \leq |v_{n+1}(M;\, t) - v_n(M;\, t)| \\
&\quad \leq \sum_{s=0}^{n+1} 2 \begin{pmatrix} n+1 \\ s \end{pmatrix} e^{-(n+2)M} (2M\sqrt{t})^s \mathrm{i}^s \mathrm{erfc}\left(\frac{2n+3}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right) \\
&\quad + \sum_{s=0}^{n} \begin{pmatrix} n \\ s \end{pmatrix} \frac{e^{-(n+2)M}}{2} (2M\sqrt{t})^{s+2} \mathrm{i}^{s+2} \mathrm{erfc}\left(\frac{2n+3}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right),
\end{aligned} \tag{3.16}$$

where we have used (3.12).

When $2n + 3 \geq Mt + 2\sqrt{t}$, the estimate

$$i^n \mathrm{erfc}(x) < \frac{2}{\sqrt{\pi}} \frac{e^{-x^2}}{(2x)^{n+1}}, \quad x > 0, \quad n \geq 0,$$

implies that the right-hand side of (3.16) is bounded above by

$$\frac{2e^{-M/2}}{\sqrt{\pi}} (1 + M\sqrt{t})^{n+2} \exp\left(-\left(\frac{(2n+3)^2}{4t} + \frac{M^2 t}{4}\right)\right). \tag{3.17}$$

Hence, $\lim_{n\to\infty} v_n(M;\, t) = v(M;\, t)$.

When $Mt + 2\sqrt{t} \leq 3$, (3.17) implies that $v(M;\, t) = v_0(M;\, t) + O\left(e^{-9/(4t)}\right)$. Expanding the five terms comprising $v_0(M;\, t)$ and using the recurrence relation (3.13) and the asymptotic formula [1, p. 298],

$$\sqrt{\pi} x e^{x^2} \mathrm{erfc}(x) \sim 1 + \sum_{m=1}^{\infty} (-1)^m \frac{1 \cdot 3 \cdots (2m-1)}{(2x^2)^m}, \qquad \text{as } z \to \infty, \tag{3.18}$$

one obtains (3.15). ∎

3.3. The behavior of $v(M; t)$ and numerical analysis. Qualitatively, we know (Theorem 3.3) that $\lim_{t\to 0} v_t(M; t) = 0$, and (Theorems 2.2 and 3.1) that $v_t(M; t)/v(M; t)$ is a decreasing function of t, tending to $-\lambda_1$ as $t \to \infty$. The asymptotic behavior of $v(M; t)$ in t-neighborhoods of 0 and ∞ is given by Theorems 3.3 and 3.1, respectively. The nonpolynomially slow initial decay given by (3.15) is well known in the case $M = 0$; see, e.g., Carslaw and Jaeger [3]. For $M > 0$, we see that this is modified by the constant $e^{-M/2}$. For large times, Theorem 3.1 states that $v(M; t) \sim c_1(M)e^{-\lambda_1(M)t}$, which is also a familiar type of result for eigenmode expansions in the selfadjoint case. However, for large M, $\lambda_1(M) \sim M^2e^{-M}$. This is so small for large M that, even for $M = 50$, it is on the limit of double precision (16-digit decimal) arithmetic to detect the decay of $v(M; t)$ over any interval of the form $[t, t + 10]$. We also remark that the value of $1/\lambda_1$ is highly sensitive to changes in M.

In the t-range between these asymptotic estimates, the behavior of $v(M; t)$ is not so immediately apparent from either (3.5) or (3.12), and a numerical summation of a truncation of one of these expansions is required. Considering small t initially, we note that Theorems 2.2 and 3.3 imply

$$v(M; t) \geq v(0, t) \geq v_0(0; t) = 1 - 2\,\mathrm{erfc}(1/(2\sqrt{t})), \qquad (M, t) \in [0, \infty)^2.$$

We deduce that the decay of $v(M; t)$ is undetectable by double precision arithmetic for $t \leq 0.007$. So, there is no point in summing (3.5) on a computer for such t. However, in the case of (3.12), because 1 is the first term in the expansion, $1 - v(M; t)$ may be well approximated by $1 - v_0(M; t)$ over this t-range.

Without descending into too detailed an analysis, for $t \in (0.007, 0.1]$ and $M \in [0, 10]$, the decay in $v(M; t)$ becomes detectable, although obviously the precise values of t depend upon M. Loosely speaking, expansion (3.12) is at an advantage over (3.5) in this range, since the latter requires many terms for accuracy. The appropriate number of terms to take in (3.12) is indicated by the error estimate (3.16), since $3 > Mt + 2\sqrt{t}$ for this parameter range. We note that instructions on the numerical calculation of $\mathrm{i}^n\mathrm{erfc}(x)$ for $x > 0$ are given in [1, Chapter 7]. Essentially, this is cheap to evaluate, using a *backward* iteration of (3.13) (to avoid cancellation errors), once the values of $\mathrm{erfc}(x)$ and $(2/\sqrt{\pi})\exp(-x^2)$ are known. However, a recalculation must be performed for each value of t.

For $M > 10$, $1 - v(M; t)$ is much smaller for $t \in [0, 1]$, and (3.12) remains appropriate over this range. Considering (3.16) again, in the case where $2n + 3 \leq Mt + 2\sqrt{t}$, the crude estimate

$$i^n\mathrm{erfc}(x) \leq 2\exp(x_-), \qquad x \in \mathbb{R}, \;\; n = 0, 1, \ldots$$

implies that

$$|v(M; t) - v_n(M; t)| \leq 4(1 + 2M\sqrt{t})^{n+2}e^{-(n+2)M}e^{M\sqrt{t}/2}. \tag{3.19}$$

Thus for $t \in [0, 1]$, $v_3(M; t)$ suffices for $M = 10$, while for $M \geq 40$, $v_0(M; t)$ is good enough. We remark that $\mathrm{i}^n\mathrm{erfc}(x)$ should be evaluated by *forward* iteration of (3.13) when $x < 0$.

For $t > 0.1$ and $M \in [0, 10]$, a truncation to m terms of (3.5) becomes more competitive. We remark that λ_n and the coefficients of $e^{-\lambda_n t}$ may be quite cheaply calculated from (3.3) and (3.4) using a Newton–Raphson iteration and that, for fixed M, the same coefficients serve for all t. As stated in Theorem 3.1, the series (3.5) is

alternating and decreasing. Thus, the error in the truncation after m terms is less in magnitude than the $(m+1)$st term. For $n > 1$ and $\theta(M, n) \in (0, 1/2)$,

$$c_n(M)e^{-\lambda_n t} = (-1)^{n+1}\frac{2e^{-M/2}(n-\theta)^2\pi^2\exp(-((n-\theta)^2\pi^2 + M^2/4)t)}{((n-\theta)^2\pi^2 + M^2/4 - M/2)((n-\theta)^2\pi^2 + M^2/4)^{1/2}}.$$

Hence, to achieve an error of less than ϵ, it is sufficient that m satisfy

$$\sqrt{\frac{[\log(m\pi\epsilon/2) + M/2]_-}{\pi^2 t}} < m, \qquad \text{where } [x]_- \equiv (|x| - x)/2.$$

So, $m = 7$ suffices for all $t > 0.1$ if $\epsilon = 10^{-16}$. As t increases, successively fewer terms are needed.

Lastly, we note that for large M and $t \in [0, 1]$,

$$1 - c_1(M)\exp(-\lambda_1 t) \approx e^{-M}(M^2 t - (M-3)),$$
$$|c_2(M)|\exp(-\lambda_2 t) \approx \frac{2\pi^2 e^{-M/2}}{\sqrt{(\pi^2 + M^2/4)}(\pi^2 + M^2/4 - M/2)}e^{-(M^2/4+\pi^2)t}.$$

Thus, when $t \geq 3/M$, the second and subsequent terms in (3.5) are entirely negligible, and the behavior of $v(M; t)$ is essentially exponential decay. This indicates that increasing M has a more pronounced effect on the large time behavior than it does on the initial decay.

4. Elliptic bounds. Here, we initially consider means of calculating the L_∞ bound given by Lemma 2.5 for a cubical domain. Subsequently, we consider bounds in a spherical geometry for somewhat more general elliptic operators, under slightly different assumptions on the advection.

4.1. Problems on the unit cube. Here, in the domain $\Omega = (-1, 1)^N$, we consider the problem

$$Lu + \lambda u \equiv -\sum_{i=1}^{N} a_i(x)\frac{\partial^2 u}{\partial x_i^2} + \sum_{i=1}^{N} b_i\frac{\partial u}{\partial x_i} + \lambda u = f, \qquad x \in \Omega, \tag{4.1}$$
$$u(x) = 0, \qquad x \in \partial\Omega.$$

It is assumed that $\lambda \geq 0$, $a \in C(\overline{\Omega})$, b, $f \in L_\infty(\Omega)$,

$$\inf_{x\in\Omega} a_i(x) \geq d_i > 0, \qquad \|b_i(x)\|_\infty \leq M_i, \tag{4.2}$$

and that f is nonnull.

THEOREM 4.1. *Suppose that $u(x)$ is the solution of a boundary value problem of the form (4.1), for some $\lambda \geq 0$, where the coefficients satisfy the above assumptions. Then,*

$$\frac{\|u\|_\infty}{\|f\|_\infty} \leq W(d, M; 0, \lambda) = \int_0^\infty e^{-\lambda t}\prod_{i=1}^{N} v(M_i/d_i; d_i t)\, dt. \tag{4.3}$$

This upper bound is attained when $a_i(x) = d_i$, $b_i(x) = M_i \operatorname{sign}[x_i]$, and f is a nonzero constant.

Proof. Lemma 2.5 implies that $W(d,\, M;\, x,\, \lambda)$ satisfies the equation

$$-\sum_{i=1}^{N} d_i \frac{\partial^2 u}{\partial x_i^2} - \sum_{i=1}^{N} M_i \left| \frac{\partial u}{\partial x_i} \right| + \lambda u = 1.$$

The rest of the proof resembles that of Theorem 2.4 and is omitted. ∎

When $N = 1$, we have already found a simple explicit formula, given by (3.2). An explicit formula is also available for $W(d,\, M;\, 0,\, \lambda)$ for $N > 1$, using the eigenmode expansion (3.5). We begin by relabelling the coefficients of (3.5) as follows:

$$v(M_i/d_i;\, t) = \sum_{n=1}^{\infty} c_n^i \exp(-\lambda_n^i t), \qquad \lambda_n^i = \lambda_n(M_i/d_i), \quad c_n^i = c_n^i(M_i/d_i). \tag{4.4}$$

From (4.3) we obtain

$$W(d,\, M;\, \lambda) = \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \cdots \sum_{n_N=1}^{\infty} \frac{c_{n_1}^1 \cdots c_{n_N}^N}{\lambda + d_1 \lambda_{n_1}^1 + \cdots + d_n \lambda_{n_N}^N}. \tag{4.5}$$

However, (4.5) is a very slowly converging series, unsuitable for computation.

To obtain practical bounds, for $N > 1$ and λ not too small, probably the most numerically efficient way of evaluating the integral is to rescale the time variable to $\tau = \lambda t$ and then apply Gauss–Laguerre quadrature—see [1, p. 923], with the values of $v(M;\, t)$ approximated in the way indicated at the end of section 3. However, the error term for this quadrature method is very difficult to evaluate in practice, so sharp guaranteed error bounds are not usually available.

Alternatively, setting $d_0 = \min[d_1,\, \ldots,\, d_N]$, we approximate

$$\begin{aligned} v(M_i/d_i;\, d_i t) &\approx 1, \qquad d_0 t \in [0,\, t^*], \\ v(M_i/d_i;\, d_i t) &\approx \sum_{n_i=1}^{m^*} c_{n_i}^i \exp(-d_i \lambda_{n_i}^i t), \qquad d_0 t > t^*, \end{aligned}$$

for some $t^* > 0$ and m^* an odd integer. Integrating exactly, one obtains

$$\begin{aligned} W(d,\, M;\, 0,\, \lambda) &\approx \frac{1 - e^{-\lambda t^*/d_0}}{\lambda} \\ &+ \sum_{n_1=1}^{m^*} \cdots \sum_{n_N=1}^{m^*} \frac{c_{n_1}^1 \cdots c_{n_N}^N \exp(-(\lambda + d_1 \lambda_{n_1}^1 + \cdots + d_N \lambda_{n_N}^N) t^*/d_0)}{\lambda + d_1 \lambda_{n_1}^1 + \cdots + d_N \lambda_{n_N}^N}. \end{aligned} \tag{4.6}$$

Since m^* is odd, (4.6) is an upper bound for (4.5). The analysis of section 3 may be used to show that if $t^* = 0.007$, $m^* = 23$, and $N \leq 10$ then the absolute error of (4.6) is less than 10^{-16}. However, for very large λ, (4.6) is insufficiently accurate to give a good approximation of $1/\lambda - W(d,\, M;\, 0,\, \lambda)$, because this corresponds in an Abelian or Tauberian sense to the subtle behavior of $1 - V(d,\, M;\, 0,\, t)$ near $t = 0$.

We note that Nm^* transcendental equations of the form (3.3) or (3.4) must be solved for (4.6), which is perfectly feasible for the parameters suggested. The sum on the right-hand side contains $(m^*)^N$ different terms, which when $m^* = 23$ is not too many to compute for the usual physical dimensions, $N \leq 3$. However, for N much bigger than 7 this method becomes impractical.

4.2. Problems on the unit ball. On $B \equiv B(\mathbb{R}^N; 0, 1)$, the unit ball in $\mathbb{R}^N$, we consider

$$(L+\lambda)u \equiv -\sum_{i,j=1}^{N} a_{ij}(x)\frac{\partial^2 u}{\partial x_i x_j} + \sum_{i=1}^{N} b_i(x)\frac{\partial u}{\partial x_i} + \lambda u = f, \qquad |x| < 1,$$
$$u(x) = 0, \qquad |x| = 1. \tag{4.7}$$

It is assumed that $\lambda \geq 0$, $a \in C(\overline{B})$, b, $f \in L_\infty(B)$,

$$\inf_{x\in\Omega} \sup_{\xi\in B} a_{ij}(x)\xi_i\xi_j \geq 1, \qquad \left\| \sum_{i=1}^{N} b_i^2 \right\|_\infty \leq K^2, \tag{4.8}$$

and that f is nonnull.

For $K \geq 0$, we define

$$L_K u \equiv -\Delta u + K\frac{x}{|x|}.\nabla u, \quad x \neq 0; \qquad L_K u \equiv -\Delta u, \quad x = 0, \tag{4.9}$$

and set $w(K; x, \lambda)$ to be the solution of (4.7) when $L = L_K$ and $f(x) \equiv 1$.

LEMMA 4.2. *$w(K; x, \lambda) = v(K; |x|, \lambda)$, where v, the solution of the problem*

$$-u_{rr} - (N-1)u_r + Ku_r + \lambda u = 1, \quad r \in (0, 1); \qquad u'(0) = u(1) = 0, \tag{4.10}$$

satisfies

$$v_r(r) < 0, \qquad v_{rr}(r) - \frac{v_r(r)}{r} < 0, \quad r \in (0, 1). \tag{4.11}$$

Furthermore, for $M(\cdot, \cdot, \cdot)$, the confluent hypergeometric Kummer function of the first kind (see [1, p. 504]*), and $r \in (0, 1)$,*

$$v(K; r, 0) = \int_r^1 \int_0^s (t/s)^{N-1} e^{K(s-t)}\, dt\, ds,$$
$$v(K; r, \lambda) = \frac{1}{\lambda}\left(1 - e^{(\gamma - K/2)(1-r)}\frac{M((N-1)\delta, N-1, 2\gamma r)}{M((N-1)\delta, N-1, 2\gamma)}\right), \quad \lambda > 0,$$
$$\gamma \equiv \sqrt{K^2/4 + \lambda}, \qquad \delta \equiv (1/2 - (K/4\gamma)). \tag{4.12}$$

Proof. We define the nonnegative sequence (b_n) by $b_1 = 0$, $b_2 = 1$,

$$b_{n+2} = \frac{K(n+1)}{(N+n)(n+2)} b_{n+1} + \frac{\lambda b_n}{(N+n)(n+2)}, \quad n \geq 1. \tag{4.13}$$

Defining $C(N, K, \lambda) = \sum_{n=2}^{\infty} b_n$, we consider the sequence (a_n), where

$$a_0 = \frac{C(N, K, \lambda)}{\lambda C(N, K, \lambda) + 2N}, \quad a_1 = 0, \quad a_2 = -\frac{a_0}{C(N, K, \lambda)}, \tag{4.14}$$

and $a_n = a_2 b_n$ for $n \geq 3$. We observe that $\sum_{n=0}^{\infty} a_n r^n$ satisfies (4.10).

The relationship between the respective solutions of (4.9) and (4.10) is clear. Since $w(K; x, \lambda)$ is unique, so is $v(K; r, \lambda)$, and thus,

$$v(r) = \sum_{n=0}^{\infty} a_n r^n. \tag{4.15}$$

(Coddington and Levinson [4] present a general theory of ODEs, including (4.9).) We note that $a_n < 0$ for $n \geq 2$, and therefore

$$v_r(r) = \sum_{n=2}^{\infty} n a_n r^{n-1} < 0; \qquad v_{rr}(r) - \frac{v_r(r)}{r} = \sum_{n=2}^{\infty} n(n+2) a_{n+2} r^n < 0, \qquad r \in (0,\, 1).$$

For $\lambda = 0$, (4.13) is elementary. For $\lambda > 0$, we define

$$y(2\gamma r) \equiv e^{(\gamma - K/2)r} \left(\frac{1}{\lambda} - v(r) \right), \quad r \in [0,\, 1].$$

$y(x)$ satisfies the following version of Kummer's equation; see [1, p. 504]:

$$xy_{xx} + ((N-1) - x)y_x - (N-1)\delta y = 0, \qquad x \in (0,\, 2\gamma), \tag{4.16}$$
$$y_x(0) = \delta y(0), \qquad y(2\gamma) = \frac{e^{\gamma - K/2}}{\lambda}.$$

(4.12) now follows for $\lambda > 0$ from the properties of Kummer's functions. ∎

THEOREM 4.3. *Suppose that $u(x)$ is the solution of* (4.1)*, under the accompanying assumptions, including* (4.8) *for some $K \geq 0$. Then,*

$$\frac{\|u\|_\infty}{\|f\|_\infty} \leq \sum_{n=0}^{\infty} \frac{K^n (N-1)!}{(n+2)(N+n)!}, \qquad \lambda = 0, \tag{4.17}$$

$$\frac{\|u\|_\infty}{\|f\|_\infty} \leq \frac{1}{\lambda} \left(1 - \frac{e^{\gamma - K/2}}{M((N-1)\delta,\, (N-1),\, 2\gamma)} \right), \qquad \lambda > 0. \tag{4.18}$$

These bounds are attained when $L = L_K$ and f is a nonzero constant.

Proof. Assuming $\|f\|_\infty = 1$, (4.9) and (4.11) imply that

$$\begin{aligned}
(L+\lambda)w &= -\sum_{i,\,j=1}^{N} a_{ij}(x) \frac{\partial^2 W}{\partial x_i \partial x_j} + \sum_{i=1}^{N} b_i(x) \frac{\partial W}{\partial x_i} + \lambda W \\
&= -\left(v_{rr}(|x|) - \frac{v_r(|x|)}{|x|} \right) \sum_{i,\,j=1}^{N} a_{ij}(x) \frac{x_i x_j}{|x|^2} - \frac{v_r(|x|)}{|x|} \sum_{1=1}^{N} a_{ii}(x) \\
&\quad + v_r(|x|) \sum_{i=1}^{N} b_i(x) x_i + \lambda v(|x|) \\
&\geq -v_{rr} - \frac{N-1}{|x|} v_r + K v_r + \lambda v = 1 \geq f(x)
\end{aligned}$$

for almost every $x \in B$. The maximum principle therefore implies that $u(x) \leq w(x)$; similarly, $u(x) \geq -w(x)$. Thus, $\|u\|_\infty \leq \|w(K;\, \cdot,\, \lambda)\|_\infty = w(K;\, 0,\, \lambda)$. The right-hand sides of (4.17) and (4.18) come from evaluations of (4.13), (4.14), and (4.12), respectively. ∎

We note that, from a computational point of view, the right-hand side is relatively easily evaluated, using the power series representation of the function $M(\cdot,\, \cdot, \cdot)$; see [1]. Lastly, we remark that it should be possible to obtain quantitative bounds on a parabolic version of (4.7) provided that a suitable expansion of the inverse Laplace transform of (4.18) can be found.

Acknowledgment. The author wishes to thank Professor J. F. Toland for suggesting this problem.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1964.
[2] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.
[3] H. S. Carslaw and J. C. Jaeger, *Operational Methods in Applied Mathematics*, 2nd ed., Oxford University Press, Oxford, 1948.
[4] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.
[5] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi, *Tables of Integral Transforms*, McGraw–Hill, New York, 1954.
[6] A. Friedman, *Partial differential equations of parabolic type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.
[7] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1983.
[8] A. T. Hill, *Estimates on the heat kernel of parabolic equations with advection*, SIAM J. Math. Anal., 28 (1997), pp. 1309–1317.
[9] A. T. Hill, *L_p Estimates on the Solutions of Elliptic Equations in Bounded Domains*, manuscript.
[10] M. H. Protter and H. F. Weinberger, *On the spectrum of general second order operators*, Bull. Amer. Math. Soc., 72 (1966), pp. 251–255.
[11] C. Pucci, *Operatori ellitici estremanti*, Ann. Mat. Pura Appl., 72 (1966), pp. 141–170.
[12] D. Sattinger, *Monotone methods in nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 21 (1972), pp. 979–1000.
[13] H. B. Stewart, *Generation of analytic semigroups by strongly elliptic operators*, Trans. Amer. Math. Soc., 199 (1974), pp. 141–162.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 736–756, May 1998

013

ANALYSIS OF CONCENTRATION AND OSCILLATION EFFECTS GENERATED BY GRADIENTS*

IRENE FONSECA†, STEFAN MÜLLER‡, AND PABLO PEDREGAL§

Abstract. A general theorem characterizing the interaction of concentrations and oscillations effects associated with sequences of gradients bounded in L^p, $p > 1$, is proved. The oscillations are recorded in the Young measure while the concentrations are encoded in the varifold.

Key words. oscillations, concentrations, Young measures, varifolds, quasiconvexity

AMS subject classifications. 49J45, 35B05

PII. S0036141096306534

1. Introduction.

Oscillatory phenomena and the characterization of limits of nonlinear quantities of oscillating sequences have been successfully analyzed by means of Young measures. These measures were first introduced by Young [38] to study nonconvex problems in optimal control theory and to provide the appropriate framework for the description of generalized minimizers in the calculus of variations. Recently Young measures have become an important tool in the study of nonlinear partial differential equations [10], [12], [13], [31], [33], [34], [35], [37] and the analysis of oscillatory behavior in nonconvex variational principles that arise in models of solid–solid phase transitions [7], [8]. Characterizations of Young measures associated with minimizing sequences of such functionals as well as with general sequences of gradients bounded in $L^p(\Omega;\mathbf{M})$ have been found in [21] and [22] (see also [29]). Here Ω is an open, bounded subset of $\mathbf{R}^N$ and $\mathbf{M} = \mathbf{M}^{m\times N}$ is the set of $m \times N$ matrices.

One of the main drawbacks of Young measures is that they miss completely concentration effects. Indeed, sequences may share the same Young measure and yet one may exhibit concentrations while the other does not. Several ways of understanding and manipulating concentrations have been proposed. We refer the reader to [14], [15], [18], [19], [24], [25], [30], [36], and [37] for some of these methods. Another possibility is using varifolds or indicator measures following the works [3], [4], [17], [28]. This is the point of view that we will take here, and we will focus on sequences that are constrained to be gradients. A similar approach has been employed in [2] for unconstrained sequences that are bounded in L^1.

The notion of a varifold has been used to describe certain nonlinear limits of oscillating measures, and it plays a role complementary to that of the Young measure.

*Received by the editors July 12, 1996; accepted for publication (in revised form) February 18, 1997.

http://www.siam.org/journals/sima/29-3/30653.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (fonseca@andrew.cmu.edu). The research of this author was partially supported by the Army Research Office and the National Science Foundation through the Center for Nonlinear Analysis, and by the National Science Foundation grants DMS-9201215 and DMS-9500531.

‡Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, 04103 Leipzig, Germany (sma@math.ethz.ch). The research of this author was partially supported by SFB256 at the University of Bonn and by the Center for Nonlinear Analysis at Carnegie Mellon University.

§ETSI Industriales, Universidad de Castilla-La Mancha, 13071 Ciudad Real, Spain (ppedregal@ind-cr.udm.es). The research of this author was supported by DGICYT (Spain) through grants PB93-0070 and PR94-304, by the Center for Nonlinear Analysis at Carnegie Mellon University, and by the Institute for Mathematics and its Applications at the University of Minnesota.

In fact, the Young measure associated with a sequence $\{f_j\}$, which is bounded in $L^p(\Omega;\mathbf{R}^d)$, describes the effect of oscillations on the limits of $\{\varphi(f_j)\}$ whenever the nonlinearity φ has growth of order strictly less than p, while the varifold describes the effect of concentrations on the limits of $\{\psi(f_j)\}$ when ψ grows asymptotically as the pth power. We will be more precise in section 3.

Our goal is to understand the relation between the varifold and the Young measure that are generated by a sequence of gradients which is bounded in $L^p(\Omega;\mathbf{M})$, $p>1$. We hope to address the case $p=1$ in a future work. A detailed description of Young measures generated by sequences of gradients bounded in $L^p(\Omega;\mathbf{M})$ was obtained in [22] (see Theorem 2.3 below).

To describe our main result we consider an open, bounded set $\Omega\subset\mathbf{R}^N$ and a sequence $\{f_j\}$ of functions bounded in $L^p(\Omega;\mathbf{R}^d)$ for some $p>1$. There exists a subsequence, still denoted $\{f_j\}$, and a family $\nu=\{\nu_x\}_{x\in\Omega}$ (called the *Young measure*) of probability measures ν_x on $\mathbf{R}^d$, as well as a nonnegative Radon measure Λ on $\Omega\times\mathcal{S}^{d-1}$ (called the *varifold*) with the following properties (see [6], [17], and section 3). For all continuous functions θ that vanish on $\partial\Omega$, $\theta\in\mathcal{C}_0(\Omega)$, all continuous functions φ on $\mathbf{R}^d$ with growth of order strictly less than p, i.e., $|\varphi(\xi)|\le C(1+|\xi|^r),\quad 1\le r<p$, and for all continuous functions ψ on $\mathbf{R}^d$ that are homogeneous of degree p, we have

$$\begin{aligned}\int_\Omega \theta(x)\varphi(f_j(x))\,dx &\to \int_\Omega \theta(x)\int_{\mathbf{R}^d}\varphi(\xi)\,d\nu_x(\xi)\,dx,\\ \int_\Omega \theta(x)\psi(f_j(x))\,dx &\to \int_{\Omega\times\mathcal{S}^{d-1}}\theta(x)\psi(\xi)\,d\Lambda(x,\xi)\\ &= \int_\Omega \theta(x)\int_{\mathcal{S}^{d-1}}\psi(\xi)\,d\lambda_x(\xi)\,d\pi(x).\end{aligned}$$

Here π is the projection of Λ onto Ω, λ_x are probability measures (for π-a.e. $x\in\Omega$), $\Lambda=\lambda\otimes\pi$, $\lambda=\{\lambda_x\}_{x\in\Omega}$ is the slicing decomposition of Λ [15], and $\mathcal{S}:=\mathcal{S}^{d-1}$ is the unit sphere in $\mathbf{R}^d$. In what follows, we will refer to (ν,Λ) as the *Young measure-varifold pair*.

If $\{u_j\}$ is a bounded sequence in $W^{1,p}(\Omega;\mathbf{R}^m)$, if $f_j=\nabla u_j$, and if the target space $\mathbf{R}^d$ is identified with the space $\mathbf{M}:=\mathbf{M}^{m\times N}$ of $m\times N$ matrices, we say that (ν,Λ) is a $W^{1,p}(\Omega)$-*Young measure-varifold pair*, and we abbreviate it by saying that (ν,Λ) is a YM-V pair. The Young measures that arise in such pairs (the so-called $W^{1,p}(\Omega)$ *Young measures*) were characterized in [22]. The following example shows that there are also restrictions Λ. Let $p=m=N$ and consider the N-homogeneous function $\psi(A):=\det A$. Then by the above

$$\int_\Omega \theta(x)\det\nabla u_j\,dx\to\int_\Omega\theta(x)\int_{\mathcal{S}}\det A\,d\lambda_x(A)\,d\pi(x).$$

On the other hand, we know that [5], [28] $\det\nabla u_j\overset{*}{\rightharpoonup}\det\nabla u$ in the sense of measures, where u is the weak limit of u_j in $W^{1,p}(\Omega;\mathbf{R}^N)$. We conclude that

$$(\det\nabla u)\,d\mathcal{L}^N=\left(\int_{\mathcal{S}}\det A\,d\lambda_x(A)\right)d\pi,$$

where $\mathcal{L}^N$ denotes the Lebesgue measure in $\mathbf{R}^N$. Therefore, if we write $\pi=\pi_a+\pi_s$, where π_a and π_s are, respectively, absolutely continuous and singular with respect to $\mathcal{L}^N$, we obtain that

$$\int_{\mathcal{S}} \det A \, d\lambda_x(A) = 0$$

for π_s a.e. $x \in \Omega$.

The main result of this paper is the following characterization theorem for YM-V pairs.

THEOREM 1.1. *Let $p > 1$. (ν, Λ) is a YM-V pair, where $\nu = \{\nu_x\}_{x\in\Omega}$, and $\Lambda = \{\lambda_x\}_{x\in\Omega} \otimes \pi$ if and only if*

1.

$$\nabla u(x) = \int_{\mathbf{M}} A \, d\nu_x(A), \quad \mathcal{L}^N \text{ a.e. } x \in \Omega,$$

for some $u \in W^{1,p}(\Omega; \mathbf{R}^m)$;

2.

$$\varphi(\nabla u(x)) \le \int_{\mathbf{M}} \varphi(A) \, d\nu_x(A), \quad \mathcal{L}^N \text{ a.e. } x \in \Omega,$$

for every quasiconvex φ for which the limit

$$\lim_{|A|\to\infty} \frac{\varphi(A)}{1 + |A|^p}$$

exists;

3.

$$\int_{\mathbf{M}} \psi(A) \, d\nu_x(A) \le \frac{d\pi}{d\mathcal{L}^N}(x) \int_{\mathcal{S}} \psi(A) \, d\lambda_x(A), \quad \mathcal{L}^N \text{ a.e. } x \in \Omega,$$

for every p-homogeneous, continuous function ψ such that $Q\psi(0) = 0$, where $Q\psi$ denotes the quasiconvexification of ψ;

4.

$$\int_{\mathcal{S}} \psi(A) \, d\lambda_x(A) \ge 0, \quad \pi_s \text{ a.e. } x \in \Omega,$$

for every p-homogeneous, continuous function ψ such that $Q\psi(0) = 0$, where π_s is the singular part of π with respect to $\mathcal{L}^N$.

We remind the reader (see [5], [11], [26]) that a function φ, defined on $\mathbf{M}$, is said to be *quasiconvex* if

$$\varphi(F) \le \frac{1}{|\Omega|} \int_\Omega \varphi(F + \nabla u(x)) \, d\mathcal{L}^N(x)$$

for all matrices F and all test functions $u \in W_0^{1,\infty}(\Omega; \mathbf{R}^m)$. If φ is not quasiconvex then its *quasiconvexification*, $Q\varphi$, is defined to be

$$Q\varphi(F) := \inf_u \frac{1}{|\Omega|} \int_\Omega \varphi(F + \nabla u(x)) \, dx$$

for all matrices F. The infimum is taken again over the set of functions u that belong to $W_0^{1,\infty}(\Omega; \mathbf{R}^m)$. In addition, if

$$|\varphi(\xi)| \le C \left(1 + |\xi|^p\right),$$

then

$$Q\varphi(F) = \inf_{u \in W_0^{1,p}(\Omega;\mathbf{R}^m)} \frac{1}{|\Omega|} \int_\Omega \varphi(F + \nabla u(x))\, dx.$$

Equivalently, $Q\varphi$ can be characterized as the largest quasiconvex function below φ [1], [11].

Parts 1 and 2 of Theorem 1.1, together with the integrability condition

$$\int_\Omega \int_{\mathbf{M}} |A|^p \, d\nu_x(A)\, dx < +\infty,$$

correspond to the characterization of the underlying Young measure and were proved in [21] and [22]. Part 3 provides the interaction between the Young measure and the absolutely continuous part of the varifold. Part 4 represents the restriction on the varifold in the set where the singular part π_s is concentrated. An interesting consequence of this result is that there are no restrictions on the singular measure π_s.

A key tool in the proof of the above theorem is the following decomposition result for sequences of gradients that are bounded in $L^p(\Omega;\mathbf{M})$ for some $p > 1$. It states, in particular, that every such sequence admits a subsequence that can be written as a sum of a sequence $\{\nabla z_j\}$ (of gradients!) whose pth power is equi-integrable and a remainder that converges to zero in measure (and hence almost uniformly). We may say that $\{\nabla z_j\}$ carries the oscillations, while the remainder accounts for the concentration effects.

LEMMA 1.2 (decomposition lemma). *Let $\Omega \subset \mathbf{R}^N$ be an open, bounded set and let $\{w_n\}$ be a bounded sequence in $W^{1,p}(\Omega;\mathbf{R}^m)$. There exists a subsequence, $\{w_j\}$, and a sequence $\{z_j\} \subset W^{1,p}(\Omega;\mathbf{R}^m)$ such that*

$$\mathcal{L}^N\left(\{z_j \neq w_j \text{ or } \nabla z_j \neq \nabla w_j\}\right) \to 0, \tag{1.1}$$

as $j \to \infty$, and $\{|\nabla z_j|^p\}$ is equi-integrable. If Ω is Lipschitz (or, more generally, an extension domain), then each z_j may be chosen to be a Lipschitz function.

Note that (1.1) implies that both sequences $\{\nabla z_j\}$ and $\{\nabla w_j\}$ generate the same Young measure.

Some remarks are in order. A similar result was derived independently by Kristensen [23]. The characterization of $W^{1,p}$-Young measures obtained in [22] (see Theorem 2.3) does not provide a way of identifying the oscillatory part and the concentrations on a given sequence $\{w_n\}$ bounded in $W^{1,p}(\Omega;\mathbf{R}^m)$; it asserts that a Young measure generated by $\{w_n\}$ will be generated also by a sequence $\{v_n\}$ with $\{|\nabla v_n|^p\}$ equi-integrable, but there is no direct relation between this and the former sequence. Also, the approach via [22] is rather indirect and implicitly relies on the lower semicontinuity results of Acerbi and Fusco [1]. In fact, once Lemma 1.2 is proved one can considerably shorten the arguments in [1] and [22] (see [27] for this point of view). Our proof of Lemma 1.2 (see section 4) still relies on essentially the same tools as [1], namely L^p estimates for maximal functions and Lipschitz extensions of $W^{1,p}$ functions off small sets, but we think that an approach that uses the decomposition result as a starting point might be more intuitive. Kristensen's proof [23], on the other hand, uses Iwaniec's estimates for perturbed Hodge decompositions. These estimates, however, in turn rely on L^p estimates involving the sharp maximal function. Finally, the result may be viewed as an L^p counterpart of a theorem by Zhang [39] which states that if $\{\nabla w_j\}$ is bounded in L^q for some $q > 1$ and generates a Young measure with

support contained in a ball $B = B(0,R) \subset \mathbf{M}$ (i.e., $\operatorname{supp}(\nu_x) \subset B$ for a.e. $x \in \Omega$) then there exists a sequence z_j with

$$|\nabla z_j| \le C(N)R \quad \text{and} \quad \mathcal{L}^N\left(\{z_j \neq w_j \text{ or } \nabla z_j \neq \nabla w_j\}\right) \to 0.$$

2. Preliminaries. For any number $p > 0$ consider the class

$$\mathcal{H}_p := \{f \in \mathcal{C}(\mathbf{M}) : f \text{ is positively homogeneous of degree } p\},$$

where $\mathcal{C}(\mathbf{M})$ is the set of continuous functions on $\mathbf{M}$. If $f \in \mathcal{H}_p$ then $f(tA) = t^p f(A)$ for all $A \in \mathbf{M}$ and $t > 0$. It is easy to show that homogeneity entails $Q\psi(0) = 0$ whenever $\psi \in \mathcal{H}_p$ and $Q\psi(0)$ is finite. Let X_p denote the set of continuous functions in $\mathbf{M}$ with growth of order at most p, i.e.,

$$X_p := \{\varphi \in \mathcal{C}(\mathbf{M}) : |\varphi(A)| \le C(1 + |A|^p)\}.$$

X_p is a Banach space under the natural norm

$$\|\varphi\| := \left\| \frac{\varphi(\cdot)}{1 + |\cdot|^p} \right\|_{L^\infty(\mathbf{M})}.$$

Finally, we consider the class

$$\mathcal{E}_p := \left\{ \varphi \in \mathcal{C}(\mathbf{M}) : \text{ there exists } f \in \mathcal{H}_p, \lim_{|A|\to\infty} \frac{\varphi(A) - f(A)}{|A|^p} = 0 \right\}.$$

Some properties of $\mathcal{E}_p$ are listed in the proposition below.

PROPOSITION 2.1.

1. *For every $\varphi \in \mathcal{E}_p$ there is a unique $f \in \mathcal{H}_p$ such that*

$$\lim_{|A|\to\infty} \frac{\varphi(A) - f(A)}{|A|^p} = 0.$$

The function f is the recession function of φ of degree p, φ_p^∞, defined by

$$\varphi_p^\infty(A) = \lim_{t\to\infty} \frac{\varphi(tA)}{t^p}.$$

2. *$\mathcal{E}_p$ is a closed, separable subspace of X_p and $\mathcal{H}_p$ is a closed subspace of $\mathcal{E}_p$.*
3. *If $f \in \mathcal{H}_p$ then*

$$\|f\| = \|f\|_{L^\infty(\mathcal{S})},$$

where $\mathcal{S}$ is the unit sphere in $\mathbf{M}$.

The proof of this proposition is elementary. The only fact that requires some comment is the separability of $\mathcal{E}_p$. Indeed, using the map

$$\mathcal{C}^\infty\left(\overline{B}(0,1)\right) \to \mathcal{E}_p,$$
$$\theta \to \left(A \mapsto \theta\left(\frac{A}{1+|A|}\right)\ |A|^p\right),$$

one can easily verify that $\mathcal{E}_p$ is isomorphic to the space of continuous functions on the unit ball of $\mathbf{M}$, equipped with the sup norm. This space is separable due to the

compactness of the unit ball. If we compare the space $\mathcal{E}_p$ with the space considered in [22],

$$E_p := \left\{ \varphi \in \mathcal{C}(\mathbf{M}) : \lim_{|A|\to\infty} \frac{\varphi(A)}{1+|A|^p} \text{ exists} \right\},$$

we see that $\mathcal{E}_p$ corresponds to the compactification of $\mathbf{M}$ by a sphere at ∞ while E_p corresponds to the one-point compactification. More general compactifications have been considered in [14], [29], [30], and [31].

We will use the following lemma, whose proof is elementary and left to the reader.

PROPOSITION 2.2. *If ψ is Lipschitz continuous on the unit sphere $\mathcal{S}$ and homogeneous of degree p, $p \geq 1$, then there is a constant $C > 0$ (depending on ψ) such that*

$$|\psi(A) - \psi(B)| \leq C \left(|A|^{p-1} + |B|^{p-1} \right) |A - B|$$

for any pair of matrices A, B.

A remark that will be used often in sections 5 and 6 is the following. Given a family of probability measures $\nu = \{\nu_x\}_{x\in\Omega}$ and a sequence of functions $\{f_j\}$ bounded in $L^p(\Omega; \mathbf{R}^d)$, it can be shown that if

$$\lim_{j\to\infty} \int_\Omega \theta(x) \varphi(f_j(x))\, dx = \int_\Omega \theta(x) \int_{\mathbf{R}^d} \varphi(\xi)\, d\nu_x(\xi)\, dx \tag{2.1}$$

for all $\theta \in \mathcal{C}_0(\Omega)$, $\varphi \in \mathcal{C}_0^\infty(\mathbf{R}^d)$, then (2.1) still holds for all $\varphi \in \mathcal{C}(\mathbf{R}^d)$ such that $\{\varphi(f_j)\}$ is equi-integrable and, in particular, for all φ on $\mathbf{R}^d$ which grow slower than $1 + |\xi|^p$. Therefore $\nu = \{\nu_x\}_{x\in\Omega}$ is *the* Young measure associated with $\{f_j\}$. We conclude that in order to identify the Young measure generated by $\{f_j\}$, it suffices to study the limits (2.1) for $\theta \in \mathcal{C}_0(\Omega)$ and $\varphi \in \mathcal{C}_0^\infty(\mathbf{R}^d)$. Also, it can be shown that

$$\int_\Omega \int_{\mathbf{R}^d} |\xi|^p \, d\nu_x(\xi)\, dx < \infty. \tag{2.2}$$

The main result in [22] is a characterization of $W^{1,p}$-Young measures in terms of Jensen's inequality for quasiconvex functions.

THEOREM 2.3. *Let $p > 1$. Then $\nu = \{\nu_x\}_{x\in\Omega}$ is a $W^{1,p}$-Young measure if and only if*

1.

$$\nabla u(x) = \int_{\mathbf{M}} A\, d\nu_x(A), \quad \mathcal{L}^N \text{ a.e. } x \in \Omega,$$

for some $u \in W^{1,p}(\Omega; \mathbf{R}^m)$;

2.

$$\varphi(\nabla u(x)) \leq \int_{\mathbf{M}} \varphi(A)\, d\nu_x(A), \quad \mathcal{L}^N \text{ a.e. } x \in \Omega,$$

for every quasiconvex φ for which the limit

$$\lim_{|A|\to\infty} \frac{\varphi(A)}{1+|A|^p}$$

exists;

3.

$$\int_\Omega \int_{\mathbf{M}} |A|^p \, d\nu_x(A)\, dx < \infty.$$

3. The representation formula. We introduced the space $\mathcal{E}_p$ in order to recover weak limits associated with sequences $\{\varphi(\nabla u_j)\}$ for $\varphi \in \mathcal{E}_p$ and any sequence $\{u_j\}$ that is bounded in $W^{1,p}(\Omega;\mathbf{R}^m)$. The representation of weak limits for such functions in terms of Young measures is only valid if one can rule out concentration effects (see [9]). To account for possible development of concentrations, we associate with $\{\nabla u_j\}$ a measure Λ on $\Omega \times \mathcal{S}$ called the *varifold associated with* $\{\nabla u_j\}$. We first recall that for an $\mathbf{M}$-valued Radon measure μ on an open set Ω the polar decomposition (see [17]) is given by $d\mu = \alpha\, d\lambda$, where λ is the total variation of μ and $\alpha : \Omega \to \mathcal{S}$ is the density of μ, i.e., the Radon–Nikodym derivative of μ with respect to its total variation λ (see [16]). A *varifold* is a nonnegative measure on $\Omega \times \mathcal{S}$. By slicing arguments (see [15]), every such measure Λ can be written in the form $\Lambda = \{\lambda_x\}_{x\in\Omega} \otimes \pi$, where π is a measure on Ω and λ_x are probability measures on $\mathcal{S}$. By the Radon–Nikodym theorem we may further write $\pi = \pi_a \mathcal{L}^N + \pi_s$, where $\pi_a := \frac{d\pi}{d\mathcal{L}^N}$ and π_s is singular with respect to $\mathcal{L}^N$. We first recall that every bounded sequence of $\mathbf{M}$-valued Radon measures has (up to a subsequence) a varifold limit (see [17]).

THEOREM 3.1. *Let $\{\mu_j\}$ be a sequence of $\mathbf{M}$-valued measures on Ω with polar decomposition $\alpha_j d\lambda_j$. Assume that $\mu_j \overset{*}{\rightharpoonup} \mu$ in the sense of measures. There exists a subsequence, still denoted $\{\mu_j\}$, and a nonnegative, finite, Radon measure $\Lambda = \{\lambda_x\}_{x\in\Omega} \otimes \pi$ on $\Omega \times \mathcal{S}$ such that for every $f \in \mathcal{C}_0(\Omega \times \mathbf{R}^d)$*

$$\begin{aligned}\lim_{j\to\infty} \int_\Omega f(x, \alpha_j(x))\, d\lambda_j(x) &= \int_{\Omega\times\mathcal{S}} f(x,y)\, d\Lambda(x,y)\\ &= \int_\Omega \int_{\mathcal{S}} f(x,y)\, d\lambda_x(y)\, d\pi(x).\end{aligned}$$

Given a sequence $\{u_n\}$, bounded in $W^{1,p}(\Omega;\mathbf{R}^m)$, we consider the bounded sequence of $\mathbf{M}$-valued Radon measures $\{|\nabla u_n|^{p-1}\, \nabla u_n \mathcal{L}^N\}$. According to Theorem 3.1, associated with a subsequence there exists a varifold, and this suggests the following definition.

DEFINITION 3.2. *A finite, Radon measure Λ supported on $\Omega \times \mathcal{S}$ is a $W^{1,p}$-varifold if there exists a bounded sequence in $W^{1,p}(\Omega;\mathbf{R}^m)$, $\{u_n\}$, such that for every $f \in \mathcal{C}_0(\Omega \times \mathbf{M})$*

$$\lim_{n\to\infty} \int_\Omega f\left(x, \frac{\nabla u_n}{|\nabla u_n|}\right) |\nabla u_n|^p\, dx = \int_{\Omega\times\mathcal{S}} f(x,A)\, d\Lambda(x,A).$$

In particular, if ψ is homogeneous of degree p and $\theta \in \mathcal{C}_0(\Omega)$ then

$$\begin{aligned}\lim_{n\to\infty} \int_\Omega \theta(x)\psi(\nabla u_n)\, dx &= \int_{\Omega\times\mathcal{S}} \theta(x)\psi(A)\, d\Lambda(x,A)\\ &= \int_\Omega \theta(x) \int_{\mathcal{S}} \psi(A)\, d\lambda_x(A)\, d\pi(x).\end{aligned}$$

In order to see how the YM-V pair determines the limits of $\{\varphi(\nabla u_n)\}$ for sequences $\{u_n\}$ bounded in $W^{1,p}(\Omega;\mathbf{R}^m)$ and having oscillatory and concentrating features, consider $\varphi \in \mathcal{E}_p$. By definition,

$$\lim_{|A|\to\infty} \frac{\varphi(A) - \varphi_p^\infty(A)}{|A|^p} = 0,$$

which implies that $\{\varphi(\nabla u_n) - \varphi_p^\infty(\nabla u_n)\}$ is weakly relatively compact in $L^1(\Omega)$. For this sequence, the representation in terms of the Young measure is valid. On the other hand, $\varphi_p^\infty \in \mathcal{H}_p$, and the limit for $\{\varphi_p^\infty(\nabla u_n)\}$ is therefore given by the varifold. Hence, we have the representation formula

$$\lim_{n\to\infty} \int_\Omega \theta(x)\varphi(\nabla u_n(x))\,dx = \int_\Omega \theta(x) \int_{\mathbf{M}} \left(\varphi(A) - \varphi_p^\infty(A)\right)\,d\nu_x(A)\,dx + \int_\Omega \theta(x) \int_{\mathcal{S}} \varphi_p^\infty(A)\,d\lambda_x(A)\,d\pi(x).$$

It is this formula that motivated our study of YM-V pairs.

Examples.

1. *Oscillations.*

Let $\Omega \subset \mathbf{R}^N$ be an open, bounded set, and let $p > 1$. Given $f \in W_0^{1,p}(Q;\mathbf{R}^m)$, we extend f to $\mathbf{R}^N$ periodically, with period $Q := (-1/2, 1/2)^N$. Fix $A \in \mathbf{M}$ and define

$$u_n(x) := Ax + \frac{1}{n} f(nx).$$

It is easy to check that $u_n \rightharpoonup Ax$ in $W^{1,p}(\Omega;\mathbf{R}^m)$, and the YM-V pair generated by this sequence is (ν, Λ), where

$$\langle \nu_x, \varphi\rangle := \int_Q \varphi(A + \nabla f(y))\,dy$$

for $\mathcal{L}^N$ a.e. $x \in \Omega$ and for all $\varphi \in C_0(\mathbf{M})$, and $\Lambda = \{\lambda_x\}_{x\in\Omega} \otimes \pi$ with $\pi := ||A + \nabla f||^p_{L^p(Q;\mathbf{M})}\,\mathcal{L}^N$ and

$$\langle \lambda_x, \psi\rangle := \frac{1}{||A + \nabla f||^p_{L^p(Q;\mathbf{M})}} \int_Q \psi\left(\frac{A + \nabla f(y)}{|A + \nabla f(y)|}\right) |A + \nabla f(y)|^p\,dy$$

for π a.e. $x \in \Omega$ and all $\psi \in C(\mathcal{S})$.

2. *Concentrations.*

We consider $f \in W_0^{1,p}(Q;\mathbf{R}^m)$ extended to $\mathbf{R}^N$ by zero, where $Q := (-1/2, 1/2)^N$. Fix $p > 1$, choose $x_0 \in \Omega$, and set

$$u_n(x) := n^{-1+N/p}\, f(n(x - x_0)).$$

Then $u_n \rightharpoonup 0$ in $W^{1,p}(\Omega;\mathbf{R}^m)$, and the YM-V pair generated by this sequence is (ν, Λ), where $\nu_x := \delta_0$ for $\mathcal{L}^N$ a.e. $x \in \Omega$, $\Lambda = \{\lambda_x\}_{x\in\Omega} \otimes \pi$, with $\pi := ||\nabla f||^p_{L^p(Q;\mathbf{M})}\,\delta_{x_0}$ and

$$\langle \lambda_x, \psi\rangle := \frac{1}{||\nabla f||^p_{L^p(Q;\mathbf{M})}} \int_Q \psi\left(\frac{\nabla f(y)}{|\nabla f(y)|}\right) |\nabla f(y)|^p\,dy$$

for π a.e. $x \in \Omega$ and all $\psi \in C(\mathcal{S})$.

Indeed, we claim that if $\varphi \in C_0(\mathbf{M})$ is such that $\varphi(0) = 0$ then

$$\lim_{n\to\infty} \int_\Omega \theta(x)\varphi(\nabla u_n(x))\,dx = 0$$

for all $\theta \in C_0(\Omega)$. Fix $\varepsilon > 0$ and $1 < q < p$. We may find $C(\varepsilon) > 0$ such that $|\varphi(A)| \le \varepsilon + C(\varepsilon)|A|^q$ for all $A \in \mathbf{M}$. Therefore,

$$\begin{aligned}\left|\int_\Omega \theta(x)\varphi(\nabla u_n(x))\,dx\right| &\le \varepsilon||\theta||_{L^\infty} + C(\varepsilon)||\theta||_{L^\infty} n^{Nq/p}\int_\Omega |\nabla f(n(x-x_0))|^q\,dx \\ &\le \varepsilon||\theta||_{L^\infty} + C(\varepsilon)||\theta||_{L^\infty} n^{N(q/p-1)}||\nabla f||^q_{L^q(Q;\mathbf{M})},\end{aligned}$$

and we conclude by letting $n \to \infty$ and then $\varepsilon \to 0$. Also, given $\psi \in C(\mathcal{S})$ and for n large enough we have

$$\begin{aligned}&\int_\Omega \theta(x)\psi\left(\frac{\nabla u_n(x)}{|\nabla u_n(x)|}\right)|\nabla u_n(x)|^p dx \\ &\qquad = \int_\Omega \theta(x)\psi\left(\frac{\nabla f(n(x-x_0))}{|\nabla f(n(x-x_0))|}\right) n^N |\nabla f(n(x-x_0))|^p\,dx \\ &\qquad = \int_Q \theta\left(x_0 + \frac{1}{n}y\right)\psi\left(\frac{\nabla f(y)}{|\nabla f(y)|}\right)|\nabla f(y)|^p\,dy.\end{aligned}$$

Letting $n \to \infty$, we deduce that

$$\int_\Omega \int_{\mathcal{S}} \theta(x)\psi(A)\,d\lambda_x(A)\,d\pi(x) = \theta(x_0)\int_Q \psi\left(\frac{\nabla f(y)}{|\nabla f(y)|}\right)|\nabla f(y)|^p\,dy.$$

3. *Generalized Solutions to PDEs.*

In the previous example it was clear that the behavior of the sequence was captured by the varifold, and little information was available through the Young measure $\nu_x = \delta_0$. Other cases where the varifold plays an important role include situations where one seeks the effective energy

$$\liminf \int_\Omega f(\nabla u_n)\,dx$$

associated with a sequence $\{u_n\}$ bounded in $W^{1,p}(\Omega;\mathbf{R}^m)$, with f nonconvex and behaving asymptotically at infinity as $C(1+|A|^p)$.

The first applications of Young measures to evolution equations, precisely to conservation laws where oscillations may be present, were provided by Tartar (see [33], [34], [35]). Later, the need to introduce a measure accounting for the development of concentrations (when L^∞ bounds are not available), possibly with conjunction with propagation of oscillations, was pointed out by Diperna and Majda (see [14]). They introduced the notion of generalized Young measure triplets (μ, ν^1, ν^2) associated with a (subsequence of a) sequence $\{v_n\}$ bounded in L^2. To establish the parallel between the YM-V (ν, Λ) and (μ, ν^1, ν^2), it suffices to set

$$\mu := \pi, \quad \nu^1 := \frac{1+|\cdot|^2}{1+\frac{d\pi}{d\mathcal{L}^N}}\,\nu_x, \quad \nu^2 := \lambda_x.$$

Diperna and Majda [14] proved that the generalized Young measure generated by a sequence of classical solutions of the two-dimensional Euler equation with uniformly bounded local kinetic energy is a measure-valued solution of the incompressible Euler equation.

4. Proof of the decomposition lemma. In this section we will prove Lemma 1.2. As mentioned in the introduction, our argument uses maximal functions and their properties, and we recall some well-known facts (see [32]).

Given a Borel measurable function $u : \mathbf{R}^N \to \mathbf{R}^d$, the *maximal function* of u is defined by

$$M(u)(x) := \sup_{r>0} \frac{1}{|B(x,r)|} \int_{B(x,r)} |u(y)|\, dy.$$

If $u \in W^{1,p}(\mathbf{R}^N; \mathbf{R}^m)$ then we set

$$M^*(u)(x) := M(u)(x) + M(\nabla u)(x),$$

and if $p > 1$, then

$$\|M^*(u)\|_{L^p(\mathbf{R}^N)} \le C(N,p) \|u\|_{W^{1,p}(\mathbf{R}^N;\mathbf{R}^m)}. \tag{4.1}$$

LEMMA 4.1. *Let $p > 1$ and let $w \in W^{1,p}(\mathbf{R}^N; \mathbf{R}^m)$. Given $\lambda > 0$ there exists a Lipschitz function z in $\mathbf{R}^N$ such that $w = z$ on $\{M(\nabla w) < \lambda\}$ and the Lipschitz constant for z is bounded by $C(N)\lambda$, where $C(N)$ is a constant depending only upon dimension.*

For the proof see, e.g., [16].

The proof of Lemma 1.2 will be divided into two steps. In the first step we consider an *extension domain* Ω, i.e., an open, bounded set Ω for which there exists an extension operator $T : W^{1,p}(\Omega; \mathbf{R}^m) \to W^{1,p}(\mathbf{R}^N; \mathbf{R}^m)$ such that

$$Tu(x) = u(x), \quad x \in \Omega, \qquad \|Tu\|_{W^{1,p}(\mathbf{R}^N;\mathbf{R}^m)} \le C \|u\|_{W^{1,p}(\Omega;\mathbf{R}^m)}.$$

In the second step we remove this restriction on Ω, generalizing the result for arbitrary open sets.

Proof of Lemma 1.2.

Step 1. Assume that Ω is an extension domain. Let $\{w_n\}$ be a bounded sequence in $W^{1,p}(\Omega; \mathbf{R}^m)$. In what follows, we identify w_n with its extension $Tw_n \in W^{1,p}(\mathbf{R}^N; \mathbf{R}^m)$.

By (4.1) the sequence $\{M(\nabla w_n)\}$ is bounded in $L^p(\mathbf{R}^n)$, so (see [6] and (2.2)) there exists a subsequence (not relabeled) and a parametrized measure $\mu = \{\mu_x\}_{x\in\Omega}$ such that

$$\int_\Omega \int_{\mathbf{R}} |s|^p \, d\mu_x(s)\, dx < \infty, \tag{4.2}$$

and whenever $\{f(M(\nabla w_n))\}$ converges weakly in $L^1(\Omega)$, its weak limit is given by

$$\overline{f}(x) := \langle \mu_x, f \rangle, \quad \mathcal{L}^N\text{a.e. } x \in \Omega.$$

Let $k \in \mathbf{N}$ and consider the truncation map $T_k : \mathbf{R} \to \mathbf{R}$ given by

$$T_k(x) := \begin{cases} x, & |x| \le k, \\ k\dfrac{x}{|x|}, & |x| > k. \end{cases}$$

Clearly $\{T_k(M(\nabla w_n))\}$ is a bounded sequence in $L^\infty(\Omega)$, therefore equi-integrable, and so given $a \in L^\infty(\Omega)$ we have

$$\begin{aligned} \lim_{k\to\infty} \lim_{n\to\infty} \int_\Omega a(x) |T_k(M(\nabla w_n))(x)|^p \, dx &= \lim_{k\to\infty} \int_\Omega a(x) \int_{\mathbf{R}} |T_k(s)|^p \, d\mu_x(s)\, dx \\ &= \int_\Omega \int_{\mathbf{R}} a(x) |s|^p \, d\mu_x(s)\, dx, \end{aligned} \tag{4.3}$$

where we have used (4.2) and the dominated convergence theorem. For every $k \in \mathbf{N}$, choose $n(k)$ with $n(k) > n(k-1)$ such that

$$\left| \lim_{n\to\infty} \int_\Omega |T_k(M(\nabla w_n))(x)|^p \, dx - \int_\Omega |T_k(M(\nabla w_m))(x)|^p \, dx \right| < \frac{1}{k}$$

whenever $m \geq n(k)$. Setting $a \equiv 1$, (4.3) reduces to

$$\lim_{k\to\infty} \int_\Omega \left|T_k(M(\nabla w_{n(k)}))(x)\right|^p \, dx = \int_\Omega \int_{\mathbf{R}} |s|^p \, d\mu_x(s)\, dx. \tag{4.4}$$

We claim that

$$\left|T_k(M(\nabla w_{n(k)}))\right|^p \rightharpoonup \overline{f} \text{ in } L^1(\Omega), \tag{4.5}$$

where

$$\overline{f}(x) := \int_{\mathbf{R}} |s|^p \, d\mu_x(s).$$

Indeed, fix $b \in L^\infty(\Omega)$, $l \in \mathbf{N}$, and let $k > l$. Clearly

$$\begin{aligned}\int_\Omega b(x) \left|T_k(M(\nabla w_{n(k)}))(x)\right|^p \, dx \leq \|b\|_{L^\infty(\Omega)} \int_\Omega \left|T_k(M(\nabla w_{n(k)}))(x)\right|^p \, dx \\ - \int_\Omega \Big(\|b\|_{L^\infty(\Omega)} - b(x)\Big) \left|T_l(M(\nabla w_{n(k)}))(x)\right|^p \, dx,\end{aligned}$$

and so, taking first the limit as $k \to \infty$, followed by the limit as $l \to \infty$, and by virtue of (4.3) and (4.4), we conclude that

$$\limsup_{k\to\infty} \int_\Omega b(x) \left|T_k(M(\nabla w_{n(k)}))(x)\right|^p \, dx \leq \int_\Omega \int_{\mathbf{R}} b(x)\, |s|^p \, d\mu_x(s)\, dx. \tag{4.6}$$

Similarly, (4.6) holds for $-b$ in place of b; hence

$$\lim_{k\to\infty} \int_\Omega b(x) \left|T_k(M(\nabla w_{n(k)}))(x)\right|^p \, dx = \int_\Omega b(x)\overline{f}(x)\, dx,$$

proving (4.5). Set

$$R_k := \left\{x \in \mathbf{R}^N : M(\nabla w_{n(k)})(x) \geq k\right\}.$$

By Lemma 4.1 there exist Lipschitz functions z_k such that

$$z_k = w_{n(k)} \quad \text{a.e. on } \mathbf{R}^N \setminus R_k, \qquad |\nabla z_k(x)| \leq C(N)k, \quad \text{a.e. } x \in \mathbf{R}^N.$$

Therefore, by (4.1) and because Ω is bounded,

$$\begin{aligned}\mathcal{L}^N\left(\Omega \cap \left\{z_k \neq w_{n(k)} \text{ or } \nabla z_k \neq \nabla w_{n(k)}\right\}\right) &\leq \mathcal{L}^N\left(R_k \cap \Omega\right) \\ &\leq \frac{1}{k^p} \int_\Omega \left|M(\nabla w_{n(k)})\right|^p \, dx,\end{aligned}$$

and this term tends to zero as $k \to \infty$. In addition, for $\mathcal{L}^N$ a.e. $x \in \Omega \setminus R_k$ we have (see [16, Theorem 3 in section 6.1] and [20, Lemma 7.7])

$$|\nabla z_k(x)| = \left|\nabla w_{n(k)}(x)\right| \leq \left|M(\nabla w_{n(k)})(x)\right| = \left|T_k(M(\nabla w_{n(k)})(x))\right|,$$

while if $x \in R_k$ then

$$|\nabla z_k(x)| \le C(N)k = C(N)\left|T_k(M(\nabla w_{n(k)})(x))\right|.$$

We conclude that

$$|\nabla z_k(x)|^p \le C\left|T_k(M(\nabla w_{n(k)})(x))\right|^p \quad \text{a.e. } x \in \Omega,$$

which, together with (4.5), yields equi-integrability of $\{|\nabla z_k|^p\}$.

Step 2. Let Ω be an open, bounded domain of $\mathbf{R}^N$, and let $\{w_j\}$ be a bounded sequence in $W^{1,p}(\Omega;\mathbf{R}^m)$. Without loss of generality we may assume that there exists $w_0 \in W^{1,p}(\Omega;\mathbf{R}^m)$ such that

$$w_j \rightharpoonup w_0 \quad \text{in } W^{1,p}(\Omega;\mathbf{R}^m), \quad w_j \to w_0 \quad \text{in } L^p_{loc}(\Omega;\mathbf{R}^m);$$

i.e., if $w_j := w_0 + \tilde{w}_j$,

$$\tilde{w}_j \rightharpoonup 0 \quad \text{in } W^{1,p}(\Omega;\mathbf{R}^m), \quad \tilde{w}_j \to 0 \quad \text{in } L^p_{loc}(\Omega;\mathbf{R}^m).$$

Let $\{\Omega_n\}$ be an increasing sequence of compactly contained subdomains of Ω, with $\mathcal{L}^N(\Omega \setminus \Omega_n) \to 0$, and choose cut-off functions $\eta_n \in \mathcal{C}_0^\infty(\Omega;[0,1])$ such that $\eta_n = 1$ if $x \in \Omega_n$. We have

$$\limsup_{n\to\infty}\limsup_{j\to\infty}\|\eta_n\tilde{w}_j\|_{L^p(\Omega;\mathbf{R}^m)} = 0$$

and

$$\begin{aligned}\limsup_{n\to\infty}\limsup_{j\to\infty}\|\nabla(\eta_n\tilde{w}_j)\|_{L^p(\Omega;\mathbf{M})} &= \limsup_{n\to\infty}\limsup_{j\to\infty}\|\tilde{w}_j\otimes\nabla\eta_n + \eta_n\nabla\tilde{w}_j\|_{L^p(\Omega;\mathbf{M})}\\ &\le \limsup_{j\to\infty}\|\nabla\tilde{w}_j\|_{L^p(\Omega;\mathbf{M})} < \infty.\end{aligned}$$

A standard diagonalization procedure yields a bounded subsequence in $W^{1,p}_0(\Omega;\mathbf{R}^m)$, $\{\eta_n\tilde{w}_{j(n)}\}$, which we extend by zero to $\mathbf{R}^N$. Now the argument used in Step 1 applies to this sequence, so we obtain a sequence $\{z_k\}$ of Lipschitz functions such that

$$\alpha_k := \mathcal{L}^N\left(\Omega \cap \left\{z_k \ne \eta_{n(k)}\tilde{w}_{j(n(k))} \text{ or } \nabla z_k \ne \nabla\left(\eta_{n(k)}\tilde{w}_{j(n(k))}\right)\right\}\right) \to 0,$$

as $k \to \infty$, and $\{|\nabla z_k|^p\}$ is equi-integrable. We conclude that $\{|\nabla(w_0+z_k)|^p\}$ is equi-integrable and

$$\mathcal{L}^N\left(\Omega \cap \left\{w_{j(n(k))} \ne w_0 + z_k \text{ or } \nabla w_{j(n(k))} \ne \nabla(w_0+z_k)\right\}\right) \le \alpha_k + \mathcal{L}^N\left(\Omega \setminus \Omega_{n(k)}\right)$$

and this term converges to zero as $k \to \infty$.

5. Characterization of YM-V: Necessary conditions. We devote this section to the proof of the necessity part of Theorem 1.1. We may assume that Ω is smooth as otherwise we can first consider smooth subsets of Ω and then exhaust Ω by such sets. Conditions 1 and 2 were established in [21] and [22]. To prove 3 and 4 we split Λ into a part, $P\nu$, that is determined by the Young measure and a remainder, $\tilde{\Lambda}$, that is related to pure concentration effects.

For $\psi \in \mathcal{C}(\mathcal{S})$ (with p-homogeneous extension $\tilde{\psi}$) and $\theta \in \mathcal{C}_0(\Omega)$, let

$$\langle P\nu_x, \psi\rangle := \langle \nu_x, \tilde{\psi}\rangle = \int_{\mathbf{M}} \tilde{\psi}(A)\, d\nu_x(A)$$

and $P\nu := \{P\nu_x\}_{x\in\Omega} \otimes \mathcal{L}^N$, i.e.,

$$\langle P\nu, \theta \otimes \psi\rangle := \int_\Omega \theta(x) \int_{\mathbf{M}} \tilde{\psi}(A)\, d\nu_x(A)\, dx.$$

Let $\tilde{\Lambda} := \Lambda - P\nu$. Suppose that $\{\nabla u_j\}$ generates the YM-V pair (ν, Λ). In Steps 2 and 3 below we will show that u_j can be decomposed as $u_j = z_j + v_j$ where $\{|\nabla z_j|^p\}$ is equi-integrable, $\{\nabla z_j\}$ generates the YM-V pair $(\nu, P\nu)$, and $\{\nabla v_j\}$ generates the YM-V pair $(\delta_0 \otimes \mathcal{L}^N, \tilde{\Lambda})$.

Step 1. *Reformulation of conditions* 3 *and* 4.

We claim that 3 and 4 are equivalent to requiring that

(i) $\tilde{\Lambda}$ is a nonnegative, finite Radon measure on $\Omega \times \mathcal{S}$;

(ii) if $\tilde{\Lambda} = \{\tilde{\lambda}_x\}_{x\in\Omega} \otimes \tilde{\pi}$ is the slicing decomposition of $\tilde{\Lambda}$, where $\tilde{\lambda}_x$ are probability measures on $\mathbf{M}$, then for $\tilde{\pi}$ a.e. $x \in \Omega$

$$\langle \tilde{\lambda}_x, \psi\rangle \geq 0 \tag{5.1}$$

for all $\psi \in \mathcal{H}_p$ such that $Q\psi(0) = 0$.

Assume first that 3 and 4 hold. Since Λ and $P\nu$ are finite, nonnegative Radon measures, it follows that $\tilde{\Lambda}$ is a finite Radon measure. In addition, if $\theta \in \mathcal{C}_0(\Omega)$, $\theta \geq 0$, and if $\psi \in \mathcal{C}(\mathcal{S})$, $\psi \geq 0$, then $Q\tilde{\psi}(0) = 0$ and we have, by 3 and 4,

$$\begin{aligned}\langle \tilde{\Lambda}, \theta \otimes \psi\rangle &= \int_\Omega \theta(x) \int_{\mathcal{S}} \psi(A)\, d\lambda_x(A)\, d\pi_s(x) \\ &\quad + \int_\Omega \theta(x) \left[\frac{d\pi}{d\mathcal{L}^N}(x) \int_{\mathcal{S}} \psi(A)\, d\lambda_x(A) - \int_{\mathbf{M}} \tilde{\psi}(A)\, d\nu_x(A)\right] dx \\ &\geq 0.\end{aligned} \tag{5.2}$$

Hence $\tilde{\Lambda} \geq 0$, proving (i).

In order to prove (ii), fix $\theta \in \mathcal{C}_0(\Omega)$, $\theta \geq 0$, $\psi \in \mathcal{H}_p$, $Q\psi(0) = 0$; using the slicing decomposition of $\tilde{\Lambda}$ and (5.2) we deduce that

$$\int_\Omega \theta(x) \langle \tilde{\lambda}_x, \psi\rangle\, d\tilde{\pi}(x) = \langle \tilde{\Lambda}, \theta \otimes \psi\rangle \geq 0. \tag{5.3}$$

The arbitrariness of θ yields the existence of a $\tilde{\pi}$-null set E_ψ such that if $x \in \Omega \setminus E_\psi$ then

$$\langle \tilde{\lambda}_x, \psi\rangle \geq 0.$$

Let $\{\psi_k\}$ be a countable, dense set in $\mathcal{H}_p$, and define

$$E := \bigcup_k \bigcup_{\{n\,:\, Q(\psi_k + (1/n)|A|^p)(0) = 0\}} E_{\psi_k + (1/n)|A|^p}. \tag{5.4}$$

It is clear that $\tilde{\pi}(E) = 0$. Fix $x \in \Omega \setminus E$, $\psi \in \mathcal{H}_p$, $Q\psi(0) = 0$, and choose a subsequence $\{\psi_{k_i}\}$ such that

$$\psi_{k_i} \to \psi \quad \text{in } L^\infty(\mathcal{S}), \qquad \|\psi_{k_i} - \psi\|_{L^\infty(\mathcal{S})} < \frac{1}{n_i},$$

where $n_i \to \infty$. Then

$$\begin{aligned}\psi_{k_i}(A) + \frac{1}{n_i}|A|^p &\geq \psi_{k_i}(A) + |A|^p \|\psi_{k_i} - \psi\|_{L^\infty(\mathcal{S})} \\ &\geq \psi_{k_i}(A) + |\psi_{k_i}(A) - \psi(A)| \\ &\geq \psi(A),\end{aligned}$$

and so

$$Q\left(\psi_{k_i} + \frac{1}{n_i}|\cdot|^p\right)(0) \geq Q\psi(0) = 0.$$

By homogeneity

$$Q\left(\psi_{k_i} + \frac{1}{n_i}|\cdot|^p\right)(0) = 0.$$

Finally, using the definition of E, $x \notin E_{\psi_{k_i}+(1/n_i)|\cdot|^p}$, therefore

$$0 \leq \lim_{i\to\infty} \left\langle \tilde{\lambda}_x, \psi_{k_i} + \frac{1}{n_i}|\cdot|^p \right\rangle = \langle \tilde{\lambda}_x, \psi\rangle, \tag{5.5}$$

concluding the proof of (ii).

Conversely, if (i) and (ii) hold, using (5.2), (5.3), and $\theta := \chi_{B(a,\rho)}$, $a \in \Omega$, $\rho > 0$, we have

$$\int_{B(a,\rho)} \langle \lambda_x, \psi\rangle \, d\pi_s + \int_{B(a,\rho)} \left(\frac{d\pi}{d\mathcal{L}^N}(x)\langle \lambda_x, \psi\rangle - \langle \nu_x, \psi\rangle \right) dx \geq 0$$

for $\psi \in \mathcal{H}_p$, $Q\psi(0) = 0$. Conditions 3 and 4 follow by virtue of the Radon–Nikodym theorem. Note that a priori the exceptional sets could depend on ψ, but the argument outlined for the definition of E above would entail the existence of π_s- and $\mathcal{L}^N$-negligible sets for which 3 and 4 hold for all $\psi \in \mathcal{H}_p$ such that $Q\psi(0) = 0$.

In light of Step 1, the rest of this section will be dedicated to proving (5.1).

Step 2. *Construction of* $\{z_j\}$.

By the decomposition lemma (Lemma 1.2) there exists a sequence of Lipschitz functions $\{z_j\}$ such that $\{|\nabla z_j|^p\}$ is equi-integrable in Ω and the set

$$R_j := \{x \in \Omega : z_j(x) \neq u_j(x), \nabla z_j(x) \neq \nabla u_j(x)\}$$

satisfies

$$\mathcal{L}^N(R_j) \to 0. \tag{5.6}$$

In particular, $\{\nabla z_j\}$ generates the YM-V pair $(\nu, P\nu)$.

Step 3. *Construction of* $\{v_j\}$.

Let $v_j := u_j - z_j$. We claim that $\{\nabla v_j\}$ generates the YM-V pair $(\delta_0 \otimes \mathcal{L}^N, \tilde{\Lambda})$. In particular, $\tilde{\Lambda} \geq 0$. The assertion regarding the Young measure follows from (5.6). To study the varifold generated by $\{\nabla v_j\}$ consider $\theta \in \mathcal{C}_0(\Omega)$ and $\psi \in \mathcal{H}_p$ such that $\psi|_{\mathcal{S}}$ is Lipschitz. In view of Proposition 2.2 and Hölder's inequality we have

$$\left| \int_\Omega \theta(x)\psi(\nabla v_j)\, dx - \int_\Omega \theta(x) \left(\psi(\nabla u_j) - \psi(\nabla z_j)\right) dx \right|$$

$$
\begin{aligned}
&= \left| \int_{R_j} \theta(x) \left(\psi(\nabla u_j - \nabla z_j) - \psi(\nabla u_j) + \psi(\nabla z_j) \right) dx \right| \\
&\le C \|\theta\|_\infty \int_{R_j} \left[\left(|\nabla u_j - \nabla z_j|^{p-1} + |\nabla u_j|^{p-1} \right) |\nabla z_j| + |\nabla z_j|^p \right] dx \\
&\le C \|\theta\|_\infty \left[\left(\int_{R_j} |\nabla z_j|^p \, dx \right)^{1/p} + \int_{R_j} |\nabla z_j|^p \, dx \right].
\end{aligned}
$$

Since $\mathcal{L}^N(R_j) \to 0$ as $j \to \infty$ and $\{|\nabla z_j|^p\}$ is equi-integrable, the last term goes to zero as $j \to \infty$ and thus, using Step 2, we conclude that

$$
\int_\Omega \theta(x) \psi(\nabla v_j) \, dx \to \langle \Lambda - P\nu, \theta \otimes \psi \rangle = \langle \tilde{\Lambda}, \theta \otimes \psi \rangle.
$$

By density, the result extends to all $\psi \in \mathcal{H}_p$ and the claim is proved.

Step 4. We prove that for $\tilde{\pi}$ a.e. $x \in \Omega$

$$
\langle \tilde{\lambda}_x, \psi \rangle \ge 0 \tag{5.7}
$$

for all $\psi \in \mathcal{H}_p$ with $Q\psi(0) = 0$.

We first make the additional assumption that ψ is Lipschitz on $\mathcal{S}$. Let $\theta \in \mathcal{C}_0^\infty(B(a,\rho))$, $0 \le \theta \le 1$. By the definition of $Q\psi$, Proposition 2.2, and Hölder's inequality, we have

$$
\begin{aligned}
0 &\le \int_{B(a,\rho)} \psi \left(\nabla(\theta v_j) \right) \\
&= \int_{B(a,\rho)} \psi \left(\theta \nabla v_j + v_j \otimes \nabla \theta \right) dx \\
&\le \int_{B(a,\rho)} \theta^p \psi(\nabla v_j) \, dx + C \int_{B(a,\rho)} \left(|\theta \nabla v_j|^{p-1} + |v_j \otimes \nabla \theta|^{p-1} \right) |v_j \otimes \nabla \theta| \, dx \\
&\le \int_{B(a,\rho)} \theta^p \psi(\nabla v_j) \, dx + C(\theta) \left[\left(\int_{B(a,\rho)} |v_j|^p \, dx \right)^{1/p} + \int_{B(a,\rho)} |v_j|^p \, dx \right].
\end{aligned}
$$

Now $v_j \rightharpoonup 0$ in $W^{1,p}(B(a,\rho))$ as $j \to \infty$, and thus $v_j \to 0$ in $L^p(B(a,\rho))$. By Step 3, the sequence $\{\nabla v_j\}$ generates the varifold $\tilde{\Lambda}$. Therefore taking the limit as $j \to \infty$ in the above inequality, we obtain

$$
0 \le \langle \tilde{\Lambda}, \theta^p \otimes \psi \rangle.
$$

The assertion follows (for $\psi \in \mathrm{Lip}(\mathcal{S})$) by taking an increasing sequence $\theta_i \to \chi_{B(a,\rho)}$ and applying the dominated convergence theorem. Hence

$$
\int_{B(a,\rho)} \langle \tilde{\lambda}_x, \psi \rangle \, d\tilde{\pi}(x) \ge 0,
$$

and the Radon–Nikodym theorem yields the existence of a set $E_\psi \subset \Omega$, $\tilde{\pi}(E_\psi) = 0$ such that (5.7) holds if $x \notin E_\psi$. Defining E as in (5.4) and following the argument (5.4)–(5.5), we finally remove the restriction that ψ be Lipschitz on $\mathcal{S}$ to conclude that (5.7) holds for $\psi \in \mathcal{H}_p$, $Q\psi(0) = 0$, proving (ii).

6. Characterization of YM-V: Sufficient conditions. Suppose that the pair (ν, Λ) satisfies the conditions of Theorem 1.1. We have to construct a sequence $\{u_j\}$, bounded in $W^{1,p}(\Omega; \mathbf{R}^m)$, such that (ν, Λ) is the YM-V pair generated by $\{\nabla u_j\}$.

As in the beginning of section 5, we write $\Lambda := P\nu + \tilde{\Lambda}$, where

$$\langle P\nu, \theta \otimes \psi \rangle := \int_\Omega \theta(x) \int_{\mathbf{M}} \tilde{\psi}(A)\, d\nu_x(A)\, dx$$

for $\theta \in \mathcal{C}_0(\Omega)$, $\psi \in \mathcal{C}(\mathcal{S})$, and with $\tilde{\psi}$ the p-homogeneous extension of ψ. From section 5, Step 1, we know that $\tilde{\Lambda}$ is a nonnegative, finite Radon measure and

$$\langle \tilde{\lambda}_x, \psi \rangle \geq 0 \tag{6.1}$$

for all $\psi \in \mathcal{H}_p$ with $Q\psi(0) = 0$, where $\{\tilde{\lambda}_x\}_{x\in\Omega} \otimes \tilde{\pi}$ denotes the slicing decomposition of $\tilde{\Lambda}$.

Step 1. We claim that it suffices to find $\{z_j\}$, $\{v_j\}$ bounded in $W^{1,p}(\Omega; \mathbf{R}^m)$ such that

$$\{|\nabla z_j|^p\} \text{ is equi-integrable}, \quad \{\nabla z_j\} \text{ generates the YM-V pair } (\nu, P\nu), \tag{6.2}$$

and

$$\{\nabla v_j\} \text{ generates the YM-V pair } (\delta_0 \otimes \mathcal{L}^N, \tilde{\Lambda}), \tag{6.3}$$

setting, as before, $u_j := z_j + v_j$. Indeed, note that since $p > 1$ then $\{|\nabla v_j|\}$ is equi-integrable, and so given $\lambda > 0$ and in view of (2.1)

$$\mathcal{L}^N\left(\{|\nabla v_j| > \lambda\}\right) \leq \frac{1}{\lambda} \int_\Omega |\nabla v_j|\, dx \to \frac{1}{\lambda} \langle \delta_0 \otimes \mathcal{L}^N, |\cdot| \rangle = 0. \tag{6.4}$$

Thus given $\theta \in \mathcal{C}_0(\Omega)$ and $\varphi \in \mathcal{C}_0^\infty(\mathbf{M})$ we have

$$\left| \int_\Omega \theta(x) \left[\varphi(\nabla u_j) - \varphi(\nabla z_j)\right] dx \right| \leq \|\theta\|_\infty C \int_\Omega |\nabla v_j|\, dx \to 0$$

as $j \to \infty$, and this implies that the Young measure associated with $\{\nabla u_j\}$ is also ν.

Similarly, if $\theta \in \mathcal{C}_0(\Omega)$, $\psi \in \mathcal{H}_p$, $\psi|_{\mathcal{S}}$ Lipschitz, by Proposition 2.2 for fixed $\lambda > 0$

$$\begin{aligned}
&\left| \int_\Omega \theta(x) \left(\psi(\nabla u_j) - \psi(\nabla z_j) - \psi(\nabla v_j)\right) dx \right| \\
&\leq C \int_{\{|\nabla v_j| \leq \lambda\}} \left(|\psi(\nabla u_j) - \psi(\nabla z_j)| + |\psi(\nabla v_j)|\right) dx \\
&\quad + C \int_{\{|\nabla v_j| \geq \lambda\}} \left(|\psi(\nabla u_j) - \psi(\nabla v_j)| + |\psi(\nabla z_j)|\right) dx \\
&\leq C \int_{\{|\nabla v_j| \leq \lambda\}} \left(|\nabla z_j|^{p-1} + |\nabla v_j|^{p-1}\right) |\nabla v_j|\, dx + C \int_{\{|\nabla v_j| \leq \lambda\}} |\nabla v_j|^p\, dx \\
&\quad + C \int_{\{|\nabla v_j| \geq \lambda\}} \left(|\nabla z_j|^{p-1} + |\nabla v_j|^{p-1}\right) |\nabla z_j|\, dx + C \int_{\{|\nabla v_j| \geq \lambda\}} |\nabla z_j|^p\, dx,
\end{aligned}$$

and so, using Hölder's inequality, (6.4), and the equi-integrability of $\{|\nabla z_j|^p\}$,

$$\limsup_{j\to\infty} \left| \int_\Omega \theta(x) \left[\psi(\nabla u_j) - \psi(\nabla z_j) - \psi(\nabla v_j)\right] dx \right| = O(\lambda).$$

Letting $\lambda \to 0^+$ and removing the regularity restrictions imposed on ψ as in (5.4)–(5.5), we conclude that the varifold associated with $\{\nabla u_j\}$ is $P\nu + \tilde{\Lambda} := \Lambda$.

Step 2. We introduce two sets of measures supported on the unit sphere $\mathcal{S}$ of M, namely

$$A := \{\mu \in \mathcal{M}(\mathcal{S}) : \mu \geq 0, \langle \mu, \psi \rangle \geq 0 \text{ if } \psi \in \mathcal{H}_p, Q\psi(0) = 0\},$$
$$H := \left\{ \overline{\delta_{\nabla u/|\nabla u|} \otimes |\nabla u|^p \, \mathcal{L}^N} : u \in W_0^{1,p}(B; \mathbf{R}^m) \right\},$$

where B is the unit ball in R^N, and the average measures of H are defined by

$$\langle \overline{\delta_{\nabla u/|\nabla u|} \otimes |\nabla u|^p \, \mathcal{L}^N}, \psi \rangle := \frac{1}{|B|} \int_B \psi \left(\frac{\nabla u}{|\nabla u|} \right) |\nabla u|^p \, dx$$

for $\psi \in \mathcal{C}(\mathcal{S})$. We do not distinguish henceforth a continuous function on $\mathcal{S}$ from its p-homogeneous extension. Note that, in view of (6.1), $\tilde{\lambda}_x \in A$ for $\tilde{\pi}$ a.e. $x \in \Omega$. It is clear that A is weak *-closed and $H \subset A$.

PROPOSITION 6.1. *A is the weak *-closure of H. Moreover, if $R > 0$ then $A \cap \{\|\mu\| \leq R\}$ is the weak *-closure of $H \cap \{\|\mu\| \leq R\}$.*

Remark. The second statement will be useful in Step 3 where we will use the fact that the weak *-topology of $\mathcal{M}(S)$ is metrizable on closed balls.

Proof of Proposition 6.1. The proof is a standard application of the Hahn–Banach theorem. We start by proving that H is convex. Fix $\theta \in (0,1)$ and let for $i = 1, 2$

$$\mu_i := \overline{\delta_{\nabla u_i/|\nabla u_i|} \otimes |\nabla u_i|^p \, \mathcal{L}^N}, \quad u_i \in W_0^{1,p}(B; \mathbf{R}^m).$$

Let $x_0 \in B$ be such that $|x_0| = 1/2$ and define

$$\tilde{u}_1(x) := k^{-1+N/p} u_1(kx), \quad \tilde{u}_2(x) := k^{-1+N/p} u_2(k(x - x_0)),$$

where $k \geq 4$. Clearly $\tilde{u}_i \in W_0^{1,p}(B; \mathbf{R}^m)$, $\tilde{u}_1$ and $\tilde{u}_2$ have disjoint supports, and a change of variables shows that $\tilde{\mu}_i = \mu_i$ for $i = 1, 2$. It follows that the function

$$\tilde{u} := \theta^{1/p} \tilde{u}_1 + (1-\theta)^{1/p} \tilde{u}_2 \in W_0^{1,p}(B; \mathbf{R}^m)$$

generates $\mu := \theta\mu_1 + (1-\theta)\mu_2$ and so $\mu \in H$.

We now show that A cannot be separated from H. Assume that $\psi \in \mathcal{C}(\mathcal{S})$ is such that $\langle \nu, \psi \rangle \geq a$ for all $\nu \in H$ and for some $a \in \mathbf{R}$. Hence, extending ψ as p-homogeneous,

$$Q\psi(0) = \inf_{u \in W_0^{1,p}(B; \mathbf{R}^m)} \frac{1}{|B|} \int_B \psi(\nabla u) \, dx \geq a,$$

and so $0 \geq Q\psi(0) \geq a$. We conclude that $Q\psi(0)$ is finite; thus $Q\psi(0) = 0$ by homogeneity, and $0 = Q\psi(0) \geq a$. By definition of A, we have that $\langle \mu, \psi \rangle \geq 0 \geq a$ for all $\mu \in A$. Hence A cannot be separated from H.

Next, we show that $A_R := A \cap \{\|\mu\| \leq R\}$ cannot be separated from $H_R := H \cap \{\|\mu\| \leq R\}$. Given $\rho > 0$ define $H_\rho := H \cap \{||\mu|| = \rho\} = H \cap \{\langle \mu, 1 \rangle = \rho\}$. We claim that

$$\overline{H_\rho} = \overline{H} \cap \{\langle \mu, 1 \rangle = \rho\}. \tag{6.5}$$

It is clear that $\overline{H_\rho} \subset \overline{H} \cap \{\langle \mu, 1\rangle = \rho\}$. Suppose that this inclusion is strict. Then there exists $\mu \in \overline{H}$ such that $\langle \mu, 1\rangle = \rho$ and $\mu \notin \overline{H_\rho}$. Since H_ρ is convex, using the Hahn–Banach theorem we may find $\psi \in C(\mathcal{S}), a \in \mathbf{R}$, such that

$$\langle \mu, \psi\rangle < a, \quad \langle \nu, \psi\rangle \geq a \quad \text{for all } \nu \in H_\rho.$$

Set $\bar{\psi} := \psi - a/\rho$. Clearly,

$$\langle \mu, \bar{\psi}\rangle = \langle \mu, \psi\rangle - \frac{a}{\rho}\langle \mu, 1\rangle < 0,$$

while $\langle \nu, \bar{\psi}\rangle \geq 0$ for all $\nu \in H_\rho$. Since H is a cone, we conclude that $\langle \nu, \bar{\psi}\rangle \geq 0$ for all $\nu \in H$; therefore $\langle \mu, \overline{\psi}\rangle \geq 0$. We have reached a contradiction, thus (6.5) is proved. Finally, since $A = \overline{H}$, it follows that $A_R \supset \overline{H_R}$, and by (6.5) we conclude that

$$\overline{H_R} = \overline{\bigcup_{0<\rho\leq R} H_\rho} \supset \bigcup_{0<\rho\leq R} \overline{H_\rho} = \bigcup_{0<\rho\leq R} \overline{H} \cap \{\langle \mu, 1\rangle = \rho\} = A \cap \{||\mu|| \leq R\}.$$

Step 3. *Construction of* $\{z_j\}$.

Using condition 3 in Theorem 1.1 with $\psi(A) := |A|^p$, we have

$$\int_\Omega \int_{\mathbf{M}} |A|^p \, d\nu_x(A)\, dx \leq \int_\Omega \frac{d\pi}{d\mathcal{L}^N}(x) \int_{\mathcal{S}} d\lambda_x(A)\, dx \leq \pi(\Omega) < \infty,$$

which, together with conditions 1 and 2 and by Theorem 2.3 (see [22] for the proof), implies that ν is a $W^{1,p}$-Young measure. Using the decomposition lemma (Lemma 1.2) (see also Step 2, Section 5) we find a sequence $\{z_j\}$ bounded in $W^{1,p}(\Omega; \mathbf{R}^m)$ and satisfying (6.2).

Step 4. *Construction of* $\{v_j\}$ *when*

$$\tilde{\Lambda} := \sum_{i=1}^{I} c_i \lambda_i \otimes \delta_{x_i}, \quad x_i \in \Omega, \lambda_i \in A, c_i > 0.$$

Here we search for a sequence $\{v_j\}$ bounded in $W_0^{1,p}(\Omega; \mathbf{R}^m)$ such that (6.3) holds, i.e., $\{\nabla v_j\}$ generates $(\delta_0 \otimes \mathcal{L}^N, \tilde{\Lambda})$ and, in addition,

$$\lim_{j\to\infty} \|\nabla v_j\|^p_{L^p(\Omega;\mathbf{M})} = \left\|\tilde{\Lambda}\right\|.$$

By Proposition 6.1 and the remark after it, there exist bounded sequences $\{w_j^{(i)}\}$ in $W_0^{1,p}(B; \mathbf{R}^m)$ such that

$$\lim_{j\to\infty} \frac{1}{|B|} \int_B \psi(\nabla w_j^{(i)})\, dx = \langle \lambda_i, \psi\rangle$$

for all $\psi \in \mathcal{H}_p$. In particular

$$\|\lambda_i\| = \lim_{j\to\infty} \frac{1}{|B|} \int_B \left|\nabla w_j^{(i)}\right|^p \, dx.$$

Now

$$v_j(x) := j^{-1+N/p} \frac{1}{|B|^{1/p}} \sum_{i=1}^{I} c_i^{1/p} w_j^{(i)}\left(j(x - x_i)\right)$$

has the desired properties.

Step 5. *Construction of* $\{v_j\}$ *in the general case.*

To obtain $\{v_j\}$ satisfying (6.3), we will use the following approximation lemma.

LEMMA 6.2. *Let* $\tilde{\Lambda}$ *be a nonnegative, finite, Radon measure on* $\Omega \times \mathcal{S}$ *with slicing decomposition* $\{\tilde{\lambda}_x\}_{x\in\Omega} \otimes \tilde{\pi}$*, let* A *be a convex set of the set of all nonnegative, finite Radon measures on* $\mathcal{S}$*, and suppose that*

$$\tilde{\lambda}_x \in A \text{ for } \tilde{\pi} \text{ a.e. } x \in \Omega.$$

Then $\tilde{\Lambda}$ *can be approximated in the weak *-topology by measures of the form*

$$\tilde{\Lambda}^{(k)} := \sum_{i=1}^{I_k} c_i^{(k)} \lambda_i^{(k)} \otimes \delta_{x_i^{(k)}}, \quad x_i^{(k)} \in \Omega, \lambda_i^{(k)} \in A, c_i^{(k)} > 0,$$

such that

$$\left\|\tilde{\Lambda}^{(k)}\right\| \le \left\|\tilde{\Lambda}\right\|.$$

Before proving the approximation lemma, we conclude the construction of $\{v_j\}$. By Lemma 6.2 we have

$$\tilde{\Lambda} = \text{ w-*-limit } \tilde{\Lambda}^{(k)}, \tilde{\Lambda}^{(k)} := \sum_{i=1}^{I_k} c_i^{(k)} \lambda_i^{(k)} \otimes \delta_{x_i^{(k)}}, \left\|\tilde{\Lambda}^{(k)}\right\| \le \left\|\tilde{\Lambda}\right\|.$$

Also, Step 4 yields the existence of sequences $\{v_j^{(k)}\}$ bounded in $W_0^{1,p}(\Omega;\mathbf{R}^m)$ generating the YM-V pair $(\delta_0 \otimes \mathcal{L}^N, \tilde{\Lambda}^{(k)})$ and such that

$$\lim_{j\to\infty} \left\|\nabla v_j^{(k)}\right\|_{L^p(\Omega;\mathbf{M})}^p = \left\|\tilde{\Lambda}^{(k)}\right\| \le \left\|\tilde{\Lambda}\right\|$$

for all k. Separability of $\mathcal{C}_0(\Omega)$, $\mathcal{C}_0(\mathbf{M})$ and $\mathcal{C}(\mathcal{S})$, and a standard diagonalization argument allow us to extract a diagonal subsequence $v_k := v_{j(k)}^{(k)}$ satisfying (6.3) and

$$\sup_k \|\nabla v_k\|_{L^p(\Omega;\mathbf{M})}^p \le \left\|\tilde{\Lambda}\right\| + 1.$$

It remains to prove Lemma 6.2.

Proof of Lemma 6.2. The result is well known to experts. We include a proof for the convenience of the reader. By Besicovitch's covering theorem, for each $k \in \mathbf{N}$ there exists a finite family of disjoint closed balls $B(x_i^{(k)}, r_i^{(k)})$ such that

$$\tilde{\pi}\left(\Omega \setminus \bigcup_{i\in I_k} B(x_i^{(k)}, r_i^{(k)})\right) < \frac{1}{k}, \quad r_i^{(k)} < \frac{1}{k}. \tag{6.6}$$

Set

$$\langle \lambda_i^{(k)}, \psi\rangle := \frac{1}{\tilde{\pi}\left(B(x_i^{(k)}, r_i^{(k)})\right)} \int_{B(x_i^{(k)}, r_i^{(k)})} \langle \tilde{\lambda}_x, \psi\rangle \, d\tilde{\pi}(x).$$

Since A is convex we have $\lambda_i^{(k)} \in A$, and we define

$$\tilde{\Lambda}^{(k)} := \sum_{i=1}^{I_k} c_i^{(k)} \lambda_i^{(k)} \otimes \delta_{x_i^{(k)}}, \quad c_i^{(k)} := \tilde{\pi}\left(B(x_i^{(k)}, r_i^{(k)})\right).$$

Then

$$\left\|\tilde{\Lambda}^{(k)}\right\| = \sum_{i=1}^{I_k} c_i^{(k)} \left\|\lambda_i^{(k)}\right\| \leq \sum_{i=1}^{I_k} \tilde{\pi}\left(B(x_i^{(k)}, r_i^{(k)})\right) \leq \tilde{\pi}(\Omega) = \left\|\tilde{\Lambda}\right\|.$$

For $\psi \in \mathcal{C}(\mathcal{S})$ and $\theta \in W_0^{1,\infty}(\Omega)$ with Lipschitz constant $\mathrm{Lip}(\theta)$ one has

$$\begin{aligned}
&\left|\langle \tilde{\Lambda}^{(k)} - \tilde{\Lambda}, \theta \otimes \psi\rangle\right| \\
&\quad = \left| \int_\Omega \langle \tilde{\lambda}_x, \psi\rangle \theta(x)\, d\tilde{\pi}(x) - \sum_{i=1}^{I_k} c_i^{(k)} \langle \lambda_i^{(k)}, \psi\rangle \theta(x_i^{(k)}) \right| \\
&\leq \left| \sum_{i=1}^{I_k} \int_{B(x_i^{(k)}, r_i^{(k)})} \langle \tilde{\lambda}_x, \psi\rangle \theta(x)\, d\tilde{\pi}(x) - \sum_{i=1}^{I_k} \int_{B(x_i^{(k)}, r_i^{(k)})} \langle \tilde{\lambda}_x, \psi\rangle\, d\tilde{\pi}(x) \theta(x_i^{(k)}) \right| \\
&\qquad + \int_{\Omega \setminus \cup B(x_i^{(k)}, r_i^{(k)})} \left|\langle \tilde{\lambda}_x, \psi\rangle\right| |\theta(x)|\, d\tilde{\pi}(x) \\
&\leq \frac{1}{k} \mathrm{Lip}(\theta) \int_\Omega \left|\langle \tilde{\lambda}_x, \psi\rangle\right| d\tilde{\pi}(x) + \|\psi\|_{L^\infty(\mathcal{S})} \|\theta\|_{L^\infty(\Omega)} \tilde{\pi}\left(\Omega \setminus \cup B(x_i^{(k)}, r_i^{(k)})\right),
\end{aligned}$$

and this expression tends to zero as $k \to \infty$. We have used (6.6). The assertion follows since test functions of the above type are dense and $\{\|\tilde{\Lambda}^{(k)}\|\}$ is bounded.

REFERENCES

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

[2] J. J. Alibert and G. Bouchitté, *Non uniform integrability and generalized Young measures*, J. Convex Anal., 4 (1997), pp. 129–147.

[3] W. Allard, *On the first variation of a varifold*, Ann. Math., 95 (1972), pp. 417–491.

[4] F. J. Almgren Jr., *Existence and regularity almost everywhere of solutions to elliptic variational problems among surfaces of varying topological type and singularity structure*, Ann. Math., 87 (1968), pp. 321–391.

[5] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[6] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDE's and Continuum Models of Phase Transitions, Lecture Notes in Physics 344, M. Rascle, D. Serre, and M. Slemrod, eds., Springer, Berlin, pp. 207–215.

[7] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[8] J. M. Ball and R. D. James, *Proposed experimental tests of a theory of fine microstructure and the two well problem*, Philos. Trans. Roy. Soc. London Ser. A, 338 (1992), pp. 389–450.

[9] J. M. Ball and F. Murat, *Remarks on Chacon's biting lemma*, Proc. Amer. Math. Soc., 107 (1989), pp. 655–663.

[10] B. Dacorogna, *Weak Continuity and Weak Lower Semicontinuity for Nonlinear Functionals*, Springer Lecture Notes 922, Springer, Berlin, 1982.

[11] B. Dacorogna, *Quasiconvexity and relaxation of non convex variational problems*, J. Funct. Anal., 46 (1982), pp. 102–118.

[12] R. J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27–70.

[13] R. J. DiPerna, *Compensated compactness and general systems of conservation laws*, Trans. Amer. Math. Soc., 292 (1985), pp. 383–420.

[14] R. J. DiPerna and A. J. Majda, *Oscillations and concentrations in weak solutions of the incompressible fluid equations*, Comm. Math. Phys., 108 (1987), pp. 667–689.

[15] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS 74, American Mathematical Society, 1990.

[16] L. C. Evans and R. F. Gariepy, *Lecture Notes on Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[17] I. Fonseca, *Lower semicontinuity of surface energies*, Proc. Roy. Soc. Edinburgh, 120A (1992), pp. 99–115.
[18] P. Gerard, *Compacité par compensation et regularité* 2*-microlocale*, Séminaire Eq. aux Dér. Part., Ecole Polytechnique, Palaiseau, exp VI, 1988-89.
[19] P. Gerard, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1989), pp. 1761–1794.
[20] D. Gilbarg and N. S. Trudinger, *Elliptic Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, New York, 1983.
[21] D. Kinderlehrer and P. Pedregal, *Characterizations of Young measures generated by gradients*, Arch. Rational Mech. Anal., 115 (1991), pp. 329–365.
[22] D. Kinderlehrer and P. Pedregal, *Gradient Young measures generated by sequences in Sobolev spaces*, J. Geom. Anal., 4 (1994), pp. 59–90.
[23] J. Kristensen, *Finite functionals and Young measures generated by gradients of Sobolev functions*, Mat-report 1994-34, Mathematical Institute, Technical University of Denmark, 1994.
[24] P. L. Lions, *The concentration-compactness principle in the calculus of variations: The locally compact case*, parts 1 and 2, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 109–145, pp. 223–283.
[25] P. L. Lions, *The concentration-compactness principle in the calculus of variations. The limit case, parts* 1 *and* 2, Rev. Mat. Iberoamericana, 1 (1) (1985), pp. 145–201, 1 (2) (1985), pp. 45–121.
[26] P. Marcellini, *Approximation of quasiconvex functions and lower semicontinuity of multiple integrals*, Manuscripta Math., 51 (1985), pp. 1–28.
[27] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.
[28] Y. Reshetnyak, *Weak convergence and completely additive vector functions on a set*, Sibirsk. Mat. Zh., 9 (1968), pp. 1039–1045.
[29] T. Roubicek, *Effective characterization of generalized Young measures generated by gradients*, Boll. Un. Mat. Itil., 9-B (1995), pp. 755–779.
[30] M. Kruzik and T. Roubicek, *On the measures of DiPerna and Majda*, Mathematica Bohemica, 122 (1997), pp. 383–399.
[31] M. E. Schonbeck, *Convergence of solutions to non-linear dispersive equations*, Comm. Partial Differential Equations, 7 (1982), pp. 959–1000.
[32] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, 1970.
[33] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Pitman Res. Notes Math. 39, R. Knops, ed., 1979, pp. 136–212.
[34] L. Tartar, *The compensated compactness method applied to systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, Nato Adv. Sci. Inst. Ser. C Math. Phys. Sci., Reidel, Dordrecht, the Netherlands, 1983.
[35] L. Tartar, *Etude des oscillations dans les équations aux dérivées partielles nonlinéaires*, Springer Lectures Notes in Physics 195, 1984, pp. 384–412.
[36] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh, 115A (1990), pp. 193–230.
[37] L. Tartar, *On mathematical tools for studying partial differential equations of continuum physics: H-measures and Young measures*, in Developments in Partial Differential Equations and Applications to Mathematical Physics, Plenum Press, New York, G. Buttazzo, Galdi, Zanghirati, eds., 1992, pp. 201–217.
[38] L. C. Young, *Lectures on Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.
[39] K. Zhang, *A construction of quasiconvex functions with linear growth at infinity*, Ann. Scuola Norm. Sup. Pisa, 19 (1992), pp. 313–326.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 757–778, May 1998

014

SOLUTIONS OF FINITELY SMOOTH NONLINEAR SINGULAR DIFFERENTIAL EQUATIONS AND PROBLEMS OF DIAGONALIZATION AND TRIANGULARIZATION*

HARRY GINGOLD† AND ALEXANDER TOVBIS‡

Abstract. It is known that existence of a formal power series solution $\hat{y}(x)$ to a system of nonlinear ordinary differential equations (ODEs) with analytic or infinitely smooth coefficients at an irregular singular point implies the existence of an actual solution $y(x)$, which possesses the asymptotic expansion $\hat{y}(x)$. In the present paper we extend this result for systems with finitely smooth coefficients. In this case one cannot speak about a formal power series solution $\hat{y}(x)$; it has therefore to be replaced by the requirement of existence of an "approximate" solution $y_0(x)$. The existence of a corresponding actual solution is a subject of certain conditions that link the smoothness of the system, the "accuracy" of the approximation $y_0(x)$, and the "degeneracy" of the system, linearized with respect to $y_0(x)$. As applications, problems of reduction of linear time dependent systems of ODEs into diagonal and triangular forms, as well as some other problems, are considered. In particular, the well-known theorem on integration of linear systems with irregular singularities is extended from analytical to finitely smooth systems. In one of the simplest cases, our result is simultaneously a consequence of the classical Levinson theorem.

Key words. irregular singularities, finitely smooth nonlinear equations, approximate solutions, diagonalization, triangularization

AMS subject classifications. 34C, 34E

PII. S0036141096307710

Introduction. In the present paper we study the existence of an actual solution to the nonlinear system of differential equations (or the vector equation)

$$\mathcal{N}y \equiv x^{1-r}y'(x) - f(x,y) = 0, \tag{0.1}$$

where $r \in \mathbb{N}$ and $x \geq x_0$ for some $x_0 \geq 0$. We assume that the entries of the n-dimensional vector-valued function $f(x,y)$ belong to the space

$$C[x_0,\infty) \times C^2(\mathcal{B}_\zeta)$$

and that $f(x,y)$ and its derivatives are bounded on $[x_0,\infty)$. Here $\mathcal{B}_\zeta$ denotes the open ball of the radius ζ in the Euclidean norm in $\mathbb{C}^n$, centered at the origin.

Similar to the analytic case, we call $x = +\infty$ an irregular singular point of (0.1) and the number r—the Poincaré rank of (0.1) at $x = +\infty$.

The classical result in the analytic case, i.e., when $f(x,y)$ is analytic at $(\infty,0) \in \bar{\mathbb{C}} \times \mathbb{C}^n$, states that the existence of a formal power series solution

$$\hat{y}(x) = \sum_{k=1}^{\infty} y_k x^{-k} \tag{0.2}$$

implies the existence of an analytic solution $y(x)$ such that

$$y(x) \sim \hat{y}(x), \quad x \to \infty,\ x \in S, \tag{0.3}$$

*Received by the editors August 2, 1996; accepted for publication (in revised form) March 25, 1997. The second author was supported in part by the NSF grant DMS 9500644.
http://www.siam.org/journals/sima/29-3/30771.html
†Department of Mathematics, West Virginia University, Morgantown, WV 26506-6310 (un051203@wvnvaxa.wvnet.edu).
‡Department of Mathematics, University of Central Florida, Orlando, FL 32816-1364 (tovbis@pegasus.cc.ucf.edu).

where S is some sector of the complex x-plane containing the positive real semiaxis (see, for example, [Hu], [Wa], [RS], [T2]). This result was extended to the C^∞-equations in [Ku]: suppose

$$f(x,y) = \sum_{|\alpha|=1}^{\infty} f_\alpha(x) y^\alpha, \tag{0.4}$$

where α is a multi-index, $|\alpha|$ denotes the length of α, all vector-valued functions $f_\alpha(x)$ are infinitely smooth on $[x_0, \infty)$ (with some positive x_0) and admit formal power series asymptotic expansions as $x \to +\infty$, and the series (0.4) is uniformly convergent for $x \in [x_0, \infty), \|y\| < \zeta$. Then the existence of the formal solution (0.2) implies the existence of a C^∞ solution $y(x)$ such that $y(x)$ satisfies (0.3), where the sector S shrinks to the positive real semiaxis. Naturally, these results are true if (0.2) is a series in fractional powers of x^{-1}.

We would like to mention here several observations that initiated our research. The first one is that the requirement of existence of a formal solution (0.2) is too strong. In fact, it suffices to show the existence of an approximate solution

$$y_N(x) = \sum_{k=1}^{N} y_k x^{-k} \tag{0.5}$$

such that

$$\mathcal{N} y_N = o(x^{-N}), \quad x \to \infty, \tag{0.6}$$

provided that N is large enough. The estimate on such N can be found in [T2].

Second, the existence of a formal solution (0.2), as shown by the following elementary example, is a luxury that sometimes cannot be afforded.

Example 0.1. Consider the Riccati equation

$$y'(x) = \frac{\theta}{x} + \frac{u(x)}{x^\beta} + \alpha y(x) + y^2(x) \tag{0.7}$$

at the singular point $x = \infty$, where θ, α are complex constants and $\beta > 0$. If one chooses $u(x)$ to be $\ln x$, $\sin x$, or any other function that does not admit a power series asymptotic expansion at $x = \infty$, then (0.7) cannot have formal power series solutions in x^{-1}, and so we are beyond the range of validity of the above mentioned results concerning the existence of an actual solution.

This example considerably restricts the range of validity of analytic methods, because it is intuitively clear that for large β the asymptotic behavior of solutions of (0.7) does not depend very much on whether $u(x) = x$ (a formal power series solution exists) or $u(x) = x \ln x$ (a formal power series solution does not exist).

Finally, it seems that a certain "gap" between the methods and results in the analytic and "nonanalytic" (i.e., finitely smooth or even L_p) theories of singular differential equations has become visible recently. This gap could be considered as a gap between formal algebraic and complex-analytic techniques on one hand, and a real-analysis technique on the other. (Compare, for example, the approach in [Wa] versus that in [Es].)

The main objective of this paper is to extend the algebraic methods used in the analytical and C^∞ cases to the finitely smooth case, i.e., to the case when the coefficients $f_\alpha(x)$ in (0.4) are only finitely smooth. Note that in this situation a formal

solution (0.2) to (0.1) cannot be defined. Therefore, the requirement of existence of (0.2) has to be replaced by the existence of an approximate solution (see Definition 0.1 below).

If the equation (0.1) is "sufficiently" smooth and if it possesses a "sufficiently" accurate approximate solution $y_0(x)$, then there exists an actual solution $y(x)$ of (0.1) that is "close" to $y_0(x)$. Theorem A below presents a rigorous formulation of this statement. However, much sharper estimates of the required amount of smoothness of (0.1) and amount of accuracy of $y_0(x)$ can be expressed in terms of "degeneracy" of the equation (0.1), linearized with respect to $y_0(x)$ (Theorem B). Theorem B, applied to Example 0.1, states that in the cases $\Re\alpha \neq 0, \beta \geq 0$ or $\Re\alpha = 0, \beta \geq 2$ the equation (0.7) possesses a smooth solution $y(x)$ such that $\lim_{x\to\infty} y(x) = 0$ for any $\theta \in \mathbb{C}$ and any smooth function $u(x)$ that is bounded at infinity.

In sections 1–3 the proof of Theorem B is given. It is also shown that Theorem A is a simple consequence of Theorem B. In section 4 these theorems are applied to problems of triangularization and of diagonalization of the $n \times n$ matrix linear differential equation

$$D_A Y \equiv x^{1-r} Y'(x) - A(x)Y(x) = 0. \tag{0.8}$$

The "amount" of smoothness of $A(x)$, which allows triangularization and diagonalization (under the additional assumption that $A(\infty)$ has n distinct eigenvalues) (0.8), has been estimated in Theorem 4.2 and Corollary 4.2, respectively. Theorems 4.1 and 4.3 contain corresponding statements for block triangularization and block diagonalization. Based on the triangularization theorem (Theorem 4.2), the classical theorem on asymptotic solution of a system of linear ODEs with irregular singularity (see, for example, [Wa, Theorem 19.1]) is extended (Corollary 4.1) to finitely smooth systems. One of the possible applications of the obtained results in the oscillation theory is discussed in Example 4.1.

DEFINITION 0.1. *For a given $m \in \mathbb{R}$ a function $y_0(x)$ is called an* m-approximate *solution of* (0.1) *if $y_0 \in C^1[x_0, \infty)$ and if there exists some $\delta > 0$ such that*

$$\mathcal{N} y_0 = O(x^{-m-\delta}) \quad as \quad x \to \infty. \tag{0.9}$$

DEFINITION 0.2. *For a given $k \in \mathbb{R}$ we say that a matrix-valued function $F(x) \in \mathcal{R}_k$ if $F(x)$ is continuous on $[x_0, \infty)$ and if there exists some $\delta > 0$ such that*

$$F(x) = A(x) + O(x^{-k-\delta}), x \to \infty, \tag{0.10}$$

where $A(x)$ is a polynomial in $\frac{1}{x}$.

We say that $F(x)$ is a C^{k+1} matrix-valued function on $[x_0, \infty]$ if $\Psi(\xi) = F(\frac{1}{\xi})$, where $\xi = \frac{1}{x}$, is a $C^{k+1}[0, x_0^{-1}]$ matrix-valued function. Then, according to the Taylor theorem,

$$\Psi(\xi) = \sum_{j=0}^{k} \Psi_j \xi^j + O(\xi^{k+1}), \quad \xi \to +0,$$

so $F(x) \in \mathcal{R}_k$. Thus, the requirement $F(x) \in \mathcal{R}_k$ is a requirement on smoothness of $F(x)$ at $x = \infty$.

THEOREM A. *Suppose there exists an m-approximate solution $y_0(x)$ of* (0.1) *such that $y_0(\infty) = 0$ and that $F(x) = \frac{\partial f}{\partial y}(x, y_0(x)) \in \mathcal{R}_k$ for some $k \in \mathbb{N}$. Then the conditions*

$$m \geq 2nr, \quad k \geq nr \tag{0.11}$$

imply the existence of an actual solution $y(x) \in C^1[\tilde{x}, \infty]$ *for some* $\tilde{x} \geq x_0$*, where*

$$y(x) - y_0(x) = o(x^{-m+nr}), \quad x \to \infty. \tag{0.12}$$

Remark 0.1. Theorem A is also valid in the case when $F(x)$ can be represented as a polynomial in fractional powers of $\frac{1}{x}$ plus a term of the order $O(x^{-k-\delta})$.

Remark 0.2. Conditions (0.11) in Theorem A guarantee the existence of an actual solution to (0.1) without any restriction on the Jacobian $F(x)$. However, these conditions can be weakened significantly if we make some assumptions about the Jacobian. For example, if all the eigenvalues of $F(\infty)$ have nonzero real parts, then the conditions (0.11) can be replaced by $m \geq 0, \quad k \geq 0$.

In what follows we introduce the notion of the *rank of degeneracy* $\nu(A)$ of the linear singular differential operator D_A defined by (0.8), where $r \in \mathbb{N}$ and $A(x)$ is an analytic matrix-valued function at $x = \infty$, and prove the following theorem.

THEOREM B. *Conditions* (0.11) *in Theorem* A *can be replaced by*

$$m \geq 2\nu(A), \quad k \geq \nu(A), \tag{0.13}$$

where the polynomial $A(x)$ (in $\frac{1}{x}$) *is defined by* $F(x)$ *according to* (0.10)*. Correspondingly,* (0.12) *becomes*

$$y(x) - y_0(x) = o(x^{-m+\nu(A)}), \ x \to \infty. \tag{0.14}$$

This theorem ties $\nu(A)$ with the values of m and k. Roughly speaking, it ties the degeneracy of the Jacobian of (0.1) with the "order" of approximation of $y_0(x)$ and with the amount of smoothness of the Jacobian $F(x)$ at $x = \infty$ in such a way that the existence of an actual solution $y(x)$, which is "close" to $y_0(x)$, is ensured. The rank of degeneracy for analytic equations was defined in [T2]. In section 1 we adjust this definition for equations of a real variable. It will follow immediately from this definition that Theorem A, as well as Remark 0.2, are particular cases of Theorem B. The latter case, in fact, coincides with Theorem 33.1 in [Wa].

Remark 0.3. In the case of equation (0.7) the rank of degeneracy $\nu(A) = 0$ if $\Re\alpha \neq 0$ and $\nu(A) = 1$ if $\Re\alpha = 0$. Correspondingly, we have to require $m = 0$ and $m = 2$. In the first case $y_0 \equiv 0$ is a 0-approximate solution provided $\beta > 0$ and $u(x)$ is bounded on $[x_0, \infty)$. In the second case the required 2-approximate solution is $y_0(x) = -\frac{\theta}{\alpha x} + \frac{\theta}{x^2\alpha^2}(1 - \frac{\theta}{\alpha})$ if $\alpha \neq 0$ and $y_0(x) = i\sqrt{\frac{\theta}{x}} - \frac{1}{4x} - \frac{3i}{32\sqrt{\theta}x^{3/2}}$ if $\alpha = 0$, provided $\beta > 2$ and $u(x)$ is bounded on $[x_0, \infty)$. It is easy to check that $F(x) \equiv \alpha + 2y_0(x)$, so the requirement on the Jacobian is satisfied in both cases.

Remark 0.4. The Riccati equation (0.7) appears in many applied problems, for example, in traveling wave solutions to the Burgers equation. The fundamental role of the latter equation in nonlinear wave phenomena is well recognized in the literature (see, for example, [Wh]). It may appear, though, that the Burgers equation is of limited value for modelling turbulence, since it is an integrable equation. One of the ways to introduce some chaos in the model is to consider a driven equation. That was done, for example, in [MK] to study shock-trains in the Burgers model (see also references there). Assuming that the driving term is time independent in some reference frame, after one integration we get the Riccati equation $y'(x) = y^2(x) + \alpha y(x) + h(x)$ for a traveling wave solution. Here $h(x)$ and α are determined by the driving term and speed of the wave, respectively. Assuming further that $h(x)$ is in the same form as in (0.7), we get conditions guaranteeing existence of a traveling wave, decaying at infinity, from Remark 0.3.

A similar question can be considered for driven nonlinear oscillators: under what conditions on the external force $g(x)$ is the oscillator

$$y'' + y = f(y) + g(x) \tag{0.15}$$

asymptotically stable at the origin? Here $f(y)$ contains only quadratic or higher order terms. It is easy to show that even in the linear case the decaying forcing $g(x) = O(x^{-1})$ does not guarantee stability. Indeed, the general solution of (0.15) with $f(y) \equiv 0,\ g(x) = \frac{\cos x}{x}$,

$$y(x) = c_1 \cos x + c_2 \sin x + \frac{1}{2} \sin x \ln x + \frac{1}{2} \int_{\infty}^{x} \frac{\sin(x-2t)}{t} dt,$$

is not bounded as $x \to \infty$ for any c_1, c_2. On the other hand, Theorem B guarantees the existence of a decaying solution if, for example, $y_0 \equiv 0$ is a 2-approximate solution to (0.15), that is, if $g(x) = O(x^{-2-\delta})$ with any $\delta > 0$. Indeed, as it follows from Definition 1.1, section 1, in the case of equation (0.15) $\nu(A) = \rho(A) = 1$, so $m \geq 2$ in Theorem B.

The proof of Theorem B is divided into

(1) preliminary normalization of the equation (0.1);

(2) contractibility of the corresponding integral operator;

(3) existence of $y(x)$ by the fixed point method.

1. Preliminary normalization.

1.1. Equation for the remainder term. Let $y_0(x)$ be an m-approximate solution of (0.1) with a corresponding $\delta > 0$ satisfying (0.9). Without loss of generality we assume that $F(x)$ satisfies (0.10) with the same δ. Then the substitution $y(x) = y_0(x) + z(x)$ yields the equation

$$x^{1-r} z'(x) = a(x) + F(x)z + h(x, z) \tag{1.1}$$

for the remainder $z(x)$, where

$$a(x) = f(x, y_0) - x^{1-r} y_0'(x) = O(x^{-m-\delta}), \quad x \to \infty; \tag{1.2}$$

$$F(x) = \frac{\partial f}{\partial y}(x, y_0) = \sum_{j=0}^{k} A_j x^{-j} + O(x^{-k-\delta}), \quad x \to \infty \tag{1.3}$$

(this follows from (0.10));

$$h(x, z) = f(x, y_0 + z) - f(x, y_0) - \frac{\partial f}{\partial y}(x, y_0) z. \tag{1.4}$$

Note that there exists some $x_1 \geq x_0$ such that

$$\|h(x, z)\| = O(\|z\|)^2 \quad \text{as} \quad z \to 0 \tag{1.5}$$

uniformly for $x \geq x_1$. Indeed, $y_0(\infty) = 0$ implies the existence of some $x_1 \geq x_0$ such that $\|y_0(x)\| < \zeta$ if $x \geq x_1$. Then (1.5) follows from the assumptions on $f(x, y)$ in (0.1).

1.2. Triangularization of linear differential operators. The problem of triangularization of holomorphic and continuous matrix-valued functions by means of similarity transformations was studied in [Br], [Lv], [Fr], [GH], [KT], [T1], and others. In the latter two papers it was shown how the triangularization of a matrix-valued function $A(x)$, analytic at $x = \infty$, can be extended to singular differential operators (0.8). We need the following statement.

THEOREM 1.1 (see [T1]). *Let the matrix-valued function $A(x)$ in* (0.8) *be analytic at $x = \infty$. Then there exists a number $p \in \mathbb{N}$ and a formal matrix series*

$$T(x) = T_0 + T_1 x^{-1/p} + T_2 x^{-2/p} + \cdots \tag{1.6}$$

(where T_j are matrices of complex numbers and T_0 is an invertible matrix) such that the transformation $Y(x) = T(x)Z(x)$ reduces the differential operator

$$D_A Y = x^{1-r}\frac{dY(x)}{dx} - A(x)Y(x) \tag{1.7}$$

into a triangular form (T-form)

$$D_B Z = x^{1-r}\frac{dZ(x)}{dx} - B(x)Z(x), \tag{1.8}$$

where $B(x)$ is an upper-triangular formal matrix series in $x^{-\frac{1}{p}}$. Moreover, the matrix $\operatorname{diag}B(x)$ *of diagonal entries of $B(x)$ is a polynomial in $x^{-1/p}$ of order not larger than pr. This matrix is invariant modulo $O(x^{-r})$ over all the T-forms of* (0.8).

Remark 1.1. Matrix $B(x)$ in (1.8) can be taken to be lower triangular as well.

Remark 1.2. The series (1.6) with an invertible matrix T_0 will be called a formal matrix series, holomorphic in $x^{-1/p}$. The series

$$V(x) = \sum_{\nu=\beta}^{\infty} V_\nu x^{\frac{-\nu}{p}},$$

where V_ν are matrices of complex numbers, $\beta \in \mathbb{Z}$, and $\det V(x) \not\equiv 0$, will be called a formal matrix series, meromorphic in $x^{-1/p}$.

Proof. The proof is based on two facts:

(a) There exists a formal matrix series $V(x)$, meromorphic in $x^{-1/p}$, that reduces (0.8) to its Jordan form D_J, where

$$J(x) = \begin{pmatrix} \lambda_1(x) & \delta_1 & 0 & \cdots & 0 \\ 0 & \lambda_2(x) & \delta_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_{n-1}(x) & \delta_{n-1} \\ 0 & \cdots & \cdots & 0 & \lambda_n(x) \end{pmatrix}. \tag{1.9}$$

Here $\lambda_j(x)$ are polynomials in $x^{-1/p}$ of order not larger than pr and the numbers δ_j are either 0 or 1 (the latter is possible only when $\lambda_j(x) \equiv \lambda_{j+1}(x)$). The polynomials $\lambda_j(x)$ are defined uniquely modulo $O(x^{-r})$. The proof of this well-known statement can be found, in fact, in [Wa, section 19.5] (see also [BJL], [KT], [T1]);

(b) Given $V(x)$, a formal matrix series meromorphic in $x^{-1/p}$, it is well known that there exists a unique factorization

$$V(x) = T(x)P(x)(x^{1/p})^K,$$

where $T(x)$ is a formal matrix series holomorphic in $x^{-1/p}$, $P(x)$ is an upper-triangular matrix polynomial in $x^{1/p}$, $P(0) = \mathrm{diag}P(x) = I$, and K is a diagonal matrix of integer numbers. This factorization was known to G. Birkhoff and could be found, for example, in [BJL].

According to (a), the transformation $Y(x) = V(x)Z(x)$ reduces (1.6) to its Jordan form D_J, where J is given by (1.9). Then

$$J(x) = V^{-1}(x)A(x)V(x) - x^{1-r}V^{-1}(x)V'(x). \tag{1.10}$$

At the same time, according to (b),

$$V(x) = T(x)U(x), \tag{1.11}$$

where $U(x) = P(x)(x^{1/p})^K$ is an upper-triangular matrix. It follows now from (1.10)–(1.11) that

$$U(x)J(x)U^{-1}(x)+x^{1-r}U'(x)U^{-1}(x) = T^{-1}(x)A(x)T(x)-x^{1-r}T^{-1}(x)T'(x). \tag{1.12}$$

So, the transformation $Y(x) = T(x)Z(x)$, applied to (1.7), yields (1.8) with

$$B(x) = U(x)J(x)U^{-1}(x) + x^{1-r}U'(x)U^{-1}(x).$$

Now one can easily verify that the matrix $B(x)$ satisfies all the assertions of the theorem. □

COROLLARY 1.1. *For any $N \in \mathbb{N}$ the transformation $Y(x) = T_N(x)Z(x)$, where*

$$T_N(x) = T_0 + T_1x^{-1/p} + \cdots + T_{Np}x^{-N} \tag{1.13}$$

is the truncated series (1.6), *reduces the operator D_A to its N-approximate triangular form D_{B_N} (T_N-form), where the matrix-valued function $B_N(x)$ is holomorphic in $x^{-1/p}$ and the first Np Taylor coefficients of $B_N(x)$ (in $x^{-1/p}$) are upper-triangular matrices.*

Proof. The fact that the formal transformation $T(x)$ reduces D_A to its triangular form D_B can be represented as an infinite series of equations on the corresponding coefficients A_n, B_n, and T_n. Then the corollary follows from the fact that the first Np equations are satisfied by the coefficients of $T_N(x)$. □

1.3. Rank of degeneracy of D_A and $\tilde{R}$-diagonal forms. Given a T-form (1.8) of the operator D_A, let us define the matrix

$$\tilde{\Lambda}(x) = \mathrm{diag}(\lambda_1(x), \dots, \lambda_n(x)), \tag{1.14}$$

where $x^r\lambda_j(x)$ consists of several consequent leading terms of the jth diagonal polynomial of $x^rB(x)$ (it is a polynomial in $x^{1/p}$ of order not greater than pr) defined as the following: the smallest term $\lambda_j x^{r+\rho_j}$ of $x^r\lambda_j(x)$ satisfies either $\Re\lambda_j \neq 0$ or $\rho_j = -r$; the coefficients of all the other terms of $x^r\lambda_j(x)$ are purely imaginary numbers. According to Theorem 1.1, the numbers ρ_j do not depend on a particular T-form D_B. Let us denote $\rho(A) = -\min_{1\le j\le n}\rho_j$ and

$$\tilde{R} = \mathrm{diag}(\rho_1, \dots, \rho_n).$$

Our aim is to reduce D_B into $D_{\hat{A}}$ with

$$\hat{A}(x) = x^{\tilde{R}}(\Lambda(x) + \tilde{B}(x)), \tag{1.15}$$

where $\tilde{B}(x) = o(1)$, $x \to \infty$, and $x^{\tilde{R}}\Lambda(x) = \tilde{\Lambda}(x)$ modulo $O(x^{-r})$. This can be achieved by the transformation $Z(x) = S(x)\tilde{Z}(x)$, where

$$S(x) = \mathrm{diag}(x^{-m_1}, x^{-m_2-\varepsilon}, \dots, x^{-m_n-(n-1)\varepsilon}). \tag{1.16}$$

This transformation is called a shearing transformation (see [Wa]) and $S(x)$ is called a shearing matrix. (For some historical remarks about shearing transformations see [Va]). The rational numbers $m_n \geq m_{n-1} \geq \cdots \geq m_1 = 0$ and the small positive number ε are to be specified below.

The action of the transformation (1.16) on the entries of $B(x)$ can be roughly described as follows: it multiplies the (i,j)th entry $B_{i,j}(x)$ of $B(x)$ by $x^{m_i-m_j+(i-j)\varepsilon}$. So, while the diagonal entries remain invariant modulo $O(x^{-r})$, the degree of each of the upper-triangular entries decreases and the degree of each of the lower-triangular entries increases.

Here and henceforth the degree of a scalar or a matrix series $a(x)$ in negative powers of the variable x means the exponent of the highest power of x occurring in the series; it is denoted by $\deg a(x)$; we set $\deg a(x) = -\infty$ if $a(x) \equiv 0$. It follows from (1.16) that $\deg S^{-1}(x) = m_n + (n-1)\varepsilon$. One can check that the minimal numbers m_j that satisfy (1.15) are recurrently defined by

$$m_k = \max\{m_{k-1}, \deg B_{1,k} - \rho_1 + m_1, \dots, \deg B_{k-1,k} - \rho_{k-1} + m_{k-1}\}, \tag{1.17}$$

where $k = 2, \dots, n$. It can be verified directly that

$$\deg \tilde{B}(x) \leq -\varepsilon \quad \text{and} \quad m_k \leq (k-1)\rho(A). \tag{1.18}$$

DEFINITION 1.1. *The rank of degeneracy $\nu(A)$ of the differential operator D_A is the minimum of*

$$\nu = \rho(A) + \deg\, S^{-1}(x) \tag{1.19}$$

over all T-forms of D_A, where ε is put to be zero.

The definition of $\nu(A)$ depends on $\tilde{R}$, that is, on the manner to choose ρ_j. The choice of ρ_j could be different if, for example, we consider x varying along the negative real semiaxis. In order to avoid confusion, $\nu(A)$ is called the $\tilde{R}$-rank of degeneracy and is denoted sometimes by $\nu(A, \tilde{R})$. In what follows we assume that (1.16) minimizes (1.19).

DEFINITION 1.2. *The operator $D_{\hat{A}}$, where $\hat{A}$ is given by* (1.15), *is called the $\tilde{R}$-diagonal form of D_A, and the matrix* (1.16) *is called its reducing shearing matrix.*

The fact that $\nu(A) = 0$ if all the eigenvalues of $A(\infty)$ have nonzero real parts follows immediately. The estimate

$$\nu(A) \leq n\rho(A) \tag{1.20}$$

is a direct consequence of (1.18). So $\nu(A) \leq nr$. Thus Theorem B implies the statement of Remark 0.2 and Theorem A.

1.4. The prenormalized equation. Let $A(x)$ be a polynomial part of the Jacobian $F(x)$ defined by (1.3), let $T_k(x)$ be a polynomial in $x^{-1/p}$ reducing D_A to its T_k-form (1.8), and let the shearing matrix $S(x)$ be defined by (1.8) according to (1.16)–(1.17). Then the transformation $z(x) = S(x)T_k(x)u(x)$ reduces (1.1) to

$$u'(x) = x^{r-1}\tilde{b}(x) + x^{\tilde{R}+r-1}[\Lambda(x) + \tilde{B}(x)]u(x) + x^{r-1}\tilde{F}(x)u + x^{r-1}\tilde{g}(x,u), \tag{1.21}$$

where

(a) $\tilde{b}(x) = T_k^{-1}(x)S^{-1}(x)a(x)$;

(b) $\tilde{R}$ is defined above. In the case $\rho(A) < r$ we put $\Lambda(x) = x^{-\tilde{R}}\tilde{\Lambda}(x)$ as in (1.15). In the opposite case we add $(m_j + (j-1)\varepsilon)\frac{1}{x}$ to the jth entry of the diagonal matrix $x^{-\tilde{R}}\tilde{\Lambda}(x)$ for every j such that $\rho_j = -r$;

(c) $\tilde{F}(x) = T_k^{-1}(x)S^{-1}(x)[F(x) - A(x)]S(x)T_k(x)$;

(d) $\tilde{g}(x,u) = T_k^{-1}(x)S^{-1}(x)h(x, S(x)T_k(x)u(x))$.

Let R denote $\tilde{R}+(r-1)I_n$, where I_n denotes the $n\times n$ identity matrix. Rewriting (1.21) as

$$u' = x^R(b(x) + [\Lambda(x) + G(x)]u + g(x,u)) \tag{1.22}$$

and using deg $x^{-\tilde{R}}S^{-1}(x) = \nu(A) + (n-1)\varepsilon$, we obtain the following:

(a)

$$b(x) = x^{-\tilde{R}}\tilde{b}(x) = O(x^{-m-\delta+\nu(A)+(n-1)\varepsilon}), \quad x \to \infty; \tag{1.23}$$

according to (1.2) and (1.21)(a),

(b) $G(x) = \tilde{B}(x) + x^{-\tilde{R}}\tilde{F}(x)$. As it follows from (1.3) and (1.21)(c),

$$x^{-\tilde{R}}\tilde{F}(x) = O(x^{-k-\delta+\nu(A)+(n-1)\varepsilon}), \quad x \to \infty. \tag{1.24}$$

The choice $\varepsilon < \frac{\delta}{n}$ together with (0.13) and (1.18) implies $x^{-\tilde{R}}\tilde{F}(x) = O(x^{-\varepsilon}), \quad x \to \infty$, and

$$G(x) = O(x^{-\varepsilon}), \quad x \to \infty; \tag{1.25}$$

(c) $g(x,u) = x^{-\tilde{R}}\tilde{g}(x,u)$, so

$$\frac{\partial g}{\partial u}(x,u) = x^{r-1}S^{-1}(x)T_{k-1}^{-1}(x)\frac{\partial h}{\partial z}\big(x, T_{k-1}(x)S(x)u\big)T_{k-1}(x)S(x).$$

According to (1.4),

$$\frac{\partial h}{\partial z}(x,z) = \frac{\partial f}{\partial y}(x, y_0 + z) - \frac{\partial f}{\partial y}(x, y_0).$$

Thus, similar to (1.5), we get $\|\frac{\partial h}{\partial z}(x,z)\| = O(\|z\|), \ z \to 0$, uniformly in $x \geq x_1$. Then, there exist constants $\hat{g} > 0$ and $\rho > 0$ such that

$$\left\|\frac{\partial g}{\partial u}(x,u)\right\| \leq x^{\nu(A)+(n-1)\varepsilon}\hat{g}\|u\| \tag{1.26}$$

for all $x \geq x_1$ and $\|z\| < \rho$.

The equation (1.22) is called the prenormalized form of (0.1) with respect to the approximate solution $y_0(x)$.

2. Contractiveness of integral operators.

2.1. Integral equation. Consider the equation (1.22) as a perturbation of the linear equation

$$U'(x) = x^R \Lambda(x) U(x). \tag{2.1}$$

This equation has the fundamental matrix solution

$$U(x) = e^{Q(x)}, \tag{2.2}$$

where $Q'(x) = x^R \Lambda(x)$. The entries $q_j(x)$ of the diagonal matrix $Q(x)$ can be represented as

$$q_j(x) = \tilde{q}_j(x) + c_j \ln x, \tag{2.3}$$

where c_j is a complex constant and $\tilde{q}_j(x)$ is a polynomial in $x^{1/p}$ of order not more than pr and $\tilde{q}_j(0) = 0$. Note that according to the construction of $\Lambda(x)$, in the case $\rho_j > -r$ all but one coefficient of $\tilde{q}_j(x)$ are purely imaginary and $c_j = 0$, while in the case $\rho_j = -r$ all these coefficients are purely imaginary.

The equation (1.22) can be converted into the integral equation

$$u(x) = e^{Q(x)} \left[C + \int^x e^{-Q(t)} t^R (b(t) + G(t)u + g(t,u)) dt \right], \tag{2.4}$$

where $C \in \mathbb{C}^n$ is an arbitrary constant vector. We have a freedom to choose a lower limit of integration for each entry of the vector integrand in (2.4) independently on that of other entries.

2.2. The integral operator $\mathcal{I}_Q$. Let Φ_γ, $\gamma \in \mathbb{R}$, denote the linear space of n-dimensional vector-valued functions $v(x) \in C[x_0, \infty)$, which are majorized by

$$\|v(x)\| \le M x^{-\gamma}, \quad x \in [x_0, \infty), \tag{2.5}$$

where the constant M depends on $v(x)$.

The lower limits of integration $s = \{s_1, \dots, s_n\}$ of the integral operator

$$\mathcal{I}_Q v = e^{Q(x)} \int_s^x e^{-Q(t)} t^R v(t) dt, \tag{2.6}$$

which acts on Φ_γ, are defined by γ and by the matrix $Q(x)$ as follows:

(a) If $\Re[\tilde{q}_j(x)] \to +\infty$ as $x \to \infty$, then $s_j = \infty$;

(b) If $\Re[\tilde{q}_j(x)] \to -\infty$ as $x \to +\infty$, then $s_j = x_0$;

(c) If $\Re[\tilde{q}_j(x)] \equiv 0$, then $s_j = x_0$ if $-\gamma - \Re c_j > 0$ or $s_j = +\infty$ if $-\gamma - \Re c_j < 0$.

We can always assume that

$$-\gamma - \Re c_j \neq 0 \tag{2.7}$$

by making a small variation of γ, if necessary.

THEOREM 2.1. *For a given diagonal matrix $Q(x)$ with entries* (2.3) *and a given $\gamma \in \mathbb{R}$ satisfying* (2.7) *there exists a constant $K > 0$ such that for any $v(x) \in \Phi_\gamma$ and satisfying* (2.5),

$$\|\mathcal{I}_Q v\| \le K M x^{-\gamma}, \quad x \in [x_0, \infty).$$

Proof. Since $Q(x)$ is a diagonal matrix, it is sufficient to consider the scalar case $n = 1$. So, let $Q(x) = q(x) + c \ln x$, where $q(x)$ is a polynomial in $x^{1/p}$.

(a, b) Suppose $\Re q(x) \not\equiv 0$, so either case (a) or (b) in the definition of lower limit s holds. Then we need to estimate the integral

$$e^{q(x)} x^c \int_s^x e^{-q(t)} t^{\bar{r}-1} t^{-c} t^{-\gamma} dt, \tag{2.8}$$

where $\lambda x^{\bar{r}}$ is the only term in $q(x)$ with nonzero real part. Since all other terms of $q(x)$ have purely imaginary coefficients, we can disregard them while estimating (2.8) for real positive x. Then (2.8) becomes

$$x^c \int_s^x e^{\alpha(x^{\bar{r}} - t^{\bar{r}})} t^{\bar{r}-1} t^{-c-\gamma} dt. \tag{2.9}$$

The assertion of the theorem for such integrals is well known (see, for example, [Wa, section 14]; the desired estimate for integrals (2.9) was derived in the course of estimating the integral (14.25) in Lemma 14.2 there).

(c) If $\Re q(x) \equiv 0$ then $|e^{q(x)}| \equiv 1$, so

$$|\mathcal{I}_Q v| = \left| x^c \int_s^x t^{-c-1} v(t) dt \right| \le M x^{\Re c} \left| \int_s^x t^{-\Re c - 1 - \gamma} dt \right|.$$

Suppose $-\gamma - \Re c < 0$. Then $s = \infty$ and

$$|\mathcal{I}_Q v| \le \frac{M}{|\gamma + \Re c|} x^{-\gamma}.$$

Suppose $-\gamma - \Re c > 0$. Then $s = x_0$ and

$$|\mathcal{I}_Q v| \le M x^{\Re c} \left. \frac{t^{-\gamma - \Re c}}{-\gamma - \Re c} \right|_{x_0}^{x} \le \frac{2M}{|\gamma + \Re c|} x^{-\gamma},$$

since $(\frac{x_0}{x})^{-\gamma - \Re c} < 1$. The statement of the theorem in this case follows from the estimate of $|\mathcal{I}_Q v|$. □

3. Fixed point method.

3.1. Convergence of iterations. The integral equation (2.4) can now be put in the operator form

$$u(x) = e^{Q(x)} C + \mathcal{I}_Q [b(x) + G(x) u(x) + g(x, u(x))], \tag{3.1}$$

where C is an arbitrary constant vector. We assume $C = 0$ and define the iterations by $u_0 \equiv 0$,

$$u_{j+1}(x) = \mathcal{I}_Q [b(x) + G(x) u_j(x) + g(x, u_j(x))], \tag{3.2}$$

where $j = 0, 1, 2, \ldots$.

Let us denote $\Delta u_j(x) = u_j(x) - u_{j-1}(x)$ and

$$\gamma = m - \nu(A) + \delta - (n-1)\varepsilon. \tag{3.3}$$

LEMMA 3.1. *Under the conditions* (0.13) *there exist some* $\tilde{x} \geq x_1$ *and* $M > 0$ *such that for every* $x \in [\tilde{x}, \infty)$ *the series*

$$u(x) = \sum_{j=1}^{\infty} \Delta u_j(x)$$

converges absolutely and uniformly and

$$\|u(x)\| \leq Mx^{-\gamma}. \tag{3.4}$$

Proof. 1. *Introduction.* The asymptotics (1.23), (1.25) can be converted into the inequalities

$$\|b(x)\| \leq \frac{1}{2}\hat{b}x^{-\gamma} \tag{3.5}$$

and

$$\|G(x)\| \leq \hat{G}x^{-\varepsilon} \tag{3.6}$$

for the appropriate constants $\hat{b}$, $\hat{G} > 0$, and $x \geq x_1$. Decreasing slightly, if necessary, the constant δ defined by (1.2)–(1.3), we can apply Theorem 2.1, with Q defined as in (3.1) and γ defined by (3.3), to estimate $u_1 = \mathcal{I}_Q b$. Then

$$\|u_1(x)\| < \frac{1}{2}K\hat{b}x^{-\gamma} \tag{3.7}$$

for $x \in [x_1, \infty)$.

Let us assume

$$\|\Delta u_l(x)\| \leq \frac{K\hat{b}x^{-\gamma}}{2^l}, \quad x \in [\tilde{x}_0, \infty), \tag{3.8}$$

for $l = 1, \dots, j$ and prove (3.8) for $l = j + 1$. Then the statement of Lemma 3.1, where $M = K\hat{b}$, will follow from (3.8) by induction.

According to (3.2)

$$\Delta u_{j+1}(x) = \mathcal{I}_Q G(x)\Delta u_j(x) + \mathcal{I}_Q \Delta g(x, u_j(x)), \tag{3.9}$$

where $\Delta g(x, u_j(x)) = g(x, u_j(x)) - g(x, u_{j-1}(x))$.

2. *Estimate of the linear term in* (3.9). The constant K, defined in Theorem 2.1, depends on Q and Φ_γ. Without loss of generality we may assume that K is the same for Φ_γ and $\Phi_{\gamma+\varepsilon}$. Let $\tilde{x} \geq (4\hat{G}K)^{1/\varepsilon}$. Then, taking into account (3.5), (3.7), (3.8), and Theorem 2.1,

$$\|\mathcal{I}_Q G(x)\Delta u_j(x)\| \leq \frac{K\hat{b}}{2^j}\hat{G}\|\mathcal{I}_Q x^{-\gamma-\varepsilon}\| \leq \frac{K\hat{b}\hat{G}}{2^j}Kx^{-\gamma-\varepsilon} \leq \frac{1}{2}\frac{K\hat{b}}{2^{j+1}}x^{-\gamma}. \tag{3.10}$$

3. *Estimate of the nonlinear term in* (3.9). We use the well-known formula

$$g(y_2) - g(y_1) = \int_0^1 \frac{\partial g}{\partial y}(x, sy_2 + (1-s)y_1)ds \cdot (y_2 - y_1) \tag{3.11}$$

to estimate the nonlinear term in (3.9). According to (3.11) we get

$$\|\Delta g(x, u_j(x))\| \leq \left\| \int_0^1 \frac{\partial g}{\partial y}(x, su_j + (1-s)u_{j-1})ds \right\| \|\Delta u_j\|. \tag{3.12}$$

Without loss of generality we may assume that $\tilde{x}$ is so large that $K\hat{b}|\tilde{x}|^{-\gamma} < \rho$, where ρ was defined in (1.26). Then

$$\|su_j(x) + (1-s)u_{j-1}(x)\| \le K\hat{b}x^{-\gamma} < \rho \tag{3.13}$$

for $x \in [\tilde{x}, \infty)$. So, according to (1.26),

$$\left\| \frac{\partial g}{\partial u}(x, su_j(x) + (1-s)u_{j-1}(x)) \right\| \le \hat{g}x^{\nu(A)+(n-1)\varepsilon} K\hat{b}x^{-\gamma}. \tag{3.14}$$

In order to estimate the right-hand side of (3.14) by $\hat{g}K\hat{b}x^{-\varepsilon}$ we need

$$-m - \delta + 2\nu(A) + 2(n-1)\varepsilon \le -\varepsilon. \tag{3.15}$$

Due to (0.13), this inequality is satisfied if $\varepsilon < \frac{\delta}{(2n-1)}$. Then (3.14) yields

$$\|\Delta g(x, u_j(x))\| \le \hat{g}K\hat{b}x^{-\varepsilon}\|\Delta u_j\|.$$

Now one can repeat the arguments of (3.10) in order to get

$$\|\mathcal{I}_Q \Delta g(x, u_j)\| \le \frac{1}{2}\frac{K\hat{b}}{2^{j+1}}x^{-\gamma}.$$

This proves (3.8) for $l = j+1$. □

3.2. Theorem B. The solution $y(x)$ to the equation (0.1) can now be represented as

$$y(x) = y_0(x) + T_k(x)S(x)u(x), \tag{3.16}$$

where $u(x)$ is a solution of the differential equation (1.22) (or of the equivalent equation (3.1)). The asymptotics (0.14) follow from (3.4). The proof of Theorem B is completed.

Remark 3.1. Suppose the diagonal matrix $e^{Q(x)}$ contains $l \le n$ entries that decrease exponentially as $x \to \infty$. Then the arguments of Lemma 3.1 will keep valid for an arbitrary constant vector C in (3.1) chosen from the corresponding l-dimensional subspace of $\mathbb{C}^n$. In this case Theorem B asserts the existence of an l-parameter family of solutions to (0.1).

Remark 3.2. In the case when the equation (0.1) is a linear nonhomogeneous equation, conditions $m \ge 2nr$ and $m \ge 2\nu(A)$ in Theorems A and B can be replaced by $m \ge nr$ and by $m \ge \nu(A)$, respectively. Indeed, in this case there is no need to estimate the nonlinear term in (3.9), so inequality (3.15) for m can be replaced by the requirement that γ in (3.3) is nonnegative.

4. Simplification of finitely smooth linear systems.

4.1. Block-diagonalization. Setting of the problem. Consider an $n \times n$ matrix differential equation

$$x^{1-r}Y'(x) = A(x)Y(x), \tag{4.1}$$

where $r \in \mathbb{N}$ and the $n \times n$ matrix-valued function $A(x) \in \mathcal{R}_k$ for some $k \in \mathbb{N}$ and $x_0 \in \mathbb{R}$.

Suppose that the spectrum of $A(\infty)$ consists of two nonempty disjoint sets σ_1 and σ_2; we can therefore assume $A(\infty)$ to be a block-diagonal matrix

$$A(\infty) = \operatorname{diag}(A_0^1, A_0^2), \quad \text{where} \quad \sigma_1 \cap \sigma_2 = \emptyset. \tag{4.2}$$

Let $l > 0$ denote the dimension of the square matrix A_0^1. Then the natural question is whether the equation (4.1) can be completely decoupled into two equations of dimensions l and $n - l$. These problems are known as block-diagonalization problems for matrix linear differential equations. Problems of block-diagonalization and of block-triangularization for finitely smooth equations are discussed below. Recurrent applications of block simplification transformations can lead eventually to full diagonalization or full triangularization of a system. These problems are also considered below.

4.2. Reduction to nonlinear equation. Let us look for a linear transformation

$$Y(x) = P(x)Z(x) \tag{4.3}$$

that reduces (4.1) to the desired block-diagonal system

$$x^{1-r}Z'(x) = B(x)Z(x). \tag{4.4}$$

Here

$$B(x) = \operatorname{diag}(B^{11}(x), B^{22}(x)) \tag{4.5}$$

and $B^{11}(\infty) = A_0^1, B^{22}(\infty) = A_0^2$.

One can immediately check that the transformation (4.3) reduces (4.1) to (4.4) iff

$$x^{1-r}P'(x) = A(x)P(x) - P(x)B(x). \tag{4.6}$$

According to (4.2), we represent

$$A(x) = \begin{pmatrix} A^{11}(x) & A^{12}(x) \\ A^{21}(x) & A^{22}(x) \end{pmatrix},$$

where $A^{11}(\infty) = A_0^1,\ A^{22}(\infty) = A_0^2,\ A^{12}(\infty) = 0,\ A^{21}(\infty) = 0$. If we are looking for $P(x)$ in the form

$$P(x) = \begin{pmatrix} I_l & P^{12}(x) \\ P^{21}(x) & I_{n-l} \end{pmatrix}, \tag{4.7}$$

then the substitution of (4.7) into (4.6) together with (4.5) yields

$$\left\{ \begin{array}{c} 0 = A^{12}(x)P^{21}(x) + A^{11}(x) - B^{11}(x), \\ x^{1-r}\dfrac{dP^{12}(x)}{dx} = A^{11}(x)P^{12}(x) - P^{12}(x)B^{22}(x) + A^{12}(x), \\ x^{1-r}\dfrac{dP^{21}(x)}{dx} = A^{22}(x)P^{21}(x) - P^{21}(x)B^{11}(x) + A^{21}(x), \\ 0 = A^{21}(x)P^{12}(x) + A^{22}(x) - B^{22}(x) \end{array} \right\} \tag{4.8}$$

(see, for instance, [Wa, section 12]). This system can be split into two decoupled nonlinear equations

$$x^{1-r}\frac{dP^{12}(x)}{dx} = A^{12}(x)+A^{11}(x)P^{12}(x)-P^{12}(x)A^{22}(x)-P^{12}(x)A^{21}(x)P^{12}(x) \tag{4.9}$$

and into a similar equation for P^{21}.

We say that the transformation (4.3) block-triangularizes (4.1) if instead of a block-diagonal matrix $B(x)$ in (4.4) we get a block-triangular matrix, say,

$$B(x) = \begin{pmatrix} B^{11}(x) & 0 \\ B^{21}(x) & B^{22}(x) \end{pmatrix}. \tag{4.10}$$

One can check directly that the transformation

$$P(x) = \begin{pmatrix} I_l & P^{12}(x) \\ 0 & I_{n-l} \end{pmatrix} \tag{4.11}$$

reduces (4.1) to (4.4), (4.10) if $P^{12}(x)$ satisfies (4.9).

4.3. Finitely smooth block-triangularization. Let us first consider problems of block-triangularization. As we will see, we can carry out the block-triangularization of equation (4.1) regardless of the condition (4.2), so this condition will be abolished in this section.

In what follows we suppose that, according to Corollary 1.1, the equation (4.1) is reduced to its k-approximate triangular form. Taking into account Remark 1.1, we get

$$A^{12}(x) = O(x^{-k-\delta}), \quad x \to \infty, \tag{4.12}$$

for some $\delta > 0$, and, consequently, that $P_0^{12}(x) \equiv 0$ is a k-approximate solution of (4.9). The fact that $\frac{\partial f}{\partial y}(x,0) \in \mathcal{R}_k$ for (4.9) can be checked directly. Then Theorem A, applied to (4.9), yields the following statement.

Statement 4.1. If the coefficient $A(x)$ of (4.1) belongs to $\mathcal{R}_k$ with

$$k \geq 2(n-l)lr, \tag{4.13}$$

then there exists a transformation (4.3), (4.11), where $P^{12}(x) = o(x^{-k+(n-l)lr})$, that reduces (4.1) to (4.4), (4.10). In the other words, the transformation (4.3), (4.11) block-triangularizes (4.1).

For a given equation (4.1) the requirement (4.13) is very crude. The more refined statement of Theorem B estimates k via the rank of degeneracy $\nu(A^{11}, A^{22})$ of the linear operator $Tr_A X = A^{11}(x)X - XA^{22}(x)$. However, instead of computing $\nu(A^{11}, A^{22})$ directly, we prefer to improve the estimate (4.13) by studying equation (4.9) in the integral form and applying to it the same arguments as in the proof of Theorem B.

Let us denote $A(x) = \tilde{A}(x) + \tilde{G}(x)$, where $\tilde{A}(x)$ is a polynomial in $x^{-1/p}$ of order not greater than pk and $\tilde{G}(x) = O(x^{-k-\delta})$. We call $\tilde{A}(x)$ and $\tilde{G}(x)$ polynomial and small parts of $A(x)$, respectively. According to Theorem 1.1, the diagonal entries $\lambda_i(x)$ of $\tilde{A}(x)$ are polynomials in $x^{-1/p}$ of order not greater than pr. Let

$$\lambda_{ij}(x) = \lambda_i(x) - \lambda_j(x), \quad 1 \leq i \leq l,\ l+1 \leq j \leq n. \tag{4.14}$$

For polynomials $\lambda_{ij}(x)$ we define the numbers ρ_{ij} in the same manner as the numbers ρ_j were defined for polynomials (1.14): ρ_{ij} is either the order of the maximal term of $\lambda_{ij}(x)$ with a nonpurely imaginary coefficient or $\rho_{ij} = -r$ if such a term fails to exist. For $i = 1, \dots, l$ and $j = l+1, \dots, n$ we denote

(4.15)

$$\rho_i = \min_{l+1\le j\le n} \rho_{ij}, \quad \rho_j = \min_{1\le i\le l} \rho_{ij}, \quad \tilde{R}_1 = \operatorname{diag}(\rho_1, \dots, \rho_l), \quad \tilde{R}_2 = \operatorname{diag}(\rho_{l+1}, \dots, \rho_n)$$

and

$$\rho(A^{11}, A^{22}) = -\min_{1\le i\le l,\ l+1\le j\le n} \rho_{ij}. \tag{4.16}$$

To study (4.9) we need to introduce differential operators

$$D_A^T Y = x^{1-r}\frac{dY(x)}{dx} + Y(x)A(x)$$

that are adjoint to (1.7).

DEFINITION 4.1. *The operator $D^T_{\hat{A}^T}$ is called an $\tilde{R}$-diagonal form of D^T_A if*

$$\hat{A}^T = (\Lambda(x) + \tilde{B}(x))x^{\tilde{R}},$$

where $\Lambda, \tilde{B}$ are defined in the same way as in Definition 1.2. *The corresponding reducing shearing transformation is defined similarly.*

Let

$$\hat{A}_1 = x^{\tilde{R}_1}(\Lambda_1(x) + \tilde{B}_1(x)) \ \text{ and } \ \hat{A}_2^T = (\Lambda_2(x) + \tilde{B}_2(x))x^{\tilde{R}_2} \tag{4.17}$$

be $\tilde{R}_1$-diagonal and $\tilde{R}_2$-diagonal forms of the operators $D_{A^{11}}$ and $D^T_{A^{22}}$, respectively. Let also $m_1 = \deg S_1^{-1}(x)$, $m_2 = \deg S_2^{-1}(x)$, where S_1 and S_2 are the corresponding shearing matrices with $\varepsilon = 0$.

We define

$$\nu = 2\rho(A^{11}, A^{22}) + m_1 + m_2. \tag{4.18}$$

Then, similar to (1.20), we get

$$\nu \le n\rho(A^{11}, A^{22}). \tag{4.19}$$

THEOREM 4.1. *The condition* (4.13) *in Statement* 4.1 *can be replaced by*

$$k \ge \nu. \tag{4.20}$$

Moreover,

$$P^{12}(x) = o(x^{-\tilde{k}-\rho(A^{11},A^{22})}), \quad x \to \infty, \tag{4.21}$$

where $\tilde{k} = k - \nu$ and

$$B^{11}(x), B^{22}(x) \in \mathcal{R}_{\tilde{k}+\rho(A^{11},A^{22})}. \tag{4.22}$$

Proof. The transformation $P^{12}(x) = S_1(x)U(x)S_2(x)$ reduces (4.9) to

$$x^{1-r}U'(x) = S_1^{-1}A^{12}S_2^{-1} + x^{\tilde{R}_1}[\Lambda_1 + \tilde{B}_1]U - U[\Lambda_2 + \tilde{B}_2]x^{\tilde{R}_2} - US_2A^{21}S_1U.$$

It can be rewritten as

(4.23)

$$U = e^{Q_1}\left\{\int_s^x e^{-Q_1}[t^{r-1}S_1^{-1}A^{12}S_2^{-1} + t^{R_1}\tilde{B}_1U - U\tilde{B}_2t^{R_2} - t^{r-1}US_2A^{21}S_1U]e^{Q_2}dt\right\}e^{-Q_2},$$

where $Q_1' = x^{R_1}\Lambda_1$, $Q_2' = x^{R_2}\Lambda_2$, $R_1 = \tilde{R}_1 + (r-1)I_l$, and $R_2 = \tilde{R}_2 + (r-1)I_{n-l}$. This equation is an $l \times (n-l)$-dimensional vector equation of the type (2.4). The set of lower limits of integration s is chosen as in (2.6).

Remark 4.1. The statement of Theorem 2.1 is valid for both integral operators

$$\mathcal{I}_1V = e^{Q_1}\left\{\int_s^x e^{-Q_1}t^{R_1}Ve^{Q_2}dt\right\}e^{-Q_2},$$

$$\mathcal{I}_2V = e^{Q_1}\left\{\int_s^x e^{-Q_1}Vt^{R_2}e^{Q_2}dt\right\}e^{-Q_2},$$

where $V(x) \in \Phi_\gamma$.

Let us rewrite (4.23) again as

$$U = e^{Q_1}\left\{\int_s^x e^{-Q_1}[t^{R_1}(b + \tilde{B}_1U - t^{-\tilde{R}_1}UhU) - U\tilde{B}_2t^{R_2}]e^{Q_2}dt\right\}e^{-Q_2}, \tag{4.24}$$

where

$$b(x) = x^{-\tilde{R}_1}S_1^{-1}A^{12}S_2^{-1} = O(x^{-k-\delta+\nu-\rho(A^{11},A^{22})+(n-2)\varepsilon}), \quad x \to \infty, \tag{4.25}$$

and

$$h(x) = S_2A^{21}S_1 = O(1), \quad x \to \infty. \tag{4.26}$$

The rest of the proof follows as in Lemma 3.1 with $\gamma = k + \delta - \nu + \rho(A^{11}, A^{22}) - (n-2)\varepsilon$. Note, however, that according to (4.26), the estimate (3.15) is replaced by

$$-k - \delta + \nu - \rho(A^{11}, A^{22}) + (n-2)\varepsilon + \rho(A^{11}, A^{22}) \ \leq -\varepsilon. \tag{4.27}$$

This yields (4.20) provided $\varepsilon < \frac{\delta}{n-1}$. The second statement follows then from (4.25). Finally, the substitution of (4.10)–(4.11) into (4.6) yields

$$B^{11} = A^{11} - P^{12}A^{21}, \quad B^{22} = A^{22} + A^{21}P^{12}, \tag{4.28}$$

so (4.22) follows from (4.21). □

Theorem 4.1 shows that the required amount of smoothness for block-triangularization is of the order $O(n)$ instead of $O(n^2)$ as could be seen from Statement 4.1.

4.4. Complete triangularization. Solutions of k-smooth systems. Let us define $\tilde{\rho}(A)$ by modifying the definition (4.16) of $\rho(A^{11}, A^{22})$ as

$$\tilde{\rho}(A) = -\min_{1\le i<j\le n} \rho_{ij}.$$

It is clear that

$$\tilde{\rho}(A) \ge \rho(A^{11}, A^{22}). \tag{4.29}$$

Then, as a consequence of Theorem 4.1, we obtain the following theorem.

THEOREM 4.2. *If the coefficient $A(x)$ of (4.1) belongs to $\mathcal{R}_k$ with*

$$k \ge 2(n-1)\tilde{\rho}(A), \tag{4.30}$$

then there exists a transformation (4.3), where $P(x)$ is an upper-triangular matrix and $\operatorname{diag} P(x) \equiv I_n$, that reduces (4.1) to (4.4), where $B(x) \in \mathcal{R}_{\tilde{k}+\tilde{\rho}(A)}$, $\tilde{k} = k - 2(n-1)\tilde{\rho}(A)$, and $B(x)$ is a lower-triangular matrix. Moreover, $P(x) = I_n + U(x)$, where

$$U(x) = o(x^{-\tilde{k}-\tilde{\rho}(A)}),\ x \to \infty. \tag{4.31}$$

Proof. Let us prove Theorem 4.2 by induction. For $n = 2$ triangularization and block-triangularization coincide. Note that in this case the statements of Theorems 4.1 and 4.2 also coincide. Suppose the statement of the theorem is true for any $n \le 2^m$, $m \in \mathbb{N}$. Let us then prove it for any n such that $2^m < n \le 2^{m+1}$.

We start to block-triangularize (4.1) by (4.3), (4.11), choosing $l = \frac{n}{2}$ if n is even or $l = \frac{n+1}{2}$ otherwise. Then, according to (4.22), (4.29), $B^{11}, B^{22} \in \mathcal{R}_{k_1}$, where $k_1 \ge k - n\rho(A^{11}, A^{22}) + \rho(A^{11}, A^{22}) \ge k - (n-1)\tilde{\rho}(A)$. Note that (4.21), (4.28) imply $\tilde{\rho}(A^{11}) = \tilde{\rho}(B^{11})$, $\tilde{\rho}(A^{22}) = \tilde{\rho}(B^{22})$ so that $\tilde{\rho}(B^{11}), \tilde{\rho}(B^{22}) \le \tilde{\rho}(A)$. Then using (4.30) we obtain

$$k_1 \ge \tilde{k} + 2(n-1)\tilde{\rho}(A) - (n-1)\tilde{\rho}(A) \ge \tilde{k} + 2(l-1)\tilde{\rho}(A).$$

Now the assertion of the theorem follows by induction arguments. ∎

In section 1.2 we described reduction of a singular differential operator D_A to its Jordan form. As a consequence of this reduction, one can get the classical Hukuhara–Turritin theorem (see, for example, [Wa, section 19.5]), stating that if $A(x)$ is a matrix-valued function, holomorphic at ∞, then (4.1) possesses a formal solution

$$Y(x) = V(x) x^H e^{Q(x)}, \tag{4.32}$$

where $V(x)$ is a matrix-valued function, holomorphic in some sector S in the complex x-plane for sufficiently large values of $|x|$, that admits the asymptotic expansion

$$\hat{V} = \sum_{j=0}^{\infty} V_j x^{-j/p}$$

in S with some $p \in \mathbb{N}$; $\det V(x) = O(x^{-m})$ for some $m \in \mathbb{N}$; $Q(x)$ is a diagonal, polynomial in $x^{1/p}$ matrix with $Q(0) = 0$; and H is a constant matrix that commutes with $Q(x)$.

Theorem 4.2 extends this result to the case $A(x) \in \mathcal{R}_k$ in the following way.

COROLLARY 4.1. *Under the assumption (4.30) there exists a solution (4.32) to (4.1), where $V(x) \in \mathcal{R}_{\tilde{k}}$, $\det V(x) = O(x^{-m})$ for some $m \in \mathbb{N}$; $Q(x)$ is a diagonal matrix, $x^{-r} Q(x) \in \mathcal{R}_{\tilde{k}+\tilde{\rho}(A)}$; and H is a constant diagonal matrix.*

Proof. According to Theorem 4.2, we can assume that $A(x)$ is a lower triangular matrix and $A(x) \in \mathcal{R}_{\tilde{k}+\tilde{\rho}(A)}$ in the notations of the theorem. Next we apply the shearing transformation $S(x) = \mathrm{diag}(x^{-m_n}, \dots, x^{-m_1})$, where $0 = m_1 \le m_2 \le \dots \le m_n$, that reduces D_A to D_B, where B is lower triangular and off-diagonal entries of B are $O(x^{-\tilde{k}-\tilde{\rho}(A)-\delta})$. The matrix B can be decomposed as $B(x) = \Lambda(x) + \tilde{B}(x)$, where $\Lambda(x) = \mathrm{diag}B(x)$ and $\tilde{B}(x)$ is strictly lower triangular. Then the change of variables

$$Y(x) = (I + Z(x))e^{Q(x)},$$

where $Z(x)$ is a strictly lower triangular matrix and $x^{1-r}Q'(x) = \Lambda(x)$, reduces the equation $D_B Y = 0$ to

$$x^{1-r}Z' = \tilde{B} + BZ - Z\Lambda.$$

The latter equation is a linear nonhomogeneous equation with $\nu(\mathcal{A}) = \rho(\mathcal{A}) = \tilde{\rho}(A)$, where the linear operator $\mathcal{A}Z = BZ - Z\Lambda$. Then, according to Remark 3.2, there exists a solution $Z(x) = O(x^{-\tilde{k}-\delta})$. That completes the proof of the corollary. Note that $H = 0$ if $\tilde{\rho}(A) < r$. □

4.5. Diagonalization. Let us consider now the problem of block-diagonalization of (4.1) provided that (4.2) holds. Considering (4.9) as an $l \times (n-l)$ vector equation of the form (0.1), we see that the Jacobian matrix $\frac{\partial f}{\partial y}(\infty, 0)$ is invertible, because the matrices $A^{11}(\infty)$ and $A^{22}(\infty)$ have a disjoint spectrum.

If the matrix-valued function $A(x)$ is analytic at $x = \infty$, then, according to Theorem 33.1 in [Wa], the equation (4.9) possesses a solution $P^{12}(x)$ such that $\lim_{x\to+\infty} P^{12}(x) = 0$ and that $P^{12}(x)$ is an analytic function in the interval $(\tilde{x}, \infty)$ for some $\tilde{x} \ge x_0$. That implies the block-diagonalization of (4.1).

Suppose, however, that the analyticity of $A(x)$ is replaced by finite smoothness. Then, according to Remark 0.2, the Jacobian matrix $\frac{\partial f}{\partial y}(\infty, 0)$ has to have eigenvalues with nonzero real part in order to guarantee the existence of a solution to (4.9). Accordingly, the original condition (4.2) of disjointness of the spectrum of $A(\infty)$ in the equation (4.1) should be replaced by the following condition: the eigenvalues λ_j, $j = 1, \dots, n$, of the matrix $A(\infty)$ can be separated into two sets $\sigma_{1,2}$ so that the difference $\lambda_\alpha - \lambda_\beta$ between any $\lambda_\alpha \in \sigma_1$ and any $\lambda_\beta \in \sigma_2$ is not purely imaginary.

Under the latter assumption various results on block-diagonalization of the finitely-smooth equation (4.1) were obtained (see, e.g., [Gi], [Co], and references there). However, we have not come across any result that guarantees such block-diagonalization under the weaker condition of disjointedness (4.2). Theorem 4.1 enables us to partially fill this gap by the following statement.

THEOREM 4.3. *Suppose that in the equation* (4.1) $A(x) \in \mathcal{R}_k$, *where*

$$k \ge \nu - \rho(A^{11}, A^{22}), \tag{4.33}$$

and the matrix $A(\infty)$ *has a disjoint spectrum (i.e.,* $A(\infty)$ *can be represented by* (4.2)*). Then there exists a matrix* $P(x)$, *given by* (4.7), *such that* $P(x) - I_n = o(1)$, $x \to \infty$, *and that the transformation* $Y(x) = P(x)Z(x)$ *block-diagonalizes* (4.1) *(i.e., reduces* (4.1) *to* (4.4)–(4.5)*).*

Proof. A consequence of the condition (4.2) is that we can block-diagonalize $\tilde{A}(x)$, which is the polynomial part of $A(x)$. Then, in addition to (4.12) we get $A^{21} = O(x^{-k-\delta})$, $x \to \infty$. The proof of the theorem coincides with that of Theorem 4.1, except that instead of (4.26) we now have

$$h(x) = S_2 A^{21} S_1 = O(x^{-k-\delta}), \quad x \to \infty.$$

This changes (4.27) to

$$-2k - 2\delta + \nu + (n-2)\varepsilon \ \le -\varepsilon. \tag{4.34}$$

This inequality is a consequence of (4.33).

The estimate (4.33) follows from the requirement $P(x) - I_n = o(1),\ x \to \infty$, where, as in Theorem 4.1, $\gamma = k + \delta - \nu + \rho(A^{11}, A^{22}) - (n-2)\varepsilon$. □

COROLLARY 4.2. *Let $A(x) \in \mathcal{R}_k$ with*

$$k \ge \tilde{\rho}(A) \tag{4.35}$$

and let all eigenvalues of $A(\infty)$ be distinct. Then there exists a matrix $P(x)$ such that $P(x) = o(1),\ x \to \infty$, and that the transformation $Y(x) = (I_n + P(x))Z(x)$ reduces (4.1) to (4.4), where $B(x)$ is diagonal and $B(x) \in \mathcal{R}_k$.

Proof. Let us choose arbitrarily some blocks A_0^1, A_0^2 in (4.2). Let $\tilde{A}(x)$ denote the polynomial part of $A(x)$. As is well known, under the assumptions of the corollary, we can reduce the operator $D_{\tilde{A}}$ to D_Λ, where the matrix $\Lambda(x)$ is diagonal modulo $O(x^{-k-\delta})$. This implies that no shearing is needed to bring $\tilde{A}^{11}$ and $\tilde{A}^{22}$ to their $\tilde{R}$-diagonal forms. Therefore, (4.18) yields $\nu = 2\rho(A^{11}, A^{22})$.

According to (4.29) and (4.35), assumptions of Theorem 4.3 are satisfied, so (4.1) can be decoupled into (4.4)–(4.5) by the transformation (4.3), (4.7). Note that (4.8) implies $B^{11} = A^{11} + A^{12}P^{21},\ \ B^{22} = A^{22} + A^{21}P^{12}$ and that the off-diagonal blocks A^{12}, A^{21} are of the order $O(x^{-k-\delta})$. So $B^{11}, B^{22} \in \mathcal{R}_k$. Each of the matrices $B^{11}(\infty), B^{22}(\infty)$ has distinct eigenvalues, so the block-diagonalization process can be continued. The proof can be completed by induction arguments. □

Remark 4.2. In the case $r = 1$ we have $\tilde{\rho}(A) = 1$. Then Corollary 4.2 is a particular case of the Levinson theorem [CL, section 3.8].

Potentially, the obtained results on triangularization and diagonalization could have various applications in such areas as oscillations, control theory, singular perturbations and others. Within the limits of the paper we would like illustrate this with the following example in oscillation theory.

Example 4.1. Oscillatory properties of solutions of the linear differential equation

$$y^{(n)}(x) = p(x)y(x) \tag{4.36}$$

on the interval $[x_0, \infty)$, where $p \in C^2[x_0, \infty)$ and x_0 is sufficiently large, were considered in [EG1], [EG2] (see also references there). By means of certain linear transformations, this equation was reduced to the system $D_A Y = 0$ with $r = 0$ and

$$\begin{aligned} A(x) = \frac{x p^{\frac{1}{n}}(x)}{s(x)} &\begin{pmatrix} 0 & s(x) & 0 & \cdots & 0 \\ 0 & 1 & s(x) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & n-2 & s(x) \\ 0 & \cdots & \cdots & 0 & n-1 \end{pmatrix} \\ &+ x[s(x) - s] \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ 1 & \cdots & \cdots & 0 & 0 \end{pmatrix} \\ &= B(x) + \tilde{B}(x), \end{aligned}$$

where $B(x), \tilde{B}(x)$ denote the first and second matrix terms, respectively, $s(x) = \frac{-np^{1+1/n}(x)}{p'(x)}$, and $s = \lim_{x\to\infty} s(x)$ provided that the limit exists. One of the most interesting cases when $0 < |s| < \infty$ and when $B(\infty)$ has a multiple eigenvalue λ was considered in [EG2]. In this case it was demonstrated in [EG1] that the multiplicity of λ is exactly two. Analysis of [EG2] shows that under the additional assumption

$$x^n p(x) \text{ is analytic at } x = \infty, \tag{4.37}$$

the equation (4.36) possesses a two-dimensional subspace S of nonoscillatory solutions, i.e., that $y \in S$ and $y \not\equiv 0$ imply existence of some $x_1 \geq x_0$ such that sign $y =$ const. on $[x_1, \infty)$.

Direct computations show that assumption (4.37) implies $s(x) - s = O(x^{-2})$ and thus the equation $D_A Y = 0$ can be considered as a perturbation of $D_B Y = 0$ for large x. The leading term of the last equation can be split into a two-dimensional block B^{11} that corresponds to the eigenvalue λ and into the remaining $(n-2)$-dimensional block B^{22}. The proof of the above mentioned statement is based on the block-diagonalization of D_A determined by B^{11}, B^{22}.

Our point is that, based on Theorem 4.3, the assumption (4.37) in [EG2] can be replaced by the less restrictive assumption $x^{n+1}p'(x) \in \mathcal{R}_1$. Indeed, under this assumption $s(x) - s = O(x^{-2})$ is still valid and $A(x) \in \mathcal{R}_0$. It remains to note that equation $D_A Y = 0$ with regular singularity at infinity $\nu = 0$; thus Theorem 4.3 guarantees block-diagonalization.

REFERENCES

[BJL] W. Balser, W.B. Jurkat, and D.A. Lutz, *A general theory of invariants for meromorphic differential equations*. I, *Formal invariants*, Funkcial. Ekvac., 22 (1979), pp. 197–221.

[Br] B.J.L. Braaksma, *Global reduction of linear differential systems involving a small singular parameter*, SIAM J. Math. Anal., 2 (1971), pp. 149–165.

[CL] E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[Co] W.A. Coppel, *Dichotomies and reducibility*, J. Differential Equations, 3 (1967), pp. 500–521.

[EG1] U. Elias and H. Gingold, *Oscillation of two-term differential equations through asymptotics*, J. Math. Anal. Appl., 186 (1994), pp. 283–305.

[EG2] U. Elias and H. Gingold, *Oscillation and block-diagonalization*, J. Math. Anal. Appl., 199 (1996), pp. 202–212.

[Es] M.S.P. Eastham, *The Asymptotic Solution of Linear Differential Systems. Applications to the Levinson Theorem*, Oxford Science Publications, Oxford, UK, 1989.

[Fr] S. Friedland, *Analytic Similarity of Matrices*, Lectures in Appl. Math. 18, Amer. Math. Soc., Providence, RI, 1980, pp. 43–85.

[GH] H. Gingold and P.-F. Hsieh, *Global analytic triangularization of a matrix function*, Linear Algebra Appl., 169 (1992), pp. 75–101.

[Gi] H. Gingold, *Dichotomies and moving singularities*, Rend. del Circ. Matematico di Palermo Ser. II, 29 (1980), pp. 61–78.

[Hu] M. Hukuhara, *Sur les points singuliers des équations différentielles linéaires*, II, J. Fac. Sci. Hokkaido Imp. Univ., 5 (1937), pp. 123–166.

[KT] S.G. Krein and A. Tovbis, *Linear singular differential equations in finite-dimensional spaces and Banach spaces*, Leningrad Math. J., 2 (1991), pp. 931–985.

[Ku] A.N. Kuznetzov, *Differentiable solutions of degenerated systems of ordinary equations*, Funct. Anal. Appl., 6 (1972), pp. 41–52.

[Lv] W.G. Leavitt, *A normal form for matrices whose elements are holomorphic functions*, Duke Univ. Math. J., 15 (1948), pp. 463–472.

[MK] M.A. Malkov, A.D. Kotelnikov, and C.F. Fennel, *Formation and regular dynamics of shock-trains in Burgers model*, Physica D, 86 (1995), pp. 480–499.

[RS] J.-P. Ramis and Y. Sibuya, *Hukuhara domains and fundamental existence and uniqueness theorems for asymptotic solutions of Gevrey type*, Asymptotic Anal., 2 (1989), pp. 39–94.

[T1] A. Tovbis, *Normal forms of holomorphic matrix-valued functions and corresponding forms for singular differential operators*, Linear Algebra Appl., 162–164 (1992), pp. 389–407.
[T2] A. Tovbis, *Nonlinear ordinary differential equations resolvable with respect to an irregular singular point*, J. Differential Equations, 109 (1994), pp. 201–221.
[Va] V.S. Varadarajan, *Linear meromorphic differential equations: A modern point of view*, Bull. Amer. Math. Soc., 33 (1996), pp. 1–42.
[Wa] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations*, Dover, New York, 1976.
[Wh] G.B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 779–793, May 1998

015

ORTHOGONAL POLYNOMIALS AND CUBATURE FORMULAE ON SPHERES AND ON BALLS*

YUAN XU†

Abstract. Orthogonal polynomials on the unit sphere in $\mathbb{R}^{d+1}$ and on the unit ball in $\mathbb{R}^d$ are shown to be closely related to each other for symmetric weight functions. Furthermore, it is shown that a large class of cubature formulae on the unit sphere can be derived from those on the unit ball and vice versa. The results provide a new approach to study orthogonal polynomials and cubature formulae on spheres.

Key words. orthogonal polynomials in several variables, on spheres, on balls, spherical harmonics, cubature formulae

AMS subject classifications. 33C50, 33C55, 65D32

PII. S0036141096307357

1. Introduction.

We are interested in orthogonal polynomials in several variables with emphasis on those orthogonal with respect to a given measure on the unit sphere S^d in $\mathbb{R}^{d+1}$. In contrast to orthogonal polynomials with respect to measures defined on the unit ball B^d in $\mathbb{R}^d$, there have been relatively few studies on the structure of orthogonal polynomials on S^d beyond the ordinary spherical harmonics which are orthogonal with respect to the surface (Lebesgue) measure (cf. [2, 3, 4, 5, 6, 8]). The classical theory of spherical harmonics is primarily based on the fact that the ordinary harmonics satisfy the Laplace equation. Recently Dunkl (cf. [2, 3, 4, 5] and the references therein) opened a way to study orthogonal polynomials on the spheres with respect to measures invariant under a finite reflection group by developing a theory of spherical harmonics analogous to the classical one. In this important theory the role of Laplacian operator is replaced by a differential-difference operator in the commutative algebra generated by a family of commuting first-order differential-difference operators (Dunkl's operators). Other than these results, however, we are not aware of any other method of studying orthogonal polynomials on spheres.

A closely related question is constructing cubature formulae on spheres and on balls. Cubature formulae with a minimal number of nodes are known to be related to orthogonal polynomials. Over the years, a lot of effort has been put into the study of cubature formulae for measures supported on the unit ball, or on other geometric domains with nonempty interior in $\mathbb{R}^d$. In contrast, the study of cubature formulae on the unit sphere has been more or less focused on the surface measure on the sphere; there is little work on the construction of cubature formulae with respect to other measures. This is partly due to the importance of cubature formulae with respect to the surface measure, which play a role in several fields in mathematics, and perhaps partly due to the lack of study of orthogonal polynomials with respect to a general measure on the sphere.

One main purpose of this paper is to provide an elementary approach towards the study of orthogonal polynomials on S^d for a large class of measures. This approach is

* Received by the editors July 26, 1996; accepted for publication (in revised form) January 9, 1997. This research was supported by the National Science Foundation under grant DMS-9500532.
http://www.siam.org/journals/sima/29-3/30735.html
†Department of Mathematics, University of Oregon, Eugene, OR 97403-1222 (yuan@math.uoregon.edu).

based on a close connection between orthogonal polynomials on S^d and those on the unit ball B^d; a prototype of the connection is the following elementary example.

For $d = 1$, the spherical harmonics of degree n are given in the standard polar coordinates by

$$Y_n^{(1)}(x_1, x_2) = r^n \cos n\theta \quad \text{and} \quad Y_n^{(2)}(x_1, x_2) = r^n \sin n\theta. \tag{1.1}$$

Under the transform $x = \cos\theta$, the polynomials $T_n(x) = \cos n\theta$ and $U_n(x) = \sin n\theta / \sin\theta$ are the Chebyshev polynomials of the first and the second kind, orthogonal with respect to $1/\sqrt{1-x^2}$ and $\sqrt{1-x^2}$, respectively, on the unit ball $[-1, 1]$ in $\mathbb{R}$. Hence, the spherical harmonics on S^1 can be derived from orthogonal polynomials on B^1.

We shall show that for a large class of weight functions on $\mathbb{R}^{d+1}$ we can construct homogeneous orthogonal polynomials on S^d from the corresponding orthogonal polynomials on B^d in a similar way. This allows us to derive properties of orthogonal polynomials on S^d from those on B^d; the latter have been studied much more extensively. Although the approach is elementary and there is no differential or differential-difference operator involved, the result offers a new way to study the structure of orthogonal polynomials on S^d.

Our approach depends on an elementary formula that links the integration on B^d to the integration on S^d. The same formula yields an important connection between cubature formulae on S^d and those on B^d; the result states roughly that a large class of cubature formulae on S^d is generated by cubature formulae on B^d and vice versa. In particular, it allows us to shift our attention from the study of cubature formulae on the unit sphere to the study of cubature formulae on the unit ball; there has been much more understanding towards the structure of the latter one. Although the result is simple and elementary, its importance is apparent. It yields, in particular, many new cubature formulae on spheres and on balls. Because the main focus of this paper is on the relation between orthogonal polynomials and cubature formulae on spheres and those on balls, we will present examples of cubature formulae in a separate paper.

The paper is organized as follows. In section 2 we introduce notation and present the necessary preliminaries, where we also prove the basic lemma. In section 3 we show how to construct orthogonal polynomials on S^d from those on B^d. In section 4 we discuss the relation between cubature formulae on the unit sphere and those on the unit ball.

2. Preliminary and basic lemma. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we let $\mathbf{x} \cdot \mathbf{y}$ denote the usual inner product of $\mathbb{R}^d$ and $|\mathbf{x}| = (\mathbf{x} \cdot \mathbf{x})^{1/2}$ the Euclidean norm of $\mathbf{x}$. Let B^d be the unit ball of $\mathbb{R}^d$ and S^d be the unit sphere on $\mathbb{R}^{d+1}$; that is,

$$B^d = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| \le 1\} \qquad \text{and} \qquad S^d = \{\mathbf{y} \in \mathbb{R}^{d+1} : |\mathbf{y}| = 1\}.$$

Polynomial spaces. Let $\mathbb{N}_0$ be the set of nonnegative integers. For $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ and $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ we write $\mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$. The number $|\alpha|_1 = \alpha_1 + \cdots + \alpha_d$ is called the total degree of $\mathbf{x}^\alpha$. We denote by Π^d the set of polynomials in d variables on $\mathbb{R}^d$ and by Π_n^d the subset of polynomials of total degree at most n. We also denote by $\mathcal{P}_n^d$ the space of homogeneous polynomials of degree n on $\mathbb{R}^d$ and we let $r_n^d = \dim \mathcal{P}_n^d$. It is well known that

$$\dim \Pi_n^d = \binom{n+d}{n} \quad \text{and} \quad r_n^d = \binom{n+d-1}{n}.$$

Orthogonal polynomials on B^d. Let W be a nonnegative weight function on B^d and assume $\int_{B^d} W(\mathbf{x}) d\mathbf{x} < \infty$. It is known that for each $n \in \mathbb{N}_0$ the set of polynomials

of degree n that are orthogonal to all polynomials of lower degree forms a vector space $\mathcal{V}_n$ whose dimension is r_n^d. We denote by $\{P_k^n\}$, $1 \le k \le r_n^d$ and $n \in \mathbb{N}_0$, one family of orthonormal polynomials with respect to W on B^d that forms a basis of Π_n^d, where the superscript n means that $P_k^n \in \Pi_n^d$. The orthonormality means that

$$\int_{B^d} P_k^n(\mathbf{x}) P_j^m(\mathbf{x}) W(\mathbf{x}) d\mathbf{x} = \delta_{j,k}\delta_{m,n}.$$

A useful notation is $\mathbb{P}_n = (P_1^n, \ldots, P_{r_n^d}^n)^T$, which is a vector with P_j^n as components (cf. [22, 24]). For each $n \in \mathbb{N}_0$, the polynomials P_k^n, $1 \le k \le r_n^d$, form an orthonormal basis of $\mathcal{V}_n$. We note that there are many bases of $\mathcal{V}_n$; if Q is an invertible matrix of size r_n^d, then the components of $Q\mathbb{P}_n$ form another basis of $\mathcal{V}_n$ which is orthonormal if Q is an orthogonal matrix. For general results on orthogonal polynomials in several variables, including some of the recent development, we refer to the survey [24] and the references therein. One family of weight functions on B^d whose corresponding orthogonal polynomials have been studied in detail is $(1 - |\mathbf{x}|^2)^{\mu-1/2}$, $\mu \ge 0$, which we will refer to as classical orthogonal polynomials on B^d (cf. [1, 6, 25]).

Ordinary spherical harmonics. The harmonic polynomials on $\mathbb{R}^d$ are the homogeneous polynomials satisfying the Laplace equation $\Delta P = 0$, where

$$\Delta = \partial_1^2 + \cdots + \partial_d^2 \qquad \text{on} \quad \mathbb{R}^d$$

and ∂_i is the ordinary partial derivative with respect to the ith coordinate. They span a subspace $\mathcal{H}_n^d = \ker \Delta \cap \mathcal{P}_n^d$ of dimension $\dim \mathcal{P}_n^d - \dim \mathcal{P}_{n-2}^d$. The spherical harmonics are the restriction of harmonic polynomials on S^{d-1}. If $Y_n \in \mathcal{H}_n^d$, then Y_n is orthogonal to $Q \in \mathcal{P}_k^d$, $0 \le k < n$, with respect to the surface measure $d\omega$ on S^{d-1}.

Dunkl's h-harmonics. For a nonzero vector $\mathbf{v} \in \mathbb{R}^d$ we define the reflection $\sigma_{\mathbf{v}}$ by

$$\mathbf{x}\sigma_{\mathbf{v}} := \mathbf{x} - 2(\mathbf{x} \cdot \mathbf{v})\mathbf{v}/|\mathbf{v}|^2, \qquad \mathbf{x} \in \mathbb{R}^d.$$

Suppose that G is a finite reflection group on $\mathbb{R}^d$ with the set $\{\mathbf{v}_i : i = 1, 2, ..., m\}$ of positive roots; assume that $|\mathbf{v}_i| = |\mathbf{v}_j|$ whenever σ_i is conjugate to σ_j in G, where we write $\sigma_i = \sigma_{\mathbf{v}_i}$, $1 \le i \le m$. Then G is a subgroup of the orthogonal group generated by the reflections $\{\sigma_{\mathbf{v}_i} : 1 \le i \le m\}$.

The h-harmonics yield orthogonal polynomials on S^{d-1} with respect to $h_\alpha^2 d\omega$, where the weight function h_α is defined by

$$h_\alpha(\mathbf{x}) := \prod_{i=1}^m |\mathbf{x} \cdot \mathbf{v}_i|^{\alpha_i}, \qquad \alpha_i \ge 0, \tag{2.1}$$

with $\alpha_i = \alpha_j$ whenever σ_i is conjugate to σ_j in G. The function h_α is a positively homogeneous G-invariant function of degree $|\alpha|_1 = \alpha_1 + \cdots + \alpha_m$. The key ingredient of the theory is a family of commuting first-order differential-difference operators, $\mathcal{D}_i$ (Dunkl's operators), defined by

$$\mathcal{D}_i f(\mathbf{x}) := \partial_i f(\mathbf{x}) + \sum_{j=1}^m \alpha_j \frac{f(\mathbf{x}) - f(\mathbf{x}\sigma_j)}{\mathbf{x} \cdot \mathbf{v}_j} \mathbf{v}_j \cdot \mathbf{e}_i, \qquad 1 \le i \le d, \tag{2.2}$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_d$ are the standard unit vectors of $\mathbb{R}^d$. The h-Laplacian is defined by (see [3])

$$\Delta_h = \mathcal{D}_1^2 + \cdots + \mathcal{D}_d^2, \tag{2.3}$$

which plays the role of Laplacian in the theory of the ordinary harmonics. In particular, the h-harmonics are the homogeneous polynomials satisfying the equation $\Delta_h Y = 0$; in other words, they are the elements of the polynomial subspace $\mathcal{H}_n^d(h^2) := \mathcal{P}_n^d \cap \ker \Delta_h$. The h-spherical harmonics are the restriction of h-harmonics on the sphere.

Basic lemma. We let $d\omega = d\omega_d$ denote the surface measure on S^d, and the surface area

$$\omega = \omega_d = \int_{S^d} d\omega_d = 2\pi^{(d+1)/2}/\Gamma((d+1)/2).$$

The standard change of variables from $\mathbf{x} \in \mathbb{R}^d$ to polar coordinates $r\mathbf{x}'$, $\mathbf{x}' \in S^{d-1}$, yields the following useful formula:

$$\int_{B^d} f(\mathbf{x}) W(\mathbf{x}) d\mathbf{x} = \int_0^1 r^{d-1} \int_{S^{d-1}} f(r\mathbf{x}') W(r\mathbf{x}') d\omega_{d-1} dr. \tag{2.4}$$

This formula connects the integral on B^d to S^{d-1} in a natural way. Our basic formula in the following establishes another relation between integrations over the unit sphere and over the unit ball.

LEMMA 2.1. *Let H be defined on $\mathbb{R}^{d+1}$. Assume that H is symmetric with respect to x_{d+1}; i.e., $H(\mathbf{x}, x_{d+1}) = H(\mathbf{x}, -x_{d+1})$, where $\mathbf{x} \in \mathbb{R}^d$. Then for any continuous function f defined on S^d,*

$$\begin{aligned}\int_{S^d} f(\mathbf{y}) H(\mathbf{y}) d\omega_d = \int_{B^d} \Big[f(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) + f(\mathbf{x}, -\sqrt{1-|\mathbf{x}|^2}) \Big] \\ \times H(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) d\mathbf{x} \Big/ \sqrt{1-|\mathbf{x}|^2}.\end{aligned} \tag{2.5}$$

Proof. For $\mathbf{y} \in S^d$, we write $\mathbf{y} = (\sqrt{1-t^2}\mathbf{x}, t)$, where $\mathbf{x} \in S^{d-1}$ and $-1 \le t \le 1$. Then it follows that (cf. [21, p. 436])

$$d\omega_d = (1-t^2)^{(d-2)/2} dt \, d\omega_{d-1}.$$

Starting from the change of variables $\mathbf{y} = (\sqrt{1-t^2}\mathbf{x}, t)$ in the integral, we get

$$\begin{aligned}\int_{S^d} f(\mathbf{y}) H(\mathbf{y}) d\omega_d &= \int_{-1}^1 \int_{S^{d-1}} f(\sqrt{1-t^2}\mathbf{x}, t) H(\sqrt{1-t^2}\mathbf{x}, t) d\omega_{d-1} (1-t^2)^{(d-2)/2} dt \\ &= \int_0^1 \int_{S^{d-1}} \Big[f(\sqrt{1-t^2}\mathbf{x}, t) + f(\sqrt{1-t^2}\mathbf{x}, -t) \Big] H(\sqrt{1-t^2}\mathbf{x}, t) d\omega_{d-1} (1-t^2)^{(d-2)/2} dt \\ &= \int_0^1 \int_{S^{d-1}} \Big[f(r\mathbf{x}, \sqrt{1-r^2}) + f(r\mathbf{x}, -\sqrt{1-r^2}) \Big] H(r\mathbf{x}, \sqrt{1-r^2}) d\omega_{d-1} r^{d-1} \frac{dr}{\sqrt{1-r^2}} \\ &= \int_{B^d} \Big[f(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) + f(\mathbf{x}, -\sqrt{1-|\mathbf{x}|^2}) \Big] H(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) \frac{d\mathbf{x}}{\sqrt{1-|\mathbf{x}|^2}},\end{aligned}$$

where in the second step we have used the symmetry of H with respect to x_{d+1}, in the third step we have changed the variable $t \mapsto \sqrt{1-r^2}$, and in the last step we have used (2.4). ∎

As a special case of this theorem, we notice that the Lebesgue measure on S^d is related to the Chebyshev weight function $1/\sqrt{1-|\mathbf{x}|^2}$ over B^d.

3. Orthogonal polynomials on spheres. Our main result in this section shows a connection between orthogonal polynomials on B^d and those on S^d, which is the surface of B^{d+1} by definition. To be more precise, we need some notation.

Throughout this section we fix the following notation: for $\mathbf{y} \in \mathbb{R}^{d+1}$, we write

$$\mathbf{y} = (y_1, \ldots, y_d, y_{d+1}) = (\mathbf{y}', y_{d+1}) = r\mathbf{x} = r(\mathbf{x}', x_{d+1}), \quad \mathbf{x} \in S^d, \quad \mathbf{x}' \in B^d, \tag{3.1}$$

where $r = |\mathbf{y}| = \sqrt{y_1^2 + \cdots + y_{d+1}^2}$ and $\mathbf{x}' = (x_1, \ldots, x_d)$.

DEFINITION 3.1. *A weight function H defined on $\mathbb{R}^{d+1}$ is called S-symmetric if it is symmetric with respect to y_{d+1} and centrally symmetric with respect to the variables $\mathbf{y}' = (y_1, \ldots, y_d)$; i.e.,*

$$H(\mathbf{y}', y_{d+1}) = H(\mathbf{y}', -y_{d+1}) \quad \textit{and} \quad H(\mathbf{y}', y_{d+1}) = H(-\mathbf{y}', y_{d+1}). \tag{3.2}$$

For examples of S-symmetric weight functions, we may take $H(\mathbf{y}) = W(\mathbf{y}')h(y_{d+1})$, where W is a centrally symmetric function on $\mathbb{R}^d$ and h is an even function on $\mathbb{R}$. There are many other examples of S-symmetric functions, including

$$H(\mathbf{y}) = \prod_{1 \le i < j \le d+1} |y_i^2 - y_j^2|^{\alpha_{ij}}, \qquad \alpha_{ij} \ge 0,$$

which becomes, when $\alpha_{ij} = \alpha$, an example of reflection invariant weight functions considered by Dunkl (associated with the octahedral group). We note that, however, the weight function $\prod_{i<j} |y_i - y_j|^\alpha$ associated with the symmetric group is not an S-symmetric function, since it is not symmetric with respect to y_{d+1}. Nevertheless, this function is centrally symmetric in $\mathbb{R}^{d+1}$. In fact, it is easy to see that S-symmetry implies central symmetry on $\mathbb{R}^{d+1}$, which we formally state in the following proposition.

PROPOSITION 3.2. *If H is an S-symmetric weight function on $\mathbb{R}^{d+1}$, then it is centrally symmetric on $\mathbb{R}^{d+1}$; that is, $H(\mathbf{y}) = H(-\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^{d+1}$.*

In association with a weight function H on $\mathbb{R}^{d+1}$, we define a weight function W_H on B^d by

$$W_H(\mathbf{x}) = H(\mathbf{x}, \sqrt{1 - |\mathbf{x}|^2}), \qquad \mathbf{x} \in B^d. \tag{3.3}$$

If H is S-symmetric, then the assumption that H is centrally symmetric with respect to the first d variables implies that W is centrally symmetric on B^d. We denote by $\{P_k^n\}$ and $\{Q_k^n\}$ systems of orthonormal polynomials with respect to the weight functions

$$W_H^{(1)}(\mathbf{x}) = 2W_H(\mathbf{x})/\sqrt{1 - |\mathbf{x}|^2} \qquad \text{and} \qquad W_H^{(2)}(\mathbf{x}) = 2W_H(\mathbf{x})\sqrt{1 - |\mathbf{x}|^2}, \tag{3.4}$$

respectively, where we keep the convention that the superscript n means that P_k^n and Q_k^n are polynomials in Π_n^d, and the subindex k has the range $1 \le k \le r_n^d$. Keeping in mind the notation (3.1) we define

$$Y_{k,n}^{(1)}(\mathbf{y}) = r^n P_k^n(\mathbf{x}') \qquad \text{and} \qquad Y_{j,n}^{(2)}(\mathbf{y}) = r^n x_{d+1} Q_j^{n-1}(\mathbf{x}'), \tag{3.5}$$

where $1 \le k \le r_n^d$, $1 \le j \le r_{n-1}^d$, and we define $Y_{j,0}^{(2)}(\mathbf{y}) = 0$. These functions are, in fact, homogeneous polynomials in $\mathbb{R}^{d+1}$.

THEOREM 3.3. *Let H be an S-symmetric weight function defined on $\mathbb{R}^{d+1}$. Assume that W_H in (3.3) is a nonzero weight function on B^d. Then the functions $Y_{k,n}^{(1)}(\mathbf{y})$ and $Y_{k,n}^{(2)}(\mathbf{y})$ defined in (3.5) are homogeneous polynomials of degree n on $\mathbb{R}^{d+1}$ and*

$$\int_{S^d} Y_{k,n}^{(i)}(\mathbf{x}) Y_{l,m}^{(j)}(\mathbf{x}) H(\mathbf{x}) d\omega_d = \delta_{k,l}\delta_{n,m}\delta_{i,j}, \qquad i,j = 1,2.$$

Proof. From the definition of W_H in (3.3), it follows that both $W_H^{(1)}$ and $W_H^{(2)}$ in (3.4) are centrally symmetric weight functions on B^d. As a consequence, the polynomials P_k^n and Q_k^n are even functions if n is even and odd functions if n is odd. In fact, recall the notation $\mathbb{P}_n$ in section 2; it is known (cf. [22]) that there exist proper matrices $D_{n,i}$ and F_n such that

$$\mathbb{P}_{n+1} = \sum_{i=1}^{d} x_i D_{n,i}^T \mathbb{P}_n + F_n \mathbb{P}_{n-1},$$

from which this conclusion follows easily from induction (cf. [23, p. 20]). This allows us to write, for example,

$$P_k^{2n}(\mathbf{x}') = \sum_{j=0}^{n} \sum_{|\alpha|_1 = 2j} a_\alpha (\mathbf{x}')^\alpha, \qquad a_\alpha \in \mathbb{R}, \quad \mathbf{x}' \in B^d,$$

where $\alpha \in \mathbb{N}_0^d$, which implies that

$$Y_{k,2n}^{(1)}(\mathbf{y}) = r^n P_k^{2n}(\mathbf{x}') = \sum_{k=0}^{n} r^{2n-2k} \sum_{|\alpha|_1 = 2k} a_\alpha (\mathbf{y}')^\alpha.$$

Since $r^2 = y_1^2 + \cdots + y_{d+1}^2$ and $\mathbf{y}' = (y_1, \ldots, y_d)$, this shows that $Y_{k,2n}^{(1)}(\mathbf{y})$ is a homogeneous polynomial of degree $2n$ in $\mathbf{y}$. Similar proof can be adopted to show that $Y_{k,2n-1}^{(1)}$ is homogeneous of degree $2n-1$ and, using the fact $r x_{d+1} = y_{d+1}$, that $Y_{k,n}^{(2)}$ are homogeneous of degree n.

Since $Y_{k,n}^{(1)}$, when restricted to S^d, is independent of x_{d+1} and $Y_{k,2n}^{(1)}$ contains a single factor x_{d+1}, it follows that

$$\int_{S^d} Y_{k,n}^{(1)}(\mathbf{x}) Y_{l,m}^{(2)}(\mathbf{x}) H(\mathbf{x}) d\omega_d = \int_{S^d} x_{d+1} P_k^n(\mathbf{x}') Q_l^m(\mathbf{x}') H(\mathbf{x}) d\omega_d = 0$$

for any (k,n) and (l,m). By the basic formula (2.6),

$$\begin{aligned}\int_{S^d} Y_{k,n}^{(1)}(\mathbf{x}) Y_{l,m}^{(1)}(\mathbf{x}) H(\mathbf{x}) d\omega_d &= 2\int_{B^d} P_k^n(\mathbf{x}') P_l^m(\mathbf{x}') H(\mathbf{x}', \sqrt{1-|\mathbf{x}'|^2}) \frac{d\mathbf{x}'}{\sqrt{1-|\mathbf{x}'|^2}} \\ &= \int_{B^d} P_k^n(\mathbf{x}') P_l^m(\mathbf{x}') W_H^{(1)}(\mathbf{x}') d\mathbf{x}' = \delta_{k,l}\delta_{n,m}\end{aligned}$$

and similarly, using the fact that $x_{d+1}^2 = 1 - |\mathbf{x}'|^2$,

$$\begin{aligned}\int_{S^d} Y_{k,n}^{(2)}(\mathbf{x}) Y_{l,m}^{(2)}(\mathbf{x}) H(\mathbf{x}) d\omega_d &= 2\int_{B^d} (1-|\mathbf{x}'|^2) Q_k^{n-1}(\mathbf{x}') Q_l^{m-1}(\mathbf{x}') H(\mathbf{x}', \sqrt{1-|\mathbf{x}'|^2}) \\ &\quad \times \frac{d\mathbf{x}'}{\sqrt{1-|\mathbf{x}'|^2}} = \int_{B^d} Q_k^{n-1}(\mathbf{x}') Q_l^{m-1}(\mathbf{x}') W_H^{(2)}(\mathbf{x}') d\mathbf{x}' = \delta_{k,l}\delta_{n,m}.\end{aligned}$$

This completes the proof. □

The assumption that H is S-symmetry in Theorem 3.3 is necessary; it is used to show that $Y_{k,n}^{(1)}$ and $Y_{k,n}^{(2)}$ in (3.5) are indeed polynomials in $\mathbf{y}$.

Example 3.4. If $H(\mathbf{y}) = 1$, then $Y_{k,n}^{(1)}$ and $Y_{k,n}^{(2)}$ are orthonormal with respect to the surface measure $d\omega$; they are the ordinary spherical harmonics. According to Theorem 3.3, the harmonics are related to the orthogonal polynomials with respect to the radial weight functions $W_0(\mathbf{x}) = 1/\sqrt{1-|\mathbf{x}|^2}$ and $W_1(\mathbf{x}) = \sqrt{1-|\mathbf{x}|^2}$ on B^d, both of which belong to the family of weight functions $W_\mu(\mathbf{x}) = W_{\mu,d}(\mathbf{x}) = w_\mu(1-|\mathbf{x}|^2)^{\mu-\frac{1}{2}}$, $\mu > -1/2$, whose corresponding orthogonal polynomials have been studied in [1, 6, 25]. For $d = 1$, the spherical harmonics are given in the polar coordinates $(y_1, y_2) = r(x_1, x_2) = r(\cos\theta, \sin\theta)$ by the formula (1.1), which can be written as

$$Y_n^{(1)}(y_1, y_2) = r^n T_n(x_1) \quad \text{and} \quad Y_n^{(2)}(y_1, y_2) = r^n x_2 U_{n-1}(x_1),$$

where, with $t = \cos\theta$, $T_n(t) = \cos n\theta$ and $U_n(t) = \sin n\theta / \sin\theta$ are the Chebyshev polynomials of the first and the second kind, which are orthogonal with respect to $1/\sqrt{1-x^2}$ and $\sqrt{1-x^2}$, respectively. It is this example that motivates our present consideration. □

DEFINITION 3.5. *We define a subspace* $\mathcal{H}_n^{d+1}(H)$ *of* $\mathcal{P}_n^{d+1}$ *by*

$$\mathcal{H}_n^{d+1}(H) = \operatorname{span}\{Y_{k,n}^{(1)}, \quad 1 \le k \le r_n^d; \quad \textit{and} \quad Y_{j,n}^{(2)}, \quad 1 \le j \le r_{n-1}^d\}.$$

THEOREM 3.6. *Let* H *be an* S*-symmetric function on* $\mathbb{R}^{d+1}$*. For each* $n \in \mathbb{N}_0$*,*

$$\dim \mathcal{H}_n^{d+1}(H) = \binom{n+d}{d} - \binom{n+d-2}{d} = \dim \mathcal{P}_n^{d+1} - \dim \mathcal{P}_{n-2}^{d+1}.$$

Proof. From the orthogonality in Theorem 3.3, the polynomials in $\{Y_{k,n}^{(1)}, Y_{j,n}^{(2)}\}$ are linearly independent. Hence, it follows readily that

$$\dim \mathcal{H}_n^{d+1}(H) = r_n^d + r_{n-1}^d = \binom{n+d-1}{n} + \binom{n+d-2}{n-1},$$

where we use the convention that $\binom{k}{j} = 0$ if $j < 0$. Using the identity $\binom{n+m}{n} - \binom{n+m-1}{n} = \binom{n+m-1}{n-1}$, it is easy to verify that

$$\dim \mathcal{H}_n^{d+1}(H) = \binom{n+d}{d} - \binom{n+d-2}{d},$$

which is the desired result. □

THEOREM 3.7. *Let* H *be an* S*-symmetric function on* $\mathbb{R}^{d+1}$*. For* $n \in \mathbb{N}_0$*,*

$$\mathcal{P}_n^{d+1} = \bigoplus_{k=0}^{[n/2]} |\mathbf{y}|^{2k} \mathcal{H}_{n-2k}^{d+1}(H);$$

that is, if $P \in \mathcal{P}_n^{d+1}$*, then there is a unique decomposition*

$$P(\mathbf{y}) = \sum_{k=0}^{[n/2]} |\mathbf{y}|^{2k} P_{n-2k}(\mathbf{y}), \qquad P_{n-2k} \in \mathcal{H}_{n-2k}^{d+1}(H).$$

Proof. Since P is homogeneous of degree n, we can write $P(\mathbf{y}) = r^n P(\mathbf{x})$, where we use the notation in (3.1) again. According to the power of y_{d+1} being even or odd and using $x_{d+1}^2 = 1 - |\mathbf{x}'|^2$ whenever possible, we can further write

$$P(\mathbf{y}) = r^n P(\mathbf{x}) = r^n [p(\mathbf{x}') + x_{d+1} q(\mathbf{x}')],$$

where p and q are polynomials of degree at most n and $n-1$, respectively, in $\mathbf{x}' \in B^d$. Moreover, if n is even, then p is even and q is odd; if n is odd, then p is odd and q is even. Since both $\{P_k^n\}$ and $\{Q_k^n\}$ form a basis for Π_n^d and since the weight functions $W_H^{(1)}$ and $W_H^{(2)}$ in (3.4) are centrally symmetric, we have the unique expansions

$$p(\mathbf{x}') = \sum_{k=0}^{[n/2]} \sum_j a_{j,k} P_j^{n-2k}(\mathbf{x}') \quad \text{and} \quad q(\mathbf{x}') = \sum_{k=0}^{[(n-1)/2]} \sum_j b_{j,k} Q_j^{n-2k-1}(\mathbf{x}'),$$

where $1 \le j \le r_{n-2k}^d$. Therefore, by the definition of $Y_{k,n}^{(1)}$ and $Y_{k,n}^{(2)}$, we have

$$P(\mathbf{y}) = \sum_{k=0}^{[n/2]} r^{2k} \sum_j a_{j,k} Y_{j,n-2k}^{(1)}(\mathbf{y}) + \sum_{k=0}^{[(n-1)/2]} r^{2k-1} \sum_j b_{j,k} Y_{j,n-2k+1}^{(2)}(\mathbf{y}),$$

which is the desired decomposition. The uniqueness follows from the orthogonality in Theorem 3.3. □

For the spherical harmonics or h-harmonics, the above theorem is usually established using the differential or differential-difference operator (cf. [19, 2]). The importance of the results in this section lies in the fact that they provide an approach to studying orthogonal polynomials on S^d with respect to a large class of measures. For example, one of the essential ingredients in the recent work of orthogonal polynomials in several variables (cf. [22, 24]) is a three-term relation in a vector-matrix form,

$$x_i \mathbb{P}_n = A_{n,i} \mathbb{P}_{n+1} + B_{n,i} \mathbb{P}_n + A_{n-1,i}^T \mathbb{P}_{n-1},$$

where $A_{n,i}$ and $B_{n,i}$ are proper matrices, which also plays a decisive role in the study of common zeros of $\mathbb{P}_n$ and cubature formulae; the results in Theorem 3.3 show that the h-spherical harmonic polynomials that are even (or odd) in x_{d+1} also satisfy such a three-term relation.

It is worthwhile to point out that the relation between orthogonal polynomials on B^d and those on S^d goes both ways. In fact, the following result holds.

THEOREM 3.8. *Let H be a weight function defined on $\mathbb{R}^{d+1}$ which is symmetric with respect to y_{d+1}. Assume that W_H in (3.3) is a nonzero weight function on B^d. Let $Y_{k,n}^{(1)}$ be the orthonormal polynomials of degree n with respect to $H(\mathbf{y})d\omega$ on S^d that are even in y_{d+1}, and write the orthonormal polynomials that are odd in y_{d+1} as $y_{d+1} Y_{k,n-1}^{(2)}$. Then*

$$P_k^n(\mathbf{x}) = Y_{k,n}^{(1)}(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) \quad \text{and} \quad Q_k^n(\mathbf{x}) = Y_{k,n}^{(2)}(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2})$$

are orthonormal polynomials of degree n in $\mathbf{x} \in B^d$ with respect to $W_H^{(1)}$ and $W_H^{(2)}$ defined in (3.4), respectively.

Proof. The orthogonality follows easily from Lemma 2.1 as in the proof of Theorem 3.3. We show that the assumption on $Y_{k,n}^{(i)}$ is justified. Since H is symmetric

with respect to y_{d+1}, we can pick the orthogonal polynomials with respect to $Hd\omega$ on S^d as either even in y_{d+1} or odd in y_{d+1} (recall the nonuniqueness of orthonormal bases). Indeed, if Y_n is a polynomial of degree n orthogonal to lower degree polynomials with respect to $Hd\omega$, so is the polynomial $Y_n(\mathbf{y}', -y_{d+1})$ by the symmetry of H with respect to y_{d+1}. Hence, if n is even, then the polynomial $Y_n(\mathbf{y}) + Y_n(\mathbf{y}', -y_{d+1})$ is an orthogonal polynomial of degree n which is even in y_{d+1}; if n is odd, then we consider $Y_n(\mathbf{y}) - Y_n(\mathbf{y}', -y_{d+1})$ instead. Therefore, the polynomials P_k^n and Q_k^n are well defined on B^d. □

It should be noted that there is no need to assume that H is S-symmetric in the above theorem; consequently, there is no assurance that $Y_{k,n}^{(i)}$ are homogeneous.

In an effort to understand Dunkl's theory of h-harmonics, we study the orthogonal polynomials on S^d associated to $h(\mathbf{y}) = |y_1|^{\alpha_1} \cdots |y_{d+1}|^{\alpha_{d+1}}$ in detail in [26]. In particular, making use of the product structure of the measure, an orthonormal basis of h-harmonics is given in terms of the orthonormal polynomials of one variable with respect to the measure $(1-t^2)^\lambda |t|^{2\mu}$ on $[-1,1]$ (which in turn can be written in terms of Jacobi polynomials). By Theorem 3.8, we can then derive an explicit basis of orthogonal polynomials with respect to $W_H(\mathbf{x}) = |x_1|^{\alpha_1} \cdots |x_d|^{\alpha_d} (1 - |\mathbf{x}|^2)^{\alpha_{d+1}}$.

The theory of the h-harmonics developed by Dunkl recently is a rich one; it has found applications in a number of fields. For numerical work, one essential problem in dealing with h-harmonics is the construction of a workable orthonormal basis for $\mathcal{H}_n^{d+1}(h^2)$. So far, such a basis has been constructed only in the case of $h(\mathbf{y}) = |y_1|^{\alpha_1} \cdots |y_{d+1}|^{\alpha_{d+1}}$, associated to the reflection group $Z_2 \times \cdots \times Z_2$. Theorem 3.3 indicates that an explicit construction of such a basis may be difficult for the reflection invariant weight functions h associated with most of other reflection groups. We illustrate by the following example.

Example 3.9. Consider the weight function h on $\mathbb{R}^3$ defined by

$$h(y_1, y_2, y_3) = |(y_1^2 - y_2^2)(y_1^2 - y_3^2)(y_2^2 - y_3^2)|^\mu,$$

which is associated to the octahedral group; the group is generated by the reflections in $y_i = 0$, $1 \le i \le 3$, and $y_i \pm y_j = 0$, $1 \le i, j \le 3$; it is the Weyl group of type B_3. This weight function is one of the simplest nonproduct weight functions on S^2. According to Theorem 3.1, the h-harmonics associated to the function $H(\mathbf{y}) = h^2(\mathbf{y})$ are related to the orthogonal polynomials on the disc $B^2 \subset \mathbb{R}^2$ with respect to the weight function $W_H^{(1)}$ and $W_H^{(2)}$ in (3.4), where the weight function $W_H^{(1)}$ is given by

$$W_H^{(1)}(x_1, x_2) = 2|(x_1^2 - x_2^2)(1 - 2x_1^2 - x_2^2)(1 - x_1^2 - 2x_2^2)|^{2\mu} / \sqrt{1 - x_1^2 - x_2^2}, \quad (x_1, x_2) \in B^2.$$

An explicit basis for the h-harmonics will mean an explicit basis for orthogonal polynomials with respect to $W_H^{(1)}$ and vice versa. However, the form of $W_H^{(1)}$ given above indicates that it may be difficult to find a closed formula for such a basis. □

4. Cubature formula on spheres and on balls. In this section we discuss the connection between cubature formulae on spheres and on balls. For a given integral $\mathcal{L}(f) := \int f d\mu$, where $d\mu$ is a nonnegative measure with support set on B^d, a cubature formula of degree M is a linear functional

$$\mathcal{I}_M(f) = \sum_{k=1}^N \lambda_k f(\mathbf{x}_k) \ , \quad \lambda_k > 0, \quad \mathbf{x}_k \in \mathbb{R}^d,$$

defined on Π^d, such that $\mathcal{L}(f) = \mathcal{I}_M(f)$ whenever $f \in \Pi^d_M$, and $\mathcal{L}(f^*) \neq \mathcal{I}_M(f^*)$ for at least one $f^* \in \Pi^d_{M+1}$. When the measure is supported on S^d, we need to replace Π^d_M by $\bigcup_{k=0}^M \mathcal{P}_k^{d+1}$ in the above formulation and require $\mathbf{x}_k \in S^d$. The points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are called *nodes* and the numbers $\lambda_1, \ldots, \lambda_N$ are called *weights*. Such a formula is called minimal if N, the number of nodes, is minimal among all cubature formulae of degree M.

Cubature formulae on the unit sphere have important applications in numerical integration and in areas ranging from coding theory to isometric embeddings between classical Banach spaces (cf. [11, 15, 16] and the references therein). Over years, construction of cubature formulae on the unit sphere with respect to the surface measure $d\omega$ has attracted a lot of attention. For example, starting from the pioneer work of Sobolev (cf. [17]), the Russian school of mathematicians have constructed various cubature formulae on S^d that are invariant under finite groups (cf. [14, 10] and the references therein). There are also important studies on Chebyshev cubature formulae, which are formulae with equal weights (cf. [9, 11, 15] and the references therein). Nevertheless, the simple results we present below on the connection between cubature formula on balls and on spheres do not seem to have been noticed before.

THEOREM 4.1. *Let H defined on $\mathbb{R}^{d+1}$ be symmetric with respect to y_{d+1}. Suppose that there is a cubature formula of degree M on B^d for W_H defined in* (3.3)*,*

$$\int_{B^d} g(\mathbf{x}) W_H(\mathbf{x}) \frac{d\mathbf{x}}{\sqrt{1-|\mathbf{x}|^2}} = \sum_{i=1}^N \lambda_i g(\mathbf{x}_i), \qquad g \in \Pi^d_M, \tag{4.1}$$

whose N nodes lie inside the unit ball B^d; that is, $|\mathbf{x}_i| \le 1$. Then there is a cubature formula of degree M on the unit sphere S^d,

$$\begin{aligned} &\int_{S^d} f(\mathbf{y}) H(\mathbf{y}) d\omega \\ &\qquad = \sum_{i=1}^N \lambda_i \Big[f(\mathbf{x}_i, \sqrt{1-|\mathbf{x}_i|^2}) + f(\mathbf{x}_i, -\sqrt{1-|\mathbf{x}_i|^2}) \Big], \quad f \in \bigcup_{k=0}^M \mathcal{P}_k^{d+1}. \end{aligned} \tag{4.2}$$

Proof. Assuming (4.1), to prove (4.2) it suffices to prove, by Lemma 2.1, that

$$\begin{aligned} &\int_{B^d} \Big[f(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) + f(\mathbf{x}, -\sqrt{1-|\mathbf{x}|^2}) \Big] W_H(\mathbf{x}) \frac{d\mathbf{x}}{\sqrt{1-|\mathbf{x}|^2}} \\ &\qquad = \sum_{i=1}^N \lambda_i \Big[f(\mathbf{x}_i, \sqrt{1-|\mathbf{x}_i|^2}) + f(\mathbf{x}_i, -\sqrt{1-|\mathbf{x}_i|^2}) \Big] \end{aligned} \tag{4.3}$$

for all polynomials $f \in \Pi^d_M$. We consider the basis of $\bigcup_{k=0}^M \mathcal{P}_k^{d+1}$ consisting of monomial $\{f_\alpha\}_{|\alpha|_1 \le M}$, where $f_\alpha(\mathbf{y}) = \mathbf{y}^\alpha$ and $\alpha \in \mathbb{N}^{d+1}$. If f_α is an odd function in y_{d+1}, then both the left side and the right side of (4.3) are zero, so the equality holds. If f_α is even in y_{d+1}, $|\alpha|_1 \le M$, then the function

$$f_\alpha(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) = \mathbf{x}^{\alpha'} (1-|\mathbf{x}|^2)^{\alpha_{d+1}/2},$$

where we write $\alpha = (\alpha', \alpha_{d+1})$, is a polynomial of degree at most M in $\mathbf{x}$. Hence, it follows from the cubature formula (4.1) that

$$\int_{B^d} f(\mathbf{x}, \pm\sqrt{1-|\mathbf{x}|^2}) W_H(\mathbf{x}) \frac{d\mathbf{x}}{\sqrt{1-|\mathbf{x}|^2}} = \sum_{i=1}^N \lambda_i f(\mathbf{x}_i, \pm\sqrt{1-|\mathbf{x}_i|^2})$$

holds. Adding the above equations for $f(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2})$ and for $f(\mathbf{x}, -\sqrt{1-|\mathbf{x}|^2})$ together proves (4.3). ∎

The theorem states that each cubature formula on the unit ball B^d leads to a cubature formula on the unit sphere S^d. The converse of this result is also true.

THEOREM 4.2. *Let H be a weight function on $\mathbb{R}^{d+1}$ which is symmetric with respect to x_{d+1}. Suppose that there is a cubature formula of degree M on the sphere S^d*

$$\int_{S^d} f(\mathbf{y}) H(\mathbf{y}) d\omega = \sum_{i=1}^N \lambda_i f(\mathbf{y}_i), \qquad f \in \bigcup_{k=0}^M \mathcal{P}_k^{d+1} \tag{4.4}$$

whose nodes are all located on S^d. Then there is a cubature formula of degree M on the unit ball B^d

$$2 \int_{B^d} g(\mathbf{x}) W_H(\mathbf{x}) \frac{d\mathbf{x}}{\sqrt{1-|\mathbf{x}|^2}} = \sum_{i=1}^N \lambda_i g(\mathbf{x}_i), \qquad g \in \Pi_M^d, \tag{4.5}$$

where $\mathbf{x}_i \in B^d$ are the first d components of $\mathbf{y}_i$; that is, $\mathbf{y}_i = (\mathbf{x}_i, x_{d+1,i})$.

Proof. By Lemma 2.1, the cubature formula (4.4) is equivalent to

$$\int_{B^d} \Big[f(\mathbf{x}, \sqrt{1-|\mathbf{x}|^2}) + f(\mathbf{x}, -\sqrt{1-|\mathbf{x}|^2}) \Big] W_H(\mathbf{x}) \frac{d\mathbf{x}}{\sqrt{1-|\mathbf{x}|^2}} = \sum_{i=1}^N \lambda_i f(\mathbf{y}_i). \tag{4.6}$$

If we write $\mathbf{y} = (\mathbf{x}, x_{d+1}) \in \mathbb{R}^{d+1}$, where $\mathbf{x} \in \mathbb{R}^d$, then for every monomial $g_\alpha(\mathbf{x}) = \mathbf{x}^\alpha \in \Pi_M^d$ the function f_α defined by $f_\alpha(\mathbf{y}) = g_\alpha(\mathbf{x})$ is a polynomial in $\mathcal{P}_k^{d+1}$, where $|\alpha|_1 = k \le M$. We can apply cubature formula (4.4) to it. Since f so defined is apparently even in x_{d+1}, the cubature (4.6) becomes cubature formula (4.5). ∎

Although these theorems are simple to state, they have important implications. They allow us to fit a large class of cubature formula on spheres into the structure of cubature formulae on balls, which suggests an alternative approach to study and construct cubature formulae.

Example 4.3. In the case $d = 1$, the formula (4.1) under the change of variable $x = \cos\theta$ becomes

$$\int_0^\pi g(\cos\theta) W_H(\cos\theta) d\theta = \sum_{i=1}^N \lambda_i g(\cos\theta_i).$$

On the other hand, we can write the integral over S^1 in the polar coordinates as

$$\int_{S^1} f(\mathbf{y}) H(\mathbf{y}) d\omega = \int_0^{2\pi} f(\cos\theta, \sin\theta) H(\cos\theta, \sin\theta) d\theta.$$

Since H is symmetric with respect to x_2, it follows that $W_H(\cos\theta) = H(\cos\theta, \sin\theta)$ in the notation of (3.3). Hence, (4.2) becomes

$$\int_0^{2\pi} f(\cos\theta, \sin\theta) W_H(\cos\theta) d\theta = \sum_{i=1}^N \lambda_i \Big[f(\cos\theta_i, \sin\theta_i) + f(\cos\theta_i, -\sin\theta_i) \Big].$$

From these formulae the relation between (4.1) and (4.2) is evident. ∎

In a separate paper we will present a number of examples on S^2 that are obtained using this approach. Here we concentrate on the theoretic side of the matter. What we are interested in is the minimal cubature formula, or cubature formula whose number of nodes is close to minimal.

We state the lower bounds on the number of nodes of cubature formulae, which are used to test whether a given cubature is minimal. Let us denote by N_{B^d} the number of nodes for a cubature formula on B^d, and by N_{S^d} the number of nodes for a cubature formula on S^d. It is well known (cf. [7, 18]) that

$$N_{B^d} \ge \dim \Pi_n^d = \binom{n+d}{n}, \qquad M = 2n \ \text{ or } \ M = 2n+1, \tag{4.7}$$

and

$$N_{S^d} \ge \sum_{k=0}^{n} \dim \mathcal{H}_k^{d+1} = \binom{n+d}{n} + \binom{n+d-1}{n-1}, \quad M = 2n \ \text{ or } \ M = 2n+1, \tag{4.8}$$

where the equal sign in (4.8) follows from the formula for $\dim \mathcal{H}_k^{d+1}$ (cf. Theorem 3.6 with $H = 1$) and simple computation. Moreover, for centrally symmetric weight functions there are improved lower bounds for cubature formula of odd degree, due to Möller for N_{B^d} and to Mysovskikh for N_{S^d} (cf. [13, 14]), which states that

$$N_{B^2} \ge \binom{n+2}{n} + \left[\frac{n+1}{2}\right], \qquad M = 2n+1, \tag{4.9}$$

and

$$N_{S^d} \ge 2\binom{n+d}{n}, \qquad M = 2n+1, \tag{4.10}$$

where, for simplicity, we have restricted the lower bound of N_{B^d} to the case $d = 2$.

Several characterizations of cubature formulas on B^d that attain the lower bound in (4.7), or (4.9), are known. For example, there is a cubature formula that attains the bound (4.7) for $M = 2n+1$ if, and only if, the corresponding orthogonal polynomials $P_1^{n+1}, \ldots, P_{r_{n+1}^d}^{n+1}$ of degree $n+1$ have $\dim \Pi_n^d$ many distinct real common zeros. The characterization for the case (4.7) with $M = 2n$ and the case (4.9) for centrally symmetric weight functions will involve common zeros of quasi-orthogonal polynomials. For these characterizations and extensions of them we refer to [12, 14, 23, 24] and the references therein. In view of the results in section 3, we see that when H is centrally symmetric, we can relate these characterizations to orthogonal polynomials on spheres.

Let us consider the number of nodes of the cubature formulae in (4.2) and (4.5).

Remark 4.4. In Theorem 4.1, the number of nodes in the cubature formula (4.2) may be less than $2N$, since if one of the nodes of (4.1), say $\mathbf{x}_i$, lies on the boundary $\partial B^d = S^{d-1}$, then $|\mathbf{x}_i| = 1$ and two nodes $(\mathbf{x}_i, \sqrt{1-|\mathbf{x}_i|^2})$ and $(\mathbf{x}_i, -\sqrt{1-|\mathbf{x}_i|^2})$ in (4.2) become one. That is,

$$\text{number of nodes of (4.2)} = 2N - \text{number of } \mathbf{x}_i \text{ on } S^{d-1}. \tag{4.11}$$

Similarly, in Theorem 4.2, the number of nodes in the cubature formula (4.5) may be less than N, since different $\mathbf{y}_i \in S^d$ may have the same first d components, which

happens when $\mathbf{y}_i$ and $\mathbf{y}_j$ form a *symmetric pair* with respect to the last component; i.e., $\mathbf{y}_i = (\mathbf{x}_i, x_{d+1})$ and $\mathbf{y}_j = (\mathbf{x}_i, -x_{d+1})$ with $x_{d+1} \neq 0$. We conclude that

(4.12) number of nodes of (4.5) = N − number of symmetric pairs among $\mathbf{y}_i$.

Clearly, the number of nodes in (4.5) satisfies a lower bound $N/2$, which is attained when the nodes of (4.4) consist of only symmetric pairs. □

It is important to remark that even if the cubature formula (4.1) on B^d in Theorem 4.1 attains the lower bound (4.7) or (4.9), the cubature formula (4.2) on S^d may not attain the lower bound (4.8) or (4.10), respectively. For example, when $d = 1$, the formula (4.1) for $M = 2n+1$ attains the lower bound (4.7) with $N_{B^1} = n+1$, which is the classical Gaussian quadrature formula. On the other hand, the corresponding formula in (4.2) attains the lower bound (4.8) with $N_{S^1} = 2n+1$ only when $x = 1$ or $x = -1$ is a node of (4.1), which does not hold in general since the nodes of a Gaussian quadrature formula on $[-1, 1]$ are zeros of orthogonal polynomials and are located in $(-1, 1)$.

As an immediate consequence of these lower bounds and Theorems 4.1 and 4.2, we formulate a corollary that seems to be of independent interest.

COROLLARY 4.5. *Let H be an S-symmetric weight function on $\mathbb{R}^3$. If there is a cubature formula of degree $2n+1$ with respect to H on S^2 that attains the lower bound in* (4.10)*, then it contains at least $2[(n+1)/2]$ nodes which are not symmetric with respect to x_3.*

Proof. Assume that a cubature formula with respect to H on S^2 exists which attains the lower bound in (4.10). Let E be the number of symmetric pairs among the nodes of the cubature. By Theorem 4.2 and (4.12), there is a cubature formula on B^2 with $N_{S^2} - E = 2\binom{n+2}{n} - E$ many nodes. Moreover, the weight function associated with the new cubature formula is centrally symmetric on B^2. Hence, by (4.9), we have the inequality that

$$N_{S^2} - E = 2\binom{n+2}{n} - E \geq \binom{n+2}{n} + \Big[\frac{n+1}{2}\Big],$$

from which we get an upper bound for E. Evidently, the number of nodes that do not contain symmetric pairs is equal to $N_{S^2} - 2E$. Hence, the upper bound for E leads to a lower bound on the number of nodes that are not symmetric with respect to x_3, which gives the desired result. □

In particular, if the cubature formula on S^2 is symmetric with respect to x_3, which means that the node of the cubature always contains the pair (x_1, x_2, x_3) and $(x_1, x_2, -x_3)$ whenever $x_3 > 0$, then there are at least $2[(n+1)/2]$ many nodes on the largest circle $x_1^2 + x_2^2 = 1$ which is perpendicular to x_3 axis. Results as such provide necessary conditions on the minimal cubature formulae; they may provide insight in the construction of the minimal formula or can be used to prove that such a formula does not exist. An analogue of Corollary 4.5 is as follows.

COROLLARY 4.6. *If there is a cubature formula on B^2 that attains the lower bound in* (4.9) *with all nodes in B^2, then it can have no more than $2[(n+1)/2]$ points on the boundary $\partial B^2 = S^1$.*

Proof. If a cubature formula on B^2 as stated exists which attains the lower bound (4.9), then by Theorem 4.1 and (4.11) there is a cubature formula on S^2 with

$$N_{S^2} = 2\binom{n+2}{n} + 2\Big[\frac{n+1}{2}\Big] - \text{number of nodes on } S^1.$$

The desired result then follows from the lower bound (4.10). ∎

We conclude this paper with another simple application dealing with Chebyshev cubature formulae, which are cubature formulae with equal weights. It is proved in [9] that the number of nodes of a Chebyshev cubature formula of degree M with respect to $1/\sqrt{1-|\mathbf{x}|^2}$on B^2 is of order $\mathcal{O}(M^3)$. Furthermore, it is conjectured there that the number of nodes of a Chebyshev cubature formula of degree M with respect to the surface measure on S^2 is of order $\mathcal{O}(M^2)$.

COROLLARY 4.7. *If there is a Chebyshev cubature formula of degree M with respect to the surface measure on S^2 whose number of nodes is of order $\mathcal{O}(M^2)$, then its nodes cannot all be symmetric with respect to a plane that contains one largest circle of S^2 and none of the nodes.*

Indeed, if such a cubature formula exists, we may assume that the plane is the coordinate plane perpendicular to the x_3 axis since the integral is invariant under rotation. Then there are an even number of nodes and all nodes form symmetric pairs. Therefore, by Theorem 4.2, there would be a Chebyshev cubature formula of degree M with respect to $1/\sqrt{1-|\mathbf{x}|^2}$ on B^2 with the number of nodes in the order of $\mathcal{O}(M^2)$, which leads to a contradiction.

Acknowledgment. The author thanks a referee for his careful review and helpful suggestions.

REFERENCES

[1] P. APPELL AND J. K. DE FÉRIET, *Fonctions hypergéométriques et hypersphériques, Polynomes d'Hermite*, Gauthier-Villars, Paris, 1926.
[2] C. DUNKL, *Reflection groups and orthogonal polynomials on the sphere*, Math. Z., 197 (1988), pp. 33–60.
[3] C. DUNKL, *Differential-difference operators associated to reflection groups*, Trans. Amer. Math. Soc., 311 (1989), pp. 167–183.
[4] C. DUNKL, *Integral kernels with reflection group invariance*, Canad. J. Math., 43 (1991), pp. 1213–1227.
[5] C. DUNKL, *Intertwining operators associated to the group S_3*, Trans. Amer. Math. Soc., 347 (1995), pp. 3347–3374.
[6] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, McGraw–Hill, New York, 1953.
[7] H. ENGLES, *Numerical quadrature and cubature*, Academic Press, New York, 1980.
[8] E.G. KALNINS, W. MILLER, JR., AND M. V. TRATNIK, *Families of orthogonal and biorthogonal polynomials on the n-sphere*, SIAM J. Math. Anal., 22 (1991), pp. 272–294.
[9] J. KOREVAAR AND J. L. H. MEYERS, *Chebyshev-type quadrature on multidimensional domains*, J. Approx. Theory, 79 (1994), pp. 144–164.
[10] V.I. LEBEDEV, *A quadrature formula for the sphere of 56th algebraic order of accuracy*, Russian Acad. Sci. Dokl. Math., 50 (1995), pp. 283–286.
[11] YU. LYUBICH AND L.N. VASERSTEIN, *Isometric embeddings between classical Banach spaces, cubature formulas, and spherical designs*, Geom. Dedicata, 47 (1993), pp. 327–362.
[12] H. M. MÖLLER, *Kubaturformeln mit minimaler Knotenzahl*, Numer. Math., 35 (1976), pp. 185–200.
[13] H. M. MÖLLER, *Lower bounds for the number of nodes in cubature formulae*, Numerical Integration, Internat. Ser. Numer. Math. Vol. 45, G. Hämmerlin, ed., Birkhäuser, Basel, 1979.
[14] I. P. MYSOVSKIKH, *Interpolatory Cubature Formulas*, "Nauka," Moscow, 1981 (in Russian).
[15] B. REZNICK, *Sums of even powers of real linear forms*, Memoir Amer. Math. Soc., 96 (1992), no. 463, viii + 155 pp.
[16] J. J. SEIDEL, *Isometric embeddings and geometric designs. Trends in discrete mathematics*, Discrete Math., 136 (1994), pp. 281–293.
[17] S. L. SOBOLEV, *Cubature formulas on the sphere invariant under finite groups of rotations*, Sov. Math. Dokl., 3 (1962), pp. 1307–1310.

[18] A. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[19] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.
[20] H. Szegő, *Orthogonal polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. 23, American Mathematical Society, Providence, RI, 1975.
[21] N. J. Vilenkin, *Special Functions and the Theory of Group Representations*, Amer. Math. Soc. Trans. Math. Monographs 22, American Mathematical Society, Providence, RI, 1968.
[22] Y. Xu, *On multivariate orthogonal polynomials*, SIAM J. Math. Anal., 24 (1993), pp. 783–794.
[23] Y. Xu, *Common Zeros of Polynomials in Several Variables and Higher Dimensional Quadrature*, Pitman Research Notes in Mathematics Series, Longman, Essex, 1994.
[24] Y. Xu, *On orthogonal polynomials in several variables*, in Special Functions, q-series, and Related Topics, Fields Institute Communications, vol. 14, 1997, pp. 247–270.
[25] Y. Xu, *Summability of Fourier orthogonal series for Jacobi polynomials on a ball in* $\mathbb{R}^d$, Trans. Amer. Math. Soc., to appear.
[26] Y. Xu, *Orthogonal polynomials for a family of product weight functions on the spheres*, Canad. J. Math., 49 (1997), pp. 175–192.

SIAM J. MATH. ANAL.
Vol. 29, No. 3, pp. 794–822, May 1998

016

CONVOLUTIONS FOR ORTHOGONAL POLYNOMIALS FROM LIE AND QUANTUM ALGEBRA REPRESENTATIONS*

H. T. KOELINK† AND J. VAN DER JEUGT‡

Abstract. The interpretation of the Meixner–Pollaczek, Meixner, and Laguerre polynomials as overlap coefficients in the positive discrete series representations of the Lie algebra $\mathfrak{su}(1,1)$ and the Clebsch–Gordan decomposition lead to generalizations of the convolution identities for these polynomials. Using the Racah coefficients, convolution identities for continuous Hahn, Hahn, and Jacobi polynomials are obtained. From the quantized universal enveloping algebra for $\mathfrak{su}(1,1)$, convolution identities for the Al-Salam and Chihara polynomials and the Askey–Wilson polynomials are derived by using the Clebsch–Gordan and Racah coefficients. For the quantized universal enveloping algebra for $\mathfrak{su}(2)$, q-Racah polynomials are interpreted as Clebsch–Gordan coefficients, and the linearization coefficients for a two-parameter family of Askey–Wilson polynomials are derived.

Key words. orthogonal polynomials, convolution, Lie algebra, quantum algebra

AMS subject classifications. 33C80, 33D80, 33C45, 33D45, 17B20, 17B37

PII. S003614109630673X

1. Introduction.

The representation theory of Lie algebras and quantum algebras, or quantized universal enveloping algebras [9], is intimately linked to special functions of (basic) hypergeometric type; see, e.g., [35], [9]. In this paper we consider especially the Lie algebra $\mathfrak{su}(1,1)$ and its quantum analogue $U_q(\mathfrak{su}(1,1))$, and we derive convolution identities for certain orthogonal polynomials which occur as overlap coefficients. The idea, which is due to Granovskii and Zhedanov [15], [16], [17], see also [34], is to consider (generalized) eigenvectors of a suitable element of the Lie algebra which is a recurrence operator in an irreducible representation of this Lie algebra. Then there is a relation between these eigenvectors and the eigenvectors of this Lie algebra element in the n-fold tensor product of irreducible representations of the Lie algebra. From the tensor product decomposition in irreducible representations for $n = 2, 3$, we obtain identities for these eigenvectors involving Clebsch–Gordan and Racah coefficients. In particular, if the overlap coefficients are known in terms of special functions, we obtain identities for these special functions in this way.

For the Lie algebra $\mathfrak{su}(1,1)$ and the positive discrete series representations, a special case of this approach is contained in Granovskii and Zhedanov [16], but the result is not worked out in detail. Elaborating the method of Granovskii and Zhedanov, Van der Jeugt [34] obtains a generalization of the classical convolution identity for the Laguerre polynomials [13, section 10.12, eq. (41)]. Van der Jeugt [34] also considers the boson Lie algebra $\mathfrak{b}(1)$, a central extension of the oscillator algebra, leading to a generalization of the convolution identity for Hermite polynomials [13, section

*Received by the editors July 17, 1996; accepted for publication (in revised form) March 3, 1997. http://www.siam.org/journals/sima/29-3/30673.html

†Vakgroep Wiskunde, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, the Netherlands (koelink@wins.uva.nl). The research of this author was supported by the Netherlands Organization for Scientific Research (NWO) under project 610.06.100.

‡Vakgroep Toegepaste Wiskunde en Informatica, Universiteit Gent, Krijgslaan 281-S9, B-9000 Gent, Belgium (Joris.VanderJeugt@rug.ac.be). The author is Senior Research Associate of the Fund for Scientific Research–Flanders (Belgium).

10.13, eq. (38)]. The last identity follows from the previous one by a well-known limit transition of Laguerre polynomials to Hermite polynomials; see, e.g., [22].

Apart from the Laguerre and the Hermite polynomials, the Meixner–Pollaczek, Meixner, and Charlier polynomials, which all fit into the Askey scheme of hypergeometric orthogonal polynomials [5], [22], also satisfy a convolution identity of the same form. This is a straightforward consequence of the existence of a generating function of a special kind; see Al-Salam [1]. It is also known that the Meixner–Pollaczek and the Meixner polynomials can be interpreted as overlap coefficients in the positive discrete series representations of $\mathfrak{su}(1,1)$; see Masson and Repka [29]. In section 3 we show how the method of Granovskii and Zhedanov for the two-fold tensor product of positive discrete series representations of $\mathfrak{su}(1,1)$ leads to a generalization of the convolution identity for Meixner–Pollaczek polynomials, from which generalized convolution formulas for Meixner, Laguerre, Charlier, and Hermite polynomials can be obtained by substitution or by limit transitions. Next, using the three-fold tensor product representation, we obtain a very general convolution identity for continuous Hahn polynomials, and similarly for the Hahn and Jacobi polynomials. These identities can also be viewed as yielding connection coefficients between two sets of orthogonal polynomials in two variables with respect to the same orthogonality measure. With this point of view, this result coincides with Dunkl's results [11], [12]. Our derivation gives an intrinsic explanation for the occurrence of balanced ${}_4F_3$-series as connection coefficients; they are Racah coefficients. Actually, the interpretation as orthogonal polynomials in two variables works in general and is an intrinsic way to determine the S-functions in [16] and [34] in terms of orthogonal polynomials instead of reducing a triple sum to a single sum.

In section 4 we apply the same idea to the quantized universal enveloping algebra $U_q(\mathfrak{su}(1,1))$ and its positive discrete series representations. Due to the noncocommutativity of the comultiplication, which is needed to define the tensor product representation, the tensor product of eigenvectors is no longer an eigenvector in the tensor product representation. This can be solved if we restrict to operators related to so-called twisted primitive elements in $U_q(\mathfrak{su}(1,1))$; see, e.g., [28], [24]. Then the whole machinery works and we obtain a generalization of the Al-Salam and Chihara [2] convolution identity for the Al-Salam and Chihara polynomials by considering the Clebsch–Gordan coefficients in the two-fold tensor product. Going to the three-fold tensor product representations yields a very general convolution identity for Askey–Wilson polynomials also involving q-Racah polynomials, and Theorem 4.10 is the key result of this paper. Overlap coefficients are also considered in somewhat more generality in Klimyk and Kachurik [21], but we have to restrict ourselves to the twisted primitive elements in order to keep the action in the tensor product representations manageable.

It is interesting to note that in this derivation we have a natural interpretation of the continuous Hahn, Hahn, and Jacobi polynomials as Clebsch–Gordan coefficients for the Lie algebra $\mathfrak{su}(1,1)$. Similarly, we have an interpretation of the Askey–Wilson polynomials as Clebsch–Gordan coefficients for the quantized universal enveloping algebra $U_q(\mathfrak{su}(1,1))$. In section 5 we shortly discuss the corresponding result for the quantized universal enveloping algebra $U_q(\mathfrak{su}(2))$, where the q-Racah polynomials then occur as Clebsch–Gordan coefficients. This case can be obtained formally from the results for $U_q(\mathfrak{su}(1,1))$. Since in the dual Hopf $*$-algebra the so-called zonal spherical elements are known in terms of a two-parameter family of Askey–Wilson polynomials, cf. [28], we obtain the explicit linearization coefficients for this subfamily of the Askey–Wilson polynomials.

It should be remarked that there does not seem to be an appropriate q-analogue of the boson Lie algebra $\mathfrak{b}(1)$. Either the Hopf $*$-algebra structure is lacking, or, as in [19], the recurrence in the two-fold tensor product representation seems unmanageable.

Instead of using generalized eigenvectors we use the spectral theory of Jacobi matrices, which we recall briefly in section 2. In particular we use this theory to interpret certain recurrence operators in $\ell^2(\mathbb{Z}_+)^{\otimes n}$, $n = 1, 2, 3$, as multiplication operators in certain weighted L^2-spaces on $\mathbb{R}^n$. This approach exploits the theory of orthogonal polynomials; cf. Propositions 3.3 and 4.3.

The notation for (basic) hypergeometric series is the standard one as in Gasper and Rahman [14]. Unexplained notions for quantized universal enveloping algebras can be found in Chari and Pressley [9].

2. Jacobi matrices and orthogonal polynomials. We recall some of the results on the spectral theory of Jacobi matrices and the relation with orthogonal polynomials. For more information we refer to Berezanskiĭ [7, Chap. VII, section 1]; see also Masson and Repka [29] and Klimyk and Kachurik [21]. The operator J acting on the standard orthonormal basis $\{e_n \mid n \in \mathbb{Z}_+\}$ of $\ell^2(\mathbb{Z}_+)$ by

$$Je_n = a_n\, e_{n+1} + b_n\, e_n + a_{n-1}\, e_{n-1}, \qquad a_n > 0,\ b_n \in \mathbb{R}, \tag{2.1}$$

is called a Jacobi matrix. This operator is symmetric, and its deficiency indices are $(0,0)$ or $(1,1)$. In particular, if the coefficients a_n and b_n are bounded, J is a bounded operator on $\ell^2(\mathbb{Z}_+)$ and thus self-adjoint. J is a self-adjoint operator, possibly unbounded, if $\sum_{n=0}^{\infty} a_n^{-1} = \infty$ by Carleman's condition. Then e_0 is a cyclic vector for J; i.e., the span of finite linear combinations of the form $J^p e_0$, $p \in \mathbb{Z}_+$, is dense in $\ell^2(\mathbb{Z}_+)$. This is the case for all Jacobi matrices considered in this paper.

Assuming this, we can use the same coefficients a_n, b_n to generate polynomials $p_n(x)$ of degree n in x by the recurrence relation

$$xp_n(x) = a_n\, p_{n+1}(x) + b_n\, p_n(x) + a_{n-1}\, p_{n-1}(x), \qquad p_{-1}(x) = 0,\ p_0(x) = 1. \tag{2.2}$$

By Favard's theorem there exists a positive measure m on the real line such that the polynomials $p_n(x)$ are orthonormal;

$$\int_{\mathbb{R}} p_n(x) p_m(x)\, dm(x) = \delta_{n,m}.$$

The measure is obtained by $m(B) = \langle E(B)e_0, e_0\rangle$, B Borel set, where E denotes the spectral decomposition of the self-adjoint operator J.

We can represent the operator J as a multiplication operator M_x on $L^2(m)$, where $M_x f(x) = xf(x)$. For this we define

$$\Lambda\colon \ell^2(\mathbb{Z}_+) \to L^2(m), \qquad \big(\Lambda e_n\big)(x) = p_n(x);$$

then Λ is a unitary operator, since it maps an orthonormal basis onto an orthonormal basis. Note that we use here that the polynomials are dense in $L^2(m)$, since the self-adjointness of J implies that the corresponding moment problem is determined. From (2.1) and (2.2) it follows that $\Lambda \circ J = M_x \circ \Lambda$, so Λ intertwines the Jacobi matrix J on $\ell^2(\mathbb{Z}_+)$ with the multiplication operator M_x on $L^2(m)$.

3. The case $\mathfrak{su}(1,1)$. The Lie algebra $\mathfrak{su}(1,1)$ is given by

$$[H,B]=2B, \qquad [H,C]=-2C, \qquad [B,C]=H.$$

There is a $*$-structure by $H^*=H$ and $B^*=-C$.

The positive discrete series representations π_k of $\mathfrak{su}(1,1)$ are unitary representations labelled by $k>0$. The representation space is $\ell^2(\mathbb{Z}_+)$ equipped with orthonormal basis $\{e_n^k\}_{\{n\in\mathbb{Z}_+\}}$. The action is given by

$$\begin{aligned}\pi_k(H)\,e_n^k &= 2(k+n)\,e_n^k,\\ \pi_k(B)\,e_n^k &= \sqrt{(n+1)(2k+n)}\,e_{n+1}^k,\\ \pi_k(C)\,e_n^k &= -\sqrt{n(2k+n-1)}\,e_{n-1}^k.\end{aligned} \tag{3.1}$$

The tensor product of two positive discrete series representations decomposes as

$$\pi_{k_1}\otimes\pi_{k_2}=\bigoplus_{j=0}^{\infty}\pi_{k_1+k_2+j}. \tag{3.2}$$

The corresponding intertwining operator can be expressed by means of the Clebsch–Gordan coefficients

$$e_n^k=\sum_{n_1,n_2} C_{n_1,n_2,n}^{k_1,k_2,k}\,e_{n_1}^{k_1}\otimes e_{n_2}^{k_2}. \tag{3.3}$$

Later we also use the notation $e_n^{(k_1k_2)k}$ for e_n^k to stress the fact that this vector arises from the decomposition $\pi_{k_1}\otimes\pi_{k_2}$ into irreducible representations. The Clebsch–Gordan coefficients are nonzero only if $n_1+n_2=n+j$, $k=k_1+k_2+j$ for $j,n_1,n_2,n\in\mathbb{Z}_+$ by considering the action of H on both sides. We normalize the Clebsch–Gordan coefficients by $\langle e_0^k, e_0^{k_1}\otimes e_j^{k_2}\rangle>0$.

For the above results Vilenkin and Klimyk [35, section 8.7] can be consulted.

3.1. Clebsch–Gordan coefficients and orthogonal polynomials. The Meixner–Pollaczek polynomials are defined by

$$P_n^{(\lambda)}(x;\phi)=\frac{(2\lambda)_n}{n!}e^{in\phi}\,{}_2F_1\left(\begin{matrix}-n,\lambda+ix\\ 2\lambda\end{matrix};1-e^{-2i\phi}\right). \tag{3.4}$$

For $\lambda>0$ and $0<\phi<\pi$, these are orthogonal polynomials with respect to a positive measure on $\mathbb{R}$; see [22], [33, App.], and references therein. The orthonormal Meixner–Pollaczek polynomials

$$p_n(x)=p_n^{(\lambda)}(x;\phi)=\sqrt{\frac{n!}{\Gamma(n+2\lambda)}}P_n^{(\lambda)}(x;\phi)$$

satisfy the three-term recurrence relation

$$\begin{aligned}2x\,\sin\phi\,p_n(x) &= a_n\,p_{n+1}(x)-2(n+\lambda)\cos\phi\,p_n(x)+a_{n-1}\,p_{n-1}(x),\\ a_n &= \sqrt{(n+1)(n+2\lambda)}.\end{aligned}$$

The orthogonality measure for Meixner–Pollaczek polynomials is absolutely continuous. Define

$$w^{(\lambda)}(x;\phi)=\frac{(2\sin\phi)^{2\lambda}}{2\pi}e^{(2\phi-\pi)x}\left|\Gamma(\lambda+ix)\right|^2;$$

then

$$\int_{\mathbb{R}} p_n^{(\lambda)}(x;\phi)p_m^{(\lambda)}(x;\phi)\, w^{(\lambda)}(x;\phi)dx = \delta_{nm}.$$

Define the self-adjoint element in $\mathfrak{su}(1,1)$;

$$X_\phi = -\cos\phi\, H + B - C. \tag{3.5}$$

PROPOSITION 3.1. $\Lambda\colon \ell^2(\mathbb{Z}_+) \to L^2(\mathbb{R}, w^{(k)}(x;\phi)dx)$, $e_n^k \mapsto p_n^{(k)}(\cdot;\phi)$, *is a unitary mapping intertwining* $\pi_k(X_\phi)$ *acting in* $\ell^2(\mathbb{Z}_+)$ *with* $M_{2x\sin\phi}$ *on* $L^2(\mathbb{R}, w^{(k)}(x;\phi)dx)$.

Here, and elsewhere, M_g denotes multiplication by the function g, so $M_g f(x) = g(x)f(x)$.

Proof. Use (3.1) and (3.5) to see that $\pi_k(X_\phi)$ is a Jacobi matrix. Next compare the coefficients with the three-term recurrence relation for the orthonormal Meixner–Pollaczek polynomials to find the result as in section 2. □

Proposition 3.1 states that $v^k(x) = \sum_{n=0}^\infty p_n^{(k)}(x;\phi)\, e_n^k$ is a generalized eigenvector for $\pi_k(X_\phi)$ for the eigenvalue $2x\sin\phi$. Next we study the action of X_ϕ in the tensor product representation $\pi_{k_1}\otimes\pi_{k_2}$. Recall that $\Delta(X_\phi) = 1\otimes X_\phi + X_\phi\otimes 1$.

PROPOSITION 3.2. $\Upsilon\colon \ell^2(\mathbb{Z}_+)\otimes\ell^2(\mathbb{Z}_+) \to L^2(\mathbb{R}^2, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)dx_1dx_2)$, *defined by* $e_{n_1}^{k_1}\otimes e_{n_2}^{k_2} \mapsto p_{n_1}^{(k_1)}(x_1;\phi)p_{n_2}^{(k_2)}(x_2;\phi)$, *is a unitary mapping intertwining* $\pi_{k_1}\otimes\pi_{k_2}(\Delta(X_\phi))$ *with* $M_{2(x_1+x_2)\sin\phi}$.

Proof. This can be seen by using the mapping Λ of Proposition 3.1 in the second tensor factor and solving the resulting three-term recurrence in the first factor. □

Proposition 3.2 states that

$$v^{k_1,k_2}(x_1,x_2) = \sum_{n_1,n_2=0}^\infty p_{n_1}^{(k_1)}(x_1;\phi)p_{n_2}^{(k_2)}(x_2;\phi)\; e_{n_1}^{k_1}\otimes e_{n_2}^{k_2}$$

are generalized eigenvectors for $\pi_{k_1}\otimes\pi_{k_2}(\Delta(X_\phi))$ for the eigenvalue $2(x_1+x_2)\sin\phi$.

So Υ maps the basis $e_{n_1}^{k_1}\otimes e_{n_2}^{k_2}$ onto orthonormal polynomials in two variables. By the Clebsch–Gordan decomposition (3.2) there exists another orthonormal basis e_n^k for the tensor product representation space. So Υe_n^k gives another set of orthonormal polynomials in two variables in $L^2(\mathbb{R}^2, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)dx_1dx_2)$.

In order to formulate the result we need the continuous Hahn polynomials introduced by Atakishiyev and Suslov [6], for which we use Askey's [3] notation; see also [22], [23];

$$p_n(x;a,b,c,d) = i^n\frac{(a+c)_n(a+d)_n}{n!}\,{}_3F_2\left(\begin{matrix}-n, n+a+b+c+d-1, a+ix\\ a+c,\ a+d\end{matrix};1\right), \tag{3.6}$$

satisfying the orthogonality relations for $\Re(a,b,c,d) > 0$,

$$\frac{1}{2\pi}\int_{\mathbb{R}}\Gamma(a+ix)\Gamma(b+ix)\Gamma(c-ix)\Gamma(d-ix)\, p_n(x;a,b,c,d)p_m(x;a,b,c,d)\,dx$$
$$= \delta_{nm}\frac{\Gamma(n+a+c)\Gamma(n+a+d)\Gamma(n+b+c)\Gamma(n+b+d)}{n!\,(2n+a+b+c+d-1)\Gamma(n+a+b+c+d-1)}.$$

The orthogonality measure is positive for $a=\bar{c}$, $b=\bar{d}$.

PROPOSITION 3.3. *In* $L^2(\mathbb{R}^2, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)dx_1dx_2)$ *we have*

$$\Upsilon e_n^k(x_1,x_2) = p_n^{(k)}(x_1+x_2;\phi)\,\Upsilon e_0^k(x_1,x_2),$$
$$\Upsilon e_0^k(x_1,x_2) = Cp_j(x_1;k_1,k_2-i(x_1+x_2),k_1,k_2+i(x_1+x_2)),$$
$$C = (-2\sin\phi)^j\sqrt{\frac{j!\,(2j+2k_1+2k_2-1)\Gamma(j+2k_1+2k_2-1)}{\Gamma(2k_1+j)\Gamma(2k_2+j)}}.$$

Note that $\Upsilon e_0^k(x_1,x_2)$ is indeed a polynomial in x_1, x_2.

Proof. The first statement follows from use of the intertwining of Proposition 3.2 and the intertwining of (3.2);

$$2(x_1+x_2)\sin\phi\,\Upsilon e_n^k(x_1,x_2) = M_{2(x_1+x_2)\sin\phi}\Upsilon e_n^k(x_1,x_2) = \big(\Upsilon\pi_k(X_\phi)\,e_n^k\big)(x_1,x_2),$$

which gives a three-term recurrence relation for Υe_n^k with respect to n of the same form as in Proposition 3.1. Taking into account the initial conditions proves the first statement.

To prove the second statement we note that for $k=k_1+k_2+j$, $l=k_1+k_2+i$,

$$\delta_{ij}\delta_{mn} = \langle e_n^k, e_m^l\rangle = \langle \Upsilon e_n^k, \Upsilon e_m^l\rangle = \int\!\!\int_{\mathbb{R}^2} p_n^{(k)}(x_1+x_2;\phi)p_m^{(l)}(x_1+x_2;\phi)$$
$$\times\Big(\Upsilon e_0^k(x_1,x_2)\Upsilon e_0^l(x_1,x_2)\Big)w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)\,dx_1dx_2$$

by the first statement and Proposition 3.2. Introduce $s=x_1+x_2$, $t=x_1$; then we find

$$\delta_{ij}\delta_{mn} = \int_{\mathbb{R}} p_n^{(k)}(s;\phi)p_m^{(l)}(s;\phi)\int_{\mathbb{R}}\Upsilon e_0^k(t,s-t)\Upsilon e_0^l(t,s-t)w^{(k_1)}(t;\phi)w^{(k_2)}(s-t;\phi)\,dtds.$$

In case $k=l$, or $i=j$, we see that the inner integral must equal the normalized orthogonality measure for the Meixner–Pollaczek polynomials $p_n^{(k)}(s;\phi)$, since the corresponding moment problem is determined. In case $k\neq l$, or $i\neq j$, we conclude that the inner integral integrated against any polynomial gives zero, so it must be zero since the polynomials are dense in $L^2(\mathbb{R}^2, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)dx_1dx_2)$. So we get

$$\delta_{ij}w^{(k)}(s;\phi) = e^{(2\phi-\pi)s}\frac{(2\sin\phi)^{2k_1+2k_2}}{4\pi^2}\int_{\mathbb{R}}\Upsilon e_0^k(t,s-t)\Upsilon e_0^l(t,s-t)$$
$$\times\,\Gamma(k_1+it)\Gamma(k_2-is+it)\Gamma(k_1-it)\Gamma(k_2+is-it)\,dt.$$

Apply Υ to (3.3) for $n=0$ to see that $\Upsilon e_0^k(t,s-t)$ is a polynomial of degree j in t. Hence, $\Upsilon e_0^k(t,s-t)$ is a multiple of a continuous Hahn polynomial of degree j with the parameters as in Proposition 3.3.

The value of the constant follows from comparing the squared norms up to a sign. The sign is determined from the condition on the Clebsch–Gordan coefficients. This implies $0<\langle\Upsilon e_0^k, \Upsilon e_0^{k_1}\otimes e_j^{k_2}\rangle$, and using the first two parts of Proposition 3.3 and Proposition 3.2 shows that the sign of C follows from the sign of a double integral of two orthogonal polynomials. Only the integral over x_2 is relevant, and the sign of C equals the sign of the leading coefficient of the continuous Hahn polynomials viewed as a polynomial in x_2, which is $(-1)^j$. ∎

So we can now apply Υ to (3.3) to find $k = k_1 + k_2 + j$,

$$\sum_{n_1+n_2=n+j} C^{k_1,k_2,k}_{n_1,n_2,n}\, p^{(k_1)}_{n_1}(x_1;\phi) p^{(k_2)}_{n_2}(x_2;\phi) = C p^{(k)}_n(x_1+x_2;\phi) \times p_j(x_1;k_1,k_2-i(x_1+x_2),k_1,k_2+i(x_1+x_2)). \tag{3.7}$$

The Clebsch–Gordan coefficients remain to be determined, and this can be done from this formula; see [34]. They can be expressed in terms of ${}_3F_2$-series, which are known as Hahn polynomials. Using the Hahn polynomials defined by

$$Q_n(x;a,b,N) = {}_3F_2\left(\begin{matrix} -n, n+a+b+1, -x \\ a+1,\ -N \end{matrix};1\right) \tag{3.8}$$

for $N \in \mathbb{Z}_+$, $0 \le n \le N$, we have, with $k = k_1 + k_2 + j$, $n_1 + n_2 = n + j$,

$$C^{k_1,k_2,k}_{n_1,n_2,n} = \sqrt{\frac{(2k_1)_{n_1}(2k_2)_{n_2}(2k_1)_j}{n!n_1!n_2!j!\,(2k_1+2k_2+2j)_n(2k_2)_j(2k_1+2k_2+j-1)_j}} \times (n+j)!\,Q_j(n_1;2k_1-1,2k_2-1;n+j);$$

see [35, section 8.7] for another proof.

Using this in (3.7) gives an identity in a weighted L^2-space, but since it is a polynomial identity it holds for all x_1, x_2. Simplifying proves the following theorem.

THEOREM 3.4. *With the notation for continuous Hahn, Meixner–Pollaczek, and Hahn polynomials as in* (3.4), (3.6), *and* (3.8), *the following convolution formula holds*:

$$\binom{n+j}{n} \sum_{l=0}^{n+j} Q_j(l;2k_1-1,2k_2-1,n+j)\, P^{(k_1)}_l(x_1;\phi)\, P^{(k_2)}_{n+j-l}(x_2;\phi) = \frac{(-2\sin\phi)^j}{(2k_1)_j}\, P^{(k_1+k_2+j)}_n(x_1+x_2;\phi)\, p_j(x_1;k_1,k_2-i(x_1+x_2),k_1,k_2+i(x_1+x_2)).$$

Remark 3.5. (i) The case $j = 0$ gives back the convolution identity for the Meixner–Pollaczek polynomials; see, e.g., [1, section 8], [2]. The case $n = 0$ gives another convolution identity for Meixner–Pollaczek polynomials, since the Hahn polynomial reduces to a summable ${}_2F_1$-series.

(ii) The polynomials on both sides of the formula in Theorem 3.4 are orthogonal polynomials in two variables for the space $L^2(\mathbb{R}^2, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)dx_1dx_2)$, so we have proved a connection coefficient formula for these polynomials. The dual connection coefficient formula follows from the orthogonality of the Clebsch–Gordan matrix, or equivalently, from the orthogonality relations for the dual Hahn polynomials.

(iii) Theorem 3.4 shows that the continuous Hahn polynomials have an interpretation as Clebsch–Gordan coefficients for $\mathfrak{su}(1,1)$. Using the generalized eigenvectors, we formally have, cf. (3.3),

$$v^{k_1,k_2}(x_1,x_2) = \sum_k C\, p_j(x_1;k_1,k_2-i(x_1+x_2),k_1,k_2+i(x_1+x_2))\, v^k(x_1+x_2),$$

with C as in Proposition 3.3. The dual relations can be written using the orthogonality measure for the continuous Hahn polynomials.

Recall the Laguerre polynomials $L_n^{(a)}(x) = (a+1)_n/n!\,{}_1F_1(-n;a+1;x)$, the Jacobi polynomials $P_n^{(a,b)}(x) = \frac{(a+1)_n}{n!}\,{}_2F_1(-n,n+a+b+1;a+1;(1-x)/2)$, and the Meixner polynomials $M_n(x;\beta;c) = {}_2F_1(-n,-x;\beta;1-c^{-1})$; see [1], [33].

COROLLARY 3.6 (see [34]). (i) *The Laguerre polynomials satisfy the following convolution identity:*

$$\sum_{l=0}^{n+j} Q_j(l;a,b,n+j)\,L_l^{(a)}(x_1)\,L_{n+j-l}^{(b)}(x_2) = \frac{(-1)^j n!\,j!}{(a+1)_j(n+j)!} \times L_n^{(a+b+1+2j)}(x_1+x_2)\,(x_1+x_2)^j P_j^{(a,b)}\left(\frac{x_2-x_1}{x_1+x_2}\right).$$

(ii) *The Meixner polynomials satisfy the following convolution identity:*

$$(c^{-1}-1)^{-j}\sum_{l=0}^{n+j}\frac{(a)_l(b)_{n+j-l}}{l!\,(n+j-l)!}Q_j(l;a-1,b-1,n+j)\,M_l(x_1;a;c)\,M_{n+j-l}(x_2;b;c)$$
$$= \frac{(a+b+2j)_n}{(n+j)!}\,M_n(x_1+x_2-j;a+b+2j;c)\,(-x_1-x_2)_j Q_j(x_1;a-1,b-1,x_1+x_2).$$

Proof. The first case follows from the limit transition of the Meixner–Pollaczek polynomials to the Laguerre polynomials; $\lim_{\phi\downarrow 0} P_n^{((a+1)/2)}(-2x/\phi;\phi) = L_n^{(a)}(x)$. In this limit transition the continuous Hahn polynomials tend to the Jacobi polynomials.

The second case follows from the substitution $\phi = \ln c/2i$ and replacing x_1 and x_2 by ik_1+ix_1 and ik_2+ix_2. For this substitution the continuous Hahn polynomials go over into the Hahn polynomials. □

Remark 3.7. (i) The case $j=0$ in both formulas gives back the convolution identities for the Laguerre and Meixner polynomials, see, e.g., [1], [2], [13, section 10.12, eq. (41)], and the case $n=0$ gives another convolution identity for the Laguerre and Meixner polynomials. Again these formulas can be viewed as connection coefficient formulas for orthogonal polynomials in two variables.

(ii) The identities of Corollary 3.6 can be obtained by considering the action of $X = -H+B-C$ in the representations π_k and $\pi_{k_1}\otimes\pi_{k_2}$ for the Laguerre case, see [34], and by considering the action of $X_c = -\big((1+c)/(2\sqrt{c})\big)H + B - C$, $0<c<1$, in the representations π_k and $\pi_{k_1}\otimes\pi_{k_2}$ for the Meixner case. The limit case $c\uparrow 1$ in the Meixner result gives the Laguerre result. In this case we can interpret the Jacobi and Hahn polynomials as Clebsch–Gordan coefficients; cf. Remark 3.5(iii).

(iii) Corollary 3.6(ii) is equivalent to Theorem 3.4 by the same substitution. Theorem 3.4 can also be obtained from Corollary 3.6(i) by a double application of the Mellin transform. For this we have to use the Laguerre polynomials that are mapped onto Meixner–Pollaczek polynomials, cf. [25, section 3], and the Jacobi polynomials that are mapped onto the continuous Hahn polynomials, cf. [23, eq. (3.4) with $\Gamma(\beta-i\lambda)$ replaced by $\Gamma(\beta+i\lambda)$].

The other hypergeometric orthogonal polynomials satisfying a convolution identity are the Charlier and Hermite polynomials; cf. [1], [2]. These identities can be obtained by taking the appropriate limits from the Meixner polynomials to the Charlier polynomials and from the Laguerre polynomials to the Hermite polynomials; cf., e.g., [22]. The Hahn polynomials tend to Krawtchouk polynomials and the Jacobi polynomials tend to Hermite polynomials. We use the notation $K_n(x;p,N) = {}_2F_1(-n,-x;-N;p^{-1})$ for Krawtchouk polynomials, $C_n(x;a) = {}_2F_0(-n,-x;-;a^{-1})$ for Charlier polynomials and $H_n(x) = (2x)^n\,{}_2F_0(-n/2,-(n-1)/2;-;-x^{-2})$ for the Hermite polynomials; see Szegő [33].

COROLLARY 3.8 (see [34]). (i) *The Hermite polynomials satisfy the following convolution identity:*

$$\sum_{l=0}^{n+j} K_j\left(l;\frac{a^2}{a^2+b^2},n+j\right)\frac{a^l}{l!}H_l(x)\frac{b^{n+j-l}}{(n+j-l)!}H_{n+j-l}(y)$$
$$=\frac{(a^2+b^2)^{(n+j)/2}}{(n+j)!}\left(\frac{b}{a}\right)^j H_n\left(\frac{ax+by}{\sqrt{a^2+b^2}}\right)H_j\left(\frac{ay-bx}{\sqrt{a^2+b^2}}\right).$$

(ii) *The Charlier polynomials satisfy the following convolution identity:*

$$\sum_{l=0}^{n+j}\binom{n+j}{l}\alpha^l\beta^{n+j-l}K_j\left(l;\frac{\alpha}{\alpha+\beta},n+j\right)C_l(x;\alpha)\,C_{n+j-l}(y;\beta)$$
$$=(-1)^j(\alpha+\beta)^nC_n(x+y-j;\alpha+\beta)\,(-x-y)_jK_j\left(x;\frac{\alpha}{\alpha+\beta},x+y\right).$$

Remark 3.9. (i) Again the case $j=0$ gives known convolution formulas; cf. [1, section 8], [2], [13, section 10.13, eq. (40)]. Corollary 3.8(ii) is derived in a different way in Vilenkin and Klimyk [35, section 8.6.5].

(ii) This time the identities have a similar interpretation, but now we have to use the Lie algebra $\mathfrak{b}(1)$, a central extension of the oscillator algebra; cf. [34]. In particular we can now interpret the Hermite and Charlier polynomials as Clebsch–Gordan coefficients.

3.2. Racah coefficients and orthogonal polynomials. In the tensor product of three positive discrete series representations $\pi_{k_1}\otimes\pi_{k_2}\otimes\pi_{k_3}$ of $\mathfrak{su}(1,1)$ we consider the following orthogonal bases:

$$e_n^{((k_1k_2)k_{12}k_3)k}=\sum_{n_{12},n_3}C^{k_{12},k_3,k}_{n_{12},n_3,n}\;e^{(k_1k_2)k_{12}}_{n_{12}}\otimes e^{k_3}_{n_3} \tag{3.9}$$

$$=\sum_{n_1,n_2,n_3,n_{12}}C^{k_1,k_2,k_{12}}_{n_1,n_2,n_{12}}C^{k_{12},k_3,k}_{n_{12},n_3,n}\;e^{k_1}_{n_1}\otimes e^{k_2}_{n_2}\otimes e^{k_3}_{n_3} \tag{3.10}$$

and

$$e_n^{(k_1(k_2k_3)k_{23})k}=\sum_{n_1,n_{23}}C^{k_1,k_{23},k}_{n_1,n_{23},n}\;e^{k_1}_{n_1}\otimes e^{(k_2k_3)k_{23}}_{n_{23}} \tag{3.11}$$

$$=\sum_{n_1,n_2,n_3,n_{23}}C^{k_2,k_3,k_{23}}_{n_2,n_3,n_{23}}C^{k_1,k_{23},k}_{n_1,n_{23},n}\;e^{k_1}_{n_1}\otimes e^{k_2}_{n_2}\otimes e^{k_3}_{n_3}. \tag{3.12}$$

Here we use the extended notation $e_n^{(k_1k_2)k}$ for the basis of the tensor product decomposition to keep track of how the decomposition is obtained.

These bases are connected by the Racah coefficients, which leads to an intertwiner for the action of $\mathfrak{su}(1,1)$. The Racah coefficients are defined by

$$e_n^{((k_1k_2)k_{12}k_3)k}=\sum_{k_{23}}U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}}\;e_n^{(k_1(k_2k_3)k_{23})k}. \tag{3.13}$$

In the previous formulas the following constraints hold:

(3.14)
$$k_{12}=k_1+k_2+j_{12},\qquad k_{23}=k_2+k_3+j_{23},$$
$$k=k_{12}+k_3+j=k_1+k_{23}+j',\qquad j_{12},j,j_{23},j'\in\mathbb{Z}_+,\ \text{and } j_{12}+j=j_{23}+j'.$$

Thus all above sums are finite sums.

Recall that $(1\otimes\Delta)(\Delta(X_\phi)) = 1\otimes 1\otimes X_\phi + 1\otimes X_\phi\otimes 1 + X_\phi\otimes 1\otimes 1$. The following proposition is proved as Proposition 3.2.

PROPOSITION 3.10. *Define the unitary mapping*

$$\Theta\colon \ell^2(\mathbb{Z}_+)\otimes\ell^2(\mathbb{Z}_+)\otimes\ell^2(\mathbb{Z}_+) \to L^2(\mathbb{R}^3, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)w^{(k_3)}(x_3;\phi)dx_1dx_2dx_3)$$

by

$$\Theta\colon e^{k_1}_{n_1}\otimes e^{k_2}_{n_2}\otimes e^{k_3}_{n_3} \mapsto p^{(k_1)}_{n_1}(x_1;\phi)p^{(k_2)}_{n_2}(x_2;\phi)p^{(k_3)}_{n_3}(x_3;\phi);$$

then Θ *intertwines* $\pi_{k_1}\otimes\pi_{k_2}\otimes\pi_{k_3}(1\otimes\Delta)(\Delta(X_\phi))$ *with* $M_{2(x_1+x_2+x_3)\sin\phi}$.

Remark 3.11. Let $\Lambda^{(k)} = \Lambda$ be the unitary mapping defined in Proposition 3.1 and $\Upsilon^{(k_1k_2)} = \Upsilon$ be the unitary mapping defined in Proposition 3.2. Using the identifications

$$\begin{aligned} &L^2(\mathbb{R}^3, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)w^{(k_3)}(x_3;\phi)dx_1dx_2dx_3)\\ &\quad= L^2(\mathbb{R}, w^{(k_1)}(x_1;\phi)dx_1)\otimes L^2(\mathbb{R}^2, w^{(k_2)}(x_2;\phi)w^{(k_3)}(x_3;\phi)dx_2dx_3)\\ &\quad= L^2(\mathbb{R}^2, w^{(k_1)}(x_1;\phi)w^{(k_2)}(x_2;\phi)dx_1dx_2)\otimes L^2(\mathbb{R}, w^{(k_3)}(x_3;\phi)dx_3), \end{aligned}$$

we have $\Theta = \Lambda^{(k_1)}\otimes\Upsilon^{(k_2k_3)} = \Upsilon^{(k_1k_2)}\otimes\Lambda^{(k_3)}$. Hence, for the orthogonal bases on the right-hand side of (3.9) and (3.12) we have

$$\begin{aligned} \Theta e^{(k_1k_2)k_{12}}_{n_{12}}\otimes e^{k_3}_{n_3} &= \left(\Upsilon^{(k_1k_2)}e^{(k_1k_2)k_{12}}_{n_{12}}\right)\left(\Lambda^{(k_3)}e^{k_3}_{n_3}\right),\\ \Theta e^{k_1}_{n_1}\otimes e^{(k_2k_3)k_{23}}_{n_{23}} &= \left(\Lambda^{(k_1)}e^{k_1}_{n_1}\right)\left(\Upsilon^{(k_2k_3)}e^{(k_2k_3)k_{23}}_{n_{23}}\right). \end{aligned}$$

And the right-hand sides are known from Propositions 3.1 and 3.2 in terms of Meixner–Pollaczek polynomials times continuous Hahn polynomials.

PROPOSITION 3.12. (i) *The following expressions hold:*

$$\begin{aligned} \Theta(e^{((k_1k_2)k_{12}k_3)k}_n)(x_1,x_2,x_3) &= p^{(k)}_n(x_1+x_2+x_3;\phi)\ \Theta(e^{((k_1k_2)k_{12}k_3)k}_0)(x_1,x_2,x_3),\\ \Theta(e^{((k_1k_2)k_{12}k_3)k}_0)(x_1,x_2,x_3) &= \Upsilon^{(k_1k_2)}e^{(k_1k_2)k_{12}}_0(x_1,x_2)\ \Upsilon^{(k_{12}k_3)}e^{(k_{12}k_3)k}_0(x_1+x_2,x_3). \end{aligned}$$

(ii) *The following expressions hold:*

$$\begin{aligned} \Theta(e^{(k_1(k_2k_3)k_{23})k}_n)(x_1,x_2,x_3) &= p^{(k)}_n(x_1+x_2+x_3;\phi)\ \Theta(e^{(k_1(k_2k_3)k_{23})k}_0)(x_1,x_2,x_3),\\ \Theta(e^{(k_1(k_2k_3)k_{23})k}_0)(x_1,x_2,x_3) &= \Upsilon^{(k_2k_3)}e^{(k_2k_3)k_{23}}_0(x_2,x_3)\ \Upsilon^{(k_1k_{23})}e^{(k_1k_{23})k}_0(x_1,x_2+x_3). \end{aligned}$$

Proof. Statement (ii) is proved analogously as statement (i). The first statement of (i) follows from Proposition 3.10 and the decomposition of the three-fold tensor product; cf. Proposition 3.3.

For the second statement we use (3.9), Remark 3.11, and Propositions 3.3 and 3.1 to find

$$\begin{aligned} &\Theta(e^{((k_1k_2)k_{12}k_3)k}_0)(x_1,x_2,x_3)\\ &\quad= \left(\Upsilon^{(k_1k_2)}e^{(k_1k_2)k_{12}}_0\right)(x_1,x_2)\sum_{n_{12}+n_3=j} C^{k_{12},k_3,k}_{n_{12},n_3,0}\ p^{(k_{12})}_{n_{12}}(x_1+x_2;\phi)p^{(k_3)}_{n_3}(x_3;\phi). \end{aligned}$$

The sum can be evaluated as $(\Upsilon^{(k_{12}k_3)}e^{(k_{12}k_3)k}_0)(x_1+x_2,x_3)$ by (3.7). ∎

Next we apply Θ to (3.13); then it follows from Proposition 3.12 that we can divide both sides by the Meixner–Pollaczek polynomial of degree n. Since Θ is unitary we obtain the Wigner–Eckart theorem [35, Chap. 8], stating that the Racah coefficients in (3.13) are independent of n. So we can restrict to the case $n = 0$ of (3.13) before applying Θ without loss of generality. We obtain

$$\begin{aligned}\sum_{j_{23}} U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}} \left(\Upsilon^{(k_2k_3)} e_0^{(k_2k_3)k_{23}}\right)(x_2,x_3) \left(\Upsilon^{(k_1k_{23})} e_0^{(k_1k_{23})k}\right)(x_1,x_2+x_3)\\ = \left(\Upsilon^{(k_1k_2)} e_0^{(k_1k_2)k_{12}}\right)(x_1,x_2) \left(\Upsilon^{(k_{12}k_3)} e_0^{(k_{12}k_3)k}\right)(x_1+x_2,x_3).\end{aligned} \tag{3.15}$$

The Racah coefficients remain to be determined, and this can actually be done from (3.15); see [34]. One can either copy the expression [34, eq. (4.8)] or use the limit $q \uparrow 1$ of the expression for the q-Racah coefficient given in Proposition 4.9. Both lead to the following expression of the Racah coefficients in terms of balanced ${}_4F_3$-series:

(3.16)

$$\begin{aligned}U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}} &= \begin{pmatrix} j+j_{12} \\ j_{23}\end{pmatrix} \frac{(2k_2)_{j_{12}}(2k_3)_j(2k_1+2k_2+2k_3+j+j_{12}-1)_{j_{23}}}{(2k_3,2k_2+2k_3+j_{23}-1)_{j_{23}}(2k_2+2k_3+2j_{23})_{j'}}\\ &\times \left(\frac{j'!(2k_1,2k_{23},2k_1+2k_{23}+j'-1)_{j'}\ j_{23}!(2k_2,2k_3,2k_2+2k_3+j_{23}-1)_{j_{23}}}{j!(2k_{12},2k_3,2k_{12}+2k_3+j-1)_j\ j_{12}!(2k_1,2k_2,2k_1+2k_2+j_{12}-1)_{j_{12}}}\right)^{1/2}\\ &\times {}_4F_3\left(\begin{matrix}2k_1+2k_2+j_{12}-1, 2k_2+2k_3+j_{23}-1, -j_{12}, -j_{23}\\ 2k_2, 2k_1+2k_2+2k_3+j+j_{12}-1, -j-j_{12}\end{matrix};1\right),\end{aligned}$$

with the convention (3.14).

The Racah coefficients can be rewritten in terms of the Racah polynomials defined by

$$R_n(\lambda(x);\alpha,\beta,\gamma,\delta) = {}_4F_3\left(\begin{matrix}-n, n+\alpha+\beta+1, -x, x+\gamma+\delta+1\\ \alpha+1,\ \beta+\delta+1,\ \gamma+1\end{matrix};1\right), \tag{3.17}$$

where $\lambda(x) = x(x+\gamma+\delta+1)$, one of the lower parameters equals $-N$, $N \in \mathbb{Z}_+$, and $0 \le n \le N$; see Wilson [36] or [22]. The orthogonality relations for the Racah polynomials follow from the fact that the Racah coefficients form a unitary matrix.

So we obtain the following theorem by simplifying (3.15) using $s = x_1+x_2+x_3$ and the explicit expression (3.16).

THEOREM 3.13. *The continuous Hahn polynomials satisfy the following convolution identity:*

$$\begin{aligned}&\sum_{l=0}^{n+j} \begin{pmatrix} j+n \\ n\end{pmatrix} \frac{(2k_2)_n(2k_3)_j(2k_1+2k_2+2k_3+j+n-1)_l}{(2k_3)_l(2k_2+2k_3+l-1)_l(2k_2+2k_3+2l)_{j+n-l}}\\ &\quad\times R_l(\lambda(n);2k_2-1,2k_3-1,-j-n-1,2k_1+2k_2+j+n-1)\\ &\quad\times p_{n+j-l}(x_1;k_1,k_2+k_3+l-is,k_1,k_2+k_3+l+is)\\ &\quad\times p_l(x_2;k_2,k_3-i(s-x_1),k_2,k_3+i(s-x_1))\\ &= p_n(x_1;k_1,k_2-i(x_1+x_2),k_1,k_2+i(x_1+x_2))\\ &\quad\times p_j(x_1+x_2;k_1+k_2+n,k_3-is,k_1+k_2+n,k_3+is),\end{aligned}$$

with the notation as in (3.6), (3.17).

Remark 3.14. (i) Theorem 3.13 can be considered as a connection coefficient formula between two systems of orthogonal polynomials for the absolutely continuous orthogonality measure with weight

$$\Gamma(k_1+ix_1)\Gamma(k_1-ix_1)\Gamma(k_2+ix_2)\Gamma(k_2-ix_2)\Gamma(k_3+i(s-x_1-x_2))\Gamma(k_3-i(s-x_1-x_2))$$

on $\mathbb{R}^2$. This follows from substituting $s = x_1 + x_2 + x_3$ in the weighted L^2-space of Proposition 3.10 and leaving out the integration with respect to s, which can be done by Proposition 3.12 and the Wigner–Eckart theorem.

(ii) Theorem 3.4 can be obtained as a limit case of Theorem 3.13 by letting $k_3 \to \infty$ and using the limit transition of the continuous Hahn polynomials to the Meixner–Pollaczek polynomials; see, e.g., [22]. Note that Theorem 3.4 is used in the derivation of Theorem 3.13.

(iii) Application of Θ to (3.9)–(3.12) gives results which are immediately derivable from Theorem 3.4.

COROLLARY 3.15 (see [34]). (i) *The Jacobi polynomials satisfy the convolution identity*

$$\begin{aligned}\sum_{l=0}^{n+j} \binom{j+n}{n} &\frac{(b+1)_n(c+1)_j(a+b+c+j+n+2)_l}{(c+1)_l(b+c+l+1)_l(b+c+2l+2)_{j+n-l}}\\ &\times R_l(\lambda(n);b,c,-j-n-1,a+b+j+n+1)\\ &\times P_{n+j-l}^{(a,b+c+2l+1)}(1-2x_1)\,(1-x_1)^l P_l^{(b,c)}\left(\frac{1-x_1-2x_2}{1-x_1}\right)\\ =\ &(x_1+x_2)^n P_n^{(a,b)}\left(\frac{x_2-x_1}{x_1+x_2}\right)\ P_j^{(a+b+2n+1,c)}(1-2(x_1+x_2)).\end{aligned}$$

(ii) *The Hahn polynomials satisfy the following convolution identity:*

$$\begin{aligned}\sum_{l=0}^{n+j} \binom{j+n}{l} &\frac{(a+1)_{n+j-l}(b+1)_l(b+1)_n(c+1)_j(a+b+c+j+n+2)_l}{(a+1)_n(c+1)_l(b+c+l+1)_l(b+c+2l+2)_{j+n-l}(a+b+2n+2)_j}\\ &\times R_l(\lambda(n);b,c,-j-n-1,a+b+j+n+1)\\ &\times (l-s)_{n+j-l}\,Q_{n+j-l}(x_1;a,b+c+2l+1,s-l)\,(x_1-s)_l\,Q_l(x_2;b,c,s-x_1)\\ =\ &(-x_1-x_2)_n\,Q_n(x_1;a,b,x_1+x_2)\,(n-s)_j\,Q_j(x_1+x_2-n;a+b+2n+1,c,s-n)\end{aligned}$$

with the notation (3.8), (3.17).

Proof. The first result follows from the limit transition of the continuous Hahn polynomials to the Jacobi polynomials. Replace x_i by sx_i and let $s \to \infty$. The second result follows by a similar substitution as in the proof of Corollary 3.6(ii). □

Remark 3.16. (i) Similar to Remark 3.7(iii) we have that Corollary 3.15(ii) and Theorem 3.13 can be obtained from each other by formal substitution. Theorem 3.13 can be obtained from Corollary 3.15(i) by a double application of the Mellin transform. Moreover, Corollary 3.15 can be proved as Theorem 3.13 by analyzing the action of X and X_c, cf. Remark 3.7(ii), in the three-fold tensor product.

(ii) Dunkl [11, Thm. 4.2, Prop. 5.4], [12, Thm. 1.7] has obtained Corollary 3.15, and hence Theorem 3.13, by a different method. Dunkl [11] obtains the two-variable Hahn polynomials by judiciously guessing solutions for a certain difference equation arising from the representation theory of the symmetric group. By symmetry considerations there are more solutions of this type, and the connection coefficients can

be calculated in terms of balanced ${}_4F_3$-series. The derivation in this paper gives an intrinsic explanation for the occurrence of the Racah polynomials as connection coefficients. See also Dunkl [11], [12] for the orthogonality relations for these two-variable Hahn and Jacobi polynomials for suitable restrictions on the parameters.

We do not obtain extensions of Corollary 3.8 in this way. For $k_1, k_2, k_3 \to \infty$ in Theorem 3.13 we obtain the same result. This is also explained by the fact that the Racah coefficients for the Lie algebra $\mathfrak{b}(1)$ are of the same form as the Clebsch–Gordan coefficients; cf. [34].

4. The case $U_q(\mathfrak{su}(1,1))$. Let $U_q(\mathfrak{sl}(2,\mathbb{C}))$ be the complex unital associative algebra generated by A, B, C, D subject to the relations

$$AD = 1 = DA, \quad AB = qBA, \quad AC = q^{-1}CA, \quad BC - CB = \frac{A^2 - D^2}{q - q^{-1}}. \tag{4.1}$$

It is a Hopf algebra. We are only concerned with the comultiplication, which is defined by

$$\begin{aligned} &\Delta(A) = A \otimes A, \quad \Delta(B) = A \otimes B + B \otimes D, \\ &\Delta(C) = A \otimes C + C \otimes D, \quad \Delta(D) = D \otimes D \end{aligned} \tag{4.2}$$

on the level of generators and extended as an algebra homomorphism. There are several possible $*$-structures on $U_q(\mathfrak{sl}(2,\mathbb{C}))$, and assuming $0 < q < 1$ we take

$$A^* = A, \quad B^* = -C, \quad C^* = -B, \quad D^* = D,$$

and the corresponding Hopf $*$-algebra is denoted by $U_q(\mathfrak{su}(1,1))$.

The positive discrete series representations π_k of $U_q(\mathfrak{su}(1,1))$ are unitary representations labelled by $k > 0$. They act in $\ell^2(\mathbb{Z}_+)$ and the action of the generators is given by

$$\begin{aligned} &\pi_k(A)\, e_n^k = q^{k+n}\, e_n^k, \qquad \pi_k(D)\, e_n^k = q^{-k-n}\, e_n^k, \\ &\pi_k(C)\, e_n^k = q^{1/2-k-n} \frac{\sqrt{(1-q^{2n})(1-q^{4k+2n-2})}}{q - q^{-1}}\, e_{n-1}^k, \\ &\pi_k(B)\, e_n^k = q^{-1/2-k-n} \frac{\sqrt{(1-q^{2n+2})(1-q^{4k+2n})}}{q^{-1} - q}\, e_{n+1}^k. \end{aligned} \tag{4.3}$$

Note that $\pi_k(D)$, $\pi_k(B)$, $\pi_k(C)$ are unbounded operators, but $\pi_k(A) \in \mathcal{B}(\ell^2(\mathbb{Z}_+))$. The operators that we consider are bounded.

Recall that the tensor product of two representations are defined by use of the comultiplication. The tensor product of two positive discrete series representation decomposes as for the Lie algebra $\mathfrak{su}(1,1)$;

$$\pi_{k_1} \otimes \pi_{k_2} \cong \bigoplus_{j=0}^{\infty} \pi_{k_1+k_2+j}. \tag{4.4}$$

So there exists a unitary matrix mapping the orthogonal basis $e_{n_1}^{k_1} \otimes e_{n_2}^{k_2}$ onto $e_n^{k_1+k_2+j}$ intertwining the action of $U_q(\mathfrak{su}(1,1))$. The matrix elements of this unitary mapping are the Clebsch–Gordan coefficients;

$$e_n^k = \sum_{n_1,n_2=0}^{\infty} C_{n_1,n_2,n}^{k_1,k_2,k}\, e_{n_1}^{k_1} \otimes e_{n_2}^{k_2}, \tag{4.5}$$

where $k = k_1 + k_2 + j$ for $j \in \mathbb{Z}_+$. The sum is finite; $n_1 + n_2 = n + j$. The Clebsch–Gordan coefficients are normalized by $\langle e_0^k, e_0^{k_1} \otimes e_j^{k_2} \rangle > 0$.

These results can be found in Burban and Klimyk [8] and Kalnins, Manocha, and Miller [19]. See Chari and Pressley [9] for general information on quantized universal enveloping algebras.

4.1. Clebsch–Gordan coefficients and orthogonal polynomials. For this section we need the Askey–Wilson polynomials and the Al-Salam and Chihara polynomials, which are a subclass of the Askey–Wilson polynomials. The Askey–Wilson polynomial is defined by

(4.6)
$$p_m(\cos\theta; a, b, c, d|q) = a^{-m}(ab, ac, ad; q)_m \, {}_4\varphi_3 \left(\begin{matrix} q^{-m}, abcdq^{m-1}, ae^{i\theta}, ae^{-i\theta} \\ ab,\ ac,\ ad \end{matrix} ; q, q \right),$$

and it is symmetric in its parameters a, b, c, and d; see [5]. The Al-Salam and Chihara polynomials, introduced originally in [2], are obtained by taking $c = d = 0$ in the Askey–Wilson polynomials;

(4.7)
$$s_m(\cos\theta; a, b|q) = p_m(\cos\theta; a, b, 0, 0|q) = a^{-m}(ab; q)_m \, {}_3\varphi_2 \left(\begin{matrix} q^{-m}, ae^{i\theta}, ae^{-i\theta} \\ ab,\ 0 \end{matrix} ; q, q \right).$$

Let $dm(\cdot; a, b, c, d|q)$ denote the normalized orthogonality measure for the Askey–Wilson polynomials, which is absolutely continuous on $[-1, 1]$ and has at most a finite number of discrete mass points outside $[-1, 1]$. We put $dm(\cdot; a, b|q) = dm(\cdot; a, b, 0, 0|q)$ for the normalized orthogonality measure for the Al-Salam and Chihara polynomials. Explicitly, let

$$w(z) = \frac{(z^2, z^{-2}; q)_\infty}{(az, a/z, bz, b/z, cz, c/z, dz, d/z; q)_\infty};$$

we use $w(z) = w(z; a, b, c, d|q)$ to stress the dependence on the parameters when needed. Let a, b, c, and d be real, or, if complex, appearing in conjugate pairs, and let all the pairwise products of a, b, c, and d not be greater than or equal to 1. Then the Askey–Wilson polynomials $p_n(x) = p_n(x; a, b, c, d|q)$ satisfy the orthogonality relations

$$\begin{aligned} &\frac{1}{2\pi h_0} \int_0^\pi p_n(\cos\theta) p_m(\cos\theta) w(e^{i\theta}) \, d\theta + \frac{1}{h_0} \sum_k p_n(x_k) p_m(x_k) w_k = \delta_{n,m} h_n, \\ &h_n = \frac{(1 - q^{n-1}abcd)}{(1 - q^{2n-1}abcd)} \frac{(q, ab, ac, ad, bc, bd, cd; q)_n}{(abcd; q)_n}, \\ &h_0 = \frac{(abcd; q)_\infty}{(q, ab, ac, ad, bc, bd, cd; q)_\infty}. \end{aligned} \tag{4.8}$$

The points x_k are of the form $\frac{1}{2}(eq^k + e^{-1}q^{-k})$ for e any of the parameters a, b, c, or d with absolute value greater than 1; the sum is over $k \in \mathbb{Z}_+$ such that $|eq^k| > 1$ and w_k is the residue of $z \mapsto w(z)$ at $z = eq^k$ minus the residue at $z = e^{-1}q^{-k}$. So the normalized orthogonality measure $dm(\cdot; a, b, c, d|q)$ can be read off from (4.8); see Askey and Wilson [5] or [14].

Let $S_m(x;a,b|q) = s_m(x;a,b|q)/\sqrt{(q,ab;q)_m}$ denote the orthonormal Al-Salam and Chihara polynomials, which satisfy the three-term recurrence relation

$$\begin{aligned} 2x\, S_n(x) &= a_{n+1}\, S_{n+1}(x) + q^n(a+b)\, S_n(x) + a_n\, S_{n-1}(x), \\ a_n &= \sqrt{(1-abq^{n-1})(1-q^n)}. \end{aligned} \tag{4.9}$$

We now define

$$Y_s = q^{1/2}B - q^{-1/2}C + \frac{s^{-1}+s}{q^{-1}-q}(A-D) \in U_q(\mathfrak{su}(1,1)). \tag{4.10}$$

Then Y_sA is a self-adjoint element in $U_q(\mathfrak{su}(1,1))$ for $s \in \mathbb{R}\backslash\{0\}$, or $s \in \mathbb{T}$. Y_s is a twisted primitive element, i.e., $\Delta(Y_s) = A \otimes Y_s + Y_s \otimes D$, meaning that Y_s is very much like a Lie algebra element.

We also use the notation $\mu(x) = (x+x^{-1})/2 = \mu(x^{-1})$ for $x \neq 0$ in this section.

PROPOSITION 4.1. *Let $\Lambda\colon \ell^2(\mathbb{Z}_+) \to L^2(\mathbb{R}, dm(\cdot;q^{2k}s,q^{2k}/s|q^2))$ be the unitary mapping defined by $\Lambda\colon e_n^k \mapsto S_n(\cdot;q^{2k}s,q^{2k}/s|q^2)$; then Λ intertwines $\pi_k(Y_sA)$ acting in $\ell^2(\mathbb{Z}_+)$ with $2(M_x-\mu(s))/(q^{-1}-q)$.*

Proof. The bounded self-adjoint operator $\pi_k(Y_sA)$ is a Jacobi matrix by (4.10) and (4.3), and the result follows upon comparing with the three-term recurrence (4.9) for the Al-Salam and Chihara polynomials as in section 2. □

Proposition 4.1 says that $v^k(x) = \sum_{n=0}^{\infty} S_n(\mu(x);q^{2k}s,q^{2k}/s|q^2)\, e_n^k$ is a generalized eigenvector of the self-adjoint operator $\pi_k(Y_sA)$ for the eigenvalue

$$\lambda_x = \frac{x+x^{-1}-s-s^{-1}}{q^{-1}-q} = 2\frac{\mu(x)-\mu(s)}{q^{-1}-q}.$$

Due to the fact that the comultiplication on $U_q(\mathfrak{su}(1,1))$ is less simple than for the Lie algebra $\mathfrak{su}(1,1)$, it takes a little more effort to determine the action of Y_sA in $\pi_{k_1} \otimes \pi_{k_2}$. The result can still be phrased using orthogonal polynomials in two variables.

PROPOSITION 4.2. *Define $\Upsilon\colon \ell^2(\mathbb{Z}_+) \otimes \ell^2(\mathbb{Z}_+) \to L^2(\mathbb{R}^2, dm(x_1,x_2))$, where*

$$dm(x_1,x_2) = dm(x_1;q^{2k_1}w_2,q^{2k_1}/w_2|q^2)\, dm(x_2;q^{2k_2}s,q^{2k_2}/s|q^2), \qquad x_2 = \mu(w_2),$$

by

$$\Upsilon\colon e_{n_1}^{k_1} \otimes e_{n_2}^{k_2} \mapsto S_{n_1}(x_1;q^{2k_1}w_2,q^{2k_1}/w_2|q^2)\, S_{n_2}(x_2;q^{2k_2}s,q^{2k_2}/s|q^2);$$

then Υ is a unitary mapping intertwining $\pi_{k_1}\otimes\pi_{k_2}(\Delta(Y_sA))$ with $2(M_{x_1}-\mu(s))/(q^{-1}-q)$ in $L^2(dm(x_1,x_2))$.

Note that $\Upsilon(e_{n_1}^{k_1} \otimes e_{n_2}^{k_2})$ forms a set of orthogonal polynomials in two variables x_1 and x_2 for $L^2(\mathbb{R}^2, dm(x_1,x_2))$, since the Al-Salam and Chihara polynomial is symmetric in its parameters.

Proposition 4.2 states that the vector

$$\begin{aligned} w(x_1;x_2) &= \sum_{n_1=0}^{\infty} S_{n_1}(\mu(x_1);q^{2k_1}x_2,q^{2k_1}/x_2|q^2)\, e_{n_1}^{k_1} \otimes v^{k_2}(x_2) \\ &= \sum_{n_1,n_2=0}^{\infty} S_{n_1}(\mu(x_1);q^{2k_1}x_2,q^{2k_1}/x_2|q^2)\, S_{n_2}(\mu(x_2);q^{2k_2}s,q^{2k_2}/s|q^2)\, e_{n_1}^{k_1} \otimes e_{n_2}^{k_2} \end{aligned}$$

is a generalized eigenvector of $\pi_{k_1} \otimes \pi_{k_2}(\Delta(Y_s A))$ for the eigenvalue λ_{x_1}. This last observation is essentially the way to obtain Proposition 4.2, since $\Delta(Y_s A) = A^2 \otimes Y_s A + Y_s A \otimes 1$ acts as a three-term recurrence operator in $e^{k_1}_{n_1} \otimes v^{k_2}(x_2)$.

Proof. We use $\Delta(Y_s A) = A^2 \otimes Y_s A + Y_s A \otimes 1$ and Proposition 4.1 to define for fixed x_2 the map $\Lambda_0 \colon \ell^2(\mathbb{Z}_+) \otimes \ell^2(\mathbb{Z}_+) \to \ell^2(\mathbb{Z}_+)$ by

$$\Lambda_0 \colon e^{k_1}_{n_1} \otimes e^{k_2}_{n_2} \mapsto S_{n_2}(x_2; q^{2k_2}s, q^{2k_2}/s|q^2)\, e^{k_1}_{n_1}$$

to obtain the recurrence in n_1

$$\begin{aligned}
&\Lambda_0\Big((q^{-1}-q)\big(\pi_{k_1} \otimes \pi_{k_2}\Delta(Y_s A)\big) + s + s^{-1}\Big)\, e^{k_1}_{n_1} \otimes e^{k_2}_{n_2} \\
&\quad = S_{n_2}(x_2; q^{2k_2}s, q^{2k_2}/s|q^2)\Big(q^{2n_1}\big((s+s^{-1})q^{2k_1} + \lambda_{w_2} q^{2k_1}(q^{-1}-q)\big)\, e^{k_1}_{n_1} \\
&\quad + \sqrt{(1-q^{2n_1+2})(1-q^{4k_1+2n})}\, e^{k_1}_{n_1+1} + \sqrt{(1-q^{2n_1})(1-q^{4k_1+2n_1-1})}\, e^{k_1}_{n_1-1}\Big).
\end{aligned}$$

Use the explicit expression for λ_{w_2} and the three-term recurrence relation (4.9) to obtain the result. ∎

We now calculate the action of $\Upsilon\, e^k_n$, which yields another set of orthonormal polynomials for $L^2(\mathbb{R}^2, dm(x_1, x_2))$.

PROPOSITION 4.3. *Let $k = k_1 + k_2 + j$ for $j \in \mathbb{Z}_+$ and $x_1 = \mu(w_1)$; then*

$$\begin{aligned}
\big(\Upsilon\, e^k_n\big)(x_1, x_2) &= S_n(x_1; q^{2k}s, q^{2k}/s|q^2)\, \big(\Upsilon\, e^k_0\big)(x_1, x_2), \\
\big(\Upsilon\, e^k_0\big)(x_1, x_2) &= C\, p_j(x_2; q^{2k_1}w_1, q^{2k_1}/w_1, q^{2k_2}s, q^{2k_2}/s|q^2), \\
C^{-1} = \big(C_j(k_1, k_2)\big)^{-1} &= \sqrt{(q^2, q^{4k_1}, q^{4k_2}, q^{4k_1+4k_2+2j-2}; q^2)_j}.
\end{aligned}$$

Proof. By Proposition 4.2, (4.4), and

$$2\frac{x_1 - \mu(s)}{q^{-1}-q}\, \Upsilon e^k_n(x_1, x_2) = \Upsilon(\pi_k(Y_s A)\, e^k_n)(x_1, x_2),$$

we obtain the three-term recurrence relation as in Proposition 4.1, but with different initial conditions. Hence, the first statement follows.

Since Υ is unitary we have the orthogonality relations $\delta_{nm}\delta_{kl} = \langle \Upsilon e^k_n, \Upsilon e^l_m \rangle =$

$$\int S_n(x_1; q^{2k}s, q^{2k}/s|q^2) S_m(x_1; q^{2l}s, q^{2l}/s|q^2) \int \Upsilon e^k_0(x_1, x_2) \Upsilon e^l_0(x_1, x_2)\, dm(x_1, x_2),$$

by our first observation. As in the proof of Proposition 3.3 we conclude $\Upsilon e^k_0(x_1, x_2) = p_j(x_2)$ is a polynomial of degree j, $k = k_1 + k_2 + j$, in x_2 satisfying the orthogonality relations

$$\int_{x_2} p_j(x_2) p_i(x_2)\, dm(x_1, x_2) = \delta_{ij} dm(x_1; q^{2k}s, q^{2k}/s|q^2)$$

as measures with respect to functions in the variable x_1.

We now assume for ease of presentation that $dm(x_1, x_2)$ is absolutely continuous. The general case can be proved similarly, or it can be obtained by analytic continuation with respect to s. The measure is absolutely continuous for $q^{2k_2} < |s| < q^{-2k_2}$, since $k_1, k_2 > 0$. Put $x_1 = \cos\theta$, $x_2 = \cos\psi$; then we obtain the explicit expression (4.8)

for the orthogonality measure;

$$\frac{1}{2\pi}\int_0^\pi p_i(\cos\psi)p_j(\cos\psi)\frac{(e^{\pm 2i\psi},e^{\pm 2i\theta};q^2)_\infty}{(q^{2k_2}se^{\pm i\psi},q^{2k_2}e^{\pm i\psi}/s,q^{2k_1}e^{\pm i\psi\pm i\theta};q^2)_\infty}d\psi$$
$$=\delta_{ij}\frac{(q^{4k_1+4k_2+4j};q^2)_\infty}{(q^2,q^{4k_1},q^{4k_2};q^2)_\infty}\frac{(e^{\pm 2i\theta};q^2)_\infty}{(q^{2k_1+2k_2+2j}se^{\pm i\theta},q^{2k_1+2k_2+2j}e^{\pm i\theta}/s;q^2)_\infty}$$

for almost all θ. The $\pm$ signs mean that we take all possible combinations in the infinite q-shifted factorials. Cancelling the $(e^{\pm 2i\theta};q^2)_\infty$ on both sides and comparing the result with (4.8), we see that p_j is a multiple of $p_j(\cdot;q^{2k_1}e^{i\theta},q^{2k_1}e^{-i\theta},q^{2k_2}s,q^{2k_2}/s|q^2)$. The constant in front follows up to a sign by comparing the squared norms. As in the proof of Proposition 3.3 the sign of C follows from the normalization of the Clebsch–Gordan coefficients, and now we obtain $C>0$. □

So we obtain a second set of orthonormal polynomials for $L^2(\mathbb{R}^2,dm(x_1,x_2))$ in terms of Al-Salam and Chihara polynomials and Askey–Wilson polynomials.

The convolution formula for the Al-Salam and Chihara polynomials is obtained by applying Υ to (4.5) using the results of Propositions 4.2 and 4.3. The results hold as an identity in a weighted L^2-space, but since it is a polynomial identity it holds for all x_1, x_2; with $x_1=\mu(w_1)$, $x_2=\mu(w_2)$, and $k=k_1+k_2+j$,

$$\sum_{n_1+n_2=n+j} C_{n_1,n_2,n}^{k_1,k_2,k}\,S_{n_1}(x_1;q^{2k_1}w_2,q^{2k_1}/w_2|q^2)\,S_{n_2}(x_2;q^{2k_2}s,q^{2k_2}/s|q^2)$$
$$=\frac{S_n(x_1;q^{2k}s,q^{2k}/s|q^2)p_j(x_2;q^{2k_1}w_1,q^{2k_1}/w_1,q^{2k_2}s,q^{2k_2}/s|q^2)}{\sqrt{(q^2,q^{4k_1},q^{4k_2},q^{4k_1+4k_2+2j-2};q^2)_j}}. \tag{4.11}$$

We have not yet calculated the Clebsch–Gordan coefficients explicitly, but we can now use (4.11) to determine $C_{n_1,n_2,n}^{k_1,k_2,k}$ by specializing to a generating function for the Clebsch–Gordan coefficients. The result is phrased in terms of q-Hahn polynomials, which are defined as follows (cf. [4]):

$$Q_n(q^{-x};a,b,N;q)={}_3\varphi_2\left(\begin{matrix}q^{-n},q^{-x},abq^{n+1}\\ aq,\ q^{-N}\end{matrix};q,q\right).$$

See, e.g., [19] for other derivations of the following lemma.

LEMMA 4.4. *With $n_1+n_2=n+j$ we get*

$$C_{n_1,n_2,n}^{k_1,k_2,k_1+k_2+j}=C\;Q_j(q^{-2n_1};q^{4k_1-2},q^{4k_2-2},n+j;q^2),$$

with the constant C given by

$$\frac{q^{2k_1(n-n_1)}(q^2;q^2)_{n+j}\sqrt{(q^{4k_1};q^2)_{n_1}(q^{4k_2};q^2)_{n_2}(q^{4k_1};q^2)_j}}{\sqrt{(q^2;q^2)_{n_1}(q^2;q^2)_{n_2}(q^2,q^{4k_1+4k_2+4j};q^2)_n(q^2,q^{4k_2},q^{4k_1+4k_2+2j-2};q^2)_j}}.$$

Proof. Observe that $C_{n_1,n_2,n}^{k_1,k_2,k}$ is independent of s, $x_1=\mu(w_1)$ and $x_2=\mu(w_2)$. Specialize $w_2=q^{2k_2}s$ and $w_1=q^{2k_1}/w_2=q^{2k_1-2k_2}/s$; then the Al-Salam and Chihara polynomials in the summand on the left-hand side of (4.11) can be evaluated explicitly, since the ${}_3\varphi_2$-series reduces to 1. For this choice the Askey–Wilson polynomial on the right-hand side can also be evaluated explicitly, and we obtain the generating function

for the Clebsch–Gordan coefficients

$$\sum_{n_1+n_2=n+j} C^{k_1,k_2,k}_{n_1,n_2,n}\, q^{2n_1(k_2-k_1)-2n_2k_2} s^{n_1-n_2} \frac{\sqrt{(q^{4k_1};q^2)_{n_1}(q^{4k_2};q^2)_{n_2}}}{\sqrt{(q^2;q^2)_{n_1}(q^2;q^2)_{n_2}}}$$
$$= \frac{q^{-2jk_2-2n(k_1+k_2+j)} s^{n-j} (q^{4k_1},q^{4k_2},q^{4k_2}s^2;q^2)_j}{\sqrt{(q^2,q^{4k_1},q^{4k_2},q^{4k_1+4k_2+2j-2};q^2)_j}} \sqrt{\frac{(q^{4k_1+4k_2+4j};q^2)_n}{(q^2;q^2)_n}}$$
$$\times\ {}_3\varphi_2\left(\begin{matrix} q^{-2n}, q^{4k_2+2j}, q^{4k_1+2j}/s^2 \\ q^{4k_1+4k_2+4j},\ 0 \end{matrix} ; q^2, q^2 \right).$$

This determines $C^{k_1,k_2,k}_{n_1,n_2,n}$, but it takes some work to find the expression in terms of q-Hahn polynomials. First, take n_1 as the summation parameter in the sum and multiply both sides by s^{n+j} to find that both sides are polynomials of degree $n+j$ in s^2. Apply [14, eq. (III.6)] to rewrite the ${}_3\varphi_2$-series as a polynomial in s^2 and the q-binomial theorem [14, eq. (II.3)] to write $(q^{4k_2}s^2;q^2)_j$ as a polynomial in s^2. Comparing next the coefficients on both sides gives an expression for the Clebsch–Gordan coefficients as a terminating ${}_3\varphi_2$-series. To put it into the required form in terms of q-Hahn polynomials, we need to apply some transformations for ${}_3\varphi_2$-series, namely [14, (III.13), (III.11)]. The constant follows by a straightforward calculation. □

Combining Lemma 4.4 with the unitarity of the intertwining operator consisting of the Clebsch–Gordan coefficients results in the orthogonality relations for the q-Hahn and dual q-Hahn polynomials; cf. [14, section 7.2].

We now have all ingredients to rewrite (4.11). Simplifying proves the following theorem.

THEOREM 4.5. *With the notation* (4.7) *and* (4.6) *for the Al-Salam and Chihara polynomials and Askey–Wilson polynomials and* $x_1=\mu(w_1)$, $x_2=\mu(w_2)$, $n,j\in\mathbb{Z}_+$, $k_1,k_2>0$ *we have*

$$(q^{4k_1};q^2)_j \sum_{l=0}^{n+j} q^{2k_1(n-l)} \begin{bmatrix} n+j \\ l \end{bmatrix}_{q^2} Q_j(q^{-2l};q^{4k_1-2},q^{4k_2-2},n+j;q^2)$$
$$\times\ s_l(x_1;q^{2k_1}w_2,q^{2k_1}/w_2|q^2)\, s_{n+j-l}(x_2;q^{2k_2}s,q^{2k_2}/s|q^2)$$
$$= s_n(x_1;q^{2k_1+2k_2+2j}s,q^{2k_1+2k_2+2j}/s|q^2)\, p_j(x_2;q^{2k_1}w_1,q^{2k_1}/w_1,q^{2k_2}s,q^{2k_2}/s|q^2).$$

Remark 4.6. (i) Theorem 4.5 is a connection coefficient formula for orthogonal polynomials in two variables, orthogonal for the same measure, where the connection coefficients are given by the q-Hahn polynomials. The dual connection coefficient problem follows from the orthogonality for the Clebsch–Gordan coefficients or, equivalently, from the orthogonality relations for the dual q-Hahn polynomials.

(ii) The case $j=0$ gives a simple convolution property for the Al-Salam and Chihara polynomials, since the q-Hahn and the Askey–Wilson polynomial reduce to 1. This was the motivation for Al-Salam and Chihara [2] to introduce the Al-Salam and Chihara polynomials as the most general set of orthogonal polynomials still satisfying a convolution property; see also Al-Salam [1, section 8]. The case $n=0$ is also of interest, since then the q-Hahn polynomial can be evaluated and the Al-Salam and Chihara polynomial on the right-hand side reduces to 1. In both cases we have a free parameter in the sum.

(iii) Formally, in the representation space $\ell^2(\mathbb{Z}_+)\otimes\ell^2(\mathbb{Z}_+)$ we have two bases of (generalized) eigenvectors for the action of Y_sA, namely $v^k(x)$ and $w(x_1;x_2)$. They

are connected by Clebsch–Gordan coefficients, which are now expressible as Askey–Wilson polynomials;

$$w(x_1;x_2)=\sum_{j=0}^{\infty}\frac{p_j(\mu(x_2);q^{2k_1}x_1,q^{2k_1}/x_1,q^{2k_2}s,q^{2k_2}/s|q^2)}{\sqrt{(q^2,q^{4k_1},q^{4k_2},q^{4k_1+4k_2+2j-2};q^2)_j}}\,v^{k_1+k_2+j}(x_1);$$

cf. [15, (23)] for the appropriate analogue in the case $U_q(\mathfrak{su}(2))$ in which the Askey–Wilson polynomials are replaced by q-Racah polynomials. The dual Clebsch–Gordan coefficient relation follows by integrating against the appropriate orthogonality measure for the Askey–Wilson polynomials.

4.2. Racah coefficients and orthogonal polynomials. In the tensor product of three positive discrete series representations $\pi_{k_1}\otimes\pi_{k_2}\otimes\pi_{k_3}$ of $U_q(\mathfrak{su}(1,1))$ we have the same orthogonal bases as in section 3.2 and we use the same notation as in (3.9)–(3.12). Similarly, we now have an intertwiner for the $U_q(\mathfrak{su}(1,1))$-action in terms of q-Racah coefficients;

$$e_n^{((k_1k_2)k_{12}k_3)k}=\sum_{k_{23}}U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}}\;e_n^{(k_1(k_2k_3)k_{23})k}. \tag{4.12}$$

Again the constraints (3.14) hold.

PROPOSITION 4.7. *Define* $\Theta\colon \ell^2(\mathbb{Z}_+)\otimes\ell^2(\mathbb{Z}_+)\otimes\ell^2(\mathbb{Z}_+)\to L^2(\mathbb{R}^3,dm(x_1,x_2,x_3))$ *by*

$$\begin{aligned}&\Theta\big(e^{k_1}_{n_1}\otimes e^{k_2}_{n_2}\otimes e^{k_3}_{n_3}\big)(x_1,x_2,x_3)\\&=S_{n_1}(x_1;q^{2k_1}w_2,q^{2k_1}/w_2|q^2)\,S_{n_2}(x_2;q^{2k_2}w_3,q^{2k_2}/w_3|q^2)\,S_{n_3}(x_3;q^{2k_3}s,q^{2k_3}/s|q^2),\end{aligned}$$

with the measure $dm(x_1,x_2,x_3)$ *given by*

$$dm(x_1;q^{2k_1}w_2,q^{2k_1}/w_2|q^2)\,dm(x_2;q^{2k_2}w_3,q^{2k_2}/w_3|q^2)\,dm(x_3;q^{2k_3}s,q^{2k_3}/s|q^2),$$

where $x_i=\mu(w_i)$. *Then* Θ *is a unitary map intertwining* $\pi_{k_1}\otimes\pi_{k_2}\otimes\pi_{k_3}(1\otimes\Delta)(\Delta(Y_sA))$ *with* $2(M_{x_1}-\mu(s))/(q^{-1}-q)$.

Proof. Observe that $(1\otimes\Delta)(\Delta(Y_sA))=A^2\otimes\Delta(Y_sA)+Y_sA\otimes\Delta(1)$. The proof now proceeds as the proof of Proposition 4.2. □

PROPOSITION 4.8. (i) *The following equality holds with* $x_i=\mu(w_i)$*:*

$$\begin{aligned}&\Theta\big(e_n^{((k_1k_2)k_{12}k_3)k}\big)(x_1,x_2,x_3)=C_j(k_{12},k_3)C_{j_{12}}(k_1,k_2)\,S_n\left(x_1;q^{2k}s,\frac{q^{2k}}{s}|q^2\right)\\&\times\,p_j\left(x_3;q^{2k_{12}}w_1,\frac{q^{2k_{12}}}{w_1},q^{2k_3}s,\frac{q^{2k_3}}{s}|q^2\right)p_{j_{12}}\left(x_2;q^{2k_1}w_1,\frac{q^{2k_1}}{w_1},q^{2k_2}w_3,\frac{q^{2k_2}}{w_3}|q^2\right).\end{aligned}$$

(ii) *The following equality holds with* $x_i=\mu(w_i)$:

$$\begin{aligned}&\Theta\big(e_n^{(k_1(k_2k_3)k_{23})k}\big)(x_1,x_2,x_3)=C_{j'}(k_1,k_{23})C_{j_{23}}(k_2,k_3)\,S_n\left(x_1;q^{2k}s,\frac{q^{2k}}{s}|q^2\right)\\&\times\,p_{j'}\left(x_2;q^{2k_1}w_1,\frac{q^{2k_1}}{w_1},q^{2k_{23}}s,\frac{q^{2k_{23}}}{s}|q^2\right)p_{j_{23}}\left(x_3;q^{2k_2}w_2,\frac{q^{2k_2}}{w_2},q^{2k_3}s,\frac{q^{2k_{23}}}{s}|q^2\right).\end{aligned}$$

The constant $C_j(k_1,k_2)$ *is defined in Proposition* 4.3.

The proof of Proposition 4.8 is slightly more complicated than the proof of its counterpart Proposition 3.12 due to the fact that we do not have a nice factorization for Θ as in Remark 3.11. This is a consequence of the noncocommutativity of the comultiplication for $U_q(\mathfrak{su}(1,1))$.

Note that the occurrence of $S_n(x_1;q^{2k}s,q^{2k}/s|q^2)$ on the right-hand side corresponds to the intertwining property of the Racah coefficients as in the proof of the first statement of Proposition 4.3.

Proof. The proof of (i) and (ii) is similar. To prove (i) we use (3.10) (for the $U_q(\mathfrak{su}(1,1))$-setting) and Proposition 4.7 to find

$$\begin{aligned}\Theta\big(e_n^{((k_1k_2)k_{12}k_3)k}\big)(x_1,x_2,x_3) &= \sum_{n_{12}+n_3=n+j} C^{k_{12},k_3,k}_{n_{12},n_3,n}\, S_{n_3}(x_3;q^{2k_3}s,q^{2k_3}/s|q^2)\\ &\times \sum_{n_1+n_2=n_{12}+j_{12}} C^{k_1,k_2,k_{12}}_{n_1,n_2,n_{12}}\, S_{n_1}(x_1;q^{2k_1}w_2,q^{2k_1}/w_2|q^2)\, S_{n_2}(x_2;q^{2k_2}w_3,q^{2k_2}/w_3|q^2)\\ = C_{j_{12}}(k_1,k_2)&p_{j_{12}}(x_2;q^{2k_1}w_1,q^{2k_1}/w_1,q^{2k_2}w_3,q^{2k_2}/w_3|q^2)\\ &\times \sum_{n_{12}+n_3=n+j} C^{k_{12},k_3,k}_{n_{12},n_3,n}\, S_{n_3}(x_3;q^{2k_3}s,q^{2k_3}/s|q^2)\, S_{n_{12}}(x_1;q^{2k_{12}}w_3,q^{2k_{12}}/w_3,|q^2)\end{aligned}$$

by (4.11). The last sum can be evaluated by another application of (4.11) leading to the result. □

The n-dependence in the right-hand sides of Proposition 4.8 is the same, so we obtain the Wigner–Eckhart theorem for the $U_q(\mathfrak{su}(1,1))$-setting by applying Θ to (4.12). So we can restrict to $n=0$ in (4.12) before applying Θ without loss of generality, and we obtain the following polynomial identity in x_2 and x_3 with w_1 as a parameter:

(4.13)

$$\begin{aligned}&C_j(k_{12},k_3)C_{j_{12}}(k_1,k_2)\,p_j(x_3;q^{2k_{12}}w_1,q^{2k_{12}}/w_1,q^{2k_3}s,q^{2k_3}/s|q^2)\\ &\times p_{j_{12}}(x_2;q^{2k_1}w_1,q^{2k_1}/w_1,q^{2k_2}w_3,q^{2k_2}/w_3|q^2)\\ &=\sum_{j_{23}=0}^{j_{12}+j} U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}}\, C_{j'}(k_1,k_{23})C_{j_{23}}(k_2,k_3)\,p_{j'}(x_2;q^{2k_1}w_1,q^{2k_1}/w_1,q^{2k_{23}}s,q^{2k_{23}}/s|q^2)\\ &\times p_{j_{23}}(x_3;q^{2k_2}w_2,q^{2k_2}/w_2,q^{2k_3}s,q^{2k_{23}}/s|q^2).\end{aligned}$$

Again we can use (4.13) in two ways. First, we specialize to a suitable formula from which the Racah coefficients can be determined explicitly. The Racah coefficients for the finite dimensional representations of $U_q(\mathfrak{sl}(2,\mathbb{C}))$, see section 5, are due to Kirillov and Reshetikhin [20]; see also [35, section 14.5]. Second, with the explicit expression for the Racah coefficients we derive a convolution identity for the Askey–Wilson polynomials.

PROPOSITION 4.9. *The Racah coefficients of* (4.12) *are given by*

$$U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}} = C\ {}_4\varphi_3\left(\begin{matrix} q^{4k_1+4k_2+2j_{12}-2}, q^{4k_2+4k_3+2j_{23}-2}, q^{-2j_{12}}, q^{-2j_{23}}\\ q^{4k_2}, q^{4k_1+4k_2+4k_3+2j+2j_{12}-2}, q^{-2j-2j_{12}}\end{matrix};q^2;q^2\right),$$

with the constant C given by

$$\frac{(q^2,q^{4k_1},q^{4k_{23}},q^{4k_1+4k_{23}+2(j+j_{12}-j_{23})-2};q^2)_{j+j_{12}-j_{23}}^{1/2}}{(q^2,q^{4k_{12}},q^{4k_3},q^{4k_{12}+4k_3+2j-2};q^2)_j^{1/2}}\frac{(q^2,q^{4k_2},q^{4k_3},q^{4k_2+4k_3+2j_{23}-2};q^2)_{j_{23}}^{1/2}}{(q^2,q^{4k_1},q^{4k_2},q^{4k_1+4k_2+2j_{12}-2};q^2)_{j_{12}}^{1/2}}$$
$$\times q^{2k_2(j-j_{23})}\begin{bmatrix} j+j_{12} \\ j_{23}\end{bmatrix}_{q^2}\frac{(q^{4k_3};q^2)_j(q^{4k_2};q^2)_{j_{12}}(q^{4k_1+4k_2+4k_3+2j+2j_{12}-2};q^2)_{j_{23}}}{(q^{4k_3},q^{4k_2+4k_3+2j_{23}-2};q^2)_{j_{23}}(q^{4k_2+4k_3+4j_{23}};q^2)_{j+j_{12}-j_{23}}}.$$

Proof. Take $w_1=s=1$ and $x_2=\mu(q^{2k_1})$ in (4.13) to find

$$\begin{aligned}&p_j(x_3;q^{2k_{12}},q^{2k_{12}},q^{2k_3},q^{2k_3}|q^2)(q^{2k_1+2k_2}w_3,q^{2k_1+2k_2}/w_3;q^2)_{j_{12}}\\&=\sum_{j_{23}}C_1U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}}p_{j_{23}}(x_3;q^{2k_1+2k_2},q^{2k_2-2k_1},q^{2k_3},q^{2k_3}|q^2)\end{aligned}$$

for C_1 an explicit constant depending upon $k_1,k_2,k_3,j_{12},j_{23}$, and j, since two Askey–Wilson polynomials can be evaluated for this choice. So the Racah coefficients occur as the coefficients when developing the polynomial of degree $j+j_{12}$ on the left-hand side into Askey–Wilson polynomials. Hence the Racah coefficients can be obtained from

$$\begin{aligned}&C_2U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}}\\&=\int p_{j_{23}}(x_3;q^{2k_1+2k_2},q^{2k_2-2k_1},q^{2k_3},q^{2k_3}|q^2)p_j(x_3;q^{2k_{12}},q^{2k_{12}},q^{2k_3},q^{2k_3}|q^2)\\&\qquad\times(q^{2k_1+2k_2}w_3,q^{2k_1+2k_2}/w_3;q^2)_{j_{12}}\,dm(x_3;q^{2k_1+2k_2},q^{2k_2-2k_1},q^{2k_3},q^{2k_3}|q^2),\end{aligned}$$

for some known constant C_2. Now observe that

$$\begin{aligned}&(q^{2k_1+2k_2}w_3,q^{2k_1+2k_2}/w_3;q^2)_{j_{12}}\,dm(x_3;q^{2k_1+2k_2},q^{2k_2-2k_1},q^{2k_3},q^{2k_3}|q^2)\\&\qquad=C_3\,dm(x_3;q^{2k_{12}},q^{2k_2-2k_1},q^{2k_3},q^{2k_3}|q^2)\end{aligned}$$

for some known constant C_3 by (3.14) and (4.8). Thus the Racah coefficients can be obtained by integration;

(4.14)
$$\begin{aligned}C_4\,U^{k_1,k_2,k_{12}}_{k_3,k,k_{23}}&=\int p_{j_{23}}(x_3;q^{2k_1+2k_2},q^{2k_2-2k_1},q^{2k_3},q^{2k_3}|q^2)\\&\qquad\times p_j(x_3;q^{2k_{12}},q^{2k_{12}},q^{2k_3},q^{2k_3}|q^2)\,dm(w_3;q^{2k_{12}},q^{2k_2-2k_1},q^{2k_3},q^{2k_3}|q^2),\end{aligned}$$

with C_4 explicitly known. Observe that three out of four of the parameters of each of the Askey–Wilson polynomials in (4.14) coincide with the parameters of the Askey–Wilson measure in (4.14). Use the connection coefficient formula for Askey–Wilson polynomials with one different parameter, cf. Askey and Wilson [5, (6.4–5)] or see [14, (7.6.8–9) with the right-hand side of (7.6.9) multiplied by $(q;q)_n$], twice to rewrite the Askey–Wilson polynomials in terms of Askey–Wilson polynomials with the same parameters as the Askey–Wilson measure in (4.14). By orthogonality the integration is then easily performed and we are left with a single sum, which can be written as a very well poised ${}_8\varphi_7$-series. This can be transformed to a balanced ${}_4\varphi_3$-series by Watson's transformation [14, (III.17)], and another application of Sears's transformation [14, (III.15)] gives the form as in the statement of the proposition. The constant follows from bookkeeping. ∎

Recall the q-Racah polynomials, see Askey and Wilson [4] or [14, section 7.2], [22],

$$R_n(\nu(x);\alpha,\beta,\gamma,\delta;q) = {}_4\varphi_3\left(\begin{matrix} q^{-n}, \alpha\beta q^{n+1}, q^{-x}, \gamma\delta q^{x+1} \\ \alpha q,\ \beta\delta q,\ \gamma q \end{matrix}; q, q\right) \tag{4.15}$$

with $\nu(x) = q^{-x} + \gamma\delta q^{x+1}$, one of the lower parameters equals q^{-N}, $N \in \mathbb{Z}_+$, and $0 \le n \le N$. The ${}_4\varphi_3$-series in Proposition 4.9 can be written in terms of a q-Racah polynomial.

We can now rewrite (4.13) to arrive at the key result of this paper. For convenience we replace q^2 by q, $(a,b,c) = (q^{k_1}, q^{k_2}, q^{k_3})$, and we relabel w_1, j_{23}, and j_{12} by t, l, and n, and finally replace x_2, x_3 by x_1, x_2. We obtain the following q-analogue of Theorem 3.13 and Corollary 3.15.

THEOREM 4.10. *With $x_1 = \mu(w_1)$, $x_2 = \mu(w_2)$, $n, j \in \mathbb{Z}_+$, we have the convolution identity for the Askey–Wilson polynomials*

$$\begin{aligned}\sum_{l=0}^{n+j} b^{j-l} \begin{bmatrix} j+n \\ l \end{bmatrix}_q \frac{(b^2;q)_n (a^2b^2c^2q^{j+n-1};q)_l (c^2;q)_j}{(c^2, b^2c^2q^{l-1};q)_l (b^2c^2q^{2l};q)_{j+n-l}} \\ \times R_l(\nu(n); b^2/q, c^2/q, q^{-j-n-1}, a^2b^2q^{n+j-1}; q) \\ \times p_{j+n-l}(x_1; at, a/t, bcq^l s, bcq^l/s|q)\, p_l(x_2; bw_1, b/w_1, cs, c/s|q) \\ = p_n(x_1; at, a/t, bw_2, b/w_2|q)\, p_j(x_2; abq^n t, abq^n/t, cs, c/s|q),\end{aligned}$$

with the notation of (4.6), (4.15).

Remark 4.11. (i) Theorem 4.5 can be obtained as a special case of Theorem 4.10 by letting $c \to 0$.

(ii) Dunkl [10, section 3] obtains a special case of Theorem 4.10 in the same spirit as his proof [11] of Corollary 3.15; see Remark 3.16(ii). The symmetric group is now replaced by the general linear group over the field of q elements. This finite group of Lie type has the symmetric group as its Weyl group; see [10] for more information.

(iii) Theorem 4.10 leads to a kind of generating function for the q-Racah and q-Hahn polynomials. Choosing $w_1 = at$ and $w_2 = cs$ in Theorem 4.10 reduces all four Askey–Wilson polynomials to a single term. The remaining free parameters s and t appear only in the combination s/t. Replacing $bcs/(at)$ by u and (a^2, b^2, c^2) by (α, β, γ) gives

$$\begin{aligned}\sum_{l=0}^{n+j} \begin{bmatrix} j+n \\ l \end{bmatrix}_q \frac{(\alpha;q)_{j+n-l} (\beta;q)_n (\alpha\beta\gamma q^{j+n-1};q)_l}{(\alpha;q)_n (\beta\gamma q^{l-1};q)_l (\beta\gamma q^{2l};q)_{j+n-l}} u^{j-l} (\beta\gamma q^l/u;q)_{j+n-l} (u;q)_l \\ \times R_l(\nu(n); \beta/q, \gamma/q, q^{-j-n-1}, \alpha\beta q^{n+j-1}; q) = (\alpha q^n u;q)_j (\beta/u;q)_n.\end{aligned}$$

For $\gamma = 0$ we obtain a similar identity for q-Hahn polynomials:

$$\begin{aligned}\sum_{l=0}^{n+j} \begin{bmatrix} j+n \\ l \end{bmatrix}_q \frac{(\alpha;q)_{j+n-l} (\beta;q)_n}{(\alpha;q)_n} Q_n(q^{-l}; \beta/q, \alpha/q, n+j; q) u^{j-l} (u;q)_l \\ = (\alpha q^n u;q)_j (\beta/u;q)_n.\end{aligned}$$

(iv) Theorem 4.10 gives the connection coefficients for two sets of orthogonal polynomials with respect to the absolutely continuous measure

$$\frac{(w_1^{\pm 2}, w_2^{\pm 2};q)_\infty}{(taw_1^{\pm 1}, aw_1^{\pm 1}/t, csw_2^{\pm 1}, cw_2^{\pm 1}/s;q)_\infty} \frac{1}{(bw_1^{\pm 1} w_2^{\pm 1};q)_\infty} \frac{dw_1}{w_1} \frac{dw_2}{w_2}$$

on the torus $\mathbb{T}^2$ for $|t|^{-1} < |a| < |t|$, $|s|^{-1} < |c| < |s|$, and $|b| < 1$. Here all possible signs for $\pm$ have to be used. (Otherwise discrete masses at points and lines have to be added.) This weight function is invariant under simultaneously interchanging w_1 with w_2, t with s, and a with c. This transforms the orthogonal polynomials on the right-hand side of Theorem 4.10 to the ones occurring in the left-hand side.

In particular, note that for $s = t$, $a = c$, the weight function is invariant under the Weyl group for B_2, i.e., the group generated by $(w_1, w_2) \mapsto (w_2, w_1)$ and $(w_1, w_2) \mapsto (w_1, w_2^{-1})$. The corresponding Weyl group invariant orthogonal polynomials are

$$p_n(\mu(w_1); at, a/t, bw_2, b/w_2|q)\, p_j(\mu(w_2); abq^n t, abq^n/t, at, a/t|q) + p_n(\mu(w_2); at, a/t, bw_1, b/w_1|q)\, p_j(\mu(w_1); abq^n t, abq^n/t, at, a/t|q)$$

for $n \geq j \geq 0$. These orthogonal polynomials do not seem directly related to the Koornwinder–Macdonald orthogonal polynomials associated with root system BC_2, see [27], although the structure of the orthogonality measure is similar.

5. Linearization coefficients for Askey–Wilson polynomials. In the results of the previous section using $U_q(\mathfrak{su}(1,1))$, especially Theorems 4.5 and 4.10, we can use analytic continuation with respect to the parameters involved in finding similar identities but with the Al-Salam and Chihara polynomials and the Askey–Wilson polynomials replaced by the dual q-Krawtchouk polynomials and the q-Racah polynomials; cf. [15], [20], [28]. These identities can be obtained by the same procedure using $U_q(\mathfrak{su}(2))$ and its representation theory instead of using $U_q(\mathfrak{su}(1,1))$. In particular we can now give an interpretation for q-Racah polynomials as Clebsch-Gordan coefficients for $U_q(\mathfrak{su}(2))$; see [15, (23)]. In this case we also have some knowledge on the structure of the dual Hopf $*$-algebra and this can be used to obtain a linearization formula for a two-parameter family of Askey–Wilson polynomials. This is an application of the results of the previous section.

We first recall $U_q(\mathfrak{su}(2))$ and its representation theory; see, e.g., [9], [24], [28], [30]. The Hopf algebra structure on $U_q(\mathfrak{su}(2))$ is the same as the Hopf algebra structure on $U_q(\mathfrak{sl}(2,\mathbb{C}))$; cf. (4.1), (4.2). The $*$-operator making $U_q(\mathfrak{su}(2))$ into a Hopf $*$-algebra for $0 < q < 1$ is given by

$$A^* = A, \quad B^* = C, \quad C^* = B, \quad D^* = D.$$

There is precisely one irreducible unitary $U_q(\mathfrak{su}(2))$-module W_N of each dimension $N+1$ with highest weight vector v_+, i.e., $A\, v_+ = q^{N/2} v_+$, $B\, v_+ = 0$. The corresponding representation is denoted by t^N. With respect to the standard orthonormal basis e_n^N, $0 \leq n \leq N$, the action of the generators is given by $t^N(A)\, e_n^N = q^{n-N/2}\, e_d^N$ and

$$t^N(B)\, e_n^N = \frac{q^{(1-N)/2}}{1-q^2} \sqrt{(1-q^{2n+2})(1-q^{2N-2n})}\, e_{n+1}^N,$$

$$t^N(C)\, e_n^N = \frac{q^{(1-N)/2}}{1-q^2} \sqrt{(1-q^{2n})(1-q^{2N-2n+2})}\, e_{n-1}^N,$$

with the convention $e_{-1}^N = 0 = e_{N+1}^N$. So e_N^N is the highest weight vector. The representation t^N, considered as a representation of $U_q(\mathfrak{sl}(2,\mathbb{C}))$, can be obtained from the discrete series representation π_k of (4.3) by formally replacing k by $-N/2$.

The Clebsch–Gordan decomposition holds; as unitary $U_q(\mathfrak{su}(2))$-modules

$$W_{N_1} \otimes W_{N_2} = \bigoplus_{j=0}^{\min(N_1,N_2)} W_{N_1+N_2-2j}.$$

The matrix coefficients of the intertwining operator give the Clebsch–Gordan coefficients;

$$e_n^N = \sum_{n_1,n_2} C_{n_1,n_2,n}^{N_1,N_2,N}\ e_{n_1}^{N_1} \otimes e_{n_2}^{N_2}. \tag{5.1}$$

Of course, the Clebsch–Gordan coefficient is zero if $N \neq N_1 + N_2 - 2j$ for $0 \leq j \leq \min(N_1, N_2)$. By considering the action of A on both sides we see that the Clebsch–Gordan coefficient is zero unless $n_1 + n_2 = n + j$, so the sum is actually a single sum. The Clebsch–Gordan coefficients are normalized by $\langle e_0^N, e_j^{N_1} \otimes e_0^{N_2} \rangle > 0$ if $N = N_1 + N_2 - 2j$, $0 \leq j \leq \min(N_1, N_2)$.

We are particularly interested in the element

$$X_p = q^{1/2}B + q^{-1/2}C - \frac{p^{1/2} - p^{-1/2}}{q - q^{-1}}(A - D) \in U_q(\mathfrak{sl}(2, \mathbb{C})), \qquad p > 0.$$

Then X_pA is self-adjoint and $\Delta(X_pA) = A^2 \otimes X_pA + X_pA \otimes 1$. Koornwinder [28] has shown that in each module W_N the action of X_pA is completely diagonalizable. To formulate this result we introduce the orthonormal dual q-Krawtchouk polynomials; for $a > 0$,

$$r_n(q^{-x} - q^{x-N}/a; a, N; q) = (-1)^n a^{n/2} q^{n(n-1)/4} \left[\begin{array}{c} N \\ n \end{array} \right]_q^{1/2} R_n(q^{-x} - q^{x-N}/a; a, N; q)$$

for $N \in \mathbb{Z}_+$ and $0 \leq x, n \leq N$. The dual q-Krawtchouk polynomials are special cases of the q-Racah polynomials (4.15) and are defined by

$$R_n(q^{-x} - q^{x-N}/a; a, N; q) = {}_3\varphi_2 \left(\begin{array}{c} q^{-n}, q^{-x}, -q^{x-N}/a \\ q^{-N}, 0 \end{array} ; q, q \right).$$

The corresponding three-term recurrence relation is

$$\begin{aligned} (q^{-x} - q^{x-N}/a)\, r_n &= A_n\, r_{n+1} + q^{n-N}(1 - a^{-1})\, r_n + A_{n-1}\, r_{n-1}, \\ A_n &= a^{-1/2} q^{-N+n/2} \sqrt{(1 - q^{n+1})(1 - q^{N-n})} \end{aligned} \tag{5.2}$$

for $0 \leq x, n \leq N$, and $r_n = r_n(q^{-x} - q^{x-N}/a; a, N; q)$.

PROPOSITION 5.1 (see [28]). *There exists an orthogonal basis $\phi_f^N = \phi_f^N(p)$, $0 \leq f \leq N$, of W_N of eigenvectors of $t^N(X_pA)$ for the eigenvalue*

$$\lambda_f^N(p) = \frac{p^{1/2} q^{N-2f} - p^{-1/2} q^{2f-N} + p^{-1/2} - p^{1/2}}{q^{-1} - q}.$$

Moreover, $\phi_f^N(p) = \sum_{n=0}^N r_n(q^{-2f} - q^{2f-2N}/p; p, N; q^2)\, e_d^N$.

Proposition 5.1 is the analogue of Proposition 4.1, and we could have formulated it in a similar fashion using the finite discrete orthogonality measure for the dual q-Krawtchouk polynomials. Actually, replacing $e^{i\theta}$, k in $S_n(\cos\theta; q^{2k}s, q^{2k}/s|q^2)$ by q^{-2f+N}/s, $-N/2$ and next taking $s^2 = -p^{-1}$ gives $r_n(q^{-2f} - q^{2f-2N}/p; p, N; q^2)$. The analogue of Proposition 4.2 is the following.

PROPOSITION 5.2. *For $0 \leq f_1 \leq N_1$, $0 \leq f_2 \leq N_2$, define in $W_{N_1} \otimes W_{N_2}$ the vector*

$$\phi_{f_1,f_2}^{N_1,N_2} = \sum_{n_1=0}^{N_1} r_{n_1}(q^{-2f_1} - q^{2f_1-2N_1-2N_2+4f_2}/p; pq^{2N_2-4f_2}, N_1; q^2)\ e_{n_1}^{N_1} \otimes \phi_{f_2}^{N_2};$$

then $t^{N_1}\otimes t^{N_2}\big(\Delta(X_pA)\big)\,\phi^{N_1,N_2}_{f_1,f_2}=\lambda^{N_1+N_2}_{f_1+f_2}(p)\,\phi^{N_1,N_2}_{f_1,f_2}$. *Moreover,* $\phi^{N_1,N_2}_{f_1,f_2}$, $0\le f_1\le N_1$, $0\le f_2\le N_2$, *constitutes an orthogonal basis of* $W_{N_1}\otimes W_{N_2}$ *of eigenvectors of* $\Delta(X_pA)$.

Proof. From $\Delta(X_pA)=A^2\otimes X_pA+X_pA\otimes 1$ it follows that there is an eigenvector of the form $\sum_{n_1=0}^{N_1}p_{n_1}\,e^{N_1}_{n_1}\otimes\phi^{N_2}_{f_2}$ by solving a three-term recurrence relation for the p_{n_1}. Now (5.2) can be used to solve this.

There are $(N_1+1)(N_2+1)$ eigenvectors in $W_{N_1}\otimes W_{N_2}$, and $\langle\phi^{N_1,N_2}_{f_1,f_2},\phi^{N_1,N_2}_{g_1,g_2}\rangle$ equals zero if $f_2\ne g_2$ by Proposition 5.1. It also equals zero if $f_1+f_2\ne g_1+g_2$ by the self-adjointness of X_pA. □

The result of Proposition 5.2 can be obtained from Proposition 4.2 by substituting k_1, k_2 by $-N_1/2$, $-N_2/2$ and w_1, w_2 by $q^{N_1+N_2-2f_1-2f_2}/s$, $q^{N_2-2f_2}/s$ and s^2 by $-p^{-1}$.

PROPOSITION 5.3. *For* $N=N_1+N_2-2j$, $0\le j\le\min(N_1,N_2)$, *we have*

$$\langle\phi^{N_1,N_2}_{f_1,f_2},e^N_n\rangle=r_n(q^{2j-2f_1-2f_2}-q^{2f_1+2f_2-2j-2N}/p;p,N;q^2)\langle\phi^{N_1,N_2}_{f_1,f_2},e^N_0\rangle$$

if $0\le f_1+f_2-j\le N$, *and* $\langle\phi^{N_1,N_2}_{f_1,f_2},e^N_n\rangle=0$ *otherwise. If nonzero, then*

$$\langle\phi^{N_1,N_2}_{f_1,f_2},e^N_0\rangle=\begin{bmatrix}N_2\\ j\end{bmatrix}_{q^2}^{1/2}\frac{p^{j/2}q^{j(2N_1+N_2)}q^{-3j(j-1)/2}}{\sqrt{(q^{2N_1},q^{2N_1+2N_2-2j+2};q^{-2})_j}}(-p^{-1}q^{2f_1+2f_2-2N_1-2N_2};q^2)_j$$
$$\times(q^{-2f_1-2f_2};q^2)_j\,{}_4\varphi_3\left(\begin{matrix}q^{-2j},q^{2j-2-2N_1-2N_2},q^{-2f_2},-p^{-1}q^{2f_2-2N_2}\\ q^{-2N_2},q^{-2f_1-2f_2},-p^{-1}q^{2f_1+2f_2-2N_1-2N_2}\end{matrix};q^2,q^2\right).$$

Note that the ${}_4\varphi_3$-series is balanced and can be written in terms of q-Racah polynomials (4.15). The ${}_4\varphi_3$-series in Proposition 5.3 equals

$$R_j(q^{-2f_2}-p^{-1}q^{2f_2-2N_2};q^{-2N_2-2},q^{-2N_1-2},q^{-2f_1-2f_2-2},-p^{-1}q^{2f_1+2f_2-2N_2};q^2).$$

The proof of Proposition 5.3 is similar to the proof of Proposition 4.3. Proposition 5.3 can also be obtained from Proposition 4.3 using the substitutions as indicated earlier. It can also be obtained by using $e^N_0=\sum C^{N_1,N_2,N}_{n_1,n_2,0}\,e^{N_1}_{n_1}\otimes e^{N_2}_{n_2}$, Propositions 5.2 and 5.1, and the explicit value for the Clebsch–Gordan coefficients for $n=0$, $N=N_1+N_2-2j$,

$$C^{N_1,N_2,N}_{n_1,n_2,0}=(-1)^{n_2}q^{n_2(N_2-j-1)}\sqrt{\frac{(q^{2n_1+2};q^2)_{n_2}(q^{2N_1-2n_1};q^{-2})_{n_2}(q^{2N_2};q^{-2})_j}{(q^2;q^2)_{n_2}(q^{2N_2};q^{-2})_{n_2}(q^{2N_1+N_2-2j+2};q^2)_j}}.$$

See, e.g., [35, section 14.3], but this simple case can also be derived as follows. Apply C to both sides of (5.1) to obtain a three-term recurrence for the Clebsch–Gordan coefficients, which reduces to a two-term recurrence for $n=0$. This can easily be solved, with the initial condition following from the unitarity and the normalization; see [34] for a similar derivation. Then we have a sum involving the product of two dual q-Krawtchouk polynomials. Upon inserting the series representation we obtain a triple sum, and after interchanging summations we can use the q-binomial theorem and the q-Chu–Vandermonde sum, see [14], to obtain a single ${}_4\varphi_3$-series.

Remark 5.4. With Proposition 5.3 at hand it is straightforward to calculate the $U_q(\mathfrak{su}(2))$-counterparts of Theorems 4.5 and 4.10. The Al-Salam and Chihara, respectively, Askey–Wilson, polynomials have to be replaced by dual q-Krawtchouk, respectively, q-Racah, polynomials. The result can also be obtained from Theorems 4.5 and 4.10 by substitution as indicated earlier, so we do not give them explicitly. These formulas give an alternative for the formulas of Groza and Kachurik [18].

In the representation space $W_{N_1} \otimes W_{N_2}$ we have two bases of eigenvectors for the action of $X_p A$, namely ϕ_f^N and $\phi_{f_1,f_2}^{N_1,N_2}$, and the corresponding Clebsch–Gordan coefficients are given by Proposition 5.3, since, with $N = N_1 + N_2 - 2j$,

$$\begin{aligned}\phi_{f_1,f_2}^{N_1,N_2} &= \sum_{j=0}^{\min(N_1,N_2)} \sum_{n=0}^{N} \langle \phi_{f_1,f_2}^{N_1,N_2}, e_n^N \rangle \, e_n^N \\ &= \sum_{j=0}^{\min(N_1,N_2)} \langle \phi_{f_1,f_2}^{N_1,N_2}, e_0^N \rangle \sum_{n=0}^{N} r_n(q^{2j-2f_1-2f_2} - q^{2f_1+2f_2-2j-2N}/p; p, N; q^2)\, e_n^N \\ &= \sum_{j=0}^{\min(N_1,N_2)} \langle \phi_{f_1,f_2}^{N_1,N_2}, e_0^N \rangle \, \phi_{f_1+f_2-j}^N.\end{aligned}$$

Here we use the convention that $\phi_f^N = 0$ for $f > N$ or $f < 0$. So introducing the notation

$$\phi_{f_1,f_2}^{N_1,N_2} = \sum_{f,j} C_{f_1,f_2,f}^{N_1,N_2,N}(p)\ \phi_f^N, \tag{5.3}$$

we see that the Clebsch–Gordan coefficients are zero unless $f_1 + f_2 = f + j$, and then $C_{f_1,f_2,f}^{N_1,N_2,N}(p) = \langle \phi_{f_1,f_2}^{N_1,N_2}, e_0^N \rangle$. So, by Proposition 5.3 we have proved that the q-Racah polynomials occur as Clebsch–Gordan coefficients for $U_q(\mathfrak{su}(2))$. This is due to Granovskii and Zhedanov [15].

Using (5.3) in a special case we can obtain the linearization coefficients for the two parameter family of Askey–Wilson polynomials occurring as spherical functions on the quantum $SU(2)$ group; cf. [28]. We consider odd-dimensional representations; $N_1 = 2l_1$, $N_2 = 2l_2$, $l_1, l_2 \in \mathbb{Z}_+$. Then the kernel of $t^{2l_1}(X_p A)$ is one dimensional and spanned by $\phi_{l_1}^{2l_1}(p)$. Moreover, $\phi_{l_1,l_1}^{2l_1,2l_2}(p) = \phi_{l_1}^{2l_1}(p) \otimes \phi_{l_2}^{2l_2}(p)$. Next we consider matrix elements as linear functionals on $U_q(\mathfrak{su}(2))$ to find

$$\begin{aligned}&\sum_{(X)} \langle t^{2l_1}(X_{(1)})\phi_{l_1}^{2l_1}(p), \phi_{l_1}^{2l_1}(r)\rangle \, \langle t^{2l_2}(X_{(2)})\phi_{l_2}^{2l_2}(p), \phi_{l_2}^{2l_2}(r)\rangle \\ &= \langle t^{2l_1} \otimes t^{2l_2}(\Delta(X))\phi_{l_1,l_1}^{2l_1,2l_2}(p), \phi_{l_1,l_1}^{2l_1,2l_2}(r)\rangle \\ &= \sum_{l=|l_1-l_2|}^{l_1+l_2} C_{l_1,l_2,l}^{2l_1,2l_2,2l}(p) C_{l_1,l_2,l}^{2l_1,2l_2,2l}(r) \langle t^{2l}(X)\phi_l^{2l}(p), \phi_l^{2l}(r)\rangle,\end{aligned} \tag{5.4}$$

where $r > 0$ is another parameter and $\Delta(X) = \sum_{(X)} X_{(1)} \otimes X_{(2)}$.

The dual Hopf $*$-algebra $A_q(SU(2))$ generated by the matrix elements of the representations t^N, $N \in \mathbb{Z}_+$, of $U_q(\mathfrak{su}(2))$, is known in terms of generators and relations, cf. [9], [24], [28], [30]. Koornwinder [28] has given an explicit expression for the element in $A_q(SU(2))$ corresponding to the linear functionals considered in (5.4), which can be considered as a zonal spherical function on the quantum $SU(2)$ group.

$$\langle t^{2l}(X)\phi_l^{2l}(p), \phi_l^{2l}(r)\rangle = \frac{q^{-l}}{(q^{2l+2};q^2)_l} \left\langle X, p_l\left(\rho; q\sqrt{\frac{p}{r}}, q\sqrt{\frac{r}{p}}, \frac{-q}{\sqrt{pr}}, -q\sqrt{pr}|q^2\right)\right\rangle, \tag{5.5}$$

where $\rho \in A_q(SU(2))$ is some fixed simple element, which is, up to an affine scaling, the linear functional $X \mapsto \langle t^2(X)\phi_1^2(p), \phi_1^2(r)\rangle$, and the last $\langle \cdot, \cdot \rangle$ denotes the duality

between $U_q(\mathfrak{su}(2))$ and $A_q(SU(2))$. Since $A_q(SU(2))$ is the dual Hopf $*$-algebra, the left-hand side of (5.4) corresponds to the multiplication of the two linear functionals. So (5.4) leads to the following identity in $A_q(SU(2))$:

$$p_{l_1}(\rho)\,p_{l_2}(\rho) = \sum_{l=|l_1-l_2|}^{l_1+l_2} q^{l_1+l_2-l}\frac{(q^{2l_1+2};q^2)_{l_1}(q^{2l_2+2};q^2)_{l_2}}{(q^{2l+2};q^2)_l}C_{l_1,l_2,l}^{2l_1,2l_2,2l}(p)C_{l_1,l_2,l}^{2l_1,2l_2,2l}(r)\,p_l(\rho)$$

with $p_l(\cdot) = p_l(\cdot; q\sqrt{\frac{p}{r}}, q\sqrt{\frac{r}{p}}, \frac{-q}{\sqrt{pr}}, -q\sqrt{pr}|q^2)$.

The only information on $A_q(SU(2))$ needed is the existence of a family of one-dimensional representations sending ρ to $\cos\theta$. Thus, applying the one-dimensional representations of $A_q(SU(2))$ and using Proposition 5.3 proves the following linearization coefficient formula.

THEOREM 5.5. *Let $p_l(x) = p_l(x; q\sqrt{\frac{p}{r}}, q\sqrt{\frac{r}{p}}, \frac{-q}{\sqrt{pr}}, -q\sqrt{pr}|q^2)$, $p, r > 0$, be defined in terms of Askey–Wilson polynomials* (4.6)*. Then the coefficients in the linearization formula*

$$p_{l_1}(x)\,p_{l_2}(x) = \sum_{j=0}^{2\min(l_1,l_2)} c_j\,p_{l_1+l_2-j}(x)$$

are given by a product of two balanced terminating ${}_4\varphi_3$-series;

$$\begin{aligned} c_j = {} & q^{-j(j-1)}q^{j+4jl_1}(q^{2l_1+2};q^2)_{l_1-j}(q^{2l_2+2};q^2)_{l_2}\begin{bmatrix} 2l_2 \\ j \end{bmatrix}_{q^2} \\ & \times\frac{(q^{2l_1+2l_2};q^{-2})_j}{(q^{2l_1+2l_2+2};q^2)_{l_1+l_2-j}}\frac{1-q^{4l_1+4l_2-4j+2}}{1-q^{4l_1+4l_2-2j+2}} \\ & \times\, p^{j/2}(-p^{-1}q^{-2l_1-2l_2};q^2)_j\ {}_4\varphi_3\left(\begin{matrix} q^{-2j}, q^{-2l_2}, q^{2j-2-4l_1-4l_2}, -p^{-1}q^{-2l_2} \\ q^{-4l_2}, q^{-2l_1-2l_2}, -p^{-1}q^{-2l_1-2l_2} \end{matrix};q^2,q^2\right) \\ & \times\, r^{j/2}(-r^{-1}q^{-2l_1-2l_2};q^2)_j\ {}_4\varphi_3\left(\begin{matrix} q^{-2j}, q^{-2l_2}, q^{2j-2-4l_1-4l_2}, -r^{-1}q^{-2l_2} \\ q^{-4l_2}, q^{-2l_1-2l_2}, -r^{-1}q^{-2l_1-2l_2} \end{matrix};q^2,q^2\right). \end{aligned}$$

Remark 5.6. (i) In particular, for $p = r$ the linearization coefficients are positive. This can already be observed without the explicit knowledge of the linearization coefficients; see [24, section 8.3], [26, section 7].

(ii) For $p = r = 1$ the Askey–Wilson polynomials $p_l(x; q, q, -q, -q|q^2)$ are the continuous q-Legendre polynomials $C_l(x; q^2|q^4)$; see [5, section 4]. This is a special case of the continuous q-ultraspherical polynomials introduced by Rogers at the end of the 19th century. Rogers calculated the linearization coefficients for the continuous q-ultraspherical polynomials, see, e.g., [5, section 4], [14, section 8.5], and we can go from Theorem 5.5 to the special case of Rogers's result by using Andrew's summation formula; see [14, (II.17)].

(iii) Not only the zonal spherical elements on the quantum $SU(2)$ group are known in terms of Askey–Wilson polynomials in the generators of the dual Hopf algebra as in (5.5), but any matrix coefficient of t^N of the form $\langle t^N(X)\phi_f^N(p), \phi_g^N(r)\rangle$ can be written in terms of Askey–Wilson polynomials on the dual Hopf algebra. This is due to Noumi and Mimachi [31], [32], [30]; see [24] for a proof. The method described here to obtain a linearization formula only works for the zonal spherical functions, as in the group case.

Acknowledgment. We thank Tom Koornwinder for useful comments. We also thank Charles Dunkl for useful comments on the previous version and pointing out [10] and Alexei Zhedanov for pointing out [15], [17]. The first author thanks the Universiteit Gent for its hospitality.

REFERENCES

[1] W. A. Al-Salam, *Characterization theorems for orthogonal polynomials*, in Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., NATO ASI series C 294, Kluwer, 1990, pp. 1–24.

[2] W. A. Al-Salam and T. S. Chihara, *Convolutions of orthonormal polynomials*, SIAM J. Math. Anal., 7 (1976), pp. 16–28.

[3] R. Askey, *Continuous Hahn polynomials*, J. Math. Phys. A: Math. Gen., 18 (1985), pp. L1017–L1019.

[4] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or $6-j$ symbols*, SIAM J. Math. Anal., 10 (1979), pp. 1008–1016.

[5] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 54 (1985), pp. 1–55.

[6] N. M. Atakishiyev and S. K. Suslov, *The Hahn and Meixner polynomials of imaginary argument and some of their applications*, J. Math. Phys. A: Math. Gen., 18 (1985), pp. 1583–1596.

[7] J. M. Berezanskiĭ, *Expansions in Eigenfunctions of Selfadjoint Operators*, Transl. Math. Monogr. 17, Amer. Math. Soc., 1968.

[8] I. M. Burban and A. U. Klimyk, *Representations of the quantum algebra $U_q(su_{1,1})$*, J. Phys. A: Math. Gen., 26 (1993), pp. 2139–2151.

[9] V. Chari and A. Pressley, *A Guide to Quantum Groups*, Cambridge University Press, Cambridge, UK, 1994.

[10] C. F. Dunkl, *Orthogonal polynomials in two variables of q-Hahn and q-Jacobi type*, SIAM J. Alg. Disc. Meth., 1 (1980), pp. 137–151.

[11] C. F. Dunkl, *A difference equation and Hahn polynomials in two variables*, Pacific J. Math., 92 (1981), pp. 57–71.

[12] C. F. Dunkl, *Orthogonal polynomials with symmetry of order three*, Canad. J. Math., 36 (1984), pp. 685–717.

[13] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions*, Vol. 2, McGraw–Hill, New York, 1953.

[14] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, UK, 1990.

[15] Y. I. Granovskii and A. S. Zhedanov, *'Twisted' Clebsch-Gordan coefficients for $SU_q(2)$*, J. Phys A: Math. Gen., 25 (1992), pp. L1029–L1032.

[16] Y. I. Granovskii and A. S. Zhedanov, *New construction of $3nj$-symbols*, J. Phys A: Math. Gen., 26 (1993), pp. 4339–4344.

[17] Y. I. Granovskii and A. S. Zhedanov, *Hidden symmetry of the Racah and Clebsch-Gordan problems for the quantum algebra $sl_q(2)$*, J. Group Theory Phys., 1 (1993), pp. 161–171.

[18] V. A. Groza and I. I. Kachurik, *Addition and product theorems for Krawtchouk, Hahn and Racah q-polynomials*, Dokl. Akad. Nauk Ukraine SSR Ser. A, 89 (1990), pp. 3–6 (in Russian).

[19] E. G. Kalnins, H. L. Manocha, and W. Miller Jr., *Models of q-algebra representations: Tensor products of special unitary and oscillator algebras*, J. Math. Phys., 33 (1992), pp. 2365–2383.

[20] A. N. Kirillov and N. Y. Reshetikhin, *Representations of the algebra $U_q(\mathfrak{sl}_2)$, q-orthogonal polynomials and invariants of links*, in Infinite-dimensional Lie Algebras and Groups, V.G. Kac, ed., World Scientific, Singapore, 1989, pp. 285–339.

[21] A. U. Klimyk and I. I. Kachurik, *Spectra, eigenvectors and overlap functions for representation operators of q-deformed algebras*, Comm. Math. Phys., 175 (1996), pp. 89–111.

[22] R. Koekoek and R. F. Swarttouw, *The Askey-Scheme of Hypergeometric Orthogonal Polynomials and its q-analogue*, Report 94-05, Technical University Delft, Delft, the Netherlands 1994; also available online from ftp.twi.tudelft.nl in directory /pub/publications/tech-reports.

[23] H. T. Koelink, *On Jacobi and continuous Hahn polynomials*, Proc. Amer. Math. Soc., 124 (1996), pp. 887–898.

[24] H. T. Koelink, *Askey-Wilson polynomials and the quantum $SU(2)$ group: Survey and applications*, Acta Appl. Math., 44 (1996), pp. 295–352.

[25] T. H. Koornwinder, *Meixner-Pollaczek polynomials and the Heisenberg algebra*, J. Math. Phys., 30 (1989), pp. 767–769.
[26] T. H. Koornwinder, *Positive convolution structures associated with quantum groups*, in Probability Measures on Groups X, H. Heyer, ed., Plenum Press, New York, 1991, pp. 249–268.
[27] T. H. Koornwinder, *Askey-Wilson polynomials for root systems of type BC*, Contemp. Math., 138 (1992), pp. 189–204.
[28] T. H. Koornwinder, *Askey-Wilson polynomials as zonal spherical functions on the $SU(2)$ quantum group*, SIAM J. Math. Anal., 24 (1993), pp. 795–813.
[29] D. R. Masson and J. Repka, *Spectral theory of Jacobi matrices in $\ell^2(\mathbb{Z})$ and the $su(1,1)$ Lie algebra*, SIAM J. Math. Anal., 22 (1991), pp. 1133–1146.
[30] M. Noumi, *Quantum groups and q-orthogonal polynomials. Towards a realization of Askey-Wilson polynomials on $SU_q(2)$*, in Special Functions, M. Kashiwara and T. Miwa, eds., ICM-90 Satellite Conference Proceedings, Springer, New York, 1991, pp. 260–288.
[31] M. Noumi and K. Mimachi, *Askey-Wilson polynomials and the quantum group $SU_q(2)$*, Proc. Japan Acad. Ser. A, 66 (1990), pp. 146–149.
[32] M. Noumi and K. Mimachi, *Askey-Wilson polynomials as spherical functions on $SU_q(2)$*, in Quantum Groups, LNM 1510, P. P. Kulish, ed., Springer, New York, 1992, pp. 98–103.
[33] G. Szegő, *Orthogonal Polynomials*, 4th ed., American Mathematical Society, Providence, RI, 1975.
[34] J. Van der Jeugt, *Coupling coefficients for Lie algebra representations and addition formulas for special functions*, J. Math. Phys., 38 (1997), pp. 2728–2740.
[35] N. J. Vilenkin and A. U. Klimyk, *Representation of Lie Groups and Special Functions*, Vols. 1–3, Kluwer Academic Publishers, Norwell, MA, 1991–3.
[36] J. A. Wilson, *Some hypergeometric orthogonal polynomials*, SIAM J. Math. Anal., 11 (1980), pp. 690–701.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 823–848, July 1998

001

QUASICONVEXIFICATION IN $W^{1,1}$ AND OPTIMAL JUMP MICROSTRUCTURE IN BV RELAXATION*

CHRISTOPHER J. LARSEN†

Abstract. An integral representation for the relaxation in $BV(\Omega;\mathbb{R}^p)$ of the functional

$$u \mapsto \int_\Omega W(\nabla u(x))dx + \mathcal{H}^{N-1}(S(u))$$

with respect to BV weak $*$ convergence is obtained. The bulk term in the integral representation reduces to the quasiconvexification of W, and we describe optimal behavior of approximating sequences along $S(u)$, for scalar valued u.

Key words. quasiconvex, lower semicontinuous, microstructure, bounded variation, relaxation

AMS subject classifications. 49K10, 49N60, 49Q15, 73V25

PII. S0036141095295991

1. Introduction.

In this paper, we study the lower semicontinuity of the functional

$$E(u) := \int_\Omega W(\nabla u)dx + \mathcal{H}^{N-1}(S(u)),$$

where $\Omega \subset \mathbb{R}^N$ is open and bounded, $u \in BV(\Omega;\mathbb{R}^p)$, ∇u is the Radon–Nikodym derivative of Du with respect to $\mathcal{L}^N$, $S(u)$, the "jump set" of u, is the complement of the set of Lebesgue points of u, and $\mathcal{H}^{N-1}$ is the $N-1$-dimensional Hausdorff measure.

This functional can be viewed as modeling materials with fracture: if u represents a deformation, the energy density W penalizes elastic deformation, and $\mathcal{H}^{N-1}(S(u))$ penalizes fracture by the size of the fracture site.

As is typical in the study of such functionals, we consider the relaxed energy I, i.e., the lower semicontinuous envelope in L^1 of the original energy E. By studying why these energies might not agree, and in particular, why corresponding energy densities might not agree, one is led to investigate the local behavior of minimizing sequences and the onset of microstructure, i.e., the development in minimizing sequences of finer and finer oscillations of their gradients, jump sets, or a combination of the two. By optimal "jump microstructure," we mean optimal oscillations involving both the gradients and jump sets of a sequence of functions that occur along the jump set of the limit of that sequence.

The jump set $S(u)$ of any BV function u is known to be $N-1$-rectifiable, and so it has a normal, ν, $\mathcal{H}^{N-1}$ almost everywhere. Furthermore, we have the decomposition

$$Du = \nabla u\mathcal{L}^N + [u] \otimes \nu\mathcal{H}^{N-1}\lfloor S(u) + C(u),$$

*Received by the editors December 11, 1995; accepted for publication (in revised form) April 1, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/29599.html

†Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609 (cjlarsen@wpi.edu). This research was partially supported by the Army Research Office and the National Science Foundation, the Center for Nonlinear Analysis (Carnegie Mellon University), and National Science Foundation grant DMS-9201215. This research was conducted while the author was at Carnegie Mellon University and at the Army Research Laboratory in Aberdeen, MD.

where $[u]$ is the jump in u, i.e., $u^+ - u^-$, where u^+ and u^- are the traces of u on either side of $S(u)$ (see, e.g., [10] and [17]), and $C(u)$ is the so-called *Cantor part*, which is singular with respect to the first two measures in the decomposition. If $C(u) = 0$, we say $u \in SBV(\Omega; \mathbb{R}^p)$, the space of *special functions of bounded variation* introduced in [9]. Since functions with $Du = C(u)$ are dense in L^1, if $\min W = W(0)$ then the relaxation of E would reduce to $\mathcal{L}^N(\Omega)W(0)$. We avoid this pathology by only considering sequences in SBV, which is equivalent to relaxing $E(\cdot)+\infty|C(\cdot)|(\Omega)$. This corresponds to allowing macroscopic states with Cantor part, but not microscopic states.

In [3], Ambrosio analyzed the energy functional on SBV given by

$$E(u) = \int_\Omega W(x, u, \nabla u)dx + \int_{S(u)} \phi(u^+, u^-, \nu)d\mathcal{H}^{N-1}(x),$$

under the hypotheses that W is Carathéodory and has superlinear growth in ∇u, and under conditions on ϕ that, in particular, allow ϕ to be any positive constant. A result is that if W is quasiconvex, and if certain assumptions on ϕ are met, then E is L^1_{loc} lower semicontinuous in SBV. The analysis of the relaxation in BV of the model where $\phi \equiv 1$ and W has superlinear growth was carried out in [11].

In this paper, we assume that W has linear growth and we take $\phi \equiv 1$. Physically, this last assumption corresponds to weighing jumps, or cracks, only by the size of the jump set $S(u)$, with no dependence on the orientation ν or size of the jump $[u]$. The linear growth of W allows interaction along the jump set: approximating sequences of u need not jump at $S(u)$, but might have much more complicated behavior consisting of combinations of smooth growth (gradients) and jumps. If W has superlinear growth, it is energetically impossible for optimal approximating sequences to develop concentrations of their gradients along the jump set of u, and so their behavior there is much simpler. The study of the case where W and ϕ both have linear growth was undertaken in [6] (see also [7]).

The main new contributions in this paper are the expressions for QW (the quasiconvexification of W) in section 3, a new method for showing the upper bound inequality for the jump density in section 4, Lemma 5.1 in section 5 which allows us to blow up in such a way that the rescaled variation measures do not lose mass as they converge weakly $*$, and finally a method for finding the optimal jump microstructure for scalar valued functions in section 6. This last result allows us to exhibit the optimal behavior of approximating sequences along $S(u)$. The method is applicable not only when the jump energy density is a constant, but also when it depends in a positive homogenous degree one way on the jump (see Theorem 6.5 and Remark 6.6).

This paper is organized as follows: in section 2 we discuss preliminaries and state the relaxation theorem, the essence of which is the integral representation

$$I(u) = \int_\Omega QW(\nabla u)dx + \int_{S(u)} h([u], \nu)d\mathcal{H}^{N-1} + \int_\Omega (QW)^\infty(dC(u)), \tag{1.1}$$

where QW is the quasiconvexification of W (see section 2), and h and $(QW)^\infty$ are defined in section 2.

In section 3 we show that, although QW is defined in terms of sequences in Sobolev spaces, there are equivalent definitions in terms of certain sequences in BV. An analogous lower semicontinuity result for superlinear W was obtained by Ambrosio in [3], Theorem 3.3.

In section 4, we show that $I(u) \leq$ the right-hand side of (1.1). We first prove that $I(u,\cdot)$ is a finite Borel regular measure, absolutely continuous with respect to $\mathcal{L}^N + |Du|$. This follows largely from [12]. The remaining issue in this section is the upper bound for $I(u,\cdot)\lfloor S(u)$, for which we introduce a new argument. There is some difficulty with this step because $\mathcal{H}^{N-1}\lfloor S(u)$ is, in general, not a Radon measure, and so taking derivatives with respect to $\mathcal{H}^{N-1}\lfloor S(u)$ is not possible. The usual method for showing upper bound inequalities for jump densities is based on [4] and [5], and involves approximating jump sets with boundaries of sets with finite perimeter. The technique here is based on looking at the intersection of the jump set with certain sets of finite perimeter. We consider level sets E_t of the components of u, such that E_t has finite perimeter and $|D_j u| := |Du|\lfloor S(u)$ concentrates on $S(u) \cap \partial_* E_t$ as we blow-up. We then see that the analysis on $S(u) \cap \partial_* E_t$ is much easier than on $S(u)$. The rest follows from constructing functions in a reasonable way, and by using a suitable covering argument.

Section 5 deals with the proof of the lower bound inequality $I(u) \geq$ the right-hand side of (1.1), which is a modified version of the corresponding argument in [6]. The changes include a lemma that allows us to choose the rescaling factors so that as the rescaled variation measures converge weakly $*$ on a cube, they do not lose any mass (see Lemma 5.1).

In section 6 we find optimal microstructure along the jump set of u, for scalar valued u. The proof relies on a coarea formula and an application of Jensen's inequality on boundaries of level sets. It turns out that the proof may be easily extended to the case where the energy density on jumps is a positive homogeneous degree one function of $[u]\nu$, and also when the energy density on jumps is just a function of the normal to $S(u)$.

2. Preliminaries and the relaxation theorem. We consider a bounded, open set $\Omega \subset \mathbb{R}^N$, and we define the Sobolev spaces $W^{1,1}(\Omega)$ and $W^{1,\infty}(\Omega)$, and the space of functions of bounded variation $BV(\Omega)$ in the usual way (see, e.g., [10] and [17]). We denote by ρ_m, or alternatively ρ_ϵ, the standard mollifier, and for $E \subset \Omega$, χ_E stands for the characteristic function of E. Given two sets A and B, we define the *symmetric difference* $A \triangle B := (A \backslash B) \cup (B \backslash A)$.

We say that a set $E \subset \Omega$ has *finite perimeter* in Ω if $\chi_E \in BV(\Omega)$. For such an E, the *measure theoretic boundary* in Ω, $\partial_* E$, is defined as

$$\left\{ x \in \Omega : \limsup_{\delta \to 0^+} \frac{\mathcal{L}^N(B(x,\delta) \cap E)}{\mathcal{L}^N(B(x,\delta))} > 0 \text{ and } \limsup_{\delta \to 0^+} \frac{\mathcal{L}^N(B(x,\delta) \backslash E)}{\mathcal{L}^N(B(x,\delta))} > 0 \right\}, \tag{2.1}$$

where $B(x,\delta)$ is the closed ball in $\mathbb{R}^N$ centered at x with radius δ. We denote by $\nu_E(x)$ the *measure theoretic normal* to E at $x \in \partial_* E$ (for properties of this normal, see [10] or [17]). The *reduced boundary* $\partial^* E$ is the set of $x \in \partial_* E$ such that x is a Lebesgue point for ν_E, with respect to the Radon measure $\mathcal{H}^{N-1}\lfloor \partial_* E$. Given a set E of finite perimeter, we define on $\partial_* E$ the following:

$$H(x) := \{ y \in \mathbb{R}^N : \nu_E(x) \cdot (y - x) = 0 \},$$
$$H^+(x) := \{ y \in \mathbb{R}^N : \nu_E(x) \cdot (y - x) \geq 0 \},$$

and

$$H^-(x) := \{ y \in \mathbb{R}^N : \nu_E(x) \cdot (y - x) \leq 0 \}.$$

For $u \in BV(\Omega;\mathbb{R}^p)$, we write $Du = D_{ac}u + D_s u$, where $D_{ac}u$ and $D_s u$ stand for, respectively, the absolutely continuous and singular part of Du with respect to $\mathcal{L}^N$. We also consider the set $S(u)$ of points which are not Lebesgue points for u, and recall that $S(u)$ is $N-1$-rectifiable, and so it has a normal, ν, $\mathcal{H}^{N-1}$-almost everywhere. We set $D_j u := D_s u \lfloor S(u)$ and use the representations $D_{ac}u = \nabla u \mathcal{L}^N$ and $D_j u = [u] \otimes \nu \mathcal{H}^{N-1} \lfloor S(u)$, so we have the decomposition

$$Du = \nabla u \mathcal{L}^N + [u] \otimes \nu \mathcal{H}^{N-1} \lfloor S(u) + C(u), \tag{2.2}$$

where $C(u) := D_s u - D_j u$, $[u]$ is the jump in u across $S(u)$, i.e., $[u] = u^+ - u^-$, where u^+ and u^- are the traces of u on either side of $S(u)$. If $C(u) = 0$, then we say u is a *special function of bounded variation*, and we write $u \in SBV(\Omega;\mathbb{R}^p)$. This space was introduced in [9].

We set $\mathbb{R}^+ := [0,\infty)$ and $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty,\infty\}$. We denote the space of $p \times N$ matrices by $\mathbb{M}^{p\times N}$, and, for $W : \mathbb{M}^{p\times N} \to \mathbb{R}$, we define the *recession function* $W^\infty : \mathbb{M}^{p\times N} \to \bar{\mathbb{R}}$ by

$$W^\infty(F) := \limsup_{t\to\infty} \frac{W(tF)}{t}.$$

We recall that a function $f : \mathbb{M}^{p\times N} \to \mathbb{R}$ is *quasiconvex* if $f(F) \le \int_A f(\nabla\phi)dx$ for all $\phi \in Fx + C_0^\infty(A;\mathbb{R}^p)$ and all $F \in \mathbb{M}^{p\times N}$, where $A \subset \mathbb{R}^N$ is any open set with $\mathcal{L}^N(A) = 1$ (see [15]). We denote by QW the quasiconvex envelope of W and by CW the convex envelope, i.e.,

$$QW(F) := \sup\{f(F) : f \le W \text{ and } f \text{ is quasiconvex}\},$$
$$CW(F) := \sup\{f(F) : f \le W \text{ and } f \text{ is convex}\}.$$

It follows from a straightforward rescaling argument (see [13]; see also [1] and [8]) that for W satisfying (H2) below, we have

$$QW(F) = \inf\left\{ \liminf_{n\to\infty} \int_A W(\nabla u_n)dx : \{u_n\} \subset W^{1,1}(A;\mathbb{R}^p), u_n \to Fx \text{ in } L^1(A;\mathbb{R}^p)\right\} \tag{2.3}$$

for any open set A with $\mathcal{L}^N(A) = 1$.

For a unit vector $\nu \in \mathbb{R}^N$, we denote by Q_ν any open unit cube centered at 0 with two faces normal to ν, and S_ν is the set $\{x \in \mathbb{R}^N : |x\cdot\nu| < 1/2\}$.

If $f : \mathbb{M}^{p\times N} \to \mathbb{R}$ is positive homogeneous of degree one and μ is a $\mathbb{M}^{p\times N}$-valued measure, we use the notation

$$\int f(d\mu) := \int f\left(\frac{d\mu}{d|\mu|}\right) d|\mu|,$$

where $|\mu|$ is the total variation measure of μ.

We consider $W : \mathbb{M}^{p\times N} \to \mathbb{R}^+$, and for $u \in BV(\Omega;\mathbb{R}^p)$, we define

$$E(u) := \int_\Omega W(\nabla u)dx + \mathcal{H}^{N-1}(S(u)).$$

The relaxed functional I is defined by

$$I(u) := \inf\left\{ \liminf_{n\to\infty} E(u_n) : \{u_n\} \subset SBV(\Omega;\mathbb{R}^p), u_n \to u \text{ in } L^1(\Omega;\mathbb{R}^p)\right\},$$

and for $\xi \in \mathbb{R}^p, \nu \in S^{N-1}$, we define the following functional:

$$h(\xi,\nu) := \inf\Big\{ \int_{Q_\nu} W^\infty(\nabla v)dx + \mathcal{H}^{N-1}(S(v)) : v \in SBV(Q_\nu;\mathbb{R}^p), \quad v = \xi \text{ if } x \in \partial Q_\nu \text{ and } x\cdot\nu \geq 0, \text{ and } v = 0 \text{ if } x \in \partial Q_\nu \text{ and } x\cdot\nu < 0\Big\}. \tag{2.4}$$

As we will see below, $h([u],\nu)$ turns out to be the energy density of I on $S(u)$.

Before giving the relaxation theorem, we state the following hypotheses on the bulk density W:

(H1) $W:\mathbb{M}^{p\times N}\to\mathbb{R}^+$ is continuous,

(H2) for some $C_0, C_1 > 0$ and all $F \in \mathbb{M}^{p\times N}$, we have

$$C_0|F| - \frac{1}{C_0} \leq W(F) \leq C_1(1+|F|),$$

(H3) there exist $m \in (0,1), L > 0$, and $C > 0$ such that

$$\left|W^\infty(F) - \frac{W(tF)}{t}\right| \leq \frac{C}{t^m} \text{ for all } F \in \mathbb{M}^{p\times N} \text{ with } |F| = 1, \text{ and all } t > L.$$

THEOREM 2.1. *If $W:\mathbb{M}^{p\times N}\to\mathbb{R}^+$ satisfies* (H1), (H2), *and* (H3), *then*

$$I(u) = \int_\Omega QW(\nabla u)dx + \int_{S(u)} h([u],\nu)d\mathcal{H}^{N-1} + \int_\Omega (QW)^\infty(dC(u)). \tag{2.5}$$

The proof of this theorem will be carried out in sections 3, 4, and 5.

3. Characterizations of QW for sequences in BV. It is useful to consider the following bulk density, which is analogous to that considered in [6]. $G:\mathbb{M}^{p\times N}\to\mathbb{R}^+$ is defined by

$$G(F) := \inf\Big\{ \liminf_{n\to\infty} \int_Q W(\nabla u_n)dx : \{u_n\} \subset SBV(Q;\mathbb{R}^p), \quad u_n \to Fx \text{ in } L^1(Q;\mathbb{R}^p), \text{ and } \mathcal{H}^{N-1}(S(u_n)) \to 0\Big\}.$$

The goal of this section is to prove that $G = QW$. Recalling (2.3), we see that the admissible class for G is larger than that for QW. The point, then, is to show that if we insist that admissible sequences for G satisfy $\mathcal{H}^{N-1}(S(u_n)) \to 0$, we might as well require the sequences to be in $W^{1,1}$, as in (2.3). In fact, the lemma below indicates that $W^{1,1}$ can be weakened to BV, with the requirement that $|D_s u_n|(Q) \to 0$.

LEMMA 3.1. *Suppose that $W:\mathbb{M}^{p\times N}\to\mathbb{R}^+$ is a Borel measurable function such that* (H2) *holds. Then*

$$QW(F) = \inf\Big\{ \liminf_{n\to\infty} \int_Q W(\nabla u_n(x))dx : \{u_n\} \subset BV(Q;\mathbb{R}^p), \quad u_n \to Fx \text{ in } L^1(Q;\mathbb{R}^p) \text{ and } |D_s u_n|(Q) \to 0\Big\}$$

for all $F \in \mathbb{M}^{p\times N}$.

Proof. We need only show $QW(F) \leq$ the right-hand side above, since the admissible class of functions for the right-hand side is broader than that in (2.3). Let $\{u_n\}$ be an admissible sequence for the right-hand side. By Theorem 2.16 in [14], we know that for each u_n, we can choose a sequence $v_{n,k} \in W^{1,1}(Q;\mathbb{R}^p)$ such that $v_{n,k} \to u_n$ in L^1 as $k \to \infty$ and

$$\int_Q QW(\nabla v_{n,k})dx \to \int_Q QW(\nabla u_n)dx + C|D_s u_n|(Q)$$

as $k \to \infty$, for some $C > 0$. Since $|D_s u_n|(Q) \to 0$, we can take a diagonal subsequence $\{v_n\}$ such that $v_n \to Fx$ and

$$\liminf_{n\to\infty} \int_Q QW(\nabla v_n)dx = \liminf_{n\to\infty} \int_Q QW(\nabla u_n)dx.$$

The lemma follows from the fact that

$$QW(F) \leq \liminf_{n\to\infty} \int_Q QW(\nabla v_n)dx. \qquad \square$$

So, it remains only to show that requiring admissible sequences to satisfy

$$\mathcal{H}^{N-1}(S(u_n)) \to 0$$

is no less restrictive than requiring $|D_s u_n|(Q) \to 0$. If the former holds, since u_n are in SBV, it is enough if u_n are uniformly bounded in L^∞. In fact, it is enough if a little less is true. In the next lemma we see that we can truncate u_n in such a way that $\liminf_{n\to\infty} \int_Q W(\nabla u_n(x))dx$ is altered by an arbitrarily small amount. It is then straightforward to show $G = QW$.

LEMMA 3.2. *Let* $W : \mathbb{M}^{p\times N} \to \mathbb{R}^+$ *be a Borel measurable function satisfying* (H2)*, and let* $f \in L^\infty(Q;\mathbb{R}^p)$ *and* $\varepsilon > 0$ *be given. Then for every sequence* $\{u_n\} \subset SBV(Q;\mathbb{R}^p)$ *such that*

$$||u_n||_{L^1(Q;\mathbb{R}^p)} + |D_{ac}u_n|(Q) \leq R$$

for all $n \in \mathbb{N}$ *and some* $R > 0$*, there exists a sequence* $\{v_n\} \subset SBV(Q;\mathbb{R}^p)$ *uniformly bounded in* $L^\infty(Q;\mathbb{R}^p)$ *such that*

$$S(v_n) \subset S(u_n), \quad ||v_n - f||_{L^1(Q;\mathbb{R}^p)} \leq ||u_n - f||_{L^1(Q;\mathbb{R}^p)}, \text{ and}$$

$$\liminf_{n\to\infty} \int_Q W(\nabla v_n(x))dx \leq \liminf_{n\to\infty} \int_Q W(\nabla u_n(x))dx + \varepsilon.$$

Proof. The proof is a simpler version of the proof of Lemma 3.7 in [6], which relies on a truncation argument proposed by De Giorgi. Set $\lambda := [\ln(||f||_\infty + 1)] + 1$, where $[\cdot]$ is integer part, and fix $k \in \mathbb{N}$ with $k \geq \lambda$. Let $i \in \{\lambda, \dots, k\}$ be given. Define $\phi_i \in W^{1,\infty}(\mathbb{R}^p;\mathbb{R}^p)$ by

$$\phi_i(x) := \begin{cases} x & \text{if } |x| \leq e^i, \\ \frac{x}{e-1}\left(\frac{e^{i+1}}{|x|} - 1\right) & \text{if } e^i < |x| < e^{i+1}, \\ 0 & \text{if } |x| \geq e^{i+1}. \end{cases}$$

Set $u_{n,i} := \phi_i \circ u_n$. Then $||u_{n,i}||_\infty \le e^i$. Since $\mathrm{Lip}(\phi_i) = 1$, we have $u_{n,i} \in SBV(Q;\mathbb{R}^p)$, $|D_{ac}u_{n,i}|(Q) \le |D_{ac}u_n|(Q)$, and $S(u_{n,i}) \subset S(u_n)$ (see [3] and [16]). Furthermore, by the choice of λ we have

$$\begin{aligned} ||u_{n,i} - f||_{L^1(Q;\mathbb{R}^p)} &= \int_{\{|u_n|<e^i\}} |u_n(x) - f(x)|dx + \int_{\{|u_n|\ge e^i\}} |\phi_i(u_n(x)) - \phi_i(f(x))|dx \\ &\le ||u_n - f||_{L^1(Q;\mathbb{R}^p)}, \end{aligned}$$

where we used the fact that $\mathrm{Lip}(\phi_i) = 1$ and $\phi_i \circ f = f$. Now, fix $n \in \mathbb{N}$ and set

$$Q_i := \{x \in Q : |u_n(x)| < e^i\}.$$

Note that we have $\int_{Q\setminus Q_i} W(\nabla u_{n,i}(x))dx \le C_1(\mathcal{L}^N(Q\setminus Q_i) + |D_{ac}u_{n,i}|(Q\setminus Q_i))$, where $\mathcal{L}^N(Q\setminus Q_i) \le \frac{R}{e^i}$ and

$$\sum_{i=\lambda}^{k} |D_{ac}u_{n,i}|(Q\setminus Q_i) \le \sum_{i=\lambda}^{k} |D_{ac}u_{n,i}|(\{e^i \le |\tilde{u}_n(x)| < e^{i+1}\}) \le |D_{ac}u_n|(Q) \le R.$$

We now have that

$$\sum_{i=\lambda}^{k} \int_{Q\setminus Q_i} W(\nabla u_{n,i}(x))dx \le \sum_{i=\lambda}^{k} C_1\frac{R}{e^i} + C_1R \le C_1R\left(\frac{1}{e^{\lambda-1}(e-1)} + 1\right)$$

so that

$$\sum_{i=\lambda}^{k} \int_{Q} W(\nabla u_{n,i}(x))dx \le (k-\lambda+1)\int_{Q} W(\nabla u_n(x))dx + C_1R\left(\frac{1}{e^{\lambda-1}(e-1)} + 1\right),$$

and by the choice of λ,

$$\begin{aligned} \frac{1}{k-\lambda+1}\sum_{i=\lambda}^{k} \int_{Q} W(\nabla u_{n,i}(x))dx \le\ & \int_{Q} W(\nabla u_n(x))dx \\ & + \frac{C_1R}{k - [\ln(||f||_\infty + 1)]}\left(1 + \frac{1}{(||f||_\infty + 1)(e-1)}\right). \end{aligned}$$

Choosing k large enough so that

$$\frac{C_1R}{k - [\ln(||f||_\infty + 1)]}\left(1 + \frac{1}{(||f||_\infty + 1)(e-1)}\right) < \varepsilon,$$

we see that there must be an $i \in \{\lambda, \dots, k\}$ so that

$$\int_{Q} W(\nabla u_{n,i}(x))dx \le \int_{Q} W(\nabla u_n(x))dx + \varepsilon$$

with $||u_{n,i}||_\infty \le e^k$, where the above choice of k does not depend on n. Hence, this can be done for all $n \in \mathbb{N}$, giving the same L^∞ bound of e^k, and the proof is complete, choosing $v_n := u_{n,i}$. ∎

We now have the following proposition.

PROPOSITION 3.3. *Suppose that* $W:\mathbb{M}^{p\times N}\to\mathbb{R}^+$ *is a Borel measurable function satisfying* (H2). *Then*

$$QW = G.$$

Proof. We need only show that

$$QW(F) \le G(F), \tag{3.1}$$

since the admissible class of $\{u_n\}$ for $G(F)$ includes that for $QW(F)$, and so $QW(F) \ge G(F)$. Choose $\{u_n\} \subset SBV(Q;\mathbb{R}^p)$ such that $u_n \to Fx$ in $L^1(Q;\mathbb{R}^p)$, $\mathcal{H}^{N-1}(S(u_n)) \to 0$, and

$$\lim_{n\to\infty}\int_Q W(\nabla u_n(x))dx = G(F).$$

Since the sequence $\{u_n\}$ is convergent in L^1, it is bounded. Furthermore, since $W(\nabla u_n(x)) \ge C_0|\nabla u_n(x)| - \frac{1}{C_0}$, we deduce that

$$\begin{aligned}|D_{ac}u_n|(Q) &= \int_Q |\nabla u_n(x)|dx\\ &\le \frac{1}{C_0}\left(\int_Q W(\nabla u_n(x))dx + \frac{1}{C_0}\right)\\ &\to \frac{1}{C_0}\Big(G(F) + \frac{1}{C_0}\Big) < \infty,\end{aligned}$$

so $\sup_{n\in\mathbb{N}}|D_{ac}u_n|(Q) < \infty$. Let $\varepsilon > 0$ and consider Lemma 3.2 with $f := Fx$, $R := \sup_{n\in\mathbb{N}}(||u_n||_{L^1} + |D_{ac}u_n|(Q))$, and the above ε and $\{u_n\}$. We now have

$$\lim_{n\to\infty}\int_Q W(\nabla v_n(x))dx \le G(F) + \varepsilon$$

for some $\{v_n\}$ with the same properties as $\{u_n\}$ and, in addition, $||v_n||_\infty \le M < \infty$ for all $n \in \mathbb{N}$. Hence, $|D_s v_n|(Q) \le 2M\mathcal{H}^{N-1}(S(v_n)) \to 0$. By Lemma 3.1, we conclude that

$$QW(F) \le \lim_{n\to\infty}\int_Q W(\nabla v_n(x))dx \le G(F) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we have (3.1). ∎

4. Upper bound. In this section, we prove an inequality leading to (2.5). Precisely,

$$I(u) \le \int_\Omega QW(\nabla u)dx + \int_{S(u)} h([u],\nu)d\mathcal{H}^{N-1} + \int_\Omega (QW)^\infty(dC(u)).$$

To do this, we first show that $I(u)$ can be considered a set function, and, in fact, a measure. For $A \subset \Omega$ open set

$$E(u,A) := \int_A W(\nabla u)dx + \mathcal{H}^{N-1}(S(u)\cap A)$$

and

$$I(u,A) := \inf\left\{\liminf_{n\to\infty} E(u_n,A) : \{u_n\} \subset SBV(A;\mathbb{R}^p), u_n \to u \text{ in } L^1(A;\mathbb{R}^p)\right\}.$$

Then we have the following proposition.

PROPOSITION 4.1. *Suppose that $u \in BV(A;\mathbb{R}^p)$, where A is a bounded, open subset of Ω. Then $I(u,\cdot)$ extends to a nonnegative, finite, Borel regular measure on A, which is absolutely continuous with respect to $\mathcal{L}^N + |Du|$.*

Proof. By an argument similar to that for Theorem 3.2 in [12], we know that $I(u,\cdot)$ is a Radon measure on A, and is, in fact, the weak $*$ limit of the measures $E(u_n,\cdot)$ for a minimizing sequence u_n. It remains to show that $I(u,\cdot)$ is finite and absolutely continuous with respect to $\mathcal{L}^N + |Du|$. Let $B \subset A$ be open. By Theorem 5.3.3 of [17] or Theorem 2 in section 5.2.2 of [10], we choose $u_n \in C^\infty(B;\mathbb{R}^p)$ such that $u_n \to u$ in $L^1(B;\mathbb{R}^p)$ and $|Du_n|(B) \to |Du|(B)$. Since the u_n are smooth, we have

$$\begin{aligned} I(u,B) &= \liminf_{n\to\infty} \int_B W(\nabla u_n)dx \\ &\le \liminf_{n\to\infty} \int_B C_1\Big[1 + |\nabla u_n|\Big]dx \\ &= C_1\Big[\mathcal{L}^N(B) + |Du|(B)\Big], \end{aligned}$$

which, in particular, implies that $I(u,A) < \infty$ for all $u \in BV(A;\mathbb{R}^p)$. ∎

Fix $A \subset \Omega$ open and $u \in BV(A;\mathbb{R}^p)$. Note that we have

$$I(u,A) \le \inf\left\{\liminf_{n\to\infty} \int_A W(\nabla u_n)dx : \{u_n\} \subset W^{1,1}(A;\mathbb{R}^p), u_n \to u \text{ in } L^1(A;\mathbb{R}^p)\right\},$$

so from [14] we know that

$$I(u,A) \le \int_A QW(\nabla u)dx + \int_{S(u)\cap A} (QW)^\infty(dD_j u) + \int_A (QW)^\infty(dC(u));$$

hence, it only remains to prove that

$$I(u,S(u)) \le \int_{S(u)} h([u](x),\nu(x))d\mathcal{H}^{N-1}(x). \tag{4.1}$$

The jump set $S(u)$ is, in general, not so easy to deal with. Indeed, there exist functions $u \in BV((0,1)^2)$ with jump set $\{(x,y) \in (0,1)^2 : x \in \mathbb{Q}\}$. Furthermore, although for such u one has $E(u) = \infty$, we know that $I(u) \le C_1[1+|Du|((0,1)^2)] < \infty$. However, measure theoretic boundaries of sets of finite perimeter are much easier to handle and, for our purposes, there are connections between $S(u)$ and certain sets of finite perimeter that we can exploit.

Let $u \in BV(\Omega)$ and let $D \subset \mathbb{R}$ be dense. Then,

$$S(u) = \bigcup_{t\in D} S(u) \cap \partial_* E_t = \bigcup_{\substack{t_1,t_2\in D \\ t_1\neq t_2}} \partial_* E_{t_1} \cap \partial_* E_{t_2},$$

where $E_t := \{x \in \Omega : u(x) > t\}$. If $u \in BV(\Omega;\mathbb{R}^p)$, we denote the t level set of u^i by E_t^i, or by E_t if it is clear that we mean the ith component. Also, if $u \in BV(\Omega)$, then E_t

has finite perimeter for $\mathcal{L}^1$ almost every t, and $\{x \in S(u) : u^-(x) < t < u^+(x)\} \subset \partial_* E_t$ (see, e.g., the proof of Theorem 1 in section 5.9 of [10]). We also point out that for $u \in BV(\Omega;\mathbb{R}^p)$, we have $S(u) = \bigcup_{i=1}^p S(u^i)$.

If $T \subset \Omega$ has finite perimeter, then $\mathcal{H}^{N-1}\lfloor\partial_* T$ is a Radon measure. Since $S(u)$ is $\mathcal{H}^{N-1}$ measurable, we conclude that $\chi_{S(u)} \in L^1(\Omega, \mathcal{H}^{N-1}\lfloor\partial_* T)$. So, for $\mathcal{H}^{N-1}$ almost every $x \in S(u) \cap \partial_* T$ we have

$$\begin{aligned}\lim_{\delta\to 0^+} \frac{\mathcal{H}^{N-1}(B(x,\delta)\cap S(u)\cap\partial_* T)}{\alpha(N-1)\delta^{N-1}} &= \lim_{\delta\to 0^+} \frac{\mathcal{H}^{N-1}(B(x,\delta)\cap S(u)\cap\partial_* T)}{\mathcal{H}^{N-1}(B(x,\delta)\cap\partial_* T)} \\ &= \lim_{\delta\to 0^+} \fint_{B(x,\delta)} \chi_{S(u)} d\mathcal{H}^{N-1}\lfloor\partial_* T = 1,\end{aligned}$$

where the first equality follows from Corollary 1 (ii) in section 5.7.2 of [10]. Hence, if $D \subset \mathbb{R}$ is countable and dense and such that E_t^i has finite perimeter for all $t \in D$ and all $i \in \{1,\dots,p\}$, then, fixing $i \in \{1,\dots,p\}$ and setting $E_t := E_t^i$, we have that for $\mathcal{H}^{N-1}$ almost every $x \in S(u)$, for all $t \in D \cap ((u^i)^-(x), (u^i)^+(x))$

(4.2)
$$\lim_{\delta\to 0^+} \frac{\mathcal{H}^{N-1}(B(x,\delta)\cap S(u)\cap\partial_* E_t)}{\alpha(N-1)\delta^{N-1}} = \lim_{\delta\to 0^+} \frac{\mathcal{H}^{N-1}(B(x,\delta)\cap S(u)\cap\partial_* E_t)}{\mathcal{H}^{N-1}(B(x,\delta)\cap\partial_* E_t)} = 1.$$

Furthermore, since $[u] \in L^1(\Omega, \mathcal{H}^{N-1}\lfloor(S(u)\cap\partial_* E_t))$, for $\mathcal{H}^{N-1}$ almost every $x \in S(u)$, for all $t \in D \cap ((u^i)^-(x), (u^i)^+(x))$ we have

$$\lim_{\delta\to 0^+} \fint_{B(x,\delta)\cap S(u)\cap\partial_* E_t} |[u](y) - [u](x)| d\mathcal{H}^{N-1}(y) = 0.$$

Note that the same is true if $B(x,\delta)$ is replaced by $Q(x,\delta) := x + \delta Q_{\nu(x)}$. Hence, for $\mathcal{H}^{N-1}$ almost every $x \in S(u)$, for all $t \in D \cap ((u^i)^-(x), (u^i)^+(x))$ we have

$$\lim_{\delta\to 0^+} \frac{1}{\delta^{N-1}} \int_{Q(x,\delta)\cap S(u)\cap\partial_* E_t} |[u](y)| d\mathcal{H}^{N-1}(y) = |[u](x)|. \tag{4.3}$$

On the other hand, for $\mathcal{H}^{N-1}$ almost every $x \in S(u)$,

$$\lim_{\delta\to 0^+} \frac{1}{\delta^{N-1}} \int_{Q(x,\delta)\cap S(u)} |[u](y)| d\mathcal{H}^{N-1}(y) = |[u](x)|,$$

which, together with (4.3), shows that for $\mathcal{H}^{N-1}$ almost every $x \in S(u)$, for all $t \in D \cap ((u^i)^-(x), (u^i)^+(x))$ we have

$$\lim_{\delta\to 0^+} \frac{|D_j u|(Q(x,\delta)\backslash\partial_* E_t)}{\delta^{N-1}} = 0. \tag{4.4}$$

Proof [Proof of (4.1)]. First, we note that $W^\infty(\xi\otimes\nu)$ is continuous since the limit W^∞ is attained uniformly (see (H3)), hence $h(\xi,\nu)$ is continuous. Note further that, for E_t as above, $[u]$ and ν are $\mathcal{H}^{N-1}\lfloor(S(u)\cap\partial_* E_t)$-measurable, and $h \le 1$, so

$$h([u](\cdot),\nu(\cdot)) \in L^1(\Omega, \mathcal{H}^{N-1}\lfloor(S(u)\cap\partial_* E_t)).$$

Let $x_0 \in S(u)\cap\Omega$ and $t \in \mathbb{R}$ be given such that $u_i^-(x_0) < t < u_i^+(x_0)$, E_t has finite perimeter,

$$\lim_{\delta\to 0^+} \frac{1}{\delta^{N-1}} \int_{Q(x_0,\delta)} |\nabla u| dx = 0, \quad \lim_{\delta\to 0^+} \frac{1}{\delta^{N-1}} |C(u)|(Q(x_0,\delta)) = 0, \tag{4.5}$$

(4.6)
$$\lim_{\delta\to 0^+}\frac{1}{\delta^{N-1}}|D_j u|(Q(x_0,\delta)\backslash\partial_* E_t)=0,\quad \lim_{\delta\to 0^+}\frac{\mathcal{H}^{N-1}(Q(x_0,\delta)\cap S(u)\cap\partial_* E_t)}{\delta^{N-1}}=1,$$

and

(4.7)
$$\lim_{\delta\to 0^+}\frac{1}{\delta^{N-1}}\int_{Q(x_0,\delta)\cap S(u)\cap\partial_* E_t}|h([u](x),\nu(x))-h([u](x_0),\nu(x_0))|d\mathcal{H}^{N-1}(x)=0.$$

Note that the above can be done for $\mathcal{H}^{N-1}$ almost every $x\in S(u)$ (the last three follow from (4.4) and (4.2)).

Since (4.1) is equivalent to

$$I(u,S(u))\le\int_{S(u)}\frac{h([u](x),\nu(x))}{|[u](x)|}|[u](x)|d\mathcal{H}^{N-1}(x),$$

and we know that

$$\lim_{\delta\to 0^+}\frac{I(u,Q(x_0,\delta))}{|D_j u|(Q(x_0,\delta))}=\lim_{\delta\to 0^+}\frac{I(u,Q(x_0,\delta))}{|[u](x_0)|\delta^{N-1}},$$

it is enough to show that

$$\limsup_{\delta\to 0^+}\frac{I(u,Q(x_0,\delta))}{\delta^{N-1}}\le h([u](x_0),\nu(x_0)).$$

Let $\varepsilon>0$ be given and choose $\delta_{x_0}\in(0,\varepsilon)$ such that if $\delta\in(0,\delta_{x_0})$, then (4.5), (4.6), and (4.7) hold to within ε. For $\delta\in(0,\delta_{x_0})$ we would like to find a sequence $\{v_n\}\subset SBV(Q(x_0,\delta);\mathbb{R}^p)$ such that $v_n\to u$ in $L^1(Q(x_0,\delta);\mathbb{R}^p)$ and

$$\liminf_{n\to\infty}\frac{E(v_n,Q(x_0,\delta))}{\delta^{N-1}}\le h([u](x_0),\nu(x_0))+O(\varepsilon).\tag{4.8}$$

Let $\delta\in(0,\delta_{x_0})$ and denote $Q(x_0,\delta)$ by Q. The idea is this: in Q, the set $S(u)\cap\partial_* E_t$ is close to the hyperplane $H(x_0)$, which has normal $\nu(x_0)$. Furthermore, on $H^+(x_0)$, u is close to the trace u^+, and on $H^-(x_0)$, u is close to u^-. Now, if $S(u)\cap\partial_* E_t$ were equal to the hyperplane $H(x_0)$, and if u were equal to its traces on each side of the hyperplane, we would do the following: choose w admissible for $h([u](x_0),\nu(x_0))$ such that

$$\int_{Q_{\nu(x_0)}}W^\infty(\nabla w)dx+\mathcal{H}^{N-1}(S(w))<h([u](x_0),\nu(x_0))+\varepsilon.\tag{4.9}$$

Extend w periodically and set

$$w_n(x):=w\Big(\frac{n}{\delta}x\Big)\in SBV_{\rm loc}\Big(\frac{\delta}{n}S_{\nu(x_0)};\mathbb{R}^p\Big).$$

Define

$$v_n(x):=\begin{cases}u^-(x_0)+w_n(x) & \text{if } x\in Q\cap(x_0+\dfrac{\delta}{n}S_{\nu(x_0)}),\\ u^+(x_0) & \text{if } x\in Q\cap H^+(x_0)\backslash(x_0+\dfrac{\delta}{n}S_{\nu(x_0)}),\\ u^-(x_0) & \text{if } x\in Q\cap H^-(x_0)\backslash(x_0+\dfrac{\delta}{n}S_{\nu(x_B)}).\end{cases}$$

Note that as $n \to \infty$, $v_n \to u$ in $L^1(Q;\mathbb{R}^p)$. Furthermore,

$$E(v_n, Q) = E\Big(v_n, Q \cap \Big(x_0 + \frac{\delta}{n} S_{\nu(x_0)}\Big)\Big) + \mathcal{L}^N\Big(Q \backslash \Big(x_0 + \frac{\delta}{n} S_{\nu(x_0)}\Big)\Big) W(0) \tag{4.10}$$

and

$$\begin{aligned} E\Big(v_n, Q \cap \Big(x_0 + \frac{\delta}{n} S_{\nu(x_0)}\Big)\Big) &= \int_{(\delta/n) S_{\nu(x_0)} \cap \delta Q_{\nu(x_0)}} W\Big(\frac{n}{\delta} \nabla w\Big(\frac{n}{\delta} x\Big)\Big) dx + \mathcal{H}^{N-1}(S(w_n) \cap Q) \\ &= \delta^{N-1} \int_{Q_{\nu(x_0)}} \frac{\delta}{n} W\Big(\frac{n}{\delta} \nabla w(x)\Big) dx + \delta^{N-1} \mathcal{H}^{N-1}(S(w)). \end{aligned}$$

By (H3),

$$\begin{aligned} \int_{Q_{\nu(x_0)}} \Big| \frac{\delta}{n} W\Big(\frac{n}{\delta} \nabla w(x)\Big) - W^\infty(\nabla w(x))\Big| dx \ &\le \int_{Q_{\nu(x_0)} \cap \{n | \nabla w| > \delta L\}} \frac{C\delta}{L^{m-1} n} dx \\ &\quad + \int_{Q_{\nu(x_0)} \cap \{n | \nabla w| \le \delta L\}} \frac{\delta}{n} C_1 [1 + L] dx \\ &\to 0 \end{aligned} \tag{4.11}$$

as $n \to \infty$. Using (4.9) and (4.10), we now have

$$\lim_{n \to \infty} \Big| E(v_n, Q) - \Big[\delta^{N-1} h([u](x_0), \nu(x_0)) + \delta^N W(0)\Big]\Big| \le \delta^{N-1} \varepsilon, \tag{4.12}$$

which gives (4.8).

We now need to consider the actual situation, where $S(u) \cap \partial_* E_t$ does not equal $H(x_0)$, and u does not equal its trace on each side of $H(x_0)$. The construction involved is quite messy, but it is a straightforward adaptation of what we have already done. Therefore, we give only an outline of the proof. Again, we need to find a sequence $\{v_n\}$ that approaches u in $L^1(Q;\mathbb{R}^p)$ such that (4.8) holds. Since u is some distance away from its trace on either side of $H(x_0)$, we cannot have a construction as simple as before. However, we can choose small disjoint balls, B, centered at points x_B in $S(u) \cap \partial_* E_t \cap Q$ that are Lebesgue points for $h([u](\cdot), \nu(\cdot))$. With a finite number of such balls, we can almost cover $S(u) \cap \partial_* E_t \cap Q$ (almost with respect to $\mathcal{H}^{N-1} \lfloor (S(u) \cap \partial_* E_t)$ and with respect to $|D_j u|$), and they can be chosen small enough so that in these balls, u is arbitrarily close to its traces on either side of $H(x_B)$. Outside them, we can take $v_n = u$, and inside, we can perform constructions of v_n as we did for x_0 above. We then need to add a transition layer around these balls, to connect v_n inside and outside the balls. There are then the two issues: i) does $v_n \to u$, and ii) does (4.8) hold?

i) This issue is easy since we can choose balls so small that essentially all of Q is outside of the balls, and we recall that there, $v_n = u$.

The real issue is ii). First, note that within each ball B, we perform a construction as we did for x_0, and so the energy there will be arbitrarily close to

$$\alpha(N-1) r_B^{N-1} h([u](x_B), \nu(x_B)) + \alpha(N) r_B^N W(0),$$

where r_B is the radius of ball B. By (4.7), we know that these energies sum to within order $\delta^{N-1}\varepsilon$ of $\delta^{N-1} h([u](x_0), \nu(x_0))$. Outside the balls, we know from (4.5), (4.6), and the choice of the balls that the energy is small. On the transition layers around

each ball, we smoothly connect v_n inside the ball to its value u outside the transition layer. There is arbitrarily small variation in the transition layers, since the balls were chosen so that $|D_j u|(Q\backslash \cup B)$ is small and in the transition layers u is close to its traces $u^+(x_B)$ and $u^-(x_B)$. Since W has linear growth, the energy in the transition layers is arbitrarily small. Therefore, (4.8) holds. □

5. Lower bound. In this section we prove that

$$I(u) \geq \int_\Omega QW(\nabla u)dx + \int_{S(u)} h([u],\nu)d\mathcal{H}^{N-1} + \int_\Omega (QW)^\infty(dC(u)).$$

As mentioned in the introduction, we rely heavily on [6], and we use the blow-up method introduced by Fonseca and Müller in [14].

Let $u_n \in SBV(\Omega;\mathbb{R}^p)$ be given such that $u_n \to u$ in $L^1(\Omega,\mathbb{R}^p)$ and

$$\begin{aligned}\liminf_{n\to\infty} &\left[\int_\Omega W(\nabla u_n)dx + \mathcal{H}^{N-1}(S(u_n))\right] \\ &= \lim_{n\to\infty}\left[\int_\Omega W(\nabla u_n)dx + \mathcal{H}^{N-1}(S(u_n))\right] < \infty.\end{aligned}$$

Define a sequence of Radon measures by $\mu_n := W(\nabla u_n)\mathcal{L}^N + \mathcal{H}^{N-1}\lfloor S(u_n)$. Since μ_n are uniformly bounded, there exists a subsequence (not relabeled) and a finite Radon measure μ such that $\mu_n \overset{*}{\rightharpoonup} \mu$. The Radon–Nikodym theorem and (2.2) allow us to write

$$\mu = d_{ac}\mathcal{L}^N + d_j|[u]|\mathcal{H}^{N-1}\lfloor S(u) + d_c|C(u)| + \mu_s,$$

where μ_s is singular with respect to the first three terms in the decomposition, and μ_s is nonnegative since μ is. In view of Proposition 3.3, we need only show that

a) $d_{ac}(x_0) \geq G(\nabla u(x_0))$ for $\mathcal{L}^N$ almost every $x_0 \in \Omega$,

b) $d_j(x_0) \geq \frac{h([u](x_0),\nu(x_0))}{|[u](x_0)|}$ for $\mathcal{H}^{N-1}$ almost every $x_0 \in S(u)$, and

c) $d_c(x_0) \geq G^\infty\left(\frac{dC(u)}{d|C(u)|}(x_0)\right)$ for $|C(u)|$ almost every $x_0 \in \Omega$.

The proofs of all three are straightforward adaptations of the corresponding arguments in [6], except that we introduce a significant simplification in the proof of c).

A complication in the proof in [6] is due to the fact that as a sequence of measures converges weakly $*$ on a cube, it might lose mass. However, the sequences we need to consider come from blowing up one particular measure. The lemma below shows that, almost everywhere, we don't have to worry about such sequences losing mass.

LEMMA 5.1. *Let λ be a Radon measure on $\Omega \subset \mathbb{R}^N$. Then, for λ almost every $x \in \Omega$, given any open, bounded convex set C containing the origin, there is a sequence $\delta_i \to 0^+$ and a Radon measure γ on C such that*

$$\lambda_{\delta_i}(\cdot) := \frac{\lambda(x+\delta_i\cdot)}{\lambda(x+\delta_i C)} \overset{*}{\rightharpoonup} \gamma \text{ on } C, \text{ and } \gamma(C) = 1.$$

Proof. We first show that for λ almost every $x \in \Omega$ we have

$$\liminf_{\delta \to 0^+} \frac{\lambda(x + \delta C)}{\delta^N} > 0 \tag{5.1}$$

for all C as in the statement of the lemma. It is sufficient to consider (5.1) for an open ball B containing the origin, since $\delta B \subset C$ for small enough δ. Put

$$A := \left\{ x \in \Omega : \liminf_{\delta \to 0^+} \frac{\lambda(x + \delta B)}{\delta^N} = 0 \right\}.$$

Let $\varepsilon > 0$ be given, and using Besicovitch's covering theorem, choose a countable family of disjoint balls $x_i + \delta_i B \subset \Omega$ such that $\lambda(A \backslash \cup (x_i + \delta_i B)) = 0$ and $\lambda(x_i + \delta_i B) < \varepsilon \delta_i^N$. It follows that $\lambda(A) < \varepsilon \frac{\mathcal{L}^N(\Omega)}{\mathcal{L}^N(B)}$, and so $\lambda(A) = 0$.

Fix $x \in \Omega$ for which (5.1) is satisfied. Without loss of generality, we can assume $x = 0$. Let $\eta \in (0, 1)$ be given and set $\delta_i := \eta^i$. Suppose that

$$\limsup_{i \to \infty} \lambda_{\delta_i}(\eta C) < \eta^N.$$

Then we can choose $j \in \mathbb{N}$ and $\alpha \in (0, \eta^N)$ such that if $i \geq j$, then

$$\frac{\lambda(\delta_i \eta C)}{\lambda(\delta_i C)} < \alpha.$$

We now have

$$\frac{\lambda(\delta_i C)}{\delta_i^N} \leq \frac{\lambda(\delta_j C)\alpha^{i-j}}{[\eta^{i-j}\delta_j]^N} \to 0$$

as $i \to \infty$, which contradicts (5.1). Hence, we may extract a subsequence, not relabeled, and choose a Radon measure γ so that $\lambda_{\delta_i} \overset{*}{\rightharpoonup} \gamma$ on C and $\lambda_{\delta_i}(\eta C) > \alpha$, where $\alpha < \eta^N$. Choose $\beta \in (\eta, 1)$ such that $\gamma(\partial \beta C) = 0$. Then, for a subsequence and a Radon measure $\bar{\gamma}$, $\lambda_{\beta\delta_i} \overset{*}{\rightharpoonup} \bar{\gamma}$ on C, and for any Borel set $A \subset C$ we have

$$\lambda_{\beta\delta_i}(A) = \frac{\lambda(\beta\delta_i A)}{\lambda(\beta\delta_i C)} < \frac{1}{\alpha} \frac{\lambda(\beta\delta_i A)}{\lambda(\delta_i C)} = \frac{1}{\alpha} \lambda_{\delta_i}(\beta A).$$

Let $\varepsilon > 0$ be given and let $D \subset C$ be an open neighborhood of $\partial \beta C$ such that $\gamma(\bar{D}) < \varepsilon$. Then,

$$\limsup_{\delta_i \to 0^+} \lambda_{\beta\delta_i}\left(\frac{\bar{D}}{\beta} \cap C\right) \leq \frac{1}{\alpha} \limsup_{\delta_i \to 0^+} \lambda_{\delta_i}(\bar{D}) \leq \frac{1}{\alpha} \gamma(\bar{D}) < \frac{\varepsilon}{\alpha}.$$

Hence,

$$\bar{\gamma}(C) \geq \bar{\gamma}\left(C \backslash \frac{D}{\beta}\right) \geq \limsup_{\delta_i \to 0^+} \lambda_{\beta\delta_i}\left(C \backslash \frac{D}{\beta}\right) \geq \liminf_{\delta_i \to 0^+} \lambda_{\beta\delta_i}\left(C \backslash \frac{\bar{D}}{\beta}\right) > 1 - \frac{\varepsilon}{\alpha}.$$

From the arbitrariness of ε, it follows that $\bar{\gamma}(C) = 1$. ∎

Proof [Proof of c]. Let $x_0 \in \Omega$ be given such that

$$\lim_{\delta \to 0^+} \frac{|Du|(Q(x_0, \delta))}{|C(u)|(Q(x_0, \delta))} = 1, \quad \lim_{\delta \to 0^+} \frac{|Du|(Q(x_0, \delta))}{\delta^{N-1}} = 0, \quad \lim_{\delta \to 0^+} \frac{|Du|(Q(x_0, \delta))}{\delta^N} = \infty,$$

$$A_0 := \lim_{\delta\to 0^+} \frac{Du(Q(x_0,\delta))}{|Du|(Q(x_0,\delta))} \text{ exists and } ||A_0|| = 1, A_0 = a\otimes\nu, \text{ and}$$

$$d_c(x_0) = \lim_{\delta\to 0^+} \frac{\mu(Q(x_0,\delta))}{|C(u)|(Q(x_0,\delta))} = \lim_{\delta\to 0^+} \frac{\mu(Q(x_0,\delta))}{|Du|(Q(x_0,\delta))} < \infty.$$

Note that the above hold for $|C(u)|$ almost every $x\in\Omega$, where the statements regarding A_0 are due to Alberti [2]. Without loss of generality, assume that $\nu = e_N$ and $|a| = 1$. Choose $\delta_k < \frac{1}{k}$ such that, setting

$$z_k(x) := \frac{\delta_k^{N-1}}{|Du|(Q(x_0,\delta_k))}\left[u(x_0+\delta_k x) - \frac{1}{\delta_k^N}\int_{Q(x_0,\delta_k)} u(y)dy\right],$$

the sequence $\{\delta_k\}$ is selected according to Lemma 5.1 so that, with $\lambda := |Du|$ and γ equal to the weak $*$ limit of $|Dz_k|$, we have $\gamma(Q) = \lim_{k\to\infty}|Dz_k|(Q) = 1$. Note that, using notations from the proof of Lemma 5.1, the δ_k chosen here are of the form $\beta\delta_i$. So, by choosing an appropriate β in the proof of Lemma 5.1, the δ_k can also be chosen such that $\mu(\partial Q(x_0,\delta_k)) = 0$ for all k. Then,

$$\begin{aligned} d_c(x_0) &= \lim_{k\to\infty}\frac{\mu(Q(x_0,\delta_k))}{|Du|(Q(x_0,\delta_k))} \\ &= \lim_{k\to\infty}\left[\frac{1}{|Du|(Q(x_0,\delta_k))}\lim_{n\to\infty}\int_{Q(x_0,\delta_k)} d\mu_n\right] \\ &= \lim_{k\to\infty}\left\{\frac{1}{|Du|(Q(x_0,\delta_k))}\lim_{n\to\infty}\left[\int_{Q(x_0,\delta_k)} W(\nabla u_n(x))dx \right.\right. \\ &\qquad \left.\left. +\mathcal{H}^{N-1}(Q(x_0,\delta_k)\cap S(u_n))\right]\right\}. \end{aligned} \tag{5.2}$$

Also,

$$\lim_{k\to\infty}\lim_{n\to\infty}\frac{\delta_k^{N-1}}{|Du|(Q(x_0,\delta_k))}\int_Q \left|u_n(x_0+\delta_k x) - \frac{1}{\delta_k^N}\int_{Q(x_0,\delta_k)} u_n(y)dy - \left[u(x_0+\delta_k x) - \frac{1}{\delta_k^N}\int_{Q(x_0,\delta_k)} u(y)dy\right]\right| dx = 0. \tag{5.3}$$

By (5.2) and (5.3), using a standard diagonalization argument, choose a subsequence $\{u_k\}$ such that

$$d_c(x_0) = \lim_{k\to\infty}\frac{1}{|Du|(Q(x_0,\delta_k))}\left[\int_{Q(x_0,\delta_k)} W(\nabla u_k(x))dx + \mathcal{H}^{N-1}(S(u_k)\cap Q(x_0,\delta_k))\right]$$

and

$$||\bar{u}_k - z_k||_{L^1(Q;\mathbb{R}^p)} \to 0, \tag{5.4}$$

where

$$\bar{u}_k(x) := \frac{\delta_k^{N-1}}{|Du|(Q(x_0,\delta_k))}\left[u_k(x_0+\delta_k x) - \frac{1}{\delta_k^N}\int_{Q(x_0,\delta_k)} u_k(y)dy\right],$$

$$z_k(x) := \frac{\delta_k^{N-1}}{|Du|(Q(x_0,\delta_k))}\left[u(x_0+\delta_k x) - \frac{1}{\delta_k^N}\int_{Q(x_0,\delta_k)} u(y)dy\right].$$

Setting $\quad t_k := \dfrac{|Du|(Q(x_0,\delta_k))}{\delta_k^N} \to \infty, \quad \theta_k := \dfrac{|Du|(Q(x_0,\delta_k))}{\delta_k^{N-1}} \to 0$

we conclude that

$$d_c(x_0) = \lim_{k\to\infty}\left[\frac{1}{t_k}\int_Q W(t_k\nabla\bar{u}_k(x))dx + \frac{1}{\theta_k}\mathcal{H}^{N-1}(S(\bar{u}_k)\cap Q)\right].$$

Since $d_c(x_0) < \infty$, we know

$$\mathcal{H}^{N-1}(S(\bar{u}_k)\cap Q) \to 0 \text{ (since } \theta_k \to 0^+) \tag{5.5}$$

and

$$d_c(x_0) \geq \limsup_{k\to\infty}\frac{1}{t_k}\int_Q W(t_k\nabla\bar{u}_k(x))dx = \limsup_{k\to\infty}\int_Q W^\infty(\nabla\bar{u}_k(x))dx \tag{5.6}$$

just as in (4.11). Since

$$\int_Q z_k(x)dx = \int_Q \bar{u}_k(x)dx = 0 \quad \text{and} \quad |D\bar{u}_k|(Q) = |Dz_k|(Q) = 1,$$

by (5.4) and Poincaré's inequality, there exist subsequences (not relabeled) $\{z_k\}, \{\bar{u}_k\}$, and there exists $u_0 \in BV(Q;\mathbb{R}^p)$ such that $z_k, \bar{u}_k \to u_0$ in $L^1(Q;\mathbb{R}^p)$.

Now,

$$Dz_k(Q) = \frac{Du(Q(x_0,\delta_k))}{|Du|(Q(x_0,\delta_k))} \to A_0 = a\otimes e_N$$

and $|Dz_k|(Q) = 1$ so, by Proposition A.1 of [14], it follows that

$$|Dz_k - (Dz_k\cdot A_0)A_0|(Q) \to 0,$$

from which we conclude that $|Dz_k\cdot e_i|(Q)\to 0$ for $i = 1,\dots,N-1$. Since

$$|Du_0\cdot e_i|(Q) \leq \liminf_{k\to\infty}|Dz_k\cdot e_i|(Q) = 0,$$

we obtain

$$u_0(x) = \hat{u}_0(x_N) \in BV\left(\left(-\frac{1}{2},\frac{1}{2}\right);\mathbb{R}^p\right).$$

Note that, in general, if $\mu_k \overset{*}{\rightharpoonup} \eta$, $|\mu_k| \overset{*}{\rightharpoonup} \gamma$, and $\gamma(Q) = \lim_{k\to\infty}|\mu_k|(Q)$, then $\eta(Q) = \lim_{k\to\infty}\mu_k(Q)$. Here we have $Dz_k \overset{*}{\rightharpoonup} \eta$ and $\gamma(Q) = \lim_{k\to\infty}|Dz_k|(Q)$, so that $\eta(Q) = A_0$. On the other hand, $z_k \to u_0$ in $L^1(Q;\mathbb{R}^p)$, which implies that $Du_0 = \eta$ in Q, and so $Du_0(Q) = A_0$. Thus, $u_0(x) - A_0(x) = p(x_N) + c$, where $p(-1/2) = p(1/2) = 0$, and $u_0(x) - A_0x$ can be extended periodically to $\mathbb{R}^N$. Without loss of generality, we can assume that the trace of $\bar{u}_k$ equals the trace of u_0, so that $\bar{u}_k - A_0x$ can be extended periodically, and we call this extension w_k. Set $v_k^j(x) := A_0x + \frac{1}{j}w_k(jx)$ and

note that, for $x \in Q_j := (-\frac{1}{2j}, \frac{1}{2j})^N$, we have $\nabla v_k^j(x) = \nabla \bar{u}_k(jx)$. By (5.5) we may choose $k(j) > j$ such that $\mathcal{H}^{N-1}(S(v^j_{k(j)}) \cap Q) < 1/j$, and we have

$$v_j := v^j_{k(j)} \to A_0 x \text{ in } L^1(Q; \mathbb{R}^p)$$

and

$$\mathcal{H}^{N-1}(S(v_j) \cap Q) \to 0.$$

Furthermore,

$$\int_Q W^\infty(\nabla v_j(x))dx = j^N \int_{Q_j} W^\infty(\nabla \bar{u}_{k(j)}(jx))dx = \int_Q W^\infty(\nabla \bar{u}_{k(j)}(x))dx$$

and so, by (5.6), we need only show that

$$G^\infty(A_0) \le \limsup_{j\to\infty} \int_Q W^\infty(\nabla v_j(x))dx.$$

By (H3) we have

$$\begin{aligned} G^\infty(A_0) &\le \limsup_{t\to\infty} \left[\frac{1}{t} \limsup_{j\to\infty} \int_Q W(t\nabla v_j(x))dx\right] \\ &\le \limsup_{t\to\infty} \limsup_{j\to\infty} \left\{ \int_{\{t|\nabla v_j|>L\}} \left[W^\infty(\nabla v_j(x)) + \frac{C}{L^{m-1}t}\right] dx \right. \\ &\qquad \left. + \int_{\{t|\nabla v_j|\le L\}} \frac{1}{t} C_1[1+L]dx \right\} \\ &\le \limsup_{j\to\infty} \int_Q W^\infty(\nabla v_j(x))dx. \qquad \square \end{aligned}$$

6. Optimal jump microstructure for scalar valued functions. We now ask the question, what behavior is it necessary to allow for admissible functions for h? That is, how do infimizing sequences behave? Below, we answer this question for scalar valued functions. The idea is based on level sets, and so it is not straightforward to extend the result to vector valued functions.

Looking at the definition of $h(\rho, \nu)$ (see (2.4)), we see that admissible functions may have both jumps and nonzero gradient. Is it possible that there is an admissible function v that jumps and has nonzero gradient, and the energy of v is below the infimum over functions that just jump, and below the infimum over functions in $W^{1,1}$? The answer to this question is "yes," and we will see that a natural example illustrates the behavior of infimizing sequences.

We first consider the two-dimensional case, and the square in Figure 6.2 represents Q_ν for $N = 2$. Suppose that $CW^\infty(\rho\nu) \gg 1$ and $W^\infty(\rho\mu) \ll 1$ for some $\rho \in \mathbb{R}^+$ and unit vectors $\nu, \mu \in \mathbb{R}^2$, where $\nu \cdot \mu > 0$. We then see that a function that is 0 below $\Gamma := \Gamma_1 \cup \Gamma_2$ and ρ above, with a jump across Γ_1 and affine growth across a narrow neighborhood of Γ_2, has lower energy than the infimum over functions that just jump (this infimum is 1), and the infimum over functions in $W^{1,1}$ (this infimum is $CW^\infty(\rho\nu)$). Note that this example fails if CW^∞ is isotropic. We show that this

behavior is optimal. The idea is this: first, we give a coarea formula which allows us to consider, for any admissible function for $h(\rho,\nu)$, the bulk energy as an integral over measure-theoretic boundaries of level sets. We may then choose a "good" level set. Next, we prove that it is energetically better for the jump part of the boundary, i.e., $S(u)$ intersected with the boundary, to be connected and flat. As we will show in Lemma 6.1, we can assume that W^∞ is convex without changing the infimum of the energy, in which case we will prove that the remainder of the boundary might as well be flat, and we conclude that Figure 6.1 captures the geometry of minimizing sequences.

We begin with the following lemma.

LEMMA 6.1. *In the scalar case,* $h_W = h_{CW}$.

Proof. Since $CW \le W$, it follows that $h_{CW} \le h_W$. Conversely, let u be an admissible function for h_{CW}. By the relaxation theorem (Theorem 2.1), we have

$$I(u,Q) \le \int_Q CW(\nabla u)dx + \mathcal{H}^{N-1}(S(u)\cap Q),$$

where we use the fact that $h \le 1$. It also follows from Theorem 2.1 that

$$I_{CW}(u,Q) = \int_Q CW(\nabla u)dx + \int_{S(u)\cap Q} h_{CW}([u],\nu)d\mathcal{H}^{N-1}.$$

By the lower semicontinuity of I, we have

$$\begin{aligned}\int_Q C\ W(\nabla u)dx &+ \int_{S(u)\cap Q} h_W([u],\nu)d\mathcal{H}^{N-1}\\ &\le \int_Q CW(\nabla u)dx + \int_{S(u)\cap Q} h_{CW}([u],\nu)d\mathcal{H}^{N-1},\end{aligned}$$

which implies $h_W \le h_{CW}$. ∎

LEMMA 6.2. *Let λ be a finite Borel regular measure on Q and let $f: Q \to \mathbb{R}^N$ be λ measurable with $||f||_\infty < \infty$. Then there is a sequence $\{f_n\} \subset C_0^\infty(Q;\mathbb{R}^N)$ such that $f_n \to f$ λ almost everywhere and $||f_n||_\infty \le ||f||_\infty$ for all n.*

We now recall some notation: for $u \in BV(Q)$, set $E_t := \{x \in Q : u(x) > t\}$. For $x \in \partial_* E_t \subset Q$ (see (2.1)), we denote by $\nu_{E_t}(x)$ the measure theoretic unit *inner* normal (see Theorem 1, section 5.8 of [10]), so that

$$\int_{E_t} \mathrm{div}\phi(x)dx = -\int_{\partial_* E_t} \phi(x)\cdot\nu_{E_t}(x)d\mathcal{H}^{N-1}(x)$$

for all $\phi \in C_0^1(Q;\mathbb{R}^N)$.

LEMMA 6.3 (Coarea formula). *Let $u \in BV(Q)$ be given, and let $f: Q\times\mathbb{M}^{1\times N}\to\mathbb{R}$ be a Carathéodory function, where measurability is Borel, and positive homogeneous of degree one in the last variable. Assume further that $f(x, \frac{dDu(x)}{d|Du|(x)}) \in L^\infty(Q,|Du|)$. Then*

$$\int_Q f(x,dDu(x)) = \int_{\mathbb{R}}\int_{\partial_* E_t} f(x,\nu_{E_t}(x))d\mathcal{H}^{N-1}(x)dt. \tag{6.1}$$

Proof. First we note that, as a consequence of Borel regularity and the coarea formula for BV functions (see Theorem 1 (ii) of section 5.5 of [10]), we have that for any set $A \subset \Omega$,

$$|Du|(A) = 0 \text{ implies } \mathcal{H}^{N-1}(A \cap \partial_* E_t) = 0 \text{ for } \mathcal{L}^1 \text{ almost every t.} \tag{6.2}$$

We know (see claim 1 in the proof of Theorem 1, section 5.5 in [10]) that

$$\int_Q u(x)\mathrm{div}\phi(x)dx = \int_{\mathbb{R}} \int_{E_t} \mathrm{div}\phi(x)dxdt$$

for all $\phi \in C_0^1(Q;\mathbb{R}^N)$. Hence,

$$\int_Q \phi(x)\cdot\sigma(x)d|Du|(x) = \int_{\mathbb{R}} \int_{\partial_* E_t} \phi(x)\cdot\nu_{E_t}(x)d\mathcal{H}^{N-1}(x)dt \tag{6.3}$$

for all $\phi \in C_0^1(Q;\mathbb{R}^N)$, where $\sigma(x) := \frac{dDu(x)}{d|Du|(x)}$.

We now show that for $\mathcal{L}^1$ almost every $t \in \mathbb{R}$, we have

$$\sigma(x) = \nu_{E_t}(x) \text{ for } \mathcal{H}^{N-1} \text{ almost every } x \in \partial_* E_t. \tag{6.4}$$

Using Lemma 6.2, choose $\sigma_n \in C_0^1(Q;\mathbb{R}^2)$ such that $\sigma_n(x) \to \sigma(x)$ $|Du|$ almost everywhere (and so, by (6.2), $\mathcal{H}^{N-1}\lfloor\partial_* E_t$ almost everywhere for $\mathcal{L}^1$ almost every t) and $|\sigma_n| \le 1$. Note that $\sigma_n \cdot \nu_{E_t}$ is $\mathcal{H}^{N-1}\lfloor\partial_* E_t$ measurable since ν_{E_t} is, and

$$t \mapsto \int_{\partial_* E_t} \sigma_n(x)\cdot\nu_{E_t}(x)d\mathcal{H}^{N-1}(x) = \int_{E_t} \mathrm{div}\sigma_n(x)dx$$

is $\mathcal{L}^1$ measurable (see, e.g., the proof of Lemma 1 in section 5.5 of [10]). Then, by (6.3) and the dominated convergence theorem,

$$\begin{aligned}
\int_{\mathbb{R}} \int_{\partial_* E_t} \sigma(x)\cdot\nu_{E_t}(x)d\mathcal{H}^{N-1}(x)dt &= \lim_{n\to\infty} \int_{\mathbb{R}} \int_{\partial_* E_t} \sigma_n(x)\cdot\nu_{E_t}(x)d\mathcal{H}^{N-1}(x)dt \\
&= \lim_{n\to\infty} \int_Q \sigma_n(x)\cdot\sigma(x)d|Du|(x) \\
&= \int_Q d|Du| \\
&= \int_{\mathbb{R}} \int_{\partial_* E_t} d\mathcal{H}^{N-1}(x)dt.
\end{aligned}$$

Since $\sigma \cdot \nu_{E_t} \le 1$, we have (6.4).

Using Lemma 6.2 once more, choose $\phi_n \in C_0^1(Q;\mathbb{R}^N)$ such that $\phi_n(x) \to f(x,\sigma(x))\sigma(x)$ $|Du|$ almost everywhere and $||\phi_n||_\infty \le ||f(\cdot,\sigma(\cdot))||_\infty$. Then, as above,

$$\begin{aligned}
\int_Q f(x, dDu(x)) &= \int_Q f(x,\sigma(x))d|Du|(x) \\
&= \lim_{n\to\infty} \int_Q \phi_n(x)\cdot\sigma(x)d|Du|(x) \\
&= \lim_{n\to\infty} \int_{\mathbb{R}} \int_{\partial_* E_t} \phi_n(x)\cdot\nu_{E_t}(x)d\mathcal{H}^{N-1}(x)dt \text{ (by (6.3))} \\
&= \int_{\mathbb{R}} \int_{\partial_* E_t} f(x,\nu_{E_t}(x))d\mathcal{H}^{N-1}(x)dt. \text{ (by (6.4))} \qquad \square
\end{aligned}$$

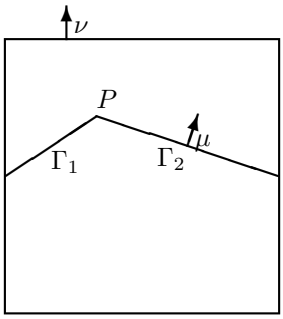

FIG. 6.1. *Admissible "function" for H.*

Now we introduce another infimum, similar to h, but which includes only very simple functions in its admissible class. Given $\nu \in S^{N-1}$, we consider the family $\mathcal{S}^2_\nu$ of squares with unit edge length, centered at zero, with two edges normal to ν (with ν in the plane of the square). Without loss of generality, we will assume $\nu = e_N$. Now consider the square $\mathcal{Q} \in \mathcal{S}^2_\nu$ with the remaining two edges having normal e_1 (in the plane of the square). We consider connected curves $\Gamma \subset \mathcal{Q}$ made up of two line segments such that Γ connects the midpoints of the edges that have normal e_1, i.e., $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_1 is the line segment from $(-1/2, 0, \dots, 0)$ to some point P in the square, and Γ_2 is the line segment connecting P to $(1/2, 0, \dots, 0)$ (see Figure 6.1). For other $\mathcal{Q} \in \mathcal{S}^2_\nu$, with two edges not having normal e_1, we consider analogous $\Gamma \subset \mathcal{Q}$. Set

$$H(\rho, \nu) := \inf\{\mathcal{H}^1(\Gamma_1) + \rho\mathcal{H}^1(\Gamma_2) CW^\infty(\mu) : \\ \mathcal{Q} \in \mathcal{S}^2_\nu, \Gamma \subset \mathcal{Q} \text{ is as above, and } \mu \text{ is the unit normal to } \Gamma_2 \text{ so that } \mu \cdot \nu \geq 0\}.$$

REMARK 6.4. Note the following:

i) $H(\rho, \nu) \leq 1$ since we can take $\Gamma = \Gamma_1$.

ii) The infimum H is attained since CW^∞ is continuous.

iii) If CW^∞ is isotropic, then the minimizing Γ equals Γ_1 or Γ_2.

THEOREM 6.5. $h = H$.

Proof. By Lemma 6.1, we may assume that $W = (CW)^\infty$. Furthermore, note that $(CW)^\infty$ is convex. Hence, in the sequel we will take W to be convex and positive homogeneous of degree one.

Step 1. We show that $h \leq H$.

Case a. $N = 2$.

Fix $\mathcal{Q} \in \mathcal{S}^2_\nu$ and $\Gamma \subset \mathcal{Q}$ as in the definition of H, and consider functions in $SBV(\mathcal{Q}; \mathbb{R})$ that are zero below Γ, ρ above, jump at Γ_1, and are affine across a narrow neighborhood of Γ_2, with another jump connection near the intersection of the boundary with Γ_2, and we see that these functions are admissible for $h(\rho, \nu)$ and their energy E approaches $\mathcal{H}^1(\Gamma_1) + \rho\mathcal{H}^1(\Gamma_2) W(\mu)$ (see Figure 6.2).

Case b. $N > 2$.

First, we point out the following:

$$h(\xi, \nu) = \inf\Big\{ \int_{Q_\nu} W^\infty(\nabla v) dx + \mathcal{H}^{N-1}(S(v) \cap Q_\nu) : v \in SBV_{loc}(S_\nu; \mathbb{R}^p), \\ v(y) = 0 \text{ if } y \cdot \nu = -\frac{1}{2}, v(y) = \xi \text{ if } y \cdot \nu = \frac{1}{2} \\ v \text{ is 1-periodic in the directions } \nu_1, \dots, \nu_{N-1} \Big\}. \tag{6.5}$$

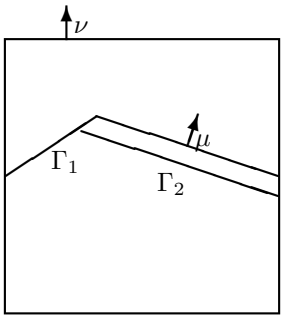

FIG. 6.2. $h \leq H$.

Clearly, $h \geq$ the right-hand side of (6) since admissible functions for h have the necessary periodicity. The other inequality follows from the fact that, after rescaling, an admissible function for the right-hand side of (6) will have the correct trace for h after altering it to jump on a set of arbitrarily small $\mathcal{H}^{N-1}$ measure.

Now, as in case a), let any $\mathcal{Q}$ and Γ as in the definition of H be given. Choose the cube Q_ν such that two sets of faces each have their normal in common with one of the faces of the square $\mathcal{Q}$. Extend Γ to Q_ν by $\bar{\Gamma} := \{x \in Q_\nu : \text{proj}(x) \text{ onto } \mathcal{Q} \in \Gamma\}$. We then construct functions as in the case $N = 2$, where there is a jump connection near the intersection of $\bar{\Gamma}_2$ with the face of Q_ν having the same normal as $\mathcal{Q}$. By (6), these functions have the necessary periodicity to be admissible for $h(\rho, \nu)$. Again, we see that the energy E approaches $\mathcal{H}^1(\Gamma_1) + \rho\mathcal{H}^1(\Gamma_2)W(\mu)$.

Step 2. We now show that $h \geq H$.

Let u be an admissible function for $h(\rho, \nu)$, i.e.,

$$u \in SBV(Q_\nu; \mathbb{R}), u = \rho \text{ if } x \in \partial Q_\nu \text{ and } x \cdot \nu \geq 0, \text{ and } u = 0 \text{ if } x \in \partial Q_\nu \text{ and } x \cdot \nu < 0.$$

For simplicity, we refer to Q_ν as Q. The goal is to find an appropriate curve $\Gamma = \Gamma_1 \cup \Gamma_2$ with H-energy no greater than the h-energy of u.

Step 2a. We first find a "good" level set. Applying Lemma 6.3 to

$$f(x, dDu(x)) := \chi_{Q \setminus S(u)}(x) W(dDu(x)),$$

so that $\left| f\left(x, \frac{dDu(x)}{d|Du|(x)}\right) \right| \leq C_1$, we get

$$\int_Q W(\nabla u(x)) dx = \int_{\mathbb{R}} \int_{\partial_* E_t \setminus S(u)} W(\nu_{E_t}(x)) d\mathcal{H}^{N-1}(x) dt.$$

Choose $t_0 \in (0, \rho)$ such that

$$\int_{\partial_* E_{t_0} \setminus S(u)} W(\nu_{E_{t_0}}(x)) d\mathcal{H}^{N-1}(x) \leq \frac{1}{\rho} \int_Q W(\nabla u(x)) dx. \tag{6.6}$$

Note that the coercivity of W guarantees that E_{t_0} has finite perimeter. Set

$$\beta := \mathcal{H}^{N-1}(\partial_* E_{t_0} \setminus S(u)) \text{ and } \bar{\nu} := \frac{1}{\beta} \int_{\partial_* E_{t_0} \setminus S(u)} \nu_{E_{t_0}}(x) d\mathcal{H}^{N-1}(x)$$

so that, by Jensen's inequality,

$$\beta W(\bar{\nu}) \leq \int_{\partial_* E_{t_0} \setminus S(u)} W(\nu_{E_{t_0}}(x)) d\mathcal{H}^{N-1}(x). \tag{6.7}$$

We can assume $\nu = e_N$ and $\bar{\nu} \cdot e_i = 0$ for $i \in \{2, \ldots, N-1\}$. Note that we can also assume that Q has e_1 normal to two of its faces, for the following reason: let Q_1 be a cube with normals e_N and e_1. We can rescale Q and u, and almost cover $\{x \in Q_1 : x_N = 0\}$ with the cubes $a_i + \delta Q$, where $a_i \in \{x \in Q_1 : x_N = 0\}$. Define $v \in SBV(Q_1)$ by

$$v(x) := \begin{cases} u\left(\dfrac{x - a_i}{\delta}\right) & \text{if } x \in a_i + \delta Q, \\ 0 & \text{if } x_N < 0 \text{ and } x \notin \cup(a_i + \delta Q), \\ \rho & \text{if } x_N \geq 0 \text{ and } x \notin \cup(a_i + \delta Q). \end{cases}$$

Using the homogeneity of W, we now have $E(v, Q_1) - E(u, Q) \leq \mathcal{H}^{N-1}(\{x \in Q_1 : x_N = 0\} \setminus \cup_i (a_i + \delta Q))$, yet $\bar{\nu}$, defined as for u, remains unchanged.

Step 2b. As suggested in our selection of t_0, we want to separate $\partial_* E_{t_0}$ into two pieces: its intersection with $S(u)$ and the rest. These will be turned into Γ_1 and Γ_2, respectively. By (6.6) and (6.7), it appears that a suitably long Γ_2 with normal in the direction $\bar{\nu}$ is energetically better than its counterpart, $\int_Q W(\nabla u) dx$. Referring to Figure 6.1, we see that the main issue is whether the segment Γ_1 needed to connect Γ_2 to the midpoint of the opposite edge has length larger than the $N-1$-dimensional area of $S(u)$, $\mathcal{H}^{N-1}(S(u))$. In fact, we can show that Γ_1 has length no larger than $\mathcal{H}^{N-1}(\partial_* E_{t_0} \cap S(u))$. What we first need to show is that $\partial_* E_{t_0}$ meets the midpoints of the opposing edges, and goes "all the way across" the cube. That is, we wish to show that

$$\int_{\partial_* E_{t_0}} \nu_{E_{t_0}}(x) \cdot e_1 d\mathcal{H}^{N-1}(x) = 0 \tag{6.8}$$

and

$$\int_{\partial_* E_{t_0}} \nu_{E_{t_0}}(x) \cdot e_N d\mathcal{H}^{N-1}(x) = 1. \tag{6.9}$$

From Theorem 1 (ii) of section 5.8 in [10], we know that if $E \subset \mathbb{R}^N$ has locally finite perimeter in $\mathbb{R}^N$, then

$$\int_E \mathrm{div}\phi(x) dx = -\int_{\partial_* E} \phi(x) \cdot \nu_E(x) d\mathcal{H}^{N-1}(x) \tag{6.10}$$

for all $\phi \in C_0^1(\mathbb{R}^N; \mathbb{R}^N)$, where, as before, ν_E is the *inner* unit normal. Define $E \subset \mathbb{R}^N$ by

$$E := E_{t_0} \cup \{x \in \mathbb{R}^N \setminus Q : x \cdot e_N > 0\}.$$

Then E is locally of finite perimeter in $\mathbb{R}^N$ and we claim that

$$\mathcal{H}^{N-1}([\partial_* E] \triangle C) = 0, \tag{6.11}$$

where $C := \partial_* E_{t_0} \cup \{x \in \mathbb{R}^N \backslash Q : x \cdot e_N = 0\}$. It is clear that $C \subset \partial_* E$ and that

$$(\partial_* E) \backslash C \subset \partial Q,$$

so the idea is to show that $\mathcal{H}^{N-1}(\partial Q \cap \partial_* E) = 0$. Let $x \in \partial Q$ be given such that $x \cdot e_N > 0$ and

$$\lim_{r \to 0} \fint_{B(x,r) \cap Q} |u(y) - \rho| dy = 0. \tag{6.12}$$

We need only show that

$$\limsup_{r \to 0} \frac{\mathcal{L}^N(B(x,r) \backslash E)}{r^N} = 0,$$

since then $x \notin \partial_* E$. We have

$$\begin{aligned} \limsup_{r \to 0} \frac{\mathcal{L}^N(B(x,r) \backslash E)}{r^N} &= \limsup_{r \to 0} \frac{1}{r^N} \mathcal{L}^N(B(x,r) \cap \{y \in Q : u(y) \leq t_0\}) \\ &\leq \limsup_{r \to 0} \frac{1}{r^N} \frac{1}{|\rho - t_0|} \int_{B(x,r) \cap Q} |u(y) - \rho| dy = 0. \end{aligned}$$

Since, by Theorem 2 of section 5.3 of [10], (6.12) holds $\mathcal{H}^{N-1}$ almost everywhere on the upper half of ∂Q, and, dealing with the case $x \cdot e_N < 0$ similarly, we find $\mathcal{H}^{N-1}(\partial Q \cap \partial_* E) = 0$.

Choose $\phi \in C_0^1(\mathbb{R}^N; \mathbb{R}^N)$ such that $\phi_i = 0$ for all $i \in \{2, \dots, N\}$, $\phi_1 = 1$ on Q. Clearly, $\mathrm{div}\phi = \partial \phi_1 / \partial x_1$, $\mathrm{div}\phi = 0$ on $\bar{Q}$, and $\nu_E = e_N$ on $\mathbb{R}^N \backslash \bar{Q}$. So, $\int_E \mathrm{div}\phi(x) dx = 0$. For example, we can take $\phi := \rho_{1/2} * \chi_{2Q} e_1$. By (6.10), (6.11), and the fact that E is locally of finite perimeter, we have

$$\begin{aligned} 0 &= \int_{\partial_* E} \phi(x) \cdot \nu_E(x) d\mathcal{H}^{N-1}(x) \\ &= \int_{\partial_* E_{t_0}} e_1 \cdot \nu_{E_{t_0}}(x) d\mathcal{H}^{N-1}(x) + \int_{\{x \cdot e_N = 0\} \backslash Q} \phi(x) \cdot e_N d\mathcal{H}^{N-1}(x) \\ &= \int_{\partial_* E_{t_0}} e_1 \cdot \nu_{E_{t_0}}(x) d\mathcal{H}^{N-1}(x), \end{aligned}$$

and we conclude (6.8).

Equation (6.9) follows by considering, for $\varepsilon > 0$, $\phi \in C_0^1((-1/2 - \varepsilon, 1/2 + \varepsilon)^N; \mathbb{R}^N)$ such that $\phi_i = 0$ for all $i < N$, $\phi = -e_N$ on Q and $-1 \leq \phi \cdot e_N \leq 0$. For example, take $\phi := -\rho_\varepsilon * \chi_{(-1/2-\varepsilon, 1/2+\varepsilon)^N} e_N$. By (6.10) and (6.11) we have

$$\int_E \mathrm{div}\phi(x) dx = \int_{\partial_* E_{t_0}} e_N \cdot \nu_{E_{t_0}}(x) d\mathcal{H}^{N-1}(x) - \int_{\{x \cdot e_N = 0\} \backslash Q} \phi(x) \cdot e_N d\mathcal{H}^{N-1}(x).$$

We see that

$$1 < \int_E \mathrm{div}\phi(x) dx < (1 + 2\varepsilon)^{N-1}$$

and

$$\left| \int_{\{x \cdot e_N = 0\} \setminus Q} \phi(x) \cdot e_N d\mathcal{H}^{N-1}(x) \right| < (1+2\varepsilon)^{N-1} - 1.$$

The arbitrariness of ε yields

$$\int_{\partial_* E_{t_0}} e_N \cdot \nu_{E_{t_0}}(x) d\mathcal{H}^{N-1}(x) = 1.$$

Step 2c. It remains to show that $\mathcal{H}^{N-1}(S(u))$ is larger than the length of Γ_1, which we construct below. The needed inequality is

$$\mathcal{H}^{N-1}(S(u)) \geq [(\beta\bar{\nu} \cdot e_1)^2 + (1 - \beta\bar{\nu} \cdot e_N)^2]^{1/2}, \tag{6.13}$$

which we now prove.

$$\begin{aligned}
\mathcal{H}^{N-1}(S(u)) &\geq \int_{\partial_* E_{t_0} \cap S(u)} |\nu_{E_{t_0}}(x)|^2 d\mathcal{H}^{N-1}(x) \\
&\quad \text{(since } |\nu_{E_{t_0}}| = 1 \ \ \mathcal{H}^{N-1} \text{ almost everywhere)} \\
&= \int_{\partial_* E_{t_0} \cap S(u)} \sum_{i=1}^{N} (\nu_{E_{t_0}}(x) \cdot e_i)^2 d\mathcal{H}^{N-1}(x) \\
&\geq \mathcal{H}^{N-1}(\partial_* E_{t_0} \cap S(u)) \left(\fint_{\partial_* E_{t_0} \cap S(u)} (\nu_{E_{t_0}}(x) \cdot e_1)^2 d\mathcal{H}^{N-1}(x) \right. \\
&\qquad \left. + \fint_{\partial_* E_{t_0} \cap S(u)} (\nu_{E_{t_0}}(x) \cdot e_N)^2 d\mathcal{H}^{N-1}(x) \right) \\
&\geq \mathcal{H}^{N-1}(\partial_* E_{t_0} \cap S(u)) \left(\left[\fint_{\partial_* E_{t_0} \cap S(u)} \nu_{E_{t_0}}(x) \cdot e_1 d\mathcal{H}^{N-1}(x) \right]^2 \right. \\
&\qquad \left. + \left[\fint_{\partial_* E_{t_0} \cap S(u)} \nu_{E_{t_0}}(x) \cdot e_N d\mathcal{H}^{N-1}(x) \right]^2 \right) \quad \text{(by Jensen's inequality)} \\
&= \mathcal{H}^{N-1}(\partial_* E_{t_0} \cap S(u))^{-1} \left(\left[\int_{\partial_* E_{t_0} \cap S(u)} \nu_{E_{t_0}}(x) \cdot e_1 d\mathcal{H}^{N-1}(x) \right]^2 \right. \\
&\qquad \left. + \left[\int_{\partial_* E_{t_0} \cap S(u)} \nu_{E_{t_0}}(x) \cdot e_N d\mathcal{H}^{N-1}(x) \right]^2 \right) \\
&= \mathcal{H}^{N-1}(\partial_* E_{t_0} \cap S(u))^{-1} \left([\beta\bar{\nu} \cdot e_1]^2 + [1 - \beta\bar{\nu} \cdot e_N]^2 \right), \\
&\qquad \text{(by (6.8) and (6.9))}
\end{aligned}$$

which gives (6.13).

Step 2d. We now construct Γ_1 and Γ_2. First, suppose that $\bar{\nu} \cdot e_N \leq 0$. Then (6.13) implies that $\mathcal{H}^{N-1}(S(u)) \geq 1$. Therefore, $E(u, Q) \geq 1 \geq H(\rho, \nu)$. Assume now

that $\bar{\nu} \cdot e_N > 0$. Consider the square in the e_1-e_N plane with normals e_1 and e_N and, suppressing e_i for $i \in \{2, \dots, N-1\}$, take $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_2 is the line segment with right endpoint $(1/2, 0)$, unit normal $\bar{\nu}/|\bar{\nu}|$, and length $|\bar{\nu}|\beta$ (if $\beta\bar{\nu}\cdot e_N \geq 1$, redefine $\beta := (1-\varepsilon)/(\bar{\nu} \cdot e_N)$). Γ_1 is then the line segment from the left endpoint of Γ_2 to $(-1/2, 0)$. Note that the length of Γ_1 is $[(\beta\bar{\nu} \cdot e_1)^2 + (1 - \beta\bar{\nu} \cdot e_N)^2]^{1/2}$, and so, by (6.13), we have $\mathcal{H}^1(\Gamma_1) \leq \mathcal{H}^{N-1}(S(u))$. Finally, by (6.6) and (6.7), we conclude that

$$\begin{aligned} H(\rho, \nu) &\leq \rho|\bar{\nu}|\beta W\left(\frac{\bar{\nu}}{|\bar{\nu}|}\right) + \mathcal{H}^1(\Gamma_1) \\ &\leq \rho\beta W(\bar{\nu}) + \mathcal{H}^{N-1}(S(u)) \\ &\leq \int_Q W(\nabla u(x))dx + \mathcal{H}^{N-1}(S(u)). \end{aligned}$$

Due to the arbitrariness of u, we have $h \geq H$. □

We conclude with the following remark, which simply says that the above argument works for more general initial jump energy densities.

REMARK 6.6. Suppose that the energy of the admissible functions for h is given by

$$E(u, Q) = \int_Q W(\nabla u)dx + \int_{S(u)} \phi([u]\nu)d\mathcal{H}^{N-1},$$

where W and ϕ are convex and positive homogeneous of degree one. Then the conclusion of Theorem 6.5 holds. Taking

$$f(x, dDu) := \chi_{Q\setminus S(u)}W(dDu) + \chi_{S(u)}\phi(dDu)$$

we have

$$\int_Q f(x, dDu) = \int_Q W(\nabla u)dx + \int_{S(u)} \phi([u]\nu)d\mathcal{H}^{N-1} = E(u, Q),$$

and, by Lemma 6.3,

$$E(u, Q) = \int_{\mathbb{R}} \left[\int_{\partial_* E_t \setminus S(u)} W(\nu_{E_t}(x))d\mathcal{H}^{N-1} + \int_{\partial_* E_t \cap S(u)} \phi(\nu_{E_t}(x))d\mathcal{H}^{N-1} \right] dt.$$

The rest of the proof of Theorem 6.5 follows with the obvious alterations.

Acknowledgments. The author would like to thank Luigi Ambrosio and the referees for comments and suggestions on earlier versions.

REFERENCES

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

[2] G. ALBERTI, *Rank-one property for derivatives of functions with bounded variation*, Proc. Royal Soc. Edinburgh, Sect. A, 123 (1993), pp. 239–274.

[3] L. AMBROSIO, *On the lower semicontinuity of quasiconvex integrals in $SBV(\Omega; \mathbb{R}^k)$*, Nonlinear Anal., 23 (1994), pp. 405–425.

[4] L. AMBROSIO AND A. BRAIDES, *Functionals defined on partitions of sets of finite perimeter* I: *Integral representation and G-convergence*, J. Math. Pures Appl., 69 (1990), pp. 285–305.

[5] L. Ambrosio and A. Braides, *Functionals defined on partitions of sets of finite perimeter* II: *Semicontinuity, relaxation and homogenization*, J. Math. Pures Appl., 69 (1990), pp. 307–333.
[6] A. C. Barroso, G. Bouchitté, G. Buttazzo, and I. Fonseca, *Relaxation of Bulk and Interfacial Energies*, Arch. Rational Mech. Anal., 135 (1996), pp. 107–173.
[7] G. Bouchitté, A. Braides, and G. Buttazzo, *Relaxation results for some free discontinuity problems*, J. Reine Angew. Math., 458 (1995), pp. 1–18.
[8] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.
[9] E. De Giorgi and L. Ambrosio, *Un nuovo tipo di funzionale del calcolo delle variazioni*, Atti. Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 82 (1988), pp. 199–210.
[10] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[11] I. Fonseca and G. Francfort, *Relaxation in BV versus quasiconvexification in $W^{1,p}$: A model for the interaction between damage and fracture*, Calc. Var., 4 (1995), pp. 407–446.
[12] I. Fonseca and J. Malý, *Relaxation of multiple integrals below the growth exponent*, Ann. Inst. H. Poincaré, 7/14 (1997), pp. 309–338.
[13] I. Fonseca and S. Müller, *Quasiconvex integrands and lower semicontinuity in L^1*, SIAM J. Math. Anal., 23 (1992), pp. 1081–1098.
[14] I. Fonseca and S. Müller, *Relaxation of quasiconvex functionals in $BV(\Omega;\mathbb{R}^p)$ for integrands $f(x,u,\nabla u)$*, Arch. Rational Mech. Anal., 123 (1993), pp. 1–49.
[15] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1966.
[16] A. I. Vol'pert, *Spaces BV and quasi-linear equations*, Math. USSR-Sb., 17 (1969), pp. 225–267.
[17] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 849–854, July 1998

002

A NOTE ON MATRIX REFINEMENT EQUATIONS*

THOMAS A. HOGAN†

Abstract. Refinement equations involving matrix masks are receiving much of attention these days. They can play a central role in the study of refinable finitely generated shift-invariant spaces, multiresolutions generated by more than one function, multiwavelets, splines with multiple knots, and matrix subdivision schemes—including Hermite-type subdivision schemes. Several recent papers on this subject begin with an assumption on the eigenstructure of the mask, pointing out that this assumption is heuristically "natural" or "preferred." In this note, we prove that stability of the shifts of the refinable function requires this assumption.

Key words. matrix refinement equation, matrix subdivision scheme, refinable function vector, stability, Riesz basis, multiwavelet, shift-invariant space, FSI space

AMS subject classifications. 39A10, 39B62, 42B99

PII. S003614109630135X

1. Introduction.

Several desirable properties are not available with compactly supported orthogonal wavelets, e.g., symmetry and piecewise polynomial structure. Presently, multiwavelets seem to offer a satisfactory alternative (see, for example, [DGHM], [GL]). Multiwavelets are wavelets constructed from a refinable function vector Φ which satisfies a matrix refinement equation of the form

$$\Phi = \sum_{\alpha \in \mathbb{Z}^d} a(\alpha)\Phi(M^T \cdot -\alpha).$$

Here, each coefficient $a(\alpha)$ is a $\Phi \times \Phi$ matrix, and $M \in \mathbb{Z}^{d\times d}$. Refinable function vectors have also appeared in the study of matrix subdivision schemes, which play an important role in the analysis of multivariate subdivision schemes (cf. [D]).

As in the case of a single refinable function, it is often impossible to study a refinable function vector directly. In such a case, its properties are analyzed indirectly via the coefficient sequence $(a(\alpha))_{\alpha\in\mathbb{Z}^d}$ (see, for example, [CDP], [HC], [HSS], [H], [P], [S]). For example, the eigenstructure of the matrix

$$A(0) = \sum_{\alpha \in \mathbb{Z}^d} a(\alpha)$$

has played an important role in such analyses. In particular, it is assumed in [HSS] and [S], that 1 is a simple eigenvalue of $A(0)$ and that all other eigenvalues have modulus strictly less than 1. In this paper, we demonstrate that this is a very reasonable assumption by proving that without such an assumption, the refinable function vector Φ cannot possibly have stable shifts.

A slightly weaker statement has already been proved by Cohen, Dyn, and Levin in [CDL] for ℓ^∞-stability. In that paper, they assumed that $a(\alpha) = 0$ for all but finitely many α and that the associated subdivision scheme is C^0, i.e., convergent. In this paper we strengthen and extend their results.

*Received by the editors April 1, 1996; accepted for publication May 27, 1997; published electronically March 25, 1998.

http://www.siam.org/journals/sima/29-4/30135.html

†Department of Mathematical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada (thogan@vega.math.ualberta.ca).

To present our results in a general setting, we recall the following definition from [JM]:

$$\mathcal{L}^p := \mathcal{L}^p(\mathbb{R}^d) := \left\{ \phi : \mathbb{R}^d \to \mathbb{C} \;\middle|\; |\phi|_p := \|\widetilde{\phi}\|_{L^p([0,1)^d)} < \infty \right\}$$

for $1 \le p \le \infty$, where $\widetilde{\phi} := \sum_{\alpha \in \mathbb{Z}^d} |\phi(\cdot - \alpha)|$. As pointed out in [JM], $\mathcal{L}^p$ is a Banach space with norm $|\cdot|_p$ and

$$\mathcal{L}^p \subset L^1 \cap L^p.$$

Now, let $M \in \mathbb{Z}^{d\times d}$ be an integer matrix satisfying $\lim_{k\to\infty} M^{-k} = 0$ and let $\phi_1, \dots, \phi_m \in \mathcal{L}^p$. We say that $\Phi := (\phi_j)_{j=1}^m$ is *M-refinable* if there exist sequences $a_{j,k} \in \ell^1(\mathbb{Z}^d) \quad (1 \le j,k \le m)$ such that

$$\phi_j = \sum_{k=1}^m \sum_{\alpha\in\mathbb{Z}^d} a_{j,k}(\alpha)\phi_k(M^T \cdot - \alpha), \qquad (j = 1, \dots, m).$$

Equivalently, Φ is refinable if

$$\widehat{\Phi}(M\xi) = A(\xi)\widehat{\Phi}(\xi) \quad \text{for all } \xi \in \mathbb{R}^d, \tag{1.1}$$

where the matrix $A := (A_{j,k})_{1\le j,k\le m}$ of (continuous 2π-periodic) functions is defined by

$$A_{j,k}(\xi) := \frac{1}{|\det M|} \sum_{\alpha\in\mathbb{Z}^d} a_{j,k}(\alpha) e^{-i\langle\alpha,\xi\rangle}.$$

The matrix A is referred to as the *(refinement) mask*.

It is already well known that equation (1.1) has only the trivial solution $\Phi = 0$ if the spectral radius $\rho(A(0)) < 1$. It is also well known that convergence of the infinite product

$$P := \prod_{j=1}^{\infty} A(M^{-j}\cdot) \tag{1.2}$$

requires that (i) $\rho(A(0)) \le 1$; (ii) 1 be the only eigenvalue of modulus 1; and (iii) the algebraic and geometric multiplicities of the eigenvalue 1 be the same. When this product does converge, the function Φ defined by $\widehat{\Phi}(\xi) = P(\xi)x$ is a solution to equation (1.1) for any $x \in \mathbb{C}^m$. Convergence of the matrix subdivision scheme associated with equation (1.1) requires similar assumptions on A (cf. [CDL]). Nonetheless, solutions to equation (1.1) may exist even without such assumptions (as pointed out in [HC] and [CDP]). However, the existence of solutions with stable shifts will require these assumptions and more.

The shifts of $\phi_1, \dots, \phi_m \in \mathcal{L}^p$ are said to be *ℓ^p-stable* if there exist constants $0 < c_1 \le c_2 < \infty$ such that

$$c_1 \sum_{j=1}^m \|a_j\|_{\ell^p} \le \| \sum_{j=1}^m \sum_{\alpha\in\mathbb{Z}^d} a_j(\alpha)\phi_j(\cdot - \alpha)\|_{L^p} \le c_2 \sum_{j=1}^m \|a_j\|_{\ell^p}$$

for any $a_1, \dots, a_m \in \ell^p(\mathbb{Z}^d)$. In [JM], Jia and Micchelli proved that the shifts of any $\phi_1, \dots, \phi_m \in \mathcal{L}^p$ are ℓ^p-stable if and only if the sequences

$$\left(\widehat{\phi}_j(\xi + 2\alpha\pi)\right)_{\alpha\in\mathbb{Z}^d}, \qquad (j = 1, \dots, m)$$

are linearly independent for every $\xi \in \mathbb{R}^d$. This will play the major role in our proofs.

In the statement of our theorems, we use the following terminology. An eigenvalue is *nondegenerate* if its algebraic and geometric multiplicities agree. A *simple* eigenvalue is a nondegenerate eigenvalue of multiplicity 1.

To facilitate our proofs, we define

$$V := \{\, v \in \mathbb{Z}^d \mid v = Mt \text{ for some } t \in [0,1)^d \,\}.$$

Then V is a complete set of representatives for the quotient group $\mathbb{Z}^d/M\mathbb{Z}^d$. In particular, $\mathbb{Z}^d$ is the disjoint union of the sets $v + M\mathbb{Z}^d$ ($v \in V$). We will actually only ever make use of the set $V' := V\backslash 0$.

2. Stability imposes structure.

THEOREM 2.1. *Let $\phi_1, \ldots, \phi_m \in \mathcal{L}^p$. Suppose $\Phi := (\phi_j)_{j=1}^m$ is M-refinable with mask A. If the shifts of $\phi_1, \ldots, \phi_m$ are ℓ^p-stable, then* 1 *is a simple eigenvalue of $A(0)$ and all other eigenvalues have modulus strictly less than* 1. *Moreover, $\widehat{\Phi}(0)$ is a right* 1*-eigenvector.*

Proof. We assume stability to demonstrate the eigenvalue assertions.

By the refinement equation (1.1), we have, for any $\xi \in \mathbb{R}^d$ and $n \in \mathbb{N}$,

$$\widehat{\Phi}(\xi) = \prod_{j=1}^{k} A(M^{-j}\xi)\widehat{\Phi}(M^{-k}\xi).$$

Since $M^{-k} \to 0$ (and since A and $\widehat{\Phi}$ are both continuous), $\rho(A(0)) < 1$ would imply that Φ is identically zero—contradicting the assumption that the shifts of Φ are stable. So, $\rho(A(0)) \geq 1$.

Now, suppose $y \in \mathbb{C}^m$ satisfies $y^T A(0) = \mu y^T \neq 0$ for some $\mu \in \mathbb{C}$ with $|\mu| \geq 1$. Then the refinement equation (1.1) implies that

$$\begin{aligned} y^T\widehat{\Phi}(2M^k(M\alpha + v)\pi) &= y^T A^k(0) A(2M^{-1}v\pi)\widehat{\Phi}(2M^{-1}v\pi + 2\alpha\pi) \\ &= \mu^k y^T A(2M^{-1}v\pi)\widehat{\Phi}(2M^{-1}v\pi + 2\alpha\pi) \end{aligned} \tag{2.1}$$

for any $k \in \mathbb{Z}_+$, $\alpha \in \mathbb{Z}^d$, $v \in V'$. Since $v \in V'$ (hence $M\alpha + v \neq 0$), our assumptions on M imply that $\lim_{k\to\infty} |M^k(M\alpha + v)| = \infty$. Since $y^T\Phi \in \mathcal{L}^p \subset L^1$, the left-hand side of equation (2.1) then tends to zero as k tends to infinity. And, since $|\mu| \geq 1$, this implies that

$$y^T A(2M^{-1}v\pi)\widehat{\Phi}(2M^{-1}v\pi + 2\alpha\pi) = 0$$

for every $\alpha \in \mathbb{Z}^d$ (which implies that $y^T A(2M^{-1}v\pi) = 0$ for every $v \in V'$, since the shifts of Φ are stable). Together with equation (2.1), this implies that $y^T\widehat{\Phi}(2\beta\pi) = 0$ for all $\beta \in \mathbb{Z}^d\backslash 0$, since every such β has a (unique) representation of the form $\beta = M^k(M\alpha + v)$ for some $k \in \mathbb{Z}_+$, $\alpha \in \mathbb{Z}^d$, $v \in V'$. The stability of the shifts of Φ then implies that $y^T\widehat{\Phi}(0) \neq 0$ and, a fortiori, $\widehat{\Phi}(0) \neq 0$. The refinement equation (1.1) then implies that $\widehat{\Phi}(0)$ is a right 1-eigenvector of $A(0)$.

Now, suppose that $y_1^T A(0) = \mu_1 y_1^T \neq 0$ and $y_2^T A(0) = \mu_2 y_2^T \neq 0$ with $|\mu_i| \geq 1$ ($i = 1, 2$). The above arguments imply that $y_i^T\widehat{\Phi}(2\beta\pi) = 0$ $\forall \beta \in \mathbb{Z}^d\backslash 0$ and $y_i^T\widehat{\Phi}(0) \neq 0$; without loss of generality, we may assume that $y_i^T\widehat{\Phi}(0) = 1$. Then $(y_2 - y_1)^T\widehat{\Phi}(2\beta\pi) = 0$ for every $\beta \in \mathbb{Z}^d$. The stability of the shifts of Φ now implies that $y_1 = y_2$.

We conclude that 1 is an eigenvalue of $A(0)$ of *geometric* multiplicity 1. It is the only eigenvalue outside of the open unit disc. Its (unique-up-to-multiplicity) right eigenvector is $\widehat{\Phi}(0)$ and its (unique-up-to-multiplicity) left eigenvector, y^T, satisfies $y^T\widehat{\Phi}(2\beta\pi) = 0$ for all $\beta \in \mathbb{Z}^d\backslash 0$. If the algebraic multiplicity of the eigenvalue 1 were greater than 1, then the (one-dimensional) left and right eigenspaces would be orthogonal one to the other (this follows easily by considering the Jordan canonical form of $A(0)$). That is, $y^T\widehat{\Phi}(0)$ would be zero—contradicting the assumption that the shifts of Φ are stable. □

Remark. The above proof in fact implies that the left eigenvector y^T of $A(0)$ actually satisfies the "sum rules"

$$y^T \sum_{\alpha\in\mathbb{Z}^d} a(\beta + M^T\alpha) = y^T \quad \forall\, \beta \in \mathbb{Z}^d,$$

as well as the so-called Strang–Fix conditions of order 1

$$y^T\widehat{\Phi}(0) \neq 0, \qquad y^T\widehat{\Phi}(2\beta\pi) = 0 \quad \forall\, \beta \in \mathbb{Z}^d\backslash 0.$$

So, stability implies accuracy of order 1 (or density) as expected.

3. Stability of matrix functions. A generalized stability notion for matrix functions has been recently considered in [CDL]. In the spirit of that paper, we will say that the shifts of any $m \times n$ matrix $\Phi = (\phi_{j,k})$ of $\mathcal{L}^p$-functions are *ℓ^p-stable* if there exist constants $0 < c_1 \le c_2 < \infty$ such that

$$c_1 \sum_{j=1}^m \|a_j\|_{\ell^p} \le \sum_{k=1}^n \left\| \sum_{j=1}^m \sum_{\alpha\in\mathbb{Z}^d} a_j(\alpha)\phi_{j,k}(\cdot - \alpha) \right\|_{L^p} \le c_2 \sum_{j=1}^m \|a_j\|_{\ell^p}$$

for any $a_1, \dots, a_m \in \ell^p(\mathbb{Z}^d)$.

Many of the results from [JM] can be generalized to cover this notion. To state some pertinent ones, we first recall some of their notation. We denote the d-dimensional torus

$$\left\{ (z_1, \dots, z_d) \in \mathbb{C}^d \;\middle|\; |z_1| = \cdots = |z_d| = 1 \right\}$$

by $\mathbb{T}^d$. Then, for any $f, g \in \mathcal{L}^2$, define

$$[f, g](z) := \sum_{\alpha\in\mathbb{Z}^d} \langle f, g(\cdot - \alpha)\rangle z^\alpha, \qquad (z \in \mathbb{T}^d),$$

where $\langle f, g\rangle := \int_{\mathbb{R}^d} f\overline{g}$ for $f, g \in L^2(\mathbb{R}^d)$. And, lastly, for $\phi_1, \dots, \phi_m \in \mathcal{L}^p$, define

$$\mathcal{S}^1(\phi_1, \dots, \phi_m) := \left\{ \sum_{j=1}^m \sum_{\alpha\in\mathbb{Z}^d} a_j(\alpha)\phi_j(\cdot - \alpha) \;\middle|\; a_j \in \ell^1(\mathbb{Z}^d) \text{ for } j = 1, \dots, m \right\}.$$

It is worth pointing out that $[f, g](z)$ is a continuous function of z on $\mathbb{T}^d$ and that $\mathcal{S}^1(\phi_1, \dots, \phi_m)$ is a subspace of $\mathcal{L}^p(\mathbb{R}^d)$.

A generalized statement of [JM, Theorem 4.1] follows.

THEOREM 3.1. *Let $\phi_{j,k} \in \mathcal{L}^2(\mathbb{R}^d)$, $(j = 1, \dots, m; k = 1, \dots, n)$. Then the shifts of $\Phi = (\phi_{j,k})$ are ℓ^2-stable if and only if one of the following conditions holds:*

(i) *For any* $\xi \in \mathbb{R}^d$, *the sequences* $(\widehat{\phi}_{j,k}(\xi + 2\alpha\pi))_{k=1,\alpha\in\mathbb{Z}^d}^n$ $(j = 1, \dots, m)$ *are linearly independent.*
(ii) *The matrix* $(\sum_{k=1}^n [\phi_{j,k}, \phi_{\ell,k}](z))_{1\le j,\ell\le m}$ *is positive definite for every* $z \in \mathbb{T}^d$.
(iii) *There exist* $g_{j,k} \in \mathcal{S}^1(\phi_{1,k}, \dots, \phi_{m,k})$ $(j = 1, \dots, m; k = 1, \dots, n)$ *such that*

$$\sum_{k=1}^{n} \langle g_{j,k}, \phi_{\ell,k}(\cdot - \alpha)\rangle = \delta_{j\ell}\delta_{0\alpha} \quad \text{for } 1 \le j, \ell \le m \quad \text{and } \alpha \in \mathbb{Z}^d,$$

and [JM, Theorem 4.2] is generalized as follows.

THEOREM 3.2. *Let* $\phi_{j,k} \in \mathcal{L}^p(\mathbb{R}^d)$, $(j = 1, \dots, m; k = 1, \dots, n)$. *Then the shifts of* $\Phi = (\phi_{j,k})$ *are* ℓ^p*-stable if and only if condition* (i) *of Theorem* 3.1 *holds.*

The proofs of these theorems are clear from the proofs of [JM, Theorems 3.3, 3.5, and 4.1]. We now state a generalization of Theorem 2.1. The proof is similar, so we provide only the major distinctions below.

THEOREM 3.3. *Let* $\phi_{j,k} \in \mathcal{L}^p(\mathbb{R}^d)$, $(j = 1, \dots, m; k = 1, \dots, n)$. *Suppose* $\Phi = (\phi_{j,k})$ *is* M*-refinable with mask* A. *If the shifts of* Φ *are* ℓ^p*-stable, then* 1 *is a nondegenerate eigenvalue of* $A(0)$*; its multiplicity is the rank of the matrix* $\widehat{\Phi}(0) = \big(\widehat{\phi}_{i,j}(0)\big)$, *and all other eigenvalues have modulus strictly less than* 1. *In particular, the columns of* $\widehat{\Phi}(0)$ *must span the right* 1*-eigenspace of* $A(0)$.

Proof. Define

$$W := \{\, y \in \mathbb{C}^m \mid y^T A(0) = \mu y^T \text{ for some } |\mu| \ge 1 \,\} \text{ and } X := \{\, x \in \mathbb{C}^m \mid A(0)x = x \,\}.$$

Then, $\operatorname{rank}\widehat{\Phi}(0) \le \dim X \le \dim W$, since every nonzero column of $\widehat{\Phi}(0)$ is a right 1-eigenvector of $A(0)$.

As in the proof of Theorem 1, if $y \in W$, then $y^T\widehat{\Phi}(2\beta\pi) = 0$ for all $\beta \in \mathbb{Z}^d\backslash 0$. If the shifts of Φ are stable, then $y^T\widehat{\Phi}(0) \ne 0$ for every $y \in W$. This implies that $\dim W \le \operatorname{rank}\widehat{\Phi}(0)$; hence both must equal the geometric multiplicity of the eigenvalue 1. In particular, all other eigenvalues have modulus strictly less than 1, and the columns of $\widehat{\Phi}(0)$ span the right 1-eigenspace.

If the algebraic multiplicity of the eigenvalue 1 is greater than its geometric multiplicity, then there exists a left 1-eigenvector y for which $y^T x = 0$ for all $x \in X$. Such y satisfies $y^T\widehat{\Phi}(2\alpha\pi) = 0$ for all $\alpha \in \mathbb{Z}^d$, and the shifts of Φ are not stable. □

We can say even more, under slightly more restrictive assumptions, on the sequences $(a_{j,k}(\alpha))_{\alpha\in\mathbb{Z}}$. Suppose, for example, that each of these sequences decays exponentially fast; then the entries of the matrix A are analytic functions. Now, if Φ is a matrix solution to the refinement equation (1.1) and the shifts of Φ are stable, then the arguments of [HC] (and the consequences of Theorem 3.3) imply that the infinite matrix product (1.2) is convergent and that the map $v \mapsto Pv$ is an isomorphism from the right 1-eigenspace of $A(0)$ onto the (vector) solution space of the refinement equation (1.1). Hence this solution space is already spanned by some N of the columns of Φ, where N is the multiplicity of the eigenvalue 1 of $A(0)$. This leads to the following theorem.

THEOREM 3.4. *Suppose some (matrix) solution of the refinement equation* (1.1) *has* ℓ^p*-stable shifts. Then a given solution* Φ *has* ℓ^p*-stable shifts if and only if the columns of* $\widehat{\Phi}(0)$ *span the right* 1*-eigenspace of* $A(0)$.

Acknowledgment. The author is grateful to Nira Dyn for discussions (at the *Program on Spline Functions and the Theory of Wavelets* in Montréal) which motivated this work.

REFERENCES

[CDL] A. Cohen, N. Dyn, and D. Levin, *Stability and inter-dependence of matrix subdivision schemes*, in Advanced Topics in Multivariate Approximation, F. Fontanella, K. Jetter, and P.-J. Laurent, eds., World Scientific Publishing Co., Singapore, 1996, pp. 33–45.

[CDP] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.

[DGHM] G. C. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1158–1192.

[D] N. Dyn, *Subdivision schemes in CAGD*, in Advances in Numerical Analysis Vol. II: Wavelets, Subdivision Algorithms and Radial Basis Functions, W. A. Light, ed., Oxford University Press, Oxford, 1992, pp. 36–104.

[GL] T. N. T. Goodman and S. L. Lee, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.

[HC] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.

[HSS] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., 73 (1996), 75–94.

[H] T. A. Hogan, *Stability and independence of the shifts of finitely many refinable functions*, J. Fourier Anal. Appl., 6 (1997), to appear.

[JM] R.-Q. Jia and C. A. Micchelli, *Using the refinement equations for the construction of pre-wavelets II: Powers of two*, in Curves and Surfaces, P.-J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.

[P] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.

[S] Z. Shen, *Refinable function vectors*, SIAM J. Math. Anal., 29 (1998), pp. 235–250.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 855–877, July 1998

003

ON THE REGULARIZATION OF FREDHOLM INTEGRAL EQUATIONS OF THE FIRST KIND *

ENRICO DE MICHELI†, NICODEMO MAGNOLI‡, AND GIOVANNI ALBERTO VIANO‡

Abstract. In this paper the problem of recovering a regularized solution of the Fredholm integral equations of the first kind with Hermitian and square-integrable kernels, and with data corrupted by additive noise, is considered. Instead of using a variational regularization of Tikhonov type, based on a priori global bounds, we propose a method of truncation of eigenfunction expansions that can be proved to converge asymptotically, in the sense of the L^2-norm, in the limit of noise vanishing. Here we extend the probabilistic counterpart of this procedure by constructing a probabilistically regularized solution without assuming any structure of order on the sequence of the Fourier coefficients of the data. This probabilistic approach allows us to use the statistical tools proper of time-series analysis, and in this way we attain a new regularizing algorithm, which is illustrated by some numerical examples. Finally, a comparison with solutions obtained by the means of the variational regularization exhibits how some intrinsic limits of the variational-based techniques can be overcome.

Key words. integral equations, inverse problems, regularization, information theory

AMS subject classification. 45B05

PII. S0036141096301749

1. Introduction.

We consider the Fredholm integral equations of the first kind

$$(Af)(x) = \int_a^b K(x,y)f(y)\,dy = g(x) \qquad (a \le x \le b) \tag{1}$$

whose kernel $K(x,y)$ is supposed to be Hermitian and square integrable; i.e.,

$$K(x,y) = \overline{K(y,x)} \tag{2}$$

and

$$\int_a^b \left\{ \int_a^b |K(x,y)|^2 dx \right\} dy < \infty. \tag{3}$$

Then $A : L^2(a,b) \to L^2(a,b)$ is a self-adjoint compact operator.

For simplicity we shall suppose hereafter that the kernel K, the function g, and the unknown function f are real-valued functions; in addition, we assume that the interval $[a,b]$ is a bounded and closed subset of the real line.

The Hilbert–Schmidt theorem guarantees that the integral operator A admits a set of eigenfunctions $\{\psi_k\}_1^\infty$ and, accordingly, a countably infinite set of eigenvalues $\{\lambda_k\}_1^\infty$. The eigenfunctions form an orthonormal basis of the orthogonal complement of the null space of the operator A and therefore an orthonormal basis of $L^2(a,b)$ when A is injective. For the sake of simplicity only this case will be considered, although this assumption can be easily relaxed with slight technical modifications.

* Received by the editors April 7, 1996; accepted for publication (in revised form) July 22, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/30174.html

† Istituto di Cibernetica e Biofisica, Consiglio Nazionale delle Ricerche, Via De Marini 6, 16149 Genova, Italy (demic@icb.ge.cnr.it).

‡ Dipartimento di Fisica, Università di Genova, Istituto Nazionale di Fisica Nucleare, sez. di Genova, Via Dodecaneso 33, 16146 Genova, Italy (magnoli@ge.infn.it, viano@ge.infn.it).

The Hilbert–Schmidt theorem also guarantees that $\lim_{k\to\infty} \lambda_k = 0$. Furthermore, we shall suppose hereafter that the eigenvalues are ordered as follows: $\lambda_1 > \lambda_2 > \lambda_3 > \cdots$ In view of the Hilbert–Schmidt theorem we associate with the integral equation (1) the following eigenfunction expansion:

$$f(x) = \sum_{k=1}^{\infty} \left(\frac{g_k}{\lambda_k} \right) \psi_k(x), \tag{4}$$

where $g_k = (g, \psi_k)$, $((\cdot,\cdot)$ denoting the scalar product in $L^2(a,b))$. The series (4) converges in the sense of L^2.

Remark. If the support of the data does not coincide with that of the solutions, i.e., $A : L^2(a,b) \to L^2(c,d)$ with $[a,b]$ different from $[c,d]$, the problem can be worked out in terms of singular values and singular functions of the operator A [6], and all of the following results can be easily reformulated.

In view of the fact that there always exists some inherent noise in the data, instead of (1) we have to deal with the following equation:

$$Af + n = \overline{g} \qquad (\overline{g} = g + n), \tag{5}$$

where n represents the noise. Therefore, instead of expansion (4) we have to consider the following expansion:

$$\sum_{k=1}^{\infty} \left(\frac{\overline{g}_k}{\lambda_k} \right) \psi_k, \tag{6}$$

where $\overline{g}_k = (\overline{g}, \psi_k)$. Expansion (6) is generally diverging because $\overline{g}$ does not belong, in general, to the range of the operator A. This is precisely a manifestation of the ill-posed character of the Fredholm integral equation of the first kind.

Several methods of regularization have been proposed (see [10, 14, 16] and references therein); all of them modify one of the elements of the triplet $\{A, X, Y\}$, where A is the integral operator defined by (1), whereas X and Y are, respectively, the solution and the data space (in our case $X \equiv Y \equiv L^2(a,b)$). Among these methods the procedure, which is probably the most popular, consists in admitting only those solutions that belong to a compact subset of the solution space X. In particular the famous method of Tikhonov leads to the construction of "regularizing operators" by the minimization of "smoothing functionals." In this latter functional the smoothing term is obtained precisely by restricting the admitted solutions to a compact subset of the space X; then the continuity of A^{-1} follows from compactness. This restriction is realized by the use of a priori bounds which can be written assuming some prior knowledge of the solution. Therefore, in addition to the inequality

$$\|Af - \overline{g}\| \leq \epsilon \tag{7}$$

which corresponds to a bound on the noise ($\|\cdot\|$ denoting the norm in $L^2(a,b)$), one also considers an a priori bound on the solution of the following form:

$$\|Cf\|_{\mathcal{Z}} \leq E, \tag{8}$$

where $\mathcal{Z}$ denotes the "constraint space" and, accordingly, C is the "constraint operator." From the bounds (7) and (8) we are led to define the regularized solution as the minimum of the following functional:

$$\Phi(f) = \|Af - \overline{g}\|^2 + \alpha^2 \|Cf\|^2_{\mathcal{Z}}, \qquad \left(\alpha = \left(\frac{\epsilon}{E}\right)\right). \tag{9}$$

In spite of several significant merits, this procedure is not free from defects. Concerning the possibility of writing suitable a priori bounds on the solution, we want to remark strongly that two different types of problems must be distinguished:

a) synthesis problems,
b) inverse problems

and to note that both are frequently solved by the use of Fredholm integral equations of the first kind. In the first class of problems, that basically consists in finding the source that produces a prescribed effect (e.g., prescribed boundary values), the a priori bounds are intrinsic of the problem itself, whereas this is not always the case for the second class. As typical examples we can consider

a′) the antenna synthesis,
b′) the signal recovery.

The problem of the antenna synthesis consists in determining, within a certain degree of approximation, the current intensity that generates a desired radiation pattern. It can be formulated in terms of Fredholm equation of the first kind [18, 24] and, consequently, it presents the typical pathology of the ill-posed problems. In this problem the a priori bound on the ohmic losses associated with the current intensity is necessary and can be regarded as a natural constraint intrinsic of the problem. Conversely, in the case of the signal recovery problem, the a priori bounds can be written only if prior knowledge of the signal is assumed. Generally, it is possible to have some a priori information regarding, for instance, the support of the signal or requiring the function representing the signal to be nonnegative. But even in these cases the prior knowledge could be insufficiently specific to be peculiar of the function to be reconstructed, and arbitrary, though reasonable, constraints must be added to solve the problem. Strictly connected with this question there is the crux of the matter: the practical choice of the regularization parameter α (see formula (9)) for a fixed $\overline{g}$ when the a priori bound (8) is unknown or is not sufficiently precise.

Moreover, let us note that the functional (9) works as a filter whose action is smoothing the Fourier components $\overline{g}_k$ for high values of k. But it is easy to exhibit examples of signals whose Fourier components are small, or even zero, for low values of k, while the significant contributions of the signal are brought by those components at intermediate values of k, which are smoothed out by the action of the filter. In these situations the standard regularization method fails, showing that the only existence of the minimum of functional (9) does not guarantee the bulk of the signal has been really recovered. This delicate point will be illustrated with numerical examples in section 4.

We suggest a different approach which is based on the following observation: for the moment, suppose that the moduli of the noiseless Fourier coefficients $|g_k|$ are monotonically decreasing as k increases; then, although the formal series (6) diverges, nevertheless the effect of the error remains limited in the beginning of the expansion, and there exists a point (a certain value of k) where divergence sets in. Thus, the idea is to stop the expansion at the point where it turns to diverge. This rough and qualitative description can be put in rigorous form by proving that even if the series (6) diverges, nevertheless it converges (in the sense of L^2-norm) as ϵ (i.e., the bound on the noise) tends to zero. This result, which has been proved by two of us (see [17]), does not give (except in very particular cases) a practical numerical method for finding out the truncation point (i.e., the value of k) at which to stop expansion (6). However, here we prove a probabilistic generalization of the results presented in [17] by removing the quite restrictive assumption that the Fourier coefficients $|g_k|$ of the signal to be

recovered are monotonically decreasing. Compared to [17] the significance of the new results is relevant. First, the hypothesis made in [17] on the order of the coefficients $|g_k|$ leads to a regularization procedure that essentially works as an ideal low-pass filter, and, as previously discussed, this does not guarantee to recover correctly the signals whose bulk is localized at intermediate frequencies. Conversely, in this paper it will be shown how to construct a regularized solution without assuming any kind of order on the coefficients $|g_k|$ by exploiting the tools supplied by the information theory. This result will lead to a more effective regularizing algorithm which is based on a suitable statistical analysis of the data and whose main feature is indeed the frequency selectivity. Second, from the application point of view, the hypothesis on the order of the coefficients $|g_k|$ is too restrictive; thus, by removing it, a much larger class of real signals can be practically analyzed. These questions are precisely the contents of sections 3 and 4. We will prove, indeed, in section 3 that it is possible to split the noisy Fourier coefficients $\overline{g}_k$ into two classes:

i) the Fourier coefficients $\overline{g}_k$ from which a significant amount of information on $f_k = (f, \psi_k)$ can be extracted;
ii) the Fourier coefficients $\overline{g}_k$ that can be regarded as random numbers because the noise prevails on the coefficients g_k.

In section 4 it will be shown how it is possible to separate practically the coefficients $\overline{g}_k$ into these two classes by the use of statistical tools supplied by the so called "time-series" analysis. Therefore, we can practically construct an approximation which converges to the real solution, and furthermore we can have some confidence that the bulk of the function f has been effectively recovered.

The paper is organized as follows. In the first part of section 2 a short sketch of the variational method based on the minimization of functional (9) is given. This will be done in order to have explicitly the formulae which will be used in section 4, where our procedure and the variational one will be compared. The second part of section 2 is devoted to the probabilistic formulation of the regularization problem in a quite general setting. In section 3 we start illustrating the asymptotic convergence of the eigenfunction expansion (in the sense of L^2-norm) as ϵ tends to zero; then this result is reconsidered from the viewpoint of probability and information theory. Here a key role will be played by the Bayes formula: it will provide the various terms of our approximation, which will be proved to be a probabilistically regularized solution of (1). The first part of section 4 is devoted to the discussion of the statistical tools that are necessary for practically recovering the regularized solution from finite samples of noisy data. Finally, some numerical examples are given in the second part of section 4.

2. Variational and probabilistic regularization.

2.1. Variational regularization. After the classical book of Tikhonov and Arsenine [23] the literature on the theory and applications of the variational regularization has been rapidly growing (see, for instance, [14]). In order to compare our algorithm with this classical one, some formulae and results of the variational regularization will be recalled here (see [5, 6, 20, 23] for proofs and details).

Let us characterize, first of all, the constraint operator C and, accordingly, the constraint space $\mathcal{Z}$. Let us take a constraint operator C such that $C^\star C$ and $A^\star A$ commute (this assumption does not restrict the theory and the applications significantly [5, 18]). Then, the space $\mathcal{Z}$ is composed by those functions $f \in L^2(a, b)$ such

that $\|Cf\|_{\mathcal{Z}}$ is finite; i.e.,

$$\|Cf\|_{\mathcal{Z}}^2 = (C^\star Cf, f) = \sum_{k=1}^{\infty} c_k^2 |f_k|^2 < \infty. \tag{10}$$

Now we consider the ball $\mathcal{U}_{\mathcal{Z}} = \{f \in \mathcal{Z} \mid \sum_{k=1}^{\infty} c_k^2 |f_k|^2 \le E^2\}$ and the restriction A_0 of the operator A (see (1)) to the ball $\mathcal{U}_{\mathcal{Z}}$. Then, the following propositions can be proved.

PROPOSITION 2.1. *If* $\lim_{k\to\infty} c_k^2 = +\infty$, *the operator* A_0^{-1} *is continuous.*

PROPOSITION 2.2. *The functional* $\Phi(f)$, *with* $\alpha = (\epsilon/E)$, *has a unique minimum which is given by*

$$f_\star = [A^\star A + \left(\frac{\epsilon}{E}\right)^2 C^\star C]^{-1} A^\star \overline{g}. \tag{11}$$

By expanding $\overline{g}$ in terms of ψ_k (eigenfunctions of the operator A), we have

$$f_\star = \sum_{k=1}^{\infty} \frac{\lambda_k \overline{g}_k}{\lambda_k^2 + c_k^2 \left(\frac{\epsilon}{E}\right)^2} \psi_k. \tag{12}$$

Next, we have the following proposition.

PROPOSITION 2.3. *The following limit holds true for any function* f *satisfying the bounds* (7) *and* (8)*:*

$$\lim_{\epsilon\to 0} \|f - f_\star\| = 0 \qquad (E \ \textit{fixed}). \tag{13}$$

In numerical computations it is often convenient to use truncated approximations. For instance, one can derive from the smoothed solution (12) the following truncated approximation:

$$f_\star^{(1)} = \sum_{k=1}^{k_\alpha} \frac{\overline{g}_k}{\lambda_k} \psi_k, \tag{14}$$

where k_α is the largest integer such that

$$\lambda_k \ge \left(\frac{\epsilon}{E}\right) |c_k|. \tag{15}$$

PROPOSITION 2.4. *The following limit holds true for any function* f *satisfying bounds* (7) *and* (8)*:*

$$\lim_{\epsilon\to 0} \left\| f - f_\star^{(1)} \right\| = 0 \qquad (E \ \textit{fixed}). \tag{16}$$

In several problems a weaker a priori bound should be used by setting $C = I$ (the identity operator). Therefore, instead of bound (8), we have

$$\|f\| = \left(\sum_{k=1}^{\infty} |f_k|^2 \right)^{1/2} \le E. \tag{17}$$

In this case the unique minimum of functional (9) is given by

$$f_\star^{(2)} = \sum_{k=1}^{\infty} \frac{\lambda_k \overline{g}_k}{\lambda_k^2 + \left(\frac{\epsilon}{E}\right)^2} \psi_k, \tag{18}$$

and, accordingly, the following truncated approximation can be introduced:

$$f_\star^{(3)} = \sum_{k=1}^{k_\beta} \frac{\overline{g}_k}{\lambda_k} \psi_k, \tag{19}$$

where k_β is the largest integer such that

$$\lambda_k \geq \frac{\epsilon}{E}. \tag{20}$$

Both $f_\star^{(2)}$ and $f_\star^{(3)}$ converge to f as $\epsilon \to 0$ in a weak sense. In fact, as shown in [20, 21], the following proposition can be proved.

PROPOSITION 2.5. *For any function f which satisfies the bounds* (7) *and* (17)*, the following limits hold true:*

$$\lim_{\epsilon \to 0} \left| \left(\left[f - f_\star^{(2)} \right], v \right) \right| = 0 \qquad (\|v\| \leq 1,\ E \text{ fixed}), \tag{21}$$

$$\lim_{\epsilon \to 0} \left| \left(\left[f - f_\star^{(3)} \right], v \right) \right| = 0 \qquad (\|v\| \leq 1,\ E \text{ fixed}). \tag{22}$$

2.2. Probabilistic regularization. Here we want to reconsider (5) from a probabilistic point of view. With this in mind we rewrite (5) in the following form:

$$A\xi + \zeta = \eta, \tag{23}$$

where ξ, ζ, and η, which correspond to f, n, and $\overline{g}$, respectively, are Gaussian weak random variables (w.r.v.) in the Hilbert space $L^2(a, b)$ [2]. A Gaussian w.r.v. is uniquely defined by its mean element and its covariance operator; in the present case we denote by $R_{\xi\xi}$, $R_{\zeta\zeta}$, and $R_{\eta\eta}$ the covariance operators of ξ, ζ, and η, respectively. Next, we make the following assumptions:

I) ξ and ζ have zero mean; i.e., $m_\xi = m_\zeta = 0$;
II) ξ and ζ are uncorrelated; i.e., $R_{\xi\zeta} = 0$;
III) $R_{\zeta\zeta}^{-1}$ exists.

The third assumption is the mathematical formulation of the fact that all the components of the data function are affected by noise. As it is shown by Franklin (see formula (3.11) of [11]), if the signal and the noise satisfy assumptions I) and II), then

$$R_{\eta\eta} = A R_{\xi\xi} A^\star + R_{\zeta\zeta} \tag{24}$$

and the cross-covariance operator is given by

$$R_{\xi\eta} = R_{\xi\xi} A^\star. \tag{25}$$

We also assume that $R_{\zeta\zeta}$ will depend on a parameter ϵ that tends to zero when the noise vanishes; i.e.,

$$R_{\zeta\zeta} = \epsilon^2 N, \tag{26}$$

where N is a given operator (e.g., $N = I$ for the white noise).

Now we are faced with the following problem.

Problem. Given a value $\overline{g}$ of the w.r.v. η find an estimate of the w.r.v. ξ.

A linear estimate of ξ will be any w.r.v. $\xi_L = L\eta$, where $L : Y \to X$ is an arbitrary linear continuous operator. Then from a value $\overline{g}$ of η one obtains the linear estimate $L\overline{g}$ of the w.r.v. ξ. Now a measure of the reliability of the estimator L is given by

$$\delta^2(\epsilon, v; L) = \mathrm{E}\left\{|(\xi - L\eta, v)|^2\right\}, \quad (v \in X = L^2(a, b)), \tag{27}$$

where $\mathrm{E}\{\cdot\}$ denotes the expectation value. Then we have the following proposition.

PROPOSITION 2.6. *If the covariance operator $R_{\zeta\zeta}$ has a bounded inverse, then there exists a unique operator L_0 that minimizes $\delta^2(\epsilon, v; L)$ for any $v \in X$, and it is given by*

$$L_0 = R_{\xi\eta} R_{\eta\eta}^{-1} = R_{\xi\xi} A^\star \left[A R_{\xi\xi} A^\star + R_{\zeta\zeta}\right]^{-1}. \tag{28}$$

Proof. See [4, 5]. □

The w.r.v. $L_0\eta$ is called the best linear estimate of ξ, and, given a value $\overline{g}$ of η, the best linear estimate $f_\star^{(4)}$ for the value of ξ is

$$f_\star^{(4)} = \frac{R_{\xi\xi} A^\star}{A R_{\xi\xi} A^\star + R_{\zeta\zeta}} \overline{g}, \qquad (A^\star = A). \tag{29}$$

If ξ and $L\eta$ have finite variance, then the global mean-square error may be defined as follows:

$$\delta^2(\epsilon, L) = \mathrm{E}\left\{\|\xi - L\eta\|^2\right\}. \tag{30}$$

When the operator L_0 which minimizes (27) does exist, it also minimizes the global error (30) if $L_0\eta$ has finite variance; i.e., if $\mathrm{Tr}\,(L_0 R_{\eta\eta} L_0^\star) < \infty$, then the following proposition can be proved.

PROPOSITION 2.7. *If the assumptions*

i) *$R_{\xi\xi}$ is an operator of trace class*;

ii) *$R_{\zeta\zeta} = \epsilon^2 N$ has bounded inverse*;

iii) *the equation $Af = 0$, where $f \in$ Range $\left(R_{\xi\xi}^{1/2}\right)$, has only the trivial solution $f = 0$*

are satisfied, then the following limit holds true:

$$\lim_{\epsilon \to 0} \delta^2(\epsilon) = 0, \tag{31}$$

where $\delta^2(\epsilon) = \inf_L \delta^2(\epsilon; L)$.

Proof. See [4, 5]. □

Let us note that $\delta^2(\epsilon) = \delta^2(\epsilon; L_0)$ when L_0 does exist and is unique.

If we want to compare the probabilistic results obtained above with the variational ones, which have been obtained by the use of eigenfunction expansions, we must expand ξ and ζ in terms of the eigenfunctions of the operator A (i.e., $\{\psi_k\}_1^\infty$). Their Fourier components are the random variables $\xi_k = (\xi, \psi_k)$ and $\zeta_k = (\zeta, \psi_k)$, whose variances are given respectively by ρ_k^2 and $\epsilon^2 \nu_k^2$. Next, in addition to the assumptions I)–III) made before, we make the following hypothesis in spite of the fact that it turns out to be completely unrealistic (see section 4):

IV) the Fourier components of ξ are mutually uncorrelated as well as the Fourier components of ζ.

Therefore, if $R_{\zeta\zeta}^{-1}$ is bounded (i.e. $\sup_k(1/\epsilon^2\nu_k^2) < \infty$), then the operator L_0 exists and the best linear estimate (29) can be written as

$$f_\star^{(4)} = \sum_{k=1}^{\infty} \frac{\lambda_k \rho_k^2}{\lambda_k^2\rho_k^2 + \epsilon^2\nu_k^2} \overline{g}_k \psi_k. \tag{32}$$

Finally, the quantities $\delta^2(\epsilon, v; L_0)$ and $\delta^2(\epsilon)$ become

$$\begin{aligned} \delta^2(\epsilon, v; L_0) &= \mathrm{E}\left\{|(\xi - L_0\eta, v)|^2\right\} = \\ &= ([R_{\xi\xi} - L_0 R_{\eta\eta} L_0^\star]v, v) = \epsilon^2 \sum_{k=1}^{\infty} \frac{\rho_k^2\nu_k^2}{\lambda_k^2\rho_k^2 + \epsilon^2\nu_k^2} |v_k|^2 \end{aligned} \tag{33}$$

and

$$\delta^2(\epsilon) = \delta^2(\epsilon; L_0) = \mathrm{Tr}\,[R_{\xi\xi} - L_0 R_{\eta\eta} L_0^\star] = \epsilon^2 \sum_{k=1}^{\infty} \frac{\rho_k^2\nu_k^2}{\lambda_k^2\rho_k^2 + \epsilon^2\nu_k^2}, \tag{34}$$

and we have the following proposition.

PROPOSITION 2.8. *The following statements hold true:*

i) *for any* $v \in X$ $(X = L^2(a, b))$

$$\lim_{\epsilon \to 0} \delta^2(\epsilon, v; L_0) = 0, \tag{35}$$

ii) *if* $Tr\, R_{\xi\xi} < \infty$, *then*

$$\lim_{\epsilon \to 0} \delta^2(\epsilon) = 0. \tag{36}$$

3. Information theory and regularization.

3.1. Asymptotic convergence, in the L^2-norm, of the eigenfunction expansion. In the variational regularization, use is made of global a priori bounds (e.g., formulae (8) or (17)), which are the natural constraints in the case of synthesis problems where the variational approach is certainly appropriate. But these global bounds are not necessarily given in the case of inverse problems where the prior knowledge on the solution can be, in several cases, rather poor. Moreover, in the truncated solutions derived by the methods of variational regularization, the point at which to stop the expansion is obtained by comparing the eigenvalues λ_k with the ratio (ϵ/E) (i.e., formula (20)) or with $(\epsilon/E)|c_k|$ (see formula (15)). In both cases this approach appears quite unnatural from the viewpoint of the experimental or physical sciences, whose methodology would rather suggest to stop the expansions at the value k_0 of k such that for $k > k_0$ the Fourier coefficients g_k of the noiseless data are smaller or at most of the same order of magnitude of ϵ, and, consequently, it is impossible to extract information from the corresponding coefficients $\overline{g}_k$. With this in mind, and assuming that the noise is represented by a bounded and integrable function $n(x)$ which satisfies the following condition:

$$\sup |n(x)| \leq \epsilon, \qquad x \in [a, b], \tag{37}$$

the following results have been proved by two of us.

LEMMA 3.1. *The following statements hold true:*

$$\sum_{k=1}^{\infty}\left(\frac{g_k}{\lambda_k}\right)^2 = \|f\|^2 = C_1 \qquad (C_1 = \textit{constant}), \tag{38}$$

$$\sum_{k=1}^{\infty}\left(\frac{\overline{g}_k}{\lambda_k}\right)^2 = +\infty \qquad \textit{if } \overline{g} \notin \textit{Range}\,(A), \tag{39}$$

$$\lim_{\epsilon\to 0}\overline{g}_k = g_k \quad \forall k. \tag{40}$$

If $k_0(\epsilon)$ is defined by

$$k_0(\epsilon) = \max\left\{m \in \mathbb{N} : \sum_{k=1}^{m}\left(\frac{\overline{g}_k}{\lambda_k}\right)^2 \le C_1\right\}, \tag{41}$$

then

$$\lim_{\epsilon\to 0} k_0(\epsilon) = +\infty. \tag{42}$$

Proof. See [17]. □

Now we can introduce the following approximation:

$$f_0^{(\epsilon)} = \sum_{k=1}^{k_0(\epsilon)}\frac{\overline{g}_k}{\lambda_k}\psi_k \tag{43}$$

and prove the following theorem.

THEOREM 3.2. *The following equality holds true:*

$$\lim_{\epsilon\to 0}\left\|f - f_0^{(\epsilon)}\right\| = 0. \tag{44}$$

Proof. See [17]. □

If we consider a sequence of noisy data $\overline{g}$ which tends to g for $\epsilon \to 0$ in the sense of the L^2-norm (i.e., $\lim_{\epsilon\to 0}\|\overline{g}-g\| = 0$), then $f_0^{(\epsilon)}$ will tend to f as $\epsilon \to 0$ in the sense of the L^2-norm (i.e., $\lim_{\epsilon\to 0}\|f_0^{(\epsilon)} - f\| = 0$). In fact, since $\|\overline{g}-g\|^2 = \sum_{k=1}^{\infty}|\overline{g}_k - g_k|^2$, the $\lim_{\epsilon\to 0}\|\overline{g}-g\| = 0$ implies that for any k, $\lim_{\epsilon\to 0}\overline{g}_k = g_k$, and in view of Lemma 3.1 and Theorem 3.2 it can be concluded that $\lim_{\epsilon\to 0}\|f_0^{(\epsilon)} - f\| = 0$. Therefore, from approximation (43) we can derive an operator $\overline{B}$ defined by

$$\overline{B}\overline{g} = \sum_{k=1}^{k_0(\epsilon)}\frac{\overline{g}_k}{\lambda_k}\psi_k, \tag{45}$$

which continuously maps, (i.e., preserving the convergence) the data $\overline{g}$ into the solution space X. Thus, continuity has been restored without requiring compactness.

Two types of difficulties still remain:

a) how to determine numerically the truncation point $k_0(\epsilon)$ if the norm of the function f (i.e., the constant $C_1 = \|f\|^2$) is unknown;
b) in any case the convergence of approximation (43) is not sufficient to guarantee that the bulk of the unknown function f has been really recovered.

We can give a satisfactory answer to these questions only in very specific and peculiar cases, as we will explain below. Suppose that the moduli of the Fourier coefficients $|g_k|$ are monotonically decreasing for increasing values of k. Since $\overline{g}_k = g_k + n_k$, it turns out that at a certain value k_0 of k we have $|g_k| \simeq |n_k| \leq \epsilon$. The Fourier coefficients of the noiseless data are of the same order of magnitude as the Fourier components of the noise, and at this point we cannot extract any information from the noisy Fourier coefficients $\overline{g}_k$. Let us now introduce the function $M(m) = \sum_{k=1}^{m} (\overline{g}_k/\lambda_k)^2$, whose relevant properties are:

1) it is an increasing function of m;
2) if ϵ is sufficiently small and the values of $|g_k|$ are monotonically decreasing for increasing k, $M(m)$ presents a "plateau" when it reaches the value C_1. Indeed, from formula (42) in Lemma 3.1 it follows that $M(m)$ remains nearly constant when it attains the value C_1. An explicit numerical example of this "plateau" is given in Figure 1D in section 4.

This "plateau" corresponds to the order–disorder transition in the coefficients $\overline{g}_k$: for $k < k_0(\epsilon)$ the data g_k prevail on n_k whereas for $k > k_0(\epsilon)$ the noise components n_k are larger or, at least, of the same order of magnitude of the noiseless data. However, it must be remarked that in practical cases to single out the plateau which does really correspond to the order–disorder transition in the coefficients $\overline{g}_k$ can be made difficult by the presence of other spurious plateaux due to the erratic behavior of the noise. Furthermore, if the coefficients g_k are negligible for low values of k, and the actual bulk of information is located only at intermediate values of k, there could be no numerical evidence of such a plateau in spite of the fact that the convergence guaranteed by Theorem 3.2 remains true. Then we are forced to look for other methods that overcome these difficulties. This issue will be investigated by means of probabilistic methods, as will be illustrated in the next subsection.

3.2. Bayes formula, information theory, and regularization. Here our goal is to find a probabilistic extension of the result of Theorem 3.2 in which the assumption requiring the Fourier coefficients $|g_k|$ to be monotonically decreasing will be removed. In fact, we will show how to construct a regularizing solution from the noisy data, disregarding the order of the coefficients $|g_k|$. For this purpose, we turn (23) into an infinite sequence of one-dimensional equations by means of orthogonal projections:

$$\lambda_k \xi_k + \zeta_k = \eta_k, \quad (k = 1, 2, ...), \tag{46}$$

where $\xi_k = (\xi, \psi_k)$, $\zeta_k = (\zeta, \psi_k)$, $\eta_k = (\eta, \psi_k)$ are Gaussian random variables. Here we retain assumptions I)–III) made in section 2.2, but we remove assumption IV). In fact, there is no reason to assume that the basis $\{\psi_k\}_1^\infty$ which diagonalizes the operator A also diagonalizes the covariance operators $R_{\xi\xi}$, $R_{\zeta\zeta}$, $R_{\eta\eta}$ [19]. Therefore, we can introduce the variances $\rho_k^2 = (R_{\xi\xi}\psi_k, \psi_k)$, $\epsilon^2\nu_k^2 = (R_{\zeta\zeta}\psi_k, \psi_k)$, $\lambda_k^2\rho_k^2 + \epsilon^2\nu_k^2 = (R_{\eta\eta}\psi_k, \psi_k)$ without assuming that the Fourier components ξ_k of ξ (and analogously also for ζ_k and η_k) are mutually uncorrelated. In view of assumptions I) and III) the following probability densities for ξ_k and ζ_k can be assumed:

$$p_{\xi_k}(x) = \frac{1}{\sqrt{2\pi}\,\rho_k} \exp\left\{-\left(\frac{x^2}{2\rho_k^2}\right)\right\}, \quad (k = 1, 2, ...) \tag{47}$$

and

$$p_{\zeta_k}(x) = \frac{1}{\sqrt{2\pi}\,\epsilon\nu_k} \exp\left\{-\left(\frac{x^2}{2\epsilon^2\nu_k^2}\right)\right\}, \quad (k = 1, 2, ...). \tag{48}$$

By the use of the (46) we can also introduce the conditional probability density $p_{\eta_k}(y|x)$ of the random variable η_k for fixed $\xi_k = x$, which reads

$$\begin{aligned} p_{\eta_k}(y|x) &= \frac{1}{\sqrt{2\pi}\,\epsilon\nu_k} \exp\left\{-\frac{(y-\lambda_k x)^2}{2\epsilon^2\nu_k^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\,\epsilon\nu_k} \exp\left\{-\frac{\lambda_k^2}{2\epsilon^2\nu_k^2}\left(x - \frac{y}{\lambda_k}\right)^2\right\}. \end{aligned} \tag{49}$$

Now let us apply the Bayes formula that provides the conditional probability density of ξ_k given η_k through the following expression:

$$p_{\xi_k}(x|y) = \frac{p_{\xi_k}(x)p_{\eta_k}(y|x)}{p_{\eta_k}(y)}. \tag{50}$$

Thus, if a realization of the random variable η_k is given by $\overline{g}_k$ (see the formulation of the problem in section 2.2), formula (50) becomes

$$p_{\xi_k}(x|\overline{g}_k) = A_k \exp\left\{-\frac{x^2}{2\rho_k^2}\right\} \exp\left\{-\frac{\lambda_k^2}{2\epsilon^2\nu_k^2}\left(x - \frac{\overline{g}_k}{\lambda_k}\right)^2\right\} \quad (A_k = \text{const.}). \tag{51}$$

Now the amount of information on the variable ξ_k which is contained in the variable η_k can be evaluated. We have [13]

$$J(\xi_k, \eta_k) = -\frac{1}{2}\log(1 - r_k^2), \tag{52}$$

where

$$r_k^2 = \frac{|\mathrm{E}\{\xi_k\eta_k\}|^2}{\mathrm{E}\{|\xi_k|^2\}\,\mathrm{E}\{|\eta_k|^2\}} = \frac{(\lambda_k\rho_k)^2}{(\lambda_k\rho_k)^2 + (\epsilon\nu_k)^2}. \tag{53}$$

Thus,

$$J(\xi_k, \eta_k) = \frac{1}{2}\log\left(1 + \frac{\lambda_k^2\rho_k^2}{\epsilon^2\nu_k^2}\right). \tag{54}$$

From equality (54) it follows that $J(\xi_k, \eta_k) < \frac{1}{2}\log 2$ if $\lambda_k\rho_k < \epsilon\nu_k$. Thus, we are naturally led to introduce the following sets:

$$\mathcal{I}_k = \{k : \lambda_k\rho_k \geq \epsilon\nu_k\}, \tag{55}$$

$$\mathcal{N}_k = \{k : \lambda_k\rho_k < \epsilon\nu_k\}. \tag{56}$$

Reverting to the conditional probability density (51), it can be regarded as the product of two Gaussian probability densities: $p_1(x) = A_k^{(1)} \exp\{-x^2/2\rho_k^2\}$ and $p_2(x) = A_k^{(2)} \exp\{-(\lambda_k^2/2\epsilon^2\nu_k^2)(x - (\overline{g}_k/\lambda_k))^2\}$, $(A_k = A_k^{(1)} \cdot A_k^{(2)})$, whose variances are respectively given by ρ_k and $(\epsilon\nu_k/\lambda_k)$. Let us note that if $k \in \mathcal{I}_k$, the variance

associated with the density $p_2(x)$ is smaller than the corresponding variance of $p_1(x)$, and vice versa if $k \in \mathcal{N}_k$. Therefore, it is reasonable to consider as an acceptable approximation of $\langle \xi_k \rangle$ the mean value given by the density $p_2(x)$ if $k \in \mathcal{I}_k$, or the mean value given by the density $p_1(x)$ if $k \in \mathcal{N}_k$. We can write the following approximation:

$$\langle \xi_k \rangle = \begin{cases} \dfrac{\overline{g}_k}{\lambda_k} & (k \in \mathcal{I}_k), \\ 0 & (k \in \mathcal{N}_k). \end{cases} \tag{57}$$

Consequently, given the value $\overline{g}$ of the w.r.v. η, we are led to consider the following estimate of ξ:

$$\widehat{B}\overline{g} = \sum_{k \in \mathcal{I}_k} \frac{\overline{g}_k}{\lambda_k} \psi_k. \tag{58}$$

However, these are only heuristic considerations based on plausible arguments. They will become rigorous statements only if it will proved that they lead to a solution $\widehat{B}\overline{g}$ which is probabilistically regularized. For this purpose, the global mean-square error associated with the operator $\widehat{B}$, i.e., $\mathrm{E}\{\|\xi - \widehat{B}\eta\|^2\}$, must be evaluated, and we have the following proposition.

PROPOSITION 3.3.

i) *If* $\lim_{k\to\infty}(\lambda_k \rho_k/\nu_k) = 0$, *then the set* $\mathcal{I}_k$ *is finite for any fixed positive value of* ϵ;

ii) *assuming that the limit stated in* i) *holds true, and, in addition, that* $R_{\xi\xi}$ *is an operator of trace class, then the following relationship holds:*

$$E\left\{\|\xi - \widehat{B}\eta\|^2\right\} = \sum_{k \in \mathcal{N}_k} \rho_k^2 + \sum_{k \in \mathcal{I}_k} \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} < \infty. \tag{59}$$

Proof. The proof of statement i) is obvious if we recall the definition of the set $\mathcal{I}_k$ (formula (55)). Statement ii) follows easily from the equality

$$\mathrm{E}\left\{\|\xi - \widehat{B}\eta\|^2\right\} = \mathrm{Tr}\,(R_{\xi\xi} - R_{\xi\xi}A^\star \widehat{B}^\star - \widehat{B}AR_{\xi\xi} + \widehat{B}R_{\eta\eta}\widehat{B}^\star) \tag{60}$$

and by the use of formulae (24), (26), and (58). □

In order to prove that approximation (58) is regularized, we need the following auxiliary lemma.

LEMMA 3.4. *Let* $k_\gamma(\epsilon)$ *be defined as follows:*

$$k_\gamma(\epsilon) = \max\left\{ m \in \mathbb{N} : \sum_{k=1}^{m} \left(\rho_k^2 + \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} \right) \le \Gamma \right\}, \tag{61}$$

where $\Gamma = Tr\, R_{\xi\xi}$ *is finite. Then the following statements hold true:*

$$\text{i)} \quad \lim_{\epsilon \to 0} k_\gamma(\epsilon) = +\infty, \tag{62}$$

$$\text{ii)} \quad \lim_{\epsilon \to 0} \left\{ \sum_{k=1}^{k_\gamma} \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} + \sum_{k=k_\gamma+1}^{\infty} \rho_k^2 \right\} = 0. \tag{63}$$

Proof. i) Let k_{γ_1} denote the sum $(k_\gamma + 1)$. Then suppose that the limit (62) does not hold. This latter assumption would imply that there exists a finite number M, which does not depend on ϵ, such that $k_{\gamma_1} < M$. Furthermore, this bound should remain true for any sequence ϵ_i tending to zero. Then we have the following inequalities:

$$\Gamma < \sum_{k=1}^{k_{\gamma_1}} \left(\rho_k^2 + \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} \right) \leq \sum_{k=1}^{M} \left(\rho_k^2 + \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} \right). \tag{64}$$

Now for any sequence ϵ_i tending to zero, we have

$$\Gamma < \sum_{k=1}^{M} \rho_k^2 \leq \sum_{k=1}^{\infty} \rho_k^2 = \Gamma, \tag{65}$$

which is contradictory. Then limit (62) holds.

ii) From $\sum_{k=1}^{\infty} \rho_k^2 = \operatorname{Tr} R_{\xi\xi} = \Gamma < \infty$, and in view of statement i), it follows that $\lim_{\epsilon \to 0} \sum_{k=k_{\gamma_1}}^{\infty} \rho_k^2 = 0$. Regarding the sum $\sum_{k=1}^{k_\gamma} (\epsilon^2 \nu_k^2 / \lambda_k^2)$, we can proceed as follows. From formula (61) we have

$$\sum_{k=1}^{k_\gamma} \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} + \sum_{k=1}^{k_\gamma} \rho_k^2 \leq \Gamma. \tag{66}$$

Then

$$\sum_{k=1}^{k_\gamma} \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} \leq \Gamma - \sum_{k=1}^{k_\gamma} \rho_k^2 = \sum_{k=k_{\gamma_1}}^{\infty} \rho_k^2, \tag{67}$$

but in view of the fact that $\lim_{\epsilon \to 0} \sum_{k=k_{\gamma_1}}^{\infty} \rho_k^2 = 0$, we have $\lim_{\epsilon \to 0} \sum_{k=1}^{k_\gamma} (\epsilon^2 \nu_k^2 / \lambda_k^2) = 0$, and statement ii) is proved. □

We can now prove the following theorem.

THEOREM 3.5. *If the covariance operator $R_{\xi\xi}$ is of trace class, and if the set $\mathcal{I}_k$ is finite (see Proposition* 3.3*), then the following limit holds true:*

$$\lim_{\epsilon \to 0} \delta^2(\epsilon, \widehat{B}) = \lim_{\epsilon \to 0} E\left\{ \|\xi - \widehat{B}\eta\|^2 \right\} = 0; \tag{68}$$

i.e., approximation (58) *is probabilistically regularized.*

Proof. In view of formula (59) in Proposition 3.3, the proof of equality (68) reduces to the proof of the following limit:

$$\lim_{\epsilon \to 0} \left\{ \sum_{k \in \mathcal{I}_k} \frac{\epsilon^2 \nu_k^2}{\lambda_k^2} + \sum_{k \in \mathcal{N}_k} \rho_k^2 \right\} = 0. \tag{69}$$

Regarding the first sum of (69), we divide the set $\mathcal{I}_k$ into two subsets defined by

$$\mathcal{I}_k^{(1)} = \{k \in \mathcal{I}_k \,:\, k \leq k_\gamma\}, \tag{70}$$

$$\mathcal{I}_k^{(2)} = \{k \in \mathcal{I}_k \,:\, k > k_\gamma\}; \qquad (\mathcal{I}_k = \mathcal{I}_k^{(1)} \cup \mathcal{I}_k^{(2)}). \tag{71}$$

Accordingly, we can write

$$\sum_{k\in\mathcal{I}_k}\frac{\epsilon^2\nu_k^2}{\lambda_k^2}=\sum_{k\in\mathcal{I}_k^{(1)}}\frac{\epsilon^2\nu_k^2}{\lambda_k^2}+\sum_{k\in\mathcal{I}_k^{(2)}}\frac{\epsilon^2\nu_k^2}{\lambda_k^2}. \tag{72}$$

Then $\sum_{k\in\mathcal{I}_k^{(1)}}(\epsilon^2\nu_k^2/\lambda_k^2)\le\sum_{k=1}^{k_\gamma}(\epsilon^2\nu_k^2/\lambda_k^2)$, and in view of Lemma 3.4 (where we proved that $\lim_{\epsilon\to0}\sum_{k=1}^{k_\gamma}(\epsilon^2\nu_k^2/\lambda_k^2)=0$) it follows that

$$\lim_{\epsilon\to0}\sum_{k\in\mathcal{I}_k^{(1)}}\frac{\epsilon^2\nu_k^2}{\lambda_k^2}=0. \tag{73}$$

Regarding the term $\sum_{k\in\mathcal{I}_k^{(2)}}(\epsilon^2\nu_k^2/\lambda_k^2)$, since $k\in\mathcal{I}_k$ then $\rho_k^2\ge(\epsilon^2\nu_k^2/\lambda_k^2)$ and therefore

$$\sum_{k\in\mathcal{I}_k^{(2)}}\frac{\epsilon^2\nu_k^2}{\lambda_k^2}\le\sum_{k=k_{\gamma_1}}^{\infty}\rho_k^2. \tag{74}$$

But, as we have seen in Lemma 3.4, $\lim_{\epsilon\to0}\sum_{k=k_{\gamma_1}}^{\infty}\rho_k^2=0$, and consequently

$$\lim_{\epsilon\to0}\sum_{k\in\mathcal{I}_k^{(2)}}\frac{\epsilon^2\nu_k^2}{\lambda_k^2}=0. \tag{75}$$

We can conclude that $\lim_{\epsilon\to0}\sum_{k\in\mathcal{I}_k}(\epsilon^2\nu_k^2/\lambda_k^2)=0$. Regarding the sum $\sum_{k\in\mathcal{N}_k}\rho_k^2$, we proceed in an analogous way by splitting the set $\mathcal{N}_k$ into two subsets defined by

$$\mathcal{N}_k^{(1)}=\{k\in\mathcal{N}_k\,:\,k\le k_\gamma\}, \tag{76}$$

$$\mathcal{N}_k^{(2)}=\{k\in\mathcal{N}_k\,:\,k>k_\gamma\};\qquad(\mathcal{N}_k=\mathcal{N}_k^{(1)}\cup\mathcal{N}_k^{(2)}). \tag{77}$$

Accordingly, we write

$$\sum_{k\in\mathcal{N}_k}\rho_k^2=\sum_{k\in\mathcal{N}_k^{(1)}}\rho_k^2+\sum_{k\in\mathcal{N}_k^{(2)}}\rho_k^2. \tag{78}$$

If $k\in\mathcal{N}_k^{(1)}$ and by the use of inequality $\rho_k^2<(\epsilon^2\nu_k^2/\lambda_k^2)$ (because $k\in\mathcal{N}_k$) we can write

$$\sum_{k\in\mathcal{N}_k^{(1)}}\rho_k^2\le\sum_{k=1}^{k_\gamma}\frac{\epsilon^2\nu_k^2}{\lambda_k^2}. \tag{79}$$

But in Lemma 3.4 we proved that $\lim_{\epsilon\to0}\sum_{k=1}^{k_\gamma}(\epsilon^2\nu_k^2/\lambda_k^2)=0$, and therefore we have $\lim_{\epsilon\to0}\sum_{k\in\mathcal{N}_k^{(1)}}\rho_k^2=0$. Regarding the second term on the right-hand side of formula (78), we have

$$\sum_{k\in\mathcal{N}_k^{(2)}}\rho_k^2\le\sum_{k=k_{\gamma_1}}^{\infty}\rho_k^2. \tag{80}$$

But, again, $\lim_{\epsilon\to 0}\sum_{k=k_{\gamma_1}}^{\infty}\rho_k^2 = 0$, and then $\lim_{\epsilon\to 0}\sum_{k\in\mathcal{N}_k^{(2)}}\rho_k^2 = 0$. □

Remarks. i) It is worth it to notice that the proof of Theorem 3.5 does not require any type of order in the sum (58). In fact, the only assumption that $\{\lambda_k\}$ is a strictly decreasing sequence does not evidently imply that the terms $(\lambda_k\rho_k/\epsilon\nu_k)$ have any type of monotonicity in k, and, consequently, the sum (58) cannot, in general, be regarded as an ordered sum of terms up to a certain maximum value of k. Thus, unlike the regularized solutions (12), (14), (18), (19), and also (45), $\widehat{B}\overline{g}$ features frequency selectivity, which is obtained by evaluating the information content of the noisy Fourier coefficients.

ii) Notice that the estimate (58) associated with the operator $\widehat{B}$ represents a probabilistically regularized solution, in the sense of the formula (68), even if, in general, it does not minimize the global mean-square error (30).

At this point in order to apply the results of this section, statistical methods that allow for splitting the coefficients $\overline{g}_k$ into the sets $\mathcal{I}_k$ and $\mathcal{N}_k$ must be investigated. These methods will be illustrated in the next section.

4. Numerical analysis: The regularizing algorithm.

4.1. The correlation function of the noisy data. The application of the results of the previous section to a Fredholm equation of the first kind would involve using statistical tools for the determination of the two sets $\mathcal{I}_k$ and $\mathcal{N}_k$. In this section this issue is discussed and the basic steps of a numerical algorithm for constructing the regularized solution $\widehat{B}\overline{g}$ from the noisy data $\overline{g}$ are outlined. For simplicity we shall work throughout only with data corrupted by white noise. However, provided the independence assumption between ξ and ζ, more general cases involving "colored" noise could be treated by using suitable methods, for instance, "prewhitening" transformations [7], whose discussion is beyond the scope of this section. Here our goal is to show that statistical estimates of the amount of information carried by the Fourier coefficients $\overline{g}_k$ can be sufficient to construct a satisfactory regularized solution. Furthermore, the direct comparison of the numerical results clearly evidentiates how some inherent limitations of the variational regularization scheme are overcome.

Following the analysis of the previous section, we are now faced with the problem of separating the Fourier coefficients $\overline{g}_k$ into two classes; one containing all the Fourier coefficients of the noisy data which are correlated, the other containing the $\overline{g}_k$ that can be regarded as random numbers. This task can be achieved by computing the correlation function of the random variables η_k; i.e., the probabilistic counterpart of the coefficients $\overline{g}_k$:

$$\Delta_\eta(k_1,k_2) = \frac{E\{[\eta_{k_1} - E\{\eta_{k_1}\}][\eta_{k_2} - E\{\eta_{k_2}\}]\}}{E\{[\eta_{k_1} - E\{\eta_{k_1}\}]^2\}^{1/2}E\{[\eta_{k_2} - E\{\eta_{k_2}\}]^2\}^{1/2}}. \tag{81}$$

In practice, just a finite realization $\{\overline{g}_k\}_1^N$ of the random variables η_k is available, from which estimates $\delta_{\overline{g}}$ of the autocorrelations can be obtained by regarding the data $\{\overline{g}_k\}_1^N$ as a finite length record of a stationary random normal series. In principle, the assumption of stationarity of the series $\{\eta_k\}$ is not correct because in general the moments of the random variables η_k will depend on k, but from the practical point of view this is usually the only possible chance. In fact, in many areas of application, it is difficult or even impossible to have multiple independent realizations $\{\overline{g}_k\}_1^N$ of the process $\{\eta_k\}$, so estimates of ensemble averages cannot be computed. Thus, we are forced to introduce the working hypothesis that the process $\{\eta_k\}$ is stationary in wide sense [9], that is, $\Delta_\eta(k_1,k_2) = \Delta_\eta(k_1 - k_2)$, and to compute the estimates

of the autocorrelation coefficients by means of the ergodic relation between ensemble and *time* (i.e., the index k in our case) averages. Of course, such a restriction can be removed whenever many independent sets of data $\{\overline{g}_k\}_1^N$ would be available for evaluating ensemble averages. Anyway, we will see later in the discussion of the algorithm how an ambiguity in the reconstruction of the regularized solution $\widehat{B}\overline{g}$ due to the assumed invariance for k-translation of $\{\eta_k\}$ will be removed.

A number of estimators of the autocorrelation function have been suggested by statisticians, and their properties are discussed in detail in [15]. An estimate which is widely used by statisticians, and in the following examples as well, is given by

$$\delta_{\overline{g}}(n) = \frac{\displaystyle\sum_{k=1}^{N-n} (\overline{g}_k - \langle \overline{g}_k \rangle)(\overline{g}_{k+n} - \langle \overline{g}_{k+n} \rangle)}{\left\{ \displaystyle\sum_{k=1}^{N-n} (\overline{g}_k - \langle \overline{g}_k \rangle)^2 \sum_{k=1}^{N-n} (\overline{g}_{k+n} - \langle \overline{g}_{k+n} \rangle)^2 \right\}^{1/2}}, \quad n = 0, ..., N-1, \tag{82}$$

where

$$\langle \overline{g}_k \rangle = \frac{1}{N-n} \sum_{k=1}^{N-n} \overline{g}_k \; ; \qquad \langle \overline{g}_{k+n} \rangle = \frac{1}{N-n} \sum_{k=1}^{N-n} \overline{g}_{k+n}. \tag{83}$$

Equation (82), which is based on the scatter diagram of $\overline{g}_{k+n}$ against $\overline{g}_k$ for $k = 1, \ldots, N-n$, represents the maximum likelihood estimate of the autocorrelation coefficients of two random variables η_k and η_{k+n} whose joint probability distribution function is bivariate normal.

In order to identify the structure of the series $\{\bar{g}_k\}_1^N$ so that we can separate correlated components from the random ones, it is necessary to have a crude test on whether $\delta_{\overline{g}}(n)$ is effectively zero. It has been shown by Anderson [1] that the distribution of an estimated autocorrelation coefficient, whose theoretical value is zero, is approximately normal. Thus, on the hypothesis that the theoretical autocorrelation $\Delta_\eta(n) = 0$, the estimate $\delta_{\overline{g}}(n)$ divided by its standard error $\sigma_\delta(n)$ will be approximately distributed as a unit normal deviate. This fact may be used to provide a rough guide as to whether theoretical autocorrelations are essentially zero. To this purpose it is usually sufficient to remember that, for normal distribution, deviations exceeding two standard errors in either direction have a probability of about 0.05, so that the 95% confidence interval of the estimate is approximately $\delta_{\overline{g}}(n) \pm 1.96\, \sigma_\delta(n)$.

Estimated autocorrelations can have rather large variances and can be highly correlated with each other [3, 12] so that care is required in the interpretation of individual autocorrelations. In particular, moderately large estimated autocorrelations can occur after the theoretical autocorrelation function has damped out, and, in any case, it must be considered that an estimated autocorrelation function always exhibits less damping than the theoretical one, as the estimated autocorrelations are inflated by sampling fluctuations (see also the following Example 1). Thus, in order to avoid a purely empirical analysis of the autocorrelations, it is necessary to assume a rough model of the series that allows us to evaluate the order of magnitude of the sampling errors $\sigma_\delta(n)$ associated with the autocorrelation estimator.

According to the discussion in section 3.2, since we expect to find the set $\mathcal{I}_k$ to be finite, we also expect that the autocorrelation function $\Delta_\eta(n)$ will vanish beyond a certain lag n_0. Thus, in what follows, it will be assumed that there exists an index n_0 such that $\Delta_\eta(n) = 0$ for $n > n_0$. In this case, if the record length N is

large enough (i.e., such that $O(1/N^2)$ terms can be neglected), use can be made of Bartlett's approximate expression for the variance of the estimated autocorrelations of a stationary normal process [3]:

$$\text{var}\,[\delta_{\overline{g}}(n)] \sim \frac{1}{N-n}\left\{1+2\sum_{v=1}^{n_0}\Delta_\eta^2(v)\right\} \quad \text{for} \quad n > n_0. \tag{84}$$

To use (84) in practice, the estimated autocorrelations $\delta_{\overline{g}}$ are substituted for the theoretical ones Δ_η, and in this case we shall refer to the square root of (84) as the *large-lag* standard error $\sigma_\delta(n; n_0)$ [7].

The index n_0 is actually recovered in a recursive way through a hypothesis generation-verification procedure. Starting from the assumption that the series is completely random, i.e., $n_0 = 0$, the standard error $\sigma_\delta(n; 0)$ is computed and the first index $\overline{n} > 0$ such that $|\delta_{\overline{g}}(\overline{n})| > 1.96\,\sigma_\delta(n; 0)$ is searched for. If there exists such an index $\overline{n}$, it becomes the new candidate to be n_0, i.e., we set $n_0 = \overline{n}$, $\sigma_\delta(n; n_0)$ is computed, and again it is tested whether the series is compatible with the hypothesis that $\Delta_\eta(n) = 0$ for $n > n_0$. The whole procedure is repeated until no new index $\overline{n}$ is found. Formally, n_0 is then defined as

$$n_0 = \max\{\overline{n} \geq 0 : \forall n \in (\overline{n}, N-1], \mid \delta_{\overline{g}}(n) \mid < 1.96\ \sigma_\delta(n, \overline{n})\}. \tag{85}$$

The set $\mathbf{Q}$ of the lags corresponding to autocorrelation values that are effectively different from zero and, consequently, indicating lack of randomness of the coefficients $\overline{g}_k$, is defined as

$$\mathbf{Q} = \{0 < n \leq n_0 : \mid \delta_{\overline{g}}(n) \mid > 1.96\ \sigma_\delta(n, 0)\}. \tag{86}$$

Let N_c be the number of elements of $\mathbf{Q}$.

As previously discussed, as a consequence of the inevitable assumption of stationarity of the process $\{\eta_k\}$, the Fourier coefficients $\overline{g}_k$ that are correlated cannot be determined in a unique way from the set $\mathbf{Q}$. In fact, an integer $n_i \in \mathbf{Q}$ just indicates a strong correlation between at least two Fourier coefficients n_i apart. This means that, in principle, any couple $(\overline{g}_{k_i}, \overline{g}_{k_i+n_i})$ for any integer $1 \leq k_i \leq (N - n_i)$ could have generated such a strong correlation at the lag n_i. Thus, from the set $\mathbf{Q}$ we can construct N_c families F_i defined as

$$F_i = \left\{(\overline{g}_{k_i}, \overline{g}_{k_i+n_i})\right\}_{k_i=1}^{(N-n_i)}, \quad i = 1, ..., N_c \tag{87}$$

from which the couples of coefficients $\overline{g}_k$ that are likely to be correlated can be selected. In theory, that is for $N \to \infty$, the N_c indices k_i and the N_c elements $n_i \in \mathbf{Q}$ are mutually dependent. In fact, any two coefficients $\overline{g}_{k_\alpha}, \overline{g}_{k_\beta}$ which are selected from the families F_i must satisfy the pairwise compatibility conditions requiring $|k_\alpha - k_\beta| \in \mathbf{Q}$. Or, in other words, it can be seen that, given the set $\mathbf{Q}$, the number $N_\mathcal{I}$ of admissible Fourier coefficients $\overline{g}_k$ is combinatorially constrained to be

$$\frac{1}{2}\,(1 + \sqrt{1 + 8N_c}) \leq N_\mathcal{I} \leq N_c + 1. \tag{88}$$

The left inequality in (88) follows directly from the observation that the maximum number of correlations among $N_\mathcal{I}$ coefficients is $\binom{N_\mathcal{I}}{2}$, then $N_c \leq \binom{N_\mathcal{I}}{2}$, whereas the

right inequality expresses that at least $(N_{\mathcal{I}}-1)$ distinct correlations can be computed among $N_{\mathcal{I}}$ coefficients (i.e., $N_c \geq N_{\mathcal{I}}-1$). For instance, if $N_c = 2$, we have from inequalities (88) that there need to be $N_{\mathcal{I}} = 3$ coefficients $\overline{g}_k$ to construct the set $\mathbf{Q}$, or, referring to (87), that the two indices k_1 and k_2 must coincide, i.e., $k_1 \equiv k_2 \geq 1$. In any case, the compatibility conditions are not sufficient to constrain in a unique way the selection of the coefficients $\overline{g}_k$ and, consequently, the construction of the regularized solution.

In practice, that is, when the record length N is finite and particularly when the signal-to-noise ratio (SNR) of the data $\overline{g}$ is small, the compatibility constraints cannot be assumed to be satisfied. In fact, because of the sampling fluctuations in the estimates $\delta_{\overline{g}}(n)$, some correlations which are actually different from zero could be incorrectly detected by the procedure discussed above. However, we shall see later in the discussion of the numerical examples how the compatibility constraints can provide us with a confidence check on the reliability of the regularized solution $\widehat{B}\overline{g}$.

In order to recover in a unique way from the set $\mathbf{Q}$ the Fourier coefficients that are likely to be correlated, we adopt the following criterion suggested by the definition itself of the autocorrelation function: for any $n_i \in \mathbf{Q}$, $i = 1, ..., N_c$, we select the pair $(\overline{g}_{k_i^\star}, \overline{g}_{k_i^\star + n_i})$ giving the maximum contribution to the autocorrelation estimate $\delta_{\overline{g}}(n_i)$; i.e., we define $k_i^\star$ as

$$k_i^\star = \arg \max_{k \in [1, N-n_i]} \{|\overline{g}_k \, \overline{g}_{k+n_i}|\}, \quad i = 1, ..., N_c, \tag{89}$$

and, accordingly, we can define the set of frequencies $\mathcal{I}_k$ exhibiting correlated Fourier coefficients as

$$\mathcal{I}_k = \{k_i^\star\}_1^{N_c} \cup \{k_i^\star + n_i\}_1^{N_c}, \tag{90}$$

where each element of $\mathcal{I}_k$ is counted only once.

4.2. Numerical examples. Throughout this section we shall consider as a sample problem the integral equation (1) with kernel

$$K(x,y) = \begin{cases} (1-x)\,y & \text{if} \quad 0 \leq y \leq x \leq 1, \\ x\,(1-y) & \text{if} \quad 0 \leq x \leq y \leq 1 \end{cases} \tag{91}$$

whose eigenfunctions and eigenvalues are, respectively,

$$\psi_k(x) = \sqrt{2}\,\sin(k\pi x), \tag{92}$$

$$\lambda_k = \frac{1}{k^2\pi^2}. \tag{93}$$

The data $g(x)$ have been noised by adding white noise $n(x)$, simulated by computer generated random numbers uniformly distributed in the interval $[-\epsilon, \epsilon]$ (see also [22] for a very preliminary numerical analysis of this problem). The examples shown hereafter differ for the choice of the input signal $f(x)$ and for the values of the noise boundary ϵ, whereas the performances of the algorithm are evaluated by direct comparison of the reconstructed signal with the true signal $f(x)$. In every example reported here, the approximations obtained through the variational scheme (see section 2.1) are computed by setting the constraint operator C such that $c_k = k$, $(k = 1, 2, ...)$, the parameter ϵ corresponding to the boundary on the noise equal to the dispersion of the noise D_ϵ (see (7)), and the boundary E on the solution equal to the norm of the unknown function, i.e., $E = \|f(x)\|$ (see (8)).

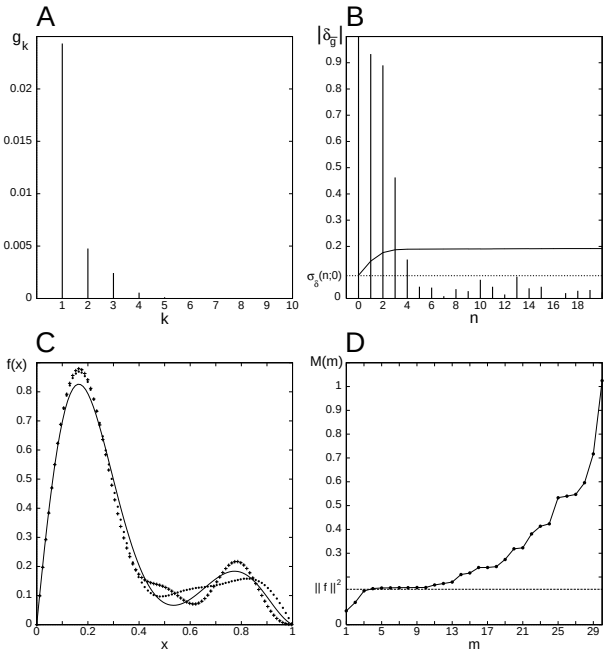

FIG. 1. *Example* 1: $f_1(x) = (1-x)\sin(3\sin(3x))$, $\epsilon = 10^{-4}$, *SNR* $\simeq 25.7dB$, $N = 512$. (A) *Noiseless Fourier coefficients* g_k. (B) *Modulus of the autocorrelation function. The horizontal dotted straight line indicates the* 95% *confidence limit* $1.96\,\sigma_\delta(n;0)$ *for a purely random sequence. The solid curved line indicates the confidence limit* $1.96\,\sigma_\delta(n;3)$, $n > 3$. *From the analysis of* $\delta_{\overline{g}}(n)$ *we have* $\mathbf{Q} = \{1,2,3\}$ *and* $\mathcal{I}_k = \{1,2,3,4\}$. (C) *Regularized solutions. The solid line represents the actual solution* $f_1(x)$. *The dots represent the reconstruction* $\widehat{B}\overline{g}$. *The crosses represent the variational solution* $f_\star^{(1)}$ *obtained by using* $c_k = k$; $k_\alpha = 8$ (*see equations* (14) *and* (15)). (D) *Plot of the function* $M(m) = \sum_{k=1}^{m} (\overline{g}_k/\lambda_k)^2$. *Notice that the value of* $M(m)$ *corresponding to its first plateau, i.e., approximately for* $4 \le m \le 10$, *is about the squared norm of the true solution.*

In Figure 1, the analysis of the sample function $f_1(x) = (1-x)\sin(3\sin(3x))$ with noise boundary $\epsilon = 10^{-4}$ is summarized. The global SNR, defined as the ratio of the mean power of the noiseless data to the noise variance, was SNR $\simeq 25.7$dB. The function $f_1(x)$ is characterized by having the bulk of information localized in the first few values of k (see the related noiseless coefficients g_k in Figure 1A) so that we expect that also a variational solution could provide a satisfactory reconstruction of the input signal. Figure 1B shows the behavior of the autocorrelation function $\delta_{\overline{g}}(n)$ along with the two lines indicating the statistical confidence limits we used to discriminate whether the autocorrelations are essentially null. The dashed horizontal straight line represents the threshold that we would have under the hypothesis of purely random sequence $\{\overline{g}_k\}$, whereas the solid line represents the threshold corresponding to the model of autocorrelation function of ideal damped type. In this example we found $n_0 = 3$, $\mathbf{Q} = \{1,2,3\}$, and the autocorrelation at $n = 4$ was rejected in spite of its quite large value (see formula (86)). The direct inspection of the values of $\epsilon\nu_k$ and g_k in repeated realizations showed that for $k = 5$ the noise was usually larger than the Fourier coefficient, confirming hence the result that the autocorrelation $\delta_{\overline{g}}(4)$ was abnormally inflated by the large autocorrelations at $n = 1,2,3$. According to the criteria (89) and (90), the set of frequencies whose corresponding Fourier coefficients

exhibit strong correlations is $\mathcal{I}_k = \{1, 2, 3, 4\}$. It is worth noticing that in this case the elements of $\mathcal{I}_k$ satisfy all the compatibility constraints; i.e., any difference between elements of $\mathcal{I}_k$ belongs to $\mathbf{Q}$, and $N_{\mathcal{I}}$ satisfies constraints (88). This complete cross consistency between $\mathbf{Q}$ and $\mathcal{I}_k$ gives a high level of confidence in the result of the whole analysis. In Figure 1C the true function f_1 (solid line), the regularized solution $\widehat{B}\overline{g}$ (crosses), and the regularized function $f_\star^{(1)}$ (dots) are compared. The truncation point of $f_\star^{(1)}$, obtained through the criterion (15), was $\alpha = 8$. Figure 1C shows how in this case both regularization methods lead to comparable results, which are quite satisfactory approximations of the "unknown" function f_1. The plot of the function $M(m)$, displayed in Figure 1D, confirms the correctness of the two approximations. In fact, it clearly exhibits a "plateau," ranging from about $m = 3$ to $m = 10$, that corresponds to the order–disorder transition of the coefficients $\overline{g}_k$. Then it could be argued that for any truncation point belonging to this "plateau" the truncated approximation will hold coefficients $\overline{g}_k$ whose information content is not completely obscured by the noise. In every example discussed here, the regularized solutions $f_\star^{(2)}$ and $f_\star^{(3)}$ (see (18) and (19)) have also been considered, providing in all cases worse results (not plotted).

The second and third examples, shown in Figure 2, are quite simple but a little tricky, and show the deep differences between our approach and the variational one. They consist of a finite linear combination of, respectively, 3 and 10 basis functions ψ_k (see the legend for numerical details), and, indeed, they have been chosen as typical signals in which the bulk of the information is not grouped in a single block of consecutive low frequencies. In these cases, setting global constraints on the solution, such as in the variational methods, leads inevitably to a failure, which is clearly evident from Figure 2C,F, since the lack of selectivity necessarily causes the regularized solution $f_\star^{(1)}$ to contain pure noisy components. On the contrary, the selectivity achieved through the analysis of the autocorrelation function overcomes this limit. In both examples the analysis of the autocorrelation function (see Figure 2A,D) led to the correct selection of the components that carry information in spite of the quite small SNR (in Example 2, SNR $\simeq 0.55$dB). Referring to the Example 2 depicted in Figure 2A,B,C, it can be observed that all the compatibility constraints are indeed satisfied; however, it is worth it to remark that, because of the sampling fluctuation of the estimates $\delta_{\overline{g}}(n)$, the autocorrelation $\delta_{\overline{g}}(6)$ was not always detected in different realizations of the noisy data $\{\overline{g}_k\}_1^N$. In these cases the set $\mathcal{I}_k$, computed from the set $\mathbf{Q} = \{4, 10\}$ missing $n = 6$, is still correct, i.e., $\mathcal{I}_k = \{3, 7, 13\}$, even though one compatibility constraint is not fulfilled.

A more complex example is shown in Figure 3. Following the trace of the previous example, here we have the input function f_4 which is characterized by having the significant Fourier components grouped in different ranges of the k axis. Consequently, the Fourier coefficients g_k that clearly emerge from the noise (in this example, $\epsilon = 10^{-4}$) are quite sparse in the range $1 \leq k \leq 12$ (see Figure 3A). The plot of the regularized solution $\widehat{B}\overline{g}$, obtained from the analysis of the autocorrelation function shown in Figure 3B, shows an acceptable agreement with the real solution f_4, even though the procedure failed in detecting the coefficient at $k = 5$. On the contrary, the "nontruncated" (in the sense that the sum runs up to N) solution $f_\star$ (see (12)), which is displayed in Figure 3D, yields a rather poor reconstruction either because the constraint operator C smooths out too many frequencies or because distortions are introduced by those coefficients which are essentially noise (e.g., $k = 2, 6, 7, 8, 9, 10$). Of course, the variational reconstruction could be considerably improved by choosing

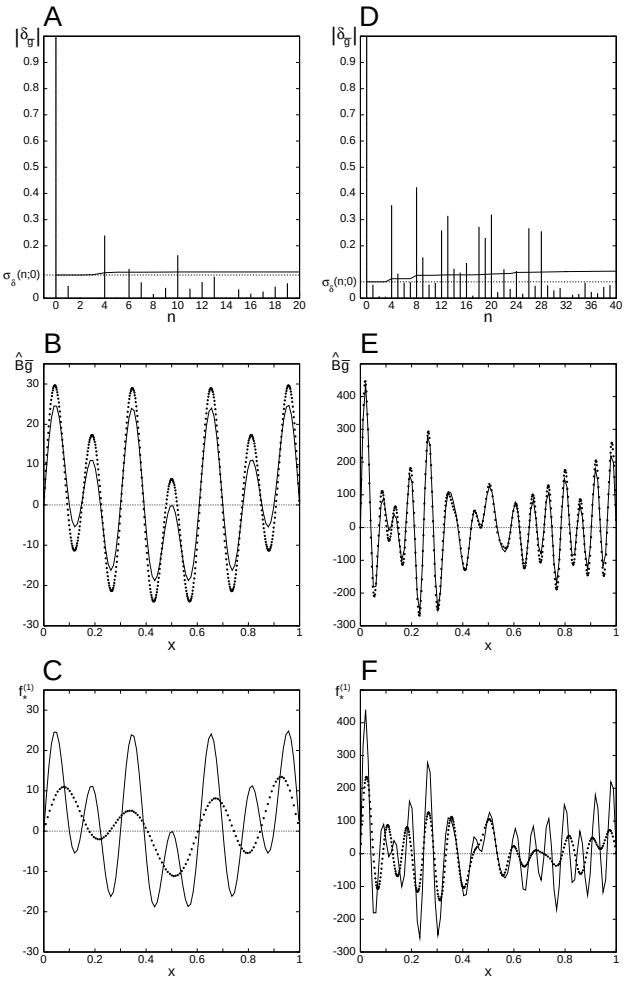

FIG. 2. *Example* 2: $f_2(x) = 5\sin(3\pi x)+10\sin(7\pi x)+15\sin(13\pi x)$, $\epsilon = 3\,10^{-3}$, $SNR \simeq 0.54dB$, $N = 512$. (A) *Modulus of the autocorrelation function.* $\mathbf{Q} = \{4, 6, 10\}$, $\mathcal{I}_k = \{3, 7, 13\}$. (B) *Comparison between the actual solution* $f_2(x)$ *(solid line) and the regularized solution* $\widehat{B\overline{g}}(x)$ *(dots).* (C) *Comparison between the actual solution* $f_2(x)$ *(solid line) and the approximated solution* $f_\star^{(1)}(x)$ *with* $k_\alpha = 9$ (*see criterion* (15)). (D) *Example* 3: *Modulus of the autocorrelation function.* $f_3(x) = \sum_{j=1}^{10} a_j \sin(k_j \pi x)$, *with* $a_j = \{17, 23, 27, 33, 43, 55, 68, 70, 77, 81\}$ *and* $k_j = \{5, 9, 13, 17, 18, 23, 24, 25, 31, 33\}$. $\epsilon = 10^{-3}$, $SNR \simeq 9.79dB$, $N = 1024$; $\mathbf{Q} = \{4, 5, 8, 9, 12, 13, 14, 15, 16, 18, 19, 20, 22, 24, 26, 28\}$, $\mathcal{I}_k = \{5, 9, 13, 17, 18, 23, 24, 25, 31, 33\}$. (E) *Comparison between the actual solution* $f_3(x)$ *(solid line) and the regularized solution* $\widehat{B\overline{g}}(x)$ *(dots).* (F) *Comparison between the actual solution* $f_2(x)$ *(solid line) and the approximated solution* $f_\star^{(1)}(x)$ *with* $k_\alpha = 27$.

a more appropriate operator C and different values for the parameters ϵ and E, but this would require more precise a priori knowledge on the actual solution.

In conclusion, some final remarks. The method of regularization based on the analysis of the correlation function of the data allows us to pick out the Fourier components of the noisy data which are likely to carry exploitable information on the unknown solution and, at the same time, for rejecting the ones dominated by the noise.

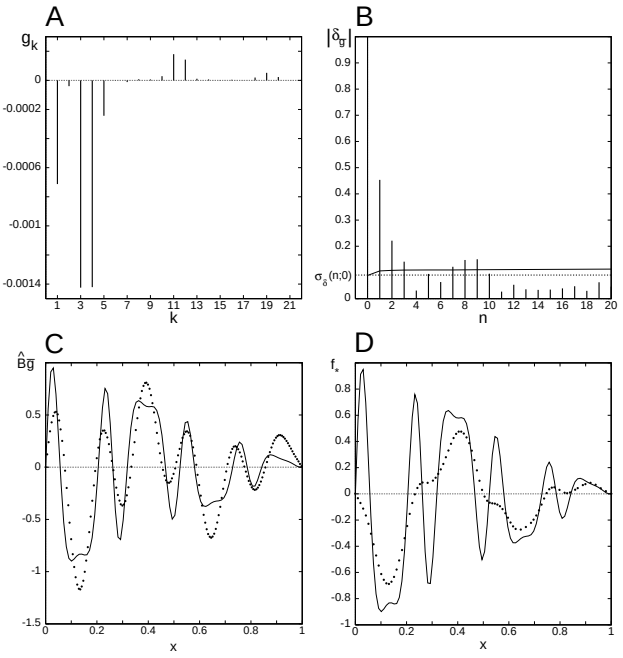

FIG. 3. *Example* 4: $f_4(x) = (1-x)\sin(5\sin(12x))$, $\epsilon = 10^{-4}$, $SNR \simeq 4.6dB$, $N = 512$. (A) *Noiseless Fourier coefficients* g_k. (B) *Modulus of the autocorrelation function.* $\mathbf{Q} = \{1,2,3,5,7,8,9\}$, $\mathcal{I}_k = \{1,3,4,9,11,12\}$. (C) *Comparison between the actual solution* $f_4(x)$ *(solid line) and the regularized solution* $\widehat{B}\overline{g}(x)$ *(dots).* (D) *Comparison between the actual solution* $f_4(x)$ *(solid line) and the variational solution* $f_\star(x)$ (*see* (12)).

Frequency selectivity is not featured by methods of regularization that basically work as low-pass filters, and we have seen this inherent limit through examples in which frequency selectivity is essential for a satisfactory reconstruction.

The regularized solution $\widehat{B}\overline{g}$ is founded only on a suitable analysis of the real data, that aims at holding only the data whose information content is significant. This approach naturally agrees with the methodology of the experimental physical science.

A moderate number of reasonable assumptions have been made in the construction of the regularized solution $\widehat{B}\overline{g}$ (see Theorem 3.5), and, more important, the solution itself does not depend on unknown parameters. Even in the variational approach, methods to reduce the dependence of the solution on free parameters have been widely investigated, and several practical strategies for choosing the regularization parameter α (see functional (9)) have been proposed (see, for instance, [8, 25] and references therein). Since the optimal parameter is impossible to determine because the exact solution is not known, many of these strategies can provide estimates of the asymptotically optimal rate of convergence of the regularized solution to the real solution when the noise vanishes.

The main difficulty of the method we have proposed regards the analysis of the correlation function. First, the correctness of the regularized solution depends on the capability of the correlation function to catch the information content of the data and to exhibit it in an effective way. Second, usually quite large data samples, i.e., N large, are necessary in order to limit sample fluctuations that could give rise to

incorrect interpretation of the correlation function itself.

REFERENCES

[1] R. L. Anderson, *Distribution of the serial correlation coefficient*, Ann. Math. Stat., 13 (1942), pp. 1–13.

[2] A. V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[3] M. S. Bartlett, *Stochastic Processes: Methods and Applications*, 3rd ed., Cambridge University Press, Cambridge, UK, 1978.

[4] M. Bertero and G. A. Viano, *On Probabilistic methods for the solution of improperly posed problems*, Boll. Un. Mat. Ital. B (5), 15 (1978), pp. 483–508.

[5] M. Bertero, C. De Mol, and G. A. Viano, *On the problems of object restoration and image extrapolation in optics*, J. Math. Phys., 20 (1979), pp. 509–521.

[6] M. Bertero, C. De Mol, and G.A. Viano, *The stability of inverse problems*, in Inverse Scattering Problems in Optics, Springer-Verlag, Berlin, 1980, pp. 161–212.

[7] G. E. P. Box and G. M. Jenkins, *Time Series Analysis*, Holden–Day, San Francisco, 1976.

[8] A. M. Davies, *Optimality in regularization*, in Inverse Problems in Scattering and Imaging, M. Bertero and E. R. Pike, eds., Adam Hilger, Bristol, UK, 1992, pp. 393–410.

[9] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[10] H. W. Engl, *Regularization methods for the stable solution of inverse problems*, Surveys Math. Indust., 3 (1993), pp. 71–143.

[11] J. N. Franklin, *Well-posed stochastic extensions of ill-posed linear problems*, J. Math. Anal. Appl., 31 (1970), pp. 682–716.

[12] W. A. Fuller, *Introduction to Statistical Time Series*, John Wiley, New York, 1976.

[13] I. M. Gel'fand and A. M. Yaglom, *Calculation of the amount of information about a random function contained in another such function*, Amer. Math. Soc. Transl. Ser. 2, 12 (1959), pp. 199–246.

[14] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Boston, 1984.

[15] G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications*, Holden–Day, San Francisco, 1968.

[16] M. Hanke, *Conjugate Gradient Type Methods for Ill-Posed Problems*, Pitman Res. Notes Math. Ser. 327, Longman Sci. Tech., Harlow, 1995.

[17] N. Magnoli and G. A. Viano, *On the eigenfunction expansions associated with Fredholm integral equations of first kind in presence of noise*, J. Math. Anal. Appl., 197 (1996), pp. 188–206.

[18] N. Magnoli and G. A. Viano, *The source identification problem in electromagnetic theory*, J. Math. Phys., 38 (1997), pp. 2366–2388.

[19] D. Middleton, *An Introduction to Statistical Communication Theory*, McGraw–Hill, New York, 1960.

[20] K. Miller, *Least square methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.

[21] K. Miller and G. A. Viano, *On the necessity of nearly-best-possible methods for analytic continuation of scattering data*, J. Math. Phys., 14 (1973), pp. 1037–1047.

[22] E. Scalas and G. A. Viano, *Resolving power and information theory in signal recovery*, J. Opt. Soc. Amer. A, 10 (1993), pp. 991–996.

[23] A. Tikhonov and V. Arsenine, *Méthodes de Rèsolution de Problémes Mal Posès*, Mir, Moscow, 1976.

[24] G. A. Viano, *On the regularization of the antenna synthesis problem*, in Partial Differential Equations and Applications, P. Marcellini, G. T. Talenti, and E. Vesentini, eds., Marcel Dekker, 1996, pp. 313–318.

[25] G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM J. Numer. Anal., 14 (1977), pp. 651–667.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 878–890, July 1998

004

"CHAOS GAMES" FOR ITERATED FUNCTION SYSTEMS WITH GREY LEVEL MAPS*

B. FORTE†, F. MENDIVIL‡, AND E. R. VRSCAY§

Abstract. Two random iteration algorithms, or "chaos games," for iterated function systems (IFS) on function spaces, namely IFS with grey level maps (IFSM), are described. The first algorithm can be interpreted as a "chaos game in code space" and is guaranteed to work only in the case of nonoverlapping IFS maps. In the second algorithm, applicable to IFSM with overlapping IFS maps but affine grey level maps, the (normalized) IFSM attractor function $\bar{u}$ serves as the density for an invariant measure $\bar{\mu}$ of an IFS with probabilities with condensation measure. As such, approximations to the attractor function of the IFSM are yielded by visitation histograms, as in the case of IFS with probabilities on measure spaces. Some computer results illustrating the convergence of this chaos game for a simple overlapping IFSM on [0,1] are also presented.

Key words. iterated function systems, chaos game, invariant measures, fractals

AMS subject classifications. 28A80, 58F08, 58F11, 60J15, 65C, 65D

PII. S0036141096306911

1. Introduction.

In this paper we formulate two *random iteration algorithms*, or "chaos games," for iterated function systems with grey level maps (IFSM). As in the case of iterated function systems with probabilities (IFSP) on probability measure spaces, the chaos game is a kind of "bin counting" algorithm which can be used to generate approximations to IFSM attractor functions. For both the IFSM and IFSP, such a random iteration algorithm represents an alternative to a *deterministic algorithm* for constructing such approximations.

In the remainder of this section, the basic definitions for IFSM are presented. In section 2, we outline a first kind of chaos game for IFSM, motivated by the chaos game for IFSP. (The important features of the latter are given in the appendix.) Its applicability in analyzing the IFSM attractor function $\bar{u}(x)$ is restricted to the case where the IFS contraction maps are "nonoverlapping." In this case, as expected, a sampling of function values by a random "chaos game" walk produces converging estimates of the average value of the attractor over a subset/pixel. The breakdown of the algorithm in the overlapping case can be understood by reformulating it as a chaos game over code space. A convergence result for this chaos game is provided by Elton's ergodic theorem. However, it can provide information about $\bar{u}(x)$ on X only in the nonoverlapping case. In section 3, we introduce another chaos game for affine IFSM in $\mathcal{L}^1$ in the general overlapping case. The (normalized) IFSM attractor function $\bar{u}$ is considered as the density of the invariant measure for an IFSP with condensation. The offset terms β_i in the affine grey level maps $\phi_i(t) = \alpha_i t + \beta_i$ will

*Received by the editors July 17, 1996; accepted for publication (in revised form) July 23, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/30691.html

†Facoltà di Scienze MM. FF. e NN. a Cà Vignal, Università Degli Studi di Verona, Strada Le Grazie, 37134 Verona, Italy (forte@biotech.sci.univr.it). The research of this author was partially supported by an NSERC collaborative projects grant.

‡Department of Applied Mathematics, Faculty of Mathematics, University of Waterloo, Waterloo, ON, N2L 3G1, Canada (mendivil@augusta.math.uwaterloo.ca).

§Department of Applied Mathematics, Faculty of Mathematics, University of Waterloo, Waterloo, ON, N2L 3G1, Canada (ervrscay@links.uwaterloo.ca). The research of this author was partially supported by an NSERC Operating Grant and NSERC Collaborative Projects Grant CPG0164670.

play the role of the condensation measure θ, while the scaling terms α_i will contribute to mixing probabilities for these condensation measures. We conclude the paper with some computer results that illustrate the convergence of the chaos game for a simple overlapping IFSM on [0,1].

1.1. Basics of IFSM. Let (X, d) denote a complete metric space, the "base space," typically $[0, 1]$ or $[0, 1]^2$ with Euclidean metric. Let $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$ be a set of one-to-one contraction maps on (X, d) with contraction factors $c_i \in [0, 1)$. For simplicity, we assume that the IFS maps are affine. Associated with the IFS maps w_i is a set of grey level maps $\Phi = \{\phi_1, \phi_2, \ldots, \phi_N\}$, $\phi_i : \mathbf{R} \to \mathbf{R}$, assumed to be Lipschitz on $\mathbf{R}$ with Lipschitz constants K_i. The set of IFS maps $\mathbf{w}$ and associated grey level maps Φ comprises an IFSM on (X, d).

Associated with the N-map IFSM $(\mathbf{w}, \Phi)$ is a *fractal transform* operator $T : \mathcal{L}^p(X) \to \mathcal{L}^p(X)$, $p \in [1, \infty)$ [6, 7]. For $u \in \mathcal{L}^p(X)$,

$$(Tu)(x) \;=\; \sum_{k=1}^{N} f_k(x), \qquad x \in X, \tag{1.1}$$

where the *fractal components* $f_k(x)$ are given by

$$f_k(x) = \begin{cases} \phi_k(u(w_k^{-1}(x))), & x \in w_k(X), \\ 0, & x \notin w_k(X). \end{cases} \tag{1.2}$$

In other words, the kth fractal component $f_k(x)$ is a modification of the grey level value of u at the preimage $w_k^{-1}(x)$ (provided this preimage exists).

For $u, v \in \mathcal{L}^p(X)$, $p \in [1, \infty)$,

$$\| Tu - Tv \|_p \leq C_p \, \| u - v \|_p, \qquad C_p = \sum_{k=1}^{N} |J_k|^{1/p} K_k, \tag{1.3}$$

where $|J_k|$ is the Jacobian associated with the transformation $x = w_k(y)$. These bounds may be improved in the μ-nonoverlapping case, where the sets $X_i = w_i(X)$ overlap only on sets of zero Lebesgue measure on X (a common assumption in the literature):

$$\| Tu - Tv \|_p \leq \bar{C}_p \, \| u - v \|_p, \qquad \bar{C}_p = \left[\sum_{k=1}^{N} |J_k| K_k^p \right]^{1/p}. \tag{1.4}$$

If $C_p < 1$, then T is contractive in $(\mathcal{L}^p, d_p)$. From the Banach contraction mapping theorem, there exists a unique fixed point $\bar{u} \in \mathcal{L}^p(X)$, i.e., $T\bar{u} = \bar{u}$. Furthermore, for any $u \in \mathcal{L}^p(X)$, $d_p(T^n u, \bar{u}) \to 0$ as $n \to \infty$. This is the basis for the deterministic algorithm to generate approximations to $\bar{u}$.

In what follows it will be useful to consider the *code space* associated with the N-map IFS $\mathbf{w}$. Recall that there exists a unique compact set $A \subseteq X$, the *attractor* of the IFS, such that $A = \cup_{i=1}^{N} w_i(A)$. Define

$$\Sigma = \{\sigma = (\sigma_1, \sigma_2, \ldots,) \mid \sigma_i \in \{1, 2, \ldots, N\} \; \forall \; i \geq 1\}. \tag{1.5}$$

Then for any $x \in A$, there exists at least one code $\sigma \in \Sigma$ such that

$$x = \lim_{n \to \infty} w_{\sigma_1} \circ w_{\sigma_2} \circ, \ldots, \circ w_{\sigma_n}(y) \tag{1.6}$$

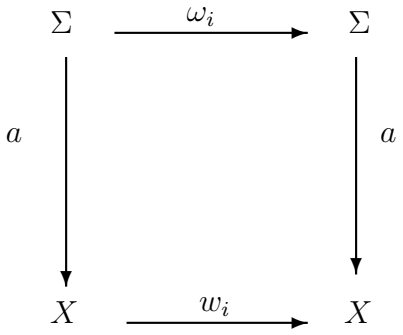

FIG. 1.1.

for any $y \in X$ [2, 1]. As well, for any $\sigma \in \Sigma$, there exists a unique point $x \in X$. We let $a : \Sigma \to X$ denote the *address map* so that $a(\sigma) = x$. If the sets $w_i(A)$ are disjoint, then a is one-to-one.

Finally, each IFS map w_i on X induces an equivalent action on Σ. The code space map corresponding to w_i is $\omega_i : (\sigma_1, \sigma_2, \dots,) \mapsto (i, \sigma_1, \sigma_2, \dots,)$. These actions are depicted schematically in Figure 1.1.

2. A simple chaos game.

2.1. Nonoverlapping sets $X_i = w_i(X)$. Let $P_K = \{B_k\}_{k=1}^K$ denote a partition of X into Borel subsets B_k. Associated with each set B_k is a "cumulative sum" S_k which will initially be set to 0. Let $(\mathbf{w}, \Phi)$ be an N-map IFSM on X satisfying the following properties:

1. The sets $X_i = w_i(X)$ cover X, i.e., $\bigcup_{k=1}^N X_i = X$.
2. As well, $m(X_i \cap X_j) = 0$ for $i \neq j$ ("nonoverlapping" sets).
3. The grey level maps ϕ_i are contractive on $\mathbf{R}$, i.e., $K_i \in [0, 1)$. (This implies that T is contractive on $\mathcal{L}^\infty(X)$ [3, 6].)

From the above assumptions, almost every point $x \in X$ (Lebesgue measure) has a unique code $\sigma = (\sigma_1, \sigma_2, \dots,) \in \Sigma$. If $\bar{u}$ is the attractor of the N-map IFSM $(\mathbf{w}, \Phi)$, then $\bar{u}(x) = \phi_{\sigma_1} \bar{u}(w_{\sigma_1}^{-1}(x))$, which may be iterated to obtain (using the contractivity of the ϕ_i)

$$\bar{u}(x) = \lim_{n \to \infty} \phi_{\sigma_1} \circ \phi_{\sigma_2} \circ, \dots, \circ \phi_{\sigma_n}(t_0), \tag{2.1}$$

where $t_0 \in \mathbf{R}$.

Let $\mathbf{p} = \{p_1, p_2, \dots, p_n\}$, $p_k > 0$, $\sum_{k=1}^N p_k = 1$ be a set of probabilities associated with each IFS grey level map pair (w_k, ϕ_k). The choice of the p_k — a crucial point — will be specified below. We now outline the first algorithm:

1. Initialize x_0 as the fixed point of w_1 (merely for convenience).
2. Initialize u_0 to be "close" to $u(x_0)$ by setting $u_0 = \phi_1^{\circ m}(1)$, where m is sufficiently large to obtain the desired accuracy. (Convergence is guaranteed by the contractiveness of the ϕ_k maps and the nonoverlapping property of the X_i.)
3. Initialize the sum $S_{j_0} = u_0$, where $x_0 \in B_{j_0}$.
4. Choose a pair $\{w_{\sigma_0}, \phi_{\sigma_0}\}$, $\sigma_0 \in \{1, 2, \dots, N\}$ according to the probabilities p_i.
5. Set $x_1 = w_{\sigma_0}(x_0)$ and $u_1 = \phi_{\sigma_0}(u_0)$.
6. Increment the sum S_{j_1} by u_1, where $x_1 \in B_{j_1}$.

7. Continue in this way by returning to 4 above, i.e.,

$$x_{n+1} = w_{\sigma_n}(x_n), \quad u_{n+1} = \phi_{\sigma_n}(u_n), \qquad \sigma_n \in \{1, 2, \ldots, N\}, \tag{2.2}$$

where the σ_k are chosen according to the probabilities p_i and then by updating the appropriate $S_{j_{n+1}}$.

PROPOSITION 2.1. *For each $k \in \{1, 2, \ldots, K\}$,*

$$\frac{S_k}{n} \to \int_{B_k} \bar{u}(x) d\bar{\mu}(x) \quad as \quad n \to \infty, \tag{2.3}$$

where $\bar{\mu}$ is the invariant measure of the IFSP $(\mathbf{w}, \mathbf{p})$. (See the appendix for the definition of IFSP.)

Proof. From the assumptions that (1) the ϕ_i maps are contractive and (2) the sets $w_i(X)$ are nonoverlapping, it follows that $u_n \approx \bar{u}(x_n)$. Let I_k denote the characteristic function of B_k. Then at the nth stage of this chaos game,

$$\frac{S_k}{n} \approx \frac{1}{n} \sum_{m=1}^{n} I_k(x_m) \bar{u}(x_m). \tag{2.4}$$

From Elton's theorem [5], in the limit $n \to \infty$ the right-hand side of the above expression becomes $\int_{B_k} \bar{u}(x) d\bar{\mu}(x)$. □

COROLLARY 2.2. *Define the probabilities to be $p_i = m(X_i)/\sum_k m(X_k)$. Then*

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \frac{S_k}{m(B_k)} &= \frac{1}{m(B_k)} \int_{B_k} \bar{u}(x) dm(x), \\ &= \bar{u}_{av}(B_k), \end{aligned} \tag{2.5}$$

the average value of $\bar{u}$ over B_k.

Proof. An easy calculation shows that the invariant measure for the IFSP $(\mathbf{w}, \mathbf{p})$ is $\mu = m$. From Eq. (2.4) the desired result follows. □

PROPOSITION 2.3. *Let $\mathcal{P}_n$ be a nested sequence of Borel partitions whose "sizes" go to zero as $n \to \infty$. Let $\bar{u}_n$ be the average value function of $\bar{u}$ associated with $\mathcal{P}_n$. Suppose that T is contractive in $\mathcal{L}^p(X, m)$ $(1 \le p < \infty)$ so that its fixed point $\bar{u} \in \mathcal{L}^p$. Then $\bar{u}_n$ converges to $\bar{u}$ in $\mathcal{L}^p$.*

Proof. Notice that $\bar{u}_n$ is the conditional expectation of $\bar{u}$ given $\mathcal{P}_n$. Thus, $\bar{u}_n$ forms a martingale sequence which is $\mathcal{L}^p$ bounded. The desired convergence then follows by the martingale convergence theorem [9]. □

Remarks. 1. The above implies that we may obtain an approximation to any accuracy (in the $\mathcal{L}^p$ sense) by using a sufficiently fine partition of X. In the particular case that $\bar{u} \in \mathcal{L}^1$, a stronger result follows from the martingale convergence theorem, namely, that $\bar{u}_n \to \bar{u}$ pointwise a.e.

2. In the general case of probabilities p_i, $\sum_i^N p_i = 1$, the limit in Corollary 2.2 becomes $\bar{u}_{av}(B_k, \bar{\mu})$, the $\bar{\mu}$-average value of $\bar{u}$ over B_k. Then Proposition 2.3 is generalized to convergence of $\bar{u}_n$ to $\bar{u}$ in $\mathcal{L}^p(\mu)$. The specific results of Corollary 2.2 and Proposition 2.3, i.e., convergence with respect to Lebesgue measure, are more relevant to computer approximations using the chaos game.

2.2. Overlapping $w_i(X)$ and a "chaos game in code space." In the case that the sets $X_i = w_i(X)$ overlap, i.e., $m(X_i \cap X_j) \neq 0$ for some pair (i, j), $i \neq j$, the chaos game of the previous section fails. One immediate consequence of overlapping

is that in Step 2 of the algorithm, it is not guaranteed that u_0 may be made "close" to $\bar{u}(x_0)$, since x_0 may have several preimages $w_i^{-1}(x_0)$. In order to further understand this problem, we formulate a chaos game algorithm on the code space Σ rather than on the base space X. This is possible from the equivalence of actions in both spaces as shown in Figure 1.1 at the end of section 1.1. Associated with the partition P_K of X into Borel subsets B_k is a partition of Σ into subsets defined by $T_k = a^{-1}(B_k)$ for all k, where $a : \Sigma \to X$ is the address map defined in section 1.1. To each T_k we now associate a cumulative sum S_k, initialized to zero.

Instead of considering functions $u : X \to \mathbf{R}$, as was done in the previous section, we consider the function $f : \Sigma \to \mathbf{R}$ defined as follows: for $\sigma \in \Sigma$, define

$$f(\sigma) = \lim_{n\to\infty} \phi_{\sigma_1} \circ \phi_{\sigma_2} \circ, \ldots, \circ \phi_{\sigma_n}(t_0), \tag{2.6}$$

where $t_0 \in \mathbf{R}$. The limit exists and is independent of t_0 by the assumption that the ϕ_i are contractive on $\mathbf{R}$. This is the "code space analogy" of the base space attractor function $\bar{u}$ as defined in Eq. (2.1).

We now modify the algorithm of the previous section to produce a chaos game on Σ instead of X. This is simply done by replacing B_{j_k} by T_{j_k}—the "bins" are now in the code space Σ instead of the base space X. Let $\sigma_1, \sigma_2, \ldots, \sigma_n, \ldots$, be the indices of the (w_i, ϕ_i) pairs chosen. Elton's ergodic theorem [5] guarantees the following result.

PROPOSITION 2.4. *For each $k \in \{1, 2, \ldots, K\}$,*

$$\frac{S_k}{n} \;\approx\; {\sum}' f(\tau_n) I_{T_j} \;\to\; \int_{T_k} f(\sigma) dP(\sigma), \quad \textit{as} \quad n \to \infty. \tag{2.7}$$

The prime indicates summation over codes $\tau_n = (\sigma_{n-l}, \sigma_{n-l+1}, \ldots, \sigma_n)$, where l is sufficiently large so that $f(\tau_n) \approx f(\sigma)$. As well, P denotes the invariant measure of the IFS with probabilities $(\mathbf{w}, \mathbf{p})$ on Σ.

In order to obtain an approximation to the IFSM attractor $\bar{u}(x)$ on X, it is necessary to interpret the above algorithm as acting on X. One may try to accomplish this by using the address map a to "push" the process onto X. From Eq. (2.7),

$$\frac{S_k}{n} \;\to\; \int_{T_k} f(\sigma) dP(\sigma). \tag{2.8}$$

This integral may involve a summation over different regions B_{k_i} which are mapped to T_k. In order to obtain a true approximation to $\bar{u}$ on X, however, we require the following quantity:

$$\int_{B_k} \left(\sum_{\sigma \in a^{-1}(x)} f(\sigma) \right) dP(a^{-1}(x)). \tag{2.9}$$

The quantities in Eqs. (2.8) and (2.9) are not necessarily identical. Equality is guaranteed only in the case that a is injective, i.e., the sets $X_i = w_i(X)$ are nonoverlapping. We illustrate the problem with the overlapping case by means of a simple example.

Example. $X = [0, 1]$ with

$$\begin{aligned} &w_1(x) = w_2(x) = \frac{1}{2}x, \quad w_3(x) = \frac{1}{2}x + \frac{1}{2}, \\ &\phi_1(t) = \phi_2(t) = \frac{1}{2}, \quad \phi_3(t) = 1. \end{aligned} \tag{2.10}$$

Then $\bar{u}(x) = 1$. Let $B_1 = [0, 1/2] = w_1(X)$ and $B_2 = [1/2, 1] = w_3(X)$. Consider the case $k = 1$ in Eqs. (2.8) and (2.9). Then $f = 1/2$ on T_1 because $\phi_1 = \phi_2 = 1/2$. The integral in Eq. (2.8) becomes

$$\begin{aligned} \int_{\sigma_1=1} \frac{1}{2} dP(\sigma) + \int_{\sigma_1=2} \frac{1}{2} dP(\sigma) &= \int_0^{\frac{1}{2}} \frac{1}{2} dP(a^{-1}(x)) \\ &= \frac{1}{2}(p_1 + p_2). \end{aligned} \tag{2.11}$$

The corresponding integral in Eq. (2.9) is

$$\begin{aligned} \int_{\sigma_1=1} \left(\frac{1}{2} + \frac{1}{2}\right) dP(\sigma) + \int_{\sigma_1=2} \left(\frac{1}{2} + \frac{1}{2}\right) dP(\sigma) &= \int_0^{\frac{1}{2}} \left(\frac{1}{2} + \frac{1}{2}\right) dP(a^{-1}(x)) \\ &= (p_1 + p_2). \end{aligned} \tag{2.12}$$

The two integrals are clearly not equal. The problem is due to the existence of "cross terms" in the integral of Eq. (2.9), which are not present in Eq. (2.8).

In the case that the address map a is injective, the integrals in Eqs. (2.8) and (2.9) are identical, and Proposition 2.1 of the previous section follows. The following generalization of Corollary 2.2 also follows:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \frac{S_k}{\bar{\mu}(B_k)} &= \frac{1}{\bar{\mu}(B_k)} \int_{B_k} \bar{u}(x) d\bar{\mu}(x) \\ &= \bar{u}_{av}(B_k, \bar{\mu}), \end{aligned} \tag{2.13}$$

the $\bar{\mu}$-average of $\bar{u}$ over B_k. Approximations to this average value may be obtained by running the standard chaos game for the invariant measure $\bar{\mu}$ of the IFSP $(\mathbf{w}, \mathbf{p})$ simultaneously: simply include another set of accumulation variables m_k, $1 \le k \le K$, and increment the appropriate m_j by 1 at each step. Then $m_k/n \to \bar{\mu}(B_k)$ as $n \to \infty$ so that $S_k/m_k \to \bar{u}_{av}(B_k, \bar{\mu})$.

In summary, the simple "chaos game" algorithms outlined above — one in the base space X and the other in the code space Σ — are guaranteed to work only in the special case of nonoverlapping $w_i(X)$. We are unable to construct comparable algorithms for the more general case of overlapping $w_i(X)$.

This is not to say that the algorithms never work in the overlapping case. They may work in "nongeneric" situations, for example, when grey level maps ϕ_i corresponding to overlapping IFS maps are identically zero. These are special cases, however. Our simple example clearly illustrates that the algorithms are not universally applicable. This serves as a motivation for the work outlined in the next section, in which a chaos game based on IFS with probabilities and condensation measures is devised.

3. Chaos game using IFSP with condensation. In what follows we assume, for simplicity of notation, that $X = [0, 1]$. Let $(\mathbf{w}, \Phi)$ denote an N-map *affine* IFSM, i.e., both IFS and grey level maps are affine:

$$w_i(x) = s_i x + a_i, \quad \phi_i(t) = \alpha_i t + \beta_i, \quad 1 \le i \le N. \tag{3.1}$$

Note that the sets $w_i(X)$ are *not* assumed to be nonoverlapping. The associated fractal transform operator T has the form

$$(Tu)(x) = \sum_{k=1}^{N} \alpha_k u\left(\frac{x - a_k}{s_k}\right) I_{X_k}(x) \; + \; \sum_{k=1}^{N} \beta_k I_{X_k}(x), \tag{3.2}$$

where $X_k = w_k(X)$. We shall write the above operation symbolically as

$$T(u) = A(u) + b, \tag{3.3}$$

where $b(x)$ is defined by the second sum in Eq. (3.2). We also assume that $\alpha_i, \beta_i \geq 0$, and that

$$C_1 = \sum_{k=1}^{N} c_i \alpha_i < 1, \tag{3.4}$$

where $c_i = |s_i|$, i.e., T is contractive in $\mathcal{L}^1$. Then the fixed point $\bar{u} = T\bar{u}$ may be written as follows:

$$\begin{aligned} \bar{u} &= b + A(\bar{u}) \\ &= \sum_{n=0}^{\infty} A^n(b). \end{aligned} \tag{3.5}$$

The iterated application of A on the function b mimics the operation of "condensation" in IFSP with condensation measures (reviewed in the appendix). The nature of this condensation is clarified if we consider the (normalized) attractor $\bar{u}$ as the density function of a probability measure $\bar{\mu}$ on X.

From the relation $T\bar{u} = \bar{u}$, one may easily compute the following integral:

$$\langle \bar{u} \rangle = \int_X \bar{u}(x)dx = \frac{\sum_k c_k \beta_k}{1 - \sum_k c_k \alpha_k}. \tag{3.6}$$

(Note that the denominator does not vanish.) As well,

$$\langle b \rangle = \int_X b(x)dx = \sum_k c_k \beta_k. \tag{3.7}$$

We also have from the relation $T\bar{u} = \bar{u}$, for any Borel set $S \subset X$,

$$\int_S \bar{u}(x)dx = \sum_{k=1}^{N} \alpha_k c_k \int_{w_k^{-1}(S)} \bar{u}(x)dx \ + \ \int_S b(x)dx. \tag{3.8}$$

Define the normalized functions $\bar{u}_1(x) = \bar{u}(x)/\langle \bar{u} \rangle$ and $b_1(x) = b(x)/\langle b \rangle$ so that $\langle \bar{u}_1 \rangle = \langle b_1 \rangle = 1$. Rewrite Eq. (3.8) in terms of these normalized functions to obtain

$$\bar{\mu}(S) = \sum_{k=1}^{N} \alpha_k c_k \bar{\mu}(w_k^{-1}(S)) \ + \ \left[1 - \sum_{k=1}^{N} \alpha_k c_k \right] \theta(S), \tag{3.9}$$

where $\bar{\mu}(S) = \int_S \bar{u}_1(x)dx$ and $\theta(S) = \int_S b_1(x)dx$. Thus, the measure $\bar{\mu} \in \mathcal{M}(X)$, with density $\bar{u}_1$, is the invariant measure of an IFSP with condensation measure $\theta \in \mathcal{M}(X)$ with density $\bar{b}_1$; cf. Eq. (A.10) in the appendix. Let $p_i = \alpha_i c_i$, $1 \leq i \leq N$ be the probabilities associated with the IFS maps w_i and let $p_0 = 1 - \sum_i^N \alpha_i c_i$ be the probability associated with the condensation measure θ. Our chaos game for affine IFSM will now be based on a chaos game for IFSP with condensation.

As in the previous section, we assume that X is partitioned into Borel subsets B_k. The cumulative sums S_k associated with the sets B_k are again initialized to zero. Here, however, they will supply information only on the visitation of the sets B_k. The algorithm is as follows:

1. Initialize x_0 as the fixed point of w_1 (merely for convenience).
2. Set $S_{j_0} = 1$, where $x_{j_0} \in B_j$.
3. Choose a $\sigma_1 = i \in \{0, 1, 2, \ldots, N\}$ according to the probabilities p_i. If
    1. $\sigma_1 \geq 1$, then define $x_1 = w_i(x_0)$. Increment the sum S_{j_1} by one, where $x_1 \in B_{j_1}$;
    2. $\sigma_1 = 0$, then choose x_1 according to the distribution with b_1 as its density. Increment the sum S_{j_1} by one, where $x_1 \in B_{j_1}$;
    3. Continue in this way, choosing the next σ_n according to the probabilities p_i, either (a) setting $x_{n+1} = w_{\sigma_n}(x_n)$ or (b) sampling from b_1. Then increment the appropriate $S_{j_{n+1}}$ accordingly.

At the nth stage, the approximation to $\bar{u}$ on B_k yielded by the above algorithm will be given by

$$\bar{u}_{av}(B_k) \approx \frac{1}{n}\left(\frac{S_j}{m(B_j)}\right)\left(\frac{\sum_k \beta_k c_k}{1 - \sum_k \alpha_i c_i}\right). \tag{3.10}$$

PROPOSITION 3.1. *The above approximations converge to the average value of $\bar{u}$ over the set B_k as $n \to \infty$.*

Proof. By Proposition 6 in [4], we have $S_k/n \to \bar{\mu}(B_k)$, where $\bar{\mu}$ is the invariant measure for the IFSP with condensation $(\mathbf{w}, \mathbf{p})$. We write the IFSP Markov operator as

$$\begin{aligned}(M\nu)(B) &= \sum_{i=1}^{N} p_i \nu(w_i^{-1}(B)) + p_0\theta(B) \\ &= (\bar{A}\nu)(B) + p_0\theta(B). \end{aligned} \tag{3.11}$$

Thus,

$$\bar{\mu} = \sum_{n}^{\infty} \bar{A}^n(\theta). \tag{3.12}$$

This shows that $\bar{\mu}$ is absolutely continuous with respect to Lebesgue measure since each term $\bar{A}^n(\theta)$ is. Since $\bar{\mu}$ is invariant with respect to M, its Radon–Nikodým derivative (i.e., its density) must also be invariant with respect to M. By scaling, we obtain the desired result. ∎

In order for the connection between IFSM and IFSP with condensation to be possible, the IFSM must be affine so that the operator T may be written as in Eq. (3.3). However, some generalizations may be made:

1. The shift or "offset" terms β_i in the grey level maps may be generalized to nonconstant functions of $x \in X$. In other words, the function $b(x)$ in Eq. (3.3) need not be a piecewise constant function. In order to define a density function for the condensation measure θ (cf. Eqs. (3.8), (3.9)) it is sufficient that the $\beta_i \in \mathcal{L}^1(X)$. The nonnegativity condition on the β_i may also be relaxed, as shown below.

2. Consider the affine IFSM operator T in Eq. (3.3), with $b(x)$ negative but bounded from below on X. Now let c be a positive constant (or a nonnegative $\mathcal{L}^1$ function) such that $b + c \geq 0$ on X. Now define the affine IFSM operator T', where

$$T'(u) = A(u) + (b + c), \qquad u \in \mathcal{L}^1(X). \tag{3.13}$$

Then T' is contractive with fixed point $\bar{u}' \in \mathcal{L}^1(X)$ given by

$$\bar{u}' = \sum_{n=0}^{\infty} A^n(b + c) \tag{3.14}$$

$$= \sum_{n=0}^{\infty} A^n(b) \; + \; \sum_{n=0}^{\infty} A^n(c)$$
$$= \bar{u} \; + \; \bar{v},$$

where $\bar{v}$ is the fixed point for the IFSM operator T_c, where $T_c(u) = A(u)+c$. Therefore $\bar{u} = \bar{u}' - \bar{v}$. "IFSP with condensation" chaos games may now be run separately for the two operators T' and T_c since they both have nonnegative condensation functions. Two accumulation sums are now computed in the algorithm, and approximations to the attractor $\bar{u}$ may then be obtained by subtracting these sums.

Example. Consider the following 3-map affine IFSM on [0,1] with overlapping IFS maps:

$$\begin{aligned} &w_1(x) = 0.5x, && \phi_1(t) = 0.6t + 0.2, \\ &w_2(x) = 0.4x + 0.3, && \phi_2(t) = 0.25t + 0.25, \\ &w_3(x) = 0.6x + 0.4, && \phi_3(t) = 0.4t + 0.6. \end{aligned} \tag{3.15}$$

The IFSM operator T is contractive in $\mathcal{L}^1(X)$. A histogram approximation of the attractor $\bar{u}$ of this IFSM is shown in Figure 3.1. This approximation was obtained by using the deterministic algorithm on an equipartition of [0,1] using 2000 subintervals. A discrete version of the IFSM operator was then iterated until satisfactory convergence was obtained. (Twenty-five iterations required about 0.03 CPU sec. All calculations were performed on an IBM RISC 6000 Model 43P-100.) Figures 3.2, 3.3, and 3.4 show histogram approximations to $\bar{u}$ yielded by the chaos game after 250,000, 2,000,000 and 10,000,000 iterations, respectively. (Again, 2000 "bins" B_k were employed.) In all three cases, less than 0.01 CPU sec. was required. It is evident that the approximations yielded by the chaos game are converging.

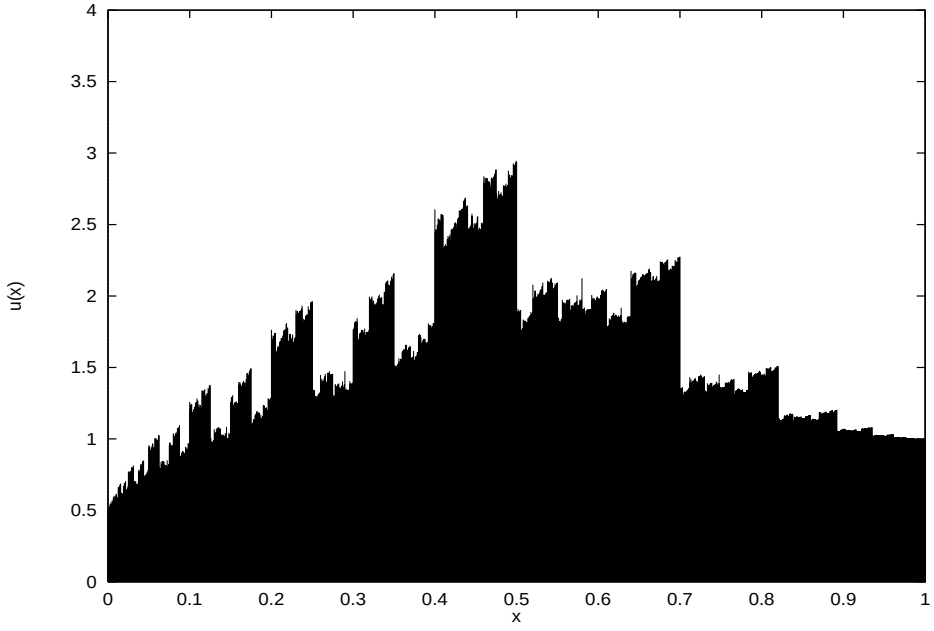

FIG. 3.1. *The attractor $\bar{u}$ to the 3-map affine IFSM.*

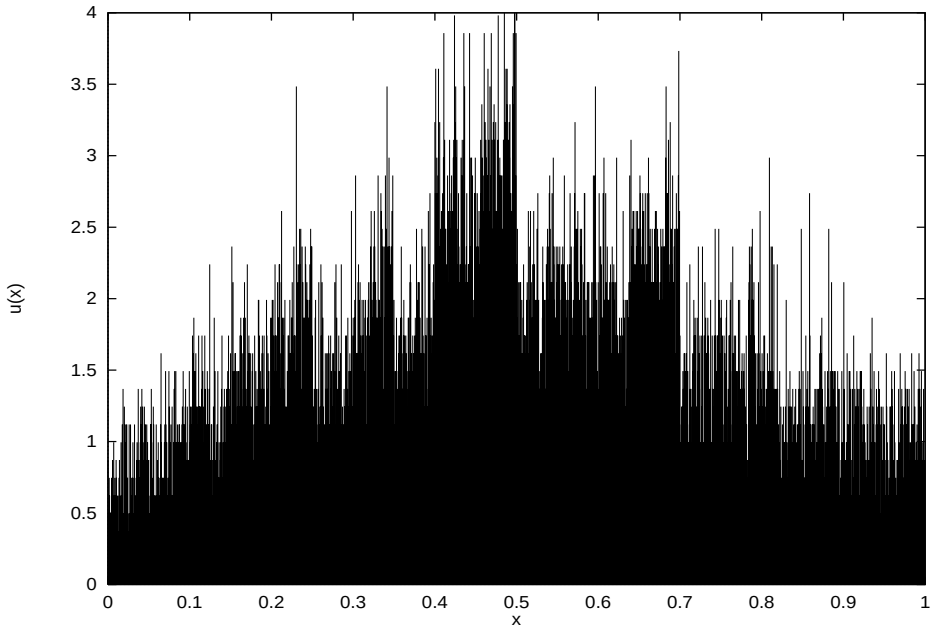

FIG. 3.2. *Chaos game approximation to* $\bar{u}$ *after* $250,000$ *iterations.*

Appendix.

IFSP and the "chaos game."

Let $(\mathbf{w}, \mathbf{p})$ be an N-map IFS with probabilities, i.e., $\mathbf{p} = \{p_1, p_2, ..., p_N\}$, $p_i \geq 0$. Let $\mathcal{B}(X)$ denote the σ-algebra of Borel subsets of X and $\mathcal{M}(X)$ the set of all probability measures on $\mathcal{B}(X)$. Associated with the N-map IFSP $(\mathbf{w}, \mathbf{p})$ is a (Markov) operator $M : \mathcal{M}(X) \to \mathcal{M}(X)$ such that for a $\mu \in \mathcal{M}(X)$ and any $S \in \mathcal{B}(X)$,

$$(M\mu)(S) = \sum_{i=1}^{N} p_i \mu(w_i^{-1}(S)). \tag{3.16}$$

M is contractive on $\mathcal{M}(X)$ [8],

$$d_H(M\mu, M\nu) \leq c d_H(\mu, \nu), \quad \forall \mu, \nu \in \mathcal{M}(X), \tag{3.17}$$

where $c = \max_{1 \leq i \leq N}\{c_i\}$ and d_H is the Monge–Kantorovich metric, referred to in the IFS literature as the "Hutchinson metric" due to its use in [8]. Thus, there exists a unique $\bar{\mu} \in \mathcal{H}(X)$ such that (1) $M\bar{\mu} = \bar{\mu}$ and (2) $d_H(M^n\mu, \bar{\mu}) \to 0$ as $n \to \infty$ for any $\mu \in \mathcal{M}(X)$. Moreover, $\text{supp}(\bar{\mu}) \subseteq A$, with the equality when all $p_i > 0$.

Given an $f \in C(X)$ and a $\mu \in \mathcal{M}(X)$, then

$$\int_X f(x) d(M\mu)(x) = \int_X (M^\dagger f) d\mu, \tag{3.18}$$

where the adjoint operator $M^\dagger : C(X) \to C(X)$ (referred to as T in [2]) is given by

$$(M^\dagger f)(x) = \sum_{k=1} p_k (f \circ w_k)(x). \tag{3.19}$$

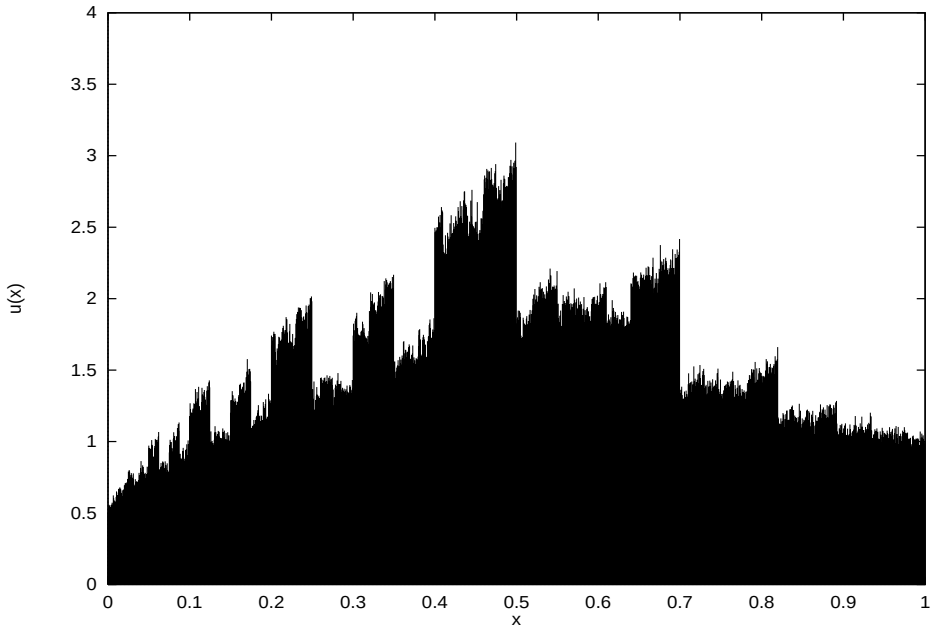

FIG. 3.3. *Chaos game approximation to* $\bar{u}$ *after* $2,000,000$ *iterations.*

The above procedure may be iterated to obtain, for $n = 1, 2, \ldots$,

$$\int_X f(x)d(M^n\mu)(x) = \sum_{i_1,\ldots,i_n}^{N} p_{i_1},\ldots,p_{i_n} \int_X (f \circ w_{i_1} \circ, \ldots, \circ w_{i_n})(x)d\mu(x). \tag{3.20}$$

Now set $\mu = \delta_{x_0}$, the Dirac unit mass at $x_0 \in X$ and $f(x) = I_S(x)$, where $S \subseteq X$, to give

$$\bar{\mu}(S) = \lim_{n\to\infty} \sum_{i_1,\ldots,i_n}^{N} p_{i_1},\ldots,p_{i_n} I_S(w_{i_1} \circ, \ldots, \circ w_{i_n}(x_0)). \tag{3.21}$$

The term involving I_S indicates whether or not the point $w_{i_1} \circ, \ldots, \circ w_{i_n}(x_0)$ lies in S. The quantity $p_{i_1}p_{i_2},\ldots,p_{i_n}$ represents the probability of choosing the finite sequence $\{\tau_{i_1}, \tau_{i_2}, \ldots, \tau_{i_n}\}$. Therefore, for each $n > 0$ the sum is equal to the probability that the point x_n lies in S.

There is a connection between Eq. (A.6) and the random iteration algorithm or "chaos game" [1], defined as follows: pick an $x_0 \in X$ and define the iteration sequence

$$x_{n+1} = w_{\tau_n}(x_n), \qquad n = 0, 1, 2, \ldots, \tag{3.22}$$

where the τ_n are chosen randomly and independently from the set $\{1, 2, \ldots, N\}$ with probabilities $P(\tau_n = i) = p_i$. A straightforward coding argument shows that for almost every code sequence $\tau = \{\tau_1, \tau_2, \ldots,)$ the orbit $\{x_n\}$ is dense on the attractor A of the IFS $\mathbf{w}$. As such, the chaos game can be used to generate computer approximations of A. However, it also provides approximations to the invariant measure $\bar{\mu}$ as a consequence of the following ergodic theorem for IFS [5]: for almost all code

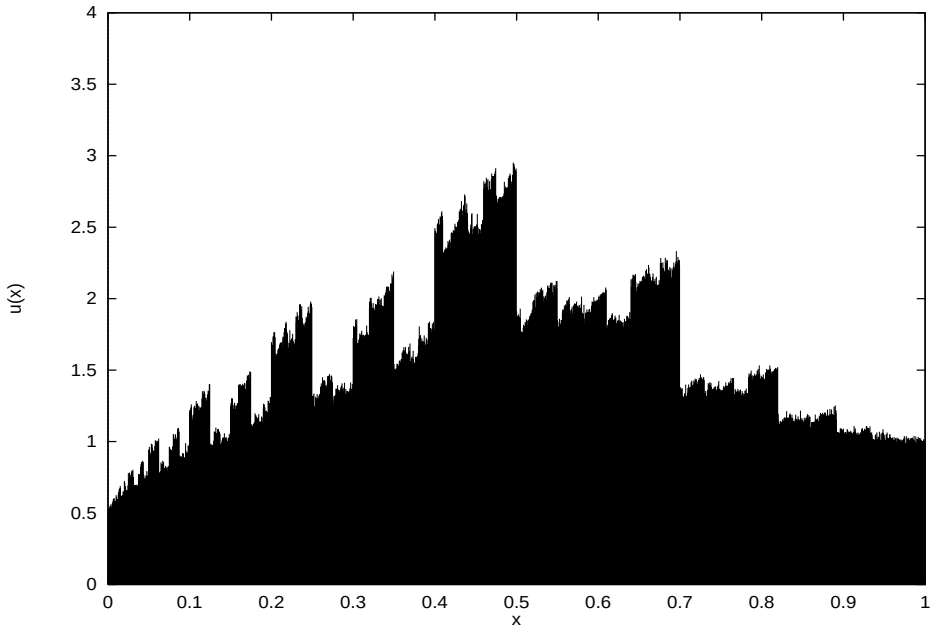

FIG. 3.4. *Chaos game approximation to* $\bar{u}$ *after* $10,000,000$ *iterations.*

sequences $\tau = (\tau_1, \tau_2, \dots,)$,

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} f(x_k) = \int_X f(x)\bar{\mu}(x) \tag{3.23}$$

for all continuous (and simple) functions $f : X \to \mathbf{R}$. By setting $f(x) = I_S(x)$ in Eq. (A.8) for an $S \subseteq X$, we obtain

$$\bar{\mu}(S) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} I_S(x_k). \tag{3.24}$$

In other words, $\bar{\mu}(S)$ is the limit of the relative visitation frequency of S during the chaos game.

IFSP with condensation. Consider an N-map IFSP on (X, d) with a condensation measure $\theta \in \mathcal{M}(X)$, i.e.,

1. $w_i : X \to X$, $1 \le i \le N$, contractive IFS maps on (X, d), with associated probabilities p_i, $1 \le i \le N$;

2. a condensation measure $\theta \in \mathcal{M}(X)$ with support $L \in \mathcal{B}(X)$ such that $\theta(L) = 1$ and $\theta(B) = \theta(B \cap L)$ for $B \in \mathcal{B}(X)$.

The Markov operator $M : \mathcal{M}(X) \to \mathcal{M}(X)$ is given by

$$(M\mu)(S) = \sum_{i=1}^{N} p_i \mu(w_i^{-1}(S)) \ + \ p_0\theta(S), \qquad S \in \mathcal{H}(X). \tag{3.25}$$

Thus

$$\int_X f(x)d(M\mu)(x) = \sum_{i=1}^{N} p_i \int_X (f \circ w_i)(x)d\mu(x) \ + \ p_0 \int_X f(x)d\theta(x). \tag{3.26}$$

Following [2], define the IFSP $\{w_1, \ldots, w_N, \hat{p}_1, \ldots, \hat{p}_N\}$, where $\hat{p}_i = p_i/(1-p_0)$, with associated Markov operator $\hat{M} : \mathcal{M}(X) \to \mathcal{M}(X)$. Define the affine map $\bar{M}_c : \mathcal{M}(X) \to \mathcal{M}(X)$, where

$$M_c\mu = (1-p_0)\hat{M}\mu + p_0\theta, \qquad \mu \in \mathcal{M}(X). \tag{3.27}$$

M_c is contractive in the d_H metric. Hence, there exists a unique $\bar{\mu}_c \in \mathcal{M}(X)$ such that $M_c\bar{\mu}_c = \bar{\mu}_c$.

REFERENCES

[1] M. F. Barnsley, *Fractals Everywhere*, Academic Press, New York, 1988.

[2] M. F. Barnsley and S. Demko, *Iterated function systems and the global construction of fractals*, Proc. Roy. Soc. London Ser. A, 399 (1985), pp. 243–275.

[3] M.F. Barnsley and L.P. Hurd, *Fractal Image Compression*, A.K. Peters, Wellesley, MA, 1993.

[4] M. Berger, *Random affine iterated function systems: mixing and encoding*, in Diffusion Processes and Related Problems in Analysis, Vol. II, Prog. Prob., 27, Birkhauser, Boston, MA, 1992, pp. 315–346.

[5] J. Elton, *An ergodic theorem for iterated maps*, J. Ergodic Theory Dynam. Systems, 7 (1987), pp. 481–488.

[6] B. Forte and E.R. Vrscay, *Solving the inverse problem for functions and image approximation using iterated function systems*, Dynam. Contin. Discrete Impuls. Systems, 1 (1995), pp. 177–231.

[7] B. Forte and E.R. Vrscay, *Theory of generalized fractal transforms*, to appear in Fractal Image Encoding and Analysis, Y. Fisher, ed., Springer-Verlag, New York, 1997.

[8] J. Hutchinson, *Fractals and self-similarity*, Indiana Univ. J. Math., 30 (1981), pp. 713–747.

[9] K. Stromberg, *Probability for Analysts*, Chapman and Hall, New York, 1994.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 891–902, July 1998

005

REARRANGEMENTS OF VECTOR VALUED FUNCTIONS, WITH APPLICATION TO ATMOSPHERIC AND OCEANIC FLOWS*

R. J. DOUGLAS†

Abstract. This paper establishes the equivalence of four definitions of two vector valued functions being rearrangements. Properties of the monotone rearrangement of a vector valued function are used to show existence and uniqueness of the minimizer of an energy functional arising from the semigeostrophic equations, a model for atmospheric and oceanic flow. At each fixed time solutions are shown to be equal to the gradient of a convex function, verifying the conjecture of Cullen, Norbury, and Purser [*SIAM J. Appl. Math.*, 51 (1991), pp. 20–31].

Key words. rearrangement of functions, semigeostrophic, variational problems, generalized solution

AMS subject classifications. 35D05, 46E30, 76C15

PII. S003614109628216X

1. Introduction.

This paper studies properties of rearrangements of vector valued functions and gives an application to atmospheric and oceanic flow. We say that two vector valued functions $f, g \in L^p(\Omega \subset \mathbf{R}^n, \lambda, \mathbf{R}^d)$, where $1 \leq p < \infty$ and Ω is bounded, are *rearrangements* if

$$\lambda\left(f^{-1}(B)\right) = \lambda\left(g^{-1}(B)\right) \tag{1.1}$$

for each Borel subset B of $\mathbf{R}^d$. (We restrict our definition of rearrangement to functions defined on measure spaces (Ω, λ) with certain properties; see section 2.1.) This is equivalent to the definition of rearrangement for scalar valued functions when $d = 1$. Rearrangement can be viewed as an equivalence relation on the space of L^p functions; therefore we can define the set of rearrangements (or equivalence class) of a given vector valued function. For a prescribed f_0 we write $R(f_0)$ to denote the set of rearrangements of f_0.

Different definitions have been given for two vector valued functions being rearrangements. Brenier [2] defined vector valued functions f and g (belonging to $L^p(\Omega \subset \mathbf{R}^n, \lambda, \mathbf{R}^d)$) to be rearrangements if $\int_\Omega F(f) = \int_\Omega F(g)$ for each F in a subclass of continuous functions from $\mathbf{R}^d$ to $\mathbf{R}$. In contrast, Cullen, Norbury, and Purser [6] made a direct extension of the definition of scalar valued rearrangement, requiring that $\lambda\{x : f(x) \geq c\} = \lambda\{x : g(x) \geq c\}$ for each $c \in \mathbf{R}^d$, where the inequalities are calculated component by component. Section 2 unifies these concepts, establishing that both are equivalent to the definition in the opening paragraph. We establish a fourth equivalent property which yields a characterization for the set of rearrangements of a prescribed vector valued function. This is a vector valued extension of the real valued characterization of Eydeland, Spruck, and Turkington [7]. These results are stated in Theorem 2.2.

*Received by the editors August 16, 1996; accepted for publication (in revised form) May 29, 1997; published electronically March 25, 1998. This research was carried out under the guidance of M. J. P. Cullen, and the meteorological significance of these results is discussed in Cullen and Douglas [*U.K. Meteorological Office Forecasting Research Scientific Paper* 47, 1997].

http://www.siam.org/journals/sima/29-4/28216.html

†Department of Mathematics, University of Reading, Whiteknights, P.O. Box 220, Reading RG6 6AX, UK (r.j.douglas@reading.ac.uk).

Section 3 studies a variational problem arising from a model for atmospheric and oceanic flow. The equations are the three-dimensional Boussinesq equations of semi-geostrophic flow, a standard model for slowly varying flows constrained by rotation and stratification. (They are recalled in section 3.2.) Cullen, Norbury, and Purser [6] interpreted solutions as a sequence of minimum energy states: at each time t, the particles arrange themselves so that geostrophic energy is minimized. The state of the fluid is known on particles; therefore we minimize geostrophic energy over the set of rearrangements of a possible state of the fluid at time t, a vector valued function. Cullen, Norbury, and Purser conjectured the existence of a unique minimizer, equal to the gradient of a convex function, which is the actual state of the fluid. We make the physically reasonable assumption that the fluid configuration belongs to L^p so that we may use the theory of monotone rearrangement of vector valued functions, which was developed by Brenier [2]. If the fluid configuration satisfies a nondegeneracy condition (see section 3.3), the Cullen–Norbury–Purser conjecture follows easily by the results of Brenier [2]. However, the nondegeneracy condition is severe, as it does not allow the function to have level sets of positive measure. Our main result is a proof of the Cullen–Norbury–Purser conjecture in Theorem 3.1; we make no restriction on the fluid configuration. We approximate functions that fail the nondegeneracy condition by a sequence of functions which satisfy it, and take appropriate limits. Uniqueness of the energy minimizer is recovered by properties of the monotone rearrangement.

The theory of rearrangements of vector valued functions is a new research area; recent advances have been made by Brenier [2]. In comparison, the theory of rearrangements of scalar valued functions is well developed; for example see Burton [3] and Alvino, Lions, and Trombetti [1]. Some results for scalar valued rearrangements do not have vector valued equivalents. For example, the monotone rearrangement of a vector valued function does not satisfy some of the inequalities which hold for the increasing rearrangement of a real valued function. (See Brenier [2] for details.) Moreover, scalar valued rearrangement of the components of a vector valued function does not imply vector valued rearrangement, although the converse is true. (See the appendix for details.)

2. Equivalent definitions of rearrangement of vector valued functions.

2.1. Introduction. In this section we establish four equivalent definitions of rearrangement for vector valued functions. We define rearrangement for vector valued functions on finite measure spaces (U, μ) which are isomorphic to $(0, \mu(U))$ endowed with Lebesgue measure λ. By isomorphic we mean there exists a measure preserving transformation $T : U \to (0, \mu(U))$. We recall the definition of measure preserving transformation in the next section. The restriction to finite measure spaces (U, μ) isomorphic to $(0, \mu(U))$ with Lebesgue measure is not severe; Royden [14] yields that any separable complete metric space U, equipped with a Borel measure μ such that $\mu(U) < \infty$ and $\mu(\{x\}) = 0$ for each $x \in U$, is isomorphic to $((0, \mu(U)), \lambda)$.

DEFINITION 2.1. *Let (U, μ) be a measure space which is isomorphic to $((0, \mu(U)), \lambda)$. Let $f, g \in L^p(U, \mu, \mathbf{R}^d)$ for $1 \le p < \infty$. Then f is a* rearrangement *of g if*

$$\mu\left(f^{-1}(B)\right) = \mu\left(g^{-1}(B)\right) \tag{2.1}$$

for every Borel subset B of $\mathbf{R}^d$.

We prove the following theorem.

THEOREM 2.2. *Let (U, μ) be as above. Let $f, g \in L^p(U, \mu, \mathbf{R}^d)$ for $1 \le p < \infty$. Then the following are equivalent.*

(i) *f is a rearrangement of g.*

(ii) *For each $F \in C(\mathbf{R}^d)$ such that $|F(\xi)| \leq K(1 + |\xi|_2^p)$ (where $|.|_2$ denotes Euclidean distance on $\mathbf{R}^d$, and K is a constant), the following equation is satisfied:*

$$\int_U F(f(x))d\mu(x) = \int_U F(g(x))d\mu(x). \tag{2.2}$$

(iii) *$\mu(f^{-1}(C)) = \mu(g^{-1}(C))$ for each set $C \in \{\prod_{i=1}^d [\alpha_i, \infty) : \alpha_i \in \mathbf{R}$ for each $i = 1, ..., d\} \bigcup \{\emptyset, \mathbf{R}^d\}$.*

(iv) *For each $\sigma \in \mathbf{R}^d$, $\alpha > 0$,*

$$\int_U (|g - \sigma|_\infty - \alpha)_+ d\mu = \int_U (|f - \sigma|_\infty - \alpha)_+ d\mu, \tag{2.3}$$

where $|.|_\infty$ denotes the infinity norm on $\mathbf{R}^d$, and the $+$ subscript denotes the positive part of the function.

Brenier [2] used property (ii) to define rearrangement of vector valued functions, while Cullen, Norbury, and Purser [6] used property (iii). This theorem shows that their definitions are equivalent. Property (iv) is a vector valued extension of the characterization of the set of rearrangements of a given real valued function by Eydeland, Spruck, and Turkington [7]. It may be shown that $R(f_0)$ is closed, and it follows from (iv) that for $w \in R(f_0)$, $\|w\|_p = \|f_0\|_p$, where

$$\|w\|_p = \left\{ \int_U |w|_\infty^p d\mu \right\}^{\frac{1}{p}}. \tag{2.4}$$

We omit the proofs, which are elementary.

2.2. Measure preserving mappings and transformations. We recall the concept of a measure preserving mapping.

DEFINITION 2.3. *A* measure preserving mapping *from a finite measure space (U, μ) to a measure space (V, ν) with $\mu(U) = \nu(V)$ is a mapping $s : U \to V$ such that for each ν-measurable set $A \subset V$, $\mu(s^{-1}(A)) = \nu(A)$.*

If a measure preserving mapping maps its domain to a measurable set, then it is surjective (up to a set of measure zero) but not necessarily injective. If a measure preserving mapping s is injective (up to a set of measure zero), and s maps μ-measurable sets to ν-measurable sets, then s^{-1} exists (almost everywhere) and is a measure preserving mapping. Such an s is called a *measure preserving transformation.*

2.3. A preliminary result.

LEMMA 2.4. *Let f, g be as in Theorem* 2.2. *Define*

$$\mathcal{M} = \{A \subset \mathbf{R}^d : \mu(f^{-1}(A)) = \mu(g^{-1}(A))\}, \tag{2.5}$$

$$H = \left\{ \prod_{i=1}^d [a_i, b_i] : a_i, b_i \in \mathbf{R}, a_i \leq b_i \text{ for } i = 1, ..., d \right\} \bigcup \{\emptyset, \mathbf{R}^d\}. \tag{2.6}$$

Suppose $H \subset \mathcal{M}$. Then $\mathcal{M}$ contains the Borel sets of $\mathbf{R}^d$.

Proof. Let $B_{cd}(H)$ denote the smallest family of sets which contains H that is closed under countable disjoint union and complementation (relative to $\mathbf{R}^d$). H is closed under finite intersection; therefore Kechris [11, page 65, Theorem 10.1(iii)]

yields that $B_{cd}(H)$ is a σ-algebra. H generates the Borel sets; therefore it follows that the Borel sets are contained in $B_{cd}(H)$. $\mathcal{M}$ is closed under countable disjoint union and complementation (relative to $\mathbf{R}^d$). Given that $H \subset \mathcal{M}$, we have $B_{cd}(H) \subset B_{cd}(\mathcal{M}) = \mathcal{M}$, so $\mathcal{M}$ contains the Borel sets. This completes the proof. □

2.4. Proof of Theorem 2.2. We begin by showing that (i) implies (ii). Let $F \in C(\mathbf{R}^d)$ satisfy $|F(\xi)| \leq K\{1 + |\xi|_2^p\}$ for each $\xi \in \mathbf{R}^d$, where K is some constant. We assume that F is nonnegative. (If not, we work with the positive and negative parts of F.) F is continuous; therefore we can choose a sequence of nonnegative functions (φ_n), each φ_n a finite linear combination of indicator functions of Borel sets, with $\varphi_n(\xi) \leq \varphi_{n+1}(\xi)$ and $\varphi_n(\xi) \to F(\xi)$ for each $\xi \in \mathbf{R}^d$. It follows from (i) that

$$\int_U \varphi_n(f(x))d\mu(x) = \int_U \varphi_n(g(x))d\mu(x) \tag{2.7}$$

for each $n \in \mathbf{N}$. Applying the dominated convergence theorem we obtain

$$\int_U F(f(x))d\mu(x) = \lim_{n\to\infty} \int_U \varphi_n(f(x))d\mu(x) \tag{2.8}$$

$$= \lim_{n\to\infty} \int_U \varphi_n(g(x))d\mu(x) \tag{2.9}$$

$$= \int_U F(g(x))d\mu(x). \tag{2.10}$$

This verifies (ii).

We show that (ii) implies (i). Let families of sets H and $\mathcal{M}$ be as in Lemma 2.4. Let $H_1 \in H$. There exists a sequence $(\varphi_n) \subset C(\mathbf{R}^d)$ such that $|\varphi_n(y)| \leq 1 + |y|_2^p$ for each $y \in \mathbf{R}^d$ and $n \in \mathbf{N}$, with $\varphi_n(y) \to 1_{H_1}(y)$ for each $y \in \mathbf{R}^d$. Noting that (ii) holds, we apply the dominated convergence theorem to obtain

$$\mu(f^{-1}(H_1)) = \int_U 1_{H_1} \circ f(x)d\mu(x) \tag{2.11}$$

$$= \lim_{n\to\infty} \int_U \varphi_n \circ f(x)d\mu(x) \tag{2.12}$$

$$= \lim_{n\to\infty} \int_U \varphi_n \circ g(x)d\mu(x) \tag{2.13}$$

$$= \int_U 1_{H_1} \circ g(x)d\mu(x) = \mu(g^{-1}(H_1)). \tag{2.14}$$

Thus $H_1 \in \mathcal{M}$. It follows that $H \subset \mathcal{M}$. Lemma 2.4 yields that $\mathcal{M}$ contains the Borel sets of $\mathbf{R}^d$; therefore f and g are rearrangements.

(i) implies (iii) is immediate. To see the converse, we show that $H \subset \mathcal{M}$, given that (iii) holds. We proceed by induction. Let $\mathcal{P}(k)$ be the proposition that all sets of the form $\prod_{i=1}^k [a_i, b_i] \times \prod_{i=k+1}^d [a_i, \infty) \in \mathcal{M}$, where $a_i, b_i \in \mathbf{R}$. We demonstrate $\mathcal{P}(1)$. Now

$$[a_1, b_1] \times \prod_{i=2}^d [a_i, \infty) = \prod_{i=1}^d [a_i, \infty) \Bigg\backslash \left(\bigcup_{n=1}^\infty \left([b_1 + 1/n, \infty) \times \prod_{i=2}^d [a_i, \infty) \right) \right), \tag{2.15}$$

and, noting that $\mathcal{M}$ is closed under countable increasing union and differences of two ordered elements (with respect to the partial order $\subset$), we obtain that $[a_1, b_1] \times$

$\prod_{i=2}^{d}[a_i,\infty) \in \mathcal{M}$. This shows $\mathcal{P}(1)$. We demonstrate that $\mathcal{P}(k+1)$ is true given that $\mathcal{P}(k)$ holds. We have that

(2.16)

$$\prod_{i=1}^{k+1}[a_i,b_i] \times \prod_{i=k+2}^{d}[a_i,\infty) = \prod_{i=1}^{k}[a_i,b_i] \times \prod_{i=k+1}^{d}[a_i,\infty) \setminus \left(\bigcup_{n=1}^{\infty} \left(\prod_{i=1}^{k}[a_i,b_i] \times [b_{k+1}+1/n,\infty) \times \prod_{i=k+2}^{d}[a_i,\infty) \right) \right).$$

We are given that $\mathcal{P}(k)$ holds; therefore $\prod_{i=1}^{k}[a_i,b_i] \times \prod_{i=k+1}^{d}[a_i,\infty) \in \mathcal{M}$, and $\prod_{i=1}^{k}[a_i,b_i] \times [b_{k+1}+1/n,\infty) \times \prod_{i=k+2}^{d}[a_i,\infty) \in \mathcal{M}$ for each $n \in \mathbf{N}$. It follows that $\prod_{i=1}^{k+1}[a_i,b_i] \times \prod_{i=k+2}^{d}[a_i,\infty) \in \mathcal{M}$. This verifies $\mathcal{P}(k+1)$. By induction $\mathcal{P}(d)$ holds, that is, all sets of the form $\prod_{i=1}^{d}[a_i,b_i] \in \mathcal{M}$ for $a_i, b_i \in \mathbf{R}$, $i = 1, ..., d$. It is immediate that $\emptyset, \mathbf{R}^d \in \mathcal{M}$; therefore $H \subset \mathcal{M}$. Lemma 2.4 yields that $\mathcal{M}$ contains the Borel sets of $\mathbf{R}^d$. This shows (i).

Let (iv) hold. The characterization of the set of rearrangements of a scalar valued function by Eydeland, Spruck, and Turkington [7] yields that $|g-\sigma|_\infty \in R(|f-\sigma|_\infty)$ in the scalar valued sense for each $\sigma \in \mathbf{R}^d$. It follows that $\mu(g^{-1}(C_\alpha(\sigma))) = \mu(f^{-1}(C_\alpha(\sigma)))$ for each $\alpha > 0$, where $C_\alpha(\sigma)$ denotes the open cube of side 2α about $\sigma \in \mathbf{R}^d$. Let K denote the set of all d-dimensional open cubes. We have shown that $K \subset \mathcal{M}$. We now demonstrate that this implies that all open subsets of $\mathbf{R}^d$ belong to $\mathcal{M}$. Recall that $\mathcal{M}$ is closed under countable decreasing intersections, increasing countable unions, and differences of ordered elements of $\mathcal{M}$. For $j = 0, \ldots, d$, every j-dimensional closed cube is a countable decreasing intersection of j-dimensional open cubes. Further, for $j = 1, \ldots, d$ every j-dimensional open cube with one $(j-1)$-dimensional open face attached is an increasing countable union of j-dimensional closed cubes. Now, for $j = 1, ..., d$, every $(j-1)$-dimensional open cube is the difference of a set of the type described in the preceding sentence and a j-dimensional open cube contained in it. It follows by induction that open and closed cubes of dimensions $0, \ldots, d$ belong to $\mathcal{M}$. Every open subset of $\mathbf{R}^d$ is a countable disjoint union of open cubes of dimensions $0, \ldots, d$; therefore such sets belong to $\mathcal{M}$. The methods of Lemma 2.4 (noting that the intersection of two open sets is open) yield that $\mathcal{M}$ contains the Borel sets. Thus (iv) implies (i). The converse follows because (i) implies that $\mu(g^{-1}(C_\alpha(\sigma))) = \mu(f^{-1}(C_\alpha(\sigma)))$ for each positive $\alpha \in \mathbf{R}$, $\sigma \in \mathbf{R}^d$. This completes the proof. □

3. Energy minimizing solutions of atmospheric and oceanic flow.

3.1. Introduction. This section studies a variational problem over the set of rearrangements of a prescribed vector valued function, which arises from an energy minimizing principle. We study the semigeostrophic equations (recalled in the next section), a standard model for slowly varying flows constrained by rotation and stratification, using the methods of Cullen, Norbury, and Purser [6]. At any given time, $\mathbf{X}$, which describes the state of the fluid, is known on particles. The *Cullen–Norbury–Purser* principle states that for a solution, the particles are arranged to minimize geostrophic energy. This yields a variational problem: minimize energy over the set of rearrangements of a prescribed fluid configuration. We verify the conjecture of

Cullen, Norbury, and Purser [6, section 5] that the energy minimum is uniquely attained and that the minimizer is equal to the gradient of a convex function. We prove the following theorem.

THEOREM 3.1. *Let Ω be a bounded, connected, closed subset of $\mathbf{R}^3$, with smooth boundary. Define, for $\mathbf{X} = (X, Y, Z) \in L^p(\Omega, \mu, \mathbf{R}^3)$, where $2 \le p < \infty$ and μ denotes three-dimensional Lebesgue measure,*

$$E(\mathbf{X}) = \frac{1}{2}\int_\Omega X^2 + x^2 + Y^2 + y^2 d\mu(\mathbf{x}) - \int_\Omega \mathbf{X}.\mathbf{x} d\mu(\mathbf{x}), \tag{3.1}$$

where $\mathbf{x} = (x, y, z) \in \Omega$. Suppose $\mathbf{X_0} \in L^p(\Omega, \mu, \mathbf{R}^3)$ for p as above. Then there exists $\mathbf{X_0}^* \in R(\mathbf{X_0})$ such that

(i) *$E(\mathbf{X_0}^*) < E(\mathbf{X})$ for each $\mathbf{X} \in R(\mathbf{X_0})\backslash\{\mathbf{X_0}^*\}$.*
(ii) *$\mathbf{X_0}^* = \nabla\Psi$ for some convex function $\Psi \in W^{1,p}(\Omega)$.*
(iii) *$\mathbf{X_0}^*$ is a cyclically monotone function.*

The functional E represents the geostrophic energy of the fluid. We define E and $\mathbf{X}$ in the next section. The unique energy minimizer is the monotone rearrangement of the prescribed function: this concept was introduced by Brenier [2] and is recalled in section 3.3. The proof uses an approximation argument, with the strict inequality following by the uniqueness of the monotone rearrangement.

3.2. The semigeostrophic equations, and the Cullen–Norbury–Purser principle. We state the three-dimensional Boussinesq equations of semigeostrophic theory on an f plane. These are a standard model for slowly varying flows constrained by rotation and stratification and are used to study front formation in meteorology. We state the equations in the form used by Hoskins [10].

$$\frac{Du_g}{Dt} - fv_{ag} = 0, \qquad \frac{Dv_g}{Dt} + fu_{ag} = 0, \tag{3.2}$$

$$\frac{D\theta}{Dt} = 0, \tag{3.3}$$

$$\nabla.\mathbf{u} = 0, \tag{3.4}$$

$$\nabla\phi = \left(fv_g, -fu_g, \frac{g\theta}{\theta_0}\right), \tag{3.5}$$

where

$$\begin{aligned} \mathbf{u} &\equiv (u, v, w) \equiv \mathbf{u_g} + \mathbf{u_{ag}}, \\ \mathbf{u_g} &\equiv (u_g, v_g, 0), \\ \frac{D}{Dt} &\equiv \frac{\partial}{\partial t} + \mathbf{u}.\nabla, \end{aligned} \tag{3.6}$$

f is the Coriolis parameter, assumed constant, g denotes the acceleration due to gravity, θ_0 is a reference value of the potential temperature θ, and ϕ is a pressure variable. Subscripts g and ag denote geostrophic and ageostrophic velocity (or wind) components, respectively, where the geostrophic velocity is defined to be the horizontal component of velocity in balance with the pressure gradient. This definition

is included in equation (3.5), as is the statement of hydrostatic balance. We solve the equations (for the velocity $\mathbf{u}$) in a closed, bounded, connected set $\Omega \subset \mathbf{R}^3$, with normal velocity $\mathbf{u}.\mathbf{n}$ given on $\partial\Omega$. For $\mathbf{x} = (x, y, z) \in \Omega$, by making the substitution

$$\mathbf{X} \equiv (X, Y, Z) \equiv (x + v_g/f, y - u_g/f, (g/f^2\theta_0)\theta), \tag{3.7}$$

it is shown in Purser and Cullen [13] that we may replace (3.2) and (3.3) by

$$\frac{D\mathbf{X}}{Dt} = \mathbf{u_g}. \tag{3.8}$$

We think of $\mathbf{X}$ as a function of the physical space coordinates $\mathbf{x}$. Rewriting in terms of $\mathbf{X}$ and $\mathbf{x}$, we have

$$\frac{DX}{Dt} = f(y - Y), \tag{3.9}$$

$$\frac{DY}{Dt} = f(X - x), \tag{3.10}$$

$$\frac{DZ}{Dt} = 0. \tag{3.11}$$

The geostrophic energy E is defined as

$$E = \int_\Omega \frac{1}{2}u_g^2 + \frac{1}{2}v_g^2 - \frac{g\theta z}{\theta_0} d\mu(\mathbf{x}) \tag{3.12}$$

$$= f^2\frac{1}{2}\int_\Omega X^2 + x^2 + Y^2 + y^2 d\mu(\mathbf{x}) - f^2\int_\Omega \mathbf{x}.\mathbf{X} d\mu(\mathbf{x}). \tag{3.13}$$

Henceforth we ignore the constant f^2. At any fixed time t, $\mathbf{X}$ is found on particles by predicting (X, Y, Z) on particles using equations (3.9), (3.10), and (3.11). The *Cullen–Norbury–Purser* principle states that for a solution, the particles are arranged to minimize geostrophic energy. Suppose one possible state of the fluid is described by values $\mathbf{X_0} = (X_0, Y_0, Z_0)$ which are known on particles. The Cullen–Norbury–Purser principle yields the energy minimization problem

$$\inf_{\mathbf{X}\in R(\mathbf{X_0})} E(\mathbf{X}), \tag{3.14}$$

where the energy minimizer (if it exists and is unique) gives the actual state of the fluid. In this way, solutions can be viewed as a sequence of minimum energy states.

We make some (physically reasonable) assumptions to enable us to use vector valued rearrangement theory. Let Ω be a closed, bounded, connected subset of $\mathbf{R}^3$, with smooth boundary. Suppose the possible fluid configuration $\mathbf{X_0} \in L^p(\Omega, \mu, \mathbf{R}^3)$, for $2 \le p < \infty$, where μ denotes three-dimensional Lebesgue measure. (Choosing $p \ge 2$ ensures finite geostrophic energy.)

3.3. Monotone rearrangement of vector valued functions. We recall the concept of the monotone rearrangement of a vector valued function: essentially, this is the vector valued analogue of the increasing rearrangement of a real valued function. Let Ω and μ be as in the last paragraph of the previous section. The following theorem is due to Brenier [2, section 1.2, Theorem 1.1].

THEOREM 3.2. *For each $u \in L^p(\Omega, \mu, \mathbf{R}^3)$, where $1 \le p < \infty$, there is a unique $u^* \in R(u)$ such that*

$$u^* \in \{\nabla\Psi : \Psi \in W^{1,p}(\Omega, \mu), \Psi \text{ convex}\}, \tag{3.15}$$

and the mapping $u \to u^*$ is continuous.

When Ω is not convex, Ψ is understood to be the restriction to Ω of a convex function defined on $\mathbf{R}^3$. We call u^* the *monotone rearrangement of* u. The name comes from the fact that u^* is a cyclically monotone function. We note that McCann [12] has generalized the first part of this result (concerning the existence of an essentially unique rearrangement equal to the gradient of a convex function) to more general measures than Lebesgue measure, and has removed the restriction that Ω must be closed, connected, and have smooth boundary.

DEFINITION 3.3. *A function $u \in L^p(\Omega, \mu, \mathbf{R}^3)$ is* nondegenerate *if $\mu(u^{-1}(E)) = 0$ for each set $E \subset \mathbf{R}^3$ with Lebesgue measure zero. We say that a function which fails to be nondegenerate is* degenerate.

Brenier established further properties of the monotone rearrangement of a nondegenerate function in the following theorem [2, section 1.2, Theorem 1.2]

THEOREM 3.4. *For each nondegenerate $u \in L^p(\Omega, \mu, \mathbf{R}^3)$, there exists a unique pair (u^*, s), where u^* is the monotone rearrangement of u and s is a measure preserving mapping from (Ω, μ) to (Ω, μ) such that*

(i) *$u = u^* \circ s$.*

(ii) *s is the unique measure preserving mapping that maximizes $\int_\Omega u(\mathbf{x}).s(\mathbf{x})d\mu(\mathbf{x})$.*

An elementary proof of the above result was found by Gangbo [8]. Note that Theorem 3.4 is not true if u has a level set of positive measure: the measure preserving mapping is not unique, nor do we have uniqueness in property (ii). Such a u is degenerate. The author is not aware of any corresponding result for degenerate functions.

3.4. Existence and uniqueness of an energy minimizer. Recall that we are studying the energy minimization problem

$$\inf_{\mathbf{X} \in R(\mathbf{X_0})} \frac{1}{2} \int_\Omega x^2 + X^2 + y^2 + Y^2 d\mu(\mathbf{x}) - \int_\Omega \mathbf{x}.\mathbf{X} d\mu(\mathbf{x}), \tag{3.16}$$

where $\mathbf{X_0} \in L^p(\Omega, \mu, \mathbf{R}^3)$ for $2 \le p < \infty$, and $\mathbf{X} = (X, Y, Z)$. We show that the first integral is conserved under rearrangements.

LEMMA 3.5. *Let $\mathbf{X_0}$ be as in Theorem* 3.1. *Let $\mathbf{X_1} \in R(\mathbf{X_0})$. Then*

$$\int_\Omega x^2 + X_1^2 + y^2 + Y_1^2 d\mu(\mathbf{x}) = \int_\Omega x^2 + X_0^2 + y^2 + Y_0^2 d\mu(\mathbf{x}), \tag{3.17}$$

where $\mathbf{X_0} = (X_0, Y_0, Z_0)$ and $\mathbf{X_1} = (X_1, Y_1, Z_1)$.

Proof. $\mathbf{X_1} \in R(\mathbf{X_0})$ implies that $X_1 \in R(X_0)$. It follows that

$$\int_\Omega X_1^2 d\mu(\mathbf{x}) = \int_\Omega X_0^2 d\mu(\mathbf{x}). \tag{3.18}$$

A similar result holds for Y_0 and Y_1. The result follows. □

To show that there is a unique energy minimizer, it remains to show that

$$\sup_{\mathbf{X} \in R(\mathbf{X_0})} \int_\Omega \mathbf{x}.\mathbf{X} d\mu(\mathbf{x}) \tag{3.19}$$

is uniquely attained. If $\mathbf{X_0}$ is nondegenerate, the result follows easily using Theorem 3.4. Our method of proof is to approximate degenerate functions with a sequence of nondegenerate functions. This shows that the monotone rearrangement is an energy

minimizer. We demonstrate that an energy minimizer is the gradient of a convex function: the monotone rearrangement is the unique such amongst the set of rearrangements; therefore the result follows.

LEMMA 3.6. *Let* $\mathbf{X} \in L^p(\Omega, \mu, \mathbf{R}^3)$ (*where* Ω, μ, *and* p *are as in section* 3.2). *Then there exists a sequence of nondegenerate functions* $(\mathbf{X_n})$ *such that* $\mathbf{X_n} \to \mathbf{X}$ *in* $L^p(\Omega, \mu, \mathbf{R}^3)$.

Proof. For each $n \in \mathbf{N}$, choose a simple function φ_n such that $\|\mathbf{X} - \varphi_n\|_p \leq 1/n$. Now for each $n \in \mathbf{N}$, define $\mathbf{X_n}$ by $\mathbf{X_n}(\mathbf{x}) = \varphi_n(\mathbf{x}) + (1/n)\mathbf{x}$ for $\mathbf{x} \in \Omega$. It is immediate that $\mathbf{X_n} \to \mathbf{X}$ in $L^p(\Omega, \mu, \mathbf{R}^3)$. It remains to show that $\mathbf{X_n}$ is nondegenerate for each $n \in \mathbf{N}$. Fix $n \in \mathbf{N}$. φ_n is a simple function; therefore it takes finitely many values which we enumerate $\{\mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_m}\}$. Define $A_i = {\varphi_n}^{-1}(\mathbf{b_i})$ for each $i = 1, ..., m$. Write ${\mathbf{X_n}}^i$ for $\mathbf{X_n}|_{A_i}$. For a given i, ${\mathbf{X_n}}^i = \mathbf{b_i} + (1/n)\mathbf{x}$. Let E be a Lebesgue negligible subset of $\mathbf{R}^3$. Then

$$\mu\left(({\mathbf{X_n}}^i)^{-1}(E)\right) = \mu\left(A_i \bigcap (nE - n\mathbf{b_i})\right) \tag{3.20}$$

$$\leq \mu(nE - n\mathbf{b_i}) \tag{3.21}$$

$$= \mu(nE) = 0. \tag{3.22}$$

By way of explanation, we used translation invariance of Lebesgue measure to obtain the first equality in (3.22), and used properties of Lebesgue measure to obtain the second. This demonstrates that ${\mathbf{X_n}}^i$ is nondegenerate (as an element in $L^p(A_i, \mu, \mathbf{R}^3)$) for each $i = 1, ..., m$.

Let E be a Lebesgue negligible subset of $\mathbf{R}^3$. Then

$$\mu\left({\mathbf{X_n}}^{-1}(E)\right) = \mu\left(\bigcup_{i=1}^{m} ({\mathbf{X_n}}^i)^{-1}(E)\right) \tag{3.23}$$

$$= \sum_{i=1}^{m} \mu\left(({\mathbf{X_n}}^i)^{-1}(E)\right) = 0. \tag{3.24}$$

To obtain (3.24) we used the countable additivity of μ and the fact that ${\mathbf{X_n}}^i$ is nondegenerate for each $i = 1, ..., m$. This shows that $\mathbf{X_n}$ is nondegenerate and completes the proof. □

LEMMA 3.7. *Let* $\mathbf{X_0}$ *be as in Theorem* 3.1. *Then*

$$\int_\Omega {\mathbf{X_0}}^*(\mathbf{x}).\mathbf{x} d\mu(\mathbf{x}) \geq \int_\Omega \mathbf{X}(\mathbf{x}).s(\mathbf{x}) d\mu(\mathbf{x}) \tag{3.25}$$

for each $\mathbf{X} \in R(\mathbf{X_0})$ *and each* $s : \Omega \to \Omega$ *a measure preserving mapping.*

Proof. Let $\mathbf{X} \in R(\mathbf{X_0})$ and let $s : \Omega \to \Omega$ be a measure preserving mapping. From the previous lemma we may choose a sequence $(\mathbf{X_n})$ of nondegenerate functions such that $\mathbf{X_n} \to \mathbf{X}$ in $L^p(\Omega, \mu, \mathbf{R}^3)$. For each $n \in \mathbf{N}$, Theorem 3.4(i) yields the existence of a unique measure preserving mapping $s_n : \Omega \to \Omega$ such that $\mathbf{X_n} = {\mathbf{X_n}}^* \circ s_n$. Applying Theorem 3.2 we have ${\mathbf{X_n}}^* \to \mathbf{X}^* = {\mathbf{X_0}}^*$. Now

$$\int_\Omega {\mathbf{X_0}}^*(\mathbf{x}).\mathbf{x} d\mu(\mathbf{x}) = \lim_{n\to\infty} \int_\Omega {\mathbf{X_n}}^*(\mathbf{x}).\mathbf{x} d\mu(\mathbf{x}) \tag{3.26}$$

$$= \lim_{n\to\infty} \int_\Omega {\mathbf{X_n}}^* \circ s_n(\mathbf{x}).s_n(\mathbf{x}) d\mu(\mathbf{x}) \tag{3.27}$$

$$= \lim_{n\to\infty} \int_\Omega \mathbf{X_n}(\mathbf{x}).s_n(\mathbf{x}) d\mu(\mathbf{x}) \tag{3.28}$$

$$\geq \lim_{n\to\infty} \int_\Omega \mathbf{X_n}(\mathbf{x}).s(\mathbf{x})d\mu(\mathbf{x}) \tag{3.29}$$

$$= \int_\Omega \mathbf{X}(\mathbf{x}).s(\mathbf{x})d\mu(\mathbf{x}) \tag{3.30}$$

as required. By way of explanation, (3.27) holds because s_n is a measure preserving map, and (3.29) follows because Theorem 3.4(ii) yields that

$$\int_\Omega \mathbf{X_n}(\mathbf{x}).s_n(\mathbf{x})d\mu(\mathbf{x}) \geq \int_\Omega \mathbf{X_n}(\mathbf{x}).s(\mathbf{x})d\mu(\mathbf{x}) \tag{3.31}$$

for each measure preserving mapping $s : \Omega \to \Omega$ and for each $n \in \mathbf{N}$. This completes the proof. □

LEMMA 3.8. *Let* $\mathbf{X_0}$ *be as in Theorem* 3.1. *Then*

$$\int_\Omega \mathbf{X_0}^*(\mathbf{x}).\mathbf{x}d\mu(\mathbf{x}) > \int_\Omega \mathbf{X}(\mathbf{x}).\mathbf{x}d\mu(\mathbf{x}) \tag{3.32}$$

for each $\mathbf{X} \in R(\mathbf{X_0})\backslash\{\mathbf{X_0}^*\}$.

Proof. Applying the previous lemma for the identity mapping, we have

$$\int_\Omega \mathbf{X_0}^*(\mathbf{x}).\mathbf{x}d\mu(\mathbf{x}) \geq \int_\Omega \mathbf{X}(\mathbf{x}).\mathbf{x}d\mu(\mathbf{x}) \tag{3.33}$$

for each $\mathbf{X} \in R(\mathbf{X_0})\backslash\{\mathbf{X_0}^*\}$. It remains to show strict inequality. Suppose there exists $\mathbf{X_1} \in R(\mathbf{X_0})$ such that $\int_\Omega \mathbf{X_1}.\mathbf{x}d\mu = \int_\Omega \mathbf{X_0}^*.\mathbf{x}d\mu$. Applying the previous lemma to $\mathbf{X_1} \in R(\mathbf{X_0})$ we obtain

$$\int_\Omega \mathbf{X_1}(\mathbf{x}).\mathbf{x}d\mu(\mathbf{x}) = \int_\Omega \mathbf{X_0}^*(\mathbf{x}).\mathbf{x}d\mu(\mathbf{x}) \tag{3.34}$$

$$\geq \int_\Omega \mathbf{X_1}(\mathbf{x}).s(\mathbf{x})d\mu(\mathbf{x}) \tag{3.35}$$

for each measure preserving mapping $s : \Omega \to \Omega$. Brenier [2, Proposition 2.1] yields that $\mathbf{X_1} \in \{\nabla\Psi : \Psi \in W^{1,2}(\Omega), \Psi \text{ convex}\}$. However Theorem 3.2 states that $\mathbf{X_0}^*$ is the unique member of $R(\mathbf{X_0})$ belonging to $\{\nabla\Psi : \Psi \in W^{1,2}(\Omega), \Psi \text{ convex}\}$; therefore $\mathbf{X_1} = \mathbf{X_0}^*$. This completes the proof. □

Proof of Theorem 3.1. The proof follows from Lemmas 3.5 and 3.8. □

Note added in revision. We can rewrite an equivalent minimization problem

$$\inf_{\mathbf{X}\in R(\mathbf{X_0})} \int_\Omega |\mathbf{X}(\mathbf{x}) - \mathbf{x}|^2 d\mu(\mathbf{x}), \tag{3.36}$$

where $|\mathbf{X}(\mathbf{x}) - \mathbf{x}|^2 = (\mathbf{X}(\mathbf{x}) - \mathbf{x}).(\mathbf{X}(\mathbf{x}) - \mathbf{x})$, as a Monge mass transport problem as follows. Define $\nu(S) = \mu(\mathbf{X_0}^{-1}(S))$ for (Borel) subsets of $\mathbf{R}^3$, and rewrite (3.36) as

$$\inf_{s\in S} \int_\Omega c(\mathbf{x}, s(\mathbf{x}))d\mu(\mathbf{x}), \tag{3.37}$$

where S is the set of measure preserving mappings between (Ω, μ) and $(\mathbf{R}^3, \nu)$, and the *cost function* $c : \mathbf{R}^3 \times \mathbf{R}^3 \to \mathbf{R}$ is defined by $c(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|^2$. (Note that $\mathbf{X_0}$ is nondegenerate if and only if ν is absolutely continuous with respect to Lebesgue measure.) Gangbo and McCann [9] showed that (3.37) has a unique minimizer by proving the existence of a unique minimizer (which has a particular form) of an appropriate Monge–Kantorovich problem. An extensive review of Monge–Kantorovich problems may be found in Cuesta-Albertos et al. [4]. This paper, developed independently, does not use this approach.

Appendix. We consider the relationship between vector valued functions which are rearrangements and vector valued functions for which corresponding components are rearrangements in the scalar valued sense. Let f, g be as in the opening paragraph of the introduction. Define, for $i = 1, ..., d$, $\Pi_i : \mathbf{R}^d \to \mathbf{R}$ to be the projection of the ith component of an element of $\mathbf{R}^d$. Write $f = (f_1, ..., f_d)$ and $g = (g_1, ..., g_d)$, where $f_i = \Pi_i \circ f$, $g_i = \Pi_i \circ g$ for $i = 1, ..., d$. The definition of vector valued rearrangement yields that if $f \in R(g)$, we have $f_i \in R(g_i)$ for each $i = 1, ..., d$ in the scalar valued sense. However, the converse is false in general. Let $f : [0,1]^2 \to \mathbf{R}^2$ be defined by

$$f(x) = \begin{cases} (1,1) & \text{if } x \in [1/2,1] \times [1/2,1], \\ (0,0) & \text{if } x \notin [1/2,1] \times [1/2,1]. \end{cases}$$

Then

$$f_1(x) = f_2(x) = \begin{cases} 1 & \text{if } x \in [1/2,1] \times [1/2,1], \\ 0 & \text{if } x \notin [1/2,1] \times [1/2,1]. \end{cases}$$

Define

$$g(x) = \begin{cases} (1,0) & \text{if } x \in [1/2,1] \times [1/2,1], \\ (0,1) & \text{if } x \in [0,1/2] \times [1/2,1], \\ (0,0) & \text{otherwise.} \end{cases}$$

Then $g_1 = f_1$ and

$$g_2(x) = \begin{cases} 1 & \text{if } x \in [0,1/2] \times [1/2,1], \\ 0 & \text{if } x \notin [0,1/2] \times [1/2,1]. \end{cases}$$

It is easily seen that $f_i \in R(g_i)$ for $i = 1, 2$, but $f \notin R(g)$. Consequently, in general we cannot apply scalar valued rearrangement theorems to components of vector valued functions and hope to obtain results pertaining to vector valued rearrangements.

Acknowledgments. The author would like to acknowledge the insight and help of G.R. Burton. The author would also like to thank J. Norbury, Y. Brenier, and J-D. Benamou for stimulating discussions. The author is grateful for the hospitality and support of the Isaac Newton Institute for Mathematical Sciences during the last stage of this work. The author is also grateful to T. Ratiu for drawing his attention to the work of Gangbo and McCann.

REFERENCES

[1] A. Alvino, P.-L. Lions, and G. Trombetti, *On optimization problems with prescribed rearrangements*, Nonlinear Anal. Theory Methods Appl., 13 (1989), pp. 185–220.

[2] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[3] G. R. Burton, *Rearrangement of functions, maximisation of convex functionals, and vortex rings*, Math. Ann., 276 (1987), pp. 225–253.

[4] J. A. Cuesta-Albertos, C. Matran, S. T. Rachev, and L. Ruschendorf, *Mass transportation problems in probability theory*, Math. Scientist, 21 (1996), pp. 34–72.

[5] M. J. P. Cullen and R. J. Douglas, *Understanding Atmospheric Dynamics using Rearrangements of Functions*, U.K. Meteorological Office Forecasting Research Scientific paper 47, 1997.

[6] M. J. P. Cullen, J. Norbury, and R. J. Purser, *Generalized lagrangian solutions for atmospheric and oceanic flows*, SIAM J. Appl. Math., 51 (1991), pp. 20–31.

[7] A. Eydeland, J. Spruck, and B. Turkington, *Multiconstrained variational problems of nonlinear eigenvalue type: New formulations and algorithms*, Math. Comp., 55 (1990), pp. 509–535.

[8] W. Gangbo, *An elementary proof of the polar factorization of vector-valued functions*, Arch. Rational Mech. Anal., 128 (1994), pp. 381–399.

[9] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.
[10] B. J. Hoskins, *The geostrophic momentum approximation and the semigeostrophic equations*, J. Atmospheric Sci., 32 (1975), pp. 233–242.
[11] A. S. Kechris, *Classical Descriptive Set Theory*, Springer-Verlag, New York, 1995.
[12] R. J. McCann, *Existence and uniqueness of monotone measure preserving maps*, Duke Math. J., 80 (1995), pp. 309–323.
[13] R. J. Purser and M. J. P. Cullen, *A duality principle in semigeostrophic theory*, J. Atmospheric Sci., 44 (1987), pp. 3449–3468.
[14] H. L. Royden, *Real Analysis*, 2nd ed., Collier–Macmillan Limited, London, 1963.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 903–912, July 1998

006

STABILIZATION OF VORTICES IN THE GINZBURG–LANDAU EQUATION WITH A VARIABLE DIFFUSION COEFFICIENT*

XU-YAN CHEN†, SHUICHI JIMBO‡, AND YOSHIHISA MORITA§

Abstract. We study equilibria of the Ginzburg–Landau equation with a variable diffusion coefficient on a bounded planar domain subject to the Neumann boundary condition. It has been previously shown that if the diffusion coefficient is constant and the ambient domain is convex, the system does not carry stable vortices in the sense that any stable equilibrium solution is a constant of modulus 1. In this article we shall prove that arbitrarily given a domain, an appropriate choice of inhomogeneous diffusion coefficient yields a stable equilibrium solution having vortices. We can even manage to make the configuration of stable vortices close to prescribed locations. Our method is to minimize the free energy functional in suitably constructed positive invariant regions for the time-dependent Ginzburg–Landau equation.

Key words. Ginzburg–Landau equation, variable diffusion, vortex, stable solution

AMS subject classifications. 35K57, 35Q99, 35J65

PII. S0036141096308752

1. Introduction.

The Ginzburg–Landau equation for a complex order parameter is a nonlinear partial differential equation arising in a macroscopic description of superconductivity (see [6]). One of the important phenomena explained particularly well by the model is the formation of topological defects or *vortices* (i.e., zeros of the complex order parameter); this has received extensive mathematical analysis in recent years (see [1, 2, 4, 14]).

In this article we are concerned with the next Ginzburg–Landau equation with a variable diffusion coefficient in a bounded domain Ω in $\mathbb{R}^2$ with the Neumann boundary condition

$$\begin{cases} \operatorname{div}(a(x)\nabla\Phi) + (1-|\Phi|^2)\Phi = 0 & \text{in} \quad \Omega, \\ \dfrac{\partial\Phi}{\partial\nu} = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where $a(x)$ is a positive smooth function, ν is the unit outward normal vector on $\partial\Omega$, and the unknown Φ is a complex valued function representing the order parameter in the superconductor model. A solution to this equation gives an equilibrium of the

* Received by the editors September 3, 1996; accepted for publication (in revised form) May 27, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/30875.html

† School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (xchen@math.gatech.edu). This research was performed while this author was with the Division of Mathematics and Informatics, Faculty of Integrated Arts and Sciences, Hiroshima University, Kagamiyama 1-7-1, Higashi-Hiroshima 739, Japan.

‡ Department of Mathematics, Faculty of Science, Hokkaido University, Sapporo 060, Japan (jimbo@euler.math.hokudai.ac.jp).

§ Department of Applied Mathematics and Informatics, Faculty of Science and Technology, Ryukoku University, Seta Ohtsu 520-21, Japan (morita@rins.ryukoku.ac.jp). This work was initiated when this author was visiting the School of Mathematics, Georgia Institute of Technology during April 1995–March 1996.

evolution equation

$$\text{(1.2)}\qquad \begin{cases} \dfrac{\partial \Phi}{\partial t} = \operatorname{div}(a(x)\nabla\Phi) + (1-|\Phi|^2)\Phi & \text{in} \quad (0,\infty)\times\Omega, \\ \dfrac{\partial \Phi}{\partial \nu} = 0 & \text{on} \quad (0,\infty)\times\partial\Omega. \end{cases}$$

Equation (1.1) can also be considered as the Euler–Lagrange equation of the Ginzburg–Landau free energy functional

$$\text{(1.3)}\qquad \mathcal{E}(\Phi) := \int_\Omega \left\{ \frac{1}{2}a(x)|\nabla\Phi|^2 + \frac{1}{4}(1-|\Phi|^2)^2 \right\} dx.$$

In this connection, the parabolic equation (1.2) is the gradient flow of $\mathcal{E}$ (see [7]).

The purpose of this article is to construct stable solutions having vortices to (1.1) by appropriately choosing the diffusion coefficient $a(x)$. Here we say that a solution to (1.1) is stable if it is Lyapunov stable as an equilibrium solution to (1.2) (see [7, 9]). In terms of the variational structure, stable equilibria correspond to local minimizers of the energy functional (1.3) in most typical cases (although not always).

Most natural candidates of stable solutions of a variational problem are global minimizers. It is however rather easy to see that the global minimizers of the energy (1.3) are exactly the constant functions with modulus 1. Therefore the solutions we are seeking—the stable equilibria with zeros—have to lie on nonminimal energy level and are not accessible by the direct method in the calculus of variation. In the case where $a(x)$ is constant, it has been shown in [10, 11, 13] that the geometry of the ambient domain plays a crucial role in the stability of equilibria; more precisely, for convex domains of $\mathbb{R}^n$ the Ginzburg–Landau equation (1.1) with constant $a(x) = \lambda^{-1} > 0$ has no stable nonconstant solutions, while in an annulus or generally in a topologically nontrivial domain (a nonsimply connected domain if $n = 2, 3$; see [13]), the equation has stable nonconstant solutions for sufficiently large λ. Those stable solutions, however, have no zero points, i.e., no vortices. Taking into account this fact, let us give a crude description of the time evolution of zeros of a solution to (1.2) in a convex domain with the constant diffusivity a. Deliver a finite number of isolated points arbitrarily in the domain and take any function Φ_0 whose zeros are identified with those points. The zeros of the solution of (1.2) with the initial data Φ_0 represent the migrating vortices. Then by the instability result of [10], under a small perturbation of initial data, all of the vortices of the solution to (1.2) eventually disappear from the domain; every vortex either is absorbed by the boundary or is annihilated through merging with other vortices. An intuitive explanation of this phenomenon is as follows: higher energy of the potential term of (1.3) at vortices drives them moving in the direction of minimizing the energy functional until they disappear. Hence to stabilize each vortex, we have to create an energy barrier around the vortex, with the aid of an appropriate choice of the diffusion coefficient $a(x)$. This is roughly what we are going to do in this article.

Now we present our main theorem.

THEOREM 1.1. *Let Ω be a bounded domain in $\mathbb{R}^2$ with C^3 boundary $\partial\Omega$ and denote by $\overline{\Omega}$ the closure of Ω. Given arbitrarily a finite number of distinct points $\{a_j\}_{j=1}^N \subset \Omega$, a map $\phi \in C^0(\overline{\Omega}\setminus\{a_1,\dots,a_N\}; S^1)$, and a positive number ρ such that*

$$\text{(1.4)}\qquad 0 < \rho < \rho_0 := \min\left\{ \min_{1\le j<k\le N} \frac{1}{2}|a_j - a_k|, \min_{1\le j\le N} \operatorname{dist}(a_j, \partial\Omega) \right\},$$

there exists a C^∞ function $a(x) > 0$ such that the equation (1.1) *has a stable solution $\Phi(x)$ such that $\Phi(x) \neq 0$ for any $x \in \overline{\Omega} \setminus \cup_{j=1}^N B_\rho(a_j)$ and that moreover $\Phi/|\Phi|$ is homotopic to ϕ in $C^0(\overline{\Omega} \setminus \cup_{j=1}^N B_\rho(a_j); S^1)$.*

If we choose in the above theorem a ϕ with nonzero local topological degree d_j at a_j, for example,

$$\phi(z) = \prod_{j=1}^N \frac{(z-a_j)^{d_j}}{|z-a_j|^{d_j}}, \qquad z \in \overline{\Omega} \subset \mathbb{C}, z \neq a_1, \ldots, a_N$$

(where $\mathbb{R}^2$ is identified with $\mathbb{C}$), then our stable solution Φ satisfies that $\deg(\Phi; \partial B_\rho(a_j)) = d_j \neq 0$ and therefore must have at least one zero point within the (small) disk $B_\rho(a_j)$. This leads to the following corollary.

COROLLARY 1.2. *Let Ω be a bounded domain in $\mathbb{R}^2$ with C^3 boundary. Given arbitrarily a finite number of distinct points $\{a_j\}_{j=1}^N \subset \Omega$ and a positive number $\rho \in (0, \rho_0)$, there exists a C^∞ function $a(x) > 0$ such that the equation* (1.1) *has a stable solution $\Phi(x)$ whose zero set*

$$Z[\Phi] := \{x \in \overline{\Omega} |\, \Phi(x) = 0\}$$

is ρ-close to the prescribed configuration in the sense that

$$Z[\Phi] \subset \bigcup_{j=1}^N B_\rho(a_j), \quad Z[\Phi] \cap B_\rho(a_j) \neq \emptyset.$$

The result can be extended to a slightly modified equation of (1.1):

$$\begin{cases} \dfrac{1}{a(x)} \operatorname{div}(a(x) \nabla \Phi) + \lambda(1 - |\Phi|^2)\Phi = 0 & \text{in} \quad \Omega, \\ \dfrac{\partial \Phi}{\partial \nu} = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{1.5}$$

For details, see section 3.

The technique we use in the proof will be essentially the same as that in [12], where it is proved that for the constant diffusivity $a(x) = 1/\lambda$ in (1.1) there is a simply connected domain in $\mathbb{R}^3$ which allows a stable solution with vortices (for sufficiently large λ). Their methods, however, are not applicable to the 2-dimensional case with constant $a(x)$ (see also [3]). It remains an interesting open problem for the constant diffusivity whether there exists a 2-dimensional (nonconvex) domain which can carry a stable solution with a vortex.

We finally remark on the physical meaning of the variable diffusion briefly. In the Ginzburg–Landau theory for superconductivity, $a(x)$ represents the coherence length for superconducting electrons in a material. When a superconductor contains impurities, it is quite natural to consider the inhomogeneous coherence length in the equation, that is, the variable diffusion coefficient. On the other hand, equation (1.5) is a model describing another physical situation, that is, a thin superconducting film with variable thickness. Indeed it can be regarded as an approximate equation in a thin domain of $\mathbf{R}^3$,

$$D(\epsilon) = \{(x, x') \in \mathbf{R}^2 \times \mathbf{R} : x \in \Omega, 0 < x' < \epsilon a(x)\} \qquad (0 < \epsilon \ll 1),$$

where the function $a(x)$ represents the geometrical variation of the thin film. It is proved in [8] that as $\epsilon \to 0$, the averaging limit of the Laplacian in $D(\epsilon)$ is equal to $a(x)^{-1}\mathrm{div}(a(x)\nabla)$ in Ω in some sense. Hence, the Ginzburg–Landau equation with constant coherence length on 3-dimensional domains $D(\epsilon)$ is approximated by equation (1.5) with variable coefficient $a(x)$ on a planar domain Ω. We also note that when a vortex is trapped by some defect of the conductor, it is called the "pinning" effect in the literature of physics, which is just analytically realized in our main result.

2. Proof of the main theorem. Let Ω be a bounded domain of $\mathbb{R}^2$ with C^3 boundary $\partial\Omega$. We always identify the complex plane $\mathbb{C}$ with $\mathbb{R}^2$. So the function spaces $C^0(\overline{\Omega};\mathbb{C})$ and $L^2(\Omega;\mathbb{C})$ are the same as $C^0(\overline{\Omega};\mathbb{R}^2)$ and $L^2(\Omega;\mathbb{R}^2)$, respectively. We will also use the Sobolev space

$$H^1(\Omega;\mathbb{C}) \equiv H^1(\Omega;\mathbb{R}^2) = \{u \in L^2(\Omega;\mathbb{R}^2) : \partial u/\partial x_j \in L^2(\Omega;\mathbb{R}^2), j = 1,\dots,n\}.$$

We often abbreviate the above spaces as $C^0(\overline{\Omega})$, $L^2(\Omega)$, and $H^1(\Omega)$, respectively.

2.1. Stability of local minimizers. We first recall a result essentially due to L. Simon (see Corollary 2 of [15]).

A function $\Phi \in H^1(\Omega)$ is called a *local minimizer* of $\mathcal{E}$ if $\mathcal{E}(\Psi) \geq \mathcal{E}(\Phi)$ for any Ψ close to Φ in the H^1-norm. It is easily seen that a local minimizer has to satisfy (1.1). A solution of Φ of (1.1) is called a *stable equilibrium* of (1.2) if for any $\epsilon > 0$ there is a $\delta > 0$ such that any solution $\Phi(x,t)$ of (1.2) with $\|\Phi(\cdot,0) - \Phi\|_{C^0(\overline{\Omega})} \leq \delta$ satisfies $\|\Phi(\cdot,t) - \Phi\|_{C^0(\overline{\Omega})} \leq \epsilon$ for all $t \geq 0$.

LEMMA 2.1. *Φ is a local minimizer of $\mathcal{E}$ if and only if it is a stable equilibrium of* (1.2).

Thanks to this lemma, we only need to find local minimizers of $\mathcal{E}$ satisfying the required properties in Theorem 1.1. The proof of Lemma 2.1 uses an infinite dimensional version of the Łojasiewicz inequality, in whose derivation the real analyticity of the nonlinear term in the equation plays a role. We refer to [15] for more detail.

2.2. Minimizers on a domain with holes. Arbitrarily given N-points $\{a_k\}_{k=1}^N \subset \Omega$ and a positive number $\rho \in (0,\rho_0)$, where ρ_0 is as in (1.4), we define a subdomain of Ω by

$$\Omega_\rho := \Omega \setminus \bigcup_{j=1}^{N} \overline{B_{\rho/2}(a_j)}. \tag{2.1}$$

The problem (with constant diffusivity) on the domain Ω_ρ

$$\begin{cases} \lambda^{-1}\Delta\Phi + (1-|\Phi|^2)\Phi = 0 & \text{in} \quad \Omega_\rho, \\ \dfrac{\partial\Phi}{\partial\nu} = 0 & \text{on} \quad \partial\Omega_\rho \end{cases} \tag{2.2}$$

has been studied in [13] and [12], where nonconstant stable solutions are constructed within any nontrivial homotopy class of $C^0(\overline{\Omega_\rho};\mathbb{R}^2\setminus\{0\})$. The corresponding restricted energy functional over Ω_ρ is

$$\mathcal{E}_\rho(\Phi) := \int_{\Omega_\rho} \left\{ \frac{\lambda^{-1}}{2}|\nabla\Phi|^2 + \frac{1}{4}(1-|\Phi|^2)^2 \right\} dx. \tag{2.3}$$

Applying Theorem 2.1 of [13] and Lemma 2.3 of [12] to the above equation yields the next lemma.

LEMMA 2.2. *Consider equation* (2.2) *under the assumption that the boundary* $\partial\Omega$ *is* C^3. *Let* $\phi : \overline{\Omega_\rho} \to S^1$ *be continuous. Then*

(i) *there is a* $\lambda^* = \lambda^*(\Omega_\rho) > 0$ *such that for any* $\lambda > \lambda^*$ (2.2) *has a stable solution* $\Phi_\rho(x)$ *satisfying*

$$|\Phi_\rho(x)| > 0 \ (x \in \overline{\Omega_\rho}) \text{ and } \phi \text{ is homotopic to } \Phi_\rho/|\Phi_\rho| \text{ in } C^0(\overline{\Omega_\rho}; S^1). \tag{2.4}$$

(ii) *Moreover, for each* $\lambda > \lambda^*$ *and* $\eta > 0$, *there are positive numbers* δ_0 *and* μ_0 (*depending on* Ω, ρ, λ, *and* η) *such that for any* $\Psi \in H^1(\Omega_\rho) \cap C^0(\overline{\Omega_\rho})$ *with*

$$\|\Psi\|_{C^0(\overline{\Omega_\rho})} \le 1 + \eta, \qquad \inf_{0 \le c < 2\pi} \|e^{ic}\Phi_\rho - \Psi\|_{L^2(\Omega_\rho)} \le \delta_0,$$

the following inequality holds:

$$\mathcal{E}_\rho(\Psi) \ge \mathcal{E}_\rho(\Phi_\rho) + \mu_0 \inf_{0 \le c < 2\pi} \|e^{ic}\Phi_\rho - \Psi\|^2_{L^2(\Omega_\rho)}. \tag{2.5}$$

From now on, the continuous map $\phi : \overline{\Omega} \setminus \{a_1, \dots, a_N\} \to S^1$ and the numbers $0 < \rho < \rho_0, \lambda > \lambda^*$, and $\eta > 0$ are fixed. We shall also assume the C^∞ smoothness of ϕ without loss of generality. (If necessary, replace ϕ by a smooth map which is homotopic to ϕ in the topology of $C^0(\overline{\Omega} \setminus \{a_1, \dots, a_N\}; S^1)$.)

2.3. Fill the holes. For resolving the original problem on the whole domain $\overline{\Omega}$, we extend suitably the coefficient λ^{-1} and Φ_ρ onto functions $a(x)$ and $\Psi(x)$ over $\overline{\Omega}$. From the intuitive discussion in the introduction, a natural choice of $a(x)$ can be a positive function equal to λ^{-1} on $\overline{\Omega}$ but much smaller near a_j in order to produce appropriate energy barriers. The extension $\Psi(x)$ should make the energy integrals over the disks $B_{\rho/2}(a_j)$ sufficiently small. Such a Ψ will be our "approximate stable candidate," one of whose positively invariant neighborhood will finally provide a stable equilibrium.

The following lemma gives the precise conditions that a and Ψ should meet.

LEMMA 2.3. *Assume that the boundary* $\partial\Omega$ *is* C^3 *and let* Φ_ρ, μ_0, *and* δ_0 *be as in Lemma* 2.2. *Suppose that* $0 < a \in C^\infty(\overline{\Omega})$ *and* $\Psi \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ *satisfy*

$$a|_{\overline{\Omega_\rho}} = \lambda^{-1}, \quad \Psi|_{\overline{\Omega_\rho}} = \Phi_\rho, \tag{2.6}$$

$$\|\Psi\|_{C^0(\overline{\Omega})} < 1 + \eta, \tag{2.7}$$

$$\sum_{j=1}^{N} \int_{B_{\rho/2}(a_j)} \left\{ \frac{1}{2} a(x) |\nabla\Psi|^2 + \frac{1}{4}(1 - |\Psi|^2)^2 \right\} dx < \mu_0\delta^2 - \gamma, \tag{2.8}$$

where $0 < \delta < \delta_0$ *and* $0 < \gamma < \mu_0\delta^2$ *are two positive numbers. Then,*

(i) *the set of functions defined by*

$$\begin{aligned} E(\gamma, \delta, a, \Psi) := \Big\{ & U \in H^1(\Omega) \cap C^0(\overline{\Omega}) : \|U\|_{C^0(\overline{\Omega})} \le 1 + \eta, \\ & \mathcal{E}(U) \le \mathcal{E}(\Psi) + \gamma, \\ & \inf_{0 \le c < 2\pi} \|U - e^{ic}\Phi_\rho\|_{L^2(\Omega_\rho)} \le \delta \Big\} \end{aligned} \tag{2.9}$$

has nonempty interior in the $C^0(\overline{\Omega}) \cap H^1(\Omega)$-topology and is positively invariant under the semiflow generated by the time-dependent Ginzburg–Landau equation (1.2);

(ii) *the energy functional $\mathcal{E}$ has a local minimizer Φ in the interior of $E(\gamma, \delta, a, \Psi)$.*

Proof. Part (i): By assumption, it is obvious that Ψ is an interior point of $E(\gamma, \delta, a, \Psi)$. Let $U \in E(\gamma, \delta, a, \Psi)$ and let $\Phi(x,t)$ be the solution of (1.2) with initial condition $\Phi(x,0) = U(x)$. We need to show that $\Phi(\cdot, t) \in E(\gamma, \delta, a, \Psi)$ for any $t > 0$. By the standard regularity theory for parabolic equations, we have $\Phi(\cdot, t) \in H^1(\Omega) \cap C^0(\overline{\Omega})$. The energy functional $\mathcal{E}$ in (1.3) is a Lyapunov function for the evolution equation (1.2); indeed

$$\frac{d}{dt}\mathcal{E}(\Phi(\cdot,t)) = -\int_\Omega \left|\frac{\partial \Phi}{\partial t}(x,t)\right|^2 dx.$$

Hence

$$\mathcal{E}(\Phi(\cdot,t)) \le \mathcal{E}(U) \le \mathcal{E}(\Psi) + \gamma, \qquad t > 0. \tag{2.10}$$

An application of the maximum principle gives

$$\|\Phi(\cdot,t)\|_{C^0(\overline{\Omega})} \le \max\left\{1, \|U\|_{C^0(\overline{\Omega})}\right\} (\le 1 + \eta), \qquad t > 0. \tag{2.11}$$

It remains to prove that $\inf_{0 \le c < 2\pi} \|\Phi(\cdot,t) - e^{ic}\Phi_\rho\|_{L^2(\Omega_\rho)} \le \delta$ for $t > 0$. Suppose by contradiction that there exists a $t_0 > 0$ such that

$$\delta < \inf_{0 \le c < 2\pi} \|\Phi(\cdot,t_0) - e^{ic}\Phi_\rho\|_{L^2(\Omega_\rho)} < \delta_0. \tag{2.12}$$

From (2.11) and (2.12) together with Lemma 2.2 (ii), we get a lower bound for the energy on the domain with holes:

$$\mathcal{E}_\rho(\Phi(\cdot,t_0)) > \mathcal{E}_\rho(\Phi_\rho) + \mu_0\delta^2. \tag{2.13}$$

Since the energy density is everywhere nonnegative, on the whole domain Ω, we have

$$\mathcal{E}_\rho(\Phi(\cdot,t_0)) \le \mathcal{E}(\Phi(\cdot,t_0)) \le \mathcal{E}(\Psi) + \gamma,$$

in view of (2.10). We thereby obtain

$$\mathcal{E}_\rho(\Phi_\rho) + \mu_0\delta^2 < \mathcal{E}(\Psi) + \gamma. \tag{2.14}$$

On the other hand, in view of the splitting

$$\mathcal{E}(\Psi) = \mathcal{E}_\rho(\Phi_\rho) + \sum_{j=1}^N \int_{B_{\rho/2}(a_j)} \left\{\frac{1}{2}a(x)|\nabla\Psi|^2 + \frac{1}{4}(1-|\Psi|^2)^2\right\} dx$$

and the assumption (2.8), we have $\mathcal{E}(\Psi) \le \mathcal{E}_\rho(\Phi_\rho) + \mu_0\delta^2 - \gamma$. This contradicts (2.14).

Part (ii): This part is very similar to the argument used in the proof of Theorem 3.1 of [12]. We only sketch it. One can look at the ω-limit set K of the positively invariant set $E := E(\gamma, \delta, a, \Psi)$ under the semiflow generated by (1.2). By the parabolic regularity theory, K is nonempty and compact. Moreover, by what we have seen in the proof of Part (i), K is contained in the interior of E. A minimizer of $\mathcal{E}$ on K will be a minimizer on the set E. ∎

2.4. Construction of extensions. In this section, we give an explicit construction of a family of $a(x)$ and $\Psi(x)$ satisfying the requirements in Lemma 2.3.

Let $\epsilon > 0$ be a small parameter. First prepare three auxiliary C^∞ functions $A_\epsilon(r)$, $\xi_\epsilon(r)$, and $\eta_\epsilon(r)$ on $[0,\infty)$, with the following properties, respectively:

$$A_\epsilon(r) = \begin{cases} \epsilon^3 & 0 \le r \le (\rho-\epsilon)/2, \\ \text{monotone increasing} & (\rho-\epsilon)/2 < r < \rho/2, \\ \lambda^{-1} & r \ge \rho/2, \end{cases}$$

$$\xi_\epsilon(r) = \begin{cases} r^2/(2\epsilon^2) & 0 \le r \le \epsilon, \\ \text{monotone increasing} & \epsilon < r < 2\epsilon, \\ 1 & 2\epsilon \le r \le (\rho-3\epsilon)/2, \\ \text{monotone decreasing} & (\rho-3\epsilon)/2 < r < (\rho-2\epsilon)/2, \\ 0 & r \ge (\rho-2\epsilon)/2. \end{cases}$$

$$\eta_\epsilon(r) = \begin{cases} 0 & 0 \le r \le (\rho-3\epsilon)/2, \\ \text{monotone increasing} & (\rho-3\epsilon)/2 < r < (\rho-2\epsilon)/2, \\ 1 & r \ge (\rho-2\epsilon)/2. \end{cases}$$

Moreover, we require that $\epsilon|\xi_\epsilon'(r)|$ and $\epsilon\eta_\epsilon'(r)$ are uniformly bounded by a constant independent of ϵ.

Now we define $a_\epsilon(x)$ and $\Psi_\epsilon(x)$. For $x \in \overline{\Omega_\rho}$, let

$$a_\epsilon(x) = \lambda^{-1}, \quad \Psi_\epsilon(x) = \Phi_\rho(x).$$

In the jth disk $B_{\rho/2}(a_j)$, for $x = a_j + r(\cos\theta, \sin\theta)$ with $0 \le r < \rho/2$, let

$$\begin{aligned} a_\epsilon(x) &= A_\epsilon(r), \\ \Psi_\epsilon(x) &= \xi_\epsilon(r)\phi\left(a_j + \frac{\rho}{2}(\cos\theta,\sin\theta)\right) + \eta_\epsilon(r)\Phi_\rho\left(a_j + \frac{\rho}{2}(\cos\theta,\sin\theta)\right). \end{aligned}$$

We claim the following lemma.

LEMMA 2.4. *There exists a constant β_0, independent of ϵ, δ, and γ, such that if $0 < \epsilon < \beta_0(\mu_0\delta^2 - \gamma)$, then the above chosen $a_\epsilon \in C^\infty(\overline{\Omega})$ and $\Psi_\epsilon \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ satisfy the requirements* (2.6), (2.7), *and* (2.8) *in Lemma* 2.3.

Proof. All conditions can be verified straightforward. Let us only check (2.8). In the jth disk, use the polar coordinates $x - a_j = re^{i\theta}$ where $0 \le r \le \rho/2, \theta \in S^1$. Recall the identity

$$|\nabla f|^2 = \left|\frac{\partial f}{\partial r}\right|^2 + \frac{1}{r^2}\left|\frac{\partial f}{\partial \theta}\right|^2.$$

Now integrating the energy integrands $J_1 = \frac{1}{2}a_\epsilon(x)|\nabla\Psi_\epsilon(x)|^2$ and $J_2 = \frac{1}{4}(1-|\Psi_\epsilon|^2)^2$ over separated pieces of $B_{\rho/2}(a_j)$, we see

$$\begin{aligned} \int_{B_{\rho/2}(a_j)} J_1 dx &= \int_0^{(\rho-\epsilon)/2}\int_{S^1} J_1 \, r dr d\theta + \int_{(\rho-\epsilon)/2}^{\rho}\int_{S^1} J_1 \, r dr d\theta \\ &\le O\left(\epsilon^3 \cdot \epsilon^{-2}\right) + O(\epsilon) = O(\epsilon) \end{aligned}$$

and

$$\int_{B_{\rho/2}(a_j)} J_2 dx = \int_0^{2\epsilon}\int_{S^1} + \int_{2\epsilon}^{(\rho-3\epsilon)/2}\int_{S^1} + \int_{(\rho-3\epsilon)/2}^{\rho}\int_{S^1}$$
$$\leq O(\epsilon^2) + 0 + O(\epsilon) = O(\epsilon).$$

For small $\epsilon > 0$, (2.8) is satisfied. □

2.5. Convergence of local minimizers. For $0 < \epsilon < \beta_0(\mu_0\delta^2 - \gamma)$, the functions $a_\epsilon(x)$ and $\Psi_\epsilon(x)$ then provide a local minimizer $\Phi_{\gamma,\delta,\epsilon} \in E(\gamma, \delta, a_\epsilon, \Psi_\epsilon)$ by the last two lemmas. In this section, we let δ and γ depend on ϵ such that $\beta_0(\mu_0\delta^2 - \gamma) = 2\epsilon$ and $\gamma = \epsilon$ and investigate the asymptotic behavior of minimizers $\tilde{\Phi}_\epsilon := \Phi_{\gamma(\epsilon),\delta(\epsilon),\epsilon}$ as $\epsilon \downarrow 0$.

LEMMA 2.5. *Let $\tilde{\Phi}_\epsilon \in E(\gamma(\epsilon), \delta(\epsilon), a_\epsilon, \Psi_\epsilon)$ be local minimizers constructed above. Then we have*

$$\lim_{\epsilon\to 0}\left[\inf_{0\leq c<2\pi} \|e^{ic}\Phi_\rho - \tilde{\Phi}_\epsilon\|_{L^2(\Omega_\rho)}\right] = 0, \tag{2.15}$$

$$\lim_{\epsilon\to 0}\left[\inf_{0\leq c<2\pi} \|e^{ic}\Phi_\rho - \tilde{\Phi}_\epsilon\|_{C^1(\overline{\Omega_{2\rho}})}\right] = 0, \tag{2.16}$$

where

$$\overline{\Omega_{2\rho}} = \overline{\Omega} \setminus \bigcup_{j=1}^{N} B_\rho(a_j).$$

From the convergence (2.16), there is an $\epsilon_0 > 0$ such that for any $0 < \epsilon < \epsilon_0$, the map $\tilde{\Phi}_\epsilon/|\tilde{\Phi}_\epsilon|$ is homotopic to $e^{ic}\Phi_\rho/|\Phi_\rho|$ for some $c \in [0, 2\pi)$ (and hence for all $c \in [0, 2\pi)$), in $C^0(\overline{\Omega} \setminus \bigcup_{j=1}^N B_\rho(a_j); S^1)$. Combined with property (2.4) in Lemma 2.2, we find that $\tilde{\Phi}_\epsilon/|\tilde{\Phi}_\epsilon|$ is homotopic to ϕ. The main Theorem 1.1 is now completely proved.

Proof of Lemma 2.5. The L^2 convergence (2.15) follows from the definition (2.9) and the fact that $\delta = O(\epsilon^{1/2})$.

The C^1 convergence (2.16) follows from (2.15), once we show the relative compactness of $\tilde{\Phi}_\epsilon$ in $C^1(\overline{\Omega_{2\rho}})$. Since $\|\tilde{\Phi}_\epsilon\|_{C^0(\overline{\Omega})}$ is uniformly bounded, we obtain that $\mathrm{div}(a_\epsilon(x)\nabla\tilde{\Phi}_\epsilon)$ are also uniformly bounded in the L^∞ norm. In view of the fact that $a_\epsilon(x) = \lambda^{-1}$ is independent of ϵ on $\overline{\Omega_\rho}$, the elliptic L^p estimates ($p > 2$) give rise to a uniform upper bound for $\|\tilde{\Phi}_\epsilon\|_{W^{2,p}(\Omega_{3\rho/2})}$ (see [5]), which implies the required relative compactness. □

3. Remark. The arguments we used in the previous section are also applicable to other similar types of energy functionals. For instance, consider

$$\mathcal{F}(\Phi) := \int_\Omega \left\{\frac{1}{2}|\nabla\Phi|^2 + \frac{\lambda}{4}(1 - |\Phi|^2)^2\right\} a(x)dx \tag{3.1}$$

for $\Phi \in H^1(\Omega)$. One has the following analogue of Theorem 1.1.

THEOREM 3.1. *Let Ω be a bounded domain in $\mathbb{R}^2$ with C^3 boundary $\partial\Omega$ and denote by $\overline{\Omega}$ the closure of Ω. Given arbitrarily a finite number of distinct points $\{a_j\}_{j=1}^N \subset \Omega$, a map $\phi \in C^0(\overline{\Omega} \setminus \{a_1, \dots, a_N\}; S^1)$, and a positive number ρ such that*

$$0 < \rho < \rho_0 := \min \left\{ \min_{1 \le j < k \le N} \frac{1}{2}|a_j - a_k|, \min_{1 \le j \le N} \operatorname{dist}(a_j, \partial\Omega) \right\},$$

there exist a C^∞ function $a(x) > 0$ and a positive number $\lambda > 0$ such that the functional $\mathcal{F}$ has a local minimizer $\Phi(x)$ such that $\Phi(x) \neq 0$ for any $x \in \overline{\Omega} \setminus \cup_{j=1}^N B_\rho(a_j)$ and that moreover $\Phi/|\Phi|$ is homotopic to ϕ in $C^0(\overline{\Omega} \setminus \cup_{j=1}^N B_\rho(a_j); S^1)$.

Recall that by Simon's result (see Lemma 2.1), the local minimizers Φ in Theorem 3.1 are stable stationary solutions of the following time-dependent equation:

$$\begin{cases} c(x)\dfrac{\partial \Phi}{\partial t} = \operatorname{div}(a(x)\nabla\Phi) + \lambda a(x)(1 - |\Phi|^2)\Phi & \text{in} \quad (0,\infty) \times \Omega, \\ \dfrac{\partial \Phi}{\partial \nu} = 0 \qquad \text{on} \quad (0,\infty) \times \partial\Omega, \end{cases}$$

where $c(x)$ is an arbitrary positive smooth function. A special case, where $c(x) = a(x)$, i.e.,

$$\begin{cases} \dfrac{\partial \Phi}{\partial t} = \dfrac{1}{a(x)}\operatorname{div}(a(x)\nabla\Phi) + \lambda(1 - |\Phi|^2)\Phi & \text{in} \quad (0,\infty) \times \Omega, \\ \dfrac{\partial \Phi}{\partial \nu} = 0 \qquad \text{on} \quad (0,\infty) \times \partial\Omega, \end{cases}$$

is often encountered in applications.

The proof of Theorem 3.1 is very similar to that of Theorem 1.1. We replace Lemma 2.3 by the following slightly generalized version.

LEMMA 3.2. *Assume that the boundary $\partial\Omega$ is C^3 and let Φ_ρ, μ_0, and δ_0 be as in Lemma 2.2. Suppose that $0 < a \in C^\infty(\overline{\Omega})$, $0 \le b \in C^\infty(\overline{\Omega})$, and $\Psi \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ satisfy*

$$a|_{\overline{\Omega_\rho}} = \lambda^{-1}, \quad b|_{\overline{\Omega_\rho}} = 1, \quad \Psi|_{\overline{\Omega_\rho}} = \Phi_\rho,$$

$$\|\Psi\|_{C^0(\overline{\Omega})} < 1 + \eta,$$

$$\sum_{j=1}^N \int_{B_{\rho/2}(a_j)} \left\{ \frac{1}{2}a(x)|\nabla\Psi|^2 + \frac{1}{4}b(x)(1 - |\Psi|^2)^2 \right\} dx < \mu_0\delta^2 - \gamma,$$

where $0 < \delta < \delta_0$ and $0 < \gamma < \mu_0\delta^2$ are two positive numbers. Let

$$\mathcal{G}(U) := \int_\Omega \left\{ \frac{1}{2}a(x)|\nabla\Phi|^2 + \frac{1}{4}b(x)(1 - |\Phi|^2)^2 \right\} dx. \tag{3.2}$$

Then

(i) *the set of functions defined by*

$$G(\gamma,\delta,a,b,\Psi) := \left\{ U \in H^1(\Omega)\cap C^0(\overline{\Omega}) : \|U\|_{C^0(\overline{\Omega})} \le 1+\eta, \ \mathcal{G}(U) \le \mathcal{G}(\Psi)+\gamma, \ \inf_{0\le c<2\pi} \|U - e^{ic}\Phi_\rho\|_{L^2(\Omega_\rho)} \le \delta \right\} \tag{3.3}$$

has nonempty interior in the $C^0(\overline{\Omega})\cap H^1(\Omega)$*-topology and is positively invariant under the semiflow generated by the following time-dependent Ginzburg–Landau equation*:

$$\begin{cases} \dfrac{\partial\Phi}{\partial t} = \operatorname{div}(a(x)\nabla\Phi) + b(x)(1-|\Phi|^2)\Phi \qquad \text{in} \quad (0,\infty)\times\Omega, \\ \dfrac{\partial\Phi}{\partial\nu} = 0 \qquad \text{on} \quad (0,\infty)\times\partial\Omega; \end{cases}$$

(ii) *the energy functional* $\mathcal{G}$ *has a local minimizer* Φ *in the interior of* $G(\gamma,\delta,a,b,\Psi)$.

The conditions in Lemma 3.2 are satisfied by $a_\epsilon(x)$ and $\Psi_\epsilon(x)$ chosen the same as in section 2.4 along with $b_\epsilon(x) := \lambda a_\epsilon(x)$. The proof of the convergences remains literally unchanged from section 2.5.

Acknowledgments. Y. Morita would like to take this opportunity to express his acknowledgment to the Georgia Institute of Technology for the secretarial support and to Ryukoku University for the financial support during April 1995–March 1996.

REFERENCES

[1] P. Baumann, N. Carlson, and D. Phillips, *On the zeros of solutions to Ginzburg–Landau type systems*, SIAM J. Math. Anal., 24 (1993), pp. 1283–1293.
[2] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices*, Birkhäuser, Boston, Cambridge, MA, 1994.
[3] N. Dancer, *Domain variation for certain sets of solutions and applications*, Topol. Methods Nonlinear Anal., 7 (1996), pp. 95–113.
[4] C. Elliott, H. Matano, and T. Qi, *Zeros of a complex Ginzburg-Landau order parameter with application to superconductivity*, European J. Appl. Math., 5 (1994), pp. 431–448.
[5] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1983.
[6] V. Ginzburg and L. Landau, *On the theory of superconductivity*, Zheksper. Teor. Fiz., 20 (1950), pp. 1064–1082.
[7] J. K. Hale, *Asymptotic Behaviour of Dissipative Systems*, Math. Surveys and Monographs 25, Amer. Math. Soc., Providence, RI, 1988.
[8] J. K. Hale and G. Raugel, *Reaction-diffusion equation on thin domains*, J. Math. Pures Appl., 71 (1992), pp. 33–95.
[9] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, New York, 1981.
[10] S. Jimbo and Y. Morita, *Stability of non-constant steady state solutions to a Ginzburg-Landau equation in higher space dimensions*, Nonlinear Anal., 22 (1994), pp. 753–770.
[11] S. Jimbo and Y. Morita, *Ginzburg–Landau equation and stable solutions in a rotational domain*, SIAM J. Math. Anal., 27 (1996), pp. 1360–1385.
[12] S. Jimbo and Y. Morita, *Stable solutions with zeros to the Ginzburg-Landau equation with Neumann boundary condition*, J. Differential Equations, 128 (1996), pp. 596–613.
[13] S. Jimbo, Y. Morita, and J. Zhai, *Ginzburg-Landau equation and stable steady state solutions in a non-trivial domain*, Comm. Partial Differential Equations, 20 (1995), pp. 2093–2112.
[14] J. Neu, *Vortices in complex scalar fields*, Physica D, 43 (1990), pp. 385–406.
[15] L. Simon, *Asymptotics for a class of non-linear evolution equations with applications to geometric problems*, Ann. of Math., 118 (1983), pp. 525–571.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 913–934, July 1998

007

STATIONARY PARTICLE SYSTEMS APPROXIMATING STATIONARY SOLUTIONS TO THE BOLTZMANN EQUATION*

S. CAPRINO†, M. PULVIRENTI‡, AND W. WAGNER§

Abstract. We show that a regularized stationary Boltzmann equation with diffusive boundary conditions can be rigorously derived from a suitable stochastic N-particle system. To do this, we prove that the L_1-distance between the k-particle density and the k-fold product of the solution to the stationary Boltzmann equation is of order $1/N$.

Key words. stationary Boltzmann equation, diffusive boundary conditions, stochastic particle system, rate of convergence

AMS subject classifications. 60K35, 76P05, 82C40

PII. S0036141096309988

1. Introduction.

Stochastic particle methods are widely used in the numerical simulation of rarefied flows. These are described at a mathematical level by the Boltzmann equation and hence, convergence results for such schemes are of practical interest. From a more fundamental point of view, in the study of these problems we are naturally led to tackle subtle difficulties related to the so-called propagation of chaos, which is an asymptotic (in the number of particles) statistical independence. Indeed, the convergence we want to establish is nothing else but a law of large numbers for (somehow weakly) dependent random variables. For this reason, results in this direction are also of interest in the field of limit theorems for large systems of interacting stochastic processes. We address the reader to [C], [BI], [LP], [W], [PWZ], [GM] for results concerning convergence of stochastic particle systems to solutions of (regularized) Boltzmann equations (see also [M], [P] for a review on these arguments and related results). Unfortunately the situation is far from being satisfactory for many reasons which we are going to illustrate.

The convergence results we mentioned above regard time-dependent problems. Namely, the empirical measure $\frac{1}{N}\sum_{i=1}^{N}\delta_{z_i(t)}(dz)$ (that is, a measure valued stochastic process), where $z_i(t)$ is the state of the ith particle at time t, is weakly converging in probability to $f(z,t)$, which is the solution of the Boltzmann equation with initial datum $f(z,0)=f_0(z)$, the distribution density of each particle at time zero, assuming also that all the particles are independently distributed. Such a convergence is not expected to hold uniformly in time. However, in most of the practical applications of these stochastic codes, we deal with stationary nonequilibrium situations, which we simulate in order to extract information on the macroscopic quantities such as profiles and fluxes. In other words, we are interested in nontrivial stationary solutions to the Boltzmann equation, but in this case the methods we have discussed so far are useless. In fact, even knowing the trend to a nonequilibrium stationary state for the Boltzmann

*Received by the editors September 29, 1996; accepted for publication (in revised form) July 23, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/30998.html
†Dipartimento di Matematica, Università di Roma, "Tor Vergata," Via della Ricerca Scientifica, 00133 Roma, Italy (caprino@axp.mat.utovrm.it).
‡Dipartimento di Matematica, Universitá di Roma, "La Sapienza," Piazzale Aldo Moro 2, 00185 Roma, Italy (pulvirenti@axcasp.caspur.it).
§Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany (wagner@wias-berlin.de).

dynamics (which is, incidentally, not known except for simplified models), we could not conclude anything on the particle approximation of this asymptotic state, being that the two limits $N \to \infty$ and $t \to \infty$ are clearly not commutable.

The systematic error of some particle simulation scheme for a stationary model Boltzmann equation was studied in [B]. An alternative approach to the construction of particle schemes for the stationary Boltzmann equation has been proposed in [BS].

In this paper we face the above-mentioned problem for a gas in a bounded domain with diffusive boundary conditions at a possibly-not-constant temperature. We consider the unique stationary measure for the N-particle system and evaluate the distance between this and the measure given by the N-fold product of the unique solution to the stationary Boltzmann equation of cutoff type, with the same boundary conditions. We show that, if the mean free path inverse is sufficiently small, the L_1-difference between the k-particle distribution functions of such two measures vanishes in the limit $N \to \infty$, for any fixed k. To do this we use a technique which we call v-functions. Such a method is used in [CDPP] and [DP] for time-dependent problems related to stochastic particle systems in a lattice, in [CP] for a one-dimensional stationary problem for a model equation, and it is indeed very effective as we shall explain in section 3.

Let us conclude by criticizing the present result. As we said, it holds for small mean free path inverse. This is a consequence of the fact that we use a constructive perturbative technique. Also, the existence and uniqueness for stationary solutions of the Boltzmann equation is proven under the same smallness assumption. We do not even know whether recent approaches to the existence problem (see, for instance, [AN] for a Boltzmann equation in a slab) can be used to obtain at least the existence of solutions for our problem without this assumption. However, the uniqueness of such solutions, which should be preliminarily known for the convergence problem we set, seems at the moment hard to be proven, even for a regularized equation as the one we consider.

In the present paper, the Boltzmann equation enjoys two regularizations. The first, and more important, is a spatial smearing, which is standard in the above-quoted literature. Actually, the existence theory for the true Boltzmann equation is, up to now, too poor to allow us to approach the real problem. In [CP] a model equation without spatial smearing has been successfully attacked. However such model is one-dimensional—that is much easier to deal with. The second type of cutoff is on the set of possible velocities, which is assumed to be bounded away from 0. This assumption is made to take a full advantage by the ergodic property of the Knudsen flow. We absolutely need this as a consequence of our ignorance of qualitative properties of the invariant measure for the N-particle system, which we only know to exist uniquely. In fact, we think that the cutoff on large velocities (see (2.6)) is only technical: it allows us to avoid difficulties which could obscure the real essence of the approach.

2. Notations and results. Let $\Omega \subset R^d$, $d = 2, 3$ be an open set with sufficiently smooth boundary in the physical space, $V = \{v \in R^d : \|v\| \geq u_{\min} > 0\}$ the velocity space, and $[0, T]$ an interval on the real line. For $(x, v, t) \in \Omega \times V \times [0, T]$ consider the following Boltzmann equation of cutoff type

$$\partial_t p(x, v, t) + (v \cdot \nabla_x) p(x, v, t) = \lambda Q(p, p)(x, v, t) \tag{2.1}$$

with initial condition

$$p(x, v, 0) = p_0(x, v) \geq 0 \tag{2.2}$$

and boundary conditions ($n(x)$ is the outward normal in $x \in \partial\Omega$)

$$p(x,v,t) = J(x,t)M(x,v) \quad x \in \partial\Omega, \quad v \cdot n(x) \leq 0. \tag{2.3}$$

Here we used the following symbols: λ is a real parameter,

$$Q(p,p)(x,v,t) = \int_\Omega dy \int_V dv_1 \int_{S_+^{d-1}} deB(v,v_1,e)h_\beta(x,y)\chi((v^*,v_1^*) \in V \times V))$$

$$\times \{p(x,v^*,t)p(y,v_1^*,t) - p(x,v,t)p(y,v_1,t)\}, \tag{2.4}$$

e is the unit vector in R^d, χ is the characteristic function of its argument,

$$v^* = v + e \cdot (v_1 - v)e, \quad v_1^* = v_1 - e \cdot (v_1 - v)e, \tag{2.5}$$

S^{d-1} is the unit sphere, and $S_+^{d-1} = \{e \in S^{d-1} | e \cdot (v - v_1) \geq 0\}$.

Concerning the collision kernel $B : R^d \times R^d \times S^{d-1} \to \mathbb{R}^+$ we assume

$$B(v,v_1,e) \leq c_1 < \infty. \tag{2.6}$$

The function h_β, which acts as a spatial mollifier, is a symmetric function belonging to L_∞, vanishing for $|x-y| \geq \beta > 0$ and such that $\int h_\beta(x,y)dy = 1$. The incoming flux J at the point x is defined as

$$J(x,t) = \int_{v \cdot n(x) \geq 0} dv \quad v \cdot n(x)p(x,v,t). \tag{2.7}$$

Finally, M is a bounded positive function defined on the set

$$\{(x,v) | x \in \partial\Omega, v \in V, v \cdot n(x) \leq 0\},$$

which we require to satisfy the following normalization condition:

$$\int_{v \cdot n(x) \leq 0} dv \quad |v \cdot n(x)|M(x,v) = 1. \tag{2.8}$$

This last assumption, together with the well-known properties of Q, ensures the conservation of the quantity

$$m(t) = \int dx \int dv \quad p(x,v,t), \tag{2.9}$$

which we assume initially to be one so that we consider normalized solutions to problem (2.1)–(2.3).

From a physical point of view, equations (2.1)–(2.3) describe a rarefied gas in a vessel with diffusive boundary conditions at possibly-not-constant temperature on the boundary. The collision operator Q differs from the usual one for the cutoff on the velocities and for the presence of the smearing function h_β. The true Boltzmann equation is recovered by removing the two cutoffs, that is, letting $h_\beta \to \delta$ (δ is the δ-function centered at the origin) and assuming $V = \mathbb{R}^d$.

It will be useful in the following to deal with the mild version of the above problem:

$$p(t,x,v) = S(t)p_0(x,v) + \lambda \int_0^t ds \quad S(t-s)Q(p,p)(x,v,s), \tag{2.10}$$

where $S(t)$ is the **Knudsen semigroup**, that is the solution to the initial boundary value problem:

$$[\partial_t + (v \cdot \nabla_x)]S(t)p_0(x,v) = 0, \tag{2.11}$$

$$(S(t)p_0)(x,v) = J(x,t)M(x,v)\,, \quad x \in \partial\Omega\,, \quad v \cdot n(x) \leq 0\,. \tag{2.12}$$

There exists a unique solution to (2.10), thanks to the Lipschitz continuity in $L_1(x,v)$ of Q, due to the presence of the smearing function h_β and (2.6).

Here we are interested in the stationary equation

$$(v \cdot \nabla_x)g(x,v) = \lambda Q(g,g)(x,v), \tag{2.13}$$

with boundary conditions corresponding to the time-dependent case (2.3) and the normalization property

$$\int dx \int dv g(x,v) = 1.$$

Existence and uniqueness of a solution for a slightly different formulation for such a problem (under a suitable smallness assumption on λ) will be established in Theorem 2.2 below. For the moment, we need a preliminary property of the Knudsen flow expressed by the following theorem which will be proven in the appendix.

THEOREM 2.1. *There exists a unique probability density $\bar{g}$ which is stationary under the action of the Knudsen flow, i.e.,*

$$S(t)\bar{g} = \bar{g}\,, \quad \forall t \in \mathbb{R}^+\,. \tag{2.14}$$

Moreover, for any $\eta > 0$ there exists $T(\eta) > 0$ such that, for any $t \geq T(\eta)$ and for any probability density f, it is

$$\|S(t)f - \bar{g}\|_{L_1} \leq \eta. \tag{2.15}$$

Remark. We stress that the assumption for the velocities to stay bounded away from 0 implies the independence of $T(\eta)$ from the probability density f, which is of great importance to prove our main result.

We now also establish existence and uniqueness for the stationary solution of the boundary value problem (2.13).

THEOREM 2.2. *If λ is sufficiently small, then there exists a unique probability density g which is invariant for the flow* (2.10)*:*

$$g = S(t)g + \lambda \int_0^t ds \quad S(t-s)Q(g,g), \qquad t \in \mathbb{R}^+. \tag{2.16}$$

Moreover, it is globally attractive; that is,

$$\|p(t) - g\|_{L_1} \leq e^{-c\,t}\,,$$

where $p(t)$ is any solution to (2.10) *and c is some constant.*

The proof of this theorem, which is essentially perturbative, is given in the appendix.

Now we introduce the N-particle process which gives the approximation, in the limit $N \to \infty$, to problem (2.1)–(2.3). Let

$$Z_N = (z_1, ..., z_N), \quad z_i = (x_i, v_i), \quad i = 1, ..., N,$$

and for the sake of simplicity put

$$q(z_1, z_2, e) = h_\beta(x_1, x_2)B(v_1, v_2, e)\chi((v^*, v_2^*) \in V \times V). \tag{2.17}$$

We define the generator of the N-particle process, for any function Φ as

$$G_N(\Phi)(Z_N) = G_N^{free}(\Phi)(Z_N) + \frac{\lambda}{N} G_N^{jump}(\Phi)(Z_N), \tag{2.18}$$

where

$$G_N^{free}(\Phi)(Z_N) = \sum_{i=1}^{N} (v_i \cdot \nabla_{x_i})\Phi(Z_N) \tag{2.19}$$

(with diffusive boundary conditions to be specified; see equations (2.24)–(2.25)) and

$$G_N^{jump}(\Phi)(Z_N) = \sum_{1 \le i < j \le N} \int_{S_+^{d-1}} de[\Phi(Z_N^{(i,j)}) - \Phi(Z_N)]q(z_i, z_j, e) \tag{2.20}$$

being

$$Z_N^{(i,j)} = (z_1, \ldots, z_{i-1}, x_i, v_i^*, \ldots, z_{j-1}, x_j, v_j^*, \ldots, z_N). \tag{2.21}$$

Note that G_N^{free} is the generator of N-independent particles moving freely. The outgoing velocity v of each particle after a collision with the boundary at the point x is distributed according to the probability density given in (2.8). In other words, $\exp\{(G_N^{free})^* t\} = S_N(t)$, where $S_N(t)$ is the product of operators acting on a single particle, namely,

$$S_N(t) = \prod_{i=1}^{N} S_{\{i\}}(t), \tag{2.22}$$

where $S_{\{i\}}(t)$ is the Knudsen semigroup associated to the particle i. Therefore, the process described by the generator G_N consists of free motion (including the diffusive boundary conditions) of the N-particle system and random collisions. These collisions take place at random times, with random impact parameter e. The particles of the pair involved in the collision have mutual distance less than β, and their outgoing velocities after the interaction follow the deterministic law (2.5). This model, introduced in [C], is sometimes called the "soft balls" model.

If the system is initially distributed according to a probability density $f^N(Z_N)$, its time evolution is given by $f^N(t) = \exp\{(G_N)^* t\} f^N$. In other words,

$$\partial_t f^N(Z_N, t) + \sum_{i=1}^{N} (v_i \cdot \nabla_{x_i}) f^N(Z_N, t) = \frac{\lambda}{N} G_N^{jump} f^N(Z_N, t), \tag{2.23}$$

with initial conditions $f^N(Z_N,0)=f^N(Z_N)$ and with boundary conditions

(2.24)
$$f^N(z_1,\dots,z_i,\dots,z_N,t)=J_i^N(x_i,t,Z_N(i))M(x_i,v_i),\quad x_i\in\partial\Omega,\quad v_i\cdot n(x_i)\le 0$$

for all $i=1,\dots,N$, with $Z_N(i)=(z_1,\dots,z_{i-1},z_{i+1},\dots,z_N)$ and

$$J_i^N(x_i,t,Z_N(i))=\int_{v_i\cdot n(x_i)\ge 0} dv_i\quad v_i\cdot n(x_i)f^N(Z_N,t). \tag{2.25}$$

Condition (2.24) can easily be generalized to the case in which more than one particle stays on $\partial\Omega$, since those particles evolve independently. However, such events have vanishing probability for the N-particle system.

If we consider the stationary version of (2.23)–(2.24), that is,

$$\sum_{i=1}^N (v_i\cdot\nabla_{x_i})\tilde f^N(Z_N)=\frac{\lambda}{N}G_N^{jump}\tilde f^N(Z_N) \tag{2.26}$$

with the boundary conditions (2.24)–(2.25), we can state the following result, which is proven in the appendix.

THEOREM 2.3. *For all $N>0$ there exists a unique probability density $\tilde f^N=\tilde f^N(Z_N)$ which is invariant under the N-particle process.*

The main goal of this paper is to compare the stationary distribution $\tilde f^N$ with the one-particle stationary distribution g constructed in Theorem 2.2. To this purpose we introduce the k-particle distribution functions associated to the probability density $\tilde f^N$:

$$\tilde f_k^N(Z_k)=\int\cdots\int \tilde f^N(Z_N)dz_{k+1}\dots dz_N,\qquad k=1,\dots,N-1. \tag{2.27}$$

Introducing analogously the k-particle distribution functions for the time-dependent distribution $f^N(Z_N,t)$, we obtain from (2.23) and (2.24) the well-known BBGKY hierarchy of equations

$$\partial_t f_k^N(Z_k,t)+G_k^{free}f_k^N(Z_k,t)$$

$$=\frac{\lambda}{N}G_k^{jump}f_k^N(Z_k,t)+\lambda\frac{N-k}{N}C_{k,k+1}f_{k+1}^N(Z_k,t),\qquad k=1,\dots,N\,, \tag{2.28}$$

with boundary conditions

(2.29)
$$f_k^N(z_1,\dots,z_i,\dots,z_k,t)=J_i^N(x_i,t,Z_k(i))M(x_i,v_i),\quad x_i\in\partial\Omega,\quad v_i\cdot n(x_i)\le 0.$$

Here,

$$C_{k,k+1}f_{k+1}^N(Z_k,t)$$

$$=\sum_{i=1}^k\int\int_{S_+^{d-1}} dz_{k+1}de\quad q(z_i,z_{k+1},e)[f_{k+1}^N(Z_{k+1}^{(i,k+1)},t)-f_{k+1}^N(Z_{k+1},t)]. \tag{2.30}$$

Note that by Theorem 2.3 it follows that the unique solutions to the stationary version of problem (2.28) are those defined in (2.27).

Now we introduce the infinite Boltzmann hierarchy; that is, the (formal) limit as $N \to \infty$ of the BBGKY hierarchy, i.e.,

$$\partial_t f_k(Z_k,t) + G_k^{free} f_k(Z_k,t) = \lambda C_{k,k+1} f_{k+1}(Z_k,t), \qquad k = 1,2,\dots, \tag{2.31}$$

with initial and usual boundary conditions.

It is useful to consider the mild form of it, that is,

$$f_k(t) = S_k(t) f_k^0 + \lambda \int_0^t ds S_k(t-s) C_{k,k+1} f_{k+1}(s) \quad k = 1,2,\dots. \tag{2.32}$$

We denote by $P(t)$ the solution operator of the infinite hierarchy (2.32) that is $(P(t)f^0)_k = f_k(t)$. $P(t)$ acts on sequences $f^0 = \{f_k^0\}_{k=1,\dots,\infty}$, $f_k^0 \in L_1((\Omega \times V)^k)$.

Analogously we can define $(P^N(t)f^N)_k = f_k^N(t)$ to be the solution operator of the following finite hierarchy of equations (mild version of (2.28)):

(2.33)

$$\begin{aligned} f_k^N(t) = S_k(t) f_k^N &+ \frac{\lambda}{N} \int_0^t ds \quad S_k(t-s) G_k^{jump} f_k^N(s) \\ &+ \lambda \frac{N-k}{N} \int_0^t ds \quad S_k(t-s) C_{k,k+1} f_{k+1}^N(s), \qquad k = 1,2,\dots,N\,. \end{aligned}$$

Notice that (2.33) for $k = N$ is the mild version of (2.23).

Since (2.33) is a finite system of linear equations, it can easily be solved uniquely in $L_1((\Omega \times V)^k)$; namely, $f_k^N(t)$ are obtained by integrating $f^N(Z_N,t)$, a unique solution to (2.23).

By iterating formula (2.32), we arrive at the following formal series expansion:

$$(P(t)f^0)_k = f_k(t) = \sum_{n=0}^{\infty} \lambda^n a_{k,n}(t) f_k^0 \tag{2.34}$$

with

$$a_{k,n}(t) f_k^0 = \int_0^t \int_0^{t_1}, \dots, \int_0^{t_{n-1}} dt_1, \dots, dt_n$$

$$\times S_k(t-t_1) C_{k,k+1}, \dots, S_{k+n-1}(t_{n-1}-t_n) C_{k+n-1,k+n} S_{k+n}(t_n) f_{k+n}^0. \tag{2.35}$$

It is possible to show that the series in (2.34) converges in L_1 if the quantity λt is sufficiently small so that, under such a hypothesis, there exists a unique solution to (2.32). The method employed is the same as in [LP] and [PWZ], inspired by the well-known result due to Lanford (see [L] and [CIP]) in a L_∞-setup for the not regularized Boltzmann equation. Here we find the additional difficulty of the diffusive boundary conditions. However, working in L_1, this is not a problem, since the only property we need of the free flow is the isometry (see (2.36) below).

We will show the convergence of the series (2.34) as well as the asymptotic equivalence (for $N \to \infty$) of the operators $P^N(t)$ and $P(t)$.

Before stating Theorem 2.4 below, we stress two fairly evident estimates of the terms in the series (2.34):

$$\|S_k(t)f_k\|_{L_1} = \|f_k\|_{L_1} \tag{2.36}$$

and

$$\|C_{k,k+1}f_{k+1}\|_{L_1} \le ka\|f_{k+1}\|_{L_1}, \tag{2.37}$$

where

$$a = 2\sup_{z,z'} \int de\, q(z,z',e). \tag{2.38}$$

THEOREM 2.4. *Suppose $\lambda t < \frac{1}{8a}$. Then, given any sequence $\{f_k^0\}_{k=1,\dots,\infty}$ such that $\|f_k^0\|_{L_1} = 1$, the series (2.34) is absolutely convergent in $L_1((\Omega \times V)^k)$ for all $k \ge 1$. Moreover, given the sequence $f^N = \{f_k^N\}_{k=1,\dots,N}$ of k-particle densities, we have*

$$\|([P^N(t) - P(t)]f^N)_k\|_{L_1} \le \frac{8^k c_2}{N} \tag{2.39}$$

for some constant c_2 independent of f^N.

Remark. Since $P(t)$ has been defined as acting on infinite sequences, in (2.39) we mean $(P(t)f^N)_k = 0$ for $k > N$.

Proof. By (2.35), using (2.36) and (2.37) we have

$$\|a_{k,n}(t)f_k^0\|_{L_1} \le \frac{k(k+1),\dots,(k+n-1)}{n!}(ta)^n\|f_{k+n}^0\|_{L_1} \le 2^k(2ta)^n. \tag{2.40}$$

Therefore the series (2.34) converges for $2ta\lambda < 1$.

Let us define

$$D^N(t) = [P^N(t) - P(t)]f^N, \tag{2.41}$$

$$B_k^N(t) = \frac{\lambda}{N}\int_0^t dsS_k(t-s)G_k^{jump}(P^N(s)f^N)_k, \tag{2.42}$$

$$E_k^N(t) = -\frac{\lambda k}{N}\int_0^t dsS_k(t-s)C_{k,k+1}(P^N(s)f^N)_{k+1}. \tag{2.43}$$

By (2.32) and (2.33) we have

$$D_k^N(t) = B_k^N(t) + E_k^N(t) + \lambda\int_0^t dsS_k(t-s)C_{k,k+1}D_{k+1}^N(s), \qquad k = 1,2,\dots,N-1. \tag{2.44}$$

Iterating (2.44) $n-1$ times, with $n \le N-k$, we obtain

$$D_k^N(t) = \sum_{m=0}^{n-1}\lambda^m\int_0^t\int_0^{t_1},\dots,\int_0^{t_{m-1}} dt_1,\dots,dt_m$$

$$\times S_k(t-t_1)C_{k,k+1},\dots,S_{k+m-1}(t_{m-1}-t_m)C_{k+m-1,k+m}(B^N_{k+m}(t_m)+E^N_{k+m}(t_m))$$

$$+\lambda^n\int_0^t\int_0^{t_1},\dots,\int_0^{t_{n-1}}dt_1,\dots,dt_n$$

$$\times S_k(t-t_1)C_{k,k+1},\dots,S_{k+n-1}(t_{n-1}-t_n)C_{k+n-1,k+n}D^N_{k+n}(t_n). \tag{2.45}$$

By (2.34), (2.40), and the assumption $\lambda t<\frac{1}{8a}$, it follows from (2.41) that

$$\|D^N_{k+n}(t)\|_{L_1}\le 1+2^{k+n+1} \tag{2.46}$$

so that, after elementary calculation, we can bound the L_1-norm of the last term in the right-hand side of (2.45) by the quantity $4\cdot 4^k(\frac{1}{2})^n$.

Moreover we have

$$\|B^N_k(t)\|_{L_1}\le\frac{k^2\lambda ta}{N} \tag{2.47}$$

and

$$\|E^N_k(t)\|_{L_1}\le\frac{k^2\lambda ta}{N} \tag{2.48}$$

so that (2.45) implies

$$\|D^N_k(t)\|_{L_1}\le\frac{2^k}{N}\sum_{m\ge 0}(2a\lambda t)^{m+1}(k+m)^2+4\cdot 4^k\left(\frac{1}{2}\right)^n \tag{2.49}$$

$$\le\frac{4^k}{N}\frac{4a\lambda t}{1-4a\lambda t}+4\cdot 4^k\left(\frac{1}{2}\right)^n.$$

The thesis follows by putting $n=N-k$. □

Remark. The above result can be used to show the convergence of the solutions of the N-particle system to the solution of our Boltzmann equation. Indeed, Theorem 2.4 shows the existence and uniqueness of the solutions to hierarchy (2.32) for short times. Assume that the initial datum is factorizing, i.e., $f^N_k=f_0^{\otimes\, k}$, where f_0 is some one-particle probability density. Then it is easy to show that the unique solution of the hierarchy (2.32) we have constructed is of the form $f_k(t)=f^{\otimes\, k}(t)$, where $f(t)$ solves the Boltzmann equation (2.1) with initial datum f_0. This property is called propagation of chaos. Thus we have shown that $f^N_k(t)\to f^{\otimes\, k}(t)$ for all $k>0$, in L_1 and for short times. On the other hand, t must be smaller than a numerical constant independent of f^0 so that the procedure can be iterated in time to show that the convergence is global (see [LP], [PWZ], [P] for details).

Coming back to the stationary problem, we conclude this section by formulating the main result of this paper which will be proven in the next section. We recall that g denotes the stationary solution to the boundary value problem (2.13) constructed in Theorem 2.2 and we set

$$g_k(Z_k)=\prod_{i=1}^k g(z_i). \tag{2.50}$$

We also recall that $\tilde{f}_k^N$ denotes the k-particle distribution of the unique invariant measure of the N-particle system. We will prove the following result.

THEOREM 2.5. *There exists $\lambda_0 > 0$ such that for any $\lambda \le \lambda_0$ and any integer $k \ge 1$ it is*

$$\|\tilde{f}_k^N - g_k\|_{L_1} \le \frac{c^k}{N}, \qquad N > k,$$

for some constant c not depending on λ, k, N.

3. Proof of Theorem 2.5. We introduce a formalism which plays a very important role in what follows. Let $I \subset \mathbb{N}$ be a bounded set of indices and let $|I|$ represent its cardinality. We consider families of symmetric functions $\Phi = \{\Phi_I\}_{I\subset\mathbb{N}}$, where Φ acts on $[\Omega \times V]^{\mathbb{N}}$, and each Φ_I on $[\Omega \times V]^I$, respectively. Given two families $\Phi = \{\Phi_I\}_{I\subset\mathbb{N}}$ and $\Psi = \{\Psi_I\}_{I\subset\mathbb{N}}$, we give the following definition of **$*$-product**:

$$(\phi * \psi)_I(Z_I) = \sum_{J\subseteq I} \phi_J(Z_J)\psi_{I\setminus J}(Z_{I\setminus J}), \tag{3.1}$$

where we are using the notation $Z_I = \{z_i | i \in I\}$. Let us put

$$\phi_I^{\perp} = (-1)^{|I|}\phi_I \tag{3.2}$$

and finally, let us define

$$v_I^N = (g^{\perp} * \tilde{f}^N)_I, \tag{3.3}$$

where we set $\tilde{f}_I^N(Z_I) = \tilde{f}_k^N(Z_I)$ if $|I| = k$. We assume that

$$\tilde{f}_{\emptyset}^N = g_{\emptyset} = v_{\emptyset}^N = 1\,. \tag{3.4}$$

We want to stress that, if it were

$$\tilde{f}_I^N(Z_I) = \prod_{i\in I} f(z_i),$$

then

$$v_I^N(Z_I) = \prod_{i=1}^{|I|}[g(z_i) - f(z_i)]. \tag{3.5}$$

This means, in a sense, that the functions v^N represent the product of the differences rather than the difference of the products which we would have to deal with.

By (3.4) it follows that the definition (3.3) can be inverted to obtain

$$\tilde{f}_I^N = (g * v^N)_I, \tag{3.6}$$

and this implies, as it can be easily seen, that

$$\|\tilde{f}_k^N - g_k\|_{L_1} \le \sum_{J\subset I,\, J\neq\emptyset} \|v_J^N\|_{L_1}\,, \tag{3.7}$$

where $I = \{1, \dots, k\}$. Therefore we will prove Theorem 2.5 by estimating v^N.

As a consequence of Theorem 2.1, we have the following.

LEMMA 3.1. *For any $\eta > 0$ there exists a $T(\eta)$ such that, for $t > T(\eta)$ and $J \subset I$, the following estimate holds:*

$$\|S_J(t)v_I^N\|_{L_1} \le \eta^j \|v_I^N\|_{L_1}\,, \tag{3.8}$$

where $S_J(t) = \prod_{i\in J} S_{\{i\}}(t)$ and $j = |J|$.

Remark. Had we considered directly the difference $\tilde f_I^N - g_I$ in place of v_I^N; at the best we would have obtained $\|S_J(t)(\tilde f_I^N - g_I)\|_{L_1} \le \eta\|\tilde f_I^N - g_I\|_{L_1}$, and this is not sufficient for our purpose.

Proof. We first prove that, for all $\eta > 0$, there exists $T(\eta)$ such that, for $t > T(\eta)$, for all $u = u(z), u \in L_1$, satisfying $\int u dz = 0$, one has

$$\|S(t)u\|_{L_1} \le \eta\|u\|_{L_1}\,. \tag{3.9}$$

Indeed, denoting by u^+ and u^- the positive and negative part of u, respectively, setting

$$A = \int u^+ dz = \int u^- dz \tag{3.10}$$

we have by Theorem 2.1

$$\begin{aligned}\|S(t)u\|_{L_1} &= \|S(t)u^+ - S(t)u^-\|_{L_1}\\ &\le A\|S(t)\left(\frac{u^+}{A}\right) - \bar g\|_{L_1} + A\|S(t)\left(\frac{u^-}{A}\right) - \bar g\|_{L_1} \le 2A\eta = \eta\|u\|_{L_1}.\end{aligned} \tag{3.11}$$

Let $J = \{i_1, \dots, i_j\}$, and define the functions

$$r_k(Z_I) = S_{\{i_1,\dots,i_k\}}(t)\, v_I^N(Z_I)\,, \qquad k = 0, 1, \dots, j\,.$$

These functions satisfy

$$\int r_{k-1}(Z_I)\, dz_{i_k} = 0\,, \qquad k = 1, \dots, j,$$

since $\int v_I^N(Z_I)\, dz_i = 0\,,\ \forall i \in I\,,$ as a consequence of the definition of v^N. Thus, we obtain from (3.9) that

$$\int |r_k(Z_I)|\, dz_{i_k} = \int |S_{\{i_k\}}(t)\, r_{k-1}(Z_I)|\, dz_{i_k} \le \eta \int |r_{k-1}(Z_I)|\, dz_{i_k}$$

and $\|r_k\|_{L_1} \le \eta\, \|r_{k-1}\|_{L_1}\,,\ k = 1, \dots, j$, so that (3.8) follows immediately. □

We recall that $(P^N(t)\tilde f^N)_I = \tilde f_I^N$ for all I such that $0 < |I| \le N$ and $(P(t)g)_I = g_I$ for all I with $|I| > 0$. We extend this invariance property to the empty set; that is, (see (3.4))

$$(P^N(t)\tilde f^N)_\emptyset = 1\,, \qquad (P(t)g)_\emptyset = 1. \tag{3.12}$$

We also put

$$(P(t)\tilde f^N)_\emptyset = 1. \tag{3.13}$$

For any finite set of indices I, we have

$$v_I^N = (g^\perp * \tilde{f}^N)_I = (g^\perp * P^N(t)\tilde{f}^N)_I \tag{3.14}$$
$$= (g^\perp * P(t)\tilde{f}^N)_I + (g^\perp * [P^N(t) - P(t)]\tilde{f}^N)_I.$$

Before going on with the estimate of v^N, we introduce a suitable norm. Given an infinite sequence of L_1-functions $\phi = \{\phi_k\}$ and a real positive number α we set

$$\|\phi\|_\alpha = \sup_{k=1,2,\dots,} \|\phi_k\|_{L_1} e^{-\alpha k}. \tag{3.15}$$

Putting

$$R_I^N(t) = (g^\perp * [P^N(t) - P(t)]\tilde{f}^N)_I \tag{3.16}$$

by Theorem 2.4 and (3.12) and (3.13), it follows that for $\alpha \geq 2\log 3$ and $\lambda t < \frac{1}{8a}$, we have

$$\|R^N(t)\|_\alpha \leq \frac{c_2}{N}\,. \tag{3.17}$$

Indeed, suppose $|I| = k$,

$$\|R_I^N(t)\|_{L_1} \leq \sum_{h=1}^{k} \binom{k}{h} \|([P^N(t) - P(t)]\tilde{f}^N)_h\|_{L_1} \leq \frac{9^k c_2}{N} \tag{3.18}$$

so that (3.17) follows.

Since $(g^\perp * g)_I = 0$ if $|I| > 0$, we have by (3.4), (3.12), and (3.14) that

$$v_I^N = (g^\perp * P(t)(\tilde{f}^N - g))_I + R_I^N(t) := (g^\perp * P(t)\psi)_I + R_I^N(t) \tag{3.19}$$

where, by (3.6) we have put

$$\psi_I = \sum_{\substack{S\subset I \\ |S|>0}} v_S^N g_{I\setminus S}, \qquad \psi_\emptyset = 0. \tag{3.20}$$

We prove the following result.

LEMMA 3.2. *Let η be a positive real number and choose $T(\eta)$ as in Lemma* 3.1. *Then, for any integer $k > 0$ and $t > T(\eta)$, the following estimate holds:*

$$\|(P(t)\psi)_k - S_k(t)\psi_k\|_{L_1} \leq \frac{\delta_k}{1 - 2a\lambda t(e^\alpha + 1)} \|v^N\|_\alpha, \tag{3.21}$$

with $\delta_k = 2^k(1 + k2^{k-1}e^\alpha)(\frac{\eta}{2} + 2a\lambda t(e^\alpha + 1))$, provided that λ and η are so small to satisfy $e^\alpha[\frac{\eta}{2} + 2a\lambda t(e^\alpha + 1)] < 1$.

Proof. To prove the lemma, we write the expansion already introduced in (2.34)–(2.35). More precisely,

$$(P(t)\psi)_k = \sum_{n\geq 0} \lambda^n \int_0^t \int_0^{t_1} \cdots \int_0^{t_{n-1}} dt_1 \dots dt_n \sum_{i_1=1}^{k} \sum_{i_2=1}^{k+1} \cdots \sum_{i_n=1}^{k+n-1}$$

$$\times S_k(t - t_1) C_{k,k+1}^{i_1}, \dots, S_{k+n-1}(t_{n-1} - t_n) C_{k+n-1,k+n}^{i_n} S_{k+n}(t_n)\psi_{k+n}, \tag{3.22}$$

where

$$\sum_{i_r=1}^{k+r-1} C^{i_r}_{k+r-1,k+r} = C_{k+r-1,k+r}, \qquad r = 1, 2, \dots, n,$$

that is, $C^{i_r}_{k+r-1,k+r}$ is the contribution due to the collision of the i_rth particle (among the $k+r-1$ particles) with the $k+r$th. Let us indicate by I_r the set of indices $\{1, 2, \dots, r\}$ and by $I(k,n)$ the set $I_{k+n} \setminus I_k \equiv \{k+1, \dots, n\}$. Then by the definition of ψ it is

$$\psi_{k+n} = \sum_{S_1 \subseteq I_k} \sum_{S_2 \subseteq I(k,n)} v^N_{S_1 \cup S_2} g_{I_{k+n} \setminus (S_1 \cup S_2)} \chi(|S_1| + |S_2| > 0). \tag{3.23}$$

We now select among the particles in S_1 those which do not interact with any other particle. To this end, we consider the set $J = S_1 \setminus \{i_1, \dots, i_n\}$ and notice that

$$\sum_{i_1,\dots,i_n} \sum_{S_1 \subseteq I_k} \sum_{S_2 \subseteq I(k,n)} = \sum_{S_1 \subseteq I_k} \sum_{S_2 \subseteq I(k,n)} \sum_{J \subseteq S_1} \sum_{i_1,\dots,i_n} \prod_{r=1}^{n} \chi(i_r \notin J)\,. \tag{3.24}$$

Defining $n(s_1, j) = \max(s_1 - j, 1)$, (3.22) and (3.23) imply

$$(P(t)\psi)_k = S_k(t)\psi_k + \sum_{s_1=0}^{k} \sum_{\substack{S_1 \subseteq I_k \\ |S_1| = s_1}} \sum_{j=0}^{s_1} \sum_{\substack{J \subseteq S_1 \\ |J| = j}} \sum_{n > n(s_1,j)} \sum_{i_1,\dots,i_n} \lambda^n \tag{3.25}$$

$$\times \prod_{r=1}^{n} \chi(i_r \notin J) \sum_{s_2=0}^{n} \sum_{\substack{S_2 \subseteq I(n,k) \\ |S_2| = s_2}} \chi(s_1 + s_2 > 0) \int_0^t \int_0^{t_1} \cdots \int_0^{t_{n-1}} dt_1 \dots dt_n$$

$$\times S_{I_k \setminus J}(t - t_1) C^{i_1}_{k-j,k-j+1} S_{I_k \setminus J \cup \{i_1\}}(t_1 - t_2) \dots C^{i_n}_{k-j+n-1,k-j+n}$$

$$\times S_J(t) v^N_{S_1 \cup S_2} g_{I_{n+k} \setminus S_1 \cup S_2}.$$

Here we are using the notation $S_A(t) = \prod_{i \in A} S_{\{i\}}(t)$ and hence, $S_A(t)$ represents the Knudsen semigroup associated with the free motion of the particles with labels in A.

Formula (3.25) follows from the fact that we have selected J as the set of particles not interacting with the rest, and hence $S_J(t)$ commutes with all other operators.

By Lemma 3.1, for $\eta > 0$ and $t > T(\eta)$ we have

$$\|S_J(t) v^N_{S_1 \cup S_2}\|_{L_1} \le \eta^j \|v^N_{S_1 \cup S_2}\|_{L_1},$$

where $j = |J|$. Moreover,

$$\|g_{I_{n+k} \setminus S_1 \cup S_2}\|_{L_1} = 1.$$

Thus, using the equality

$$\sum_{i_1=1}^{k} \sum_{i_2=1}^{k+1} \cdots \sum_{i_n=1}^{k+n-1} \prod_{r=1}^{n} \chi(i_r \notin J) = \frac{(k-j+n-1)!}{(k-j-1)!},$$

by (2.37) we arrive at the formula

$$\|(P(t)\psi)_k - S_k(t)\psi_k\|_{L_1} \le \|v^N\|_\alpha \sum_{s_1=0}^{k} \binom{k}{s_1} \sum_{j=0}^{s_1} \binom{s_1}{j} \eta^j \sum_{n>n(s_1,j)} \sum_{s_2=0}^{n} \binom{n}{s_2}$$

$$\times \frac{(2a\lambda t)^n}{n!} \frac{(k-j+n-1)!}{(k-j-1)!} e^{\alpha(s_1+s_2)} \chi(s_1+s_2>0)\,. \tag{3.26}$$

Now we separate from the rest the term corresponding to $s_1 = 0$ and obtain

(3.27)

$$\|(P(t)\psi)_k - S_k(t)\psi_k\|_{L_1} \le \|v^N\|_\alpha 2^k \sum_{n\ge 1} (2a\lambda t)^n \sum_{s_2=1}^{n} \binom{n}{s_2} e^{\alpha s_2}$$
$$+ \|v^N\|_\alpha \sum_{s_1=1}^{k} \binom{k}{s_1} e^{\alpha s_1} \sum_{j=0}^{s_1} \binom{s_1}{j} \eta^j 2^{k-j} \sum_{n>n(s_1,j)} (2a\lambda t)^n \sum_{s_2=0}^{n} \binom{n}{s_2} e^{\alpha s_2}.$$

Since

$$\sum_{h=0}^{n} \binom{n}{h} a^h = (1+a)^n,$$

it follows that

(3.28)

$$\|(P(t)\psi)_k - S_k(t)\psi_k\|_{L_1} \le \|v^N\|_\alpha 2^k \sum_{n\ge 1} (2a\lambda t)^n (1+e^\alpha)^n$$
$$+ \|v^N\|_\alpha 2^k \sum_{s_1=1}^{k} \binom{k}{s_1} e^{\alpha s_1} \sum_{j=0}^{s_1} \binom{s_1}{j} \left(\frac{\eta}{2}\right)^j \sum_{n>n(s_1,j)} (2a\lambda t)^n (1+e^\alpha)^n.$$

By the hypothesis on λ, $2a\lambda t(1+e^\alpha) < 1$ so that we have

(3.29)

$$\|(P(t)\psi)_k - S_k(t)\psi_k\|_{L_1} \le \|v^N\|_\alpha 2^k \frac{2a\lambda t(1+e^\alpha)}{1-2a\lambda t(1+e^\alpha)}$$
$$+ \|v^N\|_\alpha \frac{2^k}{1-2a\lambda t(1+e^\alpha)} \sum_{s_1=1}^{k} \binom{k}{s_1} e^{\alpha s_1} \sum_{j=0}^{s_1} \binom{s_1}{j} \left(\frac{\eta}{2}\right)^j [2a\lambda t(1+e^\alpha)]^{s_1-j}.$$

After a few simple calculations, we arrive at

(3.30)

$$\|(P(t)\psi)_k - S_k(t)\psi_k\|_{L_1} \le \|v^N\|_\alpha \frac{2^k}{1-2a\lambda t(1+e^\alpha)}$$
$$\times \left\{ 2a\lambda t(1+e^\alpha) + \left[1 + e^\alpha \left(\frac{\eta}{2} + 2a\lambda t(1+e^\alpha)\right)\right]^k - 1 \right\}.$$

Using the elementary inequality

$$(1+x)^k - 1 \le k2^{k-1}x, \qquad x \in [0,1], \tag{3.31}$$

we obtain

$$\le \|v^N\|_\alpha \frac{2^k}{1-2a\lambda t(1+e^\alpha)} \left[2a\lambda t(1+e^\alpha) + k2^{k-1}e^\alpha \left(\frac{\eta}{2} + 2a\lambda t(1+e^\alpha)\right)\right] \tag{3.30}$$

$$\le \|v^N\|_\alpha \frac{2^k(1+k2^{k-1}e^\alpha)}{1-2a\lambda t(1+e^\alpha)} \left[\frac{\eta}{2} + 2a\lambda t(1+e^\alpha)\right], \tag{3.32}$$

and the lemma is proven. □

We have by Lemma 3.1 and (2.36), recalling the definition (3.20) of ψ,

$$\|S_k(t)\psi_k\|_{L_1} \le \sum_{j=1}^k \sum_{\substack{J\subseteq I_k \\ |J|=j}} \|S_k(t)v_J^N g_{I_k\setminus J}\|_{L_1}$$

$$\le \sum_{j=1}^k \sum_{\substack{J\subseteq I_k \\ |J|=j}} \eta^j \|v_J^N\|_{L_1} \le \|v^N\|_\alpha \sum_{j=1}^k \binom{k}{j} \eta^j e^{\alpha j}$$

$$= \|v^N\|_\alpha [(1+e^\alpha\eta)^k - 1] \le \|v^N\|_\alpha k2^{k-1}e^\alpha\eta \tag{3.33}$$

for $e^\alpha\eta \le 1$.

From Lemma 3.2 and (3.33) it finally follows that

$$\|(P(t)\psi)_k\|_{L_1} \le 2\frac{\delta_k}{1-2a\lambda t(1+e^\alpha)} \|v^N\|_\alpha . \tag{3.34}$$

Now the proof of the theorem is nearly complete. The estimate (3.34), together with the fact that $(P(t)\psi)_\emptyset = 0$, imply for $2a\lambda t(1+e^\alpha) < \frac{1}{2}$ that

$$\|(g^\perp * P(t)\psi)_k\|_{L_1} \le 4\|v^N\|_\alpha \sum_{j=1}^k \binom{k}{j} \delta_j$$

$$\le 4\|v^N\|_\alpha \left(\frac{\eta}{2} + 2a\lambda t(e^\alpha+1)\right) \sum_{j=1}^k \binom{k}{j} 2^j(1+j2^{j-1}e^\alpha)$$

$$\le 4\|v^N\|_\alpha(\eta + 2a\lambda t(e^\alpha+1))2e^\alpha \sum_{j=0}^k \binom{k}{j} 2^{3j}$$

$$\le 8\|v^N\|_\alpha(\eta + 2a\lambda t(e^\alpha+1))e^\alpha 9^k. \tag{3.35}$$

Thus

$$\|(g^\perp * P(t)\psi)_k\|_{L_1} e^{-\alpha k} \le 8\|v^N\|_\alpha(\eta + 2a\lambda t(e^\alpha+1))e^\alpha e^{(\log 9-\alpha)k}. \tag{3.36}$$

Now we can fix the parameters λ, T, η. We recall that $\alpha \geq 2\log 3$, and choose $\eta \leq \frac{1}{32e^{\alpha}}$. Consequently we fix $t = T(\eta)$ as in Lemma 3.1. Finally we choose λ in such a way that $e^{\alpha} 2a\lambda t(e^{\alpha}+1) \leq \frac{1}{32}$. Then we have

$$\|g^{\perp} * P(t)\psi\|_{\alpha} \leq \frac{1}{2}\|v^N\|_{\alpha} \tag{3.37}$$

and, by (3.19) and (3.37),

$$\|v^N\|_{\alpha} \leq 2\|R^N(t)\|_{\alpha}$$

so that (3.17) concludes the proof. □

Appendix.

Proof of Theorem 2.1. Consider $S(t)$ to be the Knudsen flow and $P_t(x', v'; x, v)$ to be the transition probability densities given by

$$\int P_t(x', v'; x, v) f(x', v') dx' dv' = S(t) f(x, v)$$

for $f \in L_1(\Omega \times V)$. For any final state (x, v), trace the backward trajectories $x - s\, v$ up to the instant (say t) of the collision with the boundary. Denote $y = y(x, v) = x - v\, t \in \partial\Omega$ as the point of the collision.

We introduce the set

$$\mathcal{M}(\beta) = \{(x, v) \in \Omega \times V : |v \cdot n(y)|\, M(y, v) \geq \beta\}, \qquad \beta > 0. \tag{A.1}$$

Note that for all $(x, v) \in \mathcal{M}(\beta)$, the transition $(y, v) \to (x, v)$ is performed with positive probability density (cf. (2.8)). Let

$$t_0 = \frac{4\, d_{\max}}{u_{\min}}, \tag{A.2}$$

where $u_{\min} > 0$ denotes the modulus of the smallest velocity, while $d_{\max}$ is the diameter of Ω.

We first show that

$$\inf_{x', v'} \inf_{(x,v) \in \mathcal{M}(\beta)} P_{t_0}(x', v'; x, v) \geq \gamma > 0. \tag{A.3}$$

Tracing the forward trajectory $x' + s\, v'$ up to the instant (say t') of the collision with the boundary, we denote by $y' = y'(x', v') = x' + v'\, t'$ the hitting point. So we still have to connect the points y' and y by some trajectory within the remaining time $t_0 - t - t'$. These trajectories should be such that both the upper bound for their length and the positive lower bound for their probability density are uniform in y, y'.

Assumption A (concerning Ω). For all $y, y' \in \partial\Omega$, there exists $y_1 \in \partial\Omega$ such that

$$\min(\|y_1 - y\|, \|y_1 - y'\|) \geq \varepsilon \tag{A.4}$$

and

$$|e(y', y_1) \cdot n(y')| \geq \varepsilon, \quad |e(y_1, y) \cdot n(y_1)| \geq \varepsilon, \quad |e(y, y_1) \cdot n(y)| \geq \varepsilon, \tag{A.5}$$

where $e(y, y_1) = \frac{y_1 - y}{\|y_1 - y\|}$, and $\varepsilon > 0$ does not depend on y, y', y_1.

Remark. This assumption is fulfilled if Ω has a smooth boundary, but it is also fulfilled in other cases. Note that, for simplicity, we assume Ω to be convex.

Assumption B (concerning M). There exists $u_0 > 0$ such that

$$M(x,v) \geq M_{\min} > 0, \quad \text{if} \quad u_0 \leq \|v\| \leq 4\,u_0. \tag{A.6}$$

Remark. This assumption is fulfilled if, e.g., M is a Maxwellian.

We first go from y' to $y_1 = y_1(y,y')$ provided by Assumption A, with a velocity satisfying (A.6), and then spend the rest of the time travelling between y_1 and y. The remaining time τ satisfies

$$\frac{d_{\max}}{u_{\min}} \leq \tau \leq \frac{4\,d_{\max}}{u_{\min}}, \tag{A.7}$$

since all flight times within Ω are bounded by $\frac{d_{\max}}{u_{\min}}$.

Note that, according to (A.7), one step from y_1 to y is not enough. Therefore, we choose some flight time $t^* \in (t_{\min}, t_{\max})$, where

$$t_{\min} = t_{\min}(y_1,y) = \frac{\|y-y_1\|}{4\,u_0}, \qquad t_{\max} = t_{\max}(y_1,y) = \frac{\|y-y_1\|}{u_0}, \tag{A.8}$$

and u_0 is provided by Assumption B. The velocities $v^* = \frac{y-y_1}{t^*}$ and $-v^*$ allow us to go both directions with probability density uniformly bounded from below. Indeed, according to (A.5) and (A.6), we have (cf. (2.8))

$$|v^* \cdot n(y_1)|\, M(y_1,v^*) = \|v^*\|\, |\frac{v^*}{\|v^*\|} \cdot n(y_1)|\, M(y_1,v^*) \geq u_0\,\varepsilon\, M_{\min}.$$

The same is true for y and $-v^*$.

We go in double steps before making the last one. Thus, we have to solve the equation

$$2\,l\,t^* + s = \tau, \tag{A.9}$$

for each τ satisfying (A.7), with respect to $s \in (t_{\min}, t_{\max})$ and $l = 0,1,\dots$. To this end, the intervals

$$(2\,l\,t^* + t_{\min}, 2\,l\,t^* + t_{\max}), \qquad l = 0,1,\dots, \tag{A.10}$$

should cover the interval given by (A.7). This is fulfilled if the intervals (A.10) overlap, i.e., one needs the condition

$$2\,l\,t^* + t_{\max} > 2\,(l+1)\,t^* + t_{\min} \qquad l = 0,1,\dots,$$

or $2\,(t^* - t_{\min}) < t_{\max} - 3\,t_{\min} = t_{\min}$, which is fulfilled provided that

$$t_{\min} < t^* < \frac{3}{2}\,t_{\min}.$$

Using (A.9), (A.7), and (A.8), the number l is estimated from above uniformly in y, y_1,

$$l \leq \frac{\tau}{2\,t^*} \leq \frac{8\,d_{\max}\,u_0}{u_{\min}\,\varepsilon},$$

where ε is from (A.4).

Now let $z = (x, v)$, $z' = (x', v')$ be states of the system and t_0 be defined in (A.2). We denote

$$P(z', z) = P_{t_0}(x', v'; x, v), \qquad Sf(z) = \int P(z', z)\, f(z')\, dz' = S(t_0) f(z), \tag{A.11}$$

$$\pi(z', Z';\, z) = \min\Big(P(z', z), P(Z', z)\Big) \tag{A.12}$$

and

$$\tilde{\pi}(z', Z') = \int \pi(z', Z';\, z)\, dz\,. \tag{A.13}$$

Note that $\pi \geq 0$ and $0 \leq \tilde{\pi}(z', Z') \leq 1$.

Define a transition probability

$$\hat{P}(z', Z';\, z, Z) \tag{A.14}$$

$$= \pi(z', Z';\, z)\, \delta(z - Z) + \frac{[P(z', z) - \pi(z', Z';\, z)]\, [P(Z', Z) - \pi(z', Z';\, Z)]}{1 - \tilde{\pi}(z', Z')}$$

and the operator

$$\hat{S}\, \Phi(z; Z) = \int\int \hat{P}(z', Z'; z, Z)\, \Phi(z', Z')\, dz'\, dZ'. \tag{A.15}$$

Note that

$$\int\int \hat{P}(z', Z';\, z, Z)\, dz\, dZ = \tilde{\pi}(z', Z') + \frac{[1 - \tilde{\pi}(z', Z')]\, [1 - \tilde{\pi}(z', Z')]}{1 - \tilde{\pi}(z', Z')} = 1.$$

The corresponding Markov chain behaves as follows.

With probability $\tilde{\pi}(z', Z')$, there is a transition from (z', Z') into (z, Z) according to

$$\frac{\pi(z', Z';\, z)}{\tilde{\pi}(z', Z')}\, \delta(z - Z),$$

i.e., z is distributed according to $\frac{\pi(z', Z';\, z)}{\tilde{\pi}(z', Z')}$ and $Z = z$.

With probability $1 - \tilde{\pi}(z', Z')$, there is a transition from (z', Z') into (z, Z) according to

$$\frac{[P(z', z) - \pi(z', Z';\, z)]\, [P(Z', Z) - \pi(z', Z';\, Z)]}{[1 - \tilde{\pi}(z', Z')]^2},$$

i.e., z, Z are independent and distributed according to $\frac{P(z', z) - \pi(z', Z';\, z)}{1 - \tilde{\pi}(z', Z')}$ and $\frac{P(Z', Z) - \pi(z', Z';\, Z)}{1 - \tilde{\pi}(z', Z')}$, respectively.

Note that if $z' = Z'$ then, according to (A.12) and (A.13), we have $\pi(z', Z';\, z) = P(z', z)$ and $\tilde{\pi}(z', Z') = 1$. Thus, z is distributed according to $P(z', z)$ and $Z = z$ so that the particles remain together.

LEMMA A.1. *Let R be a* **joint representation** *of f and g, i.e.,*

$$\int R(z,Z)\,dZ = f(z) \quad \textit{and} \quad \int R(z,Z)\,dz = g(Z),$$

and let $S, \hat{S}$ be defined in (A.11) *and* (A.15). *Then, for any $n \geq 1$,*

$$\|S^n f - S^n g\|_{L_1} \leq 2 \int\int \rho(z,Z)\,\hat{S}^n R(z,Z)\,dz\,dZ,$$

where ρ is the discrete distance, i.e., $\rho(z,Z) = 1$ if $z \neq Z$ and $\rho(z,Z) = 0$ if $Z = z$.

Proof. We first show that $\hat{S}^n R$ is a joint representation of $S^n f$ and $S^n g$, for any $n \geq 1$. From (A.15) and (A.14) we obtain

$$\int \hat{S}\,R(z,Z)\,dz$$

$$= \int\int \left[\pi(z',Z';\,Z) + \frac{[1-\tilde{\pi}(z',Z')]\,[P(Z',Z) - \pi(z',Z';\,Z)]}{1-\tilde{\pi}(z',Z')}\right] R(z',Z')\,dz'\,dZ'$$

$$= \int\int P(Z',Z)\,R(z',Z')\,dz'\,dZ' = \int P(Z',Z)\,g(Z')\,dZ' = S\,g(Z).$$

Analogously one shows that $\int \hat{S}\,R(z,Z)\,dZ = S\,f(z)$, so that $\hat{S}R$ is indeed a joint representation of Sf and Sg. The general case is established by induction.

The case $n = 1$ is now obtained from the estimate

$$\|S\,f - S\,g\|_{L_1}$$

$$= \int \left|\int \hat{S}R(z,Z)\,dZ - \int \hat{S}R(Z,z)\,dZ\right| dz \leq \int\int |\hat{S}R(z,Z) - \hat{S}R(Z,z)|\,dZ\,dz$$

$$= \int\int \rho(z,Z)\,|\hat{S}R(z,Z) - \hat{S}R(Z,z)|\,dZ\,dz \leq 2\int\int \rho(z,Z)\,\hat{S}R(z,Z)\,dZ\,dz.$$

In the general case we have

$$\|S^n f - S^n g\|_{L_1}$$

$$= \|S(S^{n-1}f) - S(S^{n-1}g)\|_{L_1} \leq 2\int\int \rho(z,Z)\,\hat{S}(\hat{S}^{n-1}R)(z,Z)\,dZ\,dz,$$

which completes the proof. □

LEMMA A.2. *Let $\hat{S}$ be defined in* (A.15) *and let R be a probability density function. Then there exists $\varepsilon > 0$ such that*

$$\int\int \rho(z,Z)\hat{S}^n R(z,Z)\,dz\,dZ \leq (1-\varepsilon)^n, \quad \forall n \geq 1.$$

Proof. According to (A.11), (A.12), and (A.13), we have

$$\tilde{\pi}(z', Z') = \inf_{x', v'; x, v} \int\int \min\Big(P_{t_0}(x', v'; y, w), P_{t_0}(x, v; y, w)\Big)\, dy\, dw\,.$$

Thus, using (A.3) we obtain

$$\tilde{\pi}(z', Z')$$

$$\geq \int\int \inf_{x', v'} P_{t_0}(x', v'; y, w)\, dy\, dw \geq \int\int \chi_{\mathcal{M}(\beta)}(y, w) \inf_{x', v'} P_{t_0}(x', v'; y, w)\, dy\, dw$$

$$\geq \inf_{x', v'} \inf_{(y,w)\in\mathcal{M}(\beta)} P_{t_0}(x', v'; y, w)\, \lambda(\mathcal{M}(\beta)) \geq \gamma\, \lambda(\mathcal{M}(\beta)),$$

where the symbol χ denotes the indicator function. Consequently, the estimate

$$1 \geq \tilde{\pi}(z', Z') \geq \varepsilon > 0 \tag{A.16}$$

holds, since the set $\mathcal{M}(\beta)$, defined in (A.1), has positive Lebesgue measure.

Consider a Markov chain (ξ_n, ζ_n) with the transition function (A.14). Let the random variables ξ_0, ζ_0 have the joint distribution R. Then the random variables ξ_n, ζ_n have the joint distribution $\hat{S}^n R$, and

$$\int\int \rho(z, Z)\, \hat{S}^n R(z, Z)\, dz\, dZ = P(\xi_n \neq \zeta_n), \tag{A.17}$$

where P denotes the probability. We obtain

$$P(\xi_n \neq \zeta_n) = P(\xi_n \neq \zeta_n \,|\, \xi_{n-1} = \zeta_{n-1})\, P(\xi_{n-1} = \zeta_{n-1})$$

$$+ \int\int \rho(z, Z)\, P(\xi_n \neq \zeta_n \,|\, \xi_{n-1} = z\,,\, \zeta_{n-1} = Z)\, \hat{S}^{n-1} R(z, Z)\, dz\, dZ$$

$$\leq \int\int \rho(z, Z)\, [1 - \tilde{\pi}(z, Z)]\, \hat{S}^{n-1} R(z, Z)\, dz\, dZ\,,$$

according to (A.14). Using (A.16) and (A.17), we get

$$P(\xi_n \neq \zeta_n) \leq (1 - \varepsilon) \int\int \rho(z, Z)\, \hat{S}^{n-1} R(z, Z)\, dz\, dZ = (1 - \varepsilon)\, P(\xi_{n-1} \neq \zeta_{n-1}),$$

and the result follows by iteration. ∎

From Lemmas A.1 and A.2 we get

$$\|S^n f - S^n g\|_{L_1} \leq 2\,(1 - \varepsilon)^n, \tag{A.18}$$

and, consequently,

$$\|S^{n+m} f - S^n f\|_{L_1} = \|S^n S^m f - S^n f\|_{L_1} \leq 2\,(1 - \varepsilon)^n.$$

Because of the completeness of L_1, this implies existence of some $\bar{g} = \lim_{n\to\infty} S^n f$, for which we have $S\bar{g} = \bar{g}$. Moreover, uniqueness of such $\bar{g}$ follows from (A.18).

Note that $\lim_{n\to\infty} f_n = \bar{g}, f_n \geq 0 \|f_n\|_{L_1} = 1$ imply $\bar{g} \geq 0$ and $\|\bar{g}\|_{L_1} = 1$, since $|\int \bar{g} - 1| \leq \int |\bar{g} - f_n|$.

Finally, from (A.18) we have, for $t \in [t_0\, n, t_0\, (n+1))$,

$$\|S(t)f - \bar{g}\|_{L_1} \leq \|S^n S(t - t_0\, n)f - S^n f\|_{L_1} + \|S^n f - \bar{g}\|_{L_1} \leq 4\,(1-\varepsilon)^n \leq 4\,(1-\varepsilon)^{\frac{t}{t_0}},$$

and (2.15) follows. Moreover, it follows from (A.18) that

$$\|S(t)\bar{g} - \bar{g}\|_{L_1} = \|S(t)S^n\bar{g} - S^n\bar{g}\|_{L_1} \leq 2\,(1-\varepsilon)^n\,, \quad \forall n,$$

which implies (2.14). □

Proof of Theorem 2.2. Let $p = p(x,v,t)$ and $l = l(x,v,t)$ be two solutions of the initial boundary value problem (2.1)–(2.3), with initial conditions p_0 and l_0, respectively. Writing the evolution equation in mild form (2.10), we have

$$p(t) - l(t) = S(t)(p_0 - l_0) + \lambda \int_0^t ds\, S(t-s)Q(p(s) + l(s), p(s) - l(s)),$$

where $Q(f,g)$ is the symmetrized collision operator (2.4).

By (3.9) we have that, if $h = h(x,v)$ has the property $\int h = 0$, then

$$\|S(t)h\|_{L_1} \leq e^{-bt}\|h\|_{L_1}.$$

Since $\int Q(f,g) = 0$ for any pair of functions f and g, we have (cf. (2.38))

$$\|p(t) - l(t)\|_{L_1} \leq e^{-bt}\|p_0 - l_0\|_{L_1} + 2a\lambda \int_0^t ds\, e^{-b(t-s)}\|p(s) - l(s)\|_{L_1},$$

so that, using the Gronwall lemma,

$$\|p(t) - l(t)\|_{L_1} \leq \|p_0 - l_0\|_{L_1} e^{-(b-2a\lambda)t}.$$

In particular, we obtain

$$\|p(t+\tau) - p(t)\|_{L_1} \leq \|p(\tau) - p_0\|_{L_1} e^{-(b-2a\lambda)t} \leq 2\, e^{-(b-2a\lambda)t}\,, \quad \forall \tau \geq 0.$$

Therefore, if $\lambda < \frac{b}{2a}$, there exists a probability density g which is the unique global attracting point for the flow described by (2.1) and also the unique invariant solution for such an evolution problem. □

Remark. It is not hard to show that g solves the stationary equation $(G^{free})^* g + \lambda Q(g,g) = 0$ and also solves the boundary value problem (2.13). In particular, the trace of g on the boundary does exist. These considerations are not relevant for the present analysis, so we do not go further.

Proof of Theorem 2.3. This theorem is easily proved using the same arguments as for the Knudsen flow. Indeed, it is enough to observe that for a fixed time t, the probability of each particle of the system to perform a collisionless motion is strictly positive. We remark that in [GLP] the same result has been obtained in a more difficult context.

REFERENCES

[AN] L. Arkeryd and A. Nouri, *A compactness result related to the stationary Boltzmann equation in a slab, with applications to the existence theory*, Indiana Univ. Math. J., 44 (1995), pp. 815–839.

[B] H. Babovsky, *Time averages of simulation schemes as approximations to stationary kinetic equations*, European J. Mech. B Fluids, 11 (1992), pp. 199–212.

[BI] H. Babovsky and R. Illner, *A convergence proof for Nanbu's simulation method for the full Boltzmann equation*, SIAM J. Numer. Anal., 26 (1989), pp. 45–65.

[BS] A. V. Bobylev and J. Struckmeier, *Numerical simulation of the stationary one-dimensional Boltzmann equation by particle methods*, European J. Mech. B Fluids, 15 (1996), pp. 103–118.

[C] C. Cercignani, *The Grad limit for a system of soft spheres*, Comm. Pure Appl. Math., 36 (1983), pp. 479–494.

[CDPP] S. Caprino, A. De Masi, E. Presutti, and M. Pulvirenti, *A derivation of the Broadwell equation*, Comm. Math. Phys., 135 (1991), pp. 443–465.

[CIP] C. Cercignani, R. Illner, and M. Pulvirenti, *The Mathematical Theory of Dilute Gases*, Appl. Math. Sci. 106, Springer-Verlag, New York, 1994.

[CP] S. Caprino and M. Pulvirenti, *The Boltzmann-Grad limit for a one-dimensional Boltzmann equation in a stationary state*, Comm. Math. Phys., 177 (1996), pp. 63–81.

[DP] A. DeMasi and E. Presutti , *Mathematical Methods for Hydrodynamical Limits*, Lectures Notes in Math. 1501, Springer-Verlag, Berlin, 1991.

[GLP] S. Goldstein, J. L. Lebowitz, and E. Presutti, *Mechanical systems with stochastic boundaries*, in Random Fields, Colloq. Math. Soc. János Bolyai 27, North–Holland, Amsterdam, 1981, pp. 403–419.

[GM] C. Graham and S. Méléard, *Convergence rate on path space for stochastic particle approximations to the Boltzmann equation*, Z. Angew. Math. Mech., 76 (1996), Suppl. 1, pp. 291–294.

[L] O. Lanford III, *Time evolution of large classical systems*, in Dynamical Systems, Theory and Applications, Lecture Notes in Phys. 38, Springer-Verlag, Berlin, 1975, pp. 1–111.

[LP] M. Lachowicz and M. Pulvirenti, *A stochastic system of particles modelling the Euler equation*, Arch. Rational Mech. Anal., 109 (1990), pp. 81–93.

[M] S. Méléard, *Asymptotic behaviour of some interacting particle systems: McKean–Vlasov and Boltzmann models* in Probabilistic Models for Nonlinear Partial Differential Equations, Lecture Notes in Math. 1627, D. Talay and L. Tubaro, eds., Springer-Verlag, Berlin, 1996, pp. 42–95.

[P] M. Pulvirenti, *Kinetic limits for stochastic particle systems*, in Probabilistic Models for Nonlinear Partial Differential Equations, Lecture Notes in Math. 1627, D. Talay and L. Tubaro, eds., Springer-Verlag, Berlin, 1996, pp. 96–126.

[PWZ] M. Pulvirenti, W. Wagner, and M. B. Zavelani Rossi, *Convergence of particle schemes for the Boltzmann equation*, European J. Mech. B Fluids, 13 (1994), pp. 339–351.

[W] W. Wagner, *A convergence proof for Bird's direct simulation Monte Carlo method for the Boltzmann equation*, J. Statist. Phys., 66 (1992), pp. 1011–1044.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 935–966, July 1998

008

ISOSPECTRAL SETS FOR FOURTH-ORDER ORDINARY DIFFERENTIAL OPERATORS*

LESTER F. CAUDILL, JR.†, PETER A. PERRY‡, AND ALBERT W. SCHUELLER§

Abstract. Let $L(p)u = D^4u - (p_1u')' + p_2u$ be a fourth-order differential operator acting on $L^2[0,1]$ with $p \equiv (p_1, p_2)$ belonging to $L^2_{\mathbb{R}}[0,1] \times L^2_{\mathbb{R}}[0,1]$ and boundary conditions $u(0) = u''(0) = u(1) = u''(1) = 0$. We study the isospectral set of $L(p)$ when $L(p)$ has simple spectrum. In particular we show that for such p, the isospectral manifold is a real-analytic submanifold of $L^2_{\mathbb{R}}[0,1] \times L^2_{\mathbb{R}}[0,1]$ which has infinite dimension and codimension. A crucial step in the proof is to show that the gradients of the eigenvalues of $L(p)$ with respect to p are linearly independent: we study them as solutions of a non-self-ajdoint fifth-order system, the *Borg system*, among whose eigenvectors are the gradients.

Key words. inverse spectral problem, ordinary differential equations

AMS subject classifications. 34A55, 34L20

PII. S0036141096311198

1. Introduction.

This paper initiates a study of isospectral sets of coefficients for self-adjoint, fourth-order ordinary differential operators, in Liouville–Green normal form, on the finite interval $[0,1]$. Such operators are labelled by a pair of coefficients $p = (p_1, p_2)$. Our motivation is twofold: first, we would like to understand the inverse spectral problem for fourth-order operators such as the Euler–Bernoulli operator of mechanics; second, we would like to develop techniques of analysis which are systematic in nature and are therefore likely to be useful in the study of other singular and higher order ordinary differential operators. Our goal is to understand the set of coefficients isospectral to a given pair $p = (p_1, p_2)$ as a Hilbert submanifold of a suitable Hilbert space of coefficients, in analogy to the analysis of the second-order Sturm–Liouville problem carried out by Trubowitz and his collaborators (see [11, 18, 19, 26, 27], and see [10] for more recent results).

As in the work of Trubowitz et. al., we use methods of global analysis to study the isospectral manifold as a level set of the direct spectral map from coefficients to spectra. For the class of operators we consider, the gradient $g_n(x;p)$ of a given eigenvalue $\lambda_n(p)$ is an ordered pair consisting of an eigenfunction square and the square of its derivative, and so the gradient of the mapping from coefficients to spectra is the infinite sequence of all such ordered pairs. A crucial part of the analysis is to show that these ordered pairs form a linearly independent set.

Our approach differs from the approach to the Sturm–Liouville problem taken in [11, 18, 19, 26, 27] in two respects. First of all, we use resolvent perturbation

*Received by the editors October 28, 1997; accepted for publication (in revised form) July 23, 1997; published electronically March 25, 1998. This research was supported by the National Science Foundation and the Commonwealth of Kentucky under its EPSCoR program. A portion of this work was carried out while the first author was visiting the Department of Mathematics, University of Kentucky.

http://www.siam.org/journals/sima/29-4/31119.html

†Department of Mathematics and Computer Science, University of Richmond, Richmond, VA 23173 (lcaudill@richmond.edu).

‡Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027 (perry@ms.uky.edu).

§Department of Mathematics, Whitman College, Walla Walla, WA 99362 (schuellaw@whitman.edu).

techniques rather than integral equations and complex analysis to obtain the necessary eigenvalue and eigenfunction asymptotics: see the thesis of the third author [29], where these techniques are developed at greater length. Secondly, we study orthogonality properties of the gradients, not using special identities, but rather by studying an associated non-self-adjoint, fifth-order system, the *Borg system*, among whose eigenvectors are exactly the gradients $g_n(x,p)$. Our system is the analogue, for fourth-order differential operators, of a third-order non-self-adjoint eigenvalue problem introduced by Borg [9] in his study of completeness of eigenfunction squares in the Sturm–Liouville problem. We believe this technique to be a powerful one which admits generalization to other inverse spectral problems involving ordinary differential operators.

To describe our results in detail, we first specify the class of fourth-order operators which we will study. In order to study the isospectral set as a Hilbert manifold, we wish to study operators $L(p)$ where p ranges over a Hilbert space of coefficients; here $L(p)$ is the operator

$$L(p)u = D^4u - D(p_1Du) + p_2u \tag{1.1}$$

on $L^2[0,1]$, where $p = (p_1, p_2)$. In what follows, we will impose "double Dirichlet" boundary conditions $u(0) = u''(0) = u(1) = u''(1) = 0$, although our methods can be used to treat other separated, self-adjoint boundary conditions.

A natural choice for the Hilbert space of coefficients is $E \equiv L^2_{\mathbb{R}}[0,1] \times L^2_{\mathbb{R}}[0,1]$, where $L^2_{\mathbb{R}}[0,1]$ denotes real-valued, square-integrable, measurable functions on $[0,1]$. For such singular coefficients it is convenient to define the operator $L(p)$ by the method of sesquilinear forms (see, for example, Kato [17, Chapter 6]). Since we wish to study real analyticity of various maps on E, it will also be convenient to introduce $E_{\mathbb{C}} \equiv L^2_{\mathbb{C}}[0,1] \times L^2_{\mathbb{C}}[0,1]$ and define $L(p)$ for $p \in E_{\mathbb{C}}$. To this end, we introduce the sesquilinear form

$$q(u,v) = \int_0^1 u''(x)\overline{v''(x)} + p_1(x)u'(x)\overline{v'(x)} + p_2(x)u(x)\overline{v(x)}\,dx \tag{1.2}$$

with the form domain

$$\mathcal{Q}(q) = \{u \in H^2[0,1] : u(0) = u(1) = 0\}$$

for $p \in E_{\mathbb{C}}$. It is not difficult to see that the form q with $p = 0$ is a closed positive form. Using this fact and simple perturbative estimates, one can show that the form q with $0 \neq p \in E_{\mathbb{C}}$ is also closed and sectorial, i.e., that the set

$$\{q(u,u) : u \in \mathcal{Q}(q), \|u\|_{L^2[0,1]} = 1\}$$

is contained in a sector of the complex plane of the form $\Re(z) \geq -c$, $|\Im(z)| \leq (\Re(z)+c)$. Here c depends only on $\|p\|_E$; a complete proof is given in [29, section 5.2].

It follows from the form representation theorem (see, for example, Theorem VI.2.1 of [17]) that there is a unique sectorial operator $L(p)$, i.e., a unique closed operator with numerical range in a sector, associated with the sesquilinear form q. It follows from the same theorem that for all $p \in E_{\mathbb{C}}$, the domain of $L(p)$ is contained in the $H^3[0,1]$ functions with $u(0) = u''(0) = u(1) = u''(1) = 0$. Thus $L(p)$ is an operator with compact resolvent, and its spectrum consists of an infinite sequence, $\{\lambda_n(p)\}$, of discrete eigenvalues. Using the form representation theorem, one can also show

that if $p \in C^1([0,1];\mathbb{C}^2)$, then $L(p)$ is the operator (1.1). For more singular p the action of $L(p)$ may be understood in terms of the quasi-derivatives associated with the operator $L(p)$: see Naimark [25] for the general theory and Schueller [29, section 5.3] for its application to fourth-order operators.

We wish to study isospectral sets of $L(p)$ for $p = (p_1, p_2) \in E$. In order to apply techniques of global analysis, we need to realize the direct spectral map from coefficients to spectral data as a map between Hilbert spaces. To this end, we set

$$\mu_0(p) = \overline{p_1} = \int_0^1 p_1(x)\,dx$$

and

$$\mu_n(p) = \frac{\lambda_n(p) - \lambda_n(0) - n^2\pi^2\overline{p_1}}{n^2\pi^2}.$$

We will show that the sequence $\{\mu_n(p)\}_{n=0}^{\infty}$ belongs to the Hilbert space $F \equiv \ell^2(0 \cup \mathbb{N})$. The *direct spectral map* is the mapping $\mu : E \to F$ defined by $\mu(p) = \{\mu_n(p)\}$. The *isospectral set* $M(p)$ of a given $p \in E$ is the set of all $q \in E$ with $\mu(q) = \mu(p)$. We will say that $p \in E$ has *simple spectrum* if the sectorial operator $L(p)$ has only simple eigenvalues. First of all, we will prove the following theorem.

THEOREM 1.1. *The set of $p \in E$ with simple spectrum is open and dense in E.*

Denote this set by $\mathcal{E}$. There are physically relevant families of fourth-order problems, such as the Liouville–Green normal forms of the Euler–Bernoulli equation, which are known to have simple spectrum (see, e.g., [14]). Thus, restricting attention to $p \in \mathcal{E}$ is not unreasonable for many problems of physical interest. Our main result is as follows.

THEOREM 1.2. *For each $p \in \mathcal{E}$, $M(p) \cap \mathcal{E}$ is a real-analytic submanifold of E of infinite dimension and infinite codimension.*

We can quantify the "size" of $M(p)$ more precisely by introducing some auxiliary boundary value problems associated with the formal differential operator $L(p)$. To define these auxiliary boundary value problems, we introduce the closed sesquilinear forms q_1 and q_2 which are given by the expression (1.2) defined on the respective domains

$$\mathcal{Q}_1 = \{u \in H^2[0,1] : u(0) = 0, u'(1) = 0\}$$

and

$$\mathcal{Q}_2 = \{u \in H^2[0,1] : u(0) = 0, u(1) = u'(1) = 0\}.$$

We denote the associated sectorial operators by $L_1(p)$ and $L_2(p)$, and their corresponding eigenvalues by $\sigma_n(p)$ and $\tau_n(p)$, respectively. For $p \in C_0^\infty([0,1];\mathbb{C}^2)$, these operators carry the following boundary conditions:

$$L(p) : u(0) = u''(0) = u(1) = u''(1) = 0, \tag{1.3}$$

$$L_1(p) : u(0) = u''(0) = u'(1) = u'''(1) = 0, \tag{1.4}$$

$$L_2(p) : u(0) = u''(0) = u(1) = u'(1) = 0. \tag{1.5}$$

We conjecture that for $p \in \mathcal{E}$ and q in a dense and open subset of $M(p)$, the three sets of eigenvalues $\{\lambda_n(q)\}, \{\sigma_n(q)\}, \{\tau_n(q)\}$ give local coordinates for $M(p)$. We expect to prove this in a subsequent paper.

Theorem 1.2 involves a study of the differential of the direct spectral map μ. We will first study μ on a dense subset $\mathcal{D}$ of $\mathcal{E}$ consisting of functions $p \in C_0^\infty((0,1);\mathbb{R}^2)$ such that the spectra of each of the three boundary value problems is simple and the intersection of the sets $\{\lambda_n(p)\}$, $\{\sigma_n(p)\}$, and $\{\tau_n(p)\}$ is empty. We show that the set $\mathcal{D}$ is dense in $\mathcal{E}$ in Theorem 3.1. We will also show that eigenvalues and eigenfunctions associated with operators $L(q)$ with $q \in \mathcal{E}$ can be well approximated by those associated with operators $L(p)$ with $p \in \mathcal{D}$.

To show that the isospectral manifold is real-analytic, we wish to apply the real-analytic implicit function theorem (see, for example, [26, p. 154]). Theorem 1.2 follows from Theorem 1.3.

THEOREM 1.3. *The direct spectral map μ is a real-analytic mapping from $\mathcal{E}$ to F. For $p \in \mathcal{E}$ and each $q \in M(p)$, there is an orthogonal decomposition $T_q\mathcal{E} = \mathcal{E}_v(q) \oplus \mathcal{E}_h(q)$ such that $d\mu(p)$ is a linear isomorphism of $\mathcal{E}_v(q)$ onto F and $\mathcal{E}_h(q)$ has infinite dimension.*

We will prove Theorem 1.3 by showing that (1) the map μ is real-analytic as a map from $\mathcal{E}$ into F, (2) the differential $d\mu(q)$ for an arbitrary $q \in \mathcal{E}$ is well approximated by the differential $d\mu(p)$ of a "nearby" $p \in \mathcal{D}$, and (3) the differential $d\mu(p)$ has the required mapping properties for $p \in \mathcal{D}$.

To explain steps (2) and (3) more fully, let $z_n(\cdot; p)$ be the normalized eigenfunction corresponding to eigenvalue $\lambda_n(p)$. (Note that since $p \in \mathcal{E}$, this eigenfunction is unique up to a phase.) Let $\langle \cdot\ ,\ \cdot \rangle_E$ denote the inner product on E. A short calculation shows that for $p \in \mathcal{D}$, the differential $d\mu(p)$ is given by

$$d\mu(p)(v_1, v_2) = \{\langle g_n(\ \cdot\ ; p), v\rangle_E\}_{n=0}^{\infty}\,, \tag{1.6}$$

where the gradients g_n are given by

$$g_0(x; p) = (1, 0)$$

and

$$g_n(x; p) = \left(\frac{z_n'(x;p)^2}{n^2\pi^2} - 1, \frac{z_n(x;p)^2}{n^2\pi^2}\right), \quad n \geq 1\,. \tag{1.7}$$

We wish to take $\mathcal{E}_v$ to be the span of the g_n and $\mathcal{E}_h$ to be its orthogonal complement, the kernel of $d\mu$. If $\zeta = \sum_j c_j g_j$ and c denotes the sequence $\{c_j\}$, then $\|\zeta\|_E^2 = \langle c, A(q)c\rangle$, where A is the operator on F with matrix $\langle g_i, g_j\rangle_E$. If A is a bounded invertible operator, then the g_j form a *Riesz basis* [8] for $\mathcal{E}_v$, and the operator $T : \mathcal{E}_v \to F$ defined by $T\zeta = c$, is boundedly invertible. Moreover, $S(q) = d\mu(q) \circ T^{-1} \in \mathcal{B}(F, F)$ has matrix $A(q)$ and so is a linear isomorphism. In step (2), we show that for any $\epsilon > 0$ and each $q \in \mathcal{E}$, there is a $p \in \mathcal{D}$ such that $\|S(q) - S(p)\|_{\mathcal{B}(F,F)} < \epsilon$. In step (3) we show that $S(p)$ is boundedly invertible for each $p \in \mathcal{D}$ by proving that $A(p)$ has the same property. In order to do so we show that $A(0)$ is boundedly invertible, and, by perturbation estimates and linear independence of the g_n, that the same holds true for $A(p)$ if $p \in \mathcal{D}$. Since the boundedly invertible operators are open in $\mathcal{B}(F, F)$, this shows that $A(q)$, and hence $S(q)$, is boundedly invertible for any $q \in \mathcal{E}$.

The required perturbation estimates on the g_n show that

$$\sum_n \|g_n(\ \cdot\ ; p) - g_n(\ \cdot\ ; 0)\|_E^2$$

is finite. In order to obtain these estimates, we exploit the observation that the functions z_n^2 and $(z_n')^2$ can be recovered from the respective residues of the operators $(L(p)-z)^{-1}$ and $D(L(p)-z)^{-1}D$ at $z=\lambda_n(p)$. We use resolvent perturbation theory to estimate the differences $z_n(\ \cdot\ ;p)^2 - z_n(\ \cdot\ ;0)^2$ and $z_n'(\ \cdot\ ;p)^2 - z_n'(\ \cdot\ ;0)^2$.

To prove that the g_n are linearly independent for any $p \in \mathcal{D}$, we introduce an auxiliary fifth-order differential system, the *Borg system*, satisfied by the $g_n(p)$. An analogous third-order equation was used by Borg [9] in his study of eigenfunction-squares in the Sturm–Liouville problem. The Borg system for a fourth-order operator takes the form

$$\mathcal{M}(p)g = \lambda \mathcal{B}(p)g$$

for matrix-valued differential operators $\mathcal{M}(p)$ of fifth order and $\mathcal{B}(p)$ of third order. For $p \in \mathcal{D}$, the generalized resolvent

$$R(\lambda) = (\mathcal{M}(p) - \lambda \mathcal{B}(p))^{-1}$$

has simple poles, among which are the eigenvalues $\lambda_n(p)$, with corresponding generalized eigenfunctions

$$\widehat{g}_n(x;p) = \left(z_n(x;p)^2, z_n'(x;p)^2\right)$$

The remaining poles are associated with the two auxiliary boundary value problems; since $p \in \mathcal{D}$, these are distinct from the poles $\lambda_n(p)$, and, as we shall see, all of the poles of the generalized resolvent are simple. Using the simplicity of poles, we can then construct a biorthogonal set from the rank-one residues of the generalized resolvent $R(\lambda)$; this proves linear independence of the $\widehat{g}_n(p)$. The linear independence of the gradients $g_n(p)$ is an easy consequence. The residues of the Borg operator furnish tangent vectorfields to the isospectral manifold; we expect, but have not yet proved, that they are a basis for its tangent space.

We note that the eigenvalue equation

$$D^2\left(r(x)D^2y\right) = \mu\rho(x)y$$

for the Euler–Bernoulli beam can be transformed, by means of a Liouville transform, into Liouville–Green normal form for smooth coefficients (see, e.g., [5]). Thus our results apply to Euler–Bernoulli problems with suitable boundary conditions.

A number of results exist in the literature regarding the inverse spectral problem for fourth-order differential operators. Barcilon [1, 2, 3, 4, 5, 6, 7] proved that the density and bending stiffness of an Euler–Bernoulli beam can be recovered from three sets of spectra, showed that fewer than three spectra do not uniquely determine these coefficients, and also proved some general results on inverse spectral problems for differential equations of nth order in Liouville normal form. He also showed that three sequences of eigenvalues corresponding to certain distinct boundary conditions contain the same information as one sequence of eigenvalues together with two sets of norming constants [3]. McLaughlin developed a Gel'fand–Levitan-type reconstruction algorithm for smooth coefficients from one spectrum and two sequences of norming constants [21, 22, 23, 24]. In McLaughlin's papers [20, 21] it is shown that the isospectral set for the operator $L(0)$ with boundary conditions $u(0)=u'(0)=u(1)=u'(1)=0$ is infinite-dimensional with infinite codimension and that the isospectral set for the operator $L(p)$ with the same boundary conditions is also infinite dimensional, with

infinite codimension, provided that the eigenvalues satisfy certain asymptotic forms. Gladwell gave necessary and sufficient conditions on spectral data to produce an Euler–Bernoulli beam with strictly positive (i.e., physical) density and bending stiffness [13] and carried out numerical reconstructions of Euler–Bernoulli beams from finite spectral data [15, 16].

Our results appear to be the first systematic study of the isospectral manifold for fourth-order differential operators. It should be noted that an operator very similar to our "Borg operator" in the constant coefficient case appears in Barcilon's analysis [3] of the inverse spectral problem.

The plan of this paper is as follows. In section 2 we prove some basic results about the spectra of $L(p)$ and the two associated boundary value problems. In section 3 we prove Theorem 1.1. In section 4 we show that the sequence of functions $\{g_n(q)\}$ is stable, in $\ell^2(\mathbb{N}; E)$-sense, under small perturbations of $q \in \mathcal{E}$ and that the map μ has its range in F. In section 5, we prove that the map μ is an analytic mapping from $\mathcal{E}$ into F. In section 6, we introduce and analyze the Borg system and use it to prove linear independence of the vectors $g_n(p)$ for each fixed $p \in \mathcal{D}$. Finally, in section 7 we give the proofs of Theorems 1.2 and 1.3. In Appendix A, we collect some important estimates on the integral kernels of the resolvents, at $p = 0$, of each of the three boundary value problems considered. In Appendix B, we discuss the boundary conditions on the Borg system and prove some technical domain results needed for section 6.

The results in sections 2 and 4 are proved for a number of separated self-adjoint boundary conditions in the Ph.D. thesis of the third author [29].

2. Spectra. In this section we prove some basic results about the spectra of the operators $L(p)$, $L_1(p)$, and $L_2(p)$ for $p \in E$. The symbol $L_\#(p)$ will denote one of the operators $L(p)$, $L_1(p)$, or $L_2(p)$, and $\lambda_n^\#(p)$ will denote the nth eigenvalue of $L_\#(p)$. Similarly, $q_\#$ denotes one of the three sesquilinear forms q, q_1, or q_2.

The spectra of $L(0)$, $L_1(0)$, and $L_2(0)$ are given by explicit transcendental equations (see, for example, [29]). From these, we easily deduce that $\lambda_n(0) = n^4\pi^4$, $\sigma_n(0) = (n + \frac{1}{2})^4\pi^4$, and $|\tau_n(0)^{1/4} - (n + \frac{1}{4})\pi| \leq 4e^{-n\pi}$. We expect the eigenvalues of $L_\#(p)$ for $p \neq 0$ to approach these values asymptotically so that the three sets of spectra "separate" for n large. We will use resolvent perturbation theory to show this is the case.

The following technical lemma will enable us to prove certain resolvent estimates for coefficients $p \in C_0^\infty((0,1), \mathbb{C}^2)$ and extend them by continuity to $p \in E_\mathbb{C}$. Recall that a mapping f from an open subset of a Banach space E into a Banach space F is called *compact* if $f(p_n)$ converges strongly to $f(p)$ in F whenever p_n converges weakly to p in E.

LEMMA 2.1. *For any fixed z with $\Re(z)$ sufficiently negative, the mapping $p \mapsto (L(p) - z)^{-1}$ is a compact mapping from E into the bounded operators on $L^2[0,1]$.*

Proof. Suppose that $p_n \to p$ weakly in E. It is easy to verify that the sesquilinear forms $q_\#^n$ associated with p_n converge to the sesquilinear form $q_\#$ associated with p so that $L_\#(p_n)$ converges to $L_\#(p)$ in the strong resolvent sense (see Kato [17, Theorem VIII.3.6]). It is also easy to check that there is a fixed c, depending only on $\sup \|p_n\|_E$, such that $\Re q_\#^n(u,u) \geq -c + 1$ and that the operators $(L_\#(0) + 1)^{1/2}(L_\#(p_n) + c)^{-1}$ are bounded uniformly in n. Let $R_n = (L_\#(p_n) + c)^{-1} - (L_\#(p) + c)^{-1}$ and let P_N project onto the first N eigenvectors of $L_\#(0)$. We may estimate

$$\|R_n\| \leq \|P_N R_n\| + \|(I - P_N)R_n\|$$

$$\leq \|P_N R_n\| + \|(I - P_N)(L_\#(0) + c)^{-1/2}\| \, \|(L_\#(0) + c)^{1/2} R_n\|.$$

The second right-hand term goes to zero as $N \to \infty$ uniformly in n, and the first right-hand term goes to zero as $n \to \infty$ for each fixed N by the compactness of P_N and the fact that R_n converges strongly to zero. □

Note that the same proof works if z is only required to lie in the common resolvent set of the operators $L(p_n)$ and $L(p)$.

Let $\rho(L_\#(p))$ denote the resolvent set of the operator $L_\#(p)$. The remarks in the proof of Lemma 2.1 show that there is a fixed half-plane $\Re(z) < -c$ so that $(L_\#(p) - z)^{-1}$ exists for any z in this half-plane and any $p \in E_{\mathbb{C}}$ with $\|p\|_{E_{\mathbb{C}}} \leq M$. Thus the set

$$S_M = \cap \{\rho(L_\#(p)) : \|p\|_{E_{\mathbb{C}}} \leq M\}$$

has nonempty interior; we will shortly show that it includes the complement of a countable union of discs whose size depends on M and which are centered at the eigenvalues of $L_\#(0)$. In what follows, denote by $B_M(0)$ the set $\{p \in E : \|p\|_E < M\}$.

LEMMA 2.2. *Fix $M > 0$ and let U be the interior of the set S_M. The mapping $\Psi(z, p) = (L_\#(p) - z)^{-1}$ is a compact analytic mapping from $U \times B_M(0)$ to the bounded operators on $L^2[0, 1]$.*

Proof. Compactness is an immediate consequence of Lemma 2.1 and the first resolvent formula. The resolvent identity

$$(L_\#(p) - z)^{-1} - (L_\#(q) - z)^{-1}$$

$$= (L_\#(q) - z)^{-1}(D(q_1 - p_1)D + (q_2 - p_2))(L_\#(p) - z)^{-1}$$

holds, where $D(q_1 - p_1)D + (q_2 - p_2)$ is understood as a sesquilinear form on the form domain of $L_\#(0)$. This shows that Ψ is norm continuous. A short calculation with difference quotients shows that $(L_\#(p) - z)^{-1}$ is differentiable in the complex sense and that

$$d\Psi_{z,p}(w, h) = w(L_\#(p) - z)^{-2} + (L_\#(p) - z)^{-1}(Dh_1D + h_2)(L_\#(p) - z)^{-1}. \qquad \square$$

For numbers $R > 3$ and $\alpha \in (2, 3)$, we define a region $C^\#_{R,\alpha}$ of $\mathbb{C}$ as follows. Let N be an integer obeying the bounds

$$(8R)^{\frac{1}{3-\alpha}} < N < (16R)^{\frac{1}{3-\alpha}}, \tag{2.1}$$

let

$$D^\#_N = \{z \in \mathbb{C} : |z| < \lambda^\#_N(0) + RN^\alpha\}$$

be a disc containing the first N eigenvalues of $L_\#(0)$, and let

$$E^\#_n = \{z \in \mathbb{C} : |z - \lambda^\#_n(0)| < Rn^\alpha\},$$

a disc containing the nth eigenvalue of $L_\#(0)$. We set

$$C^\#_{R,\alpha} = D_N \cup \left(\cup_{n=N+1}^\infty E_n\right).$$

Thus, the set $C^\#_{R,\alpha}$ is the union of a large disc containing the first N eigenvalues of $L_\#(0)$ and infinitely many small discs each containing exactly one of the remaining eigenvalues (see Figure 2.1). We will show that this region still contains the spectrum of $L_\#(p)$ for R sufficiently large, depending on $\|p\|_E$.

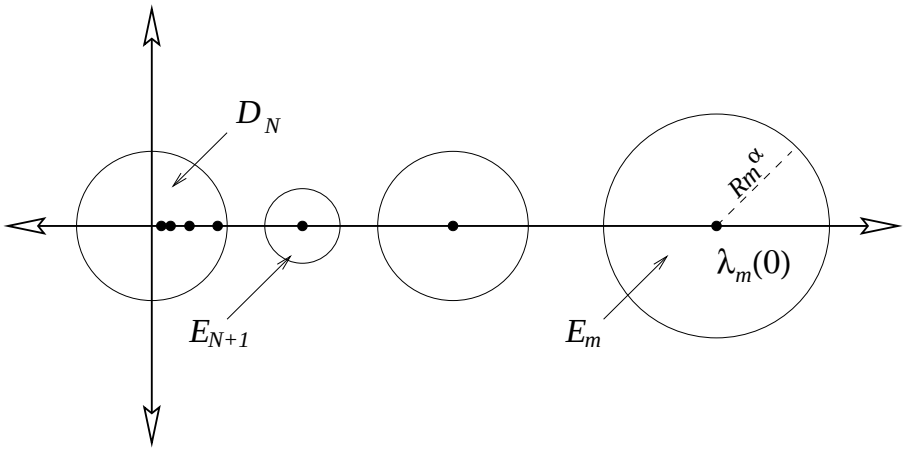

FIG. 2.1. *The set $C_{R,\alpha}$ (not drawn to scale).*

It is easy to check the following purely geometric properties of $C^{\#}_{R,\alpha}$.

LEMMA 2.3.

(i) *Fix one set of boundary conditions. For any $R > 3$, $\alpha \in (2,3)$, and $m > N$, the regions $D^{\#}_N$ and $E^{\#}_m$ are mutually disjoint.*

(ii) *Let $E^{(1)}_n$ and $E^{(2)}_n$ be the regions E_n associated with two distinct sets of boundary conditions. Then $E^{(1)}_n$ and $E^{(2)}_m$ are disjoint for $n \geq N+1$ and $m \geq N+1$.*

Proof. (i) The discs E_m have radii Rm^α, while $\lambda^{\#}_{m+1}(0) - \lambda^{\#}_m(0) \geq m^3$ if $m \geq 2$. Thus the discs will be separate if $2R(m+1)^\alpha < m^3$. This is true if $m > N$. The regions $D^{\#}_N$ and $E^{\#}_{N+1}$ will be disjoint so long as $\lambda^{\#}_{N+1}(0) - R(N+1)^\alpha - (\lambda^{\#}_N(0) + RN^\alpha) > 0$ or $N^3 > RN^\alpha + R(N+1)^\alpha$. This is guaranteed by the choice of N. (ii) Note that, by the mean value theorem, $(x+\frac{1}{4})^4 - x^4 \geq x^3$. Since the fourth roots of any two eigenvalues of the three operators are separated by a distance of at least $1/4$, it follows that the eigenvalue $\lambda^{\#}_m(0)$ associated with any one of the boundary conditions is separated from the closest eigenvalue associated with any of the three boundary conditions by at least $(m-1)^3$. Thus, it suffices to show that $(m-1)^3 > Rm^\alpha + R(m-1)^\alpha$. This will be true if $\left(\frac{m-1}{m}\right)^3 m^{3-\alpha} > R(1 + (\frac{m-1}{m})^\alpha)$. But $m > 1$ and $m^{3-\alpha} > N^{3-\alpha} > 8R$ by the choice of N. ∎

We can now state our rough bounds on the location of $\lambda^{\#}_n(p)$.

THEOREM 2.4. *Fix $M > 0$ and let $p \in E_{\mathbb{C}}$ with $\|p\|_{E_{\mathbb{C}}} < M$. There is a number $R > 3$ depending only on M so that:*

(i) *The spectrum of $L_{\#}(p)$ is contained in $C^{\#}_{R,\alpha}$.*

(ii) *The operators $L_{\#}(p)$ have exactly N eigenvalues in the region D_N.*

(iii) *The eigenvalues of $L_{\#}(p)$ with index $n \geq N+1$ are all simple.*

(iv) *The sets $\{\lambda_n(p)\}^\infty_{n=N+1}$, $\{\sigma_n(p)\}^\infty_{n=N+1}$, and $\{\tau_n(p)\}^\infty_{n=N+1}$ have empty intersection.*

Proof. The operators $L_{\#}(p)$ are sectorial with spectrum contained in a half-plane $\Re(\lambda) > -C(M)$, where $C(M)$ is a positive constant depending only on M [29]. We will construct the resolvents $(L_{\#}(p) - z)^{-1}$ perturbatively from $(L_{\#}(0) - z)^{-1}$ and estimate $\|(L_{\#}(p) - z)^{-1}\|$ uniformly in p with $\|p\|_E < M$ and $z \notin C^{\#}_{R,\alpha}$ for sufficiently large R. This will give (i); (iv) will then follow from Lemma 2.3(ii) and the following argument. Observe that for $p = 0$, the region D_N contains exactly N eigenvalues, and

the regions E_m each contain one eigenvalue. Analyticity of the resolvents as operator-valued analytic functions of $p \in E$ with $\|p\|_E < M$ will imply analyticity of the projections onto the eigenspaces of eigenvalues contained in $D_N^\#$ and $E_n^\#$; by standard perturbation theoretic arguments (see, for example, [28, section XII.2], and especially the lemma following Theorem XII.7), the corresponding spectral multiplicities must be stable under perturbation from 0 to p. This gives (ii) and (iii).

We now turn to the perturbative estimates that prove (i). We will use strong estimates on the integral kernel $G_0^\#(x,y;z)$ for the operator $(L_\#(0)-z)^{-1}$ in order to estimate $(L_\#(p)-z)^{-1}$ in operator norm. By Lemma 2.1, it suffices to prove these operator norm estimates for $p \in C_0^\infty((0,1),\mathbb{C}^2)$. This assumption allows us to bypass domain questions which might arise if the coefficients were more singular.

We begin by noting the resolvent equations, true for $\Re(z)$ sufficiently negative,

$$(L_\#(p)-z)^{-1} = (L_\#(0)-z)^{-1} - (L_\#(0)-z)^{-1}(Dp_1D+p_2)(L_\#(p)-z)^{-1}.$$

Since $(L_\#(p)-z)^{-1}$ maps into $H^3[0,1]$ and p_1 and p_2 are smooth, the composition of Dp_1D+p_2 with $(L_\#(p)-z)^{-1}$ is well defined. From this equation it is easy to derive the useful identity

$$(L_\#(p)-z)^{-1} = A(z) + B(z)(L_\#(p)-z)^{-1}\,, \tag{2.2}$$

where

$$\begin{aligned} A(z) &= (L_\#(0)-z)^{-1} \\ &\quad - (L_\#(0)-z)^{-1}Dp_1(I+C(z))^{-1}D(L_\#(0)-z)^{-1}, \end{aligned} \tag{2.3}$$

$$\begin{aligned} B(z) &= (L_\#(0)-z)^{-1}Dp_1(I+C(z)^{-1}D(L_\#(0)-z)^{-1}p_2 \\ &\quad - (L_\#(0)-z)^{-1}p_2, \end{aligned} \tag{2.4}$$

and

$$C(z) = D(L_\#(0)-z)^{-1}Dp_1. \tag{2.5}$$

These equations are valid whenever $\|C(z)\| < 1/2$. Using Lemma A.4(e), we see that this holds for $z \notin C_{R,\alpha}^\#$ so long as $CR^{-1/2}(\ln R)^{1/2}\|p_1\|_{L^2} < 1/2$, where C is a numerical constant; we can ensure this by choosing a sufficiently large R depending only on M. We wish to show that, by increasing R if necessary, we can make $\|B(z)\| < 1/2$ for $z \notin C_{R,\alpha}^\#$ and then show that $\|A(z)\| \le CR^{-1}$ for a constant C. We can then use standard analytic continuation arguments to conclude that $\|(L_\#(p)-z)^{-1}\|$ is bounded for $z \notin C_{R,\alpha}^\#$ for sufficiently large R depending on M.

First, we show how to choose R depending on M so that $\|B(z)\| < 1/2$ for $z \notin C_{R,\alpha}^\#$. From Lemma A.4(b), (d), and (e), we see that $\|B(z)\| \le CR^{-1}$ for a constant C depending on M so that $\|B(z)\| < 1$ for R sufficiently large. The estimates in Lemma A.4 also show that

$$\|A(z)\| \le \frac{C}{R} + \frac{C}{R^2}(1+\|p_1\|_{L^2}),$$

which gives an estimate of the desired form. □

3. Approximation. First of all, we give the proof of Theorem 1.1. It is not difficult to see that the set $\mathcal{E}$ is open, since all eigenvalues $\lambda_n(p)$ with $n > N$ are simple, and for each of the eigenvalues $\lambda_n(p)$ with $1 \leq n \leq N$, there is a neighborhood U_n in E of any $q \in \mathcal{E}$ such that $\lambda_n(p)$ is simple for all $p \in U_n$. Taking $U = \cap_{n=1}^N U_n$ we obtain an open neighborhood of q contained in $\mathcal{E}$.

To see that $\mathcal{E}$ is dense, we will exploit analytic perturbation theory. Let $p \in E$ with $\|p\| < M$, and suppose that one or more of the eigenvalues of $L(p)$ are degenerate. Consider the family of operators $M(t) = L(tp)$ where $|t| < 2$. By changing the values of R and N as defined in section 2 if necessary, we may assume that the conclusions of Theorem 2.4 hold for all tp with $|t| < 2$. Let

$$P(t) = \frac{1}{2\pi i} \int_{\partial D_N} (M(t) - z)^{-1}\, dz\,,$$

where D_N is as defined in the previous section. For t real, $P(t)$ is an orthogonal projection onto the first N eigenvalues of $M(t)$, counted with multiplicity. By Theorem XII.12 of [28], we can find an analytic family of holomorphically invertible operators $U(t)$ defined for $|t| < 2$ so that $U(t)$ is unitary for t real and $U(t)P(0)U(t)^{-1} = P(t)$. The operator $m(t) = U(t)^{-1}M(t)U(t)$ commutes with $P(0)$ and may be regarded, for t real, as a Hermitian matrix acting on $\mathbb{C}^N$; the first N eigenvalues of $L(tp)$ are simple if and only if the eigenvalues of $m(t)$ are simple. Observe that at $t = 0$, the first N eigenvalues of $L(0)$ are simple by explicit calculation, so the same holds for $|t|$ small. Moreover, the eigenvalues of $m(t)$ are analytic functions of t ([17, Theorem II.6.1]). Thus a given pair of eigenvalues of $m(t)$ can be degenerate for at most finitely many t between 0 and 1. Hence the same holds true of $L(tp)$, so for each $\epsilon > 0$ there is a $t \in (1-\epsilon, 1)$ so that $tp \in \mathcal{E}$. This shows that $\mathcal{E}$ is dense in E, and completes the proof of Theorem 1.1.

Next, we prove Theorem 3.1.

THEOREM 3.1. *Let $\mathcal{D}$ be the set of $p \in C_0^\infty((0,1);\mathbb{R}^2)$ such that $L_\#(p)$ has simple spectrum and the spectra of $L(p)$, $L_1(p)$, and $L_2(p)$ have empty intersection. Then $\mathcal{D}$ is dense in $\mathcal{E}$.*

Proof. First we show how small perturbations may be used to make the spectra of $L_1(p)$ and $L_2(p)$ simple, and the spectra of the three operators nonintersecting. By Theorem 2.4, we need only show that the first N eigenvalues of each operator are simple and that the union of the intersection of the three sets $\{\lambda_n\}_{n=1}^N$, $\{\sigma_n\}_{n=1}^N$, and $\{\tau_n\}_{n=1}^N$ is empty. Let $M(t) = L(tp)$, $M_1(t) = L_1(tp)$, and $M_2(t) = L_2(tp)$. By the technique used above we may associate with these operators analytic, $N \times N$ matrix-valued functions $m(t)$, $m_1(t)$, $m_2(t)$ whose eigenvalues depend holomorphically on t with $|t| < 2$. By explicit calculation, the matrices $m(0)$, $m_1(0)$, and $m_2(0)$ have simple spectra with empty intersection, so the same is true for $|t|$ small by analytic perturbation theory. Analyticity of the eigenvalues implies that the spectra of $m(t)$, $m_1(t)$, and $m_2(t)$ must be simple and have empty intersection for all but a countable set of t with no accumulation point in the region $\{t \in \mathbb{C} : |t| < 2\}$. Thus, given p and ϵ we can find a q with $\|p - q\| < \epsilon$ so that $L(q)$, $L_1(q)$, and $L_2(q)$ have simple spectrum and the three spectra have empty intersection. Since $C_0^\infty((0,1);\mathbb{R}^2)$ is norm dense in E and the simplicity and empty intersection properties involve only finitely many eigenvalues, there is an r with $\|r - q\| < \epsilon$ and $r \in \mathcal{E} \cap C_0^\infty((0,1);\mathbb{R}^2)$ so that $L(r)$, $L_1(r)$, and $L_2(r)$ have simple spectra and their spectra have empty intersection. It follows that $\mathcal{D}$ is dense in $\mathcal{E}$, as asserted. ∎

4. Stability estimates. For $m > N$, let $z_m(\,\cdot\,;p)$ denote the normalized eigenfunction of $L(p)$ corresponding to the eigenvalue $\lambda_m(p)$, and let $u_m(\cdot)$ denote the corresponding eigenfunction of $L(0)$. We wish to derive $C[0,1]$-norm estimates on the differences $z_m^2 - u_m^2$ and $(z_m')^2 - (u_m')^2$ and also on the differences $z_m^2(\,\cdot\,;p) - z_m^2(\,\cdot\,;q)$ and $(z_m')^2(\,\cdot\,;p) - (z_m')^2(\,\cdot\,;q)$. These estimates will be used to analyze the operator $A(p)$ discussed in the introduction.

Let $G(x,y;z)$ denote the integral kernel of the operator $(L(p)-z)^{-1}$. In order to estimate the above quantities, we observe that we can recover $z_m(x,q)^2$ from the diagonal of the residue of $G(x,y;z)$ at $z=\lambda_m(p)$, and $z_m'(x,q)^2$ from the diagonal of the residue of $G_{xy}(x,y;z)$. We will exploit resolvent perturbation theory to prove the following estimates.

THEOREM 4.1. *Let $M>0$ and let p and q belong to E with $\|p\|_E$, $\|q\|_E < M$. For $\alpha \in (2,3)$ choose R and N as in Theorem* 2.4. *Then there are constants C_1 and C_2 depending on M so that for any $m > N$,*

(a) $\sup_{x\in[0,1]} |z_m(x;p)^2 - z_m(x;q)^2| \le C_1 m^{2-\alpha}$ *and*

(b) $\sup_{x\in[0,1]} |z_m'(x;p)^2 - z_m'(x;q)^2| \le C_2 m^{4-\alpha}$

hold.

As an immediate corollary, setting $q=0$, we have the following theorem.

THEOREM 4.2. *Let $M>0$ and $p \in E$ with $\|p\|_E < M$. For $\alpha \in (2,3)$ choose R and N as in Theorem* 2.4. *Then there are constants C_1 and C_2 depending on M so that for any $m > N$, the estimates*

(a) $\sup_{x\in[0,1]} |z_m(x;p)^2 - z_m(x;0)^2| \le C_1 m^{2-\alpha}$ *and*

(b) $\sup_{x\in[0,1]} |z_m'(x;p)^2 - z_m'(x;0)^2| \le C_2 m^{4-\alpha}$ *hold.*

To prove these results, we first note the following lemma.

LEMMA 4.3. *Let $M>0$ and $p \in E$ with $\|p\|_E < M$, let $\alpha \in (2,3)$, and choose R and N as in Theorem* 2.4. *Then for any $m > N$, the maps $p \mapsto z_m(x;p)$ and $p \mapsto z_m'(x,p)$ are continuous as maps from E to $C[0,1]$ with the sup norm.*

We do not give the full proof of Lemma 4.3 here but refer the reader to [29, section 5.5]. One first shows the existence of fundamental solutions with the required norm continuity using a Volterra series construction. One then uses the continuity of the eigenvalue map $p \mapsto \lambda_n(p)$, together with explicit formulas for the eigenfunctions in terms of the fundamental solutions, to obtain the required continuity.

By the lemma, it is enough to prove the estimates in Theorem 4.1 for p and q belonging to $C_0^\infty((0,1);\mathbb{R}^2)$. This restriction facilitates calculations which we will carry out in what follows.

To prove Theorem 4.1 for such smooth coefficients, we exploit the fact that the difference of eigenfunction squares and derivatives can be recovered from the residues of the respective operators

$$\mathbb{A}(p,q;z) = (L(p)-z)^{-1} - (L(q)-z)^{-1}$$

and

$$\mathbb{B}(p,q;z) = D(\mathbb{A}(p,q;z))D,$$

at the appropriate eigenvalue. Here D denotes differentiation with respect to x.

We begin with the resolvent formula

$$(L(p)-z)^{-1} = (L(0)-z)^{-1} - (L(0)-z)^{-1}V_p(L(0)-z)^{-1}$$

$$+ (L(0)-z)^{-1}V_p(L(p)-z)^{-1}V_p(L(0)-z)^{-1}$$

where

$$V_p = -Dp_1D + p_2.$$

From this formula it follows that

$$\begin{aligned}
\mathbb{A}(p,q;z) &= (L(p)-z)^{-1} - (L(q)-z)^{-1} \\
&= (L(0)-z)^{-1}(-V_{p-q})(L(0)-z)^{-1} \\
&+ (L(0)-z)^{-1}V_{p-q}(L(p)-z)^{-1}V_p(L(0)-z)^{-1} \\
&+ (L(0)-z)^{-1}V_q(L(p)-z)^{-1}V_{p-q}(L(q)-z)^{-1}V_p(L(0)-z)^{-1} \\
&+ (L(0)-z)^{-1}V_q(L(q)-z)^{-1}V_{p-q}(L(0)-z)^{-1}
\end{aligned} \tag{4.1}$$

with an analogous identity for the operator $\mathbb{B}(p,q;z)$. The operator $\mathbb{B}(p,q;z)$ is initially defined on $C_0^\infty(0,1)$ and extended by density to a bounded operator from $L^2[0,1]$ to itself. Let $K_{\mathbb{A}}$ and $K_{\mathbb{B}}$ denote the respective integral kernels of $\mathbb{A}(p,q;z)$ and $\mathbb{B}(p,q;z)$. The kernels $K_{\mathbb{A}}$ and $K_{\mathbb{B}}$ can be expressed in terms of the integral kernels of $(L(p)-z)^{-1}$ and $(L(q)-z)^{-1}$, which are continuously differentiable in x and y; thus $\mathbb{A}(p,q;z)$ and $\mathbb{B}(p,q;z)$ have continuous kernels. Moreover, the formulas

$$z_m^2(x;p) - z_m^2(x;q) = \frac{1}{2\pi i}\int_{\gamma_m} K_{\mathbb{A}}(x,x;z)\,dz$$

and

$$z_m'(x;p)^2 - z_m'(x;q)^2 = \frac{1}{2\pi i}\int_{\gamma_m} K_{\mathbb{B}}(x,x;z)\,dz$$

hold, where γ_m is the contour $\{z : |z-\lambda_m(0)| = Rm^\alpha\}$.

Thus, Theorem 4.1 will follow if we can show that

$$\sup_{z\in\gamma_m}\sup_{(x,y)\in[0,1]} |K_{\mathbb{A}}(x,y;z)| \le C_1\|p-q\|_E m^{2-2\alpha} \tag{4.2}$$

and

$$\sup_{z\in\gamma_m}\sup_{(x,y)\in[0,1]} |K_{\mathbb{B}}(x,y;z)| \le C_2\|p-q\|_E m^{4-2\alpha}, \tag{4.3}$$

where C_1 and C_2 depend only on α and M. If T is an integral operator on $L^2[0,1]$ with continuous kernel $K(x,y)$, the sup norm of K is dominated by the $L^1[0,1]\to L^\infty[0,1]$ norm of the operator T. Thus it suffices to estimate the $L^1\to L^\infty$ operator norm of each of the terms in (4.1) and the corresponding identity for $\mathbb{B}(p,q;z)$. Note that each term in (4.1) contains at least one factor involving $p-q$.

Roughly speaking, each resolvent contributes a factor $m^{-\alpha}$, each derivative that occurs contributes a factor m, and each factor of p, q, or $p-q$ contributes a factor $\|p\|_E$, $\|q\|_E$, or $\|p-q\|_E$ to the estimates. This "naive power counting" gives estimates of the desired form. The power counting is justified by the following two results, which themselves depend on Lemma A.5 in Appendix A.

LEMMA 4.4. *Let $z \in \gamma_m$, let $0 \le i, j \le 1$, and let $M > 0$. Let $r \in C_0^\infty(0,1)$. Then for any p and q with $1 \le p < \infty$ and $1 \le q \le \infty$, the estimate*

$$\|D^i(L(0) - z)^{-1} D^j r\|_{p,q} \le C_{p,q} m^{i+j-\alpha} \|r\|_2$$

holds.

This lemma is a straightforward consequence of Lemma A.5; the following perturbative argument shows that an analogous result holds for the resolvent of $L(p)$ when $p \ne 0$.

LEMMA 4.5. *Let $z \in \gamma_m$, let $0 \le i, j \le 1$, and let $M > 0$. Let $r \in C_0^\infty(0,1)$. There is a positive integer $N_1 > N$ depending only on M so that for all $p \in E$ with $\|p\| < M$ and every $m > N_1$, the estimate*

$$\|D^i(L(p) - z)^{-1} D^j r\| \le C_{i+j} m^{i+j-\alpha} \|r\|_2$$

holds.

Proof. We will use equation (2.2). With R chosen sufficiently large, as in Theorem 2.4, so that (2.2) holds for $z \in \gamma_m$, we may estimate $\|A(z)\|$ and $\|B(z)\|$ for $z \in \gamma_m$ using Lemma A.5 and obtain

$$\|A(z)\| \le C m^{2-2\alpha} \|p_1\|_2$$

and

$$\|B(z)\| \le C m^{2-2\alpha} (1 + \|p_1\|_2) \|p_2\|_2.$$

Since $2 - 2\alpha < -\alpha$ for $\alpha > 2$, we recover the estimate with $i = j = 0$. If $i = 1$ and $j = 0$, we compute from the second resolvent identity that

$$\begin{aligned} &D(L(p) - z)^{-1} \\ &\quad = D(L(0) - z)^{-1} - D(L(0) - z)^{-1}(Dp_1 D + p_2)(L(p) - z)^{-1}. \end{aligned} \tag{4.4}$$

From Lemma A.5, we obtain

$$\|D(L(p) - z)^{-1}\| \le c_1 m^{1-\alpha} + c_3 \|p_1\|_2 m^{2-\alpha} \|D(L(p) - z)^{-1}\| + c_1 \|p_2\|_2 m^{1-\alpha}.$$

By choosing m so large that $c_3 \|p_1\|_2 m^{2-\alpha} < 1/2$, we can conclude that

$$\|D(L(p) - z)^{-1}\| \le C m^{1-\alpha}.$$

The proofs for $(i, j) = (0, 1)$ and $(i, j) = (1, 1)$ are similar. ∎

To finish the proof of Theorem 4.1, we use Lemmas 4.4 and 4.5 in conjunction with the identity (4.1) and the corresponding identity for the operator $D(L(p) - z)^{-1} D$ to estimate $\|\mathbb{A}(p, q; z)\|_{L^1 \to L^\infty}$ and $\|\mathbb{B}(p, q; z)\|_{L^1 \to L^\infty}$ and thereby show that (4.2) and (4.3) hold. For example, the norm of the second right-hand term in (4.1) is

$$\begin{aligned} &\|(L(0) - z)^{-1} V_{p-q} (L(p) - z)^{-1} V_p (L(0) - z)^{-1}\|_{L^1 \to L^\infty} \\ &\quad = \|(L(0) - z)^{-1} (D(p_1 - q_1)D + (p_2 - q_2)) \\ &\qquad \times (L(p) - z)^{-1} (Dp_1 D + p_2)(L(0) - z)^{-1}\|_{L^1 \to L^\infty}. \end{aligned}$$

The term of highest order in m for $z \in \gamma_m$ comes from the term involving p_1, because it involves the highest number of differentiations. In what follows, let C denote a generic constant depending on only M, a bound for $\|p\|_E$ and $\|q\|_E$, and let $\| \cdot \|_{p,q}$ denote the $\mathcal{B}(L^p[0,1], L^q[0,1])$-operator norm. We can estimate

$$\begin{aligned}
&\|(L(0)-z)^{-1}D(p_1-q_q)D(L(p)-z)^{-1}Dp_1D(L(0)-z)^{-1}\|_{L^1\to L^\infty} \\
&\quad \le \|(L(0)-z)^{-1}D(p_1-q_1)\|_{2,\infty}\ \|D(L(p)-z)^{-1}p_1\|_{2,2}\ \|D(L(0)-z)^1\|_{1,2} \\
&\quad \le C(m^{1-\alpha}\|p_1-q_1\|_{L^2[0,1]})(m^{1-\alpha}\|p_1\|_{L^2[0,1]})(m^{1-\alpha}) \\
&\quad \le Cm^{3-3\alpha}
\end{aligned}$$

using Lemma 4.4 for the first and third factors, and Lemma 4.5 for the second. Similar estimates on the remaining terms in (4.1) show that all terms can be bounded by $Cm^{2-2\alpha}\|p-q\|_E$ so that

$$\|\mathbb{A}(p,q;z)\|_{1,\infty} \le Cm^{2-2\alpha}\|p-q\|_E.$$

Analogous estimates show that

$$\|\mathbb{B}(p,q;z)\|_{1,\infty} \le Cm^{4-2\alpha}\|p-q\|_E.$$

Finally, we refine the crude eigenvalue asymptotics obtained in section 2. We let

$$p_{1,m} = \int_0^1 p_1(x)\cos(m\pi x)\,dx.$$

Note that the sequence $\{p_{1,m}\}$ belongs to $\ell^2(\mathbb{N})$.

THEOREM 4.6. *Let $p \in E$ with $\|p\|_E < M$, and $\alpha \in (2,3)$. There is a constant C depending only on α and M such that the estimate*

$$|\lambda_m(p) - \lambda_m(0) - m^2\pi^2\,(\overline{p_1} + p_{1,2m})\,| \le Cm^{4-\alpha}$$

holds for $n > N$. In particular, $m^{-2}(\lambda_m(p) - \lambda_m(0) - m^2\pi^2\overline{p_1})$ defines a sequence belonging to $\ell^2(\mathbb{N})$, and $\overline{p_1}$ may be recovered from the asymptotics of the $\lambda_m(p)$.

Proof. First suppose that $p \in C_0^\infty(0,1) \times C_0^\infty(0,1)$. It suffices to prove the estimate for such p since an arbitrary $q \in E$ can be approximated by such smooth p in norm and the eigenvalues are continuous functions of p. Let $\nu_m(t) = \lambda_m(tp)$ for $t \in [0,1]$. Then standard perturbative calculations show that

$$\begin{aligned}
&\nu_m(1) - \nu_m(0) \\
&\qquad = \int_0^1 \int_0^1 p_1(x)(z_m'(x;tp))^2\,dx\,dt \\
&\qquad + \int_0^1 \int_0^1 p_2(x)(z_m(x,tp))^2\,dx\,dt.
\end{aligned}$$

Using Theorem 4.2 together with the explicit formula

$$z_m(x;0) = \sqrt{2}\sin(m\pi x),$$

we readily obtain the claimed asymptotics. ∎

5. Analyticity. In this section, we prove the following theorem.

THEOREM 5.1. *The map μ is an analytic mapping from $\mathcal{E}$ into F.*

Proof. Theorem 4.6 already implies that the map μ has range in F. It remains to show that it has the required analyticity. It is not difficult to see that any particular μ_n is analytic in a small neighborhood of any $p \in \mathcal{E}$, where the size of the neighborhood may depend on n. To show analyticity of μ we must show, for any $p \in \mathcal{E}$, that the μ_n with $n > N$ are analytic in a fixed neighborhood of p independent of $n > N$.

To do this, we fix an $M > 0$ and choose N and R as in section 2. Let

$$P_n(p) = \frac{1}{2\pi i} \int_{\gamma_n} (L(p) - z)^{-1}\, dz\,,$$

where E_n is as defined in section 2. It follows from the analyticity of the resolvent that $P_n(p)$ is analytic in p. Moreover, for p and q with $\|p\| \le M$ and $\|q\| \le M$,

$$\begin{aligned} \|P_n(p) - P_n(q)\| &\le Rn^{\alpha} \sup_{z \in \gamma_n} \|(L(p) - z)^{-1} - (L(q) - z)^{-1}\| \\ &\le C\|p - q\|_E\, n^{2-\alpha} \end{aligned}$$

by the estimate in the proof Theorem 4.2. For $n > N$ and $\alpha > 2$ we may choose $\|p - q\| < (2CN^{2-\alpha})^{-1}$ and guarantee that $\|P_n(p) - P_n(q)\| \le 1/2$ for all $n > N$. From the formula

$$\lambda_n(q) = \frac{\langle z_n(\cdot, p), L(q)P(q)u_n(p)\rangle}{\langle z_n(\cdot, p), P_n(q)u_n(p)\rangle},$$

we see that $\lambda_n(q)$ is analytic for $\|p - q\| < (2CN^{2-\alpha})^{-1}$, which defines a fixed neighborhood of p independent of $n > N$. Thus, given $p \in \mathcal{E}$, there is a fixed neighborhood U of p so that all of the $\mu_n(q)$ are analytic for $q \in U$. This shows the required analyticity. □

6. Linear independence: The Borg system. We now consider linear independence of the functions $g_n(p)$ constructed from eigenfunctions of $L(p)$ via (1.7) when $p \in \mathcal{D}$. We shall accomplish this by displaying a set of functions which is biorthogonal to the $g_n(p)$ in E. The development of this biorthogonal set will involve the spectral theory of a non-self-adjoint fifth-order system, the *Borg system*, satisfied by the functions

$$\widehat{g}_n(x; p) = \begin{pmatrix} (z_n(x; p))^2 \\ (z_n'(x; p))^2 \end{pmatrix}, \tag{6.1}$$

where $z_n(x; p)$ is the nth eigenfunction of $L(p)$.

First, we will define the Borg system and construct a basis for its solution space from solutions of the underlying fourth-order problems. Then, we will show how one may specify boundary conditions for this system so that the spectrum of the resulting boundary value problem for the Borg system coincides with the spectra of the fourth-order operators $L(p)$, $L_1(p)$, and $L_2(p)$. Finally, we will show that the resolvent of the Borg system has simple poles and rank-one residues; this will enable us to construct the desired biorthogonal set.

6.1. The Borg system. Now, we define the Borg system and establish some of its properties.

LEMMA 6.1. *For $p \in C_0^{\infty}(0,1) \times C_0^{\infty}(0,1)$, there exist differential operators $\mathcal{M}(p)$ and $\mathcal{B}(p)$, mapping $C_0^{\infty}(0,1) \times C_0^{\infty}(0,1)$ into itself, such that*

(i)

$$\mathcal{M} = \begin{pmatrix} D^5 + M_{11}(x,D) & M_{12}(x,D) \\ M_{21}(x,D) & D^5 + M_{22}(x,D) \end{pmatrix},$$

where each M_{ij} is a linear differential operator of order not exceeding four with smooth coefficients depending on p, and

(ii)

$$\mathcal{B} = \begin{pmatrix} \frac{8}{3}D & 0 \\ \mathcal{B}_{21}(x,D) & -24D \end{pmatrix},$$

where $\mathcal{B}_{21}$ is a third-order linear differential operator with smooth coefficients depending on p.

(iii) *If u and v are solutions of $L(p)u = \lambda u$, then*

$$\mathcal{M}(p)\phi = \lambda \mathcal{B}(p)\phi, \tag{6.2}$$

where $\phi = (uv, u'v')^T$.

The proof is a direct calculation and is omitted. The explicit forms of $\mathcal{M}$ and $\mathcal{B}$ are given in Appendix B.

Next we note a purely algebraic lemma. It will be used to furnish bases of solutions from which Green's function for the fifth-order system

$$(\mathcal{M} - \lambda\mathcal{B})u = f, \tag{6.3}$$

with boundary conditions dictated by the chosen basis, can be calculated.

LEMMA 6.2. *Fix $\lambda \in \mathbb{C}$. Let $\{y_i\}_{i=1}^4$ be a fundamental set of solutions for the differential equation $L(p)u = \lambda u$. The ten products $\{y_i y_j : 1 \le i \le j \le 4\}$ have nonvanishing Wronskian.*

Lemma 6.2 is a consequence of the following abstract result about symmetric tensor products. Recall that if V is a real n-dimensional vector space with basis $\{e_i\}_{i=1}^n$, the *symmetric tensor product* $V \otimes_s V$ is the $n(n+1)/2$-dimensional real vector space spanned by the tensors $e_i \otimes_s e_j = e_i \otimes e_j + e_j \otimes e_i$. If $A : V \to V$ is a linear transformation, $A \otimes_s A$ is the linear transformation on $V \otimes_s V$ acting on basis vectors by $(A \otimes_s A)(e_i \otimes_s e_j) = (Ae_i) \otimes_s (Ae_j)$ and extended to $V \otimes_s V$ by linearity.

LEMMA 6.3. *Let $A : V \to V$ be a linear operator. Then,*

$$\det(A \otimes_s A) = 2^{\frac{n(n+1)}{2}} \det(A)^{n+1}.$$

Proof. Suppose first that A is diagonal with eigenvalues $\{\lambda_i\}_{i=1}^n$. The eigenvalues of $A \otimes_s A$ are $2\lambda_i\lambda_j$ for $1 \le i \le j \le n$. The product over these $n(n+1)/2$ numbers gives $2^{\frac{n(n+1)}{2}} \left(\prod_{i=1}^n \lambda_i\right)^{n+1}$. This proves the formula for the dense set of $n \times n$ matrices which are similar to a diagonal matrix. The general result follows by continuity of the determinant function. ∎

We denote the 10×10 matrix of products $y_i y_j$, $1 \le i \le j \le 4$, and their derivatives of up to ninth order by Y_S.

Proof of Lemma 6.2. Let Ψ be the 10×10 matrix consisting of the symmetric derivatives $D^{(k)}y_iD^{(l)}y_j + D^{(l)}y_iD^{(k)}y_j$ for $1 \le i \le j \le 4$ and $0 \le k \le l \le 3$. By Lemma 6.3,

$$\det(\Psi) = 2^{10}\left(W(y_1, y_2, y_3, y_4)\right)^5 \neq 0.$$

A direct calculation shows that there is a nonsingular constant matrix G_1 for which

$$Y_S = \Psi G_1^T, \tag{6.4}$$

which establishes the result. □

LEMMA 6.4. *Let y_i be as in Lemma* 6.2*, and let Φ be the 10×10 matrix consisting of the y_iy_j and their derivatives of up to fourth order, and the $y_i'y_j'$ and their derivatives of up to fourth order. There is a nonsingular constant matrix C such that $\Phi = Y_SC^T$.*

The proof is a direct calculation and is omitted. The nonsingular constant matrix C maps any row vector consisting of y_iy_j and its derivatives up to ninth order, where y_i and y_j are solutions of the fourth-order problem, to a corresponding row vector whose entries are y_iy_j and its first four derivatives, followed by $y_i'y_j'$ and its first four derivatives.

We conclude from Lemma 6.4 and relation (6.4) that

$$\left\{\left(y_iy_j,\, y_i'y_j'\right)^T \,:\, 1 \le i \le j \le 4\right\}$$

forms a basis for the ten-dimensional solution space of the fifth-order system (6.2).

6.2. Boundary conditions for the Borg system. We are now ready to prescribe boundary conditions on the Borg system. We do so implicitly, by specifying a basis for the desired ten-dimensional solution space. To this end, choose a basis y_j of solutions to $L(p)u = \lambda u$ to satisfy the initial conditions $D^{i-1}y_j(0) = \delta_{ij}$, $1 \le i, j \le 4$, and similarly choose a basis z_j of solutions to $L(p)u = \lambda u$ to satisfy $D^{i-1}z_j(1) = \delta_{ij}$, $1 \le i, j \le 4$. Denote by B the 4×4 matrix with

$$y_j(x,\lambda) = \sum_{i=1}^{4} B_{ij}z_i(x,\lambda), \quad 1 \le i, j \le 4;$$

the matrix B is a holomorphic function of λ with determinant 1. Bases of solutions for the fourth-order homogeneous problems $L(p)u = \lambda u$, $L_1(p)u = \lambda u$, and $L_2(p)u = \lambda u$ obeying the $x = 0$ and $x = 1$ boundary conditions are, respectively, $\{y_2, y_4, z_2, z_4\}$, $\{y_2, y_4, z_1, z_3\}$, and $\{y_2, y_4, z_3, z_4\}$. We denote the Wronskians of these sets, respectively, as $W_1(\lambda)$, $W_2(\lambda)$, and $W_3(\lambda)$; the respective zeros are exactly $\{\lambda_n(p)\}$, $\{\sigma_n(p)\}$, and $\{\tau_n(p)\}$. In terms of the matrix B,

$$\begin{aligned} W_1(\lambda) &= B_{1,4}B_{3,2} - B_{1,2}B_{3,4}, \\ W_2(\lambda) &= B_{2,4}B_{4,2} - B_{2,2}B_{4,4}, \\ W_3(\lambda) &= B_{1,2}B_{2,4} - B_{1,4}B_{2,2}. \end{aligned}$$

Let us denote, for each i and j,

$$\phi_{ij}^L = \begin{pmatrix} y_iy_j, \\ \\ y_i'y_j', \end{pmatrix} \tag{6.5}$$

and

$$\phi_{ij}^R = \begin{pmatrix} z_i z_j, \\ \\ z_i' z_j', \end{pmatrix}. \tag{6.6}$$

From Lemmas 6.2 and 6.4 it follows that either of the sets ϕ_{ij}^L or ϕ_{ij}^R form a basis for the solution space of (6.2).

In lieu of specifying explicit boundary conditions for the fifth-order system (6.2), we shall specify a ten-dimensional solution space for the Borg system by explicitly choosing a basis for the solution space. This basis, consisting of a subset $\mathcal{L}$ of the ϕ_{ij}^L and a subset $\mathcal{R}$ of the ϕ_{ij}^R, will be chosen so that the eigenfunctions of the three boundary value problems specified in section 2 will contribute to the eigenfunctions of (6.2) via (6.1). Explicitly, we choose

$$\mathcal{L} = (\phi_{22}^L, \phi_{44}^L, \phi_{24}^L)$$

and

$$\mathcal{R} = (\phi_{11}^R, \phi_{22}^R, \phi_{33}^R, \phi_{44}^R, \phi_{13}^R, \phi_{24}^R, \phi_{34}^R).$$

In what follows, we will also use the notation $\{\phi_i\}_{i=1}^{10}$ for these basis functions, where $1 \le i \le 3$ for the vectors in $\mathcal{L}$, and $4 \le i \le 10$ for the vectors in $\mathcal{R}$. We shall also designate the components of ϕ_i by

$$\phi_i(x;\lambda) = \begin{pmatrix} \zeta_i(x;\lambda), \\ \\ \eta_i(x;\lambda), \end{pmatrix}.$$

We need to verify that the ten functions in $\mathcal{L} \cup \mathcal{R}$ are linearly independent (and hence, a basis for the solution space). Computing an appropriate Wronskian determinant leads to an explicit eigenvalue condition. Recall that $\lambda_n(p)$, $\sigma_n(p)$, and $\tau_n(p)$ denote, respectively, the nth eigenvalues of $L(p)$, $L_1(p)$, and $L_2(p)$.

LEMMA 6.5. *Let $W(\lambda)$ be the Wronksian of the solution set $\mathcal{L} \cup \mathcal{R}$, and let $p \in \mathcal{D}$. Then $W(\lambda)$ is a constant multiple of $W_1(\lambda)W_2(\lambda)W_3(\lambda)$, and $W(\lambda)$ has simple zeros at the eigenvalues $\{\lambda_n(p)\}$, $\{\sigma_n(p)\}$, $\{\tau_n(p)\}$.*

Proof. By Lemma 6.4 and the remarks following it, it suffices to show that the assertion of Lemma 6.5 is true for the Wronskian determinant of the ten functions z_1^2, z_2^2, z_3^2, z_4^2, $z_1 z_3$, $z_2 z_4$, $z_3 z_4$, and y_2^2, y_4^2, $y_2 y_4$. Evaluating the determinant at $x = 1$ leads to the determinant of a block upper triangular matrix

$$A = \begin{pmatrix} I & A_{12} \\ \\ 0 & A_{22} \end{pmatrix}, \tag{6.7}$$

where I is the 7×7 identity matrix, A_{12} is a 7×3 matrix whose entries are polynomials in the $B_{i,j}$, and A_{22} is the 3×3 matrix

$$A_{22}(\lambda) = \begin{pmatrix} 2\,B_{1,2}\,B_{2,2} & B_{1,2}\,B_{2,4} + B_{2,2}\,B_{1,4} & 2\,B_{1,4}\,B_{2,4} \\ \\ 2\,B_{1,2}\,B_{4,2} & B_{1,2}\,B_{4,4} + B_{1,4}\,B_{4,2} & 2\,B_{1,4}\,B_{4,4} \\ \\ 2\,B_{2,2}\,B_{3,2} & B_{2,2}\,B_{3,4} + B_{2,4}\,B_{3,2} & 2\,B_{2,4}\,B_{3,4} \end{pmatrix}.$$

Explicit calculation gives the formula

$$\det(A_{22}(\lambda)) = W_1(\lambda)W_2(\lambda)W_3(\lambda).$$

The fact that $W(\lambda)$ has simple zeros follows from a result of Everitt (see [12]) and the fact that $p \in \mathcal{D}$. □

Thus, the spectrum of the boundary value problem for the Borg system coincides exactly with the spectra of $L(p)$, $L_1(p)$, and $L_2(p)$. With some additional calculation, we can show the following lemma.

LEMMA 6.6. *Let $p \in \mathcal{D}$. At each zero of $W(\lambda)$, the kernel of $\mathcal{M} - \lambda\mathcal{B}$ is one-dimensional.*

Proof. It is enough to show that the matrix A defined in (6.7) has a one-dimensional kernel at such points λ. To see this, note that a nonzero solution of the homogeneous equation $(\mathcal{M} - \lambda\mathcal{B})u = 0$, which satisfies the left and right boundary conditions, exists if and only if the spans of the vectors in $\mathcal{L}$ and $\mathcal{R}$ have nonempty intersection. Recalling the notation $\{\phi_i\}_{i=1}^{10}$ we see that the dimension of the kernel of $\mathcal{M} - \lambda\mathcal{B}$ is the dimension of solutions $\{\alpha_i\}$ of the equation

$$\sum_{i=1}^{10} \alpha_i \phi_i = 0\,, \tag{6.8}$$

since the sets $\{\phi_i\}_{i=1}^{3}$ and $\{\phi_j\}_{j=4}^{10}$ are linearly independent. Let $\mathbf{\Phi}$ denote the 10×10 matrix containing the components of ϕ_i and their derivatives of up to fourth order. This matrix is related by a nonsingular constant matrix to the 10×10 Wronskian matrix containing the corresponding products $y_i y_j$, $z_i z_j$ and their derivatives of up to ninth order. The solutions of (6.8) therefore correspond to the nullspace of the Wronskian matrix of the $y_i y_j$ and $z_i z_j$ so that the dimension of the space of solutions to (6.8) is exactly the dimension of the kernel of the matrix A in Lemma 6.5. Since A is upper triangular, $\dim\ker A = \dim\ker A_{22}$. The proof is completed by showing that $\dim\ker A_{22} = 1$ at each zero of $W(\lambda)$, which is the content of the next lemma. □

LEMMA 6.7. *If $W(\lambda) = 0$, then $\dim\ker(A_{22}) = 1$.*

Proof. By virtue of Lemma 6.5, $W(\lambda) = 0$ if and only if $W_j(\lambda) = 0$ for some j. Note that, since the spectra of the fourth-order problems do not overlap,

- $B_{1,2}$ and $B_{1,4}$ are never zero simultaneously (for otherwise, $W_1(\lambda) = W_3(\lambda) = 0$) and
- $B_{2,2}$ and $B_{2,4}$ are never zero simultaneously (for otherwise, $W_2(\lambda) = W_3(\lambda) = 0$).

Since the matrix A_{22} is symmetric (up to column-interchanges) with respect to the second index on the coefficients $B_{i,j}$, we may assume without loss of generality that $B_{2,2} \neq 0$ and express A_{22} equivalently as

$$A_{22} = \begin{pmatrix} 2B_{1,2} & B_{1,4} + B_{1,2}\alpha_2 & 2B_{1,4}\alpha_2 \\ 2B_{1,2}B_{4,2} & B_{1,2}B_{4,4} + B_{4,2}B_{1,4} & 2B_{1,4}B_{4,4} \\ 2B_{3,2} & B_{3,4} + B_{3,2}\alpha_2 & 2B_{3,4}\alpha_2 \end{pmatrix},$$

where

$$\alpha_2 \equiv \frac{B_{2,4}}{B_{2,2}}\,.$$

There are two possibilities: either $B_{1,2} \neq 0$ or $B_{1,4} \neq 0$. We consider the former only, the latter being essentially the same.

Assume $B_{1,2} \neq 0$. The matrix A_{22} can then be written equivalently as

$$A_{22} = \begin{pmatrix} 2 & \alpha_1 + \alpha_2 & 2\alpha_1\alpha_2 \\ 2B_{4,2} & B_{4,4} + B_{4,2}\alpha_1 & 2B_{4,4}\alpha_1 \\ 2B_{3,2} & B_{3,4} + B_{3,2}\alpha_2 & 2B_{3,4}\alpha_2 \end{pmatrix},$$

where

$$\alpha_1 \equiv \frac{B_{1,4}}{B_{1,2}}.$$

It is straightforward to show that there exist nonsingular matrices $\mathbf{C}$ and $\mathbf{E}$ for which

$$A_{22} = \mathbf{C}\mathbf{D}\mathbf{E}^T,$$

where

$$\mathbf{D} = \begin{pmatrix} 2 & \alpha_1 + \alpha_2 & 2\alpha_1\alpha_2 \\ 0 & B_{4,4} - B_{4,2}\alpha_2 & 2\alpha_1(B_{4,4} - B_{4,2}\alpha_2) \\ 0 & B_{3,4} - B_{3,2}\alpha_1 & 2\alpha_2(B_{3,4} - B_{3,2}\alpha_1) \end{pmatrix} = \begin{pmatrix} 2 & \alpha_1 + \alpha_2 & 2\alpha_1\alpha_2 \\ 0 & -\frac{W_2}{B_{2,2}} & -2\frac{\alpha_1 W_2}{B_{2,2}} \\ 0 & -\frac{W_1}{B_{1,2}} & -2\frac{\alpha_2 W_1}{B_{1,2}} \end{pmatrix}.$$

Noting that row 1 of $\mathbf{D}$ never vanishes (and is independent of the other rows) and rows 2 and 3 cannot vanish simultaneously, we see that $\dim\ker A_{22} = \dim\ker \mathbf{D} \leq 1$ and is determined by the dimension of the kernel of the 2×2 submatrix

$$AA \equiv \begin{pmatrix} -\frac{W_2}{B_{2,2}} & -2\frac{\alpha_1 W_2}{B_{2,2}} \\ -\frac{W_1}{B_{1,2}} & -2\frac{\alpha_2 W_1}{B_{1,2}} \end{pmatrix}.$$

We have

$$\det(AA) = \frac{2}{B_{1,2}B_{2,2}} W_1 W_2 (\alpha_2 - \alpha_1) = -\frac{2}{B_{1,2}^2 B_{2,2}^2} W_1 W_2 W_3,$$

and the result follows. □

As a consequence, we have not only that the eigenvalues $\{\nu_n\}$ of the boundary value problem (6.2) are precisely the eigenvalues of the three boundary value problems considered in section 2, but also that, for each such eigenvalue, the eigenfunction of (6.2) is $(z^2, (z')^2)^T$, where z is the eigenfunction of the corresponding fourth-order problem.

6.3. Biorthogonal set. We shall now construct the desired biorthogonal set from the residues of the resolvent $(\mathcal{M} - \lambda\mathcal{B})^{-1}$. It follows from explicit formulas for the Green's function in terms of the basis functions ϕ_{ij}^L, ϕ_{ij}^R and the Wronskian $W(\lambda)$ that the resolvent $(\mathcal{M} - \lambda\mathcal{B})^{-1}$ has simple poles with rank-one residues.

THEOREM 6.8. *Let $(\mathcal{M} - \lambda\mathcal{B})^{-1}$ be the resolvent of the non-self-adjoint boundary value problem* (6.2). *Then the poles of $(\mathcal{M} - \lambda\mathcal{B})^{-1}$ are simple and occur at the*

numbers ν_n. Their residue takes the form $(\chi_n, \cdot)\psi_n$, where $\psi_n \in \mathrm{Ker}(\mathcal{M} - \lambda\mathcal{B})$ and $\chi_n \in \mathrm{Ker}((\mathcal{M} - \lambda\mathcal{B})^*)$.

Proof. The integral kernel of $(\mathcal{M} - \lambda\mathcal{B})^{-1}$ is the matrix-valued function

$$\begin{pmatrix} G_{11}(x,t;\lambda) & G_{12}(x,t;\lambda) \\ G_{21}(x,t;\lambda) & G_{22}(x,t;\lambda) \end{pmatrix},$$

where the G_{ij} obey

- $G_{11}(x,t;\lambda)$ and $G_{22}(x,t;\lambda)$ are continuous in $(x,t) \in [0,1] \times [0,1]$ together with their derivatives of up to order three and have a unit jump at $x = t$ in their fourth derivative,
- $G_{12}(x,t;\lambda)$ and $G_{21}(x,t;\lambda)$ are continuous in $(x,t) \in [0,1] \times [0,1]$ together with their derivatives up to order four

and solve the differential equations

$$(\mathcal{M} - \lambda\mathcal{B})_x G_{ij}(x,t;\lambda) = 0$$

for $x \neq t$. We can find explicit expressions for the G_{ij} by setting

$$\begin{pmatrix} G_{11}, \\ G_{21}, \end{pmatrix} = \begin{cases} \sum_{i\in I} \alpha_i(t)\phi_i(x) & x < t, \\ -\sum_{j\in J} \alpha_j(t)\phi_j(x) & x > t \end{cases}$$

and

$$\begin{pmatrix} G_{12}, \\ G_{22}, \end{pmatrix} = \begin{cases} \sum_{i\in I} \beta_i(t)\phi_i(x) & x < t, \\ -\sum_{j\in J} \beta_j(t)\phi_j(x) & x > t \end{cases}$$

and solving the linear equations

$$\begin{aligned} \sum_{i=1}^{10} \alpha_i(x)\zeta_i^{(k)}(x) &= 0, \qquad 0 \le k \le 3, \\ \sum_{i=1}^{10} \alpha_i(x)\zeta_i^{(4)}(x) &= 1, \\ \sum_{i=1}^{10} \alpha_i(x)\eta_i^{(k)}(x) &= 0, \qquad 0 \le k \le 4, \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^{10} \beta_i(x)\zeta_i^{(k)}(x) &= 0, \qquad 0 \le k \le 4, \\ \sum_{i=1}^{10} \beta_i(x)\eta_i^{(k)}(x) &= 0, \qquad 0 \le k \le 3, \\ \sum_{i=1}^{10} \beta_i(x)\eta_i^{(4)}(x) &= 1. \end{aligned}$$

Using Cramer's rule to solve for the functions α_i and β_i yields an expression for Green's function in terms of the holomorphic functions $\phi_i(x;\lambda)$ and the Wronskian $W(\lambda)$. Using the known properties of $W(\lambda)$ we conclude that Green's function has simple poles.

A simple argument shows that the residue of $(\mathcal{M}-\lambda\mathcal{B})^{-1}$ has range in $\mathrm{Ker}(\mathcal{M}-\lambda\mathcal{B})$ and is therefore rank-one by Lemma 6.6. Writing the residue at $\lambda=\nu_n$ in the form $(\chi_n,\cdot)\psi_n$, it follows from the identity

$$((\mathcal{M}-\lambda\mathcal{B})^*)^{-1}=((\mathcal{M}-\lambda\mathcal{B})^{-1})^*$$

that $\chi_n\in\mathrm{Ker}((\mathcal{M}-\lambda\mathcal{B})^*)$. □

Suppose that ν_i and ν_j are distinct eigenvalues of the Borg operator. It is not difficult to see that the resolvent identity

$$(\mathcal{M}-\lambda\mathcal{B})^{-1}-(\mathcal{M}-\mu\mathcal{B})^{-1}=(\mu-\lambda)(\mathcal{M}-\lambda\mathcal{B})^{-1}\mathcal{B}(\mathcal{M}-\mu\mathcal{B})^{-1}$$

holds. Let P_i and P_j be the rank-one residues corresponding to these distinct eigenvalues. From the resolvent identity above, it is easy to see that the relations $P_i\mathcal{B}P_i=P_i$ and $P_i\mathcal{B}P_j=0$ hold for $i\neq j$. Writing $P_i=(\chi_i,\cdot)\psi_i$, we obtain the following theorem.

THEOREM 6.9. *The biorthogonality relations*

$$(\chi_i,\mathcal{B}\psi_j)=\delta_{ij} \tag{6.9}$$

hold. Consequently, the eigenfunctions $\{\psi_j\}$ form a linearly independent set in E.

Proof. The conclusion that the set $\{\psi_j\}$ is linearly independent in E follows immediately from (6.9), once it is established that

$$\chi_i\in D(\mathcal{B}^*) \tag{6.10}$$

for each i. Appendix B gives explicitly the boundary conditions which determine $D((\mathcal{M}-\lambda\mathcal{B})^*)$ and $D(\mathcal{B}^*)$. Direct comparison shows that $D((\mathcal{M}-\lambda\mathcal{B})^*)\subset D(\mathcal{B}^*)$, so (6.10) holds. □

We now order the poles, ν_n, of the Borg operator so that $\nu_{3n}=\lambda_n$, $\nu_{3n+1}=\sigma_n$, $\nu_{3n+2}=\tau_n$. The vectors ψ_{3n} are, up to normalization, exactly the vectors $\widehat{g}_n$. Thus these vectors form a linearly independent set, and their orthogonal complement is an infinite-dimensional space spanned by the vectors

$$\{\mathcal{B}^*\chi_{3n+1},\mathcal{B}^*\chi_{3n+2}\}_{n=1}^{\infty}.$$

To conclude that the same is true of the gradients g_n for the direct spectral map, we need the following lemma.

LEMMA 6.10. *The kernel of $\mathcal{B}$ is the one-dimensional subspace of $L^2[0,1]\times L^2[0,1]$ spanned by the vector $\widehat{g}_0=(0,1)^T$.*

This is a direct computation using the formulas for $\mathcal{B}$ and its boundary conditions recorded in Appendix B.

Now consider the family of vectors $\{\tilde{g}_n\}_{n=0}^{\infty}$, where $\tilde{g}_0=(0,1)$ and $\tilde{g}_n=(n^2\pi^2)^{-1}\widehat{g}_n-\tilde{g}_0$. Since $\tilde{g}_0\in\ker(\mathcal{B})$, we have the biorthogonality relations

$$\langle\mathcal{B}^*\chi_n,\tilde{g}_m\rangle=c_n\delta_{nm}$$

for $n\in\mathbb{N}$, $c_n>0$, and $m\in 0\cup\mathbb{N}$. Thus the family $\{\tilde{g}_n\}_{n=0}^{\infty}$ is linearly independent. Since the gradients g_n are obtained from $\tilde{g}_n$ by permuting the first and second entries, we have proved the following theorem.

THEOREM 6.11. *For any $p\in\mathcal{D}$, the gradients $\{g_n(x;p)\}_{n=0}^{\infty}$ are linearly independent. Moreover, the complement of their span has infinite dimension in E.*

7. Proofs of the main theorems. We now prove Theorem 1.3, first proving that $d\mu(q)$ is a linear isomorphism from $\mathcal{E}_v(q)$ onto F, and then proving that the space $\mathcal{E}_h(q)$ is complementary. In light of the discussion following the statement of Theorem 1.3, and in view of Theorem 5.1 and the estimates of section 4, the first assertion will be proved once the following result is established.

LEMMA 7.1. *For each $p \in \mathcal{D}$, $d\mu(p)$ is a linear isomorphism from $\mathcal{E}_v(p)$ onto F.*

In proving this result, we will make use of some results on Riesz bases. Recall that if $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, a basis $\{e_n\}$ is called a *Riesz basis* for $\mathcal{H}$ if there exist $a, b \in \mathbb{R}^+$ for which

$$a\|h\|^2 \leq \sum_n |\langle h, e_n \rangle|^2 \leq b\|h\|^2 \quad \forall\, h \in \mathcal{H}\,. \tag{7.1}$$

LEMMA 7.2. *Let $\mathcal{H}$ be a Hilbert space, and let $\{e_n\}$ be a Riesz basis for $\mathcal{H}$. Then, there is a unique set $\{\epsilon_m\} \subseteq \mathcal{H}$ for which*

(1) *$\langle e_n, \epsilon_m \rangle = \delta_{m,n}$ for all $m, n \in \mathbb{N}$,*

(2) *There exist $\alpha, \beta \in \mathbb{R}^+$ so that*

$$\alpha\|h\|^2 \leq \sum_m |\langle h, \epsilon_m \rangle|^2 \leq \beta\|h\|^2 \quad \forall h \in \mathcal{H}\,.$$

Proof. Let $Th = \sum_n \langle h, e_n \rangle e_n$. Then, T is self-adjoint, and

$$\langle h, Th \rangle = \left\langle h, \sum_n \langle h, e_n \rangle e_n \right\rangle = \sum_n |\langle h, e_n \rangle|^2.$$

It then follows from (7.1) that the spectrum of T is contained in the interval $[a, b]$ so that T^{-1} exists. By the spectral theorem, we have

$$b^{-2}\|h\|^2 \leq \|T^{-1}h\|^2 \leq a^{-2}\|h\|^2. \tag{7.2}$$

Let $\epsilon_n \equiv T^{-1}e_n$ for each $n \in \mathbb{N}$. Then

$$e_n = T\epsilon_n = \sum_m \langle \epsilon_n, e_m \rangle e_m,$$

so $\langle \epsilon_m, e_n \rangle = \delta_{nm}$ by the linear independence of the e_n; this establishes (1). Finally, it follows from relation (7.2) and the definition of ϵ_n that $\sum_n |\langle h, \epsilon_n \rangle|^2 = \sum_n |\langle T^{-1}h, e_n \rangle|^2$ obeys the inequality

$$ab^{-2}\|h\|^2 \leq \sum_n |\langle h, \epsilon_n \rangle|^2 \leq ba^{-2}\|h\|^2\,,$$

which establishes (2). □

LEMMA 7.3. *Let $\{d_n\}$ be a Riesz basis for $\mathcal{H}$. Then the linear map $\mathcal{A} : \mathcal{H} \mapsto \ell^2(\mathbb{N})$ defined by*

$$\mathcal{A}x = \{\langle x, d_n \rangle\}_{n \geq 1}\,,\ x \in \mathcal{H},$$

is an isomorphism.

Proof. We will show that $\mathcal{A}$ is a bounded bijection from $\mathcal{H}$ to ℓ^2. The conclusion of the lemma will then follow from the open mapping theorem. First, it is clear that

$\mathcal{A}x = 0$ only when $\langle x, d_n\rangle = 0$ for each n. Since $\{d_n\}$ is a basis for $\mathcal{H}$, we must have $x \equiv 0$ so that $\mathcal{A}$ is one-to-one. Further, for $x \in \mathcal{H}$ we have, from (7.1),

$$\|\mathcal{A}x\|^2 = \|\{\langle x, d_n\rangle\}\|_{\ell^2}^2 = \sum_n |\langle x, d_n\rangle|^2 \leq b\|x\|^2 ,$$

so $\mathcal{A}$ is bounded.

To show $\mathcal{A}$ is onto $\ell^2(\mathbb{N})$, we introduce $\{\delta_m\}$ as the Riesz basis biorthogonal to $\{d_n\}$, the existence of which is guaranteed by Lemma 7.2. Then, given $\{y_m\} \in \ell^2(\mathbb{N})$, set $y = \sum_m y_m \delta_m$. From (7.1),

$$\|y\|^2 \leq \frac{1}{a}\sum_n |\langle y, d_n\rangle|^2 = \frac{1}{a}\sum_n \left|\sum_m y_m \langle \delta_m, d_n\rangle\right|^2 = \frac{1}{a}\sum_n |y_n|^2 < \infty$$

so that $y \in \mathcal{H}$. Also,

$$\mathcal{A}y = \{\langle y, d_n\rangle\}_{n\geq 1} = \left\{\sum_m y_m \langle \delta_m, d_n\rangle\right\}_{n\geq 1} = \{y_n\}_{n\geq 1} ,$$

which shows that $\mathcal{A}$ maps onto $\ell^2(\mathbb{N})$. By the open mapping theorem, $\mathcal{A}$ is an isomorphism. □

For each $n \in \mathbb{N}$, let z_n and u_n denote, respectively, the nth eigenfunction of $L(p)$ and $L(0)$, and as before let

$$g_n(x,p) = \begin{pmatrix} \dfrac{z_n'(x,p)^2}{n^2\pi^2} - 1 \\ \\ \dfrac{z_n(x;p)^2}{n^2\pi^2} \end{pmatrix}.$$

By explicit computation

$$g_n(x,0) = \begin{pmatrix} \cos(2n\pi x) \\ \\ \dfrac{1-\cos(2n\pi x)}{n^2\pi^2} \end{pmatrix}. \tag{7.3}$$

Recalling the form (1.6) of $d\mu(p)$, it suffices, by virtue of Lemma 7.3, to show that $\{g_n(\,\cdot\,;p)\}$ is a Riesz basis for $\mathcal{E}_v(p) = \operatorname{span}\{g_n(\,\cdot\,;p)\}$. Using Fourier theory, one can show directly that $\{g_n(\,\cdot\,;0)\}$ is a Riesz basis for its span. To show that the same is true for $\{g_n(\,\cdot\,;p)\}$, we note that $\{g_n(\,\cdot\,;p)\}$ is linearly independent, by virtue of Theorem 6.11, and use the following stability result for Riesz bases.

LEMMA 7.4. *Let $\mathcal{H}$ be a Hilbert space, and let $\{e_n\} \subseteq \mathcal{H}$ be a Riesz basis for $\mathcal{H}_e \equiv \operatorname{span}\{e_n\}$. Let $\{d_n\} \subseteq \mathcal{H}$ be a linearly independent set for which*

$$\sum_n \|d_n - e_n\|^2 \equiv M < \infty.$$

Then, $\{d_n\}$ is a Riesz basis for $\mathcal{H}_d \equiv \operatorname{span}\{d_n\}$; i.e., there exist $\alpha, \beta \in \mathbb{R}^+$ so that, for each $h \in \mathcal{H}_d$,

$$\alpha\|h\|^2 \leq \sum_n |\langle h, d_n\rangle|^2 \leq \beta\|h\|^2. \tag{7.4}$$

Proof. Define a map $A : \mathcal{H}_e \longrightarrow \mathcal{H}_d$ by $Ae_n = d_n$ for each $n \in \mathbb{N}$, extended by linearity. Then, for $h \in \mathcal{H}_d$ and each n,

$$\langle h, d_n\rangle = \langle h, Ae_n\rangle = \langle A^*h, e_n\rangle. \tag{7.5}$$

Let a and b be the constants for which (7.1) holds for $\{e_n\}$ on $\mathcal{H}_e$. Then, from (7.4),

$$a\|A^*h\|^2 \le \sum_n |\langle A^*h, e_n\rangle|^2 = \sum_n |\langle h, d_n\rangle|^2 \le b\|A^*h\|^2. \tag{7.6}$$

We claim that A (and hence A^*) is boundedly invertible. If this is true, then (7.6) leads to

$$\frac{a}{\|(A^*)^{-1}\|^2}\|h\|^2 \le \sum_n |\langle h, d_n\rangle|^2 \le b\|A^*\|^2\|h\|^2,$$

which establishes (7.4).

To show A is invertible, we note that the linear independence of $\{d_n\}$ implies the injectivity of A. To see that the range of A is all of $\mathcal{H}_d$, note that any $h \in \mathcal{H}_d$ can be written as $h = \sum_n h_n d_n$, where $\{h_n\} \in \ell^2(\mathbb{N})$. Then, setting $x \equiv \sum_n h_n e_n$, one easily sees that $x \in \mathcal{H}_e$ and $Ax = h$. Hence, A maps $\mathcal{H}_e$ onto $\mathcal{H}_d$.

Finally, we show that A is bounded. Choose $x \in \mathcal{H}_e$, and write as

$$x = \sum_n x_n e_n = \sum_n \langle x, \epsilon_n\rangle e_n,$$

where $\{\epsilon_n\}$ is the Riesz basis for $\mathcal{H}_e$ which is biorthogonal to $\{e_n\}$. Then,

$$Ax = \sum_n x_n d_n = \sum_n x_n e_n + \sum_n x_n(d_n - e_n) = x + \sum_n x_n(d_n - e_n),$$

which yields

$$\begin{aligned} \|Ax\|^2 &\le \|x\|^2 + \left(\textstyle\sum_n |x_n|^2\right)\left(\textstyle\sum_n \|d_n - e_n\|^2\right) \\ &= \|x\|^2 + M\textstyle\sum_n |\langle x, \epsilon_n\rangle|^2 \\ &\le \|x\|^2 + M\beta\|x\|^2\,, \end{aligned}$$

where (7.4) was used in the last inequality. Thus,

$$\|A\|^2 \le 1 + M\beta < \infty\,,$$

and A is bounded. By the open mapping theorem, A has a bounded inverse, as asserted. □

Proof of Lemma 7.1. From the estimates of Theorem 4.1, one can show that

$$\sum_n \|g_n(\,\cdot\,;p) - g_n(\,\cdot\,;0)\|^2 < \infty\,, \tag{7.7}$$

so, by Lemma 7.4, $\{g_n\}$ is a Riesz basis for $\mathcal{E}_v(p)$. The result now follows from Lemma 7.3. □

Finally, we prove that for any $q \in \mathcal{E}$, the complementary space $\mathcal{E}_h(q)$ is infinite-dimensional. We first observe that, by the explicit formula (7.3) and Fourier analysis, the gradients $g_n(\cdot, 0)$ are orthogonal vectors, and the complementary space $\mathcal{E}_h(0)$ has infinite dimension. Since the gradients satisfy (7.7), the second part of Theorem 1.3 will follow from the next lemma.

LEMMA 7.5. *Let $\{v_n\}$ be an orthogonal set of vectors in a Hilbert space $\mathcal{H}$. Let $\{w_n\} \subset \mathcal{H}$ be a linearly independent set of vectors which satisfy $\sum_n \|v_n - w_n\|^2 < \infty$. Set $V \equiv \text{span}\{v_n\}$ and $W \equiv \text{span}\{w_n\}$. If $V^\perp$ has infinite dimension, then $W^\perp$ also has infinite dimension.*

Proof. Suppose not, and choose an infinite sequence of orthogonal unit vectors $\{e_n\}$ from $V^\perp$ so that $e_n \to 0$ weakly. Let $\epsilon > 0$ be given. Writing $e_n = P_W e_n + P_{W^\perp} e_n$ we see that $P_{W^\perp} e_n \to 0$ if $W^\perp$ has finite dimension. We will obtain a contradiction by showing that $\|P_W e_n\|$ is also small for large n. First observe that by hypothesis, for M sufficiently large and all n,

$$\sum_{m=M+1}^{\infty} |\langle e_n, w_m \rangle|^2 < \epsilon.$$

On the other hand, using the weak convergence again,

$$\sum_{m=1}^{M} |\langle e_n, w_m \rangle|^2 \to 0$$

as $n \to \infty$ for a fixed M. It follows that for any $\epsilon > 0$,

$$\limsup_{n \to \infty} \left(\sum_m |\langle e_n, w_m \rangle|^2 \right) \leq \epsilon.$$

Since $\{w_n\}$ is a Riesz basis, by virtue of Lemma 7.4, this means that $\limsup_{n\to\infty} \|P_W e_n\| \leq \epsilon$, a contradiction. ∎

Appendix A. The free Green's function and free resolvent operator. In this appendix, we prove some useful technical estimates on the Green's kernel for the differential operator $L_\#(0)$, where $L_\#(0)$ is one of the operators $L(0)$, $L_1(0)$, or $L_2(0)$ defined in (1.3)–(1.5). Let $G_0^\#(x, y; z)$ be the integral kernel of $(L_\#(0) - z)^{-1}$, and let $C^\#_{R,\alpha}$ be defined as in section 2.

In what follows, $\beta = (\beta_x, \beta_y)$, where β_x and β_y are nonnegative integers, and $\partial^\beta = \partial_x^{\beta_x} \partial_y^{\beta_y}$. We wish to derive estimates on $\partial^\beta G_0^\#(x, y; z)$. These estimates will involve series of the form

$$\sum_{n=1}^{\infty} \frac{D_x^{\beta_x} u_n^\#(x) D_y^{\beta_y} u_n^\#(y)}{\lambda_n^\#(0) - z}, \tag{A.1}$$

where $\lambda_n^\#(0)$ and $u_n^\#$ are the eigenvalues and eigenfunctions, respectively, of $L_\#(0)$. It can be verified directly that for each set of boundary conditions, the eigenfunctions obey

$$|D_x^j u_n^\#(x)| \leq C n^j, \quad 0 \leq j \leq 2,$$

for some constant C. Thus, to estimate series of the form (A.1), it suffices to majorize the numerical series

$$\sum_{n=1}^{\infty} \frac{n^{|\beta|}}{|\lambda_n^{\#}(0) - z|}.$$

In so doing, we will require the following technical result, which may be easily verified.

LEMMA A.1. *Let $R > 1$ and $\alpha \in (2,3)$. Then, for $m > (8R)^{\frac{1}{3-\alpha}}$,*

$$\int_1^{m-1} \frac{t^2}{m^4\pi^4 - t^4\pi^4}\, dt \le \frac{1}{m\pi^4} \ln m, \tag{A.2}$$

$$\int_{m+2}^{\infty} \frac{t^2}{t^4\pi^4 - [(m+1)^4\pi^4 + R(m+1)^{\alpha}]}\, dt \le \frac{C}{m}\, \ln m\,, \tag{A.3}$$

where the constant C depends only on α.

First of all, we have the following lemma.

LEMMA A.2. *Let $R > 0$, let $\alpha \in (2,3)$, and let N be an integer satisfying the bounds* (2.1). *Then for any $z \notin C_{R,\alpha}^{\#}$:*

(a) *for $|\beta| \le 1$,*

$$\sup_{x,y\in[0,1]} |\partial^{\beta} G_0^{\#}(x,y;z)| \le c_{|\beta|} R^{-1};$$

(b) *for $|\beta| = 2$,*

$$\sup_{x,y\in[0,1]} |\partial^{\beta} G_0^{\#}(x,y;z)| \le c_2 R^{-1} \ln(R),$$

where c_2 is a numerical constant depending only on α.

Proof. We consider the case $\lambda_n(0) = n^4\pi^4$, the computations for other boundary conditions being similar.

For $|\beta| \le 1$ we have the simple majorization

$$\sum_{n=1}^{\infty} \frac{n}{|n^4\pi^4 - z|} \le \sum_{n=1}^{\infty} \frac{n}{Rn^{\alpha}}\,,$$

which gives estimates of the desired form.

For $|\beta| = 2$, we seek to estimate the series

$$\sum_{n=1}^{\infty} \frac{n^2}{|n^4\pi^4 - z|}$$

for $z \notin C_{R,\alpha}$. We split the sum into

$$T_1 = \sum_{n=1}^{N} \frac{n^2}{|n^4\pi^4 - z|}$$

and

$$T_2 = \sum_{n=N+1}^{\infty} \frac{n^2}{|n^4\pi^4 - z|}\,,$$

where $N > (8R)^{\frac{1}{3-\alpha}}$.

To estimate T_1, we use the fact that for $1 \leq n \leq N$, $|z - n^4\pi^4| \geq |z_N - n^4\pi^4|$ where $z_N = N^4\pi^4 + RN^\alpha$. The integral test then yields the bound

$$\begin{aligned} T_1 &\leq \sum_{n=1}^{N-2} \frac{n^2}{|n^4\pi^4 - N^4\pi^4|} + \frac{(N-1)^2}{z_N - (N-1)^4\pi^4} \\ &\quad + \frac{N^2}{|N^4\pi^4 - z_N|} \\ &\leq \int_1^{N-1} \frac{t^2}{N^4\pi^4 - t^4\pi^4} + CR^{-1}\,. \end{aligned}$$

Using (A.2) and the fact that $N = O(R^{\frac{1}{3-\alpha}})$, we conclude that $T_1 \leq C(\alpha)R^{-1}\ln R$.

To estimate T_2, we consider the two cases $\Re(z) \leq z_N$ and $\Re(z) \geq z_N$ separately. If $\Re(z) \leq z_N$, then $|m^4\pi^4 - z| \geq |m^4\pi^4 - z_N|$, from which we obtain

$$T_2 \leq \sum_{n=N+1}^{\infty} \frac{n^2}{\pi^4 n^4 - [\pi^4 N^4 + RN^\alpha]}.$$

We may estimate this sum by

$$\frac{(N+1)^2}{\pi^4(N+1)^2 - z_N} + \int_{N+1}^{\infty} \frac{t^2}{t^4\pi^4 - [\pi^4 N^4 + CRN^\alpha]}\,dt\,,$$

which, in conjunction with (A.3), yields an estimate of the desired form.

If $\Re(z) > z_N$, we divide the half-plane $\Re(z) > z_N$ into strips

$$S_m = \{z \in \mathbb{C} : \Re(z) \in [m^4\pi^4 + Rm^\alpha, (m+1)^4\pi^4 + R(m+1)^\alpha)\}$$

and fix m so that $z \in S_m$. We can then estimate T_2 by letting $x = \Re(z)$ and splitting

$$\begin{aligned} \sum_{n=N+1}^{\infty} \frac{n^2}{|n^4\pi^4 - z|} &= \sum_{n=N+1}^{m-2} \frac{n^2}{|n^4\pi^4 - z|} \\ &\quad + \left(\frac{(m-1)^2}{z - (m-1)^4\pi^4} + \frac{m^2}{|z - m^4\pi^4|} \right. \\ &\quad \left. + \frac{(m+1)^2}{|z - (m+1)^4\pi^4|} + \frac{(m+2)^2}{|z - (m+2)^4\pi^4|} \right) \\ &\quad + \sum_{n=m+3}^{\infty} \frac{n^2}{n^4\pi^4 - x} \\ &= T_{21} + T_{22} + T_{23}\,. \end{aligned}$$

From the definition of $C_{R,\alpha}$, it is clear that

$$T_{22} \leq CR^{-1}\,,$$

which bounds T_{22}. Finally, using the fact that, for x in the interval $[m^4\pi^4+Rm^\alpha, (m+1)^4\pi^4 + R(m+1)^\alpha)$,

$$T_{21} \le \int_0^{m-1} \frac{t^2}{m^4\pi^4 - t^4\pi^4}\, dt,$$
$$T_{23} \le \int_{m+2}^{\infty} \frac{t^2}{t^4\pi^4 - [(m+1)^4\pi^4 + R(m+1)^\alpha]}\, dt\,,$$

we can use (A.2) and (A.3) to bound T_{21} and T_{23}, respectively. □

We will also need estimates on the free resolvent kernel on contours surrounding the sets $E_m^\#$.

LEMMA A.3. *Let $R > 0$, let $\alpha \in (2,3)$, let N be an integer satisfying the bounds (2.1), and let $E_m^\#$ be defined as in section 2. Let γ_m be the contour bounding $E_m^\#$. If $m > N$ and $|\beta| \le 2$, then the estimate*

$$\sup_{\substack{x,\,y\,\in\,[0,1]\\ z\,\in\,\gamma_m}} |\partial^\beta G_0^\#(x,y;z)| \le C_{|\beta|} m^{|\beta|-\alpha}$$

holds. Here $C_{|\beta|}$ is a numerical constant which diverges as $\alpha \uparrow 3$.

Proof. Here again we consider the case $\lambda_n(0) = n^4\pi^4$ and $|\beta| = 2$. We must majorize the numerical series

$$\sum_{n=1}^{\infty} \frac{n^2}{|z - n^4\pi^4|},$$

where $|z - m^4\pi^4| = Rm^\alpha$. We split the sum into

$$\sum\nolimits_{n=1}^{m-2} \frac{n^2}{|z - n^4\pi^4|}$$
$$+ \left[\frac{(m-1)^2}{|z-(m-1)^4\pi^4|} + \frac{m^2}{|z-m^4\pi^4|} + \frac{(m+1)^2}{|z-(m+1)^4\pi^4|}\right]$$
$$+ \sum\nolimits_{n=m+2}^{\infty} \frac{n^2}{|z - n^4\pi^4|}.$$

The three bracketed terms are easily estimated by $6Rm^{2-\alpha}$. We can estimate the first and last terms using (A.2) and (A.3). □

Now we derive estimates on operators involving compositions of the resolvent $(L_\#(0) - z)^{-1}$ with the operator of differentiation, D, and the operator of multiplication by a function $r \in C_0^\infty(0,1)$. These compositions are initially defined on $C_0^\infty(0,1)$ and extended by density to bounded operators. It follows from this definition and an integration by parts that the operator $D^{\beta_x}(L(0) - z)^{-1}D^{\beta_y}$ has integral kernel $(-1)^{\beta_y}(\partial^\beta G_0)(x,y;z)$. From the kernel estimates in Lemma A.2 and integration by parts, the following estimates are easily demonstrated.

LEMMA A.4. *Suppose that $r \in C_0^\infty(0,1)$. For $z \notin C_{R,\alpha}$, there exist $c_0, c_1, c_2 \in \mathbb{R}$ so that the following estimates hold:*

(a) $\|(L_\#(0) - z)^{-1}\| \le c_0 R^{-1}$.
(b) $\|(L_\#(0) - z)^{-1} r\| \le c_0 R^{-1} \|r\|_2$.
(c) $\|D(L_\#(0) - z)^{-1}\| \le c_1 R^{-1}$.

(d) $\|(L_\#(0)-z)^{-1}Dr\| \le c_1 R^{-1}\|r\|_2$.
(e) $\|D(L_\#(0)-z)^{-1}Dr\| \le c_2 R^{-1}\ln(R)\|r\|_2$.

The following bounds are used to estimate the resolvent $(L_\#(p)-z)^{-1}$ for $z \in \gamma_m$, the contour determined by the boundary of the set E_m defined in section 2. We denote by $\|A\|_{p,q}$ the norm of the linear operator A from $L^p[0,1]$ to $L^q[0,1]$, where $1 \le p,q \le \infty$. Using the strong pointwise estimates on the free resolvent and its partial derivatives from Lemma A.3, we easily obtain the following lemma.

LEMMA A.5. *Let $z \in \gamma_m$ and $r \in C_0^\infty(0,1)$, and let $0 \le i,j \le 1$. Then for any p,q with $1 \le p,q \le \infty$, the estimate*

$$\|D^i(L_\#(0)-z)^{-1}D^j\|_{p,q} \le C_{i+j} m^{i+j-\alpha}$$

holds, and for any p,q with $1 \le p \le 2$ and $1 \le q \le \infty$, the estimate

$$\|D^i(L_\#(0)-z)^{-1}D^j r\|_{p,q} \le C_{i+j} m^{i+j-\alpha}\|r\|_2$$

holds. Here C_{i+j} is the numerical constant defined in Lemma A.3.

Appendix B. Domains related to the Borg system. In this appendix, we collect some useful calculations regarding the Borg system $\mathcal{M}-\lambda\mathcal{B}$ of section 6. First of all, for $p=(p_1,p_2) \in \mathcal{D}$, the matrix-valued differential operators $\mathcal{M}$ and $\mathcal{B}$ which define the Borg system can be computed to be

$$\mathcal{M} = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} \quad \textit{and} \quad \mathcal{B} = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix},$$

where

$$\begin{aligned}
M_1 &= D^5 - p_1 D^3 - 2p_1' D^2 + \left(\frac{8}{3}p_2 - p_1''\right) D + 2p_1', \\
M_2 &= -\frac{10}{3}D^3 + \frac{4}{3}p_1 D + \frac{2}{3}p_1', \\
M_3 &= -3p_1' D^4 + 10p_2 D^3 + (3p_1 p_1' + 5p_2')D^2 + \left(p_2'' - 4p_1 p_2 + 3(p_1')^2\right) D - 6p_1' p_2, \\
M_4 &= D^5 - 5p_1 D^3 - p_1' D^2 + (4p_1^2 - 8p_1'' - 24p_2)D + (2p_1 p_1' - 2p_1''' - 6p_2'), \\
B_1 &= \frac{8}{3}D, \\
B_2 &= 0, \\
B_3 &= 10D^3 + 4p_1 D - 6p_1', \\
B_4 &= -24D.
\end{aligned}$$

Boundary conditions are to be specified so that if w is an eigenfunction (with eigenvalue λ) of one of the three underlying fourth-order problems (1.3), (1.4), (1.5), then $z=[w^2,\,(w')^2]^T$ will be in the kernel of $\mathcal{M}-\lambda\mathcal{B}$. The three fourth-order operators $L(p)$, $L_1(p)$, and $L_2(p)$ underlying the Borg system carry the following boundary conditions (see section 1):

$$L(p): w(0) = w''(0) = w(1) = w''(1) = 0,$$

$$L_1(p): w(0) = w''(0) = w'(1) = w'''(1) = 0,$$

$$L_2(p): w(0) = w''(0) = w(1) = w'(1) = 0.$$

Setting $u \equiv w^2$ and $v \equiv (w')^2$, one can compute directly that the boundary conditions which determine the domain of $\mathcal{M}$ are:

- At $x = 0$: $\quad u = u' = u''' = v' = v''' = u'' - 2v = u^{(4)} - 4v'' = 0,$
- At $x = 1$: $\quad u' = u''' = v' = 0.$

The operator $\mathcal{B}$ is to be viewed as a perturbation of $\mathcal{M}$, so we take $D(\mathcal{B}) = D(\mathcal{M})$, from which it follows that $D(\mathcal{M} - \lambda\mathcal{B}) = D(\mathcal{M})$.

It will also be useful to determine the boundary conditions which define the domains of certain adjoint operators related to the Borg system. First, consider $D(\mathcal{M}^*)$, where $\mathcal{M}^*$ denotes the Hilbert space adjoint of the unbounded operator $\mathcal{M}$.

Let $z = [u,\, v]^T \in D(\mathcal{M})$. An element $\sigma = [\phi,\, \psi]^T \in D(\mathcal{M}^*)$ must satisfy

$$\langle \mathcal{M}z, \sigma\rangle - \langle z, \mathcal{M}^\dagger\sigma\rangle = 0\,,$$

where $\mathcal{M}^\dagger$ denotes the formal adjoint of $\mathcal{M}$. One can compute directly the following boundary conditions which define $D(\mathcal{M}^*)$:

- At $x = 0$: $\quad \psi = \psi'' + \frac{2}{3}\phi = \psi^{(4)} - \frac{4}{3}\phi'' + \frac{8}{3}p_1\phi = 0,$
- At $x = 1$: $\quad \phi = \phi'' = \phi^{(4)} = \psi = \psi' = \psi'' = \psi^{(4)} = 0.$

Similarly, one can compute the following boundary conditions which define $D(\mathcal{B}^*)$:

- At $x = 0$: $\quad \psi = 0,$
- At $x = 1$: $\quad \psi = 5\psi'' + \frac{4}{3}\phi = 0.$

One immediately observes the following containment:

$$D(\mathcal{M}^*) \subset D(\mathcal{B}^*)\,.$$

Finally, proceeding as above, one may compute the following set of boundary conditions which defines $D((\mathcal{M} - \lambda\mathcal{B})^*)$:

- At $x = 0$: $\quad \psi = \psi'' + \frac{2}{3}\phi = \psi^{(4)} - \frac{4}{3}\phi'' + \frac{8}{3}p_1\phi = 0,$
- At $x = 1$: $\quad \phi = \phi'' = \phi^{(4)} = \psi = \psi' = \psi'' = \psi^{(4)} = 0\,,$

which shows that $D((\mathcal{M} - \lambda\mathcal{B})^*) = D(\mathcal{M}^*)$.

Acknowledgments. The authors would like to acknowledge the hospitality of the Institute for Mathematics and its Applications during part of the time that this work was done, and to thank David Leep and Serge Ochanine for helpful conversations on the algebraic lemmas in section 6.

REFERENCES

[1] V. Barcilon, *On the uniqueness of inverse eigenvalue problems*, J. Roy. Astr. Soc., 38 (1974), pp. 287–298.
[2] V. Barcilon, *On the solution of inverse eigenvalue problems of higher orders*, Geophys. J. R. Astr. Soc., 39 (1974), pp. 143–154.
[3] V. Barcilon, *Inverse problem for a vibrating beam*, J. Appl. Math. Phys. (ZAMP), 27 (1976), pp. 347–358.
[4] V. Barcilon, *On the multiplicity of solutions of the inverse problem for a vibrating beam*, SIAM J. Appl. Math., 37 (1979), pp. 605–613.
[5] V. Barcilon, *Inverse problem for the vibrating beam in the free-clamped configuration*, Philos. Trans. Roy. Soc. London Ser. A, 304 (1982), pp. 211–251.
[6] V. Barcilon, *Inverse Eigenvalue Problems*, in Inverse Problems, Lecture Notes in Math. 1225, Springer, New York, 1986, pp. 1–51.
[7] V. Barcilon, *Sufficient conditions for the solution of the inverse problem for a vibrating beam*, Inverse Problems, 3 (1989), pp. 181–193.
[8] N. K. Bari, *Sur les bases dans l'espace de Hilbert*, Dokl. Akad. Nauk SSSR, 54 (1946), pp. 379–382.

[9] G. Borg, *On the completeness of some sets of functions*, Acta. Math., 81 (1949), pp. 265–283.
[10] C. Coleman and J. McLaughlin, *Solution of the inverse spectral problem for an impedance with integrable derivative*, I, II, Comm. Pure Appl. Math., 46 (1993), pp. 145–184 and 185–212.
[11] B. Dahlberg and E. Trubowitz, *The inverse Sturm–Liouville problem* III, Comm. Pure Appl. Math., 37 (1984), pp. 255–267.
[12] W. N. Everitt, *The Sturm–Liouville problem for fourth-order ordinary differential equations*, Quart. J. Math, 8 (1957), pp. 146–160.
[13] G. M. L. Gladwell, *The inverse problem for the Euler–Bernoulli beam*, Proc. Roy. Soc. London Ser. A, 407 (1986), pp. 199–218.
[14] G. M. L. Gladwell, *Inverse Problems in Vibration*. M. Nijhoff, Boston, MA, 1986.
[15] G. M. L. Gladwell and S. R. A. Dods, *Examples of reconstruction of vibrating rods from spectral data*, J. Sound Vibration, 119 (1987), pp. 267–276.
[16] G. M. L. Gladwell, A. H. England, and D. Wang, *Examples of reconstruction of an Euler–Bernoulli beam from spectral data*, J. Sound Vibration, 119 (1987), pp. 81–94.
[17] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, Heidelberg, Berlin, 1976.
[18] E. Isaacson, H. McKean, and E. Trubowitz, *The inverse Sturm–Liouville problem* II, Comm. Pure Appl. Math., 37 (1984), pp. 1–11.
[19] E. Isaacson and E. Trubowitz, *The inverse Sturm–Liouville problem* I, Comm. Pure Appl. Math., 36 (1983), pp. 767–783.
[20] J. R. McLaughlin, *An inverse eigenvalue problem of order four*, SIAM J. Math. Anal., 7 (1976), pp. 646–661.
[21] J. McLaughlin, *An inverse eigenvalue problem of order four—an infinite case*, SIAM J. Math. Anal., 9 (1978), pp. 395–413.
[22] J. McLaughlin, *Fourth-order inverse eigenvalue problems*, in Spectral Theory of Differential Operators, I. W. Knowles and R. Lewis, eds., North–Holland, Amsterdam, 1981.
[23] J. McLaughlin, *Bounds for constructed solutions of second- and fourth-order inverse eigenvalue problems*, in Differential Equations (Birmingham, Alabama, 1983), North–Holland, Amsterdam, 1984, pp. 437–443.
[24] J. McLaughlin, *Analytical methods for recovering coefficients in differential equations from spectral data*, SIAM Rev., 28 (1986), pp. 53–72.
[25] M. Naimark, *Linear Differential Operators, Parts* I *and* II, Ungar, New York, 1967 and 1968.
[26] J. Pöschel and E. Trubowitz, *Inverse Spectral Theory*, Pure and Applied Mathematics 130, Academic Press, New York, 1987.
[27] J. Ralston and E. Trubowitz, *Isospectral sets for boundary value problems on the unit interval*, Ergodic Theory Dynamical Systems, 8 (1988), pp. 301–358.
[28] M. Reed and B. Simon, *Methods of Modern Mathematical Physics,* IV. *Analysis of Operators*, Academic Press, New York, San Francisco, London, 1978.
[29] A. W. Schueller, *Eigenvalue Asymptotics for Fourth-Order Self-Adjoint Ordinary Differential Operators*, Ph.D. thesis, University of Kentucky, Lexington, June, 1996. Available via anonymous ftp from http://schuelaw.whitman.edu.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 967–1001, July 1998

009

ASYMPTOTICS FOR THE SPECTRUM OF A FLUID/STRUCTURE HYBRID SYSTEM ARISING IN THE CONTROL OF NOISE*

SORIN MICU† AND ENRIQUE ZUAZUA‡

Abstract. We consider a simple model arising in the control of noise consisting of two coupled hyperbolic equations of dimensions two and one, respectively. The one-dimensional equation is assumed to be dissipative. We describe the asymptotic behavior of the eigenvalues and eigenfunctions of the system at high frequencies. Some other interesting features of the model, like the exponential decay of solutions or the compactness of the damping term, are also studied.

Key words. eigenvalues, eigenfunctions, high frequency asymptotics, hyperbolic system, aeromechanic structure interaction

AMS subject classifications. 35P20, 35L20, 73K70

PII. S0036141096312349

1. Introduction.

Recently several works both in the mathematical and technical literature have dealt with the problem of the active control of noise generated in acoustic cavities by means of the vibrations of their flexible walls. Such studies were motivated, for instance, by the development of a new class of turboprop engines which are very fuel efficient but also very loud. In this context the low frequency high magnitude acoustic fields produced by these engines cause vibrations in the fuselage which in turn generate unwanted interior noise.

In this article we analyze the spectral properties of a linear two-dimensional hybrid system arising in the development of these new technologies for noise reduction in the interior of a cavity (plane, car, etc.) which was proposed in a series of works by Banks et al. (see [3]).

Let us describe the system we study. We consider the two-dimensional square $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. We assume that Ω is filled with an elastic, inviscid, compressible fluid whose velocity field $\vec{v}$ is given by the potential $\phi = \phi(x,y,t)$, ($\vec{v} = \nabla\phi$). By linearization we assume that the potential ϕ satisfies the linear wave equation in $\Omega \times (0,\infty)$ (see [8]).

The boundary $\Gamma = \partial\Omega$ of Ω is divided in two parts: $\Gamma_0 = \{(x,0) : x \in (0,1)\}$ and $\Gamma_1 = \Gamma \backslash \Gamma_0$. The subset Γ_1 is assumed to be rigid and we impose zero normal velocity of the fluid on it. The subset Γ_0 is supposed to be flexible and occupied by a flexible string that vibrates under the pressure of the fluid on the plane where Ω lies. The displacement of Γ_0, described by the scalar function $W = W(x,t)$, obeys the one-dimensional dissipative wave equation. On the other hand, on Γ_0 we impose the continuity of the normal velocities of the fluid and the string. The string is assumed to satisfy Neumann boundary conditions on its extremes.

* Received by the editors November 20, 1996; accepted for publication (in revised form) July 23, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/31234.html

† Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense, 28040 Madrid, Spain and Facultatea de Matematica-Informatica, Universitatea din Craiova, 1100, Romania (sorin@sunma4.mat.ucm.es). The research of this author was partially supported by grant 132/1995 of CNCSU (Romania) and grant CHRX-CT94-0471 of the European Union.

‡ Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense, 28040 Madrid, Spain (zuazua@sunma4.mat.ucm.es). The research of this author was supported by grant PB93-1203 of the DGICYT (Spain) and grant CHRX-CT94-0471 of the European Union.

All deformations are supposed to be small enough so that linear theory applies.

Under natural initial conditions for ϕ and W the linear motion of this system is described by means of the following coupled wave equations:

$$
\left\{
\begin{array}{lll}
\phi_{tt} - \Delta\phi = 0 & \text{in} & \Omega \times (0,\infty), \\
\frac{\partial \phi}{\partial \nu} = 0 & \text{on} & \Gamma_1 \times (0,\infty), \\
\frac{\partial \phi}{\partial y} = -W_t & \text{on} & \Gamma_0 \times (0,\infty), \\
W_{tt} - W_{xx} + W_t + \phi_t = 0 & \text{on} & \Gamma_0 \times (0,\infty), \\
W_x(0,t) = W_x(1,t) = 0 & \text{for} & t > 0, \\
\phi(0) = \phi^0, \phi_t(0) = \phi^1 & \text{in} & \Omega, \\
W(0) = W^0, W_t(0) = W^1 & \text{on} & \Gamma_0.
\end{array}
\right. \tag{1.1}
$$

By ν we denote the unit outward normal to Ω.

In (1.1) we have chosen to take the various parameters of the system to be equal to one. This restricts the generality of our analysis. The dependence of the most interesting features of the spectrum with respect to the various parameters of the system will be studied elsewhere.

We remark also that, in (1.1), two wave equations, of dimensions two and one, respectively, and representing vibrations of different nature, are coupled. Therefore we say that (1.1) is a two-dimensional hybrid system. For examples of hybrid systems of dimension one, such as those coupling strings or beams with rigid bodies, see [10], [7], and [18].

System (1.1) is a modified version of the one introduced by Banks et al. in [3] (see also [4]). In [3] the flexible part of the boundary Γ_0 is assumed to be occupied by a flexible beam, leading to a fourth-order one-dimensional equation on Γ_0. We have chosen to consider a one-dimensional wave equation instead to simplify the exposition. However, most of the relevant spectral properties remain unchanged considering a beam equation with appropriate boundary conditions.

We also remark that we choose Neumann boundary conditions for the string. This choice allows us to separate the variables and to obtain an explicit equation for the eigenvalues. In the case of Dirichlet boundary conditions, which are considered in [3], this is no longer possible. Nevertheless, using the information we get here about the eigenfunctions of system (1.1), it can be proved that the uniform decay fails and also that there exist solutions uniformly distributed in Ω with arbitrarily small decay (see [14]).

System (1.1) is well posed in the energy space

$$\mathcal{X} = H^1(\Omega) \times L^2(\Omega) \times H^1(\Gamma_0) \times L^2(\Gamma_0)$$

for the variables (ϕ, ϕ_t, W, W_t).

The energy

$$E(t) = \frac{1}{2}\int_\Omega \left[|\nabla\phi|^2 + |\phi_t|^2 \right] dxdy + \frac{1}{2}\int_{\Gamma_0} \left[|W_x|^2 + |W_t|^2 \right] dx \tag{1.2}$$

satisfies

$$\frac{dE}{dt}(t) = -\int_{\Gamma_0} |W_t|^2 \, dx. \tag{1.3}$$

Hence, the system (1.1) is dissipative, the damping term being localized in the subset Γ_0 of the boundary.

Some of the properties of this system like existence, uniqueness, asymptotic behavior, and existence of periodic solutions were studied in previous works (see [12] and [13]).

Our aim here is to characterize the asymptotic behavior of the eigenvalues and eigenfunctions of the differential operator corresponding to system (1.1) and to describe some interesting features of the model that are direct consequences of this analysis. The study is made by using separation of variables. In this way the system is reduced to an infinite number of one-dimensional systems depending on an integer parameter k which represents the frequency of vibration in the x-direction. This allows us to obtain explicit equations for eigenvalues and to use Rouché's theorem for their localization.

Let us describe briefly the most relevant results obtained in this paper:

(a) Whenever the frequency of vibration in the $x-$direction is fixed the corresponding one-dimensional system does not decay uniformly. Indeed, at high frequencies, the real part of the one-parameter family of eigenvalues converges to zero. This is a typical situation in one-dimensional hybrid systems (see [7], [10], and [18]).

(b) The effect of the damping term on the global dynamics of the system is almost negligible at high frequencies. Indeed, most of the eigenfunctions of the system (1.1) have their energy uniformly distributed in Ω while the real part of the eigenvalues converges to zero at high frequencies.

(c) Among the two-parameter family of eigenvalues of the two-dimensional system only a one-parameter family of them is effectively damped so that their real parts remain uniformly away from zero. The corresponding eigenfunctions have their energy exponentially concentrated on the string Γ_0.

(d) As a consequence of the previous property, the difference between the semigroup generated by the damped and undamped systems is not compact. This is in contrast with the results in [18] showing that the lack of uniform decay in damped one-dimensional hybrid systems is typically due to the compactness of the damping term. Thus, the noncompactness result is genuinely two-dimensional.

Let us remark that the case we have addressed is not generic. Even in the case of surfaces of revolution the cylindrical case is a degenerate one. This was exhibited in the thesis of Allibert in the frame of the classical wave equation with Dirichlet boundary conditions (see [1]). Nevertheless, in [11] we show that, in the case of a disk-shaped cavity surrounded by a circular dissipative string, the same phenomenon is present although all rays of geometric optics meet the boundary where the losses occur. This indicates that the same behavior can be expected for different kinds of geometries or boundary conditions (see also [14]).

The rest of the paper is organized as follows.

In section 2 we present in detail the main results of this paper and we discuss some of their consequences. In section 3 we localize the eigenvalues of the undamped system corresponding to (1.1) and describe its eigenfunctions. In section 4 we obtain asymptotic estimates for the eigenvalues and eigenmodes of the damped system (1.1). This section is divided in two parts. In section 4.1 we distinguish three types of eigenvalues which, at high frequencies, approach the imaginary axis. The corresponding eigenfunctions have the property that the energy concentrated in the string vanishes asymptotically. To complete the study, in section 4.2 we prove that there exists a sequence of eigenvalues, tending to infinity, with uniformly negative real parts. The

corresponding eigenfunctions have the property that the energy localized on the string does not vanish asymptotically. Moreover, as the frequency increases the whole energy is concentrated on the string at an exponential rate. These eigenfunctions span an infinite-dimensional subspace of the energy space in which the decay rate of solutions is exponential.

In the last section we prove that the difference between the semigroup generated by the differential operator associated with the undamped system and that associated with the damped one is not compact as a consequence of the existence of an infinite-dimensional subspace in which the damping term is effective; i.e., it produces an exponential decay. We end up with an appendix that contains some technical lemmas.

2. The main results: Statements and discussion. As we said in the introduction the aim of this paper is the study of the spectrum of (1.1). In this section we state the main results concerning the eigenvalues and eigenfunctions of the system and some of their consequences.

In order to analyze the spectrum of (1.1) we look for solutions in separated variables of the form $(\phi, W) = (\psi(y,t), V(t))\cos(k\pi x)$.

We deduce that $(\psi(y,t), V(t))$ verifies the following one-dimensional system:

$$\left\{\begin{array}{lll} \psi_{tt} - \psi_{yy} + k^2\pi^2\psi = 0 & \text{in} & (0,1)\times(0,\infty), \\ \psi_y(1) = 0 & \text{for} & t \in (0,\infty), \\ \psi_y(0) = -V_t & \text{for} & t \in (0,\infty), \\ V_{tt} + k^2\pi^2 V + V_t + \psi_t(0) = 0 & \text{for} & t \in (0,\infty). \end{array}\right. \tag{2.1}$$

Now if we look for solutions of (2.1) of the form $(\psi(y,t), V(t)) = e^{\lambda t}(\psi(y), V)$, with $V \in \mathbb{R}$, it follows that the eigenvalues λ of system (1.1) are the roots of the equation

$$e^{2\sqrt{\lambda^2+k^2\pi^2}} = -\frac{\lambda^2 - \sqrt{\lambda^2+k^2\pi^2}(\lambda^2+\lambda+k^2\pi^2)}{\lambda^2 + \sqrt{\lambda^2+k^2\pi^2}(\lambda^2+\lambda+k^2\pi^2)}. \tag{2.2}$$

The corresponding eigenfunctions are $\varphi_\lambda = \psi_\lambda \cos(k\pi x)$ where ψ_λ are the eigenfunctions of (2.1):

$$\psi_\lambda = \begin{pmatrix} \dfrac{1}{\lambda}\cosh\left(\sqrt{\lambda^2+k^2\pi^2}(y-1)\right) \\ \cosh\left(\sqrt{\lambda^2+k^2\pi^2}(y-1)\right) \\ \dfrac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\sinh\left(\sqrt{\lambda^2+k^2\pi^2}\right) \\ \dfrac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\sinh\left(\sqrt{\lambda^2+k^2\pi^2}\right) \end{pmatrix}. \tag{2.3}$$

We are interested in the asymptotic behavior of the eigenvalues λ when $|\lambda| \longrightarrow \infty$. For each $k \in \mathbb{N}$ we get a sequence of eigenvalues $(\lambda_{k,m})_{m\in\mathbb{Z}^*}$ for the system (2.1) of modulus greater than $k\pi$ (that will be analyzed in section 4.1) and two eigenvalues λ_k^* and λ_k^{**} with modulus less than $k\pi$ (that will be studied in section 4.2). All these are the eigenvalues of system (1.1). For each k, $(\lambda_{k,m})_{m\in\mathbb{N}^*}$ are ordered such that $|\lambda_{k,m}|$ increases as m does and $\lambda_{k,-m} = \bar{\lambda}_{k,m}$ if $m \in \mathbb{N}^*$. The general result on the existence of eigenvalues is given in the following theorem.

THEOREM 2.1. *Let $k \in \mathbb{N}$ be fixed. The spectrum of the differential operator corresponding to system* (2.1) *consists of a sequence of eigenvalues* $(\lambda_{k,m})_{m\in\mathbb{N}^*} \cup \{\lambda_k^*\}$

with positive imaginary part and another sequence of eigenvalues $(\lambda_{k,-m})_{m\in\mathbb{N}^*}\cup\{\lambda_k^{**}\}$ *with the property that* $\lambda_{k,-m}=\bar{\lambda}_{k,m}$ *if* $m>0$ *and* $\lambda_k^{**}=\bar{\lambda}_k^*$. *All these eigenvalues are zeros of the equation* (2.2). *If* $k=0$ *then* $\lambda_k^*=\lambda_k^{**}=0$.

Remark 1. We remark that the notation λ_k^* and λ_k^{**} are used for the eigenvalues with the smallest modulus of the system. We make this distinction since the properties of these wave numbers are different from the others, as we shall see in Theorem 2.9. Actually, these eigenvalues correspond to the eigenfunctions whose energy is concentrated on the string Γ_0 and decay uniformly as $t\to\infty$.

The asymptotic properties of the wave numbers and modes depend on the relation between k and m. Therefore we divide our analysis in four cases. First, in Theorems 2.2, 2.4, and 2.6, we characterize the eigenvalues that approach the imaginary axis as the wave number increases. These are the eigenvalues $(\lambda_{k,m})_{m\in\mathbb{Z}^*}$. Then, in Theorem 2.9, we study the eigenvalues λ_k^* and λ_k^{**} which have a uniformly negative real part.

THEOREM 2.2 (eigenvalues $\lambda_{k,m}$ with $|\lambda_{k,m}|\geq\sqrt{2}k\,\pi$). *Let* $k\in\mathbb{N}$ *be fixed. The eigenvalues* $\lambda_{k,m}$ *of* (2.1) *with* $\mid\lambda_{k,m}\mid>\sqrt{2}\pi\,k$ *approach the imaginary axis when* $\mid m\mid\to\infty$ *and satisfy the following:*

$$\begin{aligned}&\mid\lambda_{k,m}-\sqrt{k^2+m^2}\,\pi\,i\mid\leq\frac{24}{\sqrt{m^2+k^2}\,\pi}\quad\text{if }\mathcal{I}m\,\lambda_{k,m}>0\qquad(m>k>0),\\&\mid\lambda_{k,m}+\sqrt{k^2+m^2}\,\pi\,i\mid\leq\frac{24}{\sqrt{m^2+k^2}\,\pi}\quad\text{if }\mathcal{I}m\,\lambda_{k,m}<0\qquad(m<-k<0).\end{aligned}\tag{2.4}$$

Remark 2. Theorem 2.2 shows that when we fix the frequency of vibration in the x-direction (k fixed) and we consider large frequencies in the y-direction (m large), the system behaves like the wave equation in Ω with homogeneous Neumann boundary conditions in all $\partial\Omega$:

$$\begin{cases}\Phi_{tt}-\triangle\Phi=0 & \text{in}\quad\Omega\times(0,\infty),\\ \frac{\partial\Phi}{\partial\nu}=0 & \text{on}\quad\partial\Omega\times(0,\infty).\end{cases}\tag{2.5}$$

Therefore, the influence of the vibrating string on Γ_0 vanishes asymptotically.

Remark 3. Note that the existence of a sequence of eigenvalues $(\lambda_{k,m})_m$ which approach the imaginary axis when $\mid m\mid\to\infty$ implies that the decay rate of the energy of solutions of (1.1) is not exponential. It is known that, for linear problems, this is equivalent to a nonuniform decay rate of solutions (see [9]).

In fact we obtain that, for each $k\in\mathbb{N}$, the system (2.1) does not have an exponential decay. This is not the case in the classical wave equation with boundary dissipation:

$$\begin{cases}\Phi_{tt}-\triangle\Phi=0 & \text{in}\quad\Omega\times(0,\infty),\\ \frac{\partial\Phi}{\partial\nu}=0 & \text{on}\quad\Gamma_1\times(0,\infty),\\ \frac{\partial\Phi}{\partial\nu}+\Phi_t=0 & \text{on}\quad\Gamma_0\times(0,\infty).\end{cases}\tag{2.6}$$

In the context of (2.6), for k fixed, the corresponding one-dimensional systems have exponential decay, but the decay rate vanishes as $k\to\infty$. This is due to the fact that the region Γ_0 in which the damping is concentrated does not satisfy the necessary geometric control condition since there are rays of geometric optics that never intersect Γ_0 (see [5] and [17]). In our case the loss of uniform decay is even worse, and it is due to the hybrid structure of the system or, equivalently, to the type of boundary condition we have imposed on Γ_0 and not only to the support Γ_0 of the damping term.

Moreover, as we mention in Remark 6, we can find a sequence of solutions of (1.1) with the energy uniformly distributed in all Ω and with arbitrarily small exponential decay rate. This is not possible in the examples given in [5] and [17] where the energy of the solutions with nonuniform decay concentrates on rays of geometric optics.

Remark 4. The fact that the eigenvalues approach the imaginary axis is a consequence not only of the localization of the dissipative region but also of the hybrid structure of the system. In [11] we show that, in the case of a disk-shaped cavity surrounded by a circular dissipative string the same phenomenon is present although all rays of geometric optics meet the boundary where the losses occur.

This indicates that the same behavior can be expected for different kinds of geometries or boundary conditions (see also [14]).

We can now analyze the eigenfunctions corresponding to the wave numbers $\lambda_{k,m}$ of Theorem 2.2. Remark 2 indicates that one can expect the first two components of the eigenfunctions of (1.1) to behave like the eigenfunctions of (2.5). Therefore we define the function

$$\psi_{k,m} = \begin{pmatrix} \dfrac{(-1)^{m+1} i}{\sqrt{k^2+m^2}\pi} \cos m\pi y \cos k\pi x \\ (-1)^{m+1} \cos m\pi y \cos k\pi x \\ 0 \\ 0 \end{pmatrix}. \tag{2.7}$$

Observe that the eigenmodes of (2.5) are the first two components of $\psi_{k,m}$.

THEOREM 2.3. *The eigenfunctions φ_λ, corresponding to the eigenvalues $\lambda = \lambda_{k,m}$ satisfying* (2.4) *have the following property:*

$$\| \varphi_\lambda - \psi_{k,m} \|_{\mathcal{X}} \leq \frac{c}{m}, \tag{2.8}$$

where c is a constant which does not depend on m and k.

Remark 5. Theorem 2.5 indicates that the last two components of the eigenfunction φ_λ (which correspond to the string located in Γ_0) vanish asymptotically when the frequency increases. This implies that, at high frequencies (in the sense of (2.4)), the string does not play an important role in the dynamics of the system.

Remark 6. The solutions of (1.1) corresponding to the eigenfunctions given by Theorem 2.3 form a sequence of solutions with the energy uniformly distributed in all Ω and with arbitrarily small exponential decay rate. This proves that the lack of the uniform decay of our system is related not only to the support of the dissipative mechanism but also to the nature of the boundary conditions or of the coupling between the different components of the system.

The second range of frequencies is studied in the following theorem.

THEOREM 2.4. *[*eigenvalues $\lambda_{k,m}$ with $k\pi \leq |\lambda_{k,m}| \leq \sqrt{2}k\pi$, first part*] For $k \in \mathbb{N}$ sufficiently large and $m = \pm 1, \pm 2, ..., \pm[\sqrt[3]{k}]$, the eigenvalues $\lambda_{k,m}$ of* (1.1) *satisfy*

$$\begin{aligned} &\left| \lambda_{k,m} - \sqrt{k^2 + \left(\frac{2m-1}{2}\right)^2}\pi\, i \right| \leq \frac{2\pi}{\sqrt[3]{k}} \text{ if } \mathcal{I}m\,\lambda_{k,m} > 0 \quad (1 \leq m \leq [\sqrt[3]{k}]), \\ &\left| \lambda_{k,m} + \sqrt{k^2 + \left(\frac{2m+1}{2}\right)^2}\pi\, i \right| \leq \frac{2\pi}{\sqrt[3]{k}} \text{ if } \mathcal{I}m\,\lambda_{k,m} < 0 \quad (-[\sqrt[3]{k}] \leq m < 0). \end{aligned} \tag{2.9}$$

Remark 7. Consider the following conservative wave equation:

$$\begin{cases} \Phi_{tt} - \triangle\Phi = 0 & \text{in} \quad \Omega \times (0,\infty), \\ \frac{\partial\Phi}{\partial\nu} = 0 & \text{on} \quad \Gamma_1 \times (0,\infty), \\ \Phi = 0 & \text{on} \quad \Gamma_0 \times (0,\infty). \end{cases} \tag{2.10}$$

Its eigenvalues are exactly $\sqrt{k^2 + \left(\frac{2m+1}{2}\right)^2}\pi\, i$. Theorem 2.4 shows that when we fix the frequency of vibration in the y-direction (m is fixed) and we consider large frequencies in the x-direction (k large), the eigenvalues of (1.1) behave like those of (2.10). The influence of the vibrating string on Γ_0 vanishes asymptotically in this range of eigenvalues. However, when comparing the behavior of these eigenvalues with those of Theorem 2.2, we observe that the boundary conditions for Φ on Γ_0 change.

Let us analyze the eigenfunctions corresponding to the eigenvalues studied in Theorem 2.4. We consider first the function

$$\tilde{\psi}_{k,m} = \begin{pmatrix} \dfrac{(-1)^{m+1}\, i}{\sqrt{k^2 + \left(\dfrac{2m+1}{2}\right)^2}\,\pi}\ \sin\ \dfrac{2m+1}{2}\pi y\ \cos\ k\pi x \\ (-1)^{m+1}\ \sin\ \dfrac{2m+1}{2}\pi y\ \cos\ k\pi x \\ 0 \\ 0 \end{pmatrix} \tag{2.11}$$

and we remark that the first two components of it correspond to eigenfunctions of problem (2.10).

THEOREM 2.5. *The eigenfunctions φ_λ corresponding to the eigenvalues $\lambda = \lambda_{k,m}$ of Theorem* 2.4 *satisfy*

$$\| \varphi_\lambda - \tilde{\psi}_{k,m} \|_{\mathcal{X}} \leq \frac{c}{\sqrt[3]{k}}, \tag{2.12}$$

where c is a constant which does not depend on k and m.

Remark 8. Remark 5 applies in this case too.

The following theorem completes the study of the eigenvalues with real part tending to zero as the wave number increases.

THEOREM 2.6 (eigenvalues $\lambda_{k,m}$ with $k\,\pi \leq |\lambda_{k,m}| \leq \sqrt{2}k\,\pi$, second part). *For all $k \in \mathbb{N}$ sufficiently large the eigenvalues $\lambda_{k,m}$ of* (1.1) *with $[\sqrt[3]{k}] < |m| \leq k$ satisfy the following estimates:*

$$\begin{aligned} &\left|\lambda_{k,m} - \sqrt{\pi^2\, k^2 + k^2\, \zeta_{k,m}^2}\, i\right| \leq \frac{1}{\sqrt[5]{k}} \ \text{if } \mathcal{I}m\, \lambda_{k,m} > 0 \quad (k \geq m > [\sqrt[3]{k}]), \\ &\left|\lambda_{k,m} + \sqrt{\pi^2\, k^2 + k^2\, \zeta_{k,m}^2}\, i\right| \leq \frac{1}{\sqrt[5]{k}} \ \text{if } \mathcal{I}m\ \lambda_{k,m} < 0 \quad (-k \leq m < -[\sqrt[3]{k}]), \end{aligned} \tag{2.13}$$

where $\zeta_{k,m} \in \mathbb{R}_+$ is the positive root of the equation

$$\tan\ k\,\zeta\ = \frac{\pi^2}{k\,\zeta^3}, \tag{2.14}$$

which belongs to $\left(\frac{m}{k}\pi, \frac{2m+1}{2k}\pi\right)$.

Remark 9. When k remains bounded and m goes to infinity the roots $\zeta_{k,m}$ of the equation (2.14) behave like $\frac{m\pi}{k}$. This corresponds to the asymptotic behavior of the eigenvalues $\lambda_{k,m}$ studied in Theorem 2.2. On the other hand, when m remains bounded and k goes to infinity, the zeros $\zeta_{k,m}$ of (2.14) behave like $\frac{(2m+1)\pi}{2k}$. This agrees with the behavior of the eigenvalues $\lambda_{k,m}$ studied in Theorem 2.4.

The eigenvalues $\lambda_{k,m}$ of Theorem 2.6 make the transition from one zone to another and still approach the imaginary axis at high frequencies.

The eigenfunctions corresponding to these eigenvalues have the same property as those of Theorems 2.3 and 2.5; i.e., the last two components vanish asymptotically.

THEOREM 2.7. *The eigenfunctions* φ_λ *corresponding to the eigenvalues of Theorem* 2.6 *satisfy*

$$\lim_{|\lambda|\to\infty} \frac{||\varphi_\lambda^3||_{H^1(\Gamma_0)}}{||\varphi_\lambda||_{\mathcal{X}}} = 0, \quad \lim_{|\lambda|\to\infty} \frac{||\varphi_\lambda^4||_{L^2(\Gamma_0)}}{||\varphi_\lambda||_{\mathcal{X}}} = 0, \tag{2.15}$$

where φ_λ^3 *and* φ_λ^4 *are the third and the fourth components of* φ_λ.

Until now we have obtained eigenvalues of system (1.1) approaching the imaginary axis when their modulus tends to infinity. The following result exhibits a sequence of eigenvalues with uniformly bounded negative real parts.

THEOREM 2.8. *[*eigenvalues λ_k with $|\lambda_k| \le k\,\pi$*] The equation* (2.2) *has, for sufficiently large* k, *two eigenvalues* λ_k^* *and* λ_k^{**} *with* $\mathcal{I}m\,\lambda_k^* > 0$ *and*

$$\left| \lambda_k^* - \sqrt{k^2(\alpha_1)^2 - k^2\pi^2} \right| \le \frac{1}{k} \quad \textit{and} \quad \lambda_k^* = \bar{\lambda}_k^{**}, \tag{2.16}$$

where α_1 *is the root of*

$$z^2 - \pi^2 + kz^3 + z\sqrt{z^2 - \pi^2} = 0 \tag{2.17}$$

with the following asymptotic behavior:

$$\alpha_1 = \sqrt[3]{\frac{\pi^2}{k}} - \frac{1}{3}\sqrt[3]{\frac{\pi}{k^2}}\, i + o\left(\frac{1}{\sqrt[3]{k^2}}\right), \ \textit{as } k \to \infty. \tag{2.18}$$

Therefore, λ_k^* *satisfies*

$$\mathcal{R}e\,\lambda_k^* \longrightarrow -\frac{1}{3} \textit{ when } k \longrightarrow \infty. \tag{2.19}$$

Remark 10. In Theorem 2.8 we prove the existence of two eigenvalues λ_k^* and λ_k^{**} with modulus less than $k\pi$. These are, for k fixed, the eigenvalues with smallest modulus and are the only ones uniformly dissipated by the system at large frequencies.

The corresponding eigenfunctions λ_k^* can be written as

$$\varphi_{\lambda_k^*} = \begin{pmatrix} \dfrac{\cosh\left(\sqrt{(\lambda_k^*)^2 + k^2\pi^2}(y-1)\right)\cos k\pi x}{\sqrt{(\lambda_k^*)^2 + k^2\pi^2}\,\sinh\left(\sqrt{(\lambda_k^*)^2 + k^2\pi^2}\right)} \\ -\dfrac{\lambda_k^*\cosh\left(\sqrt{(\lambda_k^*)^2 + k^2\pi^2}(y-1)\right)\cos k\pi x}{\sqrt{(\lambda_k^*)^2 + k^2\pi^2}\,\sinh\left(\sqrt{(\lambda_k^*)^2 + k^2\pi^2}\right)} \\ -\dfrac{1}{\lambda_k^*}\cos k\pi x \\ \cos k\pi x \end{pmatrix},$$

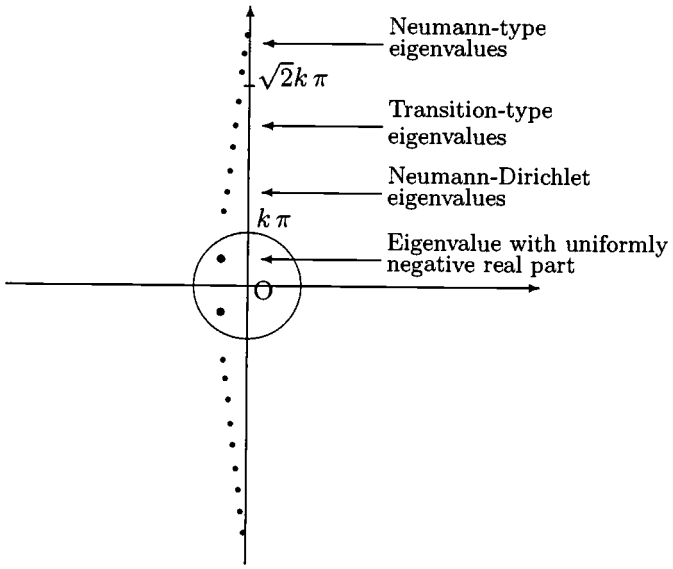

FIG. 2.1. *The sequence of eigenvalues for fixed* k.

and they have a different behavior.

THEOREM 2.9. (i) *The sequence of eigenfunctions* $\{\varphi_{\lambda_k^*}\}_k$ *converge weakly to zero in* $\mathcal{X}$ *when* k *tends to infinity.*

(ii) *The sequences* $\{\varphi^j_{\lambda_k^*}\}_k$ *do not converge strongly to zero for any* $j = 1, 2, 3, 4$ *in the corresponding norms.*

Remark 11. The eigenfunctions $\varphi_{\lambda_k^*}$ generate a subspace of the energy space of infinite dimension in which, in view of (2.19), the decay rate of the energy of the system is uniform. The energy of the solutions corresponding to the eigenfunctions of Theorem 2.9 is concentrated in the string. Indeed, the estimates of Theorem 2.9 allow us to prove that

$$\int_0^1 \int_\varepsilon^1 \left(||\varphi^1_{\lambda_k^*}||^2_{H^1(\Omega)} + ||\varphi^2_{\lambda_k^*}||^2_{L^2(\Omega)} \right) dx\, dy \leq Ce^{-2\sqrt[3]{k^2\pi^2}\varepsilon}.$$

This indicates that the energy of the acoustic wave decays exponentially fast from Γ_0 to the interior of the domain.

Figure 2.1 describes the behavior of the different families of eigenvalues for each k.

3. The conservative system. In this section we analyze the spectral properties of the conservative system corresponding to (1.1):

$$\left\{ \begin{array}{lll} \phi_{tt} - \Delta\phi = 0 & \text{in} & \Omega \times (0,\infty), \\ \frac{\partial\phi}{\partial\nu} = 0 & \text{on} & \Gamma_1 \times (0,\infty), \\ \frac{\partial\phi}{\partial y} = -W_t & \text{on} & \Gamma_0 \times (0,\infty), \\ W_{tt} - W_{xx} + \phi_t = 0 & \text{on} & \Gamma_0 \times (0,\infty), \\ W_x(0,t) = W_x(1,t) = 0 & \text{for} & t > 0. \end{array} \right. \tag{3.1}$$

In (3.1) the dissipative term W_t of the equation of displacement of the string has been dropped. The energy of this system is defined by (1.2) too, but in this case we have that $\frac{dE}{dt}(t) = 0$. This means that (3.1) is an undamped system.

The eigenvalues of (3.1) are characterized in the following theorem.

THEOREM 3.1. *System* (3.1) *has a two-parameter sequence of purely imaginary eigenvalues* $(\nu_{k,m})_{k\in\mathbb{N},\, m\in\mathbb{Z}^*}$ *given by*

$$\nu_{k,m} = \sqrt{z_{k,m}^2 + k^2\pi^2}\; i \quad \text{if } m > 0 \quad \text{and} \quad \nu_{k,m} = -\nu_{k,-m} \quad \text{if } m < 0\,, \tag{3.2}$$

where $(z_{k,m})_{m\in\mathbb{N}^*}$ *are the roots of the equation*

$$\tan z = \frac{z^2 + k^2\pi^2}{z^3}. \tag{3.3}$$

Moreover, there are another two eigenvalues of (3.1), ν_k^*, *and* ν_k^{**}, *with the modulus less than* $k\,\pi$, *given by*

$$\nu_k^* = \sqrt{k^2\pi^2 - (z_k^*)^2}\; i\,, \quad \nu_k^{**} = \bar{\nu}_k^*\,, \tag{3.4}$$

where z_k^* *is the unique positive root of the equation*

$$e^{2\,z} = \frac{z^3 - z^2 + k^2\pi^2}{z^3 + z^2 - k^2\pi^2}. \tag{3.5}$$

In the last case, $\nu_k^* = \nu_k^{**} = 0$ *when* $k = 0$.

Proof. In order to study the spectrum of (3.1) we look for solutions of this system in separated variables: $(\phi, W) = e^{\nu t}(\psi, V)\cos(n\pi x)$, where $\psi = \psi(y)$ and $V \in \mathbb{R}$. It follows that the eigenvalues ν satisfy the following transcendental equation:

$$e^{2\sqrt{\nu^2 + k^2\pi^2}} = -\frac{\nu^2 - \sqrt{\nu^2 + k^2\pi^2}(\nu^2 + k^2\pi^2)}{\nu^2 + \sqrt{\nu^2 + k^2\pi^2}(\nu^2 + k^2\pi^2)}. \tag{3.6}$$

Considering the change of variable $\nu = \sqrt{\zeta^2 - k^2\pi^2}$, equation (3.6) becomes

$$e^{2\zeta} = \frac{\zeta^3 - \zeta^2 + k^2\pi^2}{\zeta^3 + \zeta^2 - k^2\pi^2}. \tag{3.7}$$

Since the differential operator corresponding to (3.1) is conservative its eigenvalues will be all purely imaginary. Hence, we have to look only for those roots of (3.7) which are purely imaginary or real. It follows that the imaginary roots of (3.7) are the roots of the equation (3.3) and the real ones are roots of (3.5). □

We analyze now the eigenfunctions. By separation of variables, it is easy to see that the eigenfunctions have the following form:

$$\xi_\nu = \begin{pmatrix} \dfrac{-i}{\sqrt{z^2 + k^2\pi^2}}\cos\, z(y-1)\, \cos\, k\pi x \\ -\cos\, z(y-1)\, \cos\, k\pi x \\ -\dfrac{z}{z^2 + k^2\pi^2}\, \sin\, z\, \cos\, k\pi x \\ \dfrac{z\, i}{\sqrt{z^2 + k^2\pi^2}}\, \sin\, z\, \cos\, k\pi x \end{pmatrix}. \tag{3.8}$$

THEOREM 3.2. *The eigenfunctions* ξ_ν *defined by* (3.8) *corresponding to the eigenvalues* ν *given by* (3.3) *have the following property:*

$$\lim_{|\nu|\to\infty} \frac{||\xi_\nu^3||_{H^1(0,1)}}{||\xi_\nu||_{\mathcal{X}}} = 0, \quad \lim_{|\nu|\to\infty} \frac{||\xi_\nu^4||_{L^2(0,1)}}{||\xi_\nu||_{\mathcal{X}}} = 0,$$

where ξ_ν^j is the jth component of ξ_ν.

Proof. If ν is one of the eigenvalues of (3.1) with $|\nu| > k\,\pi$, it follows that $\zeta = \sqrt{\nu^2 + k^2\pi^2}$ is a purely imaginary number. Therefore $\zeta = z\,i$, where $z \in \mathbb{R}$ is a solution of the equation (3.3).

Taking into account that z satisfies (3.3), a simple calculation gives us that

$$||\xi_\nu^1||_{H^1}^2 + ||\xi_\nu^2||_{L^2}^2 = \frac{1}{2} + \frac{1}{4(z^2+k^2\pi^2)} + \frac{(1+2k^2\pi^2)\,\sin 2z}{8z(z^2+k^2\pi^2)}$$

$$= \frac{1}{2} + \frac{1}{4(z^2+k^2\pi^2)} + \frac{2z^3(z^2+k^2\pi^2)}{4(z^6+(z^2+k^2\pi^2)^2)},$$

$$||\xi_\nu^3||_{H^1}^2 = \frac{z^2(1+k^2\pi^2)\,\sin^2 z}{2(z^2+k^2\pi^2)^2} = \frac{z^2(1+k^2\pi^2)}{2(z^6+(z^2+k^2\pi^2)^2)},$$

$$||\xi_\nu^4||_{L^2}^2 = \frac{z^2\,\sin^2 z}{2(z^2+k^2\pi^2)} = \frac{z^2(z^2+k^2)}{2(z^6+(z^2+k^2\pi^2)^2)}.$$

We observe that if k remains bounded when $|\nu| \to \infty$ then, necessarily, $|z| \to \infty$. This remark allows us to conclude that

$$||\xi_\nu^1||_{H^1}^2 + ||\xi_\nu^2||_{L^2}^2 \longrightarrow \frac{1}{2} \text{ and } ||\xi_\nu^3||_{H^1}^2 + ||\xi_\nu^4||_{L^2}^2 \longrightarrow 0, \text{ when } \nu \longrightarrow \infty. \qquad \square$$

Remark 12. One can also see that ν_k^* does not have this property; i.e.,

$$\liminf_{|\nu_k^*|\to\infty} \frac{||\xi_{\nu_k^*}^3||_{H^1(0,1)}}{||\xi_{\nu_k^*}||_{\mathcal{X}}} \neq 0 \text{ and } \liminf_{|\nu|\to\infty} \frac{||\xi_{\nu_k^*}^4||_{L^2(0,1)}}{||\xi_{\nu_k^*}||_{\mathcal{X}}} \neq 0.$$

The proof of this fact is similar to that of Theorem 2.9.

4. The dissipative case. In this section we give the proofs of Theorems 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, and 2.9 which characterize the asymptotic behavior of the eigenvalues and the eigenfunctions of (1.1).

We begin with the proof of Theorem 2.1.

Proof of Theorem 2.1. Suppose first that $k \neq 0$. It is easy to see that the differential operator corresponding to (2.1) has compact resolvent (see [11]). Therefore, the spectrum of (2.1) consists of a sequence of complex eigenvalues $(\lambda_{k,m})_{m\in\mathbb{N}} \cup (\bar{\lambda}_{k,m})_{m\in\mathbb{N}}$ with the property that $\lim_{m\to\infty} |\lambda_{k,m}| = \infty$ and $\lambda_{k,m} \neq 0$ for all $m \in \mathbb{Z}$.

If $k = 0$, the operator has the same properties but the first two eigenvalues $\lambda_{0,0}$ and $\bar{\lambda}_{0,0}$ are equal to zero.

Moreover, since all the elements of the spectrum are eigenvalues of the operator it follows that they are roots of equation (2.2). $\square$

With the change of variable $\sqrt{\left(\frac{\lambda}{k}\right)^2 + \pi^2} = z$ equation (2.2) is reduced to

$$e^{2k\,z} = -\frac{z^2 - \pi^2 - kz^3 - z\sqrt{z^2-\pi^2}}{z^2 - \pi^2 + kz^3 + z\sqrt{z^2-\pi^2}}. \tag{4.1}$$

We present now four technical lemmas which give us the information we need about the poles of the function in the right-hand side of (4.1). The proofs of these lemmas will be presented in an appendix at the end of this paper.

LEMMA 4.1. *If α is a root of the equation*

$$z^2 - \pi^2 + kz^3 + z\sqrt{z^2-\pi^2} = 0, \tag{4.2}$$

then, for k large enough, we have

$$\frac{\pi}{2\sqrt[3]{k}} <\mid \alpha \mid< \frac{2\pi}{\sqrt[3]{k}}. \tag{4.3}$$

LEMMA 4.2. *For k large enough, the equation (4.2) has three roots α_i, $i=1,2,3$, with the property that*

$$\left|\alpha_i - \sqrt[3]{\frac{\pi^2}{k}}\omega_i\right| \le \frac{10}{\sqrt[3]{k^2}}, \tag{4.4}$$

where ω_i, $i=1,2,3$, are the three cubic roots of unity.

LEMMA 4.3. *The root α_1 of (4.2) satisfies*

$$\alpha_1 = \sqrt[3]{\frac{\pi^2}{k}} - \frac{1}{3}\sqrt[3]{\frac{\pi}{k^2}}\,i + o\left(\frac{1}{\sqrt[3]{k^2}}\right) \quad \text{as } k\to\infty. \tag{4.5}$$

LEMMA 4.4. *For k large enough, the equation*

$$z^2 - \pi^2 - kz^3 - z\sqrt{z^2-\pi^2} = 0 \tag{4.6}$$

has three roots β_i, $i=1,2,3$, with the property that

$$\left|\beta_i - \sqrt[3]{\frac{\pi^2}{k}}\tilde{\omega}_i\right| \le \frac{10}{\sqrt[3]{k^2}}, \tag{4.7}$$

where $\tilde{\omega}_i = -\omega_i$, $i=1,2,3$.

We can pass now to prove Theorems 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, and 2.9. In section 4.1 we analyze the case of the eigenvalues with real parts tending to zero, as the frequency increases (Theorems 2.2, 2.3, 2.4, 2.5, 2.6, and 2.7). In section 4.2 we prove the existence of eigenvalues with uniformly negative real parts (Theorems 2.8 and 2.9).

4.1. Eigenvalues with real parts tending to zero.

Proof of Theorem 2.2. If we note $\sqrt{\lambda^2 + k^2\pi^2} = \mu$, we obtain that μ satisfies the following equation:

$$e^{2\mu} = -\frac{\mu^2 - k^2\pi^2 - \mu(\mu^2 + \sqrt{\mu^2 - k^2\pi^2})}{\mu^2 - k^2\pi^2 + \mu(\mu^2 + \sqrt{\mu^2 - k^2\pi^2})}. \tag{4.8}$$

We put the equation (4.8) in the form

$$e^{2\mu} - 1 = -\frac{2(\mu^2 - k^2\pi^2)}{\mu^2 - k^2\pi^2 + \mu(\mu^2 + \sqrt{\mu^2 - k^2\pi^2})}, \tag{4.9}$$

and we localize its roots applying Rouché's theorem.

In order to do this we consider the functions

$$f(z) = e^{2z} - 1 \text{ and } g(z) = -\frac{2(z^2 - k^2\pi^2)}{z^2 - k^2\pi^2 + z(z^2 + \sqrt{z^2 - k^2\pi^2})}.$$

We remark that the equation $f(z) = 0$ has the roots $(\alpha_m)_{m\in\mathbb{Z}}$ with $\alpha_m = m\pi\, i$.

For each $m \in \mathbb{Z} \setminus \{0\}$ we define the square γ_m^1 of center α_m and side $2\varepsilon_m$ and the rectangle γ_m^2 defined by the lines $\mathcal{R}e\, z = \pm\delta_m$ and $\mathcal{I}m\, z = m\pi \pm \frac{3\pi}{4}$. Moreover, we consider the square γ^0 of center 0 and side $2M_k$ (see Fig. 4.1).

The constants ε_m, δ_m, and M_k will be chosen in such a way that

$$| f(z) | > | g(z) | \quad \text{for all } z \in \gamma_m^1 \cup \gamma_m^2 \cup \gamma^0. \tag{4.10}$$

First of all we have that, for all $z \in \mathbb{C}$,

$$\begin{aligned} | f(z) |^2 = | e^{2z} - 1 |^2 &= (e^{2\mathcal{R}e\, z} - \cos 2\,\mathcal{I}m\, z)^2 + (\sin 2\,\mathcal{I}m\, z)^2 \\ &\geq \max\{| e^{2\mathcal{R}e\, z} - 1 |, | \sin 2\,\mathcal{I}m\, z |\}. \end{aligned} \tag{4.11}$$

In order to estimate g we consider the region G_1 of the complex plane defined by

$$G_1 = \{z \in \mathbb{C} : | z | > \max\{k\,\pi, 4\}\}, \tag{4.12}$$

where $g(z)$ is analytic in view of Lemma 4.2. We deduce that, for all $z \in G_1$,

$$\begin{aligned} | g(z) | &= \left| \frac{2(z^2 - k^2\pi^2)}{z^2 - k^2\pi^2 + z(z^2 + \sqrt{z^2 - k^2\pi^2})} \right| \leq \frac{2}{| z | \left| \frac{z^2 + \sqrt{z^2 - k^2\pi^2}}{z^2 - k^2\pi^2} \right| - 1} \\ &\leq \frac{2}{| z | \frac{|z|^2 - |\sqrt{z^2 - k^2\pi^2}|}{|z|^2 + k^2\pi^2} - 1} \leq \frac{2}{| z | \frac{|z|^2 - \sqrt{2}|z|}{|z|^2 + k^2\pi^2} - 1} \leq \frac{2}{\frac{|z|}{4} - 1} \leq \frac{8}{| z | - 4}. \end{aligned} \tag{4.13}$$

We are now in condition to determine the constants ε_m, δ_m, and M_k such that (4.10) is satisfied.

If $z \in \gamma_m^1 \cap G_1$, we obtain that $| f(z) | > \varepsilon_m > | g(z) |$ if $\frac{16}{2m\pi - 9} < \varepsilon_m < \frac{1}{2}$.

Applying Rouché's theorem, it turns out that there exists a unique root of the equation (4.8) in each square γ_m^1 if $m \geq k + 1$. We denote those roots by $\mu_{k,m}$.

If $z \in \gamma_m^2 \cap G_1$, we obtain that $| f(z) | > \frac{1}{2} > | g(z) |$ if $\delta_m > \frac{1}{2}$.

Since we did not impose any upper bound for δ_m we can apply again Rouché's theorem and we obtain that, for each $m \geq k+1$ in the regions $|\mathcal{I}m\, z - m\pi| \leq \frac{3\pi}{4}$, the equation (4.8) has the same number of roots as $f(z) = 0$ does. This implies that the only roots of (4.8) in G_1 are $\mu_{k,m}$ found above.

Finally, if we choose $M_k = k\,\pi + \frac{3\pi}{4}$ we obtain, like above, that if $z \in \gamma^0 \cap G_1$, then $| f(z) | > 1/2 > | g(z) |$.

Applying Rouché's theorem, we deduce that the number of roots of (4.8) in γ^0 is equal to $2k + 2$.

In order to obtain the roots of (2.2) we return to the variable λ.

First of all we remark that if λ solves (2.2) then $\bar{\lambda}$ is a solution too. Hence, it is sufficient to look for those λ with $\mathcal{I}m\, \lambda > 0$, the other eigenvalues being conjugates of these. On the other hand, when we pass from μ to λ we are interested in those values which have the property that $\mathcal{R}e\, \lambda < 0$, since the energy of the system decreases as t increases (see (1.2) and (1.3) above and [11] for a detailed discussion on this). Those

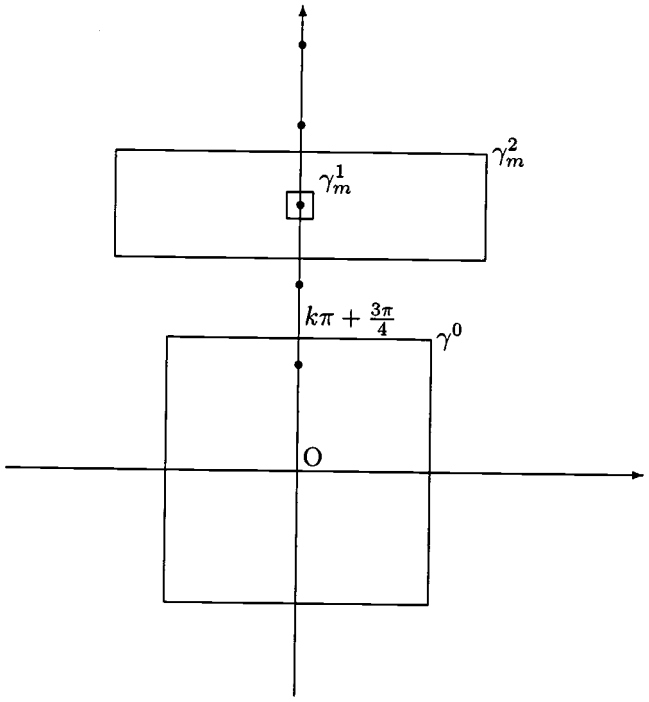

FIG. 4.1.

remarks indicate that we can establish a bijective correspondence between the zeros of the equation (4.8) and those of the equation (2.2).

Since the previous analysis gives us the roots μ of (4.8) with the property that $\mid \mu \mid > \max\{k\pi, 4\}$, we obtain all the roots $\lambda = \sqrt{\mu^2 - k^2\pi^2}$ of (2.2) with the property that $\mid \lambda \mid > \sqrt{2}k\pi$. For those eigenvalues λ with $\mathcal{I}m\,\lambda > 0$ we have

$$\mid \lambda - \sqrt{m^2+k^2}\,\pi\, i \mid = \mid \sqrt{\mu^2 - k^2\pi^2} - \sqrt{m^2+k^2}\,\pi\, i \mid$$

$$= \frac{\mid \mu - m\pi\, i \mid \mid \mu + m\pi\, i \mid}{\sqrt{\mid \mathcal{I}m\sqrt{\mu^2 - k^2\pi^2} + \sqrt{m^2+k^2}\pi \mid^2 + \mid \mathcal{R}e\sqrt{\mu^2 - k^2\pi^2} \mid^2}}$$

$$\leq \frac{\varepsilon_m \mid \mu + m\pi\, i \mid}{\mathcal{I}m\sqrt{\mu^2 - k^2\pi^2} + \sqrt{m^2+k^2}\pi} \leq \frac{\varepsilon_m(\varepsilon_m + 2m\pi)}{\sqrt{k^2+m^2}\pi} \leq \frac{3m}{\sqrt{k^2+m^2}}\varepsilon_m.$$

It turns out that, for $m > k+1$, the eigenvalues $\lambda_{k,m}$ with $\mathcal{I}m\,\lambda_{k,m} > 0$ satisfy (2.4). The corresponding result for the case $\lambda_{k,m}$ with $\mathcal{I}m\lambda_{k,m} < 0$ can be obtained in the same way. □

Remark 13. Theorem 2.2 tells us that, for each $k \in \mathbb{N}$ and for each eigenvalue $(\lambda_{k,m})_{m\in\mathbb{Z}^*}$ with $|m| \geq k+1$, the index m is given by the nearest value $\pm\sqrt{k^2+m^2}\,\pi\, i$. The other eigenvalues, which belong to the circle centered in 0 and of radius $\sqrt{2}\,k\,\pi$, are ordered in the increasing way with respect to the modulus: λ_k^*, λ_k^{**}, $\lambda_{k,\pm1}$, $\lambda_{k,\pm2}$,...,

$\lambda_{k,\pm k}$ (see Theorems 2.4, 2.6, and 2.8). Hence, for k fixed, λ_k^* and λ_k^{**} are the eigenvalues with the smallest modulus, while the modulus of $\lambda_{k,\pm k}$ approaches $\sqrt{2}k\pi$ when m increases.

We prove now Theorem 2.3.

Proof of Theorem 2.3. From (2.4) we deduce that $\sqrt{\lambda^2+k^2\pi^2}=\mu=m\pi\, i+\alpha(m)$ with $|\alpha(m)|\le\frac{1}{m}$.

The eigenfunction φ_λ can be decomposed as follows:

$$\varphi_\lambda=\begin{pmatrix} \frac{1}{\lambda}\cosh\sqrt{\lambda^2+k^2\pi^2}(y-1)\cos k\pi x \\ -\cosh\sqrt{\lambda^2+k^2\pi^2}(y-1)\cos k\pi x \\ -\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\sinh(\sqrt{\lambda^2+k^2\pi^2})\cos k\pi x \\ \frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\sinh(\sqrt{\lambda^2+k^2\pi^2})\cos k\pi x \end{pmatrix}$$

$$=\begin{pmatrix} (-1)^{m+1}\frac{i}{\lambda}\cosh\alpha(m)(y-1)\cos m\pi y\cos k\pi x \\ (-1)^m\, i\cosh\alpha(m)(y-1)\cos m\pi y\cos k\pi x \\ 0 \\ 0 \end{pmatrix}$$

$$+\begin{pmatrix} (-1)^m\frac{1}{\lambda}\sinh\alpha(m)(y-1)\sin m\pi y\cos k\pi x \\ (-1)^{m+1}\sinh\alpha(m)(y-1)\sin m\pi y\cos k\pi x \\ (-1)^{m+1}\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\sinh\alpha(m)\cos k\pi x \\ (-1)^m\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\sinh\alpha(m)\cos k\pi x \end{pmatrix}.$$

We denote by φ_1 and φ_2 the two vector-valued functions above.

We estimate first the norm of φ_2 in $\mathcal{X}$:

$$\|\varphi_2\|_{\mathcal{X}}^2=\int_0^1\int_0^1\left\{\left(\left|\frac{1}{\lambda}\cos k\pi x\right|^2+\left|\frac{k\pi}{\lambda}\sin k\pi x\right|^2\right)|\sinh\alpha(m)(y-1)\sin m\pi y|^2\right.$$

$$\left.+\left|\left(\frac{\alpha(m)}{\lambda}\cosh\alpha(m)(y-1)\sin m\pi y+\frac{m\pi}{\lambda}\sinh\alpha(m)(y-1)\cos m\pi y\right)\cos k\pi x\right|^2\right\}$$

$$+\int_0^1\int_0^1|\sinh\alpha(m)(y-1)\sin m\pi y\cos k\pi x|^2\,dx\,dy$$

$$+\int_0^1\left\{\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\sinh\alpha(m)\cos k\pi x\right|^2+\left|\frac{k\pi\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\sinh\alpha(m)\sin k\pi x\right|^2\right\}dx$$

$$+\int_0^1 \left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\ \sinh\ \alpha(m)\ \cos\ k\pi x\right|^2\ dx\ \le \int_0^1 \left\{\left|\frac{1}{\lambda}\ \sinh\ \alpha(m)(y-1)\right|^2\right.$$

$$\left.+\left|\frac{\alpha(m)}{\lambda}\ \cosh\ \alpha(m)(y-1)\right|^2+\left(\frac{(k^2+m^2)\pi^2}{|\lambda|^2}+1\right)|\ \sinh\ \alpha(m)(y-1)|^2\right\}dy$$

$$+\left(\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\right|^2+\left|\frac{k\pi\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\right|^2+\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\right|^2\right)|\ \sinh\ \alpha(m)|^2$$

$$\le 4\frac{|\ \alpha(m)\ |^2}{|\lambda|^2}+4|\alpha(m)|^2+5\frac{|\ \alpha(m)\ |^2}{|\lambda|^2}+4\frac{(k^2+m^2)\pi^2\ |\ \alpha(m)\ |^2}{|\lambda|^2}$$

$$+4|\alpha(m)|^2\left|\frac{\lambda^2+k^2\pi^2}{\lambda^4}\right|^2\left|1+k^2\pi^2+\lambda^2\right|^2\le 33|\alpha(m)|^2\le\frac{c'}{m^2},$$

where we take into account that $|\sinh\ \alpha(m)|\le 2|\alpha(m)|$ and $|\cosh\ \alpha(m)|\le 2$.

In this way we obtain that

$$\|\varphi_2\|_{\mathcal{X}}\le\frac{c'}{m}. \tag{4.14}$$

We estimate now

$$\|\ \varphi_1-\psi_{k,m}\ \|_{\mathcal{X}}=\int_0^1\int_0^1\left\{\left|\frac{i}{\sqrt{m^2+k^2\pi}}+\frac{1}{\lambda}\cosh\alpha(m)(y-1)\right|^2|\cos m\pi y\cos k\pi x|^2\right.$$

$$+\left|\frac{i}{\sqrt{m^2+k^2\pi}}+\frac{1}{\lambda}\ \cosh\ \alpha(m)(y-1)\right|^2|k\pi\ \cos\ m\pi y\ \sin\ k\pi x|^2$$

$$+\left|\frac{i}{\sqrt{m^2+k^2\pi}}+\frac{1}{\lambda}\cosh\alpha(m)(y-1)\right|^2|m\pi\ \sin m\pi y\cos k\pi x|^2$$

$$\left.+\left(\left|\frac{\alpha(m)}{\lambda}\sinh\alpha(m)(y-1)\right|^2+|1-\cosh\alpha(m)(y-1)|^2\right)|\cos m\pi y\cos k\pi x|^2\right\}dxdy$$

$$\le\int_0^1\left\{\left|\frac{i}{\sqrt{m^2+k^2\pi}}+\frac{1}{\lambda}\ \cosh\ \alpha(m)(y-1)\right|^2\right.$$

$$+\left|\frac{i(k^2+m^2)\pi^2}{\sqrt{m^2+k^2}\pi}+\frac{(k^2+m^2)\pi^2}{\lambda}\cosh\,\alpha(m)(y-1)\right|^2$$

$$+\left|\frac{\alpha(m)}{\lambda}\sinh\,\alpha(m)(y-1)\right|^2+|1-\cosh\,\alpha(m)(y-1)|^2\Bigg\}\,dy$$

$$\leq\left|\frac{i}{\sqrt{m^2+k^2}\pi}+\frac{1}{\lambda}\right|^2+\int_0^1\left|\frac{1}{\lambda}(1-\cosh\,\alpha(m)(y-1))\right|^2\,dy$$

$$+(k^2+m^2)\pi^2\left|\frac{i}{\sqrt{m^2+k^2}\pi}+\frac{1}{\lambda}\right|^2+(m^2+k^2)\pi^2\int_0^1\left|\frac{1}{\lambda}(1-\cosh\,\alpha(m)(y-1))\right|^2\,dy$$

$$+\int_0^1\left|\frac{\alpha(m)}{\lambda}\sinh\,\alpha(m)(y-1)\right|^2\,dy+\int_0^1|1-\cosh\,\alpha(m)(y-1)|^2\,dy$$

$$\leq\frac{c''}{m^2}+4|\alpha(m)|^2+2\frac{c''}{m^2}+8|\alpha(m)|^2+4|\alpha(m)|^2+4|\alpha(m)|^2\leq\frac{c'''}{m^2},$$

where we take into account that $|1-\cosh\,\alpha(m)|\leq 2\mid\alpha(m)\mid$.

We obtain that

$$\|\varphi_1-\psi_{k,m}\|_{\mathcal{X}}\leq\frac{c'''}{m}. \tag{4.15}$$

From estimates (4.14) and (4.15) we deduce that (2.8) holds. □

Next we prove Theorem 2.4 which gives estimations for the eigenvalues $\lambda_{k,\pm 1}$, $\lambda_{k,\pm 2},\ldots,\lambda_{k,\pm q}$ for $q=q(k)\leq[\sqrt[3]{k}]$. By $[\,\cdot\,]$ we denote the integer part function.

Proof of Theorem 2.4. If we consider the change of variable $\lambda=\sqrt{k^2z^2-k^2\pi^2}$ the equation (2.2) is transformed in

$$e^{2k\,z}=-\frac{z^2-\pi^2-kz^3-z\sqrt{z^2-\pi^2}}{z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2}}. \tag{4.16}$$

Let $k\in\mathbb{N}$ be sufficiently large so that Lemma 4.1 holds. We define the functions

$$f(z)=e^{2kz}+1,\quad g(z)=\frac{2(kz^3+z\sqrt{z^2-\pi^2})}{z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2}}.$$

For each integer m with $0\leq|m|\leq[\sqrt[3]{k}]$ let $\gamma^1_{k,m}$ be the square of center $\frac{2m-1}{2k}\pi\,i$ and sides $\frac{3\pi}{2k\sqrt[3]{k}}$. For all $z\in\gamma^1_{k,m}$, we have

$$\mid f(z)\mid=\mid e^{2kz}+1\mid\geq\max\{\mid e^{2k\,\mathcal{R}e\,z}-1\mid,\mid\sin\,2k\,\mathcal{I}m\,z\mid\},$$

and since $|e^x-1|>\frac{|x|}{2}$ and $|\sin x|>\frac{|x|}{2}$ for small x, we deduce that

$$|f(z)|\geq \frac{3\pi}{4\sqrt[3]{k}} \quad \forall z\in\gamma_{k,m}. \tag{4.17}$$

We now estimate g in the region $G^2=\left\{z\in\mathbb{C}:|z|\leq \frac{\pi}{\sqrt[3]{k^2}}\right\}$.

Lemma 4.1 implies that g is analytic in G^2.

For all $z\in G^2$ we have $|z|\sqrt[3]{k}\leq\pi$. Therefore we obtain that

$$\lim_{k\to\infty} kz^2=\lim_{k\to\infty} z^2=\lim_{k\to\infty} kz^3=0.$$

Hence, for all $z\in G^2$,

$$\lim_{k\to\infty}\left|\frac{kz^2+\sqrt{z^2-\pi^2}}{z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2}}\right|=\frac{1}{\pi},$$

which implies that, for k sufficiently large,

$$\left|\frac{kz^2-\sqrt{z^2-\pi^2}}{z^2-\pi^2+kz^3-z\sqrt{z^2-\pi^2}}\right|\leq 1 \quad \forall z\in G^2.$$

This result allows us to estimate the function g in G^2:

$$|g(z)|=2\,|z|\left|\frac{kz^2+\sqrt{z^2-\pi^2}}{z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2}}\right|\leq 2|z|\leq\frac{2\pi}{\sqrt[3]{k^2}}.$$

Finally, we obtain that $|f(z)|>|g(z)|$ for all $z\in\gamma^1_{k,m}$ if k is sufficiently large and $\gamma^1_{k,m}\subset G^2$. Remark that $\gamma^1_{k,m}\subset G^2$ if $|m|\leq[\sqrt[3]{k}]$.

Applying Rouché's theorem we deduce that the equation (4.16) has a root $z_{k,m}$ in each square $\gamma^1_{k,m}$ if $|m|\leq[\sqrt[3]{k}]$. This root satisfies

$$\left|z_{k,m+1}-\frac{1}{2k}(2m+1)\pi\,i\right|\leq\frac{3\sqrt{2}\pi}{4k\sqrt[3]{k}}\leq\frac{2\pi}{k\sqrt[3]{k}} \quad \text{if } m\geq 0,$$

$$\left|z_{k,m}+\frac{1}{2k}(2m+1)\pi\,i\right|\leq\frac{3\sqrt{2}\pi}{4k\sqrt[3]{k}}\leq\frac{2\pi}{k\sqrt[3]{k}} \quad \text{if } m<0.$$

We deduce that the eigenvalues $\lambda_{k,m}=\sqrt{k^2z^2_{k,m}-k^2\pi^2}$ with $0<|m|\leq[\sqrt[3]{k}]$ satisfy (2.9). □

Proof of Theorem 2.5. Estimates (2.9) imply that

$$\sqrt{\lambda^2+k^2\pi^2}=\mu=\frac{2m+1}{2}\pi\,i+\alpha(k) \text{ with } |\alpha(k)|\leq\frac{2\pi}{\sqrt[3]{k}}.$$

We write the eigenfunction φ_λ in the following form:

$$\varphi_\lambda = \begin{pmatrix} \frac{1}{\lambda}\cosh\,(\sqrt{\lambda^2+k^2\pi^2}(y-1))\,\cos\,k\pi x \\ -\cosh\,(\sqrt{\lambda^2+k^2\pi^2}(y-1))\,\cos\,k\pi x \\ -\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\,\sinh\,(\sqrt{\lambda^2+k^2\pi^2})\,\cos\,k\pi x \\ \frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\,\sinh\,(\sqrt{\lambda^2+k^2\pi^2})\,\cos\,k\pi x \end{pmatrix}$$

$$= \begin{pmatrix} (-1)^m\frac{1}{\lambda}\cosh\,\alpha(k)(y-1)\,\sin\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x \\ (-1)^{m+1}\cosh\,\alpha(k)(y-1)\,\sin\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x \\ 0 \\ 0 \end{pmatrix}$$

$$+ \begin{pmatrix} (-1)^{m+1}\frac{i}{\lambda}\sinh\,\alpha(k)(y-1)\,\cos\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x \\ (-1)^m i\,\sinh\,\alpha(k)(y-1)\,\cos\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x \\ (-1)^{m+1}i\,\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\,\cosh\,\alpha(k)\,\cos\,k\pi x \\ (-1)^m i\,\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\,\cosh\,\alpha(k)\,\cos\,k\pi x \end{pmatrix}.$$

Let φ_1 and φ_2 be the vector-valued functions appearing in the decomposition of φ_λ above.

We evaluate first the norm of φ_2 in $\mathcal{X}$:

$$\|\varphi_2\|^2_{\mathcal{X}} = \int_0^1\int_0^1 \left|\frac{1}{\lambda}\,\cos\,k\pi x\right|^2\,\sinh\,\alpha(k)(y-1)\,\cos\,\frac{2m+1}{2}\pi y$$

$$+\int_0^1\int_0^1\left\{\left|\frac{k\pi}{\lambda}\,\sinh\,\alpha(k)(y-1)\,\cos\,\frac{2m+1}{2}\pi y\,\sin\,k\pi x\right|^2\right.$$

$$+\left|\frac{\alpha(k)}{\lambda}\,\cosh\,\alpha(k)(y-1)\,\cos\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x + \frac{(2m+1)\pi}{2\lambda}\,\sinh\,\alpha(k)(y-1)\right.$$

$$\left.\left.\times\,\sin\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x\right|^2\right\} + \int_0^1\int_0^1\left|\,\sinh\,\alpha(k)(y-1)\,\cos\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x\right|^2$$

$$+\int_0^1\left\{\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\cosh\alpha(k)\cos k\pi x\right|^2 + \left|\frac{k\pi\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\cosh\alpha(k)\sin k\pi x\right|^2\right\}$$

$$+\int_0^1 \left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\cosh\,\alpha(k)\,\cos\,k\pi x\right|^2 \leq \int_0^1 \left\{\left|\frac{1}{\lambda}\,\sinh\,\alpha(k)(y-1)\right|^2\right.$$

$$\left.+\left|\frac{\alpha(k)}{\lambda}\cosh\alpha(k)(y-1)\right|^2+\left(\left(k^2+\left(\frac{2m+1}{2}\right)^2\right)\frac{\pi^2}{|\lambda|^2}+1\right)|\sinh\alpha(k)(y-1)|^2\right\}$$

$$+\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\cosh\alpha(k)\right|^2+\left|\frac{k\pi\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\cosh\alpha(k)\right|^2+\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\cosh\alpha(k)\right|^2$$

$$\leq 4\frac{|\alpha(k)|^2}{|\lambda|^2}+5\frac{|\alpha(k)|^2}{|\lambda|^2}+\left(k^2+\left(\frac{2m+1}{2}\right)^2\right)\frac{|\alpha(k)|^2\pi^2}{|\lambda|^2}+4|\alpha(k)|^2$$

$$5(k^2\pi^2+1)\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda^2}\right|^2+5\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\right|^2\leq 14|\alpha(k)|^2+60|\alpha(k)|^2\leq\frac{c'}{\sqrt[3]{k}},$$

where we take into account that, for k large enough,

$$\left|\frac{\sqrt{\lambda^2+k^2\pi^2}}{\lambda}\right|\leq 2|\alpha(k)|,\ |\sinh\,\alpha(k)|\leq 2|\alpha(k)|\text{ and }|\cosh\,\alpha(k)|\leq 5.$$

We obtain that

$$\|\varphi_2\|_{\mathcal{X}}\leq\frac{c'}{\sqrt[3]{k}}. \tag{4.18}$$

We compute now

$$\|\,\varphi_1-\tilde{\psi}_{k,m}\,\|^2_{\mathcal{X}}$$

$$=\int_0^1\int_0^1\left|\frac{i}{\sqrt{\left(\frac{2m+1}{2}\right)^2+k^2\pi}}+\frac{1}{\lambda}\,\cosh\,\alpha(k)(y-1)\right|^2\left|\,\sin\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x\right|^2$$

$$+\int_0^1\int_0^1\left\{\left|\frac{i}{\sqrt{\left(\frac{2m+1}{2}\right)^2+k^2\pi}}+\frac{1}{\lambda}\,\cosh\,\alpha(k)(y-1)\right|^2\left|k\pi\,\sin\,\frac{2m+1}{2}\pi y\,\sin\,k\pi x\right|^2\right.$$

$$\left.+\left|\frac{\alpha(k)}{\lambda}\,\sinh\,\alpha(k)(y-1)\,\sin\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x\right|^2\right.$$

$$+\left|\left(\frac{i}{\sqrt{\left(\frac{2m+1}{2}\right)^2+k^2\pi}}+\frac{1}{\lambda}\cosh\alpha(k)(y-1)\right)\frac{2m+1}{2}\pi\,\cos\frac{2m+1}{2}\pi y\cos k\pi x\right|^2$$

$$+\left|(1-\ \cosh\ \alpha(k)(y-1))\,\sin\,\frac{2m+1}{2}\pi y\,\cos\,k\pi x\right|^2\Bigg\}\,dxdy$$

$$\leq\int_0^1\left(\left(\frac{(2m+1)\pi}{2}\right)^2+k^2\pi^2+1\right)\left|\frac{i}{\sqrt{\left(\frac{2m+1}{2}\right)^2+k^2\pi}}+\frac{1}{\lambda}\,\cosh\,\alpha(k)(y-1)\right|^2$$

$$+\int_0^1\left|\frac{\alpha(m)}{\lambda}\,\sinh\,\alpha(m)(y-1)\right|^2+|1-\ \cosh\,\alpha(k)(y-1)|^2$$

$$\leq\left(\left(\frac{(2m+1)\pi}{2}\right)^2+k^2\pi^2+1\right)\left|\frac{i}{\sqrt{\left(\frac{(2m+1)}{2}\right)^2+k^2\pi}}+\frac{1}{\lambda}\right|^2$$

$$+\left(\left(\frac{2m+1}{2}\right)^2\pi^2+k^2\pi^2+1\right)\int_0^1\left|\frac{1}{\lambda}(1-\ \cosh\,\alpha(k)(y-1))\right|^2$$

$$+\int_0^1\left|\frac{\alpha(k)}{\lambda}\,\sinh\,\alpha(k)(y-1)\right|^2+\int_0^1|1-\ \cosh\,\alpha(m)(y-1)|^2$$

$$\leq\frac{2\pi}{\sqrt[3]{k}}+4|\alpha(k)|^2+2\frac{2\pi}{\sqrt[3]{k}}+8|\alpha(m)|^2+4|\alpha(k)|^2+4|\alpha(k)|^2\leq\frac{c''}{\sqrt[3]{k}}$$

since, for k large enough ($k>(2\pi)^3$), $|1-\ \cosh\,\alpha(k)|\leq 2\mid\alpha(k)\mid$.

We obtain that

$$\|\,\varphi_1-\tilde{\psi}_{k,m}\,\|_{\mathcal{X}}\leq\frac{c'''}{\sqrt[3]{k}}. \tag{4.19}$$

The estimates (4.18) and (4.19) imply that (2.12) holds. □

We pass now to the analysis of the roots of (2.2) $\lambda_{k,\pm(q+1)}$, $\lambda_{k,\pm(q+2)},\ldots,\lambda_{k,\pm k}$, with $q=[\sqrt[3]{k}]$, which make the transition from the eigenvalues studied in Theorem 2.2 to those studied in Theorem 2.4. First we prove the following lemma.

LEMMA 4.5. *For each $k\in\mathbb{N}^*$, the equation*

$$e^{2k\,z}=\frac{\pi^2+kz^3}{-\pi^2+kz^3} \tag{4.20}$$

has a sequence of roots $\pm\zeta_{k,m}\, i$, $m \in \mathbb{N}^*$, *where* $\zeta_{k,m} \in \mathbb{R}_+$ *is the positive root of the equation* (2.14) *which belongs to* $\left(\frac{m}{k}\pi, \frac{2m+1}{2k}\pi\right)$.

Proof. We look for roots of (4.20) of the form $z = \zeta\, i$. Hence, ζ is a root of the equation

$$e^{2k\,\zeta\, i} = \frac{\pi^2 - k\,\zeta^3\, i}{-\pi^2 - k\,\zeta^3\, i}. \tag{4.21}$$

Consequently, z is a root of (4.20) if ζ satisfies

$$-\pi^2 \cos k\zeta + k\zeta^3 \sin k\zeta = 0, \tag{4.22}$$

which is equivalent to (2.14).

It is easy to see that (4.22) has a zero in each interval $\left(\frac{m}{k}\pi, \frac{2m+1}{2k}\pi\right)$ that we denote by $\zeta_{k,m}$ (see [15]). □

We pass now to study the eigenvalues $\lambda_{k,\pm([\sqrt[3]{k}]+1)}$, $\lambda_{k,\pm([\sqrt[3]{k}]+2)}, \dots, \lambda_{k,\pm k}$.

Proof of Theorem 2.6. We saw that the change of variables $\lambda = \sqrt{k^2z^2 - k^2\pi^2}$ transforms (2.2) in (4.16).

We define the region of the complex plane

$$G^3 = \left\{ z \in \mathbb{C}\colon \frac{\pi}{2\sqrt[3]{k^2}} \le |z| \le 2\pi, \quad |\mathcal{R}e\, z| \le \frac{1}{k}, \quad \mathcal{I}m\, z > 0 \right\}$$

and we prove that (4.16) has a set of zeros $z_{k,m}$ in G^3 satisfying the estimate

$$|z_{k,m} - \zeta_{k,m}\, i| \le \frac{1}{k\sqrt[5]{k}}, \quad m \in \{[\sqrt[3]{k}] + 1, \dots, k\}, \tag{4.23}$$

where $\zeta_{k,m}$ are the zeros of (2.14).

We remark that if $m \in \{[\sqrt[3]{k}] + 1, \dots, k\}$, then $\zeta_{k,m}$ belongs to G^3.

We write (4.16) in the following form:

$$e^{2k\,z} - \frac{\pi^2 + kz^3}{-\pi^2 + kz^3} = -\frac{2z(kz^4 + \pi^2\sqrt{z^2 - \pi^2})}{(-\pi^2 + kz^3)(z^2 - \pi^2 + kz^3 + z\sqrt{z^2 - \pi^2})},$$

and applying Rouché's theorem we prove that the zeros of (4.16) approach those of (4.20).

We consider first the function

$$g(z) = -\frac{2z(kz^4 + \pi^2\sqrt{z^2 - \pi^2})}{(-\pi^2 + kz^3)(z^2 - \pi^2 + kz^3 + z\sqrt{z^2 - \pi^2})},$$

and we obtain an upper bound for g in G^3.

To do this we evaluate first the denominator of g:

$$|2z(kz^4 + \pi^2\sqrt{z^2 - \pi^2})| \le 2k|z|^5 + 2\pi^2|z|^2 + 4\pi^3|z|.$$

We obtain that

$$|2z(kz^4 + \pi^2\sqrt{z^2 - \pi^2})| \le \begin{cases} 6k|z|^5 & \text{if } \dfrac{\pi}{2\sqrt[3]{k^2}} \le |z| \le \dfrac{\pi}{\sqrt[4]{k}}, \\ 6\pi^3|z| & \text{if } \dfrac{\pi}{\sqrt[4]{k}} \le |z| \le 2\pi. \end{cases}$$

We estimate now the numerator of g:

$$|(-\pi^2+kz^3)(z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2})|$$

$$\geq |-\pi^2+kz^3|\,(|-\pi^2+kz^3|-|z|(|z|+\sqrt{|z|^2+\pi^2}))$$

$$\geq |-\pi^2+kz^3|^2\ -\ 5\pi|z|\,|-\pi^2+kz^3|.$$

If $\frac{\pi}{2\sqrt[3]{k^2}}\leq|z|\leq\frac{\pi}{\sqrt[4]{k}}$ and $|\mathcal{R}e\,z|\leq\frac{1}{k}$ we have that, for k sufficiently large,

$$|-\pi^2+kz^3|\geq\mathcal{R}e\,(-\pi^2+kz^3)\geq\frac{\pi}{2}.$$

If $\frac{\pi}{\sqrt[4]{k}}\leq|z|\leq 2\pi$ we have that

$$|-\pi^2+kz^3|\,(|-\pi^2+kz^3|-|z|(|z|+\sqrt{|z|^2+\pi^2}))\geq\sqrt{k}k|z|^4\left(\sqrt{k}|z|^2-\frac{2\pi}{\sqrt{k}}\right).$$

From the last two inequalities, we deduce that, for k sufficiently large, the following estimate holds:

$$|(-\pi^2+kz^3)(z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2})|\geq\begin{cases}c_1 & \text{if } \frac{\pi}{2\sqrt[3]{k^2}}\leq|z|\leq\frac{\pi}{\sqrt[4]{k}},\\ c_2\sqrt{k}\,k\,|z|^4 & \text{if } \frac{\pi}{\sqrt[4]{k}}\leq|z|\leq 2\pi,\end{cases}$$

where c_1 and c_2 are two positive constants which do not depend on k.

Going back to the function g we obtain that, for k sufficiently large,

$$|g(z)|\leq\frac{c}{\sqrt[4]{k}}\quad\text{for all } z \text{ in } G^3,$$

where c is a positive constant which does not depend on k.

We study now the function

$$f(z)=e^{2k\,z}-\frac{\pi^2+kz^3}{-\pi^2+kz^3}.$$

For each $m\in\mathbb{N}^*$ we consider the circle $\gamma^2_{k,m}$ of center $\zeta_{k,m}\,i$ and radius $r_{k,m}=\frac{1}{k\sqrt[5]{k}}$ and the circle $\widehat{\gamma}^2_{k,m}$ with the same center but with radius $R_{k,m}=\frac{1}{k}$.

In G^3 the function f is analytic. Applying Taylor's formula at $\zeta_{k,m}\,i$ we obtain that

$$\begin{aligned}f(z)&=f(\zeta_{k,m}\,i)+(z-\zeta_{k,m}\,i)f'(\zeta_{k,m}\,i)\\&+\frac{(z-\zeta_{k,m}\,i)^2}{2\pi\,i}\int_{\widehat{\gamma}^2_{k,m}}\frac{f(\zeta)\,d\zeta}{(\zeta-\zeta_{k,m}\,i)^2(\zeta-z)}.\end{aligned}\tag{4.24}$$

We look for an upper bound for the error term on the circumference $\gamma^2_{k,m}$. We have

$$\left|\frac{(z-\zeta_{k,m}\,i)^2}{2\pi\,i}\int_{\widehat{\gamma}^2_{k,m}}\frac{f(\zeta)\,d\zeta}{(\zeta-\zeta_{k,m}\,i)^2(\zeta-z)}\right|\leq\frac{r^2_{k,m}}{2\pi}\frac{2\pi R_{k,m}M}{R^2_{k,m}(R_{k,m}-r_{k,m})}=\frac{M}{\sqrt[5]{k}(\sqrt[5]{k}-1)},$$

where M is an upper bound for f on the circumference $\widehat{\gamma}^2_{k,m}$.

On the other hand

$$|f(z)| = \left| e^{2k\,z} - \frac{\pi^2 + kz^3}{-\pi^2 + kz^3} \right| \leq |e^{2k\,z}| + \left| \frac{\pi^2 + kz^3}{-\pi^2 + kz^3} \right| \leq e^{2k\,|\mathcal{R}e\,z|} + 1 + \frac{2\pi^2}{|\pi^2 - k\,z^3|}.$$

Since $|\mathcal{R}e\,z| < \frac{1}{k}$ in G^3, we obtain that $|\pi^2 - k\,z^3| > 1$ and $|f(z)| < M = e^2 + 1 + 2\pi^2$. Therefore the error term in Taylor's formula on $\gamma^2_{k,m}$ satisfies

$$\left| \frac{(z - \zeta_{k,m}\,i)^2}{2\pi\,i} \int_{\widehat{\gamma}^2_{k,m}} \frac{f(\zeta)\,d\zeta}{(\zeta - \zeta_{k,m}\,i)^2(\zeta - z)} \right| \leq \frac{M}{\sqrt[5]{k}(\sqrt[5]{k} - 1)}.$$

On the other hand,

$$|(z - \zeta_{k,m}\,i)f'(\zeta_{k,m}\,i)| = r_{k,m} \left| 2k \frac{\pi^2 - k\zeta^3_{k,m}\,i}{-\pi^2 - k\,\zeta^3_{k,m}\,i} - \frac{6\pi^2 k\zeta^2_{k,m}}{(-\pi^2 - k\,\zeta^3_{k,m}\,i)^2} \right| \geq 2kr_{k,m}.$$

Going back to Taylor's formula (4.24), we deduce that if z belongs to the circumference $\gamma^2_{k,m}$, then

$$|f(z)| \geq |(z - \zeta_{k,m}\,i)f'(\zeta_{k,m}\,i)| - \left| \frac{(z - \zeta_{k,m}\,i)^2}{2\pi\,i} \int_{\widehat{\gamma}^2_{k,m}} \frac{f(\zeta)\,d\zeta}{(\zeta - \zeta_{k,m}\,i)^2(\zeta - z)} \right|$$

$$\geq 2kr_{k,m} - \frac{20}{\sqrt[5]{k}(\sqrt[5]{k} - 1)} \geq \frac{C}{\sqrt[5]{k}}.$$

Finally, we obtain that $|f(z)| > |g(z)|$ for all z in $\gamma^2_{k,m}$.

Applying Rouché's theorem we deduce that the equation (4.16) has a unique zero $z_{k,m}$ which satisfies (4.23) in each circle $\gamma^2_{k,m}$.

Taking into account that $\lambda = \sqrt{k^2\pi^2 + k^2 z^2}$ we deduce immediately the desired result. □

Proof of Theorem 2.7. The eigenvalues $\lambda_{k,m}$ studied in Theorem 2.6 approach $\sqrt{\pi^2\,k^2 + k^2\,\zeta^2_{k,m}}\,i$, where $\zeta_{k,m}$ are the roots of the equation

$$\tan k\,\zeta = \frac{\pi^2}{k\,\zeta^3}. \tag{4.25}$$

By a similar method one can prove that $\lambda_{k,m}$ satisfy the estimates

$$\begin{aligned} &\left| \lambda_{k,m} - \sqrt{\pi^2\,k^2 + k^2\,\varrho^2_{k,m}}\,i \right| \leq \frac{1}{\sqrt[5]{k}} \text{ if } \mathcal{I}m\,\lambda_{k,m} > 0 \quad (k \geq m > [\sqrt[3]{k}]), \\ &\left| \lambda_{k,m} + \sqrt{\pi^2\,k^2 + k^2\,\varrho^2_{k,m}}\,i \right| \leq \frac{1}{\sqrt[5]{k}} \text{ if } \mathcal{I}m\,\lambda_{k,m} < 0 \quad (-k \leq m < -[\sqrt[3]{k}]), \end{aligned} \tag{4.26}$$

where $\varrho_{k,m}$ is the root of the equation

$$\tan k\,\varrho = \frac{\pi^2 + \varrho^2}{k\varrho^3}, \tag{4.27}$$

which belongs to the interval $\left(\frac{m}{k}\pi, \frac{2m+1}{2k}\pi\right)$.

Taking into account the estimates of Theorem 3.1 for the eigenvalues $\nu_{k,m}$ of the conservative problem, we deduce that, for the eigenvalues $\lambda_{k,m}$ studied in Theorem 2.6, we have:

$$|\lambda_{k,m} - \nu_{k,m}| \leq \frac{1}{\sqrt[5]{k}} \text{ for } [\sqrt[3]{k}] < |m| \leq k. \tag{4.28}$$

Since the eigenfunctions $\varphi_{\lambda_{k,m}}$ and $\xi_{\nu_{k,m}}$ have the same form, we deduce that

$$||\varphi_{\lambda_{k,m}} - \xi_{\nu_{k,m}}||_{\mathcal{X}} \leq \frac{1}{\sqrt[5]{k}}.$$

The properties of the eigenfunctions $\varphi_{\lambda_{k,m}}$ are obtained from the corresponding properties of $\xi_{\nu_{k,m}}$ (see Theorem 3.2). □

4.2. Eigenvalues with uniform negative real parts. The eigenvalues obtained in Theorems 2.2, 2.4, and 2.6 have in common the fact that their real parts tend to zero when the modulus increases. On the other hand, the last two components of the corresponding eigenfunctions vanish asymptotically.

Next we prove that there exists a sequence of eigenvalues $(\lambda_k^*)_k$ of modulus less than $k\pi$ with completely different properties.

Proof of Theorem 2.8. We consider again equation (4.16) and we look for the roots with real part going to infinity.

In the circle δ_1 of center $\sqrt[3]{\frac{\pi^2}{k}}$ and radius $\frac{10}{\sqrt[3]{k^2}}$ the function $h(z) = z^2 - \pi^2 - kz^3 - z\sqrt{z^2-\pi^2}$ does not vanish (the three roots of this function are $\sqrt[3]{\frac{\pi^2}{k}}\tilde{\omega}_i$, where $\tilde{\omega}_i$ are the cubic roots of -1 as we saw in Lemma 4.3).

We write the equation (4.16) in the form

$$e^{-2k\,z} = -\frac{z^2 - \pi^2 + kz^3 + z\sqrt{z^2-\pi^2}}{z^2 - \pi^2 - kz^3 - z\sqrt{z^2-\pi^2}}. \tag{4.29}$$

If z belongs to the circle δ_1 we have that $\mathcal{R}e\, z > \frac{\pi}{2\sqrt[3]{k}}$ and hence

$$\left|e^{-2kz}\right| = e^{-2k\mathcal{R}e\, z} \leq e^{-2k\frac{\pi}{2\sqrt[3]{k}}} = e^{-\pi\sqrt[3]{k^2}}.$$

We consider now the circle $\mathcal{C}'$ centered in α_1 and of radius $\frac{1}{k^2}$ (see Fig. 4.2).

Since the circle $\mathcal{C}'$ is contained in δ_1 we have that

$$\left|e^{-2kz}\right| =\leq e^{-2k\frac{\pi}{2\sqrt[3]{k}}} = e^{-\pi\sqrt[3]{k^2}} \quad \forall z \in \mathcal{C}'. \tag{4.30}$$

In $\mathcal{C}'$ the function $u(z) = z^2 - \pi^2 + kz^3 + z\sqrt{z^2-\pi^2}$ is analytic and it has a unique zero α_1.

Since

$$|u'(\alpha_1)| \geq |3k\alpha_1^2| - \left(\left|\sqrt{\alpha_1^2-\pi^2}\right| + |\alpha_1|\left|2 + \frac{\alpha_1}{\sqrt{\alpha_1^2-\pi^2}}\right|\right)$$

$$> 3k\frac{\pi^2}{4\sqrt[3]{k^2}} - \left(|\alpha_1| + \pi + |\alpha_1|\left|2 + \frac{\alpha_1}{\sqrt{\alpha_1^2-\pi^2}}\right|\right) > \sqrt[4]{k}$$

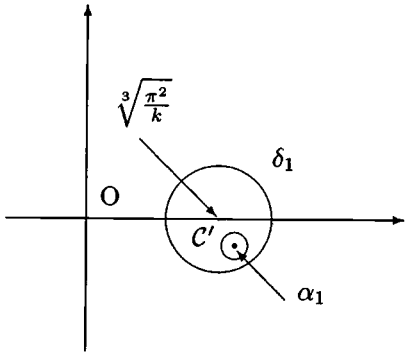

FIG. 4.2.

for k sufficiently large, by applying Taylor's theorem we obtain

$$|u(z) - u'(\alpha_1)(z-\alpha_1)| \leq a|z-\alpha_1|^2,$$

where a is a constant depending on k.

Nevertheless, we have that $|a| \leq \sup\{|u''(z)| : z \in \mathcal{C}'\} < k$.

We obtain that, if z belongs to the circumference of $\mathcal{C}'$,

$$|u(z)| \geq |u'(\alpha_1)||z-\alpha_1| - a|z-\alpha_1|^2 > \frac{1}{k^2}.$$

Hence

$$\left| -\frac{z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2}}{z^2-\pi^2-kz^3-z\sqrt{z^2-\pi^2}} \right| \geq \frac{|u(z)|}{|\,kz^3+\pi^2\,|+|z||z-\sqrt{z^2-\pi^2}|} \geq \frac{1}{k^3}.$$

Thus, for k sufficiently large and z on the circumference of $\mathcal{C}'$, we have

$$\left|e^{-2k\,z}\right| < \left| \frac{z^2-\pi^2+kz^3+z\sqrt{z^2-\pi^2}}{z^2-\pi^2-kz^3-z\sqrt{z^2-\pi^2}} \right|.$$

Applying Rouché's theorem we deduce that the equation (4.29) has a unique root z_k^* in $\mathcal{C}'$. Remark that if z_k^* is a root of (4.29), then $z_k^{**} = \bar{z}_k^*$, $-z_k^*$, and $-z_k^{**}$ are roots of this equation too.

Since $z_k^* \in \mathcal{C}'$ it follows that $z_k^* = \alpha_1 + \mathcal{O}\left(\frac{1}{k^2}\right)$. Hence, Lemma 4.3 ensures that

$$z_k^* = \sqrt[3]{\frac{\pi^2}{k}} - \frac{1}{3}\sqrt[3]{\frac{\pi}{k^2}}\,i + o\left(\frac{1}{\sqrt[3]{k^2}}\right). \tag{4.31}$$

We go back to the equation (2.2) and we obtain two roots λ_k^* and λ_k^{**} setting $\lambda_k^* = \sqrt{k^2(z_k^*)^2 - k^2\pi^2}$ and $\lambda_k^{**} = \sqrt{k^2(z_k^{**})^2 - k^2\pi^2}$.

We have

$$\left|\lambda_k^* - \sqrt{k^2(\alpha_1)^2 - k^2\pi^2}\right| = \frac{\left|(\lambda_k^*)^2 - (k^2(\alpha_1)^2 - k^2\pi^2)\right|}{\left|\lambda_k^* + \sqrt{k^2(\alpha_1)^2 - k^2\pi^2}\right|} = \frac{\left|k^2(z_k^*)^2 - k^2(\alpha_1)^2\right|}{\left|\lambda_k^* + \sqrt{k^2(\alpha_1)^2 - k^2\pi^2}\right|}$$

$$= \frac{k^2\,|(z_k^*-\alpha_1)(z_k^*+\alpha_1)|}{\left|\lambda_k^*+\sqrt{k^2(\alpha_1)^2-k^2\pi^2}\right|} \le \frac{|z_k^*+\alpha_1|}{k\left|\sqrt{(z_k^*)^2-\pi^2}+\sqrt{(\alpha_1)^2-\pi^2}\right|} \le \frac{1}{|k|}.$$

A similar result is obtained for λ_k^{**}.

We now prove (2.19). Remark first that if $\zeta = \sqrt{a+b\,i},\ a,b\in\mathbb{R}$ then $(\mathcal{R}e\,\zeta)^2 = \frac{1}{2}(a+\sqrt{a^2+b^2})$. We deduce that

$$\begin{aligned}(\mathcal{R}e\,\lambda_k^*)^2 = \frac{1}{2}\Big(&-k^2\pi^2 + k^2((\mathcal{R}e\,z_k^*)^2-(\mathcal{I}m\,z_k^*)^2)\\ &+\sqrt{(-k^2\pi^2+k^2((\mathcal{R}e\,z_k^*)^2-(\mathcal{I}m\,z_k^*)^2))^2+(2k^2\mathcal{R}e\ z_k^*\,\mathcal{I}m\,z_k^*)^2}\Big).\end{aligned} \tag{4.32}$$

Since z_k^* satisfies (4.31) we deduce from the relation (4.32) that

$$(\mathcal{R}e\,\lambda_k^*)^2 = \frac{1}{2}\Big(-k^2\pi^2 + k^2((\mathcal{R}e\,z_k^*)^2-(\mathcal{I}m\,z_k^*)^2)$$

$$+\sqrt{(-k^2\pi^2+k^2((\mathcal{R}e\,z_k^*)^2-(\mathcal{I}m\,z_k^*)^2))^2+(2k^2\mathcal{R}e\,z_k^*\,\mathcal{I}m\,z_k^*)^2}\Big)$$

$$= 2k^4(\mathcal{R}e\,z_k^*\,\mathcal{I}m\,z_k^*)^2\Big[k^2\pi^2 - k^2((\mathcal{R}e\,z_k^*)^2-(\mathcal{I}m\,z_k^*)^2)$$

$$+\sqrt{(-k^2\pi^2+k^2((\mathcal{R}e\,z_k^*)^2-(\mathcal{I}m\,z_k^*)^2))^2+(2k^2\mathcal{R}e\,z_k^*\,\mathcal{I}m\,z_k^*)^2}\ \Big]^{-1}.$$

Finally, taking into account the asymptotic expression for z_k^*, (4.31), we obtain that (2.19) holds. □

Proof of Theorem 2.9. (i) The weak convergence of $\{\varphi_{\lambda_k^*}\}_k$ is a direct consequence of the equation they satisfy.

(ii) We prove first that $\{\varphi^3_{\lambda_k^*}\}_k$ does not tend strongly to zero in $H^1(0,1)$. We have

$$||\varphi^3_{\lambda_k^*}||_{H^1(0,1)} = \frac{1}{|\lambda_k^*|^2}\left(\int_0^1 |\cos\ k\pi x|^2 + \int_0^1 |k\pi \sin\ k\pi x|^2\right) = \frac{1+k^2\pi^2}{2|\lambda_k^*|^2}.$$

Since $(\lambda_k^*)^2 = -k^2\pi^2 + k^2\alpha_1 + \mathcal{O}(k) = -k^2\pi^2 + \mathcal{O}(k)$ we obtain that $\varphi^3_{\lambda_k^*}$ does not tend to zero in $H^1(0,1)$. Evidently, $\varphi^4_{\lambda_k^*}$ does not tend to zero in $L^2(0,1)$.

We pass now to the study of $\varphi^1_{\lambda_k^*}$. We evaluate first the expression

$$|a_k|^2 = \left|\frac{1}{\sqrt{(\lambda_k^*)^2+k^2\pi^2}\,\sinh\,(\sqrt{(\lambda_k^*)^2+k^2\pi^2})}\right|^2$$

$$= \frac{1}{|(\lambda_k^*)^2+k^2\pi^2|\,\left(|\sinh\,\mathcal{R}e\,\sqrt{(\lambda_k^*)^2+k^2\pi^2}|^2 + |\sin\,\mathcal{I}m\,\sqrt{(\lambda_k^*)^2+k^2\pi^2}|^2\right)}.$$

Now,

$$\left|\left|\cosh\left(\sqrt{(\lambda_k^*)^2+k^2\pi^2}(y-1)\right)\cos k\pi x\right|\right|^2_{H^1(\Omega)}$$

$$=\frac{1}{2}\int_0^1\left(\left|\cosh\left(\sqrt{(\lambda_k^*)^2+k^2\pi^2}(y-1)\right)\right|^2+k^2\pi^2\left|\cosh\left(\sqrt{(\lambda_k^*)^2+k^2\pi^2}(y-1)\right)\right|^2\right.$$

$$\left.+((\lambda_k^*)^2+k^2\pi^2)\left|\sinh\left(\sqrt{(\lambda_k^*)^2+k^2\pi^2}(y-1)\right)\right|^2\right)$$

$$=\frac{1}{4}\int_0^1\left((-|(\lambda_k^*)^2+k^2\pi^2|+k^2\pi^2+1)\cos\left(\mathcal{I}m\sqrt{(\lambda_k^*)^2+k^2\pi^2}(y-1)\right)\right.$$

$$\left.+(|(\lambda_k^*)^2+k^2\pi^2|+k^2\pi^2+1)\sinh\left(2\mathcal{R}e\sqrt{(\lambda_k^*)^2+k^2\pi^2}(y-1)\right)\right)$$

$$=\frac{(-|(\lambda_k^*)^2+k^2\pi^2|+k^2\pi^2+1)\sin 2\mathcal{I}m\sqrt{(\lambda_k^*)^2+k^2\pi^2}}{8\mathcal{I}m\sqrt{(\lambda_k^*)^2+k^2\pi^2}}$$

$$+\frac{(|(\lambda_k^*)^2+k^2\pi^2|+k^2\pi^2+1)\sinh 2\mathcal{R}e\sqrt{(\lambda_k^*)^2+k^2\pi^2}}{8\mathcal{R}e\sqrt{(\lambda_k^*)^2+k^2\pi^2}}.$$

Taking into account that $\sqrt{(\lambda_k^*)^2+k^2\pi^2}=kz_k^*=\sqrt[3]{k^2\pi^2}-\frac{1}{3}\sqrt[3]{k\pi}\,i+o(\sqrt[3]{k})$, we obtain that

$$||\varphi^1_{\lambda_k^*}||^2_{H^1(\Omega)}\longrightarrow\frac{\sqrt[3]{\pi^2}}{4}.$$

Similarly it turns out that $||\varphi^2_{\lambda_k^*}||_{L^2(\Omega)}$ does not tend to zero. □

5. A noncompactness result. The following result is a direct application of the existence of a sequence of eigenvalues with modulus tending to infinity and uniformly negative real parts.

It is well known that, in the context of one-dimensional hybrid systems, the dissipative term is often a compact perturbation of the differential operator associated with the corresponding conservative system. This argument was used to prove that the decay rate of the energy of those systems is not uniform (see [18]). Nevertheless, in our case, this kind of argument cannot be used since the dissipative term $(0,0,0,W_t)$ is, at least apparently, a bounded but not compact perturbation of the conservative operator. It is natural to study whether this term produces a compact perturbation of the underlying conservative system.

A way to do this consists of analyzing whether the difference between the semigroup generated by the conservative operator and the semigroup generated by the dissipative one is compact or not. The existence of the sequence $(\lambda_k^*)_k$ of eigenvalues implies that the answer is negative.

THEOREM 5.1. *Let $\{S_D(t)\}_{t\geq 0}$ be the semigroup generated by the dissipative operator and let $\{S_C(t)\}_{t\geq 0}$ be the semigroup generated by the conservative system. Then, for all $t > 0$, the difference $(S_D - S_C)(t)$ is not a compact operator in $\mathcal{X}$.*

Proof. Suppose that there exists $t_0 > 0$ such that $(S_D - S_C)(t_0)$ is compact. Theorem 2.9 implies that there exists a sequence of eigenfunctions $\{\varphi_{\lambda_k^*}\}_k$, corresponding to the eigenvalues λ_k^*, which converges weakly to zero in $\mathcal{X}$. So,

$$||(S_C(t_0) - S_D(t_0))\varphi_{\lambda_k^*}||_{\mathcal{X}} \longrightarrow 0 \text{ when } k \longrightarrow \infty.$$

Since $\varphi_{\lambda_k^*}$ is an eigenfunction of the dissipative problem we have that

$$S_D(t_0)\varphi_{\lambda_k^*} = e^{\lambda_k^* t_0}\varphi_{\lambda_k^*}.$$

Hence,

$$||S_C(t_0)\varphi_{\lambda_k^*} - e^{\lambda_k^* t_0}\varphi_{\lambda_k^*}||_{\mathcal{X}} \longrightarrow 0 \text{ when } k \longrightarrow \infty. \tag{5.1}$$

Since the conservative operator generates a group of isometries (see [16]) we get that

$$||S_C(t_0)\varphi_{\lambda_k^*}||_{\mathcal{X}} = ||\varphi_{\lambda_k^*}||_{\mathcal{X}},$$

and therefore

$$||\varphi_{\lambda_k^*}||_{\mathcal{X}} = ||S_C(t_0)\varphi_{\lambda_k^*}||_{\mathcal{X}} \leq ||S_C(t_0)\varphi_{\lambda_k^*} - e^{\lambda_k^* t_0}\varphi_{\lambda_k^*}||_{\mathcal{X}} + \left|e^{\lambda_k^* t_0}\right| ||\varphi_{\lambda_k^*}||_{\mathcal{X}}. \tag{5.2}$$

In view of Theorem 2.8 we have that the sequence $(\lambda_k^*)_k$ has the property that $\mathcal{R}e\,\lambda_k^* \to -\frac{1}{3}$, when $k \to \infty$, and hence, there exists $k_1 \in \mathbb{N}$ such that $\mathcal{R}e\,\lambda_k^* < -\frac{1}{4}$ for all $k > k_1$.

We deduce that, for all $t > 0$, there exists a constant ε, depending on t but independent of k, such that

$$\left|e^{\lambda_k^* t}\right| = e^{\mathcal{R}e\,\lambda_k^* t} < 1 - \varepsilon.$$

Let us take $t = t_0$ in the last equality. Going back to (5.2), we obtain

$$\varepsilon||\varphi_{\lambda_k^*}||_{\mathcal{X}} \leq ||S_C(t_0)\varphi_{\lambda_k^*} - e^{\lambda_k^* t_0}\varphi_{\lambda_k^*}||_{\mathcal{X}}. \tag{5.3}$$

Remark that (5.1) and (5.3) imply that $||\varphi_{\lambda_k^*}||_{\mathcal{X}}$ goes to zero when $k \to \infty$ and this is a contradiction with the result of Theorem 2.9.

Finally, we obtain that $(S_D - S_C)(t)$ is not a compact operator for any $t > 0$. ∎

Remark 14. In order to compare the noncompactness result of Theorem 5.1 for our two-dimensional case with analogous one-dimensional models we consider the following problem (see [18]):

$$\begin{cases} u_{tt} - u_{xx} = 0, & x \in (0,1), \quad t > 0, \\ u(t,0) = 0, & t > 0, \\ u_{tt}(t,1) + u_t(t,1) = -u_x(t,1), & t > 0. \end{cases} \tag{5.4}$$

This is a "string-mass" model since it couples the vibrations of a string with a rigid body at the end $x = 1$ (see [9] and [18]).

The natural energy space corresponding to (5.4) is

$$\mathcal{Z} = V \times L^2(0,1) \times \mathbb{R},$$

where $V = \left\{ v \in H^1(0,1) : v(0) = 0 \right\}$.

Observe that if we define the energy of a solution u of (5.4) by

$$E(t) = \frac{1}{2} \int_0^1 \left(|u_t|^2 + |u_x|^2 \right) \, dx + \frac{1}{2} |u_t(t,1)|^2, \tag{5.5}$$

we obtain that

$$\frac{dE}{dt}(t) = -(u_t(t,1))^2 \leq 0. \tag{5.6}$$

Therefore we are dealing with a dissipative hybrid system, $u_t(t,1)$, in the last relation of (5.4), being the damping term.

Let us now consider the vector-valued unknown $U = (u, u_t, u(\cdot\,,1))$ and write equation (5.4) in the following abstract form:

$$\begin{cases} U_t + A_D(U) = 0, & t > 0, \\ U(0) = U_0. \end{cases} \tag{5.7}$$

The operator A_D in (5.7) is an unbounded operator in $\mathcal{Z}$ defined by

$$\begin{gathered} \mathcal{D} := \mathcal{D}(A_D) = \{ U \in \mathcal{Z} : A_D(U) \in \mathcal{Z} \} \\ = \left\{ U = (u,v,p) \in H^2(0,1) \cap V \times V \times \mathbb{R} : u(1) = p \right\}, \\ A_D(u,v,p) = (-v, -u_{xx}, v(1) + u_x(1)). \end{gathered}$$

It is easy to show that $(\mathcal{D}, A_D)$ is a maximal monotone operator in $\mathcal{Z}$.

Let us now consider the projection operator

$$B : \mathcal{Z} \to \mathcal{Z}, \quad B(u,v,p) = (0,0,p).$$

Observe that B is a compact operator in $\mathcal{Z}$ and $A_C = A_D - B$ is the conservative operator corresponding to (5.4).

Let $\{T_D(t)\}_{t \geq 0}$ be the strongly continuous semigroup generated by the dissipative operator A_D and let $\{T_C(t)\}_{t \geq 0}$ be the strongly continuous semigroup generated by the conservative operator A_C.

For (5.4) all the eigenvalues of the operator A_D approach the imaginary axis when the frequency increases. This is one of the consequences of the fact that A_D is obtained from A_C by a compact perturbation B. In the case of our system (1.1) this is not the case; the perturbation term is only bounded in the energy space. This is one of the major differences between one- and two-dimensional hybrid systems.

Moreover, since B is a compact operator, it can be shown that, for all $t \geq 0$, the difference $(T_D - T_C)(t)$ is a compact operator in $\mathcal{Z}$.

6. Comments. Our results indicate that the interaction between the fluid and structure in this type of model is very weak at high frequencies. As a consequence of this, if we try to change the dynamics of the system acting only on the string located on Γ_0, we have to impose very restrictive conditions on the data of the system. This

explains the results obtained in the context of the controllability of these systems and concerning the existence of periodic solutions (see [12] and [13]). The analysis of these problems was based on nonharmonic Fourier series (see [2]) and asymptotics for the spectrum.

The weak interaction of the string and the fluid is a consequence of both the hybrid structure of the system and of the localization of the string in a relatively small part of the boundary of Ω.

In [11] we analyze a slightly different model in which the domain Ω is a ball of $\mathbb{R}^2$ and the dissipation acts on the whole boundary. We prove that the energy does not decay uniformly. This clearly shows that the very weak interaction between fluid and structure at high frequencies is due to the hybrid structure of the system.

From our study the property of completeness of the eigenfunctions of the differential operator associated with (1.1) is easy to prove. The question of whether these eigenfunctions form a Riesz basis is open (for the notions of completeness and Riesz basis see [6]). For the one-dimensional systems, obtained by separation of variables fixing the number of oscillations in the x-variable, we can prove that the eigenfunctions do form a Riesz basis. However our estimates are not enough to give an answer to this question in the context of the two-dimensional problem.

We also remark that we have been able to obtain very precise information about the eigenvalues because we had the explicit equation they satisfy. We got this equation by separation of variables, which was possible since we considered Neumann boundary conditions for the string. The analysis in the case of Dirichlet boundary conditions for the string is much more difficult. Partial results, like the nonuniform decay of the energy of the system, were obtained in [11] (see also [14]).

The analysis of the rate of decay of low frequencies is a relevant problem for applications. Obviously, the techniques developed in this paper do not allow us to answer this question. This problem requires different approaches.

7. Appendix. We present here the proofs of Lemmas 4.1, 4.2, 4.3, and 4.4.

Proof of Lemma 4.1. If α is a root of (4.2), then

$$\mid k \mid\mid \alpha \mid^3 = \mid \alpha^2 - \pi^2 + \alpha\sqrt{\alpha^2 - \pi^2} \mid \leq 2 \mid \alpha \mid^2 + \pi \mid \alpha \mid + \pi^2 \leq \max\{4 \mid \alpha \mid^2, 4\pi^2\}.$$

We obtain that

$$\mid \alpha \mid \leq \max\left\{\frac{4}{k}, \sqrt[3]{\frac{4\pi^2}{k}}\right\} < \frac{2\pi}{\sqrt[3]{k}} \text{ for all } k \geq 1. \tag{7.1}$$

On the other hand we have

$$\mid k \mid\mid \alpha \mid^3 = \mid \alpha^2 - \pi^2 + \alpha\sqrt{\alpha^2 - \pi^2} \mid = \frac{\pi^2 \mid \alpha^2 - \pi^2 \mid}{\mid \alpha^2 - \pi^2 - \alpha\sqrt{\alpha^2 - \pi^2} \mid}$$

$$\geq \frac{\pi^2 \mid \alpha^2 - \pi^2 \mid}{\mid \alpha \mid^2 + \pi^2 + \mid \alpha \mid \sqrt{\mid \alpha \mid^2 + \pi^2}} \mid \geq \frac{\pi^2(\pi^2 - \mid \alpha \mid^2)}{\mid \alpha \mid^2 + \pi^2 + \mid \alpha \mid (\mid \alpha \mid + \pi)}.$$

In view of (7.1) we obtain that, if $k > 8\pi^3$, then

$$k \mid \alpha \mid^3 > \frac{\pi^2(\pi^2 - 1)}{2 + \pi^2 + \pi} > \frac{\pi^3}{8}. \qquad \square$$

Proof of Lemma 4.2. We study the relation between the roots of (4.2) and those of the equation

$$kz^3 - \pi^2 = 0. \tag{7.2}$$

The last equation has three roots $a_i = \sqrt[3]{\frac{\pi^2}{k}}\omega_i\,,\ i = 1, 2, 3$, where ω_i are the three cubic roots of unity.

We consider the functions $u(z) = kz^3 - \pi^2$ and $v(z) = z\sqrt{z^2 - \pi^2} + z^2$ defined in the circle δ_0 of center 0 and radius $\frac{2\pi}{\sqrt[3]{k}}$, where both are analytic.

In the circle δ_0 we have

$$\mid v(z) \mid = \mid z\sqrt{z^2 - \pi^2} + z^2 \mid \leq \mid z \mid (\sqrt{\mid z \mid^2 + \pi^2} + \mid z \mid) \leq \frac{10\pi^2}{\sqrt[3]{k}},$$

and hence

$$\mid v(z) \mid < \frac{10\pi^2}{\sqrt[3]{k}} \text{ if } \mid z \mid \leq \frac{2\pi}{\sqrt[3]{k}}. \tag{7.3}$$

On the other hand,

$$\mid u(z) \mid = \mid kz^3 - \pi^2 \Big| = k \mid z - \sqrt[3]{\frac{\pi^2}{k}}\omega_1 \Big| \Big| z - \sqrt[3]{\frac{\pi^2}{k}}\omega_2 \Big| \Big| z - \sqrt[3]{\frac{\pi^2}{k}}\omega_3 \Big|.$$

If z belongs to the circumference δ_0 we have that

$$\left| z - \sqrt[3]{\frac{\pi^2}{k}}\omega_i \right| \geq |z| - \left| \sqrt[3]{\frac{\pi^2}{k}}\omega_i \right| = \frac{2\pi}{\sqrt[3]{k}} - \sqrt[3]{\frac{\pi^2}{k}} > \sqrt[3]{\frac{\pi^3}{k}}, \quad i = 1, 2, 3.$$

Hence

$$\mid u(z) \mid > \pi^3 \text{ if } \mid z \mid = \frac{2\pi}{\sqrt[3]{k}}. \tag{7.4}$$

The inequalities (7.3) and (7.4) imply that $|u(z)| > |v(z)|$ for all z on the circumference δ_0.

Applying Rouché's theorem, we obtain that (4.2) has the same number of roots as (7.2) in the circle δ_0. It follows that (4.2) has three roots which satisfy (4.3). The inequality (7.3) is still valid in δ_i, $i = 1, 2, 3$. On the other hand, for all z on the circumference δ_i,

$$\mid u(z) \mid = \mid kz^3 - \pi^2 \mid = k \left| z - \sqrt[3]{\frac{\pi^2}{k}}\omega_1 \right| \left| z - \sqrt[3]{\frac{\pi^2}{k}}\omega_2 \right| \left| z - \sqrt[3]{\frac{\pi^2}{k}}\omega_3 \right|$$

$$> k\frac{10}{\sqrt[3]{k^2}} \left(\frac{\pi}{\sqrt[3]{k}} \right)^2 = \frac{10\pi^2}{\sqrt[3]{k}}.$$

Applying Rouché's theorem, we deduce that the roots of (4.2) are located in the circles δ_i and the estimate (4.5) holds. □

Proof of Lemma 4.3. Step 1: We prove first that the equation

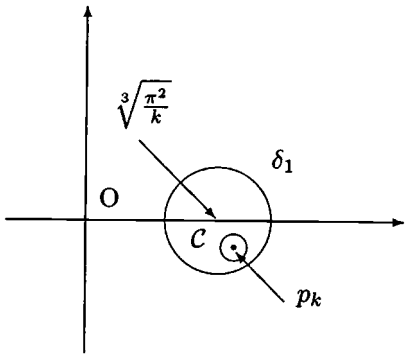

FIG. 7.1.

$$-\pi^2 + kz^3 + \pi z\, i = 0 \tag{7.5}$$

has a unique solution p_k in the circle δ_1 of center $\sqrt[3]{\frac{\pi^2}{k}}$ and radius $\frac{10}{\sqrt[3]{k^2}}$ and hence

$$p_k = \sqrt[3]{\frac{\pi^2}{k}} - \frac{1}{3}\sqrt[3]{\frac{\pi}{k^2}}\, i + o\left(\frac{1}{\sqrt[3]{k^2}}\right).$$

The existence of the root p_k in δ_1 follows by applying the estimates obtained in Lemma 4.2 to the functions $u(z) = -\pi^2 + kz^3$ and $v(z) = \pi z\, i$.

We define now $r_k = p_k - \sqrt[3]{\frac{\pi^2}{k}}$ and we deduce that r_k satisfies

$$kr_k^3 + 3kr_k^2\sqrt[3]{\frac{\pi^2}{k}} + 3kr_k\sqrt[3]{\frac{\pi^4}{k^2}} + \pi r_k\, i + \pi\sqrt[3]{\frac{\pi^2}{k}}\, i = 0.$$

Multiplying the last equation by $\sqrt[3]{k}$ we deduce that

$$\sqrt[3]{k}\left(3kr_k\sqrt[3]{\frac{\pi^4}{k^2}} + \pi\sqrt[3]{\frac{\pi^2}{k}}\, i\right) = \sqrt[3]{k}\left(-kr_k^3 - 3kr_k^2\sqrt[3]{\frac{\pi^2}{k}} - \pi r_k\, i\right).$$

Since $|r_k| = \left|p_k - \sqrt[3]{\frac{\pi^2}{k}}\right| \le \frac{10}{\sqrt[3]{k^2}}$ we have that $r_k = -\frac{1}{3}\sqrt[3]{\frac{\pi}{k^2}}\, i + o\left(\frac{1}{\sqrt[3]{k^2}}\right)$.

Hence, $p_k = \sqrt[3]{\frac{\pi^2}{k}} - \frac{1}{3}\sqrt[3]{\frac{\pi}{k^2}}\, i + o\left(\frac{1}{\sqrt[3]{k^2}}\right)$.

Step 2: We prove now that the root α_1 of (4.2) belongs to the circle $\mathcal{C}$ centered in p_k and of radius $s_k = \frac{1}{\sqrt[4]{k^3}}$ (see Fig. 7.1). This implies immediately that α_1 satisfies (4.5).

We use again Rouché's theorem considering the functions

$$u(z) = -\pi^2 + kz^3 + \pi z\, i, \quad v(z) = -z^2 - z\sqrt{z^2 - \pi^2} + \pi z\, i.$$

For z in δ_1 we have

$$|v(z)| = |z|^2\left|-1 - \frac{z}{\sqrt{z^2 - \pi^2} + \pi\, i}\right| \le 2|z|^2 \le \frac{100}{\sqrt[3]{k^2}}.$$

On the other hand, applying Taylor's formula in the point p_k, we get

$$u(z) = u'(p_k)(z-p_k) - \frac{(z-p_k)^2}{2\pi\, i}\int_{\widehat{\gamma}} \frac{u(\zeta)\, d\zeta}{(\zeta-p_k)^2(\zeta-z)},$$

where $\widehat{\gamma}$ is the circle of center p_k and radius $S_k = \frac{1}{\sqrt[3]{k}}$.

We estimate first the quantity

$$\left|\frac{(z-p_k)^2}{2\pi\, i}\int_{\widehat{\gamma}} \frac{u(\zeta) d\,\zeta}{(\zeta-p_k)^2(\zeta-z)}\right| \leq \frac{s_k^2}{2\pi}\, \frac{M}{S_k^2(S_k-s_k)}\, 2\pi S_k \leq 2M\, \sqrt[12]{\frac{1}{k^{10}}},$$

where M is an upper bound for u in $\widehat{\gamma}$.

On the other hand

$$|z-p_k|\,|u'(p_k)| = s_k|3kp_k^2 + \pi\, i| \geq s_k\,(3k|p_k|^2 - \pi) \geq \frac{1}{\sqrt[4]{k^3}}\left(3k\sqrt[3]{\frac{\pi^4}{k^2}} - \pi\right) \geq \frac{1}{2}\,\sqrt[12]{\frac{1}{k^5}}.$$

We obtain that for k sufficiently large and z on the circumference $\mathcal{C}$,

$$|u(z)| > |z-p_k|\,|u'(p_k)| - 2M\,\sqrt[12]{\frac{1}{k^{10}}} \geq \frac{1}{2}\,\sqrt[12]{\frac{1}{k^5}} - 2M\,\sqrt[12]{\frac{1}{k^{10}}} > \frac{1}{4\,\sqrt[12]{k^5}} > \frac{100}{\sqrt[3]{k^2}}.$$

We conclude that, for k sufficiently large, $|u(z)| > |v(z)|$ on the circumference of $\mathcal{C}$.

Applying Rouché's theorem, we deduce that α_1 satisfies (4.5). □

Proof of Lemma 4.4. We simply remark that, making $z = -s$, the equation (4.6) is transformed into (4.2). □

Acknowledgments. The first author wishes to thank all organizers of the Project MATAROU TEMPUS JEP 2797 and especially Professor Doina Cioranescu for their support and dedication to this program.

REFERENCES

[1] B. Allibert, *Contrôle analytique de l'équation des ondes sur des surfaces de révolution*, Ph.D. dissertation, Ecole Polytechnique, France, 1997.

[2] J. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems*, Comm. Pure Appl. Math., XXXII (1979), pp. 555–587.

[3] H. T. Banks, W. Fang, R. J. Silcox, and R. C. Smith, *Approximation Methods for Control of Acoustic/Structure Models with Piezoceramic Actuators*, J. Intelligent Material Systems Structures, 4 (1993), pp. 98–116.

[4] H. T. Banks, R. C. Smith, and Y. Wang, *Smart Material Structures. Modeling, Estimation and Control*, Research in Applied Mathematics, John Wiley & Sons, Masson, 1996.

[5] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[6] I. C. Goberg and M. G. Krein, *Introduction to the Theory of Lineal Nonselfadjoint Operators*, AMS, Providence, RI, 1969.

[7] S. Hansen and E. Zuazua, *Exact Controllability and Stabilization of a Vibrating String with an Interior Point Mass*, SIAM J. Control Optim., 33 (1995), pp. 1357–1391.

[8] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Pergamon Press, Elmsford, NY, 1987.

[9] W. Littman and L. Marcus, *Some recent results on control and stabilization of flexible structures*, in Proc. of the COMCON Workshop on Stabilization of Flexible Structures, Montpellier, France, 1987, Optimization Software, Inc., New York, pp. 151–161.

[10] W. Littman and L. Marcus, *Exact boundary controllability of a hybrid system of elasticity*, Arch. Rational Mech. Anal., 103 (1988), pp. 193–236.

[11] S. Micu, *Análisis de un modelo híbrido bidimensional fluido-estructura*, Ph.D. dissertation, Universidad Complutense de Madrid, Madrid, Spain, 1996.
[12] S. Micu and E. Zuazua, *Propriétés qualitatives d'un modèle hybride bi-dimensionnel intervenant dans le cotrôle du bruit*, C. R. Acad. Sci. Paris, 319 (1994), pp. 1263–1268.
[13] S. Micu and E. Zuazua, *Boundary controllability of a linear hybrid system arising in the control of noise*, SIAM J. Control Optim., 35 (1997), pp. 1614–1637.
[14] S. Micu and E. Zuazua, *On a weakly damped system arising in the control of noise*, Proc. 7th International Conference on Control and Estimation of Distributed Parameter System, Vorau, 1996.
[15] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[16] A. Pazy, *Semi-groups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
[17] J. Ralston, *Solutions of the wave equation with localized energy*, Comm. Pure Appl. Math., 22 (1969), pp. 807–823.
[18] B. Rao, *Uniform Stabilization of a Hybrid System of Elasticity*, SIAM J. Control Optim., 33, 2 (1995), pp. 440–454.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 1002–1021, July 1998

010

EVOLUTION OF MIXED-STATE REGIONS IN TYPE-II SUPERCONDUCTORS*

CHAOCHENG HUANG† AND THOMAS SVOBODNY†

Abstract. A mean-field model for dynamics of superconducting vortices is studied. The model, consisting of an elliptic equation coupled with a hyperbolic equation with discontinuous initial data, is formulated as a system of nonlocal integrodifferential equations. We show that there exists a unique classical solution in $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ for all $t > 0$, where Ω_0 is the initial vortex region that is assumed to be in $C^{1+\alpha}$. Consequently, for any time t, the vortex region Ω_t is of $C^{1+\alpha}$, and the vorticity is in $C^{\alpha}\left(\bar{\Omega}_t\right)$.

Key words. high-temperature superconductor, nonequilibrium superconductivity, mixed-state region, vorticity, London equations

AMS subject classifications. 82D55, 76C05

PII. S003614109731504X

1. Introduction.

One of the phenomena that characterize a superconducting material is the Meissner effect. This refers to the exclusion from the material of time-independent as well as time-varying magnetic fields.

This state of exclusion, the Meissner phase, is independent of past history. Materials are superconducting, and thus exhibit a Meissner state, only below a certain critical temperature T_c. On the other hand, at any $0 < T < T_c$, the Meissner state is destroyed and the magnetic field penetrates the whole material (normal phase) when the magnetic field exceeds some critical value $H_c(T)$. A relation between magnetic field $\mathbf{H}$ and current $\mathbf{J}$ in the material was proposed to explain the Meissner effect:

$$\lambda^2 \nabla \times \mathbf{J} + \mathbf{H} = 0, \tag{1.1}$$

where λ is a characteristic length scale. With Ampére's law,

$$\mathbf{J} = \nabla \times \mathbf{H};$$

this leads to the London equation [11]

$$\lambda^2 \nabla \times \nabla \times \mathbf{H} + \mathbf{H} = 0.$$

It follows from this equation—and this has been corroborated by experiments—that λ gives the depth of penetration of the magnetic field.

The London equations follow from the Ginzburg–Landau equations, which couple the electrodynamics to the dynamics of an order parameter, in the limit as $\kappa = \lambda/\xi$ gets arbitrarily large, where ξ, the so-called coherence length, represents the length scale on which the order parameter (density of superconductivity) varies [6]. Thus, the London equations represent a superconductor with zero stiffness in the order parameter.

*Received by the editors January 15, 1997; accepted for publication (in revised form) July 27, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/31504.html
†Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435 (chuang@math.wright.edu, svobodny@math.wright.edu).

For high-κ (type-II) superconductors, the London relation (1.1) needs to be modified [11]. It was observed that the Meissner phase obtains for magnetic fields below a certain critical field H_{c_1}, that the normal phase obtains above a higher critical field H_{c_2}, and that a different phase, the so-called Abrikosov–Shubnikov, or mixed-state, phase obtains for intermediate values of the magnetic field between H_{c_1} and H_{c_2}. In this phase the magnetic field penetrates the material in the form of quantized vortices; each vortex carries a quantum of flux, ϕ_0, known as a fluxon. These vortices interact with each other and move under the influence of applied and induced currents. As $\kappa \to \infty$, the difference between H_{c_1} and H_{c_2} increases so that for high-κ materials, which include the high temperature superconductors, this mixed state is the phase of importance.

The London equations for a single vortex filament Γ are then

$$\lambda^2 \nabla \times \nabla \times \mathbf{H} + \mathbf{H} = -\phi_0 \delta_\Gamma.$$

A mean-field model for the mixed state was arrived at in [4] by averaging the above equations over the individual vortices

$$\lambda^2 \nabla \times \nabla \times \mathbf{H} + \mathbf{H} = -\boldsymbol{\omega}.$$

The variable $\boldsymbol{\omega}$ represents the density of quantum vortices and will be referred to as the vorticity. The vorticity is assumed to be convected at a velocity $\mathbf{u}$, which is the terminal speed in the presence of Lorentz forces due to the mean field (and is perpendicular to the current). In the case that the vorticity and the magnetic field remain in a fixed direction, say $x_3 - axis$, i.e., $\boldsymbol{\omega} = (0, 0, \omega)$, $\mathbf{H} = (0, 0, H)$, the complete system then reads (see [4])

$$\omega_t + \nabla \cdot (\omega u) = 0, \tag{1.2}$$

$$\Delta H - H = -\omega, \tag{1.3}$$

$$u = -sign\,(\omega)\, \nabla H. \tag{1.4}$$

In the region $\Omega_t = \{\omega(\cdot, t) \neq 0\}$, where vortices exist, the material is in the phase of the mixed state. The boundary then represents the interface between the mixed-state phase and the superconducting phase. The evolution of such a boundary is important since any such motion is manifested as electrical resistance [7]. An approach taken in [1] is to calculate the forces experienced by any vortex due to a magnetic field formed by integration over fluxons. Among other configurations, the authors studied, via numerical simulations, the evolution of the vortex lattice starting from a configuration where vortices are concentrated in a bounded region. In the mean-field setting (1.2)–(1.4), this configuration corresponds to the case of an initially isolated mixed-state domain $\bar{\Omega}_0$ evolving in the environment of a Meissner phase. In other words, the initial data should be taken as

$$\omega(x, 0) = \omega_0(x) = \varpi_0(x)\, \chi_{\bar{\Omega}_0}(x), \tag{1.5}$$

where $\chi_{\bar{\Omega}_0}$ is the characteristic function of Ω_0, which is the initial mixed-state or vortex region, and ϖ_0 is a continuous function in the whole space. In the present paper, we are mainly concerned with the problem (1.2)–(1.4) along with the initial condition (1.5).

Since the initial vorticity (1.5) is discontinuous only on $\partial\Omega_0$, one expects that the discontinuity will evolve with the velocity u. Hence, the motion equation (1.2)

is understood in the distribution sense. In order to define solutions in appropriate spaces, it has been proposed in [4] to treat the system as the following free boundary problem:

$$\omega_t = \nabla \cdot (|\omega| \nabla H) \quad \text{in } \Omega_t, \tag{1.6}$$

$$\Delta H - H = -\omega \quad \text{in } \Omega_t, \tag{1.7}$$

$$\Delta H - H = 0 \quad \text{in } R^2 \backslash \Omega_t, \tag{1.8}$$

$$[H] = \left[\frac{\partial H}{\partial n}\right] = 0 \quad \text{on } \Gamma_t = \partial\Omega_t, \tag{1.9}$$

$$V_n = -sign\,(w)\,\frac{\partial H}{\partial n} \quad \text{on } \Gamma_t, \tag{1.10}$$

where Ω_t is the moving domain initially at Ω_0, n is the outward normal, V_n is the normal velocity of the moving boundary Γ_t, and the bracket $[\cdot]$ denotes the jump across Γ_t.

In the present paper we propose a different approach to deal with the problem (1.2)–(1.5). The system will be formulated as the following integro-differential equation:

$$\begin{aligned} &\frac{d\Phi(x,t)}{dt} = -\int_{\Omega_t} \nabla K(\Phi(x,t) - y)\left(J(\Phi)^{-1}\varpi_0\right)\left(\Phi^{-1}(y,t)\right) dy, \\ &\Phi(x,0) = x \quad \text{for } x \in \bar{\Omega}_0, \end{aligned} \tag{1.11}$$

where $\Phi : \bar{\Omega}_0 \times [0,T) \mapsto R^2$, $K(x)$ is the Green's function for the elliptic equation (1.3), J is the Jacobian, and $\Phi^{-1}(\cdot,t)$ is the inverse mapping for any fixed t. One of the advantages of the above formulation is that we can work on the fixed domain.

The main intention is to investigate classical solutions Φ for the system (1.11). We shall study uniqueness, global existence, and regularity of solutions for system (1.11).

This approach is motivated by [8] in which the authors used a system analogue to (1.11) to study motion of charged particles. We shall modify the method developed in [8] to establish short-time existence and uniqueness of the solution for (1.11). The treatment for long-time existence is partially motivated by [3]. One observes that system (1.2)–(1.5) has a certain similarity to vorticity evolution for a two-dimensional incompressible Euler system. Roughly speaking, in a two-dimensional Euler system, instead of (1.3), the relationship between the vorticity and the fluid velocity is through the Biot–Savart law [10]. When $\varpi_0(x)$ is a constant (and consequently $\varpi(x,t)$ remains constant for all t), a global smooth solution for a two-dimensional Euler system was established in [3], [5]. In our system (1.2)–(1.5), the vorticity $\varpi(x,t)$ has a more complicated structure. The main idea introduced in this paper is to estimate—instead of a C^α norm as one usually did (see [3], [8], and [5])—a C^β norm of $\varpi(\cdot,t)$ for some $0 < \beta < \alpha$. We then use this norm to bound the velocity.

The paper is organized as follows. In section 2, some preliminaries, notations, and main results will be introduced. Uniqueness and short-time existence will be investigated in section 3. Section 4 will be devoted to the derivation of some a priori estimates for solutions. Global existence will be proved in section 5.

2. Preliminaries and main results. Throughout the paper, we assume that Ω_0 is a bounded domain and that

$$\partial\Omega_0 \in C^{1+\alpha}, \quad \varpi_0(x) > 0, \quad \varpi_0 \in C^\alpha. \tag{2.1}$$

Suppose that u and ω are smooth. Then the equation (1.2) can be rewritten as

$$\omega_t + (u \cdot \nabla)\,\omega = -\omega \nabla \cdot u. \tag{2.2}$$

Let $\Phi(x,t)$ be the solution of

$$\frac{d\Phi}{dt} = u(\Phi, t), \quad \Phi(x,0) = x \quad \text{for } x \in \bar{\Omega}_0. \tag{2.3}$$

By (2.2), $\omega(\Phi(x,t),t)$ solves

$$\frac{d\omega}{dt} = -\omega \nabla \cdot u. \tag{2.4}$$

Let $J(\Phi)$ be the Jacobian of Φ. It is known that $J(\Phi)$ solves

$$\frac{dJ(\Phi)}{dt} = J(\Phi)\,\nabla \cdot u \tag{2.5}$$

so that $J(\Phi)^{-1}$ is the solution of

$$\frac{dJ(\Phi)^{-1}}{dt} = -J(\Phi)^{-1}\,\nabla \cdot u. \tag{2.6}$$

Comparing (2.4) to (2.6), it follows from the uniqueness theory of ordinary differential equations (ODE) that

$$\omega(\Phi(x,t),t) = J(\Phi)^{-1}(x,t)\,\omega_0(x) \quad \text{for } x \in \bar{\Omega}_0. \tag{2.7}$$

Since at $t = 0$, $J(\Phi) = 1$, the expression (2.7) suggests that $\omega(\Phi(x,t),t) > 0$ for $x \in \bar{\Omega}_0$, provided $J(\Phi)(x,t)$ does not vanish for t. Set

$$\Omega_t = \Phi(\Omega_0, t).$$

Then Ω_t represents the mixed-state region at time t. We extend $\omega(x,t)$ by 0 for $x \notin \bar{\Omega}_t$.

Let $K(x)$ be the fundamental solution of the elliptic equation (1.3) that has the form [2]

$$K(x) = \frac{1}{2\pi} K_0(|x|) = \frac{1}{2\pi}\left(-\ln(|x|) + S(|x|)\right), \tag{2.8}$$

where K_0 is the 0th-order modified Bessel's function of the second kind (or Hankel function of imaginary part) and S is its regular part. Hence, assuming that $H(x,t) \to 0$ as $|x| \to \infty$, we have, from (1.3),

$$H(x,t) = \int_{\Omega_t} K(x-y)\,\omega(y,t)\,dy. \tag{2.9}$$

It is easy to check that

$$\nabla H(x,t) = \nabla \int_{R^2} K(x-y)\,\omega(y,t)\,dy = \int_{\Omega_t} \nabla_x K(x-y)\,\omega(y,t)\,dy. \tag{2.10}$$

Substituting this expression, (1.4), and (2.7) into (2.3), and noting that $\omega(x,t) > 0$ in $\bar{\Omega}_t$, we arrive at the following integro-differential equation for $\Phi(x,t)$ in $\bar{\Omega}_0 \times [0,T)$:

$$\frac{d\Phi(x,t)}{dt} = -\int_{\Omega_t} \nabla K(\Phi(x,t) - y) \left(J(\Phi)^{-1} \varpi_0\right) \left(\Phi^{-1}(y,t), t\right) dy, \tag{2.11}$$
$$\Phi(x,0) = x \quad \text{for } x \in \bar{\Omega}_0.$$

Before proceeding to state our main results, we need to introduce some function spaces and notations.

For any subset $G \subseteq R^2$, multi-index $\beta = (\beta_1, \beta_2)$, $m = |\beta|$, $0 < \alpha < 1$, and any function f in G, denote by $|f|_{m+\alpha}$ and $\|f\|_{m+\alpha}$, respectively, the Hölder seminorm and norm defined as

$$|f|_{m+\alpha} = \sup_{x,y \in G,\, |\beta| = m} \frac{\left|D^\beta f(x) - D^\beta f(y)\right|}{|x-y|^\alpha}$$

and

$$\|f\|_{m+\alpha} = \sup_{x \in G,\, |\beta| \le m} \left|D^\beta f(x)\right| + |f|_{m+\alpha}.$$

Denote by $C^{m+\alpha}(G)$ the set of all functions $f(x)$ defined in G such that $\|f\|_{m+\alpha}$ is finite. If $f(x,t)$ is defined in G_t for $t < T$, where $G_t \subseteq R^2$ depends on t, we sometimes use the notation $f \in C_x^{m+\alpha}(G_t)$ to specify that $f(\cdot,t) \in C^{m+\alpha}(G_t)$ for any fixed t. For clarity, sometimes we shall also use the notations $|f(t)|_{m+\alpha,G_t}$ and $\|f(t)\|_{m+\alpha,G_t}$ to specify the dependence on the domain G_t. We also introduce the following notation:

$$|f(t)|_{\inf,\partial G_t} = \inf_{x \in \partial G_t} |f(x,t)|.$$

Definition 2.1. *A function $\Phi(x,t)$, defined for $(x,t) \in \bar{\Omega}_0 \times [0,T)$ with values in R^2, is called a $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ solution of (2.11) for $t < T$ if, for any fixed $t < T$, $\Phi(\cdot,t), D_t\Phi(\cdot,t) \in C^{1+\alpha}\left(\bar{\Omega}_0\right)$, $\Phi^{-1}(\cdot,t) \in C^{1+\alpha}\left(\bar{\Omega}_t\right)$, and $\Phi(x,t)$ solves (2.11) pointwise in $\bar{\Omega}_0 \times [0,T)$.*

We shall verify in the next section that in the class of $C^{1+\alpha}$, formulation (2.11) is equivalent to (1.2)–(1.5). We conclude this section with a statement of the main result of this paper.

Theorem 2.2. *Assume (2.1). Then there exists a unique $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ solution $\Phi(x,t)$ for (2.11) for $t > 0$. Consequently, the mixed-state region $\bar{\Omega}_t$ is of $C^{1+\alpha}$ for all $t > 0$.*

3. Short-time existence. Let $\Phi(x,t)$ be a $C^{1+\alpha}(\Omega_0)$ function for fixed t, and $J(\Phi) \neq 0$. Introduce an operator A by

$$A(\Phi)(x,t) = x - \int_0^t \int_{\Omega_s} \nabla K(\Phi(x,s) - z) \left(J(\Phi)^{-1} \omega_0\right) \left(\Phi^{-1}(z,s), s\right) dz ds. \tag{3.1}$$

In this section, we shall show that under the assumption of (2.1), this operator has a unique fixed point in $C_x^{1+\alpha}$ for $0 < t < T$ for some $T > 0$. This fixed point is obviously a $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ solution for (2.11).

Notice that at $r = 0$ the singular part of $\nabla K(x)$ (see (2.8)) is the Newtonian kernel $x/|x|$. We need the following modified version of [8, Lemma 3.1] that will be frequently used throughout the section.

LEMMA 3.1. *Let Ω be a bounded domain in R^2. Suppose that there exists a $\varphi \in C^{1+\alpha}\left(R^2\right)$ such that $\Omega = \{\varphi(x) < 0\}$ and that $\inf_{\partial\Omega} |\nabla\varphi(x)| > 0$. Define function $w(x)$ and $G(x)$, for any $g \in C^{\alpha}$, by*

$$w(x) = P_v \int_{\Omega} \nabla\left(\frac{x-z}{|x-z|^2}\right) dz,$$

$$G(x) = \int_{\Omega} \nabla\left(\frac{x-z}{|x-z|^2}\right) (g(x) - g(z))\, dz,$$

where P_v means the principal value. Then

$$|w|_{0,\Omega} \le c \ln(2 + \delta d(\Omega)), \tag{3.2}$$

$$|w|_{\alpha,\Omega} \le c\delta \ln(2 + \delta d(\Omega)), \tag{3.3}$$

$$|G|_{\alpha,\Omega} \le c\,|g|_{\alpha,\Omega} \ln(2 + \delta d(\Omega)), \tag{3.4}$$

where c is a constant depending only on α and Ω, $d(\Omega)$ is the diameter of Ω, and

$$\delta = \frac{|\nabla\varphi|_{\alpha,\partial\Omega}}{|\nabla\varphi|_{\inf,\partial\Omega}}. \tag{3.5}$$

Proof. By analyzing the proof of [3, Proposition 1], one found that $|w|_{0,\Omega} \le c$ if $\delta d(\Omega) < 1$. Assertion (3.2) then follows from [3, Proposition 1]. By carefully tracking various constants in the proof of [8, Lemma 3.1] and using (3.2), inequality (3.3) follows. Estimate (3.4) follows from (3.2) and [8, Lemma 3.2].

It follows that $\nabla A(\Phi)(x,t)$ exists for $x \in \Omega_0$ and that it has interior limit for $x \in \partial\Omega_0$. One can actually compute $\nabla A(\Phi)$ as

$$\nabla A(\Phi)(x,t) = I + A_1(\Phi)(x,t) + A_2(\Phi)(x,t), \tag{3.6}$$

for $x \in \bar{\Omega}_0$, where

$$A_1(\Phi)(x,t) = -\int_0^t P_v \int_{\Omega_s} \nabla^2 K(\Phi(x,s) - z) \cdot \left(J(\Phi)^{-1}\omega_0\right)\left(\Phi^{-1}(z,s), s\right) \nabla\Phi(x,s)\, dz ds, \tag{3.7}$$

$$A_2(\Phi)(x,t) = -\frac{1}{2\pi}\int_0^t \left(J(\Phi)^{-1}\omega_0\right)(x,s)\, \nabla\Phi(x,s)\, ds. \tag{3.8}$$

For convenience, in the following we will leave out the designation P_v; all singular integrals in the paper shall be understood as the principal values.

By (3.6)–(3.8), we can show that the $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ solution is a weak solution in the following sense.

DEFINITION 3.2. *Assume* (2.1). *A pair of functions* $(\omega, H) \in L^2\left(R^2 \times (0,T)\right) \times L^2\left(W^{2,2}\left(R^2\right),(0,T)\right)$ *is called a weak solution of* (1.2)–(1.5) *for* $0 \le t < T$ *if*

$$\int_0^T \int_{R^2} \omega\,(x,t)\,\nabla H\,(x,t)\,\nabla \xi\,(x,t)\,dxdt$$

$$= \int_0^T \int_{R^2} \omega\,(x,t)\,\xi_t\,(x,t)\,dxdt + \int_{R^2} \omega_0\,(x)\,\xi\,(x,0)\,dx, \tag{3.9}$$

$$\Delta H\,(x,t) - H\,(x,t) = -\omega\,(x,t)\,, \qquad H\,(x,t) \longrightarrow 0 \quad \text{as } |x| \longrightarrow \infty, \tag{3.10}$$

for any $\xi\,(x,t) \in C_0^\infty\left(R^2 \times [0,T)\right)$.

PROPOSITION 3.3. *Assume* (2.1). *Let* Φ *be a* $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ *solution of* (2.11) *for* $t < T$. *Define* $\omega\,(x,t)$ *by* (2.7) *for* $x \in \bar{\Omega}_t$ *and by* 0 *otherwise, and define* H *by* (2.9). *Then* (ω, H) *is a classical solution of* (1.2)–(1.5) *in the sense that the equations* (2.4), (1.3), *and* (1.5) *hold pointwise in* $\bar{\Omega}_t$, *with* $u = -\nabla H$ *and* $\nabla \cdot u\,(x,t)$ *being understood as the limit from the interior for* $x \in \partial\bar{\Omega}_t$, *and that* (1.3) *holds in* R^2. *The converse is also true. Consequently, any* $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ *solution is a weak solution.*

Proof. We only sketch the proof. Suppose that Φ is a $C^{1+\alpha}\left(\bar{\Omega}_0\right)$ solution of (2.11). By (2.7), (2.11), and Lemma 3.1, it is clear that, for any $x \in \bar{\Omega}_0$, $\omega\,(\Phi\,(x,t)\,,t)$ is differentiable in t, and that $H\,(\cdot,t) \in C^{2+\alpha}\,(\Omega_t)$. From expression (3.6)–(3.8), one can easily verify (2.4) and (1.3)–(1.5). Obviously, $(\omega, H) \in L^2 \times W^{2,2}$, and (3.10) follows from (2.9). By (2.7), for any $\xi\,(x,t) \in C_0^\infty\left(R^2 \times [0,T)\right)$ and $x \in \Omega_0$,

$$\begin{aligned}
&\frac{d}{dt}\left(\omega\left(\Phi\,(x,t)\,,t\right) J\,(\Phi)\,(x,t)\,\xi\left(\Phi\,(x,t)\,,t\right)\right)\\
&\quad = \omega\left(\Phi\,(x,t)\,,t\right) J\,(\Phi)\,(x,t)\,\frac{d}{dt}\xi\left(\Phi\,(x,t)\,,t\right)\\
&\quad = \omega\left(\Phi\,(x,t)\,,t\right) J\,(\Phi)\,(x,t)\left(\nabla\xi\left(\Phi\,(x,t)\,,t\right)\cdot\frac{d\Phi\,(x,t)}{dt} + \xi_t\left(\Phi\,(x,t)\,,t\right)\right),
\end{aligned}$$

where $\xi_t\,(x,t) = \partial\xi/\partial t$. Hence, by integration in x and t, we obtain, by (2.3),

$$\begin{aligned}
-\int_{\Omega_0} \omega_0\,(x)\,\xi\,(x,0)\,dx = &-\int_0^T \int_{\Omega_0} \omega\left(\Phi\,(x,t)\,,t\right)\nabla H\cdot\nabla\xi\left(\Phi\,(x,t)\,,t\right) J\,(\Phi)\,(x,t)\,dxdt\\
&+\int_0^T \int_{\Omega_0} \omega\left(\Phi\,(x,t)\,,t\right)\xi_t\left(\Phi\,(x,t)\,,t\right) J\,(\Phi)\,(x,t)\,dxdt.
\end{aligned}$$

Assertion (3.9) follows by changes of variables $y = \Phi\,(x,t)$ and by the fact that $\omega\,(x,t) = 0$ outside $\bar{\Omega}_t$.

We next derive some $C^{1+\alpha}$ estimates for the operator A defined in (3.1) and use them to establish a fixed point. Notice that the function K_0 in (2.8) has the specific form

$$K_0\,(r) = -\ln r + S\,(r)\,, \qquad S\,(r) = -\left(\ln\frac{r}{2} + \gamma\right) I_0 - \ln\frac{\gamma}{2} + I_1,$$

where γ is the Euler constant (≈ 0.56), and

$$I_0 = \sum_{i=1}^{\infty} \frac{(r/2)^{2i}}{i!\Gamma(i+1)}, \quad I_1 = \sum_{i=1}^{\infty} \frac{(r/2)^{2i}}{(i!)^2} \left(\sum_{j=1}^{i} \frac{1}{j} \right).$$

The following properties can be verified through direct computations [2]:

(i) $S(r)$ is smooth for $r \geq 0$;

(ii) $K_0(r) = (e^{-r}/\sqrt{r})(1 + O((1/r)))$, as $r \to \infty$;

(iii) $K_0(r) = -\ln r + O(r)$, as $r \to 0$.

We need the following estimates for solutions of the equation (1.3).

LEMMA 3.4. *Suppose that* $\omega(x) = \varpi(x)\chi_\Omega$, *where* $\varpi \in C^\alpha$, $\Omega = \{\varphi(x) < 0\}$ *with* $\varphi \in C^{1+\alpha}$, $|\nabla\varphi|_{\inf,\partial\Omega} > 0$. *Let* H *be the solution of* (1.3) *that vanishes at infinity. Then*

$$0 \leq H \leq |\varpi|_{0,\Omega}, \tag{3.11}$$

$$\|H\|_{2,\Omega} \leq c\,|\varpi|_{0,\Omega} \ln\left[(2 + \delta d(\Omega))\left(2 + |\varpi|_{0,\Omega}^{-1} |\varpi|_{\alpha,\Omega}\, d(\Omega)\right)\right], \tag{3.12}$$

$$\|H\|_{2+\alpha,\Omega} \leq c\left(\|\varpi\|_{\alpha,\Omega} + \delta\,|\varpi|_{0,\Omega}\right) \ln(2 + \delta d(\Omega)), \tag{3.13}$$

where c *is a universal constant,* $d(\Omega)$ *is the diameter of* Ω, *and*

$$\delta = \frac{|\nabla\varphi|_{\alpha,\partial\Omega}}{|\nabla\varphi|_{\inf,\partial\Omega}}.$$

Proof. The inequalities (3.11) follow from the maximum principle for elliptic equations. By (2.10), for $x \in \bar{\Omega}$, we have the integral formula for $\nabla^2 H$ (analogous to (3.6)) as follows:

$$\begin{aligned}
\nabla^2 H(x,t) &= \int_\Omega \nabla^2 K(x-z)\,\omega(z)\,dz + \frac{1}{2\pi}\omega(x) \\
&= \frac{1}{2\pi}\int_\Omega \nabla\left(\frac{x-z}{|x-z|^2}\right) dz\omega(x) \\
&\quad + \frac{1}{2\pi}\int_\Omega \nabla\left(\frac{x-z}{|x-z|^2}\right)(\omega(z) - \omega(x))\,dz \\
&\quad + \frac{1}{2\pi}\omega(x) + \int_\Omega \nabla^2 S(|x-z|)\,\omega(z)\,dz = k_1 + k_2 + k_3 + k_4.
\end{aligned} \tag{3.14}$$

By Lemma 3.1, it is easy to see that

$$\|k_1\|_\alpha + \|k_3\|_\alpha \leq c\left(\|\varpi\|_{\alpha,\Omega} + \delta\,|\varpi|_{0,\Omega}\right) \ln(2 + \delta d(\Omega)). \tag{3.15}$$

To estimate k_2, we write

$$k_2 = \frac{1}{2\pi}\left(\int_{\Omega\setminus B_\varepsilon} + \int_{B_\varepsilon}\right) \nabla\left(\frac{x-z}{|x-z|^2}\right)(\omega(z) - \omega(x))\,dz = k_{21} + k_{22},$$

where B_ε is the ball centered at x with radius ε that will be chosen later on. By integration,

$$|k_{21}| \le c\,|\varpi|_{0,\Omega} \int\limits_{\Omega\setminus B_\varepsilon} \frac{1}{|x-z|^2}\,dz \le c\,|\varpi|_{0,\Omega} \ln\left(\frac{d(\Omega)}{\varepsilon}\right)$$

if $\varepsilon \le d(\Omega)$. Otherwise, $k_{21}=0$. Since $\varpi \in C^\alpha$, we find that

$$|k_{22}| \le c\,|\varpi|_{\alpha,\Omega} \int\limits_{B_\varepsilon} \frac{1}{|x-z|^{2-\alpha}}\,dz \le c\,|\varpi|_{\alpha,\Omega}\,\varepsilon^\alpha.$$

By choosing $\varepsilon^\alpha = |\varpi|_{0,\Omega}\left(|\varpi|_{\alpha,\Omega}\right)^{-1}$, we deduce that

$$|k_2| \le c\,|\varpi|_{0,\Omega}\left(1+\left|\ln\left(2+|\varpi|_{0,\Omega}^{-1}\,|\varpi|_{\alpha,\Omega}\,d(\Omega)\right)\right|\right). \tag{3.16}$$

By Lemma 3.1, we also have

$$|k_2|_\alpha \le c\,|\varpi|_{\alpha,\Omega} \ln\left(2+\delta d(\Omega)\right). \tag{3.17}$$

Combining (3.15)–(3.17), one sees that $|k_1+k_2+k_3|_0$ is bounded by the left-hand side of (3.12) and that $\|k_1+k_2+k_3\|_\alpha$ is bounded by the left-hand side of (3.13). The assertions thus follow from the fact that $\nabla^2 S$ is smooth in R^2 and that $\nabla^2 S$ decays at the rate r^{-2} as $r\to\infty$.

THEOREM 3.5. *Suppose that the initial data satisfy* (2.1). *Then there exists a unique* $C^{1+\alpha}(\Omega_0)$ *solution* $\Phi(x,t)$ *of* (2.11) *for* $t<T$ *for some* $T>0$.

Proof. For simplicity, we assume that there exists $\varphi_0 \in C^{1+\alpha}$ with $|\nabla\varphi_0|_{\inf,\partial\Omega_0} \neq 0$ such that $\Omega_0 = \{\varphi_0(x)<0\}$. For any $M,T>0$ to be chosen later on, we define a set $W(M,T)$ of vector value functions in $\bar\Omega_0\times[0,T)$ as follows:

$$\begin{aligned} W(M,T) = \{&\Phi(x,t)\in R^2 :\ \Phi(x,0)=x,\\ &\|\Phi(t)\|_{1+\alpha,\Omega_0}\le M,\ \|\Phi(x)\|_{\alpha,[0,T)}\le M,\ \ |\nabla\Phi - I|_0 \le 1/2\}. \end{aligned}$$

Since $|\nabla\Phi - I|\le 1/2$, $\Phi^{-1}(\cdot,t)$ exists and maps Ω_t onto Ω_0 so that the mapping A defined in (3.1) is well defined for any $\Phi\in W(M,T)$. By applying Lemma 3.4 to $A(\Phi)$ with

$$\varpi = \left(J(\Phi)^{-1}\,\omega_0\right)\left(\Phi^{-1}(x,s),s\right), \tag{3.18}$$

we find that

$$\begin{aligned} \|A(\Phi)\|_{1+\alpha,\Omega_0} \le c_0 + c\int\limits_0^t &\left(\|\Phi\|_{1+\alpha,\Omega_0}\right)^{1+\alpha}\left(1+\delta_s\,|\varpi|_{0,\Omega_s}+\|\varpi\|_{\alpha,\Omega_s}\right)\\ &\cdot\left(\ln\left(2+d(\Omega_s)\,\delta_s\right)+\ln\left(2+|\varpi|_{0,\Omega}^{-1}\,\|\varpi\|_{\alpha,\Omega_s}\,d(\Omega_s)\right)\right)ds, \end{aligned} \tag{3.19}$$

where c is a constant independent of Φ, $c_0=\sup_{x\in\Omega_0}|x|$, δ_t is defined in terms of Ω_t by

$$\delta_t = \frac{|\nabla\varphi(x,t)|_{\alpha,\partial\Omega_t}}{|\nabla\varphi(x,t)|_{\inf,\partial\Omega_t}} = \frac{\left|\nabla\varphi_0\left(\Phi^{-1}(x,t)\right)\nabla\Phi^{-1}(x,t)\right|_{\alpha,\partial\Omega_t}}{|\nabla\varphi_0\left(\Phi^{-1}(x,t)\right)\nabla\Phi^{-1}(x,t)|_{\inf,\partial\Omega_t}}, \tag{3.20}$$

since $\Omega_t = \{\varphi(x,t) < 0\}$, where $\varphi(x,t) = \varphi_0\left(\Phi^{-1}(x,t)\right)$. Notice that, by direct calculation,

$$\left|\nabla\Phi^{-1}\right|_{\alpha,\Omega_t} \le |\nabla\Phi|_{\alpha,\Omega_0}\left(\left|\nabla\Phi^{-1}\right|_{0,\Omega_t}\right)^{2+\alpha},$$
$$\left\|J(\Phi)^{-1}\right\|_{\alpha,\Omega_t} \le c_0\left(|\nabla\Phi|_{\alpha,\Omega_0}+1\right)\left(\left|\nabla\Phi^{-1}\right|_{0,\Omega_t}+1\right)^{3+\alpha},$$

where c_0 is a universal constant. Since $\Phi \in W(M,T)$, it is easy to see that

$$\delta_t \le c(M), \quad d(\Omega_t) \le 2|\Phi|_0 \le 2M, \quad \|\varpi\|_{\alpha,\Omega_t} \le c(M),$$

where the last inequality is due to (3.18), and the constant $c(M)$ is a polynomial of M with coefficients depending only on initial data. Hence (3.19) results in, for $t < T$,

$$\|A(\Phi)(t)\|_{1+\alpha,\Omega_0} \le c_0 + c(M)T.$$

It is easy to derive the following estimates:

$$\|A(\Phi)(x,\cdot)\|_{\alpha} \le c(M)T^{1-\alpha}$$

and

$$|\nabla(A(\Phi)) - I| \le c(M)T.$$

We now choose $M = 1+c_0$ and $T = \min\left((2C(M))^{-1}, (1+c_0)^{1/(1-\alpha)} C(M)^{-1/(1-\alpha)}\right)$. Then $A(\Phi) \in W(M,T)$. The mapping A maps $W(M,T)$ into itself.

For any $\Phi,\ \tilde{\Phi} \in W(M,T)$. Set $\Omega_t = \Phi(\Omega_0,t)$, $\tilde{\Omega}_t = \tilde{\Phi}(\Omega_0,t)$ and define

$$\rho(t) = \left|\Phi(t) - \tilde{\Phi}(t)\right|_{0,\Omega_0}.$$

By changing variables in the expressions for $A(\Phi)$ and $A\left(\tilde{\Phi}\right)$, we obtain, from (3.1),

$$A(\Phi)(x,t) = x + \int_0^t \int_{\Omega_0} \nabla K(\Phi(x,s) - \Phi(z,s))\,\omega_0(z)\,dzds,$$
$$A\left(\tilde{\Phi}\right)(x,t) = x + \int_0^t \int_{\Omega_0} \nabla K\left(\tilde{\Phi}(x,s) - \tilde{\Phi}(z,s)\right)\omega_0(z)\,dzds.$$

It follows that

$$\left|A(\Phi)(x,t) - A\left(\tilde{\Phi}\right)(x,t)\right|$$
$$\le c\int_0^t \int_{\Omega_0} \left|\nabla K(\Phi(x,s) - \Phi(z,s)) - \nabla K\left(\tilde{\Phi}(x,s) - \tilde{\Phi}(z,s)\right)\right| dzds. \tag{3.21}$$

For $\varepsilon > 0$ to be determined later, we decompose the right-hand side of (3.21) as

$$\int_{\Omega_0} \left|\nabla K(\Phi(x,s) - \Phi(z,s)) - \nabla K\left(\tilde{\Phi}(x,s) - \tilde{\Phi}(z,s)\right)\right| dz \tag{3.22}$$
$$= \int_{\Omega_0\setminus B_\varepsilon(x)} + \int_{\Omega_0\cap B_\varepsilon(x)} = k_1 + k_2.$$

From (2.8),

$$\nabla K(z) = -\frac{z}{2\pi |z|^2} + S'(|z|)\frac{z}{2\pi |z|}.$$

Since $\nabla\Phi^{-1}$ and $\nabla\tilde{\Phi}^{-1}$ are bounded, we have

$$|\Phi(x,s) - \Phi(z,s)|, \qquad \left|\tilde{\Phi}(x,s) - \tilde{\Phi}(z,s)\right| \geq c|x-z|.$$

Therefore,

$$\begin{aligned}&\left|\nabla K(\Phi(x,s) - \Phi(z,s)) - \nabla K\left(\tilde{\Phi}(x,s) - \tilde{\Phi}(z,s)\right)\right| \\ &\qquad \leq c\rho(s)\left(1 + \frac{1}{|x-z|} + \frac{1}{|x-z|^2}\right).\end{aligned}$$

Consequently, k_1 in (3.22) is bounded by

$$|k_1| \leq c\rho(s)\int_{\varepsilon}^{d(\Omega_0)}\left(r + 1 + \frac{1}{r}\right)dr = c\rho(s)(1 + |\ln\varepsilon|). \tag{3.23}$$

Since $|\Phi(x,s) - \Phi(z,s)| \leq c\varepsilon$ for $|x-z| \leq \varepsilon$, we know that

$$S'(|\Phi(x,s) - \Phi(z,s)|) \leq c\varepsilon \quad \text{for} \quad |x-z| \leq \varepsilon.$$

It follows from the obvious estimates

$$\begin{aligned}|\nabla K(\Phi(x,t) - \Phi(z,t))| &\leq |\Phi(x,t) - \Phi(z,t)|^{-1} + S'(|\Phi(x,s) - \Phi(z,s)|) \\ &\leq c\left(1 + |x-z|^{-1}\right)\end{aligned}$$

and

$$\begin{aligned}\left|\nabla K\left(\tilde{\Phi}(x,s) - \tilde{\Phi}(z,s)\right)\right| &\leq \left|\tilde{\Phi}(x,s) - \tilde{\Phi}(z,s)\right|^{-1} + S'(|\Phi(x,s) - \Phi(z,s)|) \\ &\leq c\left(1 + |x-z|^{-1}\right)\end{aligned}$$

that the term k_2 in (3.22) can be estimated as

$$|k_2| \leq c\int_0^{\varepsilon}\left(1 + \frac{1}{r}\right)r\,dr \leq c\left(\varepsilon + \varepsilon^2\right). \tag{3.24}$$

We now choose $\varepsilon = \min(\rho(t), 1)$. From (3.21)–(3.24), it follows that

$$\left|A(\Phi)(x,t) - A\left(\tilde{\Phi}\right)(x,t)\right| \leq c\int_0^t \rho(t)(1 + |\ln\rho(t)|)\,dt. \tag{3.25}$$

Define a sequence $\Phi_n(x,t)$ by

$$\Phi_0 = x, \qquad \Phi_{n+1}(x,t) = A(\Phi_n)(x,t).$$

Since $\Phi_n \in W(M,T)$, this sequence $\{\Phi_n\}$ is precompact under the $C_{x,t}^{1+\gamma,\gamma}$ norm for any $\gamma < \alpha$. Hence we can select a subsequence, still denoting it as $\{\Phi_n\}$, and a function $\Phi \in C_{x,t}^{1+\gamma,\gamma}(\Omega_0 \times [0,T))$ such that

$$\Phi_n \longrightarrow \Phi \text{ in } C_{x,t}^{1+\gamma,\gamma}\text{norm}.$$

This implies that $\Phi \in W(M,T)$ (by checking from the definitions) and by (3.25),

$$\begin{aligned}|\Phi_{n+1}(x,t) - A(\Phi)(x,t)| &= |A(\Phi_n)(x,t) - A(\Phi)(x,t)| \\ &\le cT \sup_{0\le t\le T}\left(|\Phi_n(t) - \Phi(t)|_{0,\Omega_0}\left|\ln|\Phi_n(t) - \Phi(t)|_{0,\Omega_0}\right|\right).\end{aligned}$$

Letting $n \to \infty$, we find that Φ is a fixed point for A, i.e., $A(\Phi) = \Phi$. Next, set

$$\rho_n(t) = \sup_x |\Phi_n(x,t) - \Phi(x,t)|.$$

It follows from (3.25) that

$$\rho_{n+1}(t) \le c\int_0^t \rho_n(t)\left(1 + |\ln\rho_n(t)|\right)\,dt.$$

By [9, section 9], it follows that

$$\rho_n(t) \le c\,|T\ln T|^n.$$

For small T, the above inequality implies uniqueness of the fixed point. The proof is complete.

Using exactly the same argument, we can extend the assertions of Theorem 3.5 to the equation (2.11) with more general initial data:

$$\frac{d\Phi(x,t)}{dt} = -\int_{\Phi\left(\bar\Omega_0,t\right)} \nabla K(\Phi(x,t) - y)\left(J(\Phi)^{-1}\varpi_0\right)\left(\Phi^{-1}(y,t)\right)\,dy, \tag{3.26}$$

$$\Phi(x,0) = \Phi_0(x) \quad \text{for } x \in \bar\Omega_0.$$

COROLLARY 3.6. *In addition to the assumptions in Theorem* 3.5*, suppose also that* $\Phi_0 \in C^{1+\alpha}(\Omega_0)$, $J(\Phi_0) \ne 0$. *Then there exists a unique* $C^{1+\alpha}(\Omega_0)$ *solution* $\Phi(x,t)$ *for* (3.26) *for* $t < T$ *for some* $T > 0$. *Moreover,* T *depends only on* $\|\Phi_0\|_{1+\alpha}$, $\left\|\Phi_0^{-1}\right\|_{1+\alpha}, d(\Omega_0)$ *and* δ_0.

Corollary 3.6 will be used in the next section to extend the solutions for large time.

4. A priori estimates. From Corollary 3.6, it appears that a priori estimates on the $C^{1+\alpha}$ norms of Φ and Φ^{-1} will be sufficient to establish existence of global solutions. We first show that, actually, a uniform bound on the vorticity ω will be enough to guarantee that the solution can be extended for all $t > 0$.

LEMMA 4.1. *Suppose that* $\Phi(x,t)$ *is the* $C^{1+\alpha}$ *solution. Then, for* $x \in \bar\Omega_0$,

$$\omega(\Phi(x,t),t) = \frac{\omega_0(x)\,e^{\sigma(x,t)}}{1 + \omega_0(x)\int_0^t e^{\sigma(x,s)}ds}, \tag{4.1}$$

where

$$\sigma(x,t) = \int_0^t H(\Phi(x,s),s)\, ds. \tag{4.2}$$

Proof. From (2.4), (1.3), (1.4), and the fact that $\omega(\Phi(x,t),t) > 0$ for $x \in \bar{\Omega}_0$, one sees that $\omega(\Phi(x,t),t)$ is the solution of the following ODE:

$$\frac{d\omega}{dt} = \omega H - \omega^2. \tag{4.3}$$

Let $p(t) = \omega(\Phi(x,t),t)^{-1}$. Then p solves

$$\frac{dp}{dt} = -pH + 1.$$

Integrating this ODE directly, we obtain (4.1).

LEMMA 4.2. *Let $\Phi(x,t)$ be the $C^{1+\alpha}$ solution for $t < T$. Suppose that*

$$\eta(T) \equiv \int_0^T \|\varpi(t)\|_{0,\Omega_t}\, dt < \infty. \tag{4.4}$$

Then there exists a $0 < \beta(T) \le \alpha$ depending only on $\eta(T)$ such that for $t < T$,

$$\|\varpi(t)\|_{\beta,\Omega_t} \le c(\eta(T)), \tag{4.5}$$

$c(\eta(T))$ being a constant depending only on $\eta(T)$ and T.

Proof. From Lemma 3.4, we know that $u = -\nabla H$ is Lipschitz in Ω_t. However, the Lipschitz constant may depend on the C^α characters of ω and the domain Ω_t. We claim that ∇H is quasi-Lipschitz, with the constant depending only on $\|\varpi(t)\|_{0,\Omega_t}$, i.e., for $x, y \in \bar{\Omega}_t, |x-y| \le 1/2$,

$$|\nabla H(x,t) - \nabla H(y,t)| \le c\|\varpi(t)\|_{0,\Omega_t} |x-y| (1 + |\ln|x-y||), \tag{4.6}$$

where c is a universal constant. We point out at this moment that c in general also depends on the $L^1(\bar{\Omega}_t)$ norm of $\varpi(\cdot,t)$. However, by (2.7), our previous claim remains true.

Indeed, from (2.8) and (2.9), we have

$$\begin{aligned}\nabla H(x,t) &= \frac{1}{2\pi}\int_{\Omega_t} \frac{x-z}{|x-z|^2}\varpi(z,t)\, dz + \int_{\Omega_t} \nabla_x S(|x-z|)\varpi(z,t)\, dz \\ &= \nabla H_1(x,t) + \nabla H_2(x,t).\end{aligned}$$

The results in [10] lead to

$$|\nabla H_1(x,t) - \nabla H_1(y,t)| \le c\|\varpi(t)\|_{0,\Omega_t} |x-y| (1 + |\ln|x-y||).$$

Since ∇S is smooth and $S'(r) \sim r^{-1}$ for large r, the inequality (4.6) thus follows immediately.

Now, since

$$\frac{d\Phi}{dt} = -\nabla H\left(\Phi, t\right),$$

it follows that the quantity $\rho\left(t\right) = \left|\Phi\left(x,t\right) - \Phi\left(y,t\right)\right|^2$, for $x \neq y \in \bar{\Omega}_t$, satisfies

$$\left|\frac{d\rho\left(t\right)}{dt}\right| \leq c\left\|\varpi\left(t\right)\right\|_{0,\Omega_t} \rho\left(t\right)\left(1 + \left|\ln \rho\left(t\right)\right|\right).$$

By the Gronwall lemma, we deduce that

$$c\left|x-y\right|^{1/\tilde{\beta}(t)} \leq \rho\left(t\right) = \left|\Phi\left(x,t\right) - \Phi\left(y,t\right)\right| \leq c\left|x-y\right|^{\tilde{\beta}(t)}, \tag{4.7}$$

where $\tilde{\beta}\left(t\right) = \exp\left(-c\eta\left(t\right)\right)$ is decreasing in t. The inequality (4.7) implies

$$\left|\Phi\left(t\right)\right|_{\tilde{\beta}(t),\Omega_0} + \left|\Phi^{-1}\left(t\right)\right|_{\tilde{\beta}(t),\Omega_t} \leq c.$$

By the regularity theory for elliptic partial differential equations (PDEs), we know that the solution H of (1.3) is of the class C^1 and

$$\left\|\nabla H\left(t\right)\right\|_0 \leq c\left\|\varpi\left(t\right)\right\|_{0,\Omega_t}. \tag{4.8}$$

Hence, the function $\sigma\left(x,t\right)$ defined in (4.2) is of the class C^{β} and

$$\left|\sigma\left(t\right)\right|_{\tilde{\beta}(t),\Omega_0} \leq \int_0^t \left\|\nabla H\left(s\right)\right\|_0 \left|\Phi\left(s\right)\right|_{\tilde{\beta}(s),\Omega_0} ds \leq c\eta\left(t\right),$$

for $t < T$. The bounds on $\left|\sigma\left(t\right)\right|_{0,\Omega_0}$ can be easily derived from (4.2). Choose $\beta\left(T\right) = \min(\tilde{\beta}\left(T\right)^2, \alpha)$. It then easy to see that the assertion (4.5) hence follows from (4.1).

LEMMA 4.3. *Under the assumptions of Lemma* 4.1*, we have*

$$\left|\nabla\left(\Phi\left(\Phi^{-1}\left(x,t\right),s\right)\right)\right| \leq \exp\left(\int_s^t \left|\nabla u\left(\tau\right)\right|_0 d\tau\right),$$

for $0 < s < t$*, and*

$$\left|\nabla\varphi\left(t\right)\right|_{\inf,\partial\Omega_t} \geq \left|\nabla\varphi\left(0\right)\right|_{\inf,\partial\Omega_0} \exp\left(-\int_0^t \left|\nabla u\left(\tau\right)\right|_0 d\tau\right).$$

Proof. The first inequality follows from the dynamical property of (2.3). Since $\varphi\left(\Phi\left(x,t\right),t\right) = \varphi_0\left(x\right)$, we have

$$\nabla\varphi\left(\Phi\left(x,t\right),t\right) = \left(\left(\nabla\Phi\left(x,t\right)\right)^T\right)^{-1} \nabla\varphi_0\left(x\right).$$

By (2.3), we have

$$\frac{d}{dt}\left(\nabla\Phi\left(x,t\right)\right)^{-1} = -\left(\nabla\Phi\left(x,t\right)\right)^{-1} \nabla u\left(\Phi\left(x,t\right),t\right) = \left(\nabla\Phi\left(x,t\right)\right)^{-1} \nabla^2 H\left(\Phi\left(x,t\right),t\right).$$

Hence,

$$\frac{d}{dt}\nabla\varphi\left(\Phi\left(x,t\right),t\right)=\left(\nabla^{2}H\right)^{T}\nabla\varphi\left(\Phi\left(x,t\right),t\right).$$

The second assertion of Lemma 4.3 follows from the Gronwall lemma.

LEMMA 4.4. *Under the assumptions of Lemma* 4.1*, we have*

$$\delta_{t,\beta}\leq c\left(\eta\left(T\right)\right),$$

where $\delta_{t,\beta}$ *is defined in* (3.20) *with* α *being replaced with* $\beta=\beta\left(T\right)$.

Proof. We recall that by definition, $\Omega_t=\{\varphi<0\}$ with $\varphi=\varphi_0\left(\Phi^{-1}\left(x,t\right)\right)$. Since Φ solves (2.3), we find that φ satisfies

$$\frac{\partial\varphi}{\partial t}+u\cdot\nabla\varphi=0,\qquad u=-\nabla H.$$

Hence $\nabla\varphi$ satisfies

$$\frac{\partial\nabla\varphi}{\partial t}+\left(u\cdot\nabla\right)\nabla\varphi=-\left(\nabla u\right)^{\top}\nabla\varphi. \tag{4.9}$$

Since $\nabla\cdot u=-\Delta H=H-\omega$, this equation can be rewritten as

$$\frac{\partial\nabla^{\perp}\varphi}{\partial t}+\left(u\cdot\nabla\right)\nabla^{\perp}\varphi=\nabla u\nabla^{\perp}\varphi+\left(\omega-H\right)\nabla^{\perp}\varphi, \tag{4.10}$$

where $\nabla^{\perp}\varphi=\left(-D_2\varphi,D_1\varphi\right)$ is divergence free and tangential to $\partial\Omega_t$. By (3.14), we have

$$\begin{aligned}\nabla u&=-\nabla^{2}H\left(x,t\right)\\&=-\frac{1}{2\pi}\int\limits_{\Omega_t}\nabla\left(\frac{x-z}{\left|x-z\right|^{2}}\right)dz\omega\left(x,t\right)\\&\quad-\frac{1}{2\pi}\int\limits_{\Omega_t}\nabla\left(\frac{x-z}{\left|x-z\right|^{2}}\right)\left(\omega\left(z,t\right)-\omega\left(x,t\right)\right)dz\\&\quad-\frac{1}{2\pi}\omega\left(x,t\right)-\int\limits_{\Omega_t}\nabla^{2}S\left(\left|x-z\right|\right)\omega\left(z,t\right)\,dz=k_1+k_2+k_3+k_4.\end{aligned}$$

Since $\nabla^{\perp}\varphi$ is divergence free and tangential to $\partial\Omega_t$, we find by direct computation that

$$k_1\nabla^{\perp}\varphi=\int\limits_{\Omega_t}\nabla\left(\frac{x-z}{\left|x-z\right|^{2}}\right)\left(\nabla^{\perp}\varphi\left(x,t\right)-\nabla^{\perp}\varphi\left(z,t\right)\right)dz\omega\left(x,t\right).$$

Consequently we deduce that, by Lemma 3.1, for $t\leq T$,

$$\begin{aligned}\left\|k_1\nabla^{\perp}\varphi\left(t\right)\right\|_{\beta}&\leq c\left(\left|\omega\left(t\right)\right|_{\beta}\left|\nabla^{\perp}\varphi\left(t\right)\right|_{0}+\left|\omega\right|_{0}\left|\nabla^{\perp}\varphi\left(t\right)\right|_{\beta}\right)\\&\quad\cdot\left(\ln\left(2+\left|\omega\right|_{0}^{-1}\left|\omega\left(t\right)\right|_{\beta}d\left(\Omega_t\right)\right)+\ln\left(2+d\left(\Omega_t\right)\delta_{t,\beta}\right)\right),\end{aligned}$$

where $\beta = \beta(T)$ obtained in Lemma 4.2. In a similar manner, we have

$$\begin{aligned}\left\|k_2\nabla^{\perp}\varphi(t)\right\|_{\beta} \le c\Big(&|\omega(t)|_{\beta}\left|\nabla^{\perp}\varphi(t)\right|_0 + |\omega|_0\left|\nabla^{\perp}\varphi(t)\right|_{\beta}\Big)\\ &\cdot\Big(\ln\Big(2+|\omega|_0^{-1}|\omega(t)|_{\beta}\,d(\Omega_t)\Big)+\ln\left(2+d(\Omega_t)\,\delta_{t,\beta}\right)\Big).\end{aligned}$$

Since $\nabla^2 S$ is smooth, it then follows that

$$\begin{aligned}\left\|\nabla u\nabla^{\perp}\varphi(t)\right\|_{\beta} \le{}& c(\eta(T))\left\|\nabla^{\perp}\varphi(t)\right\|_{\beta}\\ &\cdot\Big(\ln\Big(2+|\omega|_0^{-1}|\omega(t)|_{\beta}\,d(\Omega_t)\Big)+\ln\left(2+d(\Omega_t)\,\delta_{t,\beta}\right)\Big),\end{aligned}$$

where we have used the assertion of Lemma 4.2. Since $\partial\Phi/\partial t = -\nabla H(\Phi,t)$, from (4.5) and (4.8), it is clear that $d(\Omega_t) \le 2|\Phi|_0 \le c(\eta(T))$. Hence

$$\left\|\nabla u\nabla^{\perp}\varphi(t)\right\|_{\beta} \le c(\eta(T))\left\|\nabla^{\perp}\varphi(t)\right\|_{\beta}\ln\left(2+\delta_{t,\beta}\right). \tag{4.11}$$

Next, from (4.10), we have for $x \in \Omega_t$,

$$\begin{aligned}\nabla^{\perp}\varphi(x,t) = {}&\nabla^{\perp}\varphi_0\left(\Phi^{-1}(x,t)\right)+\int_0^t\left(\nabla u\nabla^{\perp}\varphi\right)\left(\Phi\left(\Phi^{-1}(x,t),s\right),s\right)ds\\ &+\int_0^t\left((\omega-H)\nabla^{\perp}\varphi\right)\left(\Phi\left(\Phi^{-1}(x,t),s\right),s\right)ds.\end{aligned}$$

Therefore, by using Lemma 4.3, we find

$$\begin{aligned}\left|\nabla^{\perp}\varphi(t)\right|_{\beta} \le{}&\left|\nabla^{\perp}\varphi_0\right|_{\beta}\exp\left(\int_0^t|\nabla u(\tau)|_0\,d\tau\right)\\ &+c\int_0^t\left(\left|\nabla u\nabla^{\perp}\varphi(s)\right|_{\beta}\right)\exp\left(\int_s^t|\nabla u(\tau)|_0\,d\tau\right)ds\\ &+c\int_0^t\left(\left|(\omega-H)\nabla^{\perp}\varphi(s)\right|_{\beta}\right)\exp\left(\int_s^t|\nabla u(\tau)|_0\,d\tau\right)ds.\end{aligned} \tag{4.12}$$

It follows from (4.11) and (4.8) that

$$\begin{aligned}\left\|\nabla^{\perp}\varphi(t)\right\|_{\beta}\exp\left(-\int_0^t|\nabla u(\tau)|_0\,d\tau\right) \le{}&\left\|\nabla^{\perp}\varphi_0(t)\right\|_{\beta}\\ &+c(\eta(T))\int_0^t\left\|\nabla\varphi(s)\right\|_{\beta}\cdot\ln\left(2+\delta_{t,\beta}\right)\exp\left(-\int_0^s|\nabla u(\tau)|_0\,d\tau\right).\end{aligned} \tag{4.13}$$

Set

$$f(t) = \left\|\varphi(t)\right\|_{1+\beta}\exp\left(-\int_0^t|\nabla u(\tau)|_0\,d\tau\right)+2.$$

Using the second assertion in Lemma 4.3, we obtain

$$\delta_{t,\beta} \leq \frac{|\nabla\varphi(t)|_\beta}{|\nabla\varphi(0)|_{\inf,\partial\Omega_0}} \exp\left(\int_0^t |\nabla u(\tau)|_0 \, d\tau\right).$$

Therefore

$$\ln(2+\delta_{t,\beta}) \leq c\ln f + \int_0^t |\nabla u(\tau)|_0 \, d\tau.$$

It follows from (4.13) that

$$f(t) \leq c + c(\eta(T)) \int_0^t f(s) \left(\ln f(s) + \int_0^s |\nabla u(\tau)|_0 \, d\tau\right) ds.$$

Denote by $h(t)$ the function on the right-hand side of the above inequality:

$$h(t) = c + c(\eta(T)) \int_0^t f(s) \left(\ln f(s) + \int_0^s |\nabla u(\tau)|_0 \, d\tau\right) ds.$$

Then

$$h'(t) \leq c(\eta(T)) h(t) \left(\ln h(t) + \int_0^t |\nabla u(\tau)|_0 \, d\tau\right).$$

It follows that

$$\frac{d}{dt}\ln h(t) \leq c(\eta(T)) \left(\ln h(t) + \int_0^t |\nabla u(\tau)|_0 \, d\tau\right).$$

By the standard Gronwall inequality, we obtain

$$\begin{aligned}
\ln f \leq \ln h(t) &\leq ce^{c(\eta(T))} \left(1 + \int_0^t c(\eta(T)) \int_0^s |\nabla u(\tau)|_0 \, d\tau \, ds\right) \\
&\leq ce^{c(\eta(T))} \left(1 + \int_0^t c(\eta(T)) \, ds \int_0^t |\nabla u(\tau)|_0 \, d\tau\right) \qquad (4.14) \\
&\leq c(\eta(T)) \left(1 + \int_0^t |\nabla u(\tau)|_0 \, d\tau\right),
\end{aligned}$$

where in the last inequality, $c(\eta(T))$ is a constant depending only on $\eta(T)$ and T. By Lemmas 3.4 and 4.2 and equation (4.13) we deduce

$$\int_0^t |\nabla u(s)|_{0,\Omega_s} \, ds = \int_0^t \left|\nabla^2 H(s)\right|_{0,\Omega_s} ds$$

$$\leq c\int_0^t \left(\ln f + \int_0^s |\nabla u(\tau)|_0 \, d\tau\right) ds \tag{4.15}$$

$$\leq c(\eta(T)) + c(\eta(T)) \int_0^t \left(\int_0^s |\nabla u(\tau)|_0 \, d\tau\right) ds.$$

It follows that

$$\int_0^t |\nabla u(\tau)|_0 \, d\tau \leq c(\eta(T)), \quad \text{for } t < T.$$

Substituting this into (4.14), we find $\|\nabla\varphi(t)\|_{\beta(T)}$ is bounded uniformly. Therefore, $\delta_{t,\beta}$ is bounded by a constant depending on $\eta(T)$ and T for $t \leq T$. The proof of Lemma 4.4 is complete.

THEOREM 4.5. *Suppose that the assumptions in Lemma* 4.2 *hold. Then there exists a unique* $C^{1+\alpha}(\Omega_0)$ *solution* $\Phi(x,t)$ *of* (2.11) *for* $t < T + T_0$ *for some* $T_0 > 0$ *depending only on* $\eta(T)$.

Proof. By Corollary 3.6, it suffices to show that $\|\nabla\Phi(t)\|_{\alpha,\Omega_0}$ and $\left\|\nabla\Phi^{-1}(t)\right\|_{\alpha,\Omega_t}$ are uniformly bounded by $c(\eta(T))$. Recall that by differentiating (2.3), $\nabla\Phi(x,t)$ satisfies

$$\frac{\partial\nabla\Phi(x,t)}{\partial t} = \nabla u(\Phi(x,t),t)\nabla\Phi(x,t) = -\nabla^2 H(\Phi(x,t),t)\nabla\Phi(x,t).$$

Applying Lemma 3.4 (estimate (3.12)) with $\alpha = \beta(T)$, and using Lemma 4.2 and 4.4, we deduce

$$\left|\nabla^2 H(\Phi(x,t),t)\right| \leq c(\eta(T)).$$

Therefore,

$$\left|\frac{\partial|\nabla\Phi(t)|_{0,\Omega_0}}{\partial t}\right| \leq c(\eta(T))|\nabla\Phi(t)|_{0,\Omega_0}.$$

It follows that

$$ce^{-c(\eta(T))} \leq |\nabla\Phi(t)|_{0,\Omega_0} \leq ce^{c(\eta(T))}. \tag{4.16}$$

Since, recalling definitions,

$$\omega(x,t) = \left(J(\Phi)^{-1}\omega_0\right)\left(\Phi^{-1}(x,t)\right), \qquad \varphi(x,t) = \varphi_0\left(\Phi^{-1}(x,t)\right),$$

it follows that

$$\delta_{t,\alpha} + \|\omega(t)\|_{\alpha,\Omega_t} + d(\Omega_t) \leq c(\eta(T))\|\Phi(t)\|_{1+\alpha,\Omega_0}.$$

By (3.19), since $A(\Phi) = \Phi$, we obtain

$$\|\Phi(t)\|_{1+\alpha,\Omega_0} \leq c + c(\eta(T))\int_0^t \|\omega(t)\|_{0,\Omega_t}\|\Phi(t)\|_{1+\alpha,\Omega_0}\left(1+\ln\left(1+\|\Phi\|_{1+\alpha,\Omega_0}\right)\right) dt.$$

By the Gronwall inequality, it follows that $\|\Phi(t)\|_{1+\alpha,\Omega_0} \leq c(\eta(t))$ for $t < T$. Combining this estimate and (4.16), we also obtain $\left\|\Phi^{-1}(t)\right\|_{1+\alpha,\Omega_0} \leq c(\eta(t))$ for $t < T$. The proof is now complete.

5. Proof of Theorem 2.2. From Theorem 4.5, it suffices to obtain an a priori estimate for $\eta(t)$ defined in (4.4). We shall use the expression (4.1) to estimate $\eta(t)$.

LEMMA 5.1. *Suppose that Φ is a $C^{1+\alpha}$ solution for $t<T$, and ω is the vorticity. Then, for $t<T$,*

$$|\omega(t)|_{0,\Omega_s} \leq \frac{|\varpi_0|_{0,\Omega_0} e^{\sigma_h(t)}}{1+|\varpi_0|_{0,\Omega_0} \int_0^t e^{\sigma_h(s)} ds}, \tag{5.1}$$

where

$$\sigma_h(t) = \int_0^t |\omega(s)|_{0,\Omega_s} \, ds.$$

Proof. By the maximum principle, we know that $H(x,t) \leq |\omega(t)|_{0,\Omega_s}$ for $t<T$. Since $\omega \geq 0$, by (4.3), we obtain that $\omega(\Phi(x,t),t)$ satisfies

$$\frac{d\omega}{dt} = \omega H - \omega^2 \leq \omega |\omega(s)|_{0,\Omega_s} - \omega^2.$$

Integrating this differential inequality as in Lemma 4.1, we derive

$$\omega(\Phi(x,t),t) \leq \frac{\varpi_0(x) e^{\sigma_h(t)}}{1+\varpi_0(x) \int_0^t e^{\sigma_h(s)} ds}.$$

Noticing that the function $x(1+cx)^{-1}$ is increasing in x (for $c>0$), the assertion follows.

Proof of Theorem 2.2. Let σ_h be the integral defined in Lemma 5.1 and

$$f(t) = \int_0^t e^{\sigma_h(s)} ds.$$

Then

$$f'(t) = e^{\sigma_h(t)}, \qquad f''(t) = e^{\sigma_h(t)} \sigma_h'(t) = |\omega(t)|_{0,\Omega_s} f'(t).$$

It follows from (5.1) that

$$f''(t) \leq \frac{|\varpi_0|_{0,\Omega_0} (f'(t))^2}{1+|\varpi_0|_{0,\Omega_0} f(t)}$$

or equivalently

$$\frac{f''(t)}{f'(t)} \leq \frac{|\varpi_0|_{0,\Omega_0} f'(t)}{1+|\varpi_0|_{0,\Omega_0} f(t)}.$$

By integration, noticing that $f(0)=0,\ f'(0)=1$, we deduce

$$f'(t) \leq 1+|\varpi_0|_{0,\Omega_0} f(t). \tag{5.2}$$

Applying the Gronwall inequality, it follows that

$$f(t) \leq \frac{\left(e^{t|\varpi_0|_{0,\Omega_0}} - 1\right)}{|\varpi_0|_{0,\Omega_0}}.$$

Combining this with (5.2), we find

$$\exp\left(\int_0^t |\omega(s)|_{0,\Omega_s}\, ds\right) = f'(t) \leq e^{t|\varpi_0|_{0,\Omega_0}}.$$

Consequently

$$\eta(t) \leq t\,|\varpi_0|_{0,\Omega_0}.$$

The assertion of Theorem 2.2 follows from Theorem 4.5.

By differentiating the equations (1.2) and (1.3) in x, we find that $\nabla\omega$ satisfies a similar system. Applying the same methods, we can also show the following regularity result.

THEOREM 5.2. *Suppose that* $\Omega_0 \in C^{m+1+\alpha}$, $\varpi_0 \in C^{m+\alpha}$. *Then the solution* Φ *of* (2.11) *is in* $C_x^{m+1+\alpha}(\Omega_0)$.

Acknowledgment. The authors would like to thank the referee for many helpful suggestions.

REFERENCES

[1] S. AKTAS, C. P. POOLE, AND M. A. FARACH, *A numerical study of vortex interactions for high–κ superconductors*, J. Phys.: Condens. Matter, 6 (1994), pp. 7373–7384.
[2] V. ARSENIN, *Basic Equations and Special Functions of Mathematical Physics*, Iliffe Books Ltd., London, 1968.
[3] A. L. BERTOZZI AND P. CONSTANTIN, *Global regularity for vortex patches*, Comm. Math. Phys., 152 (1993), pp. 19–28.
[4] S. J. CHAPMAN, *A mean-field model of superconducting vortices in three dimensions*, SIAM J. Appl. Math., 55 (1995), pp. 1259–1274.
[5] J. Y. CHEMIN, *Persistence de structures geometriques dans les fluides incompressibles bidimensionnels*, Ann. Sci. École Norm. Sup., 26 (1993), pp. 517–542.
[6] Q. DU, M. D. GUNZBURGER, AND J. S. PETERSON, *Analysis and approximation of the Ginzburg-Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.
[7] J. EVETTS, *Concise Encyclopedia of Magnetic and Superconducting Materials*, Pergamon Press, Oxford, 1992.
[8] A. FRIEDMAN AND C. HUANG, *Averaged motion of charged particles under their self-induced electric field*, Indiana Univ. Math. J., 43 (1994), pp. 1167–1225.
[9] A. FRIEDMAN AND J. L. VELAZQUEZ, *A time-dependent free boundary problem modeling the visual image in electrophotography*, Arch. Rational Mech. Anal., 123 (1993), pp. 259–303.
[10] A. MAJDA, *Vorticity and the mathematical theory of incompressible fluid flow*, Comm. Pure Appl. Math., 39 (1986), pp. 187–220.
[11] M. TINKHAM, ED., *Introduction to Superconductivity*, McGraw–Hill, New York, 1996.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 1022–1039, July 1998

011

APPLICATIONS OF THE HOPF–LAX FORMULA FOR $u_t + H(u, Du) = 0$ *

E. N. BARRON†, R. JENSEN†, AND W. LIU‡

Abstract. The γ (sub)level set of the solution to $w_t + H(\gamma, D_x w) = 0$ is the same as the γ (sub)level set of the solution to $u_t + H(u, D_x u) = 0$, and the solution u may be built from w. This result is applied to determining upper and lower bounds for a solution of $u_t + H_1(u, Du) + H_2(u, Du) = 0$, with H_1 convex and H_2 concave, as well as $u_t + H(u, Du) = 0$, but with initial data $u(0, x) = g_1(x) \vee g_2(x)$ or $g_1(x) \wedge g_2(x)$, with g_1 quasi-convex and g_2 quasi-concave. A differential game in L^∞ is constructed giving a new proof of the Hopf formula.

Key words. Hopf–Lax formula, Hamilton–Jacobi equation, differential game, level sets

AMS subject classifications. 35F20, 35B45, 49A45

PII. S0036141097319966

1. Introduction.

Two papers have motivated the problems studied here. First, the paper by Bardi and Evans [1] used a representation theorem for the solution of a Hamilton–Jacobi equation as the value function of a differential game to prove the classical Hopf formula for $u_t + H(Du) = 0$ (see [17], [25]). Second, the paper by Bardi and Faggian [3] used the classical Hopf and Lax formulas to find sophisticated upper and lower bounds to some nonconvex and nonconcave problems. Specifically, they considered two problems. The first uses the Lax formula on the equation $u_t + H_1(Du) + H_2(Du) = 0$, but with arbitrary initial data, with H_1 convex and H_2 concave to get an upper and lower bound for u. The second uses the Hopf formula on the equation $u_t + H(Du) = 0$, with data $u(0, x) = g_1(x) + g_2(x)$ and g_1 convex, g_2 concave to also obtain upper and lower bounds for u. The objective of the present paper is to extend this result as much as possible to Hamiltonians which have u dependence and to quasi-convex–quasi-concave initial data.

We begin by looking at the connection between differential games in L^∞ and Hamilton–Jacobi equations (see [5]–[7]). In section 3, we use a representation due to Subbotin [26] for degree one homogeneous Hamiltonians to represent the solution of $u_t + H(u, Du) = 0$ with quasi-convex terminal data $u(T, x) = g(x)$ as the value function of a differential game in L^∞, see [6]. But we know that the solution of this problem is given by the Hopf formula from [10], [11]. Thus, it is the goal of this section to prove, using the differential game, that we recover the Hopf formula, i.e, that the value of the differential game is given by the Hopf formula. This is an extension of the idea of using optimal control in L^∞ to prove the Lax formula [9] as is done for the classical case in [1].

The following section contains the main result of this paper. For fixed $\gamma \in \mathbb{R}$, we consider the solution to $w_t + H(\gamma, Dw) = 0$ and let u be the solution of $u_t + H(u, Du) = 0$. We prove that

$$\{(t, x) : w(t, x) \leq \gamma\} = \{(t, x) : u(t, x) \leq \gamma\},$$

*Received by the editors April 21, 1997; accepted for publication (in revised form) July 23, 1997; published electronically March 25, 1998. The research of the authors was partially supported by NSF grant DMS-9532030. The first and second authors were also supported by a grant from Loyola University of Chicago.

http://www.siam.org/journals/sima/29-4/31996.html

†Department of Mathematical and Computer Sciences, Loyola University of Chicago, Chicago, IL 60626 (enb@math.luc.edu, rrj@math.luc.edu, wliu@math.luc.edu).

and $u(t,x) = \inf\{\gamma : w(t,x) \le \gamma\}$. This level set result is proved under a variety of hypotheses using the Hopf and Lax formulas of [9]–[11]. The referee has provided a proof (Theorem 4.2) under conditions which require only a comparison principle for the equation and $H(\gamma, p)$ nondecreasing in γ and homogeneous degree one in p. This level set result says that, in some sense, the equation with u in the hamiltonian is no more complicated than the equation without u—a remarkable fact.

Our level set result is of critical use in the proof of the upper and lower bound for solutions of $u_t + H_1(u, Du) + H_2(u, Du) = 0$, with $H_1(\cdot, p)$ convex and $H_2(\cdot, p)$ concave, and arbitrary initial data. The idea, which is due to Bardi and Faggian [3] and Bardi and Osher [2], is to double the variables, replace H_2 by a linear function in p, and apply the Lax formula to the resulting Hamiltonian, which is now convex again. The problem when one has u dependence is that the linear function will depend on u, creating a serious problem for carrying out the rest of the proof. But if we use the level set result, we may fix the u variable and the problem disappears. Then we use the construction of the actual solution from the level sets to obtain the upper and lower bounds in our case. The bounds are more complicated than those of Bardi and Faggian [3], but that is to be expected.

The final section seeks upper and lower bounds for solutions of $u_t + H(u, Du) = 0$, but with initial data for the upper bound $u(0,x) = g_1(x) \vee g_2(x)$, and $u(0,x) = g_1(x) \wedge g_2(x)$ for the lower bound, where g_1 is quasi-convex and g_2 is quasi-concave. In some ways this extends the result of [2] and [3] because, when $u(0,x) = g_1(x) + g_2(x)$, a lower bound can be found from the initial data $2(g_1 \wedge g_2)$ and an upper bound from $2(g_1 \vee g_2)$. The proof of this result is a straightforward adaptation of the doubling argument of [2], [3].

Of course the purpose of obtaining upper and lower bounds for equations is of obvious importance in numerical approximation of solutions.

One must take note that throughout this paper the Hamiltonians are assumed to be homogeneous degree one in the gradient variable. A device of Subbotin [26] allows one to dispense with this assumption in many cases, but as shown in [10], this will impose more severe convexity assumptions on the initial data to yield to our Hopf formula. The difficulty is that the sum of a quasi-convex function and a linear function is quasi-convex if and only if the quasi-convex function is convex. See [10] for a discussion of this.

2. Preliminary results. The following definitions are standard. See [18]-[20], [21], for the theory of quasi-convex duality.

DEFINITION 2.1. *A function $g : \mathbb{R}^n \to \mathbb{R}$ is quasi-convex if $\{x : g(x) \le \alpha\}$ is convex for all $\alpha \in R^1$. Equivalently, g is quasi-convex if $g(\lambda x + (1-\lambda)y) \le \max\{g(x), g(y)\}$ for all $x, y \in \mathbb{R}^n$, and $\lambda \in (0,1)$. g is quasi-concave if $-g$ is quasi-convex.*

The following quasi-convex and quasi-concave conjugates will be used in this paper.

Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ satisfying the condition introduced in [19], $\forall \gamma < \sup_x f(x)$, there exists a continuous affine functional k, such that

$$k(x) \le f(x), \quad \forall\, x \in f^{-1}([-\infty, \gamma]). \tag{2.1}$$

Define

$$f^{\#}(\gamma, p) = \sup\{p \cdot x : x \in \mathbb{R}^n, f(x) \le \gamma\}, \qquad \gamma \in R^1,\ p \in \mathbb{R}^n, \tag{2.2}$$

$$f^{\#\#}(x) = \inf\{\gamma \in R^1 : \sup_{p \in \mathbb{R}^n} (p \cdot x - f^{\#}(\gamma, p)) \le 0\}. \tag{2.3}$$

The definition of $f^{\#}(\gamma,p)$ says that $f^{\#}$ is the support function of the γ (sub)level set of f. The function f is lower semicontinuous and quasi-convex if and only if $f=f^{\#\#}$. In general, $f^{\#\#}$ is the greatest quasi-convex minorant of f.

For quasi-concave conjugates, for $-f$ satisfying (2.1) define

$$f^{c\#}(\gamma,p)=\inf\{p\cdot x : x\in\mathbb{R}^n, f(x)\geq\gamma\}, \qquad \gamma\in R^1,\ p\in\mathbb{R}^n, \tag{2.4}$$

$$f^{c\#\#}(x)=\sup\{\gamma\in R^1 : \inf_{p\in\mathbb{R}^n}(p\cdot x-f^{c\#}(\gamma,p))\geq 0\}. \tag{2.5}$$

The function f is upper semicontinuous and quasi-concave if and only if $f=f^{c\#\#}$. In general, $f^{c\#\#}$ is the greatest quasi-concave majorant of f.

Viewing a quasi-convex function as the supremum of piecewise linear function of a particular type leads to the alternative definition of quasi-convex conjugates reminiscent of the usual Legendre–Fenchel conjugate.

$$f^{*}(\gamma,p)=\sup\{p\cdot x-f(x) : x\in\mathbb{R}^n, f(x)\leq\gamma\}, \qquad \gamma\in R^1,\ p\in\mathbb{R}^n,$$
$$f^{**}(x)=\sup\{(p\cdot x-f^{*}(\gamma,p))\wedge\gamma : \gamma\in\mathbb{R}, p\in\mathbb{R}^n\}.$$

and the quasi-concave conjugates

$$f^{c*}(\gamma,p)=\inf\{p\cdot x+f(x) : x\in\mathbb{R}^n, f(x)\geq\gamma\}, \qquad \gamma\in R^1,\ p\in\mathbb{R}^n,$$
$$f^{c**}(x)=\inf\{(p\cdot x+f^{c*}(\gamma,p))\vee\gamma : \gamma\in\mathbb{R}, p\in\mathbb{R}^n\}.$$

Again, f is lower (upper) semicontinuous and quasi-convex (quasi-concave), if and only if $f=f^{**}$ (respectively, $f=f^{c**}$).

Remark 2.1. The theory of quasi-convex duality and the conjugates defined here is due to Crouzeix, Martinez-Legaz, Penot, Volle, and others. See [14],[15],[18]–[21] and the references therein, especially [19]. Some of their results were rederived in [8] under stronger hypotheses on f, coercivity, for example. The condition (2.1) due to Martinez-Legaz [19] is the weakest possible assumption for duality, i.e., $f=f^{**}$.

In the next section we discuss the Hopf formula from [10], so in this section we record the Lax formula from [9] for the viscosity solution of $u_t+H(u,Du)=0$ on $(0,\infty)\times\mathbb{R}^n$ with initial condition $u(0,x)=g(x)$. The application of the Lax formula assumes that $\gamma\mapsto H(\gamma,p)$ is nondecreasing and upper semicontinuous, $p\mapsto H(\gamma,p)$ is convex and positively homogeneous degree one and continuous, and g is, say, Lipschitz and bounded. Then viewing $H(\gamma,p)$ as the first quasi-convex conjugate of a quasi-convex function $H^{\#}(x)$, we have

$$u(t,x)=\min_{y\in\mathbb{R}^n}\left(g(y)\vee H^{\#}\left(\frac{x-y}{t}\right)\right).$$

The formula is obviously valid for g which is merely continuous, since we may approximate by a sequence of Lipschitz and bounded functions.

3. The differential game in L^∞ and Hamilton–Jacobi equation. The problem we consider here is the *backward* Hamilton–Jacobi equation

$$u_t+H(u,Du)=0, \tag{3.1}$$

on $[0,T)\times\mathbb{R}^n$, with terminal condition

$$u(T,x)=g(x), \qquad x\in\mathbb{R}^n. \tag{3.2}$$

In this section, we study the backward problem due to the desired representation of u as the value function of a differential game for which terminal data is natural.

In [10] and [11] we proved the following theorem.

THEOREM 3.1. *Assume that*

(A) *The function $g : \mathbb{R}^n \to \mathbb{R}$ is continuous and quasi-convex and satisfies* (2.1).

(B) *The Hamiltonian function $H : R^1 \times \mathbb{R}^n \to R^1$ is continuous in both variables and*

(i) $H(\gamma, \lambda p) = \lambda H(\gamma, p)$ *for all* $\gamma \in R^1,\ \lambda \geq 0,\ p \in \mathbb{R}^n$;

(ii) $|H(\gamma, p) - H(\gamma, p')| \leq K_L|p - p'|$ *for all* $\gamma \in [-L, L],\ p, p' \in \mathbb{R}^n$;

(iii) $\gamma \mapsto H(\gamma, p)$ *is nonincreasing for* $\gamma \in R^1$.

Let

$$u(t,x) = \inf\{\gamma \in R^1 : \sup_{p\in\mathbb{R}^n} \inf_{\{y:g(y)\leq\gamma\}} (p \cdot (x-y) + (T-t)H(\gamma,p)) \leq 0\}. \tag{3.3}$$

Then u is a continuous viscosity solution of (3.1)–(3.2). *In other words, using the definition of the quasi-convex conjugates,*

$$u(t,x) = (g^{\#}(\gamma,p) - (T-t)H(\gamma,p))^{\#}(x), \tag{3.4}$$

where the # operation is performed only in x.

The fact that H is homogeneous in p allows us to assume that the solution is bounded, as the next proposition shows.

PROPOSITION 3.1. *Assume* (B) *and let g be continuous. Given $L > 0$ define*

$$\beta(r) = \begin{cases} L, & \text{if } r \geq L, \\ r, & \text{if } |r| \leq L, \\ -L, & \text{if } r \leq -L, \end{cases}$$

and set $w = \beta(u)$, where u is the viscosity solution of (3.1)–(3.2). *Then w is a bounded viscosity solution of*

$$w_t + H(w, D_x w) = 0, \quad (t,x) \in [0,T) \times \mathbb{R}^n, \quad w(T,x) = \beta(g(x)).$$

Proof. We will only prove that w is a subsolution since the proof that it is a supersolution is entirely similar.

Let $w - \varphi$ achieve a unique, zero maximum at $(t_0, x_0), t_0 < T$, with φ a smooth function. If $u(t_0, x_0) > L$, then $u > L$ in a neighborhood of (t_0, x_0) and so $w(t,x) = L$ in this neighborhood. Hence, φ achieves a minimum at (t_0, x_0) and so $\varphi_t(t_0, x_0) = |D_x\varphi(t_0, x_0)| = 0$. Since $H(\gamma, 0) = 0$, we have $\varphi_t(t_0, x_0) + H(w(t_0, x_0), D_x\varphi(t_0, x_0)) = 0$. If $u(t_0, x_0) \leq -L$, then $\varphi(t_0, x_0) \geq -L$, so again $\varphi_t(t_0, x_0) = |D_x\varphi(t_0, x_0)| = 0$. If $|u(t_0, x_0)| < L$, w is a subsolution immediately by definition of β. Finally, we are reduced to the case $u(t_0, x_0) = L$.

Suppose that β_ε is a smooth approximation to β, which is strictly monotone increasing and satisfies $\beta_\varepsilon(r) = r$ when $|r| < L$, and β_ε is linear when $|r| > L + \varepsilon$, and $\beta_\varepsilon \to \beta$ uniformly, as $\varepsilon \to 0$. Let $w_\varepsilon = \beta_\varepsilon(u)$. Then, $w_\varepsilon - \varphi_\varepsilon$ achieves a zero maximum at $(t_\varepsilon, x_\varepsilon)$, where φ_ε is at most a linear translation of φ, and $(t_\varepsilon, x_\varepsilon) \to (t_0, x_0)$ (since (t_0, x_0) is the unique maximum [4]). Also, $w_\varepsilon(t_\varepsilon, x_\varepsilon) \to w(t_0, x_0) = L$ as $\varepsilon \to 0$. But then $u - \beta_\varepsilon^{-1}(\varphi_\varepsilon)$ achieves a zero maximum at $(t_\varepsilon, x_\varepsilon)$. Since u is a subsolution, this implies

$$\frac{1}{\beta_\varepsilon'(\varphi_\varepsilon(t_\varepsilon, x_\varepsilon))}(\varphi_\varepsilon)_t(t_\varepsilon, x_\varepsilon) + H(u(t_\varepsilon, x_\varepsilon), \frac{1}{\beta_\varepsilon'(\varphi_\varepsilon(t_\varepsilon, x_\varepsilon))} D_x\varphi_\varepsilon(t_\varepsilon, x_\varepsilon)) \geq 0.$$

Now we use the homogeneity property of H and the fact that $\beta'_\varepsilon(\varphi_\varepsilon(t_\varepsilon, x_\varepsilon)) > 0$ to conclude that

$$(\varphi_\varepsilon)_t(t_\varepsilon, x_\varepsilon) + H(u(t_\varepsilon, x_\varepsilon), D_x\varphi_\varepsilon(t_\varepsilon, x_\varepsilon)) \geq 0.$$

Letting $\varepsilon \to 0$, we get

$$\varphi_t(t_0, x_0) + H(u(t_0, x_0), D_x\varphi(t_0, x_0)) = \varphi_t(t_0, x_0) + H(w(t_0, x_0), D_x\varphi(t_0, x_0)) \geq 0,$$

and so w is a subsolution of (3.1). Since $w(T, x) = \beta(g(x))$ is bounded, we are done. □

Note also that if g is quasi-convex, so is $\beta(g)$.

In this section, for the purpose of representing the solution of (3.1) as the value function of a differential game in L^∞, we will assume the following strengthened condition on g:

(C) g is uniformly Lipschitz continuous, and coercive, i.e., $\lim_{|x|\to\infty} g(x) = \infty$.

In view of the preceding proposition, without loss of generality, we may replace (B)(ii) by the following assumption:

(D) $|H(\gamma, p) - H(\gamma, p')| \leq K|p - p'|$ for all $\gamma \in R^1,\ p, p' \in \mathbb{R}^n$.

We use K to denote a generic constant greater than both of the Lipschitz constants of H and g.

Remark 3.1. Under the assumptions (A)–(D) there is a unique, bounded, Lipschitz continuous viscosity solution of (3.1)–(3.2) with Lipschitz constant no greater than K. Furthermore, (C) implies that (2.1) holds for g.

Using (A)–(D), we will represent the Hamiltonian H in a form suitable for use in a differential game in L^∞. This representation is introduced in [26].

LEMMA 3.1. *Let* (A)–(D) *hold. Then setting*

$$F(z, q) = \{f \in \mathbb{R}^n : |f| \leq K\sqrt{2} = K',\ f \cdot q \geq H(z, q)\}$$

we have

$$H(\gamma, p) = \max_{q\in B(0,K')} \min_{z\leq\gamma} \min_{f\in F(z,q)} p \cdot f, \qquad |p| \leq K'. \tag{3.5}$$

Here $B(0, K')$ is the closed ball centered at the origin with radius K'.

Proof. By (C) and (B)(iii),

$$\begin{aligned} H(\gamma, p) &\geq H(\gamma, q) - K'|p - q|, \qquad \forall q \in \mathbb{R}^n \\ &= \min_{f\in B(0,K')} H(\gamma, q) + f \cdot (p - q) \\ &= \min_{z\leq\gamma} \min_{f\in B(0,K')} H(z, q) + f \cdot (p - q). \end{aligned}$$

Hence, for $p \in B(0, K')$, we obtain

$$H(\gamma, p) = \max_{q\in B(0,K')} \min_{z\leq\gamma} \min_{f\in B(0,K')} H(z, q) + f \cdot (p - q).$$

Next, using (B)(i),

$$\begin{aligned} H(\gamma, p) = |p|H(\gamma, p/|p|) &= \max_{q\in B(0,K')} \min_{z\leq\gamma} \min_{f\in B(0,K')} H(z, q)|p| + |p|f \cdot (p/|p| - q) \\ &= \max_{q\in B(0,K')} \min_{z\leq\gamma} \min_{f\in B(0,K')} p \cdot f + |p|(H(z, q) - f \cdot q) \\ &\leq \max_{q\in B(0,K')} \min_{z\leq\gamma} \min_{f\in F(z,q)} p \cdot f. \end{aligned}$$

To prove the opposite inequality, Subbotin [26] proves that if one defines the vector

$$f_0 = \begin{cases} qH(z,q) + K'(p\cdot q\ q-p)\sqrt{1-(p\cdot q)^2}, & \text{if } 0 \le p\cdot q < 1,\\ \frac{q-p}{|q-p|}K', & \text{if } -1 \le p\cdot q < 0,\\ qH(z,q), & \text{if } p\cdot q = 1, \end{cases}$$

then f_0 satisfies $|f_0| \le K'$ and both conditions $f_0\cdot q \ge H(z,q)$, $f_0\cdot p \le H(z,p)$, $\forall p,q$. Here z is fixed and p and q may be assumed to be unit vectors by (B)(i). Hence,

$$H(\gamma,p) = \min_{z\le\gamma} H(z,p) \ge \min_{z\le\gamma}\min_{f\in F(z,q)} p\cdot f,$$

and since this is true for every $q \in B(0,K')$, we are done. □

We relabel K' as K in the following.

Remark 3.2. Evans and Souganidis [16] also have a way of representing H as a max-min. We chose this representation from [26] because it gives a particularly easy proof of Theorem 3.2 below.

Now we consider the lower differential game associated with the dynamical system

$$\frac{d\xi}{d\tau} = \varphi(\tau), \quad t < \tau \le T, \qquad \xi(t) = x \in \mathbb{R}^n \tag{3.6}$$

and payoff functional

$$P((\zeta,\varphi),\eta) = g(\xi(T)) \vee \operatorname*{ess\,sup}_{t\le s\le T} \zeta(s). \tag{3.7}$$

The minimizing controls of P are (ζ,φ), while η tries to maximize P. In the lower game, given a control choice η by the maximizing player, the minimizing player uses a nonanticipating strategy Δ, i.e.,

$$\Delta : L^\infty([t,T];B(0,K)) \to L^\infty[t,T]\times L^\infty([t,T];B(0,K)),$$

such that $\Delta[\eta](\tau) = (\zeta(\tau),\varphi(\tau))$ if and only if $\varphi(\tau)\in F(\zeta(\tau),\eta(\tau))$. We shall denote the class of all such strategies by D. Then the lower value of the differential game associated with (3.6) and payoff (3.7) is given by

$$V(t,x) = \inf_{\Delta\in D}\sup_{\eta\in B(0,K)} P((\zeta,\varphi),\eta), \quad t\in[0,T], \quad x\in\mathbb{R}^n, \quad 0\le t\le T.$$

PROPOSITION 3.2. *If* (A)–(D) *hold, then* V *is the unique uniformly Lipschitz continuous viscosity solution of* (3.1) *satisfying* (3.2).

Proof. Even though this is not quite a standard differential game due to the set valued map $F(z,q)$, and the fact that the ζ controls are not in a uniformly bounded set, the particular form of the problem allows us to apply the same argument as in [6], using dynamic programming, to prove that V is the unique viscosity solution of

$$\max\{V_t + \max_{q\in B(0,K)}\min_{z\le V}\min_{f\in F(z,q)} D_xV\cdot f, \max_q\min_{z,f} z - V\} = 0,$$

and $V(T,x) = \max_q \min_{z,f} g(x)\vee z$. Since $\min_z z = -\infty$ from (3.5) we get

$$V_t + H(V, D_xV) = 0 \quad \text{and} \quad V(T,x) = g(x).$$

See also [9] for similar details involving the optimal control problem and the Lax formula. □

THEOREM 3.2. *Assume* (A)–(D). *Then the value function* V *and the function* u *in* Theorem 3.1 *are identical.*

Proof. The function u in (3.3) is continuous on $[0,T]\times\mathbb{R}^n$ with $u(T,x) = \lim_{t\to T-0} u(t,x) = g(x)$. Furthermore, u is quasi-convex in both t and x.

We need to prove that $V(t,x) = u(t,x)$ for all $t\in[0,T]$ and $x\in\mathbb{R}^n$.

First we will show that $V\le u$. Let $t\in[0,T)$ and $x\in\mathbb{R}^n$. Fix γ such that

$$\sup_{p\in\mathbb{R}^n}\inf_{\{y:g(y)\le\gamma\}}(p\cdot(x-y)+(T-t)H(\gamma,p))\le 0. \tag{3.8}$$

We will show that $V\le\gamma$ and consequently $V\le u$.

From (3.8), we have for any control, $\gamma(\cdot)$,

$$\inf_{\{y:g(y)\le\gamma\}}(\eta(\tau)\cdot(x-y)+(T-t)H(\gamma,\eta(\tau)))\le 0,\qquad \tau\in[0,T). \tag{3.9}$$

Set $E_\gamma := \{y : g(y)\le\gamma\}$ and consider $\inf_{y\in E_\gamma}\eta(\tau)\cdot(x-y)$. Since E_γ is convex, the infimum of a linear function is either achieved (on the boundary) or $-\infty$. If the infimum is bounded from below we set $y = y(\tau)\in E_\gamma$ such that $\inf_{y\in E_\gamma}\eta(\tau)\cdot(x-y) = \eta(\tau)\cdot(x-y(\tau))$, and otherwise let $y(\tau)\in E_\gamma$ be any vector such that $\eta(\tau)\cdot(x-y(\tau))+(T-t)H(\gamma,\eta(\tau))\le 0$. Define $\varphi(\tau) = \frac{y(\tau)-x}{T-t}$. Finally, we define the strategy

$$\Delta[\eta](\tau) = (\zeta(\tau),\varphi(\tau)) \equiv (\gamma,\varphi(\tau)).$$

This strategy is obviously nonanticipating, but we must show it is admissible, i.e, we must show $H(\gamma,\eta(\tau))\le\eta(\tau)\cdot\varphi(\tau)$ for $\tau\in[0,T]$. By definition of φ we have from the definition of γ,

$$\eta(\tau)\cdot\varphi(\tau) = -\eta(\tau)\cdot\frac{x-y(\tau)}{T-t}\ge H(\gamma,\eta(\tau)),$$

and hence Δ is admissible. Then, letting $\xi(\cdot)$ denote the trajectory associated with Δ and η, we have $\xi(T) = x+\int_t^T\varphi(\tau)d\tau$, and

$$\begin{aligned}
V(t,x) &\le \sup_{\eta\in B(0,K)} g(\xi(T))\vee\gamma\\
&\le \sup_{\eta\in B(0,K)} g\left(x+\int_t^T\varphi(\tau)d\tau\right)\vee\gamma\\
&\le \sup_{\eta\in B(0,K)} g\left(\frac{1}{T-t}\int_t^T x+(T-t)\varphi(\tau)d\tau\right)\vee\gamma\\
&\le \sup_{\eta\in B(0,K)}\operatorname*{ess\,sup}_{t\le s\le T} g(x+(T-t)\varphi(s))\vee\gamma\\
&= \sup_{\eta\in B(0,K)}\operatorname*{ess\,sup}_{t\le s\le T} g(y(s))\vee\gamma\\
&= \gamma.
\end{aligned}$$

The last line follows from the fact that $g(y(s))\le\gamma$ for all $s\in[0,T]$. The next to last line follows from the extended Jensen inequality for quasi-convex functions [9],

which says that $g(\int h(s)d\mu(s)) \le \operatorname{ess\,sup}_s g(h(s))$, when g is quasi-convex and μ is a probability measure.

We conclude that $V(t,x) \le u(t,x)$.

To prove the opposite inequality we first claim that

$$V(t,x) \ge \max_{q\in B(0,K)} \min_{z} \min_{f\in F(z,q)} g(x+(T-t)f) \vee z. \tag{3.10}$$

To see that this is true, let $\varepsilon > 0$ and let $\Delta^* \in D$ be $\varepsilon-$near optimal. Then, letting $(\zeta^*, \varphi^*) = \Delta^*[\eta]$ and ξ^* the associated trajectory, we have from

$$V(t,x) = \inf_{\Delta\in D} \sup_{\eta\in B(0,K)} g(\xi(T)) \vee \operatorname*{ess\,sup}_{t\le s\le T} \zeta(s)$$

that

$$\begin{aligned} V(t,x)+\varepsilon &\ge \sup_{\eta\in B(0,K)} g(\xi^*(T)) \vee \operatorname*{ess\,sup}_{t\le s\le T} \zeta^*(s) \\ &\ge \max_{q\in B(0,K)} g(\xi(T)) \vee \operatorname*{ess\,sup}_{t\le s\le T} \zeta^*(s), \qquad \xi \text{ the outcome of } (\Delta[q],q) \\ &\ge \max_{q\in B(0,K)} \min_{z} \min_{f\in F(z,q)} g(x+(T-t)f)\vee z, \end{aligned}$$

and the claim follows. Observe that given the constant control $\eta(\tau) = q \in B(0,K)$, we have $d\xi/d\tau = \varphi(\tau)$ and $\varphi(\tau) \in F(\zeta(\tau), q)$. Using (3.10),

$$\begin{aligned} V(t,x) &= \inf_{\Delta\in D} \sup_{\eta\in B(0,K)} g(\xi(T)) \vee \operatorname*{ess\,sup}_{t\le s\le T} \zeta(s) \\ &\ge \max_{q\in B(0,K)} \min_{z} \min_{f\in F(z,q)} g(x+(T-t)f)\vee z \\ &\ge \max_{q\in B(0,K)} \min_{z} \min_{\{w: q\cdot \frac{w-x}{T-t} \ge H(z,q)\}} g(w)\vee z \\ &\ge \max_{q\in B(0,K)} \min_{z} \min_{\{w: q\cdot(x-w)+(T-t)H(z,q)\le 0\}} g(w)\vee z \\ &\equiv \alpha. \end{aligned}$$

Fix $q \in B(0,K)$ and $\varepsilon > 0$. Let z^* and w^* satisfy $\alpha+\varepsilon \ge g(w^*)\vee z^*$ and $q\cdot(x-w^*)+(T-t)H(z^*,q) \le 0$. Then $\alpha+\varepsilon \ge g(w^*)$ and $\alpha+\varepsilon \ge z^*$ imply that

$$\inf_{\{w: g(w)\le \alpha+\varepsilon\}} q\cdot(x-w)+(T-t)H(\alpha+\varepsilon,q) \le q\cdot(x-w^*)+(T-t)H(z^*,q) \le 0,$$

since $H(\gamma,p)$ is nonincreasing in γ. Since ε is arbitrary and the left side is continuous from the right in α (using the coercivity of g, [8]), by definition of $u(t,x)$ we conclude that $V(t,x) \ge \alpha \ge u(t,x)$. □

Remark 3.3. Differential games in L^∞ may also be used to gives a representation theorem for more general Hamilton–Jacobi equations, say, with t and x dependence in H, as in [16]. Once the solution of the equation is written as the value function of a differential game, one has the problem of evaluating the value function—not an easy task in general. Under the conditions assumed here, we are able to determine the value function as precisely the Hopf–Lax formula of [10]. We may view this as providing yet a new proof of the Hopf–Lax formula, but we note that the conditions assumed for the differential game are more stringent than necessary (see [11]).

4. Level sets. In the previous section it was natural to consider the terminal value problem because of the differential game formulation. In the rest of this paper we consider the forward problem on $(0,\infty)\times\mathbb{R}^n$.

The next theorem is the main result of this paper. It says that the solution of the problem $u_t + H(u,Du) = 0$ may be built by looking at solutions, say, w^γ of the problem with Hamiltonian $H^\gamma(Dw^\gamma) = H(\gamma, Dw^\gamma)$. It is remarkable that these problems have the same level sets and that the function u may be constructed from w^γ.

The theorem will apply to both hypotheses for the Hopf formula of [10] and the Lax formula of [9]. In fact, following the theorem we will see that the result holds under more general circumstances.

THEOREM 4.1. *Assume that $\gamma \mapsto H(\gamma,p)$ is nondecreasing in $\gamma \in R^1$ and $p \mapsto H(\gamma,p)$ is positively homogeneous degree one. Assume that g satisfies* (2.1).

For each $\gamma \in R^1$, let w^γ denote the viscosity solution of

$$w_t^\gamma(t,x) + H(\gamma, D_x w^\gamma(t,x)) = 0, w^\gamma(0,x) = g(x).$$

Let u denote the solution of

$$u_t(t,x) + H(u(t,x), D_x u(t,x)) = 0, \qquad u(0,x) = g(x).$$

Assume that either

(a) *g is quasi-convex and continuous and H is Lipschitz continuous; or*

(b) *g is continuous and $p \mapsto H(\gamma,p)$ is convex. Then,*

$$W^\gamma = \{(t,x) : w^\gamma(t,x) \le \gamma\} = \{(t,x) : u(t,x) \le \gamma\} = U^\gamma,$$

and hence w^γ and u have the same γ (sub)level sets. Furthermore,

$$u(t,x) = \inf\{\gamma : w^\gamma(t,x) \le \gamma\}. \tag{4.1}$$

Proof. Case (a). By the Hopf formula [10], [11] for quasi-convex data, we have

$$w^\gamma(t,x) = \inf\{\alpha : \sup_p p\cdot x - g^\#(\alpha,p) - tH(\gamma,p) \le 0\}, \tag{4.2}$$

and

$$u(t,x) = \inf\{\alpha : \sup_p p\cdot x - g^\#(\alpha,p) - tH(\alpha,p) \le 0\}. \tag{4.3}$$

Then, since $\alpha \mapsto \sup_p p\cdot x - g^\#(\alpha,p) - tH(\alpha,p)$ is nonincreasing,

$$\begin{aligned} W^\gamma &= \{(t,x) : w^\gamma(t,x) \le \gamma\} \\ &= \cup_{\alpha\le\gamma}\{(t,x) : \sup_p p\cdot x - g^\#(\alpha,p) - tH(\gamma,p) \le 0\} \\ &= \{(t,x) : \sup_p p\cdot x - g^\#(\gamma,p) - tH(\gamma,p) \le 0\} \\ &= \cup_{\alpha\le\gamma}\{(t,x) : \sup_p p\cdot x - g^\#(\alpha,p) - tH(\alpha,p) \le 0\} \\ &= \{(t,x) : u(t,x) \le \gamma\}. \end{aligned}$$

Observe that for fixed (t,x), the set $\{\gamma : w^\gamma(t,x) \le \gamma\}$ is nonempty. Indeed, if we fix $\gamma' \ge u(t,x)$, then by what we just proved $(t,x) \in U^{\gamma'} = W^{\gamma'}$, and thus $w^{\gamma'}(t,x) \le \gamma'$. So, fix (t,x) and let $\gamma \ge w^\gamma(t,x)$. Then, by (4.2), replacing α by γ, we get

$$\sup_p p \cdot x - g^{\#}(\gamma,p) - tH(\gamma,p) \le 0.$$

However, by (4.3) for u, γ is admissible and so $u(t,x) \le \gamma$. Since γ was arbitrary, $u(t,x) \le v(t,x) = \inf\{\gamma : w^\gamma(t,x) \le \gamma\}$.

Suppose that $u(t,x) \le v(t,x) - \epsilon$ for some $\epsilon > 0$. By definition of v as the smallest γ, we must have $w^{v-\epsilon} > v - \epsilon$ at (t,x). But $(t,x) \in U^{v-\epsilon} = W^{v-\epsilon}$, and this is a contradiction. Hence $u(t,x) = v(t,x)$.

Case (b). By the Lax formula [9] for convex H and continuous g, we have

$$w^\gamma(t,x) = \min_{y\in\mathbb{R}^n} g(x-ty) \vee H_\gamma^{\#}(y),$$

where

$$H_\gamma^{\#}(y) = \inf\{\alpha : \sup_p p \cdot y - H(\gamma,p) \le 0\},$$

and

$$u(t,x) = \min_{y\in\mathbb{R}^n} g(x-ty) \vee H^{\#}(y),$$

where

$$H^{\#}(y) = \inf\{\alpha : \sup_p p \cdot y - H(\alpha,p) \le 0\}.$$

However,

$$H_\gamma^{\#}(y) = \inf\{\alpha : \sup_p p \cdot y - H(\gamma,p) \le 0\} = \begin{cases} +\infty, & \text{if } \sup_p p \cdot y - H(\gamma,p) > 0, \\ -\infty, & \text{if } \sup_p p \cdot y - H(\gamma,p) \le 0. \end{cases}$$

Hence,

$$w^\gamma(t,x) = \min_{y\in\mathbb{R}^n} g(x-ty) \vee H_\gamma^{\#}(y) = \min_y \{g(x-ty) : \sup_p p \cdot y - H(\gamma,p) \le 0\}.$$

Let $(t_0,x_0) \in W^\gamma$. There is a y' such that $\sup_p p\cdot y' - H(\gamma,p) \le 0$ and $g(x_0 - t_0 y') \le \gamma$. However, $H^{\#}(y') \le \gamma$ as well, and so $g(x_0 - t_0 y') \vee H^{\#}(y') \le \gamma$. Consequently, $u(t_0,x_0) \le \gamma$ and so $(t_0,x_0) \in U^\gamma$.

Now let $(t_0,x_0) \in U^\gamma$. Then, there is a point $y' \in \mathbb{R}^n$ with $g(x_0 - t_0 y') \le \gamma$ and $H^{\#}(y') \le \gamma$. Since $H^{\#}(y') \le \gamma$, there is an $\alpha' \le \gamma$ so that $\sup_p p \cdot y' - H(\alpha', p) \le 0$. Since $\alpha \mapsto H(\alpha,p)$ is nondecreasing, we have $\sup_p p \cdot y' - H(\gamma,p) \le 0$. This implies that

$$\min_y \{g(x_0 - t_0 y) : \sup_p p \cdot y - H(\gamma,p) \le 0\} \le \gamma,$$

and therefore $(t_0,x_0) \in W^\gamma$. We have shown that $W^\gamma = U^\gamma$.

Since the proof that $u(t,x) = \inf\{\gamma : w^\gamma(t,x) \le \gamma\}$ is exactly the same as in case (a), we have completed the proof of the theorem. ∎

It was pointed out to us by the referee that the previous theorem will hold under much more general conditions than stated above. As long as a comparison principle holds for the equations, we may drop any convexity or quasi-convexity assumptions and retain only the homogenous degree one property of H. To be precise we state the following whose proof is due to the referee.

THEOREM 4.2. *Assume* (B) *and g continuous. Then the conclusions of* Theorem 4.1 *hold.*

Proof. Set $u^\gamma(t,x) = u(t,x) \vee \gamma$ and $\overline{w}^\gamma(t,x) = w^\gamma(t,x) \vee \gamma$. Using the same argument as in Proposition 3.1, by the fact that $H(\gamma,p)$ is nondecreasing in γ and homogeneous degree one in p, it is not hard to see that u^γ is a subsolution and $\overline{w}^\gamma$ is a solution of

$$\vartheta_t + H(\gamma, D_x\vartheta) = 0, \qquad \vartheta(0,x) = g(x) \vee \gamma.$$

By comparison (see [4], [12], [13]), we conclude that $u^\gamma(t,x) \le \overline{w}^\gamma(t,x)$ everywhere. Consequently, $\{\overline{w}^\gamma = \gamma\} \subset \{u^\gamma \le \gamma\}$. However,

$$\{w^\gamma \le \gamma\} = \{\overline{w}^\gamma = \gamma\} \subset \{u^\gamma \le \gamma\} = \{u \le \gamma\}.$$

For the opposite inclusion we now define $u^\gamma(t,x) = u(t,x) \wedge (\gamma + \varepsilon)$ and $\overline{w}^\gamma(t,x) = w^\gamma(t,x) \wedge (\gamma+\varepsilon)$, where $\varepsilon > 0$ is fixed. By the same argument, using the nondecreasing property in γ and homogeneous property in p of $H(\gamma,p)$, one shows that u^γ is a supersolution and $\overline{w}^\gamma$ is a solution of

$$\vartheta_t + H(\gamma, D_x\vartheta) = 0, \qquad \vartheta(0,x) = g(x) \wedge (\gamma + \varepsilon).$$

By comparison we may again conclude that $u^\gamma(t,x) \ge \overline{w}^\gamma(t,x)$ everywhere. Then, if $u(t,x) \le \gamma$ we have $u^\gamma(t,x) = u(t,x) \ge w^\gamma(t,x) \wedge (\gamma + \varepsilon)$, so $\gamma \ge u(t,x) \ge w^\gamma(t,x)$ and the opposite inclusion obtains. ∎

Observe that the comparison principle applied in the proof is for bounded functions, which is applicable by Proposition 3.1. Observe also that the proof of the theorem is much more succinct than Theorem 4.1 and in principle easier, but it was the explicit formulas of Hopf and Lax which led us to the conjecture.

Remark 4.1. Consider the equation (3.1) with $u(0,x) = g(x)$.

If $H(\gamma,p)$ is *concave* and positively homogeneous degree one in p, and nondecreasing in γ, we have for any continuous g,

$$u(t,x) = \max_{y\in\mathbb{R}^n} \left\{ g(y) \wedge H^{c\#}\left(\frac{x-y}{t}\right)\right\},$$

where

$$H^{c\#}(y) = \sup\{\alpha : \inf_{p\in\mathbb{R}^n} p\cdot y - H(\alpha,p) \ge 0\}.$$

In addition, $H(\gamma,p) = \inf\{p\cdot y : H^{c\#}(y) \ge \gamma\}$.

On the other hand, if g is continuous and *quasi-concave*, but now $H(\gamma,p)$ is only positively homogeneous degree one in p and nondecreasing in γ, then

$$u(t,x) = \sup\{\gamma : \inf_p p\cdot x - g^{c\#}(\gamma,p) - tH(\gamma,p) \ge 0\},$$

with $g^{c\#}(\gamma,p)=\inf\{p\cdot x : g(x)\geq\gamma\}$. In both cases, the analogous (super)level set result of the theorem remains true:

$$W^\gamma=\{(t,x):w^\gamma(t,x)\geq\gamma\}=\{(t,x):u(t,x)\geq\gamma\}=U^\gamma(t,x),\qquad\forall\gamma.$$

In addition, $u(t,x)=\sup\{\gamma : w^\gamma(t,x)\geq\gamma\}$.

5. The case $H(\gamma,p)=H_1(\gamma,p)+H_2(\gamma,p)$. The level set theorem is applied in this section to determine an upper and lower bound for equations with continuous data but Hamiltonians which are neither convex nor concave, but rather split into the sum of a convex and concave part. This extends the result of Bardi and Faggian [3] to Hamiltonians which depend on u, but we have the added assumption of homogeneity. Throughout the remainder of this paper we assume that initial conditions satisfy (2.1).

THEOREM 5.1. *Let $H(\gamma,p)=H_1(\gamma,p)+H_2(\gamma,p)$, where $H(\gamma,p)$ is assumed to be nondecreasing in γ, and $H_i(\gamma,p),\ i=1,2$ are assumed to be continuous and positively homogeneous degree one in $p\in R^n$. $H_1(\gamma,p)$ is convex in p and nondecreasing in γ. $H_2(\gamma,p)$ is concave in p and nondecreasing in γ.*

Then the viscosity solution of $u_t+H(u,Du)=0$ with $u(0,x)=g(x), x\in R^n$, and g assumed continuous, satisfies

$$\inf\left\{\gamma:\max_{\{z_2:H_2^{c\#}(z_2)\geq\gamma\}}\min_{y_1\in\mathbb{R}^n}\Big(g(x-t(y_1+z_2))\vee H_1^{\#}(y_1)\Big)\leq\gamma\right\}\leq u(t,x),$$

$$\sup\left\{\gamma:\min_{\{z_1:H_1^{\#}(y_1)\leq\gamma\}}\max_{y_2\in\mathbb{R}^n}\Big(g(x-t(z_1+y_2))\wedge H_2^{c\#}(y_2)\Big)\geq\gamma\right\}\geq u(t,x),$$

with $x\in\mathbb{R}^n,\ t\geq 0$.

Proof. We use the following representations for H_1 and H_2:

$$H_1(\gamma,p_1)=\sup_{\{z_1:H_1^{\#}(z_1)\leq\gamma\}}p_1\cdot z_1,\quad H_2(\gamma,p_2)=\inf_{\{z_2:H_2^{c\#}(z_2)\geq\gamma\}}p_2\cdot z_2,\ p_i\in\mathbb{R}^n,\ \ i=1,2,$$

where

$$H_1^{\#}(z_1)=\inf\{\gamma:\sup_{p_1\in\mathbb{R}^n}p_1\cdot z_1-H_1(\gamma,p_1)\leq 0\},$$
$$H_2^{c\#}(z_2)=\sup\{\gamma:\inf_{p_2\in\mathbb{R}^n}p_2\cdot z_2-H_2(\gamma,p_2)\geq 0\}.$$

The function $H_1^{\#}(z_1)$ is quasi-convex and $H_2^{c\#}(z_2)$ is quasi-concave (see [8], for example). Furthermore,

$$\inf_{z_1\in\mathbb{R}^n}H_1^{\#}(z_1)=-\infty\quad\text{and}\quad\lim_{|z_1|\to\infty}H_1^{\#}(z_1)=+\infty,\tag{5.1}$$

and

$$\sup_{z_2\in\mathbb{R}^n}H_2^{c\#}(z_2)=+\infty\quad\text{and}\quad\lim_{|z_2|\to\infty}H_2^{c\#}(z_1)=-\infty.\tag{5.2}$$

It follows that for any $\gamma\in R^1$,

$$E_{\gamma,H_2^{\#}}:=\{z\in\mathbb{R}^n:H_2^{c\#}(z_2)\geq\gamma\}\neq\emptyset.$$

Fix γ and $z_2 \in E_{\gamma,H_2^\#}$, noting that z_2 depends on γ. Then, writing $p = (p_1, p_2) \in R^{2n}$ with $p_1, p_2 \in \mathbb{R}^n$ we have

$$H(\gamma, p) = H_1(\gamma, p_1) + H_2(\gamma, p_2) \le H_1(\gamma, p_1) + p_2 \cdot z_2 =: \mathcal{H}(\gamma, p; z_2).$$

Now, $\mathcal{H}(\gamma, p; z_2)$ is nondecreasing in γ, convex in $p = (p_1, p_2)$, and positively homogeneous degree one in p. Hence, it is the conjugate of a quasi-convex function given as follows:

$$\begin{aligned}
\mathcal{H}^\#(x_1, x_2; z_2) &= \inf\{\gamma : \sup_{p_1,p_2} p_1 \cdot x_1 + p_2 \cdot x_2 - \mathcal{H}(\gamma, (p_1, p_2); z_2) \le 0\} \\
&= \inf\{\gamma : \sup_{p_1,p_2} p_1 \cdot x_1 + p_2 \cdot x_2 - H_1(\gamma, p_1) - p_2 \cdot z_2 \le 0\} \\
&= \inf\{\gamma : \sup_{p_1} p_1 \cdot x_1 - H_1(\gamma, p_1) + \sup_{p_2} p_2 \cdot (x_2 - z_2) \le 0\} \\
&= \begin{cases} +\infty, & \text{if } x_2 \ne z_2, \\ H_1^\#(x_1), & \text{if } x_2 = z_2. \end{cases}
\end{aligned}$$

Now, we shall use the doubled variables argument in [3]. We consider the function $U^\gamma : [0, T] \times \mathbb{R}^n \times \mathbb{R}^n \to R^1$ which is the solution of the problem (recall that γ is fixed)

$$U_t^\gamma + H_1(\gamma, D_{x_1} U^\gamma) + H_2(\gamma, D_{x_2} U^\gamma) = 0, \quad (t, x_1, x_2) \in (0, \infty) \times R^{2n}, \tag{5.3}$$

$$U(0, x_1, x_2) = g(x_1 + x_2), \qquad x_1, x_2 \in \mathbb{R}^n. \tag{5.4}$$

Then, as in [3], $w^\gamma(t, x)$ is a viscosity solution of

$$w_t^\gamma(t, x) + H(\gamma, D_x w^\gamma(t, x)) = 0, w^\gamma(0, x) = g(x), \quad x \in \mathbb{R}^n$$

if and only if $U^\gamma(t, x_1, x_2) = w^\gamma(t, x_1 + x_2)$ is the solution of (5.3).

With $z_2 \in E_{\gamma,H_2^\#} = \{z : H_2^{c\#}(z) \ge \gamma\}$ still fixed, consider next the Cauchy problem

$$\varphi_t(t, x) + \mathcal{H}(\gamma, D_x \varphi; z_2) = 0, \qquad (t, x) \in (0, \infty) \times R^{2n}, \qquad \varphi(0, x) = g(x_1 + x_2),$$

where $x = (x_1, x_2)$. This problem has the explicit solution given by the Hopf–Lax formula in [9]:

$$\begin{aligned}
\varphi^\gamma(t, x; z_2) &= \min_{y \in R^{2n}} \left(g(y_1 + y_2) \vee \mathcal{H}^\# \left(\frac{x - y}{t}; z_2 \right) \right) \\
&= \min_{y \in R^{2n}} \left(g(x_1 - t y_1 + x_2 - t y_2) \vee \mathcal{H}^\#(y; z_2) \right).
\end{aligned}$$

Using the calculation for $\mathcal{H}^\#$ and $y = (y_1, y_2)$ we get

$$\varphi^\gamma(t, (x_1, x_2); z_2) = \min_{y_1 \in R^n} \left(g(x_1 + x_2 - t(y_1 + z_2)) \vee H_1^\#(y_1) \right). \tag{5.5}$$

Since $H(\gamma, p) \le \mathcal{H}(\gamma, p; z_2)$ we have by comparison of viscosity solutions (see, for example, Barles [4]) that

$$\varphi^\gamma(t, x; z_2) \le U^\gamma(t, x_1, x_2), \qquad \forall z_2 \in E_{\gamma,H_2^{c\#}}.$$

Using (5.5) we see that

$$\max_{z_2 \in E_{\gamma, H_2^{c\#}}} \min_{y_1 \in \mathbb{R}^n} (g(x_1 + x_2 - t(y_1 + z_2)) \vee H_1^*(y_1)) \le U^\gamma(t, x_1, x_2).$$

By Theorem 4.1, we have that $U(t, x_1, x_2) := \inf\{\gamma : U^\gamma(t, x_1, x_2) \le \gamma\}$ is the solution of

$$U_t + H_1(U, D_{x_1}U) + H_2(U, D_{x_2}U) = 0, \quad (t, x) \in (0, \infty) \times R^{2n},$$

and initial condition $U(0, x_1, x_2) = g(x_1 + x_2)$. Given $x \in \mathbb{R}^n$ write it as $x = x_1 + x_2$ for some $x_1, x_2 \in \mathbb{R}^n$. Then the solution of our original problem

$$u_t + H_1(u, D_x u) + H_2(u, D_x u) = 0, \quad (t, x) \in (0, \infty) \times R^n, \quad u(0, x) = g(x)$$

is given by $u(t, x) = U(t, x_1, x_2)$. Combining these facts results in

$$\inf \left\{ \gamma : \max_{\{z_2 : H_2^{c\#}(z_2) \ge \gamma\}} \min_{y_1 \in \mathbb{R}^n} \Big(g(x - t(y_1 + z_2)) \vee H_1^{\#}(y_1) \Big) \le \gamma \right\} \le u(t, x).$$

For the upper bound inequality in the theorem, we use the quasi-concave conjugates. Thus, in a similar manner we prove that

$$\sup \left\{ \gamma : \min_{\{z_1 : H_1^{\#}(y_1) \le \gamma\}} \max_{y_2 \in \mathbb{R}^n} \Big(g(x - t(z_1 + y_2)) \wedge H_2^{c\#}(y_2) \Big) \ge \gamma \right\} \ge u(t, x). \qquad \square$$

6. The case $g(x) = g_1(x) \vee g_2(x)$ or $g(x) = g_1(x) \wedge g_2(x)$. We first consider the case when the data splits into the maximum of a quasi-convex function g_1 and a quasi-concave function g_2. Thus, we consider the Hamilton–Jacobi equation (3.1) with the initial data

$$u(0, x) = g_1(x) \vee g_2(x).$$

We assume that $H(\gamma, p)$ is nondecreasing in γ and homogeneous degree one in p. Again, we go through the problem with doubled variables

$$U_t + H(U, D_{x_1}U + D_{x_2}U) = 0, \qquad U(0, x_1, x_2) = g_1(x_1) \vee g_2(x_2).$$

Then (with the same proof as in [3]) $u(t, x) = U(t, x, x)$. Label the solution of the doubled problem $U_{g_1 \vee g_2}$.

Now, since g_2 is assumed to be quasi-concave, it is the infimum of piecewise linear functions lying above it. Precisely, using the alternate version of the quasi-concave conjugates (see section 2) from [11], [8], we have

$$g_2(x_2) = \inf_{p_2, \gamma_2} (p_2 \cdot x_2 + g_2^{c*}(\gamma_2, p_2)) \vee \gamma_2,$$

where

$$g_2^{c*}(\gamma_2, p_2) = \inf_{\{x_2 : g_2(x_2) \ge \gamma_2\}} p_2 \cdot x_2 + g_2(x_2)$$

is the first quasi-concave conjugate of g_2 (see [11]).

Fix $\overline{\gamma}_2 \in R^1$ and $\overline{p}_2 \in \mathbb{R}^n$ with $(\overline{\gamma}_2, \overline{p}_2) \in dom(g_2^{c*})$ and set

$$h(x_1, x_2; \overline{\gamma}_2, \overline{p}_2) = g_1(x_1) \vee (\overline{p}_2 \cdot x_2 + g_2^{c*}(\overline{\gamma}_2, \overline{p}_2)) \vee \overline{\gamma}_2.$$

Consider the following problem for U_h with initial data h:

$$\begin{aligned} &U_t + H(U, D_{x_1}U + D_{x_2}U) = 0, \\ &U(0, x_1, x_2) = h(x_1, x_2; \overline{\gamma}_2, \overline{p}_2). \end{aligned}$$

Since $(p_2 \cdot x_2 + g_2^{c*}(\gamma_2, p_2)) \vee \gamma_2$ is quasi-affine, i.e., both quasi-convex and quasi-concave, and the maximum of two quasi-convex functions is quasi-convex, we now have initial data h which is quasi-convex. Thus, we may apply the Hopf–Lax formula in [10] to obtain the solution

$$U_h(t, x_1, x_2) = \inf\{\gamma : \sup_{p_1, p_2} p_1 \cdot x_1 + p_2 \cdot x_2 - h^{\#}(\gamma, p_1, p_2; \overline{\gamma}_2, \overline{p}_2) - tH(\gamma, p_1 + p_2) \le 0\}.$$

We must calculate $h^{\#}(\gamma, p_1, p_2; \overline{\gamma}_2, \overline{p}_2)$. By definition, we have, for $\overline{\gamma}_2 \le \gamma$,

$$\begin{aligned} h^{\#}(\gamma, p_1, p_2; \overline{\gamma}_2, \overline{p}_2) &= \sup_{\{x_1, x_2 : h(x_1, x_2; \overline{\gamma}_2, \overline{p}_2) \le \gamma\}} p_1 \cdot x_1 + p_2 \cdot x_2 \\ &= \sup_{\{x_1 : g_1(x_1) \le \gamma\}} p_1 \cdot x_1 + \sup_{\{x_2 : \overline{p}_2 \cdot x_2 + g_2^{c*}(\overline{\gamma}_2, \overline{p}_2) \le \gamma\}} p_2 \cdot x_2 \\ &= g_1^{\#}(\gamma, p_1) + \sup_{\{x_2 : \overline{p}_2 \cdot x_2 + g_2^{c*}(\overline{\gamma}_2, \overline{p}_2) \le \gamma\}} p_2 \cdot x_2. \end{aligned}$$

If $\overline{\gamma}_2 > \gamma$, $h^{\#}(\gamma, p_1, p_2; \overline{\gamma}_2, \overline{p}_2) = -\infty$.

Now set $S = \{x_2 : \overline{p}_2 \cdot x_2 + g_2^{c*}(\overline{\gamma}_2, \overline{p}_2) \le \gamma\}$. This is a closed convex set, in fact a half space, and we seek the support function of S, which is the supremum on the preceding line.

If we consider the half space $C = \{x : \alpha \cdot x + \beta \le 0\}$, the Legendre–Fenchel conjugate of the affine function $f(x) = \alpha \cdot x + \beta$ is

$$f^*(p) = \begin{cases} +\infty, & \text{if } p \ne \alpha, \\ -\beta, & \text{if } p = \alpha. \end{cases}$$

From Theorem 13.5, page 118 in Rockafellar [22], the support function $\sigma_C(p)$ of C is then given by

$$\begin{aligned} \sigma_C(p) &= \inf_{\lambda > 0} \lambda f^*\left(\frac{p}{\lambda}\right) \\ &= \begin{cases} +\infty, & \text{if } p \ne \lambda\alpha, \quad \forall \lambda > 0, \\ -\beta\lambda, & \text{if } p = \lambda\alpha, \quad \exists \lambda > 0. \end{cases} \end{aligned}$$

Applying this result to S we get

$$\sigma_S(p_2) = \sup_{\{x_2 : \overline{p}_2 \cdot x_2 + g_2^{c*}(\overline{\gamma}_2, \overline{p}_2) \le \gamma\}} p_2 \cdot x_2 = \begin{cases} +\infty, & \text{if } p_2 \ne \lambda\overline{p}_2, \quad \forall \lambda > 0, \\ (\gamma - g_2^{\#}(\overline{\gamma}_2, \overline{p}_2))\lambda, & \text{if } p_2 = \lambda\overline{p}_2, \quad \exists \lambda > 0. \end{cases}$$

Summarizing, we have

$$h^{\#}(\gamma,p_1,p_2;\overline{\gamma}_2,\overline{p}_2)=\begin{cases}-\infty, & \text{if } \overline{\gamma}_2>\gamma\\ +\infty, & \text{if } p_2\neq\lambda\overline{p}_2, \quad \forall\lambda>0,\\ g_1^{\#}(\gamma,p_1)+\lambda(\gamma-g_2^{c*}(\overline{\gamma}_2,\overline{p}_2)), & \text{if } p_2=\lambda\overline{p}_2, \quad \exists\lambda>0.\end{cases}$$

Putting this into the formula for U_h, we get

$$\begin{aligned}U_h(t,x_1,x_2)=\inf\{\gamma\geq\overline{\gamma}_2:\ &\sup_{p_1\in\mathbb{R}^n,\lambda>0} p_1\cdot x_1+\lambda\overline{p}_2\cdot x_2-g_1^{\#}(\gamma,p_1)\\ &-\lambda(\gamma-g_2^{c*}(\overline{\gamma}_2,\overline{p}_2))-tH(\gamma,p_1+\lambda\overline{p}_2)\leq 0\}.\end{aligned}\tag{6.1}$$

Notice that the expression in the supremum is homogeneous degree one in (p_1,p_2) together, but not individually.

Since

$$U_h(0,x_1,x_2)=h(x_1,x_2;\overline{\gamma}_2,\overline{p}_2)\geq g(x_1)\vee g(x_2)=U_{g_1\vee g_2}(0,x_1,x_2),$$

comparison for (3.1) gives us $U_h(t,x_1,x_2)\geq U_{g_1\vee g_2}(t,x_1,x_2)$ for all $t\geq 0$, $x_1,x_2\in\mathbb{R}^n$. Using (6.1) and the fact that $(\overline{\gamma}_2,\overline{p}_2)$ in $dom(g_2^{c*})$ was arbitrary, results in

$$\begin{aligned}\inf_{\overline{\gamma}_2,\overline{p}_2\in dom(g_2^{c*})}\inf\{\gamma\geq\overline{\gamma}_2:\ &\sup_{p_1,\lambda>0} p_1\cdot x_1+\overline{p}_2\cdot x_2-g_1^{\#}(\gamma,p_1)\\ &-\lambda(\gamma-g_2^{c*}(\overline{\gamma}_2,\overline{p}_2))-tH(\gamma,p_1+\lambda\overline{p}_2)\leq 0\}\geq U_{g_1\vee g_2}(t,x_1,x_2).\end{aligned}$$

Finally, since $u(t,x)=U_{g_1\vee g_2}(t,x,x)$, we conclude

$$\begin{aligned}u(t,x)\leq\inf_{\overline{\gamma}_2,\overline{p}_2\in dom(g_2^{c*})}\inf\{\gamma\geq\overline{\gamma}_2:\ &\sup_{p_1,\lambda>0}(p_1+\lambda\overline{p}_2)\cdot x-g_1^{\#}(\gamma,p_1)\\ &-\lambda(\gamma-g_2^{c*}(\overline{\gamma}_2,\overline{p}_2))-tH(\gamma,p_1+\lambda\overline{p}_2)\leq 0\}.\end{aligned}\tag{6.2}$$

Similarly, if u satisfies the initial data $u(0,x)=g_1(x)\wedge g_2(x)$, we prove that

$$\begin{aligned}u(t,x)\geq\sup_{\overline{\gamma}_1,\overline{p}_1\in dom(g_1^{*})}\sup\{\gamma\leq\overline{\gamma}_1:\ &\inf_{p_2,\lambda>0}(\lambda\overline{p}_1+p_2)\cdot x\\ &-\lambda(\gamma+g_1^{*}(\overline{\gamma}_1,\overline{p}_1))+g_2^{c\#}(\gamma,p_2)-tH(\gamma,\lambda\overline{p}_1+p_2)\geq 0\}.\end{aligned}\tag{6.3}$$

We have proved the theorem.

THEOREM 6.1. *Let g_1 and g_2 be continuous functions with g_1 quasi-convex and g_2 quasi-concave. Let $H(\gamma,p)$ satisfy (B)*(i)–(ii)*, and $\gamma\mapsto H(\gamma,p)$ nondecreasing for any fixed $p\in\mathbb{R}^n$. Then the solution u of* (3.1) *with initial data $u(0,x)=g_1(x)\vee g_2(x)$ satisfies* (6.2)*, and with initial data $u(0,x)=g_1(x)\wedge g_2(x)$, satisfies* (6.3)*.*

Remark 6.1. It is very interesting that the upper and lower bounds in this theorem use both versions of the quasi-convex or quasi-concave conjugates, i.e., # and $*$ (see section 2), and this seems unavoidable. Of course, if we needed two ## or two $**$, this remark would not be true since they both lead to the greatest quasi-convex minorant. But the first conjugates are not the same.

Remark 6.2. If we have initial data $u(0,x)=g(x)$ which may be represented as $g_1\vee g_2$ and as $h_1\wedge h_2$ for some quasi-convex functions g_1,h_1 and quasi-concave function g_2,h_2, then we obtain an upper bound for u using the g's and a lower bound using

the h's. This raises the question as to what kind of functions may be represented as the max or min of quasi-convex, quasi-concave functions. More will be said in the following remark.

Remark 6.3. If we have initial data given by $u(0,x) = g_1(x) + g_2(x)$ where g_1 is convex and g_2 is concave, as in [3], a natural way to obtain an upper bound is to replace the initial data by $2(g_1 \vee g_2)$ and $2(g_1 \wedge g_2)$ for a lower bound. Using the results of this paper we may carry out this plan not just for convex and concave g's but also for quasi-convex, quasi-concave g's.

The class of functions representable as the sum of convex and concave is very wide. Indeed, any function with bounded second derivatives may be so represented. To see this, let K denote the bound on the second derivatives of a smooth function g and consider $g_2(x) = g(x) - K|x|^2$ and $g_1(x) = K|x|^2$. Then, g_2 is concave, g_1 is convex, and $g = g_1 + g_2$.

Acknowledgments. We would like to thank Professor Bardi for the preprint [3] which motivated this paper. We also thank the anonymous referee who provided a nice proof of Theorem 4.2 and who pointed out that Theorem 4.1 is true under under more general hypotheses.

REFERENCES

[1] M. Bardi and L. C. Evans, *On Hopf formula for solutions of Hamilton–Jacobi equations*, Nonlinear Anal., 8 (1984), pp. 1373–1381.
[2] M. Bardi and S. Osher, *The nonconvex multi-dimensional Riemann problem for Hamilton–Jacobi equations*, SIAM J. Math. Anal., 22 (1991), pp. 344–351.
[3] M. Bardi and S. Faggian, *Hopf-type formulas for non-convex non-concave Hamilton–Jacobi equations*, SIAM J. Math. Anal., to appear.
[4] G. Barles, *Solutions de Viscosite des Equations de Hamilton–Jacobi*, Mathematiques and Applications 17, Springer-Verlag, New York, 1994.
[5] E. N. Barron, *Optimal control and calculus of variations in L^∞*, in Optimal Control of Differential Equations, N. H. Pavel, ed., Marcel Dekker, New York, 1994, pp. 39–47.
[6] E. N. Barron, *Differential games with maximum cost*, Nonlinear Anal., 14 (1990), pp. 971–989.
[7] E. N. Barron and H. Ishii, *The Bellman equation for minimizing the maximum cost*, Nonlinear Anal., 13 (1989), pp. 1067–1090.
[8] E. N. Barron and W. Liu, *Calculus of variations in L^∞*, Appl. Math. Optim., 35 (1997), pp. 237–263.
[9] E. N. Barron, R. Jensen, and W. Liu, *Hopf–Lax formula for $u_t+H(u,Du) = 0$*, J. Differential Equations, 126 (1996), pp. 48–61.
[10] E. N. Barron, R. Jensen, and W. Liu, *Hopf–Lax formula for $u_t + H(u,Du) = 0$: II*, Comm. PDE's, 22 (1997), pp. 1141–1160.
[11] E. N. Barron, R. Jensen, and W. Liu, *Explicit solutions for some first order PDEs*, J. Dynam. Control Systems, 3 (1997), pp. 149–164.
[12] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[13] M. G. Crandall, L. C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[14] J.-P. Crouzeix, *Continuity and differentiability properties of quasiconvex functions on $\mathbb{R}^n$*, in Generalized Concavity in Optimization and Economics, S. Schaible and W. T. Ziemba, eds., Academic Press, New York, 1981, pp. 109–130.
[15] J.-P. Crouzeix, *A duality framework in quasi-convex programming*, in Generalized Concavity in Optimization and Economics, S. Schaible and W. T. Ziemba, eds., Academic Press, New York, 1981, pp. 207–226.
[16] L. C. Evans and P. Souganidis, *Differential games and representation formulas for solutions of Hamilton Jacobi Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.
[17] E. Hopf, *Generalized solutions of nonlinear equations of first order*, J. Math. Mech., 14 (1965), pp. 951–973.

[18] J. E. Martinez-Legaz, *On lower subdifferentiable functions*, in Trends in Mathematical Optimization, K. H. Hoffman, J. B. Hiriart-Urruty, C. Lemarchal, and J. Zowe, eds., International Series Numer. Math. 84, Birkhauser, Boston, MA, 1988, pp. 197–232.
[19] J. E. Martinez-Legaz, *quasi-convex duality theory by generalized conjugation methods*, Optimization, 19 (1988), pp. 603–652.
[20] J. E. Martinez-Legaz and S. Romano-Rodriguez, *α-lower subdifferentiable functions*, SIAM J. Optim., 3 (1993), pp. 800–825.
[21] J.-P. Penot and M. Volle, *On quasi-convex duality*, Math. Oper. Res., 15 (1990), pp. 597–625.
[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[23] D. B. Silin, *Generalizing Hopf and Lax-Oleinik formulae via conjugate integral*, preprint.
[24] D. B. Silin, *Set valued integration and viscosity solutions to Hamilton Jacobi equations*, Differential Equations, 31 (1995), pp. 129–137 (in Russian).
[25] A. I. Subbotin, *Generalized Solutions of First Order PDEs*, Birkhauser, Boston, MA, 1995.
[26] A. I. Subbotin, *Existence and uniqueness results for Hamilton Jacobi equations*, Nonlinear Anal., 16 (1991), pp. 683–699.

SIAM J. MATH. ANAL.
Vol. 29, No. 4, pp. 1040–1065, July 1998

012

ORTHONORMAL WAVELET BASES ADAPTED FOR PARTIAL DIFFERENTIAL EQUATIONS WITH BOUNDARY CONDITIONS*

PASCAL MONASSE† AND VALÉRIE PERRIER†

Abstract. We adapt ideas presented by Auscher to impose boundary conditions on the construction of multiresolution analyses on the interval, as introduced by Cohen, Daubechies, and Vial. We construct new orthonormal wavelet bases on the interval satisfying homogeneous boundary conditions. This construction can be extended to wavelet packets in the case of one boundary condition at each edge. We present in detail the numerical computation of the filters and the derivative operators associated with these bases. We derive quadrature formulae in order to study the approximation error at the edge of the interval. Several examples illustrate the present construction.

Key words. wavelet, multiresolution analysis, boundary conditions

AMS subject classifications. 35C10, 42C15, 46E35

PII. S0036141095295127

1. Introduction.

Our ultimate goal is to solve numerically partial differential equations, for example, linear elliptic equations of the type

$$\left\{ \begin{array}{rcl} -\Delta u + \lambda u & = & f \text{ in } \Omega, \\ u & = & 0 \text{ on } \partial\Omega \end{array} \right. \tag{1}$$

where $\Omega = [0,1]^d$, $f \in L^2(\Omega)$, and $\lambda \geq 0$. The boundary conditions being taken by $u \in H_0^1(\Omega)$, the variational form of equation (1) is

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } \forall v \in H_0^1(\Omega), \int_\Omega \nabla u \nabla v \, d\tau + \lambda \int_\Omega uv \, d\tau = \int_\Omega fv \, d\tau. \tag{2}$$

Riesz theorem proves the existence and the uniqueness of u satisfying (2). We want to find a function close to u according to the $H_0^1(\Omega)$ norm by using a Galerkin method. Thus, we have to construct finite dimensional subspaces V_j of $H_0^1(\Omega)$ such that $\forall j, V_j \subset V_{j+1}$ and $\overline{\bigcup V_j} = H_0^1(\Omega)$. Then an approximate solution u_j is given by solving a finite dimensional problem:

$$\text{Find } u_j \in V_j \text{ such that } \forall v \in V_j, \int_\Omega \nabla u_j \nabla v \, d\tau + \lambda \int_\Omega u_j v \, d\tau = \int_\Omega fv \, d\tau. \tag{3}$$

By introducing $\{v_{j,k}\}_k$, an orthonormal basis of V_j as test functions v, (3) is reduced to a linear system. Such embedded spaces V_j can be obtained from a multiresolution analysis (MRA). Test functions in (3) can be chosen among the *scaling functions* of the MRA, as in [BNR 94]; such a method then has convergence properties similar to those of spectral methods, the precision being limited by the regularity of the MRA. In order to derive adaptive schemes based on nonlinear approximation of the exact solution u (see [DVJP 92]), we prefer to consider as test functions v the *wavelet*

* Received by the editors November 22, 1995; accepted for publication (in revised form) July 22, 1997; published electronically March 25, 1998.
http://www.siam.org/journals/sima/29-4/29512.html

† Laboratoire de Météorologie Dynamique, E.N.S., 24 rue Lhomond, 75231 Paris cedex 05, France. Permanent address of second author: Laboratoire d'Analyse, Géométrie et Applications, URA 742 Institut Galilée, Université Paris Nord, Av J.B. Clément, 93430 Villetaneuse, France (perrier@math.univ-paris13.fr).

basis; this will reduce significantly the number of degrees of freedom, owing to the compression properties of the wavelet transform. These compression properties have been practically observed in numerical experiments computing the wavelet solution of linear PDEs with periodic boundary conditions ([MPR 91], [LPT 92], [ChPe 96]). Jaffard has already proven that, by using a diagonal preconditioner, the condition number of the linear system deduced from (3) is independent of the mesh size in the *wavelet* basis, which leads to a fast resolution of such a system (see [Jaff 92]). Our present objective is then to construct an MRA on the interval $[0,1]$ and the associated orthonormal wavelet basis satisfying the boundary conditions in (1) and finally to derive the expressions for the Galerkin derivative operators.

We want to construct an MRA, of $H_0^1([0,1])$. More generally, we want to characterize subspaces of $H^s(\Omega)$ defined by vanishing values of some derivatives at the boundaries 0 and 1. Even if we do not impose conditions at the boundaries, the construction of wavelets on the interval is not trivial. The simplest solution is to use periodic wavelets (defined in [Meye 90] and implemented in [PeBa 89]) adapted to nonperiodic conditions by the Chebyshev transform $x = 1/\pi \arccos y$, which leads to "Chebyshev" wavelets (see [PST 95]) and allows to use spectral schemes based on Tau methods [MaRa 92]. In 1989, Jaffard and Meyer proposed a construction of wavelet bases on open sets of $\mathbb{R}^n$ starting from spline functions [JM 89]. However, their construction was theoretical and has not yet been implemented numerically. More recently, other constructions of wavelet bases on the interval $[0,1]$ have been proposed; they are all based on Daubechies compactly supported wavelets in $L^2(\mathbb{R})$ [Daub 88]. The first construction was done by Meyer (see [Meye 92]). It was rather theoretical and had some drawbacks, the most important being its numerical instability. Another construction avoiding these problems was then proposed independently by [CDV 93] and [AHJP 93], and extended in dimension 2 in [CDDe 95]. Nevertheless, these wavelets take arbitrary values at the boundaries and are not well adapted to the resolution of boundary value problems. This problem was theoretically solved by Auscher (see [Ausc 93]) who adapted Meyer's construction when boundary conditions are imposed.

Our goal is to construct the wavelets of [CDV 93] in a practical way using Auscher's ideas. The construction presented here was already roughly described in [MoPe 95]. Since the submission of the present paper, an alternative construction was independently developed by Chiavassa and Liandrat [ChLi 97]. First, we want to construct embedded spaces V_j as introduced in (3). For that, we will construct MRAs on the interval $[0,1]$. Section 2 recalls the basic properties of MRAs on $\mathbb{R}$, then we describe the construction on $[0,+\infty[$ in section 3, on $[0,1]$ in section 4, and the computation of the numerical filters in section 5. The expressions of the first two derivative operators are computed in section 6, an interpolation procedure is derived in section 7, and finally we give some numerical results in section 8.

2. Orthonormal wavelet bases on $\mathbb{R}$. We briefly review wavelets and MRA of $L^2(\mathbb{R})$ (for further details, see [Daub 92, Mall 89, Meye 90]). An MRA is a set $(V_j)_{j\in\mathbb{Z}}$ of closed subspaces of $L^2(\mathbb{R})$ satisfying:

1. $\{0\} = \bigcap_{j\in\mathbb{Z}} V_j \subset \cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots \subset \overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R})$;
2. $f(x) \in V_0 \Longleftrightarrow f(2^j x) \in V_j$;
3. $\exists g \in V_0$ such that $g(.-k)_{k\in\mathbb{Z}}$ is a Riesz basis of V_0.

From g, it is possible to construct a function Φ (called the scaling function) of V_0 such that $\{\Phi(.-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_0 and $\int_{-\infty}^{+\infty} \Phi(x)dx = 1$. Moreover,

because $V_0 \subset V_1$ and from point 2, there exist reals h_k such that

$$\Phi(x) = \sqrt{2} \sum_{k=-\infty}^{+\infty} h_k \Phi(2x-k). \tag{4}$$

Changing the scale, it follows that $\{2^{j/2}\Phi(2^j.-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_j. Defining W_j as the orthogonal complement of V_j in V_{j+1}:

$$W_j = V_{j+1} \ominus V_j,$$

it is easy to verify that $L^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j$. Then it is possible to construct a function Ψ (called wavelet) of V_1 such that $\{\Psi(.-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of W_0. Then $\{2^{j/2}\Psi(2^j.-k)\}_{j,k\in\mathbb{Z}}$ is an orthonormal basis of $L^2(\mathbb{R})$ and Ψ satisfies a two-scale equation:

$$\Psi(x) = \sqrt{2} \sum_{k=-\infty}^{+\infty} g_k \Phi(2x-k). \tag{5}$$

The number N of vanishing moments of the wavelet Ψ plays an important role in the approximation and in the compression of functions:

$$\int_{-\infty}^{+\infty} x^k \Psi(x)dx = 0 \quad \text{for } k = 0,\dots,N-1. \tag{6}$$

From (4) and (5) it follows that

$$\sqrt{2}\ \Phi(2x-n) = \sum_k \left(h_{n-2k}\Phi(x-k) + g_{n-2k}\Psi(x-k)\right). \tag{7}$$

Equations (4), (5), and (7) allow us to describe a fast algorithm to analyze and synthesize a given function in the wavelet basis (see [Mall 89]); it consists of a tree algorithm, each step of which is a filtering with the discrete filters h_k and g_k. In the case where only a finite number of these coefficients are nonzero, the wavelet and the scaling function are compactly supported. Daubechies was the first to construct such wavelets (see [Daub 88]) for every (finite) number of zero moments. For N vanishing moments, Φ and Ψ are supported on $[-N+1, N]$ (or whatever interval of integer boundaries and length $2N-1$) and the nonzero h_k and g_k are $(h_{-N+1},\dots,h_N)$ and $(g_{-N+1},\dots,g_N)$. Such wavelets have the minimal-length support between all the possible wavelets with N vanishing moments. Furthermore, the regularity of these functions increases asymptotically linearly with N:

$$\Phi, \Psi \in C^{\mu N} \tag{8}$$

for N sufficiently large, $\mu \simeq 0.2$.

3. Orthonormal wavelet bases on [0,+∞[. We will focus on the construction of an MRA on $[0,+\infty[$. Then, by a simple trick, it will be easy to construct an MRA on $[0,1]$.

3.1. MRA on [0,+∞[without boundary conditions. We follow first the construction of [CDV 93] by applying Auscher's ideas [Ausc 93]. We establish a few well-known properties based on Daubechies compactly supported wavelets, in order to define scaling functions at the edge 0.

3.1.1. Expression of monomials in V_0. We start from a usual MRA of $L^2(\mathbb{R})$ given by spaces V_j and a scaling function Φ. Although the monomials do not belong in the usual sense to V_0, we can write

$$\text{For } \ell = 0, \dots, N-1, \qquad \frac{x^\ell}{\ell!} = \sum_{k=-\infty}^{+\infty} P_\ell(k)\ \Phi(x-k) \tag{9}$$

with P_ℓ the polynomials defined by

$$P_\ell(X) = \sum_{n=0}^{\ell} \frac{C_{\ell-n}}{n!} X^n, \tag{10}$$

where

$$C_m = \int_{-\infty}^{+\infty} \frac{x^m}{m!} \Phi(x)dx = P_m(0). \tag{11}$$

The equality in (9) should be understood as a pointwise convergence on $\mathbb{R}$ or as a uniform or L^2 convergence on any compact set of $\mathbb{R}$. In fact, if we restrict to a compact set, the convergence is an equality since Φ is compactly supported, so that in the right-hand side of (9) only a finite number of terms participate in the sum. The proof of the existence of polynomial P_ℓ in (9) can be found, for example, in [CDV 93]. Remark that P_ℓ is a polynomial of degree ℓ and that $P_0(X) = 1$.

The coefficients C_ℓ can be computed recursively using (see [CDV 93]):

$$\begin{cases} C_0 &= 1 \\ C_\ell &= \dfrac{1}{2^\ell - 1} \displaystyle\sum_{r=1}^{\ell} M_r C_{\ell-r} \end{cases}$$

with

$$M_r = \frac{1}{\sqrt{2}} \sum_m h_m \frac{m^r}{r!}.$$

As in the case of the MRA on $\mathbb{R}$, we want the polynomials up to degree $N-1$ to remain in our new space $V_0^{[0,+\infty[}$. For that, following [Ausc 93], we define the edge scaling functions.

DEFINITION 3.1. *For $\ell = 0, \dots, N-1$, the edge scaling functions are defined by*

$$\tilde{\Phi}_\ell(x) = \sum_{k=-N+1}^{N-1-\alpha} P_\ell(k)\ \Phi(x-k)\ \chi_{[0,+\infty[}(x), \tag{12}$$

where α is a fixed parameter whose value is 0 or 1.

The interest of the parameter α will be clear later. It is linked to the (finite) dimension of the MRA spaces of $L^2([0,1])$. The functions $\tilde{\Phi}_\ell$ are such that for all x in $[0, +\infty[$

$$\frac{x^\ell}{\ell!} = \tilde{\Phi}_\ell(x) + P_\ell(N-\alpha)\ \Phi(x-(N-\alpha)) + \cdots.$$

Remark that the functions $\Phi(.-k)$ for $k \geq N-\alpha$ are all supported on $[0, +\infty[$. Moreover, according to (9), the functions $\tilde{\Phi}_\ell$ are purely polynomials on $[0,1]$ if $\alpha = 0$.

PROPOSITION 3.2. *The edge scaling functions $\tilde{\Phi}_\ell, \ell = 0, \dots, N-1$, are linearly independent and are orthogonal to the functions $\Phi_k = \Phi(. - k)$ for $k \geq N - \alpha$* (called the interior scaling functions on $[0, +\infty[$).

Proof. Knowing that the functions $\Phi_k \cdot \chi_{[0,+\infty[}$, for $k = -N+1, \dots, N-1-\alpha$ are linearly independent (see [Meye 92]) and that the degree of the polynomial P_ℓ is exactly ℓ, it is easy to see that by definition of the $\tilde{\Phi}_\ell$ these functions are independent. Moreover, for $\ell \in \{0, \dots, N-1\}$ and $k \geq N - \alpha$, we have

$$\begin{aligned}\left\langle \tilde{\Phi}_\ell \mid \Phi_k \right\rangle_{[0,+\infty[} &= \sum_{m=-N+1}^{N-1-\alpha} P_\ell(m) \int_0^{+\infty} \Phi(x-m)\Phi(x-k)dx \\ &= \sum_{m=-N+1}^{N-1-\alpha} P_\ell(m) \int_{-\infty}^{+\infty} \Phi(x-m)\Phi(x-k)dx \;=\; 0. \qquad \square\end{aligned}$$

We define now

$$V_0^{[0,+\infty[} = \overline{\operatorname{Span}\left\{(\tilde{\Phi}_\ell)_{\ell=0,\dots,N-1}, (\Phi_k)_{k\geq N-\alpha}\right\}},$$

and, more generally,

$$V_j^{[0,+\infty[} = \overline{\operatorname{Span}\left\{\left(\tilde{\Phi}_\ell(2^j.)\right)_{\ell=0,\dots,N-1}, (\Phi(2^j. - k))_{k\geq N-\alpha}\right\}}. \tag{13}$$

Then polynomials up to degree $N-1$ are in $V_j^{[0,+\infty[}$ in the same sense as in (9).

3.1.2. Two-scale equation. We will show that the set of spaces $V_j^{[0,+\infty[}$ is an MRA of $L^2([0,+\infty[)$ after establishing some preliminary results.

LEMMA 3.3. *For $\ell = 0, \dots, N-1$ and $k \in \mathbb{Z}$ we have*

$$\frac{P_\ell(k)}{2^\ell} = \sqrt{2} \sum_{m\in\mathbb{Z}} P_\ell(m)\; h_{k-2m}. \tag{14}$$

Proof. By definition of the polynomial P_ℓ

$$\frac{x^\ell}{\ell!} = \sum_m P_\ell(m)\; \Phi(x-m).$$

Changing the variable $(x \mapsto x/2)$ and using the two-scale equation (4)

$$\begin{aligned}\frac{1}{2^\ell}\frac{x^\ell}{\ell!} \;=\; \left(\frac{x}{2}\right)^\ell \frac{1}{\ell!} &= \sum_m P_\ell(m) \left(\sqrt{2}\sum_k h_{k-2m}\; \Phi(x-k)\right) \\ &= \sum_k \left(\sqrt{2}\sum_m P_\ell(m)\; h_{k-2m}\right) \Phi(x-k),\end{aligned}$$

which leads to (14), by unicity of the polynomial P_ℓ. $\square$

This lemma will be useful to prove the following two-scale equation.

PROPOSITION 3.4. *There exists a matrix b of size N x $(2N-1-\alpha)$ such that, writing $D = (d_{ij})_{1\leq i,j\leq N}$ the diagonal matrix $d_{ij} = \delta_{i-j}/2^{i-1}$,*

$$\begin{pmatrix} \tilde{\Phi}_0 \\ \vdots \\ \tilde{\Phi}_{N-1} \end{pmatrix} \left(\frac{x}{2}\right) = D \begin{pmatrix} \tilde{\Phi}_0 \\ \vdots \\ \tilde{\Phi}_{N-1} \end{pmatrix} (x) + b \begin{pmatrix} \Phi_{N-\alpha} \\ \vdots \\ \Phi_{3N-2-2\alpha} \end{pmatrix} (x). \tag{15}$$

Moreover, the general term of the matrix b is

$$b_{i+1,j-N+\alpha+1} = \frac{P_i(j)}{2^i} - \sqrt{2} \sum_{m=N-\alpha}^{\lfloor \frac{j+N-1}{2} \rfloor} P_i(m) h_{j-2m} \tag{16}$$

for $i = 0, \ldots, N-1$ and $j = N-\alpha, \ldots, 3N-2-2\alpha$, where $\lfloor x \rfloor$ is the integer part of x.

Notice that this formula has the advantage to use a *diagonal* matrix D, which is to be compared to the triangular matrix of [CDV 93]; this will be useful below. Although, contrary to [CDV 93], the supports of $\tilde{\Phi}_\ell$ are not staggered.

Proof. We know that for $x \geq 0$

$$\frac{x^i}{i!} = \tilde{\Phi}_i(x) + \sum_{k=N-\alpha}^{+\infty} P_i(k)\ \Phi(x-k).$$

Rewriting this at $x/2$ and using $2^i(x/2)^i = x^i$ leads to

$$2^i \left[\tilde{\Phi}_i(x/2) + \sqrt{2} \sum_{m=N-\alpha}^{+\infty} P_i(m) \sum_{k=2m-N+1}^{2m+N} h_{k-2m} \Phi(x-k) \right] = \tilde{\Phi}_i(x) + \sum_{k=N-\alpha}^{+\infty} P_i(k) \Phi(x-k)\ .$$

Inverting the sums in the left-hand side:

$$2^i \left[\tilde{\Phi}_i(x/2) + \sqrt{2} \sum_{k=N-2\alpha+1}^{+\infty} \left(\sum_{m=N-\alpha}^{\lfloor \frac{k+N-1}{2} \rfloor} P_i(m) h_{k-2m} \right) \Phi(x-k) \right] = \tilde{\Phi}_i(x) + \sum_{k=N-\alpha}^{+\infty} P_i(k) \Phi(x-k).$$

Thus

$$\tilde{\Phi}_i(x/2) = \frac{1}{2^i} \tilde{\Phi}_i(x) + \sum_{k=N-\alpha}^{+\infty} \left[\frac{P_i(k)}{2^i} - \sqrt{2} \sum_{m=N-\alpha}^{\lfloor \frac{k+N-1}{2} \rfloor} P_i(m) h_{k-2m} \right] \Phi(x-k)$$

with the convention $\sum_\emptyset = 0$. It just remains to show that

$$\frac{P_i(k)}{2^i} = \sqrt{2} \sum_{m=N-\alpha}^{\lfloor \frac{k+N-1}{2} \rfloor} P_i(m) h_{k-2m}$$

when $k \geq 3N-1-2\alpha$. But for such values of k, all the $h_{k-2m} = 0$ for $m < N-\alpha$ or $m > \lfloor k+N-1/2 \rfloor$, so that this is exactly (14). □

Using (14) we could also write

$$b_{i+1,j-N+\alpha+1} = \sqrt{2} \sum_{m=\lceil \frac{j-N}{2} \rceil}^{N-\alpha-1} P_i(m) h_{j-2m}.$$

The next proposition will be useful later.

PROPOSITION 3.5. *The rank of the matrix b defined in equations* (15–16) *is $N-\alpha$, and every submatrix b' obtained by extracting $N-\alpha$ rows of b and keeping one other two column, starting from the $(\alpha+1)$th is of full rank.*

Proof.

(i) For $\alpha = 0$. Note P the $N \times N$ matrix $P_{i,j} = P_{i-1}(j-1)$:

$$P = \begin{pmatrix} P_0(0) & \dots & P_0(N-1) \\ \vdots & \vdots & \vdots \\ P_{N-1}(0) & \dots & P_{N-1}(N-1) \end{pmatrix}.$$

Since *degree* $P_i = i$, the determinant of this matrix is the same as the one of a Vandermonde matrix; it is thus invertible. We can now write

$$b = \sqrt{2}\ P \begin{pmatrix} h_N & 0 & 0 & \dots & \dots & 0 \\ h_{N-2} & h_{N-1} & h_N & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ h_{-N+4} & h_{-N+5} & & \ddots & 0 & 0 \\ h_{-N+2} & h_{-N+3} & \dots & \dots & h_{N-1} & h_N \end{pmatrix},$$

and thus the rank of b is the same as the one of the right matrix. But this $N \times (2N-1)$ matrix is of rank N: extracting the square matrix by taking one other two column (including thus the first and the last), we get a triangular matrix whose diagonal is only composed of $h_N \neq 0$. It is thus invertible.

(ii) For $\alpha = 1$. Note this time P the $N \times (N-1)$ matrix $P_{i,j} = P_{i-1}(j-1)$:

$$P = \begin{pmatrix} P_0(0) & \dots & P_0(N-2) \\ \vdots & \vdots & \vdots \\ P_{N-1}(0) & \dots & P_{N-1}(N-2) \end{pmatrix}.$$

Its rank is $N-1$ and getting out whatever line we get a square matrix of rank $N-1$. Moreover,

$$b = \sqrt{2}\ P \begin{pmatrix} h_{N-1} & h_N & 0 & \dots & \dots & \dots & 0 \\ h_{N-3} & h_{N-2} & h_{N-1} & h_N & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ h_{-N+5} & h_{-N+6} & h_{-N+7} & \dots & \ddots & 0 & 0 \\ h_{-N+3} & h_{-N+4} & h_{-N+5} & \dots & \dots & h_{N-1} & h_N \end{pmatrix},$$

and the $(N-1) \times (2N-2)$ right matrix is of rank $N-1$ (extract one other two column, starting from the second column to see that). Then it becomes clear that b is of rank $N-1$ and that every extracted matrix b' taking any $N-1$ rows is also of rank $N-1$. □

Now, we can prove the main point of the section in the following.

THEOREM 3.6. *The set of spaces $\{V_j^{[0,+\infty[}\}_{j\in\mathbb{Z}}$ is an MRA of $L^2([0,+\infty[)$ in the sense:*

1. $\{0\} = \bigcap_{j\in\mathbb{Z}} V_j^{[0,+\infty[} \subset \cdots \subset V_{-1}^{[0,+\infty[} \subset V_0^{[0,+\infty[} \subset V_1^{[0,+\infty[} \subset \cdots \subset \overline{\bigcup_{j\in\mathbb{Z}} V_j^{[0,+\infty[}} = L^2([0,+\infty[)$.

2. $f(x) \in V_0^{[0,+\infty[} \Longleftrightarrow f(2^j x) \in V_j^{[0,+\infty[}$.

Proof.

(i) First verify that $V_{-1}^{[0,+\infty[} \subset V_0^{[0,+\infty[}$. Thanks to (15), we can see that $\tilde{\Phi}_\ell(x/2) \in V_0^{[0,+\infty[}$. Moreover, using (4), $\Phi_k(x/2) = \sqrt{2}\sum_{m=-N+1}^{N} h_m \Phi_{2k+m}(x)$, and for $k \geq N-\alpha$ and $m \geq -N+1$, we have $2k+m \geq N+1-2\alpha \geq N-\alpha$, so that $\Phi_k(x/2) \in V_0^{[0,+\infty[}$.

(ii) The density of $\bigcup_{j\in Z} V_j^{[0,+\infty[}$ in $L^2([0,+\infty[)$ is directly derived from the density of $\bigcup_{j\in Z} \mathrm{Span}\left\{\Phi(2^j.-k)\ ;\ k \geq N\right\}$ in $L^2([0,+\infty[)$ (see [Meye 92], [CDV 93]).

(iii) The property $\bigcap_{j\in Z} V_j^{[0,+\infty[} = \lim_{j\to-\infty} V_j^{[0,+\infty[} = 0$ results from the fact that all basis functions of $V_0^{[0,+\infty[}$ are in $L^\infty([0,+\infty[)$. □

3.2. Multiresolution analysis on [0,+∞[with boundary conditions. Let us introduce the notations for the homogeneous boundary conditions we impose to the functions.

DEFINITION 3.7. *Let $\Lambda \subset \{0,\ldots,N-1\}$ (it may be $\emptyset$). We define by $BC(\Lambda)$ the vector space of functions f in $L^2([0,+\infty[)$ that are at least $\max\Lambda$ times derivable at 0 and satisfy*

$$\forall \lambda \in \Lambda,\ f^{(\lambda)}(0) = 0.$$

We define now the spaces

$$V_j^{[0,+\infty[}(\Lambda) = V_j^{[0,+\infty[} \bigcap BC(\Lambda)\ . \tag{17}$$

Notice that in the case $\alpha = 0$, all edge functions are polynomial near 0, so there is no problem of derivability at 0. On the contrary, if $\alpha = 1$, the regularity of the edge scaling functions at 0 depends on N. In this case, we will suppose in the following that N is sufficiently large so that no problem of derivability would occur.

Remark that by definition of the edge scaling functions, we have

$$\tilde{\Phi}_\ell^{(\lambda)}(0) = \delta_{\ell-\lambda} \ \text{ for } \ell \in \{0,\ldots,N-1\}, \text{ and } \ \Phi_k^{(\lambda)}(0) = 0 \ \text{ for } k \geq N-\alpha,$$

whatever $\lambda \in \Lambda$. So that according to Proposition 3.2, a basis of $V_0^{[0,+\infty[}(\Lambda)$ is given by the family

$$\left\{\tilde{\Phi}_\ell\right\}_{\ell\notin\Lambda} \bigcup \{\Phi_k\}_{k\geq N-\alpha}\,.$$

To get an orthonormal basis of $V_0^{[0,+\infty[}(\Lambda)$, it is sufficient to orthonormalize the functions $\tilde{\Phi}_\ell$ for $\ell \notin \Lambda$. Let $\tilde{\tilde{\Phi}}_\ell, \ell \notin \Lambda$ the functions obtained by orthonormalization (the orthonormalization will be detailed in section 5). We will now construct the wavelets associated with this MRA.

3.3. Construction of the wavelets. In this section we construct the wavelet basis associated to the MRA $\{V_j^{[0,+\infty[}(\Lambda)\}$, with or without boundary conditions (this

latter case corresponds to $\Lambda = \emptyset$). We define the subspace of $L^2([0,+\infty[)$ orthogonal to $V_j^{[0,+\infty[}(\Lambda)$ in $V_{j+1}^{[0,+\infty[}(\Lambda)$:

$$W_j^{[0,+\infty[}(\Lambda) = V_{j+1}^{[0,+\infty[}(\Lambda) \ominus V_j^{[0,+\infty[}(\Lambda).$$

By scale invariance, the wavelet basis will be constructed from a basis of $W_0^{[0,+\infty[}(\Lambda)$. It is easy to see that the functions Ψ_k, for $k \geq N-\alpha$ belong to $W_0^{[0,+\infty[}(\Lambda)$. These are called interior wavelets. Let us now define the edge wavelets.

DEFINITION 3.8. *For $\ell \in \{0,\dots,N-1\}$ the edge wavelets $\tilde{\Psi}_\ell$ are defined by*

$$\text{For } x \geq 0,\ \tilde{\Psi}_\ell(x) = \sqrt{2}\left(I - P_{V_0^{[0,+\infty[}(\Lambda)}\right)\left(\tilde{\Phi}_\ell(2x) - 2^\ell \tilde{\Phi}_\ell(x)\right),$$

where I is the identity operator.

Notice that for $\ell \in \Lambda$ the equality simplifies because $\tilde{\Phi}_\ell \in V_0^{[0,+\infty[}(\Lambda)$ and therefore $\left(I - P_{V_0^{[0,+\infty[}(\Lambda)}\right)\left(\tilde{\Phi}_\ell\right) = 0$. Using (15) we can also write (using the notations of (15)):

$$\begin{pmatrix} \tilde{\Psi}_0 \\ \vdots \\ \tilde{\Psi}_{N-1} \end{pmatrix} = -\sqrt{2}\left(I - P_{V_0^{[0,+\infty[}(\Lambda)}\right) D^{-1} b \begin{pmatrix} \Phi_{N-\alpha}(2.) \\ \vdots \\ \Phi_{3N-2-2\alpha}(2.) \end{pmatrix}. \tag{18}$$

The main point of the section is the following.

THEOREM 3.9. *The edge wavelets $\tilde{\Psi}_\ell$ for $\ell = 0,\dots,N-1$ verify:*

1. *The functions $\tilde{\Psi}_\ell$ belong to $V_1^{[0,+\infty[}(\Lambda)$.*
2. *The functions $\tilde{\Psi}_\ell$ are orthogonal to $V_0^{[0,+\infty[}(\Lambda)$.*
3. *The functions $\tilde{\Psi}_\ell$ are orthogonal to the interior wavelets Ψ_k, $k \geq N-\alpha$.*
4. *The rank of the family $\left\{\tilde{\Psi}_0,\dots,\tilde{\Psi}_{N-1}\right\}$ is $N-\alpha$ and every subfamily of $N-\alpha$ functions is of full rank.*

Proof. By definition, the functions $\tilde{\Psi}_\ell$ verify the boundary conditions, and (18) leads to the point 1. The point 2 comes directly from Definition 3.8. For the point 3, let $\ell \in \{0,\dots,N-1\}$ and $k \geq N-\alpha$, then

$$\left\langle \tilde{\Psi}_\ell \mid \Psi_k \right\rangle = \left\langle \tilde{\Phi}_\ell(2.) \mid \Psi_k \right\rangle - 2^k \left\langle \tilde{\Phi}_\ell \mid \Psi_k \right\rangle - \left\langle P_{V_0^{[0,+\infty[}(\Lambda)}\left(\tilde{\Phi}_\ell(2x) - 2^\ell \tilde{\Phi}_\ell(x)\right) \mid \Psi_k \right\rangle.$$

Since Ψ_k is orthogonal to $V_0^{[0,+\infty[}(\Lambda)$, and using the two-scale equation (5), we get

$$\left\langle \tilde{\Psi}_\ell \mid \Psi_k \right\rangle = \left\langle \tilde{\Phi}_\ell(2.) \mid \Psi_k \right\rangle = \sum_m g_m \left\langle \tilde{\Phi}_\ell(2.) \mid \Phi(2.-m) \right\rangle = 0,$$

by the orthogonality between the edge scaling functions $\tilde{\Phi}_\ell$ and the interior scaling functions Φ_m.

The last point comes from a dimensional argument. By definition of the functions $\tilde{\Psi}_\ell$, and using (5), it is easy to verify that

$$\begin{aligned} &\operatorname{Span}\left\{\tilde{\Phi}_\ell\right\}_{\ell\notin\Lambda} \oplus \operatorname{Span}\left\{\tilde{\Psi}_\ell\right\}_{\ell=0,\dots,N-1} \\ &\quad = \operatorname{Span}\left\{\tilde{\Phi}_\ell(2.)\right\}_{\ell\notin\Lambda} \oplus \operatorname{Span}\left\{ b \begin{pmatrix} \Phi_{N-\alpha}(2.) \\ \vdots \\ \Phi_{3N-2-2\alpha}(2.) \end{pmatrix} \right\}; \end{aligned}$$

then

$$dim\ \mathrm{Span}\left\{\tilde{\Psi}_\ell\right\}_{\ell=0,\dots,N-1} = dim\ \mathrm{Span}\left\{ b \begin{pmatrix} \Phi_{N-\alpha}(2.) \\ \vdots \\ \Phi_{3N-2-2\alpha}(2.) \end{pmatrix} \right\} = N-\alpha,$$

since the rank of the matrix b is $N-\alpha$, and the result holds from Proposition 3.5. □

Then

$$W_0^{[0,+\infty[}(\Lambda) = \mathrm{span}\left\{ \{\tilde{\Psi}_\ell\}_{0\le\ell\le N-1}\ ;\ \{\Psi_k\}_{k\ge N-\alpha} \right\}.$$

To obtain a basis of $W_0^{[0,+\infty[}(\Lambda)$ it is sufficient to extract $N-\alpha$ functions of $\{\tilde{\Psi}_0,\dots,\tilde{\Psi}_{N-1}\}$ and to add the set of functions $\{\Psi_k\}_{k\ge N-\alpha}$. Then, orthonormalizing $N-\alpha$ of the functions $\tilde{\Psi}_\ell$ and adding the interior wavelets, we obtain an orthonormal basis of $W_0^{[0,+\infty[}(\Lambda)$.

4. Adaptation to the interval [0, 1]. The previous sections were devoted to the half line. Our goal is to construct an MRA on $[0,1]$. We will see how to adapt the above construction to this case. The boundary conditions at edge 0 are taken into account by constructing an MRA of $L^2([0,+\infty[)$ with boundary conditions $BC(\Lambda_0)$, and those at edge 1 by constructing an MRA of $L^2(]-\infty,1])$ with boundary conditions $BC(\Lambda_1)$. Then, merging the two MRAs will lead to the result.

4.1. MRA on]−∞, 1] with boundary conditions. The MRA on $]-\infty,1]$ will be constructed from an MRA on $[0,+\infty[$ by a change of variable given by the operator T:

$$\text{for } f\in L^2(\mathbb{R}),\quad Tf(x) = f(1-x). \tag{19}$$

Then Tf belongs to $L^2(\mathbb{R})$. We see that *support* $T\Phi = [-N+1, N]$, the same as Φ. $T\Phi$ satisfies a two-scale relation:

$$T\Phi(.) = \sqrt{2}\sum_{k=-N+1}^{N} \check{h}_k\ T\Phi(2.-k), \tag{20}$$

where $\check{h}_k = h_{1-k}$ for all k in $\mathbb{Z}$.

Notice that this operator is isometric from $L^2([0,+\infty[)$ to $L^2(]-\infty,1])$:

$$\int_{-\infty}^{1} Tf(x)\ Tg(x)dx = \int_0^{+\infty} f(x)g(x)dx,$$

and it is involutive on $L^2(\mathbb{R})$: $TT = I$.

Starting from the function $T\Phi$, instead of Φ, we can construct a new MRA of $L^2([0,+\infty[)$ satisfying boundary conditions defined by the set Λ_1. It provides us with the following:

(i) edge scaling functions $\tilde{\Phi}_\ell^\sharp$ for $\ell\notin\Lambda_1$,
(ii) interior scaling functions $T\Phi_k$ for $k\ge N-\alpha_1$,
(iii) edge wavelets $\tilde{\Psi}_\ell^\sharp$ for $\ell = 0,\dots,N-1-\alpha_1$,
(iv) interior wavelets $T\Psi_k$ for $k\ge N-\alpha_1$,

with a parameter α_1 which is either 0 or 1. Thus, to generate a basis of an MRA of $L^2(]-\infty,1])$ it is sufficient to consider the functions:

$$T(\tilde{\Phi}_\ell^\sharp), T(T\Phi_k), T(\tilde{\Psi}_\ell^\sharp), \text{ and } T(T\Psi_k).$$

Remark.

$$T\left(T\left[\Phi_k(2^j.)\right]\right)(x) = \Phi_{-1-k+2^j}(2^j x).$$

4.2. MRA on [0, 1] with boundary conditions. We merge the two MRAs of $L^2([0,+\infty[)$ and $L^2(]-\infty,1])$, in order to construct an MRA on $[0,1]$ verifying boundary conditions, $\{V_j^{[0,1]}(\Lambda_0,\Lambda_1)\}_{j\geq j_{min}}$. We will choose j_{min} so that edge functions at 0 and edge functions at 1 do not interact (i.e., their supports are disjoint). For a given scale j, the functions at edge 0 are supported on $[0,(2N-1-\alpha_0)/2^j]$ (we note α_0 instead of α not to confuse the parameters of both edges) and the functions at edge 1 are supported on $[1-(2N-1-\alpha_1)/2^j,1]$. Then j_{min} will be so that

$$2^{j_{min}} \geq 4N-\alpha_0-\alpha_1.$$

Since $\alpha_0,\alpha_1 \leq 1$, we get

$$j_{min} = \lceil \log_2 4N \rceil.$$

The vector space $V_j^{[0,1]}(\Lambda_0,\Lambda_1)$ for $j \geq j_{min}$ will be the span of the following functions:

(i) $\tilde{\Phi}_\ell(2^j.)$ for $\ell \notin \Lambda_0$ (edge functions at 0),
(ii) $\tilde{\Phi}_\ell^\sharp(2^j(1-.))$ for $\ell \notin \Lambda_1$ (edge functions at 1),
(iii) Φ_k for $\{k \geq N-\alpha_0; \exists \ell \geq N-\alpha_1, k=-1-\ell+2^j\}$ (interior scaling functions).

The interior scaling functions are those common in the MRA of $L^2([0,+\infty[)$ and $L^2(]-\infty,1])$. They correspond to

$$N-\alpha_0 \leq k \leq 2^j-N-1+\alpha_1.$$

We would like to renumber the scaling functions at edge 1. So we define

$$\tilde{\Phi}_{2^j-1-l} \stackrel{def}{=} \tilde{\Phi}_\ell^\sharp.$$

The indices cannot be confused with those at edge 0 when $j \geq j_{min}$.

The vector space $W_j^{[0,1]}(\Lambda_0,\Lambda_1)$ for $j \geq j_{min}$ is constructed similarly. Let us evaluate the dimensions of these spaces:

$$\begin{aligned} dim\ V_j^{[0,1]}(\Lambda_0,\Lambda_1) &= (N-\#\Lambda_0) + \big((2^j-N-1+\alpha_1)-(N-\alpha_0)+1\big) + (N-\#\Lambda_1) \\ &= 2^j - \#\Lambda_0 - \#\Lambda_1 + \alpha_0 + \alpha_1, \\ dim\ W_j^{[0,1]}(\Lambda_0,\Lambda_1) &= (N-\alpha_0) + \big((2^j-N-1+\alpha_1)-(N-\alpha_0)+1\big) + (N-\alpha_1) \\ &= 2^j. \end{aligned}$$

The interest of parameters α_0 and α_1 is now clear; when we need equal dimensions (for example, in order to construct wavelet packets), if there is at most one boundary condition at each edge, we can fix the parameters α_0 and α_1 so that $dim V_j^{[0,1]}(\Lambda_0,\Lambda_1) = 2^j$. Therefore, we will choose

$$\alpha_0 = \delta_{\#\Lambda_0-1} \text{ and } \alpha_1 = \delta_{\#\Lambda_1-1}.$$

5. Practical computations.

5.1. Orthonormalization of the scaling functions. It is often more interesting to have an orthonormal basis than a Riesz basis for the spaces $V_j^{[0,1]}(\Lambda_0,\Lambda_1)$ and $W_j^{[0,1]}(\Lambda_0,\Lambda_1)$. As we have seen before, orthonormalizing the edge functions would lead to orthonormal bases, since they are already orthogonal to interior functions. For that, we need their Gram matrix.

We use (15). Multiplying by the transpose of each term of this equality and integrating on $[0,+\infty[$ each member leads to the following proposition.

PROPOSITION 5.1. *The Gram matrix $G^{\tilde\Phi}$ of the edge scaling functions $\tilde\Phi_\ell$ for $\ell=0,\dots,N-1$ is given by*

$$2G^{\tilde\Phi} = DG^{\tilde\Phi}D + b\,{}^t\!b, \tag{21}$$

where the matrices D and b have been introduced in (15).

Actually the product $DG^{\tilde\Phi}D$ is the term-by-term multiplication of the matrix $G^{\tilde\Phi}$ by the matrix M of general term $M_{ij}=1/2^{i+j-2}$ for $1\le i,j\le N$. Denoting by M_1 the $N\times N$ matrix whose general term is 1, computing $G^{\tilde\Phi}$ is equivalent to divide term-by-term the matrix $b\,{}^t\!b$ by the matrix $2M_1-M$. Note that the computation of this matrix is easier than in [CDV 93].

To take into account the boundary conditions, we retain only the functions $\tilde\Phi_\ell$ for $\ell\notin\Lambda$. Their Gram matrix is obtained from $G^{\tilde\Phi}$ by keeping only the rows and the columns whose index is not in $1+\Lambda$. The 1 comes from the fact that we number the rows and columns of the matrix starting from 1 and not from 0: let $G_\Lambda^{\tilde\Phi}$ this matrix. The Gram procedure allows us to orthonormalize the edge scaling functions $(\tilde\Phi_\ell)_{\ell\notin\Lambda}$.

PROPOSITION 5.2. *Defining*

$$\begin{pmatrix} \tilde{\tilde\Phi}_0 \\ \vdots \\ \tilde{\tilde\Phi}_{N-1} \end{pmatrix}_{\ell\notin\Lambda} = (G_\Lambda^{\tilde\Phi})^{-\frac12} \begin{pmatrix} \tilde\Phi_0 \\ \vdots \\ \tilde\Phi_{N-1} \end{pmatrix}_{\ell\notin\Lambda} \tag{22}$$

the family of scaling functions $\{\tilde{\tilde\Phi}_\ell\}_{\ell\notin\Lambda}$ is orthonormal and satisfies the two-scale equation

$$\frac{1}{\sqrt2}\begin{pmatrix} \tilde{\tilde\Phi}_0 \\ \vdots \\ \tilde{\tilde\Phi}_{N-1} \end{pmatrix}_{\ell\notin\Lambda}\left(\frac{x}{2}\right) = H_0\begin{pmatrix} \tilde{\tilde\Phi}_0 \\ \vdots \\ \tilde{\tilde\Phi}_{N-1} \end{pmatrix}_{\ell\notin\Lambda}(x) + h_0\begin{pmatrix} \Phi_{N-\alpha} \\ \vdots \\ \Phi_{3N-2-2\alpha} \end{pmatrix}(x) \tag{23}$$

where H_0 and h_0 are the $(N-\#\Lambda)\times(N-\#\Lambda)$ and $(N-\#\Lambda)\times(2N-1-\alpha)$ matrices:

$$H_0 = \frac{1}{\sqrt2}\,(G_\Lambda^{\tilde\Phi})^{-1/2}\,D_\Lambda\,(G_\Lambda^{\tilde\Phi})^{1/2} \quad \textit{and} \quad h_0 = \frac{1}{\sqrt2}\,(G_\Lambda^{\tilde\Phi})^{-1/2}\,b_\Lambda.$$

The matrix D_Λ is extracted from D retaining only rows and columns of numbers not in $\Lambda+1$ and b_Λ from b retaining rows of numbers not in $\Lambda+1$.

Remark that $H_0\,{}^t\!H_0 + h_0\,{}^t\!h_0 = I$.

5.2. Orthonormalization of the wavelets. Similarly as for the scaling functions, orthonormalizing the wavelet basis amounts to orthonormalizing the edge wavelets $\tilde{\Psi}_\ell$. To compute the Gram matrix $G^{\tilde{\Psi}}$ of the $\tilde{\Psi}_\ell$, we need the following lemma.

LEMMA 5.3. *The edge wavelets $\tilde{\Psi}_\ell$ satisfy*

$$\frac{1}{\sqrt{2}}\begin{pmatrix}\tilde{\Psi}_0\\ \vdots\\ \tilde{\Psi}_{N-1}\end{pmatrix}\left(\frac{x}{2}\right) = g_1 \begin{pmatrix}\tilde{\tilde{\Phi}}_0\\ \vdots\\ \tilde{\tilde{\Phi}}_{N-1}\end{pmatrix}_{\ell\notin\Lambda}(x) + g_2 \begin{pmatrix}\Phi_{N-\alpha}\\ \vdots\\ \Phi_{3N-2-2\alpha}\end{pmatrix}(x) \tag{24}$$

with $g_1 = D^{-1}b\ {}^t h_0 H_0$ and $g_2 = D^{-1}b\ ({}^t h_0 h_0 - I)$.

Proof. By definition (18) of the $\tilde{\Psi}_\ell$:

$$\left(\frac{1}{\sqrt{2}}\tilde{\Psi}_\ell\left(\frac{x}{2}\right)\right)_{\ell=0,N-1} = -D^{-1}b \begin{pmatrix}\Phi_{N-\alpha}(x)\\ \vdots\\ \Phi_{3N-2-2\alpha}(x)\end{pmatrix} + D^{-1}b\ P_{V_0^{[0,+\infty[}(\Lambda)} \begin{pmatrix}\Phi_{N-\alpha}(2.)\\ \vdots\\ \Phi_{3N-2-2\alpha}(2.)\end{pmatrix}\left(\frac{x}{2}\right).$$

Let us compute, for $k = N-\alpha,\dots,3N-2-2\alpha$:

$$P_{V_0^{[0,+\infty[}(\Lambda)}\left(\Phi_k(2.)\right)\left(\frac{x}{2}\right) = \sum_{\ell\notin\Lambda}\langle\,\Phi_k(2.)\mid\tilde{\tilde{\Phi}}_\ell\rangle\,\tilde{\tilde{\Phi}}_\ell\left(\frac{x}{2}\right) + \sum_{n\geq N-\alpha}\langle\Phi_k(2.)\mid\Phi_n\rangle\,\Phi\left(\frac{x}{2}-n\right).$$

We have $\langle\Phi_k(2.)\mid\Phi_n\rangle = 0$ for $N-\alpha\leq k\leq 3N-2-2\alpha$ and $n\geq N-\alpha$.

Applying twice (23) and the orthonormality between the $\tilde{\tilde{\Phi}}_\ell$ leads to

$$\left(\langle\Phi_k(2.)\mid\tilde{\tilde{\Phi}}_\ell\rangle\right)_{\substack{N-\alpha\leq k\leq 3N-2-2\alpha\\ \ell\notin\Lambda}} = \frac{1}{\sqrt{2}}\ {}^t h_0,$$

and using (23) once more:

$$P_{V_0^{[0,+\infty[}(\Lambda)}\begin{pmatrix}\Phi_{N-\alpha}(2.)\\ \vdots\\ \Phi_{3N-2-2\alpha}(2.)\end{pmatrix}\left(\frac{x}{2}\right) = {}^t h_0\ H_0 \begin{pmatrix}\tilde{\tilde{\Phi}}_0(x)\\ \cdots\\ \tilde{\tilde{\Phi}}_{N-1}(x)\end{pmatrix}_{\ell\notin\Lambda} + {}^t h_0\ h_0 \begin{pmatrix}\Phi_{N-\alpha}(x)\\ \vdots\\ \Phi_{3N-2-2\alpha}(x)\end{pmatrix},$$

which gives the expected result. ∎

It is then easy to compute the Gram matrix $G^{\tilde{\Psi}}$of the edge wavelets:

$$G^{\tilde{\Psi}} = g_1\,{}^t g_1 + g_2\,{}^t g_2.$$

Remember that in the case $\alpha = 1$, we must skip one of the functions $\tilde{\Psi}_\ell$ then we skip the row and the corresponding column in $G^{\tilde{\Psi}}$. This matrix will be called $G^{\tilde{\Psi}}$ again. In order to simplify the notations, let us assume that the removed function is $\tilde{\Psi}_{N-1}$.

PROPOSITION 5.4. *Let* $G_0 = (G^{\tilde{\Psi}})^{-1/2}\ g_1$ *and* $g_0 = (G^{\tilde{\Psi}})^{-1/2}\ g_2$, *we can write*

$$\frac{1}{\sqrt{2}} \begin{pmatrix} \tilde{\tilde{\Psi}}_0 \\ \vdots \\ \tilde{\tilde{\Psi}}_{N-1-\alpha} \end{pmatrix} \left(\frac{x}{2}\right) = G_0 \begin{pmatrix} \tilde{\tilde{\Phi}}_0 \\ \vdots \\ \tilde{\tilde{\Phi}}_{N-1} \end{pmatrix}_{k \notin \Lambda} (x) + g_0 \begin{pmatrix} \Phi_{N-\alpha} \\ \vdots \\ \Phi_{3N-2-2\alpha} \end{pmatrix} (x). \tag{25}$$

As in [Ausc 93], the bases constructed here allow us to characterize some subspaces of $H^s([0,1])$ defined by vanishing derivatives at the edges. The characterizations provide, moreover, a diagonal preconditioner for the derivative operators, as in [Jaff 92].

Numerically, all the calculations have been done to obtain the two-scale relations (23) and (25) that give the matrices used in the fast algorithms.

5.3. Fast algorithms. As in the case of the real line, we can describe a fast algorithm for analyzing and synthesizing vectors in the spaces of the MRA. The algorithm is based on the elementary step

$$V_j^{[0,1]}(\Lambda_0, \Lambda_1) = V_{j-1}^{[0,1]}(\Lambda_0, \Lambda_1) \oplus W_{j-1}^{[0,1]}(\Lambda_0, \Lambda_1).$$

Suppose that we know the scalar products of a given function f in $V_j^{[0,1]}(\Lambda_0, \Lambda_1)$ with the corresponding scaling functions, and that we want to compute scalar products with the scaling functions of $V_{j-1}^{[0,1]}(\Lambda_0, \Lambda_1)$ and with the wavelets of $W_{j-1}^{[0,1]}(\Lambda_0, \Lambda_1)$.

We start then from a vector c^j composed of:

(i) $\langle\ f \mid 2^{j/2}\tilde{\tilde{\Phi}}_\ell(2^j .)\ \rangle$ for $\ell \in \{0, \ldots, N-1\} \backslash \Lambda_0$,

(ii) $\langle\ f \mid 2^{j/2}\Phi_k(2^j .)\ \rangle$ for $N - \alpha_0 \leq k \leq 2^j - 1 - N + \alpha_1$,

(iii) $\langle\ f \mid 2^{j/2}\tilde{\tilde{\Phi}}_\ell(2^j .)\ \rangle$ for $\ell \in \{2^j - 1, \ldots, 2^j - N\} \backslash \{2^j - 1 - \Lambda_1\}$.

To obtain the projection of the function f on $V_{j-1}^{[0,1]}(\Lambda_0, \Lambda_1)$, we have to multiply the vector c^j by the matrix

$$\begin{pmatrix} H_0 & h_0 & 0 \\ 0 & \mathcal{H}^j & 0 \\ 0 & h_1 & H_1 \end{pmatrix},$$

where:

(i) $\mathcal{H}^j$ is the square $(2^j - 2N + \alpha_0 + \alpha_1) \times (2^j - 2N + \alpha_0 + \alpha_1)$ matrix whose general term is $\mathcal{H}^j_{k,l} = h_{-N+1+\alpha_0+l-2k}$ (the classical filter h_k is defined in (4)),

(ii) h_0 (defined, as H_0, in (23) for the edge 0) is completed with columns of 0 at the right to fit the width of $\mathcal{H}^j$,

(iii) h_1 (defined, as H_1, in (23) for the edge 1) with columns of 0 at the left.

To obtain the projection on $W_{j-1}^{[0,1]}(\Lambda_0, \Lambda_1)$, we multiply c^j by the matrix

$$\begin{pmatrix} G_0 & g_0 & 0 \\ 0 & \mathcal{G}^j & 0 \\ 0 & g_1 & G_1 \end{pmatrix},$$

where

(i) $\mathcal{G}^j$ is the square $(2^j - 2N + \alpha_0 + \alpha_1) \times (2^j - 2N + \alpha_0 + \alpha_1)$ matrix whose general term is $\mathcal{G}^j_{k,l} = g_{-N+1+\alpha_0+l-2k}$ (the classical filter g_k is defined in (5)),

(ii) g_0 (defined, as G_0, in (25) for the edge 0) is completed with columns of 0 at the right to fit the width of $\mathcal{G}^j$,

(iii) g_1 (defined, as G_1, in (25) for the edge 1) with columns of 0 at the left.

The result lies in the vectors c^{j-1} and d^{j-1}.

To find c^j starting from c^{j-1} and d^{j-1}, we have to multiply the column vector

$$\begin{pmatrix} c^{j-1} \\ d^{j-1} \end{pmatrix}$$

by the matrix

$$\begin{pmatrix} {}^tH_0 & 0 & 0 & {}^tG_0 & 0 & 0 \\ {}^th_0 & \tilde{\mathcal{H}}^j & {}^th_1 & {}^tg_0 & \tilde{\mathcal{G}}^j & {}^tg_1 \\ 0 & 0 & {}^tH_1 & 0 & 0 & {}^tG_1 \end{pmatrix},$$

where the matrices $\tilde{\mathcal{H}}^j$ and $\tilde{\mathcal{G}}^j$ (which have the same size as $\mathcal{H}^j$ and $\mathcal{G}^j$) are built from the h_k and g_k and h_0, g_0, h_1, g_1 are completed with columns of zeros, as above. Notice that in practice, for this algorithm, matrix multiplication is used only for the edges; for interior points it is much faster to rely on discrete convolutions as on the real line.

5.4. Numerical filters and graphs. As an example we compute, for $N = 2$ vanishing moments, the discrete edge filters H_0, H_1, h_0, h_1, G_0, G_1, g_0, g_1 introduced in (23–25) (Tables 1, 2, and 3), for various boundary conditions and values of α. These filters are associated to the classical filters on $\mathbb{R}$ defined by [Daub 88]:

h_{-1}	=	0.48296291314453	g_{-1}	=	−0.12940952255126
h_0	=	0.83651630373781	g_0	=	−0.22414386804201
h_1	=	0.22414386804201	g_1	=	0.83651630373781
h_2	=	−0.12940952255126	g_2	=	−0.48296291314453.

In addition, we plot the graphs of the corresponding edge scaling functions and edge wavelets for Dirichlet boundary conditions and for the two possible values of the parameter α:

– for $\alpha = 0$: table 2 and Figure 1 (edge 0),

– for $\alpha = 1$: table 3 and Figure 2 (edge 0).

Note that for $\alpha = 0$ the edge functions are polynomials up to degree $N - 1$ near the boundary, this is no more true for $\alpha = 1$, but in this case $dim\ V_j^{[0,1]}(Dirichlet) = dim\ W_j^{[0,1]}(Dirichlet) = 2^j$.

6. Computation of operators. In the following, we will suppose that N is large enough so that the scaling functions and the wavelets are sufficiently differentiable, in order to compute the derivative operators.

6.1. First-order derivative operator. We focus first at edge 0. We want to compute the map:

$$P_{V_j^{[0,+\infty[}(\Lambda)} \left(\frac{d}{dx}\right) P_{V_j^{[0,+\infty[}(\Lambda)},$$

where $P_{V_j^{[0,+\infty[}(\Lambda)}$ is the orthogonal projection on $V_j^{[0,+\infty[}(\Lambda)$. Since we know an orthonormal basis of $V_j^{[0,+\infty[}(\Lambda)$, it is sufficient to estimate the operator on each

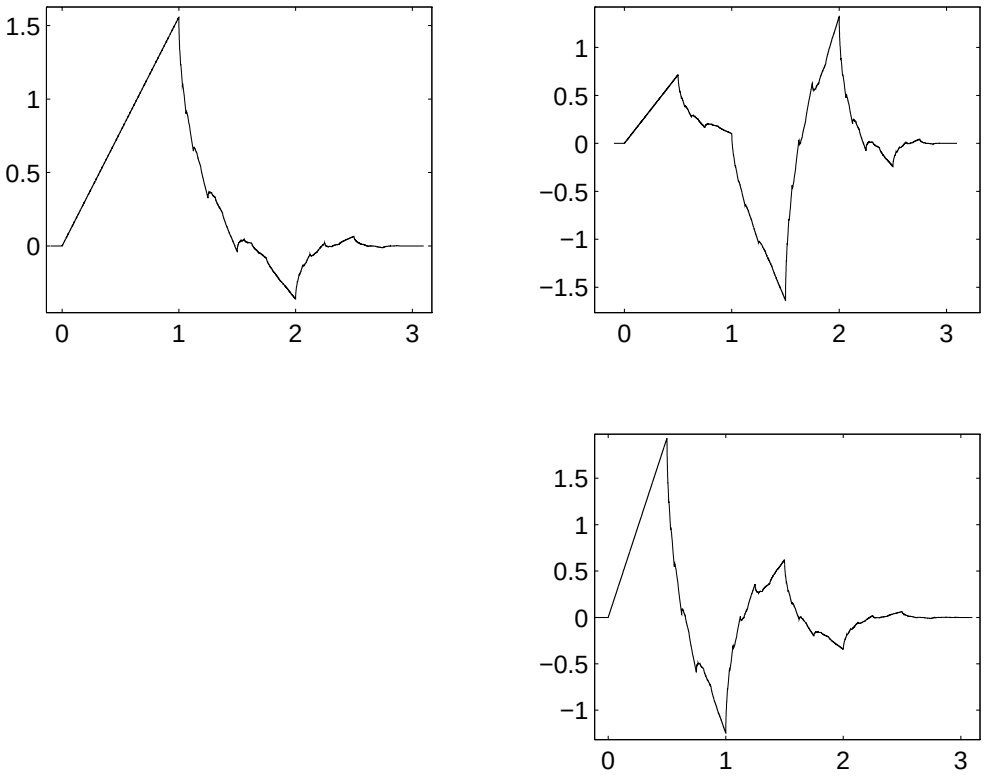

FIG. 1. *Scaling function (first column) and wavelets (second) at edge* 0 *for* $N = 2, \Lambda = \{0\}, \alpha = 0$ *(Dirichlet boundary conditions).*

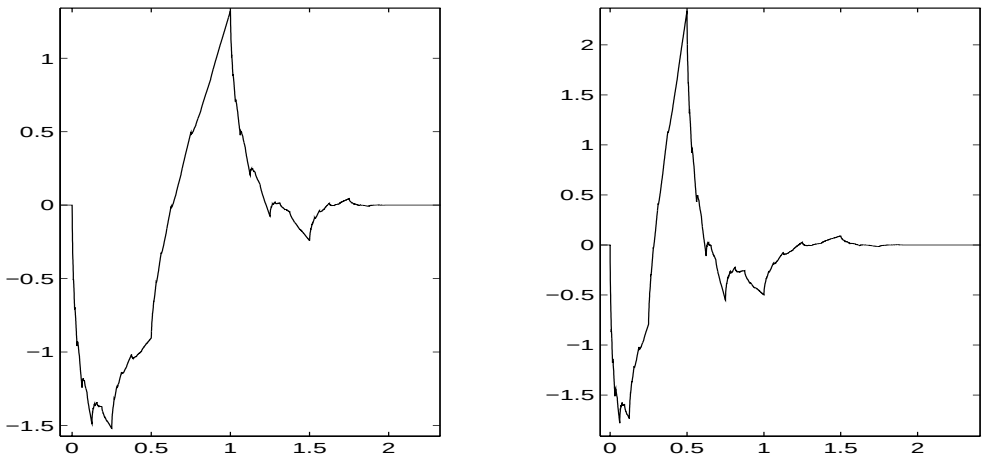

FIG. 2. *Scaling function (first column) and wavelet (second) at edge* 0 *for* $N = 2, \Lambda = \{0\}, \alpha = 1$ *(Dirichlet boundary conditions).*

TABLE 1
Discrete filters at the edges for $N = 2, \Lambda = \emptyset$ (without boundary conditions).

H_0		h_0		
0.88593511638920	0.21935410044001	0.36114692232438	0.16562528942820	-0.09562380543598
-0.43402397094665	0.17472505539062	0.86653097274512	0.15054506203086	-0.08691723208868
G_0		g_0		
0.12072248810796	-0.11030975957357	0.27289605514679	-0.82102934519411	0.47402151346040
-0.11030975957357	0.95351281616018	-0.21029673524411	-0.16067128595389	0.09276361019652

h_1			H_1	
0.44293306463308	0.76718257229668	0.36758039943505	0.18330518269759	-0.21569668742557
0.01383403229066	0.02396124680098	0.24198194598437	0.41343371796754	0.87735498908223
g_1			G_1	
-0.02810340334383	-0.04867652245712	-0.44874160052110	0.84864346408161	-0.27436479678578
0.22984585818723	0.39810470428955	-0.77779197972887	-0.27436479678578	0.32931310397683

TABLE 2
Discrete filters at the edges for $N = 2, \Lambda = \{0\}$ (Dirichlet boundary condition), $\alpha = 0$.

H_0	h_0		
0.35355339059327	0.89982198849888	0.22133750644759	-0.12778926892928
G_0	g_0		
0.32472811514904	0.13816747232024	-0.81030614623011	0.46783047165196
0.87724093111970	-0.41379963582216	0.21074548066631	-0.12167395999319

h_1			H_1
0.41483432960318	0.71851413559648	0.43202999424892	0.35355339059327
g_1			G_1
-0.24314023061746	-0.42113123279346	0.24823276949896	0.83780176961381
0.13709597819763	0.23745719975166	-0.86702397672507	0.41603869391197

vector of the basis. Using the fact that we are working on an MRA of $L^2([0,+\infty[)$ which is scale invariant, we just have to compute

$$\partial_0 = P_{V_0^{[0,+\infty[}(\Lambda)} \left(\frac{d}{dx}\right) P_{V_0^{[0,+\infty[}(\Lambda)}.$$

We have thus to compute four types of inner products:

$$r_{k,\ell} = \left\langle \Phi_k | \frac{d\Phi_\ell}{dx} \right\rangle \quad \text{for } k, \ell \geq N - \alpha, \tag{26}$$

$$B_{k,\ell} = \left\langle \Phi_k | \frac{d\tilde{\tilde{\Phi}}_\ell}{dx} \right\rangle \quad \text{for } k \geq N - \alpha,\ \ell \notin \Lambda, \tag{27}$$

$$B'_{k,\ell} = \left\langle \tilde{\tilde{\Phi}}_k | \frac{d\Phi_\ell}{dx} \right\rangle \quad \text{for } k \notin \Lambda,\ \ell \geq N - \alpha, \tag{28}$$

$$A_{k,\ell} = \left\langle \tilde{\tilde{\Phi}}_k | \frac{d\tilde{\tilde{\Phi}}_\ell}{dx} \right\rangle \quad \text{for } k, \ell \notin \Lambda. \tag{29}$$

Notice that integrating by parts we have since $\Phi_\ell(0) = 0$, $B'_{k,\ell} = -B_{\ell,k}$. The calculation of the $r_{k,\ell}$ is as follows:

$$r_{k,\ell} = \int_0^{+\infty} \Phi(x-k)\ \frac{d\Phi}{dx}(x-\ell)\ dx = \int_{\mathbb{R}} \Phi(x-(k-\ell))\ \frac{d\Phi}{dx}(x)\ dx \overset{def}{=} r_{k-\ell}.$$

As explained in [Beyl 92], the r_i can be computed by an eigenvalue problem.

TABLE 3
Discrete filters at the edges for $N = 2, \Lambda = \{0\}$ (Dirichlet boundary condition), $\alpha = 1$.

H_0	h_0	
0.35355339059327	-0.81009258730098	0.46770717334674
G_0	g_0	
0.93541434669349	0.30618621784790	-0.17677669529664

h_1		H_1
0.46770717334674	0.81009258730098	0.35355339059327
g_1		G_1
-0.17677669529664	-0.30618621784790	0.93541434669349

It remains to obtain the matrices A and B corresponding to the terms $A_{k,\ell}$ and $B_{k,\ell}$. According to the supports of the scaling functions, A is a square matrix of size $(N - \#\Lambda)^2$ and B a rectangular matrix, of size $(N - \#\Lambda)$ x $(2N - 2)$.

PROPOSITION 6.1. *The matrices A and B in* (29) and (27) *can be expressed:*

(i) *The matrix B of size $(N - \#\Lambda) \times (2N - 2 - \alpha)$ is given by*

$$B = (G_\Lambda^{\tilde\Phi})^{-1/2}\ B^{\tilde\Phi}$$

where

$$\begin{aligned} B^{\tilde\Phi}_{i+1,j-N+\alpha+1} &\overset{def}{=} \left\langle \tilde\Phi_i \mid \frac{d\Phi_j}{dx} \right\rangle \\ &= \sum_{k=-N+1}^{N-1-\alpha} P_i(k)\ r_{k-j}, \end{aligned} \tag{30}$$

for $i = 0, \dots, N-1, i \notin \Lambda$, and $N - \alpha \le j \le 3N - 3 - 2\alpha$.

(ii) *The matrix A of size $(N - \#\Lambda) \times (N - \#\Lambda)$ is given by*

$$A = (G_\Lambda^{\tilde\Phi})^{-1/2}\ A^{\tilde\Phi}\ {}^t(G_\Lambda^{\tilde\Phi})^{-1/2}$$

where $A^{\tilde\Phi}$ satisfies

$$A^{\tilde\Phi} = D_\Lambda\ A^{\tilde\Phi}\ D_\Lambda - b_\Lambda\ {}^tB^{\tilde\Phi}\ D_\Lambda + D_\Lambda\ B^{\tilde\Phi}\ {}^tb_\Lambda + b_\Lambda\ r\ {}^tb_\Lambda \tag{31}$$

(*r is the square matrix of size $2N - 1 - \alpha$ whose general term is r_{i-j}, D_Λ is the submatrix of D in* (15) *containing only the rows and columns of numbers not in $1 + \Lambda$, b_Λ the submatrix of b in* (15) *containing only the rows of numbers not in $1 + \Lambda$, and $G_\Lambda^{\tilde\Phi}$ is defined in* (21)).

Actually in the case $\alpha = 0$, the above equation has no meaning because of the size of $B^{\tilde\Phi}$: we must add a new column of 0 at the right corresponding to the zero value of the scalar products $\langle \tilde\Phi_k | d\Phi_{3N-2}/dx \rangle$ (the intersection of the supports of these functions is reduced to one point).

Proof. Equation (30) comes directly from Definition 3.1 of the edge functions $\tilde\Phi_i$. Equation (31) is derived from (15) by taking the derivative of each member, transposing, multiplying at right with (15), and integrating. $A^{\tilde\Phi}$ is the matrix of the scalar products $\langle \tilde\Phi_k | d\tilde\Phi_\ell/dx \rangle$. □

As for the Gram matrix in (21), $A^{\tilde\Phi}$ can be computed by a term by term division of matrices. Nevertheless, we can see that here, if $0 \notin \Lambda$, this does not determine the

upper left term of $A^{\tilde{\Phi}}$. Fortunately, we can compute it directly:

$$A^{\tilde{\Phi}}_{1,1} = \int_0^{+\infty} \tilde{\Phi}_0(x) \frac{d\tilde{\Phi}_0}{dx}(x)dx = -\frac{1}{2}\tilde{\Phi}_0(0)^2 = -\frac{1}{2}.$$

As for the construction of the MRA on $[0,1]$, the adaptation to the interval is given through the operator T (given in (19)) and

$$T\frac{df}{dx}(x) \stackrel{def}{=} \frac{df}{dx}(1-x) = -\frac{d\ Tf}{dx}(x),$$

which leads to the following shape for the matrix of the first-order derivative operator in $V_j^{[0,1]}(\Lambda_0, \Lambda_1)$:

$$\delta_j = \left(\begin{array}{c|c|c} A_0 & B_0 & 0 \\ \hline -{}^tB_0 & r^j & -{}^tB_1 \\ \hline 0 & B_1 & A_1 \end{array}\right)$$

where the matrices B_0 (edge 0) and B_1 (edge 1) are completed with columns of zeros, and r^j is built from the r_k. Notice that as usual, the multiplication of a column vector by r^j can be computed by a discrete convolution, which is more efficient.

It is often more interesting to know the derivative operator expressed in the basis of scaling functions at the largest scale and wavelets at intermediate scales. Actually we know the matrix of the derivative operator in the basis of $V_J^{[0,1]}(\Lambda_0, \Lambda_1)$. We use the decomposition

$$V_J^{[0,1]}(\Lambda_0, \Lambda_1) = W_{J-1}^{[0,1]}(\Lambda_0, \Lambda_1) \oplus \cdots \oplus W_{j_{min}}^{[0,1]}(\Lambda_0, \Lambda_1) \oplus V_{j_{min}}^{[0,1]}(\Lambda_0, \Lambda_1).$$

Let $\mathcal{B}$ the basis $V_J^{[0,1]}(\Lambda_0, \Lambda_1)$ composed of scaling functions at the coarsest scale and wavelets at intermediate scales (called the "wavelet" basis). To compute the derivative matrix in this "wavelet" basis, since all the bases are orthonormal, the matrix of the change of basis P from the scaling function basis of $V_J^{[0,1]}(\Lambda_0, \Lambda_1)$ to $\mathcal{B}$ satisfies $P^{-1} = {}^tP$, so that the matrix of the derivative operator in $\mathcal{B}$ is ${}^tP\delta_J P$. Figure 3 gives the shape of the first-order derivative operator and the number of nonzero coefficients, in the two bases.

6.2. Second-order derivative operator. In this section we want to compute the Galerkin operator:

$$P_{V_j^{[0,+\infty[}(\Lambda)} \left(\frac{d^2}{dx^2}\right) P_{V_j^{[0,+\infty[}(\Lambda)}.$$

Knowing this time the scalar products among the scaling functions

$${}_2r_k = \left\langle\ \Phi_k \mid \frac{d^2\Phi}{dx^2}\ \right\rangle,$$

we want to compute the matrices ${}_2A$ and ${}_2B$ of the scalar products:

$$ {}_2A_{i,j} = \left\langle\ \tilde{\tilde{\Phi}}_i \mid \frac{d^2\tilde{\tilde{\Phi}}_j}{dx^2}\ \right\rangle, \tag{32}$$

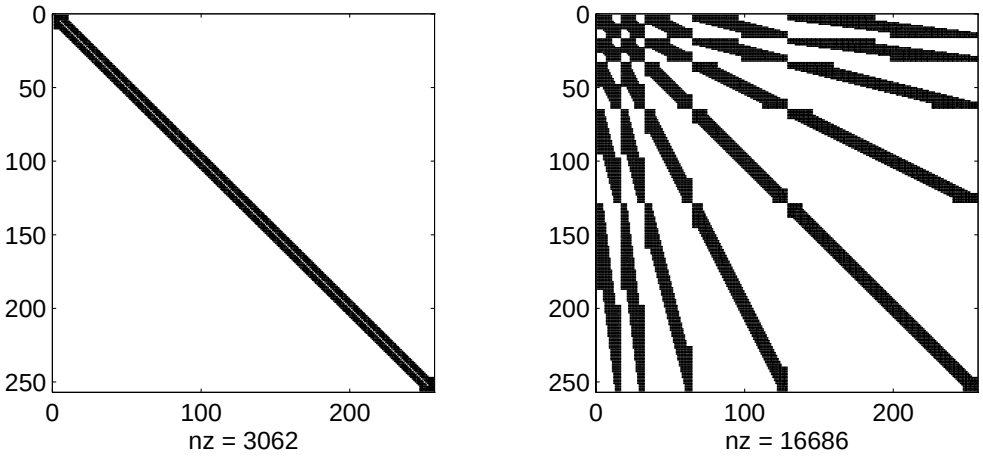

FIG. 3. *Shape of the first-order derivative operator in the basis of scaling functions (left) and in the wavelet basis (right) ($N = 4, j = 8, \Lambda = \emptyset$). We give the number (nz) of nonzero coefficients in each case.*

$$ {}_2B_{i,j} = \left\langle \tilde{\tilde{\Phi}}_i \mid \frac{d^2\Phi_j}{dx^2} \right\rangle. \tag{33} $$

Remark that integrating twice by parts we get ${}_2B_{i,j} = \langle\, \Phi_j \mid d^2\tilde{\tilde{\Phi}}_i/dx^2 \,\rangle$. Similarly, as for Proposition 6.1, we derive the following proposition.

PROPOSITION 6.2. *The matrices ${}_2A$ and ${}_2B$ in* (32–33) *can be expressed:*

(i) *The matrix ${}_2B$ of size $(N - \#\Lambda) \times (2N - 2 - \alpha)$ is given by*

$$ {}_2B = (G_\Lambda^{\tilde{\Phi}})^{-1/2}\ {}_2B^{\tilde{\Phi}} $$

where

$$ \begin{aligned} {}_2B^{\tilde{\Phi}}_{i+1,j-N+\alpha+1} &\overset{def}{=} \left\langle \tilde{\Phi}_i \mid \frac{d^2\Phi_j}{dx^2} \right\rangle \\ &= \sum_{k=-N+1}^{N-1-\alpha} P_i(k)\ {}_2r_{k-j}, \end{aligned} \tag{34} $$

for $i = 0, \ldots, N-1, i \notin \Lambda$, and $N - \alpha \le j \le 3N - 3 - 2\alpha$.

(ii) *The matrix ${}_2A$ of size $(N - \#\Lambda) \times (N - \#\Lambda)$ is given by*

$$ {}_2A = (G_\Lambda^{\tilde{\Phi}})^{-1/2}\ {}_2A^{\tilde{\Phi}}\ {}^t(G_\Lambda^{\tilde{\Phi}})^{-1/2} $$

where $A^{\tilde{\Phi}}$ satisfies

$$ \frac{1}{2}\ ({}_2A^{\tilde{\Phi}}) = D_\Lambda\ ({}_2A^{\tilde{\Phi}})\ D_\Lambda - b_\Lambda\ ({}^t{}_2B^{\tilde{\Phi}})\ D_\Lambda + D_\Lambda\ ({}_2B^{\tilde{\Phi}})\ {}^t b_\Lambda + b_\Lambda\ {}_2r\ {}^t b_\Lambda \tag{35} $$

(${}_2r$ is the square matrix of size $2N - 1 - \alpha$ whose general term is ${}_2r_{i-j}$, D_Λ is the submatrix of D in (15) *containing only the rows and columns of numbers not in $1 + \Lambda$, b_Λ the submatrix of b in* (15) *containing only the rows of numbers not in $1 + \Lambda$, and $G_\Lambda^{\tilde{\Phi}}$ is defined in* (21)).

If $0 \notin \Lambda$ and $1 \notin \Lambda$, (35) allows us to compute all the coefficients of the matrix ${}_2A^{\tilde{\Phi}}$ except the terms ${}_2A^{\tilde{\Phi}}_{1,2}$ and ${}_2A^{\tilde{\Phi}}_{2,1}$. We must compute them directly by

$$\begin{aligned}
{}_2A^{\tilde{\Phi}}_{1,2} &\stackrel{def}{=} \int_0^{+\infty} \tilde{\Phi}_0(x)\ \frac{d^2\tilde{\Phi}_1}{dx^2}(x)\ dx \\
&= \int_0^{+\infty} (1 - \sum_{k=N-\alpha}^{+\infty} \Phi(x-k)) \frac{d^2\tilde{\Phi}_1}{dx^2}(x) dx \\
&= \left[\tilde{\Phi}'_1(x)\right]_0^{+\infty} - \sum_{k=N-\alpha}^{+\infty} \left({}_2B^{\tilde{\Phi}}_{\Lambda}\right)_{2,k-N+\alpha+1} \\
&= -1 - \sum_{k=N-\alpha}^{3N-3-\alpha} \left({}_2B^{\tilde{\Phi}}_{\Lambda}\right)_{2,k-N+\alpha+1},
\end{aligned}$$

and integrating by parts,

$$\begin{aligned}
{}_2A^{\tilde{\Phi}}_{2,1} &= \left[\tilde{\Phi}'_0(x)\tilde{\Phi}_1(x)\right]_0^{+\infty} - \int_0^{+\infty} \tilde{\Phi}'_0(x)\tilde{\Phi}'_1(x) dx \\
&= -\left[\tilde{\Phi}_0(x)\tilde{\Phi}'_1(x)\right]_0^{+\infty} + \int_0^{+\infty} \tilde{\Phi}_0(x) \frac{d^2\tilde{\Phi}_1}{dx^2}(x) dx \\
&= 1 + {}_2A^{\tilde{\Phi}}_{1,2}.
\end{aligned}$$

For the edge 1, things happen in the same way since

$$\left\langle\ Tf \mid \frac{d^2Tg}{dx^2}\ \right\rangle_{L^2([0,+\infty[)} = \left\langle\ f \mid \frac{d^2g}{dx^2}\ \right\rangle_{L^2(]-\infty,1])}.$$

7. Quadrature formulae. The problem discussed here is: Given a continuous function f on $[0,1]$, how to compute its projection on the space $V_j^{[0,1]}(\Lambda_0,\Lambda_1)$? Numerically, the problem is formulated slightly differently.

1. Given the collocation values $\{f(k/2^j)\}_{0\le k\le 2^j}$, how to approximate the inner products:

$$\begin{gathered}
\{\ \langle\ f \mid 2^{j/2}\tilde{\tilde{\Phi}}_\ell(2^j.)\ \rangle\ \}_{\ell\notin\Lambda_0}, \\
\{\ \langle\ f \mid 2^{j/2}\Phi_k(2^j.)\ \rangle\ \}_{N-\alpha_0\le k\le 2^j-1-N+\alpha_1}, \\
\{\ \langle\ f \mid 2^{j/2}\tilde{\tilde{\Phi}}_\ell(2^j(1-.))\ \rangle\ \}_{\ell\notin 2^j-1-\Lambda_1}?
\end{gathered}$$

This point addresses the problem of quadrature formulae, as presented by Sweldens and Piessens in [SwPi 94]; in the case of the real line, they estimate the accuracy of the approximation obtained by the quadrature formulae presented below. We will adapt their method to the interval for our construction of scaling functions (with and without boundary condition).

2. Given a function f in $V_j^{[0,+\infty[}(\Lambda_0,\Lambda_1)$ by its scalar products on the related basis, how to compute the grid values $\{f(k/2^j)\}_{0\le k\le 2^j}$? This point is concerned with the values of scaling functions at points $k/2^j$. We will see how to compute these values in our case.

Note that these two problems are not unisolvant since we have $(2^j - \#\Lambda_0 - \#\Lambda_1 + \alpha_0 + \alpha_1)$ scalar products whereas $2^j + 1$ values of f.

Thus, to go through to solve the point 1, we introduce an integer $q \geq N$, and, as explained in [SwPi 94] and also by Masson in [Mass 96], we consider the nodes for $0 \leq p \leq q$:

$$\begin{aligned} x_p &= -\lfloor q/2 \rfloor + p, \\ y_p &= p. \end{aligned} \tag{36}$$

From the grid values of f at these points, we deduce an approximation of:

- the scalar product $\langle\ f \mid 2^{j/2}\Phi_k(2^j.)\ \rangle$ thanks to the values of f at nodes $(k + x_p)/2^j$,
- the edge scalar product $\langle\ f \mid 2^{j/2}\tilde{\tilde{\Phi}}_\ell(2^j.)\ \rangle$ thanks to the values of f at nodes $y_p/2^j$ (for the edge 0) or $(1 - y_p)/2^j$ (for the edge 1).

The weight coefficients w_p are introduced such that

$$\int_{\mathbb{R}} P(x)\ \Phi(x)\ dx = \sum_{p=0}^{q} w_p\ P(x_p) \tag{37}$$

for all polynomial function P of degree $\leq q$. To find the weights w_p, it is sufficient to solve the linear system

$$\begin{pmatrix} \frac{x_0^0}{0!} & \cdots & \frac{x_q^0}{0!} \\ \vdots & & \vdots \\ \frac{x_0^q}{q!} & \cdots & \frac{x_q^q}{q!} \end{pmatrix} \begin{pmatrix} w_0 \\ \vdots \\ w_q \end{pmatrix} = \begin{pmatrix} C_0 \\ \vdots \\ C_q \end{pmatrix}$$

(the C_k are recursively defined in (11)). Then the inner product is approximated by

$$\langle\ f \mid 2^{j/2}\Phi_k(2^j.)\ \rangle \approx \sum_{p=0}^{q} w_p\ f((k + x_p)/2^j).$$

Remarks. 1. Since the quadrature formula is exact for polynomials up to degree q, the accuracy of the previous interpolation is of order $q+1$; that is, the error between f and its interpolation in V_j is of order $2^{-j(q+1)s}$ in L^2-norm for a s-regular function. To be consistent with the accuracy order of the wavelet projection in V_j, that is 2^{jNs} for Daubechies wavelets (see [Daub 92]), it is sufficient to take $q = N - 1$. But as experimented by Sweldens and Piessens in [SwPi 94], the choice $q = N - 1$ can ruin the approximation properties of the wavelet expansion. Therefore, it is preferable to choose $q \geq N$.

2. As the nodes x_p are equally spaced, we derive no more than a Lagrangian interpolation to compute each coefficient, and for high values of q, the system is ill conditioned. As stated in [SwPi 94], it is then advisable in this case to use Chebyshev nodes.

To take into account these two remarks, we will take q close to N in numerical experiments (see section 8.1).

For the edge 0, we proceed similarly to define the weights w_p^ℓ for $\ell \notin \Lambda_0$ by

$$\int_0^{+\infty} P(x)\tilde{\tilde{\Phi}}_\ell(x)dx = \sum_{p=0}^{q} w_p^\ell\ P(y_p),$$

for all polynomial function P of degree $\leq q$. The remarks stated above concerning the choice of q are also valid for the quadrature formula at the edges, since the approximation properties of wavelets on the interval are similar (see [CDV 93] and [Ausc 93]).

As above, we have to compute the moments of the edge scaling functions. For $0 \leq p \leq q$, define the column vector X_p (vector of edge moments) by

$$X_p = \int_0^{+\infty} \begin{pmatrix} \tilde{\tilde{\Phi}}_0(x) \\ \vdots \\ \tilde{\tilde{\Phi}}_{N-1}(x) \end{pmatrix}_{\ell \notin \Lambda_0} \frac{x^p}{p!} dx.$$

Then using (23) we get

$$X_p = \frac{1}{2^{p+1/2}} \left(H_0 X_p + h_0 \int_0^{+\infty} \begin{pmatrix} \Phi(x - N + \alpha_0) \\ \vdots \\ \Phi(x - (3N - 2 - 2\alpha_0)) \end{pmatrix} \frac{x^p}{p!} dx \right),$$

so that changing of variable in the integrals and using the binomial formula:

$$(2^{p+1/2} I - H_0) X_p = h_0 \begin{pmatrix} \frac{(N-\alpha_0)^p}{p!} & \cdots & \frac{(N-\alpha_0)^0}{0!} \\ \vdots & & \vdots \\ \frac{(3N-2-2\alpha_0)^p}{p!} & \cdots & \frac{(3N-2-2\alpha_0)^0}{0!} \end{pmatrix} \begin{pmatrix} C_0 \\ \vdots \\ C_p \end{pmatrix},$$

which determines exactly X_p since H_0 is similar to $1/\sqrt{2} D_{\Lambda_0}$, so that $2^{p+1/2}$ cannot be an eigenvalue of H_0. Then the w_p^ℓ are determined by the linear systems

$$\begin{pmatrix} w_0^0 & \cdots & w_q^0 \\ \vdots & & \vdots \\ w_0^{N-1} & \cdots & w_q^{N-1} \end{pmatrix}_{\ell \notin \Lambda_0} \begin{pmatrix} \frac{x_0^0}{0!} & \cdots & \frac{x_0^q}{q!} \\ \vdots & & \vdots \\ \frac{x_q^0}{0!} & \cdots & \frac{x_q^q}{q!} \end{pmatrix} = (\, X_0 \mid \cdots \mid X_q \,),$$

and the inner product is approximated by

$$\langle \, f \mid 2^{j/2} \tilde{\tilde{\Phi}}_\ell(2^j .) \, \rangle \;\approx\; \sum_{p=0}^{q} w_p^\ell \, f(p/2^j).$$

The same procedure should be done for the edge 1.

Let us now see how to deal with problem 2. It is clear that it is sufficient to know the values of the scaling functions at integer points. To compute this for interior scaling functions, see [DaMi 93]; using the two-scale equation, we have to solve an eigenvalue problem. For edge scaling functions, we proceed as follows: The nonzero integer values of the $\tilde{\Phi}_\ell$ are at points $0, 1, \ldots, 2N - 2 - \alpha_0$. We use Definition 3.1 to obtain the values $\tilde{\Phi}_\ell(k)$ for $\ell \notin \Lambda_0$ and $0 \leq k \leq 2N - 2 - \alpha_0$, then (22) for the values $\tilde{\tilde{\Phi}}_\ell(k)$.

Notice that all these values could have been approximately obtained via the *cascade algorithm* (see [Daub 92]).

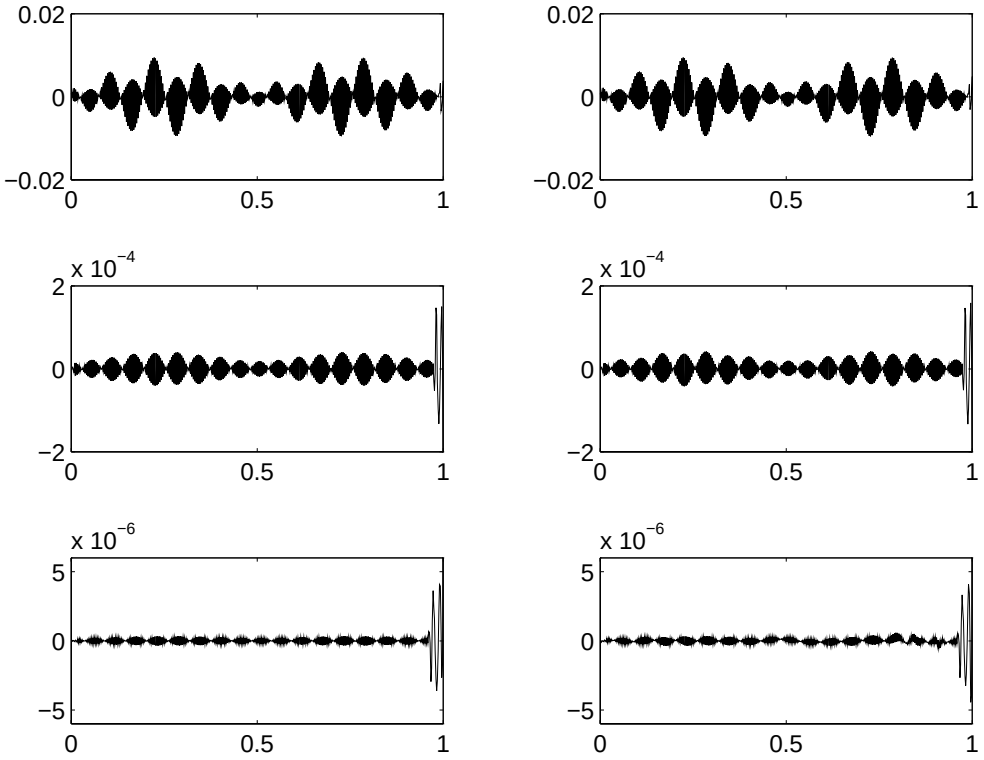

FIG. 4. *Interpolation error (left column) and projection error (right column) in space* V_8 *for* $f(x) = sin(2\pi x)sin(50x)$*, for* $N = q = 2$ *(first row),* $N = 4, q = 5$ *(second row), and* $N = 6, q = 8$ *(last row).*

8. Numerical results.

8.1. Interpolation and projection error. It becomes possible then to compute numerically the projection on $V_j^{[0,1]}(\Lambda_0, \Lambda_1)$ of a given function f.

To compare the interpolation error at the edges and in the interior, we draw the points:

$$\left(\frac{k}{2^p} \ , \ \left(f - \tilde{f}_j\right)\left(\frac{k}{2^p}\right) \right)_{0 \le k \le 2^p}$$

where $\tilde{f}_j$ is the interpolating function of f in $V_j^{[0,1]}(\Lambda_0, \Lambda_1)$, computed by the procedure described in section 7, from the values of f at node points $k/2^j$.

To compare this interpolation error to the projection error $\left(f - P_{V_j^{[0,1]}(\Lambda_0,\Lambda_1)} f\right)$, we also draw the points

$$\left(\frac{k}{2^p} \ , \ \left(f - P_{V_j^{[0,1]}(\Lambda_0,\Lambda_1)} \tilde{f}_{j+4}\right)\left(\frac{k}{2^p}\right) \right)_{0 \le k \le 2^p}.$$

In this case, the interpolation error to compute $\tilde{f}_{j+4}$ will be negligible with regard to the projection in $V_j^{[0,1]}(\Lambda_0, \Lambda_1)$.

Figure 4 represents these curves plotted for $(N = 2, q = 2)$, $(N = 4, q = 5)$, and $(N = 6, q = 8)$ ($q + 1$ is the number of nodes defined in (36) for the quadrature formulae), and $j = 8, p = 9$ and without boundary conditions, analyzing the projection of $f(x) = sin(2\pi x)sin(50x)$.

Experiments show that the interpolation procedure has the same order of accuracy as the projection error. In both cases, we can see that the error is not satisfactory at edge 1, comparatively with the interior. This difference between the error in the interior and the error at the edges grows rapidly with the order N of vanishing moments. However, numerical experiments show that the global error decreases as expected, both in the interior and at the edges, when j is growing. This (relatively) bad result at edge 1 can be interpreted in two manners.

The interpolation error at edge 1 comes from the fact that in the quadrature formula, the $N + 1$ equidistant nodes are localized near the edge. Recalling that the length of the support of the edge basis functions is $2^{-j}(2N - 1)$ at scale j, the nodes cover only half the support of the integrated function! Since usual Daubechies compactly supported wavelets present a strong asymmetry, they are more localized at the left of their support. If we had chosen the least asymmetric ones, the interpolation error would be shared by the two edges.

The projection error at the edges is due to the fact that we keep only N functions at each edge, whereas there are $2N - 1$ scaling functions of $\mathbb{R}$ that contain 0 in their support. The consequence is that there is a loss at the edges of the superconvergence phenomenon observed for the wavelet approximation on $\mathbb{R}$ [Go 95]. The fact that this problem is not visible at the edge 0 comes from the strong asymmetry of the chosen wavelets.

Note that the curves plotted in Figure 4 are very close to those obtained by Masson in [Mass 96], studying the projection error with biorthogonal wavelet bases on the interval.

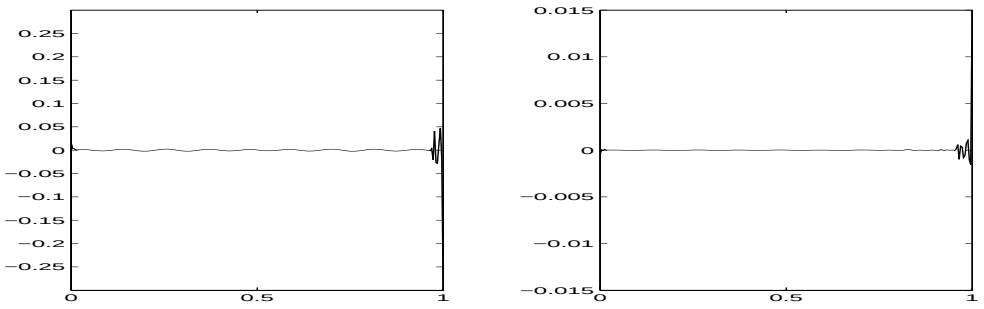

FIG. 5. *Derivative error for $N = 4$ (left column) and $N = 6$ (right column) of $f(x) = sin(2\pi x)sin(50x)$ ($j = 8$ scales).*

8.2. Derivative error. We draw on Figure 5 the derivative error for $N = 4$ and $N = 6$ vanishing moments, without boundary conditions for $f(x) = sin(2\pi x)sin(50x)$, between the exact derivative and the derivative obtained by the Galerkin procedure, as explained in section 6, for $j = 8$ scales. As previously, this derivative error being linked to the projection error, the accuracy is lost at the right edge.

REFERENCES

[AHJP 93] L. Anderson, N. Hall, B. Jawerth, and G. Peters, *Wavelets on closed subsets on the real line*, Topics in the Theory and Applications of Wavelets, L. L. Schumaker and G. Webb, eds., Academic Press, Boston, MA, 1993.

[Ausc 93] P. Auscher, *Ondelettes à support compact et conditions aux limites*, J. Funct. Anal., 111 (1993), pp. 29–43.

[BNR 94] S. Bertoluzza, G. Naldi, and J. C. Ravel, *Wavelet methods for the numerical solution of boundary value problems on the interval*, Wavelets: Theory, Algorithms, and Applications, C.K. Chui, L. Montefusco, and L. Puccio, eds., 1994, pp. 425–448.

[Beyl 92] A. G. Beylkin, *On the representation of operators in bases of compactly supported wavelets*, SIAM J. Numer. Anal., 29 (1992), pp. 1716–1740.

[ChPe 96] P. Charton and V. Perrier, *A pseudo-wavelet scheme for the two-dimensional Navier-Stokes equation*, Comp. Appl. Math., 15 (1996), pp. 139–160.

[ChLi 97] G. Chiavassa and J. Liandrat, *On the effective construction of compactly supported wavelets satisfying homogeneous boundary conditions on the interval*, Appl. Comp. Harmonic Analysis, 4 (1997), pp. 62–73.

[CDDe 95] A. Cohen, W. Dahmen, and R. DeVore, *Multiscale decompositions on bounded domains*, Trans. Amer. Math. Soc., 1997, to appear.

[CDV 93] A. Cohen, I. Daubechies, and P. Vial, *Wavelets on the interval and fast wavelet transforms*, Appl. Comp. Harmonic Analysis, 1 (1993), pp. 54–81.

[DaMi 93] W. Dahmen and C. A. Micchelli, *Using the refinement equation for evaluating integrals of wavelets*, SIAM J. Numer. Anal., 30 (1993), pp. 507–537.

[Daub 88] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 499–519.

[Daub 92] I. Daubechies, *Ten lectures on wavelets*, CBMS Lecture Notes, 61, SIAM, Philadelphia, PA, 1992.

[DVJP 92] R.-A. DeVore, B. Jawerth, and V. Popov, *Compression of wavelet decompositions*, Amer. J. Math., 114 (1992), pp. 737–785.

[Go 95] S. M. Gomes, *Convergence estimates for the Wavelet-Galerkin method: Superconvergence at the node points*, Adv. Comput. Math., 4 (1995), pp. 261–282.

[Jaff 92] S. Jaffard, *Wavelet methods for fast resolution of elliptic problems*, SIAM J. Numer. Anal., 29 (1992), pp. 965–987.

[JM 89] S. Jaffard and Y. Meyer, *Bases d'ondelettes dans des ouverts de* $\mathbb{R}^n$, J. Math. Pures Appl., 68 (1989), pp. 95–108.

[LPT 92] J. Liandrat, V. Perrier, and P. Tchamitchian, *Numerical resolution of nonlinear partial differential equations using the wavelet approach*, Wavelets and Their Applications, Ruskai et al., eds., Jones and Bartlett, Boston, MA, 1992, pp. 227–238.

[MPR 91] Y. Maday, V. Perrier, and J.-C. Ravel, *Adaptativité dynamique sur bases d'ondelettes pour l'approximation d'équations aux dérivées partielles*, C. R. Acad. Sci. Paris, t. 312, Série I, 1991, pp. 405–410.

[MaRa 92] Y. Maday and J.-C. Ravel, *Adaptativité dynamique par ondelettes : conditions aux limites et dimensions supérieures*, C. R. Acad. Sci. Paris, t. 315, Série I, 1992, pp. 85–90.

[Mall 89] S. Mallat, *Multiresolution approximation and wavelets*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–88.

[Mass 96] R. Masson, *Biorthogonal spline wavelets on the interval for the resolution of boundary problems*, M^3AS, 6 (1996), pp. 749–791.

[Meye 90] Y. Meyer, *Ondelettes et Opérateurs*, Hermann, Paris, 1990.

[Meye 92] Y. Meyer, *Ondelettes sur l'intervalle*, Rev. Mat. Iberoamericana, 7 (1992), pp. 115–133.

[MoPe 95] P. Monasse and V. Perrier, *Construction d'ondelettes sur l'intervalle pour la prise en compte de conditions aux limites*, C. R. Acad. Sci. Paris, t. 321, Série I, 1995, pp. 1163–1169.

[PeBa 89] V. Perrier and C. Basdevant, *Periodical wavelet analysis, a tool for inhomogeneous field investigation. Theory and algorithms*, Rech. Aérospat., n. 1989-3 (1989), pp. 53–67.

[PST 95] G. Plonka, K. Selig, and M. Tasche, *On the construction of wavelets on a bounded interval*, preprint.

[SwPi 94] W. Sweldens and R. Piessens, *Quadrature formulae and asymptotic error expansions for wavelet approximations of smooth functions*, SIAM J. Numer. Anal., 31 (1994), pp. 1240–1264.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1067–1086, September 1998

HOPF-TYPE ESTIMATES AND FORMULAS FOR NONCONVEX NONCONCAVE HAMILTON–JACOBI EQUATIONS*

MARTINO BARDI† AND SILVIA FAGGIAN‡

Abstract. Simple explicit estimates are presented for the viscosity solution of the Cauchy problem for the Hamilton–Jacobi equation where either the Hamiltonian or the initial data are the sum of a convex and a concave function. The estimates become equalities whenever a "minmax" equals a "maxmin" and thus a representation formula for the solution is obtained, generalizing the classical Hopf formulas as well as some formulas of Kružkov [*Functional Anal. Appl.*, 2 (1969), pp. 128–136].

Key words. Hamilton–Jacobi equations, viscosity solutions, Hopf's representation formulas, Lax–Oleinik formula, nonconvex Hamiltonians

AMS subject classifications. 35F25, 49L25, 35C99, 35L99

PII. S0036141096309629

1. Introduction.

We consider the initial value problem for the Hamilton–Jacobi equation

$$\begin{cases} (HJ) \quad u_t + H(D_x u) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ (IC) \quad u(x,0) = u_0(x) & \text{in } \mathbb{R}^N \times \{0\}, \end{cases}$$

where $H \in C(\mathbb{R}^N)$ is continuous and $u_0 : \mathbb{R}^N \to \mathbb{R}$ is uniformly continuous. This problem has a unique viscosity solution u in the space $UC_x(\mathbb{R}^N \times [0,T])$ of the continuous functions which are uniformly continuous in x uniformly in t, see [23] or [15] (for the general theory of viscosity solutions we refer to [12, 14, 5, 34, 2]).

We are interested in giving explicit pointwise upper and lower bounds for the solution, yielding in some cases a representation formula for u, and involving only a finite number of max-min operations over finite dimensional sets of parameters. The prototypes of this sort of results are the first Hopf formula (sometimes called Lax formula)

$$\begin{aligned} u(x,t) :&= \inf_{z\in\mathbb{R}^N} \sup_{y\in\mathbb{R}^N} \{u_0(z) + y\cdot(x-z) - tH(y)\} \\ &= \min_{z\in\mathbb{R}^N} \left\{u_0(z) + tH^*\left(\tfrac{x-z}{t}\right)\right\} \\ &= \min_{k\in\mathbb{R}^N} \{u_0(x-tk) + tH^*(k)\} \end{aligned} \tag{1.1}$$

holding for convex (or concave) Hamiltonian and Lipschitz initial data [22, 30, 19, 3, 18], or for Lipschitz and convex H and continuous u_0 [34, section 10.1], or also for strictly convex Hamiltonian and *lower semicontinuous (l.s.c.)* u_0 [26, 27], and the second Hopf formula

$$\begin{aligned} u(x,t) :&= \sup_{y\in\mathbb{R}^N} \inf_{z\in\mathbb{R}^N} \{u_0(z) + y\cdot(x-z) - tH(y)\} \\ &= \max_{y\in\mathbb{R}^N} \{x\cdot y - u_0^*(y) - tH(y)\} \end{aligned} \tag{1.2}$$

*Received by the editors September 23, 1996; accepted for publication (in revised form) June 2, 1997; published electronically April 14, 1998.
http://www.siam.org/journals/sima/29-5/30962.html

†Dipartimento di Matematica Pura ed Applicata, Università di Padova, via Belzoni, 7, I-35131 Padova, Italy (bardi@math.unipd.it). The research of this author was partially supported by MURST project "Problemi nonlineari nell'analisi e nelle applicazioni fisiche, chimiche, biologiche."

‡Dipartimento di Matematica, Università di Padova, via Belzoni, 7, I-35131, Padova, Italy (faggian@dm.unipi.it).

in the case of Lipschitz and convex (or concave) initial data and merely continuous Hamiltonian [22, 3, 31], or for convex u_0 and Lipschitz H [34, section 10.1]. Here H^* and u_0^* denote, respectively, the Legendre transforms (convex conjugates) of H and u_0, and in the following we denote with F^* also the Legendre transform of a concave function F, namely,

$$F^*(y) = -(-F)^*(y) = \inf_{x\in\mathbb{R}^N}\{-x\cdot y - F(x)\}.$$

Let us mention that both formulas were recently extended to the case of quasiconvex data by means of a suitable notion of quasiconvex conjugate function [8, 9, 10].

Our main results are the following extensions of the Hopf formulas.

THEOREM A. *If $H(p) = H_1(p) + H_2(p)$ with H_1 convex, H_2 concave, and u_0 is uniformly continuous, then the unique viscosity solution $u \in UC_x(\mathbb{R}^N \times [0,T])$ of $(HJ)(IC)$ satisfies*

$$\max_{w\in\mathbb{R}^N}\min_{k\in\mathbb{R}^N} g(k,w,x,t) \le u(x,t) \le \min_{k\in\mathbb{R}^N}\max_{w\in\mathbb{R}^N} g(k,w,x,t),$$

where $g(k,w,x,t) := u_0(x - t(k-w)) + tH_1^*(k) + tH_2^*(w)$.

THEOREM B. *If $u_0 = u_1 + u_2$, with u_1 convex and Lipschitz continuous, u_2 concave and Lipschitz continuous, and $H \in C(\mathbb{R}^N)$, then the unique viscosity solution $u \in UC_x(\mathbb{R}^N \times [0,T])$ of $(HJ)(IC)$ satisfies*

$$\max_{k\in\mathbb{R}^N}\min_{w\in\mathbb{R}^N} f(k,w,x,t) \le u(x,t) \le \min_{w\in\mathbb{R}^N}\max_{k\in\mathbb{R}^N} f(k,w,x,t),$$

where $f(k,w,x,t) := x\cdot(k+w) - u_1^*(k) - u_2^*(-w) - tH(k+w)$.

The main features of the estimates of both theorems are that

(i) all the inequalities are equalities at the initial time $t = 0$,

(ii) the lower and upper bound on u are, respectively, the max-min and the min-max of the same family of explicit functions, so they might be shown to coincide in many cases.

These properties suggest that one should be able to derive a representation formula for u at least for short times, by applying some min-max theorem. Indeed, we give results of this kind under some additional assumptions on the data.

Also note that the assumption that H be the sum of a convex and a concave function, is a rather mild regularity property which is satisfied, for instance, by semiconcave or semiconvex functions, e.g., by functions with second derivatives bounded either from above or below.

Theorem A contains as a special case the extension of the first Hopf formula (1.1) to the case of uniformly continuous initial data. Moreover, the representation formula we obtain for short times coincides with a formula of Kružkov [28] for the short time classical solution of $(HJ)(IC)$ in the case of smooth data.

On the other hand, Theorem B contains the second Hopf formula (1.2), as well as the estimates of Bardi and Osher [4] for the case of initial data sum of a convex function in a group of variables and a concave function in the remaining variables. Indeed we use their result to prove Theorem B. A similar pair of inequalities for some special Hamiltonians is the intermediate step in the proof of Theorem A. We borrow from Kružkov [28] an idea of "doubling the variables" to complete the proofs once this first step is established.

If the Hamiltonian depends explicitly on the variable x, i.e., $H = H(x, D_x u)$, then the solution u of $(HJ)(IC)$ can be represented as the value function of a differential

game, that is, as an inf-sup over infinite dimensional sets of controls and strategies, see [20, 25]. Here we show that for some Hamiltonians depending on x in a rather special way one can derive a much simpler representation formula by using Theorem A or B after a change of variables.

The classical Hopf formulas are important tools for the study of several properties of solutions of Hamilton–Jacobi equations, such as the asymptotic behavior as time goes to infinity [30, 6], the connection with geometric solutions in the sense of symplectic mechanics [11], uniqueness [7], and they have been used for numerical schemes [4, 33, 16, 17]. Moreover, from any formula for a viscosity solution of (HJ) one derives a representation formula for the entropy weak solution of a scalar conservation law, such as the Lax–Oleinik formula if one starts from the first Hopf formula, see, e.g., [13, 32, 18]. The results of this paper have similar applications that will be studied elsewhere.

The paper is organized as follows. In section 2 we extend the first Hopf formula to the case of uniformly continuous initial data. In section 3 we prove Theorem A and some related results. Section 4 is devoted to the proof of Theorem B. In section 5 we describe some cases where Theorems A and B give representation formulas for the solution, as well as some examples where the inequalities are strict. Finally, in section 6 we give an extension to equations with Hamiltonian depending on x.

2. First Hopf formula for uniformly continuous initial data. In this section, we revisit the first Hopf formula (1.1) for convex Hamiltonian H under the relaxed assumption $u_0 \in UC(\mathbb{R}^N)$ on the initial data. The main result of the section is in the following.

THEOREM 2.1. *If $H : \mathbb{R}^N \to \mathbb{R}$ is convex and $u_0 : \mathbb{R}^N \to \mathbb{R}$ uniformly continuous, then*

$$u(x,t) := \begin{cases} \min\limits_{z\in\mathbb{R}^N} \left\{ u_0(z) + tH^*\left(\dfrac{x-z}{t}\right)\right\} & \text{in } \mathbb{R}^N \times (0,+\infty), \\ u_0(x) & \text{in } \mathbb{R}^N \times \{0\} \end{cases} \tag{2.1}$$

gives the unique viscosity solution of $(HJ)(IC)$ in the space $UC_x(\mathbb{R}^N \times [0,+\infty))$ of continuous functions of (x,t), uniformly continuous in x, uniformly in t.

The proof of Theorem 2.1 is only slightly different from its previous editions, such as in [30, 3, 18], so we omit it. We check only that (2.1) defines a function of $UC_x(\mathbb{R}^N \times [0,+\infty))$, in particular, that $u(x,t)$ is continuous in t because the proof of this part is different. Note that in [18, 3] the authors proved that u_0 Lipschitzean implies that u defined by (2.1) is itself Lipschitzean in both x and t, which is not necessarily true in our case.

Let us observe that the min in (2.1) is attained because H^* is *l.s.c.* and superlinear, i.e., $\lim_{|p|\to+\infty} H^*(p)/|p| = +\infty$.

As done in [18], it can be shown that for any fixed $x \in \mathbb{R}^N$, and $t > 0$

$$u(x,t) = \min_{y\in\mathbb{R}^N} \left\{ u(y,s) + (t-s)H^*\left(\frac{x-y}{t-s}\right)\right\} \tag{2.2}$$

for any fixed $s \in [0,t]$, and this crucial relation is used throughout the proof of the theorem. We need the following lemma.

LEMMA 2.2. *Let $x \in \mathbb{R}^N,\ t \geq 0$. If $z^* := z^*(x,t) \in \mathbb{R}^N$ is such that*

$$u(x,t) = u_0(z^*) + tH^*\left(\frac{x-z^*}{t}\right),$$

then

$$\lim_{t\to 0+} z^*(x,t) = x. \tag{2.3}$$

Moreover, for any fixed $h \in [0,t]$, *let* $y^* := y^*(x,t,h)$ *be such that*

$$u(x,t) = u(y^*, t-h) + hH^*\left(\frac{x-y}{h}\right).$$

Then,

$$\lim_{h\to 0+} y^*(x,t,h) = x. \tag{2.4}$$

Proof. Fix any $\delta > 0, 0 < t < \delta$, and $b \in dom(H^*)$. Since $u(x,t) \le tH^*(b) + u_0(x-tb)$, we have

$$tH^*\left(\frac{x-z^*}{t}\right) \le u_0(x-tb) - u_0(z^*) + tH^*(b).$$

It is easy to check that H^* is superlinear, that is, $|v| = H^*(v)o(1)$, as $|v| \to +\infty$. Then

$$|x-z^*| \le o(1)(u_0(x-tb) - u_0(z^*) + tH^*(b)),$$

for $|x-z^*|/t \to +\infty$. Now observe that $u_0 \in UC(\mathbb{R}^N)$ implies that for all $z \in \mathbb{R}^N$, $|u(z)| \le C(1+|z|)$ for some constant $C > 0$. Hence, $\exists C_1 > 0$ such that

$$|x-z^*| \le o(1)(C|z^*| + C_1), \tag{2.5}$$

and then $\exists C_2 > 0$ such that

$$|z^*|(1 + Co(1)) \le C_2 + o(1) \tag{2.6}$$

for $|x-z^*|/t \to +\infty$. Let t_n be any sequence such that $t_n \to 0+$, as $n \to \infty$. We claim that $z_n^* := z_n^*(x,t_n)$ is bounded. In fact, if this is false, then we can find a subsequence t_{n_k} such that $|z_{n_k}^*| \to +\infty$, as $k \to \infty$. In this case $|x - z_{n_k}^*|/t_{n_k} \to +\infty$, as $k \to \infty$, then plugging $z_{n_k}^*$ into (2.6) and passing to limits as $k \to \infty$, we derive a contradiction.

In order to prove that (2.3) holds, we assume by contradiction that t_n is a sequence such that $t_n \to 0+$, as $n \to \infty$, and $|z_n^* - x| \ge \alpha > 0$ for all $n \in \mathbb{N}$ and for some $\alpha > 0$. Again, $|x - z_n^*|/t_n \to +\infty$, as $n \to \infty$, and by passing to limits in (2.5) we get a contradiction.

The second assertion can be proved in a similar way using (2.2). ∎

PROPOSITION 2.3. *In the hypotheses of Theorem* 2.1, $u(x,t)$, *defined by* (2.1), *is a continuous function of* (x,t); *it is also uniformly continuous in* x, *uniformly in* t.

Proof. It can be easily shown that $|u(x,t) - u(\bar{x},t)| \le \omega_0(|x-\bar{x}|)$ for any x, $\bar{x} \in \mathbb{R}^N$, $t > 0$, where ω_0 is a modulus of continuity for u_0 (see [18, 3]).

Then, we observe that

$$|u(y,t) - u_0(x)| \le \omega_0(|x-y|) + |u(x,t) - u_0(x)|$$

so it's enough to show that $|u(x,t) - u_0(x)| \to 0$, as $t \to 0+$. Pick $z^* = z^*(x,t)$ as in (2.2), and note that $H^*(v) \geq -H(0),\ \forall v \in \mathbb{R}^N$. Hence,

$$u_0(x) - u(x,t) \leq \omega_0(|x - z^*|) + |H(0)|t,$$

and the right-hand side tends to 0 as t tends to $0+$. Also, for any fixed $b \in dom(H^*)$, we have

$$\begin{aligned} u(x,t) - u_0(x) &\leq u_0(x - tb) + tH^*(b) - u_0(x) \\ &\leq \omega_0(|b|t) + |H^*(b)|t \end{aligned}$$

and again the right-hand side goes to 0, as $t \to 0+$.

We see last that $t \mapsto u(x,t)$ is continuous, for any fixed $x \in \mathbb{R}^N$. Suppose $\bar{t} < t$. From (2.2) we derive

$$u(x,t) \leq (t - \bar{t})H^*\left(\frac{x-y}{t-\bar{t}}\right) + u(y,t),$$

where y can be chosen in such a way that $q := x - y/(t - \bar{t}) \in dom(H^*)$. Then,

$$u(x,t) - u(x,\bar{t}) \leq \omega_0(|q||t - \bar{t}|) + |H^*(q)||t - \bar{t}|.$$

Now we choose $y^* = y^*(x,t,t-\bar{t})$ as in (2.4). Then,

$$u(x,\bar{t}) - u(x,t) \leq \omega_0(|y^* - x|) + |H(0)||t - \bar{t}|,$$

and the proof is complete by Lemma 2.2. □

Remark. It was shown by other authors that the first Hopf formula (2.1) gives the viscosity solution of $(HJ)(IC)$ under weaker assumptions on the initial data and stronger hypotheses on the Hamiltonian. Subbotin [34, Proposition 10.3] proved that if H is Lipschitz continuous and u_0 is merely continuous, then (2.1) gives the unique minimax solution, and therefore viscosity solution, of the problem. Kružkov [26, 27] proved that in the case of *l.s.c.* u_0 with at most linear growth at infinity and H of class C^2 with positive least eigenvalue of the Hessian matrix, (2.1) gives the unique generalized solution of $(HJ)(IC)$ in his sense, which is equivalent to the viscosity sense.

3. Estimates for the solution for Hamiltonians sum of a convex and a concave function. Now we relax the convexity hypothesis on H, replace it with

$$(H1) \qquad \begin{cases} H_1 : \mathbb{R}^N \to \mathbb{R} & \text{convex,} \\ H_2 : \mathbb{R}^N \to \mathbb{R} & \text{concave,} \\ H(p) = H_1(p) + H_2(p), \end{cases}$$

still assume $u_0 : \mathbb{R}^N \to \mathbb{R}$ uniformly continuous, and under these assumptions we prove sharp estimates for the solution of $(HJ)(IC)$. To this end, set $g : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \times [0,+\infty) \to \mathbb{R}$

$$g(k,w,x,t) := u_0(x - t(k-w)) + tH_1^*(k) + tH_2^*(w), \tag{3.1}$$

where H_1^*, H_2^* are, respectively, the Legendre transforms of H_1, H_2. Set also

$$u_-(x,t) := \sup_{w \in \mathbb{R}^N} \inf_{k \in \mathbb{R}^N} g(k,w,x,t), \tag{3.2}$$

$$u_+(x,t) := \inf_{k\in\mathbb{R}^N} \sup_{w\in\mathbb{R}^N} g(k,w,x,t). \tag{3.3}$$

THEOROM A. *Assume $u_0 : \mathbb{R}^N \to \mathbb{R}$ uniformly continuous and $(H1)$. Then the unique viscosity solution $u \in UC_x(\mathbb{R}^N \times [0,T])$ of $(HJ)(IC)$ satisfies*

$$u_-(x,t) \le u(x,t) \le u_+(x,t) \tag{3.4}$$

for all $(x,t) \in \mathbb{R}^N \times [0,T]$.

Remark. Note that for $t = 0$ the inequalities in (3.4) are equalities and $u_-(x,0) = u(x,0) = u_+(x,0) = u_0(x)$. Note further that if $H_2 = 0$, then

$$H_2^*(p) = \begin{cases} -\infty & \text{if } p \neq 0, \\ 0 & \text{if } p = 0, \end{cases}$$

and a simple computation shows that u_-, u_+ (and then u itself) are both given by the Hopf formula for convex Hamiltonian. Similarly, if $H_1 = 0$ we get the Hopf formula for concave Hamiltonian.

The proof of Theorem A consists essentially of two steps. First, we follow an idea of Kružkov [28], and show in Lemma 3.1 that the solution u of $(HJ)(IC)$ can be related to the solution of the following Cauchy problem in doubled state variables

$$\begin{cases} U_t + H_1(D_yU) + H_2(D_zU) = 0 & \text{in } \mathbb{R}^{2N} \times (0,T) \quad (HJ'), \\ U(y,z,0) = u_0(y+z) & \text{in } \mathbb{R}^{2N} \times \{0\} \quad (IC'), \end{cases}$$

where U is the unknown, and D_yU (respectively, D_zU) denotes the gradient of U with respect to the first (resp., the last) N space variables. Next, we prove an estimate like (3.4) in the case of H_1 depending only on the first j variables and H_2 depending only on the last $N-j$ (forthcoming Proposition 3.2). Finally, we apply this result to (HJ') (IC') and derive (3.4) from the related estimate on U.

LEMMA 3.1. *The function $u(x,t)$ is a viscosity solution of $(HJ)(IC)$ if and only if $U(y,z,t) := u(y+z,t)$ is a viscosity solution of $(HJ')(IC')$.*

Proof. Let (x_0,t_0) be fixed in $\mathbb{R}^N \times (0,T)$, and $y_0, z_0 \in \mathbb{R}^N$ such that $y_0 + z_0 = x_0$. Trivially, u satisfies (IC) if and only if U satisfies (IC'). First we show that if $U(y,z,t) = u(y+z,t)$ is a viscosity solution of (HJ') then $u(x,t)$ is a viscosity solution of (HJ). Consider a test function $\psi \in C^1(\mathbb{R}^N \times (0,T))$ such that $u - \psi$ has a local maximum at (x_0,t_0), and set $\varphi(y,z,t) := \psi(y+z,t)$. Then,

$$(U-\varphi)(y,z,t) \le (U-\varphi)(y_0,z_0,t_0)$$

for every (y,z,t) sufficiently close to (y_0,z_0,t_0). Hence, since U is a subsolution of (HJ'), we have

$$\varphi_t + H_1(D_y\varphi) + H_2(D_z\varphi) \le 0 \quad \text{at} \quad (y_0,z_0,t_0),$$

and by replacing φ_t, $D_y\varphi$, $D_z\varphi$ with the corresponding expressions involving ψ, we deduce

$$\psi_t + H(D_x\psi) \le 0 \quad \text{at} \quad (x_0,t_0).$$

Since (x_0,t_0) was arbitrary, we have proved that u is a subsolution of (HJ). By similar reasoning, we also obtain that u is a supersolution of (HJ).

Next we show that if u is a viscosity solution of (HJ), then $U(y,z,t) = u(y+z,t)$ is a viscosity solution of (HJ'). Let $(y_0,z_0,t_0) \in \mathbb{R}^N \times \mathbb{R}^N \times (0,T)$ and $\varphi \in C^1(\mathbb{R}^N \times \mathbb{R}^N \times (0,T))$ such that $U-\varphi$ has a local maximum at (y_0,z_0,t_0). Adding if necessary a constant to φ, we may assume

$$(U-\varphi)(y,z,t) \le (U-\varphi)(y_0,z_0,t_0) = 0 \tag{3.5}$$

for all (y,z,t) near (y_0,z_0,t_0). Thus, by fixing $z = z_0$ and using the definition of U, we get

$$u(y+z_0,t) - \varphi(y,z_0,t) \le u(y_0+z_0,t_0) - \varphi(y_0,z_0,t_0) \tag{3.6}$$

for all (y,t) close to (y_0,t_0). Since $\varphi(\cdot,z_0,\cdot) \in C^1(\mathbb{R}^N \times (0,T))$ and $(y,t) \mapsto u(y+z_0,t)$ is itself a viscosity solution of (HJ) (easy proof), (3.6) establishes

$$\varphi_t + H_1(D_y\varphi) + H_2(D_z\varphi) \le 0 \quad \text{at} \quad (y_0,z_0,t_0).$$

To demonstrate that U is a subsolution of (HJ') at (y_0,z_0,t_0) it is enough to show that

$$D_y\varphi(y_0,z_0,t_0) = D_z\varphi(y_0,z_0,t_0).$$

Set $z := z_0 - se_i$ and $y := y_0 + se_i$, where $s \in \mathbb{R}$, note that (3.5) holds for $|s|$ small enough and gives

$$\varphi(y_0,z_0,t_0) - \varphi(y_0+se_i, z_0-se_i, t_0) \le 0.$$

We divide both sides by $s > 0$ ($s < 0$), and by passing to limits as $s \to 0+$ (resp., as $s \to 0-$) we obtain $\partial\varphi/\partial y_i \le \partial\varphi/\partial z_i$ (resp., $\partial\varphi/\partial y_i \ge \partial\varphi/\partial z_i$) at (y_0,z_0,t_0), and then $D_y\varphi = D_z\varphi$ at (y_0,z_0,t_0).

In a similar way we can show that U is a supersolution of (HJ') at an arbitrarily fixed (y_0,z_0,t_0), and thus the proof is complete. □

From now on, we write $p = (p_A, p_B)$ and mean $p_A \in \mathbb{R}^j,\ p_B \in \mathbb{R}^{N-j}$.

PROPOSITION 3.2. *Assume $u_0 \in UC(\mathbb{R}^N)$, $H_1 : \mathbb{R}^N \to \mathbb{R}$ convex, $H_2 : \mathbb{R}^N \to \mathbb{R}$ concave, and $H(p) = H_1(p_A) + H_2(p_B)$. Then the unique viscosity solution $u \in UC_x(\mathbb{R}^N \times [0,T])$ satisfies for all $(x,t) \in \mathbb{R}^N \times [0,T]$*

$$\sup_{z_B \in \mathbb{R}^{N-j}} \inf_{z_A \in \mathbb{R}^j} G(z,x,t) \le u(x,t) \le \inf_{z_A \in \mathbb{R}^j} \sup_{z_B \in \mathbb{R}^{N-j}} G(z,x,t), \tag{3.7}$$

where $G : \mathbb{R}^N \times \mathbb{R}^N \times [0,+\infty) \to \mathbb{R}$ is either defined by

$$G(z,x,t) = u_0(x_A - tz_A, x_B + tz_B) + tH_1^*(z_A) + tH_2^*(z_B) \tag{3.8}$$

or, for $t \ne 0$, by

$$G(z,x,t) = u_0(z) + tH_1^*\left(\frac{x_A - z_A}{t}\right) + tH_2^*\left(\frac{z_B - x_B}{t}\right). \tag{3.9}$$

Proof. We prove (3.7) for G given by (3.8): the other case can be immediately derived from the change of variables $w = x - tz$. Since H_2 can be written as the Legendre transform of H_2^*, we have

$$H(p) \le H_1(p_A) - z_B \cdot p_B - H_2^*(z_B) =: \mathcal{H}(z_B,p)$$

$\forall z_B \in \mathbb{R}^{N-j}$. Let $z_B \in dom(H_2^*)$ be fixed; then $H_2^*(z_B) > -\infty$. The Legendre transform of the convex function $p \mapsto \mathcal{H}(z_B, p)$ is then given by

$$\mathcal{H}^*(z_B, p) = \begin{cases} H_2^*(z_B) + H_1^*(p_A), & \text{if } p_B = -z_B, \\ +\infty, & \text{if } p_B \neq -z_B. \end{cases}$$

Let's now consider the following Cauchy problem:

$$\begin{cases} \psi_t + \mathcal{H}(z_B, D_x\psi) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ \psi(z_B, x, 0) = u_0(x) & \text{in } \mathbb{R}^N \times \{0\}. \end{cases}$$

Its unique viscosity solution is given by the first Hopf formula that is

$$\begin{aligned} \psi(z_B, x, t) &= \min_{z_A \in \mathbb{R}^j} \{u_0(x_A - tz_A, x_B + tz_B) + tH_2^*(z_B) + tH_1^*(z_A)\} \\ &= \min_{z_A \in \mathbb{R}^j} G(z, x, t) \\ &= \min_{z_A \in dom(H_1^*)} G(z, x, t). \end{aligned} \tag{3.10}$$

Furthermore, $\psi(z_B, x, t)$ is a subsolution of $(HJ)(IC)$, since

$$\varphi_t + H(D_x\varphi) \leq \varphi_t + \mathcal{H}(z_B, D_x\varphi) \leq 0, \quad \text{at } (x,t)$$

for any $x \in \mathbb{R}^N$ and $t > 0$, provided that $\varphi \in C^1(\mathbb{R}^N \times (0,T))$ is a test function such that $\psi(z_B, \cdot, \cdot) - \varphi(\cdot, \cdot)$ has a local maximum at (x,t). Hence, a standard comparison theorem for unbounded viscosity solutions (see, e.g., [15]) gives

$$\psi(z_B, x, t) \leq u(x,t) \quad \forall z_B \in dom(H_2^*).$$

By taking the supremum for $z_B \in dom(H_2^*)$ we deduce

$$\sup_{z_B \in dom(H_2^*)} \psi(z_B, x, t) = \sup_{z_B \in dom(H_2^*)} \min_{z_A \in dom(H_1^*)} G(z, x, t) \leq u(x,t)$$

that is the first inequality in (3.7), for we note that

$$\sup_{z_B \in dom(H_2^*)} \min_{z_A \in dom(H_1^*)} G(z, x, t) = \sup_{z_B \in \mathbb{R}^{N-j}} \min_{z_A \in \mathbb{R}^j} G(z, x, t).$$

The second inequality in (3.7) can be derived by similar reasoning. We apply the first Hopf formula to compute the unique viscosity solution of

$$\begin{cases} \xi_t + \tilde{\mathcal{H}}(z_A, D_x\xi) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ \xi(z_A, x, 0) = u_0(x) & \text{in } \mathbb{R}^N \times \{0\}, \end{cases}$$

where

$$\tilde{\mathcal{H}}(z_A, p) := z_A \cdot p_A - H_1^*(z_A) + H_2(p_B) \leq H(p)$$

is a concave function of p, $\tilde{\mathcal{H}} \not\equiv -\infty$, for all fixed $z_A \in dom(H_1^*)$. Similarly, we get

$$\max_{z_B \in \mathbb{R}^{N-j}} G(z, x, t) = \xi(z_A, x, t) \geq u(x,t)$$

by means of the same comparison theorem, and draw the conclusion by taking the infimum for $z_A \in dom(H_1^*)$. □

Proof of Theorem A. We observe that the initial data $U_0(y,z) = u_0(y+z)$ and the Hamiltonian $h(p,q) = H_1(p) + H_2(q)$ of the problem $(HJ')(IC')$ satisfy the assumptions of Proposition 3.2, and then the unique viscosity solution $U(y,z,t)$ of $(HJ')(IC')$ in $UC_{y,z}(\mathbb{R}^N \times \mathbb{R}^N \times \times[0,T])$ verifies

$$U_-(y,z,t) \le U(y,z,t) \le U_+(y,z,t), \tag{3.11}$$

where

$$\begin{aligned} U_-(y,z,t) &= \sup_{w\in\mathbb{R}^N} \inf_{k\in\mathbb{R}^N} \{u_0(y+z+t(w-k)) + tH_1^*(k) + tH_2^*(w)\}, \\ U_+(y,z,t) &= \inf_{k\in\mathbb{R}^N} \sup_{w\in\mathbb{R}^N} \{u_0(y+z+t(w-k)) + tH_1^*(k) + tH_2^*(w)\}. \end{aligned}$$

Since, by Lemma 3.1, $u(y+z,t)$ is a solution of $(HJ')(IC')$, then it coincides with $U(y,z,t)$. Now we fix $z_0 \in \mathbb{R}^N$. Then, for every $x \in \mathbb{R}^N$, (3.11) holds, in particular, for $z = z_0$ and $y = x - z_0$, which gives exactly (3.4). □

Remark 1. Theorem A has an interpretation within the theory of differential games. Consider the controlled dynamical system

$$y'(s) = b(s) - a(s), \quad y(0) = x,$$

where the controls a and b belong to $\mathcal{L} = L^1([0,t],\mathbb{R}^N)$ and the cost functional

$$J(a,b,x,t) := \int_0^t \left(H_1^*\left(a(s)\right) + H_2^*\left(b(s)\right)\right) ds + u_0(y(t)),$$

which the controller of a seeks to minimize while that of b wants to maximize. Consider the set of nonanticipating strategies $\mathcal{S}$, i.e., the maps $\alpha : \mathcal{L} \to \mathcal{L}$ such that, for all $\tau > 0$, $b(s) = \tilde{b}(s)$ for all $s \le \tau$ implies $\alpha[b](s) = \alpha[\tilde{b}](s)$ for all $s \le \tau$. The value of this differential game is

$$v(x,t) := \inf_{\alpha\in\mathcal{S}} \sup_{b\in\mathcal{L}} J(\alpha[b],b,x,t) = \sup_{\beta\in\mathcal{S}} \inf_{a\in\mathcal{L}} J(a,\beta[a],x,t),$$

provided the last equality holds, and the PDE

$$u_t + H_1(Du) + H_2(Du) = 0 \ \text{ on } \ \mathbb{R}^N \times (0,\infty) \tag{3.12}$$

is the Isaacs equation associated to the game by the dynamic programming method. By the results of [20, 25] the value function v exists and it is the unique solution of (3.12) such that $u(x,0) = u_0(x)$. On the other hand, if the choice of both controllers is restricted to *constant* controls $a(\cdot) \equiv k$, $b(\cdot) \equiv w$, we have the so–called *game with simple motions*, a static game where the decision variables k and w are chosen in $\mathbb{R}^N$. In this case

$$J(a,b,x,t) = g(k,w,x,t),$$

so $u_-(x,t)$ and $u_+(x,t)$ are, respectively, the lower and the upper value of this game. Therefore, Theorem A states that the value $v(x,t)$ of the differential game is between the lower and the upper value of the game with simple motions.

In the case of a single controller, i.e., either H_1 or H_2 is null, the first Hopf formula states that constant controls generate the same value function as integrable controls (see [3]). In general, this is no longer true for two controllers, as we will see on some examples in section 5.

Remark 2. Observe that all 'inf' and 'sup' in (3.2), (3.3), (3.7) can be computed, respectively, on $dom(H_1^*)$, $dom(H_2^*)$ (see proof of Proposition 3.2), and that they are actually attained. For instance,

$$\sup_{z_B} \inf_{z_A} G(z,x,t) = \max_{z_B} \min_{z_A} G(z,x,t). \tag{3.13}$$

In fact, directly from the first Hopf formula we get that 'inf'='min,' and since the inf-envelope in (3.13) defines an *upper semicontinuous (u.s.c.)* function of z_B, and $\lim_{|p|\to+\infty} H_2^*(p)/|p| = -\infty$, we have also 'sup'='max.'

Remark 3. A simple computation shows that if H_1 (resp., H_2) depends only on the first j variables (resp., the last $N-j$ variables) then $dom(H_1) \subset \{(p_A,0) \in \mathbb{R}^N : p_A \in \mathbb{R}^j\}$ (resp., $dom(H_2) \subset \{(0,p_B) \in \mathbb{R}^N : p_B \in \mathbb{R}^{N-j}\}$). Thus the inequalities in (3.7), which we use indeed to demonstrate Theorem A, are special cases of (3.4).

Remark 4. The functions u_-, u_+ defined in Theorem A are, respectively, a sub and a supersolution of $(HJ)(IC)$ in the generalized sense of Ishii [24].

Proof. Since u_- is *l.s.c.*, we have to prove that the upper semicontinuous envelope u_- is a viscosity subsolution of (HJ) in the usual sense. By a result of Ishii [24], this is true if u_- is a (finite) sup of subsolutions. In the assertion and proof of Proposition 3.2, replace N, j, x_A, x_B, z_A, z_B with, respectively, $2N, N, y, z, k, w$, and consider for any fixed $w \in dom(H_2^*)$, the function $\psi(w,(y,z),t)$ defined by (3.10) with such substitutions. Then ψ is the solution of

$$\begin{cases} \psi_t + \mathcal{H}(w, D_y\psi, D_z\psi) = 0 & \text{in } \mathbb{R}^{2N} \times (0,T), \\ \psi(w,y,z,0) = u_0(y+z) & \text{in } \mathbb{R}^{2N} \times \{0\}, \end{cases}$$

where, for $p, q \in \mathbb{R}^N$

$$\mathcal{H}(w,p,q) := H_1(p) - w \cdot q - H_2^*(w) \geq H_1(p) + H_2(q). \tag{3.14}$$

Since the data of this problem satisfy the assumptions of Lemma 3.1, the unique viscosity solution of

$$\begin{cases} \tilde{\psi}_t + \mathcal{H}(w, D_x\tilde{\psi}, D_x\tilde{\psi}) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ \tilde{\psi}(w,x,0) = u_0(x) & \text{in } \mathbb{R}^N \times \{0\} \end{cases}$$

is given by $\tilde{\psi}(w,x,t) := \psi(w,(x_0, x-x_0),t)$, for any fixed $x_0 \in \mathbb{R}^N$. Since (3.14) holds, we observe that $\tilde{\psi}(w,\cdot,\cdot)$ is a (continuous) subsolution of $(HJ)(IC)$ for all $w \in dom(H_2^*)$, and we reach the conclusion by noting that the proof of Proposition 3.2 gives

$$u_-(x,t) = \max_{w \in \mathbb{R}^N} \psi(w,(x_0,x-x_0),t).$$

A similar argument shows that u_+ is a generalized supersolution of $(HJ)(IC)$, i.e., its lower semicontinuous envelope is a viscosity supersolution. □

Remark 5. In a special case, Kružkov proved that U_+ is Lipschitz and satisfies (HJ) almost everywhere, see Theorem 4 in [28] for the precise assumptions.

Remark 6. It can be shown that if u_0 is Lipschitz continuous, then 'inf' and 'sup' in formulas (3.2), (3.3) can be computed on particular compact subsets of $\mathbb{R}^N$, i.e.,

$$\max_{|w| \leq L_2} \min_{|k| \leq L_1} g(k,w,x,t) \leq u(x,t) \leq \min_{|k| \leq L_1} \max_{|w| \leq L_2} g(k,w,x,t),$$

where L_1 (resp., L_2) is the Lipschitz constant of H_1 (resp., H_2) on $\bar{B}(0, Lip(u_0))$. This result follows from the lemma below.

LEMMA 3.3. *Let the hypotheses of Proposition* 3.2 *be satisfied, assume* u_0 *Lipschitz continuous, and let* $L_1 > 0$ *(respectively,* $L_2 > 0$*) be such that* $|H_1(p_A) - H_1(q_A)| \le L_1|p_A - q_A|$*, for all* $|p_A|, |q_A| \le Lip(u_0)$*, (resp.,* $|H_2(p_B) - H_2(q_B)| \le L_2|p_B - q_B|$*, for all* $|p_B|, |q_B| \le Lip(u_0)$*,). Then, for any compact* $K_1 \subset \mathbb{R}^j$*, (resp.,* $K_2 \subset \mathbb{R}^{N-j}$*), such that* $K_1 \supset B(x_A, L_1 t)$ *(resp.,* $K_2 \supset B(x_B, L_2 t)$*), we have*

$$\max_{z_B \in K_2} \min_{z_A \in K_1} G(z, x, t) \le u(x, t) \le \min_{z_A \in K_1} \max_{z_B \in K_2} G(z, x, t),$$

if G *is given by* (3.9)*, and*

$$\max_{|z_B| \le L_2} \min_{|z_A| \le L_1} G(z, x, t) \le u(x, t) \le \min_{|z_A| \le L_1} \max_{|z_B| \le L_2} G(z, x, t),$$

if G *is defined by* (3.8).

Proof. Observe that, for any fixed $z_B \in \mathbb{R}^{N-j}$, the solution $w(z_B, x_A, t)$ of

$$\begin{cases} w_t + H_1(D_{x_A} w) = 0, & \text{in } \mathbb{R}^j \times (0, T), \\ w(x_A, 0) = u_0(x_A, z_B), & \text{in } \mathbb{R}^j \times \{0\} \end{cases}$$

is given by

$$\begin{aligned} w(z_B, x_A, t) &= \min_{z_A \in \mathbb{R}^j} \left\{ u_0(z_A, z_B) + tH_1^* \left(\frac{x_A - z_A}{t} \right) \right\} + tH_2^* \left(\frac{z_B - x_B}{t} \right) \\ &= \min_{z_A \in B(x_A, L_1 t)} \left\{ u_0(z_A, z_B) + tH_1^* \left(\frac{x_A - z_A}{t} \right) \right\} + tH_2^* \left(\frac{z_B - x_B}{t} \right) \end{aligned}$$

by the property of the cone of dependence. Now we take the max for $z_B \in B(x_B, L_2 t)$ and use (3.7), (3.9) to get

$$\max_{z_B \in K_2} w(z_B, x_A, t) \le \sup_{z_B \in \mathbb{R}^{N-j}} \inf_{z_A \in \mathbb{R}^j} G(z, x, t) \le u(x, t).$$

By similar reasoning we get the second inequality. The inequalities with G given by (3.8) follow from the change of variables $w = (x - z)/t$. ∎

4. Estimates for the solution with initial data sum of a convex and a concave function. In this section we give a result similar to that of Theorem A for the following type of data:

$$(H2) \qquad \begin{cases} u_1 : \mathbb{R}^N \to \mathbb{R} & \text{convex and Lipschitzean,} \\ u_2 : \mathbb{R}^N \to \mathbb{R} & \text{concave and Lipschitzean,} \\ u_0(x) = u_1(x) + u_2(x) \end{cases}$$

and $H : \mathbb{R}^N \to \mathbb{R}$ merely continuous. We set $f : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \times [0, \infty) \to \mathbb{R}$

$$f(k, w, x, t) := x \cdot (k + w) - u_1^*(k) - u_2^*(-w) - tH(k + w), \tag{4.1}$$

where u_1^*, u_2^* are, respectively, the Legendre transforms of u_1, u_2; we set also

$$v_-(x, t) := \sup_{k \in \mathbb{R}^N} \inf_{w \in \mathbb{R}^N} f(k, w, x, t), \tag{4.2}$$

$$v_+(x,t) := \inf_{w\in\mathbb{R}^N} \sup_{k\in\mathbb{R}^N} f(k,w,x,t). \tag{4.3}$$

THEOREM B. *Assume (H2) and H continuous. The unique viscosity solution $u \in UC_x(\mathbb{R}^N \times [0,T])$ of (HJ)(IC) satisfies*

$$v_-(x,t) \le u(x,t) \le v_+(x,t) \tag{4.4}$$

for all $(x,t) \in \mathbb{R}^N \times [0,T]$.

Remark. Note that for $t = 0$ the inequalities in (4.4) are equalities, and that for $u_2 = 0$ (resp., $u_1 = 0$), both u_-, u_+ are given by the second Hopf formula for convex (resp., concave) initial data.

The idea of the proof of Theorem B is similar to that of Theorem A. We study the Cauchy problem in doubled state variables

$$\begin{array}{lll} V_t + H(D_yV + D_zV) = 0 & \text{in } \mathbb{R}^{2N}\times(0,T) & (HJ''),\\ V(y,z,0) = u_1(y) + u_2(z) & \text{in } \mathbb{R}^{2N}\times\{0\} & (IC''). \end{array}$$

Let $V = V(y,z,t)$ be the unique viscosity solution of (HJ'') (IC'') in $UC_{y,z}(\mathbb{R}^{2N} \times (0,T))$. The initial data of this problem is the sum of a convex function of the first N variables and a concave function of the last N variables, so the analogue of Proposition 3.2 for this problem is known from [4] and gives the following estimate:

$$\sup_{k\in\mathbb{R}^N} \inf_{w\in\mathbb{R}^N} F(k,w,y,z,t) \le V(y,z,t) \le \inf_{w\in\mathbb{R}^N} \sup_{k\in\mathbb{R}^N} F(k,w,y,z,t), \tag{4.5}$$

where $F : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \times [0,\infty) \to \mathbb{R}$ is defined by

$$F(k,w,x,t) := y\cdot k + z\cdot w - u_1^*(k) - u_2^*(-w) - tH(k+w). \tag{4.6}$$

The next step is the connection between V and u, which is given by the following statement.

LEMMA 4.1. *Let u_0 satisfy (H2), H be continuous, and $V(y,z,t)$ be a continuous viscosity solution of $(HJ'')(IC'')$. Then, $u(x,t) := V(x,x,t)$ is a viscosity solution of (HJ)(IC).*

Proof. The initial condition (IC) is trivially satisfied by $u(x,t) = V(x,x,t)$. We claim now that u is a subsolution of (HJ). Let $\varphi \in C^1(\mathbb{R}^N \times [0,T])$ and suppose $u - \varphi$ has a local strict maximum at (x_0,t_0), that is $\exists \delta > 0$ such that

$$V(x,x,t) - \varphi(x,t) < V(x_0,x_0,t_0) - \varphi(x_0,t_0) \tag{4.7}$$

for all $(x,t) \in K := \bar{B}(x_0,\delta) \times [t_0-\delta, t_0+\delta] \setminus \{(x_0,t_0)\}$. Define the sequence of test functions

$$\Phi_n(y,z,t) := \varphi(y,t) + n|y-z|^2,\ n \in \mathbb{N},$$

and let (y_n,z_n,t_n) be the point at which $V - \Phi_n$ attains its maximum in K. Then,

$$(V-\Phi_n)(y_n,z_n,t_n) \ge (V-\Phi_n)(x_0,x_0,t_0) = V(x_0,x_0,t_0) - \varphi(x_0,t_0), \tag{4.8}$$

and by adding and subtracting $V(y_n,y_n,t_n) - \varphi(y_n,t_n)$ to the left-hand side of (4.8) and employing (4.7), we see also that

$$\begin{aligned}(V-\Phi_n)(y_n,z_n,t_n) \le\ & V(y_n,z_n,t_n) - V(y_n,y_n,t_n)\\ & + V(x_0,x_0,t_0) - \varphi(x_0,t_0) - n|y_n-z_n|^2.\end{aligned}$$

Now we combine this last inequality with (4.8) and deduce by the continuity of V that $n|y_n - z_n|^2$ is bounded on K. Thus we may assume, upon passing to subsequences and reindexing if necessary, that

$$\begin{aligned} &(y_n, z_n, t_n) \to (\bar{y}, \bar{z}, \bar{t}) \in K, \\ &n|y_n - z_n|^2 \to \alpha \in \mathbb{R}, \\ &n \to +\infty. \end{aligned}$$

In particular, $|y_n - z_n| \to 0$, $n \to +\infty$, that implies $\bar{y} = \bar{z}$. Now, by passing to limits as $n \to +\infty$ in (4.8), we derive

$$V(\bar{y}, \bar{y}, \bar{t}) - \varphi(\bar{y}, \bar{t}) - \alpha \geq V(x_0, x_0, t_0) - \varphi(x_0, t_0).$$

Then from (4.7) we get $\alpha = 0$, $(\bar{y}, \bar{t}) = (x_0, t_0)$. Provided V is a subsolution of (HJ''), we have

$$\frac{\partial \Phi_n}{\partial t} + H(D_y \Phi_n + D_z \Phi_n) \leq 0 \quad \text{at } (y_n, z_n, t_n)$$

and then, letting $n \to \infty$,

$$\frac{\partial \varphi}{\partial t} + H(D_x \varphi) \leq 0 \quad \text{at } (x_0, t_0),$$

i.e., $u(x,t)$ is a subsolution of (HJ) at (x_0, t_0). Similarly, it can be shown that u is a supersolution of (HJ) at any point (x_0, t_0), by using in this case the sequence of test functions

$$\Psi_n(y, z, t) = \xi(y, t) - n|y - z|^2,$$

where $\xi \in C^1(\mathbb{R}^N \times (0, T))$ is such that $u - \xi$ has a local strict minimum at (x_0, t_0). □

Proof of Theorem B. Just put $y = z = x$ in (4.5). □

Remark 1. The result of Bardi and Osher (4.5) is a special case of Theorem B. It can be obtained from (4.4) by assuming $u_1(x) = u_2(x_A)$ and $u_2(x) = u_2(x_B)$ and by simple calculations. A similar computation shows that (4.4) implies the second Hopf formula (1.2), provided $u_2 \equiv 0$.

Remark 2. Also in this case 'inf' and 'sup' in (4.2), (4.3) are actually attained, for they can be computed, respectively, on the bounded sets $dom(u_2^*)$, $dom(u_1^*)$.

Remark 3. The function v_-, v_+ defined in Theorem B, are, respectively, a sub and a supersolution of $(HJ)(IC)$ in the generalized sense of Ishii [24].

Proof. As proved by Bardi and Osher in their paper [4], for any fixed $k \in dom(u_1^*)$, the function

$$(y, z, t) \to \xi(k, y, z, t) := \min_{w \in \mathbb{R}^N} F(k, w, y, z, t),$$

where F is defined by (4.6), is the viscosity solution of (HJ''), with the initial condition

$$\xi(k, y, z, 0) = k \cdot y - u_1^*(k) + u_2(z) \leq u_1(y) + u_2(z) \tag{4.9}$$

(see [4], proof of Theorem 1, for details). Hence, Lemma 4.1 shows $(x,t) \to \xi(k, x, x, t)$ solves (HJ), and since (4.9) holds, this function is also a (continuous) subsolution of $(HJ)(IC)$. Finally,

$$v_-(x, t) = \max_{k \in \mathbb{R}^N} \xi(k, x, x, t)$$

yields the thesis. The proof for v_+ is symmetric. □

Remark 4. Van, Than, and Hoang [35] have recently proved that v_+, defined by (4.3), is Lipschitz continuous in $\mathbb{R}^N \times [0, T]$ and satisfies (HJ) almost everywhere.

5. Representation formulas. In this section, we study sufficient conditions for (3.4) and (4.4) to hold as equalities, and hence derive representation formulas for the solution of $(HJ)(IC)$ in both sets of assumptions: $(H1)$ and general u_0, $(H2)$ and general H.

Consider $f : \mathbb{R}^N \to \mathbb{R}$ and define the *index of nonconvexity* as

$$\mu(f) = \inf\{\alpha \in \mathbb{R} \ : \ f(x) + \alpha|x|^2 \text{ convex}\} \leq +\infty.$$

It is easy to check that if the 'inf' is finite it is actually attained, that f is convex iff $\mu(f) \leq 0$ and semiconvex iff $\mu(f)$ is finite. It is also easy to show that if $f \in C^2(\mathbb{R}^N)$ and semiconvex and λ is the infimum of the eigenvalues of $D^2 f(x)$ for $x \in \mathbb{R}^N$, then $\lambda = -2\mu(f)$.

Hereafter, we show that a representation formula can be obtained at least for small times, i.e., for $t \in [0, T]$, where T can be estimated in terms of the index of nonconvexity of the data.

We can assume with no loss of generality that $0 < \mu(u_0), \mu(-u_0)$, (if u_0 is convex or concave the second Hopf formula applies) and we will also assume $\mu(u_0), \mu(-u_0) < +\infty$, which are equivalent, respectively, to the semiconvexity and semiconcavity of u_0.

PROPOSITION 5.1. *Under the hypotheses of Theorem* A*, assume*

$$T \leq \min\left\{ \frac{|\mu(H_1^*)|}{\mu(u_0)}, \frac{|\mu(-H_2^*)|}{\mu(-u_0)} \right\},$$

and $0 < \mu(u_0), \mu(-u_0) < +\infty$. *Then the unique viscosity solution of* $(HJ)(IC)$ *in* $UC_x(\mathbb{R}^N \times [0, T])$ *is given by*

$$u(x, t) = u_-(x, t) = u_+(x, t),$$

where the functions u_-, u_+ *are defined in* (3.2), (3.3).

The proof of this proposition is based on the following theorem, which is a simplified version of the classical minimax results in [1, chapter 6, section 2].

MINIMAX THEOREM. *Let* $A \subset \mathbb{R}^n$ *and* $B \subset \mathbb{R}^m$ *be convex sets. Let also* $L : A \times B \to \mathbb{R}$ *be such that the following assumptions are satisfied:*

(i) $\forall b \in B, a \mapsto L(a, b)$ *is* l.s.c. *and convex, and* $\forall a \in A, b \mapsto L(a, b)$ *is* u.s.c. *and concave;*

(ii) $\exists b_0 \in B$ *and* $\exists a_0 \in A$ *such that* $\forall \lambda \in \mathbb{R}$ *the sets* $\{a \in A \ : \ L(a, b_0) \leq \lambda\}$ *and* $\{b \in B \ : \ L(a_0, b) \geq \lambda\}$ *are bounded.*

Then there exists a saddle point*, that is,* $(a^*, b^*) \in A \times B$ *such that*

$$L(a^*, b^*) = \min_{a \in A} \max_{b \in B} L(a, b) = \max_{b \in B} \min_{a \in A} L(a, b).$$

Proof of Proposition 5.1. We first observe that $k \mapsto g(k, w, x, t)$ is *l.s.c.* and that it is convex since it can be written as the sum of four convex functions of k as follows:

$$\begin{aligned} g(k, w, x, t) = &\{u_0(x - t(k - w)) + \mu(u_0)|x - t(k - w)|^2\} \\ &+ \ \{-\mu(u_0)\left(|x|^2 + t|w|^2 + 2tx \cdot w - 2t(x + tw) \cdot k\right) + tH_2^*(w)\} \\ &+ \ \{t(H_1^*(k) + \mu(H_1^*)|k|^2) \ \} \ + \ \{ \ - t(\mu(H_1^*) + t\mu(u_0))|k|^2\}. \end{aligned}$$

For any fixed x, t, $w \in dom(H_2^*)$ and $\lambda \in \mathbb{R}$, set

$$A_\lambda := \{k \in dom(H_1^*) \mid g(k, w, x, t) \leq \lambda\}.$$

Since H_1^* is superlinear, we have $H_1^*(k) = |k|h(k)$ with $h{:}\mathbb{R}^N \to \mathbb{R}$ such that $\lim_{|k|\to\infty} h(k) = +\infty$; moreover, let $C > 0$ be such that $u_0(x) \geq -C(|x| + 1)$. Then,

$$\begin{aligned} g(k,w,x,t) - tH_2^*(w) &= u_0(x - t(k-w)) + tH_1^*(k) \\ &\geq t|k|(h(k) - C) + K \end{aligned} \tag{5.1}$$

for a suitable $K > 0$, dependent only from x, t, C, w. Note that the right-hand side of (5.1) tends to $+\infty$ as $|k| \to \infty$, thus $\lim_{|k|\to\infty} g(k,w,x,t) = +\infty$. Then there exists $R > 0$ such that $|x| > R$ implies $g(k,w,x,t) > \lambda$, and so $A_\lambda \subset \bar{B}(0,R)$.

Similarly, it can be shown that $w \mapsto g(k,w,x,y)$ is concave and *u.s.c.*, and that for all fixed x, t, $k \in dom(H_1^*)$ and $\lambda \in \mathbb{R}$, the set

$$B_\lambda := \{w \in dom(H_2^*) \mid g(k,w,x,t) \geq \lambda\}$$

is a bounded subset of $\mathbb{R}^N$. □

Remark. In the case of smooth H and u_0, Kružkov [28, Theorem 1] proved an analogue of Proposition 5.1 for smooth solutions, where the index of nonconvexity of the data is replaced by a bound on the eigenvalues of the Hessian matrix.

While Proposition 5.1 applies to the general case, a stronger assertion can be proved in the case $H(p) = H_1(p_A) + H_2(p_B)$, where T can be arbitrarily large in some special cases. Set

$$\mu_A(u_0) = \sup_{v_B \in \mathbb{R}^{N-j}} \mu(u_0(\cdot, v_B)), \quad \mu_B(-u_0) = \sup_{v_A \in \mathbb{R}^j} \mu(-u_0(v_A, \cdot))$$

and consider the properties

$$\text{for any fixed } v_B \in \mathbb{R}^{N-j}, \quad v_A \mapsto u_0(v_A, v_B) \text{ convex}, \tag{5.2}$$

$$\text{for any fixed } v_A \in \mathbb{R}^j, \quad v_B \mapsto u_0(v_A, v_B) \text{ concave}. \tag{5.3}$$

Next, define

$$T_1 := \begin{cases} +\infty & \text{if (5.2) holds,} \\ \dfrac{|\mu(H_1^*)|}{\mu_A(u_0)} & \text{otherwise,} \end{cases}$$

$$T_2 := \begin{cases} +\infty & \text{if (5.3) holds,} \\ \dfrac{|\mu(-H_2^*)|}{\mu_B(-u_0)} & \text{otherwise,} \end{cases}$$

where $\mu(H_1^*)$ (resp., $\mu(-H_2^*)$) is the index of nonconvexity of H_1 (resp., H_2) as a function defined on $\mathbb{R}^j$ (resp., $\mathbb{R}^{N-j}$). Observe that if (5.2) (resp., (5.3)) does not hold, we have $\mu_A(u_0) > 0$ (resp., $\mu_B(-u_0) > 0$).

PROPOSITION 5.2. *Under the hypotheses of Proposition* 3.2, *for* $T < \min\{T_1, T_2\}$, *the unique viscosity solution in* $UC_x(\mathbb{R}^N \times [0,T])$ *of* $(HJ)(IC)$ *is given by*

$$u(x,t) = \sup_{z_B \in \mathbb{R}^{N-j}} \inf_{z_A \in \mathbb{R}^j} G(z,x,t) = \inf_{z_A \in \mathbb{R}^j} \sup_{z_B \in \mathbb{R}^{N-j}} G(z,x,t),$$

where G *is the function defined either by* (3.8) *or* (3.9).

Proof. Apply the minimax theorem. □

Remark 1. Note that if (5.2) and (5.3) hold, then $T_1 = T_2 = +\infty$, so we have a representation formula for all times. This result extends Theorem 2 in [28].

Remark 2. For times larger than the bounds indicated in Proposition 5.1 and 5.2, it may happen that $u_- < u_+$. We show this on some examples. The first two are taken from [29] and we report them for the reader's convenience.

Example 5.1. Consider the equation

$$u_t + u_{x_1}^2 - \frac{1}{2}u_{x_2}^2 = 0 \quad \text{in } \mathbb{R}^2 \times (0,\infty),$$

and the initial data

$$u_0(x) = -|x_1 + x_2|.$$

Here u_0 is concave, but $\mu_A(u_0) = +\infty$, so $T_1 = 0$. We use G given by (3.8), set $z = (k, w)$, and compute

$$g(k,w,x,t) = -|x_1 + x_2 - tk + tw| + \frac{tk^2}{4} - \frac{tw^2}{2}.$$

By remark (6) in section 3 (see Lemma 3.3), we can restrict the minmax procedure to compact sets and more precisely, since $\mathrm{Lip}(u_0) = 1$,

$$u_-(x,t) = \max_{|w|\le 1} \min_{|k|\le 2} g(k,w,x,t),$$

$$u_+(x,t) = \min_{|k|\le 2} \max_{|w|\le 1} g(k,w,x,t).$$

A straightforward but tedious computation gives

$$u_+(x,t) = -|x_1 + x_2| - \frac{t}{2}, \tag{5.4}$$

$$u_-(x,t) = \begin{cases} u_+(x,t) & \text{for } |x_1 + x_2| \ge t, \\ -t - (x_1 + x_2)^2/(2t) & \text{otherwise.} \end{cases} \tag{5.5}$$

Example 5.2. For the equation of Example 5.1, and for fixed $t > 0$, a smooth initial function u_0 such that $u_-(x,t) < u_+(x,t)$ can be constructed as follows. Take the initial data of the previous example, $\tilde{u}_0(x) = -|x_1 + x_2|$, and rename $\tilde{u}_+$, $\tilde{u}_-$ the corresponding super and subsolution given by (5.4), (5.5). It is easy to check that

$$|\tilde{u}_+(0,t) - u_+(0,t)| \le \sup_{[-2t,2t]\times[-t,t]} |\tilde{u}_0 - u_0|$$

and the same inequality holds for $|\tilde{u}_-(0,t) - u_-(0,t)|$. Since $\tilde{u}_+(0,t) - \tilde{u}_-(0,t) = t/2$ by (5.4), (5.5), we can take a smooth function u_0 approximating $\tilde{u}_0$ uniformly on a compact set so that $u_+(0,t) - u_-(0,t) \ge t/4$.

Example 5.3. Consider the equation

$$u_t + |u_{x_1}| - |u_{x_2}| = 0 \quad \text{in } \mathbb{R}^2 \times (0,\infty),$$

and observe that $\mu(H_1^*) = 0$, so that $T_1 = 0$. It is easy to compute

$$u_-(x,t) = \max_{|w|\le 1} \min_{|k|\le 1} u_0(x_1 + kt, x_2 + wt),$$

$$u_+(x,t) = \min_{|k|\le 1} \max_{|w|\le 1} u_0(x_1 + kt, x_2 + wt).$$

Now we choose any u_0 such that

$$\max_{|y_2|\le t} \min_{|y_1|\le t} u_0(y_1, y_2) < \min_{|y_1|\le t} \max_{|y_2|\le t} u_0(y_1, y_2).$$

Then $u_-(0,t) < u_+(0,t)$.

In the case of general Hamiltonian and u_0 satisfying (H2), we have similar results.

PROPOSITION 5.3. *Under the hypotheses of Theorem* B*, assume that* T *satisfies*

$$T \le \min\left\{ \frac{|\mu(u_1^*)|}{\mu(H)}, \frac{|\mu(-u_2^*)|}{\mu(-H)} \right\}$$

and $0 < \mu(H), \mu(-H) < +\infty$. *Then the unique viscosity solution of* $(HJ)(IC)$ *in* $UC_x(\mathbb{R}^N \times [0,T])$ *is given by*

$$u(x,t) = v_-(x,t) = v_+(x,t)$$

where the functions v_-, v_+ *are defined in* (4.2), (4.3).

Proof. Apply the minimax theorem. □

As we noted before for the dual case, we have a better result for initial data satisfying (H2) and such that $u_1(x) = u_1(x_A)$ and $u_2(x) = u_2(x_B)$. Similarly, we set

$$\mu_A(H) = \sup_{v_B \in \mathbb{R}^{N-j}} \mu(H(\cdot, v_B)), \quad \mu_B(-H) = \sup_{v_A \in \mathbb{R}^j} \mu(-H(v_A, \cdot)),$$

and

(5.6) for any fixed $v_B \in \mathbb{R}^{N-j}$, $v_A \mapsto H(v_A, v_B)$ convex,

(5.7) for any fixed $v_A \in \mathbb{R}^j$, $v_B \mapsto H(v_A, v_B)$ concave,

and then

$$T_1 := \begin{cases} +\infty & \text{if (5.6) holds,} \\ \dfrac{|\mu(u_1^*)|}{\mu_A(H)} & \text{otherwise,} \end{cases}$$

$$T_2 := \begin{cases} +\infty & \text{if (5.7) holds,} \\ \dfrac{|\mu(-u_2^*)|}{\mu_B(H)} & \text{otherwise,} \end{cases}$$

where $\mu(u_1^*)$ (resp., $\mu(-u_2^*)$) is the index of nonconvexity of u_1 (resp., u_2) as a function defined on $\mathbb{R}^j$ (resp., on $\mathbb{R}^{N-j}$).

PROPOSITION 5.4. *Assume* $H \in C(\mathbb{R}^N)$ *and* u_0 *satisfying* $(H2)$ *and such that* $u_1(x) = u_1(x_A)$ *and* $u_2(x) = u_2(x_B)$. *Then for* $T < \min\{T_1, T_2\}$, *the unique viscosity solution in* $UC_x(\mathbb{R}^N \times [0,T])$ *of* $(HJ)(IC)$ *is given by*

$$u(x,t) = \sup_{z_B \in \mathbb{R}^{N-j}} \inf_{z_A \in \mathbb{R}^j} F(z,x,t) = \inf_{z_A \in \mathbb{R}^j} \sup_{z_B \in \mathbb{R}^{N-j}} F(z,x,t),$$

where $F : \mathbb{R}^N \times \mathbb{R}^N \times [0,+\infty) \to \mathbb{R}$ *is defined by*

$$F(v,x,t) = x \cdot v - u_1^*(v_A) - u_2^*(-v_B) - tH(v).$$

Proof. Combine [4] Theorem 1 and the minimax theorem. □

6. Hamiltonians with dependence on the state variable. Hopf's formula can be extended to Cauchy problems for Hamiltonians with a particular dependence on the state variable x, i.e.,

$$\begin{cases} (HJ)_x \quad u_t + H(B(x)D_x u) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ (IC) \quad u(x,0) = u_0(x) & \text{in } \mathbb{R}^N \times \{0\}, \end{cases}$$

where $\psi : \mathbb{R}^N \to \mathbb{R}$ is a C^1-diffeomorphism and $B(x) = [D\psi(x)]^{-1}$. To this purpose, we are going to study the following associated problem:

$$\begin{cases} (HJ) \quad v_t + H(D_x v) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ (IC)_\psi \quad v(x,0) = u_0(\psi^{-1}(x)) & \text{in } \mathbb{R}^N \times \{0\}. \end{cases}$$

If H is convex and $u_0 \cdot \psi^{-1} \in UC(\mathbb{R}^N)$, its unique viscosity solution in $UC_x(\mathbb{R}^N \times [0,T])$ is, of course,

$$\begin{aligned} v(y,t) &= \min_{z \in \mathbb{R}^N} \left\{ u_0(\psi^{-1}(z)) + tH^* \left(\frac{y-x}{t} \right) \right\} \\ &= \min_{w \in \mathbb{R}^N} \{ u_0(\psi^{-1}(y - tw)) + tH^*(w) \}. \end{aligned}$$

LEMMA 6.1. *Let $\psi : \mathbb{R}^N \to \mathbb{R}$ be a C^1-diffeomorphism, $u, v \in C(\mathbb{R}^N \times [0,T])$ be such that $u(x,t) = v(\psi(x),t)$. Then v is a viscosity solution of $(HJ)(IC)_\psi$ if and only if u is a viscosity solution of $(HJ)_x(IC)$.*

Proof. Apply a change of variables for viscosity solutions [2]. □

A direct consequence of the preceding lemma is the following proposition.

PROPOSITION 6.2. *If H is convex, $u_0 \in C(\mathbb{R}^N)$, and $u_0 \cdot \psi^{-1} \in UC(\mathbb{R}^N)$, then the unique viscosity solution $u \in C(\mathbb{R}^N \times [0,T])$ of $(HJ)_x(IC)$ is given by*

$$\begin{aligned} u(x,t) &= \min_{z \in \mathbb{R}^N} \left\{ u_0(z) + tH^* \left(\frac{\psi(x) - \psi(z)}{t} \right) \right\} \\ &= \min_{w \in \mathbb{R}^N} \{ u_0(\psi^{-1}(\psi(x) - tw)) + tH^*(w) \}. \end{aligned} \tag{6.1}$$

Observe that uniqueness follows from Lemma 6.1 and the uniqueness of v as a solution of $(HJ)(IC)_\psi$ in $UC(\mathbb{R}^N \times [0,T])$.

Example. Formula (6.1) applies to the following case: consider the Hamiltonian $H = H(g_1(x_1)u_{x_1}, \dots, g_N(x_N)u_{x_N})$, where $g_i \in C(\mathbb{R}^N)$, $g_i > 0$, $g_i \in L^\infty(\mathbb{R})$. Then $\psi : \mathbb{R}^N \to \mathbb{R}^N$ defined by

$$\psi(x) = \left(\int_0^{x_1} \frac{1}{g_1(t)} dt, \dots, \int_0^{x_N} \frac{1}{g_N(t)} dt \right)$$

is the required isomorphism, and its inverse $\varphi = (\varphi_1, \dots, \varphi_N)$ is Lipschitz since it satisfies

$$\varphi_(y_i) = \int_0^{y_i} g_i(\varphi_i(t)) dt,$$

(see [21] for some physical examples).

Now we extend Theorem A to problems with Hamiltonians dependent on x. Applying Proposition 3.2, Theorem A, and Lemma 6.1 to $(HJ)(IC)_\psi$, we easily obtain the next result.

THEOREM 6.3. *Assume $u_0 \in C(\mathbb{R}^N)$, $u_0 \cdot \psi^{-1} \in UC(\mathbb{R}^N)$, $H \in C(\mathbb{R}^N)$, $H(p) = H_1(p) + H_2(p)$, with H_1 convex and H_2 concave. Then the unique (continuous) viscosity solution of $(HJ)_x(IC)$, satisfies*

$$\max_{w\in\mathbb{R}^N} \min_{k\in\mathbb{R}^N} g_\psi(k,w,x,t) \le u(x,t) \le \min_{k\in\mathbb{R}^N} \max_{w\in\mathbb{R}^N} g_\psi(k,w,x,t),$$

where $g_\psi(k,w,x,t) := u_0(\psi^{-1}(\psi(x) - t(k-w))) + tH_1^*(k) + tH_2^*(w)$.

Remark. It is clear that one can proceed similarly to extend Theorem B to $(HJ)_x(IC)$ under the assumption of section 4 and, for example, the additional condition that ψ^{-1} is Lipschitz continuous.

REFERENCES

[1] J. F. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.

[2] M. BARDI AND I. CAPUZZO DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser, Boston, MA, 1997.

[3] M. BARDI AND L. C. EVANS, *On Hopf's formulas for solutions of Hamilton–Jacobi equations*, Nonlinear Anal., 8 (1984), pp. 1373–1381.

[4] M. BARDI AND S. OSHER, *The non-convex multidimensional Riemann problem for Hamilton–Jacobi equations*, Siam J. Math. Anal., 22 (1991), pp. 344–351.

[5] G. BARLES, *Solutions de viscosité des équations de Hamilton–Jacobi*, Springer-Verlag, Berlin, Heidelberg, 1994.

[6] G. BARLES, *Asymptotic behaviour of viscosity solutions of first order Hamilton–Jacobi equations*, Ricerche Mat., 34 (1985), pp. 227–260.

[7] G. BARLES, *Uniqueness for first-order Hamilton–Jacobi equations and Hopf formula*, J. Differential Equations, 69 (1987), pp. 346–367.

[8] E. N. BARRON, R. JENSEN, AND W. LIU, *Hopf-Lax-type formula for $u_t + H(u,Du) = 0$*, J. Differential Equations, 126 (1996), pp. 48–61.

[9] E. N. BARRON, R. JENSEN, AND W. LIU, *Hopf–Lax–type formula for $u_t + H(u,Du) = 0$: II*, Comm. Partial Differential Equations, 22 (1997), pp. 1141–1160.

[10] E. N. BARRON, R. JENSEN, AND W. LIU, *Explicit solution of some first order PDE's*, J. Dynamical Control Systems, 3 (1997), pp. 149–164.

[11] F. CARDIN, *On viscosity and geometrical solutions of Hamilton–Jacobi equations*, Nonlinear Anal., 20 (1993), pp. 713–719.

[12] M. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[13] E. D. CONWAY AND E. HOPF, *Hamilton's theory and generalized solutions of the Hamilton–Jacobi equation*, J. Math. Mech., 13 (1964), pp. 939–986.

[14] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[15] M. G. CRANDALL AND P. L. LIONS, *On existence and uniqueness of solutions of Hamilton–Jacobi equations*, Nonlinear Anal., 10 (1986), pp. 353–370.

[16] L. CORRIAS, M. FALCONE, AND R. NATALINI, *Numerical schemes for conservation laws via Hamilton–Jacobi equations*, Math. Comp., 64 (1995), pp. 555–580.

[17] L. CORRIAS, *Fast Legendre–Fenchel transform and applications to Hamilton–Jacobi equation and conservation laws*, SIAM J. Numer. Anal., 33 (1996), pp. 1534–1558.

[18] L. C. EVANS, *Partial Differential Equations*, Berkeley Mathematics Lecture Notes Series 3, 1993.

[19] L. C. EVANS, *Some min–max methods for the Hamilton–Jacobi equation*, Indiana Univ. Math. J., 33 (1984), pp. 31–50.

[20] L. C. EVANS AND P. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[21] S. FAGGIAN, *Formule di tipo Hopf per soluzioni di viscosità di equazioni di Hamilton–Jacobi*, thesis, Università di Padova, Italy, 1995.

[22] E. HOPF, *Generalized solutions of non–linear equations of first order*, J. Math. Mech., 14 (1965), pp. 951–973.

[23] H. Ishii, *Uniqueness of unbounded viscosity solution of Hamilton–Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721–748.
[24] H. Ishii, *Perron's method for Hamilton–Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.
[25] H. Ishii, *Representation of solutions of Hamilton–Jacobi equations*, Nonlinear Anal., 12 (1988), pp. 121–146.
[26] S. N. Kružkov, *The Cauchy problem in the large for certain non-linear first order equations*, Soviet Math. Dokl., 1 (1960), pp. 474–477.
[27] S. N. Kružkov, *Generalized solutions of nonlinear first order equations with several independent variables*, II, Math. USSR-Sb., 1 (1967), pp. 93–116.
[28] S. N. Kružkov, *On the minmax representation of solutions of first order nonlinear equations*, Functional Anal. Appl., 2 (1969), pp. 128–136.
[29] S. N. Kružkov, *First order nonlinear equations and the differential games connected with them*, Uspekhi Mat. Nauk, 24 (1969), pp. 227–228 (in Russian).
[30] P. L. Lions, *Generalized solutions of Hamilton–Jacobi equations*, Pitman, London, 1982.
[31] P. L. Lions and J. C. Rochet, *Hopf's formula and multi-time Hamilton–Jacobi equations*, Proc. Amer. Math. Soc., 96 (1986), pp. 79–84.
[32] S. Osher, *The Riemann Problem for nonconvex scalar conservation laws and Hamilton–Jacobi equations*, Proc. Amer. Math. Soc., 89 (1983), pp. 641–646.
[33] S. Osher and C. W. Shu, *High order essentially non-oscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.
[34] A. I. Subbotin, *Generalized solutions of First-Order PDEs*, Birkhäuser, Boston, MA, 1995.
[35] T. D. Van, M. D. Than, and M. Hoang, *On the representation of Lipschitz global solutions of the Cauchy problem for Hamilton–Jacobi equations*, in Proc. Internat. Conference Analysis and Mechanics of Continuous Media (Ho Chi Minh City), Ho Chi Minh City Math. Soc., 3 (1995), pp. 428–436.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1087–1105, September 1998

002

NONEXISTENCE OF HIGHER DIMENSIONAL STABLE TURING PATTERNS IN THE SINGULAR LIMIT*

YASUMASA NISHIURA† AND HIROMASA SUZUKI‡

Abstract. When the thickness of the interface (denoted by ε) tends to zero, any *stable* stationary internal layered solutions to a class of reaction–diffusion systems cannot have a smooth limiting interfacial configuration. This means that if the limiting configuration of the interface has a smooth limit, it must become unstable for small ε, which makes a sharp contrast with the one-dimensional case. This suggests that stable layered patterns must become very fine and complicated in this singular limit. In fact we can formally derive that the rate of shrinking of stable patterns is of order $\varepsilon^{1/3}$. Using this scaling, the resulting rescaled reduced equation determines the morphology of magnified patterns. A variational characterization of the critical eigenvalue combined with the matched asymptotic expansion method is a key ingredient for the proof, although the original linearized system is not of self-adjoint type.

Key words. reaction–diffusion system, interfacial pattern, singular perturbation, matched asymptotic expansion

AMS subject classifications. 35B25, 35B35, 35K57, 35R35

PII. S0036141096313239

1. Introduction.

Dynamics of interfacial patterns attracts much interest in many fields such as population dynamics, combustion, chemical reaction, solidification, and so on. One of the pioneering works in the pattern formation problem can be traced back to Turing [9], who found that spatially inhomogeneous patterns can be formed by diffusion-driven instability if the inhibitor diffuses faster than the activator. A typical model system is of the form

$$\begin{cases} u_t = \varepsilon^2 \Delta u + f(u,v), & \\ & (x,t) \in \Omega \times (0,\infty), \\ v_t = D\Delta v + g(u,v), & \\ \dfrac{\partial u}{\partial n} = 0 = \dfrac{\partial v}{\partial n}, & (x,t) \in \partial\Omega \times (0,\infty), \end{cases} \tag{1.1}$$

where u is the activator, v is the inhibitor, Ω is a smooth bounded domain in $\mathbf{R}^N$ ($N \geq 2$), and ε is a small positive parameter. The nonlinearity f has at least two stable branches for a fixed v, and the signs of g are different on these branches, typically $(f,g) = (u(1-u)(u-a) - v, u - \gamma v)$, where $0 < a < 1$, $\gamma > 0$. More precise assumptions for (f,g) are displayed at the end of this section. Although (1.1) exhibits a variety of patterns depending on diffusion and/or reaction rates, we focus on the stationary ones in higher space dimensions; we are especially interested in layered solutions which have internal transition layers from one stable branch of the nullcline $f = 0$ to the other one (see (A.4)). The basic issue asks, *Does* (1.1) *have nonconstant stable stationary solutions up to* $\varepsilon = 0$*? And, if it does, what are the*

*Received by the editors December 9, 1996; accepted for publication (in revised form) September 15, 1997; published electronically April 14, 1998.
http://www.siam.org/journals/sima/29-5/31323.html
†Research Institute for Electronic Science, Hokkaido University, Sapporo 060, Japan (nishiura@aurora.elsip.hokudai.ac.jp).
‡Hiroshima National College of Maritime Technology, Higashino-cho, Toyota-gun, Hiroshima, 725-02, Japan (suzuki@hiroshima-cmt.ac.jp).

asymptotic configurations of them as $\varepsilon \downarrow 0$*?* As we shall see, this is closely related to finding the location of free boundary called the *interface* separating two different states. Numerically as well as experimentally, for a *fixed* $\varepsilon > 0$, a variety of stationary patterns have been observed such as spots, stripes, and labyrinthine patterns for (1.1) (see, for instance, [1] and the references therein). Hence naively one can expect that (1.1) has a lot of stable stationary solutions for small ε up to $\varepsilon = 0$.

In fact, for the one-dimensional case, it is proved rigorously (see [5]) that many stable layered solutions coexist up to $\varepsilon = 0$. It should be noted that each layer position has a definite limit and the distance between layer positions remains finite as $\varepsilon \downarrow 0$.

On the other hand, for the higher dimensional case, we know very little about the limiting configuration of *stable* stationary solutions to (1.1) when ε tends to zero. For instance, the planar front does not persist as a stable one (see [7]), and more complicated patterns take it over for small ε. We rephrase our basic question in the following way: *Does* (1.1) *have an* ε*-family of stable stationary layered solutions* $(U^\varepsilon, V^\varepsilon)$ *with smooth interface* Γ^ε *up to* $\varepsilon = 0$*?* Here Γ^ε is defined by

$$\Gamma^\varepsilon \equiv \{x \in \Omega \mid U^\varepsilon(x) = \alpha^*\},$$

where α^* is an intermediate value between two stable branches of $f = 0$ and, for instance, is equal to $1/2$ for the above specific example. Note that "smooth up to $\varepsilon = 0$" means that there exists an $(N-1)$-dimensional smooth compact connected manifold Γ^0 without boundary embedded in $\mathbf{R}^N$ such that Γ^ε converges to Γ^0 smoothly as $\varepsilon \downarrow 0$.

The goal in this paper is to give a *negative* answer to this question under the assumption that it has a matched asymptotic expansion of order 1 (see section 2 for details). Namely, we have the following theorem.

THEOREM 1.1 (main theorem). *Suppose that* (1.1) *has an* ε*-family of stationary matched asymptotic solutions of order* 1 *whose interface is smooth up to* $\varepsilon = 0$*. Then it must become unstable for small* ε*.*

We prove this in section 3 by converting the linearized eigenvalue problem around $(U^\varepsilon, V^\varepsilon)$ to a variational one (see (3.5)) which consists of a scalar elliptic part and a nonlocal term coming from the coupling of the two equations. A key idea is that in the variational characterization of the maximal eigenvalue of the system, the term associated with the (local) elliptic boundary problem is positive and it dominates the contribution from the nonlocal term for a highly oscillating admissible function in azimuthal direction. This instability result leads to the natural question, how about the fate of stable ones when $\varepsilon \downarrow 0$? The above theorem strongly suggests that stable patterns somehow must become very fine and/or complicated when $\varepsilon \downarrow 0$, and if it happens, can we characterize the domain size of those patterns and their morphologies? We shall discuss these issues in section 4; in fact we can formally derive that the rate of shrinking of stable patterns is of order $\varepsilon^{1/3}$. Using this scaling, the resulting rescaled reduced equation determines the morphology of magnified patterns (see also [5], [6], and [8]).

We prove the above theorem under the following assumptions.

(A.0) Γ^ε is an $(N-1)$-dimensional smooth compact connected manifold without boundary inside of Ω, and the domain surrounded by Γ^ε is simply connected.

(A.1) f and g are smooth functions of u and v defined on some open set $\mathcal{O}$ in $\mathbf{R}^2$ and the partial derivative f_v(resp., g_u) is a negative (resp., positive) constant function.

(A.2) (a) The nullcline of f is sigmoidal and consists of three smooth curves $u = h_-(v)$, $h_0(v)$, and $h_+(v)$ defined on the intervals I_-, I_0, and I_+, respectively. Let

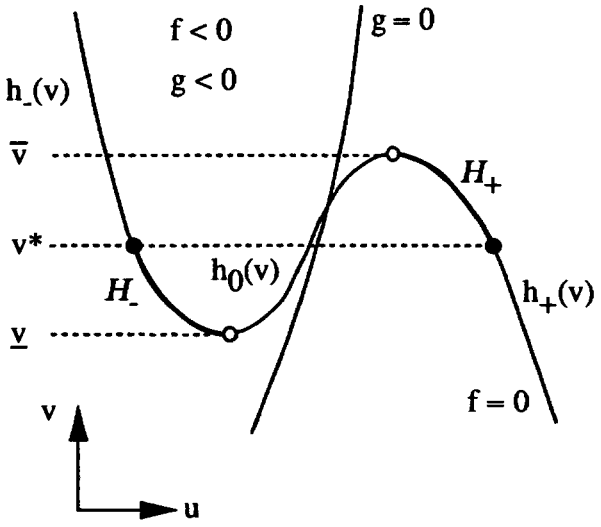

FIG. 1.1. *Functional forms of* $f = 0$ *and* $g = 0$.

$\min I_- = \underline{v}$ and $\max I_+ = \overline{v}$; then the inequality $h_-(v) < h_0(v) < h_+(v)$ holds for $v \in I^* \equiv (\underline{v}, \overline{v})$, and $h_+(v)$ (resp., $h_-(v)$) coincides with $h_0(v)$ at only one point $v = \overline{v}$ (resp., $\underline{v}$), respectively.

(b) The nullcline of g intersects with that of f at one or three points transversally as in Figure 1.1. The critical point on $u = h_-(v)$ (resp., $h_+(v)$ or $h_0(v)$), if it exists, is denoted by $P = (u_-, v_-) = (h_-(v_-), v_-)$ (resp., $Q = (u_+, v_+) = (h_+(v_+), v_+)$ or $R = (u_0, v_0) = (h_0(v_0), v_0)$).

(A.3) $J(v)$ has an isolated zero at $v = v^* \in I^*$ such that $dJ/dv < 0$ at $v = v^*$, where $J(v) = \int_{h_-(v)}^{h_+(v)} f(s, v)ds$. Moreover, we assume that $v_- < v^* < v_+$.

(A.4) $f_u < 0$ on $H_+ \cup H_-$, where H_- (resp., H_+) denotes the part of the curve $u = h_-(v)$ (resp., $h_+(v)$) defined by H_- (resp., H_+) $= \{(u, v) | u = h_-(v)$(resp., $h_+(v))$ for $v_- \le v < v^* (v^* < v \le v_+)\}$, respectively. Note that $v_- \le$ (resp., $\le v_+$) is replaced by $\underline{v} <$ (resp., $< \overline{v}$) when there are no critical points on the branch $u = h_-(v)$ (resp., $h_+(v)$). H_+ and H_- are called the stable branches of $f = 0$. See the thick solid part of $f = 0$ in Figure 1.1.

(A.5) (a)
$$g|_{H_-} < 0 < g|_{H_+},$$

(b)
$$\det\left(\frac{\partial(f, g)}{\partial(u, v)}\right)\Big|_{H_+ \cup H_-} > 0.$$

(A.6)
$$g_v|_{H_+ \cup H_-} < 0.$$

Remark 1.1. The assumption for the partial derivatives in (A.1) is necessary to make the problem (3.4) self-adjoint; however, it is plausible that our result holds true for the general case.

Remark 1.2. It is not known rigorously that, under the above assumptions, any smooth layered solution with a smooth limiting interface has a matched asymptotic

expansion of order 1; however, it is conjectured to be true at least for a large class of such layered solutions.

Remark 1.3. (A.4) and (A.5) (b) imply that

$$\frac{d}{dv}g(h_{\pm}(v),v) = \left.\frac{f_u g_v - f_v g_u}{f_u}\right|_{H_+\cup H_-} < 0. \tag{1.2}$$

We use the following notation throughout the paper: let $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ denote the usual multi-index of order $|\sigma| = \sigma_1 + \sigma_2 + \cdots + \sigma_N$ with nonnegative integers σ_i, and write $\partial_i = \partial/\partial x_i$ $(1 \le i \le N)$.

(i) Let k be a nonnegative integer and $\alpha \in (0,1)$. By $C^{k+\alpha}(\overline{\Omega})$ we mean the Banach space of all real-valued functions $u \in C^k(\overline{\Omega})$ for which the derivatives $\partial^\sigma u$ with $|\sigma| = k$ are Hölder continuous on $\overline{\Omega}$ with exponent α. The norm is

$$\|u\|_{C^{k+\alpha}(\overline{\Omega})} = \sum_{j=0}^{k} |u|_{j,\overline{\Omega}} + |u|_{k+\alpha,\overline{\Omega}},$$

where

$$|u|_{j,\overline{\Omega}} = \max_{|\sigma|=k} \sup_{x\in\overline{\Omega}} |\partial^\sigma u(x)|,$$

and

$$|u|_{k+\alpha,\overline{\Omega}} = \max_{|\sigma|=k} \sup_{x,y\in\overline{\Omega}} \frac{|\partial^\sigma u(x) - \partial^\sigma u(y)|}{|x-y|^\alpha}.$$

(ii) $C_0^{k+\alpha}(\overline{\Omega})$ is the subspace of $C^{k+\alpha}(\overline{\Omega})$ whose elements are functions vanishing on $\partial\Omega$.

(iii) $C_\varepsilon^{k+\alpha}(\overline{\Omega})$ is the Banach space of all real-valued functions in $C^{k+\alpha}(\overline{\Omega})$ but with the special norm depending on ε:

$$\|u\|_{C_\varepsilon^{k+\alpha}(\overline{\Omega})} = \sum_{j=0}^{k} \varepsilon^j |u|_{j,\overline{\Omega}} + \varepsilon^{k+\alpha} |u|_{k+\alpha,\overline{\Omega}}.$$

(iv) $C_{\varepsilon,0}^{k+\alpha}(\overline{\Omega})$ is the subspace of $C_\varepsilon^{k+\alpha}(\overline{\Omega})$ whose elements are functions vanishing on $\partial\Omega$.

2. Matched asymptotic expansion of singularly perturbed stationary solutions. In this section, we summarize the necessary conditions for the existence of the ε-family of matched asymptotic solutions with internal transition layers of the following stationary problem:

$$\begin{cases} 0 = \varepsilon^2 \Delta u + f(u,v) \\ \\ 0 = D\Delta v + g(u,v) \end{cases} \qquad \text{in } \Omega, \tag{2.1}$$

$$\frac{\partial u}{\partial n} = 0 = \frac{\partial v}{\partial n} \qquad \text{on } \partial\Omega. \tag{2.2}$$

Before presenting the precise form of matched asymptotic expansion, we need to do a change of variables near the interface. Let us assume that there exists an ε-family

of smooth solutions $(U^\varepsilon(x), V^\varepsilon(x))$ to (2.1) and (2.2) with interior transition layers such that the interface defined by

$$\Gamma^\varepsilon \equiv \{x \in \Omega \mid U^\varepsilon(x) = \alpha^*\} \tag{2.3}$$

is a compact smooth manifold of dimension $N-1$ embedded in $\mathbf{R}^N$ and has a definite limit Γ^0 with the same properties as $\varepsilon \downarrow 0$. Let (X_ϕ, ϕ) be a local chart on Γ^0, with $\phi(X_\phi)$ an open subset of $\mathbf{R}^{N-1}$. For $x_0 \in X_\phi$, $\phi(x_0) = s = (s^1, \dots, s^{N-1})$ and we denote the inverse of ϕ by

$$x_0 = (x_0^1(s), \dots, x_0^N(s)).$$

In some tubular neighborhood $U_d(\Gamma^0) = \{x \in \mathbf{R}^N \mid |y(x)| \le d\}$ of Γ^0, the local coordinate system $(s, y) = (s^1, \dots, s^{N-1}, y)$ is defined and for $x \in U_d(\Gamma^0)$,

$$x = X(s, y) \equiv x_0(s^1, \dots, s^{N-1}) + y\nu(s^1, \dots, s^{N-1}) \tag{2.4}$$

holds, where $\nu(s^1, \dots, s^{N-1})$ is the outward unit normal vector at $s = (s^1, \dots, s^{N-1})$ to Γ^0. Then, X becomes a diffeomorphism from $[-d, d] \times \Gamma^0$ to $U_d(\Gamma^0)$ if d is strictly smaller than the infimum of the radii of curvature of Γ^0. Its inverse is denoted by $(S(x), Y(x))$. Then Γ^ε can be represented by

$$\Gamma^\varepsilon = \{x_0(s) + \gamma(s, \varepsilon)\nu(s) \mid s \in \Gamma^0\},$$

where

$$\gamma(s, \varepsilon) = \sum_{k=1}^{m} \varepsilon^k \gamma_k(s) + \varepsilon^m \hat{\gamma}_{m+1}(s, \varepsilon).$$

Here we introduce the local shift variable τ by the following relation:

$$y = \tau + \omega\left(\frac{\tau}{d}\right)\gamma(s, \varepsilon), \tag{2.5}$$

where $\omega(\tau) \in C^\infty(\mathbf{R})$ is a cut-off function such that

$$\omega(\tau) = 1 \quad \text{for} \quad |\tau| \le \frac{1}{2}, \quad \omega(\tau) = 0 \quad \text{for} \quad |\tau| \ge 1,$$

$$0 \le \omega(\tau) \le 1, \quad |\omega'| \le 3.$$

Then, by the implicit function theorem, $\tau = \tau(s, y, \varepsilon)$ satisfying (2.5) is defined for sufficiently small ε. In place of x, we use a new independent variable $\hat{x}$, defined by

$$\hat{x} = \hat{X}(x, \varepsilon) = \begin{cases} x, & x \in \Omega \backslash U_d(\Gamma^0), \\ X(S(x), \tau(S(x), Y(x), \varepsilon)), & x \in U_d(\Gamma^0). \end{cases}$$

Let Ω_ε^+ (resp., Ω_0^+) be the region surrounded by Γ^ε (resp., Γ^0) and $\Omega_\varepsilon^- \equiv \Omega \backslash \overline{\Omega}_\varepsilon^+$ (resp., $\Omega_0^- \equiv \Omega \backslash \overline{\Omega}_0^+$). Then, note that $\hat{x} = \hat{X}(x, \varepsilon)$ maps Γ^ε to Γ^0, and $\Omega_\varepsilon^\pm$ to $\Omega_0^\pm$, respectively; namely, the free boundary Γ^ε becomes a fixed boundary Γ^0 in the new coordinate. Throughout the paper, we shall use the following notation:

$$u(x) = u(s, y), \quad \hat{u}(\hat{x}) = \hat{u}(s, \tau).$$

Using the above transformation, the stationary problem (2.1) with (2.2) can be rewritten as

$$\begin{cases} \varepsilon^2 M^\varepsilon \hat{u} + f(\hat{u}, \hat{v}) = 0 \\ \\ D M^\varepsilon \hat{v} + g(\hat{u}, \hat{v}) = 0 \end{cases} \qquad \text{in } \Omega, \tag{2.6}$$

$$\frac{\partial \hat{u}}{\partial n} = 0 = \frac{\partial \hat{v}}{\partial n} \qquad \text{on } \partial\Omega, \tag{2.7}$$

where $\hat{u} = \hat{u}(\hat{x})$, $\hat{v} = \hat{v}(\hat{x})$, and M^ε is the representation of the Laplacian Δ_x in $\hat{x}$. In $\Omega \backslash U_d(\Gamma^0)$, M^ε is equal to $\Delta_{\hat{x}}$. On the other hand, in the neighborhood $U_d(\Gamma^0)$, M^ε is defined in the following way: for the local coordinate system (s, y) defined by (2.4) in $\mathbf{R}^N$, let g^{ij} be the contravariant metric tensor and $g = \det(g^{ij})$. Then for $u(x) = u(s, y)$, Laplacian Δ_x is expressed by

$$\begin{aligned} (\Delta_x u)(x) &= (\Delta_{(s,y)} u)(s, y) \\ &\equiv \frac{\partial^2}{\partial y^2} u(s, y) + (N-1) H(s, y) \frac{\partial}{\partial y} u(s, y) \\ &\quad + \frac{1}{\sqrt{g}} \sum_{i=1}^{N-1} \frac{\partial}{\partial s^i} \left(\sqrt{g} \sum_{j=1}^{N-1} g^{ij} \frac{\partial}{\partial s^j} u(s, y) \right), \end{aligned} \tag{2.8}$$

where $H = H(s, y)$ is the mean curvature of the hypersurface $\Gamma(y) = \{x_0(s) + y\nu(s) \mid s \in \Gamma^0\}$ at (s, y). Using this representation, for $\hat{u}(\hat{x}) = \hat{u}(s, \tau)$, M^ε is defined by

$$(M^\varepsilon \hat{u})(\hat{x}) \equiv \Delta_{(s,y)} \hat{u}(s, \tau(s, y, \varepsilon)).$$

It follows from this definition that M^ε can be expanded as $M^\varepsilon = \sum_{k \geq 0} \varepsilon^k M_k$, where

$$M_0 \equiv \Delta_{\hat{x}}, \qquad \hat{x} \in \Omega,$$

and for $k \geq 1$,

$$M_k = \begin{cases} 0, & \hat{x} \in \Omega \backslash U_d(\Gamma^0), \\ \\ \text{at most the second order differential operator in} \\ s^i \ \ (i = 1, \ldots, N-1) \ \text{ and } \ \tau, & \hat{x} \in U_d(\Gamma^0). \end{cases}$$

In the following, we consider only (2.6) and (2.7), so we omit the hat symbol ($\hat{}$). A family of solutions $(U^\varepsilon, V^\varepsilon) \in X_\varepsilon \equiv C_\varepsilon^{2+\alpha}(\overline{\Omega}) \times C^{2+\alpha}(\overline{\Omega})$ of (2.6) and (2.7) is called an *ε-family of matched asymptotic solutions of order m* when it has the following expansions (2.9)–(2.12). Roughly speaking, $(U^\varepsilon, V^\varepsilon)$ is expanded separately in two regions $\Omega_0^\pm$ divided by the interface Γ^0, and they are matched smoothly at Γ^0. It should be recalled that the boundary condition (2.3) is always satisfied at Γ^0 for small ε owing to the change of variables (2.5). More precisely, for any positive integer m, we have

$$U^\varepsilon(x) = \begin{cases} U_+^\varepsilon(x) \equiv U_m^+(x, \varepsilon) + \Phi_m^+(x, \varepsilon) + o(\varepsilon^m), & x \in \Omega_0^+, \\ \\ U_-^\varepsilon(x) \equiv U_m^-(x, \varepsilon) + \Phi_m^-(x, \varepsilon) + o(\varepsilon^m), & x \in \Omega_0^-, \end{cases}$$

(2.9)
$$V^\varepsilon(x) = \begin{cases} V_+^\varepsilon(x) \equiv V_m^+(x,\varepsilon) + \varepsilon^2 \Psi_m^+(x,\varepsilon) + o(\varepsilon^m), & x \in \Omega_0^+, \\ V_-^\varepsilon(x) \equiv V_m^-(x,\varepsilon) + \varepsilon^2 \Psi_m^-(x,\varepsilon) + o(\varepsilon^m), & x \in \Omega_0^-, \end{cases}$$

where

$$U_m^\pm(x,\varepsilon) = \sum_{k=0}^m u_k^\pm(x)\varepsilon^k, \quad V_m^\pm(x,\varepsilon) = \sum_{k=0}^m v_k^\pm(x)\varepsilon^k, \tag{2.10}$$

$$\Phi_m^\pm(x,\varepsilon) = \begin{cases} \omega(\frac{Y(x)}{d}) \displaystyle\sum_{k=0}^m \phi_k^\pm\left(S(x), \frac{Y(x)}{\varepsilon}\right)\varepsilon^k, & x \in U_d(\Gamma^0) \cap \Omega_0^\pm, \\ 0, & x \in \Omega_0^\pm \backslash U_d(\Gamma^0), \end{cases} \tag{2.11}$$

$$\Psi_m^\pm(x,\varepsilon) = \begin{cases} \omega(\frac{Y(x)}{d}) \displaystyle\sum_{k=0}^m \psi_k^\pm\left(S(x), \frac{Y(x)}{\varepsilon}\right)\varepsilon^k, & x \in U_d(\Gamma^0) \cap \Omega_0^\pm, \\ 0, & x \in \Omega_0^\pm \backslash U_d(\Gamma^0), \end{cases} \tag{2.12}$$

$\phi_k^\pm$ and $\psi_k^\pm$ are functions of s and ξ, and ξ is the *stretched variable* $\xi \equiv \tau/\varepsilon$ (recall that $Y(\hat{x}) = \tau$). Here the topology of Landau's o is X_ε. The coefficients $u_k^\pm$, $v_k^\pm$, $\phi_k^\pm$, and $\psi_k^\pm$ satisfy the equations listed below in appropriate function spaces, which can be obtained by making *outer* and *inner* expansions and equating like powers of ε^k. The inner and outer solutions are not independent in the sense that they must satisfy the boundary conditions as well as the C^1-*matching conditions* between $(U_+^\varepsilon, V_+^\varepsilon)$ and $(U_-^\varepsilon, V_-^\varepsilon)$ on Γ^0. Let $\beta^\varepsilon(s) = v^* + \sum_{k=1}^m \beta_k(s)\varepsilon^k + \varepsilon^m \beta_{m+1}(s,\varepsilon)$ be the expansion of the value of V^ε on Γ^0. Note that the 0th order should be v^* from (A.3), since $(U^\varepsilon, V^\varepsilon)$ is a stationary solution.

We briefly explain the algorithm of the matched asymptotic expansion method and display the equations and relations up to order $O(\varepsilon^m)$. For more detailed arguments, see [2] and [6].

First we divide (2.6) into two problems as follows:

$$\begin{cases} \varepsilon^2 M^\varepsilon u^+ + f(u^+, v^+) = 0 & \\ & \text{in } \Omega_0^+, \\ DM^\varepsilon v^+ + g(u^+, v^+) = 0 & \\ u^+ = \alpha^*, \quad v^+ = \beta^\varepsilon & \text{on } \Gamma^0, \end{cases} \tag{2.13}_+$$

$$\begin{cases} \varepsilon^2 M^\varepsilon u^- + f(u^-, v^-) = 0 & \\ & \text{in } \Omega_0^-, \\ DM^\varepsilon v^- + g(u^-, v^-) = 0 & \\ u^- = \alpha^*, \quad v^- = \beta^\varepsilon & \text{on } \Gamma^0, \\ \dfrac{\partial u^-}{\partial n} = 0 = \dfrac{\partial v^-}{\partial n} & \text{on } \partial\Omega. \end{cases} \tag{2.13}_-$$

Then the interface is regarded as the boundary layer at Γ^0.

Outer expansion. Let

$$u^{\pm} = \sum_{k=0}^{m} u_k^{\pm}(x)\varepsilon^k, \quad v^{\pm} = \sum_{k=0}^{m} v_k^{\pm}(x)\varepsilon^k, \tag{2.14}$$

where both $u_k^{\pm}(x)$ and $v_k^{\pm}(x)$ belong to $C^{\infty}(\overline{\Omega}_0^{\pm})$. Substituting (2.14) into $(2.13)_{\pm}$ and equating like powers of ε, then $(u_k^{\pm}(x), v_k^{\pm}(x))$ satisfy the following equations:

$k = 0$

$$\begin{cases} f(u_0^{\pm}, v_0^{\pm}) = 0 & \\ & \text{in } \Omega_0^{\pm}, \\ DM_0 v_0^{\pm} + g(u_0^{\pm}, v_0^{\pm}) = 0 & \\ & \\ \dfrac{\partial v_0^{-}}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.15}$$

$k \geq 1$

$$\begin{cases} f_u^{0\pm} u_k^{\pm} + f_v^{0\pm} v_k^{\pm} = -\displaystyle\sum_{i+j=k-2} M_i u_j^{\pm} + P_{k-1}^{\pm} & \\ & \\ DM_0 v_k^{\pm} + g_u^{0\pm} u_k^{\pm} + g_v^{0\pm} v_k^{\pm} = -D \displaystyle\sum_{i+j=k, i\geq 1} M_i v_j^{\pm} + Q_{k-1}^{\pm} \text{ in } \Omega_0^{\pm}, & \\ & \\ \dfrac{\partial v_k^{-}}{\partial n} = 0 \qquad \text{on } \partial\Omega, & \end{cases} \tag{2.16}$$

where $f_u^{0\pm} \equiv \frac{\partial}{\partial u} f(u_0^{\pm}, v_0^{\pm})$, $f_v^{0\pm} \equiv \frac{\partial}{\partial v} f(u_0^{\pm}, v_0^{\pm})$, and so on. $P_{k-1}^{\pm}$ and $Q_{k-1}^{\pm}$ are functions determined only by $u_0^{\pm}, v_0^{\pm}, \ldots, u_{k-1}^{\pm}, v_{k-1}^{\pm}$. For $k = 1$, we define the right-hand side of (2.16) as equal to zero. Since $(U^{\varepsilon}, V^{\varepsilon})$ is a layered solution connecting two stable branches of the kinetics (f, g), it follows from (A.4) and the first equation of (2.15) that $(u_0^{\pm}, v_0^{\pm})$ lies on either H_+ or H_-. For definiteness, we assume that $u_0^{\pm} = h_{\pm}(v_0^{\pm})$. Then (2.16) can be uniquely solved recursively owing to the assumptions (A.4) and (A.5). This expansion is, however, insufficient because the layer part is not taken into account; in fact, u_0^{+} and u_0^{-} are discontinuous on Γ^0. To cope with this defect, we introduce a new variable $\xi = \tau/\varepsilon$ that rescales a neighborhood of the interface. Also note that the boundary conditions for $v_k^{\pm}$ are determined by the C^1-matching conditions discussed later.

Inner expansion. We introduce the stretched variable $\xi = \tau/\varepsilon$ and let

$$\begin{aligned} u^{\pm} &= U_m^{\pm}(x,\varepsilon) + \sum_{k=0}^{m} \phi_k^{\pm}\left(S(x), \frac{Y(x)}{\varepsilon}\right)\varepsilon^k, \\ v^{\pm} &= V_m^{\pm}(x,\varepsilon) + \varepsilon^2 \sum_{k=0}^{m} \psi_k^{\pm}\left(S(x), \frac{Y(x)}{\varepsilon}\right)\varepsilon^k, \end{aligned} \tag{2.17}$$

where $\phi_k^{\pm} = \phi_k^{\pm}(s,\xi)$ and $\psi_k^{\pm} = \psi_k^{\pm}(s,\xi)$. Since the definition domains of $\phi_k^{\pm}$ and $\psi_k^{\pm}$ are semi-infinite, these functions and the inhomogeneous terms of their equations listed below must have some decaying property for solvability. An appropriate function space for this purpose is the following.

DEFINITION 2.1. *Let $\mathcal{E}^{\pm}$ be the set of functions $E^{\pm}(s,\xi,\varepsilon)$ defined on $\Gamma_0 \times I^{\mp} \times [0,\varepsilon_0)$ with the property that for each C^{∞} linear differential operator D of any order in the variables s and ξ, there exist positive constants $C_{\pm}$ and K (possibly depending on D and $E^{\pm}$, but not on s, ξ, and ε) with $|DE^{\pm}| \le Ke^{-C_{\pm}|\xi|}$. Here $I^{-} \equiv (-\infty, 0)$ and $I^{+} \equiv (0,\infty)$.*

Substituting (2.17) into $(2.13)_{\pm}$ and equating like powers of ε, we obtain the following equations:

$k = 0$

$$\begin{cases} \ddot{\phi}_0^{\pm} + f(h_{\pm}(v^*) + \phi_0^{\pm}, v^*) = 0, \\ D\ddot{\psi}_0^{\pm} = g(h_{\pm}(v^*), v^*) - g(h_{\pm}(v^*) + \phi_0^{\pm}, v^*), \\ \phi_0^{\pm}(s, \mp\infty) = 0, \quad \psi_0^{\pm}(s, \mp\infty) = 0 = \dot{\psi}_0^{\pm}(s, \mp\infty), \end{cases}$$

$k = 1$

$$\begin{cases} \ddot{\phi}_1^{\pm} + \tilde{f}_u^{0\pm}\phi_1^{\pm} = -\tilde{M}_1\phi_0^{\pm} - \tilde{f}_u^{0\pm}\{u_1^{\pm}(s,0) + (u_0^{\pm})_{\tau}(s,0)\xi\} \\ \qquad\qquad\qquad\qquad -\tilde{f}_v^{0\pm}\{v_1^{\pm}(s,0) + (v_0^{\pm})_{\tau}(s,0)\xi\}, \\ D\ddot{\psi}_1^{\pm} = -D\tilde{M}_1\psi_0^{\pm} + \tilde{Q}_0^{\pm}, \\ \phi_1^{\pm}(s, \mp\infty) = 0, \quad \psi_1^{\pm}(s, \mp\infty) = 0 = \dot{\psi}_1^{\pm}(s, \mp\infty), \end{cases}$$

$k \ge 2$

$$\begin{cases} \ddot{\phi}_k^{\pm} + \tilde{f}_u^{0\pm}\phi_k^{\pm} = -\displaystyle\sum_{i+j=k, i\ge 1} \tilde{M}_i\phi_j^{\pm} + \tilde{P}_{k-1}^{\pm}, \\ D\ddot{\psi}_k^{\pm} = -D\displaystyle\sum_{i+j=k, i\ge 1} \tilde{M}_i\psi_j^{\pm} + \tilde{Q}_{k-1}^{\pm}, \\ \phi_k^{\pm}(s, \mp\infty) = 0, \quad \psi_k^{\pm}(s, \mp\infty) = 0 = \dot{\psi}_k^{\pm}(s, \mp\infty) \end{cases}$$

for $\xi \in I^{\mp}$ and $s \in \Gamma^0$, where $\cdot = \frac{\partial}{\partial\xi}$, $\tilde{f}_u^{0\pm} \equiv \frac{\partial}{\partial u} f(h_{\pm}(v^*) + \phi_0^{\pm}, v^*)$, $(u_0^{\pm})_{\tau}(s,0) = \frac{\partial}{\partial\tau}u_0^{\pm}(s,0)$, and so on. $\tilde{P}_{k-1}^{\pm}$ depends on $u_0^{\pm}, v_0^{\pm}, \ldots, u_k^{\pm}, v_k^{\pm}, \phi_0^{\pm}, \psi_0^{\pm}, \ldots, \phi_{k-1}^{\pm}, \psi_{k-2}^{\pm}$, and $\tilde{Q}_{k-1}^{\pm}$ does, moreover, depend on $\psi_{k-1}^{\pm}$. The solvability of the above equations in the space $\mathcal{E}^{\pm}$ can be shown in a similar way to [2], so we leave the details to the reader. $\tilde{M}^{\varepsilon}$ is the representation of M^{ε} in variables s and ξ and is expanded as

$$\tilde{M}^{\varepsilon} \equiv \frac{1}{\varepsilon^2}\sum_{k\ge 0}\varepsilon^k\tilde{M}_k.$$

Here $\tilde{M}_k$ $(k \ge 0)$ are at most second-order differential operators in s and ξ. The precise forms of $\tilde{M}_k$ are presented at the end of this section.

Boundary conditions and C^1-matching conditions. Now we describe the boundary conditions of $v_k^{\pm}$ and $\phi_k^{\pm}$ on Γ^0. Then $u_k^{\pm}$, $v_k^{\pm}$, $\phi_k^{\pm}$, and $\psi_k^{\pm}$ are determined

recursively. These conditions are given by

$$\alpha^* = \sum_{k=0}^{m} u_k^{\pm}(s,0)\varepsilon^k + \sum_{k=0}^{m} \phi_k^{\pm}(s,0)\varepsilon^k,$$

$$v^* + \sum_{k=1}^{m} \beta_k(s)\varepsilon^k = \sum_{k=0}^{m} v_k^{\pm}(s,0)\varepsilon^k + \varepsilon^2 \sum_{k=0}^{m-2} \psi_k^{\pm}(s,0)\varepsilon^k.$$

Equating like powers of ε, we have the following boundary conditions:

$k = 0$

$$\phi_0^{\pm}(s,0) = \alpha^* - u_0^{\pm}(s,0), \quad v_0^{\pm} = v^* \qquad s \in \Gamma^0, \tag{2.18}$$

$k \geq 1$

$$\phi_k^{\pm}(s,0) = -u_k^{\pm}(s,0), \quad v_k^{\pm}(s,0) = \beta_k(s) - \psi_{k-2}^{\pm}(s,0) \qquad s \in \Gamma^0.$$

In this way, we obtain the formal asymptotic solution of $(2.13)_{\pm}$. Since $(U^\varepsilon, V^\varepsilon)$ is a stationary solution of (2.6), $(U_\pm^\varepsilon, V_\pm^\varepsilon)$ must satisfy the C^1-matching conditions; that is,

$$\varepsilon \frac{\partial U_+^\varepsilon}{\partial \nu} = \varepsilon \frac{\partial U_-^\varepsilon}{\partial \nu}, \quad \varepsilon \frac{\partial V_+^\varepsilon}{\partial \nu} = \varepsilon \frac{\partial V_-^\varepsilon}{\partial \nu} \quad \text{on} \quad \Gamma^0.$$

After some computation, we have

$k = 0$

$$(v_0^+)_\tau(s,0) = (v_0^-)_\tau(s,0), \quad \dot{\phi}_0^+(s,0) = \dot{\phi}_0^-(s,0), \qquad s \in \Gamma^0. \tag{2.19}$$

$k \geq 1$

$$\begin{aligned} &(v_k^+)_\tau(s,0) + \dot{\psi}_{k-1}^+(s,0) = (v_k^-)_\tau(s,0) + \dot{\psi}_{k-1}^-(s,0), \\ &\dot{\phi}_k^+(s,0) + (u_{k-1}^+)_\tau(s,0) = \dot{\phi}_k^-(s,0) + (u_{k-1}^-)_\tau(s,0). \end{aligned} \qquad s \in \Gamma^0 \tag{2.20}$$

The second equation of (2.15) with the boundary and C^1-matching conditions (see (2.18) and (2.19)) is called the *reduced problem*; namely,

$$\begin{cases} D\Delta v_0^{\pm} + g(h_{\pm}(v_0^{\pm}), v_0^{\pm}) = 0 & \text{in} \quad \Omega_0^{\pm}, \\ v_0^{\pm} = v^*, \quad \dfrac{\partial v_0^+}{\partial \nu} = \dfrac{\partial v_0^-}{\partial \nu} & \text{on} \quad \Gamma^0, \\ \dfrac{\partial v_0^-}{\partial n} = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{2.21}$$

This is a free boundary problem for Γ^0 which determines the asymptotic configuration of stationary interfacial patterns. We define $v_0(x) \in C^1(\overline{\Omega})$ by

$$v_0(x) = \begin{cases} v_0^+(x), & x \in \Omega_0^+, \\ v_0^-(x), & x \in \Omega_0^-. \end{cases}$$

We close this section by presenting a lemma on the representations of $\tilde{M}_k$, which will become useful in the next section. The proof is delegated to [6].

LEMMA 2.2. *$\tilde{M}_0$, $\tilde{M}_1$, and $\tilde{M}_2$ have the following forms:*

$$\tilde{M}_0 \equiv \frac{\partial^2}{\partial \xi^2}, \quad \tilde{M}_1 \equiv (N-1)H_0(s)\frac{\partial}{\partial \xi},$$

$$\tilde{M}_2 \equiv \Delta^{\Gamma^0} - (P_1(s)+P_2(s))\frac{\partial}{\partial \xi} + P_3(s)\frac{\partial^2}{\partial \xi^2}$$

$$-D_s\frac{\partial}{\partial \xi} - H_1(s)(\xi+\gamma_1(s))\frac{\partial}{\partial \xi},$$

where

$$P_1(s) = \frac{1}{2G}\sum_{i=1}^{N-1} G_{s^i}\sum_{j=1}^{N-1} G^{ij}\partial_{s^j}\gamma_1,\ P_2(s) = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}[G^{ij}_{s^i}\partial_{s^j}\gamma_1 + G^{ij}\partial_{s^i s^j}\gamma_1],$$

$$P_3(s) = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} G^{ij}\partial_{s^i}\gamma_1\partial_{s^j}\gamma_1 > 0,\ D_s = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} G^{ij}\left(\partial_{s^i}\gamma_1\frac{\partial}{\partial s^j} + \partial_{s^j}\gamma_1\frac{\partial}{\partial s^i}\right),$$

$$H_1(s) \equiv \sum_{i=1}^{N-1}\kappa_i(s)^2.$$

$H_0(s)$ (resp., $\kappa_i(s)$) are the mean (resp., principal) curvature of Γ^0 at $s \in \Gamma^0$, G^{ij} is the contravariant metric tensor for the manifold Γ^0 of dimension $N-1$, $G = \det(G^{ij})$, and Δ^{Γ^0} is Laplace–Beltrami's operator defined on Γ^0. In particular, the coefficients of $\frac{\partial}{\partial s^j}$ in D_s are independent of ξ.

3. Instability result for stationary patterns as $\varepsilon \downarrow 0$. In this section we prove that the internal layered solutions in the previous section must become unstable when ε tends to zero. For this purpose, we show that the linearized eigenvalue problem around $(U^\varepsilon, V^\varepsilon)$,

$$\begin{cases} \lambda w = \varepsilon^2 M^\varepsilon w + f^\varepsilon_u w + f^\varepsilon_v z, \\ \lambda z = DM^\varepsilon z + g^\varepsilon_u w + g^\varepsilon_v z, \end{cases} \tag{3.1}$$

has an unstable eigenvalue where λ is the eigenvalue parameter, where $f^\varepsilon_u = \frac{\partial}{\partial u} f(U^\varepsilon, V^\varepsilon)$ and so on. The main result is the following.

THEOREM 3.1. *Suppose that* (1.1) *has an ε-family of stationary matched asymptotic solutions of order* 1 *(i.e., $m = 1$) whose interface is smooth up to $\varepsilon = 0$. Then,* (3.1) *has a positive (i.e., unstable) eigenvalue of $O(\varepsilon)$ for small ε.*

Proof. We divide the proof into three steps. In the first two steps we assume that the ε-family of stationary solutions has an asymptotic expansion up to order 2 (i.e., $m = 2$), which makes the proof readable, especially to understand how to handle the Laplacian part. Then it is easy to modify the proof to be valid for the order 1 case (see Step 3 below). Apparently (3.1) is not a self-adjoint problem, since $f^\varepsilon_v \neq g^\varepsilon_u$. In order to reduce the problem to a self-adjoint one, we first introduce the following

auxiliary problem:

$$\begin{cases} \lambda w = \varepsilon^2 M^\varepsilon w + f_u^\varepsilon w + f_v^\varepsilon z, \\ \eta z = DM^\varepsilon z + g_u^\varepsilon w + g_v^\varepsilon z, \end{cases} \tag{3.2}$$

where η is an auxiliary parameter. The following Step 1 deals with the case $\eta = 0$, where, by solving the second equation with respect to z (see (A.6)) and substituting it into the first equation, we have a self-adjoint problem of w. Then the existence of the unstable eigenvalue for (3.2) with $\eta = 0$ is shown via the variational method. This result can be extended to case $\eta > 0$ in Step 2, where it is proved that a positive eigenvalue of $O(\varepsilon)$ of (3.2) exists for each $\eta \neq 0$. The proof of Theorem 3.1 for $m = 2$ is an immediate consequence of these two steps. Finally in Step 3 we show that the expansion up to order 1 is sufficient for the proof of instability. Note that the linearized instability implies a nonlinear one for the class of evolutional systems like (1.1).

Step 1 ($\eta = 0$ case). Solving the second equation of (3.2) with $\eta = 0$ with respect to z as

$$z = (N^\varepsilon)^{-1} g_u^\varepsilon w, \tag{3.3}$$

where $(N^\varepsilon)^{-1} \equiv (-DM^\varepsilon - g_v^\varepsilon)^{-1}$, which is well defined from (A.6), and then substituting (3.3) into the first equation of (3.2), we obtain a scalar problem for w:

$$\lambda w = \varepsilon^2 M^\varepsilon w + f_u^\varepsilon w + f_v^\varepsilon (N^\varepsilon)^{-1} g_u^\varepsilon w. \tag{3.4}$$

From (A.1), (3.4) becomes a self-adjoint problem.

LEMMA 3.2. (3.4) *has a positive* (*i.e., unstable*) *eigenvalue of* $O(\varepsilon)$ *for small* ε.

Proof. In what follows, for simplicity of notation, we can assume that $f_v \equiv -1$ and $g_u \equiv 1$ without loss of generality. Since the linearized operator (3.4) is self-adjoint, we shall prove that the largest eigenvalue λ_0^ε of (3.4) becomes positive for small ε, which is characterized by

$$\lambda_0^\varepsilon = \sup_{w \in H^1(\Omega)} \frac{\displaystyle\int_\Omega \{-\varepsilon^2 |\nabla_{M^\varepsilon} w|^2 + f_u^\varepsilon w^2 - |(N^\varepsilon)^{-1/2} w|^2\} dx}{\displaystyle\int_\Omega w^2 dx}, \tag{3.5}$$

where ∇_{M^ε} is the representation of ∇ with respect to the coordinate $\hat{x}$ (see section 2). Recall that we write x instead of $\hat{x}$ and hence $Y(x) = \tau$ and $\xi \equiv \tau/\varepsilon$. Now we construct a suitable test function for our purpose. Let

$$Q(\xi) = \begin{cases} \omega\left(\dfrac{\varepsilon\xi}{d}\right) \dot{U}(\xi) & \text{for} \quad |\xi| \le \dfrac{d}{\varepsilon}, \\ 0 & \text{for} \quad |\xi| \ge \dfrac{d}{\varepsilon}, \end{cases}$$

where U is a solution of

$$\begin{cases} \ddot{U} + f(U, v^*) = 0, \\ U(\pm\infty) = h_\mp(v^*), \quad U(0) = \alpha^*. \end{cases} \tag{3.6}$$

We define $w(x)$ by the following product with $\Theta \in L^2(\Gamma^0)$ and $\|\Theta\|_{L^2(\Gamma^0)} = 1$:

$$w(x) = \begin{cases} Q\left(\dfrac{Y(x)}{\varepsilon}\right)\Theta(S(x)), & x \in U_d(\Gamma^0), \\ 0, & x \in \Omega \backslash U_d(\Gamma^0). \end{cases}$$

For this $w(x)$, $\varepsilon^2|\nabla_{M^\varepsilon} w|^2$ is computed as

$$\begin{aligned} \varepsilon^2|\nabla_{M^\varepsilon} w|^2 = \varepsilon^2(\dot{U}^2|\nabla^{\Gamma^0}\Theta|^2 - 2\ddot{U}^2\dot{U}^2\nabla^{\Gamma^0}\Theta \cdot \nabla^{\Gamma^0}\gamma_1 + \ddot{U}^2|\nabla^{\Gamma^0}\gamma_1|^2) \\ +\dddot{U}^2\Theta^2 + O(\varepsilon^3) \end{aligned} \tag{3.7}$$

in $U_d(\Gamma^0)$, where

$$\nabla^{\Gamma^0}\Theta_1 \cdot \nabla^{\Gamma^0}\Theta_2 \equiv \sum_{i,j=1}^{N-1} G^{ij}\frac{\partial\Theta_1}{\partial s^i}\frac{\partial\Theta_2}{\partial s^j}, \quad |\nabla^{\Gamma^0}\Theta_3|^2 \equiv \nabla^{\Gamma^0}\Theta_3 \cdot \nabla^{\Gamma^0}\Theta_3$$

for $\Theta_i \in L^2(\Gamma^0)(i = 1, 2, 3)$, and G^{ij} is the contravariant metric tensor for Γ^0. Here we used the fact that

$$|\nabla_{M^\varepsilon}\hat{u}|^2 = |\nabla_{(s,y)}u|^2 \equiv \sum_{i,j=1}^{N-1} g^{ij}\frac{\partial u}{\partial s^i}\frac{\partial u}{\partial s^j} + \left(\frac{\partial u}{\partial y}\right)^2$$

for $\hat{u}(s, \tau) = u(s, y)$. Hence, for a function $\varphi = \varphi(s, \tau(s, y, \varepsilon)/\varepsilon)$, we note that $\frac{\partial\varphi}{\partial s^i} = \varphi_{s^i} + \frac{1}{\varepsilon}\varphi_\xi\tau_{s^i}$, $\frac{\partial\varphi}{\partial y} = \frac{1}{\varepsilon}\varphi_\xi\tau_y$, $\tau_{s^i} = -\varepsilon\partial_{s^i}\gamma_1 + O(\varepsilon^2)$, and $\tau_y \equiv 1$ in a neighborhood of $\tau = 0$ (see (2.5)). Integrating (3.7) over Ω, the first term of numerator of (3.5) becomes

$$\begin{aligned} &\int_\Omega \varepsilon^2|\nabla_{M^\varepsilon} w|^2 dx \\ &= \varepsilon\int_{\Gamma^0}\int_{|\xi|\le\frac{d}{\varepsilon}} \{\varepsilon^2\dot{U}^2|\nabla^{\Gamma^0}\Theta|^2 + \varepsilon^2\ddot{U}^2\hat{P}_1(s)\Theta^2 + \dddot{U}^2\Theta^2\} d\xi dS + O(\varepsilon^4), \end{aligned} \tag{3.8}$$

where $\hat{P}_1(s) = |\nabla^{\Gamma^0}\gamma_1|^2$. The remainder term $O(\varepsilon^4)$ depends only on Γ^0 and the L^2-norm of Θ. In order to compute the second term of the numerator of (3.5), first note that

$$f_u^\varepsilon = \begin{cases} \tilde{F}_u^{0+} + \varepsilon\tilde{F}_u^{1+} + \varepsilon^2\tilde{F}_u^{2+} + O(\varepsilon^3) \ \text{ in } \ \Omega_0^+ \cap U_d(\Gamma^0), \\ \tilde{F}_u^{0-} + \varepsilon\tilde{F}_u^{1-} + \varepsilon^2\tilde{F}_u^{2-} + O(\varepsilon^3) \ \text{ in } \ \Omega_0^- \cap U_d(\Gamma^0), \end{cases} \tag{3.9}$$

where

$$\begin{aligned} \tilde{F}_u^{0\pm} &\equiv f_u(h_\pm(v^*) + \phi_0^\pm, v^*), \\ \tilde{F}_u^{1\pm} &\equiv f_{uu}(h_\pm(v^*) + \phi_0^\pm, v^*)\{\xi(u_0^\pm)_\tau(s, 0) + u_1^\pm(s, 0) + \phi_1^\pm\} \\ &\quad + f_{uv}(h_\pm(v^*) + \phi_0^\pm, v^*)\{\xi(v_0)_\tau(s, 0) + v_1^\pm(s, 0)\}, \end{aligned}$$

and the remainder term $O(\varepsilon^3)$ depends only on the stationary pattern $(U^\varepsilon, V^\varepsilon)$. The $O(1)$ term of (3.9) multiplied by w^2 combined with the third term of (3.8) vanishes, which is easily seen by differentiating (3.6) with respect to ξ. Hence we only focus on the contribution of (3.9) coming from the $O(\varepsilon)$-term and higher. The next equality is a key ingredient for the proof

$$\int_{-\infty}^{0} \tilde{F}_u^{1+}\dot{U}^2 d\xi + \int_{0}^{\infty} \tilde{F}_u^{1-}\dot{U}^2 d\xi = (v_0)_\tau(s,0)\frac{d}{dv}J(v^*) > 0. \tag{3.10}$$

In order to show (3.10), we note that $\dot{U} = \dot{\phi}_0^{\pm}$, $\ddot{U} = \ddot{\phi}_0^{\pm}$ for $\xi \in I^{\mp}$ (so we omit the superscript $\pm$ of $\dot{\phi}_0$ and $\ddot{\phi}_0$), and $p^{\pm} \equiv \dot{\phi}_1^{\pm}$ satisfy the next equation (see section 2):

$$\ddot{p}^{\pm} + \tilde{F}_u^{0\pm} p^{\pm} = \Omega^{\pm}, \tag{3.11}$$

where

$$\Omega^{\pm}(s,\xi) \equiv H^{\pm}(s,\xi) - \{(u_0^{\pm})_\tau(s,0)\tilde{f}_u^{0\pm} + (v_0)_\tau(s,0)\tilde{f}_v^{0\pm}\}$$

and

$$H^{\pm} \equiv -(N-1)H_0\ddot{\phi}_0 - \tilde{F}_u^{1\pm}\dot{\phi}_0. \tag{3.12}$$

Multiplying $\dot{\phi}_0$ on both sides of (3.11) and using the relations

$$\int_{\mp\infty}^{0} \tilde{f}_u^{0\pm}\dot{\phi}_0(z)dz = -\ddot{\phi}_0(0), \qquad \int_{\mp\infty}^{0} \tilde{f}_v^{0\pm}\dot{\phi}_0(z)dz = \int_{h_\pm(v^*)}^{\alpha^*} f_v(u,v^*)du,$$

we obtain, by integration by parts,

$$\begin{aligned}\int_{\mp\infty}^{0} H^{\pm}(s,z)\dot{\phi}_0(z)dz &= \ddot{\phi}_1^{\pm}(s,0)\dot{\phi}_0(0) - \ddot{\phi}_0(0)\{\dot{\phi}_1^{\pm}(s,0) + (u_0^{\pm})_\tau(s,0)\} \\ &\quad + (v_0)_\tau(s,0)\int_{h_\pm(v^*)}^{\alpha^*} f_v(u,v^*)du.\end{aligned} \tag{3.13}$$

On the other hand, multiplying $\dot{\phi}_0$ on both sides of (3.12) and using (3.13), we have

$$\begin{aligned}&\int_{-\infty}^{0} \tilde{F}_u^{1+}\dot{U}^2 d\xi + \int_{0}^{\infty} \tilde{F}_u^{1-}\dot{U}^2 d\xi = \int_{-\infty}^{0} \tilde{F}_u^{1+}\dot{\phi}_0^2 d\xi + \int_{0}^{\infty} \tilde{F}_u^{1-}\dot{\phi}_0^2 d\xi \\ &= -(N-1)H_0\int_{-\infty}^{0} \ddot{\phi}_0\dot{\phi}_0 d\xi - (N-1)H_0\int_{0}^{\infty} \ddot{\phi}_0\dot{\phi}_0 d\xi \\ &\quad -\ddot{\phi}_1^{+}(s,0)\dot{\phi}_0(0) + \ddot{\phi}_0\{\dot{\phi}_1^{+}(s,0) + (u_0^{+})_\tau(s,0)\} - (v_0)_\tau(s,0)\int_{h_+(v^*)}^{\alpha^*} f_v(u,v^*)du \\ &\quad +\ddot{\phi}_1^{-}(s,0)\dot{\phi}_0(0) - \ddot{\phi}_0\{\dot{\phi}_1^{-}(s,0) + (u_0^{-})_\tau(s,0)\} + (v_0)_\tau(s,0)\int_{h_-(v^*)}^{\alpha^*} f_v(u,v^*)du \\ &= (v_0)_\tau(s,0)\int_{h_-(v^*)}^{h_+(v^*)} f_v(u,v^*)du,\end{aligned}$$

which is the required result (3.10). Here we used the fact that $\ddot{\phi}_1^+(s,0) = \ddot{\phi}_1^-(s,0)$ and the C^1-matching condition of $\phi_1^\pm$ (see (2.20)). Using (3.8), (3.9), (3.10), and the Hopf boundary Lemma for v^0 on Γ^0 (see (1.2) and (2.21)), we obtain

$$\begin{aligned}\lambda_0^\varepsilon \geq C \Bigg[& \varepsilon \frac{m_*}{K_1} \frac{d}{dv} J(v^*) \\ & + \int_\Gamma \varepsilon^2 \left\{ -|\nabla^{\Gamma^0} \Theta|^2 - \frac{1}{K_1}(K_2 \hat{P}_1(s) - \hat{P}_2(s))\Theta^2 \right\} dS \\ & - \frac{1}{K_1 \varepsilon} \int_\Omega |(N^\varepsilon)^{-1/2} w|^2 dx \Bigg] + O(\varepsilon^3),\end{aligned} \tag{3.14}$$

where

$$m_* \equiv \min_{s \in \Gamma^0} (v_0)_\tau(s,0) < 0, \qquad K_1 \equiv \int_{-\infty}^{\infty} \dot{U}^2 d\xi, \qquad K_2 \equiv \int_{-\infty}^{\infty} \ddot{U}^2 d\xi,$$

$$\hat{P}_2(s) \equiv \int_{-\infty}^{0} \tilde{F}_u^{2+} \dot{U}^2 d\xi + \int_0^{\infty} \tilde{F}_u^{2-} \dot{U}^2 d\xi,$$

and C is a positive constant. The objective is to choose an appropriate test function in order to make the first term of $[\,\cdot\,]$ in (3.14) dominant, which is positive and $O(\varepsilon)$. First we choose Θ as the kth eigenfunction Θ_k of the following eigenvalue problem:

$$\Delta^\Gamma \Theta_k - \frac{1}{K_1}(K_2 \hat{P}_1(s) - \hat{P}_2(s))\Theta_k = \mu_k \Theta_k \qquad \text{on } \Gamma^0.$$

Then the second term of (3.14) is equal to $\varepsilon^2 \mu_k$. Note also that Θ_k converges to 0 as $k \to \infty$ in weak $L^2(\Gamma^0)$-sense. As for the third term of (3.14), which comes from the nonlocal part, we first note that when ε tends to zero,

$$\frac{1}{\varepsilon} Q\left(\frac{\tau}{\varepsilon}\right) = \frac{1}{\varepsilon} \dot{U}\left(\frac{\tau}{\varepsilon}\right) \omega\left(\frac{\tau}{d}\right) \longrightarrow c_0 \delta(\tau) \qquad \text{in } H^{-1}((-d,d))\text{-sense}, \tag{3.15}$$

where $\delta(\tau)$ is a Dirac's δ-function at 0 and c_0 is a positive constant. Let K_k^ε be

$$K_k^\varepsilon \equiv \int_\Omega \left| (N^\varepsilon)^{-1/2} \left(\frac{w_k}{\varepsilon}\right) \right|^2 dx = \int_\Omega \left[(N^\varepsilon)^{-1} \left(\frac{w_k}{\varepsilon}\right) \right] \left(\frac{w_k}{\varepsilon}\right) dx.$$

In view of (3.15) and considering that $(N^\varepsilon)^{-1}$ is a uniformly bounded operator mapping from $H^{-1}(\Omega)$ to $H^1(\Omega)$ with respect to ε, we see that K_k^ε is uniformly bounded with respect to ε and k and that $\int_{-d}^{d} (Q(\tau/\varepsilon)/\varepsilon) \cdot d\tau$ converges to the trace operator on Γ^0 from $H^1(\Omega)$ to $H^{1/2}(\Gamma^0)$ in operator norm sense. Therefore, by using the fact that Θ_k converges weakly to 0 as $k \to \infty$, we see that for any given small $c^* > 0$, there exists an ε_0 and k_0 such that

$$K_k^\varepsilon < c^* \quad \text{for} \quad 0 < \varepsilon \leq \varepsilon_0, \ k \geq k_0. \tag{3.16}$$

Substituting $\Theta = \Theta_k$ and $w = w_k$ into (3.14), we have

$$\lambda_0^\varepsilon \geq C\varepsilon \left[\frac{m_*}{K_1} \frac{d}{dv} J(v^*) - \frac{1}{K_1} K_k^\varepsilon + \varepsilon \mu_k \right] + O(\varepsilon^3),$$

where C is a positive constant. Using (3.16) and taking ε smaller, if necessary, we see that

$$\frac{m_*}{K_1}\frac{d}{dv}J(v^*) - \frac{1}{K_1}K_k^\varepsilon > 0. \tag{3.17}$$

Therefore the right-hand side of (3.5) becomes positive for sufficiently small $\varepsilon > 0$, which is greater or equal to the $O(\varepsilon)$ quantity. In the rest of the proof, we show that the upper bound of (3.5) is also of $O(\varepsilon)$. Since the first and the third terms of the numerator of (3.5) are nonpositive, it holds obviously that

$$\lambda_0^\varepsilon \le \sup_{w\in H^1(\Omega)} \frac{\displaystyle\int_\Omega f_u^\varepsilon w^2 dx}{\displaystyle\int_\Omega w^2 dx}.$$

In view of the expansion (3.9) and the assumptions (A.2) and (A.4), f_u^ε has a positive sign only in the ε-neighborhood along the normal direction of Γ^0. Therefore we have the estimate

$$\lambda_0^\varepsilon \le C\varepsilon|\Gamma^0|,$$

where C is a positive constant and $|\cdot|$ denotes the area, which completes the proof of Lemma 3.2. □

Step 2 ($\eta \neq 0$ case). Rewriting (3.2) as

$$\begin{cases} \lambda w = \varepsilon^2 M^\varepsilon w + f_u^\varepsilon w + f_v^\varepsilon z, \\ \\ 0 = DM^\varepsilon z + g_u^\varepsilon w + (g_v^\varepsilon - \eta)z \end{cases} \tag{3.18}$$

and noting that $g_v^\varepsilon - \eta < 0$ for $\eta \ge 0$ from (A.6), we see that all the computation in Step 1 is also valid for (3.18) with $\eta \ge 0$. Therefore we have the following lemma.

LEMMA 3.3. (3.18) *has a positive eigenvalue* $\lambda = \lambda^\varepsilon(\eta)$ *for* $\eta \ge 0$. *Moreover, there exist positive constants* C_0 *and* C_1 ($C_0 < C_1$) *which are independent of* η *and* ε *such that*

$$C_0\varepsilon < \lambda^\varepsilon(\eta) < C_1\varepsilon \tag{3.19}$$

holds for $\eta \ge 0$.

Proof of Theorem 3.1 ($m = 2$ case). It follows from Lemma 3.3 that $\lambda^\varepsilon(\eta)$ is a continuous function of η for $\eta \ge 0$. Since $\lambda^\varepsilon(\eta)$ has lower and upper bounds like (3.19), we see that $\eta = \lambda^\varepsilon(\eta)$ holds at least at one point $\eta = \eta^*$ (> 0) by the intermediate value theorem. This η^* is the required unstable eigenvalue for (3.1).

Step 3 (extension to $m = 1$ case). It suffices to show how Step 1 should be changed under the weaker assumption. When the asymptotic expansion has only terms up to $m = 1$, we see from Lemma 2.2 that we lose the precise expression for the gradient part $|\nabla_{M^\varepsilon} w|^2$, and the right-hand side of (3.7) becomes $\ddot{U}^2\Theta^2 + O(\varepsilon^2)$; namely, the gradient part is part of the remainder term $O(\varepsilon^2)$. Similarly, the first two terms of the integrand of the right-hand side of (3.8) move to the remainder term $O(\varepsilon^3)$. On the other hand, (3.10) remains valid. Hence (3.14) becomes the following:

$$\lambda_0^\varepsilon \ge C\left[\varepsilon\frac{m_*}{K_1}\frac{d}{dv}J(v^*) - \frac{1}{K_1\varepsilon}\int_\Omega |(N^\varepsilon)^{-1/2}w|^2 dx\right] + O(\varepsilon^2). \tag{3.20}$$

Note that the remainder term $O(\varepsilon^2)$ depends on Θ. In order to get the estimate (3.16), we select the function Θ in Step 1 as an eigenfunction of the elliptic operator of second order. However it is not necessary to take such an eigenfunction; in fact it suffices to employ any weakly convergent sequence to zero. Once we fix an appropriate Θ and then take ε to be sufficiently small, we have a similar estimate to (3.17). It is easy to see that Lemma 3.3 and the proof after it hold true for this case. This completes the proof of Theorem 3.1. □

4. Concluding remarks. As was mentioned in section 1, Theorem 1.1 strongly suggests that stable patterns of (1.1) become very fine and/or complicated in the limit of $\varepsilon \downarrow 0$ in higher dimensional spaces. What we discuss here seeks to find an appropriate scaling in space and time by which the resulting singular limit dynamics could have stable patterns of *finite* size. Such patterns are usually maintained by the balance of two competing forces as described below. In the course of the following formal analysis, it is intuitive why the stable patterns of (1.1) must become fine in the original scale.

Suppose there is a sharp transition layer (interface) Γ connecting two stable bulk states. There are two forces that drive the interface: one is the bulk force causing the translation of interface with certain speed $W(v|_\Gamma)$ which depends on the value of v at Γ; the other is a geometric force, i.e., mean-curvature effect.

In one word, the characteristic size of stable patterns is determined by the *balance between the above two forces* and turns out to be proportional to $\varepsilon^{1/3}$. It should be noted that the scale $\varepsilon^{1/3}$ coincides with the fastest growing wavelength of the planar front of (1.1) (see [7]). In what follows we consider a smooth subdomain $\tilde{\Omega}_\varepsilon$ ($\subset \Omega$) and assume that both u and v satisfy the Neumann boundary conditions on $\partial\tilde{\Omega}_\varepsilon$ and the diameter of $\tilde{\Omega}_\varepsilon$ shrinks to zero as $\varepsilon \downarrow 0$ with order ε^α. Here α $(0 < \alpha < 1)$ is an unknown exponent. Typically $\tilde{\Omega}_\varepsilon$ is a unit cell of some periodic structure in $\mathbf{R}^N$.

Applying a change of variable with unknown exponent α

$$\mathbf{y} = \frac{\mathbf{x}}{\varepsilon^\alpha} \qquad (0 < \alpha < 1)$$

to (1.1) ($D = 1$ for simplicity), we have

$$\text{(4.1)} \qquad \begin{cases} u_t = \varepsilon^{2(1-\alpha)}\Delta_{\mathbf{y}} u + f(u,v) \\ \\ v_t = \varepsilon^{-2\alpha}\Delta_{\mathbf{y}} v + g(u,v) \end{cases} \qquad \text{in } \hat{\Omega}_\varepsilon,$$

where $\Delta_{\mathbf{y}}$ stands for the Laplacian in $\mathbf{y}$-variable and $\hat{\Omega}_\varepsilon$ is the stretched domain of $\tilde{\Omega}_\varepsilon$. It is more convenient to rewrite (4.1) in the following form:

$$\text{(4.2)} \qquad \begin{cases} \varepsilon^{-(1-\alpha)} u_t = \varepsilon^{1-\alpha}\Delta_{\mathbf{y}} u + \varepsilon^{-(1-\alpha)} f(u,v), \\ \\ \varepsilon^{2\alpha} v_t = \Delta_{\mathbf{y}} v + \varepsilon^{2\alpha} g(u,v). \end{cases}$$

Suppose $\hat{\Omega}_\varepsilon$ has a smooth limit $\hat{\Omega}$ as $\varepsilon \downarrow 0$. Taking a limiting procedure similar to [3], we obtain the following interfacial dynamics:

$$\text{(4.3)} \qquad \begin{cases} \varepsilon^{-(1-\alpha)}\Gamma_t = \{W(v|_\Gamma) - \varepsilon^{1-\alpha}\kappa\}\mathbf{N} & \text{on } \Gamma(t), \\ \\ \varepsilon^{2\alpha} v_t^{\pm} = \Delta_{\mathbf{y}} v^{\pm} + \varepsilon^{2\alpha} g(h^{\pm}(v^{\pm}), v^{\pm}) & \text{in } \hat{\Omega}^{\pm}(t), \end{cases}$$

where $\Gamma(t)$ stands for the limiting configuration of the interface, κ denotes the mean curvature of $\Gamma(t)$, $\mathbf{N}$ is the unit normal vector at Γ pointing from $\hat{\Omega}^+$ to $\hat{\Omega}^-$, $W(\cdot)$ is the travelling velocity of the first equation of (1.1) with $\varepsilon = 1$ for a fixed v and is typically a monotone decreasing function of v, the domain $\hat{\Omega}$ is divided into two parts $\bar{\Omega}^{\pm}(t)$ by $\Gamma(t)$ where $u = h^{\pm}(v)$ on each subdomain, respectively, and v is matched in C^1-sense at $\Gamma(t)$. In view of the second equation of (4.3), $v^{\pm}$ may be expanded as

$$v^{\pm} = v_0^{\pm}(\mathbf{y}, t) + \varepsilon^{2\alpha} v_1^{\pm}(\mathbf{y}, t) + O(\varepsilon^{4\alpha}). \tag{4.4}$$

Substituting (4.4) into (4.3) and equating like powers of ε, we easily see that $v_0^{\pm} \equiv v^*$, where v^* is the equal area level of $f(u, v)$ (see (A.3)) with $W(v^*) = 0$. Expanding $W(v|_\Gamma)$ into the Taylor series, the principal part of the next order of (4.3) becomes

$$\begin{cases} \varepsilon^{-(1-\alpha)} \Gamma_t = \{\varepsilon^{2\alpha} W'(v^*) v_1|_\Gamma - \varepsilon^{1-\alpha} \kappa\} \mathbf{N} & \text{on } \Gamma(t), \\ 0 = \Delta_{\mathbf{y}} v_1^{\pm} + g(h^{\pm}(v^*), v^*) & \text{in } \hat{\Omega}^{\pm}(t). \end{cases} \tag{4.5}$$

The first term of the right-hand side of (4.5) is the bulk force and the second one is the mean-curvature effect. In order to make these two terms comparable, namely, in order that the bulk force is balanced with the curvature effect, the exponent α must be taken as $\alpha = \frac{1}{3}$. Suppose $\alpha \neq \frac{1}{3}$. Then either the bulk force or the curvature effect becomes dominant as $\varepsilon \downarrow 0$; hence there is no chance to have nontrivial stationary patterns of finite size in such an ε^{α}-rescaled domain. Employing this exponent $\alpha = \frac{1}{3}$ and introducing a new time scale $\tau \equiv \varepsilon^{4/3} t$, the rescaled interfacial dynamics is given by

$$\begin{cases} \Gamma_\tau = \{W'(v^*) v_1|_\Gamma - \kappa\} \mathbf{N} & \text{on } \Gamma(t), \\ 0 = \Delta_{\mathbf{y}} v_1^{\pm} + g(h^{\pm}(v^*), v^*) & \text{in } \bar{\Omega}^{\pm}(t). \end{cases} \tag{4.6}$$

Suppose $\tilde{\Omega}_\varepsilon$ is the unit cell of a periodic structure such as hexagonal lattice and that $\tilde{\Omega}_\varepsilon / \varepsilon^{1/3}$ has a definite limit $\hat{\Omega}$ as $\varepsilon \downarrow 0$; then the stationary problem of (4.6)

$$\begin{cases} 0 = \{W'(v^*) v_1|_\Gamma - \kappa\} \mathbf{N} & \text{on } \Gamma, \\ 0 = \Delta_{\mathbf{y}} v_1^{\pm} + g(h^{\pm}(v^*), v^*) & \text{in } \hat{\Omega}^{\pm}, \\ v_1^{\pm} \quad \text{are matched in} \quad C^1 \text{ -sense at} \quad \Gamma \end{cases} \tag{4.7}$$

is expected to give a stable morphology of the unit cell. We call (4.7) the *morphology equations* of (1.1). Note that (4.7) is exactly the same as (2.19) in [6] where Suzuki used the matched asymptotic method to obtain it. However very little is known about the existence problem for (4.7) as well as their geometric profiles.

There is another observation due to [4] for a related system to (1.1) from a different point of view, that claims that the global minimizer of the following functional must oscillate rapidly with frequency being proportional to $\varepsilon^{1/3}$. The functional is given by

$$\int_\Omega \left\{ \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} W(u) + \frac{1}{\varepsilon} |(-\Delta + \gamma I)^{-1/2} u|^2 \right\} dx, \tag{4.8}$$

where $W(u)$ is a double-well potential like $u^4/4-u^2/2$. This is related to our problem in the following sense. Suppose the relaxation time of v is much shorter than u (i.e., the quasi-static assumption for v is valid); then (1.1) can be replaced by

$$\text{(4.9)}\qquad \begin{cases} u_t = \varepsilon\Delta u + \dfrac{1}{\varepsilon} f_0(u) - \dfrac{1}{\varepsilon} v & \\ & (x,t) \in \Omega \times (0,\infty), \\ 0 = D\Delta v + u - \gamma v & \\ & \\ \dfrac{\partial u}{\partial n} = 0 = \dfrac{\partial v}{\partial n} & (x,t) \in \partial\Omega \times (0,\infty), \end{cases}$$

where $f_0(u) = u - u^3$. Solving the second equation with respect to v and substituting it to the first equation, we have a scalar equation for u with nonlocal term

$$u_t = \varepsilon\Delta u + \frac{1}{\varepsilon} f_0(u) - \frac{1}{\varepsilon}(-\Delta + \gamma I)^{-1} u,$$

which is the L^2-gradient equation of the functional (4.8). Suppose that $\Omega = Q = (0,1)^N$ (N-dimensional cube) with periodic boundary conditions. We see by employing the arguments of [4] that the global minimizer u_ε of (4.8) has to satisfy the following inequality:

$$\text{(4.10)}\qquad C_1\varepsilon^{-1/3} \le \frac{\displaystyle\int_Q |\nabla H(u_\varepsilon)| dx}{\displaystyle\int_Q |u_\varepsilon| dx} \le C_2\varepsilon^{-1/3},$$

where $H(z) = \int_0^z W^{1/2}(s)ds$ and C_1,C_2 are positive constants independent of ε. Roughly speaking, the middle term of (4.10) counts the number of interface, and hence, (4.10) means that the global minimizer has to take a fine structure, although we do not know whether u_ε is spatially periodic or not. Finally, it should be noted that the estimate (4.10) is valid only for the global minimizer and not for the local minimizers.

REFERENCES

[1] J. Boissonade, E. Dulos, and P. de Kepper, *Turing patterns: From myth to reality*, in Chemical Waves and Patterns, R. Kapral and K. Showalter, eds., Kluwer Academic Publishers, Norwell, MA, 1995, pp. 221–268.

[2] H. Ikeda, *On the asymptotic solutions for a weakly coupled elliptic boundary value problem with a small parameter*, Hiroshima Math. J., 16 (1986), pp. 227–250.

[3] X.-Y. Chen, *Dynamics of interfaces in reaction diffusion systems*, Hiroshima Math. J., 21 (1991), pp. 47–83.

[4] S. Müller and G. Weiss, private communication, 1996.

[5] Y. Nishiura, *Coexistence of Infinitely Many Stable Solutions to Reaction Diffusion Systems in the Singular Limit*, Dynam. Report. Expositions Dynam. Systems (N.S.) 3, Springer-Verlag, New York, 1994, pp. 25–103.

[6] H. Suzuki, *Asymptotic characterization of stationary interfacial patterns for reaction diffusion systems*, Hokkaido Math. J., 26 (1997), pp. 631–667.

[7] M. Taniguchi and Y. Nishiura, *Instability of planar interfaces in reaction-diffusion systems*, SIAM J. Math. Anal., 25 (1994), pp. 99–134.

[8] M. Taniguchi and Y. Nishiura, *Stability and characteristic wavelength of planar interfaces in the large diffusion limit of the inhibitor*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 117–145.

[9] A. Turing, *The chemical basis of morphogenesis*, Philos. Trans. Roy. Soc. London Ser. B, 237 (1952), pp. 37–72.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1106–1117, September 1998

003

ON MOTHER BODIES OF CONVEX POLYHEDRA*

BJÖRN GUSTAFSSON†

Abstract. If Ω is a bounded domain in $\mathbb{R}^N$ provided with a mass distribution ρ_Ω (e.g., Lebesgue measure restricted to Ω), another mass distribution μ sitting in Ω and producing the same external Newtonian potential as ρ_Ω is sometimes called a mother body of Ω, provided it is maximally concentrated in some sense. We first discuss the meaning of this and formulate five desirable properties ("axioms") of mother bodies. Then we show that convex polyhedra do have unique mother bodies in that sense made precise in the case that ρ_Ω is either Lebesgue measure on Ω, hypersurface measure on $\partial\Omega$, or any mixture of these two.

Key words. convex polyhedron, ridge, mother body, skeleton, balayage

AMS subject classifications. Primary: 31B20; Secondary: 31A20, 52B11

PII. S0036141097317918

1. Introduction.

A mother body (or maternal or materic body) in the terminology of the Bulgarian school of geophysics [Zi], [Ko1], [Ko2] is a more concentrated mass distribution sitting in a given body and producing the same external gravitational field as the latter. For example, one good mother body for a ball with constant mass density is a point mass (of appropriate strength) at the center of the ball. The meaning of a mother body being "more concentrated" is quite vague and there is no general agreement of its exact meaning.

Mother bodies are an important computational tool in geophysics (see, e.g., [Zi], [Ko2]). For solid polyhedra with constant mass density there are natural candidates of mother bodies with support on systems of hyperplanes reaching the boundary of the polyhedron at edges and corners. There is a beautiful example of D. Zidarov [Zi, Sect. III.6] (see also section 4 in the present paper) showing that mother bodies of this sort are not unique in general. One purpose of the present paper is to show that for *convex* polyhedra we do have uniqueness. (Zidarov's counterexample is a square in two dimensions with a smaller square at one corner cut away; hence, it is nonconvex.) The same result holds if, instead of constant volume density, the mass of the polyhedron is sitting on its boundary and has constant density there with respect to surface measure and even for any mixture of these two measures.

This paper however starts with a long discussion of what one should reasonably require of a mother body. This results in five "axioms" ((1)–(5) below), which we feel are fairly well motivated. In practice it is usually not possible to satisfy them all, but they could at least be looked upon as guide lines. The formulation of such a system of axioms is a secondary purpose of this paper.

SOME GENERAL NOTATION. If $A \subset \mathbb{R}^N$ we set
$A^c = \mathbb{R}^N \backslash A$,
$A^e = \mathbb{R}^N \backslash \bar{A}$ ($\bar{A}$ = closure of A),
int A = the interior of A,
$B(x,r) = \{y \in \mathbb{R}^N : |y-x| < r\}$,
$\mathcal{L}^N = N$-dimensional Lebesgue measure,

*Received by the editors March 4, 1997; accepted for publication (in revised form) September 15, 1997; published electronically April 14, 1998.
http://www.siam.org/journals/sima/29-5/31791.html
†Department of Mathematics, Royal Institute of Technology, S-10044, Stockholm, Sweden (gbjorn@math.kth.se).

$\mathcal{H}^{N-1}$ = $(N-1)$-dimensional Hausdorff measure,
$\mathcal{L}^N \lfloor \Omega$, $\mathcal{H}^{N-1} \lfloor \partial\Omega$: the above measures restricted to Ω and $\partial\Omega$, respectively,
supp μ = the closed support of a distribution μ.
$E(x) = \begin{cases} -c_2 \log|x| & (N=2), \\ c_N |x|^{2-N} & (N \geq 3) \end{cases}$
is the Newtonian kernel so that $-\Delta E = \delta$, the Dirac measure at the origin.
$U^\mu = E * \mu$ = the Newtonian potential of μ, if μ is a distribution with compact support in $\mathbb{R}^N$. Thus $-\Delta U^\mu = \mu$.

2. Discussion of mother bodies. By a "body" we shall mean a bounded domain $\Omega \subset \mathbb{R}^N$ satisfying $\Omega = \operatorname{int}(\overline{\Omega})$, $\mathcal{H}^{N-1}(\partial\Omega) < \infty$, and provided with an associated mass distribution $\rho = \rho_\Omega$. Primarily we think of the mass distribution with density one in the domain and density zero outside; i.e., $\rho = \mathcal{L}^N \lfloor \Omega$. However, the results in this paper work equally well for the case of hypersurface measure on the boundary, i.e., $\rho = \mathcal{H}^{N-1} \lfloor \partial\Omega$, or for any mixture of these two.

Thus given any two constants $a, b \geq 0$ with $a + b > 0$ we associate with any Ω as above the mass distribution

$$\rho_\Omega = a\mathcal{H}^{N-1} \lfloor \partial\Omega + b\mathcal{L}^N \lfloor \Omega.$$

Then ρ_Ω is a positive Radon measure, and we denote by U^Ω its Newtonian potential

$$U^\Omega = U^{\rho_\Omega} = E * \rho_\Omega$$

(a and b will be kept fixed throughout the discussion).

Given a body $\Omega \subset \mathbb{R}^N$, a mother body for it should be a signed measure μ having certain properties. The basic requirement is that

$$U^\mu = U^\Omega \qquad \text{in } \Omega^e. \tag{1}$$

Clearly this implies that supp $\mu \subset \bar{\Omega}$.

One possible additional requirement is that

$$U^\mu \geq U^\Omega \qquad \text{in all } \mathbb{R}^N. \tag{2}$$

Such a condition is natural if one wishes to think of ρ_Ω as being the result of applying some kind of (partial) balayage process to μ (cf. [Zi], [Sa1], [Ko1], [Ko2], [Gu-Sa1], [Gu-Sg], [Gu2]). Indeed, any balayage (or "sweeping") process we know of can be thought of as being composed of elementary steps in which point masses are swept to measures of the kind ρ_B for balls B centered at the support of the point masses, and for each such elementary step the potential of the measure decreases.

Thus in order to have a μ which is as "primitive" as possible with respect to balayage one should ask U^μ to be as large as possible. For this to be sensible one has to have a lower bound on μ because otherwise one can always increase a given U^μ. For example, as is natural, one may ask μ to be positive:

$$\mu \geq 0. \tag{3}$$

Since $\mu = \rho_\Omega$ itself satisfies (1)–(3) and the supremum of any increasing sequence of superharmonic functions is superharmonic (or $\equiv +\infty$), it follows that for any given Ω there exist (plenty of) measures μ satisfying (1)–(3) with U^μ maximal among potentials of such measures. A mother body for Ω should be one of these (cf. Proposition 2.1 at the end of this section).

In order for μ to be a good mother body it should be concentrated or minimal in some sense, such as having small support sitting deeply inside Ω. Although this is to some degree implicit in the desire that U^μ should be as large as possible, we also want to formulate such conditions in direct geometric terms. One way to be concentrated is simply to be singular with respect to Lebesgue measure. We shall find the slightly stronger requirement

$$\mathcal{L}^N(\text{supp } \mu) = 0 \tag{4}$$

convenient to work with.

It is easy to see, however, that (4) does not guarantee a good mother body. For any Ω there are an abundance of measures μ satisfying all of (1)–(4). It is just to fill Ω with (infinitely many) disjoint balls so that the remaining set has measure zero, and then replace the volume part of ρ_Ω by the sum of the appropriate point masses sitting in the centers of these balls. In other words, one writes $\Omega = \bigcup_{j=1}^\infty B(x_j, r_j) \cup (\text{null set})$, where the $B(x_j, r_j)$ are disjoint, and then takes $\mu = a\mathcal{H}^{N-1}\lfloor\partial\Omega + b\sum_{j=1}^\infty \mathcal{L}^N(B(x_j, r_j))\delta_{x_j}$, δ_x denoting the unit point mass at $x \in \mathbb{R}^N$.

One way in which a mother body μ constructed as above, by ball-packing, is not good is that supp μ typically (even if $a = 0$) contains all of $\partial\Omega$, and therefore cuts off the exterior of Ω from the interior. This must necessarily be so in general because when supp μ does not reach $\partial\Omega$ then (1) gives a harmonic continuation of U^Ω across $\partial\Omega$ into Ω, which is not possible unless $\partial\Omega$ is real analytic (roughly speaking). Nevertheless, whenever possible we desire something like the following to hold.

$$\text{Each component of } \mathbb{R}^N\backslash\text{supp } \mu \text{ intersects } \Omega^e. \tag{5}$$

This simply means that for each $x \in \Omega\backslash\text{supp } \mu$ there is a curve in $\mathbb{R}^N\backslash\text{supp } \mu$ joining x with some point in Ω^e.

The requirements (1)–(5) are the "axioms" for a mother body which we propose. As indicated earlier there is neither existence nor uniqueness of mother bodies satisfying (1)–(5) in general (Zidarov's counterexample fulfills all of (1)–(5)). Indeed, the problem of finding a mother body of a given body exhibits all features of an ill-posed problem: existence and uniqueness of solutions only under special conditions and sensitive dependence on given data when solutions do exist. Nevertheless, for certain particular classes of bodies, e.g., various kinds of polyhedra (see sections 3 and 4 below and [Gu-Sa2]) and certain types of algebraic domains [Sav-St-Sv], there are constructive algorithms for computing (candidates of) mother bodies.

For the rest of this section, we discuss in more detail the roles of the axioms (1)–(5) and various ways of relaxing or strengthening them. The axioms naturally fall into three groups: (1); (2) and (3); (4) and (5).

Axiom (1) is the most indispensable one. In the case that Ω^e has more than one component, a possible way to relax it is to require only

$$\nabla U^\mu = \nabla U^\Omega \text{ in } \Omega^e$$

(equality of the corresponding fields), which is actually more physical. An even weaker requirement is to ask (1) to hold only in the unbounded component of Ω^e.

The role of the conditions (2) and (partly) (3) is to guarantee that ρ_Ω is the result of a natural balayage operator applied to μ. When $a = 0$, $b > 0$, such an operator $\mu \mapsto \text{Bal}(\mu; b)$ can be defined by declaring that $\text{Bal}(\mu; b)$ shall be the measure which is closest to μ in the energy norm among all measures ν which satisfy $\nu \le b\mathcal{L}^N$. This

makes plain sense and defines $\mathrm{Bal}(\mu; b)$ uniquely whenever $\mu \geq 0$ has finite energy. The definition can then easily be extended to the case of infinite energy. One can show that if Ω is a body, then $\mathrm{Bal}(\mu; b) = \rho_\Omega$ holds if and only if both (1) and (2) are satisfied. In particular, it is possible to reconstruct Ω from μ when (1) and (2) hold, and both conditions are really necessary for this (there are examples of two different Ω satisfying (1), (3)–(5) for the same μ).

Thus the perhaps abstract-looking condition (2) plays a significant role in the context of balayage. It is probably more important than (3) because it is possible, to a certain extent, to allow nonpositive measures μ in $\mathrm{Bal}(\mu; b)$. We refer to [Sa1], [Gu-Sa1], [Gu2] for details on the above balayage operators.

For a general measure μ, $\mathrm{Bal}(\mu; b)$ will not necessarily be of the form ρ_Ω for some open set Ω, but if μ satisfies (3) and (4), it will. This allows for doing "continuous balayage," as follows. Suppose (1)–(4) hold for the pair (Ω, μ), and define for any $t \in \mathbb{R}$ the open set $\Omega(t)$ by $\mathrm{Bal}(e^t\mu; b) = \rho_{\Omega(t)}$. (This defines $\Omega(t)$ only up to a null set, but one naturally takes the largest possible $\Omega(t)$.) Then $\Omega(s) \subset \Omega(t)$ for $s < t$, $\Omega(0) = \Omega$, and $\Omega(t)$ shrinks down to $\mathrm{supp}\, \mu$ as $t \to -\infty$. Moreover, the pair $(\Omega(t), e^t\mu)$ satisfies (1)–(4) for each $t \in \mathbb{R}^N$.

One important point with this family $\Omega(t)$ is that its evolution can be described without reference to μ. Indeed, under some smoothness assumptions the evolution can be described by a nonlocal, but μ-independent, differential equation for the motion of $\partial\Omega(t)$: the normal velocity of the boundary $\partial\Omega(t)$ at any particular point is to be equal to the normal derivative at that point of the function $p = p_{\Omega(t)}$ which solves the Dirichlet problem $\Delta p = 1$ in $\Omega(t)$, $p = 0$ on $\partial\Omega(t)$. This is a Hele–Shaw type moving boundary problem, and by solving it (backwards) for $-\infty < t \leq 0$ with $\Omega(0) = \Omega$ as initial domain one should, in principle, get a canonical candidate of a mother body, namely by taking $\mu = \lim_{t\to-\infty} e^{-t}\rho_{\Omega(t)}$. Unfortunately, however, this moving boundary problem is badly ill-posed and existence of global solutions is not to be expected in general. Local in time solutions exist if, and basically only if, the initial domain has a real analytic boundary (see, e.g., [Re-Wo], [Ti]).

It is possible to introduce balayage operators as above and to do continuous balayage also when $a > 0$, but everything is more complicated in that case: the balayage operators are less well behaved and the evolution families are less continuous (cf. [He], [Gu-Sg]). It is not even true that μ determines Ω uniquely via (1)–(5) [He, Prop. 6.2]. Axiom (2) does not quite suffice for this, as it does in the case $a = 0$, and should therefore ideally be replaced by something stronger.

Returning now to the general case $(a, b \geq 0)$, another advantage with conditions (2), (3) is that they guarantee a certain coupling between the geometry of Ω and of $\mathrm{supp}\, \mu$. One may therefore prove [Gu-Sa1], [Sg], [Gu-Sg] that for any Ω and any point $x \in \partial\Omega$, the inward normal ray of $\partial\Omega$ at x intersects the closed convex hull of the support of any μ satisfying (1)–(3). Without (2), (3) there will be no geometric coupling whatsoever between Ω and $\mathrm{supp}\, \mu$. For any domain D and any (small) ball $B \subset D$ one can find a domain Ω approximating D arbitrarily well and a signed measure μ with $\mathrm{supp}\, \mu \subset B$ such that (1), (4), (5) hold for Ω, μ. See [Gu1], [Sa2] for the case $a = 0$.

From what has been said above it should be clear that conditions (2) and (3) have strong potential theoretic significance. There are however other points of view for which they seem less urgent. In certain complex variable and PDE approaches, see, e.g., [Eb], [Kh-Sh], [Sh], [St-Sv], one considers the search for mother bodies mainly as a problem of analytic continuation (of U^Ω), and one is happy if one can find

a distribution (or even analytic functional) μ which satisfies (1) and some (usually stronger) form of (4), (5). If supp μ is then sufficiently small there will simply be no other good candidate for a mother body.

Also for questions of uniqueness of mother bodies conditions (2) and (3) appear often to be dispensable.

The last group of axioms, (4) and (5), are requirements only on the set supp μ. They imply that supp μ is minimal as a set (see Proposition 2.1 below), and they are necessary to guarantee any reasonable degree of uniqueness of mother bodies, e.g., to exclude occurrence of continuous families of them. A sharper form of (5), which together with (1) and (4) definitely guarantees uniqueness (see Proposition 2.1), is

(6) supp μ does not disconnect any open set

(i.e., $D \setminus \text{supp } \mu$ is connected whenever D is an open connected set). Clearly (6) implies (5). However, with requirement (6) in place of (5), mother bodies will exist more rarely (polyhedra will not admit mother bodies, for example). On the other hand, in cases when one allows distributional mother bodies (5) becomes too weak to even exclude continuous families of mother bodies and therefore has to be replaced by something stronger, like (6).

The strongest reasonable requirement in the direction of (4), (5), (6) is to require supp μ to consist of only finitely many points. This is what one (classically) requires of a "quadrature domain," namely that there exists a measure or distribution μ with finite support and satisfying some form of (1). The word quadrature domain is however also used in wider senses. See [Sh] for an overview.

In two dimensions, quadrature domains in the above (classical) sense can be produced as conformal images of the unit disc under rational mapping functions. (This is for the case $a = 0$, to which we stick for a moment.) Taking for example $\Omega = f(B(0,1))$, where $f(z) = z + c_2 z^2 + \cdots + c_n z^n$ is a univalent polynomial of degree $n \geq 2$ ($z = x_1 + ix_2$), (1) will hold with μ a distribution of order $n-1$ supported at the origin. Clearly also (4)–(6) will hold then, but (2) and (3) will fail. By an argument similar to the proof of Proposition 2.1 (iii), one realizes that there cannot simultaneously exist measures satisfying (1), (4), and (5).

Thus such a simple and smooth domain as the conformal image of the unit disc under a quadratic (or higher degree) polynomial does not admit a mother body in our sense. This is of course disappointing, but one has to keep in mind that the problem of finding a mother body is ill-posed and that the requirements (1)–(5) taken all together combine several different aspects of it (balayage, analytic continuation, minimality, etc.).

Indeed, as the following proposition shows, our axioms for a mother body seem to be fairly complete in the sense that they contain or imply many of the criteria for concentration and minimality which have been used previously for similar purposes. Examples of such criteria are minimality of supp μ as a set, largeness of U^μ (e.g., Kounchev [Ko2] maximizes integrallike $\int_\Omega U^\mu dx$ among all μ satisfying (1), (3)), and μ being an extremal point in a suitable convex set [An1],[An2], [Ka-Pi]. Proposition 2.1 shows (in particular) that if (1), (3), (4), (5) hold for a measure μ then, within the class of measures satisfying (1) and (3), supp μ is minimal, U^μ is maximal, and μ is an extremal point.

PROPOSITION 2.1. *Let μ be a measure satisfying* (1), (4), (5) *with respect to a given body Ω, and let ν, μ_1, μ_2 be (possibly) other measures.*

(i) *If ν satisfies* (1) *and supp $\nu \subset$ supp μ, then $\nu = \mu$.*

(ii) *If ν satisfies* (1), (3) *and $U^\nu \geq U^\mu$ in all $\mathbb{R}^N$, then $\nu = \mu$.*

(iii) *If ν satisfies* (1), (4), (6), *then $\nu = \mu$.*

(iv) *If μ_1, μ_2 satisfy* (1), (3) *and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$, then $\mu_1 = \mu_2 = \mu$.*

Proof. We first consider statements (i)–(iii). Since $U^\nu, U^\mu \in L^1_{loc}(\mathbb{R}^N)$ it is enough to prove that $U^\nu = U^\mu$ holds almost everywhere (a.e.), or by (4) a.e. in $\mathbb{R}^N \backslash \text{supp } \mu$. So let D be a component of $\mathbb{R}^N \backslash \text{supp } \mu$, set $u = U^\nu - U^\mu$, and we shall prove that $u = 0$ a.e. in D. Note that D meets Ω^e by (5) and that $u = 0$ in $D \cap \Omega^e$.

In case (i) u is harmonic in D; hence, it follows by harmonic continuation that $u = 0$ in all D. In case (ii) u is superharmonic and nonnegative in D; hence, it is either strictly positive in all D or vanishes identically in D. But the first alternative has already been excluded, and we again get $u = 0$ in D. In case (iii) u is harmonic in $D \backslash \text{supp } \nu$. Using (6) it follows that $u = 0$ in $D \backslash \text{supp } \nu$, hence, a.e. in D.

Proof of (iv): Since $\mu_1, \mu_2 \geq 0$ we have supp $\frac{1}{2}(\mu_1 + \mu_2) =$ supp $\mu_1 \cup$ supp μ_2. Thus supp $\mu_j \subset$ supp μ, and the conclusion follows immediately from (i). ∎

Note. If, in (i)–(iii) of the proposition, one allows μ and ν to be general distributions (instead of measures), then one still gets the conclusion that $U^\nu = U^\mu$ outside a compact set K of measure zero ($K =$ supp μ in cases (i) and (ii), $K =$ supp $\mu \cup$ supp ν in case (iii)). This means that $\mu - \nu$ annihilates all functions which are harmonic in some neighborhood of K, which is about as close to the conclusion $\nu = \mu$ as one may come in the case of distributions. Note that there are distributions with support at a single point, e.g., the Laplacian of the Dirac measure, whose potential vanishes identically outside that point.

3. Mother bodies for convex polyhedra. Having formulated precise requirements for mother bodies ((1)–(5) above) one naturally wonders which bodies admit mother bodies in that precise sense and when they are unique. This is a question which is largely open, but in this section we at least start answering it by proving that convex polyhedra always have unique mother bodies.

THEOREM 3.1. *Let $\Omega \subset \mathbb{R}^N$ be a convex bounded open polyhedron provided with a mass distribution ρ_Ω as in section* 2. *Then there exists a measure μ satisfying* (1)–(5). *Its support is contained in a finite union of hyperplanes and reaches $\partial\Omega$ only at corners and edges (not at faces), it has no mass on $\partial\Omega$, and U^μ is a Lipschitz continuous function. Moreover, μ is unique among all signed measures satisfying* (1), (4), (5).

Note. The support of μ coincides with what is sometimes called the "ridge" of Ω [Ev-Ha], [Ja]. For convex polyhedra this is the set of points in Ω which have at least two closest neighbors on $\partial\Omega$. See Figure 1 for an example in two dimensions.

Proof. Write $\Omega = \bigcap_{j=1}^m H_j$, where H_j are open half spaces and m is minimal. For any j, set

$$\delta_j(x) = \text{dist}\left(x, H_j^c\right),$$
$$u_j(x) = a\delta_j(x) + \frac{b}{2}\,\delta_j(x)^2.$$

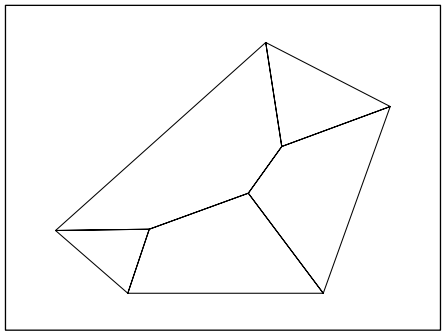

FIG. 1. *The mother body of a convex polyhedron.*

Then $u_j > 0$ in H_j, $u_j = 0$ on H_j^c, and $\Delta u_j = a\mathcal{H}^{N-1}\lfloor\partial H_j + b\mathcal{L}^N\lfloor H_j$. Set also

$$\begin{aligned}
\delta(x) &= \operatorname{dist}(x, \Omega^c) \\
&= \inf\{\delta_1(x), \dots, \delta_m(x)\}, \\
u(x) &= a\delta(x) + \frac{b}{2}\,\delta(x)^2 \\
&= \inf\{u_1(x), \dots, u_m(x)\}, \\
R &= \{x \in \Omega : \delta(x) = \delta_j(x) \text{ for at least two different } j\}, \\
D_j &= \{x \in \Omega \backslash R : \delta(x) = \delta_j(x)\} \\
&= \{x \in \Omega : \delta_j(x) < \delta_k(x) \text{ for all } k \neq j\}.
\end{aligned}$$

Note that u and u_j are strictly monotone functions of δ and δ_j, respectively (on the range $[0, +\infty)$).

We note that R (the "ridge") is contained in a finite union of hyperplanes, $\Omega = R \cup D_1 \cup \cdots \cup D_m$, u is Lipschitz continuous, $u > 0$ in Ω, $u = 0$ on Ω^c. Within Ω we have $\Delta u_j = b$ for all j, hence, $\Delta u \le b$ in Ω, using the principle that the infimum of a finite family of superharmonic functions (e.g., $u_j(x) - (b/2N)|x|^2$) is superharmonic. The same principle actually gives that $\Delta u \le b + \sum_{j=1}^m a\mathcal{H}^{N-1}\lfloor\partial H_j$ in all $\mathbb{R}^N$ and hence (since $u = 0$ on Ω^c) that $\Delta u \le b\mathcal{L}^N\lfloor\Omega + a\mathcal{H}^N\lfloor\partial\Omega = \rho_\Omega$. Outside $\overline{R}$ we have equality in this formula, as is easily seen. Thus $\Delta u = \rho_\Omega - \mu$ where μ is a positive measure with $\operatorname{supp}\mu \subset \overline{R}$. Since u vanishes at infinity and $\Delta u = \Delta\left(U^\mu - U^\Omega\right)$ we have $u = U^\mu - U^\Omega$. It now follows that μ satisfies (1)–(4) and that U^μ is Lipschitz continuous.

The latter property implies that μ is absolutely continuous with respect to $\mathcal{H}^{N-1}$. Since $\mathcal{H}^{N-1}\left(\overline{R} \cap \partial\Omega\right) = 0$ it follows that $\mu(\partial\Omega) = 0$.

To see finally that μ satisfies (5), take any $x \in \Omega$ and let $y \in \partial\Omega$ be a closest point on the boundary. Then $y \in \partial H_j$ for a unique j, and it is easy to see that the whole segment (x, y) is in D_j and that $y \notin \bar{R}$. Thus, if $x \notin \operatorname{supp}\mu$, the closed segment $[x, y + \varepsilon(y - x)]$ $(\varepsilon > 0)$ connects x with Ω^e without meeting $\operatorname{supp}\mu$, proving (5).

It remains to prove the uniqueness part of the theorem. Let μ, u, and u_j be as above, and let ν be any signed measure satisfying (1), (4), (5) (when stated for ν). Set $v = U^\nu - U^\Omega$. Then $v = 0 = u_j$ in H_j^e and $\Delta(v - u_j) = 0$ in $\operatorname{int}(H_j^c \cup \Omega) \setminus \operatorname{supp}\nu$. Set

$$\omega_j = \text{ the unbounded component of } \operatorname{int}(H_j^c \cup \Omega) \setminus \operatorname{supp}\nu.$$

It follows that

$$v = u_j \qquad \text{in} \qquad \omega_j \tag{7}$$

and also, since ω_j is open, that

$$\nabla v = \nabla u_j \qquad \text{in} \qquad \omega_j. \tag{8}$$

Assumption (5) for ν implies that

$$\omega_1 \cup \cdots \cup \omega_m = \mathbb{R}^N \setminus \operatorname{supp} \nu.$$

Since $\mathcal{L}^N(\operatorname{supp} \nu) = 0$ by (4) it follows that $\bigcup_{j=1}^m \omega_j$ is an open subset of $\mathbb{R}^N$ satisfying

$$\mathcal{L}^N \left(\mathbb{R}^N \setminus \bigcup_{j=1}^m \omega_j \right) = 0, \tag{9}$$

$$\bigcup_{j=1}^m \bar{\omega}_j = \mathbb{R}^N. \tag{10}$$

By (7), (8) v is continuously differentiable in $\bigcup_{j=1}^m \omega_j$ with

$$|\nabla v(x)| \le C < \infty \qquad \left(x \in \bigcup_{j=1}^m \omega_j \right). \tag{11}$$

Next we claim that the distributional gradient of v is a locally integrable function. To see this, note that $v = U^\nu - U^\Omega = E * (\nu - \rho_\Omega)$; hence,

$$\nabla v = (\nabla E) * (\nu - \rho_\Omega). \tag{12}$$

Here everything is to be interpreted in the sense of distributions. Now, ∇E is a locally integrable (vector) function and $\nu - \rho_\Omega$ is a signed Radon measure with compact support. It then follows (cf. [Do, Sect. 26]) from (12) that ∇v is also a locally integrable (vector) function.

Combining this information with (9), (11) we conclude that the distributional gradient ∇v is in $L^\infty\left(\mathbb{R}^N\right)$ and hence that v is a Lipschitz continuous function (i.e., has such a representative).

By continuity, for the Lipschitz continuous version of v, the relation (7) on ω_j extends to hold on all $\bar{\omega}_j$. Thus it follows from (10) that for each $x \in \mathbb{R}^N$ we have $v(x) = u_j(x)$ for some j.

Now let $x \in D_j$, and let y be the closest point on ∂H_j. Then $u_j(\xi) < u_k(\xi)$ for every $\xi \in [x, y]$ and for every $k \neq j$. On H_j^e, $v = u_j = 0$, so by continuity $v(y) = u_j(y)$. Since v is continuous and coincides everywhere with some u_k it follows that $v(\xi) = u_j(\xi)$ for all $\xi \in [x, y]$, in particular $v(x) = u_j(x)$. Thus $v = u_j = u$ in D_j. Since j was arbitrary we conclude that $v = u$ and $\nu = \mu$, completing the proof of the theorem. □

Example. Let Ω be a regular polygon in $\mathbb{R}^2$, say centered at the origin and with $n \ge 3$ corners uniformly distributed on the unit circle. Clearly Ω is convex. Let us compute its mother body μ.

The support of μ, i.e., the ridge R of Ω, consists of the n radii from the origin to the corners of Ω. The density of μ with respect to arclength on R equals the jump of the normal derivative of U^μ across R, or, what is the same, the jump of the normal derivative of $u = U^\mu - U^\Omega$. Since, in the notations of the proof above, $\nabla u = (a+b\delta)\nabla\delta$ and $\nabla\delta$ is a constant unit vector in each component of $\Omega\setminus R$, it follows that the density of μ is proportional to $a + b(1-r)$ on R, where $r = |x|$. Indeed, evaluation of the constant of proportionality gives that

$$d\mu = \frac{2\pi}{n}(a + b(1-r))dr \quad \text{on } R. \tag{13}$$

As n increases, Ω approaches the unit disc $B(0,1)$. One might hope then that μ should approach the unique mother body of the disc, namely $2\pi(a+\frac{b}{2})$ times the Dirac measure at the origin. However, one sees from (13) that, as $n \to \infty$, the μ converge towards that measure on $B(0,1)$ which has density $\frac{a}{r} + b(\frac{1}{r} - 1)$ with respect to area measure. This certainly is more concentrated than the original mass distribution $\rho_{B(0,1)}$, but less concentrated than the Dirac measure. In particular, the mother bodies of the regular polyhedra do not converge towards the mother body of the limiting disc.

This failure of convergence should not be surprising since, as was discussed in section 2, the search for mother bodies is an ill-posed problem with no continuous dependence on initial data, even when unique solutions do exist. The mother bodies for the approximating polyhedra may actually be more useful and more realistic in practical problems than the mother body for the disc itself. Consider, e.g., the case $a = 0$, $b = 1$ and think of the ill-posed Hele–Shaw model briefly discussed in section 2. In experiments with Hele–Shaw flows one never sees an initially circular blob shrinking down to a point. The predominant phenomenon always is that shrinking occurs by development of fingers of the exterior domain penetrating into the fluid (see, e.g., [Ho]). What eventually remains of the fluid domain is not a pointlike blob, but rather a kind of skeleton, which is somewhat reminiscent of the mother body of the approximating polygon Ω for a suitable n.

Thus there is a possibility that mother bodies of polyhedra could be a useful tool for handling ill-posed Hele–Shaw problems: one approximates a given initial fluid domain by a polygon, computes its mother body (uniquely determined and easily computed in the convex case), and then the whole evolution in time is obtained by balayage (section 2). The initial approximation with a polygon of course contains a degree of arbitrariness, but it is also known, for real Hele–Shaw flows, that the onset of the finger development contains a stochastic element.

4. General polyhedra and Zidarov's counterexample. By a (general) polyhedron we mean a domain which is the interior of a finite union of compact convex polyhedra. Mother bodies for general polyhedra will be treated in subsequent papers, e.g., [Gu-Sa2]. The situation in higher dimensions is not completely clear at present, but let us summarize what is known in the two-dimensional case.

When hypersurface measure is present in ρ_Ω, i.e., when $a > 0$, $b \geq 0$, nonconvex polyhedra do not admit mother bodies satisfying all of (1)–(5). Indeed, if Ω is a nonconvex polyhedron in $\mathbb{R}^2$, then Ω must have a nonconvex corner and it is well known that classical balayage of any positive measure μ in Ω onto $\partial\Omega$ will then be a measure on $\partial\Omega$ whose density with respect to $\mathcal{H}^1$ tends to infinity at the corner. When $b = 0$, requirement (1) means that ρ_Ω will have to coincide with this balayage measure; hence, a mother body μ cannot exist in this case. This argument extends

to the case $a > 0$, $b > 0$.

On the other hand, extending previous work of G. Choquet and I. Deny [Ch-De] concerning regular polyhedra, D. Siegel [Si] has constructed, in the pure hypersurface case ($a > 0$, $b = 0$), mother bodies (or skeletons, as he calls them) for general polyhedra which satisfy (1)–(2), (4)–(5). The construction actually works for general a, $b \geq 0$. Moreover the construction is canonical (involves no choices) and the shape of the mother body reflects that of the original body. Hence we feel that it is a satisfactory mother body, although the positivity requirement (3) is violated in the nonconvex case.

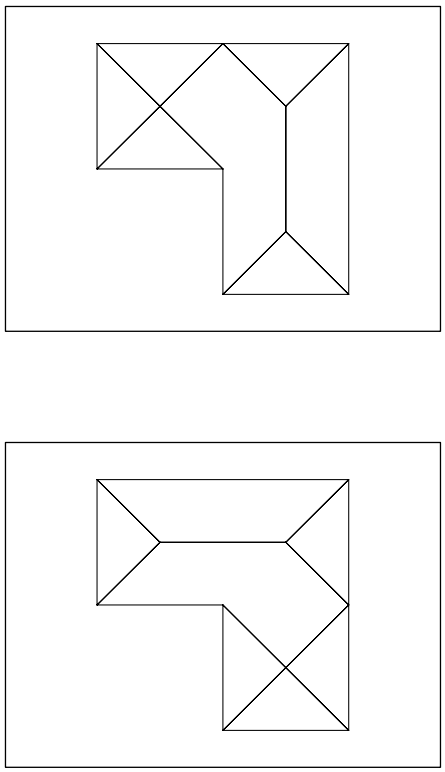

FIG. 2. *Two mother bodies for a nonconvex polyhedron in the case* $a = 0$ *(Zidarov's example).*

In the case of pure volume measure ($a = 0$, $b > 0$) the function $\Omega \mapsto \rho_\Omega$ is additive under disjoint unions (even after "removal of slits," i.e., after taking the interior of the closure). Therefore a possible way to construct a mother body for a polyhedron Ω is to decompose it into finitely many subpolyhedra, e.g., convex ones, each of which has a mother body satisfying (1)–(5). By adding these up one gets a measure μ which automatically satisfies (1)–(4) for Ω. Requirement (5) is more troublesome, but at least in the two-dimensional case it can be met by choosing the decomposition properly [Gu-Sa2].

In conclusion, mother bodies satisfying all of (1)–(5) do exist for arbitrary polyhedra when $N = 2$ and $a = 0$. However, as Zidarov discovered, there is no uniqueness of mother bodies for nonconvex polyhedra. Zidarov's example is a square in $\mathbb{R}^2$ with a smaller square at one corner removed, say $\Omega = (-1, 1)^2 \backslash (-1, 0]^2$. This can be de-

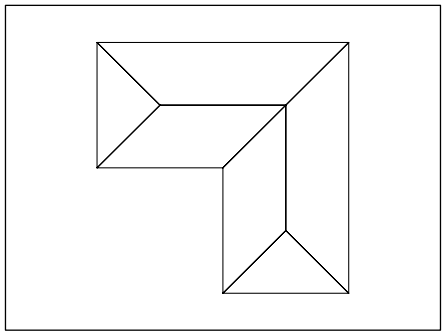

FIG. 3. *Siegel's mother body (with* (3) *violated) for the same nonconvex polyhedron* ($a, b \geq 0$).

composed into three squares (with side length one) in a natural way. Adding up the mother bodies for these, one gets a measure not satisfying (5). But, if Ω is instead decomposed into a rectangle (with side lengths one and two) and a square, the sum of the mother bodies for these will satisfy all of (1)–(5). This decomposition can be made in two different ways, and the result will be two different mother bodies.

These are depicted in Figure 2. Figure 3 shows the mother body obtained by Siegel's procedure for the same Ω. The latter does not satisfy (3), but it has other advantages, namely that supp μ meets $\partial\Omega$ only at corners not at smooth points of $\partial\Omega$, that it shares the symmetry properties of Ω, and that it is indecomposable in a certain sense.

Acknowledgments. I would like to dedicate this paper to the memory of Dimiter Zidarov, who died in 1993. His work in geophysical potential theory has for several years been one of my main sources of inspiration. I also want to thank Ognyan Kounchev and Harold S. Shapiro for many interesting discussions concerning mother bodies and Siv Sandvik for typing my manuscript.

REFERENCES

[An1] G. ANGER, *Direct and inverse problems in potential theory,* in Nonlinear Evolution Equations and Potential Theory (Proc. Summer School, Podhradí, 1973), Academia, Prague, 1975, pp. 11–44.

[An2] G. ANGER, *Lectures in potential theory and inverse problems*, in Geodätische und Geophysikalische Veroffentlichungen, Reihe 3, The National Committee for Geodesy and Geophysics, Acad. Sci. GDR, Berlin (1980), pp. 15–95.

[Ch-De] G. CHOQUET AND J. DENY, *Sur quelques proprétes de morgene caractéristic des fonctions harmoniques et polyharmoniques*, Bull. Soc. Math. France, 72 (1944), pp. 118–140.

[Do] W. DONOGHUE, *Distributions and Fourier Transforms*, Academic Press, New York, 1969.

[Eb] P. EBENFELT, *Singularities of the solution to a certain Cauchy problem and an application to the Pompeiu problem*, Duke Math. J., 71 (1993), pp. 119–142.

[Ev-Ha] W. D. EVANS AND D. J. HARRIS, *Sobolev embeddings for generalized ridged domains*, Proc. London Math. Soc., 54 (1987), pp. 141–175.

[Gu1] B. GUSTAFSSON, *A distortion theorem for quadrature domains for harmonic functions*, J. Math. Anal. Appl., 202 (1996), pp. 169–182.

[Gu2] B. GUSTAFSSON, *Direct and inverse balayage—some new developments in classical potential theory*, Nonlinear Anal., 30 (1997), pp. 2557–2565.

[Gu-Sa1] B. Gustafsson and M. Sakai, *Properties of some balayage operators with applications to quadrature domains and moving boundary problems*, Nonlinear Anal., 22 (1994), pp. 1221–1245.

[Gu-Sa2] B. Gustafsson and M. Sakai, *On Mother Bodies and Sticking Properties of Polyhedra*, Preprint, 1997.

[Gu-Sg] B. Gustafsson and H. Shahgholian, *Existence and geometric properties of solutions of a free boundary problem in potential theory*, J. Reine Angew. Math., 473 (1996), pp. 137–179.

[He] A. Henrot, *Subsolutions and supersolutions in free boundary problems*, Ark. Mat., 32 (1994), pp. 79–98.

[Ho] G. M. Homsy, *Viscous fingering in porous media*, Ann. Rev. Fluid Mech., 19 (1987), pp. 271–311.

[Ja] U. Janfalk, *On a minimization problem for vector fields in L^1*, Bull. London Math. Soc., 28 (1996), pp. 165–176.

[Ka-Pi] A. F. Karr and A. O. Pittenger, *An inverse balayage problem for Brownian motion*, Ann. Probab., 7 (1979), pp. 189–191.

[Kh-Sh] D. Khavinson and H. S. Shapiro, *The Schwarz Potential in $\mathbb{R}^n$ and Cauchy's Problem for the Laplace Equation*, Research report TRITA-MAT-1989-36, Royal Institute of Technology, Stockholm, 1989.

[Ko1] O. I. Kounchev, *The partial balayage materic bodies and optimization problems in gravimetry*, in Inverse Modeling in Exploration Geophysics, Proceedings of the 6th International Mathematical Geophysics Seminar, Berlin, 1988, A. Vogel, R. Gorenflo, B. Kummer, and C. O. Ofoegbu, eds., Vieweg & Sohn, Braunschweig/Wiesbaden.

[Ko2] O. I. Kounchev, *Obtaining materic bodies through concentration and optimization of a linear functional*, in Geophysical Data Inversion Methods and Applications, Proceedings of the 7th International Mathematical Geophysics Seminar, Berlin, 1989, A. Vogel, R. Gorenflo, C. O. Ofoegbu, and B. Ursin, eds., Vieweg & Sohn, Braunschweig/Wiesbaden.

[Re-Wo] M. Reissig and L. V. Wolfersdorf, *A simplified proof for a moving boundary problem for Hele-Shaw flows in the plane*, Ark. Mat., 31 (1993), pp. 101–116.

[Sa1] M. Sakai, *Quadrature Domains*, Lecture Notes in Math., 934, Springer-Verlag, Berlin, 1982.

[Sa2] M. Sakai, *Linear Combinations of Harmonic Measures and Quadrature Domains of Signed Measures with Small Supports*, Proc. Edinburgh Math. Soc, to appear.

[Sav-St-Sv] T. Savina, B. Sternin, and V. Shatalov, *Notes on "Mother Body" Problem in Geographics*, Preprint, 1995.

[Sg] H. Shahgholian *Quadrature surfaces as free boundaries*, Ark Mat., 32 (1994), pp. 475–492.

[Sh] H. S. Shapiro, *The Schwarz Function and Its Generalization to Higher Dimensions*, Wiley & Sons, New York, 1992.

[Si] D. Siegel, *Integration of Harmonic Functions Over Polygons*, Preprint, 1990.

[St-Sv] B. Sternin and V. Shatalov, *Differential Equations on Complex Manifolds*, Kluwer Academic Publishers, Dortrecht, 1994.

[Ti] F. R. Tian, *A Cauchy integral approach to Hele-Shaw problems with a free boundary: The case of zero surface tension*, Arch. Rational Mech. Anal., 135 (1996), pp. 175–196.

[Zi] D. Zidarov *Inverse Gravimetric Problem in Geoprospecting and Geodesy*, Elsevier, Amsterdam, 1990.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1118–1128, September 1998

004

A UNIQUENESS THEOREM FOR AN INVERSE SCATTERING PROBLEM IN AN EXTERIOR DOMAIN*

PETER HÄHNER†

Abstract. The scattering of time-harmonic acoustic waves by a simply connected, sound-soft obstacle in an inhomogeneous medium in $\mathbb{R}^2$ is considered. We prove that the Cauchy data of the scattered waves on a large circle for all incident waves and for an interval of wave numbers κ uniquely determine the obstacle and the inhomogeneity. To this end we show that products of harmonic functions satisfying a Dirichlet condition on the interior boundary of an annular plane domain are complete. Then, we use the limit $\kappa \to 0$ and obtain uniqueness of the obstacle from Schiffer's uniqueness result. Uniqueness of the inhomogeneity also follows from $\kappa \to 0$ together with the denseness result for the linear span of products of harmonic functions.

Key words. inverse obstacle scattering, inhomogeneous medium, complete set, products of harmonic functions

AMS subject classifications. 35R30, 35J05, 35J25

PII. S0036141097318614

1. Introduction.

We consider the scattering of time-harmonic waves by a simply connected obstacle in a two-dimensional inhomogeneous medium assuming a Dirichlet boundary condition. Problems of this type occur when a three-dimensional acoustic scattering problem at an infinitely long cylinder is reduced to two dimensions by assuming that the fields and the refractive index are constant along each line parallel to the cylinder. Similarly, the scattering of time-harmonic electromagnetic waves by an infinitely long cylindrical conductor leads to this problem assuming a constant magnetic permeability, the invariance of the electric permittivity in the direction of the cylinder axis and the electric field polarized in that direction.

We suppose that on a large circle, surrounding the obstacle, the Cauchy data of the scattered fields are known for an interval of wave numbers and for all possible incident waves. Our aim is to prove that these data uniquely determine the obstacle and the inhomogeneity.

Although there are many papers dealing with similar problems in the absence of an obstacle or investigating an inverse obstacle problem in a homogeneous medium to the author's knowledge there is only one paper which examines an inverse obstacle problem in an inhomogeneous medium. However, in [7] the authors assume that the inhomogeneity in the exterior domain is known and then prove uniqueness of the obstacle with techniques used in [6].

For the full space problem, i.e., without any obstacle, in three or more dimensions, Sylvester and Uhlmann and others even obtained uniqueness of the inhomogeneity assuming only the knowledge of the Cauchy data at a fixed wave number for all incident fields (see the review of results in [10]). In two dimensions there are only partial results of this type [10, 5] and no general uniqueness result using only one fixed wave number is known. Therefore, we use an interval of wave numbers κ. This means that we use more data than expected from a count of degrees of freedom. It

*Received by the editors March 24, 1997; accepted for publication (in revised form) August 27, 1997; published electronically April 14, 1998.
http://www.siam.org/journals/sima/29-5/31861.html
†Institut für Numerische und Angewandte Mathematik, Lotzestraße 16–18, D-37083 Göttingen, Germany (haehner@math.uni-goettingen.de).

is then possible to use the limit $\kappa \to 0$ to reduce the problem to two questions about harmonic functions.

First, do the Cauchy data of a suitable harmonic function satisfying a Dirichlet boundary condition uniquely determine the obstacle? The answer attributed to Schiffer is yes and can be found in [3, Theorem 5.1] in a slightly different setting.

Second, are linear combinations of products of harmonic functions, satisfying a Dirichlet condition on the interior boundary of an annular plane domain, dense in certain function spaces? Although we use it as an auxiliary result to prove uniqueness of the inverse scattering problem we think that this question is interesting in itself. A positive answer for this question without imposing the Dirichlet boundary condition was given by Calderón in [2]. We shall give a positive answer for our case in the next section. We first assume the obstacle to be a disk and use separation of variables for the density result. For the general case we use a conformal mapping.

In the third section we state existence and uniqueness of the direct scattering problem and examine the behavior of the solutions as $\kappa \to 0$. There are some difficulties due to the fact that in $\mathbb{R}^2$ the fundamental solutions for the Helmholtz equation do not converge as $\kappa \to 0$.

Finally, in the fourth section we combine the results of the previous sections to prove the uniqueness result for the inverse scattering problem.

2. Completeness of products of harmonic functions. Let $D_0 \subset \mathbb{R}^2$ be a nonempty, open, connected, and bounded set. We assume the exterior domain $D := \mathbb{R}^2 \setminus \overline{D_0}$ to be connected and to have a C^2-smooth boundary $\partial D = \partial D_0$. For $R > 0$ we define the disk $B_R := \{x \in \mathbb{R}^2 : |x| < R\}$. In the sequel we suppose that R is sufficiently large to ensure $\overline{D_0} \subset B_R$. For a bounded domain, n denotes the normal vector at the boundary directed into the exterior of the domain.

The main result of this section is the following theorem.

THEOREM 2.1. *Let $q \in L^2(D)$ satisfy $q(x) = 0$ for all $|x| \geq R$. Furthermore, assume that*

$$\int_{D \cap B_{2R}} q(x)u(x)v(x)dx = 0 \tag{2.1}$$

for all functions u, $v \in C^2(D \cap B_{2R}) \cap C^1(\overline{D} \cap B_{2R})$ satisfying $\Delta u = 0$, $\Delta v = 0$ in $D \cap B_{2R}$, $u|_{\partial D} = v|_{\partial D} = 0$, and $\int_{\partial D} \frac{\partial u}{\partial n} ds = \int_{\partial D} \frac{\partial v}{\partial n} ds = 0$. Then $q(x) = 0$ almost everywhere in D.

We shall reduce this theorem to the special case $D = \mathbb{R}^2 \setminus \overline{B_1}$ with the help of a conformal mapping. Hence, we first examine the special case $D_0 = B_1$. Using polar coordinates $x = (x_1, x_2) = (r\cos\theta, r\sin\theta)$ for $x \in \mathbb{R}^2$ it is easily seen that the functions

$$u_l(x) = (x_1 + ix_2)^l - (x_1 - ix_2)^{-l} = (r^l - r^{-l})e^{il\theta} \ , \ l \in \mathbb{N} \ , \ x \in \mathbb{R}^2 \setminus B_1 \ ,$$

$$u_l(x) = (x_1 - ix_2)^{|l|} - (x_1 + ix_2)^{-|l|} = (r^{|l|} - r^{-|l|})e^{il\theta} \ , \ -l \in \mathbb{N} \ , \ x \in \mathbb{R}^2 \setminus B_1 \ ,$$

satisfy $\Delta u_l = 0$ in $\mathbb{R}^2 \setminus \overline{B_1}$, $u_l|_{\partial B_1} = 0$, and $\int_{\partial B_1} \frac{\partial u_l}{\partial n} ds = 0$.

Equation (2.1) reads for $u = u_l$, $v = u_m$,

$$\int_1^{2R} r \int_0^{2\pi} q(r\cos\theta, r\sin\theta)e^{i(l+m)\theta} d\theta (r^{|l|} - r^{-|l|})(r^{|m|} - r^{-|m|})dr = 0 \ .$$

For a fixed $k \in \mathbb{N}_0$ we choose $l = k + j$ and $m = -j$, $j \in \mathbb{N}$, and obtain

$$\int_1^{2R} r \int_0^{2\pi} q(r\cos\theta, r\sin\theta) e^{ik\theta} d\theta (r^{|k|+j} - r^{-(|k|+j)})(r^j - r^{-j}) dr = 0 \ , \ j \in \mathbb{N} \ . \tag{2.2}$$

Choosing $l = k - j$, $m = j$, $j \in \mathbb{N}$, in the case $-k \in \mathbb{N}$ we find that the equations (2.2) hold for all $k \in \mathbb{Z}$, $j \in \mathbb{N}$. Therefore, we first examine whether a function $f \in L^2(1, \rho)$ satisfying

$$\int_1^{\rho} f(r)(r^{k+j} - r^{-(k+j)})(r^j - r^{-j}) dr = 0 \ , \ j \in \mathbb{N} \ ,$$

for a fixed $k \in \mathbb{N}_0$ must vanish almost everywhere in $(1, \rho)$.

To this end we define for $j \in \mathbb{N}$, $k \in \mathbb{N}_0$, $r \geq 1$ the functions

$$\begin{aligned} e_{j,k}(r) &:= (r^{k+j} - r^{-(k+j)})(r^j - r^{-j}) \ , \\ f_j(r) &:= (r + r^{-1})(r - r^{-1})^{2j} \ , \\ g_j(r) &:= (r - r^{-1})^{2j} \ . \end{aligned}$$

The $e_{j,k}$ appear in the above equation but they are difficult to work with. However, it turns out that the f_j, g_j are contained in $\operatorname{span}\{e_{j,1} : j \in \mathbb{N}\}$, $\operatorname{span}\{e_{j,0} : j \in \mathbb{N}\}$, respectively. Since the f_j, g_j are essentially even polynomials in the variable $(r - 1/r)$ they are more suitable for our purposes than the $e_{j,k}$. Hence, we start with the following lemma.

LEMMA 2.2. $\operatorname{span}\{g_j : j \in \mathbb{N}\} \subset \operatorname{span}\{e_{j,0} : j \in \mathbb{N}\}$ *and* $\operatorname{span}\{f_j : j \in \mathbb{N}\} \subset \operatorname{span}\{e_{j,1} : j \in \mathbb{N}\}$.

Proof. We observe with the help of

$$(r + r^{-1})(r^j - r^{-j}) = (r^{j+1} - r^{-(j+1)}) + (r^{j-1} - r^{-(j-1)}) \ , \ j \in \mathbb{N} \ ,$$

that

$$(r + r^{-1})^2 e_{1,0}(r) = e_{2,0}(r) \ , \tag{2.3}$$

$$(r + r^{-1})^2 e_{j,0}(r) = e_{j+1,0}(r) + 2e_{j,0}(r) + e_{j-1,0}(r) - 2e_{1,0}(r) \ , \ j \geq 2 \ , \tag{2.4}$$

$$(r + r^{-1})^2 e_{1,1}(r) = e_{2,1}(r) + e_{1,1}(r) \ , \tag{2.5}$$

$$(r + r^{-1})^2 e_{j,1}(r) = e_{j+1,1}(r) + 2e_{j,1}(r) + e_{j-1,1}(r) - e_{1,1}(r) \ , \ j \geq 2 \ . \tag{2.6}$$

Starting with

$$g_1(r) = e_{1,0}(r) \ , \ f_1(r) = (r + r^{-1})(r - r^{-1})^2 = e_{1,1}(r)$$

we now use induction to prove $f_N \in \operatorname{span}\{e_{j,1} : j = 1, \ldots, N\}$ and $g_N \in \operatorname{span}\{e_{j,0} : j = 1, \ldots, N\}$ for all $N \in \mathbb{N}$. From $f_{N-1} = \sum_{j=1}^{N-1} \alpha_j e_{j,1}$ for a fixed $N \geq 2$ we conclude

$$\begin{aligned} f_N(r) &= (r - r^{-1})^2 f_{N-1}(r) \\ &= ((r + r^{-1})^2 - 4) f_{N-1}(r) \\ &= \sum_{j=1}^{N-1} \alpha_j (r + r^{-1})^2 e_{j,1}(r) - 4 \sum_{j=1}^{N-1} \alpha_j e_{j,1}(r) \ , \end{aligned}$$

i.e., $f_N = \sum_{j=1}^{N} \beta_j e_{j,1}$ by (2.5) and (2.6). Using (2.3) and (2.4), a similar reasoning yields the assertion for g_N and we have proved the lemma. □

Now, we prove the density result for $\operatorname{span}\{e_{j,k}: j \in \mathbb{N}\}$ with respect to the L^2-norm. $C_0(1,\rho)$ denotes the space of continuous functions having compact support in $(1,\rho)$.

LEMMA 2.3. *Assume $k \in \mathbb{N}_0$ and $\rho > 1$. Then, $\operatorname{span}\{e_{j,k}: j \in \mathbb{N}\}$ is dense in $L^2(1,\rho)$; i.e., if $f \in L^2(1,\rho)$ satisfies*

$$\int_1^\rho f(r) e_{j,k}(r)\,dr = 0 \quad \text{for all } j \in \mathbb{N},$$

then $f = 0$.

Proof. We start with the case $k = 0$ and show that for any $f \in C_0(1,\rho)$ there is a sequence $q_N \in \operatorname{span}\{e_{j,0}: j \in \mathbb{N}\}$, $N \in \mathbb{N}$, with $\|q_N - f\|_{L^2} \to 0$, $N \to \infty$. Then, the assertion for $k = 0$ follows from the density of $C_0(1,\rho)$ in $L^2(1,\rho)$.

The transformation $t = r - 1/r$, $r > 1$, maps $(1,\infty)$ bijectively onto $(0,\infty)$. The inverse map is given by $r = z(t)$ with $z(t) := t/2 + \sqrt{1 + t^2/4}$, $t > 0$. We use this transformation to reduce the problem to an approximation problem with even polynomials. Defining $g(t) := t^{-2} f(z(|t|))$ for $0 < |t| \le \rho - 1/\rho$ and $g(0) := 0$, we have that $g \in C[-\rho + 1/\rho, \rho - 1/\rho]$ is an even function. Hence, by the Weierstrass approximation theorem there is a sequence $\tilde{q}_N$, $N \in \mathbb{N}$, of even polynomials which converge uniformly to g,

$$\max\{|\tilde{q}_N(t) - g(t)|: t \in [-\rho + 1/\rho, \rho - 1/\rho]\} \to 0\ ,\ N \to \infty\ .$$

With $q_N(r) := (r - 1/r)^2 \tilde{q}_N(r - 1/r)$, $1 \le r \le \rho$, $N \in \mathbb{N}$, we have $q_N \in \operatorname{span}\{g_j: j \in \mathbb{N}\}$ and thus $q_N \in \operatorname{span}\{e_{j,0}: j \in \mathbb{N}\}$ by Lemma 2.2. Furthermore, using the transformation $r = z(t)$ we can estimate

$$\int_1^\rho |q_N(r) - f(r)|^2\,dr = \int_0^{\rho - 1/\rho} |\tilde{q}_N(t) - g(t)|^2 t^4 z'(t)\,dt$$

$$\le c \max\{|\tilde{q}_N(t) - g(t)|^2: t \in [0, \rho - 1/\rho]\} \to 0\ ,\ N \to \infty\ ,$$

with a suitable positive constant c and we have proved the lemma for $k = 0$.

Now, we consider the case $k \in \mathbb{N}$. For $f \in C_0(1,\rho)$ we define

$$g(t) := t^{-2}\Big(z(|t|) + z(|t|)^{-1}\Big)^{-1} f((z(|t|))^{1/k})$$

for $0 < |t| \le \rho^k - 1/\rho^k$ and $g(0) := 0$. We have again that $g \in C[-\rho^k + 1/\rho^k, \rho^k - 1/\rho^k]$ is an even function which can be approximated uniformly by a sequence $\tilde{q}_N$, $N \in \mathbb{N}$, of even polynomials,

$$\max\{|\tilde{q}_N(t) - g(t)|: t \in [-\rho^k + 1/\rho^k, \rho^k - 1/\rho^k]\} \to 0\ ,\ N \to \infty\ .$$

Defining $q_N(s) := (s + 1/s)(s - 1/s)^2 \tilde{q}_N(s - 1/s)$, $1 \le s \le \rho^k$, $N \in \mathbb{N}$, we have $q_N \in \operatorname{span}\{f_j: j \in \mathbb{N}\}$ and $q_N \in \operatorname{span}\{e_{j,1}: j \in \mathbb{N}\}$ by Lemma 2.2. Hence, $p_N(r) := q_N(r^k)$, $1 \le r \le \rho$, can be represented as

$$p_N(r) = \sum_{j=1}^{M} a_j e_{j,1}(r^k) = \sum_{j=1}^{M} a_j (r^{k+jk} - r^{-(k+jk)})(r^{jk} - r^{-jk})\ ,$$

i.e., $p_N \in \operatorname{span}\{e_{j,k}: j \in \mathbb{N}\}$.

With the help of the transformations $r = s^{1/k}$ and $s = z(t)$ we can estimate

$$\int_1^\rho |p_N(r) - f(r)|^2 dr = \int_1^{\rho^k} |q_N(s) - f(s^{1/k})|^2 (1/k) s^{1/k-1} ds$$

$$= \int_1^{\rho^k} \Big| \tilde{q}_N\Big(s - \frac{1}{s}\Big) - \Big(s + \frac{1}{s}\Big)^{-1} \Big(s - \frac{1}{s}\Big)^{-2} f(s^{1/k}) \Big|^2 \Big(s + \frac{1}{s}\Big)^2 \Big(s - \frac{1}{s}\Big)^4 \frac{1}{k} s^{1/k-1} ds$$

$$= \int_0^{\rho^k - 1/\rho^k} |\tilde{q}_N(t) - g(t)|^2 w(t) dt$$

$$\le c \max\{|\tilde{q}_N(t) - g(t)|^2 : t \in [0, \rho^k - 1/\rho^k]\} \to 0 \ , \ N \to \infty \ ,$$

with a bounded function w and a positive constant c. This ends the proof of the lemma. □

Since we shall map $B_{2R} \cap D$ conformally on an annulus during the proof of Theorem 2.1 we state the needed results about the conformal mapping and its inverse in the following lemma and briefly sketch its proof.

LEMMA 2.4. *Let $G := D \cap B_{2R}$, D, and R as in Theorem 2.1. Then there exist a constant $\rho > 1$ and a conformal mapping*

$$\varphi: \{z \in \mathbb{C}: (\Re z, \Im z) \in G\} \to \{w \in \mathbb{C}: 1 < |w| < \rho\} \ .$$

Furthermore, $\varphi \in C^1(\overline{G})$ and, denoting by ψ the inverse function to φ , $|\psi'|^2$ is bounded in $\{w \in \mathbb{C}: 1 < |w| < \rho\}$, where ψ' is the complex derivative of ψ.

Proof. Let $u \in C^2(G) \cap C^1(\overline{G})$ be the unique solution of the Dirichlet problem $\Delta u = 0$ in G, $u|_{\partial D} = 0$, $u|_{\partial B_{2R}} = 1$. Uniqueness of the solution can be inferred from the maximum principle. Existence and $C^1(\overline{G})$-regularity of the solution can be obtained with the help of a single-layer ansatz [9, Theorems 7.29, 7.27]. Defining $k := \int_{\partial B_{2R}} \frac{\partial u}{\partial n} ds$, we conclude from Green's theorem

$$k = \int_{\partial G} \frac{\partial u}{\partial n} \overline{u} ds = \int_G |\nabla u|^2 dx > 0 \ .$$

Let v be a real-valued conjugate harmonic function to u, i.e., u and v satisfy the Cauchy–Riemann equations, and define

$$\varphi(x_1 + ix_2) := \exp((2\pi/k)(u(x_1, x_2) + iv(x_1, x_2))) \ \text{ for } (x_1, x_2) \in G \ .$$

Note that v is not single-valued. However, φ is a single-valued holomorphic function in $\{z \in \mathbb{C}: (\Re z, \Im z) \in G\}$ and $\varphi \in C^1(\overline{G})$. Then φ is the searched-for conformal mapping with $\rho := e^{2\pi/k} > 1$ (see the proof of [1, Theorem 10, Chapter 6, p. 247]).

If ψ is the inverse function to φ we have $\varphi'(\psi(w))\psi'(w) = 1$ for $1 < |w| < \rho$, hence $|\varphi'(z)|^2 > 0$ for $z \in G$. The boundary point lemma states $\frac{\partial u}{\partial n}(z) \neq 0$ for $z \in \partial G$ (see

[4, Lemma 3.4]). Therefore we compute $|\varphi'(z)|^2 = ((2\pi/k)e^{(2\pi/k)u(z)})^2|\nabla u(z)|^2 > 0$ for $z \in \partial G$. Then, $|\varphi'(z)|^2 \geq \delta > 0$ for all $z \in \overline{G}$ and $|\psi'(w)|^2 \leq \delta^{-1}$ for $1 < |w| < \rho$. This ends the proof of the lemma. □

We are now in a position to prove Theorem 2.1.

Proof of Theorem 2.1. Let us first consider the case $D_0 = B_1$ and assume equation (2.1) holds for all harmonic functions u, v in $\mathbb{R}^2 \setminus \overline{B_1}$ which vanish on ∂B_1 and satisfy $\int_{\partial B_1} \frac{\partial u}{\partial n} ds = \int_{\partial B_1} \frac{\partial v}{\partial n} ds = 0$. By the reasoning immediately after Theorem 2.1 we arrive at equation (2.2), which reads

$$\int_1^{2R} \left(r \int_0^{2\pi} q(r\cos\theta, r\sin\theta) e^{ik\theta} d\theta \right) e_{j,|k|}(r) dr = 0 \ , \ j \in \mathbb{N} \ , \ k \in \mathbb{Z} \ .$$

We define, for $k \in \mathbb{Z}$,

$$f_k(r) := r \int_0^{2\pi} q(r\cos\theta, r\sin\theta) e^{ik\theta} d\theta \ , \ 1 < r < 2R \ ,$$

and obtain from Lemma 2.3 for all $k \in \mathbb{Z}$: $f_k(r) = 0$ almost everywhere in $(1, 2R)$. Since the trigonometric polynomials are a complete orthogonal system in $L^2(0, 2\pi)$ we can conclude $q(r\cos\theta, r\sin\theta) = 0$ for almost all $1 < r < 2R$, $0 \leq \theta \leq 2\pi$, i.e., $q = 0$.

In the general case we use the conformal mapping φ with inverse ψ from Lemma 2.4. For any function $\tilde{u} \in C^2(\mathbb{R}^2 \setminus \overline{B_1}) \cap C^1(\mathbb{R}^2 \setminus B_1)$ with $\tilde{u}|_{\partial B_1} = 0$, $\Delta\tilde{u} = 0$ in $\mathbb{R}^2 \setminus \overline{B_1}$ and $\int_{\partial B_1} \frac{\partial \tilde{u}}{\partial n} ds = 0$, the function

$$u(x) := \tilde{u}(\Re\varphi(x_1 + ix_2), \Im\varphi(x_1 + ix_2)) \ , \ x = (x_1, x_2) \in \overline{D} \cap B_{2R} \ ,$$

satisfies $u \in C^2(D \cap B_{2R}) \cap C^1(\overline{D} \cap B_{2R})$, $u(x) = 0$, $x \in \partial D$, and

$$\Delta u(x) = (\Delta\tilde{u})(\Re\varphi(x_1 + ix_2), \Im\varphi(x_1 + ix_2))|\varphi'(x_1 + ix_2)|^2 = 0 \ \text{ in } D \cap B_{2R} \ .$$

Due to $\int_{\partial B_1} \frac{\partial \tilde{u}}{\partial n} ds = 0$ there is a single-valued conjugate harmonic function $\tilde{w} \in C^2(\mathbb{R}^2 \setminus \overline{B_1}) \cap C^1(\mathbb{R}^2 \setminus B_1)$ to $\tilde{u}$. Hence, defining

$$f(x_1 + ix_2) := \tilde{u}(\Re\varphi(x_1 + ix_2), \Im\varphi(x_1 + ix_2)) + i\tilde{w}(\Re\varphi(x_1 + ix_2), \Im\varphi(x_1 + ix_2))$$

for $x = (x_1, x_2) \in \overline{D} \cap B_{2R}$, f is a single-valued holomorphic function and we can compute

$$\int_{\partial D} \frac{\partial u}{\partial n} ds = \int_{\partial D} \frac{\partial(\Re(f))}{\partial n} ds = \int_{\partial D} \frac{d(\Im(f))}{ds} ds = 0;$$

i.e., the condition for the normal derivative is also satisfied.

Analogously, we can define v from $\tilde{v} \in C^2(\mathbb{R}^2 \setminus \overline{B_1}) \cap C(\mathbb{R}^2 \setminus B_1)$ with $\tilde{v}|_{\partial B_1} = 0$, $\int_{\partial B_1} \frac{\partial \tilde{v}}{\partial n} ds = 0$, and $\Delta\tilde{v} = 0$ in $\mathbb{R}^2 \setminus \overline{B_1}$. Using the change of variables formula and the fact that $|\psi'(y_1 + iy_2)|^2$ is the Jacobian of the transformation in $\mathbb{R}^2$ which is given by ψ, we obtain

$$0 = \int_{D \cap B_{2R}} q(x)u(x)v(x)dx$$

$$= \int\limits_{1<|y|<\rho} q((\Re\psi(y_1+iy_2),\Im\psi(y_1+iy_2)))\tilde{u}(y)\tilde{v}(y)|\psi'(y_1+iy_2)|^2 dy$$

for all harmonic functions $\tilde{u}$, $\tilde{v}$ in $\mathbb{R}^2\setminus\overline{B_1}$ which vanish on ∂B_1 and whose integral over ∂B_1 of the normal derivative is zero. Therefore, from the considerations for $D_0 = B_1$ we can conclude $q((\Re\psi(y_1+iy_2),\Im\psi(y_1+iy_2)))|\psi'(y_1+iy_2)|^2 = 0$, almost everywhere in $1 < |y| < \rho$, hence $q = 0$, and we have proved Theorem 2.1. □

3. The direct scattering problem. With D and B_R as in the previous section we shall examine the following scattering problems (P_κ): given $\kappa > 0$, a uniformly Hölder continuous function $q \in C^{0,\alpha}(\overline{D})$, $0 < \alpha < 1$, satisfying $\operatorname{supp}(1-q) \subset B_R$ and $\Im(q) \geq 0$ in D, and an incident field $u^i \in C^2(B_R)$ with $\Delta u^i + \kappa^2 u^i = 0$ in B_R, find the scattered field $u^s \in C^2(D) \cap C(\overline{D})$ satisfying $\Delta u^s + \kappa^2 q u^s = \kappa^2(1-q)u^i$ in D; the Sommerfeld radiation condition $|\hat{x}\cdot\nabla u^s(x) - i\kappa u^s(x)|^2 = o(|x|^{-1/2})$, $|x| \to \infty$, uniformly for all directions $\hat{x} := |x|^{-1}x$; and the boundary condition $u^s|_{\partial D} = -u^i|_{\partial D}$.

We are interested in the unique solvability of (P_κ) and in the behavior of the solutions as $\kappa \to 0$. Therefore, we also state (P_0): given an incident field $u^i \in C^2(B_R)$ with $\Delta u^i = 0$ in B_R, find the scattered field $u^s \in C^2(D) \cap C(\overline{D})$ satisfying $\Delta u^s = 0$ in D; $|u^s(x)| = O(1)$, $|x| \to \infty$, uniformly for all directions $\hat{x} := |x|^{-1}x$; and the boundary condition $u^s|_{\partial D} = -u^i|_{\partial D}$.

We denote by

$$\Phi_\kappa(x,y) := \frac{i}{4} H_0^{(1)}(\kappa|x-y|)\ ,\ x,y \in \mathbb{R}^2\ ,\ x \neq y\ ,\ \kappa > 0,$$

and

$$\Phi_0(x,y) := \frac{1}{2\pi}\ln\frac{1}{|x-y|}\ ,\ x,y \in \mathbb{R}^2\ ,\ x \neq y,$$

the fundamental solutions to the Helmholtz equation and Laplace equation, respectively.

THEOREM 3.1. *(P_κ) has a unique solution for all $\kappa \geq 0$.*

Let $u_0^i \in C^2(B_R)$ be a harmonic function in B_R. If for $0 < \kappa < 1$ the incident fields u_κ^i satisfy $\|u_\kappa^i - u_0^i\|_{\infty,B_R} \to 0$, $\kappa \to 0$, then the solutions u_κ^s of (P_κ) converge uniformly in $\overline{D\cap B_R}$ to the solution of (P_0) for the incident field u_0^i.

Proof. The existence and uniqueness results can be found in the literature. The case $\kappa = 0$ is treated in [9, Theorem 6.20]. Since the uniqueness proofs given for the three-dimensional case in [11, Satz 2] or in [3, Theorem 3.7] can be carried over to our case (P_κ) has at most one solution if $\kappa > 0$. Moreover, analogously to [11, Satz 10], the existence of a solution to (P_κ), $\kappa > 0$, can be established by an ansatz of a combined volume and double-layer potential. However, we are also interested in the behavior of the solutions for $\kappa \to 0$, and the second assertion of the theorem cannot be inferred from the ansatz in [11] (see [12, 8] for the case $q = 1$). Therefore, we choose a more complicated ansatz along the lines suggested in [8] which yields the desired convergence.

For $0 < \kappa < 1$, $a \in C(\overline{D\cap B_R})$, and $\varphi \in C(\partial D)$ we define

$$u^s(x) := \int\limits_{\partial D} \Big\{\frac{\partial\Phi_\kappa}{\partial n(y)}(x,y) + \Big(1 - \frac{2\pi}{\ln\kappa}\Big)\Phi_\kappa(x,y)\Big\}\varphi(y)ds(y)$$
$$-\frac{1}{|\partial D|}\int\limits_{\partial D}\Phi_\kappa(x,z)ds(z)\int\limits_{\partial D}\varphi(y)ds(y)$$

$$+\kappa \int_{D\cap B_R} \Phi_\kappa(x,y)a(y)dy \ , \ x \in D \ . \tag{3.1}$$

The normal vector n is directed into D and by $|\partial D|$ we denote the arclength of ∂D.

From the mapping properties and jump relations of potentials we can conclude that if a and φ are a solution of the integral equations

$$\begin{aligned} -a(x) + \kappa(q(x)-1)\Bigg\{ & \int_{\partial D} \Big\{ \frac{\partial \Phi_\kappa}{\partial n(y)}(x,y) + \Big(1 - \frac{2\pi}{\ln \kappa}\Big)\Phi_\kappa(x,y)\Big\}\varphi(y)ds(y) \\ & - \frac{1}{|\partial D|}\int_{\partial D} \Phi_\kappa(x,z)ds(z) \int_{\partial D} \varphi(y)ds(y) \\ & + \kappa \int_{D\cap B_R} \Phi_\kappa(x,y)a(y)dy \Bigg\} = \kappa(1-q(x))u^i(x) \ , \ x \in D \cap B_R, \end{aligned} \tag{3.2}$$

$$\begin{aligned} \varphi(x) + 2\Bigg\{ & \int_{\partial D} \Big\{ \frac{\partial \Phi_\kappa}{\partial n(y)}(x,y) + \Big(1 - \frac{2\pi}{\ln \kappa}\Big)\Phi_\kappa(x,y)\Big\}\varphi(y)ds(y) \\ & - \frac{1}{|\partial D|}\int_{\partial D} \Phi_\kappa(x,z)ds(z) \int_{\partial D} \varphi(y)ds(y) \\ & + \kappa \int_{D\cap B_R} \Phi_\kappa(x,y)a(y)dy \Bigg\} = -2u^i(x) \ , \ x \in \partial D, \end{aligned} \tag{3.3}$$

then u^s as in (3.1) is a solution to (P_κ).

Using the asymptotic behavior of $H_0^{(1)}(z)$ for $|z| \to 0$ (see [8]), for $\kappa \to 0$ we obtain the operator

$$\begin{pmatrix} -I & 0 \\ 0 & I + L_0 \end{pmatrix} : C(\overline{D\cap B_R}) \times C(\partial D) \to C(\overline{D\cap B_R}) \times C(\partial D)$$

as the uniform limit of the integral operators in equations (3.2) and (3.3). Here, I denotes the identity map on $C(\overline{D\cap B_R})$ and $C(\partial D)$, respectively, and $L_0 : C(\partial D) \to C(\partial D)$ is defined by

$$\begin{aligned} (L_0\varphi)(x) := 2 & \int_{\partial D} \Big\{ \frac{\partial \Phi_0}{\partial n(y)}(x,y) + \Phi_0(x,y) + 1\Big\}\varphi(y)ds(y) \\ & - \frac{2}{|\partial D|}\int_{\partial D} \Phi_0(x,z)ds(z) \int_{\partial D} \varphi(y)ds(y) \ , \ x \in \partial D. \end{aligned}$$

Since $I + L_0$ has a bounded inverse [8, Theorem 2.1] we know by a Neumann series argument that for sufficiently small κ the integral equations (3.2), (3.3) have a unique solution and that these solutions $(a_\kappa, \varphi_\kappa)$ converge to $(0, \varphi_0)$, φ_0 being the unique solution to $(I + L_0)\varphi_0 = -2u_0^i|_{\partial D}$. Therefore, inserting $(a_\kappa, \varphi_\kappa)$ into the potential in (3.1) and using once more the behavior of $H_0^{(1)}(z)$ for $|z| \to 0$ we see that the solutions

u_κ^s converge uniformly in $D \cap B_R$ to the function

$$u_0^s(x) := \int_{\partial D} \Big\{ \frac{\partial \Phi_0}{\partial n(y)}(x,y) + \Phi_0(x,y) + 1 \Big\} \varphi_0(y) ds(y) \\ - \frac{1}{|\partial D|} \int_{\partial D} \Phi_0(x,z) ds(z) \int_{\partial D} \varphi_0(y) ds(y) \ , \ x \in D,$$

(see [8, Theorem 2.2]). u_0^s is harmonic and bounded in D and satisfies $u_0^s|_{\partial D} = (1/2)(I + L_0)\varphi_0 = -u_0^i|_{\partial D}$ due to the jump relations. Hence, it is the solution of (P_0) for the incident field u_0^i. This ends the proof of the theorem. □

Now, we know the behavior of the scattered fields as $\kappa \to 0$. In the next section we also need some knowledge concerning the approximation of a harmonic function by incident fields or by the sum of incident and scattered fields as $\kappa \to 0$. We establish these results in the last theorem of this section.

THEOREM 3.2.

(a) *Let $u_0^i \in C^2(\overline{B_{3R/2}})$ be harmonic. Then, there are incident fields $u_\kappa^i \in C^2(B_{3R/2})$, $0 < \kappa < 1$, such that $\|u_\kappa^i - u_0^i\|_{\infty, B_R} \to 0$, $\kappa \to 0$.*

(b) *Let $u_0 \in C^2(D \cap B_{2R}) \cap C^1(\overline{D} \cap B_{2R})$ be harmonic in $D \cap B_{2R}$, $u_0|_{\partial D} = 0$, and $\int_{\partial D} \frac{\partial u_0}{\partial n} ds = 0$. Then, there are incident fields $u_\kappa^i \in C^2(B_{3R/2})$, $0 < \kappa < 1$, such that the solutions u_κ^s to (P_κ) satisfy $\|u_\kappa^i + u_\kappa^s - u_0\|_{\infty, D \cap B_R} \to 0$, $\kappa \to 0$.*

Proof. For part (a) we define $\varphi \in C(\partial B_{3R/2})$ to be the unique solution to

$$\varphi(x) - 2 \int_{\partial B_{3R/2}} \frac{\partial \Phi_0}{\partial n(y)}(x,y) \varphi(y) ds(y) = -2u_0^i(x) \ , \ x \in \partial B_{3R/2}.$$

It is well known that this integral equation has a unique solution [9, Theorem 6.16] and that

$$u_0^i(x) = \int_{\partial B_{3R/2}} \frac{\partial \Phi_0}{\partial n(y)}(x,y) \varphi(y) ds(y) \ , \ x \in B_{3R/2}.$$

Now, we set

$$u_\kappa^i(x) = \int_{\partial B_{3R/2}} \frac{\partial \Phi_\kappa}{\partial n(y)}(x,y) \varphi(y) ds(y) \ , \ x \in B_{3R/2},$$

and use the convergence of the double-layer potentials as in Theorem 3.1 to obtain $\|u_\kappa^i - u_0^i\|_{\infty, B_R} \to 0$, $\kappa \to 0$. This proves part (a).

For part (b) we choose R' with $3R/2 < R' < 2R$. Adding

$$u_0^i(x) := \int_{\partial B_{R'}} \Big\{ \frac{\partial u_0}{\partial n}(y) \Phi_0(x,y) - \frac{\partial \Phi_0}{\partial n(y)}(x,y) u_0(y) \Big\} ds(y) \ , \ x \in B_{R'},$$

and

$$u_0^s(x) := - \int_{\partial D} \frac{\partial u_0}{\partial n}(y) \Phi_0(x,y) ds(y) \ , \ x \in D,$$

and using $u_0|_{\partial D} = 0$ and Green's formula, we obtain $u_0^i(x) + u_0^s(x) = u_0(x)$, $x \in D \cap B_{R'}$. u_0^i is harmonic in $B_{R'}$. u_0^s is harmonic and bounded in D because of $\int_{\partial D} \frac{\partial u_0}{\partial n} ds = 0$. Hence, u_0^s is the solution of (P_0) for the incident field u_0^i. According to part (a) u_0^i can be approximated by u_κ^i in B_R. Because Theorem 3.1 implies that the solutions u_κ^s to (P_κ) with incident fields u_κ^i converge to u_0^s we have proved the theorem. □

4. A uniqueness theorem for the inverse scattering problem. Let B_R, D_0, D, and q satisfy the assumptions of the previous sections. Moreover, we assume that $\tilde{D}_0 \subset B_R$, $\tilde{D} := \mathbb{R}^2 \setminus \overline{\tilde{D}_0}$, and $\tilde{q} \in C^{0,\alpha}(\overline{\tilde{D}})$ satisfy the analogous assumptions. (P_κ) denotes the scattering problem with data D, q, and u_κ^i, u_κ^s is its solution. $(\tilde{P}_\kappa)$ denotes the scattering problem with data $\tilde{D}$, $\tilde{q}$, and u_κ^i, $\tilde{u}_\kappa^s$ its solution. Our final theorem states that, if for $(\tilde{P}_\kappa)$ and (P_κ) the Cauchy data of the scattered fields coincide on ∂B_R for all possible incident fields u_κ^i and for all $0 < \kappa < 1$, then $\partial D = \partial \tilde{D}$ and $q = \tilde{q}$; i.e., the Cauchy data for an interval of frequencies and all possible incident fields uniquely determine the obstacle D_0 and the inhomogeneity q.

Let us add two remarks before we prove the theorem.

The incident fields are very often chosen to be plane waves $u_\kappa^i(x) = \exp(i\kappa d \cdot x)$, $x \in \mathbb{R}^2$, propagating in the direction $d \in \mathbb{R}^2$, $|d| = 1$. In any closed subset of B_R any solution to $\Delta u + \kappa^2 u = 0$ in B_R can be uniformly approximated by elements from the span of the plane waves [6, Lemma 3.2]. Therefore we could also work with plane waves instead of the larger set of incident fields.

Furthermore, in scattering theory one usually works with the far field pattern of the scattered fields. But since the far field pattern uniquely determines the Cauchy data of the scattered field [3, Theorem 2.13] and vice versa [3, Theorem 2.5], coincidence of the far field patterns and coincidence of the Cauchy data are equivalent.

THEOREM 4.1. *If for all $0 < \kappa < 1$ and for all solutions $u_\kappa^i \in C^2(B_{3R/2})$ of $\Delta u_\kappa^i + \kappa^2 u_\kappa^i = 0$ in $B_{3R/2}$ the Cauchy data of the scattered fields for (P_κ) and $(\tilde{P}_\kappa)$ coincide, i.e., $u_\kappa^s|_{\partial B_R} = \tilde{u}_\kappa^s|_{\partial B_R}$ and $\frac{\partial u_\kappa^s}{\partial n}|_{\partial B_R} = \frac{\partial \tilde{u}_\kappa^s}{\partial n}|_{\partial B_R}$, then $\partial D = \partial \tilde{D}$ and $q = \tilde{q}$.*

Proof. We first prove $\partial D = \partial \tilde{D}$ by reducing everything to the case $\kappa = 0$ and using Schiffer's idea. We choose $u_0^i(x_1, x_2) := x_1 + ix_2$, $(x_1, x_2) \in \mathbb{R}^2$, and according to Theorem 3.2 (a) incident fields $u_\kappa^i \in C^2(B_{3R/2})$, $0 < \kappa < 1$, with $\|u_\kappa^i - u_0^i\|_{\infty, B_R} \to 0$, $\kappa \to 0$. From $u_\kappa^s|_{\partial B_R} = \tilde{u}_\kappa^s|_{\partial B_R}$, $0 < \kappa < 1$, and Theorem 3.1 we obtain $u_0^s|_{\partial B_R} = \tilde{u}_0^s|_{\partial B_R}$. Denoting by G the unbounded component of $D \cap \tilde{D}$ we have $u_0^s(x) = \tilde{u}_0^s(x)$ for all $x \in G$ because of the uniqueness for the Dirichlet problem in the exterior of B_R and because harmonic functions are analytic.

If we assume $D_0 \neq \tilde{D}_0$, without loss of generality, we can assume that $D_1 := (\mathbb{R}^2 \setminus \overline{G}) \setminus \overline{D_0}$ is nonempty. D_1 is an open subset of D and $\partial D_1 \subset \partial G \cup \partial D = (\partial G \setminus \partial D) \cup \partial D$. Since u_0^s coincides with $-u_0^i$ on ∂D and with $\tilde{u}_0^s = -u_0^i$ on $\partial G \setminus \partial D$ the maximum principle implies $u_0^s = -u_0^i$ in D_1 and hence in D by analyticity. This is a contradiction because u_0^i is unbounded in D whereas u_0^s is bounded. Then, we must have $D_0 = \tilde{D}_0$, i.e., $\partial D = \partial \tilde{D}$.

Now, we prove $q = \tilde{q}$. We first establish the relation (2.1) in Theorem 2.1 with $q - \tilde{q}$. To this end we choose harmonic functions u, $v \in C^2(D \cap B_{2R}) \cap C^1(\overline{D} \cap B_{2R})$ satisfying $u|_{\partial D} = v|_{\partial D} = 0$ and $\int_{\partial D} \frac{\partial u}{\partial n} ds = \int_{\partial D} \frac{\partial v}{\partial n} ds = 0$. According to Theorem 3.2 (b) there are incident fields $u_\kappa^i \in C^2(B_{3R/2})$, $0 < \kappa < 1$, such that the solutions u_κ^s to (P_κ) satisfy $\|u_\kappa^i + u_\kappa^s - u\|_{\infty, D \cap B_R} \to 0$, $\kappa \to 0$, and similarly for the solutions $\tilde{u}_\kappa^s$ to $(\tilde{P}_\kappa)$ with incident field u_κ^i. To shorten notation we define $u_\kappa := u_\kappa^i + u_\kappa^s$ and $\tilde{u}_\kappa := u_\kappa^i + \tilde{u}_\kappa^s$ in $\overline{D \cap B_R}$. From $u_\kappa^s|_{\partial B_R} = \tilde{u}_\kappa^s|_{\partial B_R}$ and $\frac{\partial u_\kappa^s}{\partial n}|_{\partial B_R} = \frac{\partial \tilde{u}_\kappa^s}{\partial n}|_{\partial B_R}$ we infer

the same equations for u_κ and $\tilde{u}_\kappa$ for $0 < \kappa < 1$. This yields, together with Green's theorem,

$$\begin{aligned}
0 &= \int\limits_{\partial(D\cap B_R)} \Big(v\frac{\partial}{\partial n}(u_\kappa - \tilde{u}_\kappa) - (u_\kappa - \tilde{u}_\kappa)\frac{\partial v}{\partial n}\Big) ds \\
&= \int\limits_{D\cap B_R} (\Delta u_\kappa - \Delta \tilde{u}_\kappa) v dx \\
&= \kappa^2 \int\limits_{D\cap B_R} (\tilde{q}\tilde{u}_\kappa - q u_\kappa) v dx \\
&= \kappa^2 \int\limits_{D\cap B_R} (\tilde{q} - q) u_\kappa v dx + \kappa^2 \int\limits_{D\cap B_R} \tilde{q}(\tilde{u}_\kappa - u_\kappa) v dx.
\end{aligned}$$

Using $q(x) - \tilde{q}(x) = 0$ for $|x| \geq R$, dividing by κ^2, and taking $\kappa \to 0$ we obtain

$$\int\limits_{D\cap B_{2R}} (\tilde{q} - q) u v dx = 0$$

because $\tilde{u}_\kappa$ and u_κ converge uniformly to u according to Theorem 3.2 (b). Then, from Theorem 2.1 we can conclude $q = \tilde{q}$ and we have proved the uniqueness result. ∎

REFERENCES

[1] L. V. Ahlfors, *Complex Analysis*, 2nd ed., McGraw-Hill, New York, NY, 1966.

[2] A. P. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasileira de Mathemática, Rio de Janeiro, 1980, pp. 65–73.

[3] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1992.

[4] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[5] V. Isakov and Z. Sun, *The inverse scattering at fixed frequencies in two dimensions*, Indiana Univ. Math. J., 44 (1995), pp. 883–896.

[6] A. Kirsch and R. Kress, *Uniqueness in inverse obstacle scattering*, Inverse Problems, 9 (1993), pp. 285–299.

[7] A. Kirsch and L. Päivärinta, *Some uniqueness theorems in inverse scattering theory*, Math. Methods Appl. Sci., 1998, to appear.

[8] R. Kress, *On the low wave number asymptotics for the two-dimensional exterior Dirichlet problem for the reduced wave equation*, Math. Methods Appl. Sci., 9 (1987), pp. 335–341.

[9] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.

[10] J. Sylvester and G. Uhlmann, *The Dirichlet to Neumann map and applications*, in Inverse Problems in Partial Differential Equations, D. Colton, R. Ewing, and W. Rundell, eds., SIAM, Philadelphia, PA, 1990, pp. 101–139.

[11] P. Werner, *Randwertprobleme der mathematischen Akustik*, Arch. Rational Mech. Anal., 10 (1962), pp. 29–66.

[12] P. Werner, *Low frequency asymptotics for the reduced wave equation in two-dimensional exterior spaces*, Math. Methods Appl. Sci., 8 (1986), pp. 134–156.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1129–1139, September 1998

005

CONVERGENCE AND DIVERGENCE OF DECREASING REARRANGED FOURIER SERIES*

ANTONIO CÓRDOBA† AND PABLO FERNÁNDEZ†

Abstract. In a number of useful applications, e.g., data compression, the appropriate partial sums of the Fourier series are formed by taking into consideration the size of the coefficients rather than the size of the frequencies involved. The purpose of this paper is to show the limitations of that method of summation. We use several results from the number theory to construct counterexamples to L^p-convergence for $p < 2$. We also show how to obtain positive results if we combine the two points of view, i.e., cutting on frequencies and the size of coefficients at the same time. This can be considered as a kind of uncertainty principle for Fourier sums.

Key words. partial Fourier sums, L_p-convergence, Gaussian sums

AMS subject classifications. 42A20, 11L03

PII. S0036141097320705

1. Introduction.

For any function, $f \in L^1(\mathbf{T})$, we can construct its Fourier series

$$f(x) \sim \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{2\pi i k x}.$$

The traditional way of reconstructing the function from its Fourier coefficients is to consider the partial sums

$$S_N f(x) = \sum_{|k| \le N} \hat{f}(k) e^{2\pi i k x}.$$

It is well known that there is convergence in norm. If $f \in L^p(\mathbf{T})$, for $1 < p < \infty$, then

$$S_N f \to f \text{ in } L^p(\mathbf{T}) \text{ as } N \to \infty;$$

and since Carleson [1], it is also known that we have almost everywhere convergence.

However, taking into account the interpretation of the Fourier coefficients as, for example, the x-ray diffraction pattern of a periodic electron density, it seems to be more natural to pay attention to the coefficients that give us more information, that is, those of bigger magnitude, and to reconstruct the function ordering the Fourier coefficients in decreasing order. The same comments also apply if we are interested in the application of the Fourier series to signal processing algorithms. The mathematical expression of this fact leads us to consider, for each $\lambda > 0$, partial sums

$$\tilde{S}_\lambda f(x) = \sum_{|\hat{f}(k)| > \lambda} \hat{f}(k) e^{2\pi i k x}$$

and their limit when $\lambda \to 0^+$.

*Received by the editors April 30, 1997; accepted for publication (in revised form) August 25, 1997; published electronically April 14, 1998.
http://www.siam.org/journals/sima/29-5/32070.html
†Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049-Cantoblanco, Madrid, Spain (antonio.cordoba@uam.es, pablo.fernandez@uam.es).

In a recent paper [2], Körner answered in the negative a question asked by Carleson and Coifman, proving the existence of a function $f \in L^2(\mathbf{T})$ such that

$$\limsup_{\lambda \to 0^+} \left| \sum_{|\hat{f}(k)| \geq \lambda} \hat{f}(k) e^{2\pi i k\, x} \right| = \infty \text{ for almost every } x \in \mathbf{T}.$$

Körner's proof is based on an ingenious modification of a construction due to Olevskii for the Haar system, and it also uses a probabilistic lemma of Salem and Zygmund.

Using a different method, we show in this paper that L^p-convergence, for $p < 2$, also fails for the partial sums $\tilde{S}_\lambda$. More concretely, we have the following theorem.

THEOREM 1.

a) *If we define a maximal operator*

$$\tilde{S}^* f(x) = \sup_{\lambda > 0} \left| \tilde{S}_\lambda f(x) \right|,$$

then, for all $1 \leq p < 2$, *there is a function* $f \in L^p(\mathbf{T})$ *(explicitly constructed) such that*

$$\|\tilde{S}^* f\|_p = \infty.$$

b) *For each* $p < 2$, *there exists a function* $f \in L^p$ *such that* $\limsup_{\lambda \to 0^+} \|\tilde{S}_\lambda f\|_p = \infty$.

Our arguments are of a number theory nature, and we use the Farey dissection of the interval $[0, 1)$ and the prime number theorem in the proof.

This divergence phenomenon suggests that $\tilde{S}_\lambda f$ is not the proper sum to be taken. One may argue that one reason for the failure of $\tilde{S}_\lambda f$ to converge is because we have not taken into account the uncertainty principle in the following way: it does not make sense to impose restrictions upon the size of $|\hat{f}(k)|$ and not upon $|k|$ itself.

Let us consider the modified partial sums

$$S_N^\delta f(x) = \sum_{\substack{|k| \leq N \\ |\hat{f}(k)| \geq N^{-\delta}}} \hat{f}(k) e^{2\pi i k x}, \delta > 0.$$

Then we have the following theorem.

THEOREM 2.

a) *If* $\delta < \frac{1}{2}$, *then for every* $p < 2$ *there is a function* $f \in L^p(\mathbf{T})$ *such that*

$$\| \limsup_{N \to \infty} S_N^\delta f \|_p = \infty.$$

b) *If* $\delta \geq \frac{1}{2}$, *then*

$$\sup_N \|S_N^\delta f\|_p < \infty \text{ for every } f \in L^p(\mathbf{T}), 1 < p \leq 2.$$

c) *If* $\delta > \frac{1}{2}$, *then we have*

$$\lim_{N \to \infty} \|S_N^\delta f - f\|_p = 0 \quad \text{for every } f \in L^p(\mathbf{T}), 1 < p \leq 2.$$

d) *If* $\delta = \frac{1}{2}$, *then for every* $p < 2$ *there exists* $f \in L^p(\mathbf{T})$ *such that*

$$\limsup_{N \to \infty} \|S_N^\delta f - f\|_p > 0.$$

More generally, given a decreasing function ϕ, one can consider partial sums

$$S_N^{\phi} f(x) = \sum_{\substack{|k| \le N \\ |\hat{f}(k)| \ge \phi(N)}} \hat{f}(k) e^{2\pi i k x}.$$

Our construction shows that the behavior of S_N^{ϕ} depends upon the condition

$$N \cdot \phi^2(N) = o(1).$$

Finally, we would like to say that we believe more important than the actual results presented in this paper, are the methods of construction of the examples. They illustrate, yet again, the connection between Fourier series and number theory.

2. A trigonometrical sum estimate. Take N large enough and consider

$$P_N^*(x) = \max_{1 \le j \le N} \left| P_N^j(x) \right| = \max_{1 \le j \le N} \left| \sum_{\substack{p \text{ prime} \\ N < p \le N + j}} e^{2\pi i p x} \right|.$$

LEMMA 1.

$$\|P_N^*\|_r \ge C\, N^{3/4 - 1/2r} \log^{-1-1/r}(N), \text{ for any } 1 < r < 2.$$

Proof. First, we take the primes q, $\sqrt{N} \le q < \sqrt{2\,N}$. For each a, $(a, q) = 1$, we have the Farey intervals

$$I_{a/q} = \left(\frac{a}{q} - \frac{1}{8\,q^2}\,,\, \frac{a}{q} + \frac{1}{8\,q^2} \right).$$

It is easy to see that these intervals are disjoint. Let us consider the set

$$E_N = \bigcup_{\substack{q \text{ prime} \\ \sqrt{N} \le q < \sqrt{2N}}} \bigcup_{a=1}^{q-1} I_{a/q}.$$

Then,

$$\begin{aligned} \|P_N^*\|_r^r &= \int_0^1 \left(P_N^*(x)\right)^r \, dx \ge \int_{E_N} \left(P_N^*(x)\right)^r \, dx \\ &= \sum_{\substack{q \text{ prime} \\ \sqrt{N} \le q < \sqrt{2N}}} \sum_{a=1}^{q-1} \int_{I_{a/q}} \left(P_N^*(x)\right)^r \, dx. \end{aligned}$$

We have the following fact.

FACT 1. *If* $|x - y| \le \frac{1}{8q^2}$, *with* $q \ge \sqrt{N}$, *then* $P_N^*(x) \ge C P_N^*(y)$. (The proof follows by summation in parts.)

Using this fact, we obtain

$$\begin{aligned}
\|P_N^*\|_r^r &\geq C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \sum_{a=1}^{q-1} \int_{I_{a/q}} dx \, (P_N^*(a/q))^r \\
&\geq C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \frac{1}{q^2} \sum_{a=1}^{q-1} \left| \sum_{\substack{p \text{ prime} \\ N < p < 2N}} e^{2\pi i p\, a/q} \right|^r \\
&= C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \frac{1}{q^2} \sum_{a=1}^{q-1} \left| \sum_{r=1}^{q-1} \sum_{\substack{p \text{ prime} \\ N < p < 2N \\ p \equiv r(q)}} e^{2\pi i p\, a/q} \right|^r \\
&= C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \frac{1}{q^2} \sum_{a=1}^{q-1} \left| \sum_{r=1}^{q-1} e^{2\pi i r\, a/q} \left[\pi(2N, q, r) - \pi(N, q, r) \right] \right|^r,
\end{aligned}$$

where $\pi(x, a, b)$, with $(a, b) = 1$, counts the number of primes less than or equal to x in the arithmetic progression $b, b + a, b + 2a, b + 3a, \ldots$. Let us rename the difference of π's in our expression as a coefficient $b_{N,q,r}$, such that $b_{N,q,r} \geq 0$. Then, using the inequality

$$\left(\sum |a_j|^r \right)^{1/r} \geq \left(\sum |a_j|^2 \right)^{1/2}, \text{if } r \leq 2,$$

we can write

$$\begin{aligned}
\|P_N^*\|_r^r \geq\ & C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \frac{1}{q^2} \left\{ \sum_{a=1}^{q-1} \left| \sum_{r=1}^{q-1} e^{2\pi i r\, a/q} b_{N,r,q} \right|^2 \right\}^{r/2} \\
&= C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \frac{1}{q^2} \left\{ (q-1) \sum_{r=1}^{q-1} b_{N,r,q}^2 + \sum_{r \neq s} b_{N,q,r}\, b_{N,q,s} \sum_{a=1}^{q-1} e^{2\pi i\, a\, (r-s)/q} \right\}^{r/2} \\
&= C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \frac{1}{q^2} \left\{ (q-1) \sum_{r=1}^{q-1} b_{N,q,r}^2 - \sum_{r \neq s} b_{N,q,r}\, b_{N,q,s} \right\}^{r/2}.
\end{aligned}$$

Since the inequality

$$N \sum_{j=1}^{N} a_j^2 - \sum_{j \neq k} a_j\, a_k \geq \sum_{j=1}^{N} a_j^2$$

holds for $a_k \geq 0$, we have

$$\|P_N^*\|_r^r \geq C \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \frac{1}{q^2} \left\{ \sum_{r=1}^{q-1} b_{N,q,r}^2 \right\}^{r/2} \geq \frac{C}{N} \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \left\{ \sum_{r=1}^{q-1} b_{N,q,r}^2 \right\}^{r/2}.$$

Now we want to estimate the size of the interior sum, $\sum_{r=1}^{q-1} b_{N,q,r}^2$. In general, it is quite difficult to obtain lower bounds for the size of the $b_{N,q,r}$'s, even assuming the generalized Riemann hypothesis, especially if we want to make these estimates uniform in q and r (in the considered range). Fortunately, we are dealing with the sum of the squares and this makes things easier. By Cauchy's inequality and Chebyshev's theorem, we have

$$(q-1)\sum_{r=1}^{q-1} b_{N,q,r}^2 \geq \left(\sum_{r=1}^{q-1} b_{N,q,r}\right)^2 > C\frac{N^2}{\log^2(N)}.$$

Therefore,

$$\begin{aligned}\|P_N^*\|_r^r &\geq \frac{C}{N} \sum_{\substack{q \text{ prime} \\ \sqrt{N} \leq q < \sqrt{2N}}} \left(\frac{N^2}{(q-1)\log^2(N)}\right)^{r/2} \\ &\geq \frac{C}{N}\,\frac{N^r}{N^{r/4}\log^r(N)}\,\#\{\text{primes } q\ /\sqrt{N} \leq q < \sqrt{2N}\} \\ &\geq C\,N^{3r/4-1/2}\log^{-r-1}(N) \qquad \square\end{aligned}$$

3. Basic construction. Take $\alpha > 0$ (to be determined) and consider the functions

$$f_1(x) = 1 + 2\sum_{k=0}^{\infty} \underbrace{\frac{1}{2^{k\alpha}} \sum_{n=2^k}^{2^{k+1}-1} \cos(2\pi\, n\, x)}_{f_{2^k}^1(x)},$$

$$f(x) = 1 + 2\sum_{k=0}^{\infty} \underbrace{\frac{1}{2^{k\alpha}} \sum_{n=2^k}^{2^{k+1}-1} a_n\, \cos(2\pi\, n\, x)}_{f_{2^k}(x)},$$

$$\text{where } a_n = \begin{cases} 1 + \frac{1}{n} & \text{if } n \text{ prime,} \\ 1 - \frac{1}{2^k} & \text{if not.} \end{cases}$$

We can evaluate the L^p norm, for $p > 1$, of the functions f_{2^k} using the well-known estimates for the L^p-norm of the *Dirichlet kernel*,

$$\|f_{2^k}\|_p \leq \left\|f_{2^k} - f_{2^k}^1\right\|_p + \left\|f_{2^k}^1\right\|_p \leq C\,2^{k\,(1-1/p-\alpha)}.$$

Therefore,

$$\|f\|_p \leq 1 + 2\sum_{k=0}^{\infty} \|f_{2^k}\|_p \leq 1 + C\sum_{k=0}^{\infty} 2^{k(1-1/p-\alpha)}.$$

As a result, $f \in L^p(\mathbf{T})$ whenever $\alpha > 1 - \frac{1}{p}$.

For each $k = 0, 1, 2 \dots$ and for some j, $2^k < j < 2^{k+1}$ (we will see later how to choose j), we construct the sequence

$$a_{k,j} = \frac{1}{2^{k\alpha}} \left(1 + \frac{1}{j}\right) \searrow 0 \ \text{ as } k \to \infty.$$

The operator $\tilde{S}_{a_{k,j}}$ keeps all the frequencies up to 2^k; and from the next dyadic block, it keeps only some prime frequencies (those less than j):

$$|\tilde{S}_{a_{k,j}} f(x)| \geq \frac{1}{2^{k\alpha}} \left| \sum_{\substack{\nu \text{ prime} \\ 2^k \leq \nu \leq j}} e^{2\pi i \nu x} \right| - \left| \underbrace{\sum_{|\nu| \leq 2^k} \hat{f}(\nu) e^{2\pi i \nu x}}_{I} + \underbrace{\sum_{\substack{\nu \text{ prime} \\ 2^k \leq \nu \leq j}} \frac{e^{2\pi i \nu x}}{\nu}}_{II} \right|.$$

Now, for each $x \in \mathbf{T}$, we can choose $j = j(x)$ in such a way that the first term equals $2^{-k\alpha} P^*_{2^k}(x)$ (see the previous section for the definition of P^*_N). On the other hand, both terms I and II are $O(1)$ as $k \to \infty$, so for every $x \in \mathbf{T}$, we have

$$\sup_{\lambda > 0} \left| \tilde{S}_\lambda f(x) \right| \geq \frac{1}{2^{k\alpha}} P^*_{2^k}(x) - O(1).$$

It follows that

$$\|\tilde{S}^* f\|_p \geq \frac{1}{2^{k\alpha}} \|P^*_{2^k}\|_p - O(1).$$

Recalling our basic lemma, with $N = 2^k$, we obtain

$$\|\tilde{S}^* f\|_p \geq C_p \ 2^{k(3/4 - 1/2p - \alpha)} k^{-1-1/p} - O(1).$$

So the L^p norm diverges when $\alpha < \frac{3}{4} - \frac{1}{2p}$. Therefore, for each $1 \leq p < 2$, we can find an α with

$$1 - \frac{1}{p} < \alpha < \frac{3}{4} - \frac{1}{2p}$$

such that the function f constructed above satisfies

$$\|f\|_p < \infty \text{ and } \left\| \tilde{S}^* f \right\|_p = \infty.$$

This completes the proof of Theorem 1a).

In order to prove part b), we need an extra argument. Let us begin with the well-known estimate (see [3], Khintchin inequality)

$$\left[\int_0^1 \int_0^1 \left| \sum_{k=2^n}^{2^{n+1}-1} r_k(t) e^{2\pi i k x} \right|^p dx\, dt \right]^{1/p} \sim 2^{n/2},$$

where $\{r_k(t)\}$ denotes the Rademacher system of orthonormal functions. We use it to "construct," for each $p < 2$, a polynomial

$$P_n(x) = \sum_{k=2^n}^{2^{n+1}-1} a_k \, e^{2\pi i k x}, \text{ where } a_k \text{ is either 0 or 1,}$$

with $\|P_n\|_p \geq C_p\, 2^{n/2}$, for some $C_p > 0$. Next, we consider, for some $\alpha > 0$ to be determined, the function

$$Q_n(x) = \sum_{k=2^n}^{2^{n+1}-1} b_k\, e^{2\pi i k x}, \text{ where } b_k = \begin{cases} 2^{-n\alpha} & \text{if } a_k = 1, \\ 2^{-n\alpha} - 2^{-n} & \text{if } a_k = 0. \end{cases}$$

Then,

$$\|Q_n\|_p \leq C_p\ 2^{n(1-1/p-\alpha)} +\ C'_p\ 2^{-n/p} +\ C''_p\ 2^{-n/2},$$

thus the function

$$f(x) = \sum_{n=1}^{\infty} Q_n(x)$$

satisfies

$$\|f\|_p < \infty \text{ if } \alpha > 1 - \frac{1}{p}.$$

On the other hand,

$$\sum_{|\hat{f}(k)| \geq 2^{-n\alpha}} \hat{f}(k) e^{2\pi i\, k\, x} = \sum_{|k| \leq 2^n} \hat{f}(k) e^{2\pi i\, k\, x} + 2^{-n\alpha} P_n(x),$$

and so

$$\begin{aligned} \left\| \sum_{|\hat{f}(k)| \geq 2^{-n\alpha}} \hat{f}(k) e^{2\pi i\, k\, x} \right\|_p &\geq 2^{-n\alpha}\, \|P_n\|_p - O(1) \\ &\geq C_p 2^{n(1/2-\alpha)} - O(1). \end{aligned}$$

Consequently, for each $p < 2$, we can find an α, $1 - 1/p < \alpha < 1/2$, such that

$$\|f\|_p < \infty \text{ and } \limsup_{\lambda \to 0^+} \|\tilde{S}_\lambda f\|_p = \infty.$$

This proves part b) of Theorem 1.

4. The modified partial sums. Now we consider, for each $\delta > 0$, the modified partial sums

$$S_N^\delta f(x) = \sum_{\substack{|k|\ \leq\ N \\ |\hat{f}(k)|\ \geq\ N^{-\delta}}} \hat{f}(k) e^{2\pi i k x}.$$

Case $\delta \geq \frac{1}{2}$. For $r \leq 2$, let us compare these operators with the following partial sums:

$$\begin{aligned} \|S_N^\delta f - S_N f\|_r &= \left(\int_0^1 |S_N^\delta f(x) - S_N f(x)|^r dx \right)^{1/r} \\ &= \left(\int_0^1 \left| \sum_{\substack{|k|\ \leq\ N \\ |\hat{f}(k)|\ <\ N^{-\delta}}} \hat{f}(k) e^{2\pi i\, k\, x} \right|^r dx \right)^{1/r} \end{aligned}$$

$$\leq \left(\int_0^1 \left| \sum_{\substack{|k| \leq N \\ |\hat{f}(k)| < N^{-\delta}}} \hat{f}(k) e^{2\pi i\, k\, x} \right|^2 dx \right)^{1/2}$$

$$= \left(\sum_{\substack{|k| \leq N \\ |\hat{f}(k)| < N^{-\delta}}} |\hat{f}(k)|^2 \right)^{1/2} \leq C \left(N\, N^{-2\delta} \right)^{1/2} \leq C\, N^{1/2-\delta}.$$

Therefore,

$$\|S_N^\delta f - S_N f\|_r = O(1) \text{ as } N \to \infty.$$

If $f \in L^r$, the partial sums $S_N f$ tend to f in the L^r-norm. Therefore, applying the triangular inequality, we obtain

$$\left\| S_N^\delta f - f \right\|_r = O(1) \qquad \text{as } N \to \infty.$$

This proves part b) of Theorem 2. Part c) is quite easy now, because if $\delta > \frac{1}{2}$, we get an $o(1)$ as $N \to \infty$, that is, the function is recovered, in norm, when we sum its Fourier series in this way.

Case $\delta < \frac{1}{2}$. We have to make a slight modification of the function f defined in section 3. Let us begin with

$$f_k(x) = \frac{1}{2^{k/2}} \sum_{n=2^k}^{2^{k+1}-1} a_n e^{2\pi i n x},$$

$$\text{with } a_n = \begin{cases} 1 - \frac{1}{\lfloor 2^{k/2\delta} \rfloor} + \frac{1}{n^{1/2\delta}} & \text{if } n \text{ is prime,} \\ 1 - \frac{1}{\lfloor 2^{k/2\delta} \rfloor} & \text{if not.} \end{cases}$$

Thus,

$$f(x) = \sum_k f_k(x) \quad \text{is in } L^p \text{ for any } p < 2.$$

Next, we translate these frequencies to the right:

$$g_k(x) = e^{2\pi i x \left\{ \lfloor 2^{k/2\delta} \rfloor - 2^k \right\}} f_k(x)$$

$$= \frac{1}{2^{k/2}} \sum_{l=\lfloor 2^{k/2\delta} \rfloor}^{\lfloor 2^{k/2\delta} \rfloor + 2^k} a_{l+2^k-\lfloor 2^{k/2\delta} \rfloor} e^{2\pi i l x}.$$

The function $g(x) = \sum_k g_k(x)$ is, of course, in all L^p, $p < 2$. Now, take the sequence

$$a_{k,j} = \frac{1}{\lfloor 2^{k/2\delta} \rfloor^\delta} \left(1 - \frac{1}{\lfloor 2^{k/2\delta} \rfloor} + \frac{1}{\left(j + 2^k - \lfloor 2^{k/2\delta} \rfloor \right)^{1/2\delta}} \right).$$

If we estimate the sums

$$\sum_{\substack{|\nu| \leq a_{k,j}^{-\delta} \\ |\hat{g}(\nu)| \geq a_{k,j}}} \hat{g}(\nu) e^{2\pi i \nu x},$$

it is easy to see that we are just summing

$$\sum_{|\hat{g}(\nu)|\geq a_{k,j}} \hat{g}(\nu)e^{2\pi i\nu x},$$

and we have seen that these sums diverge in norm as $k\to\infty$ with the adequate choice of j.

Case $\delta=\frac{1}{2}$. We can use the same arguments with the Rademacher functions used in the proof of part b) of Theorem 1 to construct, for each $p<2$, a polynomial P_n, with coefficients 0 or 1, such that $\|P_n\|_p \geq C_p\, 2^{\frac{n}{2}}$ for some $C_p>0$; and a polynomial Q_n (putting $\alpha=1/2$ in the definition of its coefficients) in such a way that

$$\|Q_n\|_p \leq C'_p\ 2^{-n/2}\cdot 2^{n(1-1/p)}.$$

Therefore,

$$f(x)=\sum_{n=1}^{\infty} Q_n(x) \text{ satisfies } \|f\|_p \leq C'_p \sum_{n=1}^{\infty} 2^{-n(1/p-1/2)} <\infty.$$

On the other hand,

$$\sum_{\substack{|k|\ \leq\ 2^n\\ |\hat{f}(k)|\ \geq\ 2^{-n/2}}} \hat{f}(k)e^{2\pi i\,k\,x} = \sum_{|k|\leq 2^{n-1}} \hat{f}(k)e^{2\pi i\,k\,x} + 2^{-n/2}P_n(x),$$

and so

$$\begin{aligned}\left\| \sum_{\substack{|k|\ \leq\ 2^n\\ |\hat{f}(k)|\ \geq\ 2^{-n/2}}} \hat{f}(k)e^{2\pi i\,k\,x} - f\right\|_p &\geq 2^{-n/2}\,\|P_n\|_p - \left\|\sum_{k=n}^{\infty} Q_k\right\|_p\\ &\geq \frac{C_p}{\sqrt{2}} - o(1) \qquad \text{as } n\to\infty.\end{aligned}$$

5. Final remarks. With the use of Gaussian sums, one may produce explicit examples of functions $f\in L^p(\mathbf{T})$, $p<\frac{4}{3}$, such that

$$\limsup_{\lambda\to 0^+}\left|\sum_{|\hat{f}(k)|\geq\lambda} \hat{f}(k)e^{2\pi ik\,x}\right| =\infty \quad \text{ for almost every } x\in\mathbf{T}.$$

The construction is as follows. Take

$$f_k(x)=\sum_{2^{2k}}^{2^{2k+2}-1} a_n\,\cos(2\pi nx)$$

$$\text{with } a_n=\begin{cases} 2^{-k/2+k\epsilon}-2^{-3k}, & \text{if} \quad n\neq s^2,\\ 2^{-k/2+k\epsilon}-2^{-3k}(2^{2k+2}-|n|)^{-1/2}, & \text{if} \quad n=s^2.\end{cases}$$

Clearly,

$$\|f_k\|_p \sim 2^{k(\epsilon+3/2-2/p)}.$$

For each $p < 4/3$, taking ϵ small enough, we have that

$$f(x) = \sum_{k=1}^{\infty} f_k(x) \quad \in L^p.$$

Next, we will use the following facts.

FACT 2. *If $(P,Q) = 1$, with $\frac{M}{2} \le Q \le M$ and $Q \not\equiv 2 \bmod 4$, then*

$$\sup_{1\le k\le M} \left| \sum_{j=M}^{M+k} e^{2\pi i j^2 P/Q} \right| \ge C\sqrt{Q}.$$

FACT 3. *If $|x-y| \le \frac{1}{M^2}$, then*

$$\sup_{1\le k\le M} \left| \sum_{j=M}^{M+k} e^{2\pi i j^2 x} \right| \ge C \sup_{1\le k\le M} \left| \sum_{j=M}^{M+k} e^{2\pi i j^2 y} \right|.$$

FACT 4. *For each $x \in [0,1)$, let us consider the sequence of convergents $\{P_n/Q_n\}$ of its continuous fraction expansion. Then, for all irrational x, there are infinite Q_n, which are not congruent with 2 mod 4.* Consider the function f defined by

$$f(x) = \sum_k f_k(x) = \sum_\nu \hat{f}(\nu)\, e^{2\pi i \nu x}.$$

Take an irrational x and its sequence $\{Q_n\}$ of denominators not congruent with 2 mod 4. It determines dyadic blocks,

$$2^{n-1} \le Q_n \le 2^n.$$

We introduce the coefficients

$$a_{n,j} = \frac{1}{2^{n/2-n\epsilon}} - \frac{1}{2^{3n}\sqrt{2^{2n+2}-|j|}},$$

where n identifies the dyadic block previously selected and j, $2^{2n} < j < 2^{2n+2}$, is chosen so that

$$\left| \sum_{\nu=2^n}^{k(j)} e^{2\pi i \nu^2 P_n/Q_n} \right| \ge C\sqrt{Q_n},$$

where $k(j)$ is the first integer less or equal $\sqrt{j}$. Then,

$$\begin{aligned} \left| \sum_{|\hat{f}(\nu)|\ge a_{n,j}} \hat{f}(\nu) e^{2\pi i \nu x} \right| &\ge \frac{1}{2^{n/2-n\epsilon}} \left| \sum_{s=2^n}^{k(j)} e^{2\pi i s^2 x} \right| + O(1) \\ &\ge \frac{C}{2^{n/2-n\epsilon}} \sqrt{Q_n} + O(1) \ge C\, 2^{n\epsilon} + O(1). \end{aligned}$$

Hence,

$$\sup_{\lambda\to 0^+} \left| \sum_{|\hat{f}(\nu)|\ge\lambda} \hat{f}(\nu) e^{2\pi i \nu x} \right| = \infty \quad \text{for almost all } x \in [0,1).$$

As we stated in the introduction, Körner [2] obtained almost everywhere divergence for functions f in the class $L^2(\mathbf{T})$.

REFERENCES

[1] L. Carleson, *On the convergence and growth of partial sums of Fourier series*, Acta Math., 116 (1966), pp. 135–157.
[2] T.W. Körner, *Divergence of decreasing rearranged Fourier series,* Ann. Math., 144 (1996), pp. 167–180.
[3] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1988.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1140–1156, September 1998

006

STABILITY AND LINEAR INDEPENDENCE ASSOCIATED WITH SCALING VECTORS*

JIANZHONG WANG†

Abstract. In this paper, we discuss stability and linear independence of the integer translates of a scaling vector $\Phi = (\phi_1, \ldots, \phi_r)^T$, which satisfies a matrix refinement equation

$$\Phi(x) = \sum_{k=0}^{n} P_k \Phi(2x - k),$$

where (P_k) is a finite matrix sequence. We call $P(z) = \frac{1}{2}\sum P_k z^k$ the symbol of Φ. Stable scaling vectors often serve as generators of multiresolution analyses (MRAs) and therefore play an important role in the study of multiwavelets. Most useful MRA generators are also linearly independent.

The purpose of this paper is to characterize stability and linear independence of the integer translates of a scaling vector via its symbol. A polynomial matrix $P(z)$ is said to be two-scale similar to a polynomial matrix $Q(z)$ if there is a polynomial matrix $T(z)$ such that $P(z) = T(z^2)Q(z)T^{-1}(z)$. This kind of factorization of $P(z)$ is called two-scale factorization. We give a necessary and sufficient condition, in terms of two-scale factorization of the symbol, for stability and linear independence of the integer translates of a scaling vector.

Key words. stability, linear independence, scaling vectors, multiwavelets, multiresolution analysis, two-scale similarity

AMS subject classifications. Primary, 41A63, 42C05; Secondary, 42A05, 42A38

PII. S003614109630330X

1. Introduction.

In this paper, we discuss stability and linear independence of the integer translates of a scaling vector. A distribution vector

$$\Phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_r(x))^T, \quad x \in \mathbb{R},$$

is said to be a scaling vector if it is compactly supported and satisfies a matrix refinement equation

$$\Phi(x) = \sum_{k\in\mathbb{Z}} P_k \Phi(2x - k), \tag{1.1}$$

where the matrix sequence $(P_k)_{k\in\mathbb{Z}}$ is called a *mask* of Φ. Taking the Fourier transform of both sides of (1.1), we obtain

$$\hat{\Phi}(\omega) = P(z)\hat{\Phi}(\omega/2), \qquad z = e^{-i\omega/2}, \tag{1.2}$$

where $P(z) := \frac{1}{2}\sum P_k z^k$ is a *symbol* of Φ. We call $\hat{\Phi}(0)$ the *moment* (of order 0) of Φ since $\hat{\Phi}(0) = \int_{\mathbb{R}} \Phi(x)dx$. When $\hat{\Phi}(0) = 0$, we call Φ a zero-moment scaling vector. We will characterize stability and linear independence of the integer translates of a scaling vector via a special factorization of its symbol. Our study of scaling vectors is based on shift-invariant spaces. Hence, we first introduce some notions and results in the theory of shift-invariant spaces.

*Received by the editors May 2, 1996; accepted for publication (in revised form) October 13, 1997; published electronically April 15, 1998. This research was supported by NSF grant DMS-9503282.
http://www.siam.org/journals/sima/29-5/30330.html
†Department of Mathematics and Information Sciences, Sam Houston State University, Huntsville, TX 77341 (jwang@galois.shsu.edu).

Let S be a linear space of distributions on $\mathbb{R}$. We say that S is *shift-invariant* if

$$f \in S \quad \Longrightarrow \quad f(\cdot - j) \in S \quad \forall j \in \mathbb{Z}.$$

We are interested in shift-invariant spaces generated by entries of a scaling vector [4]. In this paper, since only scaling vectors are considered, any distribution vector is assumed to be compactly supported. Let l denote the space containing all sequences of complex numbers and l_0 denote the space of all compactly supported sequences in l. For a distribution vector Φ, the semiconvolution of Φ with a vector sequence $\mathbf{a} := (\mathbf{a}_k) \in (l)^r$, denoted by $\Phi * \mathbf{a}$, is defined by

$$\Phi * \mathbf{a} = \sum_{j=1}^{r} \sum_{k \in \mathbb{Z}} a_{j,k} \phi_j(x - k).$$

We define

$$S_0(\Phi) := \{\Phi * \mathbf{a}; \quad \mathbf{a} \in (l_0)^r\},$$

$$S(\Phi) := \{\Phi * \mathbf{a}; \quad \mathbf{a} \in (l)^r\},$$

and, if $\Phi \in (L^p)^r$, $1 \le p \le \infty$,

$$S_p(\Phi) := \mathrm{clos}_{L^p}(S_0(\Phi)).$$

It is obvious that these sets are shift-invariant spaces of functions and distributions. The vector Φ is called the generator of $S(\Phi)$ ($S_0(\Phi)$, $S_p(\Phi)$).

For $\Phi \in (L^p)^r, 1 \le p \le \infty$, its L^p norm is defined by $\|\Phi\|_p = (\sum_{i=1}^r \|\phi_i\|_p^p)^{1/p}$. Similarly, for a vector sequence $\mathbf{a} \in (l^p)^r$, its l^p norm is $\|\mathbf{a}\|_p = (\sum_{j=1}^r \sum_{k \in \mathbb{Z}} |a_{j,k}|^p)^{1/p}$. The integer translates of a distribution vector $\Phi \in (L^p)^r$ are said to be *l^p-stable* ($1 \le p \le \infty$) if there exist two positive constants C_1 and C_2 such that, for any $\mathbf{a} \in (l^p)^r$,

$$C_1 \|\mathbf{a}\|_p \le \|\Phi * \mathbf{a}\|_p \le C_2 \|\mathbf{a}\|_p. \tag{1.3}$$

It is known that (1.3) holds if and only if the r sequences

$$(\hat{\phi}_l(\omega + 2k\pi))_{k \in \mathbb{Z}}, \quad l = 1, 2, \ldots, r, \tag{1.4}$$

are linearly independent for all $\omega \in \mathbb{R}$ (see [12]), where $\hat{f}$ denotes the Fourier transform of function f:

$$\hat{f}(\omega) := \int_{x \in \mathbb{R}} f(x) \exp(-ix\omega)\, dx.$$

Note that the linear independence of sequences (1.4) does not depend on p and that it does make sense even for the distribution vectors not in $(L^p)^r$. (Recall that distribution vectors in this paper are assumed to be compactly supported. Hence, their Fourier transforms are entire function vectors.) Therefore, we give the following definition.

A distribution vector Φ is said to have *stable* integer translates if the sequences in (1.4) are linearly independent for all $\omega \in \mathbb{R}$ [12].

Another important notion for distribution vectors is linear independence.

A distribution vector Φ is said to have *linearly independent* integer translates if for any $\mathbf{a} \in (l)^r$, $\Phi * \mathbf{a} = \mathbf{0} \Longrightarrow \mathbf{a} = \mathbf{0}$.

Jia and Micchelli in [12] proved the following result. The integer translates of Φ are linearly independent if and only if the sequences in (1.4) are linearly independent for all $\omega \in \mathbb{C}$. Hence, the linear independence of the integer translates of Φ implies the stability of the integer translates of Φ.

For a distribution vector, we also introduce the notion of finitely linear independence.

A distribution vector Φ is said to have *finitely linearly independent* integer translates if for any $\mathbf{a} \in (l_0)^r$, $\Phi * \mathbf{a} = \mathbf{0} \Longrightarrow \mathbf{a} = \mathbf{0}$.

It is clear that if the integer translates of Φ are stable (or linearly independent), then they also are finitely linearly independent. But the converse is not true.

For convenience, in the rest of the paper, we will simply say that Φ is *stable* (*linearly independent, finitely linearly independent*) instead of that Φ has *stable* (*linearly independent, finitely linearly independent*) integer translates.

Stability and linear independence are important properties of distribution vectors. Bases of shift-invariant spaces are often required to be stable so that the duality principle can be applied efficiently (see [1], [2]). In wavelet theory, generators of MRAs are stable. Moreover, most useful generators of MRAs also are linearly independent. Hence, it is desirable to give a criterion for stability and linear independence of scaling vectors in terms of their symbols.

Scaling vectors are discussed in several papers for different purposes. Goodman and Lee [6] and Goodman, Lee, and Tang [7] discussed scaling vectors in generality and constructed multiwavelets using MRAs generated by scaling vectors. In [8], Heil and Colella discussed the existence and uniqueness of the solution of a matrix refinement equation, and the regularity of the solution as well. Discussions of the supports of scaling vectors can be found in Massopust, Ruch, and Van Fleet [15]; Ruch, So, and Wang [19]; and So and Wang [20]. Approximation order provided by scaling vectors was studied by Heil, Strang, and Strela [9]; Jia, Riemenschneider, and Zhou [13]; and Plonka [16]. In [16], Plonka introduced two-scale similarity for the symbol of a scaling vector to characterize its approximation order. This method also was successfully applied in the construction of the scaling vectors with required regularity (see [3]) and with symmetry (see [17], [21]).

In this paper, we use two-scale similarity to characterize stability, linear independence, and finitely linear independence of scaling vectors. In section 2, we discuss stability and linear independence of a distribution vector. A result of [11] shows that any generator of a shift-invariant space can be obtained as a finitely transformed linearly independent generator in the same space. Using this result, we obtain necessary and sufficient conditions for stability (linear independence) of a scaling vector in terms of properties of the corresponding transform. In section 3, we discuss two-scale similarity of the symbol of a scaling vector. Based on the results in section 2, we derive necessary and sufficient conditions for stability (linear independence) of a scaling vector in terms of two-scale similarity. Some examples are given in section 4.

2. Transform of generators of a shift-invariant space. In this section, we discuss stability and linear independence of a distribution vector. We denote by $\dim \Phi$ the dimension (i.e., the number of the components) of a distribution vector Φ. We start with the following result in [11].

THEOREM 2.1 (Jia). *Let Φ be a distribution vector. Then there exists a linearly independent distribution vector $\Psi \in S(\Phi)$ such that $\dim \Psi \leq \dim \Phi$, $\Phi \subset S_0(\Psi)$, and*

$S(\Psi) = S(\Phi)$.

We will call $\dim \Psi$ the *index* of Φ and denote it by $\Phi^\#$. It is obvious that $\Phi^\# \le \dim \Phi$, and $\Phi^\# < \dim \Phi$ if and only if Φ is finitely linearly dependent.

Let Ψ and Φ be two generators of a shift-invariant space S. Then there is a rational function matrix $T(z)$ such that $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Psi}(\omega)$, where $T(z)$ is called the *transform* (matrix) from Ψ to Φ. By Theorem 2.1, if $\Psi \in S(\Phi)$ is linearly independent, then the transform $T(z)$ is a Laurent polynomial matrix. We now introduce some notations for transform matrices. In the rest of this paper, a polynomial always means a Laurent polynomial. We denote the polynomial ring over $\mathbb{C}$ by $\mathcal{P}$ and denote the field of rational functions over $\mathbb{C}$ by $\mathcal{R}$. We shall simply call a polynomial matrix a *P-matrix* and call a matrix with rational function entries an *R-matrix*. The set of all $r \times s$ P-matrices is denoted by $\mathcal{P}^{r\times s}$, and the set of all $r \times s$ R-matrices is denoted by $\mathcal{R}^{r\times s}$. When $r = s$, we use $\mathcal{P}^r$ ($\mathcal{R}^r$) instead of $\mathcal{P}^{r\times r}$ ($\mathcal{R}^{r\times r}$). A matrix $P(z) \in \mathcal{P}^r$ is said to be invertible if there is an R-matrix $R(z)(:= P^{-1}(z)) \in \mathcal{R}^r$ such that $R(z)P(z) = I$, where I is the identity matrix, and $P(z)$ is said to be *invertible everywhere* if $P(z)$ is invertible for any $z \in \mathbb{C} \setminus \{0\}$. Hence, an invertible polynomial matrix may not be invertible everywhere. If $P(z)$ is invertible everywhere, then $\det P(z) = cz^k$ with $c \neq 0$, which implies that its inverse $P^{-1}(z)$ is also a polynomial matrix. For a nonsquare P-matrix, its rank and its everywhere rank can be defined in the same way. For an arbitrary P-matrix, the following three types of operations are called *elementary row (column) operations*:

1. multiplying the ith row (column) by cz^k, where c is a nonzero constant and k is an integer;
2. interchanging the ith and the jth row (column);
3. adding the product of $p(z) \in \mathcal{P}$ and the jth row(column) to the ith row (column), where $i \neq j$.

A matrix that performs an elementary operation is said to be an *elementary P-matrix*, and a finite product of elementary matrices is said to be a *fundamental P-matrix*. It is clear that a P-matrix is fundamental if and only if it is invertible everywhere. Recall that a (nonsquare) everywhere full rank P-matrix can always be extended to a (square) fundamental P-matrix. Hence, for convenience, we also call a (nonsquare) everywhere full rank matrix a fundamental P-matrix.

Fundamental P-matrices play an important role in the study of transforms between the generators of a shift-invariant space. The first observation is that if $\dim \Phi = \dim \Psi$ and the transform from Ψ to Φ is a fundamental matrix, then $S(\Phi) = S(\Psi)$ and Φ is linearly independent (finitely linearly independent, stable) if and only if Ψ is linearly independent (finitely linearly independent, stable).

Generally, let Φ and Ψ be two generators of a shift-invariant space S. The transform T from Ψ to Φ is not necessarily fundamental. However, since both Φ and Ψ are compactly supported, T is at least an R-matrix, and T reduces to a P-matrix if $\Phi \in S_0(\Psi)$.

Now we begin to discuss finitely linear dependence with the following lemma.

LEMMA 2.2. *If Φ is finitely linearly dependent, then there is a finitely linearly independent distribution vector $\Psi \subset S_0(\Phi)$ such that $\dim \Psi = \Phi^\#$, $S_0(\Psi) = S_0(\Phi)$, and therefore the transform matrix from Ψ to Φ is fundamental.*

Proof. Set $r = \dim \Phi$ and $s = \Phi^\#$. By Theorem 2.1, there is a linearly independent generator $\Phi_I \subset S(\Phi)$ such that $\dim \Phi_I = s$ and $\Phi \subset S_0(\Phi_I)$. Therefore, there is a full rank matrix $T(z) \in \mathcal{P}^{r\times s}$ such that $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Phi}_I(\omega)$. Let $D(z)$ be the canonical Smith's form of $T(z)$. Since $T(z)$ has full rank, we have $D(z) = \binom{D_s}{O}$,

where $D_s(z) = \mathrm{diag}\{d_1(z), d_2(z), \ldots, d_s(z)\}$ with $d_i(z) \neq 0, 1 \leq i \leq s$, and O is the $(r-s) \times s$ zero matrix. Furthermore, there are two fundamental matrices $L(z) \in \mathcal{P}^r$ and $R(z) \in \mathcal{P}^s$ such that

$$T(z) = L(z)D(z)R(z). \tag{2.1}$$

Defining Ψ_I by

$$\hat{\Psi}_I(\omega) = D(e^{-i\omega})R(e^{-i\omega})\hat{\Phi}_I(\omega),$$

we obtain $\Psi_I = (\psi_1, \psi_2, \ldots, \psi_s, 0, \ldots, 0)^T$. Let $\Psi = (\psi_1, \psi_2, \ldots, \psi_s)^T$ and $L_s(z)$ be the $r \times s$ matrix containing the first s columns of $L(z)$. Then Ψ is finitely linearly independent and $\hat{\Phi}(\omega) = L_s(z)\hat{\Psi}(\omega)$, which implies that $\Phi \subset S_0(\Psi)$. On the other hand, we also have $\hat{\Psi}(z) = L\hat{}_s(e^{-i\omega})\hat{\Phi}(\omega)$, where $L\hat{}_s(z)$ is the $s \times r$ matrix containing the first s rows of $L^{-1}(z)$. Since $L(z)$ is fundamental, so are $L^{-1}(z)$, $L_s(z)$, and $L\hat{}_s(z)$. Hence, $\Psi \subset S_0(\Phi)$. The proof is completed. □

The following theorem characterizes finitely linear dependence of distribution vectors.

THEOREM 2.3. *A distribution vector Φ ($\dim \Phi = r$) is finitely linearly dependent if and only if there exists a nonidentity matrix $T(z) \in \mathcal{P}^r$ such that $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Phi}(\omega)$.*

Proof. "$\Longrightarrow$". Set $\Phi^{\#} = s$. Since Φ is finitely linearly dependent, $s < r$. Let Ψ_I, Ψ, $L(z)$, $L_s(z)$, and $L\hat{}_s(z)$ be defined as in the proof of the above lemma. Then we have $\hat{\Phi}(\omega) = L_s(e^{-i\omega})L_s\hat{}(e^{-i\omega})\hat{\Phi}(\omega)$ and $L\hat{}_s(e^{-i\omega})L_s(e^{-i\omega}) \neq I$.

"$\Longleftarrow$". Suppose that there is a P-matrix $T(z) \neq I$ such that

$$\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Phi}(\omega).$$

Then we have

$$0 = (I - T(e^{-i\omega}))\hat{\Phi}(\omega),$$

which implies that Φ is finitely linearly dependent. □

Now we characterize linear independence of a distribution vector.

THEOREM 2.4. *Φ is linearly dependent if and only if there is a distribution vector $\Psi \subset S(\Phi)$ and a nonfundamental P-matrix $P(z)$ such that*

$$\hat{\Phi}(\omega) = P(e^{-i\omega})\hat{\Psi}(\omega). \tag{2.2}$$

Proof. "$\Longrightarrow$". If Φ is linearly dependent, then, by [12], there is a nonzero vector (a_i) such that the distribution $\phi = \sum_{i=1}^{r} a_i\phi_i$ is linearly dependent. Without loss of generality, we may assume $a_1 = 1$. By [18], there is a polynomial $p(z)$ with at least one zero in $\mathbb{C} \setminus \{0\}$ and a distribution g in $S(\phi)$ such that $\hat{\phi}(\omega) = p(e^{-i\omega})\hat{g}(\omega)$. Therefore, we have

$$\begin{pmatrix} \hat{\phi}_1(\omega) \\ \hat{\phi}_2(\omega) \\ \vdots \\ \hat{\phi}_r(\omega) \end{pmatrix} = \begin{pmatrix} p(e^{-i\omega}) & -a_2 & \cdots & -a_r \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \begin{pmatrix} \hat{g}(\omega) \\ \hat{\phi}_2(\omega) \\ \vdots \\ \hat{\phi}_r(\omega) \end{pmatrix},$$

where $p(z)$ has at least one zero in $\mathbb{C} \setminus \{0\}$. Hence, the transform matrix is not fundamental.

"$\Longleftarrow$". Assume there exists a distribution vector $\Psi \subset S(\Phi)$ and a nonfundamental P-matrix $P(z)$ such that (2.2) holds. Write $\dim \Phi = r$ and $\dim \Psi = s$. If $r > s$, then Φ is finitely linearly dependent. We now assume $r \leq s$. Choose two square fundamental matrices $L(z)$ and $R(z)$ so that $P(z) = L(z)D(z)R(z)$, where $D(z)$ is the canonical Smith's form of $P(z)$. Setting $\hat{\Phi}_l(\omega) = L^{-1}(e^{-i\omega})\hat{\Phi}(\omega)$ and $\hat{\Psi}_r(\omega) = R(e^{-i\omega})\hat{\Psi}(\omega)$, respectively, we have $\hat{\Phi}_l(\omega) = D(e^{-i\omega})\hat{\Psi}_r(\omega)$. Let $D = \begin{pmatrix} D_r & O \end{pmatrix}$, where $D_r(e^{-i\omega}) = \operatorname{diag}\{d_1(e^{-i\omega}), \ldots, d_r(e^{-i\omega})\}$ and O is the $r \times (s-r)$ zero matrix. Since $P(z)$ is nonfundamental, at least one of $d_1(e^{-i\omega}), \ldots, d_r(e^{-i\omega})$ vanishes at an $\omega_0 \in \mathbb{C}$. Hence, one of the components of Φ_l is linearly dependent. It follows that Φ is also linearly dependent. ∎

COROLLARY 2.5. *The transform between two linearly independent generators (of a shift-invariant space) must be a fundamental matrix.*

We have similar results for stability of a distribution vector.

THEOREM 2.6. *Let* $\Gamma = \{z \in \mathbb{C}; \quad |z| = 1\}$. Φ *is unstable if and only if there is a distribution vector* $\Psi \subset S(\Phi)$ *such that the transform from* Ψ *to* Φ *is a P-matrix which is singular at a value* $z \in \Gamma$.

COROLLARY 2.7. *The transform between two stable generators (of a shift-invariant space) is an R-matrix invertible over* Γ.

3. Stability and linear independence of a scaling vector. We now characterize the linear independence (finitely linear independence, stability) of a scaling vector in terms of its symbol. We are interested in the scaling vector Φ with a P-matrix symbol. It is clear that if a scaling vector Φ has symbol $P(z)$, then its moment $\hat{\Phi}(0)$ satisfies $\hat{\Phi}(0) = P(1)\hat{\Phi}(0)$, which implies that $\hat{\Phi}(0)$ is either a zero vector or a 1-eigenvector of $P(1)$. If $P(1)$ has more than one 1-eigenvectors, then different scaling vectors may share one symbol.

We now assume that Φ is a solution of (1.1) with a nonvanished moment $\mathbf{v}$, which is a 1-eigenvector of $P(1)$. If (1.1) also has a nontrivial zero-moment solution Φ_0, then all functions $\Phi + c\Phi_0$, $c \in \mathbb{C}$, have the same symbol $P(z)$ and the same moment $\mathbf{v}$. On the other hand, if two scaling vectors Φ_1 and Φ_2 have the same symbol and the same moment, then $\Phi_1 - \Phi_2$ is a zero-moment scaling vector. Hence, we have the following conclusion. Let $\mathbf{v} \neq 0$ be a 1-eigenvector of $P(1)$. The solution of (1.1) with a given moment $\mathbf{v}$ is unique if and only if (1.1) has no nontrivial zero-moment solution.

As mentioned in [8], if (1.1) has nontrivial zero-moment solutions, then $P(1)$ must have eigenvalue 2^β, $(\beta \geq 1, \beta \in \mathbb{Z})$. We now denote by $E(M)$ the set of all eigenvalues of matrix M, and let

$$\mathcal{O}P^r = \left\{A \in \mathcal{P}^r; \quad 2^\beta \notin E(A(1)),\ \beta \in \mathbb{Z},\ \beta \geq 1\right\}$$

and

$$\mathcal{S}P^r = \left\{A \in \mathcal{O}P^r; \quad 1 \in E(A(1))\right\}.$$

Therefore, if $P \in \mathcal{S}P^r$ and $\mathbf{v}$ is a 1-eigenvector of $P(1)$, then (1.1) has a unique solution Φ with the moment $\mathbf{v}$.

In contrast with the fact that several scaling vectors may share a symbol, a scaling vector may have more than one symbol. However, the following lemma confirms the uniqueness of the symbol of a finitely linearly independent scaling vector.

LEMMA 3.1. *The symbol of a scaling vector* Φ *(with* $\dim \Phi = r$*) is unique in* $\mathcal{P}^r$ *if and only if* Φ *is finitely linearly independent.*

Proof. Assume Φ is finitely linearly independent. Then $\dim \Phi = \Phi^{\#}$. If Φ satisfies the two scaling equations

$$\Phi(x) = \sum_{k\in\mathbb{Z}} C_k \Phi\left(2x-k\right)$$

and

$$\Phi(x) = \sum_{k\in\mathbb{Z}} D_k \Phi\left(2x-k\right),$$

then

$$0 = \sum_{k\in\mathbb{Z}} (C_k - D_k)\Phi\left(2x-k\right). \tag{3.1}$$

Since $\Phi(\cdot)$ is finitely linearly independent, so is $\Phi(2\cdot)$. By $(C_k - D_k) \in (l_0)^{r\times r}$, we have $C_k - D_k = 0$. The converse is trivial. □

Since the relation between a scaling function and its symbols has been cleared, we are ready to characterize scaling vectors via their symbols. The notion of two-scale similarity plays an important role in characterizing scaling vectors [17]. A matrix $P \in \mathcal{P}^r$ is said to be *two-scale similar* to $Q \in \mathcal{P}^r$ if there exists an invertible matrix $T \in \mathcal{P}^r$ such that

$$P(z) = T(z^2)Q(z)T^{-1}(z). \tag{3.2}$$

The matrix $T(z)$ in (3.2) is called *two-scale similar transform.* Since the matrix $T^{-1}(z)$ need not necessarily be a polynomial matrix, $T(z)$ may be singular at some points in $\mathbb{C}\setminus\{0\}$. Hence, P being two-scale similar to Q does not imply Q being two-scale similar to P. However, if P is two-scale similar to Q with a fundamental transform, then Q also is two-scale similar to P. In this case, we say that P and Q are *fundamentally two-scale similar*. As shown in section 2, Ψ and Φ with $S(\Phi) = S(\Psi)$ have the same linear independence (finitely linear independence, stability) if and only if their symbols P and Q are fundamentally two-scale similar.

The following fact is often used in the rest. If P is fundamentally two-scale similar to Q with the two-scale transform T, and Φ satisfies the two-scaling equation

$$\hat{\Phi}(\omega) = P(e^{-i\omega/2})\hat{\Phi}(\omega/2),$$

then the scaling vector Ψ_I determined by

$$\hat{\Psi}_I(\omega) = T^{-1}(e^{-i\omega})\hat{\Phi}(\omega) \tag{3.3}$$

satisfies

$$\hat{\Psi}_I(\omega) = Q(e^{-i\omega/2})\hat{\Psi}_I(\omega/2). \tag{3.4}$$

The following theorem characterizes a finitely linearly dependent scaling vector.

THEOREM 3.2. *Let $P \in \mathcal{S}P^r$ be a symbol of Φ ($\dim\Phi = r$). Then Φ is finitely linearly dependent if and only if the following two conditions are satisfied.*

(1) *$P(z)$ is fundamentally two-scale similar to a matrix*

$$Q(z) = \begin{pmatrix} Q_s(z) & Y(z) \\ 0 & X(z) \end{pmatrix}, \tag{3.5}$$

where $Q_s \in \mathcal{S}P^s, X \in \mathcal{O}P^{r-s}$, *and* $s < r$.

(2) *Let* T *be the two-scale similar transform:* $P(z) = T(z^2)Q(z)T^{-1}(z)$. *Then the last* $r - s$ *components of* $T^{-1}(1)\hat{\Phi}(0)$ *vanish; that is*

$$T^{-1}(1)\hat{\Phi}(0) = (u_1, \ldots, u_s, 0, \ldots, 0)^T. \tag{3.6}$$

Furthermore, if $P(1)$ *has only a single* 1*-eigenvector, then the solution* Φ *of* (1.1) *with* $\hat{\Phi}(0) \neq 0$ *is unique, and therefore* Φ *is finitely linearly dependent if and only if* (1) *holds.*

Proof. " $\Longrightarrow$ ". Assume Φ is finitely linearly dependent with a symbol $P \in \mathcal{S}P^r$. By Lemma 2.2, there is a finitely linearly independent scaling vector Ψ such that $\hat{\Phi}(\omega) = \overline{T}(e^{-i\omega})\hat{\Psi}(\omega)$, where $\overline{T}(z)$ is an $r \times s$ fundamental matrix. Let $\Psi_I = (\psi_1, \ldots, \psi_s, 0, \ldots, 0)^T$ and $T(z)$ be an $r \times r$ fundamental matrix extended from $\overline{T}(z)$. Then $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Psi}_I(\omega)$. Let $Q(z) = T^{-1}(z^2)P(z)T(z)$. Then, setting $z = e^{-i\omega/2}$, we obtain

$$\begin{aligned}\hat{\Psi}(\omega) &= T^{-1}(z^2)\hat{\Phi}(\omega) = T^{-1}(z^2)P(z)\hat{\Phi}(\omega/2)\\ &= T^{-1}(z^2)P(z)T(z)\hat{\Psi}(\omega/2) = Q(z)\hat{\Psi}(\omega/2),\end{aligned}$$

which implies that $Q(z)$ is a symbol of Ψ_I. Writing

$$Q(z) = \begin{pmatrix} Q_s(z) & Y(z) \\ W(z) & X(z) \end{pmatrix},$$

we have

$$\begin{pmatrix} \hat{\Psi}(\omega) \\ 0 \end{pmatrix} = \begin{pmatrix} Q_s(z) & Y(z) \\ W(z) & X(z) \end{pmatrix} \begin{pmatrix} \hat{\Psi}(\omega/2) \\ 0 \end{pmatrix}, \qquad z = e^{-i\omega/2}. \tag{3.7}$$

From (3.7), we obtain $W(z)\hat{\Psi}(\omega/2) = 0$. Since Ψ is finitely linearly independent, by Lemma 3.1, $W(z) = 0$.

We now prove $Q_s \in \mathcal{S}P^s$. From (3.7), we have

$$\hat{\Psi}(\omega) = Q_s(e^{-i\omega/2})\hat{\Psi}(\omega/2), \tag{3.8}$$

which derives $\hat{\Psi}(0) = Q_s(1)\hat{\Psi}(0)$, and therefore 1 is an eigenvalue of $Q_s(1)$. Note that matrix $P(1)$ is similar to matrix $Q(1)$. Hence, they have the same set of eigenvalues. However, all eigenvalues of both $Q_s(1)$ a nd $X(1)$ are also eigenvalues of $Q(1)$. Therefore, we have the following:

$$P \in \mathcal{S}P^r \Longleftrightarrow Q \in \mathcal{S}P^r \Longleftrightarrow Q_s \in \mathcal{S}P^s \text{ and } \ X \in \mathcal{O}P^{r-s},$$

which implies (1).

Choosing $\omega = 0$ in $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Psi}_I(\omega)$, we obtain

$$T^{-1}(1)\hat{\Phi}(0) = (\hat{\psi}_1(0), \ldots, \hat{\psi}_s(0), 0, \ldots, 0)^T,$$

which is (3.6).

" $\Longleftarrow$ ". Let Φ be a scaling vector whose symbol P is fundamentally two-scale similar to a $Q(z)$:

$$P(z) = T(z^2)Q(z)T^{-1}(z),$$

where

$$Q(z) = \begin{pmatrix} Q_s(z) & Y(z) \\ 0 & X(z) \end{pmatrix}, \qquad s < r,$$

with $Q_s \in \mathcal{S}P^s$ and $X \in \mathcal{O}P^{r-s}$, and

$$T^{-1}(1)\hat{\Phi}(0) = (u_1, \ldots, u_s, 0, \ldots, 0)^T := \mathbf{u}_I.$$

We prove that Φ is finitely linearly dependent. Note that

$$Q_s \in \mathcal{S}P^s \text{ and } X \in \mathcal{O}P^{r-s} \Longrightarrow Q \in \mathcal{S}P^r.$$

Besides,

$$Q(1)\mathbf{u}_I = T^{-1}(1)P(1)T(1)\mathbf{u}_I = \mathbf{u}_I.$$

Hence, there is a unique distribution $\Psi_I = (\psi_{I,1}, \ldots, \psi_{I,r})^T$ satisfying (3.4) with the moment $\mathbf{u}_I$. Write $\mathbf{u} = (u_1, \ldots, u_s)^T$. By (3.2) and (3.5), we have $Q_s(1)\mathbf{u} = \mathbf{u}$. Let $\Psi = (\psi_1, \ldots, \psi_s)^T$ be the unique solution of (3.8) with the moment $\mathbf{u}$. Then the scaling vector $(\psi_1, \ldots, \psi_s, 0, \ldots, 0)^T$ is a solution of (3.4) with the moment $\mathbf{u}_I$. Since the solution of (3.4) with the moment $\mathbf{u}_I$ is unique, we have $\Psi_I = (\psi_1, \ldots, \psi_s, 0, \ldots, 0)^T$. By $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Psi}_I(\omega)$, Φ is finitely linearly dependent.

Finally, we prove that if $P(1)$ has only a single 1-eigenvector, then (2) will be unconditionally satisfied. In fact, if $P(1)$ has only a single 1-eigenvector, so do Q and $Q_s(z)$. Let $\mathbf{u} = (u_1, \ldots, u_s)^T$ be the 1-eigenvector of Q_s. Then $\mathbf{u}_I = (u_1, \ldots, u_s, 0, \ldots, 0)^T$ is the only 1-eigenvector (without counting the scalar multiples) of $Q(1)$. Recalling that $\hat{\Phi}(0)$ is the only 1-eigenvector (without counting the scalar multiples) of $P(1)$, we have $T^{-1}(1)\hat{\Phi}(0) = c\mathbf{u}_I = (cu_1, \ldots, cu_s, 0, \ldots, 0)^T$, which implies that (2) is true. □

Remark 1. If $P(1)$ has more than one (linearly independent) 1-eigenvector, then the solutions of (1.1) corresponding to different moments (without counting the scalar multiples) may have differently linearly independent properties. We illustrate this phenomenon in the following example.

Example 1. Let

$$P(z) = \begin{pmatrix} (\frac{1+z}{2})^2 & 0 \\ 0 & (\frac{1+z}{2})^4 \end{pmatrix}.$$

It is clear that $P(z) \in \mathcal{S}P^2$ and vector $(0, 1)$ and $(1, 1)$ are two linearly independent 1-eigenvectors of $P(1)$. Obviously, $P(z)$ is two-scale fundamentally similar to itself with transform $T(z) = I$. Denote by $N_m(x)$ the mth order cardinal B-spline [2]. The scaling vectors $\Phi_1 = (N_2(x), 0)^T$ and $\Phi_2 = (N_2(x), N_4(x))^T$ are solutions of (1.1) with the moments $\hat{\Phi}_1(0) = (1, 0)^T$ and $\hat{\Phi}_2(0) = (1, 1)^T$, respectively. Note that Φ_1 is finitely linearly dependent but Φ_2 is not. Note also that the vector $(1, 0)^T$ satisfies (3.6) while $(1, 1)^T$ does not.

Remark 2. If, in Theorem 3.2, $P(z)$ is merely assumed to be in $\mathcal{P}^r$, then the solution of (1.1) with a certain moment is not necessarily unique. In this case, if a solution of (1.1) Φ with $\hat{\Phi}(0) \neq 0$ is finitely linearly dependent, then its symbol P is still fundamentally two-scale similar to a matrix Q in the form (3.5), but Q_s is no longer necessarily in $\mathcal{S}P^r$ and X is no longer necessarily in $\mathcal{O}P^{r-s}$. On the other hand, if one of Φ's symbols is fundamentally two-scale similar to Q in the form (3.5)

with a matrix X such that the refinement equation $\hat{F}(\omega) = X(e^{-i\omega/2})\hat{F}(\omega/2)$ has only the trivial solution $F = 0$ with $\hat{F}(0) = 0$, and if $\hat{\Phi}(0)$ satisfies the condition (2) in 3.2, then Φ is finitely linearly dependent. The proof of this remark is similar to that for Theorem 3.2.

According to Theorem 3.2, from the set of symbols of a finitely linearly dependent scaling vector, we can select a relatively simple symbol for it.

COROLLARY 3.3. *If Φ ($\dim \Phi = r$) is finitely linearly dependent and $\Phi^{\#} = s$ and one of its symbols is in $\mathcal{S}P^r$, then Φ has a symbol fundamentally two-scale similar to a matrix*

$$\begin{pmatrix} Q_s(z) & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.9}$$

Proof. Let P be a symbol of Φ. By Theorem 3.2, P is fundamentally two-scale similar to a matrix Q in the form (3.5) with a two-scale transform matrix T, and $T^{-1}(1)\hat{\Phi}(0) = (u_1, \ldots, u_s, 0, \ldots, 0)^T$. Let Ψ_I be defined by $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Psi}_I(\omega)$. Since $X \in \mathcal{O}P^r$ and

$$\hat{\Psi}_I(0) = (u_1, \ldots, u_s, 0, \ldots, 0)^T,$$

we have

$$\Psi_I = (\psi_1, \ldots, \psi_s, 0, \ldots, 0)^T,$$

which satisfies

$$\begin{pmatrix} Q_s(e^{-i\omega/2}) & 0 \\ 0 & 0 \end{pmatrix} \hat{\Psi}(\omega/2) = \hat{\Psi}(\omega).$$

Setting

$$\bar{P}(z) = T^2(z) \begin{pmatrix} Q_s(z) & 0 \\ 0 & 0 \end{pmatrix} T^{-1}(z),$$

we have

$$\hat{\Phi}(\omega) = \bar{P}(e^{-i\omega/2})\hat{\Phi}(\omega/2),$$

where $\bar{P}(z)$ is a symbol of Φ. □

We now characterize linear independence of a scaling vector. First, we prove the following lemma.

LEMMA 3.4. *The symbol of a linearly independent scaling vector is a P-matrix.*

Proof. Assume the scaling vector Φ is linearly independent, as is $\Phi(2\cdot)$. By [22], a compactly supported function in $S(\Phi(2\cdot))$ is also in $S_0(\Phi(2\cdot))$. Hence, $\Phi \in S_0(\Phi(2\cdot))$, and therefore its symbol is a P-matrix. □

The following theorem characterizes linear independence of a scaling vector.

THEOREM 3.5. *Assume the scaling vector Φ with nonzero moment is finitely linearly independent and its symbol is a P-matrix $P \in \mathcal{P}^r(\mathcal{S}P^r)$. Then Φ is linearly dependent if and only if P is two-scale similar to a matrix $Q \in \mathcal{P}^r(\mathcal{S}P^r)$ with a nonfundamental two-scale similar transform T such that*

$$\operatorname{rank} T(1) = \operatorname{rank}\left(T(1), \hat{\Phi}(0)\right). \tag{3.10}$$

Remark 3. If the matrix $T(1)$ is nonsingular, then the condition (3.10) is always true. Hence, the condition (3.10) is effective whenever $T(1)$ is singular. Note that the condition (3.10) is equivalent to "there is a $\mathbf{u} \in \mathbb{C}^r$ such that $T(1)\mathbf{u} = \hat{\Phi}(0)$."

To prove this theorem, we need some notations. Let $m > 1$ be an odd integer, and let h_m be the smallest positive integer such that $2^{h_m} \equiv 1 (\mathrm{mod}\, m)$. For a primitive mth root of unit z_0, we call

$$p_m(z) = (z_0 - z)(z_0^2 - z) \cdots (z_0^{2^{h_m-1}} - z)$$

an *m-cycle polynomial*. Since $z_0^{2^{h_m}} = z_0$, it can be verified that $p_m(z_0^{2^l}) = 0$ for any integer $l \geq 0$, and therefore $p_m(z^2) = p_m(z)p_m(-z)$. We say that $p(z)$ has *m-cycle zeros* if $p_m(-z)$ (NOT $p_m(z)$!) is a factor of $p(z)$. When we will not stress the index m, m-cycle zeros and m-cycle polynomials are simply called cycle zeros and cycle polynomials, respectively. A polynomial $p(z)$ is said to have *symmetric zeros* if there is an $\alpha \in \mathbb{C}, \alpha \neq 0$, such that $p(\alpha) = p(-\alpha) = 0$.

Note that any polynomial $p(z)$ has a unique factorization

$$p(z) = cz^k \prod_{j=1}^{l} p_{m_j}(z) \prod_{l=1}^{t} (z - z_i), \tag{3.11}$$

where $p_{m_j}, j = 1, \ldots, l$, are all cycle polynomials, and $z_i \neq 0$, $i = 1, \ldots, t$.

LEMMA 3.6. *Assume the scaling vector Φ's symbol $P \in \mathcal{P}^r$ is two-scale similar to a P-matrix Q,*

$$P(z) = T_c(z^2)Q(z)T_c^{-1}(z), \tag{3.12}$$

with the two-scale transform matrix

$$T_c(z) = \mathrm{diag}(\underbrace{1, 1, \ldots,}_{j-1} c(z), \ldots, 1), \tag{3.13}$$

where $c(z)$ is either $(z - c)$, $c \neq 0$, or an m-cycle polynomial $p_m(z)$. Also assume $\hat{\phi}_j(0) = 0$ whenever $c(z) = z - 1$. Then there is a compactly supported function $\psi \in S(\Phi)$ such that $\hat{\phi}_j(\omega) = c(e^{-i\omega})\hat{\psi}(\omega)$, and therefore Φ is linearly dependent. Furthermore, if $P \in \mathcal{S}P^r$, then $Q \in \mathcal{S}P^r$.

Proof. Without loss of generality, we assume $j = 1$. By (3.12) and (3.13), we have

$$\hat{\phi}_1(\omega) = p_{11}(e^{-i\omega/2})\hat{\phi}_1(\omega/2) + \sum_{i=2}^{r} c(e^{-i\omega})q_{i1}(e^{-i\omega/2})\hat{\phi}_i(\omega/2), \tag{3.14}$$

where $p_{11}(z) = q_{11}(z)\frac{c(z^2)}{c(z)}$.

If $c(z) = z - c$, $c \neq 0$ and 1, then $c(z)$ cannot be a factor of $c(z^2)$. It follows that $c(z^2)$ is a factor of $p_{11}(z)$. Let ω_0 satisfy $e^{-i\omega_0} - c = 0$. We have, for an integer k,

$$\begin{aligned}
\hat{\phi}_1(\omega_0 + 2k\pi) &= p_{11}((-1)^k e^{-i\frac{\omega_0}{2}})\hat{\phi}_1\left(\frac{\omega_0}{2} + k\pi\right) \\
&\quad + \sum_{i=2}^{r} c(e^{-i\omega_0})q_{i1}((-1)^k e^{-i\frac{\omega_0}{2}})\hat{\phi}_i\left(\frac{\omega_0}{2} + \pi\right) \\
&= q_{11}((-1)^k e^{-i\omega_0/2})\frac{c(e^{-i\omega_0})}{c((-1)^k e^{-i\omega_0/2})}\hat{\phi}_1\left(\frac{\omega_0}{2} + k\pi\right) \\
&= 0.
\end{aligned}$$

In the case $c(z) = z - 1$, by the assumption, $\hat{\phi}_1(0) = 0$. For $k \neq 0$, writing $k = 2^l j$ with odd j, we have, by (3.14),

$$\begin{aligned}\hat{\phi}_1(2k\pi) &= p_{11}(e^{-ik\pi})\hat{\phi}_1(k\pi) = (e^{-i2^l j\pi} + 1)q_{11}(e^{-i2^l j\pi})\hat{\phi}_1(2^l j\pi) \\ &= (2q_{11}(1))^l\,(e^{-ij\pi} + 1)q_{11}(-1)\hat{\phi}_1(j\pi) = 0.\end{aligned}$$

We now prove $\hat{\phi}_1(\omega_0 + 2k\pi) = 0, k \in \mathbb{Z}$, for some $\omega_0 \in \mathbb{R}$ in the case $c(z) = p_m(z)$. The proof is similar to that one of Theorem 1 in [14]. For reader's convenience, we include it here. Let z_0 be a primitive mth root. Then z_0 has the form $e^{-i2n\pi/m}$, where n is an integer relatively prime to m, and $p_m(e^{-i2^{d+1}n\pi/m}) = 0$, for all integers $d \geq 0$. We claim that for all integers k,

$$\hat{\phi}_1(2n\pi/m + 2k\pi) = 0. \tag{3.15}$$

To prove (3.15), we write $n + km$ in the form $2^l j$, where l is a nonnegative integer and j is an odd integer. Recalling that $p_m(-z) = p_m(z^2)/p_m(z)$, we have

$$\begin{aligned}&\hat{\phi}_1\left(\frac{2n\pi}{m} + 2k\pi\right) = \hat{\phi}_1(2^{l+1}j\pi/m) \\ &= p_{11}(e^{-i2^l j\pi/m})\hat{\phi}_1\left(\frac{2^l j\pi}{m}\right) + \sum_{i=2}^{r} p_m(e^{-i2^{l+1}j\pi/m})q_{i1}(e^{-i2^l j\pi/m})\hat{\phi}_i\left(\frac{2^l j\pi}{m}\right) \\ &= q_{11}(e^{-i2^l j\pi/m})\frac{p_m(e^{-i2^{l+1}j\pi/m})}{p_m(e^{-i2^l j\pi/m})}\hat{\phi}_1\left(\frac{2^l j\pi}{m}\right) \\ &= p_{m(-}e^{-i2^l j\pi/m})q_{11}(e^{-i2^l j\pi/m})\hat{\phi}_1\left(\frac{2^l j\pi}{m}\right) \\ &= \left(\prod_{t=0}^{l} p_m(-e^{-i2^t j\pi/m})\prod_{t=0}^{l} q_{11}(-e^{-i2^t j\pi/m})\right)\hat{\phi}_1\left(\frac{j\pi}{m}\right).\end{aligned}$$

Hence, in order to prove (3.15) it suffices to show that $p_m(-e^{-ij\pi/m}) = 0$. For this purpose, we invoke Euler's theorem to find an integer $s > l$ such that $2^s \equiv 1 (\mathrm{mod}\, m)$. It follows that

$$j \equiv 2^s j \equiv 2^{s-l}(2^l j) \equiv 2^{s-l}n \pmod{m}. \tag{3.16}$$

Since j is odd, by (3.16), $j - 2^{s-l}n = (2t+1)m$ for some integer t. From this we see that

$$p_m(-e^{-ij\pi/m}) = p_m(e^{-i2^{s-l}n\pi/m}) = 0.$$

In all cases, ϕ_1 is linearly dependent, as is Φ. It is also clear that the function ψ defined by $\hat{\phi}_1(\omega) = c(e^{-i\omega})\hat{\psi}(\omega)$ is compactly supported and in the space $S(\Phi)$.

Finally, we prove $P \in \mathcal{S}P^r \Longrightarrow Q \in \mathcal{S}P^r$. When $c(z) \neq z - 1$, $Q(1)$ is similar to $P(1)$, and therefore $P \in \mathcal{S}P^r \Longleftrightarrow Q \in \mathcal{S}P^r$. When $c(z) = z - 1$, (3.12) implies

$$P(1) = \begin{pmatrix} p_{11}(1) & 0 \\ & P_{r-1}(1) \end{pmatrix}, \qquad Q(1) = \begin{pmatrix} p_{11}(1)/2 & * \\ 0 & P_{r-1}(1) \end{pmatrix}.$$

Therefore, $Q(1)$ preserves all eigenvalues of $P(1)$ except $p_{11}(1)$, which changes to $p_{11}(1)/2$. Note that $P \in \mathcal{S}P^r$. Hence, there is no positive integer β such that $p_{11}(1) =$

2^β. It follows that $p_{11}(1)/2 \notin \{2^\beta; \quad \beta \geq 0\}$ and, therefore, $Q \in \mathcal{O}P^r$. We now prove $1 \in E(Q(1))$ so that $Q \in \mathcal{S}P^r$. Let $\Phi_1 = (\phi_2, \ldots, \phi_r)^T$. Then $P(1)\hat{\Phi}(0) = \hat{\Phi}(0)$ and $\hat{\phi}_1(0) = 0$ imply that $\hat{\Phi}_1(0) \neq 0$ and $P_{r-1}(1)\hat{\Phi}_1(0) = \hat{\Phi}_1(0)$. It follows that $1 \in E(P_{r-1}(1))$ and $1 \in E(Q(1))$. □

The following is the proof of Theorem 3.5.

Proof. "$\Longrightarrow$". If Φ is linearly dependent, then, by Theorem 2.1, there exists a linearly independent scaling vector $\Psi \in S(\Phi)$ such that $\Phi \subset S_0(\Psi)$ and $S(\Psi) = S(\Phi)$. Since Φ is finitely linearly independent, $\dim \Phi = \dim \Psi$. By Lemma 3.4, Ψ's symbol Q is a P-matrix. Let T be the polynomial matrix determined by $\hat{\Phi}(\omega) = T(e^{-i\omega})\hat{\Psi}(\omega)$. Then $P(z)$ is two-scale similar to $Q(z)$ with the transform $T(z)$. By Corollary 2.5, T is nonfundamental. Besides, we have $T(1)\hat{\Psi}(0) = \hat{\Phi}(0)$. Thus, (3.10) is true.

"$\Longleftarrow$ ". We assume $P(z)$ is the symbol of Φ, and $P \in \mathcal{P}^r(\mathcal{S}P^r)$ is two-scale similar to $Q \in \mathcal{P}^r(\mathcal{S}P^r)$ with a nonfundamental two-scale transform matrix $T \in \mathcal{P}^r$. We now factor T into the form $T(z) = L(z)D(z)R(z)$, where L and R both are fundamental and D is the canonical Smith's form of T. It is clear that D is a nonfundamental diagonal matrix. We define Φ_E by $\hat{\Phi}_E(\omega) = L^{-1}(e^{-i\omega})\hat{\Phi}(\omega)$. Then Φ is linearly independent if and only if Φ_E is linearly independent. The symbol of Φ_E is $\tilde{P}(z) = L^{-1}(z^2)P(z)L(z)$. It is obvious that $\operatorname{rank} T(1) = \operatorname{rank}(T(1), \hat{\Phi}(0))$ if and only if $\operatorname{rank} D(1) = \operatorname{rank}(D(1), \hat{\Phi}_E(0))$. Let $\tilde{Q}(z) = R(z^2)Q(z)R^{-1}(z)$. Then $\tilde{P}(z) = D(z^2)\tilde{Q}(z)D^{-1}(z)$. We now factor $D(z)$ into $D(z) = \prod_s T_s(z) \prod_j U_j(z)$, where T_s is the diagonal matrix of the form (3.13) with $\det(T_s(1)) \neq 0$, while $U_j(z)$ is the diagonal matrix of the form (3.13) with $\det(U_j(1)) = 0$. If the product $\prod_s T_s(z)$ has at least one factor, by Lemma 3.6, Φ_E and hence Φ are linearly dependent. Now if $\prod_s T_s(z) = I$, then the product $\prod_j U_j(z)$ has at least one factor. Without loss of generality, we can assume $U_1(z) = \operatorname{diag}(z-1, 1, \ldots, 1)$. It follows that the first row of $D(1)$ is zero. Since there is a $\mathbf{v} \in \mathbb{C}^r$ such that $D(1)\mathbf{v} = \hat{\Phi}_E(0)$, $\hat{\phi}_{E,1}(0) = 0$. By Lemma 3.6, Φ_E and hence Φ are linearly dependent.

The proof of $P \in \mathcal{P}S^r \Longrightarrow Q \in \mathcal{S}P^r$ is similar to that one in Lemma 3.6. □

From Theorem 3.5, we derive a sufficient condition for linear independence of a scaling vector.

COROLLARY 3.7. *Assume the scaling vector Φ is finitely linearly independent and its symbol is a P-matrix $P \in \mathcal{P}^r$. Then Φ is linearly independent if* (i) *the matrix $(P(-1), \hat{\Phi}(0))$ has the full rank, and* (ii) $\det P(z)$ *has neither symmetric zeros nor cycle zeros.*

Proof. Assume that conditions (i) and (ii) hold and Φ is linearly dependent. By Theorem 3.5, there is a P-matrix Q and a nonfundamental P-matrix T such that $P(z) = T(z^2)Q(z)T^{-1}(z)$ and

$$\operatorname{rank} T(1) = \operatorname{rank}(T(1), \hat{\Phi}(0)).$$

Note that

$$\det P(z) = \det Q(z) \det T(z^2) / \det T(z). \tag{3.17}$$

Write $t(z) = \det T(z)$, $p(z) = \det P(z)$, and $q(z) = \det Q(z)$. Then (3.17) becomes $p(z) = q(z)t(z^2)/t(z)$, where t, p, and q are polynomials with $\deg(t) \geq 1$. We factor the polynomial $t(z)$ into $t(z) = (1-z)^s d(z)$ with $d(1) \neq 0$. If $\deg(d) \geq 1$, then $p(z)$ has either symmetric zeros or cycle zeros. This contradicts the condition (ii). Now if $\deg(d) = 0$, then $t(z) = c(1-z)^s$, $s \geq 1$. In this case, $T(1)$ is singular, and therefore $(T(1), \hat{\Phi}(0))$ does not have full rank. Since $P(-1) = T(1)Q(-1)T^{-1}(-1)$, the matrix $(P(-1), \hat{\Phi}(0))$ does not have full rank. This contradicts condition (i). □

Similarly, we have the following result for the stability of a scaling vector.

THEOREM 3.8. *Assume the scaling vector Φ is finitely linearly independent and its symbol is a P-matrix $P \in \mathcal{P}^r(\mathcal{S}P^r)$. Then Φ is unstable if and only if P is two-scale similar to a P-matrix $Q \in \mathcal{P}^r(\mathcal{S}P^r)$ with a nonfundamental two-scale transform matrix T such that* $\det T(z)$ *has zeros on* $|z| = 1$ *and* $\operatorname{rank} T(1) = \operatorname{rank}(T(1), \hat{\Phi}(0))$.

COROLLARY 3.9. *Assume the scaling vector Φ is finitely linearly independent and its symbol is a P-matrix $P \in \mathcal{P}^r$. Then Φ is stable if* (i) *the matrix* $(P(-1), \hat{\Phi}(0))$ *has full rank and* (ii) $\det P(z)$ *has neither symmetric zeros on* $|z| = 1$ *nor cycle zeros.*

Remark 3. When $r = 1$, the scaling vector Φ reduces to a single function. A sufficient and necessary condition for linear independence (stability) of a scaling function has been obtained by Jia and Wang [14]. Besides, Hogan [10] also obtained the same results as Corollary 3.7 and 3.9 in a different way.

4. Examples. In this section we give more examples.

Example 2. Consider the following scaling equation:

$$\Phi(x) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \Phi(2x) + \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} \Phi(2x-1) + \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \Phi(2x-2). \tag{4.1}$$

Its symbol is

$$P(z) = \frac{1}{2} \begin{pmatrix} 1+z+z^2 & -z \\ z+z^2 & -1+z^2 \end{pmatrix},$$

where matrix $P(1) = \left(\begin{smallmatrix} 3/2 & -1/2 \\ 1 & 0 \end{smallmatrix}\right)$ has eigenvalues 1 and 1/2. A right 1-eigenvector is $\mathbf{v}_0 = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$. $P(1)\mathbf{v}_0 = \mathbf{v}_0$. It is easy to check that

$$\frac{1}{2} \begin{pmatrix} 1+z+z^2 & -z \\ z+z^2 & -1+z^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ z^2 & -1 \end{pmatrix} \begin{pmatrix} (1+z)/2 & z/2 \\ 0 & (-1+z^2+z^3)/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ z & -1 \end{pmatrix},$$

and the equation $x(\cdot) = -x(2\cdot) + x(2\cdot -2) + x(2\cdot -3)$ has only the trivial compactly supported solution $x(\cdot) = 0$. Hence, by Theorem 3.2, the solution of (4.1) with the mean value $\begin{pmatrix} \hat{\phi}_1(0) & \hat{\phi}_2(0) \end{pmatrix}^T = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$ is finitely linearly dependent. In fact, this solution is $\Phi(x) = \begin{pmatrix} \chi(0,1] & \chi(1,2] \end{pmatrix}^T$. By Corollary 3.3, one of Φ's symbols is

$$\begin{pmatrix} (1+z)/2 & 0 \\ z^2(1+z)/2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ z^2 & -1 \end{pmatrix} \begin{pmatrix} (1+z)/2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ z & -1 \end{pmatrix}.$$

We can verify that $\Phi(x)$ satisfies the following equation:

$$\Phi(x) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Phi(2x) + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Phi(2x-1) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \Phi(2x-2) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \Phi(2x-3).$$

By Theorem 3.2, $\Phi(x)$ will satisfy any equation with a symbol having the form of

$$\begin{pmatrix} 1 & 0 \\ z^2 & -1 \end{pmatrix} \begin{pmatrix} (1+z)/2 & Y(z) \\ 0 & X(z) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ z & -1 \end{pmatrix},$$

where $Y(z)$ and $X(z)$ are arbitrary polynomials such that $X(1) \neq 2^{\beta}(\beta \geq 1, \beta \in \mathbb{Z})$. For example, if we set $Y(z) = -1/2$, $X(z) = z/2$, then, by Theorem 3.2, the matrix

$$\begin{pmatrix} 1/2 & 1/2 \\ 0 & (z+z^2)/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ z^2 & -1 \end{pmatrix} \begin{pmatrix} (1+z)/2 & -1/2 \\ 0 & z/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ z & -1 \end{pmatrix}$$

should be a new symbol of Φ. It can be verified that $\Phi(x)$ satisfies the following equation:

$$\Phi(x) = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \Phi(2x) + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Phi(2x-1) + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Phi(2x-2).$$

Example 3. Now we analyze the scaling vector in Example 4.6 in [15] (for the case of $L = r = 2$). The symbol of the scaling vector $\Phi = (\phi_1, \phi_2)^T$ is

$$P(z) = \left(\frac{1+z}{2}\right)^2 \begin{pmatrix} 1/2 & 1/2 \\ z/2 & 1/2 \end{pmatrix}.$$

Using the formula of Theorem 2.3 in [20], we obtain $\operatorname{supp}\phi_1 \subset [0, 2\frac{1}{3}]$ and $\operatorname{supp}\phi_2 \subset [0, 2\frac{2}{3}]$. The graphs of $(\phi_1, \phi_2)^T$, which can be found in Figure 4.1 of [15], show that $\operatorname{supp}\phi_1 = [0, 2\frac{1}{3}]$ and $\operatorname{supp}\phi_2 = [0, 2\frac{2}{3}]$. Since the right end-point of the support of any finitely linear combination of the integer translates of ϕ_1 is $k+1/3$ while the right end-point of the support of any finitely linear combination of the integer translates of ϕ_2 is $k + 2/3$, scaling vector $(\phi_1, \phi_2)^T$ is finitely linearly independent. Matrix $P(1) = \left(\begin{smallmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{smallmatrix}\right)$ has eigenvalues 1 and 0, and $\begin{pmatrix} \hat{\phi}_1(0) & \hat{\phi}_2(0) \end{pmatrix}^T = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$ is a 1-eigenvector. $P(z)$ has the following two-scale similarity:

$$P(z) = T(z^2)Q(z)T^{-1}(z),$$

where

$$T(z) = \begin{pmatrix} 1 & 0 \\ 1 & z-1 \end{pmatrix}, \quad Q(z) = \begin{pmatrix} (\frac{1+z}{2})^2 & (\frac{z-1}{2})(\frac{z+1}{2})^2 \\ \frac{1+z}{8} & 0 \end{pmatrix}.$$

Note that $T(z)$ is nonfundamental and the equation $T(1)\mathbf{u} = (1 \quad 1)^T$ has a solution $\mathbf{u} = (1, 0)^T$. By Theorem 3.5, $(\phi_1, \phi_2)^T$ is linearly dependent. Indeed, we have $\sum_{k\in\mathbb{Z}}(\phi_1(x-k) - \phi_2(x-k)) = 0$.

Example 4. Finally, we consider the orthonormal fractal scaling vector in [5]. Its symbol is

$$P(z) = \frac{\sqrt{2}}{20} \begin{pmatrix} 6\sqrt{2}(1+z) & 16 \\ -(1+z)(1-10z+z^2) & -\sqrt{2}(3-10z+3z^2) \end{pmatrix},$$

and the initial vector is $\hat{\Phi}(0) = \begin{pmatrix} \sqrt{2} & 1 \end{pmatrix}^T$. Using the formula of Theorem 2.3 in [20], we obtain that $\operatorname{supp}\phi_1 \subset [0, 1]$ and $\operatorname{supp}\phi_2 \subset [0, 2]$. We first point out that Φ is finitely linearly independent. Indeed, since $\operatorname{supp}\phi_1 \subset [0, 1]$, the integer translates of ϕ_1 are linearly independent. It follows that the entire function $\hat{\phi}_1$ has no 2π-periodic zero in $\mathbb{C}$. Assuming Φ is finitely linearly dependent, then there are two polynomials $a(z)$ and $b(z)$ such that $a(z)\hat{\phi}_1(\omega) - b(z)\hat{\phi}_2(\omega) = 0$, $z = e^{-i\omega}$. Hence, $\hat{\phi}_2(\omega) = \frac{a(z)}{b(z)}\hat{\phi}_1(\omega)$. Since $\hat{\phi}_2$ is also an entire function and $\hat{\phi}_1$ has no 2π-periodic zero in $\mathbb{C}$, $s(z) := \frac{a(z)}{b(z)}$ must be a polynomial. Furthermore, since $\operatorname{supp}\phi_1 \subset [0, 1]$ and $\operatorname{supp}\phi_2 \subset [0, 2]$, $s(z)$ must be a

linear polynomial. Assuming that $s(z) = (c + dz)$, we have $\hat{\phi}_2(\omega) = (c + de^{-i\omega})\hat{\phi}_1(\omega)$. Therefore, $T(e^{-i\omega})\hat{\Phi}(\omega) = (\hat{\phi}_1(\omega), 0)^T$, where $T(z) = \left(\begin{smallmatrix} 1 & 0 \\ c+dz & -1 \end{smallmatrix}\right)$. By Theorem 3.2, we also have $T^{-1}(z^2)P(z)T(z) = \left(\begin{smallmatrix} * & * \\ 0 & * \end{smallmatrix}\right)$; thus

$$(c + dz^2)\left(6\sqrt{2}(1 + z) + 16(c + dz)\right) + (1 + z)(1 - 10z + z^2)$$
$$+\sqrt{2}(c + dz)(3 - 10z + 3z^2) = 0.$$

However, this equation has no solution for c and d. Hence, Φ is finitely linearly independent. Finally, we confirm the linear independence of Φ from the fact that $\det P(z) = -\frac{1}{40}(z + 1)^3$ and $(P(-1), \hat{\Phi}(0)) = \left(\begin{smallmatrix} 0 & 4\sqrt{5} & \sqrt{2} \\ 0 & -8/5 & 1 \end{smallmatrix}\right)$ has full rank.

Acknowledgment. The author would like to thank the referees for their helpful comments.

REFERENCES

[1] C. K. Chui and Jianzhong Wang, *A general framework of compactly supported spline and wavelets*, J. Approx. Theory, 71 (1993), pp. 263–304.
[2] C. K. Chui and Jianzhong Wang, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–916.
[3] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of Refinable Function Vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.
[4] C. de Boor, R. DeVore, and A. Ron, *Approximation from shift-invariant subspaces of $L^2(\mathbb{R}^d)$*, J. Funct. Anal., 119 (1994), pp. 37–78.
[5] G. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1158–1192.
[6] T. N. T. Goodman and S. L. Lee, *Wavelet of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.
[7] T. N. T. Goodman, S. L. Lee, and W. S. Tang, *Wavelets in wandering subspaces*, Trans. Amer. Math. Soc., 338 (1993), pp. 639–654.
[8] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.
[9] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., 73 (1996), pp. 75–94.
[10] T. A. Hogan, *Stability and Independence of the Shifts of Finitely Many Refinable Functions*, J. Fourier Anal. Appl., to appear.
[11] Rong-Qing Jia, *Shift-invariant spaces on the real line*, Proc. Amer. Math. Soc., 125 (1997), pp. 785–793.
[12] Rong-Qing Jia and C. A. Micchelli, *On linear independence for integer translates of a finite number of functions*, Proc. Edinburgh Math. Soc., 36 (1992), pp. 69–85.
[13] Rong-Qing Jia, S. D. Riemenschneider, and Ding-Xuan Zhou, *Approximation by multiple refinable functions*, Canad. J. Math., to appear.
[14] Rong-Qing Jia and Jianzhong Wang, *Stability and linear independence associated with wavelet decompositions*, Proc. Amer. Math. Soc., 117 (1993), pp. 1115–1124.
[15] P. Massopust, D. Ruch, and P. Van Fleet, *On the support properties of scaling vectors*, App. Comput. Harmon. Anal., 3 (1996), pp. 229–238.
[16] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.
[17] G. Plonka and V. Strela, *Construction of Multi-Scaling Functions with Approximation and Symmetry*, Fachbereich Mathematik, Universität Rostock, Germany, 1996.
[18] A. Ron, *A necessary and sufficient condition for the linear independence of the integer translates of a compactly supported distribution*, Constr. Approx., 5 (1989), pp. 297–308.
[19] D. Ruch, W. So, and Jianzhong Wang, *Global Support of a Scaling Vector*, Appl. Comput. Harmon. Anal., to appear.
[20] W. So and Jianzhong Wang, *Estimating the support of a scaling vector*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 66–73.

[21] V. Strela, *Multiwavelets: Regularity, Orthogonality and Symmetry Via Two-Scale Similarity Transform*, preprint, Massachusetts Institute of Technology, Cambridge, MA, 1995.
[22] Kang Zhao, *Global linear independence and finitely supported dual basis*, SIAM J. Math. Anal., 23 (1992), pp. 1352–1355.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1157–1176, September 1998

007

ON THE REGULARITY OF MATRIX REFINABLE FUNCTIONS*

QINGTANG JIANG†

Abstract. It is shown that the transition operator $\mathbf{T_P}$ associated with the matrix refinement mask $\mathbf{P}(\omega) = 2^{-d}\Sigma_{\alpha\in[0,N]^d}\mathbf{P}_\alpha\exp(-i\alpha\omega)$ is equivalent to the matrix $(2^{-d}\mathcal{A}_{2i-j})_{i,j}$ with $\mathcal{A}_j = \Sigma_{\kappa\in[0,N]^d}\mathbf{P}_{\kappa-j}\otimes\mathbf{P}_\kappa$ and $\mathbf{P}_{\kappa-j}\otimes\mathbf{P}_\kappa$ denoting the Kronecker product of matrices $\mathbf{P}_{\kappa-j}$, $\mathbf{P}_\kappa$. Some spectral properties of $\mathbf{T_P}$ are studied and a complete characterization of the matrix refinable functions in the Sobolev space $W^n(\mathbf{R}^d)$ for nonnegative integers n is provided. The Sobolev regularity estimate of the matrix refinable function is given in terms of the spectral radius of a restricted transition operator. These estimates are analyzed in some examples.

Key words. matrix refinable function, transition operator, regularity

AMS subject classifications. 42C15, 39B62, 42B05, 41A15

PII. S003614109630817X

1. Introduction. Let $\{\mathbf{P}_\alpha\}$ be a real $r\times r$ matrix sequence with finite elements nonzero. The vectors Φ, r-dimensional column functions, used in this paper are solutions to functional equations of the type

$$\Phi = \sum_{\alpha\in\mathbf{Z}^d}\mathbf{P}_\alpha\Phi(2\cdot-\alpha). \tag{1.1}$$

Define

$$\mathbf{P}(\omega) := 2^{-d}\sum_{\alpha\in\mathbf{Z}^d}\mathbf{P}_\alpha\exp(-i\alpha\omega);$$

then, in the Fourier domain, functional equations (1.1) can be written as

$$\widehat{\Phi}(\omega) = \mathbf{P}(\omega/2)\widehat{\Phi}(\omega/2). \tag{1.2}$$

Equations of the type (1.1) or (1.2) are called *matrix* (*vector*) *refinement equations*; $\mathbf{P}$ ($\{\mathbf{P}_\alpha\}$) is called the (*matrix*) *refinement mask*, and any solution Φ of (1.1) is called a *matrix refinable function* (or *refinable vector*). Equations (1.1) are considered in the area of wavelets for the construction of multiwavelets, and there are many papers on the existence of the solutions of equations (1.1), the constructions of multiwavelets, and related topics; see, e.g., [1], [3], [7], [8], [11]–[16], [21]–[23], [25]–[27], and [29]–[31]. The present paper considers the Sobolev regularity of the matrix refinable functions.

For the case $r = 1, d = 1$, compactly supported refinable functions are solutions of the two-scale equation

$$\phi(x) = \sum_{j=0}^{J} h_j\phi(2x-j).$$

*Received by the editors August 14, 1996; accepted for publication (in revised form) October 2, 1997; published electronically April 15, 1998. This work was supported by an NSTB post-doctoral research fellowship at The National University of Singapore.

http://www.siam.org/journals/sima/29-5/30817.html

†Department of Mathematics, The National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260 and Department of Mathematics, Peking University, 100871 Beijing (qjiang@haar.math.nus.edu.sg).

Over the years, several techniques have been developed to determine the regularity of refinable functions; see [5], [9], [32], [6], [10], [17], and [2] (in [17] and [2], the refinement mask $\{h_j\}$ is not necessarily finitely supported). One of the main results is the following (see [32], [9]): assume that the refinement mask

$$m_0(\omega) = \frac{1}{2}\sum_{j=0}^{J} h_j e^{-ij\omega} \tag{1.3}$$

can be factorized as

$$m_0(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^K q(\omega), \tag{1.4}$$

where $q(\omega)$ is a trigonometric polynomial. Then the Sobolev exponent $s(\phi) := \sup\{s \geq 0 : \int(1+|\omega|^2)^s|\widehat{\phi}(\omega)|^2 d\omega < +\infty\}$ satisfies

$$s(\phi) \geq K - \log_4 \rho(\mathbf{T}_q),$$

where $\mathbf{T}_q$ is the transition operator associated with q and $\rho(\mathbf{T}_q)$ is the spectral radius of $\mathbf{T}_q$. For a trigonometric polynomial $p(\omega) = \sum_{l=0}^{L} p_l e^{-il\omega}$, the transition operator associated with p is defined by

$$\mathbf{T}_p f(\omega) := \left|p\Big(\frac{\omega}{2}\Big)\right|^2 f\Big(\frac{\omega}{2}\Big) + \left|p\Big(\frac{\omega}{2}+\pi\Big)\right|^2 f\Big(\frac{\omega}{2}+\pi\Big), \quad f \in V_L,$$

where V_L denotes the vector space of trigonometric polynomials defined by

$$V_L := \left\{ \sum_{l=-L}^{L} f_l e^{-il\omega} : \quad f_l \in \mathbf{C} \right\}.$$

Further, if refinable function ϕ is stable and $q(\pi) \neq 0$, then above regularity estimate is optimal; i.e.,

$$s(\phi) = K - \log_4 \rho(\mathbf{T}_q).$$

There is another method to give regularity estimates of refinable functions. Let ϕ be a compactly supported refinable function with corresponding mask $m_0(\omega)$ given by (1.3) for some positive integer J. Assume that $m_0(\omega)$ satisfies the vanishing moment conditions of order $K+1$; i.e., $\frac{d^\alpha}{d\omega^\alpha} m_0(\omega)|_{\omega=\pi} = 0, 0 \leq \alpha \leq K$. Equivalently, $m_0(\omega)$ can be written in the form of (1.4). Let V_J^0 denote the subspace of V_J defined by

$$V_J^0 := \left\{ f \in V_J : \quad \sum_{j=-J}^{J} j^n f_j = 0, \quad n = 0, \ldots, 2K-1 \right\}. \tag{1.5}$$

Then V_J^0 is invariant under $\mathbf{T}_{m_0}$. Let $\mathbf{T}_{m_0}|_{V_J^0}$ denote the restriction of $\mathbf{T}_{m_0}$ to V_J^0. If $\rho(\mathbf{T}_{m_0}|_{V_J^0}) < 1$, then

$$s(\phi) \geq -\log_4 \rho(\mathbf{T}_{m_0}|_{V_J^0}).$$

In fact the above two methods are completely equivalent; see [6]. The first method relies upon the factorization of the refinement mask $m_0(\omega)$. However in the higher

dimension case, the refinement mask is often irreducible. The second method was successfully used by Riemenschneider and Shen to estimate the regularities of two dimension refinable functions constructed in [28]. Further studies on the problem of the regularity in higher dimensions with dilation matrices were carried out in [20] and [4].

The regularity of the matrix refinable function Φ (for the case $d = 1$) was first studied by Cohen, Daubechies, and Plonka [3] based on the factorization of the matrix refinement mask $\mathbf{P}(\omega)$. However such estimates of regularity are usually hard to compute. There is another approach (essentially the second method for the scalar case) to the regularity estimate of the refinable vector Φ carried out by Shen in [29], and such estimates are provided in terms of the spectral radius of a restricted transition operator. More precisely, letting $\mathbf{P}(\omega) = 2^{-d}\sum_{\alpha\in[0,N]^d}\mathbf{P}_\alpha e^{-i\alpha\omega}$ be the corresponding matrix refinement mask, the *transition operator* $\mathbf{T_P}$ associated with $\mathbf{P}$ is defined by

$$\mathbf{T_P}H(\omega) := \sum_{\nu\in\mathbf{Z}^d/2\mathbf{Z}^d}\mathbf{P}\Big(\frac{\omega}{2}+\pi\nu\Big)H\Big(\frac{\omega}{2}+\pi\nu\Big)\mathbf{P}^*\Big(\frac{\omega}{2}+\pi\nu\Big), \quad H\in\mathbf{H}_N. \tag{1.6}$$

Throughout this paper $\mathbf{H}_N$ denotes the space of all $r\times r$ matrices with each entry a trigonometric polynomial whose Fourier coefficients are supported in $[-N,N]^d$; M^* and M^T denote the Hermitian adjoint and the transpose of a matrix M, respectively. The transition operator $\mathbf{T_P}$ leaves $\mathbf{H}_N$ invariant. In [29], the regularity of Φ was given in terms of the spectral radius of $\mathbf{T_P}|_{\mathbf{H}_N^0}$, the restricted operator of $\mathbf{T_P}$ to an invariant subspace $\mathbf{H}_N^0$ of $\mathbf{H}_N$ under $\mathbf{T_P}$. The smaller the invariant subspace $\mathbf{H}_N^0$, the smaller $\rho(\mathbf{T_P}|_{\mathbf{H}_N^0})$ will be and hence the better the estimate on the regularity of Φ. Thus a small $\mathbf{T_P}$ invariant subspace of $\mathbf{H}_N$ is required.

For the case $r = 1, d = 1$, let m_0 be a given refinement mask defined by (1.3) for some positive integer J; then the transition operator $\mathbf{T}_{m_0}$ is equivalent under the basis $\{e^{-ij\omega}\}_{j=-J}^{J}$ of V_J to the matrix

$$\mathcal{T}_{m_0} = (2^{-1}a_{2i-j})_{-N\le i,j\le N},$$

where a_j is the autocorrelation of $\{c_\kappa\}$ defined by $a_j := \sum_\kappa c_{\kappa-j}\overline{c_\kappa}$; see [24], [6]. We note that the invariant subspace V_J^0 defined by (1.5) can be written as

$$V_J^0 = \{f\in V_J: \quad v_n(f_{-J},\dots,f_J)^T = 0, \quad n = 0,\dots,2K-1\},$$

where

$$v_n := ((-J)^n,\dots,J^n), \quad n = 0,\dots,2K-1. \tag{1.7}$$

The row vector v_n is a generalized left 2^{-n}-eigenvector of the matrix $\mathcal{T}_{m_0}$ (see [6]). Thus to give the regularity estimates of refinable vectors, we at first change equivalently the transition operator $\mathbf{T_P}$ into its representing matrix $\mathcal{T}_\mathbf{P}$, then find left 2^{-n}-eigenvectors of the matrix $\mathcal{T}_\mathbf{P}$. Using these left eigenvectors, we construct the invariant subspace $\mathbf{H}_N^0$ and then provide the regularity estimates in terms of the spectral radius of the restricted transition operator $\mathbf{T_P}|_{\mathbf{H}_N^0}$.

This paper is organized as follows. In section 2 we show that the transition operator $\mathbf{T_P}$ is equivalent to the matrix $\mathcal{T}_\mathbf{P} = (2^{-d}\mathcal{A}_{2i-j})_{i,j\in[-N,N]^d}$, where $\mathcal{A}_j$ is the $r^2\times r^2$ matrix given by

$$\mathcal{A}_j = \sum_{\kappa\in[0,N]^d}\mathbf{P}_{\kappa-j}\otimes\mathbf{P}_\kappa$$

and $\mathbf{P}_{\kappa-j}\otimes\mathbf{P}_\kappa$ is the Kronecker product of $\mathbf{P}_{\kappa-j}$ and $\mathbf{P}_\kappa$. In section 2, we also find left eigenvectors of $\mathcal{T}_\mathbf{P}$ which will be used for the regularity estimate of refinable vectors. In the first part of section 3, we will give a characterization of refinable vectors in the Sobolev space $W^n(\mathbf{R}^d), n \in \mathbf{Z}_+$. In the second part of section 3, we provide a $\mathbf{T_P}$-invariant subspace $\mathbf{H}_N^0$ of $\mathbf{H}_N$ and give the regularity estimate of refinable vectors in terms of the spectral radius of the restricted transition operator $\mathbf{T_P}|_{\mathbf{H}_N^0}$. In the last part of this paper, section 4, we will give the estimates on the smoothness of some matrix refinable functions. About the B-splines defined by knots $0, 0, 1, 1$ and $0, 1, 1, 2$ and the GHM-orthogonal scaling functions, our estimates on their regularities are optimal.

Before going to the next section, we introduce some notation used in this paper. Let $\mathbf{Z}_+$ denote the set of all nonnegative integers and $\mathbf{Z}_+^d$ denote the set of all d-tuples of nonnegative integers. We shall adopt the multi-index notation

$$\omega^\beta := \omega^{\beta_1}\cdots\omega_d^{\beta_d}, \quad \beta! := \beta_1!\cdots\beta_d!, \quad |\beta| := \beta_1+\cdots+\beta_d$$

for $\omega = (\omega_1,\ldots,\omega_d)^T \in \mathbf{R}^d$, $\beta = (\beta_1,\ldots,\beta_d)^T \in \mathbf{Z}_+^d$. If $\alpha, \beta \in \mathbf{Z}^d$ satisfy $\beta-\alpha \in \mathbf{Z}_+^d$, we shall write $\alpha \le \beta$ and denote

$$\binom{\beta}{\alpha} := \frac{\beta!}{\alpha!(\beta-\alpha)!}.$$

For $\beta = (\beta_1,\ldots,\beta_d)^T \in \mathbf{Z}_+^d$, denote

$$D^\beta := \frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}}\cdots\frac{\partial^{\beta_d}}{\partial x_d^{\beta_d}},$$

where $\partial_j = \frac{\partial}{\partial x_j}$ is the partial derivative operator with respect to the jth coordinate, $1 \le j \le d$. For $\omega, \zeta \in \mathbf{R}^d$, we use $\zeta\omega$ to denote their scalar product.

For $j = 1,\ldots,r$, let $\mathbf{e}_j := (\delta_j(k))_{k=1}^r$ denote the standard unit vectors in $\mathbf{R}^r$. In this paper, for an $r\times 1$ vector function $f = (f_1,\ldots,f_r)^T$, *f is in a space on* $\mathbf{R}^d$ means that every component f_i of f is in this space, and we will use the notation $|f| := (\sum_{i=1}^r |f_i|^2)^{\frac{1}{2}}$.

For a matrix or an operator A, we say$A$ satisfies *Condition E* if the spectral radius of $A \le 1$, 1 is the unique eigenvalue of A on the unit circle and 1 is simple. For two matrices A, B, $A \le B$ should be understood as stating that $B-A$ is positive semidefinite.

For a finitely supported sequence s on $\mathbf{Z}^d$, its support is defined by $\mathrm{supp}s := \{\beta \in \mathbf{Z}^d : s(\beta) \ne 0\}$, and for a finitely supported $r\times r$ matrix sequence S on $\mathbf{Z}^d$, its support is defined by $\mathrm{supp}S := \cup\mathrm{supp}s_{ij}$, where s_{ij} is the (i,j)-entry of S. Throughout this paper, we assume that the matrix refinement mask $\mathbf{P}$ satisfies $\mathrm{supp}\{\mathbf{P}_\alpha\} \subset [0,N]^d$ for some positive integer N, and we use c to denote the universal constant which may be different at different occurrences.

2. Transition operator. In this section, we first show that the transition operator $\mathbf{T_P}$ defined by (1.6) is equivalent to a matrix, then we study some spectral properties of $\mathbf{T_P}$.

For any $H = \sum_{j\in[-N,N]^d} H_j e^{-ij\omega} \in \mathbf{H}_N$,

$$\mathbf{P}(\omega)H(\omega)\mathbf{P}(\omega)^* = 2^{-2d}\sum_{\ell\in[0,N]^d}\sum_{\kappa\in[0,N]^d}\mathbf{P}_\kappa e^{-i\omega\kappa}H(\omega)\mathbf{P}_\ell^T e^{i\omega\ell}$$

$$= 2^{-2d} \sum_{\ell\in[0,N]^d} \sum_{\kappa\in[0,N]^d} \mathbf{P}_\kappa H(\omega)\mathbf{P}_\ell^T e^{-i\omega(\kappa-\ell)}$$
$$= 2^{-2d} \sum_{\kappa\in[0,N]^d} \sum_{n\in[-N,N]^d} \mathbf{P}_\kappa H(\omega)\mathbf{P}_{\kappa-n}^T e^{-i\omega n}$$
$$= 2^{-2d} \sum_{j\in[-N,N]^d} \sum_{\kappa\in[0,N]^d} \sum_{n\in[-N,N]^d} \mathbf{P}_\kappa H_j \mathbf{P}_{\kappa-n}^T e^{-i\omega(n+j)}.$$

Thus

$$\mathbf{T_P} H(\omega) = 2^{-2d} \sum_{\nu\in\mathbf{Z}^d/2\mathbf{Z}^d} \sum_{j\in[-N,N]^d} \sum_{n\in[-N,N]^d} \sum_{\kappa\in[0,N]^d} \mathbf{P}_\kappa H_j \mathbf{P}_{\kappa-n}^T (-1)^{\nu(n+j)} e^{-i\frac{\omega}{2}(n+j)}.$$

For any $n \in [-N,N]^d, j \in [-N,N]^d$, write $n+j = 2\ell+\mu$ for some $\ell \in [-N,N]^d$ and $\mu \in \mathbf{Z}^d/2\mathbf{Z}^d$. By the fact that $\sum_{\nu\in\mathbf{Z}^d/2\mathbf{Z}^d}(-1)^{\nu\mu} = 2^d\delta_\mu$,

$$\sum_{\nu\in\mathbf{Z}^d/2\mathbf{Z}^d} (-1)^{\nu(n+j)} = 2^d\delta_\mu.$$

Hence

$$\mathbf{T_P} H(\omega) = 2^{-d} \sum_{j\in[-N,N]^d} \sum_{\ell\in[-N,N]^d} \sum_{\kappa\in[0,N]^d} \mathbf{P}_\kappa H_j \mathbf{P}_{\kappa-(2\ell-j)}^T e^{-i\omega\ell} \tag{2.1}$$
$$= \sum_{\ell\in[-N,N]^d} \left(2^{-d} \sum_{j\in[-N,N]^d} \sum_{\kappa\in[0,N]^d} \mathbf{P}_\kappa H_j \mathbf{P}_{\kappa-(2\ell-j)}^T \right) e^{-i\omega\ell}.$$

That is, $\mathbf{T_P}$ changes sequence $\{H_j\}_{j\in[-N,N]^d}$ into another sequence:

$$\left\{ 2^{-d} \sum_{j\in[-N,N]^d} \sum_{\kappa\in[0,N]^d} \mathbf{P}_\kappa H_j \mathbf{P}_{\kappa-(2\ell-j)}^T \right\}_{\ell\in[-N,N]^d}.$$

Let M be an $r\times r$ matrix with $M(j)$ the jth column of M. Define the $r^2\times 1$ vector $\mathrm{vec}(M)$ by

$$\mathrm{vec}(M) := (M(1)^T, \dots, M(r)^T)^T.$$

For $H = \sum_{j\in[-N,N]^d} H_j e^{-i\omega j} \in \mathbf{H}_N$, let $\mathrm{vec}(H)$ be the $(r^2(2N+1)^d)\times 1$ vectors defined by

$$\mathrm{vec}(H) := \left((\mathrm{vec}(H_j))^T|_{j=(-N,\dots,-N)}, \dots, (\mathrm{vec}(H_j))^T|_{j=(N,\dots,N)} \right)^T. \tag{2.2}$$

For the matrices of the form $\mathbf{P}_\ell H_j \mathbf{P}_\kappa^T$, we have (see [19])

$$\mathrm{vec}(\mathbf{P}_\ell H_j \mathbf{P}_\kappa^T) = (\mathbf{P}_\kappa \otimes \mathbf{P}_\ell)\mathrm{vec}(H_j), \tag{2.3}$$

where $\mathbf{P}_\kappa \otimes \mathbf{P}_\ell$ denotes the Kronecker product of matrices $\mathbf{P}_\kappa$ and $\mathbf{P}_\ell$:

$$\mathbf{P}_\kappa \otimes \mathbf{P}_\ell = (p_\kappa(\tau,i)\mathbf{P}_\ell)_{1\le\tau,i\le r}, \quad \mathbf{P}_\kappa = (p_\kappa(\tau,i))_{1\le\tau,i\le r}.$$

For $j\in\mathbf{Z}^d$, define the $r^2\times r^2$ matrices

$$\mathcal{A}_j := \sum_{\ell\in[0,N]^d} \mathbf{P}_{\ell-j} \otimes \mathbf{P}_\ell$$

and define the $(r^2(2N+1)^d) \times (r^2(2N+1)^d)$ matrix

$$\mathcal{T}_{\mathbf{P}} := \left(2^{-d}\mathcal{A}_{2i-j}\right)_{i,j\in[-N,N]^d}. \tag{2.4}$$

Then from (2.1) and (2.3) and for any $\kappa \in [-N,N]^d$,

$$\begin{aligned}\mathrm{vec}((\mathbf{T}_{\mathbf{P}}H)_\kappa) &= 2^{-d}\sum_{j\in[-N,N]^d}\sum_{\ell\in[0,N]^d}\mathrm{vec}(\mathbf{P}_\ell H_j \mathbf{P}^T_{\ell-(2\kappa-j)})\\ &= 2^{-d}\sum_{j\in[-N,N]^d}\sum_{\ell\in[0,N]^d}(\mathbf{P}_{\ell-(2\kappa-j)}\otimes\mathbf{P}_\ell)\mathrm{vec}(H_j)\\ &= \sum_{j\in[-N,N]^d} 2^{-d}\mathcal{A}_{2\kappa-j}\mathrm{vec}(H_j) = (\mathcal{T}_{\mathbf{P}}\mathrm{vec}(H))(\kappa).\end{aligned}$$

Hence we have the following theorem.

THEOREM 2.1. *The transition operator $\mathbf{T}_{\mathbf{P}}$ is equivalent to the matrix $\mathcal{T}_{\mathbf{P}}$ defined by* (2.4) *under the basis $\{e^{-i\omega\ell}\}_{\ell\in[-N,N]^d}$ of $\mathbf{H}_N$, and for any $H \in \mathbf{H}_N$,*

$$\mathrm{vec}(\mathbf{T}_{\mathbf{P}}H) = \mathcal{T}_{\mathbf{P}}\mathrm{vec}(H), \tag{2.5}$$

where $\mathrm{vec}(H)$ *is the vector defined by* (2.2).

In the rest of this section, we will find some left eigenvectors of $\mathcal{T}_{\mathbf{P}}$. These eigenvectors are associated with the vanishing moment conditions of the matrix refinement mask $\mathbf{P}$. We say that mask $\mathbf{P}(\omega)$ satisfies the *vanishing moment conditions* of order $m \in \mathbf{Z}_+$ if there exist $1\times r$ real vectors $\mathbf{l}_0^\beta$ with $\mathbf{l}_0^0 \neq 0$, $\beta \in \mathbf{Z}_+^d, |\beta| \le m-1$ such that

$$\sum_{0\le\alpha\le\beta}\binom{\beta}{\alpha}(2i)^{|\alpha-\beta|}\mathbf{l}_0^\alpha(D^{\beta-\alpha}\mathbf{P})(\nu\pi) = \delta_\nu 2^{-|\beta|}\mathbf{l}_0^\beta, \quad \nu \in \mathbf{Z}^d/2\mathbf{Z}^d. \tag{2.6}$$

Assume that $\Phi = (\phi_l)_{l=1}^r \in L^2(\mathbf{R}^d)$ is a compactly supported matrix refinable function with corresponding mask $\mathbf{P}$. Under the assumption that $\phi_l(x-j)$, $1\le l\le r$, $j\in\mathbf{Z}^d$, are linearly independent, that there exist vectors $\mathbf{l}_0^\beta, |\beta| \le m-1$ satisfying (2.6) is equivalent to that ϕ_l, $1\le l\le r$, provide approximation of order m, see [15], [27] for $d=1$. For $d=1$, (2.6) implies a matrix factorization of $\mathbf{P}(\omega)$ under the assumption that Φ is stable (see [27], [3]). It is shown in [23] that if $\det G_\Phi(\nu\pi) \neq 0, \nu \in \mathbf{Z}^d/2\mathbf{Z}^d$, then $\mathbf{P}(0)$ satisfies Condition E and $\mathbf{P}$ satisfies the vanishing moment conditions of order at least 1, where

$$G_\Phi(\omega) := \sum_{\kappa\in\mathbf{Z}^d}\widehat{\Phi}(\omega+2\pi\kappa)\widehat{\Phi}^*(\omega+2\pi\kappa).$$

Thus in what follows we will assume that $\mathbf{P}(0)$ satisfies Condition E and $m \ge 1$ in (2.6). In this case, if Φ is a compactly supported nontrivial refinable vector, then $\widehat{\Phi}(0) = c\mathbf{r}$ for some nonzero constant c, where $\mathbf{r}$ is the normalized right 1-eigenvector of $\mathbf{P}(0)$.

Let $m_0 \in \mathbf{Z}_+, m_0 \le m$ be the largest integer such that there exist row vectors $\mathbf{l}_0^\beta \in \mathbf{R}^r$, $\beta \in \mathbf{Z}_+^d, m \le |\beta| \le m+m_0-1$ satisfying

$$\sum_{0\le\alpha\le\beta}\binom{\beta}{\alpha}(2i)^{|\alpha-\beta|}\mathbf{l}_0^\alpha(D^{\beta-\alpha}\mathbf{P})(0) = 2^{-|\beta|}\mathbf{l}_0^\beta. \tag{2.7}$$

Equations (2.7) can be written as

$$\mathbf{l}_0^\beta \left(2^{-|\beta|}\mathbf{I}_r - \mathbf{P}(0)\right) = \sum_{0\le\alpha<\beta} \binom{\beta}{\alpha} (2i)^{|\alpha-\beta|} \mathbf{l}_0^\alpha (D^{\beta-\alpha}\mathbf{P})(0), \tag{2.8}$$

where $\mathbf{I}_r$ is the $r\times r$ identity matrix. Thus if each of the numbers of $2^{-m}, 2^{-m-1}, \dots, 2^{-m-m_0}$ is not an eigenvalue of $\mathbf{P}(0)$ for some $m_0 \in \mathbf{Z}_+$, then vectors $\mathbf{l}_0^\beta \in \mathbf{R}^r$, $\beta \in \mathbf{Z}_+^d, m \le |\beta| \le m+m_0-1$ can be chosen iteratively by (2.8). Since in the examples which are analyzed below $m_0 = m$, in the following we will assume that $m_0 = m$. For the case $r = 1$, since $\mathbf{P}(0) = 1$, such assumption is not needed.

Let $B(\omega) = \sum_{\kappa\in\mathbf{Z}_+^d, |\kappa|\le 2m-1} B_\kappa e^{i\kappa\omega}$ be the vector trigonometric polynomial satisfying

$$D^\beta B(0) = i^{|\beta|}\mathbf{l}_0^\beta, \quad \beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1. \tag{2.9}$$

The coefficients B_κ, $1\times r$ vectors, can be found by the following equations:

$$\sum_{|\kappa|\le 2m-1} \kappa^\beta B_\kappa = \mathbf{l}_0^\beta, \quad \beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1.$$

One can check that the vanishing moment conditions (2.6) and (2.7) can be written equivalently in the form

$$D^\beta (B(2\omega)\mathbf{P}(\omega))|_{\omega=0} = D^\beta B(0) \quad \forall \beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1, \tag{2.10}$$

and

$$D^\beta \left(B(2\omega)\,\mathbf{P}(\omega)\right)|_{\omega=\nu\pi} = 0 \quad \forall \beta \in \mathbf{Z}_+^d, |\beta| \le m-1, \nu \in \mathbf{Z}^d/2\mathbf{Z}^d\backslash\{0\}. \tag{2.11}$$

Let $\mathbf{l}_0^\beta$, $\beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1$ be the row vectors satisfying (2.6) and (2.7). For $\kappa \in \mathbf{Z}^d$, define row vectors $\mathbf{l}_\kappa^\beta$ by

$$\mathbf{l}_\kappa^\beta := \sum_{0\le\alpha\le\beta} \binom{\beta}{\alpha} \kappa^{\beta-\alpha}\mathbf{l}_0^\alpha \quad \text{for } \beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1, \tag{2.12}$$

and then define the $1\times(r^2(2N+1)^d)$ vectors $\mathbf{L}_N^\beta$ by

$$\mathbf{L}_N^\beta := (\mathbf{l}^\beta(\kappa)|_{\kappa=(-N,\dots,-N)}, \dots, \mathbf{l}^\beta(\kappa)|_{\kappa=(N,\dots,N)}) \tag{2.13}$$

with

$$\mathbf{l}^\beta(\kappa) := \sum_{0\le\alpha\le\beta} (-1)^\alpha \binom{\beta}{\alpha} \mathbf{l}_\kappa^\alpha \otimes \mathbf{l}_0^{\beta-\alpha}, \quad \kappa \in \mathbf{Z}^d.$$

For the case $d = 1$, $\mathbf{l}_\kappa^\beta, \kappa \in \mathbf{Z}^d$, are the coefficients for the reproduction of polynomials by the integer translates of Φ; see [15].

For two $1\times r$ vectors $\mathbf{v}, \mathbf{u}$ and the $r\times r$ matrix M, we have (see [19])

$$(\mathbf{v}\otimes\mathbf{u})\mathrm{vec}(M) = \mathbf{u}M\mathbf{v}^T. \tag{2.14}$$

LEMMA 2.1. *Assume that the refinement mask* $\mathbf{P}$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta$, $|\beta| \le 2m-1$, *and* B *is the vector trigonometric polynomial satisfying* (2.9). *Let* $\mathbf{L}_N^\beta$ *be the vectors defined by* (2.13)*; then for any* $H \in \mathbf{H}_N$

$$\mathbf{L}_N^\beta \mathrm{vec}(H) = (-i)^{|\beta|} D^\beta \left(B(\omega)H(\omega)B^*(\omega)\right)|_{\omega=0}, \quad \beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1,$$

where $\mathrm{vec}(H)$ *is the vector defined by* (2.2).

Proof. By (2.14), for any $\beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1$, and any $H \in \mathbf{H}_N$,

$$\begin{aligned}
\mathbf{L}_N^\beta \mathrm{vec}(H) &= \sum_\kappa \mathbf{L}_N^\beta(\kappa)\mathrm{vec}(H_\kappa) = \sum_\kappa \sum_{0\le\alpha\le\beta} (-1)^{|\alpha|}\binom{\beta}{\alpha}\mathbf{l}_0^{\beta-\alpha} H(\kappa)(\mathbf{l}_\kappa^\alpha)^T \\
&= \sum_\kappa \sum_{0\le\alpha\le\beta} (-1)^{|\alpha|}\binom{\beta}{\alpha}\mathbf{l}_0^{\beta-\alpha} H(\kappa) \sum_{0\le\gamma\le\alpha} \kappa^\gamma \binom{\alpha}{\gamma}(\mathbf{l}_0^{\alpha-\gamma})^T \\
&= \sum_\kappa \sum_{0\le\alpha\le\beta} \sum_{0\le\gamma\le\alpha} (-1)^{|\alpha|}\binom{\beta}{\alpha}\kappa^\gamma\binom{\alpha}{\gamma}(-i)^{|\beta-\alpha|} D^{\beta-\alpha}B(0)H(\kappa) i^{|\alpha-\gamma|} D^{\alpha-\gamma}B^*(0) \\
&= (-i)^{|\beta|} \sum_{0\le\alpha\le\beta} \sum_{0\le\gamma\le\alpha} \binom{\beta}{\alpha}\binom{\alpha}{\gamma} D^{\beta-\alpha}B(0) \sum_\kappa (-i\kappa)^\gamma H(\kappa) D^{\alpha-\gamma}B^*(0) \\
&= (-i)^{|\beta|} \sum_{0\le\alpha\le\beta} \sum_{0\le\gamma\le\alpha} \binom{\beta}{\alpha}\binom{\alpha}{\gamma} D^{\beta-\alpha}B(0) D^\gamma H(0) D^{\alpha-\gamma}B^*(0) \\
&= (-i)^{|\beta|} D^\beta \left(B(\omega)H(\omega)B^*(\omega)\right)|_{\omega=0}. \qquad \square
\end{aligned}$$

THEOREM 2.2. *Assume that the refinement mask* $\mathbf{P}$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta$, $|\beta| \le 2m-1$, *and* B *is the vector trigonometric polynomial satisfying* (2.9). *Let* $\mathbf{L}_N^\beta$ *be the vectors defined by* (2.13); *then*

$$\mathbf{L}_N^\beta \mathcal{T}_\mathbf{P} = 2^{-|\beta|}\mathbf{L}_N^\beta, \quad \beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1.$$

Proof. We need only to show that for any $H \in \mathbf{H}_N$, $\mathbf{L}_N^\beta \mathcal{T}_\mathbf{P}\mathrm{vec}(H) = 2^{-|\beta|}\mathbf{L}_N^\beta \mathrm{vec}(H)$. In fact, by (2.5) and Lemma 2.1,

$$\begin{aligned}
(2i)^{|\beta|}\mathbf{L}_N^\beta \mathcal{T}_\mathbf{P}\mathrm{vec}(H) &= (2i)^{|\beta|}\mathbf{L}_N^\beta \mathrm{vec}(\mathbf{T}_\mathbf{P} H) = D^\beta\left(B(2\omega)\mathbf{T}_\mathbf{P} H(2\omega)B^*(2\omega)\right)|_{\omega=0} \\
&= \sum_{\nu\in\mathbf{Z}^d/2\mathbf{Z}^d} D^\beta(B(2\omega)\mathbf{P}(\omega+\nu\pi)H(\omega+\nu\pi)\mathbf{P}(\omega+\nu\pi)^* B^*(2\omega))|_{\omega=0} \\
&= \sum_{\nu\in\mathbf{Z}^d/2\mathbf{Z}^d} \sum_{0\le\alpha\le\beta} \sum_{0\le\gamma\le\alpha} \binom{\beta}{\alpha}\binom{\alpha}{\gamma} D^\alpha\left(B(2\omega)\mathbf{P}(\omega)\right)|_{\omega=\nu\pi} \\
&\qquad \cdot D^\gamma H(\omega)|_{\omega=\nu\pi} D^{\beta-\alpha-\gamma}\left(B(2\omega)\mathbf{P}(\omega)\right)^*|_{\omega=\nu\pi}.
\end{aligned}$$

Since for any $\beta,\alpha,\gamma \in \mathbf{Z}_+^d$ with $|\beta| \le 2m-1$ and $\gamma \le \alpha \le \beta$, $\min(|\alpha|, |\beta-\alpha-\gamma|) \le m-1$, then by (2.10) and (2.11)

$$\begin{aligned}
&(2i)^{|\beta|}\mathbf{L}_N^\beta \mathcal{T}_\mathbf{P}\mathrm{vec}(H) \\
&= \sum_{0\le\alpha\le\beta} \sum_{0\le\gamma\le\alpha} \binom{\beta}{\alpha}\binom{\alpha}{\gamma} D^\alpha(B(2\omega)\mathbf{P}(\omega))|_{\omega=0} D^\gamma H(\omega)|_{\omega=0} D^{\beta-\alpha-\gamma}\left(B(2\omega)\mathbf{P}(\omega)\right)^*|_{\omega=0} \\
&= \sum_{0\le\alpha\le\beta} \sum_{0\le\gamma\le\alpha} \binom{\beta}{\alpha}\binom{\alpha}{\gamma} D^\alpha B(0) D^\gamma H(0) D^{\beta-\alpha-\gamma}B^*(0) \\
&= D^\beta\left(B(\omega)H(\omega)B^*(\omega)\right)|_{\omega=0} = i^{|\beta|}\mathbf{L}_N^\beta \mathrm{vec}(H).
\end{aligned}$$

Therefore $\mathbf{L}_N^\beta \mathcal{T}_\mathbf{P}\mathrm{vec}(H) = 2^{-|\beta|}\mathbf{L}_N^\beta \mathrm{vec}(H)$; the proof of Theorem 2.2 is completed. $\square$

Since $\mathbf{L}_N^0 = (\mathbf{l}_0^0 \otimes \mathbf{l}_0^0, \dots, \mathbf{l}_0^0 \otimes \mathbf{l}_0^0) \neq 0$, 1 is an eigenvalue of $\mathbf{T_P}$. In the case $r = 1, d = 1$, for any $n \in \mathbf{Z}_+$, $n \leq 2m-1$, the vector v_n defined by (1.7) is a generalized left eigenvector of eigenvalue 2^{-n} of $\mathcal{T}_\mathbf{P}$ (see p. 228 in [6]), and hence $2^{-n}, 0 \leq n \leq 2m-1$ are eigenvalues of $\mathbf{T_P}$. Theorem 2.2 says that for $n \in \mathbf{Z}_+, n \leq 2m-1$, if there exists $\beta \in \mathbf{Z}_+^d, |\beta| = n$, and $\mathbf{L}_N^\beta \neq 0$, then 2^{-n} is an eigenvalue of $\mathcal{T}_\mathbf{P}$ (also $\mathbf{T_P}$) with $\mathbf{L}_N^\beta$ being a corresponding left eigenvector. As the vectors v_n do for the case $r = 1, d = 1$, vectors $\mathbf{L}_N^\beta$ also play an important role in the estimate of the Sobolev regularity of refinable vector Φ, which will be shown in the next section.

3. Sobolev regularity estimates. In this section we will consider the Sobolev regularity of the matrix refinable function Φ of (1.1). For $s \geq 0$, we say $f \in W^s(\mathbf{R}^d)$ if $(1 + |\omega|^2)^{\frac{s}{2}} \widehat{f}(\omega) \in L^2(\mathbf{R}^d)$. In the first part of this section, we will provide a characterization of Φ in $W^n(\mathbf{R}^d)$ for $n \in \mathbf{Z}_+$. We need a lemma.

LEMMA 3.1. *Assume that* $\mathbf{P}(\omega)$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta$, $|\beta| \leq 2m-1$, *and* B *is the vector trigonometric polynomial satisfying* (2.9)*; then for any compactly supported solution* Φ *of* (1.1),

$$D^\beta \Big(B(\omega)\widehat{\Phi}(\omega) \Big) |_{\omega = 2\pi\ell} = 0 \quad \textit{for any } \ell \in \mathbf{Z}^d \backslash \{0\}, \beta \in \mathbf{Z}_+^d, |\beta| \leq m-1.$$

Proof. Since Φ is compactly supported, $\widehat{\Phi}(\omega)$ is analytic. For any $\ell \in \mathbf{Z}^d \backslash \{0\}$, write ℓ in the form of $\ell = 2^n\nu + 2^{n+1}\kappa$ for some $n \in \mathbf{Z}_+, \nu \in \mathbf{Z}^d/2\mathbf{Z}^d \backslash \{0\}, \kappa \in \mathbf{Z}^d$. Then

$$\begin{aligned}\widehat{\Phi}(2\pi\ell + \omega) &= \mathbf{P}\Big(\frac{2\pi\ell+\omega}{2}\Big) \cdots \mathbf{P}\Big(\frac{2\pi\ell+\omega}{2^n}\Big)\mathbf{P}\Big(\frac{2\pi\ell+\omega}{2^{n+1}}\Big)\widehat{\Phi}\Big(\frac{2\pi\ell+\omega}{2^{n+1}}\Big) \\ &= \mathbf{P}\Big(\frac{\omega}{2}\Big) \cdots \mathbf{P}\Big(\frac{\omega}{2^n}\Big)\mathbf{P}\Big(\frac{\omega}{2^{n+1}} + \nu\pi\Big)\widehat{\Phi}\Big(\frac{2\pi\ell+\omega}{2^{n+1}}\Big).\end{aligned}$$

Thus by (2.6) and (2.7), or by its equivalent forms (2.10) and (2.11),

$$\begin{aligned}&D^\beta \Big(B(\omega)\widehat{\Phi}(\omega) \Big) |_{\omega=2\pi\ell} = D^\beta \Big(B(\omega)\widehat{\Phi}(2\pi\ell+\omega) \Big) |_{\omega=0} \\ &= \sum_{0 \leq \alpha \leq \beta} \binom{\beta}{\alpha} D^\alpha \Big(B(\omega)\mathbf{P}\Big(\frac{\omega}{2}\Big) \Big) |_{\omega=0} \\ &\quad \cdot D^{\beta-\alpha} \Big(\mathbf{P}\Big(\frac{\omega}{2^2}\Big) \cdots \mathbf{P}\Big(\frac{\omega}{2^n}\Big)\mathbf{P}\Big(\frac{\omega}{2^{n+1}} + \nu\pi\Big)\widehat{\Phi}\Big(\frac{2\pi\ell+\omega}{2^{n+1}}\Big) \Big) |_{\omega=0} \\ &= \sum_{0 \leq \alpha \leq \beta} \binom{\beta}{\alpha} D^\alpha B\Big(\frac{\omega}{2}\Big) |_{\omega=0} \\ &\quad \cdot D^{\beta-\alpha} \Big(\mathbf{P}\Big(\frac{\omega}{2^2}\Big) \cdots \mathbf{P}\Big(\frac{\omega}{2^n}\Big)\mathbf{P}\Big(\frac{\omega}{2^{n+1}} + \nu\pi\Big)\widehat{\Phi}\Big(\frac{2\pi\ell+\omega}{2^{n+1}}\Big) \Big) |_{\omega=0} \\ &= D^\beta \Big(B\Big(\frac{\omega}{2}\Big)\mathbf{P}\Big(\frac{\omega}{2^2}\Big) \cdots \mathbf{P}\Big(\frac{\omega}{2^n}\Big)\mathbf{P}\Big(\frac{\omega}{2^{n+1}} + \nu\pi\Big)\widehat{\Phi}\Big(\frac{2\pi\ell+\omega}{2^{n+1}}\Big) \Big) |_{\omega=0} = \cdots \\ &= D^\beta \Big(B\Big(\frac{\omega}{2^n}\Big)\mathbf{P}\Big(\frac{\omega}{2^{n+1}} + \nu\pi\Big)\widehat{\Phi}\Big(\frac{2\pi\ell+\omega}{2^{n+1}}\Big) \Big) |_{\omega=0} = 0\end{aligned}$$

since $D^\alpha \big(B(\frac{\omega}{2^n})\mathbf{P}(\frac{\omega}{2^{n+1}} + \nu\pi) \big) |_{\omega=0} = 0$ for any $\alpha \leq \beta$. ∎

If a refinable vector Φ is contained in $W^{\frac{n}{2}}(\mathbf{R}^d)$ for some $n \in \mathbf{Z}_+$, then

$$\int_{\mathbf{R}^d} |\omega|^n |\widehat{\Phi}(\omega)|^2 d\omega < \infty. \tag{3.1}$$

For any $\beta_0 \in \mathbf{Z}_+^d$, $|\beta_0| = n$, define

$$H_{\beta_0}(\kappa) := \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} (i\omega)^{\beta_0} \widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega) e^{i\kappa\omega} d\omega, \quad \kappa \in \mathbf{Z}_+^d.$$

Let F be the matrix function defined by

$$\widehat{F}(\omega) := (i\omega)^{\beta_0} \widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega).$$

The finiteness of the integral in (3.1) implies that every entry of F is continuous and hence $H_{\beta_0}(\kappa) = F(\kappa)$. (3.1) also implies the existence of $D^{\beta_0}(\text{auto}(\Phi))(= F)$, where

$$\text{auto}(\Phi)(y) := \int_{\mathbf{R}^d} \Phi(x)\Phi^*(x-y)dx.$$

Since Φ is compactly supported on $[0, N]^d$, the support of F is contained in $[-N, N]^d$. Therefore $H_{\beta_0}(\kappa) = 0$ for $\kappa \notin [-N, N]^d$. Define

$$G^{(\beta_0)}(\omega) := \sum_{\kappa} H_{\beta_0}(\kappa) e^{-i\kappa\omega};$$

then $G^{(\beta_0)}(\omega) \in \mathbf{H}_N$ for any $|\beta_0| = n$.

PROPOSITION 3.1. *Assume that the refinement mask* $\mathbf{P}$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta$, $|\beta| \le 2m-1$. *Suppose there exists a refinable vector* Φ *contained in* $W^{\frac{n}{2}}(\mathbf{R}^d)$ *for some* $n \in \mathbf{Z}_+$ *with* $n \le 2m-1$; *then for any* $\beta_0 \in \mathbf{Z}_+^d$, $|\beta_0| = n$,

$$\mathcal{T}_{\mathbf{P}} \text{vec}(G^{(\beta_0)}) = 2^{-n} \text{vec}(G^{(\beta_0)})$$

and for any $\beta \in \mathbf{Z}_+^d, \beta \le \beta_0$,

$$\mathbf{L}_N^\beta \text{vec}(G^{(\beta_0)}) = \beta_0! \delta_{\beta_0}(\beta) |\mathbf{l}_0^0 \widehat{\Phi}(0)|^2 = \begin{cases} 0, & \beta < \beta_0, \\ \beta_0! |\mathbf{l}_0^0 \widehat{\Phi}(0)|^2, & \beta = \beta_0. \end{cases}$$

Proof. By the Poisson summation formula,

$$G^{(\beta_0)}(\omega) = \sum_{\ell \in \mathbf{Z}^d} (i\omega + i2\pi\ell)^{\beta_0} \widehat{\Phi}(\omega + 2\pi\ell)\widehat{\Phi}^*(\omega + 2\pi\ell).$$

By the definition of $\mathbf{T}_{\mathbf{P}}$,

$$\begin{aligned} \mathbf{T}_{\mathbf{P}} G^{(\beta_0)}(\omega) &= \sum_{\nu \in \mathbf{Z}^d/2\mathbf{Z}^d} \sum_{\ell \in \mathbf{Z}^d} (i\omega/2 + 2\ell\pi i + \nu\pi i)^{\beta_0} \mathbf{P}(\omega/2 + \nu\pi) \\ &\quad \cdot \widehat{\Phi}(\omega/2 + 2\ell\pi + \nu\pi)\widehat{\Phi}^*(\omega/2 + 2\ell\pi + \nu\pi)\mathbf{P}^*(\omega/2 + \nu\pi) \\ &= \frac{1}{2^n} \sum_{\nu \in \mathbf{Z}^d/2\mathbf{Z}^d} \sum_{\ell \in \mathbf{Z}^d} (i\omega + 4\ell\pi i + 2\nu\pi i)^{\beta_0} \widehat{\Phi}(\omega + 4\ell\pi + 2\nu\pi)\widehat{\Phi}^*(\omega + 4\ell\pi + 2\nu\pi) \\ &= \frac{1}{2^n} G^{(\beta_0)}(\omega), \end{aligned}$$

and hence $\mathcal{T}_{\mathbf{P}} \text{vec}(G^{(\beta_0)}) = 2^{-n} \text{vec}(G^{(\beta_0)})$ by (2.5).

By Lemma 3.1, for any $\alpha \in \mathbf{Z}_+^d, |\alpha| < 2m-1$, and $\ell \in \mathbf{Z}^d \backslash \{0\}$,

$$D^\alpha \left(B(\omega)\widehat{\Phi}(\omega + 2\ell\pi)\widehat{\Phi}^*(\omega + 2\ell\pi)B^*(\omega)\right)|_{\omega=0} = 0.$$

Therefore, by Lemma 2.1,

$$\begin{aligned}
&\mathbf{L}_N^\beta \mathrm{vec}(G^{(\beta_0)}) = (-i)^{|\beta|} D^\beta (B(\omega)G^{(\beta_0)}(\omega)B^*(\omega))|_{\omega=0} \\
&= (-i)^{|\beta|} D^\beta ((i\omega)^{\beta_0} B(\omega)\widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega)B^*(\omega))|_{\omega=0} \\
&\quad + (-i)^{|\beta|} D^\beta \left(\sum_{\ell \in \mathbf{Z}^d \backslash \{0\}} (i\omega + i2\ell\pi)^{\beta_0} B(\omega)\widehat{\Phi}(\omega + 2\ell\pi)\widehat{\Phi}^*(\omega + 2\ell\pi)B^*(\omega) \right)|_{\omega=0} \\
&= \beta_0! \delta_{\beta_0}(\beta) |\mathbf{l}_0^0 \widehat{\Phi}(0)|^2. \qquad \square
\end{aligned}$$

We note that if λ is a simple eigenvalue of a matrix, then the product of the corresponding left row eigenvector and right column eigenvector is not zero (see Lemma 6.3.10 in [18]). Thus $\mathbf{l}_0^0 \widehat{\Phi}(0) \neq 0$ since $\mathbf{P}(0)$ satisfies Condition E and $\widehat{\Phi}(0)$ is a right 1-eigenvector of $\mathbf{P}(0)$. By the fact that $\Phi \in W^{s_1}(\mathbf{R}^d)$ if $\Phi \in W^s(\mathbf{R}^d)$ and $s_1 < s$, Proposition 3.1 leads to the following corollary.

COROLLARY 3.1. *Assume that the refinement mask* $\mathbf{P}$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta$, $|\beta| \leq 2m-1$. *Suppose there exists a nontrivial refinable vector* Φ *contained in* $W^{\frac{n}{2}}(\mathbf{R}^d)$ *for some* $n \in \mathbf{Z}_+$ *with* $n \leq 2m-1$; *then for any* $\beta \in \mathbf{Z}_+^d, |\beta| \leq n$, $\mathbf{L}_N^\beta \neq 0$, *and* $1, 2^{-1}, \ldots, 2^{-n}$ *are eigenvalues of* $\mathbf{T}_\mathbf{P}$.

The next theorem will give a characterization of the refinable vector Φ in the Sobolev space $W^n(\mathbf{R}^d)$, $n \in \mathbf{Z}_+$. But first, we need another lemma. For $j \in \mathbf{Z}_+$, denote

$$\Pi_j(\omega) := \chi_{2^j \mathbf{T}^d}(\omega) \Pi_{i=1}^j \mathbf{P}(2^{-i}\omega).$$

LEMMA 3.2. *For any* $H_1(\omega), H_2(\omega) \in \mathbf{H}_N$,

$$\int_{\mathbf{T}^d} H_1(\omega)(\mathbf{T}_\mathbf{P}^j H_2)(\omega)d\omega = \int_{\mathbf{R}^d} H_1(\omega)\Pi_j(\omega)H_2(2^{-j}\omega)\Pi_j(\omega)^* d\omega. \tag{3.2}$$

Proof. The proof of (3.2) can be found in [26]. In fact for $j = 1$,

$$\begin{aligned}
&\int_{\mathbf{R}^d} H_1(\omega)\Pi_1(\omega)H_2\left(\frac{\omega}{2}\right)\Pi_1(\omega)^* d\omega \\
&= \sum_{\beta \in \mathbf{Z}^d} \int_{\mathbf{T}^d} H_1(\omega)\mathbf{P}\left(\frac{\omega}{2} + \beta\pi\right)H_2\left(\frac{\omega}{2} + \beta\pi\right)\mathbf{P}^*\left(\frac{\omega}{2} + \beta\pi\right)\chi_{\mathbf{T}^d}\left(\frac{\omega}{2} + \beta\pi\right)d\omega \\
&= \sum_{\alpha \in \mathbf{Z}^d} \sum_{\nu \in \mathbf{Z}^d/2\mathbf{Z}^d} \int_{\mathbf{T}^d} H_1(\omega)\mathbf{P}\left(\frac{\omega}{2} + \nu\pi\right)H_2\left(\frac{\omega}{2} + \nu\pi\right)\mathbf{P}^*\left(\frac{\omega}{2} + \nu\pi\right)\chi_{\mathbf{T}^d}\left(\frac{\omega}{2} + 2\alpha\pi + \nu\pi\right)d\omega \\
&= \sum_{\nu \in \mathbf{Z}^d/2\mathbf{Z}^d} \int_{\mathbf{T}^d} H_1(\omega)\mathbf{P}\left(\frac{\omega}{2} + \nu\pi\right)H_2\left(\frac{\omega}{2} + \nu\pi\right)\mathbf{P}^*\left(\frac{\omega}{2} + \nu\pi\right)d\omega \\
&= \int_{\mathbf{T}^d} H_1(\omega)\mathbf{T}_\mathbf{P}H_2(\omega)d\omega.
\end{aligned}$$

For general j, this formula can be found by induction. $\square$

THEOREM 3.1. *Assume that the refinement mask* $\mathbf{P}$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta$, $|\beta| \le 2m-1$; *then a refinable vector* Φ *is contained in* $W^n(\mathbf{R}^d)$ *for some* $n \in \mathbf{Z}_+$ *with* $n \le m-1$ *if and only if there exists a positive semidefinite* $H \in \mathbf{H}_N$ *satisfying the following conditions:*

(i) $\mathbf{T_P} H = 4^{-n} H$;

(ii) *there exist constants* $c_0, \delta > 0$ *such that*

$$H(\omega) \ge c_0 |\omega|^{2n} \mathbf{r}\mathbf{r}^T \quad \text{for } \omega \in [-\delta, \delta]^d,$$

where $\mathbf{r}$ *is the normalized right* 1*-eigenvector of* $\mathbf{P}(0)$.

Proof. "$\Longrightarrow$" If $\Phi \in W^n(\mathbf{R}^d)$, let

$$H(\omega) = (-1)^n \sum_{|\beta_0|=n} G^{(2\beta_0)}(\omega) \ge 0. \tag{3.3}$$

Then Proposition 3.1 leads to $\mathbf{T_P} H = 4^{-n} H$.

Since $\mathbf{P}(0)$ satisfies Condition E, $\widehat{\Phi}(\omega) \to c\mathbf{r}$ with $c \ne 0$ as $\omega \to 0$ (see [14], [23], and [26]). Thus there exists a constant $\delta > 0$ such that

$$\widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega) \ge \frac{c^2}{2} \mathbf{r}\mathbf{r}^T \quad \text{for } \omega \in [-\delta, \delta]^d.$$

Therefore

$$\begin{aligned} H(\omega) &= (-1)^n \sum_{|\beta_0|=n} \sum_{\ell \in \mathbf{Z}^d} (i\omega + i2\ell\pi)^{2\beta_0} \widehat{\Phi}(\omega + 2\ell\pi)\widehat{\Phi}^*(\omega + 2\ell\pi) \\ &\ge \sum_{|\beta_0|=n} \omega^{2\beta_0} \widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega) \ge \frac{c^2}{2} \mathbf{r}\mathbf{r}^T \sum_{|\beta_0|=n} \omega^{2\beta_0} = \frac{c^2 |\omega|^{2n}}{2} \mathbf{r}\mathbf{r}^T. \end{aligned}$$

"$\Longleftarrow$" Denote $g_j(\omega) := 4^{nj} \Pi_j(\omega) H(2^{-j}\omega) \Pi_j(\omega)^*$. Then

$$\begin{aligned} g_j(\omega) &\ge c_0 4^{nj} \chi_{[-\delta,\delta]^d}\left(\frac{\omega}{2^j}\right) \Pi_j(\omega) \left(\frac{|\omega|}{2^j}\right)^{2n} \mathbf{r}\mathbf{r}^T \Pi_j(\omega)^* \\ &= c_0 |\omega|^{2n} \chi_{[-\delta,\delta]^d}\left(\frac{\omega}{2^j}\right) \Pi_j(\omega)\mathbf{r}(\Pi_j(\omega)\mathbf{r})^*. \end{aligned}$$

Thus by the Fatou lemma and the fact that $\widehat{\Phi}(\omega) = \lim_{j\to\infty} \chi_{[-\delta,\delta]^d}(\frac{\omega}{2^j}) \Pi_j(\omega)\mathbf{r}$,

$$\begin{aligned} \int_{\mathbf{R}^d} |\omega|^{2n} |\widehat{\Phi}(\omega)|^2 d\omega &= c \int_{\mathbf{R}^d} \sum_{i=1}^r \mathbf{e}_i^T \liminf_{j\to\infty} |\omega|^{2n} \chi_{[-\delta,\delta]^d}\left(\frac{\omega}{2^j}\right) \Pi_j(\omega)\mathbf{r}(\Pi_j(\omega)\mathbf{r})^* \mathbf{e}_i d\omega \\ &\le c \sum_{i=1}^r \mathbf{e}_i^T \liminf_{j\to\infty} \int_{\mathbf{R}^d} |\omega|^{2n} \chi_{[-\delta,\delta]^d}\left(\frac{\omega}{2^j}\right) \Pi_j(\omega)\mathbf{r}(\Pi_j(\omega)\mathbf{r})^* d\omega \mathbf{e}_i \\ &\le c \sum_{i=1}^r \mathbf{e}_i^T \liminf_{j\to\infty} \int_{\mathbf{R}^d} g_j(\omega) d\omega \mathbf{e}_i < \infty, \end{aligned}$$

where the last inequality follows from

$$\int_{\mathbf{R}^d} g_j(\omega) d\omega = 4^{jn} \int_{\mathbf{T}^d} \mathbf{T}_\mathbf{P}^j H(\omega) d\omega = \int_{\mathbf{T}^d} H(\omega) d\omega < \infty.$$

By the continuity of $\widehat{\Phi}$, this leads to $\Phi \in W^n(\mathbf{R}^d)$. □

For $n = 0$, $W^0(\mathbf{R}^d) = L^2(\mathbf{R}^d)$. In fact the characterization of $\Phi \in L^2(\mathbf{R}^d)$ can be given in a more easy checking way. In [23], it was shown that under the assumption that $\mathbf{P}(0)$ satisfies Condition E, $\Phi \in L^2(\mathbf{R}^d)$ if and only if there exists a positive semidefinite $H \in \mathbf{H}_N$ satisfying $\mathbf{T}_{\mathbf{P}}H = H$ and $\mathbf{l}_0^0 H(0)(\mathbf{l}_0^0)^T > 0$.

If $\Phi \in W^n(\mathbf{R}^d)$, $n \le m-1$, where $H \in \mathbf{H}_N$ is defined by (3.3), then Proposition 3.1 implies that there exists a positive semidefinite H satisfying $\mathbf{T}_{\mathbf{P}}H = 4^{-n}H$ and

$$\mathbf{L}_N^{\beta}\mathrm{vec}(H) = c\beta! \sum_{|\beta_0|=n} \delta_{2\beta_0}(\beta) \tag{3.4}$$

for any $\beta \in \mathbf{Z}_+^d, |\beta| \le 2n$, where c is a nonzero constant independent of β. In the case $r = 1$, the existence of such positive semidefinite H is also sufficient for $\Phi \in W^n(\mathbf{R}^d)$. In fact by Lemma 2.1, (3.4) is equivalent to that for any $\beta \in \mathbf{Z}_+^d, |\beta| \le 2n$,

$$D^{\beta}\left(|B(\omega)|^2 H(\omega)\right)|_{\omega=0} = c \sum_{|\beta_0|=n} \delta_{2\beta_0}(\beta), \tag{3.5}$$

which implies that $D^{\beta}H(0) = c(\mathbf{l}_0^0)^{-2}\sum_{|\beta_0|=n}\delta_{2\beta_0}(\beta)$ (in this case $\mathbf{l}_0^0$ is a nonzero real number). Thus $H(\omega) = c|\omega|^{2n} + \mathrm{o}(|\omega|^{2n})$ (as $\omega \to 0$) and hence $H(\omega)$ satisfies condition (ii) of Theorem 3.1. For $r = 1, d = 1$, such results were given in [32].

Theorem 3.1 gives the characterization of refinable vectors $\Phi \in W^s(\mathbf{R}^d)$ with s being nonnegative integers. In the following, we will give an estimate of the Sobolev regularity of Φ in terms of the spectral radius of $\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}_N^0}$, the restricted operator of $\mathbf{T}_{\mathbf{P}}$ to an invariant subspace $\mathbf{H}_N^0$ of $\mathbf{H}_N$.

For $j \in \mathbf{Z}_+, 1 \le j \le r$ and $\alpha \in \mathbf{Z}_+^d, |\alpha| \le m-1$, let ${}_j\mathbf{l}_N^{\alpha}$, ${}_j\mathbf{r}_N^{\alpha}$ be the $1 \times (r^2(2N+1)^d)$ vectors defined by

$$\begin{aligned} {}_j\mathbf{l}_N^{\alpha} &:= ({}_j\mathbf{l}^{\alpha}(\kappa)|_{\kappa=(-N,\dots,-N)}, \dots, {}_j\mathbf{l}^{\alpha}(\kappa)|_{\kappa=(N,\dots,N)}), \\ {}_j\mathbf{r}_N^{\alpha} &:= ({}_j\mathbf{r}^{\alpha}(\kappa)|_{\kappa=(-N,\dots,-N)}, \dots, {}_j\mathbf{r}^{\alpha}(\kappa)|_{\kappa=(N,\dots,N)}), \end{aligned} \tag{3.6}$$

with

$${}_j\mathbf{l}^{\alpha}(\kappa) := \mathbf{e}_j^T \otimes \mathbf{l}_{-\kappa}^{\alpha}, \quad {}_j\mathbf{r}^{\alpha}(\kappa) := \mathbf{l}_{\kappa}^{\alpha} \otimes \mathbf{e}_j^T, \quad \kappa \in \mathbf{Z}^d,$$

where $\mathbf{l}_{\kappa}^{\alpha}$ are the vectors defined by (2.12).

LEMMA 3.3. *Assume that the refinement mask* $\mathbf{P}$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^{\beta}$, $|\beta| \le 2m-1$, *and* B *is the vector trigonometric polynomial satisfying* (2.9). *For* $1 \le j \le r$ *and* $\alpha \in \mathbf{Z}_+^d, |\alpha| \le m-1$, *let* ${}_j\mathbf{l}_N^{\alpha}$ *and* ${}_j\mathbf{r}_N^{\alpha}$ *be the row vectors defined by* (3.6); *then for any* $H \in \mathbf{H}_N$,

$${}_j\mathbf{l}_N^{\alpha}\mathrm{vec}(H) = i^{\alpha}D^{\alpha}\left(B(\omega)H(\omega)\mathbf{e}_j\right)|_{\omega=0}, {}_j\mathbf{r}_N^{\alpha}\mathrm{vec}(H) = (-i)^{\alpha}D_V^{\alpha}\left(\mathbf{e}_j^T H(\omega)B^*(\omega)\right)|_{\omega=0},$$

where $\mathrm{vec}(H)$ *is the vector defined by* (2.2).

Proof. For any $H \in \mathbf{H}_N$, $H(\omega) = \sum_{\kappa\in[-N,N]^d} H_{\kappa}e^{-i\kappa\omega}$,

$$\begin{aligned} D^{\alpha}\left(B(\omega)H(\omega)\mathbf{e}_j\right)|_{\omega=0} &= \sum_{0\le\gamma\le\alpha}\binom{\alpha}{\gamma}D^{\gamma}B(0)D^{\alpha-\gamma}H(0)\mathbf{e}_j \\ &= i^{\alpha}\sum_{\kappa}\sum_{0\le\gamma\le\alpha}\binom{\alpha}{\gamma}(-\kappa)^{\alpha-\gamma}\mathbf{l}_0^{\gamma}H_{\kappa}\mathbf{e}_j = i^{\alpha}\sum_{\kappa}\mathbf{l}_{-\kappa}^{\alpha}H_{\kappa}\mathbf{e}_j \\ &= i^{\alpha}\sum_{\kappa}(\mathbf{e}_j^T \otimes \mathbf{l}_{-\kappa}^{\alpha})\mathrm{vec}(H_{\kappa}) = i^{\alpha}{}_j\mathbf{l}_N^{\alpha}\mathrm{vec}(H). \end{aligned}$$

The proof of the second formula is similar and details are omitted here. ∎

Let $\mathbf{H}_N^0$ be the subspace of $\mathbf{H}_N$ defined by

$$\begin{aligned}(3.7)\quad \mathbf{H}_N^0 := \{H \in \mathbf{H}_N : \quad & \mathbf{L}_N^\beta \mathrm{vec}(H) = 0, \quad {}_j\mathbf{l}_N^\alpha \mathrm{vec}(H) = 0, \text{ and} \\ & {}_j\mathbf{r}_N^\alpha \mathrm{vec}(H) = 0 \ \forall \beta, \alpha \in \mathbf{Z}_+^d, |\beta| \le 2m-1, |\alpha| \le m-1, 1 \le j \le r\}.\end{aligned}$$

PROPOSITION 3.2. *Assume that the refinement mask* $\mathbf{P}$ *satisfies* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta$, $|\beta| \le 2m-1$. *Let* $\mathbf{H}_N^0$ *be the subspace of* $\mathbf{H}_N$ *defined by* (3.7)*; then* $\mathbf{H}_N^0$ *is invariant under* $\mathbf{T}_\mathbf{P}$.

Proof. By Theorem 2.2, for any $H \in \mathbf{H}_N^0$ and $\beta \in \mathbf{Z}_+^d, |\beta| \le 2m-1$,

$$\mathbf{L}_N^\beta \mathrm{vec}(\mathbf{T}_\mathbf{P} H) = \mathbf{L}_N^\beta \mathcal{T}_\mathbf{P} \mathrm{vec}(H) = 2^{-|\beta|} \mathbf{L}_N^\beta \mathrm{vec}(H) = 0.$$

Let B be the vector trigonometric polynomial satisfying (2.9). By Lemma 3.3, for any $\alpha \in \mathbf{Z}_+^d, |\alpha| < m$, ${}_j\mathbf{l}_N^\alpha \mathrm{vec}(H) = 0$, and ${}_j\mathbf{r}_N^\alpha \mathrm{vec}(H) = 0$ for all $j, 1 \le j \le r$, are equivalent to $D^\alpha \left(B(\omega)H(\omega)\right)|_{\omega=0} = 0$ and $D^\alpha \left(H(\omega)B^*(\omega)\right)|_{\omega=0} = 0$, respectively. One can check by (2.10) and (2.11), $D^\alpha \left(B(\omega)\mathbf{T}_\mathbf{P} H(\omega)\right)|_{\omega=0} = 0$ ($D_V^\alpha \left(\mathbf{T}_\mathbf{P} H(\omega)B^*(\omega)\right)|_{\omega=0} = 0$, respectively) for all $\alpha \in \mathbf{Z}_+^d, |\alpha| < m$ if $D^\alpha \left(B(\omega)H(\omega)\right)|_{\omega=0} = 0$ ($D_V^\alpha \left(H(\omega)B^*(\omega)\right)|_{\omega=0} = 0$, respectively) for any $\alpha \in \mathbf{Z}_+^d, |\alpha| < m$. Thus $\mathbf{H}_N^0$ is invariant under $\mathbf{T}_\mathbf{P}$. ∎

Let $\mathbf{T}_\mathbf{P}|_{\mathbf{H}_N^0}$ denote the restriction of $\mathbf{T}_\mathbf{P}$ to $\mathbf{H}_N^0$. By the fact that the product of the left and right eigenvectors of a simple eigenvalue of a matrix is not zero again, Theorem 2.2 leads to the following corollary.

COROLLARY 3.2. *If* 2^{-n}, $0 \le n \le 2m-1$, *is a simple eigenvalue of* $\mathbf{T}_\mathbf{P}$ *and there exists* $\beta \in \mathbf{Z}_+^d$ *such that* $|\beta| = n$, $\mathbf{L}_N^\beta \ne 0$, *then* 2^{-n} *is not an eigenvalue of* $\mathbf{T}_\mathbf{P}|_{\mathbf{H}_N^0}$.

For the next proposition, we need to consider the transition operators on other spaces. Let $\mathbf{P}$ ($\{\mathbf{P}_\kappa\}$) be a given matrix mask satisfying (2.6) and (2.7) for some row vectors $\mathbf{l}_0^\beta$, $|\beta| \le 2m-1$, and $\mathrm{supp}\{\mathbf{P}_\kappa\} \subset [0, N]^d$. Denote $\mathcal{N} := \max(N, 2m)$. Let $\mathbf{H}_\mathcal{N}$ denote the space of all $r \times r$ matrices with each entry a trigonometric polynomial whose Fourier coefficients are supported in $[-\mathcal{N}, \mathcal{N}]^d$, and let $\mathbf{T}_{\mathbf{P},\mathcal{N}}$ denote the transition operator on $\mathbf{H}_\mathcal{N}$ defined by

$$\mathbf{T}_{\mathbf{P},\mathcal{N}} H(\omega) := \sum_{\nu \in \mathbf{Z}^d/2\mathbf{Z}^d} \mathbf{P}\left(\frac{\omega}{2} + \pi\nu\right) H\left(\frac{\omega}{2} + \pi\nu\right) \mathbf{P}^*\left(\frac{\omega}{2} + \pi\nu\right), \quad H \in \mathbf{H}_\mathcal{N}.$$

Then $\mathbf{T}_{\mathbf{P},\mathcal{N}}$ is a linear operator on $\mathbf{H}_\mathcal{N}$ leaving $\mathbf{H}_\mathcal{N}$ and $\mathbf{H}_N$ invariant, and $\mathbf{T}_{\mathbf{P},\mathcal{N}}$ is equivalent to the matrix

$$\mathcal{T}_{\mathbf{P},\mathcal{N}} := (2^{-d} \mathcal{A}_{2i-j})_{i,j \in [-\mathcal{N},\mathcal{N}]^d},$$

where $\mathcal{A}_j = \sum_{\ell \in [0,N]^d} \mathbf{P}_{\ell-j} \otimes \mathbf{P}_\ell$.

Let $\mathbf{H}_\mathcal{N}^0$ be the subspace of $\mathbf{H}_\mathcal{N}$ defined as follows: $H \in \mathbf{H}_\mathcal{N}^0$ if and only if $\mathbf{L}_\mathcal{N}^\beta \mathrm{vec}(H) = 0$, ${}_j\mathbf{l}_\mathcal{N}^\alpha \mathrm{vec}(H) = 0$, and ${}_j\mathbf{r}_\mathcal{N}^\alpha \mathrm{vec}(H) = 0$ for all $\beta, \alpha \in \mathbf{Z}_+^d, |\beta| \le 2m-1, |\alpha| \le m-1, 1 \le j \le r$. In this case $\mathbf{L}_\mathcal{N}^\beta$, ${}_j\mathbf{l}_\mathcal{N}^\alpha$, and ${}_j\mathbf{r}_\mathcal{N}^\alpha$ are $1 \times (r^2(2\mathcal{N}+1)^d)$ vectors defined by (2.12) and (3.6), respectively, with $\mathcal{N}$ instead of N. It can be shown similarly that $\mathbf{H}_\mathcal{N}^0$ is invariant under $\mathbf{T}_{\mathbf{P},\mathcal{N}}$. Let $\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}_\mathcal{N}^0}$ denote the restriction of $\mathbf{T}_{\mathbf{P},\mathcal{N}}$ to $\mathbf{H}_\mathcal{N}^0$, and let $H_0 \in \mathbf{H}_\mathcal{N}$ defined by

$$(3.8)\qquad H_0(\omega) := \sum_{i=1}^d (1 - \cos\omega_i)^{2m} \mathbf{I}_r, \quad \omega = (\omega_1, \ldots, \omega_d)^T \in \mathbf{T}^d;$$

then $H_0(\omega) \in \mathbf{H}^0_{\mathcal{N}}$.

We note that the transition operator $\mathbf{T_P}$ defined by (1.6) is the restriction of $\mathbf{T}_{\mathbf{P},\mathcal{N}}$ to the subspace $\mathbf{H}_N$ of $\mathbf{H}_{\mathcal{N}}$, and $\mathcal{T}_{\mathbf{P}}$ defined by (2.4) is a submatrix of $\mathcal{T}_{\mathbf{P},\mathcal{N}}$. In fact, if $\mathcal{N} > N$, then $\mathcal{T}_{\mathbf{P},\mathcal{N}}$ can be written as

$$\mathcal{T}_{\mathbf{P},\mathcal{N}} = \begin{bmatrix} M_1 & \mathbf{0} & \mathbf{0} \\ * & \mathcal{T}_{\mathbf{P}} & * \\ \mathbf{0} & \mathbf{0} & M_2 \end{bmatrix},$$

where M_1 (M_2, respectively) is a strictly lower (upper, respectively) triangular matrix. Thus $\mathbf{T}_{\mathbf{P},\mathcal{N}}$ ($\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}}$, respectively) and $\mathbf{T_P}$ ($\mathbf{T_P}|_{\mathbf{H}^0_N}$, respectively) have the same nonzero eigenvalues and the eigenvectors of $\mathbf{T}_{\mathbf{P},\mathcal{N}}$ are in $\mathbf{H}_N$. Hence $\rho(\mathbf{T_P}|_{\mathbf{H}^0_N}) = \rho(\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}})$, where $\rho(\mathbf{T_P}|_{\mathbf{H}^0_N})$ and $\rho(\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}})$ denote the spectral radii of $\mathbf{T_P}|_{\mathbf{H}^0_N}$ and $\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}}$, respectively.

Choose a vector norm on space $\mathbf{H}^0_{\mathcal{N}}$ and define the operator (matrix) norm $\|\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}}\|$ with respect to this vector norm. Then

$$\lim_{n\to\infty} \|(\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}})^n\|^{1/n} = \rho(\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}}) = \rho(\mathbf{T_P}|_{\mathbf{H}^0_N}).$$

PROPOSITION 3.3. *Assume that* $\mathbf{P}$ *satisfies conditions* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}^\beta_0, |\beta| \le 2m-1$. *Let* $\mathbf{H}^0_N$ *be the subspace of* $\mathbf{H}_N$ *defined by* (3.7) *and* $\rho(\mathbf{T_P}|_{\mathbf{H}^0_N})$ *the spectral radius of* $\mathbf{T_P}|_{\mathbf{H}^0_N}$. *Then for any* $\epsilon > 0$, *for the corresponding refinable function* Φ, *there exists a constant* c *independent of* n *such that*

$$\int_{\mathbf{\Omega}_n} \left|\widehat{\Phi}(w)\right|^2 dw \le c\left(\rho(\mathbf{T_P}|_{\mathbf{H}^0_N}) + \epsilon\right)^n,$$

where $\mathbf{\Omega}_n := 2^n\mathbf{T}^d \backslash 2^{n-1}\mathbf{T}^d$, $n \in \mathbf{Z}_+$.

This proposition together with the usual Littlewood–Paley technique leads to the following Sobolev estimate of refinable vector Φ.

THEOREM 3.2. *Assume that* $\mathbf{P}$ *satisfies conditions* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}^\beta_0, |\beta| \le 2m-1$. *Let* $\mathbf{H}^0_N$ *be the subspace of* $\mathbf{H}_N$ *defined by* (3.7) *and* $\rho(\mathbf{T_P}|_{\mathbf{H}^0_N})$ *the spectral radius of* $\mathbf{T_P}|_{\mathbf{H}^0_N}$. *Then the matrix refinable function* Φ *is in* $W^s(\mathbf{R}^d)$ *for any* $s < s_0 := -\log_4 \rho(\mathbf{T_P}|_{\mathbf{H}^0_N})$.

The proofs of Proposition 3.3 and Theorem 3.2 can be carried out by modifying the proofs of Proposition 4.4 and Theorem 4.5 in [29]. For completeness, we give them here.

Proof of Proposition 3.3. Let $H_0(\omega) \in \mathbf{H}^0_{\mathcal{N}}$ defined by (3.8). Note that $H_0(\omega) \ge \mathbf{I}_r$ for $\omega \in \mathbf{T}^d \backslash (\frac{1}{2}\mathbf{T}^d)$, and $\widehat{\Phi}$ is continuous on $\mathbf{T}^d$; thus for any positive integer n,

$$\begin{aligned}\int_{\mathbf{\Omega}_n} \widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega)d\omega &= \int_{\mathbf{\Omega}_n} \Pi_n(\omega)\widehat{\Phi}(2^{-n}\omega)\widehat{\Phi}^*(2^{-n}\omega)\Pi_n^*(\omega)d\omega \\ &\le c\int_{\mathbf{\Omega}_n} \Pi_n(\omega)\Pi_n^*(\omega)d\omega \le c\int_{\mathbf{\Omega}_n} \Pi_n(\omega)H_0(2^{-n}\omega)\Pi_n^*(\omega)d\omega \\ &= c\int_{\mathbf{T}^d} (\mathbf{T}^n_{\mathbf{P},\mathcal{N}}H_0)(\omega)d\omega,\end{aligned}$$

where the last equation can be shown similar to (3.2). Since the Hilbert–Schmidt norm $\|M\|_2 = \sqrt{\mathrm{Tr}(MM^*)}$ is an equivalent norm for finite matrices, by applying the

trace operation, we obtain

$$\int_{\mathbf{\Omega}_n} |\widehat{\Phi}(\omega)|^2 d\omega = \int_{\mathbf{\Omega}_n} \mathrm{Tr}\Big(\widehat{\Phi}(\omega)\widehat{\Phi}^*(\omega)\Big)\, d\omega \\ \le c_\epsilon \left(\rho(\mathbf{T}_{\mathbf{P},\mathcal{N}}|_{\mathbf{H}^0_{\mathcal{N}}}) + \epsilon\right)^n = c_\epsilon \left(\rho(\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}^0_N}) + \epsilon\right)^n$$

with c_ϵ independent of n. ∎

Proof of Theorem 3.2. For $s < s_0$, let $\epsilon > 0$ be a constant satisfying $s < -\log_4(\rho(\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}^0_N}) + \epsilon)$.

$$\int_{\mathbf{\Omega}_n} |\widehat{\Phi}(w)|^2 d\omega \le c(\epsilon + \rho(\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}^0_N}))^n$$

for some constant c independent of n and $\widehat{\Phi}$ is continuous on $\mathbf{T}^d$; thus

$$\int_{\mathbf{R}^d} (1+|\omega|^2)^s |\widehat{\Phi}(\omega)|^2 d\omega = \int_{\mathbf{T}^d} (1+|\omega|^2)^s |\widehat{\Phi}(\omega)|^2 d\omega + \sum_{n=1}^{\infty} \int_{\mathbf{\Omega}_n} (1+|\omega|^2)^s |\widehat{\Phi}(\omega)|^2 d\omega \\ \le c + c\sum_{n=1}^{\infty} 2^{2ns} \left(\epsilon + \rho(\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}^0_N})\right)^n < \infty.$$

Therefore $\Phi \in W^s(\mathbf{R}^d)$. ∎

Let $C^\gamma(\mathbf{R}^d)$ denote the space defined in the following way: if $\gamma = n + \gamma_1$ with $n \in \mathbf{Z}_+$ and $0 \le \gamma_1 < 1$, then $f \in C^\gamma(\mathbf{R}^d)$ if and only if $f \in C^{(n)}(\mathbf{R}^d)$ and $f^{(n)}$ is uniformly Hölder continuous with exponent γ_1; i.e.,

$$|D^\beta f(x+y) - D^\beta f(x)| \le c|y|^{\gamma_1} \text{ for any } \beta \in \mathbf{Z}^d_+, |\beta| = n,$$

for some constant c independent of $x, y \in \mathbf{R}^d$. With the well-known inclusion

$$W^s(\mathbf{R}^d) \subset C^\gamma(\mathbf{R}^d) \quad \text{for } s > \gamma + \frac{d}{2},$$

Theorem 3.2 leads to the following corollary.

Corollary 3.3. *Suppose* $\mathbf{P}$ *satisfies conditions* (2.6) *and* (2.7) *for some row vectors* $\mathbf{l}_0^\beta, |\beta| \le 2m-1$. *Let* $\mathbf{H}^0_N$ *be the subspace of* $\mathbf{H}_N$ *defined by* (3.7) *and* $\rho(\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}^0_N})$ *the spectral radius of* $\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}^0_N}$; *then refinable vector* $\Phi \in C^\gamma(\mathbf{R}^d)$ *for any* $\gamma < -\log_4 \rho(\mathbf{T}_{\mathbf{P}}|_{\mathbf{H}^0_N}) - \frac{d}{2}$.

4. Examples. In this section, we will give the Sobolev regularity estimates of some refinable vectors Φ. Before doing this, we shall decide if $\Phi = (\phi_l)_{l=1}^r$ is stable or orthogonal. It was shown (see [7], [13], [21], and [26]) that Φ is stable if and only if there exists a positive constant c such that $G_\Phi(\omega) \ge c\mathbf{I}_r$ for all $\omega \in \mathbf{T}^d$ and that Φ is orthogonal if and only if $G_\Phi(\omega) = \mathbf{I}_r$ for all $\omega \in \mathbf{T}^d$ and the matrix mask $\mathbf{P}$ is a *CQF* (*conjugate quadrature filter*), i.e., $\mathbf{P}$ satisfies

$$\sum_{\nu \in \mathbf{Z}^d/2\mathbf{Z}^d} \mathbf{P}(\omega + \nu\pi)\mathbf{P}^*(\omega + \nu\pi) = \mathbf{I}_r.$$

Assume that $\mathbf{P}$ satisfies the vanishing moment conditions of order at least one, and $\mathbf{P}(0)$ satisfies Condition E. By Theorem 2.2, 1 is an eigenvalue of $\mathbf{T}_{\mathbf{P}}$. If the 1-eigenmatrix of $\mathbf{T}_{\mathbf{P}}$ is positive (or negative) definite on $\mathbf{T}^d$, then there exists a nontrivial

refinable vector Φ in $L^2(\mathbf{R}^d)$ by Theorem 3.1, and $G_\Phi(\omega)$ is also a 1-eigenmatrix of $\mathbf{T_P}$. Therefore if eigenvalue 1 is simple, then $G_\Phi(\omega)$ is the unique (up to a constant) 1-eigenmatrix of $\mathbf{T_P}$ and hence Φ is stable. If $\mathbf{P}$ is a CQF, then $\mathbf{I}_r$ is a 1-eigenmatrix of $\mathbf{T_P}$. Thus if 1 is a simple eigenvalue of $\mathbf{T_P}$, then $\Phi \in L^2(\mathbf{R}^d)$ and $G_\Phi(\omega) = c\mathbf{I}_r$ for some nonzero constant c. Hence Φ is orthogonal; i.e., the integer shifts of ϕ_l, $1 \le l \le r$, form an orthogonal basis of their closed linear span in $L^2(\mathbf{R}^d)$. Therefore to decide if the refinable vector Φ is stable (or orthogonal), we need only to check that if 1 is a simple eigenvalue of $\mathbf{T_P}$ and the corresponding eigenmatrix is positive (or negative) definite on $\mathbf{T}^d$. In fact the stability of Φ implies that $\mathbf{T_P}$ satisfies Condition E and the 1-eigenmatrix of $\mathbf{T_P}$ is positive (or negative) definite on $\mathbf{T}^d$; see [29].

Assume that Φ is a compactly supported refinable vector with refinement mask $\mathbf{P}$ satisfying (2.6) and (2.7) for some row vectors $\mathbf{l}_0^\beta$, $|\beta| \le 2m-1$. To estimate the regularity of Φ by Theorem 3.2, we need to find $\rho(\mathbf{T_P}|_{\mathbf{H}_N^0})$. We note that λ is an eigenvalue of $\mathbf{T_P}|_{\mathbf{H}_N^0}$ if and only if there exists a right eigenvector $\mathbf{v}$ of eigenvalue λ of $\mathcal{T}_\mathbf{P}$ satisfying that for any $\beta, \alpha \in \mathbf{Z}_+^d, |\beta| \le 2m-1, |\alpha| \le m-1, 1 \le j \le r$,

$$\mathbf{L}_N^\beta \mathbf{v} = 0, \quad {}_j\mathbf{l}_N^\alpha \mathbf{v} = 0, \text{ and } {}_j\mathbf{r}_N^\alpha \mathbf{v} = 0, \tag{4.1}$$

where $\mathbf{L}_N^\beta$, ${}_j\mathbf{l}_N^\alpha$, and ${}_j\mathbf{r}_N^\alpha$ are the vectors defined by (2.13) and (3.6), respectively. Let $H_0 \in \mathbf{H}_N$ be the unique matrix function such that $\mathrm{vec}(H_0) = \mathbf{v}$; then H_0 is a λ-eigenmatrix of $\mathbf{T_P}|_{\mathbf{H}_N^0}$. Thus $\rho(\mathbf{T_P}|_{\mathbf{H}_N^0})$ is the largest modulus of all such eigenvalues of $\mathcal{T}_\mathbf{P}$ that have corresponding right eigenvectors satisfying (4.1).

We say that the Sobolev regularity estimate s_0 is optimal if $\Phi \in W^s(\mathbf{R}^d)$ if and only if $s < s_0$.

Example 4.1. Let ϕ_1 and ϕ_2 be the B-splines defined by the knots $0,0,1,1$ and $0,1,1,2$, respectively; i.e., $\phi_1(x) = 2x(1-x)\chi_{[0,1]}(x)$ and $\phi_2(x) = x^2\chi_{[0,1)}(x) + (2-x)^2\chi_{[1,2]}(x)$. Then $\Phi = (\phi_1, \phi_2)^T$ satisfies the matrix refinement equation (1.1) with mask

$$\mathbf{P}(\omega) := \frac{1}{4}\begin{bmatrix} 1+e^{-i\omega} & 1 \\ e^{-i\omega}+e^{-2i\omega} & \frac{1}{2}+2e^{-i\omega}+\frac{1}{2}e^{-2i\omega} \end{bmatrix}.$$

Mask $\mathbf{P}$ satisfies the vanishing moment conditions of order 3 with $\mathbf{l}_0^0 = (1,1)$, $\mathbf{l}_0^1 = (\frac{1}{2}, 1)$ and $\mathbf{l}_0^2 = (0,1)$; see [27]. The eigenvalues of $\mathbf{P}(0)$ are $1, \frac{1}{4}$. We can find vectors $\mathbf{l}_0^3 = (-\frac{1}{4}, 1)$, $\mathbf{l}_0^4 = (-\frac{1}{10}, \frac{9}{10})$ and $\mathbf{l}_0^5 = (\frac{1}{4}, \frac{1}{2})$ satisfying (2.7). In this case, $\mathcal{T}_\mathbf{P}$ is a 20×20 matrix. For $0 \le \beta \le 5$, $\mathbf{L}_2^\beta \ne 0$. Thus $2^{-\beta}$, $0 \le \beta \le 5$, are eigenvalues of $\mathcal{T}_\mathbf{P}$. In fact the eigenvalues of $\mathcal{T}_\mathbf{P}$ or $\mathbf{T_P}$ are $1, \frac{1}{2}, \frac{1}{4}(3), \frac{1}{8}(4), \frac{1}{16}(3), \frac{1}{32}(2), 0(4)$. Here, for an eigenvalue λ, the notation $\lambda(\ell)$ means that the algebraic multiplicity of λ is ℓ. Thus $\mathbf{T_P}$ satisfies Condition E. We can find a right 1-eigenvector $\mathbf{v}$ of $\mathcal{T}_\mathbf{P}$:

$$\mathbf{v} = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 3 \quad 1 \quad 4 \quad 3 \quad 3 \quad 12 \quad 0 \quad 3 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0)^T.$$

That is,

$$H(\omega) = \begin{bmatrix} 4 & 3+3e^{i\omega} \\ 3+3e^{-i\omega} & 12+e^{i\omega}+e^{-i\omega} \end{bmatrix}$$

is a 1-eigenmatrix of $\mathbf{T_P}$. Checking directly, $H(\omega) \ge 2\mathbf{I}_2$ for all $\omega \in \mathbf{T}^d$; thus Φ is stable since $\mathbf{T_P}$ satisfies Condition E.

To estimate the regularity by our method, we need to find the largest eigenvalue module of $\mathbf{T_P}|_{\mathbf{H}_N^0}$. By Corollary 3.2, 1, $\frac{1}{2}$ are not eigenvalues of $\mathbf{T_P}|_{\mathbf{H}_N^0}$. We find $\frac{1}{8}$

is the largest eigenvalue module of $\mathbf{T_P}|_{\mathbf{H}_N^0}$ with a corresponding eigenmatrix

$$\begin{bmatrix} (e^{-i\omega}+e^{i\omega})/2 & -1-e^{i\omega} \\ -1-e^{-i\omega} & 2 \end{bmatrix}.$$

Therefore $\Phi \in W^{\frac{3}{2}-\epsilon}(\mathbf{R})$ or $\Phi \in C^{1-\epsilon}(\mathbf{R})$ for any $\epsilon > 0$, and our estimate is optimal from the definition of Φ.

Example 4.2. Let $\Phi = (\phi_1, \phi_2)^T$ be the refinable vectors treated in [11]. The mask of Φ is given by

$$\mathbf{P}(\omega) := \frac{1}{20}\begin{bmatrix} 6+6e^{-i\omega} & 8\sqrt{2} \\ (-1+9e^{-i\omega}+9e^{-2i\omega}-e^{-3i\omega})/\sqrt{2} & -3+10e^{-i\omega}-3e^{-2i\omega} \end{bmatrix}.$$

Mask $\mathbf{P}$ is a CQF and satisfies the vanishing moment conditions of order 2 with $\mathbf{l}_0^0 = (1.4142, 1)$ and $\mathbf{l}_0^1 = (.7071, 1)$; see [27]. The eigenvalues of $\mathbf{P}(0)$ are $1, -.2$ and we can find vectors $\mathbf{l}_0^2 = (.4714, .8333)$ and $\mathbf{l}_0^3 = (.3536, .5)$ satisfying (2.7). For $0 \le \beta \le 3$, vectors $\mathbf{L}_3^\beta \neq 0$; thus $2^{-\beta}$ are eigenvalues of $\mathcal{T}_\mathbf{P}$. The eigenvalues of $\mathcal{T}_\mathbf{P}$ or $\mathbf{T_P}$ are $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}(2), -.2(2), .2(2), -.1(2), -.05(4), .04$, and $0(12)$. Thus $\mathbf{T_P}$ satisfies Condition E and hence Φ is orthogonal.

By Corollary 3.2, $1, \frac{1}{2}$, and $\frac{1}{4}$ are not eigenvalues of $\mathbf{T_P}|_{\mathbf{H}_N^0}$. We find that the largest eigenvalue module of $\mathbf{T_P}|_{\mathbf{H}_N^0}$ is $\frac{1}{8}$ with a corresponding eigenmatrix $H(\omega) = \sum_{k=-3}^{3} H_k e^{-ik\omega}$ given by

$$H_0 = \begin{bmatrix} -.0875 & .0674 \\ .0674 & -.1085 \end{bmatrix}, \quad H_1 = H_{-1}^T = \begin{bmatrix} -.0042 & .0004 \\ .0674 & -.0417 \end{bmatrix}$$

and

$$H_2 = H_{-2}^T = \begin{bmatrix} 0 & 0 \\ .0004 & 0 \end{bmatrix}, \quad H_3 = H_{-3} = \mathbf{0}.$$

Thus $\Phi \in W^{\frac{3}{2}-\epsilon}(\mathbf{R})$ or $\Phi \in C^{1-\epsilon}(\mathbf{R})$ for any $\epsilon > 0$. It was shown in [11] that Φ is in the Lip space, i.e., $|\Phi(x) - \Phi(y)| \le c|x-y|$ for some constant c independent of $x, y \in \mathbf{R}$. However $\Phi \notin C^1(\mathbf{R})$ since $\frac{1}{\sqrt{2}}(\phi_1(x)+\phi_1(x-1))+\phi_2(x)$ is the hat function $x\chi_{[0,1]}(x) + (2-x)\chi_{(1,2]}(x)$ (see [31]); thus our estimate is optimal.

At last we will analyze two refinable vectors from [1].

Example 4.3. Let $\Phi = (\phi_1, \phi_2)^T$ be the refinable vector treated in [1]. The mask of Φ is given by

$$\mathbf{P}(\omega) := \frac{1}{8}\begin{bmatrix} 2+4e^{-i\omega}+2e^{-2i\omega} & 2-2e^{-2i\omega} \\ -\sqrt{7}+\sqrt{7}e^{-2i\omega} & -\sqrt{7}+2e^{-i\omega}-\sqrt{7}e^{-2i\omega} \end{bmatrix}.$$

Mask $\mathbf{P}$ is a CQF and satisfies the vanishing moment conditions of order 2 with $\mathbf{l}_0^0 = (1, 0)$ and $\mathbf{l}_0^1 = (1, .2743)$, see [1]. The eigenvalues of $\mathbf{P}(0)$ are $1, -.4114$, and we can find vectors $\mathbf{l}_0^2 = (1.0752, .5486)$ and $\mathbf{l}_0^3 = (1.2257, .7909)$ satisfying (2.7). For $0 \le \beta \le 3$, vectors $\mathbf{L}_2^\beta \neq 0$; thus $2^{-\beta}$, $0 \le \beta \le 3$ are eigenvalues of $\mathcal{T}_\mathbf{P}$. The eigenvalues of $\mathbf{T_P}$ or $\mathcal{T}_\mathbf{P}$ are $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, -.4114(2), .2318, -.2057(3), .0130(2)$, and $0(8)$. Thus $\mathbf{T_P}$ satisfies Condition E and Φ is orthogonal.

By Corollary 3.2, $1, \frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$ are not eigenvalues of $\mathbf{T_P}|_{\mathbf{H}_N^0}$. We find the largest eigenvalue module of $\mathbf{T_P}|_{\mathbf{H}_N^0}$ is .2318 with a corresponding eigenmatrix $H(\omega) =$

$\sum_{k=-2}^{2} H_k e^{-ik\omega}$ given by

$$H_0 = \begin{bmatrix} .2117 & 0 \\ 0 & .7564 \end{bmatrix}, \quad H_1 = H_{-1}^T = \begin{bmatrix} -.1059 & .1930 \\ -.1930 & .3253 \end{bmatrix}$$

and $H_2 = H_{-2} = \mathbf{0}$. Thus $\Phi \in W^{1.0545-\epsilon}(\mathbf{R})$ or $\Phi \in C^{.5545-\epsilon}(\mathbf{R})$ for any $\epsilon > 0$.

Example 4.4. Let $\Phi = (\phi_1, \phi_2)^T$ be another refinable vector treated in [1]. The mask $\mathbf{P}(\omega) := \frac{1}{2}\sum_{k=0}^{3} \mathbf{P}_k e^{-ik\omega}$ of Φ is given by

$$\mathbf{P}_0 = \frac{1}{40}\begin{bmatrix} 10-3\sqrt{10} & 5\sqrt{6}-2\sqrt{15} \\ 5\sqrt{6}-3\sqrt{15} & 5-3\sqrt{10} \end{bmatrix}, \mathbf{P}_1 = \frac{1}{40}\begin{bmatrix} 30+3\sqrt{10} & 5\sqrt{6}-2\sqrt{15} \\ -5\sqrt{6}-7\sqrt{15} & 5-3\sqrt{10} \end{bmatrix},$$

and $\mathbf{P}_2 = S_0\mathbf{P}_1 S_0, \mathbf{P}_3 = S_0\mathbf{P}_0 S_0$, where $S_0 = \mathrm{diag}(1,-1)$. Mask $\mathbf{P}$ is a CQF and satisfies the vanishing moment conditions of order 3 with $\mathbf{l}_0^0 = (1,0)$, $\mathbf{l}_0^1 = (1.5, .2372)$, and $\mathbf{l}_0^2 = (2.3063, .7117)$; see [1]. The eigenvalues of $\mathbf{P}(0)$ are $1, .0257$ and we can find vectors $\mathbf{l}_0^3 = (3.6283, 1.8980)$, $\mathbf{l}_0^4 = (6.0943, 4.9822)$, and $\mathbf{l}_0^5 = (11.5329, 13.4836)$ satisfying (2.7). Vectors $\mathbf{L}_3^\beta \neq 0$, thus $2^{-\beta}$, $0 \leq \beta \leq 5$, are eigenvalues of $\mathbf{T_P}$. The eigenvalues of $\mathbf{T_P}$ or $\mathcal{T}_\mathbf{P}$ are $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, .1357, -.0625, -.0576, .0257(2)$, $.0128(2), .0078(2), .0064(4), -.0016(4), .0032(2), .0007$, and $.0003(2)$. Thus $\mathbf{T_P}$ satisfies Condition E and Φ is orthogonal.

By Corollary 3.2, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ are not eigenvalues of $\mathbf{T_P}|_{\mathbf{H}_N^0}$. We find the largest eigenvalue module of $\mathbf{T_P}|_{\mathbf{H}_N^0}$ is .1357 with a corresponding eigenmatrix $H(\omega) = \sum_{k=-3}^{3} H_k e^{-ik\omega}$ given by

$$H_0 = \begin{bmatrix} .1180 & 0 \\ 0 & .8072 \end{bmatrix}, \quad H_1 = H_{-1}^T = \begin{bmatrix} -.0506 & .1602 \\ -.1602 & .3362 \end{bmatrix}$$

and

$$H_2 = H_{-2}^T = \begin{bmatrix} -.0084 & .0087 \\ -.0087 & .0086 \end{bmatrix}, \quad H_3 = H_{-3} = \mathbf{0}.$$

Thus $\Phi \in W^{1.4408-\epsilon}(\mathbf{R})$ or $\Phi \in C^{.9408-\epsilon}(\mathbf{R})$ for any $\epsilon > 0$.

Acknowledgments. The author would like to express his thanks to two anonymous referees for many helpful suggestions to this paper.

REFERENCES

[1] C. K. Chui and J. Lian, *A study on orthonormal multi-wavelets*, J. Appl. Numer. Math., 20 (1996), pp. 273–298.

[2] A. Cohen and I. Daubechies, *A new technique to estimate the regularity of refinable functions*, Rev. Mat. Iberoamericana, 12 (1996), pp. 527–591.

[3] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.

[4] A. Cohen, K. Gröchenig, and L. Villemoes, *Regularity of multivariable refinable functions*, preprint, 1996.

[5] J. P. Conze and A. Raugi, *Fonctions harmonique pour un opérateur de transition et applications*, Bull. Soc. Math. France, 118 (1990), pp. 273–310.

[6] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[7] C. de Boor, R. DeVore, and A. Ron, *The structure of finitely generated shift-invariant spaces in $L^2(\mathbf{R}^d)$*, J. Funct. Anal., 119 (1994), pp. 37–78.

[8] G. Donovan, J. Geronimo, D. Hardin, and P. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1791–1815.

[9] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.
[10] G. Gripenberg, *Unconditional bases of wavelets for Sobolev spaces*, SIAM J. Math. Anal., 24 (1993), pp. 1030–1042.
[11] J. Geronimo, D. Hardin, and P. Massopust, *Fractal functions and wavelet expansions based on several scaling functions*, J. Approx. Theory, 78 (1994), pp. 373–401.
[12] T. N. Goodman and S. L. Lee, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.
[13] T. N. Goodman, S. L. Lee, and W. S. Tang, *Wavelets in wandering subspaces*, Trans. Amer. Math. Soc., 338 (1993), pp. 639–654.
[14] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.
[15] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., 73 (1996), pp. 75–94.
[16] L. Hervé, *Multi-resolution analysis of multiplicity d: Application to dyadic interpolation*, Appl. Comput. Harmon. Anal., 1 (1994), pp. 299–315.
[17] L. Hervé, *Construction et régularité des fonctions d'échelle*, SIAM J. Math. Anal., 26 (1995), pp. 1361–1385.
[18] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1989.
[19] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1991.
[20] R. Jia, *Characterization of smoothness of multivariate refinable functions in Sobolev spaces*, preprint, 1996.
[21] R. Jia and C. Micchelli, *Using the refinement equation for the construction of prewavelets* II*: Powers of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.
[22] R. Jia and Z. Shen, *Multiresolution and wavelets,* Proc. Edinburgh Math. Soc., 37 (1994), pp. 271–300.
[23] Q. Jiang and Z. Shen, *On existence and weak stability of matrix refinable functions*, Constr. Approx., to appear.
[24] W. Lawton, *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys., 32 (1991), pp. 57–61.
[25] W. Lawton, S. L. Lee, and Z. Shen, *An algorithm for matrix extension and wavelet construction*, Math. Comp., 65 (1996), pp. 723–737.
[26] R. Long, W. Chen, and S. Yuan, *Wavelets generated by vector multiresolution analysis*, Appl. Comput. Harmon. Anal., 4 (1997), pp. 317–350.
[27] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.
[28] S. Riemenschneider and Z. Shen, *Multidimensional interpolatory subdivision schemes*, SIAM J. Numer. Anal., 34 (1997), pp. 2357–2381.
[29] Z. Shen, *Refinable function vectors*, SIAM J. Math. Anal, 29 (1998), pp. 234–249.
[30] W. So and J. Wang, *Estimating the support of a scaling vector*, SIAM. J. Matrix Anal. Appl., 18 (1997), pp. 66–73.
[31] G. Strang and V. Strela, *Short wavelets and matrix dilation equations*, IEEE Trans. Signal Proc., 43 (1995), pp. 108–115.
[32] L. Villemoes, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1177–1199, September 1998

008

MULTIVARIATE REFINEMENT EQUATIONS AND CONVERGENCE OF SUBDIVISION SCHEMES*

BIN HAN† AND RONG-QING JIA†

Abstract. Refinement equations play an important role in computer graphics and wavelet analysis. In this paper we investigate multivariate refinement equations associated with a dilation matrix and a finitely supported refinement mask. We characterize the L_p-convergence of a subdivision scheme in terms of the p-norm joint spectral radius of a collection of matrices associated with the refinement mask. In particular, the 2-norm joint spectral radius can be easily computed by calculating the eigenvalues of a certain linear operator on a finite dimensional linear space. Examples are provided to illustrate the general theory.

Key words. refinement equations, subdivision schemes, joint spectral radii, wavelets

AMS subject classifications. 39B12, 41A25, 42C15, 65D99

PII. S0036141097294032

1. Introduction. The purpose of this paper is to investigate functional equations of the form

$$f = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha) f(M\cdot - \alpha), \tag{1.1}$$

where f is the unknown function defined on the s-dimensional Euclidean space $\mathbb{R}^s$, a is a finitely supported sequence on $\mathbb{Z}^s$, and M is an $s \times s$ integer matrix such that $\lim_{n\to\infty} M^{-n} = 0$. Equation (1.1) is called a *refinement equation*, and the matrix M is called a *dilation matrix*. Correspondingly, the sequence a is called the *refinement mask*. Any function satisfying a refinement equation is called a *refinable function*. Refinement equations play an important role in computer graphics and wavelet analysis. See Jia and Micchelli [9] for some discussion of multiresolution and wavelet decompositions related to general dilation matrices.

If a satisfies

$$\sum_{\alpha \in \mathbb{Z}^s} a(\alpha) = m := |\det M|, \tag{1.2}$$

then it is known that there exists a unique compactly supported distribution f satisfying the refinement equation (1.1) subject to the condition $\hat{f}(0) = 1$. This distribution is said to be *the normalized solution* to the refinement equation with mask a. This fact was essentially proved by Cavaretta, Dahmen, and Micchelli in [1, Chap. 5] for the case in which the dilation matrix is two times the $s \times s$ identity matrix I. The same proof applies to the general refinement equation (1.1).

*Received by the editors March 19, 1997; accepted for publication (in revised form) September 22, 1997; posted electronically April 30, 1998. This research was supported in part by NSERC Canada under grant OGP 121336. The results of this paper were reported by the second author on August 9, 1995, in the 1995 AMS-SIAM Summer Seminar in Applied Mathematics, Park City, Utah.
http://www.siam.org/journals/sima/29-5/29403.html

†Department of Mathematical Sciences, University of Alberta, Edmonton, T6G 2G1, Canada (bhan@math.ualberta.ca, http://approx.math.ualberta.ca/~bhan; jia@xihu.math.ualberta.ca).

For $1 \le p \le \infty$, by $L_p(\mathbb{R}^s)$ we denote the Banach space of all (complex-valued) measurable functions f on $\mathbb{R}^s$ such that $\|f\|_p < \infty$, where

$$\|f\|_p := \left(\int_{\mathbb{R}^s} |f(x)|^p \, dx \right)^{1/p} \qquad \text{for } 1 \le p < \infty$$

and $\|f\|_\infty$ is the essential supremum of f on $\mathbb{R}^s$. The Fourier transform of a function $f \in L_1(\mathbb{R}^s)$ is defined to be

$$\hat{f}(\xi) := \int_{\mathbb{R}^s} f(x) e^{-ix\cdot\xi} \, dx, \qquad \xi \in \mathbb{R}^s,$$

where $x \cdot \xi$ denotes the inner product of two vectors x and ξ in $\mathbb{R}^s$.

Let f be the normalized solution to the refinement equation (1.1). Taking the Fourier transform of the functions on both sides of (1.1), we obtain

$$\hat{f}(\xi) = H\big((M^T)^{-1}\xi\big)\hat{f}\big((M^T)^{-1}\xi\big), \qquad \xi \in \mathbb{R}^s, \tag{1.3}$$

where M^T denotes the transpose of M, and

$$H(\xi) := \sum_{\alpha\in\mathbb{Z}^s} a(\alpha) e^{-i\alpha\cdot\xi}/m, \qquad \xi \in \mathbb{R}^s. \tag{1.4}$$

Obviously, (1.2) implies $H(2\beta\pi) = 1$ for all $\beta \in \mathbb{Z}^s$. Thus, it follows from (1.3) that, for all positive integers k and all $\beta \in \mathbb{Z}^s$,

$$\hat{f}\big(2\pi(M^T)^k\beta\big) = \hat{f}(2\pi\beta).$$

If, in addition, f lies in $L_1(\mathbb{R}^s)$, then by the Riemann–Lebesgue lemma we have

$$\hat{f}(2\pi\beta) = \lim_{k\to\infty} \hat{f}\big(2\pi(M^T)^k\beta\big) = 0 \qquad \forall\, \beta \in \mathbb{Z}^s \setminus \{0\}.$$

A function f is said to satisfy the *moment conditions* of order 1 if $\hat{f}(0) = 1$ and $\hat{f}(2\pi\beta) = 0$ for all $\beta \in \mathbb{Z}^s \setminus \{0\}$. Thus, if the normalized solution f of the refinement equation (1.1) lies in $L_1(\mathbb{R}^s)$, then f satisfies the moment conditions of order 1.

In order to solve the refinement equation (1.1), we start with a compactly supported function $\phi \in L_p(\mathbb{R}^s)$ $(1 \le p \le \infty)$ and use the iteration scheme $f_n := T_a^n\phi$, $n = 0, 1, 2, \ldots$, where T_a is the bounded linear operator on $L_p(\mathbb{R}^s)$ given by

$$T_a\phi := \sum_{\alpha\in\mathbb{Z}^s} a(\alpha)\phi(M\cdot - \alpha), \qquad \phi \in L_p(\mathbb{R}^s). \tag{1.5}$$

This iteration scheme is called a *subdivision scheme* (see [1]). In the literature a subdivision scheme is also referred to as a *cascade algorithm*.

Let $\ell(\mathbb{Z}^s)$ denote the linear space of all sequences on $\mathbb{Z}^s$, and let $\ell_0(\mathbb{Z}^s)$ denote the linear space of all finitely supported sequences on $\mathbb{Z}^s$. For $\beta \in \mathbb{Z}^s$ we use δ_β to denote the sequence on $\mathbb{Z}^s$ given by

$$\delta_\beta(\alpha) = \begin{cases} 1 & \text{if } \alpha = \beta, \\ 0 & \text{if } \alpha \in \mathbb{Z}^s \setminus \{\beta\}. \end{cases}$$

In particular, we write δ for δ_0. For a vector $y \in \mathbb{Z}^s$ we use ∇_y to denote the difference operator on $\ell(\mathbb{Z}^s)$ given by

$$\nabla_y u = u - u(\cdot - y), \qquad u \in \ell(\mathbb{Z}^s).$$

Let $e_1, \dots, e_s$ denote the unit coordinate vectors in $\mathbb{R}^s$. For simplicity, we write ∇_j for ∇_{e_j}, $j = 1, \dots, s$.

The *subdivision operator* S_a associated with a (see [1]) is the linear operator on $\ell(\mathbb{Z}^s)$ given by

$$S_a u(\alpha) := \sum_{\beta \in \mathbb{Z}^s} a(\alpha - M\beta) u(\beta), \qquad \alpha \in \mathbb{Z}^s, \tag{1.6}$$

where $u \in \ell(\mathbb{Z}^s)$. Then for $\phi \in L_p(\mathbb{R}^s)$ $(1 \le p \le \infty)$ we have

$$T_a \phi = \sum_{\alpha \in \mathbb{Z}^s} S_a \delta(\alpha) \phi(M\cdot - \alpha).$$

By induction on n, it is easily verified that

$$T_a^n \phi = \sum_{\alpha \in \mathbb{Z}^s} S_a^n \delta(\alpha) \phi(M^n \cdot - \alpha). \tag{1.7}$$

Taking the Fourier transform of both sides of (1.5), we obtain

$$\widehat{T_a \phi}(\xi) = H\big((M^T)^{-1}\xi\big)\, \hat{\phi}\big((M^T)^{-1}\xi\big), \qquad \xi \in \mathbb{R}^s, \tag{1.8}$$

where H is given by (1.4). It follows from (1.8) that

$$\hat{f}_n(\xi) = \prod_{k=1}^{n} H\big((M^T)^{-k}\xi\big) \hat{\phi}\big((M^T)^{-n}\xi\big), \qquad \xi \in \mathbb{R}^s.$$

Since $H(2\pi\beta) = 1$ for all $\beta \in \mathbb{Z}^s$, we have $\hat{\phi}(2\pi\beta) = \hat{f}_n(2\pi(M^T)^n\beta)$. Suppose f_n converges to the normalized solution f of (1.1) in the L_p-norm for some p $(1 \le p \le \infty)$. Then

$$\hat{\phi}(2\pi\beta) = \lim_{n\to\infty} \hat{f}_n(2\pi(M^T)^n\beta) = \delta(\beta) \qquad \forall\, \beta \in \mathbb{Z}^s.$$

In other words, ϕ must satisfy the moment conditions of order 1.

By using the Poisson summation formula, it is easily seen that a compactly supported integrable function ϕ satisfies the moment conditions of order 1 if and only if its shifts form a partition of unity; i.e.,

$$\sum_{\alpha \in \mathbb{Z}^s} \phi(\cdot - \alpha) = 1. \tag{1.9}$$

Let ϕ_0 be the function given by

$$\phi_0(x) := \prod_{j=1}^{s} \chi(x_j) \qquad \text{for } x = (x_1, \dots, x_s) \in \mathbb{R}^s, \tag{1.10}$$

where

$$\chi(t) = \begin{cases} 1+t & \text{for } t \in [-1,0), \\ 1-t & \text{for } t \in [0,1], \\ 0 & \text{for } t \in \mathbb{R} \setminus [-1,1]. \end{cases}$$

Evidently, ϕ_0 satisfies (1.9).

We say that the subdivision scheme associated with mask a *converges in the* L_p*-norm* if there is a function $f \in L_p(\mathbb{R}^s)$ such that

$$\lim_{n\to\infty} \|T_a^n \phi_0 - f\|_p = 0,$$

where ϕ_0 is the function given in (1.10). If the subdivision scheme converges in the L_∞-norm, then the limit function is continuous.

When the dilation matrix is two times the $s \times s$ identity matrix I, the uniform convergence of a subdivision scheme was studied by Cavaretta, Dahmen, and Micchelli [1]. In particular, they proved that the subdivision scheme associated with a mask a converges uniformly, provided the normalized solution of the corresponding refinement equation is continuous and has stable shifts. Dyn [4] also considered the uniform convergence of multivariate subdivision schemes and related her study to matrix subdivision schemes. Concerning general dilation matrices, Deslauriers, Dubois, and Dubuc [3] found a necessary and sufficient condition for the uniform convergence of interpolatory subdivision schemes.

In their study of nonseparable multidimensional wavelet bases, Kovaĉević and Vetterli [10] assumed the L_2-convergence of the subdivision scheme, but did not give any detail about possible characterization of the L_2-convergence. In [2], Cohen and Daubechies built orthonormal and biorthogonal wavelet bases of $L_2(\mathbb{R}^2)$ with dilation matrices of determinant 2 and established a sufficient condition for the uniform convergence of the corresponding subdivision scheme. In [13] Villemoes investigated continuity of nonseparable quincunx wavelets and the uniform convergence of related subdivision schemes.

In this paper we give a systematic and comprehensive study of multivariate refinement equations and subdivision schemes. Our main result characterizes the L_p-convergence of a subdivision scheme associated with a general dilation matrix M and a mask a. All the relevant results mentioned above are special cases of the general setting considered in this paper.

To describe our main result, we introduce the linear operators A_ε ($\varepsilon \in \mathbb{Z}^s$) on $\ell_0(\mathbb{Z}^s)$ as follows:

$$A_\varepsilon v(\alpha) = \sum_{\beta\in\mathbb{Z}^s} a(\varepsilon + M\alpha - \beta) v(\beta), \qquad v \in \ell_0(\mathbb{Z}^s),\ \alpha \in \mathbb{Z}^s. \tag{1.11}$$

We observe that the set $\mathbb{Z}^s$ is an abelian group under addition, and $M\mathbb{Z}^s$ is a subgroup of $\mathbb{Z}^s$. Let E be a complete set of representatives of the distinct cosets of the quotient group $\mathbb{Z}^s/M\mathbb{Z}^s$. For a finite subset K of $\mathbb{Z}^s$, we denote by $\ell(K)$ the linear subspace of $\ell_0(\mathbb{Z}^s)$ consisting of all sequences supported on K. It will be proved in section 2 that there exists a finite subset K of $\mathbb{Z}^s$ such that $\ell(K)$ contains $\nabla_j \delta$ for $j = 1, \ldots, s$ and is invariant under every A_ε ($\varepsilon \in E$). Let

$$V := \Big\{ v \in \ell(K) : \sum_{\alpha\in\mathbb{Z}^s} v(\alpha) = 0 \Big\}.$$

It is easily seen that V is a common invariant subspace of A_ε ($\varepsilon \in E$) if and only if $\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta) = 1$ for all $\alpha \in \mathbb{Z}^s$.

Let M be a dilation matrix with $m := |\det M|$ and let a be an element in $\ell_0(\mathbb{Z}^s)$ such that $\sum_{\alpha\in\mathbb{Z}^s} a(\alpha) = m$. Our main result states that the subdivision scheme associated with the mask a and the dilation matrix M converges in the L_p-norm ($1 \le p \le \infty$) if and only if the following two conditions are satisfied:

(a) $\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta) = 1$ for all $\alpha \in \mathbb{Z}^s$;

(b) $\rho_p(\{A_\varepsilon|_V : \varepsilon \in E\}) < m^{1/p}$, where $\rho_p(\{A_\varepsilon|_V : \varepsilon \in E\})$ denotes the joint p-norm spectral radius of the linear operators $A_\varepsilon|_V$, $\varepsilon \in E$.

This result will be proved in section 3 after a discussion of the joint spectral radius in section 2. In section 4, we demonstrate that the 2-norm joint spectral radius can be easily computed by calculating the eigenvalues of a certain linear operator on a finite dimensional linear space. Throughout the paper, examples are provided to illustrate the general theory.

2. Joint spectral radius. This section is devoted to a study of joint spectral radii of a finite collection of linear operators associated to a refinement equation.

Let $\mathcal{A}$ be a finite collection of linear operators on a *finite dimensional* vector space V. A vector norm $\|\cdot\|$ on V induces a norm on the linear operators on V as follows. For a linear operator A on V, define

$$\|A\| := \max_{\|v\|=1} \big\{\|Av\|\big\}.$$

For a positive integer n we denote by $\mathcal{A}^n$ the Cartesian power of $\mathcal{A}$:

$$\mathcal{A}^n = \big\{(A_1, \dots, A_n) : A_1, \dots, A_n \in \mathcal{A}\big\}.$$

When $n = 0$, we interpret $\mathcal{A}^0$ as the set $\{I\}$, where I is the identity mapping on V. Let

$$\|\mathcal{A}^n\|_\infty := \max\big\{\|A_1 \cdots A_n\| : (A_1, \dots, A_n) \in \mathcal{A}^n\big\}.$$

Then the uniform joint spectral radius of $\mathcal{A}$ is defined to be

$$\rho_\infty(\mathcal{A}) := \lim_{n\to\infty} \|\mathcal{A}^n\|_\infty^{1/n}.$$

The uniform joint spectral radius was introduced by Rota and Strang in [12].

The p-norm joint spectral radius of a finite collection of linear operators was introduced by Jia in [6]. We define, for $1 \le p < \infty$,

$$\|\mathcal{A}^n\|_p := \left(\sum_{(A_1,\dots,A_n)\in\mathcal{A}^n} \|A_1 \cdots A_n\|^p\right)^{1/p}.$$

For $1 \le p \le \infty$, the *p-norm joint spectral radius* of $\mathcal{A}$ is defined to be

$$\rho_p(\mathcal{A}) := \lim_{n\to\infty} \|\mathcal{A}^n\|_p^{1/n}.$$

It is easily seen that this limit indeed exists, and

$$\lim_{n\to\infty} \|\mathcal{A}^n\|_p^{1/n} = \inf_{n\ge1} \|\mathcal{A}^n\|_p^{1/n}.$$

Clearly, $\rho_p(\mathcal{A})$ is independent of the choice of the vector norm on V.

If $\mathcal{A}$ consists of a single linear operator A, then

$$\rho_p(\mathcal{A}) = \rho(A),$$

where $\rho(A)$ denotes the spectral radius of A, which is independent of p. If $\mathcal{A}$ consists of more than one element, then $\rho_p(\mathcal{A})$ depends on p in general. By some basic properties of ℓ_p spaces, we have that, for $1 \le p \le r \le \infty$,

$$(\#\mathcal{A})^{1/r-1/p}\rho_p(\mathcal{A}) \le \rho_r(\mathcal{A}) \le \rho_p(\mathcal{A}),$$

where $\#\mathcal{A}$ denotes the number of elements in $\mathcal{A}$. Furthermore, it is easily seen from the definition of the joint spectral radius that $\rho(A) \le \rho_\infty(\mathcal{A})$ for any element A in $\mathcal{A}$.

Recall that S_a is the subdivision operator given in (1.6). From (1.7) we see that, in order to study convergence of the subdivision scheme, we need to analyze the sequences $S_a^n\delta$, $n = 1, 2, \ldots$. For this purpose, we introduce the biinfinite matrices A_ε ($\varepsilon \in \mathbb{Z}^s$) as follows:

$$A_\varepsilon(\alpha, \beta) := a(\varepsilon + M\alpha - \beta), \qquad \alpha, \beta \in \mathbb{Z}^s. \tag{2.1}$$

LEMMA 2.1. *Suppose $\alpha = \varepsilon_1 + M\varepsilon_2 + \cdots + M^{n-1}\varepsilon_n + M^n\gamma$, where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, $\gamma \in \mathbb{Z}^s$. Then for any $\beta \in \mathbb{Z}^s$,*

$$S_a^n\delta(\alpha - \beta) = A_{\varepsilon_n} \cdots A_{\varepsilon_1}(\gamma, \beta).$$

Proof. The proof proceeds by induction on n. For $n = 1$ and $\alpha = \varepsilon_1 + M\gamma$, we have

$$S_a\delta(\alpha - \beta) = a(\varepsilon_1 + M\gamma - \beta) = A_{\varepsilon_1}(\gamma, \beta).$$

Suppose $n > 1$ and the lemma has been verified for $n - 1$. For $\alpha = \varepsilon_1 + M\alpha_1$, where $\alpha_1, \varepsilon \in \mathbb{Z}^s$, we have

$$S_a^n\delta(\alpha - \beta) = \sum_{\eta\in\mathbb{Z}^s} a(\alpha - \beta - M\eta)S_a^{n-1}\delta(\eta) = \sum_{\eta\in\mathbb{Z}^s} a(\varepsilon_1 + M\eta - \beta)S_a^{n-1}\delta(\alpha_1 - \eta). \tag{2.2}$$

Suppose $\alpha_1 = \varepsilon_2 + \cdots + M^{n-2}\varepsilon_n + M^{n-1}\gamma$. Then by the induction hypothesis we have

$$S_a^{n-1}\delta(\alpha_1 - \eta) = A_{\varepsilon_n} \cdots A_{\varepsilon_2}(\gamma, \eta).$$

This, in connection with (2.2), gives

$$S_a^n\delta(\alpha - \beta) = \sum_{\eta\in\mathbb{Z}^s} A_{\varepsilon_n} \cdots A_{\varepsilon_2}(\gamma, \eta)A_{\varepsilon_1}(\eta, \beta) = A_{\varepsilon_n} \cdots A_{\varepsilon_2}A_{\varepsilon_1}(\gamma, \beta),$$

thereby completing the induction procedure. ∎

The biinfinite matrices A_ε ($\varepsilon \in \mathbb{Z}^s$) defined in (2.1) may be viewed as the linear operators given in (1.11).

Now let $\mathcal{A}$ be a finite collection of linear operators on a vector space V, which is not necessarily finite dimensional. A subspace W of V is said to be $\mathcal{A}$-invariant if it is invariant under every operator A in $\mathcal{A}$. Let U be a subset of V. The intersection of all $\mathcal{A}$-invariant subspaces of V containing U is $\mathcal{A}$-invariant, and we call it the *minimal*

$\mathcal{A}$-*invariant subspace* generated by U, or the *minimal common invariant subspace* of the operators A in $\mathcal{A}$ generated by U. This subspace is spanned by the set

$$\{A_1 \cdots A_j u : u \in U,\, (A_1, \dots, A_j) \in \mathcal{A}^j,\, j = 0, 1, \dots\}.$$

If, in addition, V is finite dimensional, then there exists a positive integer k such that the set

$$\{A_1 \cdots A_j u : u \in U,\, (A_1, \dots, A_j) \in \mathcal{A}^j,\, j = 0, 1, \dots, k\}$$

already spans the minimal $\mathcal{A}$-invariant subspace generated by U.

We define, for $1 \le p < \infty$,

$$\|\mathcal{A}^n v\|_p := \left(\sum_{(A_1, \dots, A_n) \in \mathcal{A}^n} \|A_1 \cdots A_n v\|^p \right)^{1/p}$$

and, for $p = \infty$,

$$\|\mathcal{A}^n v\|_\infty := \max\bigl\{ \|A_1 \cdots A_n v\| : (A_1, \dots, A_n) \in \mathcal{A}^n \bigr\}.$$

The *symbol* of a sequence $a \in \ell_0(\mathbb{Z}^s)$ is the Laurent polynomial $\tilde{a}(z)$ given by

$$\tilde{a}(z) := \sum_{\alpha \in \mathbb{Z}^s} a(\alpha) z^\alpha, \qquad z \in (\mathbb{C} \setminus \{0\})^s,$$

where $z^\alpha := z_1^{\alpha_1} \cdots z_s^{\alpha_s}$ for $z = (z_1, \dots, z_s) \in (\mathbb{C} \setminus \{0\})^s$ and $\alpha = (\alpha_1, \dots, \alpha_s) \in \mathbb{Z}^s$.

For $\beta \in \mathbb{Z}^s$ we denote by τ^β the shift operator on $\ell_0(\mathbb{Z}^s)$ given by

$$\tau^\beta \lambda := \lambda(\cdot - \beta) \qquad \text{for } \lambda \in \ell_0(\mathbb{Z}^s).$$

Let ν be an element of $\ell_0(\mathbb{Z}^s)$. Then its symbol $\tilde{\nu}(z)$ is a Laurent polynomial, which induces the difference operator $\tilde{\nu}(\tau) := \sum_{\beta \in \mathbb{Z}^s} \nu(\beta) \tau^\beta$.

Let E be a complete set of representatives of the distinct cosets of the quotient group $\mathbb{Z}^s / M\mathbb{Z}^s$. We assume that E contains 0. Thus, each element $\alpha \in \mathbb{Z}^s$ can be uniquely represented as $\varepsilon + M\gamma$, where $\varepsilon \in E$ and $\gamma \in \mathbb{Z}^s$.

As usual, for $1 \le p \le \infty$, $\ell_p(\mathbb{Z}^s)$ denotes the Banach space of all sequences on $\mathbb{Z}^s$ such that $\|a\|_p < \infty$, where

$$\|a\|_p := \Bigl(\sum_{\alpha \in \mathbb{Z}^s} |a(\alpha)|^p \Bigr)^{1/p} \qquad \text{for } 1 \le p < \infty,$$

and $\|a\|_\infty$ is the supremum of a on $\mathbb{Z}^s$. In the following lemma, the underlying vector norm on $\ell_0(\mathbb{Z}^s)$ is chosen to be the ℓ_p-norm.

LEMMA 2.2. *Let S_a be the subdivision operator associated with a dilation matrix M and a mask a. Let $\mathcal{A} := \{A_\varepsilon : \varepsilon \in E\}$, where A_ε are the linear operators on $\ell_0(\mathbb{Z}^s)$ given by* (1.11). *Then for $1 \le p \le \infty$ and $\nu \in \ell_0(\mathbb{Z}^s)$,*

$$\|\tilde{\nu}(\tau) S_a^n \delta\|_p = \|\mathcal{A}^n \nu\|_p, \qquad n = 1, 2, \dots. \tag{2.3}$$

Proof. Suppose $\alpha = \varepsilon_1 + M\varepsilon_2 + \cdots + M^{n-1}\varepsilon_n + M^n\gamma$, where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \in E$ and $\gamma \in \mathbb{Z}^s$. Then by Lemma 2.1 we have

$$\begin{aligned} \tilde{\nu}(\tau) S_a^n \delta(\alpha) &= \sum_{\beta \in \mathbb{Z}^s} \nu(\beta) S_a^n \delta(\alpha - \beta) \\ &= \sum_{\beta \in \mathbb{Z}^s} A_{\varepsilon_n} \cdots A_{\varepsilon_1}(\gamma, \beta) \nu(\beta) = A_{\varepsilon_n} \cdots A_{\varepsilon_1} \nu(\gamma). \end{aligned}$$

Hence, (2.3) is true for $p = \infty$. When $1 \le p < \infty$ we have

$$\sum_{\alpha\in\mathbb{Z}^s} \left|\tilde{\nu}(\tau) S_a^n \delta(\alpha)\right|^p = \sum_{(\varepsilon_1,\dots,\varepsilon_n)\in E^n} \sum_{\gamma\in\mathbb{Z}^s} \left|A_{\varepsilon_n}\cdots A_{\varepsilon_1}\nu(\gamma)\right|^p.$$

This verifies (2.3) for $1 \le p < \infty$. □

Let $\mathcal{A} := \{A_\varepsilon : \varepsilon \in E\}$. We claim that, for each $\nu \in \ell_0(\mathbb{Z}^s)$, the minimal $\mathcal{A}$-invariant subspace generated by ν is finite dimensional. To establish this result, we shall introduce the concept of admissible sets. For a finite subset K of $\mathbb{Z}^s$, recall that $\ell(K)$ is the linear subspace of $\ell_0(\mathbb{Z}^s)$ consisting of all sequences supported on K. Let A be a linear operator on $\ell_0(\mathbb{Z}^s)$. A finite subset K of $\mathbb{Z}^s$ is said to be *admissible* for A if $\ell(K)$ is invariant under A. See [5] for the related notion of *good* sets. The following lemma shows that there exists a finite subset K of $\mathbb{Z}^s$ such that K contains the support of ν and is admissible for all A_ε, $\varepsilon \in E$.

LEMMA 2.3. *Suppose M is an $s \times s$ dilation matrix and a is a sequence on $\mathbb{Z}^s$ with its support $\Omega := \{\alpha \in \mathbb{Z}^s : a(\alpha) \ne 0\}$ being finite. Let A_ε ($\varepsilon \in E$) be the linear operators on $\ell_0(\mathbb{Z}^s)$ given by* (1.11). *Then a finite subset K of $\mathbb{Z}^s$ is admissible for $A := A_0$ if and only if*

$$M^{-1}(\Omega + K) \cap \mathbb{Z}^s \subseteq K. \tag{2.4}$$

Consequently, for any finite subset G of $\mathbb{Z}^s$, there exists a finite subset K of $\mathbb{Z}^s$ such that K contains G and K is admissible for all A_ε, $\varepsilon \in E$.

Proof. Suppose K is admissible for A. Let $\alpha \in M^{-1}(\Omega + K) \cap \mathbb{Z}^s$. Then we have $M\alpha = \gamma + \beta$ for some $\gamma \in \Omega$ and $\beta \in K$. It follows that

$$A\delta_\beta(\alpha) = a(M\alpha - \beta) = a(\gamma) \ne 0.$$

Since K is admissible for A, we have $A\delta_\beta \in \ell(K)$, and therefore $\alpha \in K$. This shows that (2.4) is true.

Conversely, suppose (2.4) is true. Let $v \in \ell(K)$ and $\alpha \in \mathbb{Z}^s$. Then

$$Av(\alpha) = \sum_{\beta\in\mathbb{Z}^s} a(M\alpha - \beta)v(\beta) \ne 0$$

implies that $M\alpha - \beta \in \Omega$ for some $\beta \in K$. It follows that $M\alpha \in \Omega + K$. Therefore, $\alpha \in M^{-1}(\Omega+K)\cap\mathbb{Z}^s$, and so $\alpha \in K$ by (2.4). This shows that A maps $\ell(K)$ to $\ell(K)$. In other words, K is admissible for A.

From the above proof we see that a finite subset K of $\mathbb{Z}^s$ is admissible for A_ε if and only if

$$M^{-1}(\Omega - \varepsilon + K) \cap \mathbb{Z}^s \subseteq K. \tag{2.5}$$

The set $\Omega - E$ consists of all the points $\omega - \varepsilon$, where $\omega \in \Omega$ and $\varepsilon \in E$.

Now suppose G is a finite subset of $\mathbb{Z}^s$. Let $H := MG \cup (\Omega - E) \cup \{0\}$, and let

$$K := \left(\sum_{n=1}^{\infty} M^{-n}H\right) \cap \mathbb{Z}^s.$$

In other words, an element $\alpha \in \mathbb{Z}^s$ belongs to K if and only if $\alpha = \sum_{n=1}^{\infty} M^{-n}h_n$ for some sequence of elements $h_n \in H$. Since $0 \in H$ and $M^{-1}H \supseteq G$, we have

$$K \supseteq \mathbb{Z}^s \cap M^{-1}H \supseteq \mathbb{Z}^s \cap G = G.$$

Moreover,

$$\begin{aligned} M^{-1}(\Omega - \varepsilon + K) \cap \mathbb{Z}^s &\subseteq M^{-1}(H + K) \cap \mathbb{Z}^s \\ &= (M^{-1}H + M^{-1}K) \cap \mathbb{Z}^s \\ &\subseteq (M^{-1}H + M^{-2}H + \cdots) \cap \mathbb{Z}^s = K. \end{aligned}$$

Thus, K satisfies (2.5). Hence, K is admissible for all A_ε, $\varepsilon \in E$. □

LEMMA 2.4. *Let $\mathcal{A}$ be a finite collection of linear operators on a vector space V. Let ν be a vector in V, and let $V(\nu)$ be the minimal $\mathcal{A}$-invariant subspace generated by ν. If $V(\nu)$ is finite dimensional, then*

$$\lim_{n\to\infty} \|\mathcal{A}^n \nu\|_p^{1/n} = \rho_p\big(\mathcal{A}|_{V(\nu)}\big). \tag{2.6}$$

Proof. Let $\|\cdot\|$ be a vector norm on $V(\nu)$. Since $V(\nu)$ is finite dimensional, there exists a positive integer k such that $V(\nu)$ is spanned by the set

$$Y := \{A_1 \cdots A_j \nu : (A_1, \ldots, A_j) \in \mathcal{A}^j,\ j = 0, 1, \ldots, k\}.$$

Thus, there exists a constant $C_1 > 0$ such that $\|\mathcal{A}^n y\|_p \le C_1 \|\mathcal{A}^n \nu\|_p$ for all $y \in Y$ and all $n = 1, 2, \ldots$. Moreover, there exists a positive constant C_2 such that

$$\|\mathcal{A}^n|_{V(\nu)}\|_p \le C_2 \max_{y\in Y} \|\mathcal{A}^n y\|_p, \qquad n = 1, 2, \ldots .$$

Therefore, there exists a positive constant C such that for all $n = 1, 2, \ldots$,

$$\|\mathcal{A}^n|_{V(\nu)}\|_p \le C \|\mathcal{A}^n \nu\|_p .$$

But $\|\mathcal{A}^n \nu\|_p \le \|\mathcal{A}^n|_{V(\nu)}\|_p \|\nu\|$. This proves the desired relation (2.6). □

THEOREM 2.5. *Let S_a be the subdivision operator associated with a dilation matrix M and a mask a. Let $\mathcal{A} := \{A_\varepsilon : \varepsilon \in E\}$, where E is a complete set of representatives of the distinct cosets of the quotient group $\mathbb{Z}^s/M\mathbb{Z}^s$ and A_ε are the linear operators on $\ell_0(\mathbb{Z}^s)$ given by* (1.11). *Then for $\nu \in \ell_0(\mathbb{Z}^s)$,*

$$\lim_{n\to\infty} \|\tilde{\nu}(\tau) S_a^n \delta\|_p^{1/n} = \rho_p\big(\{A_\varepsilon|_{V(\nu)} : \varepsilon \in E\}\big), \tag{2.7}$$

where $V(\nu)$ is the minimal $\mathcal{A}$-invariant subspace generated by ν. Moreover, if W is the minimal $\mathcal{A}$-invariant subspace generated by a finite set Y, then

$$\rho_p\big(\{A_\varepsilon|_W : \varepsilon \in E\}\big) = \max_{\nu\in Y} \big\{ \lim_{n\to\infty} \|\tilde{\nu}(\tau) S_a^n \delta\|_p^{1/n} \big\}. \tag{2.8}$$

Proof. By Lemma 2.3, $V(\nu)$ is finite dimensional, and so the relevant joint spectral radius in (2.7) is well defined. By Lemma 2.2 we have

$$\|\tilde{\nu}(\tau) S_a^n \delta\|_p = \|\mathcal{A}^n \nu\|_p, \qquad 1 \le p \le \infty,\ n = 1, 2, \ldots .$$

Applying Lemma 2.4 to the present situation, we obtain (2.7).

For the second part of the theorem, we let W be the minimal $\mathcal{A}$-invariant subspace generated by a finite set Y, and observe that W is a finite sum of the linear subspaces $V(\nu)$, $\nu \in Y$. Hence

$$\rho_p\big(\{A_\varepsilon|_W : \varepsilon \in E\}\big) = \max_{\nu\in Y} \big\{ \rho_p\big(\{A_\varepsilon|_{V(\nu)} : \varepsilon \in E\}\big) \big\}.$$

This, together with (2.7), verifies (2.8). □

3. Convergence of subdivision schemes. In this section we characterize the L_p-convergence ($1 \le p \le \infty$) of a subdivision scheme in terms of the corresponding refinement mask.

In our study of convergence, the concept of stability plays an important role. The shifts of a function ϕ in $L_p(\mathbb{R}^s)$ are said to be *stable* if there are two positive constants C_1 and C_2 such that

$$C_1\|\lambda\|_p \le \Big\|\sum_{\alpha\in\mathbb{Z}^s} \lambda(\alpha)\phi(\cdot-\alpha)\Big\|_p \le C_2\|\lambda\|_p \qquad \forall\, \lambda \in \ell_0(\mathbb{Z}^s). \tag{3.1}$$

It was proved by Jia and Micchelli [8] that a compactly supported function $\phi \in L_p(\mathbb{R}^s)$ satisfies the L_p-stability condition in (3.1) if and only if, for any $\xi \in \mathbb{R}^s$, there exists an element $\beta \in \mathbb{Z}^s$ such that

$$\hat{\phi}(\xi + 2\pi\beta) \ne 0.$$

It is easily seen that the shifts of the function ϕ_0 given in (1.10) are stable.

First, we give a necessary condition for the subdivision scheme to converge.

THEOREM 3.1. *Let M be a dilation matrix with $m := |\det M|$, a an element in $\ell_0(\mathbb{Z}^s)$ with $\sum_{\alpha\in\mathbb{Z}^s} a(\alpha) = m$, and S_a the subdivision operator associated with M and a. If the subdivision scheme associated with M and a converges in the L_p-norm ($1 \le p \le \infty$), then for any vector $y \in \mathbb{Z}^s$,*

$$\lim_{n\to\infty} m^{-n/p}\|\nabla_y S_a^n \delta\|_p = 0. \tag{3.2}$$

Consequently, if the subdivision scheme associated with a converges in the L_p-norm, then

$$\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta) = 1 \qquad \forall\, \alpha \in \mathbb{Z}^s. \tag{3.3}$$

Proof. Suppose ϕ is a compactly supported function in $L_p(\mathbb{R}^s)$, ϕ satisfies the moment conditions of order 1, and the shifts of ϕ are stable. For $n = 0, 1, 2, \ldots$, let $a_n := S_a^n\delta$ and $f_n := T_a^n\phi$, where T_a is the operator given in (1.5). Then by (1.7) we have

$$f_n = \sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)\phi(M^n\cdot - \alpha).$$

Hence, for $y \in \mathbb{Z}^s$ we have

$$\begin{aligned} f_n - f_n(\cdot - M^{-n}y) &= \sum_{\alpha\in\mathbb{Z}^s} \big[a_n(\alpha) - a_n(\alpha - y)\big]\phi(M^n\cdot - \alpha) \\ &= \sum_{\alpha\in\mathbb{Z}^s} \nabla_y a_n(\alpha)\phi(M^n\cdot - \alpha). \end{aligned}$$

Since the shifts of ϕ are stable, there exists a constant $C > 0$ such that

$$m^{-n/p}\|\nabla_y a_n\|_p \le C\|f_n - f_n(\cdot - M^{-n}y)\|_p. \tag{3.4}$$

In particular, the above estimate is valid for $f_n = T_a^n\phi_0$, where ϕ_0 is the function given in (1.10). If the subdivision scheme converges in the L_p-norm, then there exists

a compactly supported function f in $L_p(\mathbb{R}^s)$ ($f \in C(\mathbb{R}^s)$ in the case $p = \infty$) such that $\|f_n - f\|_p \to 0$ as $n \to \infty$. Moreover, by the triangle inequality, we have

$$\|f_n - f_n(\cdot - M^{-n}y)\|_p \le \|f - f(\cdot - M^{-n}y)\|_p + 2\|f - f_n\|_p.$$

Hence, $\|f_n - f_n(\cdot - M^{-n}y)\|_p \to 0$ as $n \to \infty$. This, together with (3.4), verifies (3.2).

For the second part of the theorem, we observe that if the subdivision scheme converges in the L_p-norm for some p with $1 \le p \le \infty$, then it also converges in the L_1-norm. Thus, we only have to deal with the case $p = 1$.

Let E be a complete set of representatives of the distinct cosets of $\mathbb{Z}^s/M\mathbb{Z}^s$. Then $\#E = m$, and $\mathbb{Z}^s$ is the disjoint union of $\alpha + M\mathbb{Z}^s$, $\alpha \in E$. Since $\sum_{\alpha\in\mathbb{Z}^s} a(\alpha) = m$, we have

$$\sum_{\alpha\in E}\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta) = m.$$

Thus, (3.3) will be proved if we can show

$$\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta) = \sum_{\beta\in\mathbb{Z}^s} a(-M\beta) \qquad \forall\, \alpha \in E. \tag{3.5}$$

To this end, we deduce from $a_n = S_a a_{n-1}$ that

$$\sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha) = \sum_{\alpha\in\mathbb{Z}^s}\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta)a_{n-1}(\beta) = m\sum_{\beta\in\mathbb{Z}^s} a_{n-1}(\beta).$$

An induction argument gives $\sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha) = m^n$. Moreover,

$$\begin{aligned}\sum_{\beta\in\mathbb{Z}^s} a_n(\alpha - M\beta) &= \sum_{\beta\in\mathbb{Z}^s}\sum_{\gamma\in\mathbb{Z}^s} a(\alpha - M\beta - M\gamma)a_{n-1}(\gamma)\\ &= \sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta)\sum_{\gamma\in\mathbb{Z}^s} a_{n-1}(\gamma - \beta) = m^{n-1}\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta).\end{aligned}$$

Thus, we have

$$\sum_{\beta\in\mathbb{Z}^s}\big[a(\alpha - M\beta) - a(-M\beta)\big] = m^{-(n-1)}\sum_{\beta\in\mathbb{Z}^s}\big[a_n(\alpha - M\beta) - a_n(-M\beta)\big].$$

It follows that

$$\left|\sum_{\beta\in\mathbb{Z}^s}\big[a(\alpha - M\beta) - a(-M\beta)\big]\right| \le m^{-(n-1)}\|\nabla_\alpha a_n\|_1. \tag{3.6}$$

If the subdivision scheme is L_1-convergent, then by the first part of the theorem we have $m^{-(n-1)}\|\nabla_\alpha a_n\|_1 \to 0$ as $n \to \infty$. This, together with (3.6), implies (3.5), as desired. □

For the case $M = 2I$, it was proved by Cavaretta, Dahmen, and Micchelli [1] that the condition in (3.3) is necessary for the subdivision scheme to converge in the L_∞-norm.

The next theorem gives a characterization of convergence of the subdivision scheme.

THEOREM 3.2. *Let M be a dilation matrix with $m := |\det M|$, a an element in $\ell_0(\mathbb{Z}^s)$ such that $\sum_{\alpha\in\mathbb{Z}^s} a(\alpha) = m$, and S_a the corresponding subdivision operator. Then the subdivision scheme associated with M and a converges in the L_p-norm $(1 \le p \le \infty)$ if and only if*

$$\lim_{n\to\infty} \|\nabla_j S_a^n \delta\|_p^{1/n} < m^{1/p} \qquad \text{for } j = 1, \dots, s. \tag{3.7}$$

Proof. Let A_ε be the linear operators on $\ell_0(\mathbb{Z}^s)$ given by (2.1), and let V be the minimal common invariant subspace of A_ε $(\varepsilon \in E)$ generated by $\nabla_j\delta$, $j = 1, \dots, s$. Then V is finite dimensional, and by Theorem 2.5 we have

$$\rho_p := \rho_p\big(\{A_\varepsilon|_V : \varepsilon \in E\}\big) = \max_{1\le j\le s} \big\{ \lim_{n\to\infty} \|\nabla_j S_a^n \delta\|_p^{1/n} \big\}.$$

Thus, (3.7) is equivalent to $\rho_p\big(\{A_\varepsilon|_V : \varepsilon \in E\}\big) < m^{1/p}$.

Let $\mathcal{A} := \{A_\varepsilon|_V : \varepsilon \in E\}$. If $\rho_p(\{A_\varepsilon|_V : \varepsilon \in E\}) \ge m^{1/p}$, then we have

$$\inf_{n\ge1} \|\mathcal{A}^n\|_p^{1/n} = \lim_{n\to\infty} \|\mathcal{A}^n\|_p^{1/n} \ge m^{1/p}.$$

It follows that

$$m^{-n/p}\|\mathcal{A}^n\|_p \ge 1 \qquad \forall\, n.$$

From the proof of Lemma 2.4 we see that there exists a positive constant C such that $\|\mathcal{A}^n\|_p \le C \max_{1\le j\le s} \|\mathcal{A}^n \nabla_j \delta\|_p$ for all n. Moreover, by Lemma 2.2, we have $\|\mathcal{A}^n\nabla_j\delta\|_p = \|\nabla_j S_a^n\delta\|_p$. Hence,

$$\rho_p \ge m^{1/p} \quad \Longrightarrow \quad \max_{1\le j\le s} \big\{ m^{-n/p} \|\nabla_j S_a^n \delta\|_p \big\} \ge 1/C.$$

Thus, the subdivision scheme associated with a is not L_p-convergent, by Theorem 3.1. This shows that (3.7) is necessary for the subdivision scheme to converge in the L_p-norm.

In order to prove the sufficiency part of the theorem, we pick a compactly supported function ϕ in $L_p(\mathbb{R}^s)$ such that ϕ satisfies the moment conditions of order 1. (In the case $p = \infty$, we assume that ϕ is continuous.) Let $f_n := T_a^n\phi_0$ and $g_n := T_a^n\phi$ for $n = 1, 2, \dots$, where T_a is the operator given in (1.5) and ϕ_0 is the hat function given in (1.10). Moreover, let b_n be the sequence given by

$$b_n(\alpha) = \max_{1\le j\le s} |\nabla_j a_n(\alpha)|, \qquad \alpha \in \mathbb{Z}^s.$$

We claim that there exists a positive constant C independent of n such that

$$\|f_{n+1} - g_n\|_p \le C\, m^{-n/p} \|b_n\|_p, \qquad 1 \le p \le \infty. \tag{3.8}$$

If $\rho_p < m^{1/p}$, then we can find r, $0 < r < 1$, such that $\rho_p < r m^{1/p}$. By the definition of ρ_p and Theorem 2.5, we see that $\|b_n\|_p^{1/n} < r m^{1/p}$ is valid for sufficiently large n. Hence, there exists a positive constant C_0 such that $\|b_n\|_p \le C_0 (r m^{1/p})^n$ for all $n \ge 1$. If we choose ϕ to be ϕ_0, then it follows from (3.8) that

$$\|f_{n+1} - f_n\|_p \le C C_0 r^n, \qquad n = 1, 2, \dots.$$

This shows that the sequence f_n converges to a function f in the L_p-norm. Furthermore, (3.8) tells us that $\|f_{n+1} - g_n\|_p \to 0$ as $n \to \infty$. However,

$$\|g_n - f\|_p \le \|g_n - f_{n+1}\|_p + \|f_{n+1} - f\|_p$$

by the triangle inequality. Therefore, $\|g_n - f\|_p \to 0$ as $n \to \infty$. Thus, it suffices to prove (3.8). For this purpose, we shall follow the lines of Jia and Lei [7].

In what follows, by $\operatorname{supp}\phi$ we denote the support of ϕ, and by $\operatorname{supp} a$ we denote the set $\{\alpha \in \mathbb{Z}^s : a(\alpha) \neq 0\}$. For a sequence $\lambda \in \ell_\infty(\mathbb{Z}^s)$ and a subset G of $\mathbb{R}^s$, we use $\|\lambda\|_\infty(G)$ to denote the supremum of λ on the set $\mathbb{Z}^s \cap G$. Moreover, for $n = 1, 2, \ldots$, let

$$X_n(\gamma) := M^{-n}\big([0,1)^s + \gamma\big), \qquad \gamma \in \mathbb{Z}^s.$$

Let x be a point in $\mathbb{R}^s$. By (1.7) we have

$$f_{n+1}(x) = \sum_{\beta\in\mathbb{Z}^s} a_{n+1}(\beta)\phi_0(M^{n+1}x - \beta) \qquad \text{and} \qquad g_n(x) = \sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)\phi(M^n x - \alpha).$$

Since $\sum_{\beta\in\mathbb{Z}^s} \phi_0(\cdot - \beta) = 1$ and $\sum_{\alpha\in\mathbb{Z}^s} \phi(\cdot - \alpha) = 1$, it follows that

$$f_{n+1}(x) - g_n(x) = \sum_{\alpha\in\mathbb{Z}^s}\sum_{\beta\in\mathbb{Z}^s} \big[a_{n+1}(\beta) - a_n(\alpha)\big]\phi(M^n x - \alpha)\phi_0(M^{n+1}x - \beta). \tag{3.9}$$

In the above sum we only have to consider those terms for which $\phi(M^n x - \alpha) \neq 0$ and $\phi_0(M^{n+1}x - \beta) \neq 0$.

Let $x \in X_n(\gamma)$, where $\gamma \in \mathbb{Z}^s$ is fixed for the time being. Suppose $\phi(M^n x - \alpha) \neq 0$. Then we have $M^n x - \gamma \in [0,1)^s$ and $M^n x - \alpha \in \operatorname{supp}\phi$. It follows that

$$\alpha = \gamma + (M^n x - \gamma) - (M^n x - \alpha) \in \gamma + [0,1)^s - \operatorname{supp}\phi. \tag{3.10}$$

Suppose $\phi_0(M^{n+1}x - \beta) \neq 0$. Then $M^{n+1}x - \beta \in \operatorname{supp}\phi_0$. This, in connection with $M^n x - \alpha \in \operatorname{supp}\phi$, yields

$$M\alpha - \beta = (M^{n+1}x - \beta) - M(M^n x - \alpha) \in \operatorname{supp}\phi_0 - M\operatorname{supp}\phi. \tag{3.11}$$

Moreover, Theorem 3.1 tells us that (3.7) implies $\sum_{\eta\in\mathbb{Z}^s} a(\beta - M\eta) = 1$ for all $\beta \in \mathbb{Z}^s$; hence, we have

$$a_{n+1}(\beta) - a_n(\alpha) = \sum_{\eta\in\mathbb{Z}^s} a(\beta - M\eta)[a_n(\eta) - a_n(\alpha)]. \tag{3.12}$$

We observe that $a(\beta - M\eta) \neq 0$ implies $\beta - M\eta \in \operatorname{supp} a$. This, together with (3.11), gives

$$M(\alpha - \eta) = (M\alpha - \beta) + (\beta - M\eta) \in \operatorname{supp}\phi_0 - M\operatorname{supp}\phi + \operatorname{supp} a. \tag{3.13}$$

In light of (3.10) and (3.13), there exists a positive integer N such that both α and η belong to $\gamma + [-N, N]^s$, provided $\phi(M^n x - \alpha) \neq 0$, $\phi_0(M^{n+1}x - \beta) \neq 0$, and $a(\beta - M\eta) \neq 0$. However, $a_n(\eta) - a_n(\alpha)$ can be written as a sum of finitely many terms of the form $\nabla_j a_n(\nu)$, where $\nu \in \gamma + [-N, N]^s \cap \mathbb{Z}^s$ and $j = 1, \ldots, s$. Therefore, (3.12) tells us that there exists a positive constant C independent of n such that

$$|a_{n+1}(\beta) - a_n(\alpha)| \le C\|b_n\|_\infty(\gamma + [-N, N]^s), \tag{3.14}$$

provided $\phi(M^n x - \alpha)\phi_0(M^{n+1}x - \beta) \neq 0$ for some $x \in X_n(\gamma)$.

We observe that $\sum_{\beta\in\mathbb{Z}^s}|\phi_0(M^{n+1}x-\beta)|=1$. Consequently, by (3.9) and (3.14) we obtain

$$|f_{n+1}(x)-g_n(x)|\le C|\phi|^\circ(M^nx)\|b_n\|_\infty(\gamma+[-N,N]^s)\qquad\text{for } x\in X_n(\gamma), \tag{3.15}$$

where $|\phi|^\circ$ denotes the 1-periodization of $|\phi|$:

$$|\phi|^\circ(x):=\sum\nolimits_{\alpha\in\mathbb{Z}^s}|\phi(x-\alpha)|,\qquad x\in\mathbb{R}^s.$$

In the case $p=\infty$, ϕ is a continuous function with compact support; hence, there exists a constant $C_1>0$ such that $|\phi|^\circ(x)\le C_1$ for all $x\in\mathbb{R}^s$. It follows from (3.15) that $\|f_{n+1}-g_n\|_\infty\le CC_1\|b_n\|_\infty$. This proves (3.8) for the case $p=\infty$.

For $1\le p<\infty$, we deduce from (3.15) that

$$\int_{X_n(\gamma)}|f_{n+1}(x)-g_n(x)|^p\,dx\le C^p\biggl(\int_{X_n(\gamma)}\bigl[|\phi|^\circ(M^nx)\bigr]^p\,dx\biggr)\sum_{\alpha\in\gamma+[-N,N]^s}|b_n(\alpha)|^p.$$

Since $\phi\in L_p(\mathbb{R}^s)$ is compactly supported, we have

$$|\phi|^\circ(x)=\sum_{\alpha\in\mathbb{Z}^s\cap([0,1)^s-\operatorname{supp}\phi)}|\phi(x-\alpha)|,\qquad x\in[0,1)^s.$$

Hence, $C_2:=\int_{[0,1)^s}|\phi|^\circ(x)^p\,dx<\infty$. Consequently,

$$\int_{X_n(\gamma)}\bigl[|\phi|^\circ(M^nx)\bigr]^p\,dx=m^{-n}\int_{\gamma+[0,1)^s}\bigl[|\phi|^\circ(x)\bigr]^p\,dx=C_2m^{-n}.$$

Finally, we obtain

$$\begin{aligned}\|f_{n+1}-g_n\|_p^p&=\sum_{\gamma\in\mathbb{Z}^s}\int_{X_n(\gamma)}|f_{n+1}(x)-g_n(x)|^p\,dx\\&\le C^pC_2m^{-n}\sum_{\gamma\in\mathbb{Z}^s}\sum_{\alpha\in\gamma+[-N,N]^s}|b_n(\alpha)|^p.\end{aligned}$$

However,

$$\sum_{\gamma\in\mathbb{Z}^s}\sum_{\alpha\in\gamma+[-N,N]^s}|b_n(\alpha)|^p=\sum_{\alpha\in\mathbb{Z}^s}|b_n(\alpha)|^p\sum_{\gamma\in\alpha+[-N,N]^s}1=(2N+1)^s\sum_{\alpha\in\mathbb{Z}^s}|b_n(\alpha)|^p.$$

The preceding discussion tells us that

$$\|f_{n+1}-g_n\|_p\le C_3m^{-n/p}\|b_n\|_p$$

for some constant $C_3>0$. The proof is complete. □

Suppose K is an admissible set for every A_ε, $\varepsilon\in E$, and $\ell(K)$ contains $\nabla_j\delta$ for $j=1,\dots,s$. Let

$$V:=\Bigl\{v\in\ell(K):\sum_{\alpha\in\mathbb{Z}^s}v(\alpha)=0\Bigr\}.$$

If $\sum_{\beta\in\mathbb{Z}^s}a(\alpha-M\beta)=1$, then V is invariant under every A_ε, $\varepsilon\in E$. Thus, we may restate Theorem 3.2 as follows.

THEOREM 3.3. *Under the conditions of Theorem* 3.2*, the subdivision scheme associated with* a *converges in the* L_p*-norm* $(1 \le p \le \infty)$ *if and only if the following two conditions are satisfied:*

(a) $\sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta) = 1$ *for all* $\alpha \in \mathbb{Z}^s$*;*

(b) $\rho_p(\{A_\varepsilon|_V : \varepsilon \in E\}) < m^{1/p}$.

Proof. For $j = 1, \dots, s$, $\nabla_j\delta \in V$. Conversely, V is spanned by vectors of the form $\tau^\beta\nabla_j\delta$, where $\beta \in \mathbb{Z}^s$, $j = 1, \dots, s$. Let $\mathcal{A} := \{A_\varepsilon|_V : \varepsilon \in E\}$. By Lemma 2.2 we have

$$\|\mathcal{A}^n\tau^\beta\nabla_j\delta\|_p = \|\tau^\beta\nabla_j S_a^n\delta\|_p = \|\nabla_j S_a^n\delta\|_p.$$

This shows that

$$\rho_p(\{A_\varepsilon|_V : \varepsilon \in E\}) = \max_{1\le j\le s}\big\{\lim_{n\to\infty}\|\nabla_j S_a^n\delta\|_p^{1/n}\big\}.$$

Thus, Theorem 3.3 follows from Theorem 3.2 at once. □

After a closer examination of the proof of Theorems 3.1 and 3.2, we obtain the following result.

THEOREM 3.4. *Let* a *be a finitely supported sequence on* $\mathbb{Z}^s$ *satisfying* (1.2) *and let* $T = T_a$ *be the linear operator given by* (1.5)*. Suppose* u *is a compactly supported function in* $L_p(\mathbb{R}^s)$ $(1 \le p \le \infty)$*,* u *satisfies the moment conditions of order* 1*, and the shifts of* u *are stable. If there exists a function* $f \in L_p(\mathbb{R}^s)$ (*a continuous function* f *in the case* $p = \infty$) *such that*

$$\lim_{n\to\infty}\|T^n u - f\|_p = 0, \tag{3.16}$$

then for any compactly supported function $v \in L_p(\mathbb{R}^s)$ *satisfying the moment conditions of order* 1 *we also have*

$$\lim_{n\to\infty}\|T^n v - f\|_p = 0. \tag{3.17}$$

Consequently, if the normalized solution f *of* (1.1) *lies in* $L_p(\mathbb{R}^s)$ (f *is a continuous function in the case* $p = \infty$)*, and if the shifts of* f *are stable, then the subdivision scheme associated with mask* a *converges to* f *in the* L_p*-norm.*

Proof. Suppose (3.16) is true for a function u that satisfies the moment conditions of order 1 and has stable shifts. Then the proof of Theorem 3.1 tells us that (3.2) is valid for every vector $y \in \mathbb{Z}^s$. Therefore, from the proof of Theorem 3.2 we see that (3.17) holds true for any compactly supported function $v \in L_p(\mathbb{R}^s)$ satisfying the moment conditions of order 1. In particular, if f itself has stable shifts, then we may choose u to be f in (3.16). Thus, in such a case, the subdivision scheme converges in the L_p-norm. □

We end this section by two examples which illustrate the general theory developed so far.

Example 3.5. Let $M = 2I$, where I is the 2×2 identity matrix. Consider the refinement equation

$$f = \sum_{\alpha\in\mathbb{Z}^2} a(\alpha) f(2\cdot - \alpha), \tag{3.18}$$

where the mask a is given by its symbol

$$\tilde{a}(z) = \frac{1}{4}\,z_1^{-1} + 1 + \frac{3}{4}\,z_1 + \frac{3}{4}\,z_1^{-1}z_2 + z_2 + \frac{1}{4}\,z_1 z_2, \qquad z = (z_1, z_2) \in \mathbb{T}^2.$$

We claim that the subdivision scheme associated with a is convergent in the L_p-norm for $1 \le p < \infty$, but it is not L_∞-convergent.

For $\varepsilon \in E := \{(0,0),(1,0),(0,1),(1,1)\}$, let A_ε be the operator on $\ell_0(\mathbb{Z}^2)$ given by

$$A_\varepsilon v(\alpha) = \sum_{\beta\in\mathbb{Z}^2} a(\varepsilon + 2\alpha - \beta)v(\beta), \qquad \alpha \in \mathbb{Z}^2,\ v \in \ell_0(\mathbb{Z}^2).$$

Let K be the set consisting of the points $(-1,0),(0,0),(1,0),(-1,1),(0,1),(1,1)$. Then K is admissible for A_ε, $\varepsilon \in E$. Let V be the linear space

$$\Big\{v \in \ell(K) : \sum_{\alpha\in\mathbb{Z}^2} v(\alpha) = 0\Big\}.$$

Then V is the minimal common invariant space of A_ε ($\varepsilon \in E$) generated by $\nabla_j\delta$, $j = 1, 2$. The dimension of V is 5. We choose a basis for V as follows:

$$v_1 = \delta - \delta_{(1,0)},\ v_2 = \delta - \delta_{(-1,0)},\ v_3 = \delta_{(0,1)} - \delta_{(1,1)},$$
$$v_4 = \delta_{(0,1)} - \delta_{(-1,1)}, \quad \text{and} \quad v_5 = r(\delta - \delta_{(0,1)}),$$

where $v = 1$ for $p = \infty$, and r is a number such that $0 < r < (3/2)^{1/p} - 1$ for $1 \le p < \infty$.

By computation, the matrix representations of $A_\varepsilon|_V$ ($\varepsilon \in E$) under this basis are given by

$$A_{(0,0)}|_V = \begin{bmatrix} 3/4 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 3/4 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad A_{(1,0)}|_V = \begin{bmatrix} 0 & -1/4 & 0 & 0 & 0 \\ 0 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3/4 & 0 \\ 0 & 0 & 0 & 1/4 & 0 \\ 0 & -r/4 & 0 & 3r/4 & 1 \end{bmatrix},$$

and

$$A_{(0,1)}|_V = \begin{bmatrix} 1/4 & 0 & 0 & 0 & 0 \\ 0 & 3/4 & 0 & 0 & 0 \\ 3/4 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad A_{(1,1)}|_V = \begin{bmatrix} 0 & -3/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 & 0 \\ 0 & -1/4 & 0 & 0 & 0 \\ 0 & 3/4 & 0 & 0 & 0 \\ 0 & -r/2 & 0 & 0 & 0 \end{bmatrix}.$$

Since $A_{(0,0)}|_V$ has an eigenvalue 1, we have

$$\rho_\infty\big(\{A_\varepsilon|_V : \varepsilon \in E\}\big) \ge 1.$$

Therefore, the subdivision scheme is not L_∞-convergent.

For the case $1 \le p < \infty$, we choose the maximum row sum norm as the matrix norm. Since $0 < r < (3/2)^{1/p} - 1$, we have

$$\sum_{\varepsilon\in E} \|A_\varepsilon|_V\|^p \le 1 + (1+r)^p + (3/4)^p + (3/4)^p < 4.$$

This shows that

$$\rho_p\big(\{A_\varepsilon|_V : \varepsilon \in E\}\big) < 4^{1/p}.$$

By Theorem 3.3, the subdivision scheme is L_p-convergent for $1 \le p < \infty$.

Example 3.6. Let

$$M = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

Then M is a dilation matrix with $\det M = 2$. Let a be the sequence on $\mathbb{Z}^2$ given by its symbol

$$\tilde{a}(z) = 1 + tz_1 + (1-t)z_2, \qquad z = (z_1, z_2) \in \mathbb{T}^2,$$

where t is a real number. We claim that the subdivision scheme associated with a and M is L_∞-convergent if and only if $0 < t < 1$.

Let A_0 and A_1 be the linear operators on $\ell_0(\mathbb{Z}^2)$ given by

$$A_0 v(\alpha) = \sum_{\beta \in \mathbb{Z}^2} a(M\alpha - \beta) v(\beta) \quad \text{and} \quad A_1 v(\alpha) = \sum_{\beta \in \mathbb{Z}^2} a((1,0) + M\alpha - \beta) v(\beta),$$

where $\alpha \in \mathbb{Z}^2$ and $v \in \ell_0(\mathbb{Z}^2)$. Let K be the set

$$[-1,2] \times [-2,2] \cap \mathbb{Z}^2 \setminus \{(-1,-2), (2,-2), (2,2), (-1,2)\}.$$

Then K is admissible for both A_0 and A_1. Let V be the linear space

$$\Big\{ v \in \ell(K) : \sum_{\alpha \in \mathbb{Z}^2} v(\alpha) = 0 \Big\}.$$

Then V is the minimal common invariant subspace of A_0 and A_1 generated by $\nabla_j \delta$, $j = 1, 2$. The dimension of V is 15. By Theorem 3.3, the subdivision scheme associated with a and M is L_∞-convergent if and only if $\rho_\infty(A_0|_V, A_1|_V) < 1$.

We observe that

$$A_0(\delta - \delta_{(0,-1)}) = t(\delta - \delta_{(0,-1)}) \quad \text{and} \quad A_0(\delta_{(0,1)} - \delta_{(1,1)}) = -(1-t)(\delta_{(0,1)} - \delta_{(1,1)}).$$

Hence, both t and $-(1-t)$ are eigenvalues of A_0. Consequently,

$$\rho_\infty(A_0|_V, A_1|_V) \geq \max\{|t|, |1-t|\}.$$

Therefore, for $t \leq 0$ or $t \geq 1$, the subdivision scheme associated with a and M is *not* convergent in the L_∞-norm.

Now assume that $0 < t < 1$. We wish to show $\rho_\infty(A_0|_V, A_1|_V) < 1$ in this case. Let

$$W := \big\{ \delta_\alpha - \delta_\beta : \alpha, \beta \in K,\ |\alpha - \beta| \leq 1 \big\},$$

where $|\alpha| := \max\{|\alpha_1|, |\alpha_2|\}$ for $\alpha = (\alpha_1, \alpha_2) \in \mathbb{Z}^2$. Then W is a spanning set for V. Any vector $v \in V$ can be represented as $\sum_{w \in W} c_w w$, where the coefficients c_w ($w \in W$) are not uniquely determined. Define

$$\|v\| := \min\Big\{ \sum_{w \in W} |c_w| : v = \sum_{w \in W} c_w w. \Big\}$$

Clearly, $\|\cdot\|$ is a norm on V.

We observe that

$$A_0(\delta-\delta_{(-1,0)})=(1-t)(\delta-\delta_{(0,1)}),\qquad A_0(\delta-\delta_{(0,-1)})=t(\delta-\delta_{(0,-1)})$$

and

$$A_1(\delta-\delta_{(1,0)})=-(1-t)(\delta-\delta_{(0,1)}),\qquad A_1(\delta-\delta_{(0,1)})=t(\delta-\delta_{(0,1)}).$$

With the help of these relations, we can verify through a simple but tedious computation that

$$\big\|A_{\varepsilon_1}A_{\varepsilon_2}A_{\varepsilon_3}A_{\varepsilon_4}A_{\varepsilon_5}w\big\|<1$$

for all $w\in W$ and all $\varepsilon_j\in\{0,1\}$, $j=1,\dots,5$. Thus, there exists a positive number $\sigma<1$ such that

$$\big\|A_{\varepsilon_1}A_{\varepsilon_2}A_{\varepsilon_3}A_{\varepsilon_4}A_{\varepsilon_5}w\big\|\le\sigma$$

for all $w\in W$ and all $\varepsilon_j\in\{0,1\}$, $j=1,\dots,5$. Let v be a unit vector in V. Choose a representation $\sum_{w\in W}c_ww$ for v such that $\sum_{w\in W}|c_w|=\|v\|=1$. Then

$$\big\|A_{\varepsilon_1}A_{\varepsilon_2}A_{\varepsilon_3}A_{\varepsilon_4}A_{\varepsilon_5}v\big\|\le\sum_{w\in W}|c_w|\big\|A_{\varepsilon_1}A_{\varepsilon_2}A_{\varepsilon_3}A_{\varepsilon_4}A_{\varepsilon_5}w\big\|\le\sigma<1.$$

This shows that $\|A_{\varepsilon_1}A_{\varepsilon_2}A_{\varepsilon_3}A_{\varepsilon_4}A_{\varepsilon_5}\|\le\sigma<1$ for all $\varepsilon_j\in\{0,1\}$, $j=1,\dots,5$. Therefore, $\rho_\infty(A_0|_V,A_1|_V)<1$ and the subdivision scheme associated with a and M is L_∞-convergent if $0<t<1$.

4. L_2-convergence. In general, the p-norm joint spectral radius is difficult to compute. However, the 2-norm joint spectral radius can be easily computed by calculating the spectral radius of a certain finite matrix.

Given $a\in\ell_0(\mathbb{Z}^s)$, the symbol $\tilde a(z)$ is well defined on the s-torus

$$\mathbb{T}^s:=\{(z_1,\dots,z_s)\in\mathbb{C}^s:|z_1|=\cdots=|z_s|=1\}.$$

For $a,b\in\ell_0(\mathbb{Z}^s)$, the discrete convolution of a and b, denoted $a*b$, is given by

$$a*b(\alpha):=\sum_{\beta\in\mathbb{Z}^s}a(\alpha-\beta)b(\beta),\qquad\beta\in\mathbb{Z}^s.$$

It is easily seen that

$$\widetilde{a*b}(z)=\tilde a(z)\tilde b(z),\qquad z\in(\mathbb{C}\setminus\{0\})^s.$$

For $z\in\mathbb{C}$, we use $\overline{z}$ to denote the complex conjugate of z. Note that for $z\in\mathbb{T}^s$ and $\alpha\in\mathbb{Z}^s$, we have $\overline{z^\alpha}=z^{-\alpha}$. For $a\in\ell_0(\mathbb{Z}^s)$, we denote by a^* the sequence given by $a^*(\alpha):=\overline{a(-\alpha)}$, $\alpha\in\mathbb{Z}^s$. Then for $z\in\mathbb{T}^s$ we have

$$\widetilde{a^*}(z)=\sum_{\alpha\in\mathbb{Z}^s}\overline{a(-\alpha)}z^\alpha=\sum_{\alpha\in\mathbb{Z}^s}\overline{a(-\alpha)z^{-\alpha}}=\overline{\tilde a(z)}.$$

If $b=a*a^*$, then we have

$$\tilde b(z)=\tilde a(z)\widetilde{a^*}(z)=|\tilde a(z)|^2\qquad\text{for }z\in\mathbb{T}^s.$$

THEOREM 4.1. *Let M be an $s \times s$ dilation matrix. For $a \in \ell_0(\mathbb{Z}^s)$, let $b := a*a^*$ and denote by S_a and S_b the subdivision operators associated with a and b, respectively. Then for $\nu \in \ell_0(\mathbb{Z}^s)$,*

$$\lim_{n\to\infty} \|\tilde{\nu}(\tau) S_a^n \delta\|_2^{1/n} = \sqrt{\rho(B|_W)}$$

and

$$\lim_{n\to\infty} \|\tilde{\mu}(\tau) S_b^n \delta\|_\infty^{1/n} = \rho(B|_W),$$

where $\mu := \nu*\nu^*$, B is the linear operator on $\ell_0(\mathbb{Z}^s)$ given by

$$Bw(\alpha) = \sum_{\beta\in\mathbb{Z}^s} b(M\alpha - \beta) w(\beta), \qquad \alpha \in \mathbb{Z}^s,\ w \in \ell_0(\mathbb{Z}^s), \tag{4.1}$$

and W is the minimal B-invariant subspace generated by μ.

Proof. For $n = 1, 2, \ldots$, write a_n for $S_a^n \delta$ and b_n for $S_b^n \delta$. Note that the symbol of $\tilde{\nu}(\tau) a_n$ is $\tilde{\nu}(z)\tilde{a}_n(z)$, and the symbol of $\tilde{\mu}(\tau) b_n$ is $\tilde{\mu}(z)\tilde{b}_n(z)$. By the Parseval identity we have

$$\begin{aligned} \|\tilde{\nu}(\tau) a_n\|_2^2 &= \sum_{\alpha\in\mathbb{Z}^s} \left|\tilde{\nu}(\tau) a_n(\alpha)\right|^2 \\ &= \frac{1}{(2\pi)^s} \int_{[0,2\pi)^s} \left|\tilde{\nu}(e^{i\xi})\tilde{a}_n(e^{i\xi})\right|^2 d\xi = \frac{1}{(2\pi)^s} \int_{[0,2\pi)^s} \tilde{\mu}(e^{i\xi})\tilde{b}_n(e^{i\xi})\, d\xi. \end{aligned}$$

Since $\tilde{\mu}(e^{i\xi})\tilde{b}_n(e^{i\xi}) \ge 0$ for all $\xi \in \mathbb{R}^s$, it follows that

$$\begin{aligned} \tilde{\mu}(\tau) b_n(0) \le \|\tilde{\mu}(\tau) b_n\|_\infty &\le \frac{1}{(2\pi)^s} \int_{[0,2\pi)^s} \left|\tilde{\mu}(e^{i\xi})\tilde{b}_n(e^{i\xi})\right| d\xi \\ &= \frac{1}{(2\pi)^s} \int_{[0,2\pi)^s} \tilde{\mu}(e^{i\xi})\tilde{b}_n(e^{i\xi})\, d\xi = \tilde{\mu}(\tau) b_n(0). \end{aligned}$$

From the proof of Lemma 2.2 we see that $\tilde{\mu}(\tau) b_n(0) = B^n \mu(0)$. Hence,

$$\|\tilde{\nu}(\tau) S_a^n \delta\|_2^2 = \|\tilde{\mu}(\tau) S_b^n \delta\|_\infty = B^n \mu(0).$$

It follows that

$$\lim_{n\to\infty} \|\tilde{\mu}(\tau) S_b^n \delta\|_\infty^{1/n} \le \lim_{n\to\infty} |B^n \mu(0)|^{1/n} \le \rho(B|_W).$$

Moreover, since W is the minimal B-invariant subspace generated by μ, Theorem 2.5 tells us that

$$\rho(B|_W) \le \lim_{n\to\infty} \|\tilde{\mu}(\tau) S_b^n \delta\|_\infty^{1/n}.$$

This completes the proof. □

We remark that Goodman, Micchelli, and Ward in [5] established a result similar to Theorem 4.1 for the special case $\nu = \delta$.

The following theorem discusses the relationship among the spectra of B when it is restricted to different invariant subspaces.

THEOREM 4.2. *Let M be an $s \times s$ dilation matrix. For an element $b \in \ell_0(\mathbb{Z}^s)$, let B be the linear operator on $\ell_0(\mathbb{Z}^s)$ given by (4.1). Suppose Ω is the support of b. Then the set K_0 given by*

$$K_0 := \Big(\sum_{n=1}^{\infty} M^{-n}\Omega\Big) \cap \mathbb{Z}^s \tag{4.2}$$

is admissible for B. Moreover, if W is a finite dimensional B-invariant subspace, then the eigenvalues of $B|_{W\cap\ell(K_0)}$ are also eigenvalues of $B|_W$, and all the other eigenvalues of $B|_W$ are 0.

Proof. Let K_0 be the set given in (4.2). Then

$$M^{-1}(\Omega + K_0) \subseteq \sum_{n=1}^{\infty} M^{-n}\Omega.$$

This shows that K_0 satisfies (2.4). Hence, K_0 is an admissible set for B, by Lemma 2.3. Since $\ell(K_0)$ is an invariant subspace of B, the eigenvalues of $B|_{W\cap\ell(K_0)}$ also are eigenvalues of $B|_W$.

Let K be an admissible set for B such that $\ell(K) \supseteq W$. In order to prove that all the other eigenvalues of $B|_W$ are 0, it suffices to show that there exists a positive integer N such that

$$B^N\lambda \in \ell(K_0) \qquad \forall\, \lambda \in \ell(K). \tag{4.3}$$

To see this, suppose (4.3) is valid and σ is an eigenvalue of $B|_W$ with an eigenvector $\lambda \in W \setminus \ell(K_0)$. Then by (4.3) we have $\sigma^N\lambda = B^N\lambda \in \ell(K_0)$. However, $\lambda \notin \ell(K_0)$. Hence, this happens only if $\sigma = 0$. Thus, it remains to prove (4.3). For this purpose, it suffices to prove that for each $\beta \in K \setminus K_0$, there exists a positive integer N such that $B^N\delta_\beta \in \ell(K_0)$.

Let j be a positive integer. For $\lambda \in \ell(K)$, we have

$$B^j\lambda(\alpha) = \sum_{\gamma\in\mathbb{Z}^s} b(M\alpha - \gamma)B^{j-1}\lambda(\gamma).$$

Hence, $B^j\lambda(\alpha) \neq 0$ only if $M\alpha - \gamma \in \Omega$ for some $\gamma \in \mathbb{Z}^s$ with $B^{j-1}\lambda(\gamma) \neq 0$. Let n be a positive integer and let $\alpha, \beta \in \mathbb{Z}^s$. Then $B^n\delta_\beta(\alpha) \neq 0$ holds true only if there exist $\alpha_0, \alpha_1, \dots, \alpha_n \in \mathbb{Z}^s$ such that $\alpha_0 = \beta$, $\alpha_n = \alpha$, and

$$M\alpha_j - \alpha_{j-1} \in \Omega \qquad \text{for } j = 1, \dots, n.$$

Hence, $B^n\delta_\beta(\alpha) \neq 0$ implies

$$\alpha \in M^{-1}\Omega + M^{-2}\Omega + \cdots + M^{-n}\Omega + M^{-n}K =: \Gamma_n.$$

Let $\Gamma := \sum_{n=1}^{\infty} M^{-n}\Omega$. Then $K_0 = \mathbb{Z}^s \cap \Gamma$ and $(\mathbb{Z}^s \setminus K_0) \cap \Gamma = \emptyset$. We shall show that Γ is a compact set. Let H be an infinite subset of Γ. Note that Ω is a finite set. By induction on n we can find a sequence of elements $\omega_n \in \Omega$ $(n = 1, 2, \dots)$ such that

$$(M^{-1}\omega_1 + \cdots + M^{-n}\omega_n + M^{-n-1}\Omega) \cap H$$

is an infinite set. Then the element $\gamma := \sum_{n=1}^{\infty} M^{-n}\omega_n$ is a limit point of H. Since $\mathbb{Z}^s \setminus K_0$ is closed and Γ is compact, $\eta := \operatorname{dist}(\mathbb{Z}^s \setminus K_0, \Gamma)$, the distance between two

sets $\mathbb{Z}^s \setminus K_0$ and Γ, is positive. Note that $M^{-n} \to 0$ as $n \to \infty$. Thus, there exists a positive integer N such that

$$B^N \delta_\beta(\alpha) \neq 0 \quad \Longrightarrow \quad \operatorname{dist}(\alpha, \Gamma) < \eta.$$

From $\operatorname{dist}(\alpha, \Gamma) < \eta$ and $\alpha \in \mathbb{Z}^s$ we deduce that $\alpha \in K_0$. This shows $B^N \delta_\beta \in \ell(K_0)$, as desired. □

The L_2-convergence of a subdivision scheme can be determined by using Theorems 3.2 and 4.1. The following theorem gives another form of characterization for the L_2-convergence.

THEOREM 4.3. *Let M be an $s \times s$ dilation matrix with $m := |\det M|$. For $a \in \ell_0(\mathbb{Z}^s)$, let $b := a*a^*/m$ and let B be the linear operator on $\ell_0(\mathbb{Z}^s)$ given by*

$$Bw(\alpha) = \sum_{\beta \in \mathbb{Z}^s} b(M\alpha - \beta) w(\beta), \qquad \alpha \in \mathbb{Z}^s,\ w \in \ell_0(\mathbb{Z}^s).$$

Denote by K_0 the set $\mathbb{Z}^s \cap \sum_{n=1}^\infty M^{-n}\Omega$, where Ω is the support of b. Let V be the linear space

$$\Big\{ w \in \ell(K_0) : \sum_{\alpha \in \mathbb{Z}^s} w(\alpha) = 0 \Big\}.$$

Then the subdivision scheme associated with a converges in the L_2-norm if and only if the following two conditions are satisfied:

(a) $\sum_{\beta \in \mathbb{Z}^s} a(\alpha - M\beta) = 1$ *for all* $\alpha \in \mathbb{Z}^s$;

(b) $\rho(B|_V) < 1$.

Proof. First, assuming that conditions (a) and (b) are satisfied, we shall prove that the subdivision scheme associated with a converges in the L_2-norm. Let W be the minimal B-invariant subspace generated by $-\delta_{e_j} + 2\delta - \delta_{e_j}$, $j = 1, \ldots, s$. By Theorem 4.2,

$$\rho(B|_W) = \rho(B|_{W \cap \ell(K_0)}).$$

It follows from condition (a) that $\sum_{\beta \in \mathbb{Z}^s} b(\alpha - M\beta) = 1$ for all $\alpha \in \mathbb{Z}^s$. Consequently, if w is an element in $\ell_0(\mathbb{Z}^s)$ such that $\sum_{\alpha \in \mathbb{Z}^s} w(\alpha) = 0$, then $\sum_{\alpha \in \mathbb{Z}^s} Bw(\alpha) = 0$. This shows $W \cap \ell(K_0) \subseteq V$. Hence, $\rho(B|_W) \le \rho(B|_V)$. By Theorem 4.1 we have that, for $j = 1, \ldots, s$,

$$\lim_{n \to \infty} \|\nabla_j S_a^n \delta\|_2^{1/n} \le \sqrt{m\rho(B|_W)} \le \sqrt{m\rho(B|_V)} < \sqrt{m}.$$

By Theorem 3.2 we conclude that the subdivision scheme associated with a converges in the L_2-norm.

Next, suppose that the subdivision scheme associated with a converges in the L_2-norm. By Theorem 3.1, condition (a) is satisfied. It remains to prove $\rho(B|_V) < 1$. Let ϕ_0 be the function given in (1.10) and let $f_n := T_a^n \phi_0$, where T_a is the operator given in (1.5). Then there exists a function $f \in L_2(\mathbb{R}^s)$ such that $\|f_n - f\|_2 \to 0$ as $n \to \infty$. For a function f defined on $\mathbb{R}^s$, let f^* be the function given by $f^*(x) = \overline{f(-x)}$ for $x \in \mathbb{R}^s$. Let $\phi := \phi_0 * \phi_0^*$ be the convolution of ϕ_0 and ϕ_0^*. Similarly, let $g_n := f_n * f_n^*$ and $g = f * f^*$. It is easily seen that $g_n = T_b^n \phi$, where T_b is the operator given by $T_b\phi = \sum_{\alpha \in \mathbb{Z}^s} b(\alpha)\phi(M \cdot - \alpha)$. Then we have

$$\begin{aligned} \|g_n - g\|_\infty &= \|f_n * f_n^* - f * f^*\|_\infty \\ &\le \|f_n * (f_n^* - f^*)\|_\infty + \|(f_n - f) * f^*\|_\infty \\ &\le (\|f_n\|_2 + \|f\|_2)\|f_n - f\|_2. \end{aligned}$$

Note that ϕ is a continuous function, ϕ satisfies the moment conditions of order 1, and the shifts of ϕ are stable. Thus, by Theorem 3.4, the subdivision scheme associated with b converges in the L_∞-norm. By Theorem 3.3, we conclude that $\rho(B|_V) < 1$. This finishes the proof of the theorem. □

In the case $s = 1$ and $M = (2)$, Theorem 4.3 was established by Jia [6]. In the multivariate case, Theorem 4.3 was also obtained independently by Lawton, Lee, and Shen [11].

We finish this paper by an example about the L_2-convergence of a subdivision scheme.

Example 4.4. Let $M = 2I$, where I is the 2×2 identity matrix. Consider the refinement equation (3.18), where the mask a is given by its symbol

$$\tilde{a}(z) = 1 + (1/2 + t)(z_1 + z_2 + z_1 z_2) + (1/2 - t)(z_1^{-1} + z_2^{-1} + z_1^{-1} z_2^{-1}),$$

with t being a real number. The normalized solution of the refinement equation is the standard linear element if $t = 0$, and is the characteristic function of the unit square $[0, 1)^2$ if $t = 1/2$. Let $b := a * a^*$. By computation we find that

$$\begin{aligned}\tilde{b}(z) = |\tilde{a}(z)|^2 = {} & (5/2 + 6t^2) + (3/2 + 2t^2)(z_1 + z_2 + z_1^{-1} + z_2^{-1}) \\ & + (3/2 - 2t^2)(z_1 z_2 + z_1^{-1} z_2^{-1}) + (1/2 + 2t^2)(z_1 z_2^{-1} + z_1^{-1} z_2) \\ & + (1/4 - t^2)(z_1^2 + z_1^2 z_2^2 + z_2^2 + z_1^{-2} + z_1^{-2} z_2^{-2} + z_2^{-2}) \\ & + (1/2 - 2t^2)(z_1^2 z_2 + z_1 z_2^2 + z_1^{-2} z_2^{-1} + z_1^{-1} z_2^{-2}).\end{aligned}$$

With this b, the operator B is given by

$$Bw(\alpha) = \sum_{\beta \in \mathbb{Z}^2} b(2\alpha - \beta) w(\beta), \qquad w \in \ell_0(\mathbb{Z}^2),\ \alpha \in \mathbb{Z}^2.$$

For $j = 1, 2$, let $\nu_j = \delta - \delta_{e_j}$ and $\mu_j = \nu_j * \nu_j^* = -\delta_{e_j} + 2\delta - \delta_{-e_j}$. Then

$$B\mu_j = B(-\delta_{e_j} + 2\delta - \delta_{-e_j}) = (1 + 4t^2)(-\delta_{e_j} + 2\delta - \delta_{-e_j}) = (1 + 4t^2)\mu_j.$$

Thus, the minimal B-invariant subspace W_j generated by μ_j is the one-dimensional subspace spanned by μ_j. By Theorem 4.1 we conclude that

$$\lim_{n \to \infty} \|\nabla_j S_a^n \delta\|_2^{1/n} = \sqrt{\rho(B|_{W_j})} = \sqrt{1 + 4t^2}\,.$$

By Theorem 3.2, the subdivision scheme associated with a converges in the L_2-norm if and only if $\sqrt{1 + 4t^2} < \sqrt{4}$; that is, $|t| < \sqrt{3}/2$.

REFERENCES

[1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, Memoirs of Amer. Math. Soc., 93 (1991), p. 186.

[2] A. Cohen and I. Daubechies, *Non-separable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–137.

[3] G. Deslauriers, J. Dubois, and S. Dubuc, *Multidimensional iterative interpolation*, Canad. J. Math., 43 (1991), pp. 297–312.

[4] N. Dyn, *Subdivision schemes in computer-aided geometric design*, in Advances in Numerical Analysis II—Wavelets, Subdivision Algorithms and Radial Functions, W. A. Light, ed., Clarendon Press, Oxford, 1991, pp. 36–104.

[5] T. N. T. Goodman, C. A. Micchelli, and J. D. Ward, *Spectral radius formulas for subdivision operators*, in Recent Advances in Wavelet Analysis, L. L. Schumaker and G. Webb, eds., Academic Press, New York, 1994, pp. 335–360.
[6] R. Q. Jia, *Subdivision schemes in L_p spaces*, Adv. Comput. Math., 3 (1995), pp. 309–341.
[7] R. Q. Jia and J. J. Lei, *Approximation by multiinteger translates of functions having global support*, J. Approx. Theory, 72 (1993), pp. 2–23.
[8] R. Q. Jia and C. A. Micchelli, *On linear independence of integer translates of a finite number of functions*, Proc. Edinburgh Math. Soc., 36 (1992), pp. 69–85.
[9] R. Q. Jia and C. A. Micchelli, *Using the refinement equation for the construction of prewavelets* V: *Extensibility of trigonometric polynomials*, Computing, 48 (1992), pp. 61–72.
[10] J. Kovačević and M. Vetterli, *Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathbb{R}^n$*, IEEE Trans. Inform. Theory, 38 (1992), pp. 533–555.
[11] W. Lawton, S. L. Lee, and Z. W. Shen, *Convergence of cascade algorithms*, Numer. Math., to appear.
[12] G.-C. Rota and W. G. Strang, *A note on the joint spectral radius*, Indag. Math., 22 (1960), pp. 379–381.
[13] L. F. Villemoes, *Continuity of nonseparable quincunx wavelets*, Appl. Comput. Harmon. Anal., 1 (1994), pp. 180–187.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1200-1223, September 1998

009

CONVERGENCE RESULTS FOR SOME CONSERVATION LAWS WITH A REFLUX BOUNDARY CONDITION AND A RELAXATION TERM ARISING IN CHEMICAL ENGINEERING*

FRANÇOIS JAMES†

Abstract. This paper deals with a system of $2N$ semilinear transport equations with a boundary condition of imposed flux. The right-hand side models some kinetic exchange between two phases. It is thus a stiff term involving a small parameter which will tend to 0. Using compensated compactness, one proves, under some assumptions on the flux, that the solution to this system converges to a solution to a system of N quasilinear equations, a solution which satisfies a set of entropy inequalities. Thus the reflux boundary condition for the quasi-linear system is given a meaning.

Key words. hyperbolic systems, boundary conditions, relaxation, entropy, compensated compactness, chromatography, distillation

AMS subject classifications. 35L65, 35L67, 35Q20

PII. S003614109630793X

1. Introduction. We are interested in the following system of $2N$ equations, $N \geq 1$,

$$\begin{cases} \partial_t \mathbf{c}_\varepsilon^1 + \partial_x u \mathbf{c}_\varepsilon^1 = \frac{1}{\varepsilon}\big(\mathbf{c}_\varepsilon^2 - \mathbf{h}(\mathbf{c}_\varepsilon^1)\big), \\ \partial_t \mathbf{c}_\varepsilon^2 + \partial_x v \mathbf{c}_\varepsilon^2 = -\frac{1}{\varepsilon}\big(\mathbf{c}_\varepsilon^2 - \mathbf{h}(\mathbf{c}_\varepsilon^1)\big), \end{cases} \tag{1.1}$$

which is a simplified model of diphasic propagation arising in chemical engineering. In this kind of problem, two phases labelled 1 and 2 are in motion with respective velocities $u > 0$ and $v \leq 0$, which are assumed here to be constant. The case $v = 0$ corresponds to a model of chromatography (a mobile phase and a stationary one), and the case $v < 0$ corresponds to distillation (two phases moving countercurrent).

In equations (1.1), $\mathbf{c}_\varepsilon^1$ and $\mathbf{c}_\varepsilon^2$ are related to the concentrations in phase 1 and 2, respectively, and therefore should be nonnegative. The right-hand side rules the matter exchanges between the two phases. Without motion, the two phases would reach a state of thermodynamical equilibrium: the concentration in phase 2 is therefore related to the concentration in phase 1 by the function $\mathbf{h}$, which enjoys several properties coming from the thermodynamics.

In the case we are considering, the equilibrium cannot be reached because of the motion. The time needed to reach the equilibrium is not negligible with respect to the characteristic times induced by the velocities u and v. This phenomenon is known as a finite exchange kinetic: the actual concentration $\mathbf{c}_\varepsilon^2$ in phase 2 differs from $\mathbf{h}(\mathbf{c}_\varepsilon^1)$. The right-hand side of the equations quantifies the attraction of the system to the equilibrium state: it is a pulling-back force, and the constant parameter $1/\varepsilon$ is the "velocity" of exchange between the two phases.

A natural question arises here: how do the solutions of (1.1) behave when ε tends to 0, that is, when the exchange kinetic becomes instantaneous (the process is then quasi-static)? The limit system is obtained in a natural way by summing the $2N$

*Received by the editors August 7, 1996; accepted for publication (in revised form) October 15, 1997; published electronically June 2, 1998.
http://www.siam.org/journals/sima/29-5/30793.html
†Mathématiques, Applications et Physique Mathématique d'Orléans, UMR CNRS 6628, Université d'Orléans, BP 6759, 45067 Orléans Cedex 2, France (james@cmapx.polytechnique.fr).

equations in (1.1) and by putting $\mathbf{c}_\varepsilon^1 = \mathbf{c}$, $\mathbf{c}_\varepsilon^2 = \mathbf{h}(\mathbf{c})$, which means indeed that the concentration in phase 2 is actually the equilibrium concentration. We are led to the following nonlinear hyperbolic system, which expresses the conservation of matter:

$$\partial_t \left(\mathbf{c} + \mathbf{h}(\mathbf{c})\right) + \partial_x \left(u\mathbf{c} + v\mathbf{h}(\mathbf{c})\right) = 0. \tag{1.2}$$

The aim of this paper is to analyze the behavior of the solutions of (1.1) when ε tends to 0, when it is provided with boundary conditions

$$\mathbf{c}_\varepsilon^1(0,t) = \mathbf{a}(t) \in L^\infty(]0,+\infty[)^N, \quad u\mathbf{c}_\varepsilon^1(1,t) + v\mathbf{c}_\varepsilon^2(1,t) = \mathbf{b}(t) \in L^\infty(]0,+\infty[)^N, \tag{1.3}$$

together with Cauchy data in $L^\infty(]0,1[)^N$. To avoid any initial layer, we shall assume that the initial data are at equilibrium, that is, $\mathbf{c}_\varepsilon^1(\cdot,0) = \mathbf{c}^0 \in L^\infty(]0,1[)^N$ and $\mathbf{c}_\varepsilon^2(\cdot,0) = \mathbf{h}(\mathbf{c}^0)$. From the point of view of distillation, the boundary conditions are natural: the first one is a Dirichlet-like "injection" at one end of a column and acts only on the incoming variable ($u > 0$); the second one looks like a Neumann condition on the other end and imposes $v < 0$ (it is a simplified model of the "reflux" in a distillation column).

Concerning the standard Cauchy problem in the scalar case, i.e., $c(0,x) = c^0(x)$, $x \in \mathbb{R}$, $c^0 \in L^\infty$, the analysis is straightforward, and the solution of (1.1) tends to the entropy solution of (1.2), thus providing an alternative to the artificial viscosity method. Such results were obtained, for instance, by Tveito and Winther in [29], where the rate of convergence is estimated, and by Natalini [22]. Let us mention also the work by Katsoulakis and Tzavaras [19], where they give contraction properties for the solution of the system with relaxation. For systems of conservation laws, we refer to Chen, Levermore, and Liu [7], where a convergence result is proved for a 2×2 genuinely nonlinear system. This point of view can be successfully used for numerical purpose, see Jin and Xin [16] for a general setting for systems and Aregba-Driollet and Natalini [2] for convergence results in the scalar case.

On the other hand, the problem with boundary conditions is not as well behaved when ε tends to 0: it is well known that the setting of a Dirichlet boundary condition for a nonlinear hyperbolic scalar equation is difficult. Bardos, Leroux, and Nédélec [3] gave such a setting in the Kružkov sense, using the artificial viscosity method in the context of BV functions. We shall not recover this formulation here, since the Dirichlet data act only on incoming variables. For systems, the first existence result was given by Benabdallah and Serre [4] for systems of two equations. We refer also to works by Dubois and LeFloch [8], where the Dirichlet boundary condition appears as a Riemann problem on a half-plane, Gisclon [10], and Gisclon and Serre [11]. We mention also Goodman's work [12], where global existence is proved for strictly hyperbolic systems of conservation laws with initial and boundary data of small BV norm. The solutions also have small total variation and therefore have strong traces on the boundary. On the other hand, in [18] Kan, Santos, and Xin consider a general system of conservation laws and compare various notions of boundary conditions (vanishing viscosity, half-space Riemann problem). Their solution is built by a Godunov method. In the same spirit, we also mention the paper by Joseph and LeFloch [17], who also compare different approximations and the resulting boundary layers.

The reflux boundary condition at $x = 1$ seems to have been very little studied. For the scalar Burgers equation with the boundary condition $u^2(.,t) = 0$, Gisclon proved in [9] that the solution satisfies $u(.,t) \leq 0$ on the boundary (which coincides with the solution in the sense of [3]).

Finally, let us mention one work which is concerned with both relaxation and boundary conditions. Wang and Xin [30] consider a 2×2 system with relaxation. The boundary condition is chosen so that uniform BV estimates hold, and they prove convergence to a scalar conservation law satisfying a boundary-entropy condition, for which uniqueness holds.

We are going to prove that, under suitable conditions on the flux $u\mathbf{c} + v\mathbf{h}(\mathbf{c})$ with respect to $\mathbf{b}$, there exists a subsequence of solutions of (1.1) which converges to a weak solution of (1.2). This solution is characterized by a set of entropy inequalities. Since we have no BV estimates for the solution with $\varepsilon > 0$, we are led to work with bounded measurable functions, and use the compensated compactness method. This can be done in two cases: first for scalar equations with any smooth function $\mathbf{h}$ and then for a system of N equations, for a specific $\mathbf{h}$, the so-called Langmuir isotherm. Notice that the Langmuir system is not hyperbolic on the whole physical domain of interest. However, we use a specific set of entropies, namely the so-called kinetic entropies, which were introduced in [14], that allows us to achieve compactness.

Finally, we prove that the weak solutions are indeed solutions in the sense of distributions and that they satisfy in a strong sense the initial condition as well as the reflux boundary condition at $x = 1$. The incoming boundary condition seems to be lost in the limiting system. This is not very surprising, since we fall from a $2N$ equations system to N equations. Some boundary layer phenomena probably occur at $x = 0$, which we do not investigate here. This may indicate that the system of conservation laws with the reflux boundary condition is well-posed, but the precise study of this is left for future research.

The paper is organized as follows. In section 2 we state a few results and notations which hold for both the scalar equation and the system. Section 3 and 4 are devoted to the proof of a priori estimates and compactness, respectively, for the scalar equation and the system. Section 5 deals with boundary conditions.

2. Preliminary results.

We state here a few results and remarks that are common to both the scalar equation and the Langmuir model. Namely, we prove that equation (1.1) is well-posed for $\varepsilon > 0$, and we also define a particular set of entropies, which appears to be natural from the structure of the equations. In the following, we shall set $\Omega \stackrel{\text{def}}{=}]0,1[\times]0,T[$.

2.1. Existence for $\varepsilon > 0$.

THEOREM 2.1. *For a given $T > 0$, assume that $\mathbf{a}$ and $\mathbf{b}$ are in $L^\infty(]0,T[)^N$, $\mathbf{c}^0 \in L^\infty \cap L^1(]0,1[)^N$, and that the function $\mathbf{h}$ is of class C^1. Then there exists a unique solution to* (1.1), *which lies in $L^\infty(]0,T[; L^1(]0,1[))$.*

Proof. We first rewrite (1.1) in an equivalent integral form by using Duhamel's principle; then we prove a contraction estimate to apply a fixed point theorem. This is rather tedious, because of the initial and boundary conditions. The set $[0,1]\times[0,+\infty[$ is indeed divided into four zones, namely, $Z_1 = \{(x,t) \quad | \quad x \geq ut, x \leq 1 + vt\}$, $Z_2 = \{(x,t) \quad | \quad x \geq ut, x \geq 1 + vt\}$, $Z_3 = \{(x,t) \quad | \quad x \leq ut, x \leq 1 + vt\}$, $Z_4 = \{(x,t) \quad | \quad x \leq ut, x \geq 1 + vt\}$, depending upon whether the characteristics encounter $\{t = 0\}$, $\{x = 0\}$, or $\{x = 1\}$.

We shall fully write the contraction estimate for t large enough so that $(x,t) \in Z_4$ for every $x \in [0,1]$. We omit in this proof the dependence in ε. Taking into account the reflux boundary condition on $x = 1$, Duhamel's principle writes, for almost every

$(x,t) \in Z_4$,

(2.1)
$$\begin{cases} \mathbf{c}^1(x,t) = \mathbf{a}\left(t - \frac{x}{u}\right) + \frac{1}{\varepsilon}\int_{t-\frac{x}{u}}^{t} \left[\mathbf{c}^2(x+u(s-t),s) - \mathbf{h}(\mathbf{c}^1(x+u(s-t),s))\right] ds, \\ \mathbf{c}^2(x,t) = \frac{1}{v}\mathbf{b}\left(t + \frac{1-x}{v}\right) - \frac{u}{v}\mathbf{a}\left(t + \frac{1-x}{v} - \frac{1}{u}\right) \\ -\frac{u}{v}\frac{1}{\varepsilon}\int_{t+\frac{1-x}{v}-\frac{1}{u}}^{t+\frac{1-x}{v}} \left[\mathbf{c}^2(1+u(s-t-\frac{1-x}{v}),s) - \mathbf{h}(\mathbf{c}^1(1+u(s-t-\frac{1-x}{v}),s))\right] ds \\ \qquad -\frac{1}{\varepsilon}\int_{t+\frac{1-x}{v}}^{t} \left[\mathbf{c}^2(x+v(s-t),s) - \mathbf{h}(\mathbf{c}^1(x+v(s-t),s))\right] ds. \end{cases}$$

Denote by $\mathcal{T}$ the application from $X = L^\infty(]0,T[;L^1_x)^{2N}$ into itself which associates the right-hand side of the equations in (2.1) with a pair $C = (\mathbf{c}^1, \mathbf{c}^2) \in X$. For two elements C and $\hat{C}$ in X, with the same initial and boundary data, the terms involving $\mathbf{a}$ and $\mathbf{b}$ in (2.1) disappear when computing $\mathcal{T}(C) - \mathcal{T}(\hat{C})$, so, for a given (x,t), we have

$$|\mathcal{T}(C)(x,t) - \mathcal{T}(\hat{C})(x,t)| \le \frac{1}{\varepsilon}\max(|T_1(x,t)|, |T_2(x,t)|),$$

where T_i follows from the difference of the integral terms and $|\cdot|$ is a norm on $\mathbb{R}^{2N}$. One has easily

$$\begin{aligned} |T_1(x,t)| &\le \int_{t-\frac{x}{u}}^{t} \left|\mathbf{c}^2(x+u(s-t),s) - \hat{\mathbf{c}}^2(x+u(s-t),s)\right| ds \\ &+ \int_{t-\frac{x}{u}}^{t} \left|\mathbf{h}(\mathbf{c}^1(x+u(s-t),s)) - \mathbf{h}(\hat{\mathbf{c}}^1(x+u(s-t),s))\right| ds \\ &\le \int_{t-\frac{x}{u}}^{t} \left|\mathbf{c}^2(x+u(s-t),s) - \hat{\mathbf{c}}^2(x+u(s-t),s)\right| ds \\ &+ K\int_{t-\frac{x}{u}}^{t} \left|\mathbf{c}^1(x+u(s-t),s) - \hat{\mathbf{c}}^1(x+u(s-t),s)\right| ds \end{aligned}$$

if K is the Lipschitz constant of $\mathbf{h}$. We can estimate $\|T_1(.,t)\|_{L^1_x}$ by Fubini's theorem, which gives

$$\begin{aligned} \|T_1(.,t)\|_{L^1_x} &\le \int_{t-1/u}^{t} \|\mathbf{c}^2(.,s) - \hat{\mathbf{c}}^2(.,s)\|_{L^1_x} ds + K\int_{t-1/u}^{t} \|\mathbf{c}^1(.,s) - \hat{\mathbf{c}}^1(.,s)\|_{L^1_x} ds \\ &\le t\left(\max_{s\in[0,t]} \|\mathbf{c}^2(.,s) - \hat{\mathbf{c}}^2(.,s)\|_{L^1_x} + K\max_{s\in[0,t]} \|\mathbf{c}^1(.,s) - \hat{\mathbf{c}}^1(.,s)\|_{L^1_x}\right). \end{aligned}$$

A similar formula can be obtained for T_2, involving the quantity $u/|v|$.

Now, if (x,t) changes zone with x, we proceed in the same way in each zone and separate the integral for the L^1_x norm. We do not write these straightforward computations, which lead to the existence of a constant $M > 0$, which depends on K and $u/|v|$, such that

$$\|\mathcal{T}(C) - \mathcal{T}(\hat{C})\|_{L^\infty(]0,t[;L^1_x)^{2N}} \le \frac{tM}{\varepsilon}\|C - \hat{C}\|_{L^\infty(]0,t[;L^1_x)^{2N}}.$$

Now, choose T_0 such that $T_0 M/\varepsilon < 1$, and apply the fixed point theorem on $L^\infty(]0,T_0[;L^1_x)^{2N}$. This gives existence and uniqueness of the solution on $[0,T_0]$.

Since the contraction estimate does not depend on the initial data, we can perform again the same argument on $[T_0, 2T_0]$, and so on, to finally reach any prescribed $T > 0$. Thus the theorem is proved. □

2.2. Diphasic entropies. We introduce here a set of entropies which is quite natural in view of the structure of the equations. They are actually a discrete version (with two velocities only) of the kinetic entropies introduced by Perthame and Tadmor in [23].

DEFINITION 2.1. *We shall say that a function $\eta : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a "diphasic entropy" for* (1.2) *if there exist two convex functions $\eta_1, \eta_2 : \mathbb{R}^N \longrightarrow \mathbb{R}$ satisfying*

$$\nabla_{\mathbf{c}}\eta_1(\mathbf{c}) = \nabla_{\mathbf{c}}\eta_2\big(\mathbf{h}(\mathbf{c})\big) \qquad \forall \mathbf{c} \in \mathbb{R}^N, \tag{2.2}$$

such that $\eta(\mathbf{c}) = \eta_1(\mathbf{c}) + \eta_2\big(\mathbf{h}(\mathbf{c})\big)$.

Remark 2.1. The function $\mathbf{h}$ itself is, in general, defined by such a pair of functions, which are given, for instance, by statistical thermodynamics models (see [15] and the quoted references therein for examples and more information). Actually, the pair $(\mathbf{c}, \mathbf{h}(\mathbf{c}))$ is a stable state of equilibrium for the diphasic system. Thus it achieves the infimum of $\eta_1(\mathbf{c}_1) + \eta_2(\mathbf{c}_2)$ under the constraint that the total amount of matter $\mathbf{c}_1 + \mathbf{c}_2$ is constant. The relation (2.2) is nothing but the characterization of the minimum and is a generalized version of the well-known "chemical potential equalities" in thermodynamics. Consequently, $\mathbf{h}'(\mathbf{c})$ is positive in the scalar case and is diagonable with positive eigenvalues for a system. This leads also to the existence of a natural "physical" entropy for such systems.

Our main concern in the following is to obtain a priori estimates on the solution $(\mathbf{c}_\varepsilon^1, \mathbf{c}_\varepsilon^2)$ to (1.1) which are uniform in ε. The classical method here is to prove that the entropy production associated with (1.1) is nonpositive for several well-chosen entropies. Consider any pair (η_1, η_2) satisfying (2.2); multiply the two equations in (1.1), respectively, by $\nabla_{\mathbf{c}}\eta_1(\mathbf{c}_\varepsilon^1)$ and $\nabla_{\mathbf{c}}\eta_2(\mathbf{c}_\varepsilon^2)$; sum; then use (2.2). We formally obtain the following law for the entropy production:

$$\begin{aligned} &\partial_t \left(\eta_1(\mathbf{c}_\varepsilon^1) + \eta_2(\mathbf{c}_\varepsilon^2)\right) + \partial_x \left(u\eta_1(\mathbf{c}_\varepsilon^1) + v\eta_2(\mathbf{c}_\varepsilon^2)\right) \\ &\qquad = \frac{1}{\varepsilon}\left[\left(\nabla_{\mathbf{c}}\eta_2\big(h(\mathbf{c}_\varepsilon^1)\big) - \nabla_{\mathbf{c}}\eta_2(\mathbf{c}_\varepsilon^2)\right) \cdot \big(\mathbf{c}_\varepsilon^2 - \mathbf{h}(\mathbf{c}_\varepsilon^1)\big)\right]. \end{aligned} \tag{2.3}$$

It remains to notice that the right-hand side is always nonpositive, since η_2 is convex. Integrating on $[0,1]$ therefore gives, at least formally,

$$\frac{d}{dt}\int_0^1 \big[\eta_1(\mathbf{c}_\varepsilon^1(x,t)) + \eta_2(\mathbf{c}_\varepsilon^2(x,t))\big]dx \leq -\big[u\eta_1(\mathbf{c}_\varepsilon^1(.,t)) + v\eta_2(\mathbf{c}_\varepsilon^2(.,t))\big]\big|_0^1, \tag{2.4}$$

and all the technical work is now to estimate the boundary terms. To give a precise meaning to this differential inequality, we have to rewrite it in a weak form by multiplying by a test function $\varphi \geq 0$ and integrating by parts.

Provided we have enough entropies, (2.4) will give a priori estimates as well as compactness of a subsequence of solutions to (1.1). We can exhibit such entropies in the scalar case on the one hand and for the system of chromatography with the Langmuir isotherm on the other hand. In both cases, the local solution of Theorem 2.1 is therefore global for fixed ε.

2.3. Subcharacteristic condition. Before proceeding to estimates, we would like to relate system (1.1) with the usual form of systems with relaxation. This is done easily in the particular case $v = -u$ by setting $\mathbf{U}^\varepsilon = \mathbf{c}^1_\varepsilon + \mathbf{c}^2_\varepsilon \in \mathbb{R}^N$ and $\mathbf{V}^\varepsilon = u\mathbf{c}^1_\varepsilon - u\mathbf{c}^2_\varepsilon \in \mathbb{R}^N$. System (1.1) is therefore rewritten as

$$\partial_t \mathbf{U}^\varepsilon + \partial_x \mathbf{V}^\varepsilon = 0, \qquad \partial_t \mathbf{V}^\varepsilon + u^2 \partial_x \mathbf{U}^\varepsilon = \frac{2}{\varepsilon}\left(\mathbf{V}^\varepsilon - u(\mathbf{c}^1_\varepsilon - \mathbf{h}(\mathbf{c}^1_\varepsilon))\right). \tag{2.5}$$

Now, we notice that, since by Remark 2.1 $\mathbf{h}'(\mathbf{c})$ has positive eigenvalues, the function $\mathbf{c} + \mathbf{h}(\mathbf{c})$ is one-to-one. Let us denote $\mathbf{U} = \mathbf{c} + \mathbf{h}(\mathbf{c})$, its inverse by $\mathbf{c} = \mathbf{g}(\mathbf{U})$, and $\mathbf{F}(\mathbf{U}) \overset{\text{def}}{=} u[\mathbf{g}(\mathbf{U}) - \mathbf{h}(\mathbf{g}(\mathbf{U}))]$. The usual form of this kind of system should involve $\mathbf{F}(\mathbf{U})$ instead of $u(\mathbf{c}^1_\varepsilon - \mathbf{h}(\mathbf{c}^1_\varepsilon))$ in (2.5). This discrepancy appears because the right-hand side of system (1.1) is not symmetric with respect to $\mathbf{c}^1$ and $\mathbf{c}^2$. Another possible writing would make use of the "Maxwellians" $\mathbf{M}_1(\mathbf{U}) = \mathbf{g}(\mathbf{U})$ and $\mathbf{M}_2(\mathbf{U}) = \mathbf{h}(\mathbf{g}(\mathbf{U}))$. The convergence results would not be affected by this change.

In [21], Liu introduced a necessary condition on $\mathbf{F}'(\mathbf{U})$ to ensure the convergence of a subsequence of solutions of (1.1) to a solution of (1.2). This condition is known as the subcharacteristic condition, and we would like to point out that it is satisfied here because the function $\mathbf{h}$ is, in some sense, monotone (see Remark 2.1). Indeed, we have

$$\mathbf{F}'(\mathbf{U}) = [I_N + \mathbf{h}'(\mathbf{g}(\mathbf{U}))]\,\mathbf{g}'(\mathbf{U}),$$

where I_N stands for the identity matrix in $\mathbb{R}^N$. However, $\mathbf{g}'(\mathbf{U}) = (I_N + \mathbf{h}'(\mathbf{g}(\mathbf{U})))^{-1}$, so $\mathbf{F}'(\mathbf{U})$ is diagonable, and its eigenvalues are given for general values of u and v by

$$\lambda_i(\mathbf{U}) = \frac{u + v\mu_i(\mathbf{U})}{1 + \mu_i(\mathbf{U})},$$

where $\mu_i > 0$ are the eigenvalues of $\mathbf{h}'$, $1 \le i \le N$. It is readily seen that $v < \lambda_i(\mathbf{U}) < u$, which is the specific version of Liu's condition in this context.

3. Scalar equation. This section is devoted to the proof of the strong convergence of a subsequence of solutions to (1.1) in the scalar case. The function h is therefore a scalar function, which satisfies

$$h(0) = 0 \qquad \text{and} \quad h'(c) > 0 \quad \forall c. \tag{3.1}$$

Also, for any given convex η_2, we can define η_1 by $\eta_1(c) = \int_0^c \eta_2'\big(h(\sigma)\big)\,d\sigma$. We have, obviously, $\eta_1'(c) = \eta_2'\big(h(c)\big)$ and $\eta_1''(c) = \eta_2''\big(h(c)\big)h'(c) > 0$. Two particular cases are interesting. These are nonsmooth entropies, but a classical regularization argument, omitted in the following, allows us to deal with them.

- "Kružkov-like" entropies. For $k \in \mathbb{R}$, we set

$$\varphi_1^k(c^1) = |c^1 - k|, \qquad \varphi_2^k(c^2) = |c^2 - h(k)|.$$

It is easily checked that φ_1^k and φ_2^k satisfy (2.2), since h is increasing.

- L^∞ entropies. For $k \in \mathbb{R}$, we define

$$\psi_1^k(c^1) = (c^1 - k)^+, \qquad \psi_2^k(c^2) = \big(c^2 - h(k)\big)^+.$$

With these last entropies, the entropy estimates on c^i_ε become L^∞ estimates.

We begin in a classical way with some entropy and a priori estimates and first notice that, to prove entropy estimates for a given pair (η_1, η_2) giving a diphasic entropy, we need the condition

$$f(c) \overset{\text{def}}{=} uc + vh(c) \leq \min_{t>0} b(t) \qquad \text{for } c \geq M. \tag{3.2}$$

This is not a very satisfactory condition to impose, since it is not satisfied by such an usual isotherm as the Langmuir one,

$$h(c) = Kc/(1+c), \qquad K > 0. \tag{3.3}$$

Condition (3.2) actually implies some restrictions on the initial and boundary data, which lead to uniform L^∞ estimates for the solution to (1.1), for a broader class of fluxes.

THEOREM 3.1. *Assume $a \geq 0$, $b \leq 0$, $c^0 \geq 0$, and*

$$c^\star \overset{\text{def}}{=} \sup\{c \geq 0; \exists c' \leq c, f(c') \leq \min b(t)\} \geq \max[\|a\|_\infty, \|c^0\|_\infty]. \tag{3.4}$$

Then there exists a constant C depending only on c_1^0, a, and b such that $0 \leq c_\varepsilon^i(t,.) \leq C$, $i = 1, 2$.

Remark 3.1. Of course the result is meaningful only if $c^\star > 0$. This occurs only if $f(c)$ becomes nonpositive for some c. For instance, consider again the case of the Langmuir isotherm (3.3). It is easily seen that for $b = 0$, $c^\star > 0$ only if $u/|v| < K$. More generally, f achieves its minimum for $c_{\min} = \sqrt{K|v|/u} - 1$, which is positive if $u/|v| < K$, and $c^\star > 0$ if we have $f(c_{\min}) \leq b$, that is, $(\sqrt{K|v|} - \sqrt{u})^2 \geq -b$.

Remark 3.2. The choice $k = \max(\|a\|, \|c^0\|)$ is possible only if $f(a(t)) - b(t) \geq 0$ and $f(c^0(x)) - b(x) \geq 0$ for all $t > 0$ and a.e. $x \in]0,1[$. Otherwise, the L^∞ norm of the solution may not be bounded by the initial and Dirichlet boundary data.

From the above L^∞ estimate, we can easily obtain a weak convergence result by considering the weak form of the first equation in (1.1). Since c_ε^1 is L^∞ bounded uniformly in ε, and ε tends to 0, then $c_\varepsilon^2 - h(c_\varepsilon^1)$ tends to 0 in the sense of distributions on Ω when ε tends to 0. But we actually have the following stronger result.

LEMMA 3.2. *Under the assumptions of Theorem* 3.1, *ensuring the L^∞ bounds on the solution, $c_\varepsilon^2 - h(c_\varepsilon^1)$ tends to 0 in $L^2_{\text{loc}}(\Omega)$.*

From this result we can deduce, using the compensated compactness method, the following main result of this section.

THEOREM 3.3. *Consider $a, b \in L^\infty(]0,T[)$, and $c^0 \in L^1 \cap L^\infty(]0,1[)$. Under the assumptions of Theorem* 3.1, *that is,*

$$a \geq 0, \quad b \leq 0, \quad c^0 \geq 0, \quad c^\star \geq \max[\|a\|_\infty, \|c^0\|_\infty],$$

there then exists a subsequence of solutions to (1.1), *still denoted by c_ε^1, which converges a.e. and strongly in $]0,1[\times]0,T[$ to $c \in L^\infty(]0,T[; L^1(]0,1[))$. Moreover, c satisfies, for any $\varphi \in \mathcal{D}(\overline{\Omega})$, $\varphi \geq 0$, $k \in \mathbb{R}$,*

$$\begin{aligned} -\int_0^T \int_0^1 \Big[\big(|c-k| + |h(c)-h(k)|\big)\partial_t\varphi + \big(u|c-k| + v|h(c)-h(k)|\big)\partial_x\varphi \Big] dxdt \\ \leq \int_0^T u|a(t)-k|\varphi(0,t)dt + \int_0^T |b(t)-f(k)|\varphi(1,t)dt \\ - \int_0^1 \big(|c^0(x)-k| + |h(c^0(x))-h(k)|\big)\,\varphi(x,0)dx. \end{aligned} \tag{3.5}$$

Section 3.1 contains the entropy estimates and the proof of Theorem 3.1, while section 3.2 is devoted to the convergence results. Finally, we give a few remarks on viscous regularization is section 3.3.

3.1. A priori estimates. First we briefly show how condition (3.2) gives general entropy estimates. Consider a pair (η_1, η_2) which satisfies (2.2), and assume for simplicity that η_2 is bounded from below by 0. We start from equation (2.4) and estimate the boundary terms.

At $x = 0$, we have $c_\varepsilon^1(0, \cdot) = a(\cdot)$. We have, since $v < 0$, $\big[u\eta_1(c_\varepsilon^1) + v\eta_2(c_\varepsilon^2)\big]\big|_0 \le u\eta_1(a) \le C$ if $a \in L^\infty$. Next at $x = 1$, we rewrite the boundary condition in the form

$$c_\varepsilon^2 = \frac{u}{|v|} c_\varepsilon^1 - \frac{b}{|v|}.$$

We want to make $-[u\eta_1(c_\varepsilon^1) + v\eta_2(\frac{u}{|v|}c_\varepsilon^1 - \frac{b}{|v|})] \le K$, K being a constant. A sufficient condition to ensure this is

$$\zeta(c) \stackrel{\text{def}}{=} -\left[u\eta_1(c) + v\eta_2\left(\frac{u}{|v|}c - \frac{b}{|v|}\right)\right] \le K,$$

or $\zeta'(c) \le 0$, for c large. Differentiating ζ and using (2.2) shows that this occurs if $\eta_1'(c) = \eta_2'(h(c)) \le \eta_2'(\frac{u}{|v|}c - \frac{b}{|v|})$. Now, the fact that η_2' is nondecreasing and condition (3.2) lead to

$$\frac{d}{dt}\int_0^1 \big[\eta_1(c_\varepsilon^1(x,t)) + \eta_2(c_\varepsilon^2(x,t))\big]dx \le C(a, b, \eta_1, \eta_2, M, K),$$

where $K = \sup_{0 \le c \le M} \zeta(c)$. By integration, this leads to

$$\int_0^1 \big[\eta_1(c_\varepsilon^1(x,t)) + \eta_2(c_\varepsilon^2(x,t))\big]dx \le \int_0^1 \big[\eta_1(c^0(x)) + \eta_2(h(c^0)(x))\big]dx + C(a, b, \eta, M, K)t.$$

We point out again the fact that condition (3.2) is not to be used as it stands, since it depends on the flux. We prefer to put restrictions on the initial and boundary data, as in Theorem 3.1, which we are going to prove now. Actually, we perform the same computations as above, with two particular choices for η_i.

Proof of Theorem 3.1. (i) We first take $\eta_j(c^j) = [c^j]^-$, which happens to be a diphasic entropy since h is increasing. With this choice, the right-hand side of (2.4) is clearly bounded by $ua(t)^- - [uc_\varepsilon^1(1,t) + vc_\varepsilon^2(1,t)]^- = ua(t)^- + [-b(t)]^-$ (by using the boundary condition at $x = 1$). This becomes nonnegative provided $a \ge 0$ and $b \le 0$. Integrating in time now gives

$$\int_0^1 \big[\eta_1(c_\varepsilon^1(x,t)) + \eta_2(c_\varepsilon^2(x,t))\big]dx \le \int_0^1 \big[(c_1^0(x))^- + (c_2^0(x))^-\big]dx \le 0$$

if the initial data are nonnegative. Thus (1.1) preserves the positivity.

(ii) We now choose $\eta_i = \psi_i^k$, for an adequate k, which will give the upper bound. Indeed, the ψ_i^k are bounded from below (by 0!), and for $k \ge \|a\|_\infty$, the term on $x = 0$ becomes nonpositive. Now, for $x = 1$, we have with our choice for η_i,

$$\zeta(c) = -u(c-k)^+ - v\left[\frac{u}{v}c + \frac{b}{v} - h(k)\right]^+ \le [f(k) - b]^+$$

by a triangle inequality. To make the right-hand side nonpositive, we must find k such that $f(k) \le b$. This implies $k \le c^\star$ and is compatible with the constraint at $x = 0$ only if $\|a\|_\infty$ is less than $c^\star$.

Finally, by integration, provided k satisfies $c^\star \ge k \ge \|a\|_\infty$, we have

$$\int_0^1 \left[(c_\varepsilon^1(x,t)-k)^+ + \big(c_\varepsilon^2(x,t)-h(k)\big)^+\right] dx \\ \le C(b,k)t + \int_0^1 \left[(c^0(x)-k)^+ + \big(h(c^0(x))-h(k)\big)^+\right] dx.$$

Now, provided $c^\star \ge \max[\|a\|_\infty, \|c^0\|_\infty]$, we can choose k such that the right-hand side is nonpositive. ∎

3.2. Strong convergence. We turn to the proof of the convergence results.

Proof of Lemma 3.1. We begin from (2.3) with the diphasic entropy given by $\eta_2(c_2) = (1/2)c_2^2$, $\eta_1(c_1) = \int_0^{c_1} h(\sigma)\,d\sigma$, then multiply by φ with compact support in $]0,1[\times]0,T[$, and integrate by parts:

$$-\int_0^T\int_0^1 \big([\eta_1(c_\varepsilon^1)+\eta_2(c_\varepsilon^2)]\partial_t\varphi + [u\eta_1(c_\varepsilon^1)+v\eta_2(c_\varepsilon^2)]\partial_x\varphi\big)\,dx\,dt \\ = -\frac{1}{\varepsilon}\int_0^T\int_0^1 \big(c_\varepsilon^2 - h(c_\varepsilon^1)\big)^2\,\varphi dx\,dt.$$

Since c_ε^1 and c_ε^2 are L^∞-bounded uniformly in ε, multiplying this relation by ε gives the result. ∎

We now wish to prove a strong convergence property on c_ε^i by using Murat–Tartar's compensated compactness argument [27].

Proof of Theorem 3.2. *Step* 1. First we prove that, up to a subsequence, c_ε^1 converges strongly. Since c_ε^1 is L^∞ bounded uniformly in ε, and the functions h and η_i are smooth, the sequences c_ε^1, $h(c_\varepsilon^1)$, and $\eta_i(c_\varepsilon^i)$ converge in $L^\infty - w*$, respectively, to $\bar c$, $\bar h$, and $\bar\eta_i$, $i = 1, 2$. Now consider the following two quantities:

$$S^\varepsilon \stackrel{\text{def}}{=} \partial_t\left(c_\varepsilon^1 + h(c_\varepsilon^1)\right) + \partial_x\left(uc_\varepsilon^1 + vh(c_\varepsilon^1)\right),$$
$$T^\varepsilon \stackrel{\text{def}}{=} \partial_t\left(\eta_1\big(c_\varepsilon^1\big) + \eta_2\big(h(c_\varepsilon^1)\big)\right) + \partial_x\left(u\eta_1\big(c_\varepsilon^1\big) + v\eta_2\big(h(c_\varepsilon^1)\big)\right).$$

We want to apply the classical div-curl lemma, which asserts that the quantity

$$(c_\varepsilon^1 + h(c_\varepsilon^1))[u\eta_1\big(c_\varepsilon^1\big) + v\eta_2\big(h(c_\varepsilon^1)\big)] - [\eta_1\big(c_\varepsilon^1\big) + \eta_2\big(h(c_\varepsilon^1)\big)](uc_\varepsilon^1 + vh(c_\varepsilon^1))$$

passes to the L^∞ weak-$*$ limit (see [27]), provided S^ε and T^ε are compact in $H^{-1}_{\text{loc}}(\Omega)$. But, for any pair (η_1,η_2) of diphasic entropies (in particular for the trivial entropies $(c_\varepsilon^1, c_\varepsilon^2)$ which give back S^ε), $T^\varepsilon = \mu^\varepsilon + g^\varepsilon$, where

$$\mu^\varepsilon \stackrel{\text{def}}{=} \partial_t\left(\eta_1(c_\varepsilon^1) + \eta_2(c_\varepsilon^2)\right) + \partial_x\left(u\eta_1(c_\varepsilon^1) + v\eta_2(c_\varepsilon^2)\right),$$
$$g^\varepsilon \stackrel{\text{def}}{=} \partial_t\left(\eta_2(h(c_\varepsilon^1)) - \eta_2(c_\varepsilon^2)\right) + \partial_x v\left(\eta_2(h(c_\varepsilon^1)) - \eta_2(c_\varepsilon^2)\right).$$

Now, T^ε is bounded in $W^{-1,\infty}$ since c_ε^1 is bounded in L^∞, and μ^ε is a nonpositive measure (it is actually 0 for the trivial entropies). By Lemma 3.1, we have that

$c_\varepsilon^2 - h(c_\varepsilon^1)$ tends to 0 in $L^2_{\rm loc}(\Omega)$; hence $\eta_2(c_\varepsilon^2) - \eta_2(h(c_\varepsilon^1))$ also tends to 0 in $L^2_{\rm loc}(\Omega)$. Since the operators ∂_t and ∂_x are continuous from $L^2_{\rm loc}(\Omega)$ to $H^{-1}_{\rm loc}(\Omega)$, g^ε tends to 0, and hence is compact, in $H^{-1}_{\rm loc}(\Omega)$. Thus, by Murat's lemma, T^ε is compact in $H^{-1}_{\rm loc}(\Omega)$.

With the obvious notation denoting the weak-$*$ limit with an overline, we obtain, after trivial simplifications,

$$\overline{h(c_1)\eta_1(c_1)} - \overline{h}\;\overline{\eta_1} = \overline{c_1\eta_2\big(h(c_1)\big)} - \bar{c}_1\overline{\eta_2}. \tag{3.6}$$

We now proceed classically by introducing the Young measure $\nu = \nu_{x,t}$ associated with the sequence c_ε^1: for every function α,

$$\alpha(c_\varepsilon^1) \rightharpoonup \overline{\alpha} = \int_{\mathbb{R}} \alpha(\xi)\, d\nu(\xi) = \langle \alpha(\xi), \nu\rangle \qquad \text{in } L^\infty - w*.$$

Equation (3.6) therefore becomes

$$\Big\langle (\xi - \bar{c})\eta_2\big(h(\xi)\big) - \Big(h(\xi) - \overline{h(c)}\Big)\,\eta_1(\xi), \nu\Big\rangle = 0.$$

If we now introduce the aforementioned Kružkov-like entropies $\eta_1(c_1) = |\xi - c_1|$, $\eta_2(c_2) = |h(\xi) - c_2|$, the preceding equality becomes

$$\Big\langle (\xi - \bar{c})|h(\xi) - h(\bar{c})| - \big(h(\xi) - \overline{h(c)}\big)|\xi - \bar{c}|, \nu\Big\rangle = 0.$$

But the fact that h is increasing implies easily that $(\xi - \bar{c})|h(\xi) - h(\bar{c})| = |\xi - \bar{c}|\big(h(\xi) - h(\bar{c})\big)$, so we finally obtain

$$\big(\overline{h(c)} - h(\bar{c})\big)\,\langle |\xi - \bar{c}|, \nu\rangle = 0.$$

The conclusion now follows exactly in the same way as in [27]: ν is a Dirac mass, except where h is affine.

Proof of Theorem 3.2. *Step* 2. First notice that any solution $(c_\varepsilon^1, c_\varepsilon^2)$ to (1.1) with the boundary conditions (1.3) satisfies

$$\begin{aligned} -\int_0^T \int_0^1 \Big[\big(|c_\varepsilon^1 - k| + |c_\varepsilon^2 - h(k)|\big)\partial_t\varphi + \big(u|c_\varepsilon^1 - k| + v|c_\varepsilon^2 - h(k)|\big)\partial_x\varphi\Big]dxdt \\ \le \int_0^T u|a(t) - k|\varphi(0,t)dt + \int_0^T |b(t) - f(k)|\varphi(1,t)dt \\ - \int_0^1 \big(|c^0(x) - k| + |h(c^0(x)) - h(k)|\big)\,\varphi(x,0)dx. \end{aligned} \tag{3.7}$$

Indeed, rewrite (2.3) with $\eta_1(c^1) = |c^1 - k|$ and $\eta_2(c^2) = |c^2 - h(k)|$, multiply by $\varphi(x,t) \ge 0$, and integrate by parts with respect to x and t. We obtain, using the

boundary condition on $x = 0$ and the fact that $v < 0$,

$$
\begin{aligned}
-\int_0^T \int_0^1 \Big[\big(|c_\varepsilon^1 - k| + |c_\varepsilon^2 - h(k)|\big)\partial_t\varphi + \big(u|c_\varepsilon^1 - k| + v|c_\varepsilon^2 - h(k)|\big)\partial_x\varphi \Big] dxdt \\
\leq \int_0^T u|a(t) - k|\varphi(0,t)dt \\
- \int_0^T \big(u|c_\varepsilon^1(1,t) - k| + v|c_\varepsilon^2(1,t) - h(k)|\big)\,\varphi(1,t)dt \\
- \int_0^1 \big(|c^0(x) - k| + |h(c^0)(x) - h(k)|\big)\,\varphi(x,0)dx.
\end{aligned}
$$

For $x = 1$, we use the boundary condition to get

$$
\begin{aligned}
&\int_0^T (u|c_\varepsilon^1(1,t) - k| + v|c_\varepsilon^2(1,t) - h(k)|)\varphi(1,t)dt \\
&\qquad = \int_0^T \left(u|c_\varepsilon^1(1,t) - k| + v\frac{1}{|v|}|b(t) - uc_\varepsilon^1(1,t) - vh(k)| \right) \varphi(1,t)dt.
\end{aligned}
$$

Again since $v < 0$, $v/|v| = -1$, we add and subtract uk in the second term of the right-hand side, and we use the triangle inequality to conclude. Finally, the first step of this proof allows us to pass to the limit in the left-hand side of (3.7). □

3.3. Remarks on viscous regularization. We consider here another possible perturbation of the hyperbolic equation, by means of a viscous regularization. We go back to the classical form of conservation law,

$$
\partial_t w + \partial_x f(w) = \varepsilon\partial_{xx} w, \qquad x < 1, \tag{3.8}
$$

provided with a perturbed Neumann condition on $x = 1$:

$$
-\varepsilon\partial_x w(t,1) + f\big(w(t,1)\big) = b(t).
$$

This is exactly the context considered by Gisclon in [9] for the Burgers equation.

We drop the Dirichlet condition on $x = 0$: it has been fully considered by Bardos, Leroux, and Nédélec in [3] and cannot be treated without a priori BV estimates, since the entropy condition on the boundary involves the trace of the solution. Notice that our boundary condition differs from the one in [3], since we do not impose the equilibrium at the boundary.

We are going to formally recover the L^∞ estimate from this perturbation, under the same assumptions as in Theorem 3.1, that is, $b \leq 0$ and condition (3.4). After that, classical compactness arguments can be performed in order to obtain strong convergence of the sequence w^ε to a weak solution.

Indeed, multiply (3.8) by $\eta'(w)$, where (η, q) is any pair entropy-flux; then integrate in x. We obtain

$$
\begin{aligned}
\frac{d}{dt}\int_{-\infty}^1 \eta(w(x,t))\,dx + q(w(1,t)) = (\varepsilon\eta'(w(1,t))\partial_x w(1,t)) \\
-\varepsilon\int_{-\infty}^1 \eta''(w(x,t))(\partial_x w(x,t))^2\,dx
\end{aligned} \tag{3.9}
$$

Now, the term involving η'' is nonnegative since η is convex, and we wish to control the quantity $q(w) - \varepsilon\eta'(w)\partial_x w$ on the boundary. Using the boundary condition, we have $q(w) - \varepsilon\eta'(w)\partial_x w = q(w) + \eta'(w)\,(b - f(w))$. But, assuming $f(0) = 0$, we can write

$$q(w) = \int_0^w \eta'(v)f'(v)\,dv = -\int_0^w \eta''(v)f(v)\,dv + \eta'(w)f(w),$$

so that

$$\frac{d}{dt}\int_{-\infty}^1 \eta(w(x,t))\,dx \le \int_0^{w(1,t)} \eta''(v)\,[f(v) - b]\,dv. \tag{3.10}$$

Now, for a general η, if we assume that

$$f(w) \le \min_{t>0} b(t) \qquad \text{if} \quad |w| \ge M, \tag{3.11}$$

for some $M > 0$ then, since $\eta'' \ge 0$, the right-hand side in (3.10) is bounded by

$$\inf_{w\le M}\int_0^w \eta''(v)\,(f(v) - b)\,dv \stackrel{\text{def}}{=} C.$$

This proves an entropy estimate for any entropy η, provided (3.11) is satisfied. Notice that, in the particular case $f(w) = ug(w)+vh\big(g(w)\big)$, condition (3.11) is exactly (3.2). Notice also that such a flux condition on a Burgers-like equation does not satisfy the assumption, since the function $w \mapsto w^2$ is not bounded.

To recover the L^∞ estimates, we first consider $\eta(w) = w^-$. Then we have $\eta''(w) = -\delta_0(w)$, so that (3.10) becomes

$$\frac{d}{dt}\int_{-\infty}^1 w(x,t)^-\,dx \le b(t) \le 0.$$

Hence $w(x,t) \ge 0$ if $w(x,0) \ge 0$. For the upper bound, we choose $\eta(w) = (w-k)^+$, for a given $k \in \mathbb{R}$, which gives $\eta''(w) = \delta_k(w)$. Thus

$$\frac{d}{dt}\int_{-\infty}^1 [w(x,t) - k]^+\,dx \le \begin{cases} 0 & \text{if } k \notin [0, w(1,t)], \\ f(k) - b(t) & \text{if } k \in [0, w(1,t)]. \end{cases}$$

If one can choose k such that $f(k) - b(t) \le 0$, then we are done. This can be done precisely if condition (3.4) is satisfied.

4. The Langmuir model. We now consider an $N \times N$ system which appears in chemical engineering both in chromatography and distillation. The unknowns are N functions $c_i(x,t)$ solutions of

$$\begin{gathered}\partial_t\big(c_i + h_i(\mathbf{c})\big) + \partial_x\big(uc_i + vh_i(\mathbf{c})\big) = 0, \quad t \ge 0,\ x \in]0,1[,\ 1 \le i \le N, \\ c_i(x,0) = c_i^\circ(x) \ge 0,\end{gathered} \tag{4.1}$$

where the vector-valued function $\mathbf{h}$ is the so-called Langmuir isotherm (see [20]),

$$h_i(\mathbf{c}) = \frac{k_i c_i}{D}. \tag{4.2}$$

The k_i's given here are numbers $0 < k_1 < k_2 < \cdots < k_N$ and $D = 1+c_1+c_2+\cdots+c_N$. Function $\mathbf{h}$ is defined for $D > 0$, which contains the "physical domain" $\{c_i \geq 0, 1 \leq i \leq N\}$. We set in the following $\mathbf{c}(x,t) = (c_1(x,t), \ldots, c_N(x,t))$.

System (4.1) of partial differential equations has been treated by Rhee, Aris, and Admundson in [24] for chromatography, which corresponds to $v = 0$, and in [25] for a countercurrent model of chromatography, which is very close to the system we deal with. Canon and James also studied both systems [5], [6], respectively, for distillation and chromatography. Serre [26] studied a variant of this system, which emphasizes the structure of the function $\mathbf{h}$. On the same variant, a kinetic formulation was obtained in [14], which led to L^∞ estimates and strong convergence properties for bounded sequences of solutions, even though system (4.1) is not hyperbolic on the whole physical domain. The entropies we are about to use are very similar to those in [14], and before defining them, we recall without proof some fundamental algebraic properties of $\mathbf{h}$ (see [5], [24], [26]).

LEMMA 4.1. (i) *If $c_i \geq 0$ for $1 \leq i \leq N$, then $A(\mathbf{c}) = \nabla_{\mathbf{c}}\mathbf{h}(\mathbf{c})$ has N real eigenvalues $\mu_i(\mathbf{c})$, and $w_i \overset{\text{def}}{=} D\mu_i$ satisfies*

$$0 < w_1 \leq k_1 \leq w_2 \leq k_2 \leq \cdots \leq k_{N-1} \leq w_N \leq k_N;$$

(ii) *w_i is a strong i-Riemann invariant, in the sense that $\nabla_{\mathbf{c}} w_i$ is a left eigenvector of $A(\mathbf{c})$;*

(iii)
$$D = \prod_{i=1}^{N} \frac{k_i}{w_i};$$

(iv)
$$c_i \prod_{j \neq i} \left(1 - \frac{k_i}{k_j}\right) = -\prod_{j=1}^{N} \left(1 - \frac{k_i}{w_j}\right);$$

(v)
$$\sigma_0 \overset{\text{def}}{=} \prod_{i=1}^{N} k_i, \qquad \sigma_j(\mathbf{c}) \overset{\text{def}}{=} \sum_{1 \leq i_1, \ldots, i_j \leq N} \frac{1}{w_{i_1} \ldots w_{i_j}} \quad \text{for } 1 \leq j \leq N,$$

are $N+1$ independent affine functions of $(c_1, \ldots, c_N)$.

These properties are very strong. (i) and (ii) give the so-called richness (Serre [26]): system (4.1) admits a diagonal form for smooth solutions, namely,

$$(1 + \mu_i)\partial_t w_i + (u + v\mu_i)\partial_x w_i = 0. \tag{4.3}$$

Moreover, this system also belongs to the Temple class [28] for which some existence and uniqueness results are known in BV when they are strictly hyperbolic (see [26], [13]).

Remark 4.1. Let us point out an important point (see [6] for further details). Property (i) allows a degeneracy of the system (two equal eigenvalues). This can happen only for $w_i = w_{i+1} = k_i$, then $\mu_i = \mu_{i+1}$, and $c_i = 0$. It requires that, initially, $w_i^0(x) = w_{i+1}^0(x) = k_i$ for some $x \in \mathbb{R}$.

Section 4.1 is devoted to some technical devices to generalize the kinetic entropies of [14] for system (4.1). Next, we establish some invariants regions in section 4.2. In

particular, we prove that the domain $\{c_i \geq 0\}$ is invariant. Finally, we prove strong convergence results in section 4.3. In the following, we shall say that a vector $\mathbf{z}$ is nonnegative, $\mathbf{z} \in \mathbb{R}^N_+$ (respectively, nonpositive, $\mathbf{z} \in \mathbb{R}^N_-$), if all its components are nonnegative (respectively, nonpositive). We denote by w_1 (respectively, $w_i^{\mathbf{a}}$, w_i^0) the i-Riemann invariant associated by Lemma 4.1(ii) with $\mathbf{c}^1_\varepsilon$ (respectively, with the data on $x = 0$, with the initial data).

4.1. Some specific entropies. Now, we define a first trivial (i.e., affine) diphasic entropy for system (4.1), from which we shall build a specific family of nontrivial (i.e., convex) diphasic entropies. This set of entropies was already mentioned by Serre [26]. For $\xi \in \mathbb{R}_+$ and $\mathbf{c}^1 \in \mathbb{R}^N_+$, we set

$$E_0(\xi; \mathbf{c}^1) = \prod_{i=1}^N \left(1 - \frac{\xi}{w_i^1}\right), \qquad \gamma(\xi) = E_0(\xi; 0) = \prod_{i=1}^N \left(1 - \frac{\xi}{k_i}\right),$$

where w_i^1 are the Riemann invariants corresponding to $\mathbf{c}^1$.

LEMMA 4.2. *The function E_0 is affine with respect to $\mathbf{c}^1$. Let $\nabla_{\mathbf{c}} E_0(\xi)$ denote its gradient. We now define, for $\xi \in \mathbb{R}^+$ and $\mathbf{c}^2 \in \mathbb{R}^N$,*

$$F_0(\xi; \mathbf{c}^2) = \nabla_{\mathbf{c}} E_0(\xi) \cdot \mathbf{c}^2 + \xi \gamma(\xi). \tag{4.4}$$

Then the pair of functions (E_0, F_0) defines a diphasic entropy for (4.1)*, and we have*

$$F_0(\xi; \mathbf{h}(\mathbf{c}^1)) = \frac{\xi E_0(\xi; \mathbf{c}^1)}{D}. \tag{4.5}$$

Proof. First notice that, if E_0 is affine and F_0 is given by (4.4), then obviously the pair (E_0, F_0) defines a diphasic entropy, since $\nabla_{\mathbf{c}^2} F_0(\xi; \mathbf{h}(\mathbf{c}^1)) = \nabla_{\mathbf{c}} E_0(\xi)$.

We are going to prove that E_0 satisfies

$$\left\{ \begin{array}{ll} E_0(k_i; \mathbf{c}^1) & = \beta_i c_i^1, \qquad \text{where} \quad \beta_i = \prod_{j \neq i} (1 - k_i/k_j), \\ E_0(\xi; \mathbf{c}^1) & = -\gamma(\xi) \left[\sum_{i=1}^N \frac{k_i c_i^1}{k_i - \xi} - D \right] \\ & = -\gamma(\xi) \left[\xi \sum_{i=1}^N \frac{c_i^1}{k_i - \xi} - 1 \right] \quad \text{for } \xi \neq k_i, \end{array} \right. \tag{4.6}$$

so that for $\xi \neq k_i$, $\nabla_{\mathbf{c}} E_0(\xi) = -\xi \gamma(\xi) (\frac{1}{k_i - \xi})_{1 \leq i \leq N}$. To prove (4.6), recall that the Riemann invariants w_i^1 are the roots of the algebraic equation $\varphi(\xi) = 0$, where

$$\varphi(\xi) = \sum_{i=1}^N \frac{k_i c_i}{k_i - \xi} - D = \xi \sum_{i=1}^N \frac{c_i}{k_i - \xi} - 1. \tag{4.7}$$

But φ is also a rational fraction with poles k_i and roots w_i^1; thus an easy computation gives

$$\varphi(\xi) = -D \prod_{i=1}^N \frac{\xi - w_i^1}{\xi - k_i} = -\prod_{i=1}^N \frac{\frac{\xi}{w_i^1} - 1}{\frac{\xi}{k_i} - 1} = -\frac{E_0}{\gamma(\xi)} \tag{4.8}$$

by Lemma 4.1(iii) and the definitions of E_0 and $\gamma(\xi)$. Putting together (4.7) and (4.8) gives (4.6). Finally, (4.5) is obtained by playing with the two definitions of E_0, since

$$\nabla_{\mathbf{c}} E_0(\xi) \cdot \mathbf{h}(\mathbf{c}^1) = -\xi \gamma(\xi) \sum_{i=1}^N \frac{k_i c_i^1}{D} \frac{1}{k_i - \xi} = \frac{\xi}{D} E_0(\xi; \mathbf{c}^1) - \xi \gamma(\xi),$$

and this completes the proof. ∎

Remark 4.2. We state here a few useful properties of F_0. First, it is a polynomial of degree $N+1$ in ξ, very similar to E_0: if $c_i^2 \in \mathbb{R}_+^N$, it has roots $0, w_1^2, \dots, w_N^2$, with $0 < w_1^2 \le k_1 \le \cdots \le w_N^2 \le k_N$. We easily obtain also that, for any $\mathbf{z} \in \mathbb{R}^N$,

$$F_0(k_i; \mathbf{z}) = \nabla_{\mathbf{c}} E_0(k_i) \cdot \mathbf{z} = \beta_i z_i, \qquad 1 \le i \le N. \tag{4.9}$$

A crucial point now is to remark that $E_0(\xi; \mathbf{c})$ and $F_0(\xi; \mathbf{h}(\mathbf{c}))$ vanish simultaneously for $\xi = w_i^1 = w_i^2$, $1 \le i \le N$. Thus, taking the convention $w_0^j = 0$ and $w_{N+1}^j = +\infty$, we easily deduce the following.

COROLLARY 4.3. *For $0 \le i \le N$, $\mathbf{c}^1 \in \mathbb{R}_+^N$, $\mathbf{c}^2 \in \mathbb{R}_+^N$, let w_i^1 (respectively, w_i^2) be the roots of E_0 (respectively, the nonzero roots of F_0). Define*

$$\begin{aligned} \chi_i^1(\xi; \mathbf{c}^1) &= |E_0(\xi; \mathbf{c})| \mathbb{1}_{\{\xi \in]w_i^1, w_{i+1}^1[\}}, \\ \chi_i^2(\xi; \mathbf{c}^2) &= |F_0(\xi; \mathbf{c}^2)| \mathbb{1}_{\{\xi \in]w_i^2, w_{i+1}^2[\}}. \end{aligned} \tag{4.10}$$

Then the pair (χ_i^1, χ_i^2) defines a diphasic entropy for (4.1).

Notice that χ_i^1 (respectively, χ_i^2) is actually convex with respect to $\mathbf{c}^1$ (respectively, to $\mathbf{c}^2$), as the absolute value of an affine function. Thus the function $\eta_i(\xi; \mathbf{c}) \overset{\text{def}}{=} \chi_i^1(\xi; \mathbf{c}) + \chi_i^2(\xi; \mathbf{h}(\mathbf{c}))$ is indeed a nontrivial convex diphasic entropy for (4.1).

The class of entropies we consider now is defined as follows. Set, for $j = 1, 2$, $\mathbf{c}^j \in \mathbb{R}_+^N$, and a fixed $0 \le i \le N$,

$$S^j(\mathbf{c}^j) = \int_{\mathbb{R}_+} g(\xi) \chi_i^j(\xi; \mathbf{c}^j)\, d\xi, \qquad j = 1, 2.$$

The functions $S(\mathbf{c}) = S^1(\mathbf{c}) + S^2(\mathbf{h}(\mathbf{c}))$ are diphasic entropies for (1.2), for any nonnegative function g such that $g\chi_i^j$ is integrable at $+\infty$ in ξ (recall that χ_i^j is a polynomial in ξ). The corresponding entropy flux is $Q(\mathbf{c}) = uS^1(\mathbf{c}) + vS^2(\mathbf{h}(\mathbf{c}))$. We have to complement these functions by using for g a Dirac mass, $g(\xi) = \delta_{\xi^*}(\xi)$. To justify this, consider a sequence of nonnegative g's which converge to such a Dirac mass. These entropies will appear in the proof of the maximum principle below. Let us denote by $\mathcal{E}$ the set of all these entropies for $0 \le i \le N$.

Remark 4.3. The entropies in $\mathcal{E}$ are defined only on $\mathbb{R}_+^N$ and therefore cannot be used to prove the invariance of $\mathbb{R}_+^N$. But it is easily checked that the pairs $([c_i^1]^-, [c_i^2]^-)$, where r^- is the negative part of $r \in \mathbb{R}$, define diphasic entropies on the domain $D > 0$.

4.2. Invariant regions. In this subsection, we shall prove that the solution $(\mathbf{c}_\varepsilon^1, \mathbf{c}_\varepsilon^2)$ to (4.1) is bounded in L^∞ uniformly in ε, thus giving rise to a weakly convergence subsequence. In the next subsection, we prove that this subsequence actually converges almost everywhere to a solution in the sense of (4.14) below.

THEOREM 4.4. *Assume $\mathbf{c}^0 \in L^1 \cap L^\infty(]0,1[)^N$, $\mathbf{a} \in L^\infty(\mathbb{R}_+)^N$, $\mathbf{b} \in L^\infty(\mathbb{R}_+)^N$, $\mathbf{c}^0, \mathbf{a}$ nonnegative, and $\mathbf{b}$ nonpositive. Let $0 < w^- \le k_1$ satisfy $w^- \le w_1^{\mathbf{a}}(t), w_1^0(x) \le k_1$ for all (t, x). Define*

$$\psi(\xi) \overset{\text{def}}{=} \nabla_{\mathbf{c}} E_0(\xi) \cdot \mathbf{b} + (u + v\xi)\gamma(\xi),$$

and assume that

$$\xi^\star \overset{\text{def}}{=} \inf\{\xi \le k_1; \exists \xi' \le \xi, \psi(\xi') \le 0\} > w^-. \tag{4.11}$$

Let $(\mathbf{c}^1_\varepsilon, \mathbf{c}^2_\varepsilon)$ *be a solution of* (1.1). *Then there exists a constant* C *independent of* ε *such that* $0 \le \mathbf{c}^i_\varepsilon(x,t) \le C$, $1 \le i \le N$, $\forall (t,x) \in [0,T] \times [0,1]$.

Remark 4.4. Once again, one can choose $\xi_0 = w^-$ only if w^- satisfies $\psi(w^-) \le 0$.

Remark 4.5. The existence of ξ^* relies on the nonpositivity of the polynomial ψ on $[0,k_1]$ (one has $\psi(0) = u > 0$ and $\psi(k_1) = -k_1 b_1 \prod_{i>1}(k_i - k_1) \ge 0$, so this is not trivially satisfied). This leads to a condition on $u, v, \mathbf{b}$, and k_1, which is actually not very explicit, except for $N = 1$ (see Remark 3.1). However, one can rewrite things as follows. For $0 < \xi < k_1$, define $\mathbf{c}(\xi) \in \mathbb{R}^N_+$ by

$$c_i(\xi) = \frac{k_i - \xi}{N\xi}, \qquad 1 \le i \le N.$$

Then a few easy algebraic computations prove

$$\nabla_{\mathbf{c}} E_0(\xi) \cdot \mathbf{c}(\xi) = -\gamma(\xi), \quad \nabla_{\mathbf{c}} E_0(\xi) \cdot \mathbf{h}(\mathbf{c}(\xi)) = -\xi\gamma(\xi),$$

so that $\psi(\xi) \le 0$ rewrites $\nabla_{\mathbf{c}} E_0(\xi) \cdot [\mathbf{b} - (u\mathbf{c}(\xi) + \xi\gamma(\xi))] \le 0$. Thus condition (4.11) can be compared to (3.4) in a more consistent way. Notice that this can also be read as an entropy inequality, since $\nabla_{\mathbf{c}} E_0(\xi) \cdot [\mathbf{b} - (u\mathbf{c}(\xi) + \xi\mathbf{h}(\mathbf{c}(\xi)))] = E_0(\xi; \mathbf{b}) - E_0(\xi; u\mathbf{c}(\xi) + \xi\mathbf{h}(\mathbf{c}((\xi))) = F_0(\xi; \mathbf{b}) - F_0(\xi; u\mathbf{c}(\xi) + \xi\mathbf{h}(\mathbf{c}((\xi)))$.

Proof of Theorem 4.1. To lighten the notations a bit, we omit the index ε in this proof. First notice that w^- exists since $\mathbf{a}$ and $\mathbf{c}^0$ are nonnegative and uniformly bounded.

Let us prove first that for a given index i, if $a_i \ge 0$, $b_i \le 0$, and $c_i^0 \ge 0$, then $c_i^j \ge 0$ for $j = 1, 2$. For this purpose we make use of the entropy introduced in Remark 4.3. Inequality (2.4) can be rewritten here as

$$\frac{d}{dt}\int_0^1 \left([c_i^1(x,t)]^- + [c_i^2(x,t)]^-\right) dx \le [c_i^1(0,t)]^- + v[c_i^2(0,t)]^- - u[c_i^1(1,t)]^- - v[c_i^2(1,t)]^-.$$

Now, as in the scalar case, we notice that $v < 0$ and $(c_i^j)^- \ge 0$, so $u[c^1(0,t)]^- + v[c^2(0,t)]^- \le u[a_i(t)]^-$ by the boundary condition at $x = 0$. Since $a_i(t) \ge 0$ for $1 \le i \le N$, $[a_i(t)]^- = 0$, and the same occurs for the initial data.

For $x = 1$, we have to prove that $F \overset{\text{def}}{=} -u[c^1(1,t)]^- - v[c^2(1,t)]^- \le 0$. This clearly occurs if $(b_i - uc_i^1(1,t))/v \ge 0$. If this is not the case, we have $0 \ge b_i(t) \ge uc_i(t)$ since $v < 0$ so that $F = b_i(t)|\beta_i| \le 0$. Hence the following differential inequality holds:

$$\frac{d}{dt}\int_0^1 \left([c^1(x,t)]^- + [c^2(x,t)]^-\right) dx \le 0.$$

The conclusion now follows easily: the components of $\mathbf{c}^1_\varepsilon$ and $\mathbf{c}^2_\varepsilon$ remain nonnegative for any $t > 0$.

We turn now to the proof of the upper bound. For simplicity, we assume the nonnegativity. In view of formula (iv) in Lemma 4.1, we have to prove that there exists $\xi_0 > 0$ such that $w_1^j \ge \xi_0$ for all (t,x). We consider the diphasic entropy (S^1, S^2),

$$S^1(\mathbf{c}^1) = \int_{w_1^1}^{w_2^1} |E_0(\xi; \mathbf{c}^1)| g(\xi) d\xi, \quad S^2(\mathbf{c}^2) = \int_{w_1^2}^{w_2^2} |F_0(\xi; \mathbf{c}^2)| g(\xi) d\xi.$$

The usual trick of convexity of S^1 and S^2 leads to

$$\frac{d}{dt}\int_0^1 [S^1(\mathbf{c}^1(x,t))+S^2(\mathbf{c}^2(x,t))]\,dx \leq -[uS^1(\mathbf{c}^1(x,t))+vS^2(\mathbf{c}^2(x,t))]\Big|_{x=0}^{x=1}. \tag{4.12}$$

Set $H_0 = uS^1(\mathbf{c}^1(0,t))+vS^2(\mathbf{c}^2(0,t))$ and $H_1 = -[uS^1(\mathbf{c}^1(1,t))+vS^2(\mathbf{c}^2(1,t))]$. We have

$$H_0 = \int_{w_1^1}^{w_2^1} u|E_0(\xi;\mathbf{c}^1)|g(\xi)d\xi + \int_{w_1^2}^{w_2^2} v|F_0(\xi;\mathbf{c}^2)|g(\xi)d\xi \leq \int_{w_1^{\mathbf{a}}}^{w_2^{\mathbf{a}}} u|E_0(\xi;\mathbf{a})|g(\xi)d\xi,$$

since $v<0$. For any $\xi_0 \leq w^-$, choosing $g=\delta_{\xi_0}$ cancels the right-hand side of the preceding inequality.

Concerning H_1, we want to take $g=\delta_{\xi_0}$ for a carefully chosen $\xi_0 \leq w^-$ such that

$$H_1 = -\int_{w_1^1}^{w_2^1} u|E_0(\xi;\mathbf{c}^1)|\delta_{\xi_0}(\xi)d\xi - \int_{w_1^2}^{w_2^2} v\left|F_0\left(\xi;\frac{1}{v}[\mathbf{b}-u\mathbf{c}^1]\right)\right|\delta_{\xi_0}(\xi)d\xi \leq 0. \tag{4.13}$$

We know, since everything is nonnegative, that $0<w_1^1,w_1^2\leq k_1\leq w_2^1,w_2^2$, so that necessarily $\xi_0\leq w_2^1,w_2^2$. Now, if $\xi_0<w_1^2$, then $H_1\leq 0$ by (4.13). If $\xi_0\geq w_1^2$, we have by construction $F_0(\xi_0;(\mathbf{b}-u\mathbf{c}^1)/v)\leq 0$ (indeed one can check that $F_0(\xi=0)=0$ and $\partial_\xi F_0(\xi=0)\geq 0$). On the other hand, an easy computation shows

$$vF_0\left(\xi;\frac{1}{v}[\mathbf{b}-u\mathbf{c}^1]\right) = \psi(\xi_0) - uE_0(\xi_0;\mathbf{c}^1).$$

Since $w^-\geq\xi^\star$, one can choose any $\xi^\star\leq\xi_0\leq w^-$ such that $\psi(\xi_0)\leq 0$. The preceding equality therefore gives $E_0(\xi_0;\mathbf{c})\leq 0$, so that $\xi_0\in[w_{2p+1}^1,w_{2p+2}^1]$ for some $p\geq 1$, by assertion (iv) in Lemma 4.1. Since $\xi_0\leq k_1\leq w_2^1$, necessarily $\xi_0\in[w_1^1,w_2^1]$ so that finally, H_1 can be rewritten, by simple consideration of sign on E_0 and F_0,

$$H_1 = uE_0(\xi_0;\mathbf{c}^1) + vF_0\left(\xi_0;\frac{1}{v}[\mathbf{b}-u\mathbf{c}^1]\right) = \psi(\xi_0)\leq 0.$$

The preceding choice of ξ_0 cancels the right-hand side of (4.12). When integrating in t, we introduce the initial data, but the choice of $g=\delta_{\xi_0}$ for $\xi_0\leq w^-$ gives also $S^1(\mathbf{c}^0(x)) = S^2(\mathbf{h}(\mathbf{c}^0(x))=0$, so finally (4.12) gives

$$\int_0^1 [S^1(\mathbf{c}^1(x,t))+S^2(\mathbf{c}^2(x,t))]\,dx \leq 0,$$

which leads to $S^1(\mathbf{c}^1(x,t))=S^2(\mathbf{c}^2(x,t))=0$, $\forall t>0$. A simple contradiction argument then gives $w_1^1(x,t)\geq\xi_0$ and $w_1^2(x,t)\geq\xi_0$ for a.e. x, $\forall t>0$. □

4.3. Strong convergence. The L^∞ estimate leads obviously to the following weak convergence result: $\mathbf{c}_\varepsilon^2 - h(\mathbf{c}_\varepsilon^1)$ tends to 0 in $\mathcal{D}'(\Omega)^N$ when ε tends to 0. We actually have a stronger convergence result.

LEMMA 4.5. *Under the above assumptions ensuring the L^∞ bounds on the solution, $\mathbf{c}_\varepsilon^2 - h(\mathbf{c}_\varepsilon^1)$ tends to 0 in $L^2_{\mathrm{loc}}(\Omega)^N$.*

From (1.1), we can obtain the following inequality for the entropies (χ_i^1,χ_i^2):

$$\partial_t\left(\chi_i^1(\xi;\mathbf{c}_\varepsilon^1)+\chi_i^2(\xi;\mathbf{c}_\varepsilon^2)\right) + \partial_x\left(u\chi_i^1(\xi;\mathbf{c}_\varepsilon^1)+v\chi_i^2(\xi;\mathbf{c}_\varepsilon^2)\right)\leq 0.$$

The negative sign holds since $\nabla_{\mathbf{c}}\chi_i^2(\xi;\cdot)$ is a monotone operator, as before. Now, multiply this inequality by any nonnegative $\varphi \in \mathcal{D}(\overline{\Omega})$, integrate by parts, and treat the boundary conditions as in the above proof. One obtains

$$\begin{aligned}
&-\int_0^T\int_0^1 \left[\partial_t\varphi\left(\chi_i^1(\xi;\mathbf{c}_\varepsilon^1(x,t))+\chi_i^2(\xi;\mathbf{c}_\varepsilon^2(x,t))\right)\right.\\
&\left.+\partial_x\varphi\left(u\chi_i^1(\xi;\mathbf{c}_\varepsilon^1(x,t))+v\chi_i^2(\xi;\mathbf{c}_\varepsilon^2(x,t))\right)\right]\,dx\,dt\\
&\le \int_0^T \varphi(0,t)u\chi_i^1(\xi;\mathbf{a}(t))\,dt-\int_0^1\varphi(x,0)S(\mathbf{c}^0(x))\,dx\\
&\quad-\int_0^T\left[u\chi_i^1(\xi;\mathbf{c}_\varepsilon^1(1,t))+v\chi_i^2(\xi;\mathbf{c}_\varepsilon^2(1,t))\right]\varphi(1,t)\,dt.
\end{aligned}$$

Once again, some considerations of sign allow us to prove that for the boundary term on $x=1$, we have for any ξ, since $\mathbf{c}_\varepsilon^2(1,t)=(\mathbf{b}(t)-u\mathbf{c}_\varepsilon^1(1,t))/v$,

$$u\chi_i^1(\xi;\mathbf{c}_\varepsilon^1(1,t))+v\chi_i^2(\xi;\mathbf{c}_\varepsilon^2(1,t))\le|\nabla_{\mathbf{c}}E_0(\xi)\cdot\mathbf{b}(t)-(u+v\xi)\gamma(\xi)|\stackrel{\text{def}}{=}B(t).$$

The resulting entropy estimate is analogous to (3.7). Now, following the lines of [14], we can apply compensated compactness to obtain the following result of strong convergence.

THEOREM 4.6. *We make the same assumptions as in Theorem* 4.1. *Then there exists a subsequence of solutions to* (1.1), *still denoted by* $\mathbf{c}_\varepsilon^1$, *which converges almost everywhere and strongly in* $]0,1[\times]0,T[$ *to* $\mathbf{c}\in L^\infty(]0,T[;L^1(]0,1[))^N$. *Moreover,* $\mathbf{c}$ *satisfies, for any* $\varphi\in\mathcal{D}(\overline{\Omega})$, $\varphi\ge0$, $\xi>0$,

(4.14)
$$\begin{aligned}
&-\int_0^T\int_0^1\left[S(\mathbf{c})\partial_t\varphi+Q(\mathbf{c})\partial_x\varphi\right]\,dx\,dt\\
&\quad\le\int_0^T u\chi_i^1(\mathbf{a}(t))\varphi(0,t)\,dt-\int_0^T B(t)\varphi(1,t)\,dt-\int_0^1 S(\mathbf{c}^0(x))\varphi(x,0)\,dx,
\end{aligned}$$

with $B(t)=|\nabla_{\mathbf{c}}E_0(\xi)\cdot\mathbf{b}(t)-(u+v\xi)\gamma(\xi)|$, *for* $S(\mathbf{c})=\chi_i^1(\mathbf{c})+\chi_i^2(\mathbf{h}(\mathbf{c}))$, $Q(\mathbf{c})=u\chi_i^1(\mathbf{c})+v\chi_i^2(\mathbf{h}(\mathbf{c}))$, *and* χ_i^j *being defined by* (4.10).

Proof of Lemma 4.3. Consider the pair of entropies (η_1,η_2) obtained by choosing, for a given i, $g=\mathbb{1}_{[0,k_i]}$. Their gradients are given by

$$\begin{aligned}
\nabla_{\mathbf{c}}\eta_1(\mathbf{c}^1)&=\int_{w_i^1}^{k_i}\operatorname{sign}(E_0(\xi;\mathbf{c}^1))\nabla_{\mathbf{c}}E_0(\xi)d\xi,\\
\nabla_{\mathbf{c}}\eta_2(\mathbf{c}^2)&=\int_{w_i^2}^{k_i}\operatorname{sign}(F_0(\xi;\mathbf{c}^2))\nabla_{\mathbf{c}}E_0(\xi)d\xi.
\end{aligned}$$

Omitting here the dependence in ε, we take the scalar product of the two equations in (1.1), respectively, by $\nabla_{\mathbf{c}}\eta_1(\mathbf{c}^1)$ and $\nabla_{\mathbf{c}}\eta_2(\mathbf{c}^2)$, sum the two equations, and integrate $dx\,dt$ with a nonnegative test function $\varphi\in\mathcal{D}(\overline{\Omega})$. We obtain, after integration by

parts and multiplication by ε,

$$A^\varepsilon \stackrel{\text{def}}{=} \varepsilon \int_0^T \int_0^1 \left(\partial_t \varphi(x,t)\left[\mathbf{c}^1(x,t)+\mathbf{c}^2(x,t)\right] + \partial_x \varphi(x,t)\left[u\mathbf{c}^1(x,t)+v\mathbf{c}^2(x,t)\right]\right) dx\,dt$$
$$= -\int_0^T \int_0^1 \left[\int_{w_i^1}^{k_i} \operatorname{sign}(E_0(\xi;\mathbf{c}^1))\nabla_{\mathbf{c}} E_0(\xi) d\xi - \int_{w_i^2}^{k_i} \operatorname{sign}(F_0(\xi;\mathbf{c}^2))\nabla_{\mathbf{c}} E_0(\xi) d\xi\right] \cdot (\mathbf{c}^2 - \mathbf{h}(\mathbf{c}^1)) dx\,dt.$$

Notice that $A^\varepsilon \geq 0$ by the second equality and the convexity of η_i. Obviously, since $\mathbf{c}^1$ and $\mathbf{c}^2$ are bounded in L^∞, A^ε tends to 0 when ε goes to zero. We have to work from now on with

$$P(x,t) \stackrel{\text{def}}{=} -\left[\int_{w_i^1}^{k_i} \operatorname{sign}(E_0(\xi;\mathbf{c}^1))\nabla_{\mathbf{c}} E_0(\xi) d\xi - \int_{w_i^2}^{k_i} \operatorname{sign}(F_0(\xi;\mathbf{c}^2))\nabla_{\mathbf{c}} E_0(\xi) d\xi\right] \cdot (\mathbf{c}^2 - \mathbf{h}(\mathbf{c}^1)).$$

It is easy to check that $\operatorname{sign} E_0(\xi;\mathbf{c}^1) = \operatorname{sign} F_0(\xi,\mathbf{h}(\mathbf{c}^1)) = \operatorname{sign} F_0(\xi;\mathbf{c}^2)$ for $\xi \in [w_i^1,k_i] \cap [w_i^2,k_i]$. We are thus left with an integral over $[\min(w_i^1,w_i^2),\max(w_i^1,w_i^2)]$. Let us assume that $w_i^1 \leq w_i^2$; the computations are the same if the converse holds. We have, by considerations of sign on F_0,

$$P(x,t) = 2\int_{w_i^1}^{w_i^2} \left[|F_0(\xi;\mathbf{c}^2)| + |F_0(\xi;\mathbf{h}(\mathbf{c}^1))|\right] d\xi.$$

Now, we write for $N \geq 4$,

$$|F_0(\xi;\mathbf{c}^2)| = (\xi - w_{i-1}^2)(\xi - w_i^2)(w_{i+1}^2 - \xi)\frac{\prod_{j\notin\{i-1,i,i+1\}} |w_j^2 - \xi|}{\prod_{i=1}^N w_j^2}.$$

The fourth term is greater than some $K > 0$ (K depending on $k_1,\dots,k_N$ and ξ_0), since either $w_j^2 \leq k_{i-2}$ or $w_j^2 \geq k_{i+1}$, and $k_{i-1} \leq \xi \leq k_i$. For the first three terms, we simply write $(\xi - w_{i-1}^2)(\xi - w_i^2)(w_{i+1}^2 - \xi) \geq (\xi - w_i^2)^2(w_{i+1}^2 - w_i^2)$, which leads by integration to

$$\int_{w_i^1}^{w_i^2} |F_0(\xi;\mathbf{c}^2)| d\xi \geq \frac{K}{3}(w_i^1 - w_i^2)^3(w_{i+1}^2 - w_i^2) \geq \frac{K}{3}(w_i^1 - w_i^2)^4.$$

For $N = 3$, we have a similar estimate, since the fourth term reduces to $K/(w_1^2 w_2^2 w_3^2)$. Because the same holds for $|F_0(\xi;\mathbf{h}(\mathbf{c}^1))|$, we have finally that for some $C > 0$, depending only on $k_1,\dots,k_N$,

$$\int_0^T \int_0^1 |w_i^1(x,t) - w_i^2(x,t)|^4 \varphi(x,t) dx\,dt \leq C\int_0^T \int_0^1 P(x,t)\varphi(x,t) dx\,dt = A^\varepsilon$$

tends to zero. Thus $|w_i^1 - w_i^2|$ tends to 0 in $L^4_{\text{loc}}(\Omega)$ and therefore in $L^2_{\text{loc}}(\Omega)$. For $N = 2$, the same computations lead to convergence in $L^3_{\text{loc}}(\Omega)$ and hence in $L^2_{\text{loc}}(\Omega)$. Finally, if $N = 1$, we directly obtain $L^2_{\text{loc}}(\Omega)$. Since the function $(w_1,\dots,w_N) \mapsto (c_1,\dots,c_N)$ is Lipschitz continuous, we are done. ∎

Proof of Theorem 4.2. We merely give the sketch of the proof, referring to [14] for the detailed computations, which are identical. Summing the equations for $0 \le i \le N$, we obtain, with the same notations as in the scalar case,

(4.15)
$$T^\varepsilon(\xi) = \partial_t[G_0(\xi, \mathbf{c}^1) + H_0(\xi, \mathbf{h}(\mathbf{c}^1))] + \partial_x[uG_0(\xi, \mathbf{c}^1) + vH_0(\xi, \mathbf{h}(\mathbf{c}^1))] = \mu^\varepsilon(\xi) + g^\varepsilon(\xi),$$

with $G_0 = |E_0|$ and $H_0 = |F_0|$. Since η_1 and η_2 are convex, the usual computation proves that $\mu^\varepsilon(\xi)$ is a nonpositive measure. By Lemma 4.3, $g^\varepsilon(\xi)$ is compact in $H^{-1}_{\mathrm{loc}}(\Omega)$; thus, again applying Murat's lemma, we can apply the compensated compactness lemma to (4.15), for two different values ξ and ξ'. We obtain, after some easy simplifications,

$$\overline{G_0(\xi)}\ \overline{\xi' G_0(\xi')/D} - \overline{G_0(\xi')}\ \overline{\xi G_0(\xi)/D} \ = \ (\xi' - \xi)\overline{G_0(\xi)G_0(\xi')/D}.$$

Dividing by $\overline{G_0(\xi)}\ \overline{G_0(\xi')}\ (\xi' - \xi)$ and letting ξ' go to ξ, we get

$$\partial_\xi \frac{\xi\overline{G_0(\xi)/D}}{\overline{G_0(\xi)}} = \frac{\overline{G_0(\xi)^2/D}}{\overline{G_0(\xi)}^2}. \tag{4.16}$$

Of course (4.16) has to be justified at points where $\overline{G_0(\xi)} = 0$. This occurs when $G_0(\xi_0, w) = 0$ for all w in the support of ν, that is, $w_j = \xi_0$ for some j. If w_j is a simple eigenvalue, the formula is justified by applying l'Hospital's rule in a neighborhood of ξ_0 to G_0', which is not zero since the root is simple. When we have a double root, that is, $\xi_0 = k_j$, the same technique can be used with G_0'', which in no case can be zero, since the root cannot be triple.

Equation (4.16) is not completely satisfactory because its right-hand side does not vanish when $d\nu$ is a Dirac mass. Therefore, we again apply compensated compactness to (4.15) for a given ξ and

$$\partial_t\left(D + \frac{\alpha}{D}\right) + \partial_x\left(uD + v\frac{\alpha}{D}\right) = 0, \qquad \alpha = k_1u_1 + \cdots + k_Nu_N.$$

This yields $\overline{G_0(\xi)}\ \overline{\alpha/D} - \xi\overline{G_0(\xi)/D}\ \overline{D} = \overline{G_0(\xi)\alpha/D} - \xi\overline{G_0(\xi)}$. After dividing it by $\overline{D}\,\overline{G_0(\xi)}$, we can combine it with the left-hand side of (4.15) to get

$$-\partial_\xi \frac{\overline{G_0(\xi)\alpha/D}}{\overline{G_0(\xi)}} = \frac{\overline{G_0(\xi)^2/D}}{\overline{G_0(\xi)}^2} - \frac{1}{\overline{D}} \ \ge 0, \tag{4.17}$$

this last inequality being just Cauchy–Schwarz. Inequality (4.17) is the keystone to proving that $d\nu$ is in fact a Dirac mass.

Indeed, if in (4.17) the inequality is strict at some point, we obtain a contradiction by comparing the values of the nonincreasing function

$$\frac{\overline{G_0(\xi)\alpha/D}}{\overline{G_0(\xi)}}$$

at the points $\xi = 0$ and $\xi = +\infty$, then $\xi = k_i$ and $\xi = \infty$. This means that the inequality in (4.17) is just an equality for all $\xi \ge 0$, so that the equality case in Cauchy–Schwarz applies. We obtain the existence of a function $\lambda(\xi)$ such that, for all $\xi \ge 0$, $G_0(\xi, \mathbf{c}^1) = \lambda(\xi)D(\mathbf{c}^1)$ a.e. in the support of $d\nu(\mathbf{c}^1)$. From this we deduce

that, for two possible elements $\mathbf{c}^1$, $\mathbf{c}^{'1}$ of the support of $d\nu(\mathbf{c}^1)$, we have necessarily $G_0(\xi, \mathbf{c}^1) = G_0(\xi, \mathbf{c}^{'1})$. Thus $\sigma_j(\mathbf{c}^1) = \sigma_j(\mathbf{c}^{'1})$, and by Lemma 4.1 (v) this proves that $\mathbf{c}^1 = \mathbf{c}^{'1}$ and the support of ν is a single point. ∎

Remark 4.6. Notice that formula (4.14) is exactly the kinetic formulation obtained in [14], but the boundary terms forbid us to write it in the usual way, with some nonnegative measure on the right-hand side.

5. Boundary conditions. So far, we have defined in Theorems 3.2 and 4.2 kinds of weak solutions. The aim of this section is to prove that these solutions are actually solutions to (1.2) in the sense of distributions, and to give a meaning to the reflux boundary condition at $x = 1$. It seems that we lose the Dirichlet-like boundary condition at $x = 0$ when passing to the limit. This is not really surprising, since we pass from $2N$ equations to N equations: the system becomes overdetermined.

Before precisely stating our results, we need to introduce some material. Indeed, we want to precisely state the meaning of the boundary conditions. But we deal with L^∞ functions, which usually do not have any trace on the boundary. The following result, which we state as a lemma, follows easily by choosing the test functions $\varphi \in \mathcal{D}(\Omega)$ in (3.5) or (4.14).

LEMMA 5.1. *Let (η_1, η_2) be any pair of convex functions defining a diphasic entropy. Let $\mathbf{c} \in L^\infty(\Omega)$ be a weak solution as in Theorems* 3.2 *or* 4.2. *Then the vector-valued function $\psi = (\psi_1, \psi_2) \overset{\text{def}}{=} (\eta_1(\mathbf{c}) + \eta_2(\mathbf{h}(\mathbf{c})), u\eta_1(\mathbf{c}) + v\eta_2(\mathbf{h}(\mathbf{c})))$ is in $L^\infty(\Omega)$, and $\operatorname{div}\psi = \partial_t\psi_1 + \partial_x\psi_2$ is a nonnegative measure in Ω.*

We are thus in a position to apply a result by Anzellotti [1, Theorems 1.2 and 1.9], which essentially states that ψ has a trace on $\partial\Omega$, in some sense. We recall this result here without proof.

THEOREM 5.2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with locally Lipschitz boundary $\partial\Omega$. Set $X(\omega) = \{\psi \in L^\infty(\Omega; \mathbb{R}^n); \operatorname{div}\psi$ is a bounded measure in $\Omega\}$. Then there exists a trace operator*

$$\gamma : X(\Omega) \to L^\infty(\partial\Omega),$$

such that, for any $\varphi \in BV(\Omega) \cap L^\infty(\Omega) \cap C^0(\Omega)$,

$$\int_\Omega \varphi \operatorname{div}\psi \, dx + \int_\Omega (\psi, \varphi) \, dx = \int_{\partial\Omega} \gamma\psi \, \varphi \, d\sigma, \tag{5.1}$$

where σ is the superficial measure on $\partial\Omega$.

In this result, (ψ, φ) has to be defined as a measure (Definition 1.4 in [1]). We denote by γ^0 the trace on $]0,1[\times\{0\}$, by γ_0 and γ_1 the traces, respectively, on $\{0\}\times]0,T[$ and $\{1\}\times]0,T[$. Since $\gamma\psi$ is, by construction, a weak trace on $\partial\Omega$ of the normal component of ψ, we have

at $t = 0$, $\quad \gamma\psi = \gamma^0[\eta_1(\mathbf{c}) + \eta_2(\mathbf{h}(\mathbf{c}))]$,

at $x = 0, 1$, $\quad \gamma\psi = \gamma_{0,1}[u\eta_1(\mathbf{c}) + v\eta_2(\mathbf{h}(\mathbf{c}))]$.

In particular, for the trivial entropies, we recover the conservative variables so that, for $1 \le i \le N$, $c_i + h_i(\mathbf{c})$ has a trace on $t = 0$, and $uc_i + vh_i(\mathbf{c})$ has traces on $x = 0$ and $x = 1$. Notice that this trace is attained in a weak sense (see [18]), in contrast with the traces of BV functions, which are attained in L^1.

THEOREM 5.3. (i) *Let $\mathbf{c}$ be a solution as in Lemma* 5.1. *Then it is a solution to* (1.2) *in $\mathcal{D}'(\Omega)$, and we have, for $1 \le i \le N$,*

$$\begin{cases} \gamma_1[uc_i + vh_i(\mathbf{c})] & = b_i, \quad \text{a.e. } t \in]0,T[; \\ \gamma^0[c_i + h_i(\mathbf{c})] & = c_i^0 + h_i(\mathbf{c}^0), \quad \text{a.e. } x \in]0,1[. \end{cases} \tag{5.2}$$

(ii) *For any pair* (η_1, η_2) *denoting the Kružkov entropies in the scalar case, the kinetic entropies for the Langmuir system, define* ψ *as in Lemma* 5.1. *Then the following entropy inequalities hold for a.e.* $t \in]0,T[$:

$$\left\{ \begin{array}{ll} \gamma_0[u\eta_1(\mathbf{c}) + v\eta_2(\mathbf{h}(\mathbf{c}))] & \leq u\eta_1(\mathbf{a}), \\ \gamma_1[u\eta_1(\mathbf{c}) + v\eta_2(\mathbf{h}(\mathbf{c}))] & \leq B(t), \end{array} \right. \tag{5.3}$$

where $B(t) = |b(t) - f(k)|$ *in the scalar case and is defined in Theorem* 4.2 *for the Langmuir system.*

Remark 5.1. This theorem shows that the initial condition and the reflux boundary condition are satisfied in a strong sense (in $L^\infty(\partial\Omega)$, actually). We have no information about the input boundary condition at $x = 0$, except for the entropy inequalities (5.3). Notice that, even for the conservative variables themselves, we loose some information. Indeed, we know that there is a trace for $u\mathbf{c} + v\mathbf{h}(\mathbf{c})$ at $x = 0$, but this function is not one-to-one, so we cannot compare $\mathbf{c}$ to $\mathbf{a}$. Moreover, even if $u\mathbf{c} + v\mathbf{h}(\mathbf{c})$ is one-to-one, a boundary layer phenomenon will very likely occur here, as the following easy computation shows.

Consider a stationary solution to (1.1) in the scalar case, for a linear function $f(c) = (u + vk)c$, with $k > u/|v|$. The system boils down to the single ordinary differential equation

$$\frac{dc}{dx} = \frac{1}{\varepsilon uv}[b - f(c)], \qquad c(0) = a.$$

There exists a unique equilibrium point c^* such that $f(c^*) = b$, and it is attractive. The solution c_ε is computed explicitly:

$$c_\varepsilon(x) = \frac{b}{u+kv} + \left(a - \frac{b}{u+kv}\right) \exp\left(-\frac{u+kv}{\varepsilon uv}x\right) = c^* + (a - c^*) \exp\left(-\frac{u+kv}{\varepsilon uv}x\right).$$

Obviously, the trace of the limit solution is c^*, which has no reason to coincide with a. We do not wish to investigate this boundary layer now, and leave it for future work.

Proof of Theorem 5.2. To prove part (i) of the theorem, we sum the two equations in (1.1), which gives the conservation of matter, and proceed exactly as in the proof of the convergence theorems. Provided we choose a test function $\varphi \in \mathcal{D}(]0,1] \times [0,T[)$, that is, if the test function does not see the boundary condition at $x = 0$, we obtain a weak formulation with an equality sign:

$$\begin{aligned} -\int_0^1 \int_0^T [(\mathbf{c} + \mathbf{h}(\mathbf{c}))\partial_t\varphi + (u\mathbf{c} + v\mathbf{h}(\mathbf{c}))\partial_x\varphi]dx\,dt \\ = \int_0^1 [\mathbf{c}^0 + \mathbf{h}(\mathbf{c}^0)]\varphi(x,0)dx - \int_0^T \mathbf{b}(t)\varphi(1,t)dt, \end{aligned} \tag{5.4}$$

since the boundary condition at $x = 1$ is satisfied exactly.

As a first consequence, we obtain, by taking $\varphi \in \mathcal{D}(\Omega)$, that $\mathbf{c}$ is actually a solution to (1.2) in $\mathcal{D}'(\Omega)$. Therefore we can apply (5.1) with $\psi = (c_i + h_i(\mathbf{c}), uc_i + vh_i(\mathbf{c}))$, $1 \leq i \leq N$, and $\varphi \in \mathcal{D}(]0,1] \times [0,T[)$. The left-hand side of (5.4) is exactly $\int_\Omega (\psi, \varphi)$, and $\operatorname{div} \psi = 0$, so we are left with

$$\begin{aligned} \int_0^1 \gamma^0[c_i + h_i(\mathbf{c})]\varphi(x,0)dx - \int_0^T \gamma_1[uc_i + vh_i(\mathbf{c})]\varphi(1,t)dt \\ = \int_0^1 [c_i^0 + h_i(\mathbf{c}^0)]\varphi(x,0)dx - \int_0^T b_i(t)dt. \end{aligned}$$

Since this holds for any φ, we obtain (5.2).

Now, (5.3) follows from (3.5) or (4.14). By Lemma 5.1, for any pair (η_1, η_2), $\psi = (\eta_1(\mathbf{c})+\eta_2(\mathbf{h}(\mathbf{c})), u\eta_1(\mathbf{c})+v\eta_2(\mathbf{h}(\mathbf{c})))$ satisfies that div ψ is a nonnegative measure. Thus we can apply (5.1) in both formulae, with $\varphi \in \mathcal{D}([0,1] \times [0,T[)$, and obtain

$$\int_0^1 \gamma^0[\eta_1(\mathbf{c}) + \eta_2(\mathbf{h}(\mathbf{c}))]\varphi(x,0)dx - \int_0^T \gamma_1[u\eta_1(\mathbf{c}) + v\eta_2(\mathbf{h}(\mathbf{c}))]\varphi(1,t)dt \leq \int_0^1 [\eta_1(\mathbf{c}^0) + \eta_2(\mathbf{h}(\mathbf{c}^0))]\varphi(x,0)dx - \int_0^T B(t)dt.$$

Since this holds for any φ, we obtain (5.3). □

Acknowledgment. The author wishes to thank Benoît Perthame for numerous helpful discussions and remarks on this paper.

REFERENCES

[1] G. Anzellotti, *Pairings between measures and bounded functions and compensated compactness*, Ann. Mat. Pura Appl., 135 (1983), pp. 293–318.

[2] D. Aregba-Driollet and R. Natalini, *Convergence of relaxation for conservation law*, Appl. Anal., 61 (1996), pp. 163–193.

[3] C. Bardos, A. Y. Leroux, and J.-C. Nédélec, *First order quasilinear equations with boundary conditions.*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[4] A. Benabdallah and D. Serre, *Problèmes aux limites pour des systèmes hyperboliques non linéaires de deux équations à une dimension d'espace*, C. R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 677–680.

[5] E. Canon and F. James, *Resolution of the Cauchy problem for several hyperbolic systems arising in chemical engineering*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 3 (1992), pp. 219–238.

[6] E. Canon and F. James, *Resolution of the Cauchy problem for non strictly hyperbolic systems arising in chemical engineering*, in Proceedings 3rd International Conference on Hyperbolic Problems, Uppsala, 1990, B. Engquist and B. Gustafsson, eds., Chartwell Bratt, Studentlitteratur, Lund, Sweden, 1991, pp. 212–225.

[7] G. Q. Chen, D. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[8] F. Dubois and P. LeFloch, *Boundary conditions for nonlinear hyperbolic systems of conservation laws*, J. Differential Equations, 71 (1988), pp. 93–122.

[9] M. Gisclon, *Comparaison de deux perturbations singulières pour l'équation de Burgers avec conditions aux limites*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 1011–1014.

[10] M. Gisclon, *Etude des conditions aux limites pour un système strictement hyperbolique, via l'approximation parabolique*, J. Math. Pures Appl., 75 (1996), pp. 485–508.

[11] M. Gisclon and D. Serre, *Conditions aux limites pour un système strictement hyperbolique fournies par le schéma de Godunov*, Mod. Math. Anal. Num., 31 (1997), pp. 359–380.

[12] J. B. Goodman, *Initial Boundary Value Problems for Hyperbolic Systems of Conservation Laws*, Ph.D. Thesis, Stanford University, 1981.

[13] A. Heibig, *Existence and uniqueness of solutions for some hyperbolic systems of conservation laws*, Arch. Rat. Mech. Anal., 126 (1994), pp. 79–101.

[14] F. James, Y.-J. Peng, and B. Perthame, *Kinetic formulation of the chromatography and some other hyperbolic systems*, J. Math. Pures Appl., 74 (1995), pp. 367–385.

[15] F. James, M. Sepúlveda, and P. Valentin, *Statistical thermodynamics models for multicomponent diphasic isothermal equilibria*, Math. Models Methods Appl. Sci., 7 (1997), pp. 1–29.

[16] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–277.

[17] K. T. Joseph and P. G. LeFloch, *Boundary Layers in Weak Solutions to Hyperbolic Conservation Laws*, preprint, C.M.A.P. 341, École Polytechnique, 1996.

[18] P. T. Kan, M. Santos, and Z. Xin, *Initial boundary value problem for conservation laws*, Comm. Math. Phys., 186 (1997), pp. 701–730.

[19] M. Katsoulakis and A. Tzavaras, *Contractive relaxation systems and the scalar multidimensional conservation law*, Comm. Partial. Differential Equations, 22 (1997), pp. 195–233.
[20] I. Langmuir, *The adsorption of gases on plane surfaces of glass, mica and platinum,* J. Amer. Chem. Soc., 40 (1918), pp. 1361–1403.
[21] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.
[22] R. Natalini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 795–823.
[23] B. Perthame and E. Tadmor, *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys., 136 (1991), pp. 501–517.
[24] H. K. Rhee, R. Aris, and N. R. Amundson, *On the theory of multicomponent chromatography*, Phil. Trans. Roy. Soc. London, Ser. A, 267 (1970), pp. 419–455.
[25] H. K. Rhee, R. Aris, and N. R. Amundson, *Multicomponent exchange in continuous countercurrent exchangers*, Phil. Trans. Roy. Soc. London, Ser. A, 269 (1971), pp. 187–215.
[26] D. Serre, *Richness and the classification of quasilinear hyperbolic systems*, in Multidimensional Hyperbolic Problems and Computations, Math. Appl. 29, J. Glimm and A. Majda, eds., Springer-Verlag, Heidelberg, 1991, pp. 315–333.
[27] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Heriot-Watt Sympos., vol. 4, Pitman Res. Notes in Math. Ser. 39, R. J. Knopps, ed., Pitman Press, London, Boston, 1975, pp. 136–211.
[28] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. AMS, 280 (1983), pp. 781–795.
[29] A. Tveito and R. Winther, *On the rate of convergence to equilibrium for a system of conservation laws including a relaxation term*, SIAM J. Math. Anal., 28 (1997), pp. 136–161.
[30] W. C. Wang and Z. Xin, *Asymptotic limit of the initial boundary value problem for conservation laws with relaxational extensions*, Comm. Pure Appl. Math., 1998, to appear.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1224–1237, September 1998

010

SOME INEQUALITIES FOR THE GROWTH OF ELLIPTIC INTEGRALS*

S.-L. QIU†, M. K. VAMANAMURTHY‡, AND M. VUORINEN§

Abstract. Growth of the complete elliptic integral $\mathcal{K}(r)$ near the singularity $r = 1$ is studied and new bounds in terms of elementary functions are obtained. In particular, two recent conjectures are solved and various convexity properties of $\mathcal{K}(r)$ are given.

Key words. concave, convex, elliptic integrals, modulus, ring

AMS subject classifications. Primary, 33E05; Secondary, 30C62

PII. S0036141096310491

1. Introduction.

For $r \in [0, 1]$, Legendre's complete elliptic integrals of the first and second kind are defined by

$$\mathcal{K} = \mathcal{K}(r) = \int_0^{\pi/2} (1 - r^2 \sin^2 t)^{-1/2} dt,\ \mathcal{K}' = \mathcal{K}'(r) = \mathcal{K}(r'), \tag{1.1}$$

$$\mathcal{E} = \mathcal{E}(r) = \int_0^{\pi/2} (1 - r^2 \sin^2 t)^{1/2} dt,\ \mathcal{E}' = \mathcal{E}'(r) = \mathcal{E}(r'), \tag{1.2}$$

respectively, where $r' = \sqrt{1 - r^2}$. Note that $\mathcal{K}(0) = \mathcal{E}(0) = \pi/2$ and $\mathcal{K}(1) = \infty$, $\mathcal{E}(1) = 1$.

These integrals have been studied extensively by several authors from different points of view. Asymptotic approximation of symmetric normal forms of these integrals appears in [C, CG]. Behavior of $\mathcal{K}(r)$ near the singularity $r = 1$ has been studied in [K]. Monotonicity properties of certain combinations of $\mathcal{E}(r)$ and $\mathcal{K}(r)$, with applications to Robin capacity, appear in [DP, ADV]. Relationships among these elliptic integrals, the Gauss arithmetic-geometric mean, and other mean values have been studied in [BB2, G, VV].

The present study is motivated by the theory of quasi-conformal maps, where elliptic integrals occur in the formulas for the moduli of plane ring domains (cf. (1.3)). In this context, the following two conjectures were stated in [AVV2, Conjectures 3.1].

Conjecture 1.1. For $r \in (0, 1)$,

$$r'^2 < \frac{r' \exp(\mathcal{K}(r)) - 4}{\exp(\pi/2) - 4} < 2\frac{\sqrt{1 - r}}{2 - r}.$$

*Received by the editors October 14, 1996; accepted for publication (in revised form) September 22, 1997; published electronically June 2, 1998. This work was begun during the first author's visit to the Mathematics Department, University of Helsinki, with support from CIMO and the Finnish Ministry of Education, and completed at Auckland during his visit to the University of Auckland with support from its Department of Mathematics.

http://www.siam.org/journals/sima/29-5/31049.html

†School of Science and Arts, Hangzhou Institute of Electronics Engineering, Hangzhou 310037, People's Republic of China.

‡Department of Mathematics, University of Auckland, P.B. 92019, Auckland, New Zealand (vamanamu@math.auckland.ac.nz).

§Department of Mathematics, P. O. Box 4 (Yliopistonkatu 5), University of Helsinki, Helsinki FIN-00014, Finland (vuorinen@csc.fi).

Conjecture 1.2. For $r \in (0,1)$,

$$\mathcal{K}(r) < \log(1 + (4/r')) - (\log 5 - \pi/2)(1 - r).$$

These conjectures are proved in Theorem 1.7. We shall also establish monotonicity and concavity properties of certain combinations of complete elliptic integrals, from which sharp functional inequalities for these integrals follow. As a by-product, we shall provide affirmative answers to the following two questions.

Question 1.3. Is the function $\mathcal{E}'(r)[\mathcal{K}(r) - \mathcal{E}(r)]/\{r^2[2\log(4/r') - 1]\}$ increasing from (0, 1) onto $(\pi/[4(\log 16 - 1)], \pi/4)$?

Question 1.4. Is the function $[\mathcal{E}(r)\,\mathcal{E}'(r) - \mathcal{K}(r)\,\mathcal{E}'(r) + r^2\,\mathcal{K}(r)\,\mathcal{K}'(r)]/[r^2\,\mathcal{K}'(r)]$ decreasing from $(0,1)$ onto $(1, \pi/2)$?

As an application, sharp lower and upper bounds are obtained for the modulus of the Grötzsch ring $B\backslash[0,r]$ defined by

$$\mu(r) = \frac{\pi\,\mathcal{K}'(r)}{2\,\mathcal{K}(r)}, \quad 0 < r < 1, \tag{1.3}$$

where B is the unit disk in the plane (cf. [LV]).

Throughout this paper, r' denotes $\sqrt{1-r^2}$ for all $r \in [0,1]$, and c denotes the constant $e^{\pi/2} - 4 = 0.81047\ldots$.

We now state some of our main results.

THEOREM 1.5. *There exists a unique $r_0 \in (0,1)$ such that the function $f(r) \equiv \mathcal{K}'(r)/\log((4/r) + cr)$ is strictly increasing on $(0, r_0]$ and decreasing on $[r_0, 1)$, with $f(0^+) = f(1^-) = 1$. In particular, for $r \in (0,1)$,*

$$\mathcal{K}(r) > \log(cr' + (4/r')). \tag{1.4}$$

This inequality is sharp as r tends to 0 or 1.

THEOREM 1.6. *For $a \in (0,\infty)$, define the function f on $[0,1]$ by*

$$f(r) = \log(1 + (a/r')) - \mathcal{K}(r)$$

for $r \in [0,1)$, and $f(1) = \lim_{r\to 1^-} f(r) = \log(a/4)$. Then

(1) For $0 < a \le \pi/(4-\pi)$, f is strictly decreasing from $[0,1]$ onto $[c_1, c_2]$, where $c_1 = \log(a/4)$ and $c_2 = \log(1+a) - \pi/2$. In particular, for $r \in (0,1)$,

$$\log\left(\frac{4}{a} + \frac{4}{r'}\right) - b_1 < \mathcal{K}(r) < \log\left(\frac{4}{a} + \frac{4}{r'}\right),$$

where $b_1 = \log 4 - (\pi/2) + \log(1 + (1/a)) \ge \log(16/\pi) - \pi/2 = 0.05706\ldots$.

(2) For $a > \pi/(4-\pi)$, there exists a unique $r_0 = r_0(a) \in (0,1)$ such that f is strictly increasing on $(0, r_0]$ and decreasing on $[r_0, 1)$. In particular, for $r \in (0,1)$,

$$\mathcal{K}(r) < \log\left(\frac{4}{a} + \frac{4}{r'}\right) + b_2, \tag{1.5}$$

where

$$b_2 = \begin{cases} 0 \text{ if } a \le 4/c, \\ \pi/2 - \log(4 + (4/a)) \text{ if } a > 4/c. \end{cases}$$

The next theorem settles Conjectures 1.1 and 1.2.

THEOREM 1.7. *For $r \in (0,1)$,*

$$r'^2 < \frac{r' \exp(\mathcal{K}(r)) - 4}{\exp(\pi/2) - 4} < r' < 2\frac{\sqrt{1-r}}{2-r}, \tag{1}$$

$$\mathcal{K}(r) < \log(1 + (4/r')) - (\log 5 - \pi/2)(1-r), \tag{2}$$

$$\frac{\pi \log((4/r) + cr)}{2\log((4/r') + c)} < \mu(r) < \frac{\pi \log((4/r) + c)}{2\log((4/r') + cr')}. \tag{3}$$

The inequalities in (1) *are sharp as r tends to* 0, *while those in* (2) *and* (3) *are sharp as r tends to* 0 *or* 1.

The next result answers the Question 1.3 in the affirmative.

THEOREM 1.8. *The function $f(r) \equiv \mathcal{E}'(r)[\mathcal{K}(r) - \mathcal{E}(r)]/\{r^2[2\log(4/r') - 1]\}$ is strictly increasing from* (0, 1) *onto* $(\pi/(4a), \pi/4)$, *where $a = 4\log 2 - 1$.*

Our next result shows that the answer to the Question 1.4 is also affirmative.

THEOREM 1.9. *The function*

$$f(r) \equiv \frac{[\mathcal{E}(r)\,\mathcal{E}'(r) - \mathcal{K}(r)\,\mathcal{E}'(r) + r^2\,\mathcal{K}(r)\,\mathcal{K}'(r)]}{[r^2\,\mathcal{K}'(r)]}$$

is strictly decreasing on $(0,1)$, *with $f(0^+) = \pi/2$ and $f(1^-) = 1$.*

2. Preliminary results. In this section, we prove monotonicity properties for certain combinations of $\mathcal{K}(r)$ and $\mathcal{E}(r)$. These are needed for the proofs of the theorems stated in section 1.

LEMMA 2.1. *The function*

$$f_1(r) \equiv [16 + 32c - (144c + 16c^2)r^2 - 25c^2r^4]\,\mathcal{K}'(r) + (48c + 40c^2r^2)\,\mathcal{E}'(r)$$

is strictly decreasing from $(0,1)$ *onto* (a_1, ∞), *where* $a_1 = \pi(16 - 64c - c^2)/2 = -57.377146\ldots$, *so that f_1 has a unique zero $r_1 \in (0,1)$ such that $f_1(r) > 0$ for $r \in (0, r_1)$ and $f_1(r) < 0$ for $r \in (r_1, 1)$.*

Proof. By [AVV2, Theorem 1.2], $f_1(0^+) = \infty$ and $f_1(1) = a_1$. Next, rearranging the terms, we have

$$f_1(r) = 48c(\mathcal{K}' + \mathcal{E}') + 16(1-c)\,\mathcal{K}' - 40c^2r^2(2\,\mathcal{K}' - \mathcal{E}') - 16c(9-4c)r^2\,\mathcal{K}' - 25c^2r^4\,\mathcal{K}'. \tag{2.1}$$

Since

$$\frac{d}{dr}[r'^2(2\,\mathcal{K} - \mathcal{E})] = -\frac{1+3r^2}{r}(\mathcal{K} - \mathcal{E}),$$

which is clearly negative for all $r \in (0,1)$, the function $r^2(2\,\mathcal{K}' - \mathcal{E}')$ is strictly increasing on $(0,1)$. Hence, from [AVV2, Theorems 2.1(3) and 1.2] we see that the right side of (2.1) is a sum of strictly decreasing functions, so that the result follows. □

LEMMA 2.2. *Let r_1 be as in Lemma* 2.1. *Then the function*

$$f_2(r) \equiv [16(1+c) - 4c(12+c)r^2 - 5c^2r^4]r^2\,\mathcal{K}'(r) - (16 - 32cr^2 - 9c^2r^4)\,\mathcal{E}'(r)$$

is strictly increasing on $(0, r_1]$ and decreasing on $[r_1, 1)$, so that, on $(0,1)$, f_2 has a unique zero r_2 such that $f_2(r) < 0$ for $r \in (0, r_2)$ and $f_2(r) > 0$ for $r \in (r_2, 1)$.

Proof. Since

$$f'_2(r) = r f_1(r),$$

where f_1 is as in Lemma 2.1, the piecewise monotonicity of f_2 follows from Lemma 2.1.

Clearly, $f_2(0^+) = -16$ and $f_2(1) = 0$. Hence, the assertion for the zero of f_2 follows from the first conclusion. □

LEMMA 2.3. *Let r_2 be as in Lemma 2.2. Then the function*

$$f_3(r) \equiv [16 + 16(1+c)r^2 - c(16+c)r^4 - c^2r^6]\,\mathcal{K}'(r) - [16(2+c) - 16cr^2 - 2c^2r^4]\,\mathcal{E}'(r)$$

is strictly decreasing on $(0, r_2]$ and increasing on $[r_2, 1)$, so that, on $(0,1)$, f_3 has a unique zero r_3 such that $f_3(r) > 0$ for $r \in (0, r_3)$ and $f_3(r) < 0$ for $r \in (r_3, 1)$.

Proof. The piecewise monotonicity of f_3 follows from the derivative

$$f'_3(r) = f_2(r)/r,$$

where f_2 is as in Lemma 2.2.

Since $f_3(0^+) = \infty$ and $f_3(1) = 0$, the assertion for the zero of f_3 follows from the first conclusion. □

LEMMA 2.4. *Let r_3 be as in Lemma 2.3. Then the function*

$$f_4(r) \equiv \frac{4 - cr^2}{4 + cr^2}\,\frac{r'^2\,\mathcal{K}'(r)}{\mathcal{E}'(r) - r^2\,\mathcal{K}'(r)} - \log\left(\frac{4}{r} + cr\right)$$

is strictly increasing on $(0, r_3]$ and decreasing on $[r_3, 1)$, with $f_4(0^+) = 0$ and $f_4(1) = 16e^{-\pi/2} - 2 - \pi/2 = -0.24472\ldots$, so that, on $(0,1)$, f_4 has a unique zero r_4 such that $f_4(r) > 0$ for $r \in (0, r_4)$ and $f_4(r) < 0$ for $r \in (r_4, 1)$.

Proof. By differentiation,

$$(4 + cr^2)^2(\,\mathcal{E}' - r^2\,\mathcal{K}')^2 f'_4(r) = r\,\mathcal{K}' f_3(r),$$

where f_3 is as in Lemma 2.3. Hence, the piecewise monotonicity of f_4 follows from Lemma 2.3.

Hence, $f_4(1) = 16e^{-\pi/2} - 2 - \pi/2$. By l'Hôpital's rule and [AVV2, Theorem 1.2], we have

$$\lim_{r\to 0}\left\{r'^2 - \left(\,\mathcal{E}' - r^2\,\mathcal{K}'\right)\right\}/r = 0,$$

and hence, by [AVV2, Theorem 1.2] and $\lim_{r\to 0}(\,\mathcal{K}'(r) - \log(4/r)) = 0$ [WW, p. 521],

$$f_4(0^+) = \lim_{r\to 0}\left\{\left(\,\mathcal{K}' - \log\frac{4}{r}\right) - \log\left(1 + \frac{c}{4}r^2\right) + \frac{r\,\mathcal{K}'}{\mathcal{E}' - r^2\,\mathcal{K}'}\left[\frac{r'^2 - (\,\mathcal{E}' - r^2\,\mathcal{K}')}{r} - \frac{2crr'^2}{4 + cr^2}\right]\right\} = 0.$$

The assertion for the zero of f_4 is clear. □

LEMMA 2.5. *The function $f(r) \equiv \sqrt{r'}\,\frac{(2-r^2)\,\mathcal{K}(r) - 2\,\mathcal{E}(r)}{(1-r')^2}$ is strictly decreasing and concave from $(0,1)$ onto $(0, \pi/4)$. In particular, for $r \in (0,1)$,*

$$\pi(1-r)/4 < f(r) < \pi/4.$$

Proof. Put $r = 2\sqrt{x}/(1+x)$. Then $r' = (1-x)/(1+x), x = (1-r')/(1+r')$, and by the Landen transformation [BB1, Theorem 1.2],

$$f(r) = x'[\,\mathcal{K}(x) - \mathcal{E}(x)]/x^2,$$

which, as a function of x, is strictly decreasing and concave from $(0,1)$ onto $(0,\pi/4)$ [QV1, Theorem 1.8(1)]. Since x is a convex function of r, the result follows from [AQV, Lemma 2.1(2)]. □

THEOREM 2.6. *For $a \in (-\infty,\infty)$, define the function f on $[0,1]$ by*

$$f(r) = (a+r')[\,\mathcal{E}(r) - r'^2\,\mathcal{K}(r)]/r^2$$

for $r \in (0,1), f(0) = \pi(a+1)/4$, and $f(1) = a$. Then

(1) *f is strictly decreasing on $[0,1]$ iff $a \le 3$. Moreover, for $a \in [1,3]$, f is concave on $(0,1)$.*

(2) *For $a > 3$, there exists an $r_0 = r_0(a) \in (0,1)$ such that f is strictly increasing on $(0,r_0]$ and decreasing on $[r_0,1)$.*

Proof. Since $f(r) = (a-1)(\,\mathcal{E} - r'^2\,\mathcal{K})/r^2 + (\,\mathcal{E} - r'^2\,\mathcal{K})/(1-r')$,

$$\begin{aligned}\frac{r'(1-r')^2}{r(\,\mathcal{E} - r'\,\mathcal{K})} f'(r) &= g_1(r)\\ &\equiv (a-1)\frac{\sqrt{r'}}{1+r'}\cdot\sqrt{r'}\frac{(2-r^2)\,\mathcal{K} - 2\,\mathcal{E}}{(1-r')^2}\cdot\frac{(1-r')^3}{r^2(\,\mathcal{E} - r'\,\mathcal{K})} - 1\\ &\equiv (a-1)g_2(r) - 1,\end{aligned} \tag{2.2}$$

with $g_1(0) = (a-3)/2$ and $g_1(1^-) = -1$. Clearly, $\sqrt{r'}/(1+r')$ is strictly decreasing on (0, 1). Hence, g_2 is a product of three positive and strictly decreasing functions on $(0,1)$, by Lemma 2.5 and [QV1, Theorem 1.7(1)].

(1) If $a \le 1$, then g_1 is negative on $(0,1)$, and hence f is strictly decreasing on $(0,1)$.

If $1 < a \le 3$, then g_1 is strictly decreasing from (0, 1) onto $(-1,(a-3)/2)$. Hence, it follows from (2.8) that f is strictly decreasing on (0, 1).

Conversely, if f is strictly decreasing on (0, 1), then $g_1(0) = (a-3)/2 \le 0$, and hence $a \le 3$.

For $a \in [1,3]$, g_1 is negative and decreasing on (0, 1). Since

$$-f'(r) = (-g_1(r))\cdot\frac{r}{r'(1+r')}\cdot r^2\frac{\mathcal{E} - r'\,\mathcal{K}}{(1-r')^3},$$

and since $r^2(\,\mathcal{E} - r'\,\mathcal{K})(1-r')^{-3}$ is a positive and increasing function on (0, 1) [QV1, Theorem 1.7 (1)], we see that f' is decreasing on (0, 1). This yields the concavity of f on (0, 1) for $a \in [1,3]$.

(2) If $a > 3$, then g_1 is strictly decreasing on (0, 1) with $g_1(0^+) = (a-3)/2 > 0$ and $g_1(1^-) = -1 < 0$, and hence, part (2) follows. □

COROLLARY 2.7. *For $a \in [1,3]$, let $a_1 = (\pi(1+a)/4) - a$. Then*

$$(\pi(1+a)/4) - a_1 r < (a+r')[\,\mathcal{E}(r) - r'^2\,\mathcal{K}(r)]/r^2 < \pi(a+1)/4$$

for $r \in (0,1)$.

Proof. The result follows from Theorem 2.6(1). □

The next theorem complements the well-known result that the function $\mathcal{K}(r) + \log r'$ is decreasing and concave from $(0,1)$ onto $(\log 4, \pi/2)$ [AVV1, Theorem 2.2(2)].

THEOREM 2.8. (1) *The function* $f(r) \equiv \mathcal{K}'(r) + \log(r/(1+r))$ *is strictly decreasing and convex from* $(0,1)$ *onto* $(\pi/2 - \log 2, \log 4)$. *In particular,*

$$\frac{\pi}{2} - \log 2 + \log(1 + 1/r) < \mathcal{K}'(r) < \frac{\pi}{2} - \log 2 + \left(\log 8 - \frac{\pi}{2}\right)(1-r) + \log(1+1/r)$$

for $r \in (0,1)$.

(2) *The function* $g(r) \equiv \mathcal{K}(r) + \log(r'/(1+r'))$ *is strictly increasing and convex from* $(0,1)$ *onto* $(\pi/2 - \log 2, \log 4)$.

Proof. For part (1), let $F_1(r) = [1 - r - (\mathcal{E}' - r^2\mathcal{K}')]/r$, $F_2(r) = r'^2$, $F_3(r) = \mathcal{E}' - 1$ and $F_4(r) = r^3$. Then $F_1(1) = F_2(1) = F_3(0) = F_4(0) = 0$, and

$$f'(r) = F_1(r)/F_2(r), 2F_1'(r)/F_2'(r) = -F_3(r)/F_4(r).$$

Since $3F_3'(r)/F_4'(r) = (\mathcal{K}' - \mathcal{E}')/(rr'^2)$ is strictly decreasing on (0, 1) [AVV2, Theorem 2.1(6)], f' is negative and strictly increasing on (0, 1) by [VV, Lemma 1.1]. The value $f(1)$ is clear, and the limit $f(0^+)$ follows from $\lim_{r\to 0}(\mathcal{K}'(r) - \log(4/r)) = 0$ [WW, p. 521].

Part (2) follows from part (1) and [AQV, Lemma 2.1(2)]. □

Remark 2.9. It is natural to ask if the function $f(r) \equiv \mathcal{K}'(r) + \log r$ is convex on $(0,1)$. By differentiation we get

$$f'(r) = \frac{1}{r}\left(1 - \frac{\mathcal{E}' - r^2\mathcal{K}'}{r'^2}\right), \quad f''(r) = \frac{(1+r^2)\mathcal{K}' - 2\mathcal{E}'}{r'^4} + \frac{1}{r^2}\left(\frac{\mathcal{E}' - r^2\mathcal{K}'}{r'^2} - 1\right).$$

Thus by l'Hôpital's rule, $f''(0)^+ = \infty$ and $f''(1^-) = (5\pi/16) - 1 < 0$, so that f is neither convex nor concave.

LEMMA 2.10. (1) *The function* $f(r) \equiv 2\mathcal{E}(r) - r'^2\mathcal{K}(r)$ *is strictly increasing and convex from* $(0,1)$ *onto* $(\pi/2, 2)$.

(2) *There exists a unique* $r_1 \in (0.251, 0.252)$ *such that the function* $g(r) \equiv [1 - 2r^2\log(4/r)]/[2\log(4/r) - 1]^2$ *is strictly increasing on* $(0, r_1]$ *and decreasing on* $[r_1, 1]$ *with* $g(1) = -1/(\log 16 - 1) = -0.56414\ldots$, *so that* g *has a unique zero* $r_2 \in (0.487, 0.488)$ *such that* $g(r) > 0$ *for* $r \in (0, r_2)$ *and* $g(r) < 0$ *for* $r \in (r_2, 1]$.

(3) *The function* $h(r) \equiv \{[2\mathcal{E}(r) - r'^2\mathcal{K}(r)]/\mathcal{E}(r)\} + \{2r'^2/[2\log(4/r) - 1]\}$ *is strictly increasing from* $(0,1)$ *onto* $(1,2)$.

Proof. The value $f(0) = 2$ is clear, while the limit $f(1^-) = \pi/2$ follows from [AVV2, Theorem 1.2]. Then part (1) follows since

$$f'(r) = (\mathcal{E} - r'^2\mathcal{K})/r$$

is positive and strictly increasing from (0,1) onto (0,1) by [AVV1, Theorem 2.2(7)].

For part (2), we have $rg'(r)(2\log(4/r) - 1)^3 = 2g_1(r)$, where $g_1(r) = 2 - r^2 - 4[r\log(4/r)]^2$. Then g_1 is strictly decreasing from (0, 1) onto $(1 - 16(\log 2)^2, 2)$. Since $g_1(0.251) = 0.0053\ldots > 0$, while $g_1(0.252) = -0.0049\ldots < 0$, g_1 has a unique zero $r_1 \in (0.251, 0.252)$ such that $g_1(r) > 0$ for $r \in (0, r_1)$ and $g_1(r) < 0$ for $r \in (r_1, 1)$. Hence, the piecewise monotonicity of g follows, and g has a unique zero $r_2 \in (0,1)$ such that $g(r) > 0$ for $r \in (0, r_2)$ and $g(r) < 0$ for $r \in (r_2, 1)$.

Let $g_2(r) = 1 - 2r^2\log(4/r)$. Then $g_2(0.487) = 0.00114\ldots$ and $g_2(0.488) = -0.00198\ldots$. Hence, $0.487 < r_2 < 0.488$.

For part (3), let $h_1(r) = \mathcal{K}(2\,\mathcal{E} - r'^2\,\mathcal{K})/\,\mathcal{E}^2 + 4g(r) - 1$. Then

$$rh'(r) = h_1(r). \tag{2.3}$$

By part (1), $\mathcal{K}(2\,\mathcal{E} - r'^2\,\mathcal{K})/\,\mathcal{E}^2$ is strictly increasing from $(0,1)$ onto $(1,\infty)$. Hence, it follows from part (2) that $h_1(r) > 0$ for $r \in (0, r_2]$, and for $r \in [a,b) \subset (0.487, 1)$,

$$h_1(r) \geq h_2(a,b) \equiv \frac{\mathcal{K}(a)[2\,\mathcal{E}(a) - a'^2\,\mathcal{K}(a)]}{\mathcal{E}(a)^2} + 4g(b) - 1.$$

Computation gives

$$h_2(0.487, 0.64) = 0.006\ldots, \quad h_2(0.64, 0.73) = 0.028\ldots,$$

$$h_2(0.73, 0.8) = 0.031\ldots, \quad h_2(0.8, 0.87) = 0.001\ldots,$$

$$h_2(0.87, 0.94) = 0.075\ldots, \quad h_2(0.94, 1) = 0.642\ldots.$$

Hence, $h_1(r) > 0$ also for $r \in [0.487, 1)$. Consequently, the monotonicity of h follows from (2.3).

The limiting values of h are clear. □

LEMMA 2.11. (1) *The function* $f(r) \equiv [(4 - 3r^2)\,\mathcal{K}(r) - 4\,\mathcal{E}(r)]/r^2$ *is strictly increasing and convex from* $(0,1)$ *onto* $(-\pi/2, \infty)$.

(2) *The function* $g(r) \equiv [(8-7r^2)\,\mathcal{E}(r) - (8-3r^2)(r')^2\,\mathcal{K}(r)]/r^6$ *is strictly increasing and convex from* $(0,1)$ *onto* $(3\pi/32, 1)$.

Proof. The limiting values are clear. Next, the power series expansions for $\mathcal{K}$ and $\mathcal{E}$ [BF, eqs. 900.00 and 900.07] give

$$f(r) = \frac{\pi}{2}\left\{-1 + \sum_{n=1}^{\infty} \frac{n-1}{n+1}\left[\frac{1.3\cdots(2n-1)}{2.4\cdots(2n)}\right]^2 r^{2n}\right\}$$

and

$$g(r) = \frac{9\pi}{16}\sum_{n=1}^{\infty}\left[\frac{1.3\cdots(2n+1)}{2^n}\right]^2 \frac{1}{n!(n+3)!}r^{2n},$$

so that the assertions follow. □

THEOREM 2.12 (cf. [Q2, Lemma 7]). (1) *The function* $f(r) \equiv [(2 - r^2)\,\mathcal{K} - 2\,\mathcal{E}]/[2\log(1/r') - r^2]$, *is increasing from* $(0,1)$ *onto* $(\pi/8, 1/2)$.

(2) *The function* $g(r) \equiv [(3-r^2)\,\mathcal{K} - 3\,\mathcal{E}]/\log(1/r')$ *is increasing from* $(0,1)$ *onto* $(\pi/2, 1)$.

(3) *The function* $h(r) \equiv [(4-r^2)\,\mathcal{K} - 4\,\mathcal{E}]/\log(1/r')$ *is decreasing from* $(0,1)$ *onto* $(3,\pi)$.

(4) *The function* $F(r) \equiv [(8-7r^2)\,\mathcal{E} - (8-3r^2)(r'^2)\,\mathcal{K}]/[(r^2)(1+r'^2) - 4(r'^2)\log(1/r')]$ *is increasing from* $(0,1)$ *onto* $(9\pi/32, 1)$.

Proof. (1) Denote the numerator and denominator of $f(r)$ by $f_1(r)$ and $f_2(r)$, respectively. Then $f_1(0) = f_2(0) = 0$, and $f_1'(r)/f_2'(r) = [\,\mathcal{E} - (r'^2)\,\mathcal{K}]/(2r^2)$. By (1.1), (1.2), this is increasing from $(0,1)$ onto $(\pi/8, 1/2)$. Hence, the assertion follows from [VV, Lemma 1.1].

(2) $g(r) = g_1(r)/g_2(r)$, with $g_1(0) = g_2(0) = 0$. Then $g_1'(r)/g_2'(r) = 2\,\mathcal{E} - (r'^2)\,\mathcal{K}$, which is easily seen to be increasing from (0,1) onto $(\pi/2, 2)$. Hence, the assertion follows from [VV, Lemma 1.1].

(3) $h(r) = h_1(r)/h_2(r)$, with $h_1(0) = h_2(0) = 0$. Then $h_1'(r)/h_2'(r) = 3\,\mathcal{E} - (r'^2)\,\mathcal{K}$, which is easily seen to be decreasing from (0,1) onto $(3, \pi)$. Hence, the assertion follows from [VV, Lemma 1.1].

(4) $F(r) = F_1(r)/F_2(r)$, with $F_1(0) = F_2(0) = 0$. Then $F_1'(r)/F_2'(r) = (9/4)f(r)$. Hence, the assertion follows from (1) and [VV, Lemma 1.1]. □

3. Proofs of the main theorems. In this section, we prove the main theorems stated in section 1.

3.1. Proof of Theorem 1.5. Differentiation gives

$$r[r' \log(cr + (4/r))]^2 f'(r) = (\,\mathcal{E}' - r^2\,\mathcal{K}')f_4(r), \tag{3.1}$$

where f_4 is as in Lemma 2.4.

Take $r_0 = r_4$, the zero of f_4 on (0, 1). Then the piecewise monotonicity of f follows from (3.1) and Lemma 2.4.

Next, the limiting values $f(0^+) = f(1) = 1$ are clear. Hence, the inequality (1.4) and its sharpness follow. □

3.2. Proof of Theorem 1.6. By differentiation,

$$\begin{aligned} r'^2(a + r')f'(r)/r &= g(r) \\ &\equiv a - (a + r')(\,\mathcal{E} - r'^2\,\mathcal{K})/r^2, \end{aligned} \tag{3.2}$$

with $g(0^+) = a - \pi(a + 1)/4$ and $g(1^-) = 0$.

For (1), we investigate two cases.

Case 1. $0 < a \le 3$.

In this case, it follows from Theorem 2.6(1) that g is strictly increasing on (0, 1), and hence $g(r) < 0$ for $r \in (0, 1)$, so that f is strictly decreasing on (0, 1) by (3.2).

Case 2. $3 < a \le \pi/(4 - \pi) = 3.65979\ldots$.

In this case, by Theorem 2.6(2) there exists an $r_0 = r_0(a) \in (0, 1)$ such that g is strictly decreasing on $(0, r_0]$ and increasing on $[r_0, 1)$. Since $g(0^+) = a(4-\pi)/4-\pi/4 \le 0$ and $g(1^-) = 0$, it follows that $g(r) < 0$ for all $r \in (0, 1)$. Hence, f is strictly decreasing on $(0, 1)$ by (3.2).

For part (2), we note that $g(0) > 0$ when $a > \pi/(4 - \pi)$. Hence, the piecewise monotonicity of f follows from (3.2) and Theorem 2.6(2), and

$$f(r) > \min\{f(0), f(1^-)\} = \begin{cases} \log(a/4) & \text{if } a \le 4/c, \\ \log(1 + a) - \pi/2 & \text{if } a > 4/c \end{cases}$$

for $r \in (0, 1)$, so that the inequality (1.5) holds. □

COROLLARY 3.5. (1) *For* $r \in (0, 1)$,

$$\log((4/r') + cr') < \mathcal{K}(r) < \log((4/r') + c). \tag{3.3}$$

These inequalities are sharp as r *tends to* 0 *or* 1.

(2) *There exists a unique* $r_1 \in (0, 1)$ *such that the inequality*

$$\mathcal{K}(r) \le \log(1 + (4/r')) - (\log 5 - \pi/2) \tag{3.4}$$

holds for $r \in (0, r_1]$, with equality iff $r = r_1$. The inequality is reversed if $r \in [r_1, 1)$.

Proof. The first inequality in (3.3) and its sharpness were obtained in Theorem 1.5. Taking $a = 4/c$ in (1.5), we get the second inequality in (3.3), with

$$\lim_{r\to 0} \mathcal{K}/\log(c + (4/r')) = \lim_{r\to 1} \mathcal{K}/\log(c + (4/r')) = 1.$$

Part (2) follows from the piecewise monotonicity of f in Theorem 1.6(2) with $a = 4$. □

COROLLARY 3.2. *There exists a unique $r_2 \in (\sin 70^\circ, \sin 71^\circ)$ such that the function $F(r) \equiv \log((4/r') + c) - \mathcal{K}(r)$ is strictly increasing on $(0, r_2]$ and decreasing on $[r_2, 1)$ with $F((0,1)) = (0, c_1]$, where $c_1 = F(r_2) < 0.067809628$. In particular, for $r \in (0, 1)$,*

$$\log\left(\frac{4}{r'} + c\right) - 0.067809628 < \log\left(\frac{4}{r'} + c\right) - c_1 < \mathcal{K}(r) < \log\left(\frac{4}{r'} + c\right). \tag{3.5}$$

Proof. Put $a = 4/c$. Then $F(r) = \log c + \log(1 + (a/r')) - \mathcal{K}$, and the piecewise monotonicity of F follows from Theorem 1.6(2).

Let $F_1(r) = cg(r)$, where g is as in (3.2). Then

$$r'^2(4 + cr')F'(r) = rF_1(r) = r\left\{4 - (4 + cr')(\mathcal{E} - r'^2\mathcal{K})/r^2\right\}.$$

Since $F_1(\sin 70^\circ) = 0.001\ldots$ and $F_1(\sin 71^\circ) = -0.002\ldots$, $r_2 \in (\sin 70^\circ, \sin 71^\circ)$ by Theorem 2.6(2), and

$$\begin{aligned} c_1 = F(r_2) &< \log((4/\cos 71^\circ) + c) - \mathcal{K}(\sin 70^\circ) \\ &< \log((4/\cos 71^\circ) + c) - 2.50455 = 0.067809627\ldots < 0.067809628. \end{aligned}$$

The inequalities in (3.5) are clear. □

3.3. Proof of Theorem 1.7. The first and second inequalities in part (1) follow from Corollary 3.5(1). Since $\sqrt{1+r}(2-r)$ is strictly decreasing from (0, 1) onto $(\sqrt{2}, 2)$,

$$r' = \sqrt{1+r}(2-r) \cdot \frac{\sqrt{1-r}}{2-r} < 2\frac{\sqrt{1-r}}{2-r}.$$

Clearly, the inequalities in (1) are all sharp as r tends to 0.

For part (2), by Corollary 3.5(1), we need only to prove that

$$\log((4/r') + c) \le \log((4/r') + 1) - (\log 5 - (\pi/2))(1 - r)$$

or, equivalently,

$$\log(1 + (cr'/4)) \le \log(1 + (r'/4)) - (\log 5 - (\pi/2))(1 - r) \tag{3.6}$$

for $r \in (0, 1)$. For this purpose, we investigate the minimum of the function

$$H(r) \equiv \log(1 + (r'/4)) - \log(1 + (cr'/4)) - (\log 5 - (\pi/2))(1 - r)$$

for $r \in [0, 1]$. It is obvious that $H(0) = H(1) = 0$.

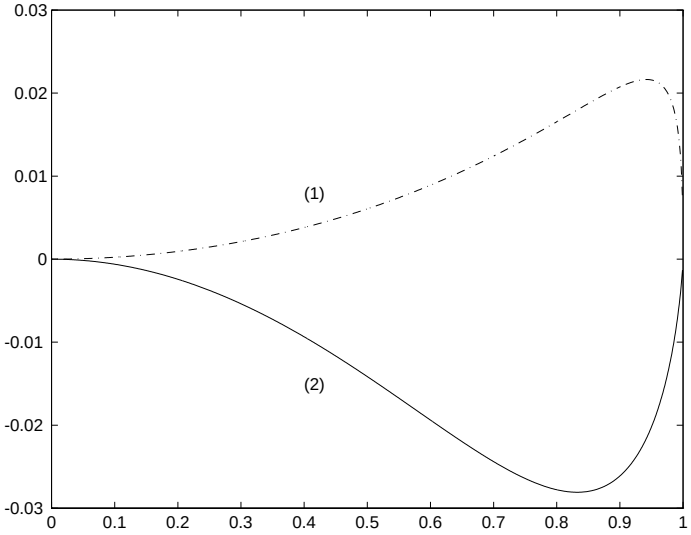

FIG. 1.

By differentiation, we get

$$H'(r) = \log 5 - \frac{\pi}{2} - \frac{4(1-c)r}{r'(4+r')(4+cr')},$$

which is clearly strictly decreasing from (0, 1) onto $(-\infty, \log 5 - \pi/2)$. Since $\log 5 - \pi/2 > 0$, there exists a unique $r_0 \in (0,1)$ such that H is strictly increasing on $(0, r_0]$ and decreasing on $[r_0, 1)$. This yields

$$\min\{H(r); 0 \le r \le 1\} = \min\{H(0), H(1)\} = 0,$$

from which inequality (3.6) follows.

Next, the inequality (2) is sharp as r tends to 0 or 1, since

$$\begin{aligned} &\lim_{r\to 0} \mathcal{K}/\{\log(1+(4/r')) - (\log 5 - (\pi/2))(1-r)\} \\ =&\lim_{r\to 1} \mathcal{K}/\{\log(1+(4/r')) - (\log 5 - (\pi/2))(1-r)\} = 1. \end{aligned}$$

Finally, part (3) follows from Corollary 3.5(1). □

Remark 3.3. (1) From the proof of Theorem 1.7(2), one can see that the upper bound of $\mathcal{K}(r)$ given in Corollary 3.5(1) is better than that in Theorem 1.7(2). Thus, Conjecture 1.2 is proved.

(2) In order to illuminate Corollary 3.5(1) we have graphed in Figure 1 the functions $\log((4/r') + c) - \mathcal{K}(r)$ and $\log((4/r') + cr') - \mathcal{K}(r)$ labeled by (1) and (2), respectively.

In order to illuminate Theorem 1.7(2) we have graphed in Figure 2 the function $\log(1 + (4/r')) - (\log 5 - (\pi/2))(1-r) - \mathcal{K}(r)$.

3.4. Proof of Theorem 1.8. By differentiation,

$$\{r'[2\log(4/r') - 1]\}^2 f'(r) = r(2\,\mathcal{E}' - r^2\,\mathcal{K}')F_1(r')F(r), \tag{3.7}$$

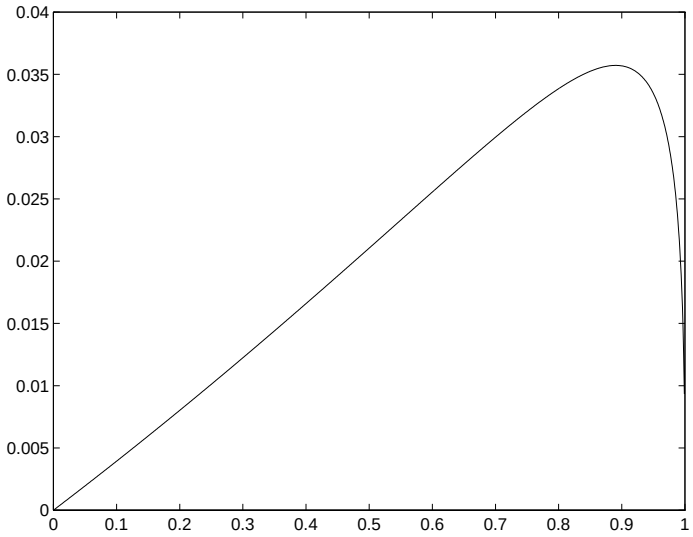

FIG. 2.

where

$$\begin{aligned} F_1(r) &\equiv 2\log(4/r) + 2r'^2\,\mathcal{E}/(2\,\mathcal{E} - r'^2\,\mathcal{K}) - 1,\\ F(r) &\equiv \mathcal{K}\,\mathcal{K}'/[2\,\mathcal{E}'F_2(r')] - \mathcal{K}\,\mathcal{E}'/[(2\,\mathcal{E}' - r^2\,\mathcal{K}')F_1(r')]\\ &\quad - [(2-r^2)\,\mathcal{K} - 2\,\mathcal{E}]/(2r^4),\\ F_2(r) &\equiv (2\,\mathcal{E} - r'^2\,\mathcal{K})/\,\mathcal{E} + 2r'^2/[2\log(4/r) - 1]. \end{aligned}$$

By Lemma 2.10(1) and (3), F_1 and F_2 are positive, and decreasing and increasing on $(0,\,1)$, respectively. Hence, for $r \in (a, b] \subset (0, 1)$,

$$F(r) > \frac{\mathcal{K}'(r)\,\mathcal{K}(a)}{2\,\mathcal{E}'(r)F_2(a')} - \frac{\mathcal{K}(r)\,\mathcal{E}'(r)}{[2\,\mathcal{E}'(r) - r^2\,\mathcal{K}'(r)]F_1(a')} \tag{3.8}$$

$$-[(2-r^2)\,\mathcal{K}(r) - 2\,\mathcal{E}(r)]/(2r^4).$$

By [QV1, Corollary 3.12] and [AVV2, Theorem 2.1(7)], the last term is strictly decreasing on $(0, 1)$, while the second term is decreasing by Lemma 2.10(1), so the right side of (3.8) is decreasing on $(0, 1)$. Therefore, it follows from (3.8) that

$$F(r) > F_3(a,b) \equiv \frac{\mathcal{K}(a)\,\mathcal{K}'(b)}{2F_2(a')\,\mathcal{E}'(b)} \tag{3.9}$$

$$-\frac{\mathcal{K}(b)\,\mathcal{E}'(b)}{[2\,\mathcal{E}'(b) - b^2\,\mathcal{K}'(b)]F_1(a')} - \frac{(2-b^2)\,\mathcal{K}(b) - 2\,\mathcal{E}(b)}{2b^4}$$

for $r \in (a, b] \subset (0, 1)$. Computation gives

$$\begin{aligned} F_3(0, \sin 28^\circ) &= 0.019\ldots, \quad F_3(\sin 28^\circ, \sin 37^\circ) = 0.053\ldots,\\ F_3(\sin 37^\circ, \sin 45^\circ) &= 0.005\ldots, \quad F_3(\sin 45^\circ, \sin 51^\circ) = 0.007\ldots,\\ F_3(\sin 51^\circ, \sin 56^\circ) &= 0.005\ldots, \quad F_3(\sin 56^\circ, \sin 60^\circ) = 0.012\ldots. \end{aligned}$$

Hence, it follows from (3.7) and (3.9) that

$$f'(r) > 0 \quad \text{for } r \in (0, \sqrt{3}/2]. \tag{3.10}$$

Next, making use of Legendre's relation [BB, p. 24], [WW, p. 52], we can write (3.7) as

$$r\{r'[2\log(4/r') - 1]\}^2 f'(r) = F_4(r)F_5(r), \tag{3.11}$$

where

$$F_4(r) \equiv 2\log(4/r') - 2\,\mathcal{E}'(\mathcal{K} - \mathcal{E})/F_5(r) - 1, \quad F_5(r) \equiv 2\,\mathcal{E}'(\mathcal{E} - r'^2\,\mathcal{K})/r^2 - (\pi/2).$$

This is seen as follows.

By differentiation and Legendre's relation [BB, p. 24]

$$\begin{aligned}\left[r^2\left(2\log\left(\frac{4}{r'}\right) - 1\right)\right]^2 f'(r) &= \left(\frac{r}{r'^2}\right)\left[r^2\left(2\log\left(\frac{4}{r'}\right) - 1\right)((\mathcal{K} - \mathcal{E})(\mathcal{K}' - \mathcal{E}') + \mathcal{E}\,\mathcal{E}')\right.\\ &\qquad \left. - 2\;\mathcal{E}'(\mathcal{K} - \mathcal{E})\left(r^2 + \left(2\log\left(\frac{4}{r'}\right) - 1\right)r'^2\right)\right]\\ &= \left(\frac{r}{r'^2}\right)\left[r^2\left(2\log\left(\frac{4}{r'}\right) - 1\right)\left(2\,\mathcal{E}\,\mathcal{E}' - \left(\frac{\pi}{2}\right)\right)\right.\\ &\qquad \left. - 2\,\mathcal{E}'(\mathcal{K} - \mathcal{E})\left(r^2 + \left(2\log\left(\frac{4}{r'}\right) - 1\right)r'^2\right)\right]\\ &= \left(\frac{r}{r'^2}\right)\left[\left(2\log\left(\frac{4}{r'}\right) - 1\right)\left\{2\,\mathcal{E}'(\mathcal{E} - r'^2\,\mathcal{K}) - \left(\frac{\pi}{2}\right)r^2\right\} - 2\,\mathcal{E}'(\mathcal{K} - \mathcal{E})r^2\right].\end{aligned}$$

Hence,

$$\begin{aligned}r^3 r'^2\left(2\log\left(\frac{4}{r'}\right) - 1\right)^2 f'(r) &= r^2 F_5(r)\left[2\log\left(\frac{4}{r'}\right) - 1 - 2\,\mathcal{E}'(\mathcal{K} - \mathcal{E})/F_5(r)\right]\\ &= r^2 F_5(r)F_4(r),\end{aligned}$$

which proves (3.11).

Clearly, F_5 is strictly increasing from $(0,\,1)$ onto $(0, \pi/2)$ [AVV2, Theorem 2.1(7)].

By differentiation,

$$r^3 r'^2 F_5'(r) = 2\left\{r^2(\mathcal{K}' - \mathcal{E}')(\mathcal{E} - r'^2\,\mathcal{K}) + r'^2\,\mathcal{E}'[(2 - r^2)\,\mathcal{K} - 2\,\mathcal{E}]\right\},$$

and hence, by Legendre's relation [BB, p. 24],

$$\begin{aligned}F_4'(r) = \frac{2r}{r'^2} - \frac{2}{F_5(r)^2}\left\{\frac{r}{r'^2}F_5(r)\,[2\,\mathcal{E}\,\mathcal{E}' + \mathcal{K}\,\mathcal{K}' - \mathcal{E}\,\mathcal{K}' - \mathcal{K}'\,\mathcal{E} + F_5(r) - F_5(r)]\right.\\ \left. - \mathcal{E}'(\mathcal{K} - \mathcal{E})F_5'(r)\right\}\\ = \frac{4}{rF_5(r)^2}\,\mathcal{E}'(\mathcal{K} - \mathcal{E})\left[\frac{r}{2}F_5'(r) - F_5(r)\right]\end{aligned}$$

so that

$$rF_5(r)^2 F_4'(r) = 4\,\mathcal{E}'(\mathcal{K} - \mathcal{E})F_6(r), \tag{3.12}$$

where

$$\begin{aligned} F_6(r) &\equiv \frac{r}{2}F_5'(r) - F_5(r) \\ &= [(\mathcal{K}' - \mathcal{E}')(\mathcal{E} - r'^2\mathcal{K})/r'^2] + \mathcal{E}'F_7(r) + \pi/2, \\ F_7(r) &\equiv [(4-3r^2)\mathcal{K} - 4\mathcal{E}]/r^2. \end{aligned}$$

Next, let $F_8(r) \equiv \mathcal{E}'F_7(r)$. Then

$$F_7'(r) = \left[(8-7r^2)\mathcal{E} - (8-3r^2)r'^2\mathcal{K}\right]/(r^3r'^2)$$

and hence,

$$\frac{r'^2F_8'(r)}{r(\mathcal{K}' - \mathcal{E}')} = F_9(r) \equiv F_7(r) + \frac{\mathcal{E}'}{\mathcal{K}' - \mathcal{E}'} \cdot \frac{(8-7r^2)\mathcal{E} - (8-3r^2)r'^2\mathcal{K}}{r^4}. \tag{3.13}$$

By Lemma 2.11(2), F_9 is strictly increasing from (0, 1) onto $(-\pi/2, \infty)$. Since $F_9(\sqrt{3}/2) = 1.32\ldots > 0$, it follows from (3.13) that F_8 is strictly increasing on $[\sqrt{3}/2, 1)$, and hence, by [QV2, Theorem 2.1(6)], F_6 is strictly increasing on $[\sqrt{3}/2, 1)$. Since $F_6(\sqrt{3}/2) = 0.06\ldots > 0$, it follows from (3.12) that F_4 is strictly increasing on $[\sqrt{3}/2, 1)$. Since $F_4(\sqrt{3}/2) = 0.53\ldots > 0$, it follows from (3.11) that

$$f'(r) > 0, \quad \text{for } r \in [\sqrt{3}/2, 1). \tag{3.14}$$

The monotonicity of f now follows from (3.10) and (3.14).

The remaining conclusions are clear. □

3.5. Proof of Theorem 1.9. The limit $f(0^+)$ is clear, while $f(1^-) = 1$ follows from [AVV2, Theorems 1.2, 2.1(7)]. Next, $f(r) = g(r)/h(r)$, where $g(r) = \mathcal{E}(r)\mathcal{E}'(r) - \mathcal{K}(r)\mathcal{E}'(r) + r^2\mathcal{K}(r)\mathcal{K}'(r), h(r) = r^2\mathcal{K}'(r)$. Hence, by [AVV2, Theorem 1.2], $g(0^+) = h(0^+) = 0$. By differentiation and simplification,

$$\frac{g'(r)}{h'(r)} = \frac{2\mathcal{E}(r)}{1 + (G(r')/H(r'))}, \tag{3.15}$$

where $G(r) = r^2\mathcal{K}(r), H(r) = \mathcal{K}(r) - \mathcal{E}(r)$. Again $G(0) = H(0) = 0$, and

$$\frac{G'(r)}{H'(r)} = 1 + \frac{r'^2\mathcal{K}(r)}{\mathcal{E}(r)}.$$

Hence, the result follows from [AVV2, Theorems 1.2, 1.3] and [VV, Lemma 1.1]. □

Some of the above results (e.g., Theorem 1.7), together with those in [Q1, Q2, QV2] solve all the conjectures raised in [AVV2, Conjecture 3.1].

Acknowledgments. The authors are grateful to the referee for many helpful suggestions.

REFERENCES

[ADV] G. D. Anderson, P. Duren, and M. K. Vamanamurthy, *An inequality for complete elliptic integrals*, J. Math. Anal. Appl., 182 (1994), pp. 257–259.

[AQV] G. D. Anderson, S.-L. Qiu, and M. K. Vamanamurthy, *Elliptic integral inequalities, with applications*, Constr. Approx., 14 (1998), pp. 195–207.

[AVV1] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Functional inequalities for complete elliptic integrals and their ratios*, SIAM J. Math. Anal., 21 (1990), pp. 536–549.

[AVV2] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Functional inequalities for hypergeometric functions and complete elliptic integrals*, SIAM J. Math. Anal., 23 (1992), pp. 512–524.

[BB1] J. M. Borwein and P. B. Borwein, *Pi and the AGM*, John Wiley, New York, 1987.

[BB2] J. M. Borwein and P. B. Borwein, *Inequalities for compound mean iterations with logarithmic asymptotes*, J. Math. Anal. Appl., 177 (1993), pp. 572–582.

[BF] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Physicists*, Grundlehren Math. Wiss. 57, Springer-Verlag, Berlin, 1954.

[C] B. C. Carlson, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[CG] B. C. Carlson and J. L. Gustafson, *Asymptotic approximations for symmetric elliptic integrals*, SIAM J. Math. Anal., 25 (1994), pp. 288–303.

[DP] P. Duren and J. Pfaltzgraff, *Robin capacity and extremal length*, J. Math. Anal. Appl., 179 (1993), pp. 110–119.

[G] F. Garvan, *Cubic modular identities of Ramanujan, hypergeometric functions and analogues of the arithmetic-geometric mean iteration*, in The Rademacher Legacy to Mathematics, Contemp. Math. 166, G. E. Andrews, D. M. Bressoud, and L. A. Parson, eds., AMS, Providence, RI, 1994, pp. 245–264.

[K] R. Kühnau, *Eine Methode, die Positivität einer Funktion zu prüfen*, Z. Angew. Math. Mech., 74 (1994), pp. 140–142.

[LV] O. Lehto and K. I. Virtanen, *Quasiconformal Mappings in the Plane*, 2nd ed., Grundlehren Math. Wiss. 126, Springer-Verlag, New York, 1973.

[Q1] S.-L. Qiu, *Proof of a conjecture on the first elliptic integrals*, J. Hangzhou Inst. Electronics Engrg., 3 (1993), pp. 30–36.

[Q2] S.-L. Qiu, *On two conjectures concerning elliptic integrals*, J. Hangzhou Inst. Electronics Engrg., 3 (1994), pp. 11–18.

[QV1] S.-L. Qiu and M. K. Vamanamurthy, *Elliptic integrals and the modulus of Grötzsch ring*, Panamer. Math. J., 5 (1995), pp. 41–60.

[QV2] S.-L. Qiu and M. K. Vamanamurthy, *Sharp estimates for complete elliptic integrals*, SIAM J. Math. Anal., 27 (1996), pp. 823–834.

[VV] M. K. Vamanamurthy and M. Vuorinen, *Inequalities for means*, J. Math. Anal. Appl., 183 (1994), pp. 155–166.

[WW] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge Univ. Press, London, 1958.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1238–1255, September 1998

011

ON THE EQUATIONS DESCRIBING A RELAXATION TOWARD A STATISTICAL EQUILIBRIUM STATE IN THE TWO-DIMENSIONAL PERFECT FLUID DYNAMICS*

ANDRO MIKELIĆ† AND RAOUL ROBERT†

Abstract. The large scale evolution of a two-dimensional (2D) incompressible ideal fluid can be modeled by introducing eddy-viscosity terms. This procedure introduces a new convection-diffusion equation for vorticity. Such relaxation equations have a structural similarity with the 2D Navier–Stokes equations in the "stream function-vorticity" formulation but also contain an additional degenerate transport term being essential for conserving the kinetic energy. Using the negative entropy as the Lyapunov functional and after performing the precise estimates for the degenerate transport, we prove existence and uniqueness of solutions to the relaxation equation for a large class of initial data. Furthermore, we study the long time dynamics of the solution, making a link with the statistical equilibrium theory.

Key words. relaxation equations, two-dimensional incompressible perfect fluid, statistical equilibrium, small scale turbulence

AMS subject classifications. 76F99, 35K55, 35Q35, 76C99

PII. S0036141096306509

1. Introduction.

We are concerned here with the behavior of a 2D incompressible perfect fluid in a periodic domain $\mathcal{Y} = (\mathbb{R}/\mathbb{Z})^2$. The motion of the fluid is described by Euler equations, which we write in the classical stream function-vorticity formulation:

$$(E) \quad \begin{cases} \dfrac{\partial \omega}{\partial t} + \operatorname{div}\,(\omega \nabla \times \psi) \;= 0 \quad \text{in } \mathcal{Y}, \\ -\Delta \psi = \omega \quad \text{in } \mathcal{Y}, \quad \displaystyle\int_{\mathcal{Y}} \psi \, dy = 0, \quad \psi \text{ is periodic,} \end{cases}$$

where ω is the scalar vorticity of the flow (satisfying of course $\int_{\mathcal{Y}} \omega \, dy = 0$) and ψ is the corresponding stream function.

The global existence and uniqueness of the solution for a given initial datum $\omega_0 \in L^\infty(\mathcal{Y})$ are well known (it is the famous Youdovitch's theorem [18]).

A further step is to describe the long-time dynamics of the flow: how does $\omega(t)$ behave when $t \to +\infty$? We know that, in general, the function $\omega(t)$ develops oscillations at smaller and smaller scales so that, from a practical point of view, an exhaustive deterministic description soon fails. On the other hand, when observed at a large scale (taking local averages of $\omega(t)$), the flow displays the formations of large structures (the so-called coherent structures in meteorology). In previous works [9], [11], [12], [14], [15] we gave a description of this phenomenon in terms of statistical mechanics, showing that the formation of these structures corresponds to the tendency of the system to reach its statistical equilibrium (see also [5]).

To briefly summarize these works let us say that to each initial datum ω_0 we can associate an equilibrium state ω^* (or more generally, an equilibrium set) that the system is likely to reach (in the weak sense) when $t \to +\infty$.

*Received by the editors July 10, 1996; accepted for publication (in revised form) September 15, 1997; published electronically June 18, 1998.
http://www.siam.org/journals/sima/29-5/30650.html
†C.N.R.S. U.M.R. 5585, Analyse Numérique, Bât. 101, Université Lyon 1, 43, Bd. du onze novembre, 69622 Villeurbanne Cedex, France (andro@iris.univ-lyon1.fr, robert@iris.univ-lyon1.fr).

In this paper we shall consider the simplest case, where the initial vorticity ω_0 can take only the two values $+1, -1$. In this case we have the following simple description of ω^*. It is obtained by solving the variational problem $(\mathcal{V}.\mathcal{P}.)$:

Find the minimum value of the (negative) entropy functional

$$\mathcal{J}(\omega) = \int_{\mathcal{Y}} \Big(\frac{1+\omega}{2} \ln(1+\omega) + \frac{1-\omega}{2} \ln(1-\omega) \Big)\, dy$$

under the constraints

$$(\mathcal{V}.\mathcal{P}.1) \qquad \int_{\mathcal{Y}} \omega \, dy = 0,$$

$$(\mathcal{V}.\mathcal{P}.2) \qquad \frac{1}{2} \int_{\mathcal{Y}} \psi\omega \, dy = \frac{1}{2} \int_{\mathcal{Y}} \psi_0\omega_0 \, dy \quad \text{(the energy corresponding to } \omega_0\text{)}.$$

It is easily seen that this problem always has a solution (not necessarily unique).

Once a statistical equilibrium state (or set) is defined, we can study the relaxation process of the system toward the equilibrium. This issue was addressed in [13] and [17], where we have proposed a simple model of convection–diffusion equation to describe this relaxation process. In this model $\omega(t, y)$ denotes the local mean value of the "microscopic" vorticity which oscillates at small scale between $+1$ and -1.

Then we get for ω the equation

$$\frac{\partial \omega}{\partial t} + \text{div } (\omega \nabla \times \psi) - A \text{ div } \big(\nabla\omega + \beta(\omega)\, (1-\omega^2) \nabla\psi \big) = 0, \tag{1.1}$$

with

$$\beta(\omega) = -\frac{\int_{\mathcal{Y}} \omega^2 \, dy}{\int_{\mathcal{Y}} (1-\omega^2)\, |\nabla\psi|^2 \, dy},$$

where $A > 0$ is a viscosity-type coefficient.

Equation (1.1) looks like Navier–Stokes equations, with the only difference being in the supplementary term $\beta(\omega)(1-\omega^2)\nabla\psi$, which ensures the conservation of the energy.

Our approach raises a natural question: Is (1.1) physically realistic?

The derivation of (1.1) is based upon the two following strong physical assumptions:

First, the adequacy of the entropy functional given by the statistical equilibrium theory of the perfect fluid. Some experiments and results of numerical simulations indicate that it is a rather reasonable hypothesis [10], [16], [17].

Second, the validity of the empirical principle of nonequilibrium thermodynamics used to derive (1.1). This issue, which is out of the scope of this paper, is discussed in [17]. It is known from [17] that the case $A =$ constant, which is considered here, is only a rough approximation; a detailed study of the vorticity diffusion mechanism yields a function $A(\omega)$.

Let us emphasize that the interest of this approach is to obtain relaxation equations like (1.1) with an eddy-viscosity term which is explicitly calculated on the basis of a clear physical hypothesis. Moreover such equations are found to be efficient in performing numerical simulations (see [17]).

One may wonder why (1.1) is not Galilean invariant. This follows naturally from the fact that the statistical equilibrium states for the Euler equation are defined

globally on the whole domain and not locally. Let us note here a close similarity with the Vlasov–Fokker–Planck equation introduced by Chandrasekhar [2] to describe the relaxation of stellar systems (see also [3]).

We give here a first mathematical study of (1.1). In our opinion the interest of the study goes beyond this particular case, and our approach is extendible to other equations of the same kind, modeling relaxation processes.

The reader acquainted with the classical techniques of nonlinear parabolic problems readily sees that for a given fixed β the compacity method works straightforwardly and gives a solution of (1.1) for all time. But β is not fixed and the specific difficulty of the problem is to get some estimate on β. We will show in this paper that such an estimate can be obtained by introducing some restriction on the initial datum ω_0; then the solution exists for all time. Moreover we prove a uniform (in time) H^1 estimate on a solution. As a consequence, we can apply the method from Dafermos [4] to get some information on the asymptotic behavior of the solution: under some condition on the initial datum ω_0 we can prove that the Ω-limit set associated with ω_0 is included in the set of the critical points of the variational problem $(\mathcal{V}.\mathcal{P}.)$.

2. Some auxiliary results. In this section we define the operators connecting velocity and vorticity (stream function and vorticity, resp.) and prove some auxiliary inequalities.

Our flow domain is the unit periodic cell $\mathcal{Y} =]-\frac{1}{2}, \frac{1}{2}[^2$. Let $L^2_0(\mathcal{Y}) = \{\varphi \in L^2(\mathcal{Y}) : \int_{\mathcal{Y}} \varphi = 0\}$ and $\tilde{H}^\alpha_{per}(\mathcal{Y}) = H^\alpha_{per}(\mathcal{Y}) \cap L^2_0(\mathcal{Y}), 0 \le \alpha \le 2$, where $H^\alpha_{per}(\mathcal{Y})$ is the usual Sobolev space of periodic functions.

Let $\mathcal{A}u = -\Delta u$ for $u \in C^2_{per}(\mathcal{Y}) \cap L^2_0(\mathcal{Y})$.

Using Friedrichs theorem we extend $\mathcal{A}$ to a self-adjoint densely defined linear operator in $L^2_0(\mathcal{Y})$. In this case $\mathcal{D}(\mathcal{A}) = \{\varphi \in L^2_0(\mathcal{Y}) \ | \mathcal{A}\varphi \in L^2_0(\mathcal{Y})\} = \tilde{H}^2_{per}(\mathcal{Y})$ and the spectrum $\sigma(\mathcal{A})$ consists of the eigenvalues $\lambda_{k,l} = 4(k^2 + l^2)\pi^2$, $\lambda_{k,l} \neq 0$. Since its resolvent set meets the right half-plane, $\mathcal{A}$ is sectorial (in the sense of Henry [7]).

In the next step we introduce the system

$$\left\{ \begin{array}{rrrrr} & & \text{curl} & u = \omega & \text{in} \quad \mathcal{Y}, \\ & & \text{div} & u = 0 & \text{in} \quad \mathcal{Y}, \\ u & \text{is} & 1-\text{periodic}, & \displaystyle\int_{\mathcal{Y}} u dx = 0. & \end{array} \right. \tag{2.1}$$

Then we set $u = G\omega$. Obviously G is continuous as a linear operator $G : L^2_0(\mathcal{Y}) \to \tilde{H}^1_{per}(\mathcal{Y})^2$ and $G : \tilde{H}^\alpha_{per}(\mathcal{Y}) \to \tilde{H}^{1+\alpha}_{per}(\mathcal{Y})^2$, $0 \le \alpha \le 2$.

Throughout the paper the stream function ψ is defined by

$$\left\{ \begin{array}{rrrr} & & -\Delta\psi = \omega & \text{in} \quad \mathcal{Y}, \\ \psi & \text{is} & 1-\text{periodic}, & \displaystyle\int_{\mathcal{Y}} \psi dx = 0. \end{array} \right. \tag{2.2}$$

Obviously, $\psi = \mathcal{A}^{-1}\omega$ and $\mathcal{A}^{-1}$ is continuous as an operator $\mathcal{A}^{-1} : L^2_0(\mathcal{Y}) \to \tilde{H}^2_{per}(\mathcal{Y})$.

In our considerations an important role is played by the function (negative entropy)

$$S(x) = \frac{1+x}{2}\ln(1+x) + \frac{1-x}{2}\ln(1-x), \qquad x \in]-1, 1[. \tag{2.3}$$

Obviously

$$S'(x) = \frac{1}{2}\ln\frac{1+x}{1-x}, \quad S''(x) = \frac{1}{1-x^2}, \quad S(x) = S(-x),$$

and $S \in C^\infty(]-1,1[)$. It is convex on $]-1,1[$ and monotone and increasing on $]0,1[$.

It is of some importance to compare $S(x)$ with $|x|^p, p \geq 2$, on $]-1,1[$. We have the following.

LEMMA 2.1. *Let $p \geq 2$ and $x \in]-1,1[$. Then*

$$S(x) \geq \frac{1}{2(p-1)}\left(\frac{p}{p-2}\right)^{\frac{p-2}{2}}|x|^p. \tag{2.4}$$

Furthermore for $0 < \delta << 1$ we have

$$S(\delta x) = \frac{1}{2}\delta^2 x^2 + \mathcal{O}(\delta^3). \tag{2.5}$$

We continue by obtaining explicit constants in various embedding and interpolation inequalities.

LEMMA 2.2. *Let $\varphi \in \tilde{H}^1_{per}(\mathcal{Y})$. Then we have $|\varphi(x)| \leq \int_{\mathcal{Y}} |x-y|^{-1}|\nabla\varphi(y)|\,dy$ and*

$$\|\varphi\|_{L^{2q}(\mathcal{Y})} \leq \big(q+1\big)^{\frac{q+1}{2q}}\sqrt{\pi}\|\nabla\varphi\|_{L^2(\mathcal{Y})^2} \qquad \forall q \in]1,+\infty[. \tag{2.6}$$

Proof. The first inequality is classical (see, e.g., Chapter 7 in Gilbarg and Trudinger [6]). Inequality (2.6) is a consequence of the theory of the Riesz potentials V_μ, $(V_\mu f)(x) \equiv \int_{\mathcal{Y}} |x-y|^{2(\mu-1)} f(y)\,dy$, $f \in L^1(\mathcal{Y})$, applied to the case $\mu = \frac{1}{2}$. We refer to Gilbarg and Trudinger [6] for more details. □

LEMMA 2.3. *Let $\varphi \in \tilde{H}^1_{per}(\mathcal{Y})$. Then we have*

$$\|\varphi\|_{L^2(\mathcal{Y})} \leq \frac{1}{2\pi}\|\nabla\varphi\|_{L^2(\mathcal{Y})^2} \tag{2.7}$$

and

$$\|\varphi\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} \leq \frac{1}{2\pi}\|\varphi\|_{L^2(\mathcal{Y})}. \tag{2.8}$$

Proof. $4\pi^2$ is the first eigenvalue of the operator $\mathcal{A}$. Hence the second Poincaré's inequality and the condition $\int_{\mathcal{Y}} \varphi = 0$ imply (2.7). Estimate (2.8) is proved using the definition of the norm in $\tilde{H}^{-1}_{per}(\mathcal{Y}) = (\tilde{H}^1_{per}(\mathcal{Y}))'$ and (2.7):

$$\|\varphi\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} = \sup_{\zeta \in \tilde{H}^1_{per}(\mathcal{Y})} \frac{\int_{\mathcal{Y}} \varphi\zeta\,dy}{\|\nabla\zeta\|_{L^2(\mathcal{Y})^2}} \leq \frac{1}{2\pi}\|\varphi\|_{L^2(\mathcal{Y})}.$$

Note that

$$\|\varphi\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})} = \int_{\mathcal{Y}} |\nabla\theta|^2\,dy$$

for any periodic function θ such that $\mathcal{A}\theta = \varphi$. □

LEMMA 2.4. *Let $\varphi \in \tilde{H}^1_{per}(\mathcal{Y})$. Then we have*

$$\|\varphi\|_{L^{2+2\delta}(\mathcal{Y})} \leq \frac{\sqrt{1+\delta}}{\sqrt[4]{\pi}} \left(2\pi^{\frac{3}{2}}\right)^{\frac{\delta}{2(1+\delta)}} \|\nabla\varphi\|_{L^2(\mathcal{Y})^2}, \qquad 0 < \delta. \tag{2.9}$$

Proof. An estimate for the $L^{2+2\delta}$-norm is obtained by interpolation between $L^2(\mathcal{Y})$ and $L^{2+4\delta}(\mathcal{Y})$. Then inequalities (2.6) and (2.7) imply (2.9). ∎

We will need a particular form of Gronwall's inequality.

LEMMA 2.5. *Assume a, b, α, β are nonnegative constants with $\alpha < 1, \beta < 1$, and $0 < T < +\infty$. Then there exists a constant $M = M(b, \alpha, \beta, T) < +\infty$ such that for any integrable function $h : [0, T] \to \mathbb{R}$ satisfying*

$$0 \leq h(t) \leq at^{-\alpha} + b\int_0^t (t-s)^{-\beta} h(s)ds \quad \textit{for almost every} \quad t \in [0, T]$$

we have

$$0 \leq h(t) \leq aMt^{-\alpha} \quad \textit{a.e. on} \quad]0, T].$$

Proof. See, e.g., Henry [7, pp. 188–190]. ∎

Finally, for the comfort of the reader we recall the classic existence of the unique local mild solution from Henry [7]:

Let us consider the nonlinear Cauchy problem

$$\begin{cases} \dfrac{dx}{dt} + Bx = f(t, x), \quad t > t_0, \\ x(t_0) = x_0, \end{cases} \tag{2.10}$$

where it is assumed that B is a sectorial operator in the Banach space X and that f maps some open set $\mathcal{U}$ in $\mathbb{R} \times X^\alpha$ into X for some $\alpha \in [0, 1[$ and X^α being the domain of definition for $(B + \lambda I)^\alpha$, for some $\lambda > 0$. Suppose that f is locally Hölder continuous in t and locally Lipschitz in x on $\mathcal{U}$.

Then for any $(t_0, x_0) \in \mathcal{U}$ there exists $T = T(t_0, x_0) > 0$ such that the problem (2.10) has a unique mild solution x on $]t_0, t_0 + T[$ with initial value $x(t_0) = x_0$. More precisely, there exists a unique continuous function $x : [t_0, t_0 + T[\to X$ such that $x(t_0) = x_0$, and on $]t_0, t_0 + T[$ we have that $(t, x(t)) \in \mathcal{U}, x(t) \in D(B)$, $\frac{dx}{dt}(t)$ exists, $t \to f(t, x(t))$ is locally Hölder continuous, and $\int_{t_0}^{t_0+\rho} \|f(t, x(t))\|_X dt < +\infty$ for some $\rho > 0$, and the differential equation (2.10) is satisfied on $]t_0, t_0 + T[$.

3. Existence and uniqueness of a mild solution. We start with the position of the problem. Because of the singular nature of the parameter β we introduce the set $\mathcal{V}_\gamma, \gamma > 1$, by

$$\mathcal{V}_\gamma = \{z \in \tilde{H}^\gamma_{per}(\mathcal{Y}) : \int_{\mathcal{Y}} (1 - z^2(y))|\nabla \mathcal{A}^{-1} z|^2 dy \neq 0\}. \tag{3.1}$$

Obviously $\mathcal{V}_\gamma$ is an open set in $\tilde{H}^\gamma_{per}(\mathcal{Y})$.

Then we have the following.

DEFINITION 3.1. *Let $\omega_0 \in \mathcal{V}_\gamma$. Then a mild solution of the problem (1.1) on $[0, T[$ is a function $\omega \in C([0, T[; \mathcal{V}_\gamma)$ such that $\omega(0) = \omega_0$, $\omega \in C(]0, T[; \tilde{H}^2_{per}(\mathcal{Y}))$, $\frac{d\omega}{dt} \in C(]0, T[; L^2_0(\mathcal{Y}))$, and*

(3.2)

$$\frac{d\omega}{dt} + A\mathcal{A}\omega = -\operatorname{div}\{\omega G(\omega)\} + A\beta(\omega)\operatorname{div}\{(1-\omega^2)\nabla(\mathcal{A}^{-1}\omega)\} \equiv F(\omega) \quad on \quad]0,T[,$$

where

(3.3)

$$\beta(\xi) = \begin{cases} -\dfrac{\displaystyle\int_{\mathcal{Y}} \xi^2(y)dy}{\displaystyle\int_{\mathcal{Y}} (1-\xi^2(y))|\nabla\mathcal{A}^{-1}\xi|^2 dy} \text{ for } \xi \in L_0^2(\mathcal{Y}), \quad \text{such that} \\ \displaystyle\int_{\mathcal{Y}} (1-\xi^2)|\nabla\mathcal{A}^{-1}\xi|^2 \neq 0, \\ \\ -\infty \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

Remark 3.2. For $z \in L_0^{2+\delta}(\mathcal{Y}), \delta > 0$, we have $\nabla\mathcal{A}^{-1}z \in L^\infty(\mathcal{Y})^2$, and

$$\int_{\mathcal{Y}} (1-z^2(y))|\nabla\mathcal{A}^{-1}z|^2\, dy$$

is finite. □

Our first goal is to establish a local existence of a mild solution for (3.2). After that initial step we will prove the global existence.

In order to apply the abstract local existence theorem we have to establish the properties of F. We write the nonlinear mapping F from (3.2) in the form

$$F(\xi) = -\nabla\xi \cdot G(\xi) - A\beta(\xi)\xi(1-\xi^2) - 2A\beta(\xi)\xi\nabla\xi \cdot \nabla(\mathcal{A}^{-1}\xi), \tag{3.4}$$

and straightforward estimates give the following.

PROPOSITION 3.3. *$F : \mathcal{V}_\gamma \to L_0^2(\mathcal{Y})$ is locally Lipschitzian.*

Since we know that $\mathcal{A}$ is a sectorial operator and $F : \mathcal{V}_\gamma \to L_0^2(\mathcal{Y})$ is locally Lipschitz continuous, we are in the situation to use the general local existence result for nonlinear parabolic equations and obtain the following.

THEOREM 3.4. *Let A be a positive constant, and let $\omega_0 \in \mathcal{V}_\gamma, \gamma > 1$. Then there exists $T = T(\omega_0, \gamma) > 0$ such that* (3.2) *has a unique mild solution $\omega \in C([0,T[;\mathcal{V}_\gamma) \cap C^1(]0,T[;L_0^2(\mathcal{Y}))$ on $]0,T[$, with initial value $\omega(0) = \omega_0$.*

Remark 3.5. Using Theorem 3.3.4. from Henry [7] we have that either $T = +\infty$ or else there exists a sequence $t_n \to T$ as $n \to +\infty$ such that $\int_{\mathcal{Y}}(1-\omega^2(t_n))|\nabla\mathcal{A}^{-1}\omega(t_n)|^2 \to 0$ as $n \to +\infty$ or $\|\omega(t_n)\|_{\tilde{H}^\gamma_{per}(\mathcal{Y})} \to +\infty$. □

Now it is important to extend the existence to any time interval.

The usual approach is to get some a priori estimates for some norms, which stay bounded. Here the essential difficulty will be controlling the singular parameter $\beta(\omega)$.

Let us use the physical structure of our problem and obtain a uniform $L^\infty-$ bound on ω.

First we introduce an appropriate weak formulation for (3.2):

$$\int_0^t \int_{\mathcal{Y}} \frac{\partial\omega}{\partial\tau}\varphi\, dyd\tau + A\int_0^t \int_{\mathcal{Y}} \nabla\omega \cdot \nabla\varphi\, dyd\tau = \int_0^t \int_{\mathcal{Y}} \omega G(\omega) \cdot \nabla\varphi\, dyd\tau$$
$$- \int_0^t \int_{\mathcal{Y}} A\beta(\omega)(1-\omega^2)\nabla(\mathcal{A}^{-1}\omega) \cdot \nabla\varphi\, dyd\tau \quad \forall\varphi \in L^2(0,T;H^1_{per}(\mathcal{Y})). \tag{3.5}$$

We use (3.5) in proving the following result.

PROPOSITION 3.6. *Let $\omega_0 \in \mathcal{V}_\gamma$ and $-1 \le \omega_0 \le 1$ in $\mathcal{Y}$. Then $-1 \le \omega(t) \le 1$ in $\mathcal{Y}$ $\forall t \in [0,T[$.*

Proof. We use the truncation method (see, for example, Artola [1]). Let $w_\eta^+ = \sup\{(\omega-1)^+ - \eta, 0\} \equiv (\omega - \eta - 1)^+, \eta > 0$.

Then we insert $\varphi = \frac{1}{\eta} - \frac{1}{\eta + w_\eta^+}$ as a test function for (3.5). It follows that

$$\int_{\mathcal{Y}} \left\{ \frac{w_\eta^+(t)}{\eta} - \ln\left(1 + \frac{w_\eta^+(t)}{\eta}\right)\right\} dy + A\int_0^t \int_{\mathcal{Y}} \left| \frac{\nabla w_\eta^+}{\eta + w_\eta^+}\right|^2 dy d\tau$$
$$- A\int_0^t \beta(\omega) \int_{\mathcal{Y}} \{1 + \eta + w_\eta^+\} \left\{ w_\eta^+ + 2\ln\left(1 + \frac{w_\eta^+(t)}{\eta}\right)\right\} dy d\tau = 0.$$

Since $\beta(\omega) \le 0$, we get

$$\int_{\mathcal{Y}} \left\{ \frac{w_\eta^+(t)}{\eta} - \ln\left(1 + \frac{w_\eta^+(t)}{\eta}\right)\right\} dy \le 0 \quad \forall t.$$

Finally, $w_\eta^+(t) = 0$, implying $(\omega(t) - 1)^+ \le \eta$ $\forall \eta > 0$. Therefore $\omega(t) \le 1$ in $\mathcal{Y}$ $\forall t \in [0,T[$.

Inequality $\omega(t) \ge -1$ is proved analogously. ∎

We continue by proving that energy is conserved.

PROPOSITION 3.7. *Let ω be a mild solution for* (3.2), *corresponding to the initial datum $\omega_0 \in \mathcal{V}_\gamma$.*

Then we have

$$\|\omega(t)\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} = \|\omega_0\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} \quad \forall t \in [0,T[, \tag{3.6}$$

where we recall that

$$\|\zeta\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} \overset{def}{=} \|\nabla(\mathcal{A}^{-1}\zeta)\|_{L^2(\mathcal{Y})^2} \quad \forall \zeta \in L^2_0(\mathcal{Y}).$$

Proof. Using the regularity of $\psi = \mathcal{A}^{-1}\omega$, we have

$$-\Delta \frac{\partial \psi}{\partial t} = \frac{\partial \omega}{\partial t} \qquad \text{in} \quad \mathcal{Y},$$

which implies $\displaystyle\int_{\mathcal{Y}} \nabla \frac{\partial \psi}{\partial t} \cdot \nabla \psi \, dy = \int_{\mathcal{Y}} \frac{\partial \omega}{\partial t} \psi \, dy$. Finally,

$$\frac{1}{2}\int_{\mathcal{Y}} |\nabla(\mathcal{A}^{-1}\omega(t))|^2 \, dy = \frac{1}{2}\int_{\mathcal{Y}} |\nabla(\mathcal{A}^{-1}\omega_0)|^2 \, dy + \int_0^t \int_{\mathcal{Y}} \frac{\partial \omega}{\partial t} \psi \, dy d\tau. \tag{3.7}$$

Now using $\psi = \mathcal{A}^{-1}\omega$ as a test function in (3.5) and the definition of $\beta(\omega)$, we obtain

$$\int_0^t \int_{\mathcal{Y}} \frac{\partial \omega}{\partial t} \psi \, dy d\tau + A\int_0^t \int_{\mathcal{Y}} \nabla\omega \cdot \nabla\psi \, dy d\tau - A\int_0^t \int_{\mathcal{Y}} \omega^2 \, dy d\tau = 0. \tag{3.8}$$

Obviously, $\int_0^t \int_{\mathcal{Y}} \nabla\omega \cdot \nabla\psi \, dy d\tau = \int_0^t \int_{\mathcal{Y}} \omega^2 \, dy d\tau$ and (3.7) and (3.8) imply (3.6). ∎

In the next step we would like to use the fact that negative entropy should be decreasing during the evolution in time. We have the following.

PROPOSITION 3.8. *Let ω be a mild solution for* (3.2), *corresponding to the initial datum $\omega_0 \in \mathcal{V}_\gamma$, and $-1 \le \omega_0 \le 1$ in $\mathcal{Y}$. Then we have*

$$\int_{\mathcal{Y}} S(\omega(t))\,dy \le \int_{\mathcal{Y}} S(\omega_0)\,dy \quad \forall t \in [0,T[\tag{3.9}$$

with the entropy function S given by (2.3).

Proof. The idea is to use $S'(\omega)$ as a test function in (3.5). However, this is not possible since we do not know whether $S'(\omega) \in L^2(0,T;H^1_{per}(\mathcal{Y}))$ or not.

We introduce "regularized entropy functions" S_δ, $\delta > 0$, by

$$S_\delta(\xi) = \frac{1+\delta+\xi}{2}\ln(1+\delta+\xi) + \frac{1+\delta-\xi}{2}\ln(1+\delta-\xi) \tag{3.10}$$

$\forall \xi \in L^2_0(\mathcal{Y})$, $-1 \le \xi \le 1$. Obviously, $S_\delta \in C^\infty[-1,1]$ $\forall \delta > 0$. Now we choose $\varphi = S'_\delta(\omega)$ as a test function for (3.5). It gives

$$\begin{aligned}\int_{\mathcal{Y}} S_\delta(\omega(t))\,dy - \int_{\mathcal{Y}} S_\delta(\omega_0)\,dy + A(1+\delta)\int_0^t\int_{\mathcal{Y}} \frac{|\nabla\omega|^2}{(1+\delta)^2-\omega^2}\,dyd\tau \\ = A(1+\delta)\int_0^t(-\beta(\omega))\int_{\mathcal{Y}}\left\{\omega^2 - \frac{\delta(2+\delta)}{2(1+\delta)}\omega\ln\frac{1+\delta+\omega}{1+\delta-\omega}\right\}dyd\tau.\end{aligned} \tag{3.11}$$

It should be noticed that $-\beta(\omega) \ge 0$ and $x\ \ln\frac{1+\delta+x}{1+\delta-x} \ge 0$ on $[-1,1]$. Therefore, we only have to estimate $\int_0^t(-\beta(\omega))\int_{\mathcal{Y}}\omega^2\,dyd\tau$.

Before estimating $(-\beta)$ we remark that passing to the limit $\delta \to 0$ in (3.11) and using the theorem of B. Levi imply

$$A\int_0^t\int_{\mathcal{Y}}\frac{|\nabla\omega|^2}{1-\omega^2}\,dyd\tau = \int_{\mathcal{Y}} S(\omega_0)\,dy - \int_{\mathcal{Y}} S(\omega(t))\,dy + A\int_0^t(-\beta(\omega))\int_{\mathcal{Y}}\omega^2\,dyd\tau,$$

and, consequently, $\frac{|\nabla\omega|}{\sqrt{1-\omega^2}} \in L^2(]0,t[\times\mathcal{Y})$.

Let us now estimate $-\beta(\omega)$. We start with an obvious inequality:

$$\frac{1}{\int_{\mathcal{Y}}(1-\omega^2)|\nabla(\mathcal{A}^{-1}\omega)|^2} \le \frac{\int_{\mathcal{Y}}(1-\omega^2)|\mathcal{B}'(\omega)|^2|\nabla\omega|^2}{\{\int_{\mathcal{Y}}(1-\omega^2)\nabla(\mathcal{A}^{-1}\omega)\cdot\nabla\mathcal{B}(\omega)\}^2}$$

$$= \frac{\int_{\mathcal{Y}}(1-\omega^2)|\mathcal{B}'(\omega)|^2|\nabla\omega|^2}{\{\int_{\mathcal{Y}}\omega\int_{-1}^{\omega}(1-\xi^2)\mathcal{B}'(\xi)d\xi\}^2} \quad \forall\mathcal{B}\in C^1[-1,1].$$

We take $\mathcal{B}(\xi) = S'_\delta(\xi)$ and obtain

$$\int_{\mathcal{Y}}\omega\int_{-1}^{\omega}(1-\xi^2)S''_\delta(\xi)d\xi\,dy = (1+\delta)\int_{\mathcal{Y}}\left\{\omega^2 - \frac{\delta(2+\delta)}{2(1+\delta)}\omega\ \ln\frac{1+\delta+\omega}{1+\delta-\omega}\right\}dy.$$

Consequently,

$$0 \le -\beta(\omega) \le \frac{\int_{\mathcal{Y}}\omega^2\,dy\cdot\int_{\mathcal{Y}}\frac{|\nabla\omega|^2}{(1+\delta)^2-\omega^2}\,dy}{\left\{\int_{\mathcal{Y}}\left[\omega^2 - \frac{\delta(2+\delta)}{2(1+\delta)}\omega\ \ln\frac{1+\delta+\omega}{1+\delta-\omega}\right]dy\right\}^2}.$$

After inserting this inequality in (3.11) and passing to the limit $\delta \to 0$, we get the result. ∎

COROLLARY 3.9. *Let ω be a mild solution for* (3.2), *corresponding to the initial datum $\omega_0 \in \mathcal{V}_\gamma, \gamma > 1$, and $-1 \le \omega_0 \le 1$ in $\mathcal{Y}$. Then we have*

$$\int_{\mathcal{Y}} S(\omega(t_2))\,dy \le \int_{\mathcal{Y}} S(\omega(t_1))\,dy \quad \forall t_1, t_2 \in [0, T[, \quad t_1 < t_2. \tag{3.12}$$

Up to now, we have obtained an estimate on the L^∞-norm of the vorticity but no information on the derivatives. We come now to prove that it is equivalent to get an estimate on $\|\nabla\omega\|^2_{L^2(]0,t[\times\mathcal{Y})^2}$ or on $\|\beta(\omega)\|_{L^1(0,t)}$. We have the following result.

PROPOSITION 3.10. *Let ω be a mild solution for* (3.2) *with $\omega_0 \in \mathcal{V}_\gamma$, and $-1 \le \omega_0 \le 1$ in $\mathcal{Y}$. Then we have*

$$\|\nabla\omega\|^2_{L^2(]0,t[\times\mathcal{Y})^2} \le \frac{1}{2A}\|\omega_0\|^2_{L^2(\mathcal{Y})} + \frac{2}{3}\|\beta(\omega)\|_{L^1(0,t)} \tag{3.13}$$

and

$$4\pi^2 \le \frac{\int_{\mathcal{Y}} \omega^2(t)\,dy}{\|\omega_0\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})}} \le -\beta(\omega) \le \left(\frac{4}{\pi}\right)^2 \frac{\int_{\mathcal{Y}} |\nabla\omega(t)|^2\,dy}{\int_{\mathcal{Y}} \omega^2(t)\,dy} \le \frac{4}{\pi^4}\frac{\int_{\mathcal{Y}} |\nabla\omega(t)|^2\,dy}{\|\omega_0\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})}} \tag{3.14}$$

for (a.e.) t belonging to the interval of existence.

This result indicates that it is crucial to estimate β. Also, due to the inequalities (3.13) and (3.14), it is unlikely that some estimate for $\nabla\omega$ could be obtained by some clever choice of the test function. Finally, differentiating (3.2) does not seem to be promising either.

We try another approach. Now, in order to get the crucial estimate on β, we construct an invariant set of initial values.

Let $+\infty > p > 2$, and let $1 > \epsilon > 0$. We introduce the set $\mathcal{U}(p,\varepsilon)$ by

$$\begin{aligned}\mathcal{U}(p,\varepsilon) = \Big\{ & z \in L^2_0(\mathcal{Y}) : -1 \le z \le 1 \text{ a.e. in } \mathcal{Y}, \quad \|z\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})} \\ & \ge \frac{\mathcal{C}(p)}{(1-\varepsilon)} \left\{ \int_{\mathcal{Y}} S(z)\,dy \right\}^{1+2/p} \Big\},\end{aligned} \tag{3.15}$$

where $\mathcal{C}(p) > 0$ is a constant which will be prescribed later. Of course, since the energy is conserved and the (negative) entropy $\int_{\mathcal{Y}} S(z)\,dy$ is decreasing in time, the set $\mathcal{U}(p,\varepsilon)$ is conserved by the flow of (1.1).

The significance of $\mathcal{U}(p,\varepsilon)$ will appear in the following results.

PROPOSITION 3.11. *$\mathcal{U}(p,\varepsilon) \setminus \{0\}$ and $\mathcal{U}(p,\varepsilon) \cap \tilde{H}^\gamma_{per}(\mathcal{Y}) \setminus \{0\}$ are nonempty.*

Proof. Let $z_0 \in \tilde{H}^\gamma_{per}(\mathcal{Y}), -1 \le z_0 \le 1, z_0 \ne 0$, and let $\delta > 0$. We would like to prove $\delta z_0 \in \mathcal{U}(p,\varepsilon)$ for $\delta \le \delta_0$. Using (2.5) we get

$$\left| S(\delta z_0) - \frac{1}{2}\delta^2 z_0^2 \right| \le C\delta^3.$$

As $\frac{2p}{p+2} \in]1,2[$ we conclude that

$$\int_{\mathcal{Y}} S(\delta z_0)\,dy \le \frac{\delta^2}{2}\int_{\mathcal{Y}} z_0^2\,dy + C\delta^3 \le \delta^{\frac{2p}{p+2}} \|z_0\|^{2p/(p+2)}_{\tilde{H}^{-1}_{per}(\mathcal{Y})} \left\{ \frac{1-\varepsilon}{\mathcal{C}(p)} \right\}^{p/(p+2)}$$

for $\delta \le \delta_0$. □

For the function $z \in \mathcal{U}(p,\varepsilon)$ we have a natural estimate on $\beta(\omega)$. More precisely we have the following.

PROPOSITION 3.12. *Let $z \in \mathcal{U}(p,\varepsilon)$ for some $p > 2$ and $1 > \varepsilon > 0$, and let us take*

$$\mathcal{C}(p) = \frac{2}{\sqrt{\pi}} \left\{ \frac{4p(p-1)\pi^{3/2}}{p-2} \right\}^{2/p}. \tag{3.16}$$

Then we have

$$\int_{\mathcal{Y}} (1-z^2)|\nabla \mathcal{A}^{-1} z|^2 \, dy \ge \varepsilon \|z\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})}. \tag{3.17}$$

Proof. Let us estimate from below the term $\int_{\mathcal{Y}} (1-z^2)|\nabla \mathcal{A}^{-1} z|^2 \, dy$. We have

$$\begin{aligned} \int_{\mathcal{Y}} (1-z^2)|\nabla \mathcal{A}^{-1} z|^2 \, dy &= \int_{\mathcal{Y}} |\nabla \mathcal{A}^{-1} z|^2 \, dy - \int_{\mathcal{Y}} z^2(y)|\nabla \mathcal{A}^{-1} z|^2 \, dy \\ &= \|z\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})} - \int_{\mathcal{Y}} z^2(y)|\nabla \mathcal{A}^{-1} z|^2 \, dy \\ &\ge \|z\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})} - \|z\|^2_{L^p(\mathcal{Y})} \| |\nabla \mathcal{A}^{-1} z| \|^2_{L^{2p/(p-2)}(\mathcal{Y})}. \end{aligned} \tag{3.18}$$

Now (2.4) implies

$$\|z\|_{L^p(\mathcal{Y})} \le \{2(p-1)\}^{1/p} \left(\frac{p-2}{p} \right)^{(p-2)/2p} \left\{ \int_{\mathcal{Y}} S(z) \, dy \right\}^{1/p},$$

and for $p > 2$ (see Lemma 2.4)

$$\| |\nabla \mathcal{A}^{-1} z| \|^2_{L^{2p/(p-2)}(\mathcal{Y})} \le \frac{1}{\sqrt{\pi}} \frac{p}{p-2} \left(2\pi^{3/2}\right)^{2/p} \sum_{i=1}^{2} \int_{\mathcal{Y}} \left| \frac{\partial}{\partial y_i} \nabla \mathcal{A}^{-1} z \right|^2 dy$$

so that

$$\begin{aligned} \int_{\mathcal{Y}} (1-z^2)|\nabla \mathcal{A}^{-1} z|^2 \, dy &\ge \|z[\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})} \\ &- \frac{1}{\sqrt{\pi}} \left(\frac{4\pi^{3/2} p(p-1)}{p-2} \right)^{2/p} \left\{ \int_{\mathcal{Y}} S(z) \, dy \right\}^{2/p} \sum_{i=1}^{2} \int_{\mathcal{Y}} \left| \frac{\partial}{\partial y_i} \nabla \mathcal{A}^{-1} z \right|^2 dy. \end{aligned} \tag{3.19}$$

Using regularity we have

$$-\triangle \frac{\partial \psi}{\partial y_i} = \frac{\partial z}{\partial y_i}, \quad i = 1, 2 \text{ (where } \psi = \mathcal{A}^{-1} z),$$

which implies

$$\sum_i \int_{\mathcal{Y}} \left| \nabla \frac{\partial \psi}{\partial y_i} \right|^2 dy = \sum_i \int_{\mathcal{Y}} \frac{\partial z}{\partial y_i} \frac{\partial \psi}{\partial y_i} \, dy = -\sum_i \int_{\mathcal{Y}} z \frac{\partial^2 \psi}{\partial y_i^2} \, dy = \int_{\mathcal{Y}} z^2 \, dy.$$

Hence we have

$$\sum_i \int_{\mathcal{Y}} \left| \frac{\partial}{\partial y_i} \nabla \mathcal{A}^{-1} z \right|^2 dy = \int_{\mathcal{Y}} z^2 \, dy. \tag{3.20}$$

Now (2.4) implies $\|z\|^2_{L^2(\mathcal{Y})} \le 2\int_{\mathcal{Y}} S(z)\,dy$ so that $z \in \mathcal{U}(p,\varepsilon)$ gives

$$\int_{\mathcal{Y}} (1-z^2)|\nabla \mathcal{A}^{-1} z|^2\,dy \ge \|z\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})} - \frac{2}{\sqrt{\pi}}\left(\frac{4\pi^{3/2}p(p-1)}{p-2}\right)^{2/p}\left\{\int_{\mathcal{Y}} S(z)\right\}^{1+2/p} \ge \varepsilon\|z\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})},$$

and the proposition is proved. ∎

Now we are in a situation to state a global existence result.

THEOREM 3.13. *Let $\omega_0 \in \mathcal{U}(p,\varepsilon) \cap \tilde{H}^{\gamma}_{per}(\mathcal{Y}) \setminus \{0\}$ for some $p > 2, 1 > \varepsilon > 0$, and $\gamma > 1$. Then the unique mild solution for (3.2) exists for all $T > 0$.*

Proof. Using Propositions 3.12 and 3.7 we immediately find that the theorem can fail only if there exists a sequence $t_n \to T_{max} < +\infty$ such that $\|\omega(t_n)\|_{\tilde{H}^{\gamma}_{per}(\mathcal{Y})} \to +\infty$.

Let us use the classical expression for the solution

$$\omega(t) = e^{-t\mathcal{A}}\omega_0 + \int_0^t e^{-(t-s)\mathcal{A}} F(\omega(s))\,ds,$$

and take the norm

$$\|\omega\|_{\tilde{H}^{\gamma}_{per}(\mathcal{Y})} = \|\mathcal{A}^{\gamma/2}\omega\|_{L^2(\mathcal{Y})},$$

which gives

$$\|\omega(t)\|_{\tilde{H}^{\gamma}_{per}(\mathcal{Y})} \le \|e^{-\mathcal{A}t}\omega_0\|_{\tilde{H}^{\gamma}_{per}(\mathcal{Y})} + \int_0^t \|\mathcal{A}^{\gamma/2}e^{-\mathcal{A}(t-s)}\|_{op}\|F(\omega(s))\|_{L^2(\mathcal{Y})}\,ds,$$

where $\|\cdot\|_{op}$ denotes the operator norm on $\mathcal{L}\big(L^2(\mathcal{Y}), L^2(\mathcal{Y})\big)$.

Straightforward estimates give

$$\|F(\omega(s))\|_{L^2(\mathcal{Y})} \le C(1+|\beta(s)|)(1+\|\nabla\omega(s)\|_{L^2(\mathcal{Y})^2}).$$

But from Proposition 3.12 we know that $|\beta(s)|$ remains bounded so that we finally get

$$\|\omega(t)\|_{\tilde{H}^{\gamma}_{per}(\mathcal{Y})} \le \frac{C_1}{t^{\gamma/2}} + C_2\int_0^t \frac{1+\|\omega(s)\|_{\tilde{H}^{\gamma}_{per}(\mathcal{Y})}}{(t-s)^{\gamma/2}}\,ds. \tag{3.21}$$

After applying the generalization of Gronwall's inequality from Lemma 2.5 to (3.21), we conclude that $\|\omega(t)\|_{\tilde{H}^{\gamma}_{per}(\mathcal{Y})}$ remains bounded as $T \to T_{max}$.

Now Theorem 3.3.4. from Henry [7] implies global existence. ∎

Finally, we will discuss the case of nonsmooth ω_0. We have the following result.

THEOREM 3.14. *Let $\omega_0 \in \mathcal{U}(p,\varepsilon) \setminus \{0\}$ for some $p > 2$ and $1 > \varepsilon > 0$. Then for all $T > 0$ there exists a unique $\omega \in C([0,T]; L^2_0(\mathcal{Y})) \cap L^2(0,T;\tilde{H}^1_{per}(\mathcal{Y}))$, $\frac{d\omega}{dt} \in L^2(0,T;\tilde{H}^{-1}_{per}(\mathcal{Y}))$ such that*

$$\begin{cases} -1 \le \omega(t,x) \le 1 \quad \text{a.e. on} \quad]0,T[\times\mathcal{Y}, \\ \omega(0) = \omega_0, \quad \beta(\omega) \in L^\infty(0,T), \end{cases} \tag{3.22}$$

and it satisfies

(3.23)
$$\begin{cases} \displaystyle\int_0^T \left\langle \frac{d\omega}{dt}, \varphi \right\rangle_{\tilde{H}^{-1}_{per}(\mathcal{Y}), \tilde{H}^{1}_{per}(\mathcal{Y})} dt + A \int_0^T \int_{\mathcal{Y}} \nabla\omega \cdot \nabla\varphi \, dydt \\ \displaystyle - \int_0^T \int_{\mathcal{Y}} \omega G(\omega) \cdot \nabla\varphi \, dydt + A \int_0^T \beta(\omega) \int_{\mathcal{Y}} (1-\omega^2) \nabla \mathcal{A}^{-1} \omega \cdot \nabla\varphi \, dydt = 0 \\ \forall \varphi \in L^2(0,T; \tilde{H}^1_{per}(\mathcal{Y})). \end{cases}$$

Furthermore, the energy is conserved, i.e.,

$$\|\omega(t)\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} = \|\omega_0\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} \qquad \forall t \in [0,T], \tag{3.24}$$

and the (negative) entropy is decreasing, i.e.,

$$\int_{\mathcal{Y}} S(\omega(t_1)) \leq \int_{\mathcal{Y}} S(\omega(t_2)) \quad \textit{for} \quad t_1 \geq t_2. \tag{3.25}$$

Proof. We consider problem (3.2) with initial condition $\omega_0^\delta = \omega_0 * \rho_\delta$, where ρ_δ is a regularizing sequence such that

$$\begin{cases} \rho_\delta \in \mathcal{D}(\mathbb{R}^2), \quad \displaystyle\int_{\mathbb{R}^2} \rho_\delta(x)\, dx = 1, \quad \rho_\delta(x) \geq 0, \\ \text{supp } \rho_\delta \subset B(0, r_\delta), \qquad r_\delta \to 0 \quad \text{as} \quad \delta \to 0. \end{cases}$$

Since we have $\|\omega_0^\delta - \omega_0\|_{L^2(\mathcal{Y})} \to 0$ when $\delta \to 0$, we deduce easily that $\omega_0^\delta \in \mathcal{U}(p, \varepsilon/2) \cap \mathcal{V}_\gamma$, for $\delta \leq \delta_0$, δ_0 small enough.

Consequently, for δ small enough, we can apply Theorem 3.4 and get a unique mild solution ω^δ for all $T > 0$.

Now we take $\varphi = \omega^\delta$ in the variational equation (3.23) and obtain the standard a priori estimates

$$\begin{aligned} &\|\omega^\delta\|_{L^\infty(0,T;L^2_0(\mathcal{Y}))} \leq C\{1 + \|\omega_0^\delta\|_{L^2_0(\mathcal{Y})}\}, \\ &\|\nabla\omega^\delta\|_{L^2(0,T;L^2_0(\mathcal{Y}))} \leq C\{1 + \|\omega_0^\delta\|_{L^2_0(\mathcal{Y})}\}, \\ &\left\| \frac{\partial \omega^\delta}{\partial t} \right\|_{L^2(0,T;\tilde{H}^{-1}_{per}(\mathcal{Y}))} \leq C\{1 + \|\omega_0^\delta\|_{L^2_0(\mathcal{Y})}\}. \end{aligned}$$

Using the above a priori estimate we subtract a subsequence converging weakly star in the above functional spaces. The application of Aubin's lemma and passing once again to the subsequence gives, in addition, strong convergence in $L^2(]0,T[\times\mathcal{Y})$. Now the standard variational argument completes the proof of existence. Identity (3.24) and estimate (3.25) follow straightforwardly.

Uniqueness easily follows from Proposition 3.3 and the estimates (3.22), (3.24), and (3.25). ∎

Remark 3.15. The condition $\omega_0 \in \mathcal{U}(p, \varepsilon)$ gives a relation between the initial energy and the initial entropy which represents a sufficient condition for solvability for problem (3.2). It is an open problem to determine if it is the necessary condition for existence of a solution global in time. ∎

4. Asymptotic behavior. In this section we study the asymptotic behavior of the solutions of problem (1.1) given by Theorem 3.14.

The entropy functional $\int_{\mathcal{Y}} S(\omega)$ defines a natural Lyapunov functional for our evolution problem. However, we have uniform time estimates only for $\|\omega(t)\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})}$, $\|\omega(t)\|_{L^\infty(\mathcal{Y})}$, and $\int_{\mathcal{Y}} S(\omega(t))$. Since the functional $\omega_0 \longmapsto \int_{\mathcal{Y}} S(\omega_0)\,dy - \int_{\mathcal{Y}} S(\omega(t))\,dy$ is not weakly lower semicontinuous on $\mathcal{U}(p,\varepsilon)$, we need L^2-compactness to apply classical results on the asymptotic behavior; we are going to prove a new H^1- estimate uniform in time.

We start with the corresponding regularity result.

PROPOSITION 4.1. *Let $\omega_0 \in \mathcal{U}(p,\varepsilon) \cap \tilde{H}^1_{per}(\mathcal{Y})$, $\omega_0 \neq 0$. Then the unique solution given by Theorem* 3.14 *satisfies*

$$\omega \in L^2(0,T;\tilde{H}^2_{per}(\mathcal{Y})), \quad \frac{d\omega}{dt} \in L^2(0,T;L^2_0(\mathcal{Y}))$$

so that $\omega \in C([0,T];\tilde{H}^1_{per}(\mathcal{Y}))$.

Proof. We need only to go back to the proof of Theorem 3.14. In addition to the choice $\varphi = \omega^\delta$, we also take $\varphi = \frac{d\omega^\delta}{dt}$. Then

$$\int_0^T dt \int_{\mathcal{Y}} \left|\frac{d\omega^\delta}{dt}\right|^2 dy + \frac{A}{2}\int_{\mathcal{Y}} |\nabla\omega^\delta(T)|^2\,dy - \frac{A}{2}\int_{\mathcal{Y}} |\nabla\omega_0^\delta|^2\,dy = \int_0^T dt \int_{\mathcal{Y}} F(\omega^\delta(t))\frac{d\omega^\delta}{dt}\,dy.$$

Now, since $\displaystyle\int_{\mathcal{Y}} |\nabla\omega_0^\delta|^2\,dy$ is bounded, from the inequalities

$$\|F(\omega^\delta(s))\|_{L^2(\mathcal{Y})} \le C\big(1 + \|\nabla\omega^\delta(s)\|_{L^2(\mathcal{Y})^2}\big),$$
$$\int_0^T dt \int_{\mathcal{Y}} |\nabla\omega^\delta|^2\,dy \le C$$

we get

$$\left\|\frac{d\omega^\delta}{dt}\right\|^2_{L^2(]0,T[\times\mathcal{Y})} \le C_1 + C_2\left\|\frac{d\omega^\delta}{dt}\right\|_{L^2(]0,T[\times\mathcal{Y})}.$$

Thus, $\frac{d\omega^\delta}{dt}$ is bounded in $L^2(0,T;L^2_0(\mathcal{Y}))$, which implies $\frac{d\omega}{dt} \in L^2(0,T;L^2_0(\mathcal{Y}))$. Finally from $\frac{\partial\omega}{\partial t} - A\triangle\omega = F(\omega)$ we deduce $\omega \in L^2(0,T;\tilde{H}^2_{per}(\mathcal{Y}))$. ∎

Our next step is obtaining an $H^1(\mathcal{Y})-$estimate which is uniform in time.

We have the following result.

PROPOSITION 4.2. *Let $\omega_0 \in \mathcal{U}(p,\varepsilon) \cap \tilde{H}^1_{per}(\mathcal{Y}) \setminus \{0\}$. Then we have the estimate*

$$\|\nabla\omega(t)\|_{L^2(\mathcal{Y})^2} \le R. \tag{4.1}$$

Proof. Let ω be the solution given by Proposition 4.1. We have

$$\begin{aligned}\frac{\partial\omega}{\partial t} - A\triangle\omega = F(\omega) \equiv -\nabla\omega\cdot G(\omega) - A\beta(\omega)\omega(1-\omega^2)\\ - 2A\beta(\omega)\omega\nabla\omega\cdot\nabla(\mathcal{A}^{-1}\omega) \quad \text{a.e. on} \quad]0,T[\times\mathcal{Y}.\end{aligned} \tag{4.2}$$

Let us choose $-\triangle\omega$ as a test function. Then we get

$$\frac{d}{dt}\frac{1}{2}\|\nabla\omega(t)\|^2_{L^2(\mathcal{Y})^2} + A\|\triangle\omega(t)\|^2_{L^2(\mathcal{Y})} \leq \|F(\omega)(t)\|_{L^2(\mathcal{Y})}\|\triangle\omega\|_{L^2(\mathcal{Y})^2} \tag{4.3}$$

a.e. on $]0,T[$.

We need an estimate on $\|F(\omega)(t)\|_{L^2(\mathcal{Y})}$:

$$\begin{aligned}\|F(\omega)\|_{L^2(\mathcal{Y})} &\leq \|\nabla\omega\cdot G(\omega)\|_{L^2(\mathcal{Y})}\\ &+ A|\beta(\omega)|\||\omega(1-\omega^2)\|_{L^2(\mathcal{Y})} + 2A|\beta(\omega)|\|\omega\nabla\omega\cdot\nabla(\mathcal{A}^{-1}\omega)\|_{L^2(\mathcal{Y})}.\end{aligned} \tag{4.4}$$

Since $\|\omega\|_{L^\infty(\mathcal{Y})} \leq 1$, we have $\||\nabla(\mathcal{A}^{-1}\omega)|\|_{L^\infty(\mathcal{Y})} = \||G(\omega)|\|_{L^\infty(\mathcal{Y})} \leq C$, and from the interpolation inequality

$$\|\nabla\omega\|_{L^2(\mathcal{Y})} \leq C\|\omega\|^{1/2}_{L^2(\mathcal{Y})}\|\triangle\omega\|^{1/2}_{L^2(\mathcal{Y})}$$

we get

$$\|\nabla\omega\|_{L^2(\mathcal{Y})} \leq C\|\triangle\omega\|^{1/2}_{L^2(\mathcal{Y})}. \tag{4.5}$$

(For sake of simplicity we will denote by the same letter C the different constants that we will encounter.)

Then we have

$$\|\nabla\omega\cdot G(\omega)\|_{L^2(\mathcal{Y})} \leq C\|\triangle\omega\|^{1/2}_{L^2(\mathcal{Y})}, \tag{4.6}$$

$$|\beta(\omega)| \leq \frac{\int_{\mathcal{Y}}\omega^2(t)\,dy}{\varepsilon\|\omega_0\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})}} \leq \frac{1}{\varepsilon\|\omega_0\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})}} \leq C, \tag{4.7}$$

$$2A|\beta(\omega)|\|\omega\nabla\omega\nabla(\mathcal{A}^{-1}\omega)\|_{L^2(\mathcal{Y})} \leq C\|\triangle\omega\|^{1/2}_{L^2(\mathcal{Y})}, \tag{4.8}$$

and thus

$$\|F(\omega)(t)\|_{L^2(\mathcal{Y})} \leq C\big(1+\|\triangle\omega\|^{1/2}_{L^2(\mathcal{Y})}\big). \tag{4.9}$$

Now setting $v(t) = \|\nabla\omega\|^2_{L^2(\mathcal{Y})^2}$, (4.3) becomes

$$\frac{1}{2}\frac{d}{dt}v(t) + A\|\triangle\omega(t)\|^2_{L^2(\mathcal{Y})} \leq C\big(\|\triangle\omega\|^{3/2}_{L^2(\mathcal{Y})} + \|\triangle\omega\|_{L^2(\mathcal{Y})}\big).$$

A straightforward application of Young's inequality then gives

$$\frac{1}{2}\frac{d}{dt}v(t) + \frac{A}{2}\|\triangle\omega(t)\|^2_{L^2(\mathcal{Y})} \leq C,$$

but we have $\|\triangle\omega\|^2_{L^2(\mathcal{Y})^2} \geq \|\nabla\omega\|^2_{L^2(\mathcal{Y})^2}$ so that

$$\frac{d}{dt}v(t) + Av(t) \leq C, \tag{4.10}$$

which yields the bound (4.1). □

Our next step is to construct a continuous nonlinear semigroup connected with our evolution problem.

Let us consider the complete metric space $X = \mathcal{U}(p,\varepsilon), p \in]2,+\infty[$, endowed with the $L^2_0(\mathcal{Y})-$metric (it is not a linear space) and $X^* = X \setminus \{0\}$. It should be noticed that $\omega_0 \in X^*$ implies $\omega(t) \in X^*\ \forall t \geq 0$.

Let us define a family of maps $\{\mathcal{T}(t) : X^* \to X^*, \quad t \geq 0\}$ by setting

$$\mathcal{T}(t)\omega_0 = \omega(t,x) \quad \forall \omega_0 \in X^*, \tag{4.11}$$

where $\omega \in C\big([0,+\infty[,\, L^2_0(\mathcal{Y})\big)$ is the unique solution given by Theorem 3.14.

LEMMA 4.3. *The family of maps* $\{\mathcal{T}(t) : X^* \to X^*, \quad t \geq 0\}$ *is a continuous nonlinear semigroup.*

Proof. It is straightforward, and we leave it to the reader. □

Remark 4.4. It is easy to see that if $\omega_0^n \to 0$ in X, then $\mathcal{T}(t)(\omega_0^n) \to 0\ \forall t > 0$. Hence the semigroup $\mathcal{T}$ is extended continuously to X by defining $\mathcal{T}(t)0 = 0$. □

Let us consider the functional

$$\mathcal{J}(\varphi) = \int_{\mathcal{Y}} S(\varphi)\, dy.$$

Obviously, $\mathcal{J}$ is continuous on X, and we have seen that $\mathcal{J}\big(\mathcal{T}(t)(\omega_0)\big)$ is a decreasing function of t so that it is a Lyapunov functional for $\mathcal{T}$.

We define the $\Omega-$limit set for ω_0 by

$$\Omega(\omega_0) = \{z \in X : \exists t_n \to +\infty \quad \text{such that} \quad \mathcal{T}(t_n)\omega_0 \to z\} \equiv \bigcap_{s\geq 0} \overline{\bigcup_{t\geq s} \mathcal{T}(t)}\omega_0. \tag{4.12}$$

PROPOSITION 4.5. *For every* $\omega_0 \in X \cap \tilde{H}^1_{per}(\mathcal{Y})$, $\Omega(\omega_0)$ *is a compact connected and nonempty subset of* X. *Furthermore,* $\Omega(\omega_0)$ *is* $\mathcal{T}$*-invariant and* $\mathcal{J}$ *has a constant value on it.*

Proof. First, because of Proposition 4.2, the orbit of ω_0 defined by $\gamma(\omega_0) \equiv \big\{\mathcal{T}(t)\omega_0 : \quad t \geq 0\big\}$ is relatively compact in X. Consequently, Proposition 4.5 is a direct consequence of Propositions 2.1 and 2.2 from Dafermos [4]. □

Our next goal is to describe $\Omega(\omega_0)$ in terms of statistical equilibrium. We will prove that any $\overline{\omega} \in \Omega(\omega_0)$ is a Gibbs state (as described in [11], [12]).

THEOREM 4.6. *Let us consider* $\omega_0 \in X \cap \tilde{H}^1_{per}(\mathcal{Y})$, *and let* $\overline{\omega}$ *be any element of* $\Omega(\omega_0)$. *Then there exist constants* $\lambda > 0$ *and* $C \in \mathbb{R}$ *such that*

$$\overline{\omega} = \tanh\big(\lambda\overline{\psi} + C\big) \quad \text{on} \quad \mathcal{Y}, \tag{4.13}$$

where $\overline{\psi}$ *is the stream function associated with* $\overline{\omega}$.

Furthermore, taking any sequence $\{t_n\}, t_n \to +\infty$ *such that* $\omega(t_n) \rightharpoonup \overline{\omega}$ *weakly in* $\tilde{H}^1_{per}(\mathcal{Y})$, *we have* $\beta(\omega(t_n)) \to \beta_\infty = -\lambda$.

Proof. From Proposition 4.2 we deduce that $\overline{\omega} \in X \cap \tilde{H}^1_{per}(\mathcal{Y})$. Let us now take $\overline{\omega}$ as initial datum and consider $\omega(t) = \mathcal{T}(t)\overline{\omega}$; we have $\omega \in C\big([0,+\infty[,\, \tilde{H}^1_{per}(\mathcal{Y})\big)$.

Let us now consider the function

$$F_\delta(t,y) = \frac{\nabla\omega(t,y)}{\sqrt{(1+\delta^2) - \omega^2}}, \quad \delta > 0.$$

As $\delta \to 0$, $|F_\delta|$ is obviously increasing, and we know from the proof of Proposition 3.8 that

$$\int_0^t \int_{\mathcal{Y}} |F_\delta(t,y)|^2 \, dydt \le C \quad (\text{when } \delta \to 0).$$

We deduce that there exists a function $\zeta(t,y) \in L^2(]0,T[\times\mathcal{Y})$ such that $F_\delta \to \zeta$ a.e. on $]0,T[\times\mathcal{Y}$ and strongly in $L^2(]0,T[\times\mathcal{Y})$.

Moreover ζ satisfies

$$\sqrt{1-\omega^2}\zeta = \nabla\omega \qquad \text{a.e. on }]0,T[\times\mathcal{Y}.$$

Now taking the limit $\delta \to 0$, we get from (3.11)

$$A\int_0^t \int_{\mathcal{Y}} \zeta^2 \, dyd\tau = \int_{\mathcal{Y}} S(\overline{\omega})\, dy - \int_{\mathcal{Y}} S(\omega(t))\, dy - A\int_0^t \int_{\mathcal{Y}} \beta(\omega)\omega^2 \, dyd\tau \quad \forall t \ge 0.$$

Since $\mathcal{J}$ is constant on $\Omega(\omega_0)$, we have for all t

$$\int_0^t \int_{\mathcal{Y}} \zeta^2 \, dyd\tau = -\int_0^t \int_{\mathcal{Y}} \beta(\omega)\omega^2 \, dyd\tau.$$

Taking the derivative with respect to t, we get

$$\int_{\mathcal{Y}} \zeta^2 \, dy = -\int_{\mathcal{Y}} \beta(\omega)\omega^2 \, dy \quad \text{for almost all } t \ge 0.$$

Since

$$\int_{\mathcal{Y}} \omega^2 \, dy = \int_{\mathcal{Y}} \nabla\omega \cdot \nabla\psi \, dy = \int_{\mathcal{Y}} \zeta\sqrt{1-\omega^2}\nabla\psi \, dy$$

and

$$-\beta(\omega) = \frac{\int_{\mathcal{Y}} \omega^2 \, dy}{\int_{\mathcal{Y}} (1-\omega^2)|\nabla\psi|^2 \, dy},$$

we get

$$\left(\int_{\mathcal{Y}} \zeta^2 \, dy\right)^{1/2} \left(\int_{\mathcal{Y}} (1-\omega^2)|\nabla\psi|^2 \, dy\right)^{1/2} = \int_{\mathcal{Y}} \sqrt{1-\omega^2}\zeta\cdot\nabla\psi \, dy \quad \text{for almost all } t \ge 0.$$

Due to the Cauchy–Schwarz inequality, this is possible if and only if ζ and $\sqrt{1-\omega^2}\nabla\psi$ are (positively) colinear for almost all $t \ge 0$, which implies that $\nabla\omega$ and $(1-\omega^2)\nabla\psi$ are (positively) colinear a.e. on $\mathcal{Y}$ for almost all $t \ge 0$. But $\omega \in C\big([0,+\infty[, \tilde{H}^1_{per}(\mathcal{Y})\big)$, and we deduce that $\nabla\overline{\omega}$ and $(1-\overline{\omega}^2)\nabla\overline{\psi}$ are (positively) colinear. Now, since $\nabla\overline{\omega} \not\equiv 0$ and $(1-\overline{\omega}^2)\nabla\psi \not\equiv 0$, we have

$$\nabla\overline{\omega} = \lambda(1-\overline{\omega}^2)\nabla\overline{\psi} \quad \text{a.e. on } \mathcal{Y}, \text{ with } \lambda > 0.$$

This implies that $\nabla\overline{\omega} \in L^\infty(\mathcal{Y})^2$ so that $\overline{\omega}$ is Lipschitz continuous.

Let us now prove that $-1 < \overline{\omega} < 1$ on $\mathcal{Y}$. We define $\mathcal{Y}^* = \{y \in \mathcal{Y} : -1 < \overline{\omega}(y) < 1\}$, $\mathcal{Y}^*$ is an open set of $\mathcal{Y}$. Let us assume that $\mathcal{Y}^* \ne \mathcal{Y}$ and take $y^* \in \partial\mathcal{Y}^*$. We denote by $\mathcal{Y}^*_1$ a connected component of $\mathcal{Y}^*$ such that $y^* \in \partial\mathcal{Y}^*_1$.

On $\mathcal{Y}_1^*$ we have

$$\lambda\nabla\overline{\psi} = \frac{\nabla\overline{\omega}}{1-\overline{\omega}^2},$$

i.e.,

$$\nabla\left(\lambda\overline{\psi} - \frac{1}{2}\ln\frac{1+\overline{\omega}}{1-\overline{\omega}}\right) = 0 \quad \text{on} \quad \mathcal{Y}_1^*;$$

that is,

$$\overline{\omega} = \tanh\bigl(\lambda\overline{\psi} + C\bigr),$$

but $\overline{\psi}$ is bounded so that $\overline{\omega}(y)$ cannot converge to 1 or -1 when $y \to y^*$. Thus $\mathcal{Y}^* = \mathcal{Y}$, and the above relationship holds on all $\mathcal{Y}$.

The proof of the last assertion is straightforward. ∎

As a straightforward consequence of Theorem 4.6, we have the following.

COROLLARY 4.7. *Any $\overline{\omega} \in \Omega(\omega_0)$ is a critical point of the variational problem $(\mathcal{V}.\mathcal{P}.)$.*

Remark 4.8. The variational problem $(\mathcal{V}.\mathcal{P}.)$ always has solutions. Indeed, let us take $\omega_0 \in L^2_0(\mathcal{Y})$ such that $-1 \le \omega_0 \le 1$ a.e. We define the set

$$\mathcal{E} = \bigl\{\omega \in L^2_0(\mathcal{Y}) : \quad -1 \le \omega_0 \le 1 \quad \text{a.e. and } \|\omega\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})} = \|\omega_0\|_{\tilde{H}^{-1}_{per}(\mathcal{Y})}\bigr\}.$$

$\mathcal{E}$ is a compact subset of $L^2_0(\mathcal{Y})$ for the weak L^2-topology. Thus, the convex, positive, and weakly lower semicontinuous functional $\mathcal{J}(\omega)$ reaches its infimum on $\mathcal{E}$.

Theorem 4.7 shows that for $\omega_0 \in X \cap \tilde{H}^1_{per}(\mathcal{Y})$, $(\mathcal{V}.\mathcal{P}.)$ has critical points such that $\|\omega\|_{L^\infty(\mathcal{Y})} < 1$. This consequence is not trivial since nothing ensures a priori that the solution ω^* of $(\mathcal{V}.\mathcal{P}.)$ does not reach the values ± 1. ∎

Remark 4.9. Let us consider any critical point $\overline{\omega} \neq 0$ such that

$$\overline{\omega} = \tanh\bigl(\lambda\overline{\psi} + C\bigr), \quad \lambda > 0.$$

Then, we have $\nabla\overline{\omega} = \lambda(1-\overline{\omega}^2)\nabla\overline{\psi}$ on $\mathcal{Y}$ so that

$$\int_{\mathcal{Y}} \nabla\overline{\omega}\cdot\nabla\overline{\psi}\,dy = \lambda\int_{\mathcal{Y}}(1-\overline{\omega}^2)|\nabla\overline{\psi}|^2\,dy < \lambda\int_{\mathcal{Y}}|\nabla\overline{\psi}|^2\,dy,$$

and integrating by parts yields

$$\lambda > \frac{\int_{\mathcal{Y}}\overline{\omega}^2\,dy}{\int_{\mathcal{Y}}\overline{\omega}\overline{\psi}\,dy} \ge 4\pi^2 = \lambda_1,$$

where λ_1 is the first eigenvalue of the operator $\mathcal{A}$.

Since $\overline{\omega}$ is a function of $\overline{\psi}$, we know that it is a stationary solution of incompressible Euler equations. But due to the above inequality we are in a situation where Arnold's stability criterion (see, e.g., chapter 3 in Marchioro and Pulvirenti [8]) does not apply.

Another related consequence is that the classical appeal to the free energy functional for studying $(\mathcal{V}.\mathcal{P}.)$ fails.

Indeed, let us consider the free energy functional

$$\mathcal{F}(\xi) \equiv \mathcal{J}(\xi) - \frac{\lambda}{2}\|\xi\|^2_{\tilde{H}^{-1}_{per}(\mathcal{Y})},$$

defined on $\{\omega \in L^2_0(\mathcal{Y}) : \quad -1 \leq \omega \leq 1 \quad \text{a.e.}\}$.

It is easy to see that for $\lambda > \lambda_1$, $\mathcal{F}(\omega)$ need not be convex.

This situation is a special feature of the periodic geometry which, on one hand, simplifies the study of (1.1), getting rid of some intricacies connected with the boundary conditions. On the other hand it introduces a great complexity in the study of $(\mathcal{V}.\mathcal{P}.)$, describing the statistical equilibrium: we *never* have a case where a simple criterion ensures existence of a unique critical point. The case of a bounded domain with boundary is quite different (see [11], [12], [17]). □

REFERENCES

[1] M. Artola, *Sur une classe de problèmes paraboliques quasi-linéaires*, Boll. Un. Mat. Ital. B, 5 (1986), pp. 51–70.

[2] S. Chandrasekhar, *Stochastic problems in physics and astronomy*, Rev. Modern Phys., 15 (1943), pp. 1–89.

[3] P. H. Chavanis, J. Sommeria, and R. Robert, *Statistical mechanics of* 2D *vortices and stellar systems*, Astrophys. J., 471 (1996), pp. 385–399.

[4] C. Dafermos, *Asymptotic behavior of solutions of evolution equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978, pp. 103–127.

[5] M. A. Denoix, J. Sommeria, and A. Thess, *Two-dimensional turbulence: The prediction of coherent structures by statistical mechanics*, in Progress in Turbulence Research, H. Branover and Y. Unger, eds., American Institute of Aeronautics and Astronautics, New York, 1994, pp. 88–107.

[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Heidelberg, 1983.

[7] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1993.

[8] C. Marchioro and M. Pulvirenti, *Mathematical Theory of Incompressible Nonviscous Fluids*, Springer-Verlag, New York, 1994.

[9] J. Michel and R. Robert, *Large deviations for Young measures and statistical mechanics of infinite dimensional dynamical systems with conservation law*, Comm. Math. Phys., 159 (1994), pp. 195–215.

[10] J. Miller, P. B. Weichman, and M. C. Cross, *Statistical mechanics: Euler equation and Jupiter's red spot*, Phys. Rev. A, 45 (1992), pp. 2328–2359.

[11] R. Robert and J. Sommeria, *Statistical equilibrium states for two-dimensional flows*, J. Fluid Mech., 229 (1991), pp. 291–310.

[12] R. Robert, *A maximum entropy principle for two-dimensional Euler equations*, J. Statist. Phys., 65 (1991), pp. 531–553.

[13] R. Robert and J. Sommeria, *Relaxation towards a statistical equilibrium state in two-dimensional perfect fluid dynamics*, Phys. Rev. Lett., 69 (1992), pp. 2276–2279.

[14] R. Robert, *Statistical mechanics and hydrodynamical turbulence*, in Proceedings of the International Congress of Mathematicians, Zürich, 1994, Birkhäuser-Verlag, Basel, Switzerland, 1995, pp. 1523–1531.

[15] R. Robert, *Relaxation towards a statistical equilibrium state in two-dimensional perfect fluid dynamics*, in Proc. XIth Internat. Congress of Math. Physics, Paris, 1994, International Press, Cambridge, MA, 1995, pp. 583–592.

[16] R. Robert, C. Rosier, *The modelling of small scales in* 2D *turbulent flows*, J. Statist. Phys., 86 (1997), pp. 481–515.

[17] J. Sommeria, C. Staquet, and R. Robert, *Final equilibrium state of a two-dimensional shear layer*, J. Fluid Mech., 233 (1991), pp. 661–689.

[18] V. I. Youdovitch, *Non-stationary flow of an incompressible liquid*, Zh. Vychisl. Mat. i Mat. Fiz., 3 (1963), pp. 1032–1066.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1256–1267, September 1998

012

SCATTERING THEORY FOR THE HARTREE EQUATION*

NAKAO HAYASHI†, PAVEL I. NAUMKIN‡, AND TOHRU OZAWA§

Abstract. We study the scattering problem for the Hartree equation

$$i\partial_t u = -\frac{1}{2}\Delta u + f(|u|^2)u, \quad (t,x) \in \mathbf{R}\times\mathbf{R}^n,$$

with initial data $u(0,x) = u_0(x), x \in \mathbf{R}^n$, where $f(|u|^2) = V * |u|^2, V(x) = \lambda|x|^{-1}, \lambda \in \mathbf{R}, n \geq 2$. We prove that for any $u_0 \in H^{0,\gamma} \cap H^{\gamma,0}$, with $\frac{1}{2} < \gamma < \frac{n}{2}$, such that the value $\epsilon = \|u_0\|_{0,\gamma} + \|u_0\|_{\gamma,0}$ is sufficiently small, there exist unique $u_\pm \in H^{\sigma,0} \cap H^{0,\sigma}$ with $\frac{1}{2} < \sigma < \gamma$ such that for all $|t| \geq 1$

$$\left\| u(t) - \exp\left(\mp i f(|\hat{u}_\pm|^2)\left(\frac{x}{t}\right)\log|t|\right) U(t)u_\pm \right\|_{L^2} \leq C\epsilon|t|^{-\mu+7\nu},$$

where $\mu = \min(1, \frac{\gamma}{2})$, $0 < \nu < \min(1, \frac{\gamma-\sigma}{12})$, $\hat{\varphi}$ denotes the Fourier transform of φ, $U(t)$ is the free Schrödinger evolution group, and $H^{m,s}$ is the weighted Sobolev space defined by

$$H^{m,s} = \{\varphi \in \mathcal{S}'; \|\varphi\|_{m,s} = \|(1+|x|^2)^{s/2}(1-\Delta)^{m/2}\varphi\|_{L^2} < \infty\}.$$

Key words. asymptotic behavior, Hartree equation, scattering

AMS subject classifications. 35Q55

PII. S0036141096312222

1. Introduction.

This paper is devoted to the study of the asymptotic behavior for large time of solutions to the Cauchy problem for the Hartree equation

$$\begin{cases} i\partial_t u = -\frac{1}{2}\Delta u + f(|u|^2)u, \quad (t,x) \in \mathbf{R}\times\mathbf{R}^n, \\ u(0,x) = u_0(x), \quad x \in \mathbf{R}^n, \end{cases} \tag{1.1}$$

where

$$f(|u|^2) = V * |u|^2 = \int V(x-y)|u|^2(y)dy, \quad V(x) = \lambda|x|^{-1}, \quad \lambda \in \mathbf{R}, \quad \text{and} \quad n \geq 2.$$

There is a large amount of literature on the Cauchy problem (1.1) and the asymptotic behavior in time of solutions for (1.1) with $\lambda > 0$ (see [2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 22]). The nonlinearity in equation (1.1) is critical from the point of view of large time asymptotic behavior of solutions since the L^2 norm of the nonlinear term is not integrable in time at infinity. The scattering problem for (1.1) has been studied in the framework of the nonexistence of scattering states [14] and of the existence of modified wave operators [6]. If the initial data is $u_0, xu_0 \in L^2$. Some time decay estimates of the nonlinearity $f(|u|^2)$ in L^p norms were obtained in [3] by using the pseudoconformal conservation law, and it was proved in papers [18, 19] that for large time the potential $tf(|u(t,tx)|^2)$ behaves as the Coulomb potential $\frac{1}{|x|}$. It

*Received by the editors November 18, 1996; accepted for publication (in revised form) July 23, 1997; published electronically June 18, 1998.
http://www.siam.org/journals/sima/29-5/31222.html

†Department of Applied Mathematics, Science University of Tokyo 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162, Japan (nhayashi@rs.kagu.sut.ac.jp).

‡Instituto de Física y Matemáticas, Universidad Michoacana AP 2-82, CP 58040, Morelia, Michoacan, Mexico (naumkin@ifm1.ifm.umich.mx). The work of this author was supported by Consejo Nacional de Ciencia y Tecnologia de Mexico (Conacyt).

§Department of Mathematics, Hokkaido University, Sapporo 060, Japan.

seems that the decay rates of solutions to (1.1) obtained in [3, 4, 8, 9, 11, 13] through the pseudoconformal conservation law are not sufficient to obtain the existence and uniqueness of the modified scattering states (see [16] for a study in this direction). The only exception is [10], where the existence of solutions with the same decay rate as in the free case is proved and the asymptotic profile of the solutions is obtained. In the present paper we propose a new setting for the study of large time behavior of solutions to (1.1) to make clear the connection with the theory of long range scattering for (1.1). Although the method of the present paper follows [10], the argument here is different from the previous one in some respects and requires a number of sharp estimates. Our approach here is based on the sharp L^p estimates of the time decay rate of the solutions to the Cauchy problem (1.1). As far as we know there are no other results concerning the sharp L^p time decay estimates of solutions to the Cauchy problem (1.1) for the critical case in higher space dimensions under consideration. To derive the desired L^p estimates of the solutions we have to introduce a certain phase function since the previous methods [4, 11, 12] based solely on the a priori estimates of the L^2 norm of $(x + it\nabla)u(t)$ without specifying any phase function do not work for the critical case. Also we extensively use an explicit representation of the free Schrödinger evolution group (see formula (1.4) below). We note that the method presented here is general enough since it is also applicable to a wide class of nonlinear Schrödinger equations with derivatives in the nonlinear term.

We denote by $\mathcal{F}\varphi$ or $\hat{\varphi}$ the Fourier transform of φ defined by

$$\mathcal{F}\varphi(\xi) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int e^{-ix\xi}\varphi(x)dx$$

and let $\mathcal{F}^{-1}\varphi(x)$ be the inverse Fourier transform of φ:

$$\mathcal{F}^{-1}\varphi(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int e^{ix\xi}\varphi(\xi)d\xi.$$

The free Schrödinger evolution group $U(t) = e^{\frac{it}{2}\Delta}$ is given by

$$U(t)\varphi = \frac{1}{(2\pi it)^{\frac{n}{2}}} \int e^{\frac{i(x-y)^2}{2t}}\varphi(y)dy = \mathcal{F}^{-1}e^{-\frac{it}{2}\xi^2}\mathcal{F}\varphi.$$

We define the weighted Sobolev space $H^{m,s}$ by $H^{m,s} = \{\varphi \in \mathcal{S}'; \|\varphi\|_{m,s} = \|(1 + |x|^2)^{s/2}(1-\Delta)^{m/2}\varphi\| < \infty\}$, where $m, s \in \mathbf{R}$ and $\|\cdot\|$ denotes the usual L^2 norm.

We now state our results in this paper.

THEOREM 1.1. *We assume that $u_0 \in H^{\gamma,0} \cap H^{0,\gamma}$ and $\epsilon = \|u_0\|_{\gamma,0} + \|u_0\|_{0,\gamma}$ is sufficiently small, where $\frac{1}{2} < \gamma < \frac{n}{2}$, $n \geq 2$.*

Then there exists a unique global solution u of the Hartree equation (1.1) *such that $u \in C(\mathbf{R}; H^{\gamma,0} \cap H^{0,\gamma})$ and*

$$\sup_{\alpha\in[\frac{1}{2},\sigma]} \sup_{t\in\mathbf{R}} (1 + |t|)^{\alpha}\|u(t)\|_{p(\alpha)} \leq C\epsilon,$$

where $\frac{1}{2} < \sigma < \gamma$, $p(\alpha) = \frac{2n}{n-2\alpha}$.

THEOREM 1.2. *Let u be the solution of* (1.1) *obtained in Theorem* 1.1.

Then for any initial data u_0 satisfying the conditions of Theorem 1.1, *there exist unique functions $u_\pm \in H^{\sigma,0} \cap H^{0,\sigma}$, $\frac{1}{2} < \sigma < \gamma$, such that for all $|t| \geq 1$*

$$\|u(t) - \exp\left(\mp if(|\hat{u}_\pm|^2)\left(\frac{x}{t}\right)\log|t|\right)U(t)u_\pm\| \leq C\epsilon|t|^{-\mu+7\nu}, \tag{1.2}$$

where $\mu = \min(1, \frac{\gamma}{2})$, $0 < \nu < \min\left(1, \frac{\gamma-\sigma}{12}\right)$.

In the previous paper [10] the following results were shown. When the initial data $u_0 \in H^{\gamma,0} \cap H^{0,\gamma}$, where $\gamma > \frac{n}{2}$, are such that the norm $\epsilon = \|u_0\|_{\gamma,0} + \|u_0\|_{0,\gamma}$ is sufficiently small, then there exists a unique global solution u of the Hartree equation (1.1) such that $u \in C(\mathbf{R}; H^{\gamma,0} \cap H^{0,\gamma})$ and $\|u(t)\|_\infty \le C\epsilon(1+|t|)^{-n/2}$. Moreover there exist unique functions $\Phi \in L^\infty$ and $\hat{u}_+ \in L^\infty \cap L^2$ such that $\| \int_1^t f(|\hat{u}(\tau)|^2)\frac{d\tau}{\tau} - f(|\hat{u}_+|^2)\log t - \Phi\|_\infty \le C\epsilon t^{-\xi\eta}$ and $\|\mathcal{F}(U(-t)u)(t)\exp\left(i\int_1^t f(|\hat{u}(\tau)|^2)\frac{d\tau}{\tau}\right) - \hat{u}_+ e^{i\Phi}\|_k \le C\epsilon t^{-\xi}$ for $t \ge 1$, where $k = 2$ or ∞, $0 < \eta < \frac{2}{n}$, $2\xi + \frac{n}{2} < \gamma$, and $0 < \xi < 1$. Furthermore the following asymptotic formula is valid for large time t uniformly with respect to $x \in \mathbf{R}^n$:

$$u(t,x) = \frac{1}{(2\pi i t)^{\frac{n}{2}}}\hat{u}_+\left(\frac{x}{t}\right)\exp\left(i\frac{x^2}{2t} - if(|\hat{u}_+|^2)\left(\frac{x}{t}\right)\log t\right) + O\left(\epsilon t^{-\frac{n}{2}-\xi\eta}\right) \tag{1.3}$$

with the estimate $\|\mathcal{F}\left(U(-t)u\right)(t) - \hat{u}_+\exp\left(-if\left(|\hat{u}_+|^2\right)\log t\right)\|_k \le C\epsilon t^{-\xi\eta}$, where $k = 2$ or ∞. If we write formula (1.3) in the form

$$u(t,x) = \exp\left(-if(|\hat{u}_+|^2)\left(\frac{x}{t}\right)\log t\right)U(t)u_+ + O\left(\epsilon t^{-\frac{n}{2}-\xi\eta}\right),$$

then we get the inequality

$$\|u(t) - \exp\left(-if(|\hat{u}_+|^2)\left(\frac{x}{t}\right)\log t\right)U(t)u_+\|_\infty \le C\epsilon t^{-\frac{n}{2}-\xi\eta},$$

which is similar to (1.2) but gives the estimate in the uniform norm. The setting in Theorem 1.2 fits more closely to the theory of long-range scattering developed in [6], where the existence of modified wave operators for (1.1) has been proved. Roughly speaking, the result of Theorem 1.2 implies the asymptotic completeness of the modified wave operators. Another advantage of the setting in Theorem 1.2 over the previous one in [10] is that for the L^2 theory of scattering one needs only the requirement $\gamma > \frac{1}{2}$ for the index of the weighted Sobolev spaces as compared to the previous assumption, namely $\gamma > \frac{n}{2}$.

We organize our paper as follows. Below we explain the necessary notations and then in section 2 we give some preliminary results. In Lemma 2.1 we formulate well-known embedding results for the Sobolev spaces. Lemma 2.2 gives the sharp time decay estimate of the L^p norm of the function in terms of the free evolution group $U(t)$. In Lemma 2.3 we prove the estimates of the nonlinearity of equation (1.1) in the weighted Sobolev spaces. In section 3 we prove Theorems 1.1 and 1.2 by using a priori estimates of the solutions obtained in Lemma 3.2 in a space X_T. The function space X_T is the following:

$$X_T = X_T^\nu = \Bigg\{\varphi \in C([-T,T];\mathcal{S}'); |||\varphi|||_{X_T} = \sup_{t\in[-T,T]}(1+|t|)^{-\nu}\|\varphi(t)\|_{\gamma,0}$$
$$+ \sup_{t\in[-T,T]}(1+|t|)^{-\nu}\|U(-t)\varphi(t)\|_{0,\gamma} + \sup_{\alpha\in[\frac{1}{2},\sigma]}\sup_{t\in[-T,T]}(1+|t|)^{\alpha}\|\varphi(t)\|_{p(\alpha)} < \infty\Bigg\},$$

where $p(\alpha) = \frac{2n}{n-2\alpha}$, $\frac{1}{2} < \sigma < \gamma < \frac{n}{2}$, $0 < \nu < \min(1, \frac{\gamma-\sigma}{12})$.

We consider below the case $t > 0$ since the opposite case is treated analogously.

Notation and function spaces. We let $\partial_j = \partial/\partial x_j, \partial^l = \partial_1^{l_1}\cdots\partial_n^{l_n}$, $l \in (\mathbf{N}\cup\{0\})^n$, $M = M(t) = \exp(ix^2/2t)$, $J_j = J_j(t) = (x_j + it\partial_j) = U(t)x_jU(-t)$, $J =$

$(J_1,\dots,J_n)=U(t)xU(-t)$, and $|J|^\zeta=U(t)|x|^\zeta U(-t)$, $\zeta\in[0,\infty)$. We introduce some function spaces. As usual, $L^p=\{\varphi\in\mathcal{S}';\|\varphi\|_p<\infty\}$, where $\|\varphi\|_p=(\int|\varphi(x)|^p dx)^{\frac{1}{p}}$ if $1\le p<\infty$ and $\|\varphi\|_\infty=\mathrm{ess.sup}\{|\varphi(x)|;x\in\mathbf{R}^n\}$ if $p=\infty$. For simplicity we let $\|\varphi\|=\|\varphi\|_2$. The weighted Sobolev space $H_p^{m,s}$ is defined by $H_p^{m,s}=\{\varphi\in\mathcal{S}';\|\varphi\|_{m,s,p}=\|(1+|x|^2)^{s/2}(1-\Delta)^{m/2}\varphi\|_p<\infty\}$, $m,s\in\mathbf{R}$, $1\le p\le\infty$; also for simplicity we denote $H^{m,s}=H_2^{m,s}$, $\|\cdot\|_{m,s}=\|\cdot\|_{m,s,2}$. We let $(\psi,\varphi)=\int\psi\cdot\overline{\varphi}dx$. Denote by $\dot{B}^s_{p,q}$ the homogeneous Besov space with the seminorm

$$\|\psi\|_{\dot{B}^s_{p,q}}=\Big(\int_0^\infty y^{-1-\xi q}\sup_{|z|\le y}\sum_{|k|\le[s]}\|\partial^k(\psi_{(z)}-\psi)\|_p^q dy\Big)^{\frac{1}{q}},$$

where $s=[s]+\xi$, $0<\xi<1$, $\psi_{(z)}(x)=\psi(x+z)$, and $[s]$ is the largest integer less than s. We note that the seminorm of $\dot{B}^\gamma_{2,2}$ is equivalent to that of the homogeneous Sobolev space $\dot{H}^{\gamma,0}$, where $\dot{H}^{s,m}=\{\varphi\in\mathcal{S}';\|\varphi\|_{\dot{H}^{s,m}}=\||x|^s(-\Delta)^{m/2}\varphi\|<\infty\}$ (see [1]). We let $C(I;E)$ be the space of continuous functions from an interval I to a Banach space E. Different positive constants might be denoted by the same letter C.

Note that the free Schrödinger evolution group can be represented as $U(t)=M(t)D(t)\mathcal{F}M(t)$, where $D(t)$ is the dilation operator defined by $(D(t)\psi)(x)=(it)^{-\frac{n}{2}}\psi(\frac{x}{t})$ and

$$U(-t)=M(-t)\mathcal{F}^{-1}D(t)^{-1}M(-t)=M(-t)i^n\mathcal{F}^{-1}D\left(\frac{1}{t}\right)M(-t), \tag{1.4}$$

since $D(t)^{-1}=i^nD(\frac{1}{t})$. By using the above identities we easily get

$$\begin{aligned}J_j(t)&=U(t)x_jU(-t)=M(t)D(t)\mathcal{F}M(t)x_jM(-t)\mathcal{F}^{-\infty}\rangle\backslash\mathcal{D}\left(\frac{\infty}{\sqcup}\right)\mathcal{M}(-\sqcup)\\&=M(t)D(t)i^n(i\partial_j)D\left(\frac{1}{t}\right)M(-t)=M(t)D(t)i^nD\left(\frac{1}{t}\right)(it\partial_j)M(-t)\\&=M(t)(it\partial_j)M(-t)\text{ and }|J|^\zeta(t)=M(t)(-t^2\Delta)^{\frac{\zeta}{2}}M(-t),\zeta\in[0,\infty).\end{aligned}$$

2. Preliminaries.

LEMMA 2.1. *Let q,r be any numbers satisfying $1\le q,r\le\infty$, and let j,m be any real numbers satisfying $0\le j<m$. If $u\in H_r^{m,0}(\mathbf{R}^n)\cap L^q(\mathbf{R}^n)$, then the following inequality is valid:*

$$\|(-\Delta)^{j/2}u\|_p\le C\|(-\Delta)^{m/2}u\|_r^a\|u\|_q^{1-a}, \tag{2.1}$$

where C is a constant depending only on n,m,j,q,r,a. Here $p\ge1$ is such that $\frac{1}{p}=\frac{j}{n}+a(\frac{1}{r}-\frac{m}{n})+\frac{1-a}{q}$ and the parameter a is any from the interval $\frac{j}{m}\le a\le1$, with the following exception: if the value $m-j-\frac{n}{r}$ is a nonnegative integer, then the parameter a is any from the interval $\frac{j}{m}\le a<1$.

For proof of Lemma 2.1, see, e.g., [5, 21].

LEMMA 2.2. *We let $u(t,x)$ be a smooth function. Then we have the estimate*

$$\|u(t)\|_{p(\alpha)}\le C|t|^{-\alpha}\|\mathcal{F}U(-t)u(t)\|_{p(\alpha)}+C|t|^{-\alpha-\varrho}\|U(-t)u(t)\|_{0,\gamma}\text{for}|t|\ge1,$$

where $p(\alpha)=\frac{2n}{n-2\alpha}$, $\alpha\in[\frac{1}{2},\frac{n}{2})$, $n\ge2$, $\varrho\in[0,1]$, $\gamma=\alpha+2\varrho$.

Remark. We will show (see Lemma 3.2 below) that the norm $\|\mathcal{F}U(-t)u(t)\|_{p(\alpha)}$ does not grow with time and the norm $\|U(-t)u(t)\|_{0,\gamma}$ grows a little with time; i.e., it obeys the estimate $\|U(-t)u(t)\|_{0,\gamma} \le C(1+|t|)^{\nu}$, where $0 < \nu < \min(1, \frac{\gamma-\alpha}{12})$. Therefore Lemma 2.2 gives us the estimate of the decay rate of the solution u of the Cauchy problem (1.1).

Proof. We have the identity, with $v(t) = U(-t)u(t)$ and $w(t,x) = (e^{\frac{ix^2}{2t}} - 1)v(t,x)$,

$$\begin{aligned} u(t) = U(t)v(t) &= M(t)D(t)\mathcal{F}v(t) + M(t)D(t)\mathcal{F}\left(M(t)-1\right)v(t) \\ &= \frac{e^{\frac{ix^2}{2t}}}{(2\pi it)^{\frac{n}{2}}} \int e^{-iy\frac{x}{t}} v(t,y) \left(1 + \left(e^{\frac{iy^2}{2t}} - 1\right)\right) dy \\ &= \frac{e^{\frac{ix^2}{2t}}}{(it)^{\frac{n}{2}}} \left(\hat{v}\left(t, \frac{x}{t}\right) + \hat{w}\left(t, \frac{x}{t}\right)\right). \end{aligned} \tag{2.2}$$

We get the estimate

$$\left|e^{\frac{iy^2}{2t}} - 1\right| = 2\left|\sin\frac{y^2}{4t}\right| \le \min\left(2, \frac{|y|^2}{2|t|}\right) \le 2^{1-2\varrho}\frac{|y|^{2\varrho}}{|t|^{\varrho}} \tag{2.3}$$

for any ϱ satisfying $0 \le \varrho \le 1$, and by a direct calculation we see that

$$|f\left(\frac{\cdot}{t}\right)\|_p = |t|^{n/p}\|f\|_p. \tag{2.4}$$

Applying equality (2.4) and estimate (2.3) to identity (2.2) and using Lemma 2.1 with $p = p(\alpha), a = 1, r = 2, j = 0, m = \alpha$, we get

$$\begin{aligned} \|u(t)\|_p &\le C|t|^{-\alpha}(\|\hat{v}(t)\|_p + \|\hat{w}(t)\|_p) \le C|t|^{-\alpha}(\|\hat{v}(t)\|_p + \|\hat{w}(t)\|_{\alpha,0}) \\ &\le C|t|^{-\alpha}(\|\hat{v}(t)\|_p + \|w(t)\|_{0,\alpha}) \le C|t|^{-\alpha}(\|\hat{v}(t)\|_p + |t|^{-\varrho}\|v(t)\|_{0,\gamma}). \end{aligned}$$

This implies the lemma. □

LEMMA 2.3. *We let $u(t,x)$ be a smooth function and $0 < \gamma < \frac{n}{2}$. Then the following estimates are valid*:

$$\left|Im\left(|x|^{\gamma}U(-t)f(|u|^2)u(t), |x|^{\gamma}U(-t)u(t)\right)\right| \le C\|u\|^2_{2n/(n-1)}\|U(-t)u\|^2_{\dot{H}^{0,\gamma}}$$

and

$$\left|Im\left((-\Delta)^{\gamma/2}f(|u|^2)u(t), (-\Delta)^{\gamma/2}U(-t)u(t)\right)\right| \le C\|u\|^2_{2n/(n-1)}\|u\|^2_{\dot{H}^{\gamma,0}}.$$

Proof. Let us only consider the case $0 < \gamma < 1$ since the other cases are treated analogously. By the relation $M(t)(-t^2\Delta)^{\gamma/2}M(-t) = U(t)|x|^{\gamma}U(-t)$, we have, with $g = M(-t)u$ and $f = f(|u|^2) = f(|g|^2)$,

$$\begin{aligned} &|\ \mathrm{Im}\left(|x|^{\gamma}U(-t)f(|u|^2)u, |x|^{\gamma}U(-t)u\right)\ | = \left|\mathrm{Im}\left((-t^2\Delta)^{\gamma/2}fg, (-t^2\Delta)^{\gamma/2}g\right)\right| \\ &= \left|\mathrm{Im}\left((-t^2\Delta)^{\gamma/2}fg - f(-t^2\Delta)^{\gamma/2}g, (-t^2\Delta)^{\gamma/2}g\right)\right| \\ &\le C\|g\|_{2n/(n-2\gamma)}\|(-t^2\Delta)^{\gamma/2}f\|_{n/\gamma}\|(-t^2\Delta)^{\gamma/2}g\|, \end{aligned}$$

where we have used the fractional Leibniz rule which is proved in [17]. Since $f = V * |u|^2 = C(-\Delta)^{-(n-1)/2}|g|^2$ with a certain constant C (see [21]) we obtain by

Lemma 2.1, with $j=0$, $m=n-1-\gamma$, $p=\frac{n}{\gamma}$, $r=\frac{n}{n-1}$, $a=1$,

$$\begin{aligned}&\left|\mathrm{Im}\left(|x|^\gamma U(-t)(V*|u|^2)u, |x|^\gamma U(-t)u\right)\right|\\&\le C|t|^\gamma\|g\|_{2n/(n-2\gamma)}\|(-\Delta)^{-(n-1-\gamma)/2}|g|^2\|_{n/\gamma}\||x|^\gamma U(-t)u\|\\&\le C|t|^\gamma\|g\|_{2n/(n-2\gamma)}\|u\|^2_{2n/(n-1)}\||x|^\gamma U(-t)u\|\\&\le C\|u\|^2_{2n/(n-1)}\|U(-t)u\|^2_{H^{0,\gamma}}.\end{aligned}$$

The second estimate of the lemma follows from the same argument as in the proof of the first one, so we omit it. □

3. Proofs of Theorems 1.1 and 1.2. Below we follow the notation in the statement of Theorems 1.1 and 1.2 without further comments. To clarify the idea of the proof of the theorems we only show a priori estimates of local solutions to (1.1). For that purpose we use the following local existence theorem.

THEOREM 3.1. *We assume that the initial data $u_0\in H^{\gamma,0}\cap H^{0,\gamma}$, where $\frac{1}{2}<\gamma<\frac{n}{2}$, are such that the norm $\epsilon=\|u_0\|_{\gamma,0}+\|u_0\|_{0,\gamma}$ is sufficiently small. Then there exists a finite time interval $[-T,T]$ with $T>1$ such that there exists a unique solution $u\in C([-T,T];H^{\gamma,0}\cap H^{0,\gamma})$ of the Cauchy problem* (1.1).

For the proof of Theorem 3.1, see, e.g., [3, 8, 15].

LEMMA 3.2. *Let u be the local solutions to* (1.1) *stated in Theorem* 3.1. *Then for all $t\in[-T,T]$ we have the following estimates*:

$$(1+|t|)^{-\nu}(\|u(t)\|_{\gamma,0}+\|U(-t)u(t)\|_{0,\gamma})<2\epsilon \tag{3.1}$$

and

$$\sup_{\alpha\in[\frac{1}{2},\sigma]}(1+|t|)^{\alpha}\|u(t)\|_{p(\alpha)}<\sqrt{\epsilon}, \tag{3.2}$$

where $\frac{1}{2}<\sigma<\gamma$, $p(\alpha)=\frac{2n}{n-2\alpha}$.

Proof. On the contrary, let at least one of the estimates (3.1) or (3.2) be violated in the whole time interval $[-T,T]$. Via the continuity of the norms in the left-hand sides of (3.1) and (3.2), we can find a maximal-time interval $[-T_0,T_0]$ such that

$$\begin{aligned}&(1+|t|)^{-\nu}(\|u(t)\|_{\gamma,0}+\|U(-t)u(t)\|_{0,\gamma})\le 2\epsilon \text{ and}\\&\qquad\sup_{\alpha\in[\frac{1}{2},\sigma]}(1+|t|)^{\alpha}\|u(t)\|_{p(\alpha)}\le\sqrt{\epsilon}\end{aligned} \tag{3.3}$$

for all $t\in[-T_0,T_0]$. By using the commutation relation $[L,|J|^\gamma]=0$, where $L=i\partial_t+\frac{1}{2}\Delta$, we get from the Hartree equation (1.1) $L|J|^\gamma u=|J|^\gamma f(|u|^2)u$. Multiplying both sides of this equation by $\overline{|J|^\gamma u}$ and using Lemma 2.3, we obtain

$$\||J|^\gamma u(t)\|^2\le\||x|^\gamma u_0\|^2+C\int_0^t\|u(s)\|^2_{2n/(n-1)}\||J|^\gamma u(s)\|^2ds.$$

We use the estimate $\|u(t)\|_{p(\frac{1}{2})}\le 2\sqrt{\epsilon}(1+|t|)^{-\frac{1}{2}}$, which follows from (3.3), to get

$$\||J|^\gamma u(t)\|^2\le\||x|^\gamma u_0\|^2+C\epsilon\int_0^t(1+s)^{-1}\||J|^\gamma u(s)\|^2ds,$$

whence via Gronwall's inequality we find the estimate $\||J|^\gamma u(t)\| \le \||x|^\gamma u_0\|(1+t)^\nu$, which implies

$$(1+t)^{-\nu}\||x|^\gamma U(-t)u(t)\| \le \||x|^\gamma u_0\|. \tag{3.4}$$

Here ν is given explicitly by $\nu = C\epsilon$, where C is the same as above and therefore dependent only on n and γ. We may therefore take ν as small as we like by taking ϵ sufficiently small. In particular, we may always take ν in the range as in the statement of the theorem at the cost of taking ϵ accordingly small. In the same way as in the proof of (3.4), we have $(1+t)^{-\nu}\|u(t)\|_{\gamma,0} \le \|u_0\|_{\gamma,0}$, whence the first estimate (3.1) of the lemma follows on the time interval$[-T_0, T_0]$. By Lemma 2.1 and estimate (3.3) we have

$$\sup_{\alpha\in[\frac12,\sigma]}\sup_{t\in[-1,1]}(1+|t|)^\alpha\|u(t)\|_{p(\alpha)} \le C\epsilon < \sqrt{\epsilon}. \tag{3.5}$$

Now let us consider $t \ge 1$. From Lemma 2.2 and estimate (3.3) it follows that

$$\|u(t)\|_{p(\alpha)} \le C\epsilon t^{-\alpha-\beta+\nu} + Ct^{-\alpha}\|\mathcal{F}U(-t)u(t)\|_{p(\alpha)} \tag{3.6}$$

for any $\alpha \in [\frac12, \sigma]$, where $\beta = \min(1, \frac{\gamma-\sigma}{2})$, $0 < \nu < \beta/6$. Multiplying both sides of (1.1) by $U(-t)$, we obtain $i(U(-t)u(t))_t + U(-t)f(|u|^2)u = 0$, whence in view of identity (1.4) we have

$$i\hat v_t - t^{-1}f(|\hat v|^2)\hat v = t^{-1}\left(I_1(t) + I_2(t)\right), \tag{3.7}$$

where

$$I_1(t) = \mathcal{F}\left(M(-t) - 1\right)\mathcal{F}^{-1}f\left(|\mathcal{F}M(t)v|^2\right)\mathcal{F}M(t)v,$$

$$I_2(t) = f\left(|\mathcal{F}M(t)v|^2\right)\mathcal{F}M(t)v - f(|\hat v|^2)\hat v.$$

Introducing a new dependent variable $\hat w = \hat v B(t)$, where $B(t) = \exp(i\int_1^t f(|\hat v|^2)\frac{d\tau}{\tau})$, we write (3.7) in the form $i\hat w_t = B(t)t^{-1}\left(I_1(t) + I_2(t)\right)$, whence, integrating with respect to t from 1 to t, we get

$$\hat w(t) = \hat w(1) - i\int_1^t B(\tau)\left(I_1(\tau) + I_2(\tau)\right)\frac{d\tau}{\tau}. \tag{3.8}$$

In the same way as in the proof of Lemma 2.3, we have, with $\frac12 < \theta \le \gamma$ (we now let $\gamma < 1$ since the case $1 \le \gamma < \frac{n}{2}$ is treated analogously),

$$\begin{aligned}
\|f(h_1h_2)h_3\|_{\theta,0} &\le C(\|f\|_\infty\|h_3\| + \|(-\Delta)^{\theta/2}(fh_3)\|)\\
&\le C(\|f\|_\infty\|h_3\| + \|(-\Delta)^{\theta/2}(fh_3) - f(-\Delta)^{\theta/2}h_3\| + \|f(-\Delta)^{\theta/2}h_3\|)\\
&\le C(\|f\|_\infty\|h_3\|_{\theta,0} + \|h_3\|_{p(\theta)}\|(-\Delta)^{\theta/2}f\|_{n/\theta})\\
&\le C(\|f\|_\infty\|h_3\|_{\theta,0} + \|h_3\|_{p(\theta)}\|(-\Delta)^{-(n-1-\theta)/2}h_1h_2\|_{n/\theta})\\
&\le C\|h_3\|_{\theta,0}(\|f\|_\infty + \|h_1\|_{p(\frac12)}\|h_2\|_{p(\frac12)})\\
&\le C\|h_3\|_{\theta,0}(\|h_1\|_q\|h_2\|_q + \|h_1\|_r\|h_2\|_r + \|h_1\|_{\frac12,0}\|h_2\|_{\frac12,0}) \le C\prod_{j=1}^3\|h_j\|_{\theta,0},
\end{aligned} \tag{3.9}$$

where $q < p(\frac{1}{2}) < r < p(\theta)$. We use (3.9) and (2.3) to obtain, with $h = M(t)v$ and $0 \le \theta \le \sigma$,

$$\begin{aligned}\|I_1(t)\|_{\theta,0} &\le C\|\mathcal{F}\left(M(-t)-1\right)\mathcal{F}^{-1}f(|\hat{h}|^2)\hat{h}\|_{\theta,0} \le Ct^{-\omega}\|\mathcal{F}^{-1}f(|\hat{h}|^2)\hat{h}\|_{0,\gamma}\\ &\le Ct^{-\omega}\|f(|\hat{h}|^2)\hat{h}\|_{\gamma,0} \le Ct^{-\omega}\|\hat{h}\|_{\gamma,0}^3 \le Ct^{-\omega}\|v\|_{0,\gamma}^3, \qquad (3.10)\end{aligned}$$

where $\omega = \min(1, \frac{\gamma-\theta}{2})$. We easily see that

$$f(|\hat{h}|^2)\hat{h} - f(|\hat{v}|^2)\hat{v} = f(|\hat{h}|^2)(\hat{h}-\hat{v}) + f\left((\hat{h}-\hat{v})\overline{\hat{h}}\right)\hat{v} + f\left((\overline{\hat{h}}-\overline{\hat{v}})\hat{v}\right)\hat{v},$$

whence, by virtue of (3.10) and using (2.3) and Lemma 2.1, we obtain

$$\begin{aligned}\|I_2(t)\|_{\sigma,0} &= \|f(|\hat{h}|^2)\hat{h} - f(|\hat{v}|^2)\hat{v}\|_{\sigma,0} \le C\|\hat{h}-\hat{v}\|_{\sigma,0}\left(\|\hat{h}\|_{\sigma,0}^2 + \|\hat{v}\|_{\sigma,0}^2\right)\\ &\le C\|\left(M(t)-1\right)v\|_{0,\sigma}\|v\|_{0,\sigma}^2 \le Ct^{-\beta}\|v\|_{0,\gamma}^3, \qquad (3.11)\end{aligned}$$

where $\frac{1}{2} < \sigma < \gamma < \frac{n}{2}$, $\beta = \min(1, \frac{\gamma-\sigma}{2})$. By (3.8), (3.10), (3.11), and estimate (3.3) we have

$$\begin{aligned}\|\mathcal{F}U(-t)u\|_{p(\alpha)} &= \|\hat{v}\|_{p(\alpha)} = \|\hat{w}\|_{p(\alpha)} \le C\epsilon + C\int_1^t \left(\|I_1(\tau)\|_{p(\alpha)} + \|I_2(\tau)\|_{p(\alpha)}\right)\frac{d\tau}{\tau}\\ &\le C\epsilon + C\int_1^t \left(\|I_1(\tau)\|_{\alpha,0} + \|I_2(\tau)\|_{\alpha,0}\right)\frac{d\tau}{\tau} \le C\epsilon\int_1^t \tau^{-1-\beta+3\nu}d\tau \le C\epsilon. \qquad (3.12)\end{aligned}$$

We apply (3.12) to (3.6) to get the estimate (3.2) on the interval $[-T_0, T_0]$. The contradiction obtained proves the estimates (3.1) and (3.2) on the whole-time interval $[-T, T]$ of the existence of solutions u to the Cauchy problem (1.1). Lemma 3.2 is proved. □

We are now in a position to prove Theorems 1.1 and 1.2.

Proof of Theorem 1.1. We have by Lemma 3.2

$$|||u|||_{X_T} \le 2\epsilon + \sqrt{\epsilon}.$$

Then the standard continuation argument yields the result. □

Proof of Theorem 1.2. By (3.8), (3.10), (3.11), and Theorem 1.1 we have

$$\|\hat{w}(t) - \hat{w}(s)\|_{\sigma,0} \le C\epsilon s^{-\beta+3\nu} \qquad (3.13)$$

for $t > s \ge 1$, where $\frac{1}{2} < \sigma < \gamma$, $\beta = \min(1, \frac{\gamma-\sigma}{2})$, $0 < \nu < \beta/6$. Therefore we find that there exists a unique function $\hat{w}_+ \in H^{\sigma,0}$ such that

$$\|\hat{w}(t) - \hat{w}_+\|_{\sigma,0} \le C\epsilon t^{-\beta+3\nu}. \qquad (3.14)$$

In the same way as in the proof of (3.9) and (3.11) we have

$$\|I_1(t)\|_{0,\sigma} + \|I_2(t)\|_{0,\sigma} \le Ct^{-\beta}(\|v\|_{0,\gamma} + \|v\|_{\gamma,0})^3 \le C\epsilon t^{-\beta+3\nu}. \qquad (3.15)$$

Hence from (3.8) we see that $\hat{w}_+ \in H^{0,\sigma}$ and

$$\|\hat{w}(t) - \hat{w}_+\|_{0,\sigma} \le C\epsilon t^{-\beta+3\nu}. \qquad (3.16)$$

By Lemma 2.1, with $p = 2n$, $j = 0$, $a = 1$, $m = n-1$, $r = \frac{2n}{2n-1}$, we find that

$$\begin{aligned} \|f(h_1h_2)\|_{2n} &= C\|(-\Delta)^{-(n-1)/2}(h_1h_2)\|_{2n} \\ &\le C\|h_1\|\|h_2\|_{p(\frac12)} \le C\|h_1\|\|h_2\|_{\frac12,0}. \end{aligned} \tag{3.17}$$

Let us now consider the case $\frac12 < \sigma < 1$. The other cases can be treated analogously. We use estimate (3.17) to get the following estimate with $\frac12 < \sigma < \gamma$:

$$\begin{aligned} \|f(h_1h_2)\|_{\dot B^\sigma_{2n,2}} &= \bigg(\int_0^\infty y^{-1-2\sigma} \sup_{|z|\le y} \|(f_{(z)} - f)\|_{2n}^2 dy\bigg)^{1/2} \\ &\le C\|h_2\|_{\frac12,0}\bigg(\int_0^\infty y^{-1-2\sigma} \sup_{|z|\le y} \|(h_{1(z)} - h_1)\|^2 dy\bigg)^{1/2} \\ &\quad + C\|h_1\|_{\frac12,0}\bigg(\int_0^\infty y^{-1-2\sigma} \sup_{|z|\le y} \|(h_{2(z)} - h_2)\|^2 dy\bigg)^{1/2} \\ &\le C\|h_1\|_{\sigma,0}\|h_2\|_{\sigma,0}, \end{aligned} \tag{3.18}$$

whence we obtain

$$\begin{aligned} &\|f(|\hat w(t)|^2) - f(|\hat w(s)|^2)\|_{\dot B^\sigma_{2n,2}} = \Big\| f\left((\hat w(t) - \hat w(s))\,\overline{\hat w(t)}\right) \\ &\quad + f\left((\overline{\hat w(t)} - \overline{\hat w(s)})\hat w(s)\right)\Big\|_{\dot B^\sigma_{2n,2}} \le C\|\hat w(t) - \hat w(s)\|_{\sigma,0}(\|\hat w(t)\|_{\sigma,0} + \|\hat w(s)\|_{\sigma,0}) \\ &\le Cs^{-\beta+6\nu}(\|u_0\|_{0,\gamma} + \|u_0\|_{\gamma,0}) \end{aligned} \tag{3.19}$$

for all $1 < s < t$, where $\beta = \min(1, \frac{\gamma-\sigma}{2})$, $0 < \nu < \beta/6$.

We now let $\Psi(t) = \int_1^t \left(f(|\hat w(\tau)|^2) - f(|\hat w(t)|^2)\right)\frac{d\tau}{\tau}$. Then

$$\begin{aligned} \Psi(t) - \Psi(s) &= \int_s^t \left(f(|\hat w(\tau)|^2) - f(|\hat w(t)|^2)\right)\frac{d\tau}{\tau} \\ &\quad - \left(f(|\hat w(t)|^2) - f(|\hat w(s)|^2)\right)\log s, \end{aligned} \tag{3.20}$$

where $1 < s < \tau < t$. We apply (3.19) to (3.20) to get, for all $1 < s < t$,

$$\|\Psi(t) - \Psi(s)\|_{\dot B^\sigma_{2n,2}} \le C\epsilon s^{-\beta+6\nu}. \tag{3.21}$$

This implies that there exists a unique real-valued function $\Phi \in \dot B^\sigma_{2n,2}$ such that $\lim_{t\to\infty}\Psi(t) = \Phi$ in $\dot B^\sigma_{2n,2}$. We let $t \to \infty$ in (3.21). Then

$$\|\Phi - \Psi(t)\|_{\dot B^\sigma_{2n,2}} \le C\epsilon t^{-\beta+6\nu}. \tag{3.22}$$

We note that $\hat u_+ = \hat w_+ \exp(-i\Phi)$. (For the function $\hat u_-$, analogously we have $\hat u_- = \overline{\hat w_+}\exp(i\Phi)$.) By the facts that $\Phi \in \dot B^\sigma_{2n,2}$, $\hat w_+ \in H^{\sigma,0}\cap H^{0,\sigma}$, and Lemma 2.1, we see that $\hat u_+ \in H^{\sigma,0}\cap H^{0,\sigma}$, and we also have the following by Theorem 1.1, (3.14), (3.21), and (3.22):

$$\begin{aligned} \|u_+\|_{0,\sigma} &= \|\hat u_+\|_{\sigma,0} = \|\exp(i\Phi)\hat w_+\|_{\sigma,0} \le C(\|w_+\| + \|\hat w_+\exp(i\Phi)\|_{\dot B^\sigma_{2,2}}) \\ &\le C\epsilon + C\bigg(\int_0^\infty y^{-1-2\sigma}\sup_{|z|\le y}\|(\hat w_{+(z)} - \hat w_+)\exp(i\Phi_{(z)})\|^2 dy\bigg)^{\frac12} \\ &\quad + C\|\hat w_+\|_{p(\frac12)}\bigg(\int_0^\infty y^{-1-2\sigma}\sup_{|z|\le y}\|\Phi_{(z)} - \Phi\|_{2n}^2 dy\bigg)^{\frac12} \\ &\le C\epsilon + C\|\hat w_+\|_{\sigma,0} + C\|\hat w_+\|_{\frac12,0}\|\Phi\|_{\dot B^\sigma_{2n,2}} \le C\epsilon. \end{aligned} \tag{3.23}$$

We easily find that the following identity holds:

$$\int_1^t f(|\hat{w}(\tau)|^2)\frac{d\tau}{\tau} = f(|\hat{w}_+|^2)\log t + \Phi + (\Psi(t)-\Phi) + \left(f(|\hat{w}(t)|^2) - f(|\hat{w}_+|^2)\right)\log t. \tag{3.24}$$

In the same way as in the proof of (3.11), (3.17), we have

$$\begin{aligned}\|I_2(t)\| &= \|f(|\hat{h}|^2)\hat{h} - f(|\hat{v}|^2)\hat{v}\| \le \|\hat{h}-\hat{v}\|\|f(|\hat{h}|^2)\|_\infty \\ &\quad + \|\hat{v}\|_{p(\frac{1}{2})}\left(\|f\left((\hat{h}-\hat{v})\overline{\hat{h}}\right)\|_{2n} + \|f\left((\overline{\hat{h}}-\overline{\hat{v}})\hat{v}\right)\|_{2n}\right) \\ &\le C\|\hat{h}-\hat{v}\|\left(\|\hat{h}\|_{\sigma,0}^2 + \|\hat{v}\|_{\sigma,0}^2\right) \le C\|\left(M(t)-1\right)v\|\|v\|_{0,\sigma}^2 \\ &\le Ct^{-\mu}\|v\|_{0,\gamma}^3 \le Ct^{-\mu+3\nu}\left(\|v\|_{0,\gamma}+\|v\|_{\gamma,0}\right)^3 \le C\epsilon t^{-\mu+3\nu}, \end{aligned} \tag{3.25}$$

where $\mu = \min(1, \frac{\gamma}{2})$. Hence by virtue of (3.10) with $\theta = 0$ and (3.25) we see from (3.8) that

$$\|\hat{w}(t) - \hat{w}_+\| \le C\epsilon t^{-\mu+3\nu}. \tag{3.26}$$

Using (3.17) and (3.26), analogously to (3.19) we get

$$\begin{aligned}\|f(|\hat{w}(t)|^2) - f(|\hat{w}(s)|^2)\|_{2n} &= \|f\left((\hat{w}(t)-\hat{w}(s))\overline{\hat{w}(t)}\right) + f\left((\overline{\hat{w}(t)}-\overline{\hat{w}(s)})\hat{w}(s)\right)\|_{2n} \\ &\le C\|\hat{w}(t)-\hat{w}(s)\|\left(\|\hat{w}(t)\|_{p(\frac{1}{2})} + \|\hat{w}(s)\|_{p(\frac{1}{2})}\right) \\ &\le Cs^{-\mu+6\nu}\left(\|u_0\|_{0,\gamma} + \|u_0\|_{\gamma,0}\right), \end{aligned} \tag{3.27}$$

and therefore we obtain

$$\|\Phi - \Psi(t)\|_{2n} \le C\epsilon t^{-\mu+6\nu}. \tag{3.28}$$

By (3.24), (3.27), and (3.28) we have

$$\left\|\int_1^t f(|\hat{w}(\tau)|^2)\frac{d\tau}{\tau} - f(|\hat{w}_+|^2)\log t - \Phi\right\|_{2n} \le C\epsilon t^{-\mu+7\nu}. \tag{3.29}$$

Since $\hat{w}(t) = B(t)\mathcal{F}U(-t)u(t) = \exp(i\int_1^t f(|\hat{w}(\tau)|^2)\frac{d\tau}{\tau})\mathcal{F}U(-t)u(t)$, we have, in view of (3.14), (3.26), and (3.29),

$$\begin{aligned}&\|\mathcal{F}U(-t)u(t) - \hat{u}_+\exp\left(-if(|\hat{u}_+|^2)\log t\right)\| \\ &= \left\|\hat{w}(t)\exp\left(-i\int_1^t f(|\hat{w}(\tau)|^2)\frac{d\tau}{\tau}\right) - \hat{u}_+\exp\left(-if(|\hat{u}_+|^2)\log t\right)\right\| \\ &\le \left\|\hat{w}(t)-\hat{w}_+\| + C\|\left(\int_1^t f(|\hat{w}(\tau)|^2)\frac{d\tau}{\tau} - f(|\hat{u}_+|^2)\log t - \Phi\right)\hat{w}_+\right\| \\ &\le C\epsilon t^{3\nu-\mu} + C\left\|\hat{w}_+\|_{p(\frac{1}{2})}\|\int_1^t f(|\hat{w}(\tau)|^2)\frac{d\tau}{\tau} - f(|\hat{w}_+|^2)\log t - \Phi\right\|_{2n} \\ &\le C\epsilon t^{7\nu-\mu}. \end{aligned} \tag{3.30}$$

By (2.3) and Theorem 1.1 we get

$$\begin{aligned}\|u(t) - M(t)D(t)\mathcal{F}U(-t)u(t)\| &= \|M(t)D(t)\mathcal{F}\left(M(t)-1\right)U(-t)u(t)\| \\ &= \|\left(M(t)-1\right)U(-t)u(t)\| \le Ct^{-\mu}\|U(-t)u(t)\|_{0,\gamma} \le C\epsilon t^{-\mu+\nu}. \end{aligned} \tag{3.31}$$

Via (3.23), (3.30), and (3.31) it follows that

$$\begin{aligned}
&\|u(t) - \exp\left(-if\left(|\hat{u}_+|^2\right)\left(\frac{x}{t}\right)\log t\right)U(t)u_+\| \\
&\quad = \|u(t) - M(t)D(t)\exp\left(-if(|\hat{u}_+|^2)\log t\right)\mathcal{F}M(t)u_+\| \\
&\quad \le \|u(t) - M(t)D(t)\mathcal{F}U(-t)u(t)\| \\
&\quad + \|M(t)D(t)\left(\mathcal{F}U(-t)u(t) - \hat{u}_+\exp\left(-if(|\hat{u}_+|^2)\log t\right)\right)\| \\
&\quad + \|M(t)D(t)\exp\left(-if(|\hat{u}_+|^2)\log t\right)\mathcal{F}\left(M(t)-1\right)u_+\| \\
&\quad \le C\epsilon t^{7\nu-\mu} + Ct^{7\nu-\mu}\|u_+\|_{0,\sigma} \le C\epsilon t^{7\nu-\mu}. \qquad (3.32)
\end{aligned}$$

From (3.32), Theorem 1.2 follows. □

Acknowledgments. The authors would like to thank the referee for letting us know of some recent works on the asymptotic behavior of solutions of Schrödinger–Poisson equation [3, 18, 19] and for useful comments. P. Naumkin is grateful to Instituto de Física y Matemáticas de Universidad Michoacana for kind hospitality.

REFERENCES

[1] J. Bergh and J. Löfström, *Interpolation Spaces*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.
[2] J. P. Chadam and R. T. Glassey, *Global existence of solutions to the Cauchy problem for time dependent Schrödinger Hartree equations*, J. Math. Phys., 16 (1975), pp. 1211–1230.
[3] F. Castella, *L^2-solutions to the Schrödinger-Poisson system: Existence, uniqueness, time behavior, and smoothing effects*, M^3 AS, 8 (1997), pp. 1051–1083.
[4] J. P. Dias and M. Figueira, *Conservation laws and time decay for the solutions of some nonlinear Schrödinger-Hartree equations and systems*, J. Math. Anal. Appl., 84 (1981), pp. 486–508.
[5] A. Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
[6] J. Ginibre and T. Ozawa, *Long range scattering for nonlinear Schrödinger and Hartree equations in space dimension $n \ge 2$*, Comm. Math. Phys., 151 (1993), pp. 619–645.
[7] R. T. Glassey, *Asymptotic behavior of solutions to certain nonlinear Schrödinger-Hartree equations*, Comm. Math. Phys., 53 (1977), pp. 9–18.
[8] J. Ginibre and G. Velo, *On a class of nonlinear Schrödinger equations with non-local interactions*, Math. Z., 170 (1980), pp. 109–136.
[9] N. Hayashi, *Asymptotic behavior of solutions to time-dependent Hartree equations*, Nonlinear Anal., 12 (1988), pp. 313–319.
[10] N. Hayashi and P. I. Naumkin, *Asymptotics for large time behavior of solutions to the nonlinear Schrödinger and Hartree equations*, Amer. J. Math., 120 (1998), pp. 369–389.
[11] N. Hayashi and T. Ozawa, *Time decay of solutions to the Cauchy problem for time-dependent Schrödinger-Hartree equations*, Comm. Math. Phys., 110 (1987), pp. 467–478.
[12] N. Hayashi and T. Ozawa, *Scattering theory in the weighted $L^2(\mathbf{R}^n)$ spaces for some Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 48 (1988), pp. 17–37.
[13] N. Hayashi and T. Ozawa, *Time decay for some Schrödinger equations*, Math. Z., 200 (1989), pp. 467–483.
[14] N. Hayashi and Y. Tsutsumi, *Scattering theory for Hartree type equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987), pp. 187–213.
[15] H. Hirata, *The Cauchy problem for Hartree type Schrödinger equation in weighted Sobolev space*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 38 (1991), pp. 567–588.
[16] H. Hirata, *Large time behavior of solutions for Hartree equation with long range interaction*, Tokyo J. Math., 18 (1995), pp. 167–177.
[17] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg-de Vries equation via contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.
[18] J. L. López and J. Soler, *Scaling limits in the 3-D Schrödinger–Poisson system*, Appl. Math. Lett., 5 (1997), pp. 61–65.

[19] J. L. López and J. Soler, *Asymptotic behaviour to the* 3*-D Schrödinger–Poisson and Wigner–Poisson systems*, submitted.
[20] H. Nawa and T. Ozawa, *Nonlinear scattering with nonlocal interaction*, Comm. Math. Phys., 146 (1992), pp. 259–275.
[21] E. M. Stein, *Singular Integral and Differentiability Properties of Functions*, Princeton Math. Ser. 30, Princeton University Press, Princeton, NJ, 1970.
[22] T. Wada, *Asymptotic behavior of the solutions of the Hartree type equations*, Adv. Math. Sci. Appl., 6 (1996), pp. 67–77.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1268–1278, September 1998

013

POROUS MEDIUM EQUATION WITH ABSORPTION*

CATHERINE BANDLE†, TOKUMORI NANBU‡, AND IVAR STAKGOLD§

Abstract. We study the approach to the steady state for the porous medium equation with absorption under positive, time-independent, Dirichlet boundary conditions. Special attention is given to the case where the solution of the steady-state problem vanishes in an interior region (known as a dead core.) The results are compared to those for the heat equation with absorption.

Key words. reaction-diffusion, porous medium, dead core

AMS subject classifications. 35K57, 76S05

PII. S0036141096311423

1. Introduction.

We shall study the asymptotic behavior of the solution of the initial-boundary value problem for a generalization of the porous medium equation with absorption. The porous medium occupies a bounded domain Ω and the solution obeys Dirichlet boundary conditions independent of time. Our goal is to describe how the solution $u(x,t)$ of the evolution problem tends to its steady state $\phi(x)$. In applications, u is either a concentration or a temperature required to be nonnegative; either physical interpretation will be used as is convenient.

Although we are mainly interested in the case of positive Dirichlet conditions, let us begin with some remarks for the case of zero boundary conditions. The steady state then vanishes identically and the possibility of *extinction in finite time* arises: is there a time T such that $u(x,t) \equiv 0$ for all x and $t \geq T$? It is known (see [13] and [16]; for the heat equation, see [11] and [15]) that extinction in finite time occurs when the absorption is strong or the diffusion is fast (see below for the definition of these terms).

By contrast, when the boundary values are positive, the steady state does not vanish identically and one can no longer expect $u(x,t)$ to coincide with $\phi(x)$ beyond some finite time. If, however, $\phi(x)$ vanishes in an interior region D, known as a *dead core*, then $u(x,t)$ may also vanish in a time-dependent dead core $D(t)$ whose relationship to D is one of the objects of our study. For the heat equation with absorption, we showed in [3] that a dead core for the stationary problem can occur only if the absorption is strong, and then the corresponding evolution problem for large t always has a dead core $D(t)$ which, for typical initial and boundary conditions, expands to D as $t \to \infty$; whether or not there is a steady core, Ricci [18] proved that the supnorm $\|u(x,t) - \phi(x)\|$ decays exponentially in time when the absorption is of power-law type. For the porous medium equation with absorption, we shall show that the existence of a nonempty D no longer guarantees that $D(t)$ is nonempty for large t, but if $D(t)$ is nonempty, then again $D(t) \to D$ as $t \to \infty$.

*Received by the editors October 30, 1996; accepted for publication (in revised form) October 1, 1997; published electronically June 18, 1998.
http://www.siam.org/journals/sima/29-5/31142.html
†Mathematisches Institut, Universitat Basel, Rheinsprung Basel, Rheinsprung 21, CH4-051 Basel, Switzerland (bandle@math.unibas.ch).
‡Mathematics Department, Toyama Medical and Pharmaceutical University, 2630 Sugitani, Toyama 930-01, Japan (toku@ms.toyama-mpu.ac.jp).
§Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (stakgold@math.udel.edu).

In the next section we formulate the problem more precisely and introduce the familiar model problem with power-law diffusion and absorption. We briefly discuss existence and uniqueness and define the notions of sub- and supersolution. In section 3 we state some monotonicity theorems and use the nondiffusive (lumped-parameter) problem and the steady-state problem to obtain the bounds needed in section 4, where we prove the principal theorem relating the evolutionary and stationary dead cores. In section 5 we establish decay estimates and find that there are cases when $\|u(x,t) - \phi(x)\|$ does not decay exponentially in time, in contrast with the one-dimensional treatment given by Ricci and Tarzia [19], where the assumptions made always led to exponential decay.

2. Formulation and model problem. We consider the following initial-boundary value problem for $u(x,t)$:

$$u_t - \Delta(A(u)) = -\lambda f(u) \qquad \text{in } Q = \Omega \times R^+ \tag{2.1}$$

subject to the boundary condition

$$u(x,t) = \chi(x) \geq 0 \qquad \text{on } \Gamma = \partial\Omega \times R^+ \tag{2.2}$$

and the initial condition

$$u(x,0) = u_0(x) \text{ in } \Omega, \qquad \text{with } 0 \leq u_0(x) \leq 1. \tag{2.3}$$

Here $\chi(x)$ is continuous on Γ and $u_0(x)$ can be extended to a continuous function on $\bar{\Omega}$, satisfying the compatibility condition

$$u_0(x) = \chi(x), \ x \in \partial\Omega.$$

The physical domain Ω is either an open interval in R^1 or a bounded, arcwise connected domain in $R^N (N > 1)$ whose boundary is of class C^3; $\lambda \in R^+$; Δ denotes the N-dimensional Laplace operator.

The function $A \geq 0$ characterizes the diffusion and $f \geq 0$ characterizes the absorption. Writing

$$\Delta(A(u)) = \text{ div } (A'(u)\text{grad}u),$$

we recognize the diffusion coefficient (or thermal conductivity) as $A'(u)$. We shall choose A and f to generalize the *model* equation

$$u_t - \Delta(u^m) = -\lambda u^p, \quad m > 0, \quad p > 0. \tag{2.4}$$

The following terminology for (2.4) is now fairly standard and stems from the behavior of the diffusion and absorption near $u = 0$:

$m < 1$	fast-diffusion equation,
$m = 1$	heat equation,
$m > 1$	porous-medium equation,
$p < 1$	strong absorption,
$p \geq 1$	weak absorption.

Note that if $p < 1$, the absorption is non-Lipschitz at $u = 0$; if $m < 1$, the diffusion coefficient mu^{m-1} is infinite at $u = 0$. For (2.1) we make the following definitions

consistent with the model problem:

$$\int_0^\epsilon \frac{1}{A(s)} ds < \infty \quad \Longleftrightarrow \text{fast diffusion}, \tag{2.5}$$

$$\int_0^\epsilon \frac{1}{f(s)} ds < \infty \quad \Longleftrightarrow \text{strong absorption}. \tag{2.6}$$

If the integrals are infinite, we speak, respectively, of slow diffusion and weak absorption. Thus, the heat equation would fall in the category of slow diffusion.

When $\chi(x) \equiv 0$ in (2.2), the steady state vanishes and $u(x,t)$ tends to zero as $t \to \infty$. Fast diffusion *or* strong absorption yields extinction in finite time (see [13] for a survey; the best necessary condition can be found in [16]). We can perhaps understand this phenomenon by the following qualitative argument. For small u, the absorption is still relatively large in the case of strong absorption and will tend to drive the solution more quickly to zero than in the case of weak absorption. Fast diffusion has the same effect: for small u, the diffusion is large and creates a large flux of concentration directed toward the lower concentration (= zero) on the boundary. If $\chi(x)$ is not identically zero in (2.2), the corresponding steady state $\phi(x)$ does not vanish identically, but it is still true that $u(x,t)$ tends to $\phi(x)$ as t approaches infinity. We see that fast diffusion no longer tends to lower the interior concentration. Indeed, suppose $\chi(x) \equiv 1$; then diffusion will generate a flux from the boundary to the smaller interior concentration and fast diffusion will serve to counteract absorption rather than to reinforce it, as was the case when $\chi(x) \equiv 0$. We shall therefore be more interested in slow diffusion and we shall impose the following conditions on $A(u)$ that generalize the case $m \geq 1$ for the model problem:

$$(P_A) \qquad \begin{cases} A(u) \in C^1[0,\infty) \cap C^2(0,\infty); \\ A'(u) > 0, \quad A''(u) \geq 0 \quad (u > 0); \\ A(0) = 0, \quad A'(0) \geq 0, \quad A(1) = 1. \end{cases}$$

Note that $A'(0)$ is finite so that the integral in (2.5) is indeed infinite.

For the absorption f, we impose the following conditions, which generalize the model problem:

$$(P_f) \qquad \begin{cases} f \in C[0,\infty) \cap C^2(0,\infty), \\ f(0) = 0, \quad f(1) = 1, \quad f'(s) > 0 \quad (s > 0). \end{cases}$$

The function

$$g(s) = f(A^{-1}(s)) \tag{2.7}$$

plays a role in the analysis of the steady-state problem

$$-\Delta(A(\phi)) = -\lambda f(\phi) \text{ in } \Omega, \quad \phi(x) = \chi(x) \text{ on } \partial\Omega. \tag{2.8}$$

Setting $\Phi = A(\phi)$, (2.8) reduces to

$$-\Delta\Phi = -\lambda g(\Phi) \text{ in } \Omega, \quad \Phi = A^{-1}(\chi(x)) \text{ on } \partial\Omega, \tag{2.9}$$

where g is defined in (2.7). Thus, the steady-state problem depends only on the single combined function g rather than on the individual functions f and A. In the model

problem $g(\Phi) = \Phi^{p/m}$. In any event, (2.9) is the same type of steady-state problem that arises for the heat equation with absorption. This problem was analyzed in [2], [9], [12], and [21]. In [2], we found that (2.9) can have a dead core—for sufficiently large λ—only if

$$(*) \qquad \int_0^1 \frac{ds}{\sqrt{G(s)}} < \infty, \text{ where } G(s) = \int_0^s g(\xi)d\xi.$$

For the model problem this reduces to $\frac{p}{m} < 1$, which is then also sufficient (see [2]). The sufficiency of $(*)$ for more general g satisfying

$$(P_g) \qquad \begin{cases} g \in C^2(0,1),\ g(0) = 0,\ g(1) = 1,\ g'(s) \geq 0 \text{ on } (0,1), \\ \quad \text{either } g''(s) \geq 0 \text{ on } (0,1) \text{ or } g''(s) \leq 0 \text{ on } (0,1), \end{cases}$$

will be shown in section 3. Most of our results can be proved under the less restrictive assumption of g being concave or convex only in some neighborhood of the origin, but for simplicity we have taken g to be either convex or concave in (0,1). The function g can easily be extended to $[0,2]$ without sacrificing smoothness, convexity, or concavity. Such an extension is needed in some proofs and will be used without further discussion.

Proofs of existence and uniqueness for problem (2.1)–(2.3) are based on a suitable notion of weak solution (see [5], and also [1], [6], [14]), which we include here for the sake of completeness.

Definition. Let $Q_T = \Omega \times (0,T)$ and let n denote the outward unit normal to Ω. A function $u \in C([0,T] : L^1(\Omega)) \cap L^\infty(Q_T)$ is called a weak solution of the problem (1.1)–(1.3) if it satisfies

$$(2.10) \quad \int_\Omega u(x,T)\sigma(x,T)dx - \int_{Q_T} [u\sigma_t + A(u)\triangle\sigma]dxdt + \int_0^T \int_{\partial\Omega} A(\chi)\frac{\partial\sigma}{\partial n}ds = \int_\Omega u_0\sigma(x,0)dx + \int_{Q_T} -f(u)\sigma dxdt$$

for all $\sigma \in C^2(\bar{Q}_T)$ with $\sigma = 0$ on $\partial\Omega \times (0,T)$. Equation (2.10) is obtained easily by multiplying (2.1) by σ, integrating over Q_T, and using the divergence theorem. The proof of existence and uniqueness of a weak solution to (2.1)–(2.3) now follows the lines of Bertsch [5].

A weak supersolution of (2.1)–(2.3) is defined by replacing the equal sign in (2.10) by $\geq$ and restricting σ to be nonnegative. Similarly, one can define a weak subsolution. For our purposes, it suffices to consider the usual super- and subsolutions (easily seen to be weak super- and subsolutions) defined as follows.

We say that $\bar{u} \geq 0$ is a supersolution of (2.1)–(2.3) if

$$(2.11) \qquad \bar{u}_t - \Delta A(\bar{u}) + \lambda f(\bar{u}) \geq 0\ , \quad \bar{u}(x,0) \geq u_0(x), \quad \bar{u}|_{\partial\Omega} \geq \chi(x).$$

Similarly, $u \geq 0$ is a subsolution if all the inequalities in (2.11) are reversed. If $\underline{u} \leq \bar{u}$, then the unique weak solution u of (2.1)–(2.3) satisfies (see [5])

$$\underline{u} \leq u \leq \bar{u}, \text{ almost everywhere (a.e.) in } Q_T.$$

Note that in general u is not continuous in Q_T. However, if χ is smooth, say differentiable, then a bounded weak solution is also continuous. This follows from

the arguments of [17], together with the very general result of [7]. We, henceforth, assume that u is continuous so that we shall write $\underline{u} \leq u$ or $u \leq \bar{u}$, dropping the a.e. qualification. Similar definitions apply to the steady-state problem (2.8) if g satisfies (P_g).

It follows from the maximum principle that the solutions of (2.1)–(2.3) and of (2.8) satisfy $u(x,t) \leq 1$ in $Q, \phi(x) \leq 1$ in Ω.

To show monotonicity of $u(x,t)$ in time, we need to impose a natural condition on the initial value $u_0(x)$:

$$\Delta(A(u_0)) - \lambda f(u_0) \leq a < 0, \quad x \in \Omega. \tag{2.12}$$

This condition holds automatically if $u_0(x)$ is a positive constant. If $u_0(x)$ satisfies (2.12), it is an upper solution to (2.1)–(2.3) so that $u(x,t) \leq u_0(x)$ for any t. Now let $v(x,t) = u(x,t+\tau)$; then $v(x,t)$ satisfies

$$v_t - \Delta(A(v)) = -\lambda f(v), \quad v(x,0) = u(x,\tau), \;\; v(\partial\Omega, t) = \chi(x).$$

Since $u(x,\tau) \leq u_0(x)$, v is a subsolution of (2.1)–(2.3), and hence $u(x,t+\tau) \leq u(x,t)$. Hence, $u(\cdot,t)$ is monotonically decreasing. Standard theorems can be used to show that $u(x,t)$ tends to the steady state $\phi(x)$ as $t \to \infty$.

3. Monotonicity and other comparison theorems. Consider problem (2.1)–(2.3) when only one part of the data is changed. We then have the following monotonicity properties:

(a) Let u_1 and u_2 be the solutions corresponding to λ_1, λ_2, respectively, with $\lambda_1 \leq \lambda_2$; then $u_2 \leq u_1$ in Q.

(b)If $f_1 \leq f_2$ on [0, 1], then $u_2 \leq u_1$ in Q.

(c) If the initial or the boundary value is decreased so is the solution.

(d) Let $u_0(x) \equiv 1, \chi(x) \equiv 1$ and consider two domains $\Omega_1 \subset \Omega_2$. Then $u_2 \leq u_1$ on Q_1.

These all are easy to prove using super- and subsolution techniques. Let us prove (d), for instance. Since u_2 satisfies the differential equation on Ω_2, it also satisfies it on Ω_1. Moreover, by the maximum principle, $u_2 \leq 1$ in Q_2 and hence on $\partial\Omega_1$. Clearly $u_2(x,0) = 1$ on Ω_1. Therefore $u_2(x,t)$ is a subsolution of (2.1)–(2.3) on Q_1 and $u_2 \leq u_1$ on Q_1 as required.

Remarks.

1. There is no straightforward comparison theorem with respect to $A(u)$ or $A'(u)$. (See, however, [4] for some partial results when $f = 0$.)
2. Similar results to (a)–(d) hold for the steady-state problem (see [2]).

Next we look at the "lumped-parameter" problem and the steady-state problem with a view to using them as comparison problems for (2.1)–(2.3). The lumped parameter problem has no diffusion term. It can be obtained as a special case of (2.1) with initial value $u_0(x) \equiv 1$ and boundary condition of vanishing *normal derivative*. We can then seek a solution $z(t)$ independent of x:

$$z_t = -\lambda f(z), \quad t > 0; \quad z(0) = 1. \tag{3.1}$$

As long as $z(t) > 0$, we integrate (3.1) to obtain

$$\lambda t = \int_{z(t)}^{1} \frac{ds}{f(s)}. \tag{3.2}$$

Thus, if $\int_0^1 \frac{ds}{f(s)} = \infty$ (weak absorption), (3.2) provides a solution $z(t) > 0$ for all t, with z tending to zero as $t \to \infty$. If, however,

$$\int_0^1 \frac{ds}{f(s)} = I < \infty \text{ (strong absorption)}, \tag{3.3}$$

then $z(t) > 0$ for $\lambda t < I$ and $z(t) \equiv 0$ for $\lambda t \geq I$. Strong absorption therefore leads to extinction in finite time. Note that the phenomenon is independent of λ, although the *time* of extinction does depend on λ. Similar results hold if the initial value for z is any positive number.

In the model problem, the solution of (3.1) is given explicitly by

$$z(t) = \begin{cases} [1 + \lambda\,(p-1)t]^{-\frac{1}{p-1}}, & p > 1; \\ e^{-\lambda t}, & p = 1; \\ [1 - \lambda\,(1-p)t]_+^{\frac{1}{1-p}}, & p < 1. \end{cases} \tag{3.4}$$

Therefore, extinction occurs in finite time if and only if $p < 1$. Then, $I = \frac{1}{1-p}$.

Comparison with (3.1) leads immediately to two results for (2.1)–(2.3).

THEOREM 3.1. (a) *If $\chi(x) = 0$, then $z(t)$ is a supersolution of* (2.1)–(2.3) *so that $u(x,t) \leq z(t)$ and, if the absorption is strong, there is extinction in finite time for $u(x,t)$.*

(b) *If $\min_{\overline{\Omega}} u_0(x) = \theta > 0$, then the solution $z(t,\theta)$ of* (3.1) *with initial value θ is a subsolution of* (2.1)–(2.3) *so that $u(x,t) \geq z(t,\theta)$. If the absorption is weak, then $z(t,\theta) > 0$ for all t, and hence $u(x,t) > 0$ in Q.*

We also need information on the steady-state problem.

The following theorem shows that $(*)$ is necessary and sufficient for the existence of a dead core for sufficiently large λ. We are grateful to J. Ildefonso Diaz for pointing out the sufficiency (see also [8]).

THEOREM 3.2 (see [2]). *Let g satisfy (P_g) and let $\chi(x) > 0$. If $(*)$ is not satisfied, there is no dead core for any λ; if $(*)$ is satisfied, then a dead core exists for sufficiently large λ.*

Proof. (a) Suppose $\int_0^1 \frac{ds}{\sqrt{G(s)}} = \infty$. Then $w(x)$ defined implicitly by

$$\int_{w(x)}^a \frac{ds}{\sqrt{G(s)}} = \sqrt{2\lambda}\, x \tag{3.5}$$

is positive for all $x > 0$, is a decreasing function of x, and satisfies the ordinary differential equation

$$w'' = \lambda g(w),\ x > 0;\ w(0) = a.$$

Now let Ω be a bounded domain and consider problem (2.9) with $\Phi(\partial\Omega) = A^{-1}(\chi) \geq a$. Then choosing the coordinate system so that Ω lies in the half-space $x > 0$, we see that $w(x)$ is a lower solution and, since $w > 0, \Phi > 0$ and so is ϕ. Thus, no dead core is possible for any λ.

(b) If $\int_0^1 \frac{ds}{\sqrt{G(s)}} < \infty$, then $g'(0) = \infty$ and hypothesis (P_g) implies g is concave. Hence, $g(v) \geq vg(1), 0 \leq v \leq 1$, and

$$G(v) = \int_0^v g(s)ds \leq vg(v) \leq \frac{1}{g(1)}\, g^2(v), \quad 0 \leq v \leq 1. \tag{3.6}$$

We shall construct, for λ sufficiently large, an upper solution $v(r)$ to (2.9) for a ball B_R, with $v \equiv 0, 0 < r < R/2$. We begin by observing that on the positive real line, the function $w(x)$ defined implicitly by

$$\int_0^{w(x)} \frac{ds}{\sqrt{G(s)}} = \sqrt{2\mu}\ x \tag{3.7}$$

satisfies $w'' = \mu g(w)$, $x > 0; w(0) = w'(0) = 0$. We note that w is an increasing function of x and μ. Now *choose* μ so that $w\left(\frac{R}{2}\right) = 1$ and consider the function

$$v = w\left(r - \frac{R}{2}\right)$$

in the ball $r \le R$ in R^N (having extended v to be zero for $r < R/2$). We then obtain

$$\begin{aligned} \triangle v = v'' + \frac{N-1}{r}\, v' &= \mu g(v) + \frac{N-1}{r}\sqrt{2\mu G(v)} \quad \text{in } \left(\frac{R}{2}, R\right) \\ &\le \mu g(v) + \frac{2(N-1)}{R}\sqrt{2\mu}\ \sqrt{G(v)} \\ &\le \left[\mu + \frac{2(N-1)}{R}\sqrt{\frac{2\mu}{g(1)}}\right] g(v) \doteq \nu g(v) \qquad \text{(by (3.6))}. \end{aligned}$$

Since $v\left(\frac{R}{2}\right) = 0, v'\left(\frac{R}{2}\right) = 0$, and $g(0) = 0$, the last inequality can be extended to $(0, R)$. Furthermore, $v(R) = 1$ so that $v(r)$ is an upper solution of the elliptic problem (2.9) for $\lambda \ge \nu$. Since v vanishes for $r < \frac{R}{2}$, so do Φ and ϕ. Now consider (2.9) on an arbitrary domain Ω with $\Phi(\partial\Omega) = A^{-1}(\chi) \le 1$. Then Ω contains a ball B_R on whose boundary $\Phi \le 1$. Therefore, $\Phi \le v$ and Ω contains a dead core for $\lambda \ge \nu$. □

For any $x_0 \in \Omega$, we can take $R = r_0 =$ the distance from x_0 to the boundary; part (b) of Theorem 3.2 then shows that for λ large enough, x_0 belongs to the dead core. This suggests making the following definition.

DEFINITION 3.3. *Let $x_0 \in \Omega$ and let g satisfy $(*)$ and (P_g). Define*

$$\lambda_0 = \inf_{\lambda} \{\Phi(x_0, \lambda) = 0\}\,;\ \lambda^* = \inf_{x_0} \lambda_0. \tag{3.8}$$

Remarks.

(a) Of course $\Phi(x_0, \lambda)$ and $\phi(x_0, \lambda)$ either both vanish or are both positive.

(b) Particularly simple estimates for λ_0 and $\lambda^\star$ are available in the model problem (see [13], [20]). In that case we are considering (2.9) with $g(s) = s^\alpha$, $\alpha < 1$. We claim that

$$\lambda_0 \le \frac{P_{n,\alpha}}{r_0^2}, \quad \lambda^\star \le \frac{P_{n,\alpha}}{\rho^2}, \tag{3.9}$$

where $P_{n,\alpha} = \frac{4+2(n-2)(1-\alpha)}{(1-\alpha)^2}, r_0 = \text{dist } (x_0, \partial\Omega); \rho =$ inradius of Ω. The proof consists in noting that the function $w = (\frac{|x-x_0|}{r_0})^{2/1-\alpha}$ satisfies $-\Delta w + \frac{P_{n,\alpha}}{r_0^2} w^\alpha = 0$ in Ω and $w(\partial\Omega) \ge 1$. Thus, w is a supersolution of (2.9) for any $\lambda \ge \frac{P_{n,\alpha}}{r_0^2}$. Since w vanishes at x_0, so does Φ, and hence ϕ. This proves the first part of (3.9) and the second follows at once.

4. The evolution problem whose steady state has a dead core. Suppose $(*)$ is satisfied with λ large enough for the steady state to have a dead core. Does the corresponding evolution problem have a dead core and, if so, how does it behave for large t? The answer is given by the following theorem.

THEOREM 4.1. *Let $u(x,t)$ be the solution of* (2.1)–(2.3) *under properties* $(P_A), (P_f), (*)$, *and* (P_g). *For fixed $x_0 \in \Omega$, choose $\lambda > \lambda_0$, where λ_0 is defined in* (3.8). *Then*

(a) *if f satisfies* (3.3), $u(x_0,t) = 0$ *for*

$$t \geq t_0 \doteq \frac{I}{\lambda - \lambda_0};$$

(b) *if the integral in* (3.3) *is infinite and* $\min_{\overline{\Omega}} u_0(x) > 0$, *then* $u(x_0,t) > 0$ *for all t.*

Proof. Part (b) is equivalent to Theorem 3.1(b). To prove part (a), it suffices to exhibit a supersolution $w(x,t)$ such that $w(x_0,t) = 0$ for $t \geq t_0$. Modifying the idea in [3] (see also [10]), we are led to try a function w of the form

$$w = A^{-1}(Az + A\phi) \qquad \text{(i.e., } Aw = Az + A\phi\text{)},$$

where $z(t,\gamma)$ is the solution of (3.1) with a parameter γ to be chosen and ϕ is the solution of (2.8) with $\lambda = \lambda_0$. Note that w vanishes at x_0 for $\gamma t \geq I$. Since $w \geq z$ and $w \geq \phi$, it is clear that $w(x,0) \geq u_0(x)$ and $w(\partial\Omega,t) \geq \chi(x)$; it remains only to choose γ to satisfy the differential inequality for a supersolution. From the definition of w, we have

$$(Aw)_t = A'(w)w_t = A'(z)z_t$$

so that

$$w_t = \frac{A'(z)}{A'(w)} z_t \geq z_t = -\gamma f(z),$$

where we have used the fact that $A'' \geq 0$ and $z_t \leq 0$. It follows that

$$\begin{aligned} w_t - \Delta(A(w)) + \lambda f(w) &\geq -\gamma f(z) - \Delta(A(\phi)) + \lambda f(w) \\ &= -\gamma f(z) - \lambda_0 f(\phi) + \lambda f(w) \\ &\geq (\lambda - \lambda_0 - \gamma) f(w). \end{aligned}$$

By choosing $\gamma = \lambda - \lambda_0$ we obtain the desired result. □

Remarks.

1. If f satisfies (3.3) we may be interested in the time of onset of the dead core. Then, if $\lambda > \lambda^*$ (see (3.8)), a nonempty dead core $D(t)$ will exist for $t \geq I/(\lambda - \lambda^*)$.
2. In the model problem, condition $(*)$ means $p/m < 1$, and then the steady state has a dead core for $\lambda > \lambda^*$. Suppose $\lambda > \lambda^*$. Then Theorem (4.1) states that if $p < 1$ (so that (3.3) is satisfied), the evolution problem for large times contains a dead core which ultimately covers any interior point of the steady dead core; if, however, $p \geq 1$ (so that the integral in (3.3) is infinite), then $u(x,t)$ is positive for all $x \in \Omega, t > 0$.

3. Often only an upper bound $\overline{\lambda}_0$ to λ_0 is known explicitly (see, for instance, (3.9)). Then, if $\lambda > \overline{\lambda}_0$ and f satisfies (3.3), we have $u(x_0, t) = 0$ for $t \geq I/(\lambda - \overline{\lambda}_0)$.
4. Note that our proof does not go through if $A'' < 0$, which corresponds to $m < 1$ in the model problem.

5. Decay estimates. We consider (2.1)–(2.3) with $u_0(x) \geq \phi(x)$, where $\phi(x)$ is the solution of the corresponding steady-state problem (2.8). It then follows, since ϕ is a lower solution, that $u(x,t) \geq \phi(x)$. If we assume in addition that $u_0 \in L^\infty(\Omega)$, then u is bounded a.e. in Q_∞. From a result of [17] it then follows that $A(u(\cdot,t)) \to A(\phi)$ in $L^2(\Omega)$ as $t \to \infty$. Hence, $\lim_{t\to\infty} |u(x,t) - \phi(x)|_\infty = 0$.

Our estimates hold for almost all $x \in \Omega$. If the data are smooth (cf. section 2), the solution is continuous and the estimates hold pointwise. We seek decay estimates for $\delta(x,t) \doteq u(x,t) - \phi(x)$ when conditions $(P_A), (P_f)$, and (P_g) are satisfied. For the model problem (2.4) this would mean $m \geq 1, p > 0$. Our principal results are contained in the following theorem.

THEOREM 5.1.

(a) *if g is convex, then*

$$0 \leq \delta(x,t) \leq z(t),$$

where $z(t)$ is the solution of (3.1).

(b) *if g is concave, then*

$$0 \leq \delta(x,t) \leq \zeta(t),$$

where $\zeta(t)$ is the solution of

$$\zeta_t = -\gamma A(\zeta),\ \zeta(0) = 1,\ \gamma = \lambda g'(1). \tag{5.1}$$

Proof. (a) As candidate for a supersolution to (2.1)–(2.3), choose $w(x,t)$ defined by

$$Aw = Az + A\phi \qquad (w = A^{-1}(Az + A\phi)),$$

where $z(t)$ satisfies (3.1) and ϕ satisfies (2.8). Then, as in the preceding section, we find

$$(Aw)_t = A'(w)w_t = A'(z)z_t, \quad w_t \geq z_t = -\lambda f(z).$$

Thus,

$$\begin{aligned} w_t - \triangle(A(w)) + \lambda f(w) \quad &\geq \lambda\left[f(w) - f(z) - f(\phi)\right] \\ &= \lambda\left[g(Aw) - g(Az) - g(A\phi)\right], \end{aligned}$$

and, since $Aw = Az + A\phi$ and g is convex, the right side is nonnegative. Because we also have $w \geq z$ and $w \geq \phi$, we see that w is a supersolution to (2.1)–(2.3). The convexity of A implies the concavity of A^{-1} so that

$$\delta = u - \phi \leq w - \phi = A^{-1}Aw - A^{-1}A\phi \leq A^{-1}Az = z.$$

(b) As candidate for an upper solution we now choose

$$v = A^{-1}(A\zeta + A\phi) \text{ or } Av = A\zeta + A\phi,$$

where ζ satisfies (5.1) and ϕ satisfies (2.8). By a similar calculation, as in part (a) we find $v_t \geq \zeta_t$ so that

$$v_t - \triangle(A(v)) + \lambda f(v) \geq \lambda\,[g(Av) - g(A\phi)] - \gamma A\zeta. \tag{5.2}$$

The concavity of g yields

$$g(Av) - g(A\phi) \geq g'(Av)\,[Av - A\phi] = g'(Av)\,[A\zeta]\,.$$

Since $A(1) = 1$, we have $Av \leq 2$ and $g(Av) - g(A\phi) \geq g'(2)A\zeta$. The values of g in [1,2] being at our disposal (subject to preserving smoothness and concavity), we can take $g'(2)$ as close to $g'(1)$ as we please so that the last inequality is also valid with $g'(1)$ replacing $g'(2)$. Thus, if $\gamma = \lambda g'(1)$, the right side of (5.2) is nonnegative, and, since the boundary and initial inequalities in (2.11) are clearly satisfied, $v(x,t)$ is an upper solution to (2.1)–(2.3). We therefore find, using the concavity of A^{-1},

$$\delta = u - \phi \leq v - \phi \leq A^{-1}Av - A^{-1}A\phi \leq A^{-1}A\zeta = \zeta,$$

thereby proving part (b) of our theorem. □

Remarks.

1. These results can be generalized to the case where g is only convex or concave in a neighborhood of the origin.

2. For the model problem, $p \geq m$ corresponds to convex g and $p \leq m$ corresponds to concave g. Thus, using the explicit form (3.4), our results take the form

$$0 \leq u - \phi \leq \begin{cases} [1 + \lambda(p-1)t]^{-\frac{1}{p-1}}\,, & p \geq m \geq 1 \text{ with } p > 1, \\ e^{-\lambda t}, & p \geq m = 1. \end{cases} \tag{5.3}$$

$$0 \leq u - \phi \leq \begin{cases} [1 + \gamma(m-1)t]^{-\frac{1}{m-1}}\,, & p \leq m \text{ with } m > 1, \\ e^{-\gamma t}, & p \leq m = 1. \end{cases} \tag{5.4}$$

Here $\gamma = \lambda g'(1) = \frac{\lambda p}{m}$. Note that for $p = m = 1$, the two estimates agree. For the heat operator ($m = 1$) and arbitrary p, Ricci [18] showed that $u(x,t)$ decays exponentially to $\phi(x)$, whereas our method gives this result only if $p \leq 1$. Of course if $m \neq 1$, we have new results for the porous medium equation with absorption. Even though our estimates are not optimal, they suggest the possibility of nonexponential decay to the steady state. We now exhibit such a case. Suppose $u_0(x) \equiv 1, 1 < p < m$, and $\lambda > \lambda^*$ (see (3.8)); then the stationary problem has a dead core D and, by combining (5.4) with Theorem 3.1(b), we find for $x \in D$

$$[1 + \lambda(p-1)t]^{-\frac{1}{p-1}} \leq u(x,t) \leq [1 + \gamma(m-1)t]^{-\frac{1}{m-1}}\,.$$

These two estimates are compatible and show that $u(x,t)$ does not decay exponentially to zero in the dead core.

Acknowledgments. The authors are grateful to the referees for their helpful suggestions.

REFERENCES

[1] D. G. Aronson, M. Crandall, and L. A. Peletier, *Stabilization of solutions of a degenerate diffusion problem*, Nonlinear Anal., 6 (1982), pp. 1001–1022.
[2] C. Bandle, R. P. Sperb, and I. Stakgold, *Diffusion-reaction with monotone kinetics*, Nonlinear Anal., 8 (1984), pp. 321–333.
[3] C. Bandle and I. Stakgold, *The formation of the dead core in parabolic reaction-diffusion equations*, Trans. Amer. Math. Soc., 286 (1984), pp. 275–293.
[4] P. Benilan and J. I. Diaz, *Comparison of solutions of nonlinear evolution problems with different nonlinear terms*, Israel J. Math, 42 (1982), pp. 241–257.
[5] M. Bertsch, *A class of degenerate diffusion equations with singular nonlinear term*, Nonlinear Anal., 7 (1983), pp. 117–127.
[6] X.-Y. Chen, H. Matano, and M. Mimura, *Finite-point extinction and continuity of interfaces in a nonlinear diffusion equation with strong absorption*, J. Reine Angew. Math., 459 (1995), pp. 1–36.
[7] E. DiBenedetto, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–118.
[8] J. I. Diaz, *Nonlinear Partial Differential Equations and Free Boundaries*, Pitman, London, 1985.
[9] J. I. Diaz and J. Hernandez, *On the existence of a free boundary for a class of reaction diffusion systems*, SIAM J. Math. Anal., 15 (1984), pp. 670–685.
[10] J. I. Diaz and J. Hernandez, *Qualitative properties of free boundaries for some nonlinear degenerate parabolic problems*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Pitman, London, 1987, pp. 85–93.
[11] A. Friedman and M. Herrero, *Extinction properties of semilinear heat equations with absorption*, J. Math. Anal. Appl., 124 (1987), pp. 530–546.
[12] A. Friedman and D. Phillips, *The free boundary of a semilinear elliptic equation*, Trans. Amer. Math. Soc., 282 (1984), pp. 153–182.
[13] A. S. Kalashnikov, *Some problems of the qualitative theory of nonlinear degenerate second-order parabolic equations*, Russian Math. Surveys, 42 (1987), pp. 169–222.
[14] R. Kersner, *Degenerate parabolic equations with general nonlinearities*, Nonlinear Anal., 4 (1980), pp. 1043–1062.
[15] R. Kersner, *Nonlinear heat conduction with absorption: Space localization and extinction in finite time*, SIAM J. Appl. Math., 43 (1983), pp. 1274–1285.
[16] A. V. Lair and M. E. Oxley, *Extinction in finite time of solutions to nonlinear absorption-diffusion equations*, J. Math. Anal. Appl., 182 (1994), pp. 857–866.
[17] M. Langlais and D. Phillips, *Stabilization of solutions of nonlinear and degenerate evolution equations*, Nonlinear Anal., 9 (1985), pp. 321–333.
[18] R. Ricci, *Large time behavior of the solution of the heat equation with nonlinear strong absorption*, J. Differential Equations, 79 (1989), pp. 1–13.
[19] R. Ricci and D. A. Tarzia, *Asymptotic behavior of the solutions of the dead-core problem*, Nonlinear Anal., 13 (1989), pp. 405–411.
[20] I. Stakgold, *Partial extinction in reaction-diffusion*, Confer. Sem. Mat., Univ. Bari, 224 (1987), pp. 1–21.
[21] J. L. Vasquez, *A strong maximum principle for some quasilinear elliptic equations*, Appl. Math. Optim., 12 (1984), pp. 191–202.

SIAM J. MATH. ANAL.
Vol. 29, No. 5, pp. 1279–1300, September 1998

014

SHARP ESTIMATES FOR THE EIGENVALUES OF SOME DIFFERENTIAL EQUATIONS*

SAMIR KARAA†

Abstract. We present optimal upper and lower bounds for the eigenvalues of the differential equations $y'' - q(x)y + \lambda\rho(x)y = 0$ and $(q(x)y')' + \lambda\rho(x)y = 0$ on a finite interval with Dirichlet boundary conditions when the coefficient functions $q(x)$ and $\rho(x)$ are nonnegative and are subjected to some kind of additional constraints. One of the basic ideas used in our work consists in reducing the problem of maximizing $\lambda(q,\rho)$ to an elementary problem of calculus of variations. This allows us to establish sufficient optimality conditions for our problems. We establish in the last part of this paper some comparison results for eigenvalues via symmetrization.

Key words. eigenvalue, Lagrange multiplier, rearrangement, isoperimetric inequalities

AMS subject classifications. 34, 49

PII. S0036141096307849

1. Introduction. Let $\lambda(q,\rho)$ denote the first eigenvalue of the boundary-value problem

$$y'' - q(x)y + \lambda\rho(x)y = 0, \tag{1.1}$$

$$y(0) = y(l) = 0, \tag{1.2}$$

with l being a positive real number. Let also H, A, and B be positive numbers such that $Hl > B$ and define the sets

$$U = \left\{ q \in L^1(0,l) \,/\, q(x) \geq 0,\ \int_0^l q(x)\,dx = A \right\}$$

and

$$V = \left\{ \rho \in L^\infty(0,l) \,/\, 0 \leq \rho(x) \leq H,\ \int_0^l \rho(x)\,dx = B \right\}.$$

The first aim of this paper is to study the extremal eigenvalue problem

$$\text{maximize } \lambda(q,\rho) \quad \text{subject to} \quad (q,\rho) \in U \times V. \tag{1.3}$$

If the function q is identically zero, then the eigenvalues of the new system

$$y'' + \lambda\rho(x)y = 0, \quad y(0) = y(l) = 0,$$

which will be denoted by $\lambda_n^0(\rho)$, characterize the frequencies of a string of density $\rho(x)$, fixed at its endpoints $x = 0$ and $x = l$ and having a unit tension. Krein [16] determined the densities which maximize $\lambda_n^0(\rho)\,(n = 1, 2, ...)$ among all measurable

*Received by the editors August 5, 1996; accepted for publication (in revised form) September 26, 1997; published electronically June 18, 1998.
http://www.siam.org/journals/sima/29-5/30784.html
†Laboratoire des Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France (samir@mip.ups-tlse.fr).

functions $\rho(x)$ from V. He has shown that the extremal functions are step periodic functions, having the period l/n and taking only the values h and H. The minimum (resp., maximum) of λ_n^0 occurs for the periodic step function which in each interval $[(k-1)l/n, kl/n]$, $k = 1, ..., n$, is symmetrically decreasing (resp., increasing) i.e., symmetric with respect to (w.r.t.) $(2k-1)l/(2n)$ and increasing (resp., decreasing) in $[(k-1)l/n, (2k-1)l/(2n)]$. He has also shown that unlike the minimal function, the maximal function is not unique in the case $n \geq 2$. We recall his sharp estimates

$$\frac{4n^2H}{B^2}\, X\left(\frac{B}{Hl}\right) \leq \lambda_n^0(\rho) \leq \frac{n^2\pi^2 H}{B^2}, \tag{1.4}$$

where $X(t)$ is the smallest positive root of the equation $X^{1/2}\, tg\, X^{1/2} = t(1-t)^{-1}$. However, Banks [3] has determined upper and lower bounds for the eigenvalues of the fourth-order differential equation

$$y^{(4)} + \lambda\rho(x)y = 0,$$

$$y(0) = y''(0) = y(l) = y''(l) = 0$$

under the condition that $\rho \in V$.

The problem when q is not zero and $\rho(x) = 1$ has been solved in [20] and [12] by different methods. Later, the authors of [9] studied the same problem when $q(x)$ satisfies the condition $\int_0^1 q(x)^\alpha\, dx = 1$, α being a nonzero real. In [14] and [15], sharp estimates for the first eigenvalue of problem (1.1)–(1.2) are obtained when the coefficients $q(x)$ and $\rho(x)$ are subjected to a general kind of constraints. For other extremal problems concerning eigenvalues see [4], [5], [6], [7], [8], [11], [18], [19], and the references quoted there.

To solve our problem (1.3), we are led to study the following problem:

$$\text{maximize} \quad \mu(\rho) \quad \text{subject to} \quad \rho \in V,$$

where $\mu(\rho) = \inf_y G[\rho, y]$ and

$$G[\rho, y] = \frac{\int_0^l y'^2\, dx + A \max|y|^2}{\int_0^l \rho y^2\, dx}.$$

The inf above is taken in the class of nonzero functions from $H_0^1(0, l)$. Thus, the major purpose of the first part of this work is to provide answers to the following problems.
Problem I. Find a function $\rho_0 \in V$ for which $\mu(\rho) \leq \mu(\rho_0) \quad \forall \rho \in V$.
Problem II. Find a pair of functions $(q_0, \rho_0) \in U \times V$ for which

$$\lambda(q, \rho) \leq \lambda(q_0, \rho_0) \qquad \forall (q, \rho) \in U \times V.$$

It is clear that standard compactness arguments do not enable us to establish the existence of solutions for Problem II. Before presenting our analysis, we give the following propositions

PROPOSITION 1.1. (i) *If $H = \infty$, then $\lambda(q, \rho)$ cannot be estimated from above.*
(ii) *If H is finite, then $\lambda(q, \rho) \leq (1 + Al)\pi^2 HB^{-2}$ for all $(q, \rho) \in U \times V$.*
(iii) *If $q \in U$ and if ρ is subjected to the constraints*

$$\int_0^l \rho(x)\, dx = B, \qquad \rho(x) \geq h,$$

where h is a positive number satisfying $hl < B$, then

$$\lambda(q,\rho) \le (1+Al)l^{-2}\pi^2/h^3.$$

Proof. Suppose that $H = \infty$ and put $q(x) \equiv Al^{-1}$ and $\rho(x) = \varepsilon^{-1}B.\chi_{[0,\varepsilon]}(x)$. (Here and throughout the paper $\chi_I(x)$ denotes the characteristic function of the set $I \subset R$.) Then we have $q \in U$, $\rho \in V$, and

$$\int_0^l \rho(x)y(x)^2\,dx = \varepsilon^{-1}B\int_0^\varepsilon y(x)^2\,dx \le \varepsilon B\int_0^\varepsilon y'(x)^2\,dx$$

for all $y \in H_0^1(0,l)$. It follows that $\lambda(q,\rho) \ge \varepsilon^{-1}B^{-1}$, and therefore $\lambda(q,\rho)$ can be arbitrarily big. The second part of the proposition follows from (1.4) and from the fact that for every function $q \in U$,

$$\int_0^l q(x)y(x)^2\,dx \le Al\int_0^l y'(x)^2\,dx$$

for all $y \in H_0^1(0,l)$. Finally we establish (iii) by using the last inequalities and the fact that $\int_0^l \rho(x)y(x)^2\,dx \ge h\int_0^l y(x)^2\,dx$ for all $y \in H_0^1(0,l)$ since $\rho(x) \ge h$. □

PROPOSITION 1.2. *Let (q,ρ) be in $U \times V$. Then there exist two functions $\bar q \in U$ and $\bar\rho \in V$, even w.r.t. $x = l/2$, such that $\lambda(q,\rho) \le \lambda(\bar q,\bar\rho)$.*

Proof. For given admissible functions q and ρ, let us define the functions $\bar q$ and $\bar\rho$ by

$$\bar q(x) = \frac{1}{2}\left[q(x)+q(l-x)\right] \quad \text{and} \quad \bar\rho(x) = \frac{1}{2}\left[\rho(x)+\rho(l-x)\right].$$

It is clear that $\bar q \in U$ and $\bar\rho \in V$. Let $\bar y$ denote the even first eigenfunction corresponding to $\lambda(\bar q,\bar\rho)$. Then we have

$$\int_0^l q(x)\bar y(x)^2\,dx = \int_0^l q(l-x)\bar y(x)^2\,dx = \int_0^l \bar q(x)\bar y(x)^2\,dx$$

and, similarly, $\int_0^l \rho(x)\bar y(x)^2\,dx = \int_0^l \bar\rho(x)\bar y(x)^2\,dx$. Now suppose that $\int_0^l \rho(x)\bar y(x)^2\,dx = 1$. Then we have

$$\begin{aligned}\lambda(q,\rho) &\le \int_0^l \bar y'(x)^2\,dx + \int_0^l q(x)\bar y(x)^2\,dx\\ &= \int_0^l \bar y'(x)^2\,dx + \int_0^l \bar q(x)\bar y(x)^2\,dx \le \lambda(\bar q,\bar\rho).\end{aligned}$$ □

2. Sufficient optimality conditions. The book of Hestenes [13, p. 215], provides a theorem used for solving problems of the following type.

Problem III. Minimize $\int_0^l F_0(x,\rho(x))\,dx$ subject to the constraints $h \le \rho \le H, 0 \le h < H$, and

$$\int_0^l F_i(x,\rho(x))\,dx = M_i, \quad i = 1,...,N,$$

where $F_0, F_1, ..., F_N$ are given functions, each continuous on $[0,l]\times[h,H]$, and $M_1, ..., M_N$ are given constants. There is a dual theorem used for maximizing the functional $\int_0^l F_0(x,\rho(x))\,dx$, which we do not state here.

THEOREM 2.1. *If $\rho_0(x)$ is a solution of Problem* III, *then there exist constants (Lagrange multipliers) $\nu_0 \geq 0, \nu_1, \nu_2, ..., \nu_N$, not all zero such that for every $x \in [0, l]$*

$$\min_{h \leq \rho \leq H} [\nu_0 F_0(x,\rho) + \nu_1 F_1(x,\rho) + \cdots + \nu_N F_N(x,\rho)] \tag{2.1}$$

$$= \nu_0 F_0(x,\rho_0(x)) + \nu_1 F_1(x,\rho_0(x)) + \cdots + \nu_N F_N(x,\rho_0(x)).$$

Conversely, if a function $\rho_0(x)$ and constants $\nu_0 > 0, \nu_1, ..., \nu_N$ exist which satisfy (2.1) *and if the conditions*

$$\int_0^l F_i(x,\rho_0(x))\,dx = M_i$$

hold for $i = 1, ..., N$, then $\rho_0(x)$ solves Problem III.

Owing to the sufficiency part of the theorem, once we have determined a function $\rho_0(x)$ satisfying the conditions of the theorem, we can affirm that this function is optimal. Our Problems I and II are clearly not of the form III. To see how Theorem 2.1 applies to them we need Propositions 2.2 and 2.3 below. These propositions give sufficient conditions for a function $\rho_0(x)$ (a couple of functions $(\tilde{q}(x), \tilde{\rho}(x))$) to be a solution of Problem I (II). Having these propositions and Theorem 2.1 at hand we can solve Problem I (II) by constructing a function (a couple of functions) satisfying the conditions of the propositions.

PROPOSITION 2.2. *Let $\rho_0(x)$ be a function of V and $y_0(x)$ be a minimizer of functional $G[\rho_0, y]$ over $H_0^1(0,l)$. If*

$$\int_0^l \rho_0(x) y_0(x)^2\,dx \leq \int_0^l \rho(x) y_0(x)^2\,dx$$

for every function $\rho(x) \in V$, then $\rho_0(x)$ is a solution of Problem I.

Proof. Let $\rho(x)$ be an arbitrary member in V. Then we have

$$\begin{aligned}
\mu(\rho) &= \inf_{y \in H_0^1} \left[\int_0^l y'^2\,dx + A \max |y|^2\right] \Big/ \int_0^l \rho y^2\,dx \\
&\leq \left[\int_0^l y_0'^2\,dx + A \max |y_0|^2\right] \Big/ \int_0^l \rho y_0^2\,dx \\
&\leq \left[\int_0^l y_0'^2\,dx + A \max |y_0|^2\right] \Big/ \int_0^l \rho_0 y_0^2\,dx \\
&= \mu(\rho_0).
\end{aligned}$$

Therefore ρ_0 is a solution of Problem I. □

PROPOSITION 2.3. *Let $(\tilde{q}, \tilde{\rho})$ be in $U \times V$ and $\tilde{y}$ be any first eigenfunction of the problem $y'' - \tilde{q}(x)y + \lambda \tilde{\rho}(x) y = 0$, $y(0) = y(l) = 0$. If*

$$\int_0^l \tilde{\rho}(x) \tilde{y}(x)^2\,dx \leq \int_0^l \rho(x) \tilde{y}(x)^2\,dx, \tag{2.2}$$

$$\int_0^l q(x) \tilde{y}(x)^2\,dx \leq \int_0^l \tilde{q}(x) \tilde{y}(x)^2\,dx \tag{2.3}$$

for every couple $(q, \rho) \in U \times V$, then $(\tilde{q}, \tilde{\rho})$ solves Problem II.

Another way of stating Proposition 2.2 is to say that a sufficient condition for a function $\rho_0(x) \in V$ to be a solution of Problem I is that there exist constants $\nu_0 > 0$ and ν_1 such that for every $x \in [0, l]$

$$\min_{0 \le \rho \le H} [\nu_0 y_0^2(x)\rho + \nu_1 \rho] = \nu_0 y_0^2(x)\rho_0(x) + \nu_1 \rho_0(x),$$

where y_0 is a minimizer of functional $G[\rho_0, y]$ over $H_0^1(0, l)$. This is in fact the reason for choosing the trial function given by (3.1) below. Of course, this remark may be applied to Proposition 2.3 and the couple $(\tilde{q}, \tilde{\rho})$.

A similar approach was used by Barnes [6] to solve other kinds of extremal eigenvalue problems. For example, he studied the problem of determining the shape of the strongest column in the class of all columns of length l, volume V and having similar cross-sectional areas $A(x)$ satisfying $a \le A(x) \le b$, where a and b are prescribed positive bounds.

Remark 2.4. Conditions (2.2) and (2.3) are necessary for any extremal couple. In fact, suppose that $(\tilde{q}, \tilde{\rho})$ is a solution of Problem II. Then $\tilde{q}$ maximizes the first eigenvalue of the problem $y'' - q(x)y + \lambda\tilde{\rho}(x)y = 0$, $y(0) = y(1) = 0$ in the set U. Since U is convex, a standard argument of calculus of variations yields

$$\int_0^l [q(x) - \tilde{q}(x)]\, \tilde{y}(x)^2\, dx \le 0$$

for all functions q in U. Similarly, since $\tilde{\rho}$ maximizes the first eigenvalue of the problem $y'' - \tilde{q}(x)y + \lambda\rho(x)y = 0$, $y(0) = y(1) = 0$ in the convex set V, we deduce as above that

$$\int_0^l [\rho(x) - \tilde{\rho}(x)]\, \tilde{y}(x)^2\, dx \ge 0$$

for all functions ρ in V.

3. Optimal solutions. As shown above, Problem I (II) will be solved by finding a function ρ_0 (a couple of functions $(\tilde{q}, \tilde{\rho})$) and y_0 $[\tilde{y}]$ satisfying the conditions of Proposition 2.2 (2.3). Let ρ_0 be the function defined by

$$\rho_0(x) = \begin{cases} H & \text{if } 0 \le x \le a_0, \\ 0 & \text{if } a_0 < x < l - a_0, \\ H & \text{if } l - a_0 \le x \le l. \end{cases} \tag{3.1}$$

The number a_0 is chosen so that $\rho_0(x) \in V$. This means that

$$a_0 = (2H)^{-1}B. \tag{3.2}$$

LEMMA 3.1. *Let* $m_0 = \inf_{y \in H_0^1} G[y]$, *where*

$$G[y] = \frac{\int_0^l y'^2\, dx + A \max |y|^2}{\int_0^l \rho_0 y^2\, dx}. \tag{3.3}$$

Then m_0 *is attained on a nonnegative function* $y_0 \in H_0^1(0, l)$ *and*

$$m_0 = \frac{\pi^2 H}{B^2} \left[\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{AB}{H\pi^2}} \right]^2. \tag{3.4}$$

Proof. A similar result when $\rho_0 \equiv 1$ has been obtained by Talenti. Our proof will use some arguments of his paper [20].

Step 1. First, it is easily seen that G has a nonnegative minimizer y_0 in $H_0^1(0,l)$, which is symmetric with respect to $l/2$. Suppose now that y_0 is convex in an interval (x_1,x_2), where $0 \le x_1 < x_2 \le l$. If z is given by

$$z(x) = y_0(x_1) + \delta(x - x_1), \quad \text{where} \quad \delta = [y_0(x_2) - y_0(x_1)]/(x_2 - x_1)$$

for $x \in (x_1,x_2)$ and coincides with y_0 out of (x_1,x_2), then the value of $G[y_0]$ is greater than $G[z]$. Indeed we have $y_0(x_1) = z(x_1)$, $y_0(x_2) = z(x_2)$, and

$$\int_{x_1}^{x_2} \rho_0(x) y_0^2(x)\,dx \le \int_{x_1}^{x_2} \rho_0(x) z^2(x)\,dx.$$

On the other hand, by Hölder's inequality

$$\int_{x_1}^{x_2} y_0'\,dx \le \left(\int_{x_1}^{x_2} {y_0}'^2\,dx\right)^{1/2} \left(\int_{x_1}^{x_2} dx\right)^{1/2};$$

$$[y_0(x_2) - y_0(x_1)]^2 \le \left(\int_{x_1}^{x_2} {y_0}'^2\,dx\right)(x_2 - x_1);$$

$$\int_{x_1}^{x_2} z'^2\,dx \le \int_{x_1}^{x_2} {y_0}'^2\,dx.$$

Since we can do the same for all other intervals where y_0 is convex, one deduces that the function y_0 is concave in the interval $[0,l]$. Now put

$$z_0(x) = \begin{cases} y_0(a_0) & \text{if } x \in [a_0, l - a_0], \\ y_0(x) & \text{otherwise.} \end{cases}$$

Therefore, $z_0 \in H_0^1(0,l)$, $\max|z_0| \le \max|y_0|$, $\int_0^l z_0'^2\,dx \le \int_0^l y_0'^2\,dx$, and $\int_0^l \rho_0 z_0^2\,dx = \int_0^l \rho_0 y_0^2\,dx$, which implies that $G[z_0] \le G[y_0]$. Hence y_0 must be constant in the interval $[a_0, l - a_0]$.

Step 2. Now we shall show that y_0' is everywhere continuous in $]0,l[$.

Let E denote the set of all $x \in [0,l]$ such that $y_0(x) = \max|y_0|$. Since y_0 is concave and symmetric about $x = l/2$, E is exactly an interval of the form $[b_0, l - b_0]$, where $0 < b_0 \le a_0$. Besides, the concavity of y_0 implies also that $y_0'(\xi^-)$ and $y_0'(\xi^+)$ exist and are finite at every point ξ from $]0,l[$ and

$$y_0'(\xi^+) \le y_0'(\xi^-). \tag{3.5}$$

Let z be a function of $H_0^1(0,l)$, achieving its maximum at $x = l/2$. Then we have

$$\max_x |y_0(x) + tz(x)| = y_0(l/2) + tz(l/2)$$

for all $t > 0$ small enough. Since y_0 minimizes G over $H_0^1(0,l)$, the directional derivative $G'[y_0,z]$ must be nonnegative. $G'[y_0,z]$ is by definition the limit of

$$\frac{1}{t}(G[y_0 + tz] - G[y_0])$$

as t approaches zero through positive values. Now pick $\xi \in]0, b_0]$ and choose

$$z(x) = \begin{cases} 1 - n|x - \xi| & \text{for } x \in I_n, \\ 0 & \text{otherwise,} \end{cases}$$

where $n > \xi^{-1}$ and $I_n = \{x \in [0, l],\, |x - \xi| < 1/n\}$. As mentioned above, we have $G'[y_0, z] \geq 0$ for all $n > \xi^{-1}$, which yields

$$-n \int_{\xi - 1/n}^{\xi} y_0' \, dx + n \int_{\xi}^{\xi + 1/n} y_0' \, dx \geq -HG[y_0] \int_{I_n} y_0 \, dx$$

for all $n > \xi^{-1}$. Letting $n \to \infty$ gives $y_0'(\xi^+) \geq y_0'(\xi^-)$. From (3.5) we deduce that $y_0'(\xi^+) = y_0'(\xi^-)$ for all $\xi \in]0, b_0]$ and therefore y_0' is everywhere continuous in $]0, l[$. Thus we have in particular $y_0'(b_0) = y_0'(l - b_0) = 0$.

Step 3. Let

$$\mathcal{O} = \{x \in]0, l[,\, y_0(x) < \max |y_0|\}.$$

Then as is shown in [20], G is Gâteaux differentiable at y_0 and

$$G'[y_0][z] = 2 \left(\int_0^l \rho_0 y_0^2 \, dx \right)^{-1} \left(\int_0^l y_0' z' \, dx - G[y_0] \int_0^l \rho_0 y_0 z \, dx \right)$$

for all test functions $z \in H_0^1(0, l)$ such that

$$\textit{support of } z \ \subset \ \mathcal{O}.$$

Hence it follows that y_0 satisfies the differential equation

$$y_0''(x) + HG[y_0] y_0(x) = 0 \ \ \textit{in } \ \mathcal{O}.$$

Integrating this equation, we arrive at the following representation:

$$y_0(x) = \begin{cases} C \sin \sqrt{m_0 H} x & \text{for } x \in [0, b_0], \\ y_0(b_0) & \text{for } x \in]b_0, l - b_0[, \\ C \sin \sqrt{m_0 H} (l - x) & \text{for } x \in [l - b_0, l], \end{cases}$$

where $m_0 = G[y_0]$. Since y_0 is concave and $y_0'(b_0) = 0$, we must have $\sqrt{m_0 H} b_0 = \pi/2$ and hence

$$b_0 = \frac{\pi}{2} (m_0 H)^{-1/2}. \tag{3.6}$$

Now the constant C is known. In fact, $C = y_0(b_0)$. On the other hand, plugging y_0 into (3.3) gives

$$m_0 = \frac{m_0 b_0 H + A}{(2a_0 - b_0) H}, \tag{3.7}$$

which yields

$$B m_0 - \pi H^{1/2} m_0^{1/2} - A = 0.$$

The unique solution of this equation is given by (3.4). □

THEOREM 3.2. *The function ρ_0 defined by* (3.1) *is a solution of Problem* I.

Proof. Let ν_0 and ν_1 be two numbers such that $\nu_0 = 1$ and $\nu_1 = -y_0^2(b_0)$, where the function y_0 and the number b_0 are as above. Then ρ_0 satisfies condition (2.1) of Theorem 2.1. Indeed, for all $x \in [0, l]$,

$$\min_{0\le\rho\le H}[y_0^2(x)\rho - y_0^2(b_0)\rho] = \min_{0\le\rho\le H}[y_0^2(x) - y_0^2(b_0)]\rho_0(x),$$

which implies that $\int_0^l \rho_0(x)y_0(x)^2\,dx \le \int_0^l \rho(x)y_0(x)^2\,dx$ for all $\rho \in V$. By Proposition 2.2, ρ_0 is a solution of Problem I. □

THEOREM 3.3. *Put $q_0(x) = \frac{1}{2}A(a_0 - b_0)^{-1}\chi_I(x)$, where a_0 and b_0 are given by* (3.2) *and* (3.6)*, respectively, and $I = [b_0, a_0] \cup [l - a_0, l - b_0]$. Then the pair (q_0, ρ_0) is a solution of Problem* II.

Proof. It is clear that $q_0 \in U$. Moreover we have for all $q \in U$

$$\int_0^l q(x)y_0^2(x)\,dx \le A\max_x |y_0^2(x)| = \int_0^l q_0(x)y_0^2(x)\,dx,$$

since $y_0(x) = \max|y_0|$ for $x \in [b_0, l - b_0]$. In order to apply Proposition 2.3, it remains to show that y_0 satisfies

$$y'' - \frac{1}{2}A(a_0 - b_0)^{-1}\chi_I\, y + m_0 H\chi_J\, y = 0 \tag{3.8}$$

everywhere in $]0, l[$, where $J = [0, a_0] \cup [l - a_0, l]$. From (3.7), we get

$$2m_0 H = A(a_0 - b_0)^{-1}. \tag{3.9}$$

On the other hand, since the differential equation

$$y_0''(x) + m_0 H y_0(x) = 0$$

holds everywhere in $\mathcal{O} = \,]0, b_0[\, \cup \,]l - b_0, l[$, relation (3.9) implies that y_0 is a solution of (3.8). Hence (q_0, ρ_0) satisfies the conditions of Proposition 2.3 and therefore is a solution of Problem II. □

Remark 3.4. For a given function $f(x)$ defined on the interval $[0, l]$, we shall use the notation $f^+(x)$ and $f^-(x)$ to denote the symmetrically increasing rearrangement and the symmetrically decreasing rearrangement of f, respectively. We recall that the function f^+ is characterized by the condition that f^+ is symmetric about $x = l/2$, decreasing on the interval $[0, l/2]$ and equimeasurable to $f(x)$ on $[0, l]$; that is,

$$\text{meas}\{x/\, f^+(x) \ge t\} = \text{meas}\{x/\, f(x) \ge t\}$$

for all $t \ge 0$. The function f^- is symmetric about $x = l/2$ and satisfies $f^-(x) = f^+(l/2 + x)$ for all x in $[0, l/2]$. Theorem 3.3 shows that an estimate from above of $\lambda(q, \rho)$ of either the form

$$\lambda(q, \rho) \le \lambda(q^-, \rho^+)$$

or the form

$$\lambda(q, \rho) \le \lambda(q^+, \rho^+)$$

is in general impossible. In fact, suppose for instance that $\lambda(q_0, \rho_0) \le \lambda(q_0^-, \rho_0)$; i.e., the couple (q_0^-, ρ_0) is also a solution of Problem II. Then Remark 2.4 tells us

that $\int_0^l q_0^-(x)\bar{y}(x)^2\,dx \geq \int_0^l q(x)\bar{y}(x)^2\,dx$ for all functions q in U, where $\bar{y}$ is the first eigenfunction of the problem $y'' - q_0^-(x)y + \lambda\rho_0(x)y = 0$, $y(0) = y(l) = 0$. Since q_0^- is symmetrically decreasing and takes only the values 0 and $m_0.H$, the function $\bar{y}$ is concave in a neighborhood of $l/2$. Consequently, one can construct a function q belonging to U, taking only the values 0 and $m_0.H$ and such that $\int_0^l q(x)\bar{y}(x)^2\,dx > \int_0^l q_0^-(x)\bar{y}(x)^2\,dx$. This, however, implies that $\lambda(q_0^-, \rho_0) < \lambda(q_0, \rho_0)$. Besides, one can easily check that $\lambda(q_0^+, \rho_0) < \lambda(q_0, \rho_0)$.

Remark 3.5. We have

$$\inf_{U\times V} \lambda(q,\rho) = 4HB^{-2}\,X\left(\frac{B}{Hl}\right), \tag{3.10}$$

where $X(t)$ is the smallest positive root of the equation $X^{1/2}\,tg\,X^{1/2} = t(1-t)^{-1}$. Moreover the inf in (3.10) is not attained. In fact, from [16] we deduce that for every $(q,\rho) \in U \times V$

$$4HB^{-2}\,X\left(\frac{B}{Hl}\right) = \mu(\bar{\rho}) \leq \mu(\rho) \leq \lambda(q,\rho),$$

where $\bar{\rho} = H.\chi_{[a,l-a]}$, $a = (H-B)/(2H)$, and $\mu(\rho)$ is the first eigenvalue of

$$\begin{cases} y'' + \mu\rho(x)y = 0, \\ y(0) = y(1) = 0. \end{cases}$$

Let y_0 be a first eigenfunction of this problem when $\rho = \bar{\rho}$. Let $\{y_n\}$ be a sequence of $C_0^\infty[0,l]$ converging to y_0. Then it is possible to find a sequence of nonnegative functions $\{q_n\}$ such that

$$\int_0^l q_n\,dx = 1, \quad \int_0^1 q_n\,y_n^2\,dx = 0.$$

Hence equality (3.10) follows since $\lambda(q_n,\bar{\rho}) \to \mu(\bar{\rho})$ as $n \to \infty$. On the other hand, if there exists a couple $(\tilde{q},\tilde{\rho})$ in $U \times V$ such that $\mu(\bar{\rho}) = \lambda(\tilde{q},\tilde{\rho})$, then we must have $\tilde{q} \equiv 0$. This function is of course not admissible. Therefore the infimum in (3.10) is not achieved in $U \times V$.

4. Generalization of the preceding results. Results analogous to the preceding ones may be obtained in the more general situation in which the function ρ is subjected to the constraints

$$\int_0^l \rho(x)\,dx = B, \qquad 0 < h \leq \rho(x) \leq H, \tag{4.1}$$

where $hl < B < Hl$. Suppose now that H is a given positive constant and consider the function ρ_h defined by

$$\rho_h(x) = \begin{cases} H & \text{if } 0 \leq x \leq a, \\ h & \text{if } a < x < l-a, \\ H & \text{if } l-a \leq x \leq l, \end{cases}$$

where $a = \frac{1}{2}(B-lh)/(H-h)$. Then we have the following lemma.

LEMMA 4.1. *Let* $\delta = (B\sqrt{m_0 H} - \pi H)/(hl\sqrt{m_0 H} - \pi)$, $0 \le \delta < H$, *and* $m_h = \inf_{y \in H_0^1} G_h[y]$, *where*

$$G_h[y] = \frac{\int_0^l y'^2\,dx + A \max |y|^2}{\int_0^l \rho_h y^2\,dx}.$$

Then m_h *is attained on a nonnegative function* $y_h \in C_0^1[0,l]$ *and there exists a positive constant* $b < l/2$ *depending on* h *such that* $y_h(x) = \max |y_h|$ *for* $x \in [b, l-b]$. *Moreover if* $h \le \delta$, *then* $m_h = m_0$, *and if* $h > \delta$, *then* m_h *is strictly less than* m_0 *and equals the least positive root of the equation*

$$\frac{\cot(\sqrt{mh}a) + \sqrt{h/H}\tan(\sqrt{mH}a)}{\sqrt{h/H}\cot(\sqrt{mh}a)\tan(\sqrt{mH}a) - 1} = \tan\frac{1}{2}\left(\sqrt{mh} - \frac{A}{\sqrt{mh}}\right).$$

The proof of this lemma is similar to that of Lemma 3.1. First of all it is easily verified that for each positive $h < H$, y_h is symmetrically decreasing and concave and belongs to $C_0^1[0,l]$. On the other hand, the set $E_h = \{x \in [0,l],\ y_h(x) = \max |y_h|\}$ must have a nonzero measure; otherwise y_h will satisfy the Euler–Lagrange equation $y_h'' + m_h \rho_h y_h = 0$ on the whole interval $(0,l)$ and minimize the functional $y \longrightarrow \int_0^l y'^2\,dx / \int_0^l \rho_h y^2\,dx$ over $H_0^1(0,l)$, which is impossible. The remainder of the lemma is proved by exploiting the differential equation above, which holds outside E_h, and by taking into account the relation $m_h = G_h[y_h]$.

THEOREM 4.2. *For all* $q \in U$ *and for all* ρ *satisfying* (4.1),

$$\lambda(q,\rho) \le m_h. \tag{4.2}$$

If $h < \delta$, *then the equality is attained if* $\rho = \rho_h$ *and if* q *is defined by*

$$q(x) = \begin{cases} m_0 H & \text{if } b_0 \le x \le a \text{ or} \\ & l - a \le x \le l - b_0, \\ m_0 h & \text{if } a < x < l - a, \\ 0 & \text{otherwise.} \end{cases}$$

If $h = \delta$, *then* (4.2) *becomes equality if* $\rho = \rho_h$ *and if* $q(x) = m_0 h \chi_{[b_0, l-b_0]}$. *If* $h > \delta$, *then* (4.2) *becomes equality if* $\rho = \rho_h$ *and if* $q(x) = m_h h \chi_{[b, l-b]}$. *Here* δ *is as in Lemma* 4.1, b_0 *is given by* (3.6),

$$a = \frac{B - lh}{2(H-h)}, \qquad b = \frac{1}{2}\left(l - \frac{A}{h m_h}\right).$$

To prove this theorem it is sufficient to verify that for each h such that $0 < h < H$, the pair (q,ρ) constructed above and the function y_h found in Lemma 4.1 satisfy the conditions of Proposition 2.3. Notice that Theorem 4.2 implies that for $h < \delta$ (in particular for $h = 0$) Problem II possesses infinitely many solutions.

It should also be mentioned that our method applies to the case when the function ρ satisfies the conditions

$$\left(\int_0^l \rho(x)^s\,dx\right)^{1/s} = B, \qquad 0 \le \rho(x) \le H, \tag{4.3}$$

where s is a number > 1. In this case the upper bound for the first eigenvalue $\lambda(q,\rho)$ is

$$\Lambda_s = \frac{\pi^2 H^{2s-1}}{B^{2s}} \left[\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{AB^s}{H^s\pi^2}}\right]^2,$$

which is achieved when

$$\rho(x) = \begin{cases} H & \text{if } 0 \le x \le a_s, \\ 0 & \text{if } a_s < x < l - a_s, \\ H & \text{if } l - a_s \le x \le l, \end{cases}$$

$a_s = B^s(2H^s)^{-1}$, and when $q(x) = \Lambda_s H \chi_{I_s}$, where $I_s = [b_s, a_s] \cup [l - a_s, b - b_s]$ and $b_s = (\Lambda_s H)^{-1/2}\pi/2$.

5. Another eigenvalue problem. Now we shall be concerned with an eigenvalue problem of the form

$$(q(x)y')' + \lambda\rho(x)y = 0, \tag{5.1}$$

$$y(0) = y(l) = 0. \tag{5.2}$$

As before we would like to find sharp upper bound for the first eigenvalue $\lambda(q,\rho)$ of problem (5.1)–(5.2) under the conditions that $\rho \in V$ and $q \in U_\alpha$, where

$$U_\alpha = \left\{ q \in L^\alpha(0,l) \,/\, q \ge 0, \int_0^l q^\alpha(x)\,dx = A^\alpha \right\}$$

and $\alpha \ge 1$. Estimates of $\lambda(q,\rho)$ are obtained in [10] when $q(x)$ and $\rho(x)$ are nonnegative and are subjected to the constraints

$$\int_0^l q^\alpha(x)\,dx = 1, \qquad \int_0^l \rho^\beta(x)\,dx = 1,$$

where α and β are nonzero real numbers. It has been shown that $\lambda(q,\rho)$ cannot be estimated from above in the case $\alpha \ge 1$ and $\beta \ge 1$. Another problem has been considered by Bandle [2], who has given upper bounds for the eigenvalues of the problem

$$(q(x)y')' + \lambda y = 0, \qquad (qy')(0) = (qy')(l) = 0,$$

when $q \in U_\alpha$ and $0 < q(x) < \sigma$ for $x \in (0,l)$. σ is a given positive constant. In [11], Egorov and Kondratiev have studied the problem of determining the shape of the column clamped at both of its extremities and having the largest buckling load among all columns of length l and volume V. This problem is equivalent to that of finding a nonnegative function (*cross-sectional area*) $q(x)$ which maximizes the first eigenvalue of the following problem:

$$(q(x)y'')'' + \lambda y'' = 0,$$

$$y(0) = y'(0) = 0, \quad y(l) = y'(l) = 0,$$

under the condition that

$$\int_0^l q^\beta(x)\,dx = V,$$

where $\beta > -1/2$ and $V > 0$. Now let us consider the following problem.

Problem IV. Find a pair of functions $(\tilde{q}, \tilde{\rho}) \in U_\alpha \times V$ for which

$$\lambda(q,\rho) \le \lambda(\tilde{q},\tilde{\rho}) \qquad \forall (q,\rho) \in U_\alpha \times V.$$

The variational principle holds, which says

$$\lambda(q,\rho) = \inf_y R[q,\rho,y],$$

where

$$R[q,\rho,y] = \frac{\int_0^l qy'^2\,dx}{\int_0^l \rho y^2\,dx},$$

and the inf is taken over all nonzero functions y from $C_0^1[0,l]$. Put now

$$M_\alpha = \sup_{q,\rho} \lambda(q,\rho),$$

where the supremum is taken in the class of all pairs $(q,\rho) \in U_\alpha \times V$.

PROPOSITION 5.1. *For any couple (q,ρ) in $U_\alpha \times V$, there exist two functions $\hat{q} \in U$ and $\hat{\rho} \in V$, both symmetric w.r.t. $l/2$ such that $\lambda(q,\rho) \le \lambda(\hat{q},\hat{\rho})$.*

Proof. For given admissible functions q and ρ, consider the functions

$$\bar{q}(x) = \frac{1}{2}\left[q(x) + q(l-x)\right] \quad \text{and} \quad \bar{\rho}(x) = \frac{1}{2}\left[\rho(x) + \rho(l-x)\right].$$

Arguing as in the proof of Proposition 1.2 gives $\lambda(q,\rho) \le \lambda(\bar{q},\bar{\rho})$. Besides,

$$\begin{aligned}\int_0^l \bar{q}^\alpha(x)\,dx &= \int_0^l \left[\frac{q(x)+q(l-x)}{2}\right]^\alpha dx \\ &\le \frac{1}{2}\left[\int_0^l q^\alpha(x)\,dx + \int_0^l q^\alpha(l-x)\,dx\right] = A^\alpha,\end{aligned}$$

since $\alpha \ge 1$. Therefore one can choose $\hat{q}(x) = A\left(\int_0^l \bar{q}^\alpha(x)\,dx\right)^{-1/\alpha} \bar{q}(x)$ and $\hat{\rho}(x) = \bar{\rho}(x)$. □

The main result of this section is the following theorem.

THEOREM 5.2. *Let $\alpha > 1$ and $p = 2\alpha/(\alpha-1)$. Let $\tilde{y}$ be the first eigenfunction of the nonlinear problem*

$$(A||y'||_p^{2-p}|y'|^{p-2}y')' + \lambda\tilde{\rho}y = 0, \quad y(0) = y(l) = 0,$$

where $\tilde{\rho}\,(= \rho_0)$ is given by (3.1). Put $\tilde{q} = A||y'||_p^{2-p}|y'|^{p-2}$. Then the couple $(\tilde{q},\tilde{\rho})$ is a solution of Problem IV. *Moreover we have*

$$M_\alpha = 2A\frac{H^{1+1/\alpha}}{B^{2+1/\alpha}}\left(\frac{\alpha+1}{\alpha}\right)^{1/\alpha-1}\left(\frac{5\alpha+1}{4\alpha}\right)^{1/\alpha}\mathcal{B}^2\left(\frac{1}{2},\frac{1}{2}+\frac{1}{2\alpha}\right), \tag{5.3}$$

where $\mathcal{B}$ is Euler's beta function.

To prove this theorem, we will use three lemmas. The first one is analogous to Proposition 2.3.

LEMMA 5.3. *Let $(\tilde{q},\tilde{\rho})$ be a couple from $U_\alpha \times V$ and $\tilde{y}$ be any first eigenfunction of the problem*

$$(\tilde{q}(x)y')' + \lambda\tilde{\rho}(x)y = 0,\ \ y(0) = y(l) = 0.$$

If

$$\int_0^l \tilde{\rho}(x)\tilde{y}(x)^2\,dx \le \int_0^l \rho(x)\tilde{y}(x)^2\,dx,$$

$$\int_0^l q(x)\tilde{y}'(x)^2\,dx \le \int_0^l \tilde{q}(x)\tilde{y}'(x)^2\,dx$$

for every couple $(q,\rho) \in U_\alpha \times V$, then $(\tilde{q},\tilde{\rho})$ is a solution of Problem IV.

LEMMA 5.4. *Let p be a number > 1. Let $m = \inf_y G[y]$, where*

$$G[y] = \frac{(\int_0^l |y'|^p\,dx)^{2/p}}{\int_0^l \tilde{\rho}y^2\,dx}$$

and $\tilde{\rho}$ is given by (3.1). *The infimum above is taken in the class of all nonzero functions y from $W_0^{1,p}(0,l)$. Then G has a nonnegative minimizer $\tilde{y}$ in $W_0^{1,p}(0,l)$ with the following properties:*

(i) *$\tilde{y}$ is concave and symmetric about $x = l/2$.*

(ii) *$I = \{x \in [0,l], \tilde{y}(x) = \max|\tilde{y}|\}$ is exactly the interval $[a, l-a]$; $a = (2H)^{-1}B$.*

(iii) *$\tilde{y}'$ is everywhere continuous and $\tilde{y}$ satisfies the equation*

$$(||y'||_p^{2-p}|y'|^{p-2}y')' + m\tilde{\rho}y = 0$$

at every point of the interval $]0,l[$.

(iv)

$$m = 2\frac{H^{2-2/p}}{B^{3-2/p}}\left(\frac{p-1}{p}\right)^{-2/p}\left(\frac{3p-1}{2p}\right)^{1-2/p}\mathcal{B}^2\left(\frac{1}{2}, 1-\frac{1}{p}\right).$$

(v) *We have*

$$\tilde{y}(x) = \tilde{y}(a) + c_1(a-x)^{p/(p-1)}[1+o(1)] \ \ as\ \ x \to a^-,$$

$$\tilde{y}'(x) = c_2(a-x)^{1/(p-1)}[1+o(1)] \ \ as\ \ x \to a^-,$$

$$\tilde{y}(x) = c_3 x[1+o(1)] \ \ as\ \ x \to 0^+,$$

where c_1, c_2, and c_3 are nonzero numbers.

The proof of this lemma is similar to that of Lemma 3.1 and will therefore be omitted.

LEMMA 5.5. *Let $p > 2$ and let the functions $\tilde{\rho}$ and $\tilde{y}$ be as in Lemma* 5.4. *Put $\tilde{q}(x) = ||\tilde{y}'||_p^{2-p}|\tilde{y}'(x)|^{p-2}$. Let*

$$\mu = \inf_{y \in H_0^1(0,l)} \frac{\int_0^l \tilde{q}y'^2\,dx}{\int_0^l \tilde{\rho}y^2\,dx}.$$

Then μ is attained at $\tilde{y}$ and $\mu = m$.

Proof. Let $\{y_k\} \subset H_0^1(0,l)$ be a minimizing sequence normalized so that $\int_0^l \tilde{\rho} y_k^2\,dx = 1$. Then the integrals $\int_0^{a-\varepsilon} y_k'^2\,dx$ and $\int_{l-a+\varepsilon}^l y_k'^2\,dx$ are bounded and so there exists a subsequence converging uniformly in $[0,a-\varepsilon]$ and in $[l-a+\varepsilon,l]$ and weakly in $H^1(0,a-\varepsilon)$ and in $H^1(l-a+\varepsilon,l)$. Using a diagonalization one can find a subsequence converging almost everywhere in $(0,a)$ and in $(l-a,l)$ to a function $\bar{y}(x)$. We shall reason only in the interval $(0,a)$. We shall show that $\int_{a-\varepsilon}^a y_k^2\,dx \le C_0\varepsilon^\beta$, where C_0 and β are positive numbers independent of k. Taking into account the fact that

$$\tilde{q}(x) = c(a-x)^\gamma[1+o(1)] \ \ as\ x \to a^-,$$

where $\gamma = (p-2)/(p-1)$, $0<\gamma<1$, we deduce from the equality $y_k(a-\varepsilon) = \int_0^{a-\varepsilon} y_k'(t)\,dt$ that

$$y_k^2(a-\varepsilon) \le \int_0^{a-\varepsilon} \tilde{q}(x) y_k'^2(x)\,dx \int_0^{a-\varepsilon} \tilde{q}^{-1}(x)\,dx \le C\varepsilon^{1-\gamma}.$$

Similarly, for every x in $(a-\varepsilon,a)$, we have $y_k(x) = y_k(a-\varepsilon) + \int_{a-\varepsilon}^x y_k'(t)\,dt$ and therefore

$$\begin{aligned} y_k^2(x) &\le 2y_k^2(a-\varepsilon) + 2\left(\int_{a-\varepsilon}^x y_k'(t)\,dt\right)^2 \\ &\le 2C\varepsilon^{1-\gamma} + \frac{2C}{1-\gamma}\varepsilon^{1-\gamma} \le 4C\varepsilon^{1-\gamma} \end{aligned}$$

for ε small enough. It follows that $\int_{a-\varepsilon}^a y_k^2\,dx \le 4C\varepsilon^{2-\gamma}$. Repeating the same argument in the interval $(l-a,a)$ we obtain that $\int_{l-a}^{l-a+\varepsilon} y_k^2(x)\,dx \le 4C\varepsilon^{2-\gamma}$. If we extend the function $\bar{y}$ to the whole interval $(0,l)$ by a smooth function (the obtained function will also be denoted by $\bar{y}$), we obtain

$$\int_0^l \tilde{\rho}(x)\bar{y}^2(x)\,dx = 1 \ \text{ and } \ \int_0^l \tilde{q}(x)\bar{y}'^2(x)\,dx \le \mu.$$

This implies that $\int_0^l \tilde{q}(x)\bar{y}'^2(x)\,dx = \mu$. Since $\bar{y}$ is a minimizer in $H_0^1(0,l)$, it satisfies the equation

$$(\tilde{q}(x)\bar{y}')' + \mu\tilde{\rho}\bar{y} = 0 \tag{5.4}$$

and the boundary conditions $\bar{y}(0) = \bar{y}(l) = 0$. Similarly, $\tilde{y}$ satisfies

$$(\tilde{q}(x)\tilde{y}')' + m\tilde{\rho}\tilde{y} = 0 \tag{5.5}$$

and $\tilde{y}(0) = \tilde{y}(l) = 0$. Multiplying (5.4) (resp., (5.5)) by $\tilde{y}$ (resp., $\bar{y}$), integrating, and taking the difference, we obtain that

$$(m-\mu)\int_0^l \tilde{y}(x)\bar{y}(x)\,dx = \tilde{q}(x)\tilde{y}(x)\bar{y}'(x) - \tilde{q}(x)\tilde{y}'(x)\bar{y}(x)|_{x=0}^{x=l} = 0.$$

Since we can assume that $\tilde{y}$ and $\bar{y}$ take positive values in $(0,l)$, we deduce that $m=\mu$ and that $\tilde{y} = c\bar{y}$ in $(0,a)\cup(l-a,l)$. □

Note that the spectrum of (5.4) subject to Dirichlet boundary conditions is equal to that of the problem

$$(\bar{q}(x)y')' + \lambda H y = 0, \ \ y(0) = y(a) = 0, \tag{5.6}$$

where the function $\bar{q}$ is the restriction of $\tilde{q}$ to the interval $(0, a)$. But unlike (5.6), the eigenvalues of (5.4) all have multiplicity of infinite order.

Proof of Theorem 5.2. Let $\alpha > 1$ and $p = 2\alpha/(\alpha-1)$. Put $\tilde{q}(x) = A||\tilde{y}'||_p^{2-p}|\tilde{y}'(x)|^{p-2}$, where $\tilde{y}$ is given in Lemma 5.4. We have $\int_0^l \tilde{q}(x)^\alpha\,dx = A^\alpha$, which implies that $\tilde{q} \in U_\alpha$. Hölder's inequality implies

$$\int_0^l q(x)\tilde{y}'(x)^2\,dx \le \left(\int_0^l |\tilde{y}'(x)|^p\,dx\right)^{2/p}\left(\int_0^l q(x)^\alpha\,dx\right)^{1/\alpha}$$
$$= A\left(\int_0^l |\tilde{y}'(x)|^p\,dx\right)^{2/p} = \int_0^l \tilde{q}(x)\tilde{y}'(x)^2\,dx$$

for all $q \in U_\alpha$. On the other hand, since for each $x \in [0, l]$

$$\min_{0\le\rho\le H}[\tilde{y}^2(x)\rho - \tilde{y}^2(a_0)\rho] = \tilde{y}^2(x)\tilde{\rho}(x) - \tilde{y}^2(a_0)\tilde{\rho}(x),$$

Theorem 2.1 shows that $\int_0^l \tilde{\rho}(x)\tilde{y}(x)^2\,dx \le \int_0^l \rho(x)\tilde{y}(x)^2\,dx$ for every function $\rho \in V$. Furthermore, since the differential equation $(||\tilde{y}'||_p^{2-p}|\tilde{y}'|^{p-2}\tilde{y}')' + m\tilde{\rho}\tilde{y} = 0$ holds everywhere in $(0, l)$, Lemma 5.3 and Lemma 5.5 imply that $(\tilde{q}, \tilde{\rho})$ is a solution of Problem IV, and hence Theorem 5.2 is proved. □

COROLLARY 5.6. *Let q and ρ be two functions not identically zero such that $q \in L^\alpha(0, l)$ and $\rho \in L^\infty(0, l)$, where $\alpha > 1$. Then*

$$\lambda(q, \rho) \le C(\alpha)\,\frac{||q||_\alpha||\rho||_\infty^{1+1/\alpha}}{||\rho||_1^{2+1/\alpha}}, \tag{5.7}$$

where $C(\alpha)$ is a constant depending only on α.

Proof. To prove (5.7), it is sufficient to take

$$A = ||q||_\alpha, \qquad B = ||\rho||_1, \qquad H = ||\rho||_\infty$$

and use (5.3). Thus one obtains that

$$C(\alpha) = 2\left(\frac{\alpha+1}{\alpha}\right)^{1/\alpha-1}\left(\frac{5\alpha+1}{4\alpha}\right)^{1/\alpha}\mathcal{B}^2\left(\frac{1}{2}, \frac{1}{2}+\frac{1}{2\alpha}\right). \tag{5.8}$$

□

THEOREM 5.7. *If $\alpha = 1$, then for all $(q, \rho) \in U_1 \times V$*

$$\lambda(q, \rho) \le 12AH^2B^{-3}.$$

The equality is attained if ρ equals the function $\tilde{\rho}$ defined by (3.1) *and*

$$\tilde{q}(x) = \begin{cases} M_1H(a^2 - x^2)/2 & \text{if } 0 \le x \le a, \\ 0 & \text{if } a < x \le l/2, \\ \tilde{q}(l-x) & \text{if } l/2 \le x \le l, \end{cases}$$

where $M_1 = 12AH^2B^{-3}$ and $a = (2H)^{-1}B$.

Proof. The key to the proof is choosing a test function $\tilde{y}$ defined by

$$\tilde{y}(x) = \begin{cases} ax & \text{if } 0 \le x \le a, \\ a^2 & \text{if } a < x \le l/2, \\ \tilde{y}(l-x) & \text{if } l/2 \le x \le l, \end{cases}$$

for which we have for all $(q,\rho)\in U_1\times V$

$$\lambda(q,\rho)\le \frac{\int_0^l q\tilde{y}'^2\,dx}{\int_0^l \rho\tilde{y}^2\,dx}\le \frac{a^2A}{\int_0^l \rho\tilde{y}^2\,dx}. \tag{5.9}$$

By using the same argument as in the proof of Theorem 5.2, we obtain that

$$\int_0^l \tilde{\rho}(x)\tilde{y}(x)^2\,dx\le\int_0^l \rho(x)\tilde{y}(x)^2\,dx$$

for all $\rho\in V$. Hence from (5.9) we get

$$\lambda(q,\rho)\le a^2A\left(\int_0^l \tilde{\rho}(x)\tilde{y}(x)^2\,dx\right)^{-1}=M_1.$$

Now we shall construct a function $\tilde{q}$ such that $\tilde{y}$ satisfies the equation

$$(\tilde{q}\tilde{y}')'+M_1\tilde{\rho}\tilde{y}=0. \tag{5.10}$$

To do this we solve at first the differential equation

$$q'+M_1Hx=0,\qquad 0\le x\le a,$$

with the boundary condition $q(a)=0$. This gives

$$q(x)=M_1H(a^2-x^2)/2,\qquad 0\le x\le a.$$

Then we construct the function $\tilde{q}$ on the whole interval $[0,l]$ in the following way: $\tilde{q}(x)=q(x)$ for $0\le x\le a$, $\tilde{q}(x)=0$ for $a\le x\le l/2$, and $\tilde{q}(x)=\tilde{q}(l-x)$ for $l/2\le x\le l$. Thus $\tilde{q}\in U_1$ and the differential equation (5.10) holds everywhere in $]0,l[$, which implies that $\lambda(\tilde{q},\tilde{\rho})=M_1$ (by means of a result analogous to Lemma 5.5). Therefore the couple $(\tilde{q},\tilde{\rho})$ is extremal. □

Observe that Theorem 5.7 can be seen as a limiting case of Theorem 5.2 as $\alpha\to 1$. In fact, we have

$$M_1=\lim_{\alpha\to 1}M_\alpha.$$

Above we have not discussed the uniqueness of the obtained solutions. However, it can be proved that for each $\alpha\ge 1$ the extremal couple $(\tilde{q},\tilde{\rho})$, solution of Problem IV, is unique. Finally, we mention that, as for Problems I and II, our method applies to Problem IV when ρ satisfies conditions (4.1) or (4.3).

6. Estimates of all the eigenvalues. We now give an upper bound for the nth eigenvalue of problem (5.1)–(5.2) considered in the last section. The main result presented here is the following theorem.

THEOREM 6.1. *Let $\lambda_n(q,\rho)$ be the nth eigenvalue of problem* (5.1)–(5.2). *Suppose that the coefficients q and ρ are not identically zero and satisfy $q\in L^\alpha(0,l)$ and $\rho\in L^\infty(0,l)$, where α is a number ≥ 1. Then*

$$\lambda_n(q,\rho)\le n^2\bar{C}(\alpha)\,||q||_\alpha||\rho||_\infty^{1+1/\alpha}||\rho||_1^{-(2+1/\alpha)},$$

where $\bar{C}(\alpha)$ is given by (5.8) *if $\alpha>1$ and $\bar{C}(1)=12$. Moreover the equality is attained by two periodic functions q and ρ with period l/n and such that $q(x)=$*

$\tilde{q}(nx)$, $\rho(x) = \tilde{\rho}(nx)$ *for all* $x \in (0, l/n)$, *where* $\tilde{q}$ *and* $\tilde{\rho}$ *are the optimal functions indicated in Theorems* 5.1 *and* 5.2.

To prove this theorem, we shall use the following lemma.

LEMMA 6.2. *Let* r, s *be two real numbers such that* $rs < 0$ *and* $r + s \geq 1$. *Denote by* E *the set of all vectors* $X = (x_1, x_2, ..., x_n) \in (]0, 1[)^n$ *satisfying* $\sum_{i=1}^n x_i < 1$ *and define the function* $F : E \times E \longrightarrow R$ *by*

$$F(X, Y) = x_1^r y_1^s + x_2^r y_2^s + \cdots + x_n^r y_n^s$$

$$+ (1 - x_1 - x_2 - \cdots - x_n)^r (1 - y_1 - y_2 - \cdots - y_n)^s.$$

Then the function F *attains its minimum value* $F_{min} = (n+1)^{1-(r+s)}$ *when* $X = Y = (n+1)^{-1}(1, 1, ..., 1)$. *Moreover the minimum point is unique in the case* $r + s > 1$.

This lemma generalizes a result in [11], regarding the minimal value of the function

$$G(x, y) = x^{2/p} y^{3-2/p} + (1 - x)^{2/p} (1 - y)^{3-2/p}$$

defined on the square $0 < x < 1$, $0 < y < 1$, where p is a real number satisfying $p < 2/3$ and $p \neq 0$. We refer to [11] for the proof. We note that the minimum point in Lemma 6.2 is not unique in general. Indeed, for the case $r + s = 1$, one can easily verify that the function F achieves its minimum $F_{min} = 1$ at every couple $(X, Y) \in E \times E$ satisfying $x_i = y_i$ for $i = 1, ..., n$. If $rs < 0$ and $r + s < 1$, then the function F can take arbitrary small positive values.

Proof of Theorem 6.1. We begin with the case $n = 2$. The general case follows from similar arguments. Let q and ρ be two measurable functions such that $||q||_\alpha = 1$, $\alpha \geq 1$, $||\rho||_1 = 1$ and $0 \leq \rho(x) \leq H$ for all $x \in [0, l]$. Let y_2 (resp., $\lambda_2(q, \rho)$) be the second eigenfunction (resp., eigenvalue) of the problem

$$(q(x)y')' + \lambda \rho(x) y = 0, \quad y(0) = y(l) = 0.$$

As is well known, the function y_2 admits precisely one zero (call it x_1) in the interval $(0, l)$. The number $\lambda_2(q, \rho)$ is the first eigenvalue of the two problems

$$(q(x)y')' + \lambda \rho(x) y = 0, \quad y(0) = y(x_1) = 0,$$

$$(q(x)y')' + \lambda \rho(x) y = 0, \quad y(x_1) = y(l) = 0.$$

According to Corollary 5.6 and Theorem 5.7, we have

$$\lambda_2(q, \rho) \leq \bar{C}(\alpha) H^{1+1/\alpha} \left(\int_0^{x_1} q(x)^\alpha \, dx \right)^{1/\alpha} \left(\int_0^{x_1} \rho(x) \, dx \right)^{-(2+1/\alpha)}, \tag{6.1}$$

where $\bar{C}(\alpha)$ is given by (5.8) if $\alpha > 1$ and $\bar{C}(1) = 12$. Inequality (6.1) can be written as

$$\left(\int_0^{x_1} q(x)^\alpha \, dx \right)^{-1/\alpha} \left(\int_0^{x_1} \rho(x) \, dx \right)^{2+1/\alpha} \leq \bar{C}(\alpha) \, H^{1+1/\alpha} \, [\lambda_2(q, \rho)]^{-1}. \tag{6.2}$$

Similarly, we have

$$\left(\int_{x_1}^{l} q(x)^\alpha \, dx \right)^{-1/\alpha} \left(\int_{x_1}^{l} \rho(x) \, dx \right)^{2+1/\alpha} \leq \bar{C}(\alpha) \, H^{1+1/\alpha} \, [\lambda_2(q, \rho)]^{-1}. \tag{6.3}$$

Put $a = \int_0^{x_1} q(x)^\alpha\,dx$ and $b = \int_0^{x_1} \rho(x)\,dx$. By summing (6.2) and (6.3), we obtain

$$a^{-1/\alpha}b^{2+1/\alpha} + (1-a)^{-1/\alpha}(1-b)^{2+1/\alpha} \le \bar{C}(\alpha)H^{1+1/\alpha}\,[\lambda_2(q,\rho)]^{-1}. \tag{6.4}$$

By Lemma 6.2 (applied for $n = 1$), the left side of (6.4) is not less than 2^{-1} since a and b are both in the interval $(0,1)$. Therefore

$$1/2 \le 2\,\bar{C}(\alpha)\,H^{1+1/\alpha}\,[\lambda_2(q,\rho)]^{-1},$$

which shows that $\lambda_2(q,\rho) \le 4\,\bar{C}(\alpha)\,H^{1+1/\alpha}$. If $||q||_\alpha = A$ and $||\rho||_1 = B$, we obtain

$$\lambda_2(q,\rho) \le 4\,\bar{C}(\alpha)\,H^{1+1/\alpha}\,A\,B^{-(2+1/\alpha)}.$$

To prove the general case $n \ge 2$, we argue as above and take into account the fact that the nth eigenfunction of (5.1)–(5.2) has exactly $(n-1)$ zeros between 0 and l. □

7. Some isoperimetric inequalities. For a given function f defined on the interval $[0,l]$, we denote by f_n^+ (resp., f_n^-) the symmetrically increasing (resp., decreasing) rearrangement of f of degree n. We recall that the function f_n^+ is uniquely defined by the following conditions (see [17]):

(i) f_n^+ and f are equimeasurable on $[0,l]$. That is, for all $t \ge 0$

$$\text{meas}\{x/f_n^+(x) \ge t\} = \text{meas}\{x/f(x) \ge t\}.$$

(ii) f_n^+ is periodic on $[0,l]$ with a period equal to l/n.

(iii) f_n^+ is symmetric in $[0,l/n]$ about $x = l/(2n)$.

(iv) f_n^+ is decreasing in the interval $[0,l/(2n)]$.

Similarly, f_n^- is (uniquely) defined by (i)–(iii) and (iv)′: f_n^- is increasing in the interval $[0,l/(2n)]$. We will use throughout the notation f^+ and f^- to denote f_1^+ and f_1^-, respectively. For more information on rearrangements see [1] and [17].

We now return to problem (1.1)–(1.2). Let q and ρ be arbitrary measurable and nonnegative functions. We have already seen in Remark 3.4 that in general we cannot have an estimate for the first eigenvalue $\lambda(q,\rho)$ of problem (1.1)–(1.2) of the form $\lambda(q,\rho) \le \lambda(q^-,\rho^+)$, neither of the form $\lambda(q,\rho) \le \lambda(q^+,\rho^+)$. In this section we shall derive some conditions on q and on ρ so that the inequality $\lambda(q,\rho) \le \lambda(q^-,\rho^+)$ holds. For rearrangements of higher degree we shall prove the inequality $\lambda_n(q,\rho) \ge \lambda_n(q_n^+,\rho_n^-)$ for all integers n, where here and throughout $\lambda_n(q,\rho)$ is the nth eigenvalue of problem (1.1)–(1.2); $[\lambda_1(q,\rho) = \lambda(q,\rho)]$.

Some results concerning other differential equations have been obtained by Barnes [5]. He determined lower bounds for the first eigenvalues of the equations $(p(x)y')' + q(x)y + \lambda\rho(x)y = 0$ and $(p(x)y'')''q(x)y\mu\rho(x)y = 0$, subject to Dirichlet boundary conditions. He found that if $\lambda_1(p,q,\rho)$ (resp., $\mu_1(p,q,\rho)$) is positive, then

$$\lambda_n(p,q,\rho) \ge \lambda_n(p_n^-,q_n^-,\rho_n^-), \quad \text{resp.,} \quad \mu_n(p,q,\rho) \ge \mu_n(p_n^+,q_n^-,\rho_n^-).$$

These inequalities do not require additional conditions aside from some regularity assumptions on the coefficients p, q, and ρ and the positivity of the eigenvalues $\lambda_1(p,q,\rho)$ and $\mu_1(p,q,\rho)$. In what follows we suppose that the function ρ satisfies $h \le \rho(x) \le H$ for all $x \in [0,l]$, where h and H are two positive numbers.

LEMMA 7.1. *Let $I(\rho)$ denote the Lebesgue integral of $\rho(x)$ over $[0,l]$. If*

$$H\max_x q(x) - h\min_x q(x) \le \max\left\{h\pi^2 l^{-2},\, 4hHl^{-1}I(\rho)^{-1}\right\}, \tag{7.1}$$

then $q(x) - \lambda_1(q,\rho)\rho(x) \le 0$ for all $x \in [0,l]$. Consequently, if $y_1(x)$ is a positive

first eigenfunction of problem (1.1)–(1.2) *and if condition* (7.1) *holds, then* $y_1(x)$ *is concave.*

Proof. A variational principle implies that

$$\lambda(q,\rho)=\inf_y \frac{\int_0^l y'^2\,dx+\int_0^l qy^2\,dx}{\int_0^l \rho y^2\,dx},$$

where the infimum is taken in the class of all nonzero functions y from $C_0^1[0,l]$. By taking into account the inequality (see, for example, [11])

$$\max_{x\in[0,l]}|y(x)|^2\le\frac{l}{4}\int_0^l y'^2\,dx$$

valid for all $y\in H_0^1(0,l)$, we obtain

$$\begin{aligned}\lambda(q,\rho)&\ge\inf_y\left[\int_0^l y'^2\,dx\Big/\int_0^l\rho y^2\,dx\right]+\min_x q(x)/H\\&\ge\max\left\{H^{-1}(\pi/l)^2,\,4l^{-1}I(\rho)^{-1}\right\}+\min_x q(x)/H.\end{aligned}$$

If (7.1) holds, from the last inequality we get $\lambda(q,\rho)\ge\max_x q(x)/h$, which implies that $\lambda(q,\rho)\rho(x)-q(x)\ge 0$ for all x in $[0,l]$. If $y_1(x)$ is a positive first eigenfunction of problem (1.1)–(1.2) corresponding to $\lambda(q,\rho)$, then $y_1''(x)=[q(x)-\lambda_1(q,\rho)\rho(x)]y_1(x)\le 0$ in $[0,l]$. This means that $y_1(x)$ is a concave function. □

The following result generalizes a theorem of Schwarz [17].

THEOREM 7.2. *If condition* (7.1) *holds, then*

$$\lambda(q,\rho)\le\lambda(q^-,\rho^+).\tag{7.2}$$

Moreover if $q=q^-$ *(resp.,* $\rho=\rho^+$*) and equality holds in* (7.2)*, then* $\rho=\rho^+$ *(resp.,* $q=q^-$*).*

Proof. Suppose that (7.1) is fulfilled, and let $y_1(x)$ be a positive first eigenfunction of problem (1.1)–(1.2) corresponding to q^- and ρ^+. Since q^- and ρ^+ satisfies condition (7.1), Lemma 7.1 shows that $y_1(x)$ is a concave function. In view of the symmetry of q^- and ρ^+, $y_1(x)$ is also symmetric and symmetrically decreasing. Thus using the inequality

$$\int_0^l f^-\,g^+\,dx\le\int_0^l f\,g\,dx\le\int_0^l f^-\,g^-\,dx,$$

where f and g are two measurable nonnegative functions defined on $[0,l]$, we see that

$$\int_0^l qy_1^2\,dx\le\int_0^l q^-y_1^2\,dx\quad\text{and}\quad\int_0^l\rho^+y_1^2\,dx\le\int_0^l\rho y_1^2\,dx.$$

Hence it follows that

$$\begin{aligned}\lambda(q^-,\rho^+)&\ge\frac{\int_0^l y_1'^2\,dx+\int_0^l qy_1^2\,dx}{\int_0^l\rho y_1^2\,dx}\\&\ge\lambda(q,\rho).\end{aligned}$$

Suppose now that equality holds in (7.2) for a couple (q,ρ) satisfying (7.1) and such that $q = q^-$. Then $\lambda(q,\rho^+) = \lambda(q,\rho)$ and the reasoning above implies that $\lambda(q,\rho^+)$ and $\lambda(q,\rho)$ have the same first eigenfunction. This means that $\rho = \rho^+$. □

Remark. Suppose that ρ vanishes at some $x \in [0,l]$. To simplify this remark we suppose that q and ρ are both continuous and not identically zero. Then $q^-(l/2) > 0$ and ρ^+ vanishes at the point $x = l/2$. Let y_1 be a positive first eigenfunction of problem (1.1)–(1.2) corresponding to q^- and ρ^+. Since y_1'' is continuous and $y_1''(l/2) > 0$, y_1'' takes positive values in a neighborhood of $l/2$. Therefore y_1 is not a concave function. Thus in the case where ρ has zeros in $[0,l]$ there are no conditions assuring the concavity of the first eigenfunction. This interpretation leads us to the question of whether the concavity of the first eigenfunction corresponding to q^- and ρ^+ is necessary for estimate (7.2) to hold.

COROLLARY 7.3. *Let $\lambda(q)$ denote the first eigenvalue of the problem*

$$y'' - q(x)y + \lambda y = 0, \quad y(0) = y(l) = 0.$$

If $\max_x q(x) - \min_x q(x) \le \pi^2/l^2$, then

$$\lambda(q) \le \lambda(q^-).$$

Moreover equality holds only if $q = q^-$.

THEOREM 7.4. *For any nonzero integer n, we have*

$$\lambda_n(q,\rho) \ge \lambda_n(q_n^+, \rho_n^-). \tag{7.3}$$

Proof. Unlike (7.2), inequality (7.3) does not require condition (7.1). In fact, for $n = 1$ we have

$$\begin{aligned}\lambda(q,\rho) &\ge \frac{\int_0^l (y_1^-)'^2\,dx + \int_0^l q^+(y_1^-)^2\,dx}{\int_0^l \rho^-(y_1^-)^2\,dx}\\ &\ge \lambda(q^+,\rho^-),\end{aligned}$$

where $y_1(x)$ is a first eigenfunction of (1.1) and (1.2) corresponding to $\lambda(q,\rho)$. Let us prove the case $n = 2$. For this we shall need some concepts from [17]. The general case may be obtained via similar arguments. Let a be the zero in $(0,l)$ of the second eigenfunction of problem (1.1)–(1.2) (it is well known that a is unique for such a eigenfunction). Put $I_1 = [0,a]$ and $I_2 = [a,l]$. We consider the restriction of $\rho(x)$ and $q(x)$ in each of those intervals and we denote by $\rho^{-(i)}$ (resp., $q^{+(i)}$) the symmetrically decreasing (resp., increasing) rearrangement of first degree of the restriction of $\rho(x)$ (resp., $q(x)$) to I_i, $i = 1,2$. Let $\lambda^{(1)}$ and $\lambda^{(2)}$ be the first eigenvalues to the problems

$$y'' - q^{+(1)}(x)y + \lambda\rho^{-(1)}(x)y = 0, \quad y(0) = y(a) = 0, \tag{7.4}$$

and

$$y'' - q^{+(2)}(x)y + \lambda\rho^{-(2)}(x)y = 0, \quad y(a) = y(l) = 0, \tag{7.5}$$

respectively. By the first part of the proof we obtain

$$\lambda_2(q,\rho) \ge \max\left\{\lambda^{(1)}, \lambda^{(2)}\right\}.$$

Without loss of generality we assume that $\lambda^{(1)} \geq \lambda^{(2)}$ and define two functions P and Q as follows:

$$P(x) = \rho^{-(1)}(x) \text{ in } I_1, \quad P(x) = \rho^{-(2)}(x) \text{ in } I_2,$$

and

$$Q(x) = \begin{cases} q^{+(1)}(x) & \text{for } x \in I_1, \\ q^{+(2)}(x) + (\lambda^{(1)} - \lambda^{(2)})\rho^{-(2)}(x) & \text{for } x \in I_2. \end{cases}$$

Let $y_{(1)}(x)$ (resp., $y_{(2)}(x)$) be the first eigenfunction of problem (7.4) (resp., (7.5)). Define the function $Y(x)$ by $Y(x) = y_{(1)}(x)$ in I_1 and $Y(x) = \xi y_{(2)}(x)$ in I_2. The real ξ is chosen such that $Y'(x)$ is continuous at $x = a$. Thus $\lambda^{(1)}$ (resp., $Y(x)$) is the second eigenvalue (resp., eigenfunction) of the eigenvalue problem

$$Y'' - QY + \lambda PY = 0, \quad Y(0) = Y(l) = 0.$$

Let α be the midpoint of I_1. We may assume that $Y(x) > 0$ in the interior of I_1, and we define the new functions $\tilde{P}$, $\tilde{Q}$, and $\tilde{Y}$ in the interval $[0, l/2]$ as follows:

$$\tilde{P}(x) = \begin{cases} P(x) & \text{for } x \in [0, \alpha], \\ P(x + l/2) & \text{for } x \in [\alpha, l/2], \end{cases}$$

$$\tilde{Q}(x) = \begin{cases} Q(x) & \text{for } x \in [0, \alpha], \\ Q(x + l/2) & \text{for } x \in [\alpha, l/2], \end{cases}$$

and

$$\tilde{Y}(x) = \begin{cases} Y(x) & \text{for } x \in [0, \alpha], \\ \zeta Y(x + l/2) & \text{for } x \in [\alpha, l/2]. \end{cases}$$

The nonpositive real ζ is chosen such that $\tilde{Y}$ is continuous at $x = \alpha$. Thus $\tilde{Y}$ is the first eigenfunction of the problem

$$Y'' - \tilde{Q}Y + \lambda \tilde{P}Y = 0, \quad Y(0) = Y(l/2) = 0, \tag{7.6}$$

and $\lambda^{(1)}$ is the first eigenvalue of (7.6). Now denote by $\tilde{P}^-$ (resp., $\tilde{Q}^+$) the symmetrically decreasing (resp., increasing) rearrangement of $\tilde{P}$ (resp., $\tilde{Q}$) of first degree in $[0, l/2]$. Then, the beginning of the proof shows that

$$\lambda^{(1)} \geq \lambda_1(\tilde{Q}^+, \tilde{P}^-). \tag{7.7}$$

On the other hand, it is easy to see that the extension of $\tilde{P}^-$ (resp., $\tilde{Q}^+$) to the whole interval $[0, l]$ with a period equal to $l/2$ is exactly the symmetrically decreasing (resp., increasing) rearrangement of second degree of P (resp., Q) in $[0, l]$. Hence (7.7) and the equality $\lambda_1(\tilde{Q}^+, \tilde{P}^-) = \lambda_2(Q_2^+, P_2^-)$ yield

$$\lambda^{(1)} \geq \lambda_2(Q_2^+, P_2^-).$$

On the other hand, as $Q_2^+(x) \geq q_2^+(x)$ and $P_2^-(x) = \rho_2^-(x)$ in $[0, l]$, Sturm comparison theorem implies that

$$\lambda_2(Q_2^+, P_2^-) \geq \lambda_2(q_2^+, \rho_2^-).$$

Hence inequality (7.3) holds for $n = 2$ since $\lambda_2(q, \rho) \geq \lambda^{(1)}$. □

We finally mention that, unlike (7.3), it is not clear whether the inequality $\lambda_n(q,\rho) \le \lambda_n(q_n^-, \rho_n^+)$ holds for $n > 1$, even when q and ρ satisfy condition (7.1). However, by modifying the proof of Theorem 7.2, we can prove the inequality

$$\lambda_n(q,\rho) \le \frac{n^2\pi^2 H}{B^2}\left[\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{AB}{n^2 H\pi^2}}\right]^2$$

for all $(q,\rho) \in U \times V$, where the sets U and V are as in section 1. The equality is reached by two periodic functions q and ρ having period l/n and such that $q(x) = q_0(nx)$, $\rho(x) = \rho_0(nx)$ for all $x \in (0, l/n)$, where q_0 and ρ_0 are the optimal functions found in Theorem 3.3.

REFERENCES

[1] C. Bandle, *Isoperimetric Inequalities and Applications*, Pitman, Boston, London, Melbourne, 1980.

[2] C. Bandle, *Extremal problems for eigenvalues of the Sturm-Liouville type*, General Inequalities 5, Internat. Schriftenreihe Numer. Math., 80, Birkhäuser, Basel-Boston, MA, 1987, pp. 319–336.

[3] D. O. Banks, *Bounds for the eigenvalues of a nonhomogeneous hinged vibrating rod*, J. Math. Mech., 16 (1967), pp. 949–966.

[4] D. C. Barnes, *Extremal problems for eigenvalues with applications to buckling, vibration and sloshing*, SIAM J. Math. Anal., 16 (1985), pp. 341–357.

[5] D. C. Barnes, *Rearrangement of functions and lower bounds for eigenvalues of differential equations*, Appl. Anal., 13 (1982), pp. 237–248.

[6] E. R. Barnes, *The shape of the strongest column and some related extremal eigenvalue problems*, Quart. Appl. Math., 34 (1977), pp. 393–409.

[7] C. Bennewitz and E. J. M. Veling, *Optimal bounds for the spectrum of a one-dimensional Schrödinger operator*, General Inequalities 6, Internat. Schriftenreihe Numer. Math., 103, Birkhäuser, Basel-Boston, MA, 1992, pp. 257–268.

[8] H. Egnell, *Extremal properties of the first eigenvalue of a class of elliptic problems*, Ann. della Scuola Norm. Sup. Pisa Cl. Sci., 14 (1987), pp. 1–48.

[9] Yu. V. Egorov and S. Karaa, *Optimisation de la première valeur propre de l'opérateur de Sturm-Liouville*, C. R. Acad. Sci. Paris Sér. I. Math., 319 (1994), pp. 793–798.

[10] Yu. V. Egorov and V. A. Kondratiev, *On an estimate of the principal eigenvalue of the Sturm-Liouville operator*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 6 (1991), pp. 5–11.

[11] Yu. V. Egorov and V. A. Kondratiev, *On estimates of the first eigenvalue in some Sturm-Liouville problems*, preprint, Max-Planck-Arbeitsgruppe "Partielle Differentialgleichungen und Komplexe Analysis," 1994, pp. 1–71.

[12] M. Essén, *On estimating eigenvalues of a second order linear differentiel operator*, General Inequalities 5, Internat. Schriftenreihe Numer. Math., 80, Birkhäuser, Basel-Boston, MA, 1987, pp. 347–366.

[13] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[14] S. Karaa, *Valeurs propres extremales dans des problèmes de Sturm-Liouville*, C. R. Acad. Sci. Paris, Sér. I. Math., 321, (1995), pp. 265–270.

[15] S. Karaa, *Extremal eigenvalues and their associated nonlinear equations*, Boll. Un. Mat. Ital., 10 (1996), pp. 625–649.

[16] M. G. Krein, *On certain problems on the maximum and minimum of characteristic values and the Lyapunov zones of stability*, Trans. Amer. Math. Soc., 2 (1955), pp. 163–187.

[17] B. Schwarz, *On the extrema of a nonhomogeneous string with equimeaserable density*, J. Math. Mech., 10 (1961), pp. 401–422.

[18] I. Tadjbakhsh and J. B. Keller, *Strongest columns and isoperimetric inequalities for eigenvalues*, J. Appl. Mech., 29 (1962), pp. 159–164.

[19] R. Tahraoui, *Quelques remarques sur le controle des valeurs propres*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. VIII, J. L. Lions and H. Brézis, eds., Pitman, London, 1986, pp. 176–213.

[20] G. Talenti, *Estimates for eigenvalues of Sturm-Liouville problems*, General Inequalities 4, Internat. Schriftenreihe Numer. Math. 71, Birkhäuser, Basel, 1984, pp. 341–350.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1301–1334, November 1998

BLOW-UP IN NONLOCAL REACTION-DIFFUSION EQUATIONS*

PHILIPPE SOUPLET†

Abstract. We present new blow-up results for reaction-diffusion equations with nonlocal nonlinearities. The nonlocal source terms we consider are of several types, and are relevant to various models in physics and engineering. They may involve an integral of the unknown function, either in space, in time, or both in space and time, or they may depend on localized values of the solution. For each type of problems, we give finite time blow-up results which significantly improve or extend previous results of several authors. In some cases, when the nonlocal source term is in competition with a local dissipative or convective term, optimal conditions on the parameters for finite time blow-up or global existence are obtained.

Our proofs rely on comparison techniques and on a variant of the eigenfunction method combined with new properties on systems of differential inequalities. Moreover, a unified local existence theory for general nonlocal semilinear parabolic equations is developed.

Key words. nonlinear parabolic equations, nonlocal source, partial integrodifferential equations, finite-time blow-up, global existence

AMS subject classifications. Primary, 35K55, 35K57, 35B40, 35K10; Secondary, 45K05

PII. S0036141097318900

1. Introduction and examples.

The purpose of this paper is to study the blow-up behavior of nonnegative solutions for some classes of reaction-diffusion equations, where the reaction term may have a nonlocal, functional dependence either in space or in time (or possibly in both space and time). Given a (smoothly) bounded domain Ω in $\mathbb{R}^N$, such an equation may be written in the following general form:

$$u_t - \Delta u = F^t\big(R^t u\big)(x), \quad t > 0, \quad x \in \Omega, \tag{NLRD}$$

where, for each $t > 0$, $F^t : C([0,t] \times \overline{\Omega}) \to C(\overline{\Omega})$, and the past time restriction operator R^t is defined by $R^t u = u_{|[0,t]\times\overline{\Omega}}$. We here consider the initial-boundary value problem associated with (NLRD), with homogeneous Dirichlet boundary conditions. The structure of the right hand of equation (NLRD) means that the "heat source" at a point (t,x) may possibly depend on the whole past history of the temperature in the whole domain Ω.

A very large literature has been devoted to the study of more "classical" diffusion equations with local reaction terms:

$$u_t - \Delta u = f\big(t,x,u(t,x)\big), \quad t > 0, \quad x \in \Omega,$$

subject to appropriate initial and boundary conditions. However, numerous problems, which fall into the form of (NLRD), have been studied, and many of them arise in applications. (Note that nonlocal partial integrodifferential equations not of parabolic type are also widely encountered in applications; see, e.g., [G], [MR].)

To motivate our study, we first give a short review of examples of such parabolic equations studied in the literature, for which existence of nonglobal solutions may

*Received by the editors March 26, 1997; accepted for publication (in revised form) October 13, 1997; published electronically June 22, 1998.
http://www.siam.org/journals/sima/36-5/31890.html
†Laboratoire Analyse Géométrie et Applications, UMR CNRS 7539, Institut Galilée, Université Paris-Nord, 93430 Villetaneuse, France (souplet@math.univ-paris13.fr).

occur. (For other properties of nonlocal equations, see the references in the appendix and in the cited works.) One can distinguish, at least, four types of problems of the form (NLRD).

1.1. Problems with nonlocal reaction terms in space. They take the form

$$u_t - \Delta u = F\big((t,x,u(t,.)\big), \quad t>0, \quad x\in\Omega. \tag{NLS}$$

We distinguish two types within this category. In the first one, the functional F involves an integral of some function of $u(t,.)$ over Ω.

1.1.1. Nonlocal terms induced by an integral over Ω. For instance, Bebernes and Bressan [BB] (see also [BE], [P2]) consider the equation

$$u_t - \Delta u = f\big(t,u(t,x)\big) + \int_\Omega g\big(t,u(t,y)\big)\,dy, \quad t>0, \quad x\in\Omega. \tag{1.1}$$

These authors pay special attention to the case $f(t,u)=e^u$, $g(t,u)=ke^u$ $(k>0)$, for which (1.1) represents an ignition model for a compressible reactive gas, and prove that solutions blow up in the whole domain. Wang and Wang [WW] study the blow up behavior of solutions of the equation

$$u_t - \Delta u = \int_\Omega u^p(t,y)\,dy - ku^q(t,x), \quad t>0, \quad x\in\Omega, \tag{1.2}$$

with $p,\ q>1$, while Budd, Dold, and Stuart [BDS] and Hu and Yin [HY] consider the symmetric of problem (1.2) in the case $q=p$,

$$u_t - \Delta u = u^p - \frac{1}{|\Omega|}\int_\Omega u^p(t,y)\,dy, \quad t>0, \quad x\in\Omega,$$

for which the L^1 energy of the solutions is conserved (under Neumann boundary conditions).

A nonlocal problem where the nonlinearity also involves spatial derivatives of u has been considered by Deng, Kwang, and Levine [DKL]. They investigate the blow-up/global existence properties of the following one-dimensional convection-reaction-diffusion equation:

$$u_t - \Delta u = f\big(u(t,x),\|u(t,.)\|_k\big) + uu_x, \quad t>0, \quad x\in\Omega, \tag{1.3}$$

where

$$\|u(t,.)\|_k = \left(\int_\Omega |u|^k(t,y)\,dy\right)^{1/k} \quad \text{and} \quad f(u,\|u\|_k) = \Big(a\|u\|_k^{p-1}+b\Big)\,u,$$

with $p>1$, $1\le k<\infty$, $a>0$, and $b\in\mathbb{R}$. This problem is related to a turbulence model proposed by Burgers (see the references in [DKL]). On the other hand, Deng [D] and Y. Yin [Yi] recently studied related equations with singular nonlocal nonlinearities, typically

$$u_t - \Delta u = \frac{1}{1-\|u(t,.)\|_k}, \quad t>0, \quad x\in\Omega,$$

with zero initial and boundary conditions. For this problem, the quenching of the solution is proved [Yi]; i.e., $\|u(t,.)\|_k$ reaches 1 and u_t blows up in finite time.

Problems of the form

$$u_t - \Delta u = \frac{f(u(t,x))}{\left(\int_\Omega g(u(t,y))\,dy\right)^p}, \quad t>0, \quad x\in\Omega,$$

with $p>0$, and, e.g., $f(u)=g(u)=e^u$, arise in the modelization of shear banding in metals (see [BT], [F], and the references therein). Similar problems arise in a simplified model for Ohmic heating in a thermistor (see [L]; the full model is a coupled system of a parabolic and an elliptic equation — see [AC1], [AC2], and the references therein).

1.1.2. Localized reaction terms. Another kind of nonlocal problems in space occurs when the intensity of the source at the point x depends on the value of u at a single point x_0, different from x. Physical phenomena where the reaction is driven by the temperature at a single site can be described by equations of the type

$$u_t - \Delta u = f\big(u(t,x_0(t))\big), \quad t>0, \quad x\in\Omega. \tag{LRT}$$

Cannon and Yin [CaY] study the local solvability of (LRT), and Chadam, Peirce, and Yin [CPY] investigate the blow-up property for equation (LRT) with superlinear convex nonlinearities f, in the case when $x_0(t)$ is a constant point x_0.

Problem (LRT) is also related to some local equations with a singular source term, of the form

$$u_t - \Delta u = \delta(x-x_0(t))f(u(t,x)), \quad t>0, \quad x\in\Omega,$$

where δ is the Dirac delta distribution, studied by Olmstead and Roberts (see, e.g., [OR], [O]).

1.2. Problems with nonlocal reaction terms in time. The source term here involves an integral over $[0,t]$ of some function of the solution and of the independent variables:

$$u_t - \Delta u = f\left(u(t,x), \int_0^t g\big(t,x,s,u(s,x)\big)\,ds\right), \quad t>0, \quad x\in\Omega. \tag{NLT}$$

Such equations model diffusion phenomena with memory effects. A special case widely encountered in population dynamics are Volterra diffusion equations, where the nonlocal "hereditary" term takes the form of a convolution with a kernel:

$$g\big(t,x,s,u(s,x)\big) = k(t-s)h(u(s,x)).$$

See [Y1], [Y2], and the references therein for the study of global solutions and their stability properties in the case

$$f \equiv \big(a-bu(t,x)\big)u(t,x) - \int_0^t k(t-s)u(s,x)\,ds,$$

$a, b>0$, $k(t-s)\geq 0$. See also [BG] for some examples relevant to the thermodynamics of phase transition.

Some blow-up results were obtained for equations of the type (NLT) with "explosive" nonlinearities, of the form

$$u_t - \Delta u = \int_0^t k(t,x,s)g(u(s,x))\,ds - f(u(t,x)), \quad t>0, \quad x\in\Omega,$$

in the case $g(u)\sim u^p$ $(p>1)$, $f(u)=\lambda u$ and, e.g., $k(t,s,u)\geq k>0$ (see [Be], [Ko], and [So]).

1.3. Problems with nonlocal reaction terms in space and time. They usually take the form

(NLST)

$$u_t - \Delta u = f\left(t, x, u(t,x), \int_0^t \int_\Omega k(t,x,s,y) g(u(s,y))\, dy\, ds\right), \quad t > 0, \quad x \in \Omega.$$

Let us mention the special example

$$u_t - \Delta u = \mu(x)\left\{p \exp\left[\int_0^t \int_\Omega \beta(y) u(s,y)\, dy\, ds\right] - 1\right\}, \quad t > 0, \quad x \in \Omega, \tag{1.4}$$

which plays an important role in the theory of nuclear reactor dynamics (see the numerous references in [P1] for physical motivation). The blow up behavior was studied by Pao [P1], who proved the finite-time blow-up for large positive initial data when $\mu \equiv \text{Const.} > 0$, $\beta \geq 0$, $\beta \not\equiv 0$, and $p > 1$. His result was later improved by Guo and Su [GS], who proved the blow-up of all nonnegative solutions when $\mu \geq 0$, $\mu \not\equiv 0$, $\beta \geq 0$, $\beta \not\equiv 0$, and $p > 1$.

2. Main results. The purpose of this paper is to present new blow-up results, which, for each type of problems, significantly improve and/or generalize several of the above cited results on blow-up for nonlocal parabolic equations. In many of the equations under consideration, the nonlinearity involves both local and nonlocal terms. In particular, the competition between nonlocal source terms and local dissipative or convective terms will be rather extensively investigated, and we will determine some sharp critical exponents for blow-up or global existence. Let us illustrate our results by some typical examples of each type. (Further results will be stated in the following sections. Local existence-uniqueness and comparison theorems are given in the appendix.)

First consider the following class of nonlocal problems with integral term in space.

$$u_t - \Delta u = u^m(t,x) \|u(t,.)\|_k^p - a u^q(t,x) - b.\nabla u^r(t,x), \quad t > 0, \quad x \in \Omega, \tag{2.1}$$

$$u(t,x) = 0, \quad t > 0, \quad x \in \partial\Omega, \tag{2.2}$$

$$u(0,x) = u_0(x), \quad x \in \Omega. \tag{2.3}$$

We have the following theorem.

THEOREM A. *Assume that p, q, $r \geq 1$, $m = 0$ or $m \geq 1$, $1 \leq k \leq \infty$, $a \in \mathbb{R}$, and $b \in \mathbb{R}^N$. Let $\phi \in C(\overline{\Omega})$, with $\phi \geq 0$, $\phi \not\equiv 0$, $\phi_{|\partial\Omega} = 0$, $u_0 = \lambda\phi$, $\lambda > 0$, and let u (≥ 0) be the solution of* (2.1)–(2.3).

(i) *If $m + p > \max(q, r)$, then there exists $\Lambda(\phi) > 0$, such that u blows up in finite time in L^∞ norm if $\lambda > \Lambda(\phi)$.*

(ii) *If $m + p \leq q$ and $a > 0$ (with a large enough in case of equality), or if $m + p \leq r$ and $b \neq 0$ (with $|b|$ large enough in case of equality), then u is global and bounded.*

Theorem A extends the results of Deng, Kwang, and Levine [DKL] and Wang and Wang [WW], who proved the possibility of blow up, in the cases $m = q = 1$, $r = 2$, and $m = 0$, $b = 0$, $p = k$, respectively. The authors of [DKL] also proved that $p = 1$ is the critical blow-up exponent in the case $m = 1$, $r = 2$ (if $q = 1$). Theorem A shows that, more generally, the critical blow-up exponents are given by $r = m + p$ and $q = m + p$. Let us point out that the methods used in [DKL] and [WW] do not seem applicable to the present case. The argument in [DKL] is a comparison with a subsolution of separated form, $z(t)w(x)$, with $z(t)$ growing unbounded in finite time,

and one easily checks that $m \le 1$ is a necessary condition for a function of this form to be a subsolution. The authors of [WW] use an eigenfunction argument which does not apply for a more complicated nonlocal term such as in (2.1).

Next, we turn to nonlocal problems with moving localized source:

$$u_t - \Delta u = f\big(u(t, x_0(t))\big), \quad t > 0, \quad x \in \Omega. \tag{LST}$$

THEOREM B. *Assume that $x_0 : \mathbb{R}^+ \to \Omega$ is Hölder continuous, and let $f : \mathbb{R} \to \mathbb{R}$, locally Lipschitz, be such that*

$$\lim_{s\to\infty} f(s)/s = \infty, \quad f \text{ nondecreasing}, \quad \int^{\infty} \frac{ds}{f(s)} < \infty, \quad \text{and} \quad f(0) \ge 0. \tag{2.4}$$

Let $\phi \in C(\overline{\Omega})$, with $\phi \ge 0$, $\phi \not\equiv 0$, $\phi_{|\partial\Omega} = 0$, $u_0 = \lambda\phi$, $\lambda > 0$, and let u (≥ 0) be the solution of (LST), (2.2), (2.3). *Then there exists $\Lambda(\phi) > 0$, such that u blows up in finite time in L^∞ norm if $\lambda > \Lambda(\phi)$.*

Theorem B improves the result of Chadam, Peirce, and Yin [CPY] by allowing the localization point $x_0(t)$ to describe an arbitrary curve in the domain Ω, instead of being fixed. Moreover, the authors of [CPY] need the additional assumptions that f is convex, and that the initial data is large in the neighborhood of x_0, which we do not require. Here again, the previous argument—a comparison with the solution of the corresponding local, radially symmetric problem—does not seem to apply under the weakened assumptions of Theorem B, and some different approach is necessary.

We then consider the following nonlocal equations in time and in space and time:

$$u_t - \Delta u = \mu(x) \int_0^t u^p(s, x)\, ds - a u^q(t, x) \quad t > 0, \quad x \in \Omega, \tag{2.5}$$

and

$$u_t - \Delta u = \mu(x) \int_0^t \int_\Omega \beta(y) u^p(s, y)\, dy\, ds - a u^q(t, x), \quad t > 0, \quad x \in \Omega, \tag{2.6}$$

where $p, q \ge 1$, $a > 0$, μ is Hölder continuous in $\overline{\Omega}$, $\mu \ge 0$, $\mu \not\equiv 0$ and (in case of (2.6)) $\beta \in C(\overline{\Omega})$, $\beta \ge 0$, $\beta \not\equiv 0$. We obtain the following theorem.

THEOREM C. *Let $u_0 \in C^1(\overline{\Omega})$, with $u_0 \ge 0$, $u_0 \not\equiv 0$, $u_{0|\partial\Omega} = 0$, and let u (≥ 0) be the solution of* (2.5) *or* (2.6), *with conditions* (2.2), (2.3). *In case of* (2.6), *also assume that $\mu\beta \not\equiv 0$.*

(i) *If $p > q$, then u blows up in finite time in L^∞ norm.*

(ii) *If $p \le q$, then u is global and unbounded, that is,* $\limsup_{t\to\infty} |u(t)|_\infty = \infty$.

Theorem C improves the results of Bellout [Be], Kozhanov [Ko], and Souplet [So]. In these papers (which were concerned only with the case of (2.5)), blow-up was obtained only for sufficiently large initial data u_0 and under the strong restriction that the local dissipative term grow at most linearly. Here we prove that blow up occurs whenever $p > q$ and *for all nonnegative nontrivial* u_0. Moreover, $p = q$ is proved to be the critical blow up exponent, so that the result is optimal. Interestingly, no matter how large q is, the problems (2.5) (or (2.6)), (2.2), (2.3) admit no bounded positive solutions at all.

Concerning the nonlocal equation in space and time (1.4), with μ and β as above, we also obtain the following result.

THEOREM D. *Let $p \ge 1$ and assume that $\mu\beta \not\equiv 0$. Let $u_0 \in C^1(\overline{\Omega})$, with $u_0 \ge 0$, $u_0 \not\equiv 0$, $u_{0|\partial\Omega} = 0$, and let u (≥ 0) be the solution of* (1.4), (2.2), (2.3). *Then u blows up in finite time in L^∞ norm.*

Theorem D partially improves the results of Pao [P1] and Guo and Su [GS], who obtained the same conclusion for $p > 1$ (without assuming $\mu\beta \not\equiv 0$ in the case of [GS]). The case $p = 1$ is more difficult, in particular because the nonlinearity vanishes for $u = 0$, so that the equation admits the trivial solution. Moreover, the comparison arguments with a subsolution of separated form in [P1] and [GS] do not apply when $p = 1$.

Our results rely on three different methods. The first one is an extension of the method of self-similar subsolutions, used extensively in a recent work of Souplet and Weissler [SW1], in order to handle general parabolic equations with local, gradient-dependent nonlinearity. Since this is a comparison method, it applies under the restriction that the functional in the nonlocal reaction term be monotonically nondecreasing with respect to the unknown function, so that the comparison principle apply (see the appendix). Although this method can work also for nonlocal problems in time (see Theorem 5.4), the best results are obtained for nonlocal problems in space, either of integral type (Theorems A and 3.1) or of localized type (Theorem 4.1), yielding the optimal critical exponents.

For Theorem B, we use a comparison with a subsolution of separated form, along with some strong maximum principle arguments.

The third method is a variant of the eigenfunction method, introduced by Kaplan [Ka], combined with some new properties on systems of differential inequalities. This approach turns out to be especially powerful for nonlocal problems in time or in space and time (Theorems C, D, 5.1, and 5.2), also leading to critical exponents and to the proof of blow up of all nonnegative nontrivial solutions. Let us recall that the usual eigenfunction method introduces a single auxiliary function (namely, the first generalized Fourier coefficient of $u(t,.)$) and reduces the problem to proving global nonexistence for a differential inequality. The main difference in our approach is that we introduce *two* independent auxiliary functions (see formula (5.5)), leading to a system of two coupled differential inequalities, whose analysis is more delicate (Lemma 5.3). The classical approach with a single function, which was used in [Be, Kh, So], seems to enable one to obtain blow up only for large initial data.

Sections 3 and 4 treat nonlocal problems in space, respectively of integral and localized types. Section 5 is concerned with nonlocal problems in time or in space and time. The necessary local existence-uniqueness and comparison theorems are given in the appendix, where we construct in some detail a unified local theory for the general equation (NLRD).

3. Nonlocal reaction in space: Integral source terms. The blow up part of Theorem A is a consequence of the following more general result, concerning equations of the form

$$u_t - \Delta u = u^m(t,x)\|u(t,.)\|_k^p - F\big(u(t,x), \nabla u(t,x)\big), \quad t > 0, \quad x \in \Omega. \tag{3.1}$$

THEOREM 3.1. *Let p, $q \geq 1$, $m = 0$ or $m \geq 1$, $1 \leq k \leq \infty$, $a > 0$. Assume that $F : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ is locally Lipschitz continuous, with $F(0,0) \leq 0$, and satisfies*

$$F\big(u,U\big) \leq a\big(|u|^q + |u| + |U|^q + |U|\big), \quad u \in \mathbb{R}, \quad U \in \mathbb{R}^N, \tag{3.2}$$

and that

$$m + p > q.$$

Let $\phi \in C^1(\overline{\Omega})$, $\phi \geq 0$, $\phi \not\equiv 0$, $\phi_{|\partial\Omega} = 0$, and let u (≥ 0) be the solution of (3.1), (2.2), (2.3). *Then there exists $\Lambda(\phi) > 0$, such that u blows up in finite time in C^1 norm if $\lambda > \Lambda(\phi)$.*

Our arguments rely on a comparison with a *self-similar subsolution* which blows up in finite time. This method was used in a previous work [SW1] in order to solve the general problem of the critical blow up exponent for nonlinear parabolic equations with local gradient-dependent nonlinearity, of the form

$$u_t - \Delta u = F(u, \nabla u).$$

One of the main results of [SW1] was, in rough terms, that blow up is possible whenever F grows positively with respect to u faster than its negative growth with respect to ∇u. This principle can be illustrated on the simple model equation

$$u_t - \Delta u = u^p - \mu|\nabla u|^q$$

(with $\mu > 0$, in a bounded domain) for which (i) blow-up occurs for large initial data if $p > q$; (ii) all solutions are global and bounded if $p \leq q$. (See also [SW2], [STW], and the references therein, for related results.)

The method of self-similar subsolutions here confirms its ability to handle a wide variety of blow up problems (see also the case of spatially localized or time-integral source terms in the next sections). Let us point out that this method can be extended to handle degenerate parabolic problems (replacing the Laplacian in equation (2.1) with any nonnegative, semidefinite, possibly nonlinear, elliptic operator, e.g., the porous medium operator; see [SW1]). However, for simplicity, we have here restricted ourselves to the semilinear case.

Proof of Theorem 3.1. Since (3.1) does not a priori make sense for negative values of u, we actually consider the equation

$$Pu(t,x) \equiv u_t - \Delta u - u_+^m(t,x)\|u_+(t,.)\|_k^p + F\big(u(t,x), \nabla u(t,x)\big) = 0. \tag{3.3}$$

Existence-uniqueness of a maximal in time solution, with blow-up alternative in C^1 norm, and the validity of the comparison principle, then follow from Theorem A.12. In particular, since $F(0,0) \leq 0$, $u \geq 0$ as long as it exists.

By translation, one may assume without loss of generality that $0 \in \Omega$ and $\phi(0) > 0$. We seek a blowing up subsolution in the self-similar form

$$v(t,x) = \frac{1}{(T-t)^\gamma}\, V\left[\frac{|x|}{(T-t)^\sigma}\right],$$

with

$$V(y) = 1 + \frac{A}{2} - \frac{y^2}{2A}, \quad y \geq 0, \tag{3.4}$$

where γ, $\sigma > 0$, $A > 1$, and $0 < T < 1$ are to be determined. First note that, for T small enough,

$$\mathrm{Supp}(v_+(t,.)) = \overline{B}(0, R(T-t)^\sigma) \subset \overline{B}(0, RT^\sigma) \subset \Omega, \quad 0 \leq t < T, \tag{3.5}$$

with $R = \big(A(2+A)\big)^{1/2}$.

Next, we estimate the different terms from $Pv(t,x)$. We have

$$-\Delta v(t,x) = \frac{N/A}{(T-t)^{\gamma+2\sigma}}.$$

For all $(t,x)\in(0,T)\times\Omega$, we find

$$|v|(t,x)\le\frac{1+A+\mathrm{diam}^2(\Omega)}{(T-t)^{\gamma+2\sigma}}$$

and

$$|\nabla v|(t,x)\le\frac{\mathrm{diam}(\Omega)}{(T-t)^{\gamma+2\sigma}}.$$

The remaining terms are estimated in two different ways according to the size of $y=|x|/(T-t)^\sigma$. If $0\le y\le A$, we have $1\le V(y)\le 1+A/2$ and $V'(y)\le 0$, hence

$$v_t(t,x)=\frac{(\gamma V(y)+\sigma yV'(y))}{(T-t)^{\gamma+1}}\le\frac{\gamma(1+A)}{(T-t)^{\gamma+1}}.$$

In the case $k<\infty$, setting $M=\int_{B(0,R)}V^p(|\xi|)\,d\xi>0$ and $\alpha=p/k$, we have, after a change of variable and using (3.5),

$$\begin{aligned}v_+^m(t,x)\|v_+(t,.)\|_k^p&=\frac{V_+^m(y)}{(T-t)^{\gamma(m+p)}}\left(\int_{B(0,R(T-t)^\sigma)}V^k\left[\frac{|z|}{(T-t)^\sigma}\right]dz\right)^\alpha\\&=\frac{M^\alpha V_+^m(y)}{(T-t)^{\gamma(m+p)-N\sigma\alpha}}\\&\ge\frac{M^\alpha}{(T-t)^{\gamma(m+p)-N\sigma\alpha}}.\end{aligned}$$

(This formula remains valid for $k=\infty$ with $\alpha=0$.) Combining these estimates and using (3.2) yield, for all $(t,x)\in(0,T)\times\Omega$ such that $0\le y\le A$,

$$\begin{aligned}Pv(t,x)\le{}&\frac{\gamma(1+A)}{(T-t)^{\gamma+1}}+\frac{N}{(T-t)^{\gamma+2\sigma}}-\frac{M^\alpha}{(T-t)^{\gamma(m+p)-N\sigma\alpha}}\\&+a\frac{2\big[1+A+\mathrm{diam}^2(\Omega)\big]^q+2\,\mathrm{diam}^q(\Omega)+1}{(T-t)^{(\gamma+2\sigma)q}}.\end{aligned}\tag{3.6}$$

On the other hand, if $y\ge A$, we have $V(y)\le 1$ and $V'(y)\le -1$, so that

$$v_t(t,x)\le\frac{\gamma-\sigma A}{(T-t)^{\gamma+1}}.$$

Therefore, for all $(t,x)\in[0,T)\times\Omega$ such that $y\ge A$, we obtain

$$Pv(t,x)\le\frac{\gamma-\sigma A}{(T-t)^{\gamma+1}}+\frac{N}{(T-t)^{\gamma+2\sigma}}+a\frac{2\big[1+A+\mathrm{diam}^2(\Omega)\big]^q+2\,\mathrm{diam}^q(\Omega)+1}{(T-t)^{(\gamma+2\sigma)q}}.\tag{3.7}$$

Now the assumptions allow one to choose

$$\frac{1}{m+p-1}<\gamma<\frac{1}{q-1}\quad\big(\infty\text{ if }q=1\big).$$

Next, we may choose $\sigma>0$ so small that

$$\gamma(m+p)-N\sigma\alpha>\gamma+1>(\gamma+2\sigma)q,$$

and last we take

$$A > \max(1, \gamma/\sigma).$$

Then, for $T > 0$ sufficiently small, (3.6), (3.7) imply that $Pv \le 0$ in $(0,T) \times \Omega$.

Since $\phi(0) > 0$ and ϕ is continuous, there exist two real numbers ρ and $\epsilon > 0$, such that $\phi(x) \ge \epsilon$ for all $x \in B(0,\rho) \subset \Omega$. Taking smaller T if necessary, we have $B(0, RT^\sigma) \subset B(0,\rho) \subset \Omega$ and hence $v \le 0$ on $(0,T) \times \partial\Omega$, and from (3.5) it follows that $v(0,x) \le \lambda\phi(x)$ in $\overline{\Omega}$ for all sufficiently large λ. The comparison principle then implies that the solution u can exist no later than $t = T$, and the proof is complete. □

Remark 3.1. A careful reading of the above proof yields the following upper estimate on the blow up time:

$$T^*(\lambda\phi) \le C_\epsilon [\lambda |\phi|_\infty]^{-(m+p-1)+\epsilon} \quad \text{as } \lambda \to \infty \quad \text{for all } \epsilon > 0.$$

On the other hand, by comparing with a homogeneous supersolution obtained as a solution of the ODE $u' = u^{m+p}$, one obtains the lower bound

$$T^*(\lambda\phi) \ge C' \big(\lambda |\phi|_\infty\big)^{-(m+p-1)}, \quad \lambda > 0.$$

Proof of Theorem A. (i) The nonlinearity $f(u, \nabla u) = (b.\nabla u) r u^{r-1} = b.\nabla(u^r)$ is not locally Lipschitz continuous if $1 < r < 2$, so that the uniqueness theorem A.4 does not apply immediately. However, using the special conservative form of the gradient term, local existence and uniqueness of classical, nonnegative solutions of (2.1)–(2.3), with blow-up alternative in L^∞ norm, is proved in [DKL, p. 195] for all $\phi \in C(\overline{\Omega})$, $\phi \ge 0$, $\phi_{|\partial\Omega} = 0$, in dimension $N = 1$, and the arguments there extend without difficulty to any dimension $N \ge 1$. On the other hand, since $|(b.\nabla u) r u^{r-1}| \le C(|u|^r + |\nabla u|^r)$, finite-time blow-up for large $\lambda > 0$ is a consequence of Theorem 3.1 and of the maximum principle.

(ii) Let e_1 be the first unit vector and x_1 be the first coordinate. If $m + p \le r$, by rotation, we may assume without loss of generality that $a = |a| e_1$. A straightforward calculation then shows that $v(x) = Ce^{x_1}$ is a (bounded, stationary) supersolution for all sufficiently large $C > 0$ if $m + p \le q$ (with k large in case of equality), or if $m + p \le r$ (with $|a|$ large in case of equality). □

Remark 3.2. Under the assumptions of Theorem A, with $m + p > 1$ and $r > 1$, the solution of (2.1)–(2.3) is global and bounded if the initial data is small in L^∞ norm. Indeed, the function $v(x) = \epsilon(K - |x|^2)$ is a bounded stationary supersolution for $K = 1 + \sup_\Omega |x|^2$ and $\epsilon > 0$ sufficiently small.

4. Nonlocal reaction in space: Localized source terms. Before proving Theorem B, we state another result for a different equation, which is a localized analogue of equations (1.3) and (3.1) (studied in [DKL] and in Theorem A). The nonlocal (localized) term is involved in a product with a local one, and the source term may in addition be in competition with a local damping term. Namely, we consider the equation

$$u_t - \Delta u = u^m(t,x) u^p\big(t, x_0(t)\big) - \mu u^q(t,x), \quad t > 0, \quad x \in \Omega. \tag{4.1}$$

For this equation, we obtain a sharp critical blow-up exponent.

THEOREM 4.1. *Let p, $q \ge 1$, $m = 0$ or $m \ge 1$, $\mu > 0$, and assume that $x_0 : \mathbb{R}^+ \to \Omega$ is a function of class C^1. Let $\phi \in C(\overline{\Omega})$, $\phi \ge 0$, $\phi \not\equiv 0$, $\phi_{|\partial\Omega} = 0$, $u_0 = \lambda\phi$, $\lambda > 0$, and let u (≥ 0) be the solution of* (4.1), (2.2), (2.3).

(i) *If $m+p>q$ or $m+p=q>1$ and $\mu<1$, then there exists $\Lambda(\phi)>0$, such that u blows up in finite time in L^∞ norm if $\lambda>\Lambda(\phi)$.*

(ii) *If $m+p<q$ or $m+p=q$ and $\mu\geq 1$, then u is global and bounded.*

As in (1.3) and (3.1), a convective or dissipative gradient term may be added, which, for simplicity, we have not considered in the above statement (see Remark 4.3).

Proof of Theorem B. The existence and uniqueness of a maximal in time solution, with blow-up alternative in L^∞ norm, and the validity of the comparison principle follow from Example A.2 in section A.5 of the appendix. In particular, since $f(0)\geq 0$, $u\geq 0$ as long as it exists.

Let $T>1$, ϕ as in the hypotheses, $\lambda>1$, and assume that $T^*(\lambda\phi)>T$. Denote by v the solution of the linear heat equation

$$\begin{aligned} &v_t-\Delta v=0, \quad t>0, \quad x\in\Omega\\ &v(t,x)=0, \quad t>0, \quad x\in\partial\Omega\\ &v(0,x)=\phi(x), \quad x\in\overline{\Omega}. \end{aligned}$$

By the strong maximum principle, we have $V(x)=v(1,x)>0$ in Ω. Since x_0 is continuous into Ω, we have

$$d=\operatorname{dist}\big(x_0\langle[0,T]\rangle,\Omega^C\big)>0.$$

Since $K=\{x\in\Omega;\ \operatorname{dist}(x,\Omega^C)\geq d\}$ is compact, there exists a constant $\epsilon>0$, such that

$$V(x)\geq\epsilon, \quad x\in K. \tag{4.2}$$

On the other hand, since $f\geq 0$ on $\mathbb{R}^+$, the maximum principle implies that

$$u(1,x)\geq\lambda V(x), \quad x\in\Omega.$$

We now seek an unbounded subsolution of the form

$$w(t,x)=\lambda z(t)V(x), \quad 1\leq t<T_1<T, \quad x\in\Omega,$$

with $z\in C^1([1,T_1);\mathbb{R}^+)$ to be determined, such that $z(1)=1$ and z is nondecreasing. We have

$$\begin{aligned} Pw(t,x)&=w_t-\Delta w-f(w(t,x_0(t)))\\ &=\lambda z'(t)V(x)-\lambda z(t)\Delta V(x)-f\big[\lambda z(t)V(x_0(t))\big]. \end{aligned}$$

Using (4.2) and the fact that f is nondecreasing, we get $f\big[\lambda z(t)V(x_0(t))\big]\geq f(\lambda\epsilon z(t))$ on $[1,T_1)$. Let $A=\max(|V|_\infty,|\Delta V|_\infty)>0$. Using $(2.4)_1$, since $z(t)\geq 1$, we obtain

$$\begin{aligned} Pw(t,x)&\leq\lambda A\big(z'(t)+z(t)\big)-f\big[\lambda\epsilon z(t)\big]\\ &\leq\lambda Az'(t)-\frac{1}{2}f\big[\lambda\epsilon z(t)\big] \end{aligned}$$

if one takes λ sufficiently large. Now choose z to be the solution of the ODE

$$z'(t)=\frac{1}{2\lambda A}f\big[\lambda\epsilon z(t)\big], \quad t>1, \quad \text{with } z(1)=1.$$

Using $(2.4)_3$, we find that $\lim_{t\to T_1} z(t) = +\infty$, with

$$\begin{aligned} T_1 &= 1 + 2\lambda A \int_1^\infty \frac{z'(t)}{f\big[\lambda\epsilon z(t)\big]}\, dt \\ &= 1 + \frac{2A}{\epsilon}\int_{\lambda\epsilon}^\infty \frac{ds}{f(s)}, \end{aligned}$$

and $T_1 < T$ if λ is large enough. But clearly, w is a subsolution of (LRT), (2.2), (2.3), and the comparison principle implies that $u(t,x) \geq w(t,x)$ as long as u and w exist. Therefore u can exist no later than $t = T_1 < T$, which is a contradiction. The result follows. □

The method of proof of Theorem 4.1 is different from that of Theorem B. The blow up part relies on a method of self-similar subsolutions with moving center.

Proof of Theorem 4.1. Similarly as in the proof of Theorem 3.1, instead of (4.1), we actually consider the equation

$$Pu \equiv u_t - \Delta u - u_+^m(t,x)u_+^p(t,x_0(t)) + \mu u_+^q(t,x) = 0. \tag{4.3}$$

The properties of local solutions are then the same as in the proof of Theorem B.

(i) We construct a blowing up self-similar subsolution, whose support is centered at the point $x_0(t)$. Namely, we set

$$v(t,x) = \frac{C}{(T-t)^\gamma}\, V\left[\frac{|x - x_0(t)|}{(T-t)^\sigma}\right], \tag{4.4}$$

where V is given by (3.4), and $C > 0$, $A > 1$, and $0 < T < 1$ are to be determined. We set

$$d = \operatorname{dist}\big(x_0\langle[0,1]\rangle, \Omega^C\big) > 0$$

and

$$K = \sup_{t\in[0,1]} |x_0'(t)| < \infty.$$

We now set $y = |x - x_0(t)|/(T-t)^\sigma$, and we compute

$$\begin{aligned} Pv(t,x) =& \frac{C(\gamma V(y) + \sigma y V'(y))}{(T-t)^{\gamma+1}} + \frac{C(x - x_0(t)).x_0'(t)V'(y)}{(T-t)^{\gamma+\sigma}} \\ &+ \frac{NC}{(T-t)^{\gamma+2\sigma}} - \frac{C^{p+m}V_+^m(y)V_+^p(0)}{(T-t)^{\gamma(m+p)}} + \mu\frac{C^qV_+^q(y)}{(T-t)^{\gamma q}}. \end{aligned} \tag{4.5}$$

This time, we choose

$$\gamma = \frac{1}{m+p-1} < \frac{1}{q-1},$$

so that $\gamma + 1 = \gamma(m+p)$, and we take $0 < \sigma < 1/2$ and $A > \max(1, \gamma/\sigma)$.

First consider the case $m + p > q$; hence $\gamma q < \gamma + 1$. By taking C so large that $C^{p+m} > \gamma C(1+A)$, and then $T > 0$ suitably small, we get

$$Pv(t,x) \leq \frac{\gamma C(1+A)}{(T-t)^{\gamma+1}} + \frac{KC\,\mathrm{diam}^2(\Omega)}{(T-t)^{\gamma+\sigma}} + \frac{NC}{(T-t)^{\gamma+2\sigma}} - \frac{C^{p+m}}{(T-t)^{\gamma+1}} + \mu\frac{C^q(1+A)^q}{(T-t)^{\gamma q}} \leq 0$$

if $0 \le y \le A$ and

$$Pv(t,x) \le \frac{C(\gamma-\sigma A)}{(T-t)^{\gamma+1}} + \frac{KC\,\mathrm{diam}^2(\Omega)}{(T-t)^{\gamma+\sigma}} + \frac{NC}{(T-t)^{\gamma+2\sigma}} + \mu\frac{C^q(1+A)^q}{(T-t)^{\gamma q}} \le 0$$

if $y \ge A$.

Now suppose that $m+p=q$ and $\mu<1$. We rewrite the last two terms in (4.5) as

$$-\frac{C^{m+p}V_+^m(y)V_+^p(0)}{(T-t)^{\gamma(m+p)}} + \mu\frac{C^qV_+^q(y)}{(T-t)^{\gamma q}} = -\frac{C^{m+p}V_+^m(y)\big(1-\mu V_+^p(y)\big)}{(T-t)^{\gamma+1}},$$

so that

$$-\frac{C^{m+p}V_+^m(y)V_+^p(0)}{(T-t)^{\gamma(m+p)}} + \mu\frac{C^qV_+^q(y)}{(T-t)^{\gamma q}} \le \begin{cases} \dfrac{-(1-\mu)C^{m+p}}{(T-t)^{\gamma+1}}, & 0 \le y \le A, \\ 0, \quad y \ge A. \end{cases}$$

We then easily find again that

$$Pv(t,x) \le 0, \quad 0<t<T, \quad x \in \Omega.$$

On the other hand, thanks to the strong maximum principle ($q \ge 1$), by starting from some $t=t_0>0$, we may reduce to the case when $\phi(x)>0$ in Ω. In particular, we may assume that $\phi(x) \ge \epsilon > 0$ in some neighborhood of $x(0)$. The end of the proof is then identical to that of Theorem 3.1.

(ii) It suffices to notice that $v \equiv C > 0$ is a supersolution (with C large in the case $m+p<q$). □

Remark 4.1. Similarly as in Remark 3.1, one can derive the following estimate on the blow-up time:

$$c_1\big(\lambda|\phi|_\infty\big)^{-(p-1)} \le T^*(\lambda\phi) \le c_2\big(\lambda|\phi|_\infty\big)^{-(p-1)} \quad \text{as } \lambda \to \infty.$$

Remark 4.2. As for (2.1) (see Remark 3.2), the solution of (4.1), (2.2), (2.3) is global and bounded if the initial data is small and $m+p>1$.

Remark 4.3. The result of Theorem 4.1 (i) remains valid if a local term $-F(u,\nabla u)$ is added to the right-hand side of equation (4.1), where F satisfies the assumptions of Theorem 3.1 (the exponent r in (3.2) being replaced with some $q \in [1, m+p)$).

5. Nonlocal reaction terms in time and in space and time. The blow-up part in Theorem C will be a special case of the following results. Consider the parabolic inequalities

$$u_t - \Delta u \ge \mu(x)\left(\int_0^t u^p(s,x)\,ds\right)^\alpha - au^q(t,x), \quad t>0, \quad x \in \Omega, \tag{5.1}$$

and

$$u_t - \Delta u \ge \mu(x)\left(\int_0^t \int_\Omega \beta(y)u^p(s,y)\,dy\,ds\right)^\alpha - au^q(t,x), \quad t>0, \quad x \in \Omega, \tag{5.2}$$

where $p, \alpha \ge 1$, $q>0$, $a \ge 0$, $\mu \in C^\gamma(\overline{\Omega})$ for some $\gamma>0$, $\mu \ge 0$, $\mu \not\equiv 0$ and (in case of (5.2)) $\beta \in C(\overline{\Omega})$, $\beta \ge 0$, $\beta \not\equiv 0$.

THEOREM 5.1. *Assume that either*

$$\alpha = 1, \quad p > \max(q, 1) \tag{5.3}$$

or

$$\alpha > 1, \quad p \geq \max(q, 1). \tag{5.4}$$

Let $u \in C([0,T) \times \overline{\Omega}) \cap C^{1,2}((0,T) \times \overline{\Omega})$, $u \geq 0$, *satisfy* (5.1), *with* $u(0,x) \not\equiv 0$. *If* $0 < q < 1$, *also assume that either* $u(0,x) > 0$ *in* Ω *or* $\mu(x) \geq \mu_0 > 0$ *in* $\overline{\Omega}$. *Then* $T < \infty$.

THEOREM 5.2. *Assume that* $\mu\beta \not\equiv 0$ *and that either* (5.3) *or* (5.4) *holds. Let* $u \in C([0,T) \times \overline{\Omega}) \cap C^{1,2}((0,T) \times \overline{\Omega})$, $u \geq 0$, *satisfy* (5.2), *with* $u(0,x) \not\equiv 0$. *If* $0 < q < 1$, *also assume that either* $u(0,x) > 0$ *in* Ω *or* $\beta(x) \geq \beta_0 > 0$ *in* $\overline{\Omega}$. *Then* $T < \infty$.

Our argument is a variant of the eigenfunction technique, combined with some new properties on systems of differential inequalities. (Note that we do not require $q \geq 1$: since we do not rely on a comparison argument, no Lipschitz assumption is necessary. Theorems 5.1 and 5.2 give an a priori information on any solution, which does not suppose uniqueness of local solutions.)

Proof of Theorem 5.1. There exists some $\delta > 0$, such that $\mu(x) \geq \delta$ in some open subset $\omega \subset\subset \Omega$ (for instance a ball). Let $\lambda > 0$ be the first eigenvalue of $-\Delta$ in ω with Dirichlet boundary conditions, and ψ be the corresponding normalized positive eigenfunction, that is: $-\Delta\psi = \lambda\psi$ and $\psi > 0$ in ω, $\int_\omega \psi(x)\,dx = 1$, and $\psi = 0$ on $\partial\omega$. We also have $\partial\psi/\partial\nu \leq 0$ on $\partial\omega$ (with ν the outward normal of $\partial\omega$). Define the functions

(5.5)
$$y(t) = \int_\omega u(t,x)\psi(x)\,dx \quad \text{and} \quad z(t) = \int_0^t \int_\omega u^p(s,x)\psi(x)\,dx\,ds, \quad 0 \leq t < T.$$

Multiplying (5.1) by ψ, integrating over ω, and applying Green's formula, we get the following (denoting $' = d/dt$):

$$y' = \int_\omega u\Delta\psi\,dx + \int_{\partial\omega} \left(\psi\frac{\partial u}{\partial\nu} - \nu\frac{\partial\psi}{\partial\nu}\right) d\sigma + \int_\omega \left(\int_0^t u^p(s,x)\,ds\right)^\alpha \mu(x)\psi(x)\,dx - a\int_\omega u^q(t,x)\psi(x)\,dx$$

for all $0 < t < T$. Hence, by the properties of ψ,

$$y' + \lambda y \geq \delta \int_\omega \left(\int_0^t u^p(s,x)\,ds\right)^\alpha \psi(x)\,dx - a\int_\omega u^q(t,x)\psi(x)\,dx.$$

By Jensen's inequality ($\alpha \geq 1$) and Fubini's theorem, it follows that

$$y' + \lambda y \geq \delta z^\alpha - a\int_\omega u^q(t,x)\psi(x)\,dx.$$

Since $p \geq q$, letting $r = q/p \leq 1$, another application of Jensen's inequality yields the differential inequality

$$y' + \lambda y + az'^r \geq z^\alpha, \tag{5.6}$$

where the parameter $\delta > 0$ has been scaled out.

If $q \geq 1$, the strong maximum principle implies that $u(t,x) > 0$ for all $t \in (0,T)$, $x \in \Omega$, so that

$$z(t) > 0, \quad 0 < t < T. \tag{5.7}$$

If $0 < q < 1$ and $u_0(x) > 0$ in Ω, then (5.7) is obviously also valid. If $0 < q < 1$ and $\mu(x) \geq \mu_0$ in $\overline{\Omega}$, we replace ω with Ω and δ with μ_0 in the above argument, which yields inequality (5.6) again, and we also have (5.7) since $u(0,x) \not\equiv 0$.

Let us first handle the easier case $p = 1$ (hence $\alpha > 1$ and $q \leq 1$). We have $z' = y$, and z satisfies

$$z'' + \lambda z' + a z'^r \geq z^\alpha, \quad 0 < t < T.$$

Pick $t_0 \in (0,T)$. Since $r \leq 1$, there exists a constant $C > 0$, such that $a z'^r \leq C z' + (1/2) z^\alpha(t_0)$, so that

$$z'' + (C + \lambda) z' \geq (1/2) z^\alpha, \quad t_0 \leq t < T,$$

and [So, Theorem 1.1] enables one to conclude that $T < \infty$.

To treat the case $p > 1$, we first note that, by Jensen's inequality again,

$$z' \geq y^p.$$

The result then follows from the following lemma on systems of differential inequalities, which is of independent interest.

LEMMA 5.3. *Assume $p > 1$, $r > 0$, $a \in \mathbb{R}$, and*

$$\alpha = 1 > r \quad \text{or} \quad \alpha > 1 \geq r. \tag{5.8}$$

Let y, z be some functions in $C^1(0,T)$, with $y \geq 0$ and $z > 0$ on $(0,T)$, such that

$$\begin{cases} z' \geq y^p, \\ y' + \lambda y + a z'^r \geq z^\alpha, \end{cases} \quad 0 < t < T. \tag{5.9}$$

Then $T < \infty$.

Proof of Lemma 5.3. By translating the origin of time, we may assume that actually y, $z \in C^1([0,T))$ and $z(0) > 0$. Choose $\gamma = 1$ if $r = 1$, and $\max(r, 1/p) < \gamma < 1$ if $0 < r < 1$. It follows from (5.9) that, for all $\epsilon > 0$, there exists a constant $C_\epsilon > 0$, such that

$$C_\epsilon z'^\gamma \geq y^{p\gamma} + (3\lambda + 1) y - \epsilon \quad \text{and} \quad C_\epsilon z'^\gamma \geq 3 a z'^r - \epsilon;$$

hence

$$2 C_\epsilon z'^\gamma + 3 y' \geq 3 \Big[\lambda y + a z'^r + y' \Big] + y^{p\gamma} + y - 2\epsilon,$$

so that

$$2 C_\epsilon z'^\gamma + 3 y' \geq 3 z^\alpha + y^{p\gamma} + y - 2\epsilon. \tag{5.10}$$

We then consider two cases.

Case 1. $r = 1$ (hence $\gamma = 1$ and $\alpha > 1$). Choosing $\epsilon < z^\alpha(0)$ and setting $\nu = \min(\alpha, p) > 1$, it follows that

$$(2C_\epsilon + 3)\big[z + y\big]' \geq z^\alpha + y^p + y \geq \big[z(0)\big]^{\alpha-\nu} z^\nu + y^\nu, \quad 0 \leq t < T,$$

where we have used the fact that z is nondecreasing. Using the inequality

$$a^\nu + b^\nu \geq C(\nu)(a+b)^\nu, \quad a, b \geq 0, \tag{5.11}$$

it then follows that

$$\big[z + y\big]' \geq C[z + y]^\nu, \quad 0 \leq t < T,$$

for some $C > 0$, so that $T < \infty$.

Case 2. $r < 1$ (hence $0 < \gamma < 1$ and $\alpha \geq 1$). Picking $m \in (0, \gamma)$, by Young's inequality, we have, for some large constant $C''_\epsilon > 3$,

$$2C_\epsilon z'^\gamma = 2C_\epsilon \frac{z'^\gamma}{z^m} z^m \leq \epsilon z^{m/(1-\gamma)} + C'_\epsilon \frac{z'}{z^{m/\gamma}} = \epsilon z^{m/(1-\gamma)} + C''_\epsilon \big[z^{1-(m/\gamma)}\big]'.$$

One may assume m to be so small that $m/(1-\gamma) < 1$ and put $\theta = 1 - m/\gamma \in (0,1)$. By substituting into (5.10), with $\epsilon < 1$, we get

$$C''_\epsilon \big[z^\theta + y\big]' \geq 3z^\alpha + y^{p\gamma} + y - 2\epsilon - \epsilon z^{m/(1-\gamma)} \geq 2z^\alpha + y^{p\gamma} + y - 3\epsilon, \quad 0 \leq t < T.$$

Choosing $\epsilon < z^\alpha(0)/3$ and setting $\nu = \min(p\gamma, 1/\theta) > 1$, it follows that

$$C''_\epsilon \big[z^\theta + y\big]' \geq z^\alpha + y^{p\gamma} + y \geq \big[z(0)\big]^{\alpha-\theta\nu} z^{\theta\nu} + y^\nu, \quad 0 \leq t < T,$$

where we have used the fact that z is nondecreasing. Using (5.11), it then follows that

$$\big[z^\theta + y\big]' \geq C[z^\theta + y]^\nu, \quad 0 \leq t < T,$$

for some $C > 0$, so that $T < \infty$. □

Remark 5.1. The assumption (5.8) in Lemma 5.3 is essential (at least if $a > 0$). Indeed, if $\alpha = r > 0$, then $z(t) = Ce^t$, $y(t) = C^p e^{pt}$ is a global positive solution of (5.9) if $C > 0$ is large.

Proof of Theorem 5.2. The proof is almost identical to that of Theorem 5.1, up to the following changes. Since $\mu\beta \not\equiv 0$, we may assume $\beta(x) \geq \delta$ in ω. Proceeding as before, it follows that

$$y' + \lambda y \geq \int_\omega \mu(x)\psi(x)\,dx \left(\int_0^t \int_\Omega \beta(y) u^p(s,y)\,dy\,ds\right)^\alpha - a\int_\omega u^q(t,x)\psi(x)\,dx.$$

We then obtain the differential inequality (5.6), where δ is replaced with

$$\delta' = \big(\delta/|\psi|_\infty\big)^\alpha \int_\omega \mu(x)\psi(x)\,dx > 0,$$

and the hypotheses again imply that $z(t) > 0$ for $t \in (0,T)$. The rest then is unchanged. □

Proof of Theorem C. By replacing u with u_+ in the right-hand side of equations (5.1) and (5.2), the existence and uniqueness of a maximal in time solution, with

blow-up alternative in L^∞ norm, and the validity of the comparison principle, follow from Theorem A.13. In particular, $u \geq 0$ as long as it exists.

(i) This is a special case of Theorems 5.1 and 5.2.

(ii) If $p < q$, a simple calculation shows that $v(t,x) = C(1+t)^{1/(q-p)}$ is a supersolution for all large $C > 0$. If $p = q$, the same holds with $v(t,x) = Ce^{Ct}$. Taking $C > |u_0|_\infty$, it follows from the comparison principle that u must exist globally.

Last, if $q \geq p \geq 1$, assume for contradiction that u is globally bounded by a constant $M > 0$. Then u satisfies

$$u_t - \Delta u \geq \mu(x) \int_0^t u^p(s,x)\, ds - aM^{q-1}u(t,x), \quad t > 0, \quad x \in \Omega$$

(in the case of (5.1), and an analogous inequality in the case of (5.2)). If $p > 1$, this immediately implies finite-time blow-up by Theorems 5.1 and 5.2: a contradiction. In the case $p = 1$, by arguing as in the beginning of the proof of Theorem 5.1, one is reduced to the differential inequality

$$z'' + 2kz' \geq z, \quad t > 0, \quad \text{with } z(0) = 0 \text{ and } z' > 0,$$

for some $k > 0$. By setting $w(t) = z(t)\exp(kt)$, we see that

$$w'' \geq (1+k^2)w, \quad t > 0, \quad \text{with } w(0) = 0 \text{ and } w' > 0,$$

so that $w' \geq (1+k^2)^{1/2}w$. It follows that $z(t) \geq C\exp[((1+k^2)^{1/2} - k)t]$, for all $t \geq 1$ and some $C > 0$, and we obtain again a contradiction. We conclude that u must be global and unbounded if $q \geq p \geq 1$. □

Proof of Theorem D. Since $e^z - 1 \geq z^2/2$, $z \geq 0$, the result follows from Theorem 5.2 (with $p = 1$, $\alpha = 2$, $a = 0$ in (5.2)), in view of the maximum principle. □

Some blow up results can also be stated for the nonlocal in time analogue of (3.1) (or (1.3)) and (4.1):

$$u_t - \Delta u = u^m(t,x)\|u(.,x)\|^p_{L^k(0,t)} - F\big(u(t,x), \nabla u(t,x)\big), \quad t > 0, \quad x \in \Omega, \tag{5.12}$$

where $\|u(.,x)\|_{L^k(0,t)} = \big(\int_0^t |u|^k(s,x)\, ds\big)^{1/k}$.

The structure of the nonlocal term now does not allow the application of the previous eigenfunction technique. However, the method of self-similar subsolutions, suitably adapted, can still be used. Although the result is probably not optimal, we give it also as an illustration of this method, which seems to be the only applicable one here (if $m > 1$). We mention that similar results can be obtained along the same lines for analogous nonlocal nonlinearities in space and time.

THEOREM 5.4. *Let p, $q \geq 1$, $m = 0$ or $m \geq 1$, $1 \leq k \leq \infty$, $a > 0$. Assume that $F : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ is locally Lipschitz continuous, with $F(0,0) \leq 0$, and satisfies*

$$F\big(u,U\big) \leq a\big(|u|^q + |u| + |U|^q + |U|\big) \tag{5.13}$$

and that

$$m + p > q + \frac{p}{k}(q-1). \tag{5.14}$$

Assume that $\phi \in C^{2+\alpha}(\overline{\Omega})$, $\phi \geq 0$, $\phi \not\equiv 0$ satisfies $\phi_{|\partial\Omega} = 0$ and $\Delta\phi + F(0, \nabla\phi) = 0$ on $\partial\Omega$. Let $u_0 = \lambda\phi$, $\lambda > 0$, and let u (≥ 0) be the solution of (5.12), (2.2), (2.3). *Then there exists $\Lambda(\phi) > 0$, such that u blows up in finite time in C^1 norm if $\lambda > \Lambda(\phi)$.*

Proof of Theorem 5.4. Again, we modify (5.12) into

$$Pu(t,x) \equiv u_t - \Delta u - u_+^m(t,x)\|u_+(.,x)\|_{L^k(0,t)}^p + F\big(u(t,x), \nabla u(t,x)\big) = 0.$$

The existence and uniqueness of a maximal in time solution, with blow-up alternative in C^1 norm, and the validity of the comparison principle follow from Theorems A.1, A.2, A.3, and A.13 (ii).

Assume, without loss of generality, that $B(0,\rho) \subset \Omega$, and that $\phi(x) \geq \epsilon > 0$ in $B(0,\rho)$. The blowing up subsolution is sought under the form

$$v(t,x) = \frac{1}{(T-t)^\gamma}\, V\left[\frac{|x|}{(T-t)^\sigma}\right] + w(t,x), \tag{5.15}$$

with

$$w(t,x) = \frac{M}{T^\gamma}\left(3 - \frac{3|x|^2}{\rho^2} - 2\frac{t}{T}\right),$$

where V is given by (3.4), and A, $M > 1$, γ, $\sigma > 0$, and $0 < T < 1$ are to be fixed later.

Since the source term involves an integral over $[0,t]$, its effect for small t is not strong enough to compensate the action of the dissipative terms at the early stage of the time evolution. This is the reason for the adjunction of the polynomial part $w(t,x)$, in addition to the self-similar one. This additional term is designed to provide a negative contribution under the action of the linear part of the parabolic operator P. Namely, assuming $T < \rho^2/6N$, we have

$$w_t - \Delta w = \frac{M}{T^{\gamma+1}}(-2 + 6NT/\rho^2) \leq -\frac{M}{T^{\gamma+1}}.$$

On the other hand, to estimate the gradient term, we compute, for $0 < t < T$,

$$\nabla v(t,x) = \frac{-x}{A(T-t)^{\gamma+2\sigma}} - \frac{6Mx}{\rho^2 T^\gamma},$$

hence, since $T < 1$ and A, $M > 1$,

$$|\nabla v|(t,x) \leq \frac{\mathrm{diam}(\Omega)(1 + 6M/\rho^2)}{(T-t)^{\gamma+2\sigma}} \leq \frac{CM}{(T-t)^{\gamma+2\sigma}}.$$

(Here and in the rest of the proof, C denotes a generic constant which depends only on Ω and ρ.)

Setting $y = |x|/(T-t)^\sigma$, we have to distinguish three cases.

• First case: $0 \leq t \leq T/2$. We find

$$|v|(t,x) \leq \frac{C + A}{(T-t)^{\gamma+2\sigma}} + \frac{CM}{T^\gamma} \quad \text{and} \quad v_t(t,x) \leq \frac{\gamma(1+A)}{(T-t)^{\gamma+1}}.$$

Assuming that

$$\sigma < 1/2 \quad \text{and} \quad \gamma + 1 > (\gamma + 2\sigma)q, \tag{5.16}$$

letting

$$\delta = \gamma + 1 - (\gamma + 2\sigma)q > 0,$$

and using (5.13), we obtain, after some computations similar to those in the proof of Theorem 3.1, that

$$Pv(t,x) \le \frac{-M + 2^{\gamma+1}\big[\gamma(1+A) + N + 4a(CM+A)^q T^\delta\big]}{T^{\gamma+1}}. \tag{5.17}$$

• Second case: $T/2 \le t < T$ and $y \ge A$. We have

$$|v|(t,x) \le \frac{C}{(T-t)^{\gamma+2\sigma}} + \frac{CM}{T^\gamma} \quad \text{and} \quad v_t(t,x) \le \frac{\gamma - \sigma A}{(T-t)^{\gamma+1}}.$$

Thus we find

$$Pv(t,x) \le \frac{\gamma - \sigma A + N + 4a(CM)^q T^\delta}{(T-t)^{\gamma+1}}. \tag{5.18}$$

(Note that in cases 1 and 2, we did not use the growth of the nonlocal term, only its nonnegativeness.)

• Third case: $T/2 \le t < T$ and $0 \le y \le A$. Assuming that $T^\sigma A < \rho/2$, hence $|x| \le (T-t)^\sigma A \le \rho/2$, it follows that

$$\frac{1}{(T-s)^\gamma} \le v(s,x) \le \frac{3M}{T^\gamma} + \frac{1+A}{(T-s)^\gamma}, \quad 0 \le s \le t.$$

Denote by Bv the nonlocal term in Pv. In the case $k < \infty$, assuming that $\gamma k \ne 1$, and setting $\alpha = p/k$, we then have

$$\begin{aligned} Bv(t,x) &\ge \frac{1}{(T-t)^{\gamma m}} \left(\int_0^t \frac{ds}{(T-s)^{\gamma k}} \right)^\alpha = \frac{1}{(T-t)^{\gamma m}} \left(\frac{(T-t)^{1-\gamma k} - T^{1-\gamma k}}{\gamma k - 1} \right)^\alpha \\ &\ge (T-t)^{(1-\gamma k)\alpha} \left(\frac{1 - 2^{1-\gamma k}}{\gamma k - 1} \right)^\alpha \frac{1}{(T-t)^{\gamma m}} \\ &\ge \frac{C(\gamma,p,k)}{(T-t)^{\gamma(m+p)-\alpha}} \end{aligned}$$

for some $C(\gamma,p,k) > 0$, where we have used the fact that $T > 2(T-t)$ (consider separately the cases $\gamma k > 1$ and $\gamma k < 1$). This formula remains valid for $k = \infty$ with $\alpha = 0$. It follows that

$$Pv(t,x) \le \frac{\gamma(1+A) + N + 4a(CM+A)^q}{(T-t)^{\gamma+1}} - \frac{C(\gamma,p,k)}{(T-t)^{\gamma(m+p)-\alpha}}. \tag{5.19}$$

By the hypothesis (5.14), it is possible to choose $\gamma \ne 1/k$, such that

$$\frac{\alpha+1}{m+p-1} < \gamma < \frac{1}{q-1},$$

so that in particular

$$\gamma(m+p) - \alpha > \gamma + 1,$$

and next to take $\sigma > 0$ so small that (5.16) holds. Now, by choosing

$$A > (\gamma + N)/\sigma,$$

$$M > 2^{\gamma+1}\big(\gamma(1+A)+N\big),$$

and then T sufficiently close to 0, it follows from formulas (5.17), (5.18), and (5.19), that $Pv(t,x) \le 0$ in the three cases, that is, for all $t \in [0,T)$, $x \in \Omega$.

Taking still smaller T if necessary, we have $v(t,x) \le 0$ for x outside of $B(0,\rho)$, hence on $[0,T) \times \partial\Omega$, and $\lambda\phi(x) \ge \lambda\epsilon \ge v(0,x)$ in Ω for all sufficiently large λ. The comparison principle then implies that the solution u can exist no later than $t = T$, and the Theorem is proved. □

Appendix A. Local theory for general nonlocal semilinear parabolic equations. The local solvability for various classes of functional parabolic equations has been studied in many articles. See, e.g., [A], [BB], [CaY] for nonlocal problems in space. For nonlocal problems in time, see, e.g., [Y1], [Y2], and the references therein. Also, fully nonlinear equations with integral terms have been investigated (see, e.g., [LS], [CP], [Sf] for time-integrals and [CY] for space-integrals), and semilinear parabolic equations with constant delay have been treated in [TW], both subject matters which are out of the scope of the present article.

However, as far as we know, there does not seem to be in the literature a local theory that would apply to the nonlocal semilinear parabolic equation (NLRD) in general form. We wish to construct such a local theory. In the same time, since many (though not all) of the proofs of our blow up and global existence results rely on comparison arguments, we need a version of the comparison principle adapted to the case of nonlocal problems. Such results are proved in many of the cited works, for the specific types of nonlocal equations they consider. We here prove a general comparison principle for the equation (NLRD) in abstract form, which recovers most of these specific results. (Since we consider nonlocal problems, some care is needed. In addition to the usual Lipschitz condition, we have to require a certain assumption of nondecreasing monotonicity on the nonlocal term; see section A.4 and Example A.5 in section A.5.)

Our results enable one to handle in a same formalism all of the four types of problems described in the introduction (and actually much more general ones). On the other hand, we want to allow the nonlinearity to depend on u and ∇u as well. For these reasons, the formalism developed in this Appendix needs to be a bit abstract, although the arguments of the proofs will be rather standard, relying on Schauder a priori estimates and fixed point theorem. Let us point out that what is going to be said remains valid, as usual, if the Laplacian in (NLRD) is replaced with a strongly elliptic operator with sufficiently smooth coefficients. Moreover, other kinds of (linear) boundary conditions may be treated similarly, under suitable assumptions.

The outline of the appendix is as follows. Section A.1 contains some preliminaries and notations for the precise formulation of general nonlocal problems. The abstract existence, uniqueness, and continuation results are stated in section A.2 and proved in section A.3. Section A.4 is devoted to the comparison principle. Finally, in section A.5, the abstract results are illustrated by examples which include, in particular, the specific cases considered in the main text of the article.

A.1. Preliminaries and notations. Let Ω be a bounded domain in $\mathbb{R}^N$. In what follows, we fix some $\alpha \in (0,1)$. We will assume that

(A1) $\partial\Omega$ is of class $C^{2+\alpha}$

or, in some cases, that

(A1)′ $\partial\Omega$ is of class $C^{3+\alpha}$.

We set $Q_T = (0,T]\times\Omega$, $S_T = (0,T]\times\partial\Omega$, and $Q_{t_0,T} = (t_0,T]\times\Omega$, $0 \le t_0 < T < \infty$.

Let us set some notations concerning the needed function spaces. By $C(\overline{\Omega})$ (resp., $C^1(\overline{\Omega})$), we denote the space of real valued continuous (resp., continuously differentiable) functions on $\overline{\Omega}$, endowed with the norm

$$|\phi|_{L^\infty(\Omega)} = \sup_{x\in\overline{\Omega}} |\phi(x)| \quad \left(\text{resp.,} \quad |\phi|_{C^1(\overline{\Omega})} = |\phi|_{L^\infty(\Omega)} + \sum_{i=1}^N |\partial_{x_i}\phi|_{L^\infty(\Omega)}\right).$$

For each $T > 0$, $C\big(\overline{Q_T}\big)$ is the space of real valued continuous functions on $\overline{Q_T}$, endowed with the norm

$$|u|_{L^\infty(Q_T)} = \sup_{(t,x)\in\overline{Q_T}} |u(t,x)|.$$

$C^{0,1}\big(\overline{Q_T}\big)$ is the space of functions u of $\overline{Q_T}$, such that $\partial_{x_i}u$, $i = 1,\dots,N$, belong to $C(\overline{Q_T})$, endowed with the norm

$$|u|_{C^{0,1}\left(\overline{Q_T}\right)} = |u|_{L^\infty(Q_T)} + \sum_{i=1}^N |\partial_{x_i}u|_{L^\infty(Q_T)}.$$

$C^{1,2}\big(Q_T\big)$ is the space of functions u on Q_T, such that u, $\partial_t u$, $\partial_{x_i}u$, $\partial^2_{x_ix_j}u$, $i,j = 1,\dots,N$, are continuous in Q_T.

In addition, we will use some Hölder spaces of functions of t and x, whose definition we recall for convenience. $C^\alpha(\overline{Q_{t_0,T}})$ is the space of functions u which are uniformly Hölder continuous on $\overline{Q_{t_0,T}}$, with exponents $\alpha/2$ in t and α in x, endowed with the norm

$$|u|_{C^\alpha(\overline{Q_{t_0,T}})} = |u|_{L^\infty(Q_{t_0,T})} + \sup_{(t,x)\ne(s,y)\in Q_{t_0,T}} \frac{|u(t,x)-u(s,y)|}{(|t-s|+|x-y|^2)^{\alpha/2}}.$$

$C^{1+\alpha}(\overline{Q_{t_0,T}})$ (resp., $C^{2+\alpha}(\overline{Q_{t_0,T}})$) is the space of those u such that u, $\partial_{x_i}u$, $i = 1,\dots,N$ (resp., u, $\partial_t u$, $\partial_{x_i}u$, $\partial^2_{x_ix_j}u$, $i,j = 1,\dots,N$) belong to $C^\alpha(\overline{Q_{t_0,T}})$, the norms being defined by the sum of the C^α norms of u and of the corresponding derivatives.

Next we define the nonlocal operators making up the equation. For all $t \ge 0$, we define the past time restriction operator R^t by setting, for all $T \ge t$ and $u \in C(\overline{Q_T})$,

$$R^t u = u_{\big|[0,t]\times\overline{\Omega}} \in C(\overline{Q_t}).$$

We take a collection of functionals $\{F^t\}_{t\ge0}$, with $F^t : C(\overline{Q_t}) \to C(\overline{\Omega})$, or $F^t : C^{0,1}(\overline{Q_t}) \to C(\overline{\Omega})$. (By convention, we identify $C(\overline{Q_0})$ with $C(\overline{\Omega})$ and $C^{0,1}(\overline{Q_0})$ with $C^1(\overline{\Omega})$.) We then consider a general nonlocal problem in the form

$$\text{(A2)} \qquad \begin{cases} u_t - \Delta u = F^t\big(R^t u\big)(x), & (t,x)\in Q_T,\\ u(t,x) = 0, & (t,x)\in S_T,\\ u(0,x) = u_0(x), & x\in\overline{\Omega}. \end{cases}$$

For convenience, when T is implicitly understood, we will use the operator $F : u \mapsto Fu$, where

$$Fu(t,x) = F^t(R^t u)(x).$$

A.2. Local existence, uniqueness, and continuation results. We first give a result of local existence of classical solutions up to $t = 0$ for smooth initial data (with first order compatibility conditions). Let $\gamma > 0$ be fixed. We will assume that, for all $T > 0$, F possesses the following properties:

(A3) $$\forall u \in C^{1+\gamma}(\overline{Q_T}), Fu \in C^{\alpha}(\overline{Q_T});$$

(A4) $$F \text{ is continuous from } C^{1+\gamma}(\overline{Q_T}) \text{ to } C(\overline{Q_T}).$$

THEOREM A.1. *Assume* (A1), (A3), (A4), *and let* $u_0 \in C^{2+\alpha}(\overline{\Omega})$ *satisfy the compatiblity conditions*

(A5) $$u_0|_{\partial\Omega} = 0 \quad \textit{and} \quad \Delta u_0 + F^0 u_0 = 0 \textit{ on } \partial\Omega.$$

Then there exists a (nonnecessarily unique) maximal in time function u, *defined on* $[0, T^*) \times \overline{\Omega}$ *for some* $T^* \in (0, \infty]$, *such that for all* $T \in (0, T^*)$, $u \in C^{2+\alpha}(\overline{Q_T})$ *and* u *is a (classical) solution of* (A2).

We next give a continuation result, which states that any nonglobal solution must blow up in finite time in C^1 norm. Instead of (A3), we will assume the stronger condition that, for all $T > 0$,

(A6) $$F \text{ is bounded on bounded sets from } C^{1+\gamma}(\overline{Q_T}) \text{ to } C^{\alpha}(\overline{Q_T}),$$

and in addition to (A4), we will suppose that, for all $T > 0$,

(A7) $$F \text{ is bounded on bounded sets from } C^{0,1}(\overline{Q_T}) \text{ to } C(\overline{Q_T}).$$

On the other hand, as in classical problems with local nonlinearities, if one is concerned with nonlocal problems where the nonlinearity "does not depend on the (spatial) gradient of u," it is possible to prove blow up in L^∞ norm instead of C^1 norm. To express the fact that the nonlocal functionals F^t do not depend on the gradient of u, we consider, instead of (A7), the stronger assumption

(A7)′ $$F \text{ is bounded on bounded sets from } C(\overline{Q_T}) \text{ to } C(\overline{Q_T})$$

for all $T > 0$.

THEOREM A.2. *Assume* (A1), (A4), (A6), *and* (A7) (*resp.*, (A7)′), *and let* u *be any solution of* (A2) *as in Theorem* A.1. *If* $T^* < \infty$, *then* u *blows up in finite time in* C^1 *norm* (*resp., in* L^∞ *norm*) *in the sense that*

(A8) $$\limsup_{t \to T^*} |u(t)|_{C^1(\overline{\Omega})} = \infty \quad (\textit{resp.}, \limsup_{t \to T^*} |u(t)|_{L^\infty(\Omega)} = \infty).$$

Now, assume one of the following Lipschitz conditions, for all $T > 0$:

(A9) $$F \text{ is Lipschitz continuous on bounded sets from } C^{0,1}(\overline{Q_T}) \text{ to } C(\overline{Q_T})$$

or, alternatively,

(A9)′ $$F \text{ is Lipschitz continuous on bounded sets from } C(\overline{Q_T}) \text{ to } C(\overline{Q_T}).$$

Then one obtains a local uniqueness result.

THEOREM A.3. *Assume* (A1), *and* (A9) (*resp.*, (A9)′). *Then, for all* $T > 0$, *there exists at most one (classical) solution of* (A2) *in the class* $C^{0,1}(\overline{Q_T}) \cap C^{1,2}(Q_T)$ (*resp., in the class* $C(\overline{Q_T}) \cap C^{1,2}(Q_T)$).

Next, we wish to prove the local solvability of (A2) for less smooth initial data. For this purpose, we need reinforced assumptions on F. Namely, we may consider C^1 data (with zero order compatibility conditions) if we assume that, for all $T > 0$,

(A10) F is continuous from $C^{0,1}(\overline{Q_T})$ to $C(\overline{Q_T})$ and bounded on bounded sets,

(A11) for all sequence $(u_n) \in C^{0,1}(\overline{Q_T})$ such that the sequence $|u_n|_{C^{1+\gamma}(\overline{Q_{\epsilon,T}})}$ is bounded for all $\epsilon \in (0,T)$, then the sequence $|Fu_n|_{C^{\alpha}(\overline{Q_{\epsilon,T}})}$ is bounded for all $\epsilon \in (0,T)$.

In the same way, we may consider merely continuous data, under the analogous "gradient-independent" assumptions, i.e.,

(A10)′ F is continuous from $C(\overline{Q_T})$ to $C(\overline{Q_T})$ and bounded on bounded sets,

(A11)′ for all sequence $(u_n) \in C(\overline{Q_T})$ such that the sequence $|u_n|_{C^{\gamma}(\overline{Q_{\epsilon,T}})}$ is bounded for all $\epsilon \in (0,T)$, then the sequence $|Fu_n|_{C^{\alpha}(\overline{Q_{\epsilon,T}})}$ is bounded for all $\epsilon \in (0,T)$.

THEOREM A.4. *Assume* (A1)′, (A10), (A11), *and let*

$$u_0 \in C^1(\overline{\Omega}), \qquad \text{with } u_0{}_{|\partial\Omega} = 0.$$

(i) *Then there exists a (nonnecessarily unique) maximal in time function* u, *defined on* $[0,T^*) \times \overline{\Omega}$ *for some* $T^* \in (0,\infty]$, *such that for all* $0 < \epsilon < T < T^*$, $u \in C^{0,1}(\overline{Q_T}) \cap C^{2+\alpha}(\overline{Q_{\epsilon,T}})$, *and* u *is a (classical) solution of* (A2). (ii) *If* $T^* < \infty$, *then* u *blows up in finite time in* C^1 *norm in the sense of* (A8).

THEOREM A.4′. *Assume* (A1), (A10)′, (A11)′, *and let*

$$u_0 \in C(\overline{\Omega}), \qquad \text{with } u_0{}_{|\partial\Omega} = 0.$$

(i) *Then there exists a (nonnecessarily unique) maximal in time function* u, *defined on* $[0,T^*) \times \overline{\Omega}$ *for some* $T^* \in (0,\infty]$, *such that for all* $0 < \epsilon < T < T^*$, $u \in C(\overline{Q_T}) \cap C^{2+\alpha}(\overline{Q_{\epsilon,T}})$, *and* u *is a (classical) solution of* (A2).

(ii) *If* $T^* < \infty$, *then* u *blows up in finite time in* L^∞ *norm in the sense of* (A8).

The assumptions (A11) or (A11)′ hold if the nonlinearity does not depend on the past (nonlocal problems in space) or, more generally, if it does not involve the past up to $t = 0$, e.g., for $F^t(R^t u)(x) = \int_{t/2}^t f(u(s,x))\,ds$. On the contrary, if the nonlinearity involves the whole past, for instance for $F^t(R^t u)(x) = \int_0^t f(u(s,x))\,ds$, then these assumptions will not be satisfied a priori. However, thanks to the regularizing effect of the time integral, for this type of nonlocal terms, with no dependence on the gradient, it is still possible to prove local existence for C^1 data and to get a blow up alternative in L^∞ norm. The kind of general assumption satisfied by such F is the following:

(A12) F is continuous from $C^{0,1}(\overline{Q_T})$ to $C^{\alpha}(\overline{Q_T})$ and bounded on bounded sets

for all $T > 0$. We then have the following theorem.

THEOREM A.5. *Assume* (A1)′, (A12), *and let*

$$u_0 \in C^1(\overline{\Omega}), \quad \text{with } u_0|_{\partial\Omega} = 0.$$

(i) *Then there exists a (nonnecessarily unique) maximal in time function* u, *defined on* $[0, T^*) \times \overline{\Omega}$ *for some* $T^* \in (0, \infty]$, *such that for all* $0 < \epsilon < T < T^*$, $u \in C^{0,1}(\overline{Q_T}) \cap C^{2+\alpha}(\overline{Q_{\epsilon,T}})$, *and* u *is a (classical) solution of* (A2).

(ii) *If* $T^* < \infty$, *then* u *blows up in finite time in* C^1 *norm in the sense of* (A8). *If, in addition* (A10)′ *holds, then the blow up occurs in* L^∞ *norm.*

In the special case when the nonlocal operator F does not depend on the past, that is, if F^t is of the form

$$F^t(R^t u) \equiv f(t, u(t, .)), \quad \text{with } f : \mathbb{R}^+ \times C(\overline{\Omega}) \to C(\overline{\Omega}), \tag{A13}$$

and under a Lipschitz condition, then the blow up alternative can be made more precise, with the "lim sup" becoming a limit. (See Remark A.1 at the end of section A.3 for a variant without Lipschitz condition.)

PROPOSITION A.6. *Let* $\{F^t\}_{t\geq 0}$ *be of the form* (A13), *and assume* (A1), (A3), *and* (A9) (*resp.,* (A9)′). *Assume that* $u \in C^{0,1}(\overline{Q_T}) \cap C^{2+\alpha}(\overline{Q_{\epsilon,T}})$ (*resp.,* $u \in C(\overline{Q_T}) \cap C^{2+\alpha}(\overline{Q_{\epsilon,T}})$), $0 < \epsilon < T < T^*$, *is a maximal, classical solution of* (A2) (*which is necessarily unique by Theorem* A.3). *If* $T^* < \infty$, *then*

$$\lim_{t\to T^*} |u(t)|_{C^1(\overline{\Omega})} = \infty \quad (\textit{resp.,} \lim_{t\to T^*} |u(t)|_{L^\infty(\Omega)} = \infty). \tag{A14}$$

A.3. Proofs of the local results. We first state two preliminary results on the heat equation that we will use repeatedly. The first one is the $C^{1+\delta}$ estimate for the heat equation (see [Fr, Theorem 4, p. 191, and formula (3.23), p. 200]).

LEMMA A.7. *Assume* (A1) *and let* $W \in C(\overline{Q_T}) \cap C^{1,2}(Q_T)$. *Let* $g \in C(\overline{Q_T})$ *satisfy* $g(0, x) = 0$ *on* $\partial\Omega$, *and assume that*

$$\begin{aligned} W_t - \Delta W &= g(t, x), \quad (t, x) \in Q_T, \\ W(t, x) &= 0, \quad (t, x) \in S_T, \\ W(0, x) &= 0, \quad x \in \overline{\Omega}. \end{aligned}$$

Then

$$|W|_{C^{1+\delta}(\overline{Q_T})} \leq K T^\sigma |g|_{L^\infty(Q_T)}$$

for all $0 < \delta < 1$, *where* $\sigma > 0$ *depends only on* δ, *and* K *depends only on* Ω *and* δ *for bounded* T.

The second one gives an estimate on the heat semigroup in C^1 spaces and requires higher ($C^{3+\alpha}$) regularity on $\partial\Omega$ (see [M, Theorem 2.3, p. 39, and formula (2.39), p. 40] and also [Bel]).

LEMMA A.8. *Assume* (A1)′ *and denote by* $e^{t\Delta}$ *the heat semigroup on* $C_0(\overline{\Omega})$. *Then, for all functions* ϕ *in* $C^1(\overline{\Omega})$ *which vanish on* $\partial\Omega$, *we have*

$$|e^{t\Delta}\phi|_{C^{0,1}(\overline{Q_T})} \leq C_1 |\phi|_{C^1(\overline{\Omega})},$$

where C_1 *depends only on* Ω *for bounded* T.

Proof of Theorem A.1. It can be proved in much the same way as in [Fr, Theorem 8, p. 204 and Theorem 10, p. 206]. However, the result there is proved for local nonlinearities. Therefore, we give the proof for sake of completeness. Let $0 < T < 1$ and $M > 0$ (to be fixed later), and consider the set C_M of functions $v \in C^{1+\gamma}(\overline{Q_T})$, satisfying

$$|v|_{C^{1+\gamma}(\overline{Q_T})} \le M, \quad v(0,.) = u_0, \quad v_{|(0,T]\times\partial\Omega} = 0.$$

We define a transformation Z on C_M as follows. $w = Zv$ is the unique solution of the linear problem

$$\begin{aligned} w_t - \Delta w &= Fv, \quad (t,x) \in Q_T, \\ w(t,x) &= 0, \quad (t,x) \in S_T, \\ w(0,x) &= u_0(x), \quad x \in \overline{\Omega}. \end{aligned}$$

(Using (A3), the existence and uniqueness of w follows from [Fr, Theorem 7, p. 65], and $w \in C^{2+\alpha}(\overline{Q_T})$.) We prove that Z admits a fixed point, if $T < 1$ is sufficiently small, by employing the Schauder fixed point theorem (see, e.g., [Fr, Theorem 2, p. 189]).

We first claim that Z maps C_M into itself if T is small. Thanks to (A5), we may take $W(t,x) \equiv w(t,x) - u_0(x)$, hence $g = Fv + \Delta u_0$, in Lemma A.7, which yields

$$|w|_{C^{1+\delta}(\overline{Q_T})} \le |u_0|_{C^{1+\delta}(\overline{\Omega})} + KT^\sigma\big(k_M + |\Delta u_0|_{L^\infty(Q_T)}\big), \tag{A15}$$

where $k_M = \sup\big\{|Fv|_{L^\infty(Q_1)};\ |v|_{C^{1+\gamma}(\overline{Q_1})} \le M\big\}$, which is finite by (A4). Now, choosing

$$M = 1 + |u_0|_{C^{1+\gamma}(\overline{\Omega})}, \quad T < \max\big[1, \big[K\big(k_M + |\Delta u_0|_{L^\infty(Q_T)}\big)\big]^{-1/\sigma}\big],$$

with $\delta = \gamma$ in formula (A15), it follows that $w \in C_M$.

Applying formula (A15) for some $\delta \in (\gamma, 1)$, by the compactness of the injection $C^{1+\delta}(\overline{Q_T}) \subset C^{1+\gamma}(\overline{Q_T})$, it follows that $Z(C_M)$ is a compact subset of C_M.

To check the continuity of Z, assume that v, $v_m \in C_M$ are such that $v_m \to v$ in $C^{1+\gamma}(\overline{Q_T})$, as m goes to ∞. Then $z_m = Zv_m - Zv$ satisfies

$$z_{mt} - \Delta z_m = Fv_m - Fv, \quad (t,x) \in Q_T,$$

with null initial and boundary conditions. Thus, by Lemma A.7 and assumption (A4), we obtain

$$|z_m|_{C^{1+\gamma}(\overline{Q_T})} \le KT^\sigma |Fv_m - Fv|_{L^\infty(Q_T)} \to 0, \text{ as } m \to \infty,$$

that is, $Zv_m \to Zv$ in $C^{1+\gamma}(\overline{Q_T})$.

Since C_M is a closed convex set of the Banach space $C^{1+\gamma}(\overline{Q_T})$, by the previous properties and Schauder's fixed point theorem, there exists a fixed point u of the map Z, which is a solution of the problem (A2).

Finally, by Zorn's lemma, follows as usual the existence of a solution (still denoted by u), which extends u and is maximal in time (i.e., which cannot be extended to a solution on a larger time interval). ∎

Proof of Theorem A.2. (i) Under hypothesis (A7), assume that, for some $T_0 \in (0,\infty)$ and all $T \in (0,T_0)$, $u \in C^{2+\alpha}(\overline{Q_T})$, u is a solution of (A2), and u is such that

$$\sup_{0\le t<T_0} |u(t)|_{C^1(\overline{\Omega})} < \infty.$$

(A7) then implies

$$|Fu|_{L^\infty(Q_T)} \le M_1, \quad 0 < T < T_0.$$

By Lemma A.7, with $W(t,x) = u(t,x) - u_0(x)$ and $g = Fu + \Delta u_0$, it follows that

$$|u|_{C^{1+\gamma}(\overline{Q_T})} \le |u_0|_{C^{1+\gamma}(\overline{\Omega})} + KT^\sigma\big(M_1 + |\Delta u_0|_{L^\infty(\Omega)}\big) = M_2, \quad 0 < T < T_0.$$

From (A6), we infer that

$$|Fu|_{C^\alpha(\overline{Q_T})} \le M_3, \quad 0 < T < T_0.$$

By the $C^{2+\alpha}$ parabolic estimate [Fr, Theorem 6 p. 65], we get

$$|u|_{C^{2+\alpha}(\overline{Q_T})} \le K'\big(M_3 + |u_0|_{C^{2+\alpha}(\overline{\Omega})}\big) = M_4, \quad 0 < T < T_0.$$

(The constant K' depends only on Ω and α for bounded T; see [Fr, p. 123].) From this estimate, it is easily seen that u can be (uniquely) extended to a function $\tilde{u} \in C^{2+\alpha}(\overline{Q_{T_0}})$. To prove that u may be extended to a solution of (A2) for some $T > T_0$, it then suffices to show the existence for small τ of a solution w of the problem

$$\begin{aligned} w_t - \Delta w = G^t\big(R^t w\big)(x), &\quad (t,x) \in Q_\tau, \\ w(t,x) = 0, &\quad (t,x) \in S_\tau, \\ w(0,x) = \tilde{u}(T_0,x), &\quad x \in \overline{\Omega}, \end{aligned}$$

where G^t is defined as

$$\forall z \in C^{0,1}(\overline{Q_t}),\ G^t(z) = F^{T_0+t}(z_1), \quad \text{with } z_1(s,.) = \begin{cases} u(s,.),\ 0 \le s < T_0, \\ z(s-T_0,.),\ T_0 \le s \le T_0 + t. \end{cases}$$

The existence of such w follows from Theorem A.1, since, as is readily verified, the functionals G^t satisfy the assumptions (A3) and (A4) for all z such that $z(0,.) = \tilde{u}(T_0)$, and since $\tilde{u}(T_0)$ satisfies the corresponding compatibility conditions. From the above, we deduce that $(\mathrm{A8})_1$ holds whenever $T^*(u) < \infty$.

(ii) Under assumption (A7)′, the proof of $(\mathrm{A8})_2$ is exactly similar. □

Proof of Theorem A.3. We just need to prove the result under assumption (A9), the other case being similar. Let $T > 0$ and assume that u and v are two solutions of (A2), with $u,\ v \in C^{0,1}(\overline{Q_T}) \cap C^{1,2}(Q_T)$. Set $h = T/p$ with p a positive integer to be chosen, and $D_n = \overline{Q_{nh,(n+1)h}}$ for $n = 0, \cdots, p-1$. By Lemma A.7 with $W = u - v$ (noting that $g(0,.) = F^0u(0,.) - F^0v(0,.) = 0$), and using (A9), we obtain

$$\begin{aligned} |W|_{C^{0,1}(D_n)} &\le Kh^\sigma\big(|Fu - Fv|_{L^\infty(D_n)} + |\Delta W(nh,.)|_{L^\infty(\Omega)}\big) \\ &\le Kh^\sigma\big(L_{M,T}|W|_{C^{0,1}(\overline{Q_{(n+1)h}})} + |\Delta W(nh,.)|_{L^\infty(\Omega)}\big). \end{aligned}$$

Choosing p so large that $Kh^\sigma L_{M,T} < 1$, since $W(0,.) = 0$, it follows easily by induction that $|W|_{C^{0,1}(D_{p-1})} = 0$, that is $u = v$ on $\overline{Q_T}$. □

Proof of Theorem A.4. (i) Assume $u_0 \in C^1(\overline{\Omega})$, $u_0{}_{|\partial\Omega} = 0$. From [M, Theorem 3.1, p. 49], $\partial\Omega$ being of class $C^{3+\alpha}$, we know that there exists a sequence $u_0^n \in C^{2+\alpha}(\overline{\Omega})$, with $u_0{}_{|\partial\Omega} = 0$ and $\Delta u_0{}_{|\partial\Omega} = 0$, such that $u_0^n \to u_0$ in $C^1(\overline{\Omega})$, as $n \to \infty$. Let $\xi \in C^\infty(\mathbb{R})$, with $\xi(t) = 0$, $t \le 1/2$, $\xi(t) = 1$, $t \ge 1$, and $0 \le \xi \le 1$. Let $\xi_n(t) = \xi(nt)$, and consider the approximating problems

$$\text{(A16)} \qquad \begin{cases} u_t^n - \Delta u^n = \xi_n(t) F^t\big(R^t u^n\big)(x) \equiv F_n u^n, \quad (t,x) \in Q_T, \\ u^n(t,x) = 0, \quad (t,x) \in S_T, \\ u^n(0,x) = u_0^n(x), \quad x \in \overline{\Omega}. \end{cases}$$

The assumptions (A10) and (A11) imply that the map $u^n \mapsto F_n u^n$ satisfies (A3) and (A4). Since the compatibility conditions (A5) are verified by u_0^n, that is,

$$u_0^n{}_{|\partial\Omega} = 0 \quad \text{ and } \quad -\Delta u_0^n(x) + F_n u^n(0,x) = 0 \quad \text{ on } \partial\Omega,$$

by Theorem A.1, there exists a (maximal) solution u^n of (A16), of existence time $T_n^* \in (0,\infty]$, with $u^n \in C^{2+\alpha}(\overline{Q_T})$ for all $0 < T < T_n^*$. We then have the following lemma.

LEMMA A.9. *With the notation above, there exists $T_0 \in (0,1]$ and $C_0 > 0$ such that*

$$T_n^* > T_0 \quad \text{ and } \quad |u^n|_{C^{0,1}(\overline{Q_{T_0}})} \le C_0$$

for all n.

Proof of Lemma A.9. Let $M = \sup_{n\ge1} |u_0^n|_{C^1(\overline{\Omega})} < \infty$. For each n and for each $T \in (0, T_n^*)$, $T \le 1$, we may write u^n under the form $u^n = e^{t\Delta} u_0^n + w^n$, where $w^n \in C^{2+\alpha}(\overline{Q_T})$ is the solution of the linear problem

$$\text{(A17)} \qquad w_t^n - \Delta w^n = F_n u^n, \quad (t,x) \in Q_T,$$

with null initial and boundary conditions. Since $\partial\Omega$ is of class $C^{3+\alpha}$, from Lemma A.8 it follows that

$$\text{(A18)} \qquad |e^{t\Delta} u_0^n|_{C^{0,1}(\overline{Q_T})} \le C_1 |u_0^n|_{C^1(\overline{\Omega})} \le C_1 M.$$

Set

$$k'_M = \sup\big\{ |Fv|_{L^\infty(Q_1)};\ |v|_{C^{0,1}(\overline{Q_1})} \le (1+C_1)M \big\},$$

which is finite by (A10). (By an easy extension argument, it can be seen that k'_M stands also for an upper bound of the sets $\big\{ |Fv|_{L^\infty(Q_\tau)};\ |v|_{C^{0,1}(\overline{Q_\tau})} \le (1+C_1)M \big\}$, for all $\tau \in (0,1)$.) Suppose that $T_n^* \le 1$, or that $T_n^* > 1$ and $|u^n|_{C^{0,1}(\overline{Q_1})} > (1+C_1)M$. Then one can set $T'_n = \min\big\{ t \in (0, T_n^*);\ |u^n(t,.)|_{C^1(\overline{\Omega})} \ge (1+C_1)M \big\}$, and we have $0 < T'_n < T_n^*$, $T'_n < 1$. On the other hand, by Lemma A.7, we get

$$|w^n|_{C^{0,1}(\overline{Q_{T'_n}})} \le |w^n|_{C^{1+\gamma}(\overline{Q_{T'_n}})} \le K {T'_n}^{\sigma} k'_M;$$

hence, by (A18),

$$(1+C_1)M = |v^n|_{C^{0,1}(\overline{Q_{T'_n}})} \le C_1 M + K {T'_n}^{\sigma} k'_M,$$

so that

$$T_n' \geq \big(M/(Kk_M')\big)^{1/\sigma}.$$

The lemma follows, with $T_0 = \min\big[1, \big(M/(Kk_M')\big)^{1/\sigma}\big]$ and $C_0 = (1+C_1)M$. □

Proof of Theorem A.4 (i) (continued). From Lemma A.9 and (A10), we deduce the uniform bound

$$|F_n u^n|_{L^\infty(\overline{Q_{T_0}})} \leq M_1. \tag{A19}$$

But (A17) and Lemma A.7 then imply that $|w^n|_{C^{1+\delta}(\overline{Q_{T_0}})} \leq KM_1T_0^\sigma$ so that, by Ascoli–Arzelà's theorem, (some subsequence of) w^n converges in $C^{0,1}(\overline{Q_{T_0}})$ to some $w \in C^{0,1}(\overline{Q_{T_0}})$. Moreover, by Lemma A.8 applied to $u_0^n - u_0$, it follows that $e^{t\Delta}u_0^n$ converges to $e^{t\Delta}u_0$ in $C^{0,1}(\overline{Q_{T_0}})$; hence

$$u^n \to u = e^{t\Delta}u_0 + w \quad \text{ in } C^{0,1}(\overline{Q_{T_0}}) \quad \text{as } n \to \infty.$$

By assumption (A10), this implies that $Fu^n \to Fu$ in $C(\overline{Q_{T_0}})$, and in particular

$$F_n u^n \to Fu \quad \text{everywhere in } (0,T_0] \times \overline{\Omega}. \tag{A20}$$

On the other hand, picking $\epsilon \in (0,T_0)$, the function $z^n = \xi(t/\epsilon)u^n$ now solves

$$z_t^n - \Delta z^n = \xi(t/\epsilon)F_n u^n + (1/\epsilon)\xi'(t/\epsilon)u^n \equiv h_n, \quad (t,x) \in Q_{T_0},$$

with null initial and boundary conditions, so that, applying Lemma A.9, (A19), and Lemma A.7 again, one obtains

$$|u^n|_{C^{1+\delta}(\overline{Q_{\epsilon,T_0}})} \leq |z^n|_{C^{1+\delta}(\overline{Q_{T_0}})} \leq C(\epsilon).$$

By assumption (A11), we deduce the bound

$$|F_n u^n|_{C^\alpha(\overline{Q_{\epsilon,T_0}})} \leq C_1(\epsilon), \quad 0 < \epsilon < T_0,$$

and then, by the $C^{2+\alpha}$–parabolic estimate (see [Fr, Theorem 6, p. 65]) and the definition of ξ,

$$\begin{aligned} |u^n|_{C^{2+\alpha}(\overline{Q_{\epsilon,T_0}})} &\leq |z^n|_{C^{2+\alpha}(\overline{Q_{T_0}})} \leq C|h_n|_{C^\alpha(\overline{Q_{T_0}})} \\ &\leq C|F_n u^n|_{C^\alpha(\overline{Q_{\epsilon/2,T_0}})} + C_2(\epsilon)|u^n|_{C^\alpha(\overline{Q_{\epsilon/2,T_0}})} \leq C_3(\epsilon). \end{aligned} \tag{A21}$$

By Ascoli–Arzelà's theorem and a diagonal procedure, (some subsequence of) u^n must converge to v in $C^{1,2}(\overline{Q_{\epsilon,T_0}})$ for all $0 < \epsilon < T_0$, which, along with (A20), implies that

$$u_t - \Delta u = Fu, \quad (t,x) \in Q_{T_0}.$$

Moreover, from the bound (A21), it follows that $u \in C^{2+\alpha}(\overline{Q_{\epsilon,T_0}})$ for all $0 < \epsilon < T_0$. The existence of a maximal solution follows as in Theorem A.1.

(ii) Assume that, for some $T_0 \in (0,\infty)$ and all $0 < \epsilon < T < T_0$, $u \in C^{2+\alpha}(\overline{Q_{\epsilon,T}})$, $u \in C^{0,1}(\overline{Q_T})$, u is a solution of (A2), and u is such that

$$\sup_{0 \leq t < T_0} |u(t)|_{C^1(\overline{\Omega})} < \infty.$$

(A10) then implies

$$|Fu|_{L^\infty(Q_T)} \le M_1, \quad 0 < T < T_0.$$

By Lemma A.7, with $W(t,x) = u(t,x) - u(\epsilon,x)$ and $g = Fu + \Delta u(\epsilon,.)$, it follows that

$$|u|_{C^{1+\gamma}(\overline{Q_{\epsilon,T}})} \le |u(\epsilon,.)|_{C^{1+\gamma}(\overline{\Omega})} + KT^\sigma\big(M_1 + |\Delta u(\epsilon,.)|_{L^\infty(\Omega)}\big) = M_2(\epsilon), \quad 0 < \epsilon < T < T_0.$$

From (A11), we infer that

$$|Fu|_{C^\alpha(\overline{Q_{T_0/2,T}})} \le M_2, \quad T_0/2 < T < T_0.$$

By the $C^{2+\alpha}$ parabolic estimate, we get

$$|u|_{C^{2+\alpha}(\overline{Q_{T_0/2,T}})} \le K'\big(M_3 + |u(T_0/2,.)|_{C^{2+\alpha}(\overline{\Omega})}\big) = M_4, \quad T_0/2 < T < T_0,$$

so that u can be (uniquely) extended to a function $\tilde{u} \in C^{2+\alpha}(\overline{Q_{T_0/2,T_0}})$. The end of the proof is then identical to that of Theorem A.2 (i) (the fact that the functionals G^t satisfy the assumptions (A3) and (A4) for all z such that $z(0,.) = \tilde{u}(T_0)$ being now a consequence of (A10) and (A11)). □

Proof of Theorem A.4′. (i) The result can be proved along the lines of Theorem A.4, up to some natural changes. In particular, we need only approximate u_0 by a sequence $u_0^n \in C^{2+\alpha}(\overline{\Omega})$ in L^∞ norm. From [M, Theorem 3.1 p. 49], this can be done, assuming only $\partial\Omega$ of class $C^{2+\alpha}$. On the other hand, instead of Lemma A.8, we simply use the estimate

$$|e^{t\Delta}u_0^n|_{C(\overline{Q_T})} \le |u_0^n|_{C(\overline{\Omega})},$$

which is a consequence of the maximum principle (and hence does not require the $C^{3+\alpha}$ regularity).

(ii) The proof is exactly similar to that of Theorem A.4 (ii). □

Proof of Theorem A.5. The proof is similar to that of Theorem A.4. The main difference is that (A12) implies that $F_n u^n \to Fu$ in $C^\alpha(\overline{Q_T})$, instead of (A20). One can then obtain (A21) directly. □

Proof of Proposition A.6. We just need to prove the result under assumption (A9), the other case being similar. Suppose that (A14) is false. Then there exists $M > 0$ such that for all $\epsilon > 0$, $|u(t_1,.)|_{C^1(\overline{\Omega})} \le M$, for some $t_1 \in (T^* - \epsilon, T^*)$. Taking t_1 as a new origin of time, since $u(t_1,.) \in C^{2+\alpha}(\overline{\Omega})$ and satisfies the compatiblity conditions (A5), there exists a local solution v with initial data $u(t_1,.)$ by Theorem A.1. But since F^t is independent of the past (that is, of the form (A13)), it is easily seen from the proof of Lemma A.9 that the existence time of v is bounded from below by a positive constant which depends only on M. Moreover, the solution obtained by extending u by v after $t = t_1$ has the same regularity as u. This contradicts the maximality of u and the local uniqueness property of Theorem A.3, if one chooses ϵ small enough. □

Remark A.1. In the "gradient-independent" case, Proposition A.6 remains valid without Lipschitz assumption, supposing only (A4), (A7)′ instead of (A9)′. This can be proved in a similar way as in [B, Theorem 3.1, p. 477, and Remark, p. 478], by estimating $|u(t+\epsilon,.)|_{L^\infty(\Omega)}$ from $|u(t,.)|_{L^\infty(\Omega)}$ for small $\epsilon > 0$. In particular, $(A14)_2$ holds even if local uniqueness is not assured.

However, we do not know whether the analogue is true in the gradient-dependent case (replacing (A9) by (A4), (A7)). Indeed, $|u(t+\epsilon,.)|_{C^1(\overline{\Omega})}$ can be estimated from $|u(t,.)|_{C^1(\overline{\Omega})}$ via the C_0–C^1 estimates of the heat semigroup (see [M]), but this requires that the nonlocal nonlinear term $f(t,u(t,.))$ vanish on $\partial\Omega$, which has no reason to be satisfied, especially if f is nonlocal in space and depends on the gradient of u.

A.4. Comparison principle. We present a general version of the comparison principle, which is valid for a large class of problems of the form (A2). The nonlinearity may depend on the (local) values of $u(t,x)$ and $\nabla u(t,x)$ in any (locally Lipschitz) way, and may functionally depend on $R^t u = u(s,y)$, $0 \le s \le t$, $y \in \overline{\Omega}$, in a "Lipschitz-monotone" way, with respect to the L^∞ norm. To be more precise, we consider the operator

$$Pu \equiv Pu(t,x) = u_t - \Delta u - F^t\big(R^t u\big)(x), \quad (t,x) \in \overline{Q_T}, \tag{A22}$$

where F^t is of the form

$$F^t\big(R^t u\big)(x) = G^t\big(u(t,x), \nabla u(t,x), R^t u\big)(x), \quad G^t : \mathbb{R} \times \mathbb{R}^N \times C(\overline{Q_t}) \to C(\overline{\Omega}), \quad t \ge 0,$$

and we assume that, for all $M, T > 0$, there exists $L_{M,T} > 0$ such that for all p, $\overline{p} \in \mathbb{R}$, q, $\overline{q} \in \mathbb{R}^N$, $u \in C(\overline{Q_T})$, with $|p|$, $|\overline{p}|$, $|q|$, $|\overline{q}|$, $|u|_{L^\infty(Q_T)} \le M$,

$$\big|G^t(p,q,R^t u) - G^t\big(\overline{p},\overline{q},R^t u\big)\big|_{L^\infty(\Omega)} \le L_{M,T}\big(|p-\overline{p}| + |q-\overline{q}|\big), \tag{A23}$$

and for all $p \in \mathbb{R}$, $q \in \mathbb{R}^N$, u, $\overline{u} \in C(\overline{Q_T})$, with $|p|$, $|q|$, $|u|_{L^\infty(Q_T)}$, $|\overline{u}|_{L^\infty(Q_T)} \le M$,

$$\big|\big(G^t\big(p,q,R^t u\big) - G^t\big(p,q,R^t \overline{u}\big)\big)_+\big|_{L^\infty(\Omega)} \le L_{M,T}|(u-\overline{u})_+|_{L^\infty(Q_T)}, \quad 0 \le t \le T. \tag{A24}$$

THEOREM A.10. *Let P be defined by* (A22) *and assume that* (A23)–(A24) *are fulfilled. Let u, $v \in C(\overline{Q_T}) \cap C^{1,2}(Q_T)$ satisfy*

$$\begin{aligned} Pu(t,x) &\le Pv(t,x), \quad (t,x) \in Q_T, \\ u(t,x) &\le v(t,x), \quad (t,x) \in S_T \cup \big(\{0\} \times \overline{\Omega}\big). \end{aligned}$$

Then $u \le v$ in $\overline{Q_T}$.

Proof. Let $M = \max(|u|_{L^\infty(Q_T)}, |v|_{L^\infty(Q_T)})$. By (A23)–(A24), the function $w = u - v$ satisfies

$$\begin{aligned} w_t - \Delta w &= F^t\big(u(t,x), \nabla u(t,x), R^t u\big)(x) - F^t\big(v(t,x), \nabla v(t,x), R^t v\big)(x) \\ &= \Big(F^t\big(u(t,x), \nabla u(t,x), R^t u\big)(x) - F^t\big(v(t,x), \nabla v(t,x), R^t u\big)(x)\Big) \\ &\qquad + \Big(F^t\big(v(t,x), \nabla v(t,x), R^t u\big)(x) - F^t\big(v(t,x), \nabla v(t,x), R^t (v)\big)(x)\Big) \\ &\le L_{M,T}\Big(|w(t,x)| + |\nabla w(t,x)| + \sup_{\overline{Q_T}} w\Big). \end{aligned}$$

Take $\alpha > 2L_{M,T}$. The function $z(t,x) = w(t,x)e^{-\alpha t}$ satisfies

$$z_t - \Delta z \le L_{M,T}\Big(|z(t,x)| + |\nabla z(t,x)| + \sup_{\overline{Q_T}} z\Big) - \alpha z(t,x).$$

It follows that z cannot attain a positive maximum at a point $(\tau, c) \in Q_T$, since otherwise

$$0 \le z_t(\tau,c) - \Delta z(\tau,c) \le (2L_{M,T} - \alpha) z(\tau,c) < 0,$$

which is a contradiction. The result follows. □

We will see in section A.5 that the assumption (A24) is verified for many nonlocal nonlinearities induced by integral norms, in space, time or space-time as well. This will be essentially a consequence of the following lemma.

LEMMA A.11. *Let Ω be a bounded open set of $\mathbb{R}^N$, $1 \le k \le \infty$, and ϕ, $\psi \in L^\infty(\Omega)$. Then*

$$|\phi_+|_k - |\psi_+|_k \le C(k, |\Omega|)\ |(\phi - \psi)_+|_\infty. \tag{A25}$$

Proof of Lemma A.11. Since $\phi_+ - \psi_+ \le (\phi - \psi)_+$, we may assume ϕ and $\psi \ge 0$ a.e., without loss of generality. We may also restrict ourselves to $1 < k < \infty$, the other cases being trivial. We start from the elementary inequality

$$x^\alpha - y^\alpha \le C(\alpha)(x+y)^{\alpha-1}(x-y), \quad x,\ y > 0, \quad \alpha > 0. \tag{A26}$$

(By homogeneity, one may assume $y = 1$, and (A26) then follows from the fact that the function $x \mapsto (x^\alpha - 1)[(x+1)^{\alpha-1}(x-1)]^{-1}$ is continuous positive on $[0,\infty) \setminus \{1\}$ and has finite, positive limits at $x = 1$ and ∞.) Thus, applying (A26) for $\alpha = 1/k$ and $\alpha = k$ and Hölder's inequality yields

$$\begin{aligned}
|\phi|_k - |\psi|_k &\le C(1/k) \left(\int \phi^k(x)\,dx + \int \psi^k(x)\,dx \right)^{(1/k)-1} \int [\phi^k - \psi^k](x)\,dx \\
&\le C(1/k)C(k) \left(\int \phi^k(x)\,dx + \int \psi^k(x)\,dx \right)^{(1/k)-1} \\
&\qquad \times \int (\phi+\psi)^{k-1}(x)\,dx\ |(\phi-\psi)_+|_\infty \\
&\le C(1/k)C(k) \left(\int \phi^k(x)\,dx + \int \psi^k(x)\,dx \right)^{-\frac{k-1}{k}} |\Omega|^{1/k} \\
&\qquad \times \left(\int (\phi+\psi)^k(x)\,dx \right)^{\frac{k-1}{k}} |(\phi-\psi)_+|_\infty.
\end{aligned}$$

Then, by the inequality $(x+y)^k \le 2^{k-1}(x^k + y^k)$, $x,\ y \ge 0$, we finally conclude that

$$|\phi|_k - |\psi|_k \le C(1/k)C(k)\, 2^{(k-1)^2/k}\, |\Omega|^{1/k} |(\phi-\psi)_+|_\infty. \qquad \square$$

A.5. Examples and applications. We illustrate the previous abstract results by examples which include the specific cases under consideration in the main text of the article.

Example A.1. Consider spatially nonlocal problems of integral type, defined by

$$F^t\big(R^t u\big)(x) = f\big(t, x, u(t,x), \nabla u(t,x), |u_+(t,.)|_k\big), \tag{A27}$$

where $|.|_k = \|.\|_{L^k(\Omega)}$, $1 \le k \le \infty$, and $f : \mathbb{R}^+ \times \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^+ \to \mathbb{R}$. We have the following theorem.

THEOREM A.12. *Assume* (A1)′ *and let* F^t *be defined by* (A27)*, where* f *is locally Lipschitz continuous.*

(i) *Let*

$$u_0 \in C^1(\overline{\Omega}), \qquad \text{with } u_{0|\partial\Omega} = 0.$$

Then there exists a unique, maximal in time function u*, defined on* $[0,T^*)\times\overline{\Omega}$ *for some* $T^* = T^*(u_0) \in (0,\infty]$*, such that for all* $0 < \epsilon < T < T^*$*,* $u \in C^{0,1}(\overline{Q_T}) \cap C^{2+\alpha}(\overline{Q_{\epsilon,T}})$*, is a (classical) solution of* (A2)*. Moreover, if* $T^* < \infty$*, then*

$$\lim_{t\to T^*} |u(t,.)|_{C^1(\overline{\Omega})} = \infty.$$

(ii) *If* f *is nondecreasing with respect to its last argument, then the comparison principle* (*Theorem* A.10) *is valid.*

Remark A.2. If f does not depend on $\nabla u(t,x)$, the previous result holds (under the weaker assumption (A1)), for all $u_0 \in C(\overline{\Omega})$, $u_{0|\partial\Omega} = 0$, with blow up alternative in L^∞ norm (but $u \in C(\overline{Q_T})$ instead of $C^{0,1}(\overline{Q_T})$).

Part (i) of Theorem A.12 is a corollary of Theorems A.4 (i) and A.3 and Proposition A.6. To prove part (ii), we just need to check the validity of property (A24). To do this, we note that

$$f\big(t,x,p,q,|\phi_+|_k\big) - f\big(t,x,p,q,|\psi_+|_k\big) \le \begin{cases} 0 & \text{if } |\phi_+|_k \le |\psi_+|_k, \\ L_{M,T}\big(|\phi_+|_k - |\psi_+|_k\big) & \text{otherwise,} \end{cases}$$

where $L_{M,T}$ stands for a local Lipschitz constant of f. Property (A24) then follows from Lemma A.11.

Example A.2. The results of Theorem A.12 and of Remark A.2 remain valid without change for spatially nonlocal problems of localized type, defined by

$$F^t\big(R^t u\big)(x) = f\big(t,x,u(t,x),\nabla u(t,x),u(t,x_0(t))\big),$$

where $f : \mathbb{R}^+\times\overline{\Omega}\times\mathbb{R}\times\mathbb{R}^N\times\mathbb{R} \to \mathbb{R}$ is locally Lipschitz continuous and $x_0 : [0,\infty) \to \overline{\Omega}$ is locally Hölder continuous.

Example A.3. Consider time nonlocal problems defined by

$$F^t\big(R^t u\big)(x) = f\big(t,x,u(t,x),|u_+(.,x)|_{L^k(0,t)}\big), \tag{A28}$$

where $1 \le k \le \infty$ and $f : \mathbb{R}^+\times\overline{\Omega}\times\mathbb{R}\times\mathbb{R}^N\times\mathbb{R}^+ \to \mathbb{R}$ is locally Lipschitz continuous. We may also consider space-time nonlocal problems defined by

$$F^t\big(R^t u\big)(x) = f\big(t,x,u(t,x),I(t,u)\big), \tag{A29}$$

where

$$I(t,u) = \left(\int_0^t \int_\Omega \beta(y) u_+^k(s,y)\,dy\,ds\right)^{1/k}, \quad 1 \le k < \infty, \quad \beta \in C(\overline{\Omega}),$$

or

$$I(t,u) = |u_+|_{L^\infty((0,t)\times\Omega)}.$$

We have the following theorem.

THEOREM A.13. *Assume* (A1)′ *and let* F^t *be defined by* (A28) *or* (A29), *with* f *as above.*

(i) *Let*

$$u_0 \in C^1(\overline{\Omega}), \qquad \text{with } u_0{}_{|\partial\Omega} = 0.$$

Then there exists a unique, maximal in time function u, *defined on* $[0,T^*)\times\overline{\Omega}$ *for some* $T^* = T^*(u_0) \in (0,\infty]$, *such that for all* $0<\epsilon<T<T^*$, $u \in C^{0,1}(\overline{Q_T})\cap C^{2+\alpha}(\overline{Q_{\epsilon,T}})$, *and* u *is a (classical) solution of* (A2). *Moreover, if* $T^*<\infty$, *then* $\limsup_{t\to T^*}|u(t,.)|_{L^\infty(\Omega)}=\infty$.

(ii) *If* f *is nondecreasing with respect to its last argument, then the comparison principle* (*Theorem* A.10) *is valid.* (*This is still true if* f *depends also on* $\nabla u(t,x)$.)

Part (i) is a corollary of Theorems A.5 and A.3. Part (ii) follows from Lemma A.11 with $(0,T)$ or $(0,T)\times\Omega$ instead of Ω.

Of course, if F^t is a more general, gradient-dependent, time or space-time nonlocal operator, then local existence of solutions is valid for smoother ($C^{2+\alpha}$) initial data by Theorem A.1. (Observe that, even for (1.4), which falls within the range of Theorem A.13, local existence is proved in [P1] and [GS] only for $C^{2+\alpha}$ data.)

On the other hand, Theorem A.13 remains valid for problems with localization in time, defined by

$$F^t(R^t)(x) = f\big(t,x,u(t_0(t,x),x)\big),$$

with t_0 locally Hölder continuous $[0,\infty)\times\overline{\Omega}\to[0,\infty)$, and $0\le t_0(t,x)\le t$. Still further, there would be no obstruction to localizing both in space and time.

Remark A.3. In the case of equation (1.4), it is proved in [GS] that the blowing up solutions actually satisfy $\lim_{t\to T^*}|u(t,.)|_{L^\infty(\Omega)}=\infty$, and an estimate on the blow-up rate is also given.

Example A.4. In Examples A.1 and A.3, it is also possible to replace the L^k-norm with an expression of the form $[\int_\Omega K(t,x,y)u_+^k(t,y)\,dx]^{1/k}$ or $\int_0^t K(t,x,s)u_+^k(s,x)\,ds$ under suitable assumptions on the weight function $K\ge 0$. Nonlocal terms of the form $\int_{t/2}^t u_+^k(s,x)\,ds$, for instance, can be treated in this way.

Example A.5. To see that an assumption of nondecreasing monotonicity on the nonlocal term cannot be avoided for the comparison principle to hold, let us recall the following counterexample [WW, p. 1143]: $N=1$, $\Omega=(-1,1)$, $F^t(R^tu) = -18\int_{-1}^1 u(t,y)\,dy$, and $w(t,x)=x^2-t$, which satisfies

$$\begin{aligned} w_t-\Delta w = -3 \ge -18\textstyle\int_{-1}^1 w(t,y)\,dy, \quad 0<t<1/4, \quad -1<x<1,\\ w(t,\pm1) = 1-t>0, \quad 0<t<1/4,\\ w(0,x)=x^2\ge 0, \quad -1<x<1, \end{aligned}$$

but $w(t,0)=-t<0$. In the same direction, see the example from [CaY, p. 287], which shows that the maximum principle can be false if the gradient of u is involved in a nonlocal way in the equation.

REFERENCES

[A] H. Amann, *Parabolic evolution equations and nonlinear boundary conditions*, J. Differential Equations, 72 (1988), pp. 201–269.

[AC1] S. N. Antontsev and M. Chipot, *The thermistor problem: Existence, smoothness, uniqueness, blow-up*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.

[AC2] S. N. Antontsev and M. Chipot, *The analysis of blow-up for the thermistor problem*, Siberian Math. J., 38 (1997), pp. 827–841.

[B] J. M. Ball, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math. Oxford, 28 (1977), pp. 473–486.

[BB] J. W. Bebernes and A. Bressan, *Thermal behaviour for a confined reactive gas*, J. Differential Equations, 44 (1982), pp. 118–133.

[BE] J. W. Bebernes and R. Ely, *Comparison techniques and the method of lines for a parabolic functional equation*, Rocky Mountain J. Math., 12 (1982), pp. 723–733.

[BT] J. W. Bebernes and P. Talaga, *Nonlocal problems modelling shear banding*, Comm. Appl. Nonlinear Anal., 3 (1996), pp. 79–103.

[Be] H. Bellout, *Blow-up of solutions of parabolic equations with nonlinear memory*, J. Differential Equations, 70 (1987), pp. 42–68.

[Bel] V. S. Belonosov, *Estimates of solutions of parabolic systems in weighted Hölder classes and some applications*, Math. Sb., 110 (1979), pp. 163–188; Math. USSR Sb., 38 (1979), pp. 151–173.

[BG] D. Blanchard and H. Ghidouche, *A nonlinear system for irreversible phase changes*, European J. Appl. Math., 1 (1990), pp. 91–100.

[BDS] C. Budd, B. Dold, and A. Stewart, *Blowup in a partial differential equation with conserved first integral*, SIAM J. Appl. Math., 53 (1993), pp. 718–742.

[CaY] J. R. Cannon and H.–M. Yin, *A class of non-linear non-classical parabolic equations*, J. Differential Equations, 79 (1989), pp. 266–288.

[CPY] J. M. Chadam, A. Peirce, and H.–M. Yin, *The blow-up property of solutions to some diffusion equations with localized nonlinear reactions*, J. Math. Anal. Appl., 169 (1992), pp. 313–328.

[CY] J. M. Chadam and H.–M. Yin, *An iteration procedure for a class of integrodifferential equations of parabolic type*, J. Integral Equations Appl., 2 (1989), pp. 31–47.

[CD] P. Clément and G. Da Prato, *Some results on nonlinear heat equations for materials of fading memory type*, J. Integral Equations Appl., 2 (1990), pp. 375–391.

[D] K. Deng, *Dynamical behavior of solutions of a semilinear parabolic equation with nonlocal singularity*, SIAM J. Math. Anal., 26 (1995), pp. 98–111.

[DKL] K. Deng, M. K. Kwang, and H. A. Levine, *The influence of nonlocal nonlinearities on the long time behaviour of solutions of Burgers equation*, Quart. Appl. Math., 50 (1992), pp. 173–200.

[F] M. Fila, *Boundedness of global solutions of nonlocal parabolic equations*, Nonlinear Anal., 30 (1997), pp. 877–885.

[Fr] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

[G] R. Gardner, *Solutions of a nonlocal conservation law arising in combustion theory*, SIAM J. Math. Anal., 18 (1987), pp. 173–183.

[GS] J.–S. Guo and H.–W. Su, *The blow-up behaviour of the solution of an integrodifferential equation*, Differential Integral Equations, 5 (1992), pp. 1237–1245.

[HY] B. Hu and H.–M. Yin, *Semilinear parabolic equations with prescribed energy*, Rend. Circ. Mat. Palermo, 44 (1995), pp. 479–505.

[Ka] S. Kaplan, *On the growth of solutions of quasilinear parabolic equations*, Comm. Pure Appl. Math., 16 (1963), pp. 327–343.

[Ko] A. Khozanov, *Parabolic equations with nonlocal nonlinear source*, Siberian Math. J., 35 (1994), pp. 545–556.

[L] A. A. Lacey, *Thermal runaway in a non-local problem modelling ohmic heating.* I: *Model derivation and some special cases*, European J. Appl. Math., 6 (1995), pp. 127–144; II: *General proofs of blow-up and asymptotics of runaway*, European J. Appl. Math., 6 (1995), pp. 201–224.

[LS] A. Lunardi, E. Sinestrari, *Fully nonlinear integrodifferential equations in general Banach spaces*, Math Z., 190 (1985), pp. 225–248.

[MR] A. Majda and R. Rosales, *Resonantly interacting nonlinear hyperbolic waves*, Stud. Appl. Math., 71 (1984), pp. 149–179.

[M] X. Mora, *Semilinear parabolic equations define semiflows on C^k spaces*, Trans. Amer. Math. Soc., 278 (1983), pp. 21–55.

[O] W. E. Olmstead, *Critical speed for avoidance of blow-up in a reactive-diffusive medium*, Z. Angew. Math. Phys., 48 (1997), pp. 701–710.

[OR] W. E. Olmstead and C. A. Roberts, *Explosion in a diffusive strip due to a concentrated nonlinear source*, Methods Appl. Anal., 1 (1994), pp. 434–445.

[P1] C. V. Pao, *Nonexistence of global solutions for an integrodifferential system in reactor dynamics*, SIAM J. Math. Anal., 11 (1980), pp. 559–564.

[P2] C. V. Pao, *Blowing-up of solution for a nonlocal reaction-diffusion problem in combustion theory*, J. Math. Anal. Appl., 166 (1992), pp. 591–600.

[Sf] D. Sforza, *Parabolic integrodifferential equations with singular kernels*, J. Integral Equations Appl., 3 (1991), pp. 601–623.

[So] Ph. Souplet, *Nonexistence of global solutions to some differential inequalities of the second order and applications*, Portugal. Math., 52 (1995), pp. 289–299.

[STW] Ph. Souplet, S. Tayachi, and F. B. Weissler, *Exact self-similar blow-up of solutions of a semilinear parabolic equation with a nonlinear gradient term*, Indiana Univ. Math. J., 48 (1996), pp. 655–682.

[SW1] Ph. Souplet and F. B. Weissler, *Self-similar sub-solutions and blow-up for nonlinear parabolic equations*, J. Math. Anal. Appl., 212 (1997), pp. 60–74.

[SW2] Ph. Souplet and F. B. Weissler, *Poincaré's inequality and global solutions of a nonlinear parabolic equation*, Ann. Inst. Henri Poincaré Anal. Non Linéaire, to appear.

[TW] C. C. Travis and G. F. Webb, *Existence, stability, and compactness in the α–norm for partial functional differential equations*, Trans. Amer. Math. Soc., 240 (1978), pp. 129–143.

[WW] M. Wang and Y. Wang, *Properties of positive solutions for non-local reaction-diffusion problems*, Math. Methods Appl. Sci., 19 (1996), pp. 1141–1156.

[Y1] Y. Yamada, *On a certain class of semilinear Volterra diffusion equations*, J. Math. Anal. Appl., 88 (1982), pp. 433–457.

[Y2] Y. Yamada, *Asymptotic stability for some systems of semilinear Volterra diffusion equations*, J. Differential Equations, 52 (1984), pp. 295–326.

[Yi] Y. Yin, *Quenching for solutions of some parabolic equations with singular nonlocal terms*, Dynam. Systems Appl., 5 (1996), pp. 19–30.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1335–1360, November 1998

002

ADIABATIC INVARIANT OF THE HARMONIC OSCILLATOR, COMPLEX MATCHING AND RESURGENCE *

CARLES BONET†, DAVID SAUZIN‡, TERE SEARA†, AND MARTA VALÈNCIA†

Abstract. The linear oscillator equation with a frequency slowly dependent on time is used to test a method to compute exponentially small quantities. This work presents the matching method in the complex plane as a tool to obtain rigorously the asymptotic variation of the action of the associated Hamiltonian *beyond all orders*.

The solution in the complex plane is approximated by a series in which all terms present a singularity at the same point. Following matching techniques near this singularity one is led to an equation which does not depend on any parameter, the so-called inner equation, of a Riccati-type. This equation is studied by resurgence methods.

Key words. Adiabatic invariants, exponentially small, matching theory, resurgence theory

AMS subject classifications. 30B, 34E, 40C, 58F

PII. S0036141097321516

1. Introduction.

We consider one degree of freedom Hamiltonian system depending on a parameter that changes slowly with time modelled by a Hamiltonian of the form (see [1])

$$H(I,\varphi,\varepsilon t) = H_0(I,\lambda(\varepsilon t)) + \varepsilon\lambda'(\varepsilon t)H_1(I,\varphi,\varepsilon t)\,,$$

where $\lambda(\varepsilon t)$ is a function with definite limits at $\pm\infty$ and such that $\lambda^{k)}(\tilde{t}) \to 0$ when $\tilde{t} \to \pm\infty$, for all $k \in \mathbf{N}$. The equations of the motion are given by

$$\left\{\begin{array}{rcl} \dot{I} &=& -\varepsilon\lambda'\dfrac{\partial H_1}{\partial\varphi}, \\ \dot{\varphi} &=& \dfrac{\partial H_0}{\partial I} + \varepsilon\lambda'\dfrac{\partial H_1}{\partial I}\,. \end{array}\right. \tag{1.1}$$

This is a quasi-integrable system in the sense that we can apply the classical averaging procedure looking for a change of variables, close to the identity in powers of ε,

$$\left\{\begin{array}{rcl} I &=& J + \varepsilon u_1(J,\psi,t) + \varepsilon^2 u_2(J,\psi,t) + \cdots, \\ \varphi &=& \psi + \varepsilon v_1(J,\psi,t) + \varepsilon^2 v_2(J,\psi,t) + \cdots, \end{array}\right. \tag{1.2}$$

in order to eliminate the angle variables of the Hamiltonian.

If we truncate the formal series (1.2) at order n, the system obtained is of the form

$$\left\{\begin{array}{rcl} \dot{J} &=& \varepsilon^n\lambda'(\varepsilon t)\ldots, \\ \dot{\psi} &=& \dfrac{\partial\mathcal{H}}{\partial I}(I,\varepsilon) + \varepsilon^n\ldots. \end{array}\right. \tag{1.3}$$

*Received by the editors May 12, 1997; accepted for publication September 15, 1997; published electronically June 22, 1998. This work forms part of the Projects PR9614 of the UPC, PB95-0629 and PB94-0215 of the DGICYT, and the EC grant ERBCHRXCT-940460.

http://www.siam.org/journals/sima/29-6/32151.html

†Department Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain (bonet@ma1.upc.es, tere@ma1.upc.es, valencia@ma1.upc.es).

‡Bureau des Longitudes, CNRS, 3, rue Mazarine, 75006 Paris, France (sauzin@bdl.fr).

Poincaré proved that even though the series (1.2) are divergent they are asymptotic. In our case that means that the actions of systems (1.1) and (1.3) satisfy

$$|I(t) - J(t) - \varepsilon u_1(J(t), \psi(t), t) - \cdots - \varepsilon^{n-1} u_{n-1}(J(t), \psi(t), t)| \leq K\varepsilon^n,$$

for all $t \in \mathbf{R}$. As a consequence, $I(t)$ is an adiabatic invariant for system (1.1), in the sense that its variation is small for a long time interval. Moreover, due to the asymptotic properties of λ it happens that $u_n(J, \psi, t) \to 0$ and $v_n(J, \psi, t) \to 0$, as $t \to \pm\infty$. Then, one can see that I and J have limits at $\pm\infty$ verifying $I(\pm\infty, \varepsilon) = J(\pm\infty, \varepsilon) + O(\varepsilon^n)$, for all $n \in \mathbf{N}$. Moreover, from (1.3) and taking into account that $\lambda(\varepsilon t)$ is bounded, one has that $J(+\infty, \varepsilon) = J(-\infty, \varepsilon) + O(\varepsilon^n)$, $\forall n \in \mathbf{N}$. Hence, it follows that

$$\Delta I(\varepsilon) := I(+\infty, \varepsilon) - I(-\infty, \varepsilon) = O(\varepsilon^n), \quad \forall n \in \mathbf{N}. \tag{1.4}$$

That is, $I(t, \varepsilon)$ is a perpetual adiabatic invariant at all orders.

Nevertheless, this discrepancy is nonzero (otherwise, system (1.1) would be integrable) but it cannot be viewed directly from the asymptotic series (1.2). The goal of this paper is to present a method to compute the asymptotic expansion of the adiabatic invariant "beyond all orders."

1.1. Matching and resurgence. In fact, we have an asymptotic development of I uniformly valid for all $t \in \mathbf{R}$ and the problem is to catch the part of I invisible in the series (1.2). Matching theory principle says that in order to see the *hidden* properties of a function defined by an asymptotic series we must go to the regions, called *boundary layers*, where these series are no longer asymptotic. Boundary layers can be found by two fundamental methods: the first one is an a priori knowledge of its location provided by heuristic arguments and the second one is to look for the singularities of the series terms.

If we follow the second method for (1.2), we see that the terms of these series do not have singularities in $\mathbf{R}$ (due to the asymptotic properties of λ) and therefore, we are led to look for the boundary layers in $\mathbf{C}$. This is the principal reason why these problems are formulated using complex numbers and why the equation requires analyticity properties. Furthermore, working with analytic functions and complex asymptotic theory gives us more chances to obtain refined results.

Among others, we use as a basic tool in this paper resurgence theory for understanding the nature of the divergence of the series. But instead of analyzing the outer expansion (1.2), we apply resurgence theory to the *inner expansion* (the series near the boundary layer) to compute $\Delta I(\varepsilon)$ given in (1.4). These techniques have been used by V. Hakim and K. Mallick in [7] to compute formally the separatrix splitting of the standard map.

In the present paper we use their aproach to compute the behavior of the adiabatic invariant for a simple oscillator

$$\ddot{x} + \phi^2(\varepsilon\tau)x = 0, \tag{1.5}$$

obtaining rigorously an asymptotic expression for the adiabatic invariant $\Delta I(\varepsilon)$ *beyond all orders*. This problem is quite well understood [3] but we think useful and clarifying to treat it joining matching techniques and the resurgence theory. We have followed closely Wasow [18] and Meyer [12] formulation reducing (1.5) to a Riccati equation.

1.2. Wasow formulation and reduction to a Riccati equation. Following [18] and taking $t = \varepsilon\tau$ in (1.5), let us consider the equation

$$\varepsilon^2\ddot{u} + \phi^2(t)u = 0, \quad t \in \mathbf{R} \tag{1.6}$$

where $\phi(t)$ satisfies
H1 $\phi(t) > 0$, $\forall t \in \mathbf{R}$,
H2 $\lim_{t\to\pm\infty} \phi(t) = \phi_\pm > 0$,
H3 $\phi \in \mathcal{C}^\infty(\mathbf{R})$ and $\phi^{k)} \in L_1((-\infty, +\infty))$, $k \in \mathbf{N}$, (i.e., $\dot{\phi}$ is a gentle function).
Now, given any solution $u(t,\varepsilon)$ of (1.6), let us denote by $I(t,\varepsilon)$ the function

$$I^2(t,\varepsilon) := \phi(t)u^2(t,\varepsilon) + \varepsilon^2\frac{\dot{u}^2(t,\varepsilon)}{\phi(t)}$$

(when ϕ is a constant, $I(t,\varepsilon)$ is the action variable of the integrable Hamiltonian system associated to (1.6)). Littlewood proved in [9] that for any solution $u(t,\varepsilon)$ the limits $I(\pm\infty,\varepsilon)$ exist, and

$$\Delta I^2(\varepsilon) = I^2(+\infty,\varepsilon) - I^2(-\infty,\varepsilon) = O(\varepsilon^n), \quad \forall n \in \mathbf{N}.$$

Moreover, Wasow proved that ΔI^2 satisfies

$$\Delta I^2(\varepsilon) = 2\varepsilon\mathrm{Re}\left[\left(\sqrt{\phi(0)}u_0 + i\frac{\varepsilon}{\sqrt{\phi(0)}}\dot{u}_0\right)^2 \hat{p}(+\infty,\varepsilon)\right](1 + O(\varepsilon)), \tag{1.7}$$

where $(u_0, \dot{u}_0)$ are the initial conditions of $u(t,\varepsilon)$, and $\hat{p}(t,\varepsilon) = e^{-(2i/\varepsilon)\int_0^t \phi(s)ds}p(t,\varepsilon)$, with $p(t,\varepsilon)$ being the solution of the Riccati equation

$$\begin{cases} \varepsilon\dot{p} = 2i\phi(t)p + \dfrac{\dot{\phi}(t)}{2\phi(t)}(1 - \varepsilon p^2), \\ p(-\infty,\varepsilon) = 0 \end{cases} \tag{1.8}$$

for all $\varepsilon > 0$. Looking for the solution as a power series in ε, one can prove Littlewood's results, but in order to obtain more acurate estimates for $\Delta I^2(\varepsilon)$ we will need to extend our problem to the complex domain for the variable t.

By the change of variable $x = \int_0^t \phi(s)ds$ (1.8) becomes

$$\varepsilon w' = 2iw + f(x)(1 - \varepsilon^2 w^2), \tag{1.9}$$

where $f(x) = \frac{\dot{\phi}(t)}{2\phi^2(t)}$. Now, due to hypotheses **H1**, **H2**, **H3**, on ϕ, it is clear that $f(x)$ is a real function with gentle properties. But in order to study the problem on **C**, let us make the following extra hypotheses on f:

H4 f is real analytic in $\bar{\Gamma} - \{x_0\}$, where $x_0 \in \mathbf{C}$, such that Im $(x_0) < 0$ and $\Gamma = \{x \in \mathbf{C} : \text{Im } (x_0) < \text{Im } (x) \le 0\}$, and for $|x - x_0| \le 1$ one has

$$f(x) = \frac{1}{6(x - x_0)}\left[1 + \tilde{f}((x - x_0)^{2/3})\right]$$

with $\tilde{f}(u)$ being an holomorphic function such that $\tilde{f}(0) = 0$.

H5 f is **C**-gentle in the sense that for all $x \in \bar{\Gamma} - \{x_0\}$ one has

$$\lim_{\mathrm{Re}\, x\to\pm\infty} \int_{C_\pm(x)} \left|f^{k)}(s)\right| ds = 0, \quad k \in \mathbf{N},$$

uniformly on x, where

$$C_+(x) = \{t \in \mathbf{C} : \text{Im } (t) = \text{Im } (x), \quad \text{Re } (t) \geq \text{Re } (x) \}$$

and

$$C_-(x) = \{t \in \mathbf{C} : \text{Im } (t) = \text{Im } (x), \quad \text{Re } (t) \leq \text{Re } (x) \}.$$

Although our hypotheses **H4** and **H5** of f can seem capricious, they are deduced from the more natural hypotheses on ϕ made by Wasow in [19], namely, ϕ^2 has an analytic continuation to the complex domain and has a simple zero in $\mathbf{C}$ noted t_0, with Im $(t_0) < 0$, such that $x_0 = \int_0^{t_0} \phi(s)ds$ (the case Im $(t_0) > 0$ can be studied in an analogous way).

The aim of this paper is to compute $w(+\infty, \varepsilon)$, where $w(x, \varepsilon)$ is the solution of the Riccati equation (1.9) such that $w(-\infty, \varepsilon) = 0$. The rest of this paper is structured as follows: First of all, in section 2 we seek for $w(x, \varepsilon)$ as a power series in ε, for complex values of x. We will study its asymptotic validity until some neighborhood of the singularity x_0 which is called the *inner region*. As is usual in matching methods, in the inner region a change of variables will be needed in order to enlarge the validity of the solution. This is done in section 3, obtaining as a first approximation in this region the solution of the so-called *inner equation*. This inner equation is studied by the help of resurgence theory in a self-contained way in section 5. In the inner region we can catch some terms of our solution hidden in the power series, and in section 4 we prove that they are going to be exponentially small on ε (but not zero!) at $+\infty$. Finally, in section 6 we make some remarks for more general nonlinear inner equations. We defer for another paper the general study of (1.1) in a Hamiltonian form (see [16]). Recently, Ramis and Schäfke [14] have obtained upper bounds for $\Delta I(\varepsilon)$ showing the Gevrey-1 character (see footnote 1 in section 5) of the series (1.2) in the general case.

All of this is summarized in the following theorem.

THEOREM 1.1 (main theorem). *Let $w(x, \varepsilon)$ be the solution of the Riccati equation* (1.9) *such that* $\lim w(x, \varepsilon) = 0$ *when* $x \to -\infty$. *Then, if hypotheses* **H1**,...,**H5** *are satisfied one has*

$$\lim_{x \to +\infty} \hat{w}(x, \varepsilon) = -\frac{i}{\varepsilon} e^{-2ix_0/\varepsilon}(1 + O(\varepsilon^{2\gamma/3}))$$

where $\hat{w}(x, \varepsilon) := e^{-2ix/\varepsilon} w(x, \varepsilon)$ and γ is any number verifying $0 < \gamma < 1/2$. Moreover, the variation of the action of the Hamiltonian system associated to (1.6) *is given by*

$$\Delta I^2(\varepsilon) = -2\phi(0)u_0^2 e^{\frac{2\text{Im } (x_0)}{\varepsilon}} \sin\left(\frac{2\text{Re } (x_0)}{\varepsilon}\right)(1 + O(\varepsilon^{2\gamma/3}));$$

therefore, it is a quantity exponentially small in ε.

2. The solution in the outer left domain. In this section we prove the existence of the solution $w(x, \varepsilon)$ of the Riccati equation (1.9),

$$\varepsilon w' = 2iw + f(x)(1 - \varepsilon^2 w^2),$$

such that $\lim w(x, \varepsilon) = 0$, for Re $(x) \to -\infty$ and $x \in \Gamma$ (where Γ is defined in hypothesis **H4**), and we give an asymptotic expression of the solution in a suitable subdomain of Γ.

First of all, we look for a formal solution of (1.9) in the following proposition.

PROPOSITION 2.1. *There exists a series* $\sum_{n\geq 0}\varepsilon^n w_n(x)$ *that formally satisfies the Riccati equation* (1.9). *The functions* $w_n(x)$,

i. *verify the recurrence*

$$\text{(2.1)}\qquad \begin{cases} w_0(x) &= \dfrac{-f(x)}{2i} \\ w_1(x) &= \dfrac{w_0'(x)}{2i} \\ w_n(x) &= \dfrac{w_{n-1}'(x)}{2i} + \dfrac{f(x)}{2i}\sum_{i+j=n-2} w_i(x)w_j(x), \quad n>1; \end{cases}$$

ii. *are* **C**-*gentle functions (see hypothesis* **H5**);

iii. *are analytic functions in* $\bar{\Gamma}-\{x_0\}$ *with a singularity at* $x=x_0$ *such that*

$$\text{(2.2)}\qquad |w_n^{k)}(x)| \leq C_{n,k}\ |x-x_0|^{-(n+k+1)}\ , \quad \text{if } |x-x_0|\leq 1,\ k\in\mathbf{N}.$$

Remark. Due to the fact that $w_n(x)$ are **C**-gentle functions uniformly bounded for $|x-x_0|\geq 1$, we can choose the constants $C_{n,k}$ such that

$$\text{(2.3)}\qquad |w_n^{k)}(x)| \leq C_{n,k}\ , \quad \text{if } |x-x_0|\geq 1,\ k\in\mathbf{N}\,.$$

Proof. The recurrence is obtained directly by the substitution of the series into (1.9) and the properties of $w_n(x)$ follow from hypotheses **H4** and **H5** on $f(x)$. □

Now we will prove that if we are not close to the singularity x_0, the formal series of Proposition 2.1 is asymptotic to a **C**-gentle function $\hat{w}(x,\varepsilon)$. Unfortunately, $\hat{w}(x,\varepsilon)$ will not be a solution of (1.9) but, nevertheless, it will help us to prove the existence of the solution of (1.9) and its asymptoticity to the formal series.

Let $\Gamma_{\varepsilon^\gamma}$ be the following subdomain of Γ for a suitable $\gamma>0$.

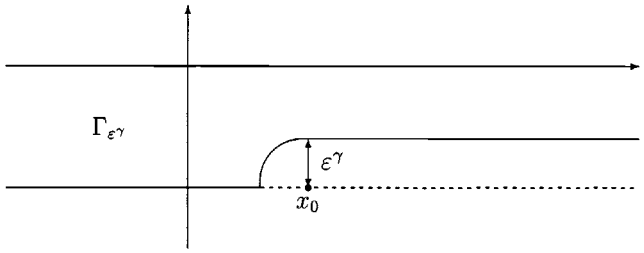

FIG. 1.

PROPOSITION 2.2. *Let* $w_n(x)$, $n\geq 0$, *be functions defined in* Γ *verifying* ii) *and* iii) *of Proposition* 2.1. *Then, for* $0\leq\gamma<1$ *and* $\varepsilon>0$ *sufficiently small, there exists an analytic function* $\hat{w}(x,\varepsilon)$ *defined for* $x\in\Gamma_{\varepsilon^\gamma}$, *such that*

i. *for any* $\delta>0$ *and* $k\in\mathbf{N}$,

$$|\hat{w}^{k)}(x,\varepsilon) - \sum_{n=0}^{N}\varepsilon^n w_n^{k)}(x)| \leq \hat{C}_{N,k}\varepsilon^{-(\gamma+\delta)}\varepsilon^{(N+1)(1-\gamma)-\gamma k}\,,$$

for all $x\in\Gamma_{\varepsilon^\gamma}$, *where* $\hat{C}_{N,k}$ *are constants independent of* ε *and* δ,

ii. $\hat{w}(x,\varepsilon)$ *is a* $\mathbf{C}$*-gentle function for* $\mathrm{Im}\,(x) \geq \mathrm{Im}\,(x_0) - \varepsilon^\gamma$.

Remark. Let us note that if $\gamma = 0$ (this means that we are far away from the singularity) this proposition says that the series $\sum_{n\geq 0} \varepsilon^n w_n(x)$ is asymptotic to the function $\hat{w}(x,\varepsilon)$ on Γ_1. In this sense we can look at i) as a weak form of asymptotic expansion near the singularity.

Proof. First of all, let us define

$$K_n(\varepsilon) := \max_{0\leq k\leq n} \left\{ \sup_{\Gamma_{\varepsilon^\gamma}} |w_n^{k)}(x)|, \sup_{\Gamma_{\varepsilon^\gamma}} \int_{C_\pm(x)} |w_n^{k)}(s)ds| \right\}.$$

From bounds (2.2) and (2.3), as $x \in \Gamma_{\varepsilon^\gamma}$ it follows that

$$K_n(\varepsilon) \leq C_n \varepsilon^{-\gamma(2n+1)},$$

where $C_n := \max\{C_{n,k} \,:\, 0 \leq k \leq n\}$ are independent of ε. Secondly, let us define, for any $\delta > 0$,

$$\alpha_n(\varepsilon) := 1 - e^{-1/(\varepsilon^\delta C_n)},$$

and note that $\alpha_n(\varepsilon) < \frac{1}{\varepsilon^\delta C_n}$.

Then, let us consider

$$\hat{\omega}^{k)}(x,\varepsilon) = \sum_{n\geq 0} \alpha_n(\varepsilon) w_n^{k)}(x) \varepsilon^n,$$

and let us define

$$L_k := \max\left\{1; \frac{C_{0,k}}{C_0}; \ldots; \frac{C_{k-1,k}}{C_{k-1}}; C_0; \ldots; C_k; C_{0,k}; \ldots; C_{k,k}\right\}.$$

Using bounds (2.2) and (2.3) it follows that

$$|\alpha_n(\varepsilon) w_n^{k)}(x)| \leq \frac{C_{n,k}}{C_n} \varepsilon^{-(\delta+\gamma)} \varepsilon^{-\gamma(n+k)}.$$

Thus, for any $k \geq 0$ and $n \geq 0$, we have that

$$|\alpha_n(\varepsilon) w_n^{k)}(x)| \leq L_k \varepsilon^{-(\gamma+\delta)} \varepsilon^{-\gamma(n+k)}. \tag{2.4}$$

So, from (2.4) and taking ε small enough, we obtain that

$$|\hat{w}^{k)}(x,\varepsilon)| \leq \sum_{n=0}^{\infty} |\alpha_n(\varepsilon) w_n^{k)}(x) \varepsilon^n| \leq 2L_k \varepsilon^{-\gamma-\delta} \varepsilon^{-k\gamma}, \tag{2.5}$$

and thus, that $\hat{w}^{k)}(x,\varepsilon)$ converges uniformly in $\Gamma_{\varepsilon^\gamma}$, for $0 \leq \gamma < 1$, $k \in \mathbf{N}$, and ε sufficiently small. Furthermore, if we define $\hat{w}(x,\varepsilon) := \hat{w}^{0)}(x,\varepsilon)$, we have that $\hat{w}^{k)}(x,\varepsilon)$ are the k-derivatives of $\hat{w}(x,\varepsilon)$.

Now, in order to see i) let us take $N > 0$ and let us again use the bounds (2.2), (2.3), and (2.4). It follows

$$|\hat{w}^{k)}(x,\varepsilon) - \sum_{n=0}^{N} w_n^{k)}(x)\varepsilon^n| = |\hat{w}^{k)}(x,\varepsilon) - \sum_{n=0}^{N} \alpha_n(\varepsilon) w_n^{k)}(x)\varepsilon^n + \sum_{n=0}^{N} (\alpha_n(\varepsilon) - 1) w_n^{k)}(x)\varepsilon^n|$$

$$\leq \sum_{n=N+1}^{\infty} |\alpha_n(\varepsilon) w_n^{k)}(x)\varepsilon^n| + \sum_{n=0}^{N} |(\alpha_n(\varepsilon)-1) w_n^{k)}(x)\varepsilon^n|$$
$$\leq L_k \sum_{n=N+1}^{\infty} \varepsilon^{n-\gamma(n+k+1)-\delta} + e^{-1/(L_N\varepsilon^\delta)} L_N \sum_{n=0}^{N} \varepsilon^{n-\gamma(n+k+1)}$$
$$\leq \hat{C}_{N,k}\varepsilon^{-(\delta+\gamma)}\varepsilon^{(N+1)(1-\gamma)-\gamma k}.$$

(we have used that $e^{-1/(L_N\varepsilon^\delta)}$ is exponentially small in ε).

By an analogous argument, using the integrals of $\hat{w}(x,\varepsilon)$ on $C_{\pm}(x)$, we can prove ii). □

Finally, the following theorem proves the existence of the solution $w(x,\varepsilon)$ and give us estimates on its domain of definition. In this domain we will also prove that the series of Proposition 2.1 is weakly asymptotic to $w(x,\varepsilon)$.

THEOREM 2.3. *Let us take* $0<\delta<1$ *and* $0\leq\gamma<1-\delta$. *Then, if* $\varepsilon>0$ *is small enough, the Riccati equation* (1.9) *defined for* $x\in\Gamma_{\varepsilon^\gamma}$, *has a unique solution* $w(x,\varepsilon)$ *such that* $\lim w(x,\varepsilon)=0$, *when* $\mathrm{Re}\,(x)\to-\infty$. *Furthermore, the solution* $w(x,\varepsilon)$ *satisfies that*

$$|w^{k)}(x,\varepsilon) - \sum_{n=0}^{N}\varepsilon^n w_n^{k)}(x)| \leq K_{N,k}\ \varepsilon^{-(\gamma+\delta)}\varepsilon^{(N+1)(1-\gamma)-\gamma k}, \tag{2.6}$$

for all $x\in\Gamma_{\varepsilon^\gamma}$ *and* $k,N\in\mathbf{N}$, *where the* $K_{N,k}$ *are constants independent of* ε *and* δ.

Proof. Let us take $w(x,\varepsilon)=\hat{w}(x,\varepsilon)+Q(x,\varepsilon)$, where $\hat{w}(x,\varepsilon)$ is the gentle function obtained in Proposition 2.2. Then, $w(x,\varepsilon)$ will be the solution of (1.9) if $Q(x,\varepsilon)$ is the solution of the equation

$$\varepsilon Q' = 2iQ - f(x)\varepsilon^2(\hat{w}Q - Q^2) + q(x,\varepsilon)\,, \tag{2.7}$$

where $q(x,\varepsilon):=2i\hat{w}+f(x)(1-\varepsilon^2\hat{w}^2)-\varepsilon\hat{w}'$ is an analytic $\mathbf{C}$-gentle function in $\Gamma_{\varepsilon^\gamma}$ such that $q(x,\varepsilon)=O(\varepsilon^n)$, for all $n\in\mathbf{N}$ and for all $x\in\Gamma_{\varepsilon^\gamma}$. Let us note that this implies that q verifies that, for any $n\in\mathbf{N}$

$$\int_{C_-(x)}\left|\frac{q(s,\varepsilon)}{\varepsilon}\right| \leq K_n\varepsilon^n \tag{2.8}$$

for some constant K_n.

Let us now consider the operator

$$T(Q)=\int_{C_-(x)} e^{2i(x-s)/\varepsilon}\left(\frac{q(s,\varepsilon)}{\varepsilon} - f(s)\varepsilon\left[\hat{w}(s,\varepsilon)Q(s,\varepsilon)-Q^2(s,\varepsilon)\right]\right)ds$$

defined in the Banach space of continuous bounded functions, with the supremum norm. Then, using bounds (2.5) and (2.8), and hypothesis **H4** we have, for ε small enough, that

1) if $||Q||\leq 1$,

$$||T(Q)||\leq K_n\varepsilon^n + \varepsilon\ln\varepsilon^\gamma(2L_0\varepsilon^{-(\gamma+\delta)}+1)\leq 1,$$

2) if $||Q_i|| \leq 1$, for $i = 1, 2$,

$$||T(Q_1) - T(Q_2)|| \leq ||Q_1 - Q_2||\varepsilon(||\hat{w}|| + 2) \int_{C_-(x)} |f(s)|ds$$
$$\leq ||Q_1 - Q_2||\varepsilon(2L_0\varepsilon^{-(\gamma+\delta)} + 2)|\ln \varepsilon^\gamma| \leq \frac{1}{2}||Q_1 - Q_2||.$$

So, by the fixed-point theorem the integral equation $T(Q) = Q$, and thus the differential equation (2.7) have a unique solution Q. Moreover, using again (2.8), one has, for any $n \in \mathbf{N}$

$$||Q|| = ||T(Q)|| \leq 2||T(0)|| \leq 2 \int_{C_-(x)} \left| \frac{q(s,\varepsilon)}{\varepsilon} \right| ds \leq 2K_n\varepsilon^n .$$

Finally, using that $Q(x,\varepsilon) = w(x,\varepsilon) - \hat{w}(x,\varepsilon)$ and the bound of $\hat{w}(x,\varepsilon)$ given in Proposition 2.2 we obtain the desired result. To obtain the bounds for the derivatives $w^{k)}(x,\varepsilon)$ we only have to use (2.7) to see that all the derivatives of Q are asymptotic to zero. □

Unfortunately, with Theorem 2.3 we have proved that $\lim w(x,\varepsilon) = O(\varepsilon^n)$ when Re $(x) \to +\infty$, for all $n \in \mathbf{N}$, but we cannot obtain a more refined description of it at infinity. So, if we want to obtain an asymptotic expression for this limit, we will need to study the solution near the singularity $x = x_0$ of $w_n(x)$. In order to simplify the exposition, we will assume from now on $0 < \gamma < 1/2$.

3. The solution in the inner domain. The goal of this section is to obtain an asymptotic representation of $w(x,\varepsilon)$ near the singularity $x = x_0$ of $w_n(x)$. Of course, we cannot obtain it at $x = x_0$ but as we will see in section 4 it will be sufficient to work at a distance of order ε of this singularity. So, we will extend $w(x,\varepsilon)$ of Theorem 2.3 from a point x^* such that $|x - x^*| = \varepsilon^\gamma$, Im $(x^*) \geq$ Im $(x_0) + \varepsilon$, and Re $(x^*) \leq$ Re (x_0) (i.e., x^* belongs to the boundary of the left domain) up to the point $\tilde{x}^*$ symmetric of x^* with respect to the line $\{\text{Re } (x) = \text{Re } (x_0)\}$. From $\tilde{x}^*$ we will continue the solution in the next section.

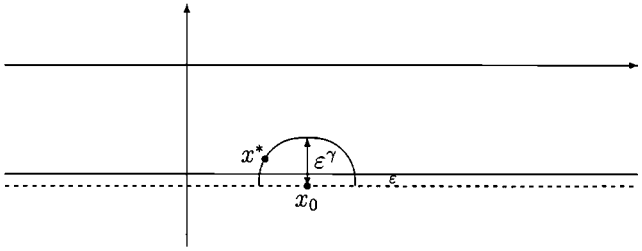

FIG. 2.

Note that, taking into account the bound (2.6), for $N = 0$, the asymptotic expression of f given by hypothesis **H4** and that $0 < \gamma < 1/2$, the initial condition of $w(x,\varepsilon)$ in the inner domain verifies

$$|w(x^*,\varepsilon) - \frac{1}{12i(x^* - x_0)}| \leq \bar{K}\varepsilon^{-\gamma}(\varepsilon^{2/3\gamma} + \varepsilon^{1-\gamma-\delta}) \leq K^*\varepsilon^{-\gamma/3},$$

where $\bar{K}$ and K^* are constants independent of ε.

Then, if we consider the change of variable and function

$$\tau = \frac{x - x_0}{\varepsilon} \qquad p(\tau, \varepsilon) = \varepsilon w(x_0 + \varepsilon\tau, \varepsilon)$$

(1.9) is transformed into

$$p' = 2ip + \varepsilon f(x_0 + \varepsilon\tau)(1 - p^2) \tag{3.1}$$

and defining $\tau^* = \frac{x^* - x_0}{\varepsilon}$, the initial condition for $p(\tau, \varepsilon)$ in the inner domain must verify

$$\left| p(\tau^*, \varepsilon) - \frac{1}{12i\tau^*} \right| \leq K^* \varepsilon^{1-\gamma/3}. \tag{3.2}$$

So, we have to study the solution of (3.1) verifying (3.2) for $\tau \in \mathbf{C}$ such that Im $\tau \geq 1$ and Re $(\tau^*) <$ Re $(\tau) <$ Re $(\tilde{\tau}^*)$, where $\tilde{\tau}^* = \frac{\tilde{x}^* - x_0}{\varepsilon}$ (we will establish in Theorem 3.2 the unicity of such a solution). In order to do this, we will compare $p(\tau, \varepsilon)$ with the solution of

$$p_0' = 2ip_0 + \frac{1}{6\tau}(1 - p_0^2) \tag{3.3}$$

such that $\lim p_0(\tau) = 0$ when Re $(\tau) \to -\infty$. This equation is obtained from (3.1) when ε tends to 0 where the initial condition is obtained "matching" the inner solution with the outer solution at τ^*.

It is easy to see, as in Proposition 2.1, that there exists a formal solution $\sum_{n\geq 0} a_n \tau^{-n-1}$ of (3.3). Moreover, looking at (2.1) and the behavior of f assumed in hypothesis **H4** one can see that the a_n are the principal parts of the terms of the outer series near the singularity, that is

$$w_n(x) = \frac{a_n}{(x - x_0)^{n+1}}(1 + O((x - x_0)^{2/3})).$$

In the next theorem existence, analytic properties, and asymptoticity of $p_0(\tau)$ are described. The proof, done with resurgence theory methods, is given in section 5.

THEOREM 3.1. i. *Equation* (3.3) *admits a unique solution* $p_0(\tau)$ *analytic in a sectorial neighborhood of* $-\infty$ *such that*

$$\lim_{\text{Re } \tau \to -\infty} p_0(\tau) = 0.$$

Moreover, this function is analytic in $\mathbf{C} - \mathbf{R}^+$*, and is asymptotic to the formal solution of* (3.3) *in every proper subsector of this set.*

ii. *Equation* (3.3) *admits a unique solution* $\tilde{p}_0(\tau)$ *analytic in a sectorial neighborhood of* $+\infty$ *such that*

$$\lim_{\text{Re } \tau \to +\infty} \tilde{p}_0(\tau) = 0.$$

Moreover, this function is analytic in $\mathbf{C} - \mathbf{R}^-$*, and is asymptotic to the formal solution of* (3.3) *in every proper subsector of this set.*

iii. *If* Re $\tau > 0$ *and* Im $\tau > 0$*,*

$$p_0(\tau) - \tilde{p}_0(\tau) = -ie^{2i\tau}(1 + O(\tau^{-1}))\,. \tag{3.4}$$

Now, if we compare $p(\tau,\varepsilon)$ with $p_0(\tau)$ we have the following theorem.

THEOREM 3.2. *The problem* (3.1), (3.2) *has a unique solution* $p(\tau,\varepsilon)$ *defined for* $D_{\tau^*} = \{\tau \in \mathbf{C} : \mathrm{Re}\,(\tau^*) \le \mathrm{Re}\,(\tau) \le \mathrm{Re}\,(\tilde{\tau}^*)\,, \mathrm{Im}\,\tau \ge 1\}$. *Moreover,* $p(\tau,\varepsilon)$ *satisfies that*

$$|p(\tau,\varepsilon) - p_0(\tau)| \le L\varepsilon^{(2/3)\gamma}$$

for all $\tau \in D_{\tau^*}$, *where* L *is independent of* ε.

For the proof of this theorem we will need the following lemma.

LEMMA 3.3. *There exists a constant* B, *independent of* ε, *such that for* τ, τ_1, *and* $\tau_2 \in D_{\tau^*}$:

i. $|p_0(\tau)| \le B$,

ii. $|\int_{\tau_1}^{\tau_2} \frac{p_0(s)}{s} ds| \le B$,

iii. $\int_{\tau_1}^{\tau_2} |\frac{\tilde{f}((\varepsilon s)^{2/3})}{s}| ds \le B\varepsilon^{(2/3)\gamma}$, *where* $\tilde{f}$ *is defined in hypothesis* **H4**.

Proof. Let us take $p_0(\tau)$ the unique solution of Theorem 3.1, and $0 < \alpha < \pi/2$ some fixed angle. Then there exists some constant C_α such that for $\tau \in \mathbf{C}$,

a. if $|\arg(\tau)| > \alpha$,

$$\left| p_0(\tau) + \frac{1}{12i\tau} \right| \le C_\alpha \frac{1}{\tau^2} \tag{3.5}$$

(use that $p_0(\tau)$ is asymptotic to the series $\sum_{n\ge 0} a_n \tau^{-n-1}$, where $a_0 = -\frac{1}{12i}$);

b. if $-\pi + \alpha \le \arg(\tau) \le \pi - \alpha$,

$$\left| \tilde{p}_0(\tau) + \frac{1}{12i\tau} \right| \le C_\alpha \frac{1}{\tau^2}$$

(use the same argument as before for $\tilde{p}_0(\tau)$);

c. if $|\arg(\tau)| < \alpha$,

$$|p_0(\tau) + ie^{2i\tau} - \tilde{p}_0(\tau)| \le C_\alpha \frac{1}{\tau}$$

(use (3.4)).

From these inequalities i) follows inmediately. In order to prove ii) we only need to integrate by parts and show that $\int_{\tau_1}^{\tau_2} \frac{e^{2is} ds}{s}$ is bounded for any τ_1, τ_2 in D_{τ^*}. Finally, iii) follows from hypothesis **H4** taking into account that $|\tau_i| \le \varepsilon^{\gamma-1}$. ∎

Proof (of Theorem 3.2). If we consider $v := p - p_0$, the problem that we have to study is

$$\begin{cases} v' &= \left[2i - \dfrac{p_0(\tau)}{3\tau}(1 + \tilde{f}((\varepsilon\tau)^{2/3}))\right] v - \dfrac{1}{6\tau}[1 + \tilde{f}((\varepsilon\tau)^{2/3}))]v^2 \\ & \quad -\dfrac{1}{6\tau}\tilde{f}((\varepsilon\tau)^{2/3})(1 - p_0^2), \\ v(\tau^*,\varepsilon) &= p(\tau^*,\varepsilon) - p_0(\tau^*). \end{cases} \tag{3.6}$$

Taking into account (3.2) and (3.5) (note that $|\arg\tau^*| > \alpha$) we have that

$$|v(\tau^*,\varepsilon)| \le K^*\varepsilon^{1-\gamma/3} + C_\alpha\varepsilon^{2(1-\gamma)} \le 2K^*\varepsilon^{1-\gamma/3}\,. \tag{3.7}$$

Now, let us consider the operator

$$\begin{aligned} T(v) = {} & e^{2i(\tau-\tau^*)} e^{-\int_{\tau^*}^{\tau} (p_0(r)/3r)(1+\tilde{f})dr} v(\tau^*, \varepsilon) \\ & - \int_{\tau^*}^{\tau} e^{2i(\tau-s)} e^{-\int_{s}^{\tau} (p_0(r)/3r)(1+\tilde{f})dr} \frac{1}{6s} v^2(s,\varepsilon) ds \\ & + \int_{\tau^*}^{\tau} e^{2i(\tau-s)} e^{-\int_{s}^{\tau} (p_0(r)/3r)(1+\tilde{f})dr} \frac{1}{6s} \tilde{f}(1-p_0^2(s)) ds, \end{aligned}$$

defined in the Banach space of continuous functions on D_{τ^*} with the supremum norm, such that $\|v\| \le L\varepsilon^{2/3\gamma}$ with $L = \frac{1}{2} e^{2B/3} B(1+B^2)$. Taking into account the bounds (3.7), (3.3), (3.3), and (3.3) of Lemma 3.3, for ε small enough, we have that

1) if $\|v\| \le L\varepsilon^{2/3\gamma}$,

$$\|T(v)\| \le \frac{1}{6} e^{2B/3} \left(12K^* \varepsilon^{1-\gamma} + 2L^2 \varepsilon^{(2/3)\gamma} \ln \varepsilon^{\gamma-1} + B(1+B^2) \right) \varepsilon^{(2/3)\gamma} \le L\varepsilon^{(2/3)\gamma};$$

2) if $\|v_i\| \le L\varepsilon^{(2/3)\gamma}$, for $i = 1, 2$,

$$\|T(v_1) - T(v_2)\| \le 2e^{2B/3} L\varepsilon^{(1/3)\gamma} \ln \varepsilon^{\gamma-1} \|v_1 - v_2\| \le \frac{1}{2} \|v_1 - v_2\|.$$

So, by the fixed-point theorem, the integral equation $T(v) = v$ and thus the differential equation (3.6) has a unique solution. Moreover, this solution can be bounded by

$$|v(\tau, \varepsilon)| \le L\varepsilon^{(2/3)\gamma},$$

for $\tau \in D_{\tau^*}$.

Finally, using that $v = p - p_0$ we finish the proof of the theorem. □

As we have seen, Theorem 3.2 gives us a bound of the function $w(x, \varepsilon)$ on the right side of the inner domain. In fact, at the point $\tilde{x}^*$ symmetric of x^*, we have

$$\left| w(\tilde{x}^*, \varepsilon) - \frac{1}{\varepsilon} p_0 \left(\frac{\tilde{x}^* - x_0}{\varepsilon} \right) \right| \le L\varepsilon^{(2/3)\gamma - 1}, \tag{3.8}$$

which will be used in the next section.

4. The solution in the outer right domain. In this section, we will extend the solution $w(x, \varepsilon)$ from the end point $\tilde{x}^*$ of the inner domain up to $+\infty$. We will do this comparing $w(x, \varepsilon)$ with the solution $\tilde{w}(x, \varepsilon)$ of (1.9) such that $\lim \tilde{w}(x, \varepsilon) = 0$, for Re $(x) \to +\infty$. The existence and the properties of $\tilde{w}(x, \varepsilon)$ are analogous to $w(x, \varepsilon)$ now considering x belonging to the outer right domain $\tilde{\Gamma}_{\varepsilon^\gamma}$.

All of this is summarized in the following theorem.

THEOREM 4.1. *Let us take $\delta > 0$. The Riccati equation* (1.9) *defined for $x \in \tilde{\Gamma}_{\varepsilon^\gamma}$, $0 < \gamma < 1 - \delta$, and $\varepsilon > 0$ sufficiently small has a unique solution $\tilde{w}(x, \varepsilon)$ such that $\lim \tilde{w}(x, \varepsilon) = 0$ when* Re $(x) \to +\infty$. *Furthermore, the solution $\tilde{w}(x, \varepsilon)$ satisfies that*

$$\left| \tilde{w}^{k)}(x, \varepsilon) - \sum_{n=0}^{N} \varepsilon^n w_n^{k)}(x) \right| \le K_{N,k}\ \varepsilon^{-(\gamma+\delta)} \varepsilon^{(N+1)(1-\gamma)-\gamma k}, \quad k \in \mathbf{N},$$

for all $x \in \tilde{\Gamma}_{\varepsilon^\gamma}$, where $K_{N,k}$ are constants independent of ε and of δ, and $w_n(x)$ are the functions given in Proposition 2.1.

Proof. It is analogous to the proof of Theorem 2.3. □

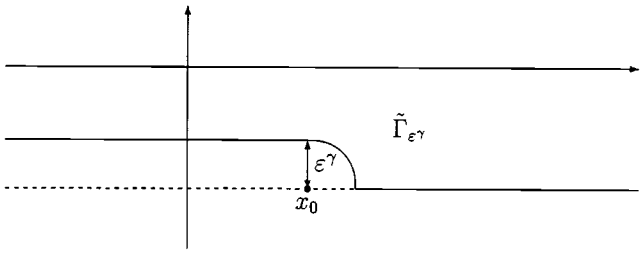

FIG. 3.

Remark. As in the previous section, in order to simplify the exposition, we will take from now on $0 < \gamma < 1/2$.

Now, let us take again $\tilde{x}^* = x_0 + \varepsilon\tilde{\tau}^*$. We want to have an estimate of $w(\tilde{x}^*, \varepsilon)$ in order to consider it as an initial condition to extend $w(x, \varepsilon)$. As we have seen in (3.8),

$$\left| w(\tilde{x}^*, \varepsilon) - \frac{1}{\varepsilon} p_0 \left(\frac{\tilde{x}^* - x_0}{\varepsilon} \right) \right| \leq L\varepsilon^{(2/3)\gamma - 1},$$

and from (3.4) with $\tau = \tilde{\tau}^*$ it follows that

$$\left| p_0 \left(\frac{\tilde{x}^* - x_0}{\varepsilon} \right) + ie^{2i(\tilde{x}^* - x_0)/\varepsilon} - \tilde{p}_0 \left(\frac{\tilde{x}^* - x_0}{\varepsilon} \right) \right| \leq C_\alpha \varepsilon^{1-\gamma}.$$

On the other hand, using Theorem 3.1 and 4.1, the asymptotic expression of f, and that $0 < \gamma < 1/2$ we obtain an analogous formula to (3.7), i.e.,

$$|\tilde{w}(\tilde{x}^*, \varepsilon) - \frac{1}{\varepsilon} \tilde{p}_0 \left(\frac{\tilde{x}^* - x_0}{\varepsilon} \right)| \leq 2\tilde{K}^* \varepsilon^{-\gamma/3}.$$

Considering this altogether, one can obtain that

$$\left| w(\tilde{x}^*, \varepsilon) - \tilde{w}(\tilde{x}^*, \varepsilon) + \frac{i}{\varepsilon} e^{2i(\tilde{x}^* - x_0)/\varepsilon} \right| \leq 3L\varepsilon^{(2/3)\gamma - 1}. \tag{4.1}$$

Now we are willing to prove the following theorem.

THEOREM 4.2. *The solution $w(x, \varepsilon)$ of (1.9) exists for $x \in \mathbf{C}$ such that* Re $(\tilde{x}^*) \leq$ Re (x) *and* Im $(x_0) + \varepsilon \leq$ Im $(x) \leq 0$ *and verifies*

$$|w(x, \varepsilon) - \tilde{w}(x, \varepsilon) + \frac{i}{\varepsilon} e^{(2i/\varepsilon)(x - x_0)}| \leq Ce^{-2/\varepsilon \text{Im } (x - x_0)} \varepsilon^{(2/3)\gamma - 1}. \tag{4.2}$$

In order to prove this theorem we need the following lemma.

LEMMA 4.3. *For $x \in \mathbf{C}$ such that* Re $(x) \geq$ Re $(\tilde{x}^*)$ *and* Im $(x) \geq$ Im $(x_0 + \varepsilon)$ *the following bounds hold:*

i. $|\int_{\tilde{x}^*}^{x} \varepsilon f(t)\tilde{w}(t, \varepsilon)dt| \leq C\varepsilon^{1-\gamma}$,

ii. $\left| \int_{\tilde{x}^*}^{x} e^{-(2i/\varepsilon)(\tilde{x}^* - s) + \int_s^{\tilde{x}^*} 2\varepsilon f(t)\tilde{w}(t, \varepsilon)dt} \varepsilon f(s)ds \right| \leq \varepsilon^{2-\gamma}$,

where C is a constant independent of ε.

Proof. The first bound follows inmediately from hypothesis **H4** and the asymptoticity of $\tilde{w}$ given in Theorem 4.1. For the second one it is sufficient to integrate by parts. □

Proof (of Theorem 4.2). Let us define $z(x,\varepsilon) := w(x,\varepsilon) - \tilde{w}(x,\varepsilon)$. From (1.9) and (4.1), $z(x,\varepsilon)$ verifies

$$\varepsilon z'(x,\varepsilon) = (2i - 2\varepsilon^2 f(x)\tilde{w}(x,\varepsilon))z(x,\varepsilon) - \varepsilon^2 f(x) z^2$$

and

$$\left| z(\tilde{x}^*,\varepsilon) + \frac{i}{\varepsilon} e^{2i(\tilde{x}^* - x_0)/\varepsilon} \right| \leq 3L\varepsilon^{2/3\gamma - 1}.$$

Thus, noting that z is the solution of a Bernoulli equation one can obtain the following integral expression for z:

$$z(x,\varepsilon) = \frac{e^{(2i/\varepsilon)(x-\tilde{x}^*) - \int_{\tilde{x}^*}^{x} 2\varepsilon f(t)\tilde{w}(t,\varepsilon)dt} z(\tilde{x}^*,\varepsilon)}{1 + z(\tilde{x}^*) \displaystyle\int_{\tilde{x}^*}^{x} e^{-(2i/\varepsilon)(\tilde{x}^* - s) + \int_{s}^{\tilde{x}^*} 2\varepsilon f(t)\tilde{w}(t,\varepsilon)dt} \varepsilon f(s) ds}. \tag{4.3}$$

Now, using the bounds given by Lemma 4.3 and (4.3), there exist some constants $\tilde{C}_i$, $i = 1, 2, 3$, independent of ε such that

$$|z(x,\varepsilon) - e^{(2i/\varepsilon)(x-\tilde{x}^*)} z(\tilde{x}^*,\varepsilon)| \leq \tilde{C}_1 \varepsilon^{-\gamma} |e^{(2i/\varepsilon)(x-\tilde{x}^*)}|$$

and then

$$\left| z(x,\varepsilon) + \frac{i}{\varepsilon} e^{(2i/\varepsilon)(x-x_0)} \right| \leq \tilde{C}_2 \varepsilon^{2/3\gamma - 1} |e^{(2i/\varepsilon)(x-\tilde{x}^*)}|.$$

Now, using that $\text{Im}\ (\tilde{x}^*) = \text{Im}\ (x_0) + \varepsilon$, we obtain

$$\left| z(x,\varepsilon) + \frac{i}{\varepsilon} e^{(2i/\varepsilon)(x-x_0)} \right| \leq \tilde{C}_3 \varepsilon^{2/3\gamma - 1} |e^{(2i/\varepsilon)(x-x_0)}| = \tilde{C}_3 \varepsilon^{2/3\gamma - 1} e^{-2/\varepsilon \text{Im}\ (x-x_0)}.$$

Finally, taking into account that $z(x,\varepsilon) = w(x,\varepsilon) - \tilde{w}(x,\varepsilon)$ we obtain the theorem. ∎

Now we are in a position to prove the main Theorem 1.1 by taking the limit as $\text{Re}\ (x) \to +\infty$ in inequality (4.2):

$$\left| \lim_{\text{Re}\ (x) \to +\infty} e^{-(2i/\varepsilon)x} w(x,\varepsilon) + \frac{i}{\varepsilon} e^{-(2i/\varepsilon)x_0} \right| \leq e^{2\text{Im}\ (x_0)/\varepsilon} \varepsilon^{(2/3)\gamma - 1}.$$

5. Resurgence of the solutions of the inner equation – Proof of Theorem 3.1.

5.1. Introduction. In this part of the article we provide a self-contained introduction to Écalle's theory of resurgent functions, and we show how our inner problem (3.3) fits within this framework.

We already know a formal solution to it:

$$\sum_{n \geq 0} a_n \tau^{-n-1},$$

and it is easy to see that there is no other formal solution. Our goal is to prove Theorem 3.1.

After the change of variable

$$z = 2i\tau,$$

(3.3) may be viewed as a particular case of singular Riccati equation of the type

$$\frac{dY}{dz} = Y + H^-(z) + H^+(z)Y^2 \tag{5.1}$$

where $H^\pm \in z^{-1}\mathbf{C}\{z^{-1}\}$ (analytic germs at infinity, vanishing at infinity); in our case, $H^-(z) = -H^+(z) = 1/6z$.

Now, resurgence is a good tool for analyzing all the equations of this kind; in fact, Écalle's theory allows the analytic classification of local equations in far more general contexts [4, 5, 2] (Of course, resurgence is not the only possible approach; see [10, 11] for another method of classifying singular local equations.), but the study of (5.1) provides a nice elementary introduction to some aspects of Écalle's work, even if many simplifications arise in the case of Riccati equations.

5.2. Singular Riccati equations and resurgence.

5.2.1. Resurgence of the formal solution. In the sequel, H^+ and H^- are two fixed analytic germs, vanishing at infinity. As (3.3), (5.1) admits a unique solution among formal expansions in negative powers of the variable; let's denote $Y_- \in z^{-1}\mathbf{C}[[z^{-1}]]$ this unique formal solution. We shall show that it is generically divergent using formal Borel transform $\mathcal{B}$.

The linear mapping $\mathcal{B}$ is defined by

$$\begin{cases} z^{-1}\mathbf{C}[[z^{-1}]] & \to & \mathbf{C}[[\zeta]], \\ z^{-n-1} & \mapsto & \zeta^n/n! \end{cases}$$

and it induces an isomorphism between the multiplicative algebra of Gevrey-1[1] formal series (without constant term) and the convolutive algebra of analytic germs at the origin $\mathbf{C}\{\zeta\}$, that is,

$$\begin{aligned} \varphi_1(z) &\mapsto \hat{\varphi}_1(\zeta), \\ \varphi_2(z) &\mapsto \hat{\varphi}_2(\zeta), \\ \varphi_1(z)\varphi_2(z) &\mapsto \hat{\varphi}_1 * \hat{\varphi}_2(\zeta) = \int_0^\zeta \varphi_1(\zeta_1)\varphi_2(\zeta-\zeta_1)d\zeta_1. \end{aligned}$$

Moreover, a formal series $\varphi(z)$ converges for $|z| > \rho$ if and only if its Borel transform is an entire function of exponential type at most ρ : $|\hat{\varphi}(\zeta)| \leq \text{const } e^{\rho|\zeta|}$. Hence, finite radius of convergence for $\hat{\varphi}$ (i.e., existence of singularities in ζ-plane) means divergence for φ.

We shall call *resurgent function* a Gevrey-1 formal series φ whose Borel transform has the following property: on any broken line issuing from the origin, there is a finite set of points such that $\hat{\varphi}$ may be continued analytically along any path that closely follows the broken line in the forward direction, while circumventing (to the right or to the left) those singular points. A nontrivial fact is the stability under convolution of this property. Indeed, resurgent functions form an algebra which can be considered either as a subalgebra of $\mathbf{C}[[z^{-1}]]$ (*formal model*) or, via $\mathcal{B}$, as a subalgebra of $\mathbf{C}\{\zeta\}$

[1]Let us recall that a formal power series $\sum_{n\geq 0} a_n\tau^{-n-1}$ is said to be Gevrey-1 if there exist two positive constants M, K such that $|a_n| \leq Mn!K^n$.

(*convolutive model*). The Borel transform of a given resurgent function is often called its *minor*.

PROPOSITION 5.1. *The formal solution of* (5.1) *is a resurgent function, with singularities in the convolutive model over the negative integers only.*

Proof. We start by performing Borel transform on (5.1) itself; differentiation with respect to z yields multiplication by $-\zeta$ and we obtain an equation for $\hat{Y}_-$:

$$-(\zeta+1)\hat{Y}(\zeta) = \hat{H}^- + \hat{H}^+ * \hat{Y}^{*2}, \tag{5.2}$$

where $\hat{H}^+$ and $\hat{H}^-$ are some entire series with infinite radius of convergence.

Let's define inductively a sequence of $\mathbf{C}[[\zeta]]$ by

- $\hat{Y}_0(\zeta) = -\hat{H}^-(\zeta)/(\zeta+1)$,
- $\forall n \geq 1$,

$$\hat{Y}_n(\zeta) = \frac{-1}{\zeta+1}\left(\hat{H}^+ * \sum_{n_1+n_2=n-1} \hat{Y}_{n_1} * \hat{Y}_{n_2}\right).$$

The valuation of $\hat{Y}_n$ being at least $2n$, the series $\sum_{n\geq 0}\hat{Y}_n$ converges formally in $\mathbf{C}[[\zeta]]$ towards the unique solution $\hat{Y}_-$ of (5.2). Now we observe that $\hat{H}^+$ and $\hat{H}^-$ define entire functions of at most exponential growth in any direction; $\hat{Y}_0$ defines thus a meromorphic function with a simple pole at -1, and, by successive convolutions, we only get for the $\hat{Y}_n$'s other simple poles at the negative integers together with ramification (logarithmic singularities).

In particular, for each integer n, $\hat{Y}_n$ is analytic in the universal covering of $\mathbf{C}\setminus(-\mathbf{N}^*)$; with some technical but easy work, one can prove that the series of holomorphic functions $\sum\hat{Y}_n$ is uniformly convergent in every compact subset of this universal covering. Therefore, $\hat{Y}_-$ is convergent at the origin and satisfies the required property of Borel transforms of resurgent functions. □

Remark 1. The definition of a general resurgent function doesn't impose anything on the nature of singularities one may encounter in following the analytic continuation of its minor and visiting the various leaves of its Riemann surface. We shall call *simple resurgent function* a resurgent function $\varphi(z)$ whose minor $\hat{\varphi}(\zeta)$ has only singularities of the form

$$\hat{\varphi}(\omega+\zeta) = \frac{c}{2\pi i\zeta} + \hat{\psi}(\zeta)\frac{\log\zeta}{2\pi i} + \hat{R}(\zeta)$$

with $c\in\mathbf{C}$ and $\hat{\psi},\hat{R}\in\mathbf{C}\{\zeta\}$. Simple resurgent functions form a subalgebra, which contains Y_- and all the other resurgent functions that appear in the sequel.

Remark 2. When writing in details the proof of Proposition 5.1, one obtains, in fact, exponential bounds in any sector $S_\alpha^+ = \{\zeta\in\mathbf{C}^*/-\pi+\alpha\leq\arg\zeta\leq\pi-\alpha\}$ (α being a small positive angle):

$$\forall\zeta\in S_\alpha^+,\ |\hat{Y}_-(\zeta)|\leq \text{const}\ e^{\rho|\zeta|},$$

where the positive number ρ depends on the radii of convergence of H^+ and H^- and on α. This allows us to apply *Laplace transform* in any direction different from the direction of $\mathbf{R}^-$.

Laplace transform in a direction θ is defined by

$$\mathcal{L}^\theta :\ \hat{\varphi}(\zeta)\mapsto\varphi^\theta(z) = \int_0^{e^{i\theta}\infty}\hat{\varphi}(\zeta)e^{-z\zeta}d\zeta.$$

When applied to an analytic function of exponential type at most ρ in direction θ, it yields a function φ^θ analytic in a half-plane bisected by the conjugate direction: Re $(ze^{i\theta}) > \rho$. If $\hat{\varphi}$ has at most exponential growth and no singularity in a sector of aperture α (in ζ-plane), by moving the direction of integration and using Cauchy theorem, we get a function analytic in a sectorial neighborhood of infinity of aperture $\pi+\alpha$ (in z-plane)[2]; moreover, in this neighborhood, $\varphi^\theta(z)$ is asymptotic in Gevrey-1 sense[3] to the formal series $\varphi = \mathcal{B}^{-1}\hat{\varphi}$ (a series which is the result of termwise application of Laplace transform to the Taylor series of $\hat{\varphi}$: $\mathcal{B}^{-1}$ is, in fact, the *formal Laplace transform*).

So, by choosing different values for θ, it is possible to associate to the formal series $\varphi(z)$ a family of sectorial germs $\{\varphi^\theta(z)\}$. When the series φ is convergent, the different φ^θ's yield the same analytic germ at infinity: the sum of φ. In general, the passage from φ to φ^θ through $\mathcal{L}^\theta \circ \mathcal{B}$ may be considered as a *resummation* process, since multiplication of formal series is taken to multiplication of sectorial germs, and differentiation w.r.t. z is respected too. We sum up this Borel–Laplace process in the following diagram.

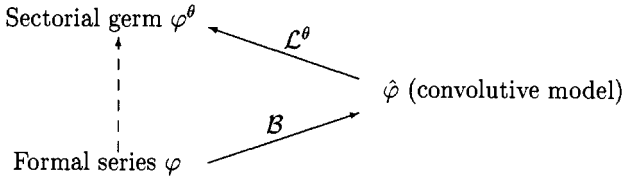

FIG. 4.

Applying Laplace transform $\mathcal{L}^\theta$ to $\hat{Y}_-$ with $\theta \in]-\pi, \pi[$, we get an analytic function defined in a sectorial neighborhood of infinity of aperture 3π in z-plane, which is a solution of (5.1). In particular, we have two possible summations of the formal solution Y_- in the half-plane $\{\text{Re } z < 0\}$ near infinity: Y_-^θ with θ close to π, and $Y_-^{\theta'}$ with θ' close to $-\pi$. These functions correspond respectively to the solutions $p_0(\tau)$ and $\tilde{p}_0(\tau)$ Theorem 3.1 refers to.

The question now is to compute the difference $Y_-^{\theta'} - Y_-^\theta$; we shall do it by analyzing the singularities of the minor $\hat{Y}_-$.

5.2.2. Formal integral. Before that, we will study a formal object, more general than the formal solution Y_-, which solves (5.1) too: the *formal integral*. We shall see that Y_- is the first term of a sequence $(\phi_n(z))$ of simple resurgent functions such that

$$Y(z,u) = \sum_{n\geq 0} u^n e^{nz} \phi_n(z) \in \mathbf{C}[[z^{-1}, ue^z]]$$

[2]This is a subset of $\mathbf{C}$ which contains, for all $\delta \in]0, \pi+\alpha[$, a sector $\{z \in \mathbf{C} / |\arg(ze^{i\theta})| < \delta/2\,,\ |z| > \rho\}$ for some positive ρ.

[3] If $\varphi(z) = \sum \varphi_n z^{-n-1}$, this means that in every closed subsector $\bar{S}$ of the domain, there exist $C, K > 0$ such that:

$$\forall N \geq 1, \forall z \in \bar{S},\ |z|^{N+1} |\varphi^\theta(z) - \sum_{n=0}^{N-1} \varphi_n z^{-n-1}| \leq CK^N N!.$$

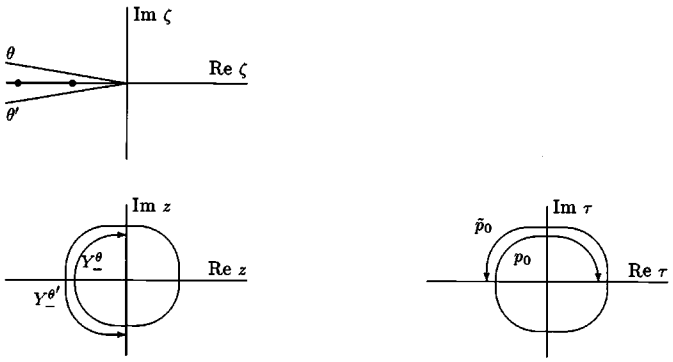

FIG. 5.

formally satisfies the equation. This means that (5.1) is formally conjugated to

$$\frac{dX}{dz} = X \tag{5.3}$$

through the formal diffeomorphism

$$Y = \Phi(z, X) = \sum_{n \geq 0} X^n \phi_n(z) \in \mathbf{C}[[z^{-1}, X]] \,.$$

Due to the fact that we deal with a Riccati equation, the formal integral admits a simple expression.

PROPOSITION 5.2. *There are formal series* $Y_+ \in z^{-1}\mathbf{C}[[z^{-1}]]$ *and* $Y_0(z) \in 1 + z^{-1}\mathbf{C}[[z^{-1}]]$ *such that*

$$Y(z,u) = \frac{ue^z Y_0(z) + Y_-(z)}{ue^z Y_0(z) Y_+(z) + 1}$$

formally solves (5.1). *Like* Y_-, *these formal series are simple resurgent functions;* [4] *their Borel transforms have singularities over* $\mathbf{Z}$ *only, and at most exponential growth at infinity.*

Proof. First we observe that our equation is equivalent to

$$-\frac{d}{dz}(1/Y) = 1/Y + H^+(z) + H^-(z)(1/Y)^2 \,. \tag{5.4}$$

Thus, using the same arguments that we used for proving Proposition 5.1, we see that there is a unique formal series $Y_+ \in z^{-1}\mathbf{C}[[z^{-1}]]$ whose inverse solves (5.1), and that it is a simple resurgent function whose Borel transform has singularities over the positive integers only and at most exponential growth.

Expecting a linear fractional dependence on the free parameter, we perform the change of unknown function

$$Y = \frac{a + Y_-(z)}{aY_+(z) + 1}.$$

[4] The definition of resurgent functions can be extended to allow them to have a constant term. Being the unity of the convolution, the Borel transform of 1 may be considered as the Dirac distribution δ at $\zeta = 0$. If $\varphi = c + \psi$ is a resurgent function of constant term c, its Borel transform is $\mathcal{B}\varphi = c\delta + \mathcal{B}\psi$, but we still call minor the germ $\hat{\varphi} = \mathcal{B}\psi$. See section 5.3 for one further generalization.

It yields the equation $da/dz = a(1 + H^+Y_- + H^-Y_+)$: the general solution is $a = ue^{z+\alpha(z)}$, where α is the unique formal series without constant term of derivative $H^+Y_- + H^-Y_+$. The series α is a simple resurgent function; its Borel transform

$$\hat{\alpha}(\zeta) = -\frac{1}{\zeta}(\hat{H}^+ * \hat{Y}_- + \hat{H}^- * \hat{Y}_+)$$

has singularities over $\mathbf{Z}^*$ only and at most exponential growth. Its exponential $Y_0 = e^\alpha$ inherits this property, by general properties of exponentiation of resurgent functions ([4, 2] : Y_0 has constant term 1 and its minor $\hat{Y}_0(\zeta) = \sum_{n\geq 1} \hat{\alpha}^{*n}/n!$ is analytic in the universal covering of $\mathbf{C} \setminus \mathbf{Z}$, with no singularity at the origin on the main sheet). □

So, we have $Y(z,u) = \sum_{n\geq 0} u^n e^{nz}\phi_n(z)$ with $\phi_0 = Y_-$, and for positive n,

$$\phi_n = (-1)^{n-1}Y_0^n Y_+^{n-1}(1 - Y_-Y_+)\,.$$

If we apply Laplace transform in a nonsingular direction θ, we obtain a one-parameter family of analytic solutions of (5.1):

$$Y^\theta(z,u) = \sum_{n\geq 0} u^n e^{nz}\mathcal{L}^\theta\hat{\phi}_n = \frac{ue^z Y_0^\theta(z) + Y_-^\theta(z)}{ue^z Y_0^\theta(z)Y_+^\theta(z) + 1},$$

defined for $\mathrm{Re}\,(ze^{i\theta}) - \rho > \text{const.}|ue^z|$ (a condition meant to ensure that the Laplace transforms of $\hat{Y}_0$, $\hat{Y}_-$, $\hat{Y}_+$ are defined and that the denominator in $Y^\theta(z,u)$ does not vanish).

In the convolutive model, we can apply Cauchy theorem and move the direction of integration in the upper or lower half-plane (depending on the value of θ). This provides an analytic continuation of $Y^\theta(z,u)$ allowing z to vary in a sectorial neighborhood of infinity of aperture 2π, that we call $Y^+(z,u)$ or $Y^-(z,u)$ as illustrated in the following diagram.

FIG. 6.

Thus, we essentially have two one-parameter families of analytic solutions of (5.1), characterized by their asymptotic behavior in the above-mentioned domains in z-plane. There must be some connection between them: a member $Y^+(.\,,u)$ of the first family has to coincide with some member $Y^-(.\,,u')$ of the other family for values of z with negative real part, and with some solution $Y^-(.\,,u'')$ for values of z with positive real part. These connection formulae will be computed in the following sections.

We are especially concerned with the functions $Y^+(z,0)$ and $Y^-(z,0)$ which correspond respectively to $p_0(\tau)$ and $\tilde{p}_0(\tau)$. At this stage, the first two statements of Theorem 3.1 are proved; the unicity assertion for the second one, for instance, is a consequence of the following easy lemma.

LEMMA 5.3. *If $u \in \mathbf{C}^*$, $Y^-(z,u)$ is defined for* $\mathrm{Re}\ z \leq 0$, $\mathrm{Im}\ z \geq 0$, *and $|z|$ big enough, and*

$$Y^-(z,u) - Y^-(z,0) = ue^z(1 + O(z^{-1})).$$

Indeed any solution of (5.1) analytic in a neighborhood of $i.\infty$ on the imaginary axis has to coincide with some function $Y^-(z,u)$; only one tends to 0 as Im z tends to infinity, and it corresponds to $u=0$.

And now we see that in order to prove the last statement of the theorem, we simply need to compute the value u_0 of the parameter such that $Y^+(z,0)=Y^-(z,u_0)$ for Re $z<0$ and to apply the lemma.

5.2.3. Alien calculus. It is essential to be able to analyze the singularities that appear in the convolutive model, for they are responsible for the divergence in the formal model. This is done by means of *alien calculus*, one of the main features of Écalle's theory, which relies on a new family of derivations: the *alien derivations*. Let's introduce them in the case of simple resurgent functions.

Let ω be in $\mathbf{C}^*$. We define an operator Δ_ω in the following way: given a simple resurgent function $\varphi(z)$, let's try to follow the analytic continuation of its minor $\hat\varphi(\zeta)$ along the half-line issuing from the origin and passing by ω (the minor is defined by the Borel transform of φ without taking into account the constant term if there is any); on this line, there is an ordered sequence $(\omega_1,\omega_2,\ldots)$ of singular points to be circumvented. If $r\geq 1$, we obtain in this way 2^{r-1} determinations of the minor in the segment $]\omega_{r-1},\omega_r[$ (with the convention $\omega_0=0$ if $r=1$ — in this case, there is only one determination), and we denote them by

$$\hat\varphi^{\varepsilon_1,\ldots,\varepsilon_{r-1}}_{\omega_1,\ldots,\omega_{r-1}},$$

each ε_i being a plus or minus sign indicating whether ω_i is circumvented to the right or to the left:

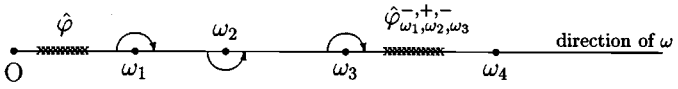

FIG. 7.

- If $\omega\notin\{\omega_1,\omega_2,\ldots\}$, we set

$$\Delta_\omega\varphi=0.$$

- If $\omega=\omega_r$ for $r\geq 1$, each of the above-mentioned determinations may have a singularity at ω:

$$\hat\varphi^{\varepsilon_1,\ldots,\varepsilon_{r-1}}_{\omega_1,\ldots,\omega_{r-1}}(\omega+\zeta)=\frac{c^{\varepsilon_1,\ldots,\varepsilon_{r-1}}_{\omega_1,\ldots,\omega_{r-1}}}{2\pi i\zeta}+\hat\psi^{\varepsilon_1,\ldots,\varepsilon_{r-1}}_{\omega_1,\ldots,\omega_{r-1}}(\zeta)\frac{\log\zeta}{2\pi i}+\text{ regular function},$$

and we set

$$\Delta_{\omega_r}\varphi=\sum_{\varepsilon_1,\ldots,\varepsilon_{r-1}}\frac{p(\varepsilon)!q(\varepsilon)!}{r!}\left(c^{\varepsilon_1,\ldots,\varepsilon_{r-1}}_{\omega_1,\ldots,\omega_{r-1}}+\mathcal{B}^{-1}\hat\psi^{\varepsilon_1,\ldots,\varepsilon_{r-1}}_{\omega_1,\ldots,\omega_{r-1}}\right),\tag{5.5}$$

where the integers p and $q=r-1-p$ are the numbers of plus signs and of minus signs in the sequence $(\varepsilon_1,\ldots,\varepsilon_{r-1})$.

It is easy to check the consistency of this definition. In some sense, $\Delta_\omega\varphi$ is a well-balanced average of the singularities of the determinations of the minor over ω; adding or removing false singularities in the list $(\omega_1,\omega_2,\ldots)$ would not affect the result, which

is a simple resurgent function (the definitions of section 5.2.1 were formulated exactly for this purpose).

The definition of operators Δ_ω for more general algebras of resurgence is given in [4, 5, 6]. In fact, these operators encode the whole singular behavior of the minor; given a sequence of points $(\omega_1, \dots, \omega_n)$ in $\mathbf{C}^*$, not necessarily on the same line, the composed operator $\Delta_{\omega_n} \cdots \Delta_{\omega_1}$ measures singularities over the point $\omega_1 + \cdots + \omega_n$.

The main property that makes these operators very useful in practice is the following: *the Δ_ω are derivations of the algebra of resurgent functions, i.e.,*

$$\forall \omega \in \mathbf{C}^*, \forall \varphi_1, \varphi_2 \text{ resurgent functions}, \ \Delta_\omega(\varphi_1\varphi_2) = (\Delta_\omega\varphi_1)\varphi_2 + \varphi_1(\Delta_\omega\varphi_2)\,.$$

By contrast with the natural derivation $\frac{d}{dz}$, they are called *alien derivations*.

Alien derivations interact with natural derivations according to the rule

$$\frac{d}{dz}\Delta_\omega\varphi = \Delta_\omega\frac{d\varphi}{dz} + \omega\Delta_\omega\varphi,$$

which reads

$$\frac{d}{dz}\dot{\Delta}_\omega\varphi = \dot{\Delta}_\omega\frac{d\varphi}{dz} \tag{5.6}$$

when one introduces the symbol $\dot{\Delta}_\omega = e^{-\omega z}\Delta_\omega$ (*pointed alien derivation*). But the $(\Delta_\omega)_{\omega\in\mathbf{C}^*}$ generate a free Lie algebra.

Finally, let's state the other property that we shall use: suppose that all the singularities of the minor of φ in a sector $\{\theta < \arg\zeta < \theta'\}$ form an ordered sequence $(\omega_1, \omega_2, \dots)$ on a half-line inside the sector

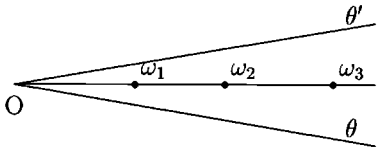

FIG. 8.

and that we can apply Borel–Laplace summation process, then

(5.7)

$$\varphi^\theta = \varphi^{\theta'} + \sum_{r\geq 1;\, i_1,\dots,i_r\geq 1} \frac{1}{r!} e^{-(\omega_{i_1}+\dots+\omega_{i_r})z}(\Delta_{\omega_{i_1}}\cdots\Delta_{\omega_{i_r}}\varphi)^{\theta'} = \left[\left(\exp\sum_{i\geq 1}\dot{\Delta}_{\omega_i}\right).\varphi\right]^{\theta'}$$

if we systematically use the notation ψ^θ for $\mathcal{L}^\theta\mathcal{B}\psi$, and $(e^{-\omega z}\psi)^\theta$ for $e^{-\omega z}\psi^\theta$.

5.2.4. Bridge equation. Coming back to our Riccati equation (5.1), let's try to compute the alien derivatives of the various simple resurgent functions that appear in the formal integral of Proposition 5.2. We shall use the generating series

$$Y(z,u) = \sum_{n\geq 0} u^n e^{nz}\phi_n(z)\,,\ \Delta_\omega Y = \sum_{n\geq 0} u^n e^{nz}\Delta_\omega\phi_n,$$

and we shall assume $\omega \in \mathbf{Z}^*$, since we already know that $\Delta_\omega Y$ vanishes if it is not the case.

Of course, it is equivalent to look for $\dot{\Delta}_\omega Y$, and it turns out that one can easily derive from (5.1) a deep relation, simply by applying $\dot{\Delta}_\omega$ to the equation itself. One obtains a linear equation for $\dot{\Delta}_\omega Y$

$$\frac{d}{dz}(\dot{\Delta}_\omega Y) = (1 + 2H^+ Y)\dot{\Delta}_\omega Y$$

(because pointed alien derivations commute with natural derivation, vanish on convergent series like $H^\pm$, and satisfy Leibniz rule), which admits $\partial Y/\partial u$ as a nontrivial solution, so that there must be some proportionality relationship

$$\dot{\Delta}_\omega Y = A_\omega(u)\frac{\partial Y}{\partial u}.$$

Simple arguments show that the coefficient $A_\omega(u)$ must be zero if $\omega \leq -2$ (because $\phi_1 \neq 0$, so the valuation of $\partial Y/\partial u$ w.r.t. e^z is exactly 1), that it is of the form $A_\omega(u) = A_\omega u^{\omega+1}$ (for homogeneity reasons), and finally that it is zero if $\omega \geq 2$ (Because one can repeat everything with (5.4); it's only here that we use the fact that (5.1) is a Riccati equation and not a more general nonlinear equation). We end up with the following proposition.

PROPOSITION 5.4. *There exist* $A^-, A^+ \in \mathbf{C}$ *such that*

$$\begin{aligned}
\dot{\Delta}_{-1} Y &= A^- \partial \frac{Y}{\partial u}, \\
\dot{\Delta}_{+1} Y &= -A^+ u^2 \partial \frac{Y}{\partial u}, \\
\dot{\Delta}_\omega Y &= 0 \text{ if } \omega \notin \{-1, +1\}.
\end{aligned}$$

So, the action of alien derivations on the formal integral is equivalent to the action of some differential operator: this important and very general result was called *bridge equation* by Écalle, since it throws a bridge between alien and ordinary calculus. When interpreted in the convolutive model, it expresses a strong link between an analytic germ at the origin and its singularities: in some ways, the germ reproduces itself at singular points, and this was the reason for naming "resurgent" such an object. Of course, with our definitions, not all resurgent functions have this property, but Écalle observed that it holds for all resurgent functions that arise "naturally" (as solutions of some analytic problem).

For instance, bridge equation holds for more general nonlinear equations, but in contrast with Riccati case, there can then be an infinity of numbers A_ω, $\omega \in \{-1, 1, 2, \ldots\}$; they are called *analytic invariants* of the equation, because it can be proven that two such equations are analytically (not only formally) conjugated if and only if they have the same set of A_ω's.

Thus, in our problem, inside the class of formally conjugated equations (5.1) (they are all conjugated to (5.3)), analytic classes are parametrized by a pair of two numbers.

Alien derivatives can be computed explicitly in terms of the two analytic invariants A^- and A^+:

$$\begin{aligned}
\Delta_{-1} Y_- &= A^- Y_0 (1 - Y_- Y_+), & \Delta_{+1} Y_- &= 0, \\
\Delta_{-1} Y_+ &= 0, & \Delta_{+1} Y_+ &= A^+ Y_0^{-1} (1 - Y_- Y_+), \\
\Delta_{-1} Y_0 &= -A^- Y_0^2 Y_+, & \Delta_{+1} Y_0 &= A^+ Y_-.
\end{aligned}$$

In particular, $\Delta_{\pm 1} Y_\pm = A^\pm + O(z^{-1})$, which means that $A^\pm$ is the residuum of $\hat{Y}_\pm(\zeta)$ at the point ± 1. These two numbers are transcendent functions of the convergent germs H^+ and H^-; we shall see later how to compute them in special cases. The vanishing of alien derivatives at integer points other than ± 1 does not mean that there is no singularity at those points: these other singularities can be detected by iterating bridge equation.

Bridge equation may be used for other purposes than analytic classification: formula (5.7) can be justified with $\varphi = Y(.\,, u)$, and we are now in a position to compare $Y^-(z,u)$ and $Y^+(z,u)$, the two families of solutions of (5.1) we obtained by resummation at the end of section 5.2.2. In the sequel, we shall take various values of z with big enough modulus and appropriated values of u in order to have $Y^\pm(z,u)$ defined.

If Re $z < 0$, applying formula (5.7) with $\theta < \pi < \theta'$ (and these angles both close to π) yields

$$Y^\theta(z,u) = [\exp \dot{\Delta}_{-1}]^{\theta'}(z,u) = \left[\exp\left(A^- \frac{\partial}{\partial u}\right)\right]^{\theta'}(z,u) = Y^{\theta'}(z, u + A^-)\,,$$

that is

$$Y^+(z,u) = Y^-(z, u + A^-)\,; \tag{5.8}$$

and similarly, if Re $z > 0$, choosing $\theta < 0 < \theta'$,

$$Y^-(z,u) = Y^+\left(z, \frac{u}{1 + A^+ u}\right). \tag{5.9}$$

Let's take $u = 0$: we already knew that $Y^+(z,0)$ and $Y^-(z,0)$ coincide for Re $z > 0$ (as was noticed at the end of section 5.2.1), but now, formula (5.8) and Lemma 5.3 show that, for Re $z < 0$ and Im $z > 0$,

$$Y^+(z,0) - Y^-(z,0) = A^- e^z (1 + O(z^{-1}))\,.$$

Finally, when the original variable $\tau = -iz/2$ has positive real and imaginary parts,

$$p_0(\tau) - \tilde{p}_0(\tau) = A^- e^{2i\tau}(1 + O(\tau^{-1}))\,. \tag{5.10}$$

5.3. Computation of analytic invariants. To complete the proof of Theorem 3.1, we only need to compute the coefficient A^- associated with an equation

$$\frac{dY}{dz} = Y + \frac{1}{6z}(1 - Y^2) \tag{5.11}$$

deduced from (3.3) by the change of variable $z = 2i\tau$, and to put it inside formula (5.10).

We shall, in fact, compute the pair of analytic invariants for all equations

$$\frac{dY}{dz} = Y - \frac{1}{2\pi i z}(B^- + B^+ Y^2), \tag{5.12}$$

where $B^\pm \in \mathbf{C}$. The result is proved in the second volume of [4], but we present here an alternative method.

PROPOSITION 5.5. *The analytic invariants of* (5.12) *are given by*

$$A^- = B^- \sigma(B^- B^+)\,,\ A^+ = -B^+ \sigma(B^- B^+),$$

where $\sigma(b) = \frac{2}{b^{1/2}} \sin \frac{b^{1/2}}{2}$. Note that this implies $A^- = -i$ in the case of (5.11), as required for ending the proof of the theorem.

Proof. Let's begin with a few simple remarks. The coefficient $A^{\pm}$ is the residuum at ± 1 of the Borel transform of $Y_{\pm}$, where $Y_-(z)$ is the unique formal solution of (5.12) and $Y_+(z)$ the unique formal solution of the equation corresponding to (5.4). It is easy to see that

$$Y_-(z) = B^- y(z)\,,\ Y_+(z) = -B^+ y(-z)\,,$$

where y is the unique formal solution of an equation depending only on $b = B^- B^+$:

$$\frac{dy}{dz} = y - \frac{1}{2\pi i z}(1 + by^2).$$

If we solve the Borel transform of this equation like we did in the proof of Proposition 5.1 (by expanding everything in powers of b), we find for the residuum of $\hat{y}(\zeta)$ at -1 an analytic function $\sigma(b)$ such that $\sigma(0) = 1$, and we obtain $A^- = B^- \sigma(b)$ and $A^+ = -B^+ \sigma(b)$.

Rather than studying the power expansion of $\sigma(b)$ (this is more or less what is done in [4, Vol. 2, pp. 476–480], but in a very efficient manner through the theory of *moulds*), we prefer to perform the change of unknown function

$$y(z) = \frac{2\pi i}{b} \cdot \frac{z q'(z)}{q(z)}\,,$$

which leads us to a second-order linear equation

$$z^2 q'' + (z - z^2) q' - \beta^2 q = 0$$

where $\beta^2 = \frac{b}{4\pi^2}$. We assume in the the sequel that $\operatorname{Re} \beta > 0$ (excluding real nonpositive values of b is innocuous since the function σ is analytic).

We exploit the peculiar form of this new equation, and write its unique formal solution with constant term 1 as the product of a monomial and expansion in fractional powers of z:

$$q(z) = z^{\beta} r(z) = 1 + O(z^{-1})\,,\ \ r(z) \in z^{-\beta}\mathbf{C}[[z^{-1}]]. \tag{5.13}$$

The series $r(z)$ may be called resurgent if we extend the definition of Borel transform by

$$z^{-\nu} \mapsto \zeta^{\nu-1}/\Gamma(\nu)\ ,\ \text{if}\ \ \nu \in \mathbf{C}\ \ \text{and}\ \ \operatorname{Re} \nu > 0,$$

and admit among resurgent functions all formal series (with possibly fractional powers) whose Borel transform, which may now be ramified at the origin and has endless analytic continuation. The convolution of minors is defined as before and we are still dealing with an algebra.

The point is that alien derivatives of r are easy to compute, for the equation it satisfies

$$r'' + (-1 + (2\beta + 1)z^{-1}) r' - \beta z^{-1} r = 0 \tag{5.14}$$

can be solved explicitly in the convolutive model.

LEMMA 5.6. *The Borel transform $\hat{r} = \mathcal{B}r$ is given by*

$$\hat{r}(\zeta) = \frac{\zeta^{\beta-1}}{\Gamma(\beta)}(1+\zeta)^{\beta}.$$

Proof. The Borel transform of (5.14)

$$(\zeta^2+\zeta)\hat{r} - (2\beta+1)1 * (\zeta\hat{r}) - \beta(1*\hat{r}) = 0$$

is equivalent to a first-order linear equation obtained by differentiation with respect to ζ (this was the only purpose of the change of unknown function (5.13)),

$$\zeta(\zeta+1)\frac{d\hat{r}}{d\zeta} = [(2\beta-1)\zeta+\beta-1]\hat{r}. \qquad \square$$

Alien calculus applies in this slightly generalized context. We only need to be careful about the determination of ζ^β we use (let's say it is the principal one) and about the sheet of the Riemann surface of the logarithm we look at. In particular, alien derivations are now indexed by points in this Riemann surface rather than by points in $\mathbf{C}^*$, and in order to compute $\Delta_{e^{i\pi}} r$ we perform a translation,

$$\hat{r}(e^{i\pi}+\zeta) = e^{i\pi(\beta-1)}\frac{\zeta^\beta}{\Gamma(\beta)}(1-\zeta)^{\beta-1},$$

and take the variation of the resulting singular germ (just as we were asked to retain the coefficient of $\log\zeta/2\pi i$ in the case of pure logarithmic singularities, according to formula (5.5)),

$$\mathcal{B}\,\Delta_{e^{i\pi}} r = -e^{i\pi\beta}(1-e^{2i\pi\beta})\frac{\zeta^\beta}{\Gamma(\beta)}(1-\zeta)^{\beta-1}.$$

We deduce that

$$\Delta_{e^{i\pi}} r = -e^{i\pi\beta}(1-e^{2i\pi\beta})\beta z^{-\beta-1}(1+O(z^{-1}))$$

and

$$\Delta_{-1}\, q = z^\beta \Delta_{e^{i\pi}} r = -2i\beta\sin(\pi\beta)z^{-1}(1+O(z^{-1})).$$

Finally we use Leibniz rule and formula (5.6):

$$\Delta_{-1}\, y = \frac{2\pi i}{b}\, z\, \Delta_{-1}(q'/q) = \frac{2}{b^{1/2}}\sin\frac{b^{1/2}}{2} + O(z^{-1}).$$

The constant term of the alien derivative is the residuum $\sigma(b)$. $\square$

6. Remarks on general nonlinear inner equations: The Kruskal–Segur strategy. Formula $p_0(\tau) - \tilde{p}_0(\tau) = -ie^{2i\tau}(1+O(\tau^{-1}))$ obtained in Theorem 3.1 for the inner equation (3.3) has been crucial for determining $w(+\infty,\varepsilon)$ and thus $\Delta I^2(\varepsilon)$, and this was the aim of the previous section. There (3.3) is studied by the resurgence theory obtaining the formula (5.10): $p_0(\tau)-\tilde{p}_0(\tau) = A^- e^{2i\tau}(1+O(\tau^{-1}))$. In order to compute exactly the coefficient $A^-(=-i)$ in the subsection 5.3, it is essential to be assured that (3.3) can be transformed into a second-order linear equation. However, in most applications (for example, for standard maps, see [7]) the inner equation is

not a Riccati equation and the method outlined in section 5 does not give quantitative results. Nevertheless, in the general case we can join the resurgence method with the Kruskal and Segur strategy [8]. This strategy adapted to our case would have the following form.

The main idea is that A^- can be computed looking at the coefficients of the formal solution of (3.3). Following the Kruskal–Segur strategy one can see that the growth of a_n is controlled comparing them with the coefficients b_n of the associated linear problem.

In this sense let $\sum_{n\geq 0} a_n \tau^{-n-1}$ be the formal solution of (3.3), which vanish at $-\infty$ (and, in fact, at $+\infty$), and let $\sum_{n\geq 0} b_n \tau^{-n-1}$ be the associated formal solution of the linear part of (3.3) $q_0' = 2iq_0 + (1)/6\tau$. We obtain that $b_n = (-1)^{n+1}(1)/12i(n!)/(2i)^n$ and as a first step in this method we would have to prove the following.

PROPOSITION 6.1. *$a_n = k_n b_n$, where $k_n = 3/\pi + O(\frac{1}{n})$, as $n \to \infty$.*

In our case we have numerical evidences of this result, and probably using the special form of our equation it could be proved analytically (see [17]). In this paper, we have not adopted this strategy because of the special form of our equation where this result is a consequence of the computation of $\Delta_{-1} Y_-$ in section 5.

Proposition 6.1 would be essential to control the behavior of the Borel transform of our solution. Let $\varphi(\xi) := \sum_{n\geq 0} (a_n \xi^n)/n!$ be the Borel transformation of $\sum_{n\geq 0} a_n \tau^{-n-1}$ and $\varphi_0(\xi) := 3/\pi \sum_{n\geq 0} (b_n \xi^n)/n! = -(1/2\pi)(1/2i + \xi)$ the Borel transformation of $3/\pi \sum_{n\geq 0} b_n \tau^{-n-1}$, and let us call $\varphi_1 := \varphi - \varphi_0$. Finally, let $f(\tau)$ be the Borel resummation of $\varphi(\xi)$, that is, its Laplace transform in some direction of the upper plane $Im\xi > 0$. As we know exactly $\varphi_0(\xi)$ in all the complex plane, this method allows us to compute its contribution to the resummation $f(\tau)$. Nevertheless, it would remain to prove that $\varphi_1(\xi)$ contributes to $f(\tau)$ only with higher-order terms. In order to prove this, it would be necessary to know the behavior of $\varphi_1(\xi)$, and, for our case, this is done in Proposition 6.2.

PROPOSITION 6.2.

i. *φ_0 has a unique singularity at $-2i$, which is a pole with residue $-\frac{1}{2\pi}$,*
ii. *φ_1 has logarithmic singularities at $-2i, -4i, -6i, \ldots$,*
iii. *Moreover, $f(\tau)$ is Gevrey-1 asymptotic to the formal solution $\sum_{n\geq 0} a_n \tau^{-n-1}$ in the sector $-3\pi/2 + \alpha \leq \arg \tau \leq \pi/2 - \alpha$, when $|\tau| \to \infty \ldots$*

An analogous proposition is studied in section 5 with the help of resurgent theory for our equation, and it seems that this can be generalized to other equations, as in [15]. Resurgent theory gives us the location of the singularities of φ_1, as well as their type and, consequently, their contribution to the "resummation" $f(\tau)$.

Finally, as a last step of this method, putting together Propositions 6.1 and 6.2 we would have the following.

PROPOSITION 6.3.

i. $f(\tau) = p_0(\tau)$ *if* $\pi/2 < \arg \tau < 2\pi$,
ii. $f(\tau) = \tilde{p}_0(\tau)$ *if* $-\pi < \arg \tau < \pi/2$.

Then, for τ such that $0 \leq \arg \tau < \pi/2$, one can use the analytic continuation of $f(\tau)$, and using these propositions and the Cauchy theorem, one can see that in our equation:

$$p_0(\tau) - \tilde{p}_0(\tau) = \int_{-\infty}^{+\infty} e^{-\tau s} \varphi(s) ds$$

$$= \int_{-\infty}^{+\infty} e^{-\tau s}\varphi_0(s)ds + \int_{-\infty}^{+\infty} e^{-\tau s}\varphi_1(s)ds$$
$$= e^{2i\tau}(-i + O(\tau^{-1})).$$

Observe that following the Kruskal–Segur strategy we can compute the coefficient $A^-/(2\pi i)$ as the residue of the Borel transform φ_0 at its pole $-2i$. We are convinced that the link between the resurgence approach and Kruskal–Segur strategy rests on the fact that, in general, all the successive approximations of the corresponding (5.2) have a pole at $-2i$. Then φ_0 would be the summation of all the polar parts at $-2i$ of those approximations, and its residue the sum of their residues (which corresponds to the coefficient A^- computed in Proposition 5.5).

REFERENCES

[1] V. I. Arnold, V. V. Kozlov, and A. I. Neishtadt, *Mathematical aspects of classical and celestial mechanics*, Dynamical Systems III, Encyclopaedia of Mathematical Sciences, Vol. 3, V. I. Arnold, ed., Springer-Verlag, Berlin, New York, 1988.

[2] B. Candelpergher, J.-C. Nosmas, and F. Pham, *Approche de la résurgence*, Actualités Math., Hermann, Paris, 1993.

[3] A. M. Dykhne, *Quantum transitions in the adiabatic approximation*, Sov. Phys. JETP, 11 (1960), pp. 411–415.

[4] J. Écalle, *Les fonctions résurgentes*, Prépub. Math., Université Paris-Sud, Orsay, 3 Vols., 1981, 1985.

[5] J. Écalle, *Cinq applications des fonctions résurgentes*, Prépub. Math., Université Paris-Sud, Orsay, 1984.

[6] J. Écalle, *Six lectures on transseries, analyzable functions and the constructive proof of Dulac's conjecture*, Bifurcations and Periodic Orbits of Vector Fields, D. Schlomiuk, ed., Kluwer Academic Publishers, Dordrecht, 1993, pp. 75–184.

[7] V. Hakim and K. Mallick, *Exponentially small splitting of separatrices, matching in the complex plane and Borel summation*, Nonlinearity, 6 (1992), pp. 57–70.

[8] M. Kruskal and H. Segur, *Asymptotics beyond all orders in a model of crystal growth*, Stud. Appl. Math., 85 (1991), pp. 129–181.

[9] J. E. Littlewood, *Lorentz's pendulum problem*, Ann. Physics, 21 (1963), pp. 233–242.

[10] J. Martinet and J.-P. Ramis, *Problèmes de modules pour des équations différentielles non linéaires du premier ordre*, IHES Publ. Math., 55 (1982), pp. 63–164.

[11] J. Martinet and J.-P. Ramis, *Classification analytique des équations différentielles non linéaires résonantes du premier ordre*, Ann. Sci. Ecole Norm. Sup., 4 (1983), pp. 571–621.

[12] R. E. Meyer, *Exponential asymptotics*, SIAM Rev., 22 (1980), pp. 213–224.

[13] R. E. Meyer, *Gradual reflection of short waves*, SIAM J. Appl. Math., 29 (1975), pp. 481–492.

[14] J. P. Ramis and R. Schäfke, *Gevrey separation of fast and slow variables*, Nonlinearity, 9 (1983), pp. 353–384.

[15] D. Sauzin, *Résurgence parametrique et exponentielle petitesse de l'écart des separatrices du pendule rapidement force*, Ann. Inst. Fourier (Grenoble), 2 (1995), pp. 453–511.

[16] A. A. Slutskin, *Motion of a one-dimensional nonlinear oscillator under adiabatic conditions*, Soviet Phys. JETP, 18 (1964), pp. 676–682.

[17] Y. B. Suris, *On the complex separatrices of some standard-like maps*, Nonlinearity, 7 (1994), pp. 1225–1236.

[18] W. Wasow, *Adiabatic invariance of a simple oscillator*, SIAM J. Math. Anal., 4 (1973), pp. 78–88.

[19] W. Wasow, *Calculation of an adiabatic invariant by turning point theory*, SIAM J. Math. Anal., 5 (1974), pp. 673–700.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1361–1380, November 1998

003

LOCALIZED SPATIAL HOMOGENIZATION AND LARGE DIFFUSION*

ANÍBAL RODRÍGUEZ-BERNAL†

Abstract. We analyze singular perturbations in elliptic equations, subjected to various boundary conditions, in which the diffusion is going to infinity in localized regions inside the domain and therefore solutions undergo a localized spatial homogenization. The limiting elliptic operator is analyzed as well as convergence of solutions, eigenvalues, and eigenvectors.

Key words. singular perturbation, eigenvalue problems, large diffusion

AMS subject classifications. 35J25, 35P, 73K20

PII. S003614109731864X

1. Introduction.

When dealing with the modeling of some physical problem it is often the case that there are some physical properties that differ very strongly from one part of the physical system to the other. As a consequence the constituting laws, although obeying a general formulation, have some significant differences depending on what part of the system one is looking at. This, in turn, implies that the differential equations describing the behavior of the system may also have different properties all along the system. In some cases this leads to the existence of some singularities in the equations and/or in the solutions. It is also often the case that one must study some singular limit problem because some parameter of the problem is varied.

For example, in past years numerous results have been given concerning the behavior of semilinear parabolic equations (or systems)

$$u_t - D\Delta u + f(u, \nabla u) = 0$$

with Neumann boundary conditions, where $u = (u_1, \dots, u_m)^T$ and $D = \mathrm{diag}(D_1 \dots, D_m)$ are the (constant) diffusion coefficients, [4, 9, 11, 12, 15]. In these cases the singular limit problem is obtained when one considers the case in which the diffusivity of the component species is very large, i.e., when $D_0 = \inf_i\{D_i\}$ is very large. Then one tries to understand the properties and behavior of the solutions as $D_0 \to \infty$. The results in those papers assert the physically clear fact that large diffusivity implies a very rapid redistribution of the spatial inhomogeneities of u and, therefore, u approaches, as time increases, a function constant in space. The dynamics of the "asymptotic set of states," i.e., constant functions, is given explicitly by the kinetic equation

$$\dot{u} + f(u, 0) = 0,$$

which is a system of ordinary differential equations (ODEs) in $\mathbb{R}^m$ that reflects the dynamics of the full reaction-diffusion system as D_0 increases. Therefore, the ODE is the true limiting problem for the singular limit $D_0 \to \infty$.

*Received by the editors March 19, 1997; accepted for publication (in revised form) October 17, 1997; published electronically June 25, 1998. Part of this research was carried out while the author was visiting the Instituto de Ciencias Matematicas de São Carlos-USP, São Paulo, Brazil. This research was supported by FAPESP grant 93/4733-7 and was partially supported by DGICYT grants PB93-0438 and PB96-0648 and EEC grant SC1-CT91-0732.

http://www.siam.org/journals/sima/29-6/31864.html

†Departamento de Matemática Aplicada, Universidad Complutense de Madrid, Madrid 28040, Spain (arober@sunma4.mat.ucm.es).

The same problem was studied in [13], when only some of the diffusion coefficients are sufficiently large. Thus in this case there are strong differences in the physical properties of the species (the diffusivities). Now the equations describing the behavior of the solutions, as some of the diffusion coefficients go to infinity, consist of a system of partial differential equations (PDEs) coupled with a system of ODEs for the components that tend to be constant in space functions. Those equations are called "shadow systems." Similar situations have been considered in [6] for the case of delayed semilinear parabolic equations or in [10, 5], where they described the large time dynamics of parabolic equations in one-dimensional domains with the diffusion coefficient being large except in a neighborhood of a finite number of points where it becomes small. The asymptotic dynamics of the parabolic problem are again described by a system of ODEs.

For the case of semilinear wave equations describing, for example, the vibrations of an elastic body [1, 8] when the wave speed is large everywhere, one also expects that the solutions will converge to constant in space functions, reflecting the property that the body tends to behave as a rigid body and then all points oscillate in phase with the same amplitude and frequency. In fact the trend to spatial homogeneity has also been established [18, 3] provided that the damping is large enough. In this case the result states that if vibrations propagate very rapidly, then in the limit an elastic body behaves as a rigid structure.

An indication of this property of spatial homogenization, when the diffusion is very large, can be obtained from the linear elliptic problem

$$\begin{cases} -\frac{1}{\epsilon}\Delta u^\epsilon + \lambda u^\epsilon = f, \\ \frac{\partial u^\epsilon}{\partial \vec{n}} = 0, \end{cases}$$

where after multiplying by ϵ one gets that u^ϵ must converge to the solution of

$$\begin{cases} -\Delta u = 0, \\ \frac{\partial u}{\partial \vec{n}} = 0. \end{cases}$$

Hence u must be constant. But integrating we get $\lambda \int_\Omega u^\epsilon = \int_\Omega f$, and in the limit $u = \frac{1}{\lambda} \fint_\Omega f$, where $\fint_\Omega = \frac{1}{|\Omega|}\int_\Omega$. Also, by looking at the eigenvalue problem

$$\begin{cases} -\frac{1}{\epsilon}\Delta u^\epsilon + \lambda u^\epsilon = \mu u^\epsilon, \\ \frac{\partial u^\epsilon}{\partial \vec{n}} = 0, \end{cases}$$

the first eigenvalue is $\mu_1^\epsilon = \lambda$ and the first eigenfunction is $\phi_1^\epsilon = |\Omega|^{-1/2}$. But the second eigenvalue is

$$\mu_2^\epsilon = \inf_{\substack{u \in H^1(\Omega) \\ \int u = 0 \\ \|u\| = 1}} \frac{1}{\epsilon}\int_\Omega |\nabla u|^2 + \lambda \int_\Omega |u|^2 \geq \frac{1}{\epsilon} c(\Omega) + \lambda \to \infty,$$

where we have used the Poincaré inequality for zero mean functions. These computations prove that only the first mode, i.e., the average of the solution, matters as ϵ goes to zero. Note that the previous arguments, applied to the case with Dirichlet boundary conditions, imply the solution converges to zero and all the eigenvalues converge to infinity. Therefore, boundary conditions also play a role in the determination of the limiting problem.

On the other hand, in this paper we are concerned with the case in which the diffusion coefficient becomes large in a localized region inside the physical domain of the differential equation, while it remains bounded away from zero in the rest. That situation can be found, for example, in composite materials, where the heat diffusion properties differ very strongly from one part to another of the material, or in population dynamics in which one species diffuses much faster than the others in some determined regions. Also, one can consider the vibrations of an elastic membrane and assume that some part of the membrane is made of a more rigid material than the rest. In the limit, one obtains a multistructure vibrating system composed of a rigid plate bound from its boundary to a membrane. In all these cases, passing to the singular limit can be regarded as a process describing a transition phenomena of rigidization or solidification in a composite material. In this situation one expects that in the "large diffusion" region, Ω_0, the solution tends to be space homogeneous and satisfies an ODE, while in the other part it must satisfy a PDE. There must be also some coupling between the two equations. Our goal is then to obtain the equations that describe the limiting problem and to study the properties and relationships between the approximating problems and the limit one.

We will be concerned with linear elliptic problems, under suitable boundary conditions, and we will analyze the behavior of solutions as $\epsilon \to 0$. We will also consider the corresponding eigenvalue problems. In this way we will obtain a description of the asymptotic behavior, as $\epsilon \to 0$, of the eigenmodes and eigenfrequencies of the system. To be more precise, we will consider the family of elliptic problems

$$(P_\epsilon)\quad \begin{cases} -Div(d_\epsilon(x)\nabla u^\epsilon) + (\lambda + V_\epsilon(x))u^\epsilon = f^\epsilon & \text{on } \Omega, \\ \frac{\partial u^\epsilon}{\partial \vec{n}_\epsilon} + b_\epsilon(x)u^\epsilon = g^\epsilon & \text{on } \Gamma, \end{cases} \tag{1.1}$$

where $0 < \epsilon \le \epsilon_0$, $\Omega \subset \mathbb{R}^N$ is a bounded regular open connected set, and Γ is the boundary of Ω. Note that $\frac{\partial u}{\partial \vec{n}_\epsilon}$ denotes the conormal derivative relative to the diffusion operator $-Div(d_\epsilon(x)\nabla u)$, i.e., $\frac{\partial u}{\partial \vec{n}_\epsilon} = d_\epsilon(x)\langle \nabla u, \vec{n}\rangle$. Also, $\lambda \in \mathbb{R}$ and the potentials $V_\epsilon(x)$ and $b_\epsilon(x)$ are given functions on Ω and Γ, respectively. Let $\Omega_0 = \cup_{i=1}^m \Omega_{0,i}$ be an open subset of Ω with boundary $\Gamma_0 = \cup_{i=1}^m \Gamma_{0,i}$ such that $\Omega_{0,i}$ is connected with boundary $\Gamma_{0,i}$, $\Gamma \cap \Gamma_0 = \emptyset$, and $\Gamma_{0,i} \cap \Gamma_{0,j} = \emptyset$ for $i \neq j$. We denote $\Omega_1 = \Omega \setminus \overline{\Omega_0}$. Note that the boundary of Ω_1 is $\Gamma \cup \Gamma_0$. The diffusion coefficients $d_\epsilon(x)$, $0 \le \epsilon \le \epsilon_0$, are assumed to be smooth, strictly positive functions on Ω, such that

$$0 < m_0 \le d_\epsilon(x) \le M_\epsilon \tag{1.2}$$

for every $x \in \Omega$ and $0 \le \epsilon \le \epsilon_0$. We also assume that the diffusion is very large on Ω_0 as ϵ approaches zero. More precisely, we assume

$$d_\epsilon(x) \longrightarrow \begin{cases} \infty & \text{uniformly on compact subsets of } \Omega_0, \\ d_0(x) & \text{uniformly on } \Omega_1 \end{cases} \tag{1.3}$$

as $\epsilon \to 0$. Finally, the potentials verify

$$V_\epsilon \to V \in L^{q_0}(\Omega) \quad \text{with } q_0 \begin{cases} > N/2 & \text{if } N \ge 3, \\ > 1 & \text{if } N = 2, \\ \ge 1 & \text{if } N = 1, \end{cases} \tag{1.4}$$

$$b_\epsilon \to b \in L^{q_1}(\Gamma) \quad \text{with } q_1 \begin{cases} > N-1 & \text{if } N \ge 3, \\ > 1 & \text{if } N = 2, \\ \ge 1 & \text{if } N = 1. \end{cases} \tag{1.5}$$

We will show that the family of solutions to these problems, $\{u^\epsilon\}_\epsilon$, converges, in some sense, to a function u that is constant in Ω_0 and that verifies a "limiting elliptic problem," the shadow system for the elliptic problem, which is of nonlocal nature. We will show that in the process of convergence there is a loss of regularity for the limiting solution by showing that typically the normal derivative of u jumps across Γ_0. Concerning the eigenvalue problem, we will show that the spectral properties of the approximate and limiting problem are close by showing that the eigenvalues and eigenfunctions converge.

The form of the shadow system is formally guessed in section 2. After introducing some notations in section 3, in section 4 we prove the convergence of solutions and in section 5 we prove convergence of eigenvalues and eigenfunctions. Finally, in section 6, we will consider the case of mixed or Dirichlet boundary conditions and different types of diffusion coefficients for which the same results hold true.

It is important to note here that the assumption

$$\Gamma \cap \Gamma_0 = \emptyset \tag{1.6}$$

that is, the diffusion is large in the interior of Ω, is crucial in the development of our analysis. The case $\Gamma \cap \Gamma_0 \neq \emptyset$ requires a different analysis that will be pursued somewhere else. In that case the interaction between diffusion and boundary conditions becomes more subtle.

In [7] the problem of convergence of solutions is analyzed for the corresponding parabolic equations.

2. The formal limiting problem. We will show some heuristic argument showing what the limiting, or "shadow" problem, should be. In doing this we will interpret (P_ϵ) in terms of a heat distribution equilibrium. Hence, $(\lambda + V_\epsilon(x))u$ accounts for the heat absorption in Ω, while f^ϵ accounts for the heat sources and $b_\epsilon(x)u - g^\epsilon(x)$ for the heat flow across the boundary.

Assume Ω_0 is connected, i.e., $m = 1$, f, g and V, b are fixed regular functions, i.e., they do not depend on ϵ. And assume that one has already obtained that the family of solutions to these problems, $\{u^\epsilon\}_\epsilon$, converges, in some sense, to a function u which is constant in Ω_0. Assume all functions live in the Sobolev space $H^1(\Omega)$. Note that if the limit function u is also in $H^1(\Omega)$, then its constant value on Ω_0, denoted u_{Ω_0}, cannot be arbitrary.

Since on Ω_1 we have $-Div(d_\epsilon(x)\nabla u^\epsilon) + (\lambda + V(x))u^\epsilon = f$ from the convergence properties of $d_\epsilon(x)$ in Ω_1, it seems reasonable to have in the limit

$$-Div(d_0(x)\nabla u) + (\lambda + V(x))u = f \quad \text{on } \Omega_1.$$

Analogously, on Γ, $\frac{\partial u^\epsilon}{\partial \vec{n}_\epsilon} + b(x)u^\epsilon = d_\epsilon(x)\langle \nabla u^\epsilon, \vec{n}\rangle + b(x)u^\epsilon = g(x)$ and then we expect to have in the limit

$$\frac{\partial u}{\partial \vec{n}_0} + b(x)u = d_0(x)\langle \nabla u, \vec{n}\rangle + b(x)u = g(x) \quad \text{on } \Gamma.$$

Also on Γ_0 we must have $u_{|\Gamma_0} = u_{\Omega_0}$.

On the other hand, integrating on Ω_0 we have $\int_{\Gamma_0} \frac{\partial u^\epsilon}{\partial \vec{n}_\epsilon} + \int_{\Omega_0} (\lambda + V(x))u^\epsilon = \int_{\Omega_0} f$, where we use the direction of the inward unit normal to Ω_0 in the surface integral. Therefore, passing to the limit we have

$$\int_{\Gamma_0} \frac{\partial u}{\partial \vec{n}_0} + u_{\Omega_0} \int_{\Omega_0} (\lambda + V(x)) = \int_{\Omega_0} f,$$

which is an equation relating the total heat flow from Ω_1 to Ω_0 through Γ_0 with the total heat absorbed in Ω_0 and the total heat input in Ω_0. It also relates the value of u_{Ω_0} with the values of u in Ω_1 through the integral term along Γ_0.

With this heuristic consideration and assuming that in the limit we will work with a space of functions constant on Ω_0, we can write the previous equalities as an "elliptic equation" for the function $u = u\mathcal{X}_{\Omega_1} + u_{\Omega_0}\mathcal{X}_{\Omega_0}$,

$$B_0(u) = (-Div(d_0(x)\nabla u) + (\lambda + V(x))u)\,\mathcal{X}_{\Omega_1} + \left(\frac{1}{|\Omega_0|}\int_{\Gamma_0}\frac{\partial u}{\partial \vec{n}_0} + \left(\fint_{\Omega_0}\lambda + V\right)u_{\Omega_0}\right)\mathcal{X}_{\Omega_0} = \hat{f}, \tag{2.1}$$

where $\fint_{\Omega_0} = \frac{1}{|\Omega_0|}\int_{\Omega_0}$ and $\hat{f} = f\mathcal{X}_{\Omega_1} + (\fint_{\Omega_0} f)\mathcal{X}_{\Omega_0}$, and boundary condition $\frac{\partial u}{\partial \vec{n}_0} + bu = g$. Note that f has been substituted in the limit by a "projection" $\hat{f}$. Also note the nonlocal coupling between the equation for $u\mathcal{X}_{\Omega_1}$ and u_{Ω_0}, given by the term $\frac{1}{|\Omega_0|}\int_{\Gamma_0}\frac{\partial u}{\partial \vec{n}_0}$, that represents the total heat flow from Ω_1 to Ω_0.

In the next sections we will show that all these formal computations can be justified, and, moreover, we will prove that the limiting "elliptic operator" has good functional properties.

3. Functional setting. Now we will introduce notation and several function spaces that will be used henceforth. Concerning functional spaces, we will use the standard Sobolev spaces $H^1(\Omega)$ and $H^{1/2}(\Gamma)$. Also, we will denote by H^{-s} the dual space of H^s, either on Ω or Γ. Note that this symbol is usually reserved to denote the dual space of H_0^s. However, this notation should produce no confusion. The duality pairing between the spaces above will be denoted $\langle\cdot,\cdot\rangle_{-s,s}$. In particular, the scalar product in L^2 will be denoted by $\langle\cdot,\cdot\rangle$. If there is no possible confusion, we will not indicate if the spaces or duality products are referred to as functions on Ω or Γ. When required, we will write $\langle\cdot,\cdot\rangle_\Omega$ and $\langle\cdot,\cdot\rangle_\Gamma$ to differentiate both cases.

We will denote by γ the trace operator defined on $H^1(\Omega)$, with values in $H^{1/2}(\Gamma)$. Analogously, we will also denote by γ_0 the trace operator defined in $H^1(\Omega)$ that restricts functions to Γ_0, i.e., $\gamma_0 : H^1(\Omega) \to H^{1/2}(\Gamma_0)$. Similarly, $\gamma_{0,i}$ will denote the trace operator on $\Gamma_{0,i}$. Moreover, for a given function $f \in H^1(\Omega)$, we will identify its trace, $\gamma(f) \in H^{1/2}(\Gamma)$, with the linear form $\gamma(f) \in H^{-1/2}(\Gamma) \subset H^{-1}(\Omega)$ such that for every $\phi \in H^1(\Omega)$,

$$\langle\gamma(f),\phi\rangle_{-1,1} \stackrel{def}{=} \langle f,\phi\rangle_\Gamma \stackrel{def}{=} \int_\Gamma f\phi;$$

that is, we use the embedding $H^{1/2}(\Gamma) \subset L^2(\Gamma) \subset H^{-1/2}(\Gamma) \subset H^{-1}(\Omega)$. The same holds for γ_0, i.e., $\gamma_0(f) \in H^{-1/2}(\Gamma_0) \subset H^{-1}(\Omega)$.

We will also consider the normal derivative operator, relative to the diffusion operator $-Div(d_\epsilon(x)\nabla u)$, defined as follows: let 2^* be the critical Sobolev exponent for the inclusion $H^1(\Omega) \subset L^p(\Omega)$; i.e.,

$$2^* = \begin{cases} \frac{2N}{N-2} & \text{if } N \geq 3, \\ \text{any positive number} & \text{if } N = 2, \\ \infty & \text{if } N = 1. \end{cases} \tag{3.1}$$

Hence $H^1(\Omega) \subset L^p(\Omega)$ with continuous inclusion if $p \leq 2^*$ and compact inclusion if $p < 2^*$. For any $p \leq 2^*$ denote $Y_{\epsilon,p} \stackrel{def}{=} \{z \in H^1(\Omega),\ -Div(d_\epsilon(x)\nabla z) \in L^{p'}(\Omega)\}$, with

$1/p + 1/p' = 1$. Then for $u \in Y_{\epsilon,p}$, $\frac{\partial u}{\partial \vec{n}_\epsilon} \in H^{-1/2}(\Gamma)$ is defined as

$$\left\langle \frac{\partial u}{\partial \vec{n}_\epsilon}, \gamma(v) \right\rangle_{-1/2,1/2} = \int_\Omega Div(d_\epsilon(x)\nabla u)v + \int_\Omega d_\epsilon(x)\nabla u \nabla v \tag{3.2}$$

for every $v \in H^1(\Omega)$. Observe that $Y_{\epsilon,p}$ is a Banach space for the norm

$$\|z\|_{Y_{\epsilon,p}} = \|z\|_{H^1(\Omega)} + \|Div(d_\epsilon \nabla z)\|_{L^{p'}(\Omega)},$$

which is, moreover, a Hilbert space if $p = 2$, and the mapping $z \mapsto \frac{\partial z}{\partial \vec{n}_\epsilon}$ is continuous between $Y_{\epsilon,p}$ and $H^{-1/2}(\Gamma)$. In the sequel we will often find the normal derivative operator on Γ_0 for functions defined either on Ω or Ω_1. When doing this, we will always take the direction of the *inner* normal to Ω_0; that is, whenever an expression of the form $\int_{\Gamma_0} \frac{\partial u}{\partial \vec{n}_\epsilon}$ appears we will assume the direction of the outward normal to Ω_1.

We now introduce a linear positive operator, L_ϵ, between $H^1(\Omega)$ and its dual, $H^{-1}(\Omega)$, such that for every $u, \phi \in H^1(\Omega)$,

$$a_\epsilon(u,\phi) = \langle L_\epsilon(u), \phi \rangle_{-1,1} = \int_\Omega d_\epsilon(x)\nabla u \nabla \phi. \tag{3.3}$$

Note that we can then write (3.2) as

$$\langle L_\epsilon(u), v \rangle_{-1,1} = \langle -Div(d_\epsilon(x)\nabla u), v \rangle + \left\langle \frac{\partial u}{\partial \vec{n}_\epsilon}, \gamma(v) \right\rangle_{-1/2,1/2} \tag{3.4}$$

for $u \in Y_{\epsilon,p}$ and $v \in H^1(\Omega)$. Also, for any $\lambda > 0$ we consider in $H^1(\Omega)$ the scalar product

$$\langle u, v \rangle_\epsilon = a_\epsilon(u,v) + \lambda \int_\Omega uv = \int_\Omega d_\epsilon \nabla u \nabla v + \lambda \int_\Omega uv = \langle L_\epsilon u + \lambda u, v \rangle_{-1,1}, \tag{3.5}$$

which, for fixed $0 < \epsilon \le \epsilon_0$, gives a norm, $\|\cdot\|_\epsilon$, equivalent to the usual one and defines the isomorphism between $H^1(\Omega)$ and its dual, $H^{-1}(\Omega)$, $L_\epsilon + \lambda I$. Note, however, that if we take the norm $\|u\|^2_{H^1(\Omega)} = \int_\Omega |\nabla u|^2 + \lambda \int_\Omega |u|^2$, then for $0 < \epsilon \le \epsilon_0$ we have from (1.2)

$$m_0 \|u\|^2_{H^1(\Omega)} \le \|u\|^2_\epsilon \le M_\epsilon \|u\|^2_{H^1(\Omega)}, \tag{3.6}$$

but $M_\epsilon \to \infty$ as $\epsilon \to 0$. Throughout the paper we will find a special class of elements $h \in H^{-1}(\Omega)$ defined as

$$\langle h, \phi \rangle_{-1,1} = \langle f, \phi \rangle_\Omega + \langle g, \gamma(\phi) \rangle_\Gamma$$

for every $\phi \in H^1(\Omega)$, where $f \in L^{p'}(\Omega)$, for some $p \le 2^*$, and $g \in H^{-1/2}(\Gamma)$. So, in brief, $h \overset{def}{=} f_\Omega + g_\Gamma$. The space $L^{p'}(\Omega) + H^{-1/2}(\Gamma)$ is endowed with the natural norm $\|f\|_{L^{p'}(\Omega)} + \|g\|_{H^{-1/2}(\Gamma)}$, and for this norm the inclusion $L^{p'}(\Omega) + H^{-1/2}(\Gamma) \subset H^{-1}(\Omega)$ is continuous.

Denote by q_0^* the number in the right-hand side of (1.4); i.e.,

$$q_0^* = \begin{cases} N/2 & \text{if } N \ge 3, \\ 1 & \text{if } N = 1, 2. \end{cases} \tag{3.7}$$

Then for any $q_0 \geq q_0^*$ if $N = 1$ or $N \geq 3$ or $q_0 > q_0^*$ if $N = 2$, let $p_0 \leq 2^*$ be such that $2/p_0 + 1/q_0 = 1$. Then the mapping $(V, u) \mapsto Vu$ is continuous from $L^{q_0}(\Omega) \times H^1(\Omega)$ into $L^{p_0'}(\Omega)$. Moreover, for every $\eta > 0$ there exists C_η such that for every $u, \phi \in H^1(\Omega)$,

$$\left| \int_\Omega Vu\phi \right| \leq \eta \|u\|_{H^1(\Omega)} \|\phi\|_{H^1(\Omega)} + C_\eta \|u\|_{L^2(\Omega)} \|\phi\|_{L^2(\Omega)}; \tag{3.8}$$

see [2].

Analogously, for the terms on the boundary, denote by 2_Γ^* the critical Sobolev exponent for the inclusion $H^{1/2}(\Gamma) \subset L^p(\Gamma)$; i.e.,

$$2_\Gamma^* = \begin{cases} \frac{2(N-1)}{N-2} & \text{if } N \geq 3, \\ \text{any positive number} & \text{if } N = 2, \\ \infty & \text{if } N = 1. \end{cases} \tag{3.9}$$

Hence $H^{1/2}(\Gamma) \subset L^p(\Gamma)$ with continuous inclusion if $p \leq 2_\Gamma^*$ and compact inclusion if $p < 2_\Gamma^*$. Also, denote by q_1^* the number in the right-hand side of (1.5); i.e.,

$$q_1^* = \begin{cases} N - 1 & \text{if } N \geq 3, \\ 1 & \text{if } N = 1, 2. \end{cases} \tag{3.10}$$

Then for any $q_1 \geq q_1^*$, if $N = 1$ or $N \geq 3$, or $q_1 > q_1^*$ if $N = 2$, let $p_1 \leq 2_\Gamma^*$ be such that $2/p_1 + 1/q_1 = 1$. Then $(b, u) \mapsto b\gamma(u)$ is continuous from $L^{q_1}(\Gamma) \times H^1(\Omega)$ into $L^{p_1'}(\Gamma)$ and for every $\eta > 0$ there exists some C_η such that for every $u, \phi \in H^1(\Omega)$,

$$\left| \int_\Gamma bu\phi \right| \leq \eta \|u\|_{H^1(\Omega)} \|\phi\|_{H^1(\Omega)} + C_\eta \|u\|_{L^2(\Gamma)} \|\phi\|_{L^2(\Gamma)}. \tag{3.11}$$

Also, for every $\eta > 0$,

$$\|u\|_{L^2(\Gamma)}^2 \leq \eta \|u\|_{H^1(\Omega)}^2 + C_\eta \|u\|_{L^2(\Omega)}^2. \tag{3.12}$$

Finally, we consider the Hilbert space

$$L^2_{\Omega_0}(\Omega) = \{u \in L^2(\Omega),\ u \text{ is constant on } \Omega_{0,i},\ i = 1, \ldots, m\}$$

and denote $H^1_{\Omega_0}(\Omega) = L^2_{\Omega_0}(\Omega) \cap H^1(\Omega)$. As a general notation, for $u \in L^2_{\Omega_0}(\Omega)$, we denote by $u_{\Omega_{0,i}}$ the constant value of u on $\Omega_{0,i}$ for $i = 1, \ldots, m$. Therefore, $u \in L^2_{\Omega_0}(\Omega)$ can be written as $u = u\mathcal{X}_{\Omega_1} + \sum_{i=1}^m u_{\Omega_{0,i}} \mathcal{X}_{\Omega_{0,i}}$.

For $\lambda \in \mathbb{R}$ we define the symmetric bilinear form

$$a_0(u, \phi) + \lambda \int_\Omega u\phi = \int_{\Omega_1} d_0(x) \nabla u \nabla \phi + \lambda \int_\Omega u\phi, \tag{3.13}$$

which is continuous and coercive on $H^1_{\Omega_0}(\Omega)$ if $\lambda > 0$. Hence (3.13) induces an operator $L_0 + \lambda I$ between $H^1_{\Omega_0}(\Omega)$ and its dual $H^{-1}_{\Omega_0}(\Omega)$, which for $\lambda > 0$ is an isomorphism.

4. The elliptic problem. In this section we analyze the solvability of problem (1.1) and the limiting one, and we make precise in what sense the solutions converge. In doing this, we will make rigorous all the formal manipulations leading to (2.1). First we study the solvability of the Neumann problem

$$\begin{cases} -Div(d_\epsilon(x)\nabla u) + (\lambda + V(x))u = f & \text{on } \Omega, \\ \frac{\partial u}{\partial \vec{n}_\epsilon} + b(x)u = g & \text{on } \Gamma \end{cases} \tag{4.1}$$

that we treat in a variational way in $H^{-1}(\Omega)$. Note that if we take test functions $\phi \in H^1(\Omega)$ and multiply in (4.1) and formally integrate by parts, we get

$$\int_\Omega d_\epsilon(x)\nabla u\nabla\phi + \int_\Omega (\lambda + V)u\phi + \int_\Gamma b(x)u\phi = \int_\Omega f\phi + \int_\Gamma g\phi;$$

see [2] for the case of Dirichlet boundary conditions.

Assume $V \in L^{q_0}(\Omega)$ and $b \in L^{q_1}(\Gamma)$ with $q_0 \geq q_0^*$, $q_1 \geq q_1^*$ as defined in the previous section. Then denoting by V_- and b_- the negative parts of V and b, respectively, we have the following theorem.

THEOREM 4.1. (i) *There exists a constant $c = c(\|V_-\|_{L^{q_0}}, \|b_-\|_{L^{q_1}})$ such that $c(0,0) = 0$ and for $\lambda > c$ the operator $T_\epsilon = L_\epsilon + (\lambda I + V)_\Omega + (b\gamma)_\Gamma$, defined by the bilinear form*

$$\tau_\epsilon(u,v) = \langle T_\epsilon u, v\rangle_{-1,1} = \int_\Omega d_\epsilon(x)\nabla u\nabla v + \int_\Omega (\lambda + V(x))uv + \int_\Gamma b(x)uv, \tag{4.2}$$

is an isomorphism between $H^1(\Omega)$ and its dual and between Y_{ϵ,p_0} and $L^{p_0'}(\Omega) + H^{-1/2}(\Gamma)$. Moreover, between the latter spaces T_ϵ is given by

$$T_\epsilon = (-Div(d_\epsilon(x)\nabla\cdot) + (\lambda + V(x))\cdot)_\Omega + \left(\frac{\partial\cdot}{\partial\vec{n}_\epsilon} + b(x)\gamma(\cdot)\right)_\Gamma.$$

By restriction to $L^2(\Omega)$, T_ϵ induces an unbounded self-adjoint positive operator with compact resolvent, B_ϵ, with domain $D(B_\epsilon) = \{u \in H^1(\Omega),\ -Div(d_\epsilon\nabla u) + Vu \in L^2(\Omega),\ \frac{\partial u}{\partial\vec{n}_\epsilon} + b(x)u = 0 \text{ on } \Gamma\}$ and for $u \in D(B_\epsilon)$,

$$B_\epsilon u = -Div(d_\epsilon(x)\nabla u) + (\lambda + V(x))u. \tag{4.3}$$

(ii) *If $\lambda \in \mathbb{R}$ and if $u \in H^1(\Omega)$ is a solution of*

$$T_\epsilon u = h = f_\Omega + g_\Gamma \tag{4.4}$$

with $f \in L^{p_0'}(\Omega)$ and $g \in H^{-1/2}(\Gamma)$, then $u \in Y_{\epsilon,p_0}$ and

$$-Div(d_\epsilon(x)\nabla u) + (\lambda + V(x))u = f \text{ in } L^{p_0'}(\Omega) \quad \text{and} \quad \frac{\partial u}{\partial\vec{n}_\epsilon} + b(x)u = g \text{ in } H^{-1/2}(\Gamma).$$

Proof. The proof is rather standard but is given for completeness. We first show that the bilinear form τ_ϵ in (4.2) is coercive on $H^1(\Omega)$ for λ large. Since $\tau_\epsilon(u,u) \geq m_0\int_\Omega|\nabla u|^2 + \int_\Omega(\lambda - V_-)|u|^2 - \int_\Gamma b_-|u|^2$ and $V_- \in L^{q_0}(\Omega)$, $b_- \in L^{q_1}(\Gamma)$, from (3.8), (3.11), and (3.12), we get coerciveness for λ large. Also, if $V \geq 0$ and $b \geq 0$, then τ_ϵ is coercive for $\lambda > 0$ and this gives $c(0,0) = 0$. Now if $h = f_\Omega + g_\Gamma \in L^{p_0'}(\Omega) + H^{-1/2}(\Gamma) \subset H^{-1}(\Omega)$ and $u \in H^1(\Omega)$ verifies $T_\epsilon u = h$, then for every $\phi \in H^1(\Omega)$,

$$\int_\Omega d_\epsilon\nabla u\nabla\phi + \int_\Omega (\lambda + V)u\phi + \int_\Gamma bu\phi = \int_\Omega f\phi + \langle g, \gamma(\phi)\rangle_\Gamma.$$

If $\phi \in \mathcal{D}(\Omega)$, then we get $-Div(d_\epsilon(x)\nabla u) + (\lambda + V)u = f$ a.e. on Ω and in particular $u \in Y_{\epsilon,p_0}$. Using this and $\phi \in H^1(\Omega)$, from (3.4), we get $\frac{\partial u}{\partial\vec{n}_\epsilon} + b(x)u = g$ on Γ.

Conversely, from (3.4) we have that T_ϵ maps Y_{ϵ,p_0} into $L^{p_0'}(\Omega) + H^{-1/2}(\Gamma)$ and on this space $T_\epsilon = (-Div(d_\epsilon(x)\nabla\cdot) + (\lambda + V(x))\cdot)_\Omega + (\frac{\partial\cdot}{\partial\vec{n}_\epsilon} + b(x)\gamma(\cdot))_\Gamma$. From this the continuity of T_ϵ and its inverse follows easily and the rest is obvious. ∎

Observe that if V, b, and Ω are smooth enough, then $D(B_\epsilon) = \{u \in H^2(\Omega),\ \frac{\partial u}{\partial \vec{n}_\epsilon} + b(x)u = 0 \text{ on } \Gamma\}$ and the graph norm in $D(B_\epsilon)$ is equivalent to that of $H^2(\Omega)$. For this a sufficient regularity condition is that $Vu \in L^2(\Omega)$ for $u \in H^1(\Omega)$, i.e., $V \in L^{q_0}$ for $q_0 \geq N$ if $N \geq 3$, or $q_0 > 2$ if $N = 2$ or $q_0 \geq 2$ if $N = 1$, and b is the restriction to Γ of a function $b \in C^1(\overline{\Omega})$ since in this case, for $u \in H^1(\Omega)$, $bu \in H^{1/2}(\Gamma)$ and elliptic regularity results give the $H^2(\Omega)$ regularity. The same situation holds for the solution of (4.1) if $g \in H^{1/2}(\Gamma)$. In other words, under suitable regularity assumptions, solutions of (4.1) are smooth in Ω and depend smoothly on f and g.

Now we analyze the solvability of the "limiting problem" described in terms of the operator L_0 defined in (3.13).

THEOREM 4.2. (i) *Assume $V \in L^{q_0}(\Omega)$ and $b \in L^{q_1}(\Gamma)$, with q_0, q_1 as above; then there exists a constant $c = c(\|V_-\|_{L^{q_0}}, \|b_-\|_{L^{q_1}})$ such that $c(0,0) = 0$ such that, for $\lambda > c$, the bilinear form*

$$\tau_0(u,\phi) = \int_{\Omega_1} d_0 \nabla u \nabla \phi + \int_\Omega (\lambda + V) u\phi + \int_\Gamma bu\phi \tag{4.5}$$

with $u, \phi \in H^1_{\Omega_0}(\Omega)$, defines an isomorphism, T_0, from $H^1_{\Omega_0}(\Omega)$ into its dual $H^{-1}_{\Omega_0}(\Omega)$. The restriction

$$\begin{aligned} T_0 : Y_{\Omega_0,p_0} &= \{z \in H^1_{\Omega_0}(\Omega),\ Div(d_0(x)\nabla z) \in L^{p_0'}(\Omega_1)\} \\ &\longrightarrow L^{p_0'}_{\Omega_0}(\Omega) + H^{-1/2}(\Gamma) \subset H^{-1}_{\Omega_0}(\Omega) \end{aligned}$$

is also an isomorphism, and on this space

$$\begin{aligned} T_0(u) = \Bigg[&(-Div(d_0(x)\nabla u) + (\lambda + V)u)\mathcal{X}_{\Omega_1} \\ &+ \sum_{i=1}^m \left(\frac{1}{|\Omega_{0,i}|} \int_{\Gamma_{0,i}} \frac{\partial u}{\partial \vec{n}_0} + \left(\fint_{\Omega_{0,i}} \lambda + V \right) u_{\Omega_{0,i}} \right) \mathcal{X}_{\Omega_{0,i}} \Bigg]_\Omega + \left(\frac{\partial u}{\partial \vec{n}_0} + bu \right)_\Gamma , \end{aligned}$$

where $u_{\Omega_{0,i}}$ denotes the constant value of u on $\Omega_{0,i}$ for $i = 1, \ldots, m$.

By restriction, the operator T_0 induces an unbounded, positive, self-adjoint operator with compact resolvent, B_0, in $H = L^2_{\Omega_0}(\Omega)$ with domain $D(B_0) = \{u \in H^1_{\Omega_0}(\Omega),\ -Div(d_0(x)\nabla u) + Vu \in L^2(\Omega_1),\ \frac{\partial u}{\partial \vec{n}_0} + bu = 0 \text{ on } \Gamma\}$ and for $u \in D(B_0)$,

$$\begin{aligned} B_0(u) = &(-Div(d_0(x)\nabla u) + (\lambda + V)u)\,\mathcal{X}_{\Omega_1} \\ &+ \sum_{i=1}^m \left(\frac{1}{|\Omega_{0,i}|} \int_{\Gamma_{0,i}} \frac{\partial u}{\partial \vec{n}_0} + \left(\fint_{\Omega_{0,i}} \lambda + V \right) u_{\Omega_{0,i}} \right) \mathcal{X}_{\Omega_{0,i}}. \end{aligned} \tag{4.6}$$

(ii) *If $\lambda \in \mathbb{R}$ and $u \in H^1_{\Omega_0}(\Omega)$ is a solution of*

$$T_0(u) = h = f_\Omega + g_\Gamma \tag{4.7}$$

with $f \in L^{p_0'}_{\Omega_0}(\Omega)$ and $g \in H^{-1/2}(\Gamma)$, then $u \in Y_{\Omega_0,p_0}$ and then (4.7) is equivalent to

$$\begin{aligned} &-Div(d_0(x)\nabla u) + (\lambda + V(x))u = f && \text{on } \Omega_1, \\ &\frac{\partial u}{\partial \vec{n}_0} + b(x)u = g && \text{on } \Gamma, \\ &\gamma_{0,i}(u) = u_{\Omega_{0,i}} && \text{on } \Gamma_{0,i}, \quad i = 1, \ldots, m, \\ &\frac{1}{|\Omega_{0,i}|} \int_{\Gamma_{0,i}} \frac{\partial u}{\partial \vec{n}_0} + \left(\fint_{\Omega_{0,i}} \lambda + V \right) u_{\Omega_{0,i}} = f_{\Omega_{0,i}}, && i = 1, \ldots, m, \end{aligned} \tag{4.8}$$

where $u_{\Omega_{0,i}}$ and $f_{\Omega_{0,i}}$ denote, respectively, the constant values of u and f on $\Omega_{0,i}$ for $i = 1, \dots, m$.

Proof. First note that the assumptions on V and b, (3.8), (3.11), and (3.12), imply that the bilinear form $\tau_0(\cdot,\cdot)$ is coercive on $H^1_{\Omega_0}(\Omega)$ for λ large enough. Hence T_0 is an isomorphism. From here one easily gets that B_0 is self-adjoint and positive in $L^2_{\Omega_0}(\Omega)$, and the compactness of the resolvent follows from the compactness of the inclusion $H^1_{\Omega_0}(\Omega) \subset L^2_{\Omega_0}(\Omega)$, since Ω is bounded.

Now assume $h = f_\Omega + g_\Gamma \in L^{p_0'}_{\Omega_0}(\Omega) + H^{-1/2}(\Gamma)$ and $u \in H^1_{\Omega_0}(\Omega)$ such that $T_0 u = h$; then taking in (4.5) first $\phi \in \mathcal{D}(\Omega_1)$ and then $\phi \in H^1_{\Omega_0}(\Omega)$ such that it vanishes on Ω_0, we get $-Div(d_0(x)\nabla u) + (\lambda + V)u = f$ on Ω_1, which implies $u \in Y_{\Omega_0,p_0}$, and $\frac{\partial u}{\partial \vec{n}_0} + bu = g$ on Γ. Using this information on Ω_1 and integrating by parts, again using (3.4), for an arbitrary $\phi \in H^1_{\Omega_0}(\Omega)$ we get

$$\sum_{i=1}^m \phi_{\Omega_{0,i}} \left(\int_{\Gamma_{0,i}} \frac{\partial u}{\partial \vec{n}_0} + \left(\int_{\Omega_{0,i}} \lambda + V \right) u_{\Omega_{0,i}} |\Omega_{0,i}| \right) = \sum_{i=1}^m \phi_{\Omega_{0,i}} |\Omega_{0,i}| f_{\Omega_{0,i}},$$

and we get the expression for T_0 on Y_{Ω_0,p_0}. From this one easily gets that T_0 is an isomorphism between Y_{Ω_0,p_0} and $L^{p_0'}_{\Omega_0}(\Omega) + H^{-1/2}(\Gamma)$. The rest is obvious. ∎

Remark. (i) Observe that if $V \in L^{q_0}(\Omega)$ and $b \in L^{q_1}(\Gamma)$, with q_0, q_1 as above, then for $u, \phi \in H^1_{\Omega_0}(\Omega)$ we have

$$\int_\Omega Vu\phi = \int_{\Omega_1} Vu\phi + \sum_{i=1}^m \left(\int_{\Omega_{0,i}} V \right) u_{\Omega_{0,i}} \phi_{\Omega_{0,i}} = \int_\Omega \hat{V} u\phi,$$

where $\hat{V} = V\mathcal{X}_{\Omega_1} + \sum_{i=1}^m (⨍_{\Omega_{0,i}} V)\mathcal{X}_{\Omega_{0,i}} \in L^{q_0}_{\Omega_0}(\Omega)$. Consequently, on $H^1_{\Omega_0}(\Omega)$, T_0 coincides with the operator $L_0 + (\lambda + \hat{V})_\Omega + (b\gamma)_\Gamma$.

Also note that since $H^1_{\Omega_0}(\Omega) \subset H^1(\Omega)$, if $h \in H^{-1}(\Omega)$ then we have $h_{|H^1_{\Omega_0}(\Omega)} \in H^{-1}_{\Omega_0}(\Omega)$. In particular, if $h = f_\Omega + g_\Gamma$, with $f \in L^{p_0'}(\Omega)$ and $g \in H^{-1/2}(\Gamma)$, we have for every $\phi \in H^1_{\Omega_0}(\Omega)$,

$$\int_\Omega f\phi + \langle g, \gamma(\phi)\rangle_\Gamma = \sum_{i=1}^m \phi_{\Omega_{0,i}} \int_{\Omega_{0,i}} f + \int_{\Omega_1} f\phi + \langle g, \gamma(\phi)\rangle_\Gamma = \int_\Omega \hat{f}\phi + \langle g, \gamma(\phi)\rangle_\Gamma,$$

where $\hat{f} = f\mathcal{X}_{\Omega_1} + \sum_{i=1}^m (⨍_{\Omega_{0,i}} f)\mathcal{X}_{\Omega_{0,i}} \in L^{p_0'}_{\Omega_0}(\Omega)$. That is, $h_{|H^1_{\Omega_0}(\Omega)} = \hat{f}_\Omega + g_\Gamma$.

(ii) Note that if Ω, Ω_1, V, and b are smooth enough, say, V bounded and $b \in C^1(\Omega_1)$, then $D(B_0) = H^1_{\Omega_0}(\Omega) \cap H^2(\Omega_1)$ both algebraically and topologically, since for $u \in D(B_0)$ the part in Ω_1 satisfies $-Div(d_0(x)\nabla u) \in L^2(\Omega_1)$, $\frac{\partial u}{\partial \vec{n}_0} + bu = 0$ on Γ and $\gamma_{0,i}(u) = u_{\Omega_{0,i}}$ on $\Gamma_{0,i}$, $i = 1, \dots, m$ and elliptic regularity results give the $H^2(\Omega_1)$ regularity. Also, since u is constant on Ω_0 it is in $H^2(\Omega_0)$. The same applies to the solution of (4.8) if $g \in H^{1/2}(\Gamma)$ and then $u \in H^1_{\Omega_0}(\Omega) \cap H^2(\Omega_1)$.

However, $D(B_0)$ is not included with continuous inclusion in $H^2(\Omega)$, since typically there is a jump on the normal derivative $\frac{\partial u}{\partial \vec{n}_0}$ across Γ_0 for $u \in D(B_0)$.

To see this, we will show that the mapping $L^2_{\Omega_0}(\Omega) \ni f \to u = B_0^{-1}(f) \in D(B_0)$ is not continuous with the topology of $H^2(\Omega)$. In fact, assume that for some $f \in L^2_{\Omega_0}(\Omega)$

we have $u = B_0^{-1}(f) \in H^2(\Omega)$. Then $\frac{\partial u}{\partial \vec{n}_0} = 0$ on Γ_0 and the restriction to Ω_1 satisfies

$$\begin{aligned} &-Div(d_0(x)\nabla u) + (\lambda + V(x))u = f \quad \text{on } \Omega_1, \\ &\frac{\partial u}{\partial \vec{n}_0} + b(x)u = 0 \quad \text{on } \Gamma, \\ &\frac{\partial u}{\partial \vec{n}_0} = 0 \quad \text{on } \Gamma_0. \end{aligned}$$

Hence, on Ω_1, u is uniquely determined by the restriction of f to Ω_1 and so are the traces $\gamma_{0,i}(u)$, $i = 1, \dots, m$. On the other hand, on $\Omega_{0,i}$ we have $\fint_{\Omega_{0,i}} \lambda + V) u_{\Omega_{0,i}} = f_{\Omega_{0,i}}$ and therefore $u_{\Omega_{0,i}}$ can be changed at will by changing $f_{\Omega_{0,i}}$. As a consequence, the matching conditions $\gamma_{0,i}(u) = u_{\Omega_{0,i}}$ on $\Gamma_{0,i}$, $i = 1, \dots, m$, cannot be verified and this is a contradiction. In fact, the same argument gives that $D(B_0)$ is not included with continuous inclusion in $H^{3/2+\delta}(\Omega)$ for any $\delta > 0$. The same argument applies to the solution of (4.8) when $g \neq 0$, and hence u does not depend continuously on f and g in the topology of $H^{3/2+\delta}(\Omega)$.

In other words, under suitable regularity assumptions, solutions of (4.8) are smooth on Ω_1, and obviously on Ω_0, and depend smoothly on f and g, but they are not smooth or do not depend smoothly on f and g on the entire Ω. The problem of course comes from the normal derivative across Γ_0.

The next theorem, which is the main result of this section, in particular states rigorously that, as $\epsilon \to 0$, the family of solutions of (1.1) converges to a solution of (4.8). Note that from the previous results we cannot expect convergence in a very strong topology on the entire Ω. However, we will show that convergence in $H^1(\Omega)$ always holds true. We start with the following lemma.

LEMMA 4.3. *Assume as* $\epsilon \to 0$,

$$d_\epsilon(x) \to d_0(x), \text{ uniformly on } \Omega_1 \quad \text{and} \quad u^\epsilon \to u, \text{ weakly in } H^1(\Omega_1).$$

Then

$$\int_{\Omega_1} d_0 |\nabla u|^2 \leq \liminf_\epsilon \int_{\Omega_1} d_\epsilon |\nabla u^\epsilon|^2.$$

If, moreover, $u^\epsilon \to u$ *in* $L^2(\Omega_1)$, *then* $u^\epsilon \to u$ *in* $H^1(\Omega_1)$ *if and only if*

$$\int_{\Omega_1} d_0 |\nabla u|^2 = \lim_\epsilon \int_{\Omega_1} d_\epsilon |\nabla u^\epsilon|^2.$$

Proof. From lower semicontinuity, we have $\int_{\Omega_1} d_0 |\nabla u|^2 \leq \liminf_\epsilon \int_{\Omega_1} d_0 |\nabla u^\epsilon|^2$ and then

$$\int_{\Omega_1} d_\epsilon |\nabla u^\epsilon|^2 - \int_{\Omega_1} d_0 |\nabla u|^2 = \int_{\Omega_1} (d_\epsilon - d_0) |\nabla u^\epsilon|^2 + \int_{\Omega_1} d_0 |\nabla u^\epsilon|^2 - \int_{\Omega_1} d_0 |\nabla u|^2.$$

But $| \int_{\Omega_1} (d_\epsilon - d_0) |\nabla u^\epsilon|^2 | \leq \|d_\epsilon - d_0\|_{L^\infty(\Omega_1)} \int_{\Omega_1} |\nabla u^\epsilon|^2 \to 0$ and we get the first statement. If, moreover, $u^\epsilon \to u$ in $L^2(\Omega_1)$, we have that the strong convergence in $H^1(\Omega_1)$ is equivalent to $\int_{\Omega_1} d_0 |\nabla u|^2 = \lim_\epsilon \int_{\Omega_1} d_0 |\nabla u^\epsilon|^2$ and the lemma is proved □

THEOREM 4.4. (i) *Assume that for* $0 < \epsilon \leq \epsilon_0$, $\{u^\epsilon\}_\epsilon \subset H^1(\Omega)$ *are given such that they are bounded in* $L^2(\Omega)$ *and* $0 \leq \int_\Omega d_\epsilon |\nabla u^\epsilon|^2 \leq M$ *for some constant independent*

of ϵ. Then, by taking subsequences if necessary, u^ϵ converges to $u \in H^1_{\Omega_0}(\Omega)$ weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$.

Even more,

$$\lim_\epsilon \int_\Omega d_\epsilon |\nabla u^\epsilon|^2 = \int_\Omega d_0 |\nabla u|^2 = \int_{\Omega_1} d_0 |\nabla u|^2$$

if and only if u^ϵ converges strongly in $H^1(\Omega)$ to u and $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla u^\epsilon|^2 = 0$.

(ii) *Assume that for $0 < \epsilon \le \epsilon_0$, V_ϵ and b_ϵ verify (1.4), (1.5), respectively, $\lambda \in \mathbb{R}$ and $\{h^\epsilon\}_\epsilon \subset H^{-1}(\Omega)$ is bounded and weakly convergent to $h \in H^{-1}(\Omega)$. Assume $\{u^\epsilon\}_\epsilon$ are bounded in $L^2(\Omega)$ and verify*

$$T_\epsilon u^\epsilon = (L_\epsilon + (\lambda I + V_\epsilon)_\Omega + (b_\epsilon \gamma)_\Gamma) u^\epsilon = h^\epsilon.$$

Then $0 \le \int_\Omega d_\epsilon |\nabla u^\epsilon|^2 \le M$ for some constant independent of ϵ and by taking subsequences if necessary, u^ϵ converges to $u \in H^1_{\Omega_0}(\Omega)$ weakly in $H^1(\Omega)$ and $V_\epsilon u^\epsilon$ and $b_\epsilon u^\epsilon$ converge strongly in $L^{p_0'}(\Omega)$ and $L^{p_1'}(\Gamma)$, respectively, to Vu and bu. Moreover, $u \in H^1_{\Omega_0}(\Omega)$ verifies

$$T_0 u = (L_0 + (\lambda I + V)_\Omega + (b\gamma)_\Gamma) u = h_{|H^1_{\Omega_0}(\Omega)} \quad \text{in } H^{-1}_{\Omega_0}(\Omega). \tag{4.9}$$

If the function u above is unique, for example, when λ is sufficiently large, then the full family u^ϵ converges to u.

Finally, u^ϵ converges strongly in $H^1(\Omega)$ to u if and only if $\langle h^\epsilon, u^\epsilon \rangle_{-1,1} \to \langle h, u \rangle_{-1,1}$, and this happens, for example, if $h^\epsilon \to h$ in $H^{-1}(\Omega)$. In such a case, we also have

$$\lim_\epsilon \tau_\epsilon(u^\epsilon, u^\epsilon) = \tau_0(u, u) \tag{4.10}$$

and in particular $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla u^\epsilon|^2 = 0$.

Proof. (i) From (1.2), we have $m_0 \int_\Omega |\nabla u^\epsilon|^2 \le \int_\Omega d_\epsilon |\nabla u^\epsilon|^2 \le M$ and then, from the boundedness in $L^2(\Omega)$, u^ϵ is bounded in $H^1(\Omega)$ and, taking subsequences if necessary, it converges weakly to u in $H^1(\Omega)$ and strongly in $L^2(\Omega)$. Using lower semicontinuity, we get $\int_K |\nabla u|^2 \le \liminf_\epsilon \int_K |\nabla u^\epsilon|^2$ for any compact set $K \subset \Omega_0$. On the other hand,

$$\inf_{x \in K} \{d_\epsilon(x)\} \int_K |\nabla u^\epsilon|^2 \le \int_K d_\epsilon |\nabla u^\epsilon|^2 \le M$$

and therefore from (1.3), we obtain $\lim_\epsilon \int_K |\nabla u^\epsilon|^2 = 0$ and $u \in H^1_{\Omega_0}(\Omega)$. From Lemma 4.3, we have

$$\liminf_\epsilon \int_\Omega d_\epsilon |\nabla u^\epsilon|^2 \ge \liminf_\epsilon \int_{\Omega_1} d_\epsilon |\nabla u^\epsilon|^2 \ge \int_{\Omega_1} d_0 |\nabla u|^2 = \int_\Omega d_0 |\nabla u|^2$$

and, in fact, $\lim_\epsilon \int_\Omega d_\epsilon |\nabla u^\epsilon|^2 = \int_\Omega d_0 |\nabla u|^2$ if and only if $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla u^\epsilon|^2 = 0$ and $\lim_\epsilon \int_{\Omega_1} d_\epsilon |\nabla u^\epsilon|^2 = \int_{\Omega_1} d_0 |\nabla u|^2$. Again, Lemma 4.3 gives the result.

(ii) Since $T_\epsilon u^\epsilon = (L_\epsilon + (\lambda I + V_\epsilon)_\Omega + (b_\epsilon \gamma)_\Gamma) u^\epsilon = h^\epsilon$, we have

$$\int_\Omega d_\epsilon |\nabla u^\epsilon|^2 + \int_\Omega (\lambda + V_\epsilon) |u^\epsilon|^2 + \int_\Gamma b_\epsilon |u^\epsilon|^2 = \langle h^\epsilon, u^\epsilon \rangle_{-1,1}. \tag{4.11}$$

From (3.8), (3.11), and (3.12) we get $\int_\Omega d_\epsilon |\nabla u^\epsilon|^2 \le M$ and the above applies. Also, since the inclusion $H^1(\Omega) \subset L^{p_0}(\Omega)$ is compact, taking subsequences if necessary, we

can assume $V_\epsilon u^\epsilon$ converges strongly in $L^{p_0'}(\Omega)$. The same holds for $b_\epsilon u^\epsilon$ in $L^{p_1'}(\Gamma)$, since $H^{1/2}(\Gamma) \subset L^{q_1}(\Gamma)$ is also compact. Moreover, for every $\phi \in H^1_{\Omega_0}(\Omega)$ it holds that

$$\tau_\epsilon(u^\epsilon, \phi) = \int_{\Omega_1} d_\epsilon \nabla u^\epsilon \nabla \phi + \int_\Omega (\lambda + V_\epsilon) u^\epsilon \phi + \int_\Gamma b_\epsilon u^\epsilon \phi = \langle h^\epsilon, \phi \rangle_{-1,1},$$

and we can pass to the limit to get

$$\tau_0(u, \phi) = \int_{\Omega_1} d_0 \nabla u \nabla \phi + \int_\Omega (\lambda + V) u \phi + \int_\Gamma b u \phi = \langle h, \phi \rangle_{-1,1},$$

i.e., $T_0 u = h_{|H^1_{\Omega_0}(\Omega)}$. In particular, with $\phi = u$,

$$\int_{\Omega_1} d_0 |\nabla u|^2 + \int_\Omega (\lambda + V)|u|^2 + \int_\Gamma b|u|^2 = \langle h, u \rangle_{-1,1}. \tag{4.12}$$

Furthermore, from (4.11) and (4.12), $\lim_\epsilon \int_\Omega d_\epsilon |\nabla u^\epsilon|^2 = \int_{\Omega_1} d_0 |\nabla u|^2$ if and only if $\langle h^\epsilon, u^\epsilon \rangle_{-1,1} \to \langle h, u \rangle_{-1,1}$, which holds true if u^ϵ or h^ϵ converges strongly. The rest follows easily. □

From this theorem we get the following consequence.

COROLLARY 4.5. *Assume that in Theorem* 4.4, $h^\epsilon = f^\epsilon_\Omega + g^\epsilon_\Gamma$ *and* $f^\epsilon \to f$ *weakly in* $L^{p_0'}(\Omega)$ *and* $g^\epsilon \to g$ *weakly in* $H^{-1/2}(\Gamma)$. *Then* (i) *From inside* Ω_1

$$Div(d_\epsilon \nabla u^\epsilon) \to Div(d_0 \nabla u) \text{ weakly in } L^{p_0'}(\Omega_1),$$

$$\frac{\partial u^\epsilon}{\partial \vec{n}_\epsilon} \to \frac{\partial u}{\partial \vec{n}_0} \text{ weakly in } H^{-1/2}(\Gamma \cup \Gamma_0).$$

(ii) *From inside* Ω_0,

$$Div(d_\epsilon \nabla u^\epsilon) \to f - (\lambda + V) \sum_{i=1}^m u_{\Omega_{0,i}} \mathcal{X}_{\Omega_{0,i}} \text{ weakly in } L^{p_0'}(\Omega_0),$$

$$\fint_{\Omega_{0,i}} -Div(d_\epsilon \nabla u^\epsilon) \to \frac{1}{|\Omega_{0,i}|} \int_{\Gamma_{0,i}} \frac{\partial u}{\partial \vec{n}_0}.$$

Also, if $f^\epsilon \to f_{\Omega_{0,i}}$, *weakly in* $L^{p_0'}(\Omega_{0,i})$ *and* $V_\epsilon \to V_{\Omega_{0,i}}$ *strongly in* $L^{q_0}(\Omega_{0,i})$, *then*

$$-Div(d_\epsilon \nabla u^\epsilon) \to \frac{1}{|\Omega_{0,i}|} \int_{\Gamma_{0,i}} \frac{\partial u}{\partial \vec{n}_0} \text{ weakly in } L^{p_0'}(\Omega_{0,i}).$$

If f^ϵ *and* g^ϵ *converge strongly, then all the convergences above are strong.*

Furthermore, u^ϵ *converges strongly to* u *in* $H^1(\Omega)$ *if and only if* $\langle g^\epsilon, u^\epsilon \rangle_\Gamma$ *converges to* $\langle g, u \rangle_\Gamma$, *and this happens, for example, if* $g^\epsilon \to g$ *strongly in* $H^{-1/2}(\Gamma)$. *In such a case we have* $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla u^\epsilon|^2 = 0$.

Proof. The convergence on Ω_1 and Ω_0 follows easily, since u^ϵ verifies

$$\begin{cases} -Div(d_\epsilon(x) \nabla u^\epsilon) + (\lambda + V_\epsilon(x)) u^\epsilon = f^\epsilon & \text{on } \Omega, \\ \frac{\partial u^\epsilon}{\partial \vec{n}_\epsilon} + b_\epsilon(x) u = g^\epsilon & \text{on } \Gamma, \end{cases}$$

and the convergence of $V_\epsilon u^\epsilon$ and $b_\epsilon u^\epsilon$ are obtained in the theorem. The convergence of the normal derivative on Γ_0 follows from the above and (3.2) applied to Ω_1. Integrating on $\Omega_{0,i}$ we get the convergence of the average of $-Div(d_\epsilon \nabla u^\epsilon)$ and the rest is obvious.

Since, from the theorem, u^ϵ converges strongly in $L^2(\Omega)$ then the strong convergence in $H^1(\Omega)$ is equivalent to the convergence $\langle g^\epsilon, u^\epsilon \rangle_\Gamma \to \langle g, u \rangle_\Gamma$. □

5. The eigenvalue problem. Consider the operators B_ϵ and B_0 of the previous section. Since they have compact resolvent their spectrum consists solely of eigenvalues of finite multiplicity, i.e., $\sigma(B_\epsilon)=\{\mu_n^\epsilon\}_n$ and $\sigma(B_0)=\{\mu_n\}_n$, which form increasing sequences converging to ∞. Note that the eigenvalue problem for B_ϵ is given by

$$\left\{\begin{array}{ll} -Div(d_\epsilon(x)\nabla u)+(\lambda+V_\epsilon)u=\mu u & \text{on } \Omega,\\ \frac{\partial u}{\partial \vec{n}_\epsilon}+b_\epsilon u=0 & \text{on } \Gamma \end{array}\right. \tag{5.1}$$

in $H^1(\Omega)$, while for B_0 the eigenvalue problem is given by

$$\begin{array}{ll} -Div(d_0(x)\nabla u)+(\lambda+V)u=\mu u & \text{on } \Omega_1,\\ \dfrac{\partial u}{\partial \vec{n}_0}+bu=0 & \text{on } \Gamma,\\ \gamma_{0,i}(u)=u_{\Omega_{0,i}} & \text{on } \Gamma_{0,i},\quad i=1,\dots,m,\\ \dfrac{1}{|\Omega_{0,i}|}\displaystyle\int_{\Gamma_{0,i}}\frac{\partial u}{\partial \vec{n}_0}+\left(\fint_{\Omega_{0,i}}\lambda+V\right)u_{\Omega_{0,i}}=\mu u_{\Omega_{0,i}}, & i=1,\dots,m, \end{array} \tag{5.2}$$

in $H^1_{\Omega_0}(\Omega)$, where $u_{\Omega_{0,i}}$ denotes the constant value of u on $\Omega_{0,i}$ for $i=1,\dots,m$.

The next result, which is the main result in this section, asserts that the spectral properties of B_ϵ and B_0 are close. Note that since B_ϵ and B_0 and their inverses are not defined on the same space, then one cannot use classical results, e.g., [14, Theorem VIII.1.14] or [16, Theorems V.9.10 and V.11.1], to analyze the behavior of eigenfunctions and eigenvalues.

THEOREM 5.1. *Assume that V_ϵ, b_ϵ verify* (1.4), (1.5) *and assume the spectrum of B_ϵ is given by*

$$\mu_1^\epsilon\le\cdots\le\mu_n^\epsilon\le\mu_{n+1}^\epsilon\cdots$$

counting multiplicities and, for each n, let ϕ_n^ϵ be an eigenfunction of μ_n^ϵ such that $\|\phi_n^\epsilon\|_{L^2(\Omega)}=1$ and such that $\{\phi_n^\epsilon\}_n$ is a Hilbert basis of $L^2(\Omega)$. Also, assume that the spectrum of B_0 is given by

$$\mu_1\le\cdots\le\mu_n\le\mu_{n+1}\cdots$$

counting multiplicities. Then the following conditions hold.

(i) *For each $n\in\mathbb{N}$,*

$$\mu_n^\epsilon\to\mu_n \text{ as } \epsilon\to 0.$$

(ii) *For each $n\in\mathbb{N}$, and for any sequence converging to zero that we still denote $\epsilon\to 0$, there exists a subsequence ϵ_j such that*

$$\phi_n^{\epsilon_j}\to\phi_n\in H^1_{\Omega_0}(\Omega) \text{ as } j\to\infty$$

strongly in $H^1(\Omega)$ and $\lim_{j\to\infty}\int_{\Omega_0}d_{\epsilon_j}|\nabla\phi_n^{\epsilon_j}|^2=0$, where ϕ_n is an eigenfunction of B_0 corresponding to μ_n and $\{\phi_n\}_n$ is a Hilbert basis of $L^2_{\Omega_0}(\Omega)$.

Proof. Note that it is sufficient to prove the result for λ large enough such that B_ϵ is an isomorphism, i.e., such that τ_ϵ in (4.2) is coercive. Then we proceed by induction in n. From the min-max characterization of the eigenvalues, we first have

$$\mu_1^\epsilon=\inf_{\substack{u\in H^1(\Omega)\\ \|u\|=1}}\int_\Omega d_\epsilon|\nabla u|^2+\int_\Omega(\lambda+V_\epsilon)|u|^2+\int_\Gamma b_\epsilon|u|^2=\inf_{\substack{u\in H^1(\Omega)\\ \|u\|=1}}\tau_\epsilon(u,u)>0$$

and

$$\mu_1 = \inf_{\substack{u\in H^1_{\Omega_0}(\Omega)\\ \|u\|=1}} \int_{\Omega_1} d_0|\nabla u|^2 + \int_\Omega (\lambda+V)|u|^2 + \int_\Gamma b|u|^2 = \inf_{\substack{u\in H^1_{\Omega_0}(\Omega)\\ \|u\|=1}} \tau_0(u,u) > 0,$$

where $\|\cdot\|$ represents the $L^2(\Omega)$ norm. Observe that for the Neumann boundary condition, $V_\epsilon = 0$, $b_\epsilon = 0$, $\mu_1^\epsilon = \mu_1 = \lambda$, and the first eigenfunction is constant on Ω. Also note that in any case the inf is attained.

On the one hand, if we restrict u to be in $H^1_{\Omega_0}(\Omega)$ with $\|u\| = 1$, we get, since u is constant on Ω_0 and from the convergence of d_ϵ, V_ϵ, b_ϵ,

$$\mu_1^\epsilon \le \tau_\epsilon(u,u) \to \tau_0(u,u)$$

and in particular $\limsup_\epsilon \mu_1^\epsilon \le \mu_1$.

On the other hand, by taking the family of first eigenfunctions, $\{\phi_1^\epsilon\}_\epsilon$, we have

$$\tau_\epsilon(\phi_1^\epsilon,\phi_1^\epsilon) = \int_\Omega d_\epsilon|\nabla\phi_1^\epsilon|^2 + \int_\Omega(\lambda+V_\epsilon)|\phi_1^\epsilon|^2 + \int_\Gamma b_\epsilon|\phi_1^\epsilon|^2 = \mu_1^\epsilon \le M$$

for some constant independent of ϵ. Therefore, by Theorem 4.4, and by taking subsequences if necessary, we have $\mu_1^\epsilon \to \mu^0$ and $\phi_1^\epsilon \to \phi \in H^1_{\Omega_0}(\Omega)$, where $\|\phi\| = 1$ and the latter convergence is weak in $H^1(\Omega)$ and strong in $L^2(\Omega)$. Even more, we can also assume $V_\epsilon\phi_1^\epsilon \to V\phi$ in $L^{p_0'}(\Omega)$ and $b_\epsilon\phi_1^\epsilon \to b\phi$ in $L^{p_1'}(\Gamma)$. But $B_\epsilon(\phi_1^\epsilon) = \mu_1^\epsilon\phi_1^\epsilon$ and then, since $\mu_1^\epsilon\phi_1^\epsilon$ converges in $H^{-1}(\Omega)$ to $\mu^0\phi$, by Theorem 4.4, ϕ satisfies $B_0(\phi) = \mu^0\phi$ and we have $\phi_1^\epsilon \to \phi \in H^1_{\Omega_0}(\Omega)$ strongly in $H^1(\Omega)$ and, even more, $\lim_\epsilon \int_{\Omega_0} d_\epsilon|\nabla\phi_1^\epsilon|^2 = 0$.

Consequently, $\mu^0 = \tau_0(\phi,\phi) \ge \mu_1$. Since the argument is independent of the subsequence we take, we really have proven $\liminf_\epsilon \mu_1^\epsilon \ge \mu_1$ and the function ϕ above, which may depend on the subsequence, is a normalized eigenfunction of μ_1. Therefore, the theorem is proved for $n = 1$.

Assume the result is proved for the first n eigenvalues. Then from the min-max characterization we have

$$\mu_{n+1}^\epsilon = \inf_{\substack{u\in H^1(\Omega)\\ u\perp\phi_1^\epsilon,\dots,\phi_n^\epsilon\\ \|u\|=1}} \tau_\epsilon(u,u).$$

Then take any sequence converging to zero that we still denote $\epsilon \to 0$, and, therefore, there exists a subsequence ϵ_j such that for every $i = 1,\dots,n$,

$$\phi_i^{\epsilon_j} \to \phi_i \in H^1_{\Omega_0}(\Omega) \text{ as } j\to\infty$$

strongly in $H^1(\Omega)$ and $\lim_{j\to\infty}\int_{\Omega_0} d_{\epsilon_j}|\nabla\phi_i^{\epsilon_j}|^2 = 0$, and ϕ_i is an eigenfunction of μ_i and the ϕ_i are mutually orthogonal and have $L^2(\Omega)$ norm equal to 1. With this we have

$$\mu_{n+1} = \inf_{\substack{u\in H^1_{\Omega_0}(\Omega)\\ u\perp\phi_1,\dots,\phi_n\\ \|u\|=1}} \tau_0(u,u).$$

We define in $L^2(\Omega)$ and $H^1(\Omega)$ the projections $P_n^{\epsilon_j} = \sum_{i=1}^n \langle\phi_i^{\epsilon_j},\cdot\rangle_\Omega\phi_i^{\epsilon_j}$, $Q_n^{\epsilon_j} = I - P_n^{\epsilon_j}$, and $P_n = \sum_{i=1}^n\langle\phi_i,\cdot\rangle_\Omega\phi_i$, $Q_n = I - P_n$. Hence from the strong convergence of $\phi_i^{\epsilon_j}$ in $H^1(\Omega)$ we have

$$P_n^{\epsilon_j} \to P_n, \qquad Q_n^{\epsilon_j} \to Q_n$$

strongly in $\mathcal{L}(L^2(\Omega), H^1(\Omega))$.

Therefore, for a given $u \in H^1_{\Omega_0}(\Omega)$, we take $u^{\epsilon_j} = Q_n^{\epsilon_j}(u)$ which is orthogonal in $L^2(\Omega)$ to $\phi_1^{\epsilon_j}, \dots, \phi_n^{\epsilon_j}$ and converges strongly in $H^1(\Omega)$ to $u^0 = Q_n(u)$, and then we claim that

$$\mu_{n+1}^{\epsilon_j} \le \frac{\tau_{\epsilon_j}(u^{\epsilon_j}, u^{\epsilon_j})}{\|u^{\epsilon_j}\|^2} \to \frac{\tau_0(u^0, u^0)}{\|u^0\|^2} \text{ as } j \to \infty.$$

Since $u^0 \perp \{\phi_1, \dots, \phi_n\}$ is arbitrary, we obtain $\limsup_j \mu_{n+1}^{\epsilon_j} \le \mu_{n+1}$. Also, since this upper bound is independent of the sequence ϵ and the subsequence ϵ_j, we have $\limsup_\epsilon \mu_{n+1}^{\epsilon} \le \mu_{n+1}$.

Now we prove our claim. The first inequality is obvious and then $\mu_{n+1}^{\epsilon_j} \le \frac{\tau_{\epsilon_j}(u^{\epsilon_j}, u^{\epsilon_j})}{\|u^{\epsilon_j}\|^2}$. But since $u^{\epsilon_j} = Q_n^{\epsilon_j}(u) = u - P_n^{\epsilon_j}(u)$, we have $\tau_{\epsilon_j}(u^{\epsilon_j}, u^{\epsilon_j}) = \tau_{\epsilon_j}(u, u) + \tau_{\epsilon_j}(P_n^{\epsilon_j}(u), P_n^{\epsilon_j}(u)) - 2\tau_{\epsilon_j}(u, P_n^{\epsilon_j}(u))$. Now, since u is constant on Ω_0, we get $\tau_{\epsilon_j}(u, u) \to \tau_0(u, u)$ and

$$\tau_{\epsilon_j}(u, P_n^{\epsilon_j}(u)) = \int_{\Omega_1} d_{\epsilon_j} \nabla u \nabla P_n^{\epsilon_j}(u) + \int_\Omega (\lambda + V_{\epsilon_j}) u P_n^{\epsilon_j}(u) + \int_\Gamma b_{\epsilon_j} u P_n^{\epsilon_j}(u)$$

$$\to \int_{\Omega_1} d_0 \nabla u \nabla P_n(u) + \int_\Omega (\lambda + V) u P_n(u) + \int_\Gamma b u P_n(u) = \tau_0(u, P_n(u)).$$

Finally,

$$\tau_{\epsilon_j}(P_n^{\epsilon_j}(u), P_n^{\epsilon_j}(u)) = \sum_{i=1}^n |\langle \phi_i^{\epsilon_j}, u\rangle|^2 \tau_{\epsilon_j}(\phi_i^{\epsilon_j}, \phi_i^{\epsilon_j})$$

$$\to \sum_{i=1}^n |\langle \phi_i, u\rangle|^2 \tau_0(\phi_i, \phi_i) = \tau_0(P_n(u), P_n(u)),$$

where we have used (4.10). Consequently, since $u^0 = Q_n(u) = u - P_n(u)$, then $\tau_\epsilon(u^\epsilon, u^\epsilon) \to \tau_0(u^0, u^0)$, and the claim is proved.

As a consequence of the above, by taking the family of $(n+1)$th eigenfunctions, $\{\phi_{n+1}^\epsilon\}_\epsilon$, we have

$$\tau_\epsilon(\phi_{n+1}^\epsilon, \phi_{n+1}^\epsilon) = \mu_{n+1}^\epsilon \le M$$

for some constant independent of ϵ. Therefore, $\int_\Omega d_\epsilon |\nabla \phi_{n+1}^\epsilon|^2$ is bounded and then by Theorem 4.4, and by taking subsequences if necessary, we have $\mu_i^\epsilon \to \mu_i$, for $i = 1, \dots, n$, $\mu_{n+1}^\epsilon \to \mu^0$, and $\phi_i^\epsilon \to \phi_i \in H^1_{\Omega_0}(\Omega)$, $i = 1, \dots, n$, and $\phi_{n+1}^\epsilon \to \phi \in H^1_{\Omega_0}(\Omega)$, where $\|\phi_i\| = \|\phi\| = 1$, ϕ_i are eigenfunctions of μ_i, and the latter convergence is weak in $H^1(\Omega)$ and strong in $L^2(\Omega)$. In addition, we can assume $V_\epsilon \phi_{n+1}^\epsilon \to V\phi$ in $L^{p_0'}(\Omega)$ and $b_\epsilon \phi_{n+1}^\epsilon \to b\phi$ in $L^{p_1'}(\Gamma)$. But since $B_\epsilon(\phi_{n+1}^\epsilon) = \mu_{n+1}^\epsilon \phi_{n+1}^\epsilon$ and since $\mu_{n+1}^\epsilon \phi_{n+1}^\epsilon$ converges in $H^{-1}(\Omega)$ to $\mu^0 \phi$, then by Theorem 4.4 ϕ satisfies $B_0(\phi) = \mu^0 \phi$ and we have $\phi_{n+1}^\epsilon \to \phi \in H^1_{\Omega_0}(\Omega)$ strongly in $H^1(\Omega)$ and, furthermore, $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla \phi_{n+1}^\epsilon|^2 = 0$.

Since we also have $\langle \phi_{n+1}^\epsilon, \phi_j^\epsilon\rangle_\Omega = 0$, for $j = 1, \dots, n$, then in the limit, $\langle \phi, \phi_j\rangle_\Omega = 0$. Consequently, $\mu^0 = \tau_0(\phi, \phi) \ge \mu_{n+1}$. Since the argument is independent of the subsequence we take, we really have proved $\liminf_\epsilon \mu_{n+1}^\epsilon \ge \mu_{n+1}$. Hence $\lim_\epsilon \mu_{n+1}^\epsilon = \mu_{n+1}$ and the function ϕ above, which may depend on the subsequence, is a normalized eigenfunction of μ_{n+1}. Therefore, the theorem is proved for $n+1$. ∎

COROLLARY 5.2. *Under the above conditions* (i) *If* $\mu \in \rho(B_0)$ *then for sufficiently small* ϵ, $\mu \in \rho(B_\epsilon)$. (ii) *If* $(a,b) \subset \mathbb{R}$ *is a bounded interval and* $K_\epsilon = \#\{\mu_n^\epsilon \in \sigma(B_\epsilon),\ \mu_n^\epsilon \in (a,b)\}$ *(counting multiplicities), then for sufficiently small* ϵ, K_ϵ *is independent of* ϵ. (iii) *If* $\mu \in \mathbb{R}$ *is such that* $\mu \in \rho(B_0)$, *and for* $0 \leq \epsilon$ *we take* $M_\epsilon = \{\mu_n^\epsilon \in \sigma(B_\epsilon),\ \mu_n^\epsilon < \mu\}$ *and we define* P_μ^ϵ *as the orthogonal projection onto the subspace* $\bigoplus_{\lambda \in M_\epsilon} E(\lambda, \epsilon)$, *then*

$$P_\mu^\epsilon \to P_\mu^0, \ \text{as } \epsilon \to 0$$

in the norm of $\mathcal{L}(H^1(\Omega))$, *where* $E(\lambda,\epsilon)$ *denotes the eigenspace of the eigenvalue* λ *of* B_ϵ.

Proof. Parts i and ii follow from the pointwise convergence of the eigenvalues. Now, for part iii note that the dimension of $\bigoplus_{\lambda \in M_\epsilon} E(\lambda,\epsilon)$ is finite and independent of ϵ. Therefore, after suitable choice of eigenfunctions we have for $\epsilon > 0$, $P_\mu^\epsilon = \sum_{i=1}^M \langle \phi_i^\epsilon, \cdot \rangle_\Omega \phi_i^\epsilon$ for some fixed $M \in \mathbb{N}$. From the theorem we get that for any sequence that we still denote $\epsilon \to 0$ there exists a subsequence ϵ_j such that $P_\mu^{\epsilon_j}$ converges to P_μ^0 strongly in $\mathcal{L}(L^2(\Omega), H^1(\Omega))$. Since the limit is independent of the subsequence we get that the full family converges. Finally, note that we are in a position to apply [16, Lemma V.9.12], and hence we get the convergence in the operator norm. ∎

6. Further comments and remarks. Now we discuss some different situations that can be handled similarly. Therefore, we just give the main arguments.

Concerning other boundary conditions, assume now $\Gamma = \Gamma_1 \cup \Gamma_2$ is a partition of the boundary, where $\Gamma_2 \neq \emptyset$ and Γ_1 could be empty, and consider the mixed problem with respect to this partition; that is, we impose the Neumann condition on Γ_1 and the Dirichlet condition on Γ_2. Note that when $\Gamma_1 = \emptyset$ we are solving the Dirichlet case. Therefore, we consider

$$\begin{cases} -Div(d_\epsilon(x)\nabla u^\epsilon) + (\lambda + V_\epsilon)u^\epsilon = f^\epsilon & \text{on } \Omega, \\ \frac{\partial u^\epsilon}{\partial \vec{n}_\epsilon} + b_\epsilon u^\epsilon = g^\epsilon & \text{on } \Gamma_1, \\ u^\epsilon = j^\epsilon & \text{on } \Gamma_2 \end{cases} \tag{6.1}$$

for which a weak formulation, using the bilinear form τ_ϵ defined in (4.2), is given as follows. Consider the Sobolev space

$$H^1_{\Gamma_2}(\Omega) = \{u \in H^1(\Omega),\ u = 0 \text{ on } \Gamma_2\}$$

whose dual space is denoted $H^{-1}_{\Gamma_2}(\Omega)$. Note that when $\Gamma_1 = \emptyset$, then $H^1_{\Gamma_2}(\Omega) = H^1_0(\Omega)$. Assume V_ϵ as in (1.4) and $b_\epsilon \in L^{q_1}(\Gamma_1)$ and converges to b. Then, for $h^\epsilon \in H^{-1}_{\Gamma_2}(\Omega)$ and $j^\epsilon \in H^{1/2}(\Gamma_2)$, a solution of the mixed problem is $u^\epsilon \in H^1(\Omega)$ such that

$$\begin{cases} \tau_\epsilon(u^\epsilon, \phi) = \displaystyle\int_\Omega d_\epsilon \nabla u^\epsilon \nabla \phi + \int_\Omega (\lambda + V_\epsilon) u^\epsilon \phi + \int_{\Gamma_1} b_\epsilon u^\epsilon \phi = \langle h^\epsilon, \phi \rangle_{-1,1}, \\ \gamma_2(u^\epsilon) = j^\epsilon \text{ on } \Gamma_2, \end{cases} \tag{6.2}$$

where the first equation holds for every $\phi \in H^1_{\Gamma_2}(\Omega)$.

For the homogeneous case, i.e., when $j^\epsilon = 0$, all the results of the previous sections apply by working on spaces with subscripts Γ_2 indicating that all functions have traces that vanish on Γ_2. Now, for the full problem (6.1) when $j^\epsilon \neq 0$, we construct a function J^ϵ in Ω with trace j^ϵ on Γ_2, and we write an equation for $v^\epsilon = u^\epsilon - J^\epsilon$ thus reducing

the problem to the homogeneous case. For this we have to control J^ϵ on Ω_0 and the way it depends on j^ϵ. For this assume that j^ϵ is such that

$$\hat{j}^\epsilon = \begin{cases} j^\epsilon & \text{on } \Gamma_2 \\ 0 & \text{on } \Gamma_1 \end{cases} \in H^{1/2}(\Gamma).$$

This condition holds true if $j^\epsilon \in H^{1/2}(\Gamma_2)$ and if $\bar{\Gamma}_1 \cap \bar{\Gamma}_2 = \emptyset$.

LEMMA 6.1. *With the above assumption, assume $\lambda > 0$ and let J^ϵ be defined as the unique solution of*

$$\begin{cases} -Div(d_\epsilon(x)\nabla J^\epsilon) + \lambda J^\epsilon = 0 & \text{on } \Omega_1, \\ \gamma(J^\epsilon) = 0 & \text{on } \Gamma_0 \cup \Gamma_1, \\ \gamma(J^\epsilon) = j^\epsilon & \text{on } \Gamma_2, \end{cases}$$

and we extend J^ϵ by zero to Ω_0; hence, $J^\epsilon \in H^1_{\Omega_0}(\Omega)$. Then the following conditions hold.

(i) *If $\{\hat{j}^\epsilon\}_\epsilon$ is bounded in $H^{1/2}(\Gamma)$, then $\{J^\epsilon\}_\epsilon$ is bounded in $H^1(\Omega)$. Moreover, if $\hat{j}^\epsilon \to j$ weakly in $H^{1/2}(\Gamma)$, then $J^\epsilon \to J$ weakly in $H^1(\Omega)$.*

(ii) *If $\hat{j}^\epsilon \to j$ strongly in $H^{1/2}(\Gamma)$, then $J^\epsilon \to J$ strongly in $H^1(\Omega)$.*

With this special choice for J^ϵ we have the following theorem.

THEOREM 6.2. *Assume V_ϵ, b_ϵ verify* (1.4) *and* (1.5)*, respectively, $h^\epsilon \to h$ weakly in $H^{-1}_{\Gamma_2}(\Omega)$, $\hat{j}^\epsilon \to j$ weakly in $H^{1/2}(\Gamma)$, λ is large enough, and u^ϵ is the unique solution of* (6.2)*. Then*

$$u^\epsilon \to u \in H^1_{\Omega_0}(\Omega)$$

weakly in $H^1(\Omega)$, where u verifies

$$\begin{cases} \tau_0(u,\phi) = \displaystyle\int_{\Omega_1} d_0 \nabla u \nabla \phi + \int_\Omega (\lambda + V) u\phi + \int_{\Gamma_1} bu\phi = \langle h, \phi\rangle_{-1,1}, \\ \gamma_2(u) = j \ \text{ on } \Gamma_2 \end{cases} \tag{6.3}$$

for every $\phi \in H^1_{\Gamma_2,\Omega_0}(\Omega) = H^1_{\Omega_0}(\Omega) \cap H^1_{\Gamma_2}(\Omega)$. Moreover, if the convergence for h^ϵ and $\hat{j}^\epsilon$ is strong, then u^ϵ converges strongly. Furthermore, $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla u^\epsilon|^2 = 0$.

In particular, if $h = f_\Omega + g_{\Gamma_1}$ with $f \in L^{p_0'}(\Omega)$ and $g \in H^{-1/2}(\Gamma_1)$, then u verifies

$$\begin{array}{ll} -Div(d_0(x)\nabla u) + (\lambda + V(x))u = f & \text{on } \Omega_1, \\ \frac{\partial u}{\partial \vec{n}_0} + bu = g & \text{on } \Gamma_1, \\ \gamma_2(u) = j & \text{on } \Gamma_2, \\ \gamma_{0,i}(u) = u_{\Omega_{0,i}} & \text{on } \Gamma_{0,i}, \quad i = 1, \dots, m, \\ \displaystyle\frac{1}{|\Omega_{0,i}|} \int_{\Gamma_{0,i}} \frac{\partial u}{\partial \vec{n}_0} + \left(\fint_{\Omega_{0,1}} \lambda + V \right) u_{\Omega_{0,i}} = \fint_{\Omega_{0,i}} f, & i = 1, \dots, m, \end{array} \tag{6.4}$$

where $u_{\Omega_{0,i}}$ denotes the constant value of u on $\Omega_{0,i}$ for $i = 1, \dots, m$.

Proof. If u^ϵ is a solution of (6.2), then the function $v^\epsilon = u^\epsilon - J^\epsilon \in H^1_{\Gamma_2}(\Omega)$ verifies

$$\tau_\epsilon(v^\epsilon, \phi) = \langle h^\epsilon, \phi\rangle_{-1,1} - \int_{\Omega_1} d_\epsilon \nabla J^\epsilon \nabla \phi - \int_{\Omega_1} (\lambda + V_\epsilon) J^\epsilon \phi - \int_{\Gamma_1} b_\epsilon J^\epsilon \phi \overset{def}{=} \langle H^\epsilon, \phi\rangle_{-1,1}$$

for every $\phi \in H^1_{\Gamma_2}(\Omega)$. From the lemma, we get that if $h^\epsilon \to h$ weakly (strongly) in $H^{-1}_{\Gamma_2}(\Omega)$ and $\hat{j}^\epsilon \to j$ weakly (strongly) in $H^{1/2}(\Gamma)$, then $H^\epsilon \to H$ weakly (strongly)

in $H^{-1}_{\Gamma_2}(\Omega)$, where H is defined by $\langle H,\phi\rangle_{-1,1} = \langle h,\phi\rangle_{-1,1} - \int_{\Omega_1} d_0 \nabla J \nabla\phi - \int_{\Omega_1}(\lambda + V)J\phi - \int_{\Gamma_1} bJ\phi$.

From the result for the homogeneous case above, we get $v^\epsilon \to v \in H^1_{\Gamma_2,\Omega_0}(\Omega)$ weakly in $H^1_{\Gamma_2}(\Omega)$ (strongly and $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla v^\epsilon|^2 = 0$). Also, v satisfies $T_0(v) = H_{|H^1_{\Gamma_2,\Omega_0}(\Omega)}$ in $H^{-1}_{\Gamma_2,\Omega_0}(\Omega)$. Therefore, $u^\epsilon \to u = v + J \in H^1_{\Omega_0}(\Omega)$ weakly in $H^1(\Omega)$ (strongly and $\lim_\epsilon \int_{\Omega_0} d_\epsilon |\nabla u^\epsilon|^2 = 0$). The rest is obvious. □

We can also consider, without any change in the analysis, the same problem for diffusion operators of the form

$$-Div(D_\epsilon(x)\nabla u),$$

where $D_\epsilon(x)$ is a positive definite symmetric matrix for each $x \in \Omega$ with smooth coefficients and such that if $\lambda^\epsilon_1(x)$ denotes the first eigenvalue of $D_\epsilon(x)$, then $0 < m_0 \le \lambda^\epsilon_1(x)$ for every $x \in \Omega$ and $0 \le \epsilon \le \epsilon_0$, and $\lambda^\epsilon_1(x) \longrightarrow \infty$ uniformly on compact subsets of Ω_0, as $\epsilon \to 0$, and $D_\epsilon(x) \longrightarrow D_0(x)$ uniformly on Ω_1, as $\epsilon \to 0$. In this case, the conormal derivative is given by $\frac{\partial u}{\partial \vec{n}_\epsilon} = \langle D_\epsilon(x)\nabla u, \vec{n}\rangle = \langle \nabla u, D^*_\epsilon(x)\vec{n}\rangle$. Also, the case of discontinuous diffusion coefficients d_ϵ can be dealt with no change, since no elliptic regularity results are needed in most of our analysis; see [17].

Acknowledgments. The author wishes to acknowledge the hospitality of the Instituto de Ciencias Matematicas de São Carlos-USP, São Paulo, Brazil, and to acknowledge useful discussions with Professor Alexandre Carvalho.

REFERENCES

[1] S. S. Antman, *Nonlinear Problems of Elasticity*, Springer-Verlag, New York, 1995.
[2] H. Brezis and T. Kato, *Remarks on the Schrödinger operator with complex potentials,* J. Math. Pures Appl., 58 (1979), pp. 137–151.
[3] A. N. Carvalho, *Spatial homogeneity in damped hyperbolic equations,* Dynam. Systems Appl., 1 (1992), pp. 221–250.
[4] A. N. Carvalho and J. K. Hale, *Large diffusion with dispersion,* Nonlinear Anal., 17 (1991), pp. 1139–1151.
[5] A. N. Carvalho and A. L. Pereira, *A scalar parabolic equation whose asymptotic behavior is dictated by a system of ordinary differential equations,* J. Differential Equations, 112 (1994), pp. 81–130.
[6] A. N. Carvalho and L. A. F. Oliveira, *Delay–partial differential equations with some large diffusion,* Nonlinear Anal., 22 (1994), pp. 1057–1095.
[7] A. N. Carvalho and A. Rodriguez-Bernal, *Upper Semicontinuity of Attractors for Parabolic Equations with Localized Large Diffusion and Nonlinear Boundary Conditions,* in preparation.
[8] P. G. Ciarlet, *Mathematical Elasticity,* North–Holland, Amsterdam, 1988.
[9] E. Conway, D. Hoff, and J. Smoller, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations,* SIAM J. Appl. Math., 35 (1978), pp. 1–16.
[10] G. Fusco, *On the explicit construction of an ODE which has the same dynamics as a scalar parabolic PDE,* J. Differential Equations, 69 (1987), pp. 85–110.
[11] J. Hale, *Large diffusivity and asymptotic behavior in parabolic systems,* J. Math. Anal. Appl., 118 (1986), pp. 455–466.
[12] J. Hale and C. Rocha, *Varying boundary conditions and large diffusivity,* J. Math. Pures Appl., 66 (1987), pp. 139–158.
[13] J. K. Hale and K. Sakamoto, *Shadow systems and attractors in reaction-diffusion equations,* Appl. Anal., 32 (1989), pp. 287–303.
[14] T. Kato, *Perturbation Theory of Linear Operators,* Springer-Verlag, New York, 1980.
[15] A. Rodriguez-Bernal, *On the construction of inertial manifolds under symmetry constraints* II: $O(2)$ *constraint and inertial manifolds on thin domains,* J. Math. Pures Appl., 72 (1993), pp. 57–79.

[16] J. Sánchez-Hubert and E. Sánchez-Palencia, *Vibration and Coupling of Continuous Systems. Asymptotic Methods,* Springer-Verlag, New York, 1989.
[17] G. Stampacchia, *Le probléme de Dirichlet por les équations elliptiques du second ordre à coefficients discontinus,* Ann. Inst. Fourier-Grenoble, 15 (1965), pp. 189–258.
[18] M. Valencia and J. Solá-Morales, *Trend to spatial homogeneity for solutions of semilinear damped wave equations,* Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 117–126.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1381–1398, November 1998

004

FINITE SPEED OF PROPAGATION FOR THE POROUS MEDIA EQUATION*

S. BONAFEDE†, G. R. CIRMI†, AND A. F. TEDEEV‡

Abstract. We consider the Cauchy–Dirichlet problem for the equation

$$u_t = \Delta u^m, \quad m > 1, \quad \text{on } D = \Re_k^N \times (t > 0),$$

where $\Re_k^N = \Re^N \cap \{x_1, ..., x_k > 0\}$, $1 \le k \le N$, $N \ge 1$. Sharp bounds of the interface (or free boundary) are obtained. We use a weighted energy method, which allows us to consider more general equations.

Key words. porous media equation, energy solution, free boundary

AMS subject classifications. 35K60, 35B05, 35K65

PII. S0036141096298072

1. Introduction. This paper is concerned with the qualitative properties of the solution of the following initial-boundary value problem:

$$\begin{cases} \frac{\partial u}{\partial t} = \Delta u^m & \text{in } D = \Re_k^N \times (0, \infty), \\ u(x,t) = 0 & \text{on } \partial\Re_k^N \times (0, \infty), \\ u(x,0) = u_o(x) & \text{on } \Re_k^N, \end{cases} \tag{1}$$

where $\Re_k^N = \Re^N \cap \{x_1, ..., x_k > 0\}$, $1 \le k \le N$, $N \ge 1$, $m > 1$. In [5], [7], [11], [13] (see also the references therein) the authors investigated the qualitative properties of the solution of the Cauchy problem for more general parabolic equations, including the porous media equation. Moreover, it is well known that the energy solutions of the initial-boundary value problem for the porous media equation in any unbounded open set of $\Re^N$ have the property of so-called finite speed of propagation of perturbations, provided that the initial data have compact support.

In [7], estimates on the growth of the free boundary defined by support $u(., t)$ are found. As a matter of fact, assuming u_o compactly supported and setting

$$\zeta(T) = \text{supremum}\,\{|x| : x \in \text{support } u(., T)\},$$

Corollary 6.1 of [7] guarantees that $\zeta(T) = O(T^\beta)$, where $\beta = \frac{1}{N(m-1)+2}$ is the Barenblatt exponent.

Of course, this result still holds in our case.

Our aim is to find new bounds on $\zeta(T)$. Indeed, we shall prove that if the support of the initial datum is bounded, then for the solution of problem (1) we have the following:

$$\gamma_1 \mu_k(0)^{\frac{\sigma(m-1)}{N+k}} T^{\frac{\sigma}{N+k}} \le \zeta(T) \le \gamma_2 \mu_k(0)^{\frac{\sigma(m-1)}{N+k}} T^{\frac{\sigma}{N+k}} \quad \forall T > T_o, \tag{2}$$

*Received by the editors January 25, 1996; accepted for publication (in revised form) July 8, 1997; published electronically July 7, 1998.
http://www.siam.org/journals/sima/29-6/29807.html
†Department of Mathematics, University of Catania, Viale A. Doria 6, 95125 Catania, Italy (bonafede@dipmat.unict.it, cirmi@dipmat.unict.it).
‡Institute of Applied Mathematics and Mechanics of National Academy of Sciences of Ukraine, 340114, Roza Luxemburg st.74, Donetsk, Ukraine (tedeev@iamm.ac.donetsk.ua).

where γ_1, γ_2 are positive constants, $\mu_k(0)$ is the moment defined by

$$\mu_k(0) = \int_{\Re_k^N} x_1 \cdots x_k u_o(x) dx,$$

σ is the number defined by

$$\sigma = \frac{N+k}{(N+k)(m-1)+2},$$

and T_o is a sufficiently large constant.

A result of this type has been obtained in [11], but in the case $k = 1$. The method used in [11] is different from ours since it is based on the construction of the so-called dipole solutions and doesn't permit consideration of more general situations, as problem (1).

In order to get the above result, we use a weighted energy method, adapted from [5], [7], combined with the moment technique of [15], [16]. The main tools are some technical inequalities together with an estimate on the $\|u(.,t)\|_{L^\infty(\Re_k^N)}$.

The second result of our paper deals with the finite speed of propagation of the solution of the following initial-boundary value problem:

$$\begin{cases} \frac{\partial}{\partial t}(|u|^{r-1}u) + (-1)^l u^{(2l)} = 0 & \text{in } Q = \Re_+ \times (t > 0), \\ u^{(j)} = 0,\ j = 0, 1, ..., l-1, & \text{on } \{0\} \times (t > 0), \\ u(x, 0) = u_o(x) & \text{on } \Re_+, \end{cases} \tag{3}$$

where $0 < r < 1$, $u^{(j)} = \frac{\partial^j u}{\partial x^j}$, $l > 1$ is an integer. In section 4, we shall prove that if the support of u_o is bounded, then for any $T > 0$, sufficiently large, the following estimate holds:

$$\zeta(T) \le \gamma_3 T^{\frac{r+1}{2(l-1)(r+1)+4}}, \tag{4}$$

with γ_3 positive constant.

If, in addition,

$$\sup_{0<t<\infty} \int_{\Re_+} x|u|^r dx = \mu_o < \infty, \tag{5}$$

then

$$\zeta(T) \le \gamma_4 T^{\frac{r}{2(1+(l-1)r)}}, \tag{6}$$

where γ_4 is a positive constant. We point out that estimates (4) and (6) improve the results of [7, Corollaries 2.2 and 6.1]. Moreover, our results can be extended to the case of equations with lower order terms in multidimensional half-space (see [3]).

2. Statement of main results. Throughout the paper we will assume that in problem (1), $u_o \ge 0$ on $\Re_k^N$,

$$u_o \in L_1(\Re_k^N) \cap L_\infty(\Re_k^N),$$

and in problem (3),

$$u_o \in L_{r+1}(\Re_+).$$

Let us denote by W the closure of $C_o^\infty(\Re_k^N)$ with respect to the seminorm

$$\left(\int_{\Re_k^N} |Dv|^2 dx\right)^{\frac{1}{2}}.$$

Let us set $q = \frac{1}{m}$.

DEFINITION 1. *Given u_o in* (1)*, we say that u is an energy solution to* (1) *if $u \geq 0$ in D,*

$$u^m \in C([0,\infty); L_{q+1}(\Re_k^N)) \cap L_2(0,T;W) \cap L_\infty(0,T; L_\infty(\Re_k^N)),$$

$\forall$ finite $T > 0$, and u satisfies the equation of problem (1) *in $\mathcal{D}'(D)$, together with the initial-boundary conditions.*

Let us denote by V the closure of $C_o^\infty(\Re_+)$ with respect to the seminorm

$$\left(\int_{\Re_+} \|u^{(l)}\|^2 dx\right)^{\frac{1}{2}}.$$

DEFINITION 2. *Given u_o in* (3)*, we say that u is an energy solution to* (3) *if*

$$u \in C([0,\infty); L_{r+1}(\Re_+)) \cap L_2(0,T;V)$$

$\forall$ finite $T > 0$, and u satisfies the equation of (3) *in $\mathcal{D}'(\Re_+)$, together with the initial-boundary conditions.*

The existence of an energy solution to problems (1) and (3) follows from the results of [6]. Related existence results can be found in [9], [1], and [12].

Our main goal is to prove the following.

THEOREM 1. *Let $u(x,t)$ be an energy solution of problem* (1)*. Assume that*

$$\textit{support } u_o \subset B_{R_o}^k,$$

where

$$B_{R_o}^k = \left\{x \in \Re_k^N, |x| < R_o\right\}.$$

Then the estimates (2) *hold.*

Remark 1. We recall that, due to Corollary 6.1 of [7], $\zeta(T) = O(T^\beta)$, where $\beta = \frac{1}{N(m-1)+2}$ is the Barenblatt exponent, while estimate (2) states that $\zeta(T) = O(T^{\frac{\sigma}{N+k}})$. As we have remarked in the introduction, Theorem 1 improves Corollary 6.1 of [7], since $\frac{\sigma}{N+k} < \beta$.

In section 4 we shall prove the following.

THEOREM 2. *Let $u(x,t)$ be an energy solution of problem* (3)*. Assume that*

$$\textit{support } u_o \in (0, R_o).$$

Then estimate (4) *holds. If, in addition, $u(x,t)$ satisfies condition* (5)*, then estimate* (6) *is also true.*

Remark 2. As a consequence of Corollary 2.1 of [7], for $T > 0$ sufficiently large,

$$\zeta(T) \leq c_o T^{\beta_o}, \tag{7}$$

where $\beta_o = \frac{r+1}{(2l-1)(r+1)+2}$, so the upper bound (4) improves (7). Furthermore, if $u \in L^\infty(0,T;L^r(\Omega))$, due to Corollary 6.1 of [7], we deduce

$$\zeta(T) \le c_1 T^{\beta_1} \quad \text{for } T>0 \text{ sufficiently large,} \tag{8}$$

with $\beta_1 = \frac{r}{(2l-1)(r+1)+1}$. Since $r \in (0,1)$, estimate (6) is better than (8).

However, we must remark that assumption (5) is stronger than assumption $u \in L^\infty(0,T;L^r(\Omega))$ and also, that we don't know under which conditions the energy solutions of problem (3) satisfy assumption (5).

3. Some technical lemmas. We recall here some technical inequalities which we shall employ in the proof of our theorems.

LEMMA 1 (Hardy's inequality). *Let $\theta>0$, $R>0$, and*

$$\int_R^\infty y(y-R)^{\theta+1}u_y^2 dy < \infty.$$

Then

$$\int_R^\infty y(y-R)^{\theta-1}u^2 dy \le \frac{4}{\theta^2}\int_R^\infty y(y-R)^{\theta+1}u_y^2 dy. \tag{9}$$

For the proof, see [10, Theorem 330]. Let us set

$$X_k = x_1 \cdots x_2 \cdots x_k.$$

LEMMA 2 (weighted Nirenberg–Gagliardo inequality). *Let $\omega \in W_2^1(\Re_k^N)\cap L_\beta(\Re_k^N)$ and $0<\beta<1<\lambda\le 2$. Then,*

$$\int_{\Omega_R^+} X_k|\omega|^\lambda dx$$

$$\le (\gamma(N,k,\beta))^\lambda \left(\int_{\Omega_R^+} X_k|\nabla\omega|^2 dx\right)^{\theta_1}\left(\int_{\Omega_R^+} X_k|\omega|^\beta dx\right)^{\theta_2}, \tag{10}$$

where

$$\theta_1 = \frac{(\lambda-\beta)(N+k)}{(N+k)(2-\beta)+2\beta}$$

and

$$\theta_2 = \frac{(2-\lambda)(N+k)+2\lambda}{(N+k)(2-\beta)+2\beta}.$$

Proof. Let

$$\Omega_{N+k} = \Re_k^N \times \{(z_1,...,z_k) : 0<z_i<x_i, \quad i=1,...,k\}$$

and $\overline{\omega} \in W_2^1(\Omega_{N+k}) \cap L_\beta(\Omega_{N+k})$. Let us put

$$\nabla_{x,z}\overline{\omega} = \left(\frac{\partial\overline{\omega}}{\partial x_1},...,\frac{\partial\overline{\omega}}{\partial x_N},\frac{\partial\overline{\omega}}{\partial z_1},...,\frac{\partial\overline{\omega}}{\partial z_N}\right),$$

$$\overline{E}_q(\overline{\omega}) = \int_{\Omega_{N+k}} |\overline{\omega}|^q dxdz,$$

$$\overline{J}(\overline{\omega}) = \int_{\Omega_{N+k}} |\nabla_{x,z}\overline{\omega}|^2 dxdz,$$

$$E_{k,q}(\omega) = \int_{\Re_k^N} X_k |\omega|^q dx,$$

$$J(\omega) = \int_{\Re_k^N} X_k |\nabla\omega|^2 dx.$$

By virtue of the Nirenberg–Gagliardo inequality (see Theorem 58.XII of [14]), there exists a constant γ, which depends only on $N+k, \beta$ such that

$$\overline{E}_\lambda(\overline{\omega}) \le \gamma^\lambda (\overline{J}(\overline{\omega}))^{\theta_1} (\overline{E}_\beta(\overline{\omega}))^{\theta_2}. \tag{11}$$

If we take $\overline{\omega}(x,z) = \omega(x)$, we have

$$\overline{E}_q(\omega) = E_{k,q}(\omega), \qquad \overline{J}(\omega) = J(\omega),$$

and so, from (11) we complete the proof.

LEMMA 3. *Assume u is an energy solution of problem* (1) *and $\rho(x) \in W^1_\infty(\Re_k^N)$, $\rho \ge 0$ on $\Re_k^N$. Then, for all T_1, T_2, $0 \le T_1 < T_2$ the following inequality is true:*

$$\frac{q}{q+1}\int_{\Re_k^N} \rho(x) v^{q+1}(x,T_2)dx + \int_{T_1}^{T_2}\int_{\Re_k^N} DvD(\rho v)dxdt$$

$$= \frac{q}{q+1}\int_{\Re_k^N} \rho(x) v^{q+1}(x,T_1)dx, \tag{12}$$

where $v = u^m$.

For the proof, we refer to Proposition 3.3 of [7].

LEMMA 4. *Let u be a solution of problem* (1) *and*

$$\zeta_1(T) = \textit{supremum}\,\{x_1 : x \in \textit{support}\ u(.,T)\}.$$

Then, for all $T_1, T_2 > 0$, $T_1 < T_2 < T$ we have

$$\zeta_1(T_2) - \zeta_1(T_1) \le \gamma (T_2 - T_1)^{\beta_o} \left(\int_{\Re_k^N} X_k u^{m+1}(x,T_1)dx \right)^{\lambda_0}, \tag{13}$$

where

$$\beta_o = \frac{m+1}{(N+k)(m-1)+2(m+1)},$$

$$\lambda_o = \frac{m-1}{m+1}\beta_o,$$

and γ is a positive constant which depends on N, k, m.

Proof. In view of the known results for the Cauchy–Dirichlet problem for the porous media equation in any unbounded domain (see [7, Corollary 6.1]), we deduce

$$\text{support } u(x,t) \subset B_{R(T)},$$

where

$$R(T) \le R_o + \gamma T^{\frac{1}{N(m-1)+2}}.$$

Let us denote

$$(x_1 - R)_+^s = \begin{cases} (x_1 - R)^s & \text{for} \quad x_1 > R, \\ 0 & \text{for} \quad x_1 \le R. \end{cases}$$

Then, choosing in (12),

$$\rho(x) = X_k (x_1 - R)_+^s, \quad T_1 = 0, \quad T_2 = T > 0, \quad R > R_o,$$

with s a sufficiently large positive number, we infer

$$\operatorname*{ess\,sup}_{0 \le t \le T} \int_{\Omega_R^+} X_k (x_1 - R)^s v^{q+1} dx + \int_0^T \int_{\Omega_R^+} X_k (x_1 - R)^s |\nabla v|^2 dx dt$$

$$\le \gamma \int_0^T \int_{\Omega_R^+} X_k (x_1 - R)^{s-1} |\nabla v| v dx dt$$

$$+ \int_0^T \int_{\Omega_R^+} \sum_{i=1}^k X_k^i v_{x_i} v (x_1 - R)^s dx dt = I_1 + I_2, \tag{14}$$

where

$$X_k^i = x_1 \cdots x_{i-1} x_{i+1} \cdots x_k$$

and

$$\Omega_R^+ = \Re_k^N \setminus \{x_1 \le R\}.$$

Let us estimate I_1, I_2. Integrating by parts, we obtain

$$I_2 = -\frac{s}{2} \int_0^T \int_{\Omega_R^+} x_2 \cdots x_k v^2 (x_1 - R)^{s-1} dx dt$$

$$\le \frac{s}{2} \int_0^T \int_{\Omega_R^+} X_k (x_1 - R)^{s-2} v^2 dx dt. \tag{15}$$

Moreover, the application of Young's inequality yields

$$I_1 \le \frac{\epsilon}{2} \int_0^T \int_{\Omega_R^+} X_k (x_1 - R)^s |\nabla v|^2 dx dt$$

$$+\frac{1}{2\epsilon}\int_0^T\int_{\Omega_R^+} X_k(x_1-R)^{s-2}v^2dxdt. \tag{16}$$

Thus, from (14)–(16) and for a sufficiently small ϵ, it follows that

$$\operatorname*{ess\,sup}_{0\le t\le T}\int_{\Omega_R^+} X_k(x_1-R)^s v^{q+1}dx+\int_0^T\int_{\Omega_R^+} X_k(x_1-R)^s|\nabla v|^2dxdt$$

$$\le \gamma\int_0^T\int_{\Omega_R^+} X_k(x_1-R)^{s-2}v^2dxdt. \tag{17}$$

Now, we would like to estimate the right-hand side of the last inequality. First, by virtue of Hardy's inequality (9), we can establish the following estimate:

$$\int_0^T\int_{\Omega_R^+} X_k(x_1-R)^{s-2}v^2dxdt\le\gamma\int_0^T\int_{\Omega_R^+} X_k(x_1-R)^s|\nabla v|^2dxdt.$$

So, if we set

$$\mathcal{F}_s=\mathcal{F}_s(T,R)=\operatorname*{ess\,sup}_{0\le t\le T}\int_{\Omega_R^+} X_k(x_1-R)^s v^{q+1}dx,$$

$$\mathcal{I}_s=\mathcal{I}_s(T,R)=\int_0^T\int_{\Omega_R^+} X_k(x_1-R)^s|\nabla v|^2dxdt,$$

from (17) we can deduce

$$\mathcal{F}_s\le\gamma\mathcal{I}_s. \tag{18}$$

Now, we can use the following weighted Nirenberg–Gagliardo inequality (see Lemma 2):

$$\int_{\Omega_R^+} X_k^1|\omega|^2dx$$

$$\le\gamma\left(\int_{\Omega_R^+} X_k^1|\nabla\omega|^2dx\right)^a\left(\int_{\Omega_R^+} X_k^1|\omega|^{q+1}dx\right)^{\frac{2(1-a)}{1+q}}, \tag{19}$$

where

$$X_k^1=x_2\cdots x_k$$

and

$$a=\frac{(N+k-1)(1-q)}{(N+k-1)(1-q)+2(q+1)}.$$

Setting in this inequality

$$\omega=(x_1(x_1-R)^{s-2})^{\frac{1}{2}}v,$$

we have

$$\int_{\Omega_R^+} X_k(x_1-R)^{s-2}v^2dx \le \gamma\left(\int_{\Omega_R^+} X_k(x_1-R)^{s-2}|\nabla v|^2dx\right.$$

$$\left.+\int_{\Omega_R^+} X_k(x_1-R)^{s-4}v^2dx\right)^a\left(\int_{\Omega_R^+} X_k^1(x_1(x_1-R)^{s-2})^{\frac{q+1}{2}}v^{q+1}dx\right)^{\frac{2(1-a)}{q+1}},$$

from which, using also Hardy and Young's inequalities, we obtain

$$\int_{\Omega_R^+} X_k(x_1-R)^{s-2}v^2dx$$

$$\le \gamma\left(\int_{\Omega_R^+} X_k(x_1-R)^{s-2}|\nabla v|^2dx\right)^a\left(\int_{\Omega_R^+} X_k v^{q+1}dx\right)^{\frac{\theta 2(1-a)}{q+1}}$$

$$\cdot\left(\int_{\Omega_R^+} X_k(x_1-R)^s v^{q+1}dx\right)^{\frac{2(1-\theta)(1-a)}{q+1}}, \tag{20}$$

where

$$\theta=\frac{s(1-q)+2(q+1)-q+1}{2s}.$$

Let us denote

$$\mathcal{E}_s=\mathcal{E}_s(t,R)=\int_{\Omega_R^+} X_k(x_1-R)^s v^{q+1}dx,$$

$$\mathcal{J}_s=\mathcal{J}_s(t,R)=\int_{\Omega_R^+} X_k(x_1-R)^s|\nabla v|^2dx.$$

Then, with the help of (20), inequality (17) may be rewritten as follows:

$$\mathcal{I}_s\le\gamma\int_0^T \mathcal{J}_{s-2}^a(t).\mathcal{E}_o^{\frac{2\theta(1-a)}{q+1}}.\mathcal{E}_s^{\frac{2(1-\theta)(1-a)}{q+1}}dt.$$

Estimating the right-hand side of the last inequality by Holder's inequality and also using (18), we can obtain

$$\mathcal{I}_s\le\gamma\mathcal{I}_{s-2}^a\mathcal{F}_o^{\frac{2\theta(1-a)}{q+1}}\mathcal{F}_s^{\frac{2(1-\theta)(1-a)}{q+1}}T^{1-a},$$

$$\le\gamma\mathcal{I}_{s-2}^a\mathcal{I}_o^{\frac{2\theta(1-a)}{q+1}}\mathcal{I}_s^{\frac{2(1-\theta)(1-a)}{q+1}}T^{1-a}. \tag{21}$$

Again, by virtue of Holder's inequality, it easily follows that

$$\begin{aligned}&\mathcal{I}_{s-2}\le\mathcal{I}_s^{\frac{s-2}{s}}\mathcal{I}_o^{\frac{2}{s}},\\ &\mathcal{I}_1\le\mathcal{I}_s^{\frac{1}{s}}\mathcal{I}_o^{\frac{s-1}{s}}.\end{aligned} \tag{22}$$

Estimating in (21) $\mathcal{I}_{s-2}$ from above and $\mathcal{I}_s$ from below, after easy manipulations we get

$$\mathcal{I}_1^{\lambda_1}(R,T) \le \gamma \mathcal{I}_o T^{\beta_1}, \tag{23}$$

where

$$\lambda_1 = \frac{(N+k)(1-q)+2(q+1)}{(N+k)(1-q)+2(q+1)+1-q}$$

and

$$\beta_1 = \frac{q+1}{(N+k)(1-q)+2(q+1)+1-q}.$$

Noticing that

$$\mathcal{I}_o(R,T) = -\frac{d\mathcal{I}_1}{dR},$$

from (23) we deduce

$$T^{-\beta_1}\mathcal{I}_1^{\lambda_1} \le -\gamma \frac{d\mathcal{I}_1}{dR}.$$

Hence

$$\mathcal{I}_1^{1-\lambda_1}(R,T)$$

$$\le \mathcal{I}_1^{1-\lambda_1}(R_o,T) - \gamma(R-R_o)T^{-\beta_1}$$

$$\le \gamma \mathcal{I}_o^{\frac{1-\lambda_1}{\lambda_1}}(R_o,T)T^{\frac{(1-\lambda_1)\beta_1}{\lambda_1}} - \gamma(R-R_o)T^{-\beta_1},$$

and therefore

$$\zeta_1(T) \le R_o + \gamma \mathcal{I}_o^{\lambda_o} T^{\beta_o}, \tag{24}$$

where λ_o, β_o are defined in Lemma 5.

By a slight modification of the proof, we can also obtain that for any T_1, T_2, with $0 \le T_1 < T_2$,

$$\zeta_1(T_2) - \zeta_1(T_1) \le \gamma(T_2-T_1)^{\beta_o} \left(\int_{T_1}^{T_2} \int_{\Re_k^N} X_k |\nabla u^m|^2 dx dt \right)^{\lambda_o}. \tag{25}$$

Finally, taking in (12) $\rho(x) = X_k$, it follows that

$$\int_{T_1}^{T_2} \int_{\Re_k^N} X_k |\nabla u^m|^2 dx \le \frac{q+1}{q} \int_{\Re_k^N} X_k u^{m+1}(x,T_1) dx, \tag{26}$$

and so from (25) and (26) we get (13).

In order to improve inequality (13), we need an estimate on the $\|u(.,t)\|_{\infty,\Re_k^N}$-norm; for this aim, we follow the method used in [15], combined with the iterative technique of [2].

LEMMA 5. *Let u be an energy solution of problem* (1). *Then, for any $t > 0$, the following estimate holds:*

$$\|u(.,t)\|_{L^\infty(\Re_k^N)} \leq \gamma \mu_k(0)^{\frac{2\sigma}{N+k}} t^{-\sigma}, \tag{27}$$

where

$$\mu_k(0) = \int_{\Re_k^N} X_k u_o(x) dx$$

and

$$\sigma = \frac{N+k}{(N+k)(m-1)+2}.$$

Proof. Using a standard method, from the weak formulation of (1), for any $p \geq m$ we obtain

$$\frac{d}{dt} \int_{\Re_k^N} X_k u^{p+1} dx$$

$$= -\frac{4mp(p+1)}{(m+p)^2} \int_{\Re_k^N} X_k |\nabla u^{\frac{p+m}{2}}|^2 dx - (p+1) \int_{\Re_k^N} \sum_{i=1}^{k} (u^m)_{x_i} X_k^i u^p dx dt.$$

By virtue of Green's formula and taking into account the boundary condition, the second term on the right-hand side is zero. Therefore, for a.e. $t > 0$, we have

$$\frac{d}{dt} \int_{\Re_k^N} X_k u^{p+1} dx = -\frac{4mp(p+1)}{(m+p)^2} \int_{\Re_k^N} X_k |\nabla u^{\frac{p+m}{2}}|^2 dx. \tag{28}$$

Moreover, since support $u(x,t)$ is bounded, for any $t > 0$ we obtain

$$\mu_k(t) = \int_{\Re_k^N} X_k u(x,t) dx = \int_{\Re_k^N} X_k u_o dx = \mu_k(0). \tag{29}$$

For any n, positive integer, let us denote by

$$p_n = m + 2^n - 2,$$

$$E_n(t) = \int_{\Re_k^N} X_k u(x,t)^{p_n+1} dx.$$

We shall obtain estimate (27) by proving that, for any $t > 0$, there exist two constants θ_1, θ_2, independent of n, such that

$$E_n(t) \leq A_n \mu_k(0)^{\sigma \frac{(N+k)(m-1)+2(p_n+1)}{N+k}} t^{-\sigma p_n}, \tag{30}$$

where

$$A_n = \left(\prod_{i=1}^{n} (\theta_1 p_i)^{\frac{\theta_2}{p_i+1}} \right)^{p_n+1}.$$

We shall prove (30) by induction on n. For $n = 1$, it follows from (28) and the weighted Nirenberg–Gagliardo inequality (10), provided $\theta_1 \geq \max(1, \sigma, \gamma(N, k, \frac{1}{m}))$ and $\theta_2 \geq 3(m+1)\sigma$, where $\gamma(N, k, \frac{1}{m})$ is the constant which appears in the inequality (10). Assume that (30) holds for n and let us prove it for $n+1$. From (28), with $p = p_{n+1} + 1$ we deduce

$$\frac{d}{dt} \int_{\Re_k^N} X_k u^{p_{n+1}+1} dx \leq - \int_{\Re_k^N} X_k |\nabla u^{p_n+1}|^2 dx; \tag{31}$$

moreover, from inequality (10), we have

$$\int_{\Re_k^N} X_k |\nabla u^{p_n+1}|^2 dx \geq \frac{[E_{n+1}(t)]^{\alpha_n}}{[E_n(t)]^{\beta_n} [\gamma(N, k, 1)]^{\delta_n}}, \tag{32}$$

where

$$\alpha_n = \frac{(N+k+2)(p_n+1)}{(p_{n+1} - p_n)(N+k)},$$

$$\beta_n = \frac{(N+k)(m-1) + 2(p_{n+1}+1)}{(p_{n+1} - p_n)(N+k)},$$

$$\delta_n = \alpha_n \frac{p_{n+1}+1}{p_n+1}.$$

Thus, from (31) and (32) we obtain

$$\frac{d}{dt} E_{n+1}(t) \leq - \frac{[E_{n+1}(t)]^{\alpha_n}}{[E_n(t)]^{\beta_n} [\gamma(N, k, 1)]^{\delta_n}}.$$

Integrating this inequality and using (30), we get

$$E_{n+1}(t) \leq [\gamma(N, k, 1)]^{\frac{\delta_n}{\alpha_n - 1}} [\sigma p_{n+1}]^{\frac{1}{\alpha_n - 1}} A_n^{\frac{\beta_n}{\alpha_n - 1}} \mu_k(0)^{\sigma \frac{(N+k)(m-1)+2(p_{n+1}+1)}{N+k}} t^{-\sigma p_{n+1}}.$$

Thus, after easy manipulations, we obtain

$$E_{n+1}(t)$$

$$\leq [\max(1, \gamma(N, k, 1), \sigma) p_{n+1}]^{2(N+k+2)} \cdot \left[\prod_{i=1}^{n} (\theta_1 p_i)^{\frac{2\theta_2}{p_i+1}} \right]^{p_{n+1}+1}$$

$$\cdot \mu_k(0)^{\sigma \frac{(N+k)(m-1)+2(p_{n+1}+1)}{N+k}} \cdot t^{-\sigma p_{n+1}}.$$

Finally, we get the proof of (30), taking

$$\theta_1 \geq \max\left(1, \sigma, \gamma(N,k,1), \gamma\left(N,k,\frac{1}{m}\right)\right)$$

and

$$\theta_2 \geq \max\left(3(m+1)\sigma, \frac{N+k+2}{4}\right).$$

Letting $n \to \infty$ in

$$[E_n(t)]^{\frac{1}{p_n+1}} \leq \left[\prod_{i=1}^{n}(\theta_1 p_i)^{\frac{\theta_2}{p_i+1}}\right] \mu_k(0)^{\sigma\frac{(N+k)(m-1)+2(p_n+1)}{(N+k)(p_n+1)}} t^{-\sigma\frac{p_n}{p_n+1}},$$

we complete the proof of Lemma 5, choosing $\gamma = \exp\left(\sum_{i=1}^{+\infty}\frac{\theta_2}{p_i+1}\log\theta_1 p_i\right)$.

The following lemma, due to F. Bernis (see [7, Lemma 6.1]), allows us to prove Theorem 1.

LEMMA 6. *Let $f : \Re \to \Re$ be a function such that*

$$f(t) - f(s) \leq C(t-s)^a s^{-b} \quad \forall t > s \geq t_o > 0,$$

where $a > b > 0$.

Then $\forall T \geq t_o$,

$$f(T) - f(t_o) \leq C(1-2^{b-a})^{-1} T^{a-b}.$$

Now we are able to prove Theorem 1.

Proof of Theorem 1. From (13) and (27) we get

$$\zeta_1(T_2) - \zeta_1(T_1) \leq \gamma\mu_k(0)^{(\frac{2\sigma}{N+k}m+1)\lambda_o}(T_2-T_1)^{\beta_o} T_1^{-\lambda_o m\sigma}.$$

Thanks to the last inequality, we can apply Lemma 6.1 of [7], which gives

$$\zeta_1(T) \leq \gamma\mu_k(0)^{\frac{\sigma}{N+k}(m-1)} T^{\frac{\sigma}{N+k}}$$

for a sufficiently large $T > 0$. In the same way, we can obtain the previous estimate for $\zeta_j(T)$, $j = 2, ..., N$, and hence,

$$\zeta(T) \leq \gamma\mu_k(0)^{\frac{\sigma}{N+k}(m-1)} T^{\frac{\sigma}{N+k}}$$

for a sufficiently large $T > 0$. Now we are able to estimate $\zeta(T)$ from below. Using (29) and the last inequality, we get

$$\mu_k(0) = \mu_k(t) = \int_{\Re_k^N} X_k u(x,t)dx$$

$$\leq \gamma\zeta(t)^{N+k}\|u(.,t)\|_{\infty,\Re_k^N}$$

$$\leq \gamma\zeta(t)^{N+k}\mu_k(0)^{\frac{2\sigma}{N+k}} t^{-\sigma},$$

or

$$\zeta(t) \geq \gamma\mu_k(0)^{\frac{\sigma(m-1)}{N+k}} t^{\frac{\sigma}{N+k}},$$

for a sufficiently large $t > 0$. The proof of Theorem 1 is complete.

4. The higher order case for $N = 1$. In order to prove Theorem 2, we need the following lemma.

LEMMA 7 (Nirenberg–Gagliardo inequality). *Given $l \geq 2$, $s \geq 2$, $0 < q < 1$, the following inequality holds:*

$$\int_R^\infty x(x-R)^{s-2}(u^{(l-1)})^2 dx \tag{33}$$
$$\leq \gamma \left(\int_R^\infty x(x-R)^{s-2}(u^{(l)})^2 dx \right)^{\beta_1} \left(\int_R^\infty x|u|^{q+1} dx \right)^{\beta_2} \left(\int_R^\infty x(x-R)^s |u|^{q+1} dx \right)^{\beta_3},$$

where

$$\beta_1 = \frac{l-1}{l} + \frac{\theta_1}{l(l-(l-1)\theta_1)},$$

$$\beta_2 = \frac{2\theta_2(1-\theta_1)}{(q+1)(l-(l-1)\theta_1)},$$

$$\beta_3 = \frac{2(1-\theta_2)(1-\theta_1)}{(q+1)(l-(l-1)\theta_1)},$$

$$\theta_1 = \frac{1-q}{1-q+2(q+1)},$$

$$\theta_2 = \frac{s(1-q)+2(q+1)+q-1}{2s}.$$

Proof. First of all, we will prove the following inequalities:

$$\int_R^\infty x(x-R)^{s-2}(u^{(l-1)})^2 dx$$

$$\leq \gamma \left(\int_R^\infty x(x-R)^{s-2}(u^{(l)})^2 dx \right)^{\frac{l-1}{l}} \left(\int_R^\infty x(x-R)^{s-2}u^2 dx \right)^{\frac{1}{l}}, \tag{34}$$

$$\int_R^\infty x(x-R)^{s-2}u_x^2 dx$$

$$\leq \gamma \left(\int_R^\infty x(x-R)^{s-2}(u^{(l)})^2 dx \right)^{\frac{1}{l}} \left(\int_R^\infty x(x-R)^{s-2}u^2 dx \right)^{\frac{l-1}{l}}. \tag{35}$$

Let us prove (34). Using [7], we can write

$$\int_R^\infty x(x-R)^{s-2}(u^{(l-1)})^2 dx$$

$$= \int_R^\infty (x-R)^{s-1}(u^{(l-1)})^2 dx + R\int_R^\infty (x-R)^{s-2}(u^{(l-1)})^2 dx$$

$$\leq \gamma \left(\int_R^\infty (x-R)^{s-1}(u^{(l)})^2 dx\right)^{\frac{1}{l}} \left(\int_R^\infty (x-R)^{s-1}u^2 dx\right)^{\frac{l-1}{l}}$$

$$+\gamma R \left(\int_R^\infty (x-R)^{s-2}(u^{(l)})^2 dx\right)^{\frac{1}{l}} \left(\int_R^\infty (x-R)^{s-2}u^2 dx\right)^{\frac{l-1}{l}}$$

$$\leq \gamma \left(\int_R^\infty x(x-R)^{s-2}(u^{(l)})^2 dx\right)^{\frac{1}{l}} \left(\int_R^\infty x(x-R)^{s-2}u^2 dx\right)^{\frac{l-1}{l}}.$$

Inequality (35) may be proved in the same way. Now, from (19) with $k = 1$, it follows that

$$\int_R^\infty x(x-R)^{s-2}u^2 dx$$

$$\leq \gamma \left(\int_R^\infty x(x-R)^{s-2}u_x^2 dx\right)^{\theta_1} \left(\int_R^\infty x|u|^{q+1} dx\right)^{\frac{2(1-\theta_1)}{q+1}\theta_2},$$

$$\left(\int_R^\infty x(x-R)^s |u|^{q+1} dx\right)^{\frac{2(1-\theta_1)}{q+1}(1-\theta_2)}$$

$$\leq \gamma \left(\left(\int_R^\infty x(x-R)^{s-2}(u^{(l)})^2 dx\right)^{\frac{1}{l}} \left(\int_R^\infty x(x-R)^s u^2 dx\right)^{\frac{l-1}{l}}\right)^{\theta_1},$$

$$\left(\int_R^\infty x|u|^{q+1}\right)^{\frac{2(1-\theta_1)\theta_2}{q+1}} \left(\int_R^\infty x(x-R)^s |u|^{q+1} dx\right)^{\frac{2(1-\theta_1)(1-\theta_2)}{q+1}}.$$

From this inequality, we can estimate

$$\int_R^\infty x(x-R)^{s-2}u^2 dx,$$

and substituting this bound into (35) we get (33). Thus, the lemma is proved.

LEMMA 8. *Let u be an energy solution of problem* (3) *and $\rho(x) \in W_\infty^l(0,\infty)$, $\rho(x) \geq 0$. Then, for any $T_1, T_2 : 0 \leq T_1 < T_2$, the following equality is true:*

$$\frac{1}{r+1}\int_0^\infty \rho(x)|u(x,T_2)|^{r+1} dx - \frac{1}{r+1}\int_0^\infty \rho(x)|u(x,T_1)|^{r+1} dx$$

$$+\int_{T_1}^{T_2}\int_0^\infty u^{(l)}(\rho(x)u)^{(l)} dx dt = 0. \tag{36}$$

For the proof, see [7].

LEMMA 9. *Assume that u is an energy solution of problem* (3). *Then, for any $T_1, T_2 : 0 \le T_1 < T_2$, we have*

$$\zeta(T_2) - \zeta(T_1)$$

$$\le \gamma (T_2 - T_1)^{\frac{r+1}{2(l-1)(r+1)+4}} \left(\int_{T_1}^{T_2} \int_0^\infty x (u^{(l)})^2 dx dt \right)^{\frac{1-r}{2(l-1)(r+1)+4}}. \tag{37}$$

Proof. Substituting $T_1 = 0$, $T_2 = T$, $\rho(x) = x(x-R)_+^s$ with $R > R_o$ and s sufficiently large positive number, and taking into account that

$$(x(x-R)^s u)^{(l)} = x((x-R)^s u)^{(l)} + l((x-R)^s u)^{(l-1)}$$

and

$$\left| \int_0^T \int_R^\infty u^{(l)} ((x-R)^s u)^{(l-1)} dx \right|$$

$$= \left| \sum_{j=0}^{l-1} C_j \int_0^T \int_R^\infty (u^{(l-j-1)})^2 (x-R)^{s-2j-1} dx dt \right|$$

$$\le \gamma \sum_{j=0}^{l-1} \int_0^T \int_R^\infty x (x-R)^{s-2j-2} (u^{(l-j-1)})^2 dx dt$$

from (36), it follows that

$$\operatorname*{ess\,sup}_{0<t\le T} \int_R^\infty x(x-R)^s |u|^{r+1} dx + \int_0^T \int_R^\infty |u^{(l)}|^2 x (x-R)^s dx dt$$

$$\le \gamma \int_0^T \int_R^\infty x \sum_{j=0}^{l-1} (x-R)^{s-2j-2} (u^{(l-j-1)})^2 dx dt$$

$$+\gamma \int_0^T \int_R^\infty \left| x \sum_{j=0}^{l-1} (x-R)^{s-j-1} u^{(l-j-1)} u^{(l)} \right| dx dt. \tag{38}$$

Using Young's inequality, the second term on the right-hand side may be estimated by

$$\frac{\epsilon}{2} \int_0^T \int_R^\infty x (x-R)^s (u^{(l)})^2 dx dt$$

$$+\frac{1}{2\epsilon} \int_0^T \int_R^\infty x \sum_{j=0}^{l-1} (x-R)^{s-2j-2} (u^{(l-j-1)})^2 dx dt.$$

So, choosing a sufficiently small $\epsilon > 0$, from (38) we have

$$\operatorname*{ess\,sup}_{0<t\le T} \int_R^\infty x(x-R)^s |u|^{r+1} dx + \int_0^T \int_R^\infty (u^{(l)})^2 x(x-R)^s dxdt$$

$$\le \gamma \int_0^T \int_R^\infty x \sum_{j=0}^{l-1} (x-R)^{s-2j-2} (u^{(l-j-1)})^2 dxdt.$$

The use of Hardy's inequality yields

$$\mathcal{F}_s + \mathcal{I}_s \le \gamma \int_0^T \int_R^\infty x(x-R)^{s-2} (u^{(l-1)})^2 dxdt, \tag{39}$$

where

$$\mathcal{F}_s = \operatorname*{ess\,sup}_{0<t\le T} \int_R^\infty x(x-R)^s |u|^{r+1} dx,$$

$$\mathcal{I}_s = \int_0^T \int_R^\infty x(x-R)^s (u^{(l)})^2 dxdt.$$

Again applying Hardy's inequality, we obtain

$$\mathcal{F}_s \le \gamma \mathcal{I}_s. \tag{40}$$

Moreover, estimating the right-hand side of (39) by the Nirenberg–Gagliardo inequality (33) and using (40), we deduce

$$\mathcal{I}_s \le \gamma \int_0^T \mathcal{J}_{s-2}^{\beta_1} \mathcal{E}_o^{\beta_2} \mathcal{E}_s^{\beta_3} dt \le \gamma \mathcal{I}_{s-2}^{\beta_1} \mathcal{I}_o^{\beta_2} \mathcal{I}_s^{\beta_3} T^{1-\beta_1}, \tag{41}$$

where

$$\mathcal{J}_{s-2} = \int_R^\infty x(x-R)^{s-2} (u^{(l)})^2 dx,$$

$$\mathcal{E}_s = \int_R^\infty x(x-R)^{s-2} |u|^{r+1} dx,$$

$$\mathcal{E}_0 = \int_R^\infty x|u|^{r+1} dx.$$

From (41), it follows that

$$\mathcal{I}_1^{\beta_4} \le -\frac{d\mathcal{I}_1}{dR} T^{\beta_5},$$

with

$$\beta_4 = \frac{2(l-1)(r+1)+4}{2(l-1)(r+1)+4+(1-r)}, \qquad \beta_5 = \frac{r+1}{2(l-1)(r+1)+4+(1-r)}.$$

From the last inequality, we derive, in the usual way, estimate (37).

Proof of Theorem 2. Let us substitute in (36) $\rho(x) = x$. Since the support of u is bounded, and taking into account the identity

$$xu^{(l)} + lu^{(l-1)} = (xu)^{(l)},$$

we obtain

$$\int_{T_1}^{T_2} \int_0^\infty x(u^{(l)})^2 dxdt \le \gamma \int_0^\infty x|u(x,T_1)|^{r+1} dx. \tag{42}$$

Thus, using (42) in (37) easily follows (4).

In order to prove inequality (6), let us note that

$$\frac{1}{r+1}\frac{d}{dt}\int_0^\infty x|u(x,t)|^{r+1}dx = -\int_0^\infty x(u^{(l)})^2 dx. \tag{43}$$

Now, we remember that the following weighted Nirenberg–Gagliardo inequality holds (see [7]):

$$\left(\int_0^\infty x|u|^{r+1}dx\right)^{\frac{1}{r+1}} \le \gamma \left(\int_0^\infty x(u^{(l)})^2 dx\right)^{\frac{\theta}{2}} \left(\int_0^\infty x|u|^q dx\right)^{\frac{1-\theta}{r}},$$

where

$$\theta = \frac{2}{(r+1)(2+(l-1)r)}.$$

Thus, using this inequality and assumption (5) from (43), we derive

$$\frac{d}{dt}\int_0^\infty x|u|^{r+1}dx \le -\gamma\mu_o^{l+1}\left(\int_0^\infty x|u|^{r+1}dx\right)^{\chi},$$

where $\chi = 2 + (l-1)r$. The last inequality yields

$$\int_0^\infty x|u|^{r+1}dx \le \gamma\mu_o^{\frac{l-1}{1+(l-1)r}} t^{\frac{-1}{1+(l-1)r}}. \tag{44}$$

From (37), (42), and (44) we get

$$\zeta(T_2) - \zeta(T_1) \le \gamma(T_2 - T_1)^{\frac{r+1}{2(l-1)(r+1)+4}} T_1^{-\frac{1-r}{(1+(l-1)r)(2(l-1)(r+1)+4)}}.$$

Therefore, Lemma 1.6 of [7] gives estimate (6). The proof of Theorem 2 is complete.

Acknowledgment. The authors thank Professor F. Nicolosi for helpful suggestions.

REFERENCES

[1] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[2] N. D. Alikakos, *L^p bounds of solutions of reaction-diffusion equations*, Comm. Partial Differential Equations, 4 (1979), pp. 827–868.

[3] S. Bonafede, G. R.Cirmi, and A. F.Tedeev, *Finite Speed of Propagation for the Porous Media Equation with Lower Order Terms*, preprint.

[4] G. I. Barenblatt and V. B. Zel'dovich, *On dipole solutions in problems of nonstationary filtration of gas under polytropic regime*, Prikl. Mat. Mekh., 21 (1957), pp. 718–720.

[5] F. Bernis, *Finite speed of propagation and asymptotic rates for some higher order parabolic equations with absorption*, Proc. Roy. Soc. Edinburgh Sect. A, 104 (1986), pp. 1–19.

[6] F. Bernis, *Existence results for "doubly" nonlinear higher order parabolic equations on unbounded domains*, Math. Ann., 279 (1988), pp. 373–394.

[7] F. Bernis, *Qualitative properties for some nonlinear higher order degenerate parabolic equations*, Houston J. Math., 14 (1988), pp. 319–352.

[8] B. H. Gilding and L. A. Peletier, *On a class of similarity solutions of the porous media equation*, J. Math. Anal. Appl., 55 (1976), pp. 351–364.

[9] O. Grange and F. Mignot, *Sur la resolution d'une inequation paraboliques non lineaires*, J. Funct. Anal., 11 (1972), pp. 77–92.

[10] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities,* Cambridge Univ. Press, Cambridge, UK, 1952.

[11] J. Hulshof and J. L. Vazquez, *The dipole solution for the porous medium equation in several space dimensions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 193–217.

[12] Y. Jingxue, *Solutions with compact support for non linear diffusion equations*, Nonlinear Anal., 19 (1992), pp. 309–321.

[13] A. S. Kalashnikov, *Some problems of the qualitative theory of non-linear degenerate second-order parabolic equations*, Uspekhi Mat. Nauk, 42 (1987), pp. 135–176 (in Russian); translated in Russian Math. Surveys, 42 (1987), pp. 169–222.

[14] C. Miranda, *Istituzioni di analisi funzionale lineare,* UMI, ed. CNR, Gubbio, 1979.

[15] A. F. Tedeev, *Estimate of finite speed of stabilization of the solution of the initial-boundary problem for the porous media equation on unbounded domain*, Mat. Zametki, 3 (1995), pp. 473–476; translated in Math. Notes, 3 (1995), pp. 329–331.

[16] A. F. Tedeev, *Local and global properties of solutions of a Cauchy-Dirichlet problem for quasilinear second order parabolic equation in unbounded domain*, in Differentsial'nye Uravneniya, to appear.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1399–1418, November 1998

005

BIFURCATION TO SPIRAL WAVES IN REACTION-DIFFUSION SYSTEMS*

ARND SCHEEL†

Abstract. For a large class of reaction-diffusion systems on the plane, we show rigorously that m-armed spiral waves bifurcate from a homogeneous equilibrium when the latter undergoes a Hopf bifurcation. In particular, we construct a finite-dimensional manifold which contains the set of small rotating waves close to the homogeneous equilibrium. Examining the flow on this center-manifold in a very general example, we find different types of spiral waves, distinguished by their speed of rotation and their asymptotic shape at large distances of the tip. The relation to the special class of λ-ω systems and the validity of these systems as an approximation is discussed.

Key words. spiral waves, center-manifolds, Ginzburg–Landau equations, λ-ω systems

AMS subject classifications. 35B32, 58F39, 35K57, 35J60

PII. S0036141097318948

1. Introduction.

We study reaction-diffusion systems

$$U_t = D\triangle U + F(\lambda, U), \qquad U(x,t) \in \mathbb{R}^N, \quad \lambda \in \mathbb{R}^p, \tag{1.1}$$

on the plane $x \in \mathbb{R}^2$. The N-dimensional vector U typically describes a set of chemical concentrations and temperature, depending on time $t \in \mathbb{R}$ and the space variable x. The parameter λ is a p-dimensional control parameter which shall allow us to create instabilities of spatially homogeneous equilibria. We shall be interested in rotating wave solutions $U(t,x) = U(0, R_{ct}x)$, $c \neq 0$, where R_φ is the rotation in $\mathbb{R}^2$ around the origin by the angle $\varphi \in \mathbb{R}/2\pi\mathbb{Z}$. Our analysis shows that this class of solutions appears naturally via some type of Hopf bifurcation. Moreover the spatial structure resembles n-armed spiral waves.

Experimentally this type of spatiotemporal pattern has been observed frequently in chemical, biological, physiological, and physical experiments (e.g., the Belousov–Zhabotinsky reaction, the catalysis on platinum surfaces, electrochemical waves in the cortex of the brain, signaling patterns of the slime mold, and the Rayleigh–Bénard convection). Nevertheless a rigorous treatment of existence and creation is still not available—for various reasons.

Spiral waves are typically observed in spatially extended oscillatory processes. Near Hopf bifurcation points the dynamics of these processes are approximated by Ginzburg–Landau equations or λ-ω-systems [1, 10]. This has been shown using formal asymptotic methods; see [1] for example. Recently a rigorous proof of the approximation property of Ginzburg–Landau equations has been given by Schneider; see [21]. The important property of the approximating equations is a decoupling of Fourier modes, which was exploited by several authors in order to construct spiral wave solutions [1, 5, 6, 7, 11], although the methods are still formal or do not cover the typical nonlinearities appearing close to bifurcation points.

*Received by the editors March 24, 1997; accepted for publication (in revised form) September 15, 1997; published electronically August 3, 1998.
http://www.siam.org/journals/sima/29-6/31894.html
†Institut für Mathematik I, Freie Universität Berlin, Arnimallee 2-6, 14195 Berlin, Germany (scheel@math.fu-berlin.de).

Of course a treatment of Hopf bifurcation using classical bifurcation methods with symmetry is not possible because either center-manifolds do not exist or Lyapunov–Schmidt reduction cannot be applied, due to presence of continuous spectrum. Another explanation is provided by the fact that the symmetry of the reaction-diffusion equation, the Euclidean symmetry, does not have bounded finite-dimensional representations; see [12] for an approach to Ginzburg–Landau equations exploiting symmetry.

In spatially extended systems including only one unbounded spatial variable, typically cylindrical domains, the continuous spectrum can be avoided by restricting to steady state solutions and considering the unbounded spatial variable as a new time direction. This approach was introduced by Kirchgässner [9] and applied to various interesting problems in mechanics, fluid dynamics, and physics; see, for example, [13, 19]. Unfortunately, considering systems with several unbounded directions, this method becomes less successful; see, however, [14].

We adopt the idea of spatial dynamics, now considering the radial direction in polar coordinates as a new time variable, in order to describe the spatial structure of rotating waves by a surprisingly finite-dimensional, nonautonomous ordinary differential equation (ODE) on a center-manifold. In particular any small bounded solution to this ODE corresponds to a rotating wave of the original reaction-diffusion system. This approach allows us to describe systematically the creation of rotating waves from homogeneous equilibria of reaction-diffusion systems.

The paper is organized as follows. In the next section we fix the abstract functional analytic setting in which we formulate a center-manifold reduction theorem. This theorem, our main result, is stated in section 3 and proved in the two subsequent sections. The key to the proof are exponential dichotomies which are proved to exist in our functional analytic framework in section 4. In section 6 we comment on a localization of our main theorem using cutoff procedures and we briefly discuss regularity of solutions in section 7. The last important abstract result is stated in section 8, where we formulate and prove the existence of a larger manifold, containing solutions which might be singular at the origin. This manifold is tangent to a subspace which is independent of time τ (alias the distance to the origin in $\mathbb{R}^2$) and therefore allows us to derive explicit bifurcation equations. We conclude by applying our results to a model problem in the remaining sections.

2. The abstract setting. Introducing polar coordinates $x = (r\cos\varphi, r\sin\varphi)$, the equation for rotating wave solutions of (1.1) becomes

$$D\left(U'' + \frac{1}{r}U' + \frac{1}{r^2}U_{\varphi\varphi}\right) + F(\lambda, U) = cU_\varphi, \tag{2.1}$$

where $' = \frac{\partial}{\partial r}$, $\varphi \in S^1 = \mathbb{R}/2\pi\mathbb{Z}$, $r \in (0,\infty)$, and c is again the speed of rotation.

We suppose that for an initial parameter value λ_0 we are given a spatially homogeneous equilibrium $U_0(\lambda_0)$ which solves $F(\lambda_0, U_0(\lambda_0)) = 0$. As we are merely interested in Hopf bifurcation, we suppose that this solution can be continued in λ to a branch $U_0(\lambda)$, which we can assume without loss of generality to be the zero solution.

To sum up we suppose that $F(\lambda, 0) = 0$ and $D = D^* > 0$.

Note, however, that the first assumption is merely a simplification for clarity of the statements and the second assumption might be generalized slightly.

As a most elementary example, which is discussed in section 9, the reader might think of $D = \mathrm{id}$ and $F(\lambda, U) = iU + \mathrm{O}(|U|^2), U \in \mathbb{R}^2 \simeq \mathbb{C}$.

Next we multiply (2.1) by D^{-1},

$$U'' + \frac{1}{r}U' + \frac{1}{r^2}U_{\varphi\varphi} = -D^{-1}F(\lambda, U) + cD^{-1}U_\varphi, \tag{2.2}$$

and linearize around $U = 0$

$$U'' + \frac{1}{r}U' + \frac{1}{r^2}U_{\varphi\varphi} = -D^{-1}F_U(\lambda, 0)U + cD^{-1}U_\varphi. \tag{2.3}$$

We work in the function spaces $H^l(S^1, \mathbb{R}^N), l \geq 0$. Functions $u \in H^l$ may be represented by their Fourier series

$$u(\varphi) = \sum_{k\in\mathbb{Z}} u_k e^{ik\varphi},$$

with $u_{-k} = \bar{u}_k$ and $\|u\|^2_{H^l} = \sum_{k\in\mathbb{Z}} |u_k|^2 k^{2l} < \infty$.

The operator $A = -\partial_{\varphi\varphi}$ is self-adjoint and positive in H^l with spectrum $\{k^2;\ k \in \mathbb{Z}\}$ and domain of definition H^{l+2}. The spectral information on our bifurcation problem is contained in the operator

$$\begin{aligned} B_{\lambda,c} : H^{l+1} \subset H^l &\to H^l, \\ u(\cdot) &\mapsto -D^{-1}F_U(\lambda, 0)u(\cdot) + cD^{-1}u_\varphi(\cdot). \end{aligned}$$

As $c \neq 0$, the operator $B_{\lambda,c}$ can be considered as a bounded perturbation of the closed, antisymmetric, unbounded operator $C_c = cD^{-1}\partial_\varphi$ on H^l. Therefore $B_{\lambda,c}$ has point spectrum, and any strip $\{z \in \mathbb{C};\ |\operatorname{Im} z| < M\}$ contains only a finite number of eigenvalues, each of finite multiplicity.

We denote by $\tilde{P}^c_+(\lambda, c)$ the spectral projection to the operator $B_{\lambda,c}$ on $(-\infty, 0] \subset \mathbb{C}$. Of course $\tilde{P}^c_+(\lambda, c)$ can be constructed via Dunford's integral.

We rewrite (2.2) as an equation in function space,

$$u'' + \frac{1}{r}u' - \frac{1}{r^2}Au = -\tilde{F}(\lambda, u) + C_c u, \tag{2.4}$$

and linearize around $u = 0$,

$$u'' + \frac{1}{r}u' - \frac{1}{r^2}Au = B_{\lambda,c}u. \tag{2.5}$$

We suppose that $\tilde{F} \in C^K(\mathbb{R}^p \times H^{l+1/2}, H^l)$, $K \geq 1$, which can be achieved assuming $F \in C^K$ and either $l > 0$ such that $H^{l+1/2} \hookrightarrow C^0$ or suitable growth conditions on F.

We say that u is a solution of (2.4) (or (2.5)) on an interval $I \subset (0, \infty)$ if

$$u \in C^0(\operatorname{clos} I, H^{l+1}) \cap C^1(\operatorname{clos} I, H^l) \cap C^2(\operatorname{int} I, H^l) \cap C^1(\operatorname{int} I, H^{l+1}) \cap C^0(\operatorname{int} I, H^{l+2})$$

and if (2.4) (or (2.5), respectively) is satisfied in H^l.

Furthermore we introduce a new time

$$\tau(r) = \begin{cases} \log r & \text{if } r \leq \bar{r}, \\ r & \text{if } r \geq 2\bar{r}, \end{cases}$$

defined by smooth, monotone interpolation on $(\bar{r}, 2\bar{r})$ such that $\tau \in C^\infty(\mathbb{R}^+, \mathbb{R})$. The exact value of the positive constant $\bar{r}$ is of no importance for the statement of the results.

For $r < \bar r$ the differential equation (2.4) becomes

$$u_{\tau\tau} - Au = -e^{2\tau}(\tilde F(\lambda, u) - C_c u).$$

Of course (2.4) remains unchanged for $r \geq 2\bar r$.

We conclude this section emphasizing the symmetry of (2.4). The rotation $\varphi \to \varphi + \bar\varphi$ on $S^1 \simeq \mathcal{SO}(2)$ acts on $H^l(S^1, \mathbb{R}^N)$ by shifting functions on the circle: $u(\varphi) \to u(\varphi - \bar\varphi)$. Of course this symmetry is inherited from the Euclidean symmetry of the original problem. Note that translations are ruled out by the ansatz for equilibria in a fixed rotating coordinate system. Translations of any solution found with this ansatz would be periodic solutions in our coordinate system.

3. Main results. We write (2.4) as a first-order differential equation,

$$\frac{d}{d\tau}\underline{u}(\tau) = \mathcal{A}(\tau)\underline{u}(\tau) + \mathcal{G}(\underline{u}(\tau)), \qquad \underline{u} = \left(u, \frac{d}{d\tau}u(\tau)\right), \tag{3.1}$$

on the space $\underline{u} \in X = H^{l+1}(S^1, \mathbb{R}^N) \times H^l(S^1, \mathbb{R}^N)$. Here

$$\mathcal{A}(\tau) = \begin{pmatrix} 0 & 1 \\ A + e^{2\tau}B_{\lambda,c} & 0 \end{pmatrix}, \qquad \mathcal{G}(\underline{u}) = -e^{2\tau}\begin{pmatrix} 0 \\ \tilde F(\lambda, u) + B_{\lambda,c}u - C_c u \end{pmatrix}$$

if $\tau \leq \log \bar r$ and

$$\mathcal{A}(\tau) = \begin{pmatrix} 0 & 1 \\ \tau^{-2}A + B_{\lambda,c} & -\tau^{-1} \end{pmatrix}, \qquad \mathcal{G}(\underline{u}) = -\begin{pmatrix} 0 \\ \tilde F(\lambda, u) + B_{\lambda,c}u - C_c u \end{pmatrix}$$

if $\tau \geq 2\bar r$. For $\tau \in (\log \bar r, 2\bar r)$, the exact form of the equation is not important, as solutions arise from solutions of (2.4) by a bounded diffeomorphic rescaling of time. The linearization of (3.1) along $\underline{u} = 0$ is

$$\frac{d}{d\tau}\underline{u}(\tau) = \mathcal{A}(\tau)\underline{u}(\tau). \tag{3.2}$$

We let

$$Y_\delta = \{\underline{u} \in C^0(\mathbb{R}, X);\ \|u\|_{Y_\delta} < \infty\}, \quad \text{where } \|u\|_{Y_\delta} := \sup_{\tau\in\mathbb{R}} e^{-\delta|\tau|}|\underline{u}(\tau)|_{X_\tau},$$

and, similarly,

$$Y_\delta^\pm = \{\underline{u} \in C^0(\overline{\mathbb{R}}^\pm, X);\ \|u\|_{Y_\delta} < \infty\}, \quad \text{where } \|u\|_{Y_\delta^\pm} := \sup_{\tau\in\mathbb{R}^\pm} e^{-\delta|\tau|}|\underline{u}(\tau)|_{X_\tau}.$$

The norm in X_τ of $\underline{u}(\tau) = (u(\tau), v(\tau))$ is defined as

$$|\underline{u}(\tau)|_{X_\tau} := \begin{cases} \tau^{-1}|u|_{H^{l+1}} + |u|_{H^{l+1/2}} + |v|_{H^l} & \text{if } \tau \geq 2\bar r, \\ |u(\tau)|_{H^{l+1}} + |v(\tau)|_{H^l} & \text{if } \tau < 2\bar r. \end{cases}$$

Let us denote by $E^c(\tau)$ the (possibly empty) linear subspace of initial values of the linear equation (3.2) at time τ, which give rise to Y_δ-bounded solutions. Of course E^c depends also on δ.

THEOREM 1. *Suppose the superposition operator $\tilde F$ to the nonlinearity F belongs to the class $C^K(\mathbb{R}^p \times H^{l+1/2}(S^1, \mathbb{R}^N), H^l(S^1, \mathbb{R}^N))$, $1 \leq K < \infty$. Suppose, furthermore, that for $\lambda = \lambda_0$ and $c = c_0$ we have $\tilde P_+^c(\lambda, c) \neq 0$.*

Then there are $\epsilon, \delta > 0$ such that, if

$$\mathrm{Lip}_u[\tilde{F} - \tilde{F}_u(\lambda, 0)] + |\lambda - \lambda_0| + |c - c_0| < \epsilon,$$

there exists a unique finite-dimensional C^K-center-manifold $\mathcal{M} \subset X \times \mathbb{R}$ which contains all solutions of (3.1) *which are bounded in Y_δ. The manifold $\mathcal{M}$ is given as a graph over $\{E^c(\tau);\ \tau \in \mathbb{R}\}$ and depends smoothly on λ, c. In any section $\tau = \tau_0$, it is tangent to $E^c(\tau)$ at $\lambda = \lambda_0, c = c_0$.*

Moreover we have

(i) flow property: *for any $\underline{u}_0 = \underline{u}(\tau_0, c, \lambda) \in \mathcal{M}$, there is a unique Y_δ-bounded solution $\underline{u}(\tau), \tau \in \mathbb{R}$, to* (3.1) *with $\underline{u}(\tau_0) = \underline{u}_0$;*

(ii) invariance: *this unique solution $\underline{u}(\tau)$ lies on $\mathcal{M}$ for all times τ and depends C^K on $\underline{u}_0, \tau, \lambda$, and c;*

(iii) dimension: *the dimension of $E^c(\tau)$ is* $\dim R(\tilde{P}^c_+(\lambda_0, c_0)) + \dim \tilde{P}^c_+(\lambda_0, 0)\mathbb{R}^N$, *where the second summand is the dimension of the range when restricted on the homogeneous N-dimensional subspace of H^l;*

(iv) symmetry: *the manifold $\mathcal{M}$ is invariant and the flow on $\mathcal{M}$ is equivariant under the diagonal action of $\mathcal{SO}(2)$ on $X = H^{l+1} \times H^l$.*

Remarks.

(i) Let us emphasize that the operator $D\triangle + \partial_\varphi$ has continuous spectrum close to the imaginary axis, which makes a standard, finite-dimensional bifurcation approach to the dynamical reaction-diffusion problem impossible.

(ii) We will later give expansions for the spaces $E^c(\tau)$ at $\tau = \infty$ and describe how to obtain expansions for $\mathcal{M}$.

(iii) It is possible to treat the case of F depending on ∇u with the same methods. Indeed, both components of the gradient, u_r and $\frac{1}{r}u_\varphi$, are bounded with respect to $|(u, u_r)|_{X_\tau}$.

(iv) A slight generalization could be obtained by the use of interpolation spaces between X_τ and $D(\mathcal{A}(\tau))$. We avoided these additional technical difficulties for the sake of clarity.

(v) Making δ larger it is possible to allow singularities of the rotating waves at the origin, a phenomenon which is frequently attributed in the literature to spiral waves. The manifold will be larger if we allow for this type of solution but will still be finite-dimensional. However, the point in this work is that even spiral wave-like solutions without singularities at the tip are created via Hopf bifurcation.

4. The linearized equation. The key to a center-manifold theorem is the construction of exponential dichotomies for the linear equation. Background information on exponential dichotomies might be found in the textbooks [2], [8], in [15], or in a non-evolutionary, elliptic context in [16].

4.1. Bounded solutions for $\tau \to -\infty$. We construct a family of projections $P^{cu}_-(\tau)$ which project on the initial values of bounded solutions to (3.2) on $(-\infty, \tau]$. In a more general context this problem has been studied in [16]. The main theorems there (Theorem 1 and Theorem 3), applied to our setting, state the following lemma.

LEMMA 2. *Under the conditions of Theorem* 1, *suppose $\tau_0, \tau_1, \tau \le 2\bar{r}$. Then there are families of evolution operators, smoothly depending on λ and c,*

$$\Phi^u_-(\tau, \tau_0): \quad X \to X, \quad \tau \le \tau_0,$$
$$\Phi^s_-(\tau, \tau_0): \quad X \to X, \quad \tau \ge \tau_0,$$

and constants $C > 0$, $\eta^u_- > \eta^s_- > 0$ such that

(i) $\Phi_-^{u/s}(\cdot,\tau_0)\underline{u}$ *is a solution of* (3.2) *for any* $\underline{u}\in X$,
(ii) $\Phi_-^{u/s}(\cdot,\cdot)\underline{u}$ *is continuous in* X,
(iii) $\Phi_-^{u}(\tau_0,\tau_0)+\Phi_-^{s}(\tau_0,\tau_0)=\mathrm{id}$,
(iv) $\Phi_-^{u/s}(\tau,\tau_1)\Phi_-^{u/s}(\tau_1,\tau_0)=\Phi_-^{u/s}(\tau,\tau_0)$, $\Phi_-^{u/s}(\tau,\tau_1)\Phi_-^{s/u}(\tau_1,\tau_0)=0$, *and*
(v) $|\Phi_-^{u}(\tau,\tau_0)|_{L(X,X)}\le Ce^{-\eta_-^u(\tau-\tau_0)}$, $|\Phi_-^{s}(\tau,\tau_0)|_{L(X,X)}\le Ce^{-\eta_-^s(\tau-\tau_0)}$,
and we can choose any $\eta_-^u>0$. *We define* $P_-^{cu}(\tau):=\Phi_-^u(\tau,\tau)$.

We will later see how we can give a more explicit representation of the evolution operators Φ in terms of Bessel functions. This also will show why the uniqueness assumption from [16] is automatically satisfied in our context because the linear equation splits into an infinite product of ODEs, which are all uniquely solvable in forward and in backward time.

4.2. Bounded solutions for $\tau\to+\infty$. The situation at $\tau=+\infty$ is considerably more difficult as $B_{\lambda,c}$ is no more τ-uniformly bounded with respect to $\tau^{-2}A$. It is due to our careful choice of norms in X_τ that we still have an analogous result to Lemma 2.

LEMMA 3. *Under the conditions of Theorem* 1, *suppose* $\tau_0,\tau_1,\tau\ge 2\bar r$. *Then there are families of evolution operators, smoothly depending on* λ,c,

$$\begin{aligned}\Phi_+^u(\tau,\tau_0):&\quad X_{\tau_0}\to X_\tau,\quad \tau\le\tau_0,\\ \Phi_+^s(\tau,\tau_0):&\quad X_{\tau_0}\to X_\tau,\quad \tau\ge\tau_0,\end{aligned}$$

and constants $C>0$, $\eta_+^u>\eta_+^s>0$ *such that*
(i) $\Phi_+^{u/s}(\cdot,\tau_0)\underline{u}$ *is a solution of* (3.2) *for any* $\underline{u}\in X$,
(ii) $\Phi_+^{u/s}(\cdot,\cdot)\underline{u}$ *is continuous in* X,
(iii) $\Phi_+^{u}(\tau_0,\tau_0)+\Phi_+^{s}(\tau_0,\tau_0)=\mathrm{id}$,
(iv) $\Phi_+^{u/s}(\tau,\tau_1)\Phi_+^{u/s}(\tau_1,\tau_0)=\Phi_+^{u/s}(\tau,\tau_0)$, $\Phi_+^{u/s}(\tau,\tau_1)\Phi_+^{s/u}(\tau_1,\tau_0)=0$, *and*
(v) $|\Phi_+^{u}(\tau,\tau_0)|_{L(X_{\tau_0},X_\tau)}\le Ce^{\eta_+^u(\tau-\tau_0)}$, $|\Phi_+^{s}(\tau,\tau_0)|_{L(X_{\tau_0},X_\tau)}\le Ce^{\eta_+^s(\tau-\tau_0)}$,
and we can choose any $\eta_+^s>0$. *We define* $P_+^{cs}(\tau):=\Phi_+^s(\tau,\tau)$.

Proof.

Step 1. *Fourier ansatz.* The proof of this lemma is the central part of our analysis. Complexifying X, the subspaces

$$E^k=\{(ue^{ik\varphi},ve^{ik\varphi})\in X;\ \underline{u}=(u,v)\in(\mathbb{C}^N)^2\}\le X_\tau$$

are invariant under (3.2). Of course, we are primarily interested in the real subspace, where we have a relation between the vectors in E^k and E^{-k}. In E^k the differential equation reads

$$u''+\frac{1}{\tau}u'-\frac{k^2}{\tau^2}u=-D^{-1}(F_u(\lambda,0)+cik)u=:B_{\lambda,c}^k u.$$

If we expand $\underline{u}(\tau)=\sum_{k\in\mathbb{Z}}\underline{u}^k(\tau)e^{ik\varphi}$, then $|\underline{u}(\tau)|_{X_\tau}$ is equivalent to $(\sum_{k\in\mathbb{Z}}|\underline{u}^k|_{E_\tau^k}^2)^{1/2}$, where

$$|\underline{u}^k|_{E_\tau^k}=k^l\left(\frac{1}{\tau}|ku^k|_{\mathbb{C}^N}+|k^{1/2}u^k|_{\mathbb{C}^N}+|v|_{\mathbb{C}^N}\right)$$

if $k\ne0$ and $|\underline{u}^0|_{E_\tau^k}=|\underline{u}^0|_{(\mathbb{C}^N)^2}$.

By the above considerations we see that it is sufficient to construct the evolution operators on E_τ^k, and uniform exponential bounds on the norms in E_τ^k will carry over to X_τ.

Step 2. *Projections.* According to the remarks in section 2, we decompose E^k into $E^k_{c,+} = P^c_+ E^k$ and $E^k_{h,+} = (1 - P^c_+)E^k$, where $P^c_+ = \mathrm{diag}\,(\tilde{P}^c_+(\lambda_0, c_0), \tilde{P}^c_+(\lambda_0, c_0))$ and $\tilde{P}^c_+(\lambda_0, c_0)$ projects on the negative part of the spectrum of B_{λ_0,c_0}.

Step 3. *Stable projections, estimates.* We show that all solutions in $E^k_{c,+}$ are exponentially bounded in E^k_τ with an arbitrarily small exponent δ, keeping λ, c sufficiently close to λ_0, c_0.

As the range of $\tilde{P}^c_+(\lambda_0, c_0)$ is finite-dimensional, only finitely many modes k are involved in the computation. We therefore use the equivalent, standard, k- and τ-independent norm on $(\mathbb{C}^N)^2$. Furthermore decomposing $B^k_{\lambda,c}$ into Jordan blocks, it is sufficient to consider

$$u'' + \frac{1}{\tau}u' - \frac{k^2}{\tau^2}u + \Lambda(k, \lambda, c)u = 0,$$

where $\Lambda(k, \lambda_0, c_0)$ is a Jordan block. The eigenvalue of $\Lambda(k, \lambda_0, c_0)$ belongs to $\overline{\mathbb{R}}_+$ as $u \in R(\tilde{P}^c_+(\lambda_0, c_0))$. If we add $\alpha' = -\alpha^2$, then $\tau = 1/\alpha$ and we see that at $\lambda = \lambda_0$, $c = c_0$, the origin $u = 0, u' = 0$, and $\alpha = 0$ (alias $\tau = \infty$) is an equilibrium with all eigenvalues of the linearization being situated on the imaginary axis. Exponential growth with rate $\eta^s_+ > 0$ arbitrarily small now follows from standard Gronwall estimates for bounded α, that is, choosing $\bar{r}$ bounded away from zero, and λ, c sufficiently close to λ_0, c_0. This proves the second inequality in (v).

Step 4. *Unstable projections, estimates.* Now let $\underline{u} \in E^k_{h,+}$. Our aim is to decompose $E^k_{h,+}$ in subspaces of exponentially decaying and exponentially growing solutions. We set

$$\tilde{u}(\tau) = \left(\frac{k^2}{\tau^2} + B^k_{\lambda_0,c_0}\right)^{1/2} u(\tau).$$

As here $B^k_{\lambda,c}$ does not have eigenvalues on $\overline{\mathbb{R}}_-$, we can use the standard square root cut along $\mathbb{R}_-$. Moreover the norm $|\underline{u}|_{E^k_\tau}$ is equivalent to $|\tilde{u}|_{\mathbb{C}^N} + |v|_{\mathbb{C}^N}$. Note that here we omitted the factor k^l, as it is independent of time and does not change the equations to be considered below. We write $\alpha = 1/\tau$ and $L(\alpha) = (k^2\alpha^2 + B^k_{\lambda_0,c_0})^{1/2}$. In the new variables, the differential equation on $E^k_{h,+}$ reads

$$\begin{aligned} \tilde{u}' &= L(\alpha)v + \partial_\alpha L(\alpha)\alpha' u \\ &= L(\alpha)v - \alpha^3 k^2 L^{-2}(\alpha)\tilde{u}, \\ v' &= -\alpha v + L(\alpha)\tilde{u}, \\ \alpha' &= -\alpha^2. \end{aligned} \tag{4.1}$$

Next, we set $|L^{-1}(\alpha)|\frac{d}{d\tau} = \frac{d}{ds}$ and obtain

$$\begin{aligned} \frac{d\tilde{u}}{ds} &= L|L^{-1}|v - \alpha^3 k^2 L^{-2}|L^{-1}|\tilde{u}, \\ \frac{dv}{ds} &= -\alpha|L^{-1}|v + L|L^{-1}|\tilde{u}, \\ \frac{d\alpha}{ds} &= -\alpha^2|L^{-1}|, \end{aligned} \tag{4.2}$$

with $L = L(\alpha)$. The linearization at $\tilde u = v = 0$, $\alpha = 0$,

$$\begin{aligned} \frac{d\tilde u}{ds} &= L|L^{-1}|v, \\ \frac{dv}{ds} &= L|L^{-1}|\tilde u, \\ \frac{d\alpha}{ds} &= 0, \end{aligned} \tag{4.3}$$

admits a projection $P(\tilde u, v) = \frac{1}{2}(\tilde u + v, \tilde u + v)$, which is independent of k and α. Therefore the flow $\tilde\Phi_0$ of (4.3) possesses uniform exponential dichotomies at $\tilde u = v = 0$. To see this, we first observe that for $s \le s_0$,

$$|\tilde\Phi_0(s, s_0)P|_{L(\mathbb{C}^{2N})} \le |e^{L|L^{-1}|(s-s_0)}|_{L(\mathbb{C}^{2N})}.$$

Now remember that by definition of the square root, the spectrum of L lies in the right half plane and is, for $|k| \to \infty, \alpha = 0$, asymptotic to $k^{1/2}e^{\pm i\pi/4}$. For finitely many k, we therefore obtain

$$\left|e^{-L|L^{-1}|t}\right|_{L(\mathbb{C}^{2N})} \le C_1 e^{-\eta_1 t}, \quad t > 0,$$

with some constants $C_1, \eta_1 > 0$, independent of k, α. As $k \to \infty$, we consider first $\tilde L = k^{-1/2}L$. Of course $\tilde L|\tilde L^{-1}| = L|L^{-1}|$. For k large, the operator $\tilde L_0 = (k\alpha^2 + D^{-1}ci)^{1/2}$ is a small (uniformly in α, k) perturbation of $\tilde L$. As $D > 0$, the spectrum of $D^{-1}i$ lies on $i\mathbb{R}^+$. Therefore the spectrum of $\tilde L_0$ lies in the right half plane, uniformly bounded away from the imaginary axis, and we can diagonalize $\tilde L_0$ by a transformation which is independent of α and k to obtain

$$\left|e^{-\tilde L_0|\tilde L_0^{-1}|t}\right|_{L(\mathbb{C}^{2N})} \le C_2 e^{-\eta_2 t}, \quad t > 0,$$

for some constants $C_2, \eta_2 > 0$, independent of α, k. By perturbation arguments, the same estimate holds true for $\tilde L$ and L, and we conclude

$$|\tilde\Phi_0(s; s_0)P|_{L(\mathbb{C}^{2N})} \le Ce^{\eta(s-s_0)}, \quad s \le s_0,$$

for some $C, \eta > 0$, independent of α and k. The calculation on $R(1-P)$ is the same and we obtain

$$|\tilde\Phi_0(s; s_0)(1-P)|_{L(\mathbb{C}^{2N})} \le \left|e^{-L|L^{-1}|(s-s_0)}\right|_{L(\mathbb{C}^{2N})} \le Ce^{-\eta(s-s_0)}, \quad s \ge s_0.$$

These two estimates together guarantee an exponential dichotomy for (4.3). Equation (4.2) is a perturbation of (4.3). We show that the perturbation of the vector field is $\mathrm{O}(\alpha)$, uniformly in k. By standard perturbation results on exponential dichotomies [2, 8], this then proves that (4.2) possesses an exponential dichotomy with projection $\tilde P(k, \alpha)$ and constants $\tilde C, \tilde\eta > 0$, independent of k, α as long as α is bounded.

The error terms we have to deal with are $\alpha^3k^2L^{-2}|L^{-1}|$ and $\alpha|L^{-1}|$. Of course for finite k these terms are $\mathrm{O}(\alpha)$. Consider now the first expression for large k:

$$\begin{aligned} \alpha^3k^2|L^{-3}| &= \alpha^3k^2|[\alpha^2k^2 + B^k_{\lambda_0,c_0}]^{-3/2}| \\ &\le \alpha^3k^2|[\alpha^2k^2 + D^{-1}cik + \mathrm{O}(1)]^{-1}| \cdot |D^{-1}cik + \mathrm{O}(1)|^{-1/2} \\ &= \left|\left[1 + D^{-1}ci\frac{1}{\alpha^2k}(1 + \mathrm{O}(1/k))\right]^{-1}\right| \cdot \mathrm{O}(\alpha k^{-1/2}). \end{aligned}$$

As $|[1 + D^{-1}ci\frac{1}{\alpha^2 k}]^{-1}| \le C_3$ uniformly in α, k, the above expression is $\mathrm{O}(\alpha k^{-1/2})$, uniformly in k. Next we consider the second error term $\alpha|L^{-1}|$:

$$\begin{aligned}\alpha|L^{-1}| &= \alpha|[\alpha^2k^2 + D^{-1}cik + \mathrm{O}(1)]^{-1/2}| \\ &= \alpha k^{-1/2}|\tilde{L}_0^{-1}(1 + \mathrm{O}(1/k))| \\ &= \mathrm{O}(\alpha k^{-1/2}).\end{aligned}$$

This proves uniform smallness of the perturbation. It remains to translate the exponential dichotomy rate $\tilde{\eta}$ into the correct time $\tau = \tau(s)$.

As $\frac{ds}{d\tau} = |L^{-1}(\alpha)|^{-1}$, it is sufficient to get α, k-uniform bounds $|L^{-1}(\alpha)|^{-1} \ge \eta_0 > 0$. This is precisely the type of estimate we developed above for $\alpha|L^{-1}|$. Indeed we showed that

$$|L^{-1}| = k^{-1/2}|\tilde{L}_0^{-1}|(1 + \mathrm{O}(1/k));$$

therefore η_0 can be chosen $\mathrm{O}(k^{1/2})$ as $k \to \infty$. This proves the lemma with $\eta_+^u = \eta_0\eta$ and η_+^s from step 3. □

4.3. Matching at $\bar{\tau} = 2\bar{r}$, the center space $E^c(\tau)$. We define $E^c(\tau) = \Phi_+^s(\tau,\tau)\Phi_-^u(\tau,\tau)X$, which, by the previous two lemmata, coincides with the definition of $E^c(\tau)$ as the initial values for Y_δ-bounded solutions if we only choose δ small enough. In order to prove the claim on the dimension of $E^c(\tau)$, we need a transversality result from the theory of Bessel functions. Suppose first that $\underline{u}(\bar{\tau}) \in E^0$, the subspace of radially homogeneous functions. For $\tau \to -\infty$ the linear equation in E^0 is $u_{\tau\tau} = e^{2\tau}\Lambda u$ with some matrix Λ, and clearly any solution is Y_δ-bounded as exponential rates of solutions coincide with the rates of the autonomous part $u_{\tau\tau} = 0$. So the negative orbit of $\underline{u}(\bar{\tau})$ is Y_δ-bounded. The positive orbit is bounded in Y_δ if and only if $\underline{u}(\bar{\tau}) \in P_+^c(\lambda_0, 0)E^0$; therefore $\dim E^c \cap E^0 = 2\dim \tilde{P}_+^c(\lambda_0,0)\mathbb{R}^N$. For the rest of this section we restrict to $(E^0)^\perp$, the nonhomogeneous Fourier modes. Recall that $P_+^c = \mathrm{diag}\,(\tilde{P}_+^c(\lambda_0,c_0), \tilde{P}_+^c(\lambda_0,c_0))$ and $P_+^h = 1 - P_+^c$ projects on the hyperbolic part of (3.2) at $r = \infty$. We claim that

$$\Phi_+^s(\bar{\tau},\bar{\tau})\Phi_-^u(\bar{\tau},\bar{\tau})P_+^h\underline{u} = 0 \tag{4.4}$$

for $\underline{u} \in (E^0)^\perp$. We decompose into Fourier modes $e^{ik\varphi}$ and minimal Jordan blocks Λ, and we consider

$$u'' + \frac{1}{\tau}u' - \frac{k^2}{\tau^2}u + \Lambda(k,\lambda_0,c_0)u = 0.$$

If Λ is semisimple, that is, $\Lambda \in \mathbb{C} \setminus \overline{\mathbb{R}}_+$, then the solutions of this scalar ODE are the Bessel functions. Indeed, we can write this equation as

$$r^2u'' + ru' + (-k^2 + (r\sqrt{\Lambda})^2)u = 0;$$

therefore

$$u(r) = u_0J_k(r\sqrt{\Lambda}) + u_1Y_k(r\sqrt{\Lambda}).$$

As $J_k(r) = r^k(1 + \mathrm{O}(r))$ and $Y_k(r) = r^{-k}(1 + \mathrm{O}(r))$ for $r \to 0$, if $k \ne 0$, solutions bounded close to $r = 0$ satisfy $u_1 = 0$. At infinity the J_k behave like

$$J_k(r) = \sqrt{\frac{2}{\pi r}}\left[\cos\left(r - \frac{k\pi}{2} - \frac{\pi}{4}\right) + \mathrm{O}(1/r)\right].$$

Solutions $u(r) = u_0 J_k(r\sqrt{\Lambda})$ can stay exponentially bounded by $e^{\delta r}$ as $r \to \infty$, for a small fixed δ, only if $\sqrt{\Lambda}$ is real. But then Λ is real and positive, that is, $u \in \tilde{P}^c_+ X$. This proves the required transversality result (4.4) for semisimple eigenvalues.

If Λ is a Jordan block we can rescale the principal vectors—without changing the angle between stable and unstable subspaces—to make it a small perturbation of its semisimple part. The transverse intersection persists for the nonsemisimple Jordan block.

Now suppose $u \in \tilde{P}^c_+ X$. Then the above reasoning showed that for any such u there is exactly one Y_δ-bounded solution. This implies $\dim (E^c(\tau) \cap (E^0)^\perp) = \dim R(\tilde{P}^c_+(\lambda_0, c_0)(E^0)^\perp)$ and proves the claim (iii) in Theorem 1 on the dimension of the invariant manifold, once it is constructed as a graph over $\{E^c(\tau);\ \tau \in \mathbb{R}\}$.

5. Nonlinear equations, proof of Theorem 1. With the estimates on the linearized equation at hand, it is fairly standard to construct invariant manifolds for the nonlinear equation. We consider (3.1).

PROPOSITION 4. *Under the conditions of Theorem* 1*, any Y^-_δ-bounded (or Y^+_δ-bounded) solution $\underline{u}(\tau, \tau_0)$ on $(-\infty, \tau_0]$ (or $[\tau_0, +\infty)$, respectively) satisfies*

$$\underline{u}(\tau, \tau_0) = \Phi^u_-(\tau, \tau_0)\underline{u}(\tau_0, \tau_0) + \int_{\tau_0}^{\tau} \Phi^u_-(\tau, \sigma)\mathcal{G}(\underline{u}(\sigma, \tau_0))d\sigma + \int_{-\infty}^{\tau} \Phi^s_-(\tau, \sigma)\mathcal{G}(\underline{u}(\sigma, \tau_0))d\sigma,$$

or

$$\underline{u}(\tau, \tau_0) = \Phi^s_+(\tau, \tau_0)\underline{u}(\tau_0, \tau_0) + \int_{\tau_0}^{\tau} \Phi^s_+(\tau, \sigma)\mathcal{G}(\underline{u}(\sigma, \tau_0))d\sigma + \int_{+\infty}^{\tau} \Phi^u_+(\tau, \sigma)\mathcal{G}(\underline{u}(\sigma, \tau_0))d\sigma,$$

respectively. On the other hand, the above integral equations possess for any $\underline{u}(\tau_0, \tau_0)$ a unique solution $\underline{u}(\tau, \tau_0)$ in $Y^\pm_\delta$ which depends C^K on $\underline{u}(\tau_0, \tau_0), \lambda, c, \tau$ and τ_0.

Proof. The integral operators are bounded operators on $Y^\pm_\delta$ and the Lipschitz constant of the nonlinearity $\mathcal{G}$ is small. Indeed

$$\mathrm{Lip}_{X_\tau} \mathcal{G} \le \mathrm{Lip}_{H^{l+1/2} \to H^l}[\tilde{F} - \tilde{F}_u],$$

which was supposed to be sufficiently small. Regularity of the unique fixed point can be proved as usually for center-manifolds; see [24] for example. □

We call the set

$$\left\{ \Phi^u_-(\tau, \tau)\underline{u} + \int_{-\infty}^{\tau} \Phi^s_-(\tau, \sigma)\mathcal{G}(\underline{u}(\sigma, \tau))d\sigma =: \Psi_-(\Phi^u_-(\tau, \tau)\underline{u});\ \underline{u} \in X \right\}$$

the center-unstable manifold $\mathcal{M}^{cu}_-(\tau)$ at $-\infty$ and the set

$$\left\{ \Phi^s_+(\tau, \tau)\underline{u} + \int_{\infty}^{\tau} \Phi^u_+(\tau, \sigma)\mathcal{G}(\underline{u}(\sigma, \tau))d\sigma =: \Psi_+(\Phi^s_+(\tau, \tau)\underline{u});\ \underline{u} \in X \right\}$$

the center-stable manifold $\mathcal{M}^{cs}_+(\tau)$ at $+\infty$, and we define

$$\mathcal{M}(\tau) = \mathcal{M}^{cu}_-(\tau) \cap \mathcal{M}^{cs}_+(\tau).$$

By definition, $\mathcal{M}(\tau) = \{\text{initial values at time } \tau \text{ of } Y_\delta\text{-bounded solutions}\}$. We have to show that $\mathcal{M}(\tau)$ is a smooth manifold, parameterized over $E^c(\tau)$.

Therefore, we have to solve $\Psi_+ - \Psi_- = 0$. The linearization is given by $\Phi^u_- - \Phi^s_+ = 0$. We already know that the kernel of this equation is exactly E^c, thus finite-dimensional. In order to apply the implicit function theorem we have to show that $\Phi^u_- - \Phi^s_+$ is surjective. We have to decompose $\underline{u} \in E^k$ into two vectors belonging to the range of Φ^s_+ and Φ^u_-, respectively, with estimates on the norms uniform with respect to k. The fact that we can decompose follows simply from the linear independence of the Bessel functions of the first and second kind J_k and Y_k (actually, we merely refer to purely imaginary arguments, the hyperbolic case, where the notation is I_k for the Bessel function bounded at $r = 0$ and K_k for the solution bounded at $r = \infty$). Estimates on the norms—for a fixed time τ—follow from uniform estimates on the Wronski determinant

$$\det \begin{pmatrix} I_k(\tau) & K_k(\tau) \\ I_k'(\tau) & K_k'(\tau) \end{pmatrix},$$

which in turn are an immediate consequence of the Taylor expansions at $r = 0$ of the Bessel functions; see for example [25]. As in section 4.3, Jordan blocks can be considered as a small perturbation. By Lyapunov–Schmidt reduction we can now solve $\Psi_+ - \Psi_- = 0$, parameterizing the set of solutions over the kernel of the linearization $E^c(\tau)$. This proves Theorem 1.

6. Local center-manifolds. If the nonlinearity $\tilde{F}$ does not have a small Lipschitz constant, which is usually the case in applications, we have to modify $\tilde{F}$.

We cut off $\tilde{F}$ outside a small neighborhood B_{ϵ_0} of zero with a smooth cutoff function in $H^{l+1/2}$, for example, the norm, which is invariant under the action of $\mathcal{SO}(2)$. Therefore let $\chi \in C^\infty([0,\infty), \mathbb{R})$ with $\chi(t) = 1$ if $t \le 1$ and $\chi(t) = 0$ if $t \ge 2$. Then define

$$\tilde{F}_{mod}(\lambda, u) = \chi(|u|^2_{H^{l+1/2}}/\epsilon_0)(\tilde{F}(\lambda, u) - \tilde{F}_u(\lambda, 0)u) + \tilde{F}_u(\lambda, 0)u.$$

The nonlinear part of $\tilde{F}_{mod}$ has an arbitrarily small Lipschitz constant if we make ϵ_0 sufficiently small, and thereby satisfies the conditions of Theorem 1. Any solution on the center-manifold to the modified nonlinearity $\tilde{F}_{mod}$, which has norm $\sup_\tau |\underline{u}(\tau)|_{X_\tau}$ small enough, will have $\sup_\tau |u(\tau)|_{H^{l+1/2}}$ small such that the modified nonlinearity coincides with the original nonlinearity on the solution $u(\tau)$, which is in consequence a solution to the original equation. Note that bounds on the norm in X_τ are by construction of $\mathcal{M}$ equivalent to bounds on the norms of the projection of the solution on $\{E^c(\tau);\ \tau \in \mathbb{R}\}$.

7. Regularity of solutions. The solutions $u(r, \varphi)$ we obtain are bounded in X_τ. By the smoothing property of the equation (which can be considered for any l, without changing $\mathcal{M}$), any solution is actually of class C^∞ with respect to $r > 0$ and φ, if F is—although $\mathcal{M}$ is not C^∞ in general! As $r \to \infty$, the angular derivatives $\partial^m_\varphi u(r, \varphi)$ are bounded for any m, which implies that the derivatives along curves $r \equiv$ const with respect to arclength $r d\varphi$ are of order $1/r^m$: patterns are slowly varying in the angular direction far away from the origin.

At $r = 0$ we have to be careful about smoothness of the solution. Suppose first that $E^c(\tau)$ does not contain solutions in the angular homogeneous subspace E^0. Then solutions in $E^c(\tau)$ are $\mathrm{O}(r) = \mathrm{O}(e^\tau)$ as $r \to 0$ and smooth in a neighborhood of the origin by interior elliptic regularity.

The homogeneous subspace can be—and has been—treated separately studying the ODE on $\mathrm{Fix}(\mathcal{SO}(2))$. Indeed there is a subspace of dimension N with solutions which actually stay bounded, whereas solutions outside this subspace have singularities of order $\log r$.

On the other hand, considering again τ-dynamics in $\mathcal{M}$, this subspace of homogeneous functions is fibered by strongly unstable fibers such that any solution in $\mathcal{M}$ converges with rate $\mathrm{O}(e^\tau)$ to a solution in the homogeneous subspace and inherits its regularity.

8. Center-manifolds at infinity. We construct a finite-dimensional invariant manifold which contains all solutions which are bounded at $\tau = +\infty$ but do not decay too rapidly. Recall that $P^c_+ = \mathrm{diag}\,(\tilde P^c_+, \tilde P^c_+)$ projects on the center part of (3.2) at $\alpha = 1/r = 1/\tau = 0$.

PROPOSITION 5. *Under the conditions of Theorem* 1*, consider equation* (3.1) *close to* $\underline{u} = 0$*. Fix* $\delta > 0$ *sufficiently small and* $K < \infty$*.*

Then there is an invariant C^K*-manifold* $\mathcal{M}^c_+$*, contained in* $\mathcal{M}^{cs}_+$ *and containing* $\mathcal{M}$*, given as a graph over* $\{R(P^c_+);\ \tau \in \mathbb{R}\}$*, smoothly depending on* λ, c*.*

Moreover there is a C^K*-flow on* $\mathcal{M}^c_+$ *such that any orbit is a solution of* (3.1) *and any solution* $\underline{u}(\tau)$ *of* (3.1) *with*

$$\sup_{\tau_0 \geq \tau > 2\bar r} e^{-\delta|\tau-\tau_0|} \frac{|\underline{u}(\tau)|_{X_\tau}}{|\underline{u}(\tau_0)|_{X_{\tau_0}}} < \infty$$

is contained in $\mathcal{M}^c_+$*.*

Proof. We start by constructing $\mathcal{M}^c_+$ for $0 < \alpha = 1/r \leq 1/2\bar r$ bounded. The manifold $\mathcal{M}^c_+$ is the union of center-unstable fibers of the zero-solution in the center-stable manifold $\mathcal{M}^{cs}_+$. These fibers can easily be shown to exist, using graph transformation (we have a smooth semiflow on $\mathcal{M}^{cs}_+$) or a Lyapunov–Perron approach as in [16]. The dependence on time $\alpha = 1/\tau$ is smooth as fibers are mapped into each other by the flow.

We have to ensure that we can arrange to have $\mathcal{M}$ included in $\mathcal{M}^c_+$. This can be achieved by either starting the graph transformation with graphs that contain $\mathcal{M}$ (and "feeding in" such graphs appropriately) or, referring to the Lyapunov–Perron approach of [16], including the manifold $\mathcal{M}$ in the fixed initial unstable fiber at $\tau = 2\bar r$ (see, for example, [16, at the end of section 3]).

We next have to continue this manifold for $\alpha > 1/2\bar r$ or, equivalently, for $\tau \to -\infty$. This will again be done using the methods from [16]. If we had an evolution-type equation we would propagate the manifold $\mathcal{M}^c_+$ with the flow. Here we do not have a flow! By [16], the equation possesses an exponential dichotomy which permits us to prove the existence of the center-unstable manifold $\mathcal{M}^{cu}_-$ (the union of unstable fibers over time τ), as pointed out in Lemma 2, and, furthermore, the existence of stable fibrations to any solution in $\mathcal{M}^{cu}_-$ for any fixed initial fiber at $\tau = 2\bar r$ (which is transversely intersecting $\mathcal{M}^{cu}_- \cap \{\tau = 2\bar r\}$). We are interested in the stable fibration induced by the manifold $\mathcal{M}^c_+$, which is of course not complementary to $\mathcal{M}^{cu}_-$. However, the methods from [16] can be adapted in order to guarantee precisely the existence of such a manifold. In the following we indicate how to make the necessary changes.

We solve the integral equation for stable and unstable fibrations with the restriction that the fiber at the initial time $\tau = 2\bar r$ belongs to a fixed manifold transverse to $\mathcal{M}^{cu}_-$ which we can choose to contain $\mathcal{M}^c_+ \cap \{\tau = 2\bar r\}$. On this smaller subspace the fixed-point equation for stable and unstable fibers still defines a contraction mapping and the solution is the desired global continuation of $\mathcal{M}^c_+$. The smoothness of the

union of the fibers as a manifold follows, because we can differentiate the fixed-point equation with respect to the base solution in the center-unstable manifold $\mathcal{M}^{cu}_{-}$. The exponential properties of the new fixed-point equation allow for a setting in the usual scale of exponentially weighted spaces [23], because the equation for the stable fiber at a fixed time τ involves only the finite time interval $[\tau, 1/2\bar{r}]$. We do not carry out the details, which include only straightforward modifications of smoothness proofs for fibrations (note, however, that we do not care about the limit $\tau = -\infty$—alias $r = 0$—of the fibration, which would lead to limitation in regularity of the fibration).

Of course the projected vector field is also smooth and thereby defines a smooth flow on the finite-dimensional manifold $\mathcal{M}^{c}_{+}$. □

The hypothesis $\tilde{F}(\lambda, 0) = 0$ was only needed in order to fix a reference solution in $\mathcal{M}^{cs}_{+}$, notably the zero solution. In general we could construct smooth fibrations along any solution in $\mathcal{M}^{cs}_{+}$.

The manifold $\mathcal{M}^{c}_{+}$ we constructed is very useful in order to describe bounded solutions near infinity, although most solutions on $\mathcal{M}^{c}_{+}$ are not bounded at the origin $r = 0$.

9. Hopf bifurcation and (λ, ω)-systems. We give the most simple nontrivial application of our main theorem. Suppose $D = \mathrm{id}$, $\tilde{F}(\lambda, 0) = 0$, $\lambda \in \mathbb{R}$, and $N = 2$; that is, $U \in \mathbb{R}^2$, which we identify with $\mathbb{C}$. Suppose that the homogeneous zero state undergoes a nondegenerate Hopf bifurcation in the space of homogeneous solutions:

$$\frac{d}{dU}\tilde{F}(\lambda, 0) = i\omega + \lambda, \quad \omega \neq 0.$$

We write U as a complex Fourier series $U(r, \varphi) = \sum_{k \in \mathbb{Z}} U^k(r) e^{ik\varphi}$. The spaces E^k are just the complex two-dimensional spans $\langle (e^{ik\varphi}, 0), (0, e^{ik\varphi}) \rangle$. The operator $B_{\lambda,c}$ acts on E^k as multiplication $B^k(\lambda, c) : U^k \to (cik - i\omega - \lambda)U^k$. Thereby $E^c(r) \leq E^{k_0}$ if $c_0 k_0 = \omega$. In other words, for any k(-armed spiral) there is a rotation speed $c = \omega/k$ such that rotating waves with this speed may bifurcate. Our analysis has shown that for other wave speeds, the homogeneous state is isolated as a rotating wave.

Let us comment on the symmetry. The flow on $\mathcal{M}$ projected on $E^c(r) \leq E^k$ is equivariant with respect to the action of $\mathcal{SO}(2)$:

$$(U, U') \to (Ue^{i\psi}, U'e^{i\psi}), \qquad \psi \in \mathbb{R}/2\pi\mathbb{Z} \simeq \mathcal{SO}(2).$$

This is exactly the same symmetry that authors usually *assumed* to be present in bifurcation equations, the so-called (λ, ω)-systems, modeling the creation of spiral waves; see [1]. We showed *rigorously* that the symmetry of (λ, ω)-systems, without any error term, is present in this type of bifurcation.

The actual solutions $U = U^k(r)e^{ik\varphi}$ of the linearized system in $E^c(\tau)$ are easily calculated: they solve $(U^k)'' + \frac{1}{r}(U^k)' = (k^2/r^2)U^k$ and are given as $U(r, \varphi) = U^c r^k e^{ik\varphi}, U^c \in \mathbb{C}$. Note that the invariant complement in E^k, spanned by $\tilde{U}(r, \varphi) = \tilde{U}^c r^{-k} e^{ik\varphi}, \tilde{U}^c \in \mathbb{C}$, converges as $E^c(r)$ to the same limit $\{(U, U');\ U' = 0\}$ as $\varphi \to \infty$. This is the reason why we constructed $\mathcal{M}^{c}_{+}$ tangent to E^k in section 8. The equation on $\mathcal{M}^{c}_{+}$ is a nonautonomous, $\mathcal{SO}(2)$-equivariant ODE in $\mathbb{C}^2$ with the linear part given by Bessel's differential equation. It can be smoothly extended to time $\tau = \infty (\alpha = 0)$, where the equation becomes autonomous. In order to determine existence and shape of rotating waves at $r = \infty$, we have to calculate expansions of the vector field on $\mathcal{M}^{c}_{+}$ and determine the ω-limit set of the two-dimensional slice $\mathcal{M}(\tau)$ in $\mathcal{M}^{c}_{+}$. We examine a simple model problem in the next sections.

10. An example. As an example we study the following reaction-diffusion system:

$$\begin{aligned} u_t &= d_1 \triangle u + \kappa u - v - a u^3, \\ v_t &= d_2 \triangle v + b u - \gamma v \end{aligned}$$

in the plane $x \in \mathbb{R}^2$. When $\kappa = \gamma$ and $b - \gamma\kappa > 0$, the pure reaction system undergoes a Hopf bifurcation in the origin $u = v = 0$. Rescaling u, v, t, and x, we may assume the system to be in the particular form

$$\begin{aligned} u_t &= \triangle u + \alpha u - \beta v - a u^3, \\ v_t &= \nu \triangle v + \beta u - \alpha v + \lambda v, \end{aligned}$$

with $\beta^2 - \alpha^2 = 1$ and $\alpha, \beta > 0$. We assume in the following that λ is close to zero; that is, we are close to a Hopf bifurcation with eigenvalues $\pm i$ of the linearized reaction system. The rotating wave ansatz yields

$$\begin{aligned} c u_\varphi &= \triangle_{r,\varphi} u + \alpha u - \beta v - a u^3, \\ c u_\varphi &= \nu \triangle_{r,\varphi} v + \beta u - \alpha v + \lambda v, \end{aligned} \tag{10.1}$$

where $\triangle_{r,\varphi} = \partial_{rr} + \frac{1}{r}\partial_r + \frac{1}{r^2}\partial_{\varphi\varphi}$. The linearization at $\lambda = 0, u = v = 0$ is

$$\begin{aligned} c u_\varphi &= \triangle_{r,\varphi} u + \alpha u - \beta v, \\ c v_\varphi &= \nu \triangle_{r,\varphi} v + \beta u - \alpha v. \end{aligned} \tag{10.2}$$

We now expand the solutions in Fourier series with respect to φ

$$(u, v) = \sum_{k \in \mathbb{Z}} (u^k, v^k) e^{ik\varphi}, \quad (u^{-k}, v^{-k}) = (\overline{u^k}, \overline{v^k}).$$

The linearization (10.2) then becomes an uncoupled system of ODEs for the Fourier coefficients

$$\begin{aligned} \triangle_{r,k} u^k &= (cik - \alpha) u^k + \beta v^k, \\ \nu \triangle_{r,k} v^k &= -\beta u^k + (cik + \alpha) v^k, \end{aligned}$$

where $\triangle_{r,k} = \partial_{rr} + \frac{1}{r}\partial_r - \frac{k^2}{r^2}$. The right side has a kernel as a linear operator on $\mathbb{C}^2$ whenever $ck = 1$. We therefore set $c = 1/k_0 + \mu$ with μ close to zero, having fixed $k_0 \in \mathbb{N}$ for the sequel.

Remember that together with the above equations we should write the equations for the complex conjugates, which are just the conjugate equations.

The eigenvector in the kernel is easily calculated as

$$w_0 = \beta u^k + \nu(i - \alpha) v^k, \quad \triangle_{r,k} w_0 = 0,$$

and

$$w_1 = -\beta u^k + (i + \alpha) v^k, \quad \triangle_{r,k} w_1 = \left(i - \alpha + \frac{i + \alpha}{\nu} \right) w_1$$

is the complementary eigenvector to the eigenvalue $i - \alpha + (i + \alpha)/\nu$.

Proposition 5 implies the existence of a center-manifold $\mathcal{M}^c_+$ with a smooth vector field, tangent to the span of $w_0 e^{ik_0\varphi}$ and $\partial_r w_0 e^{ik_0\varphi}$ at any "time" r. The vector field is obtained up to third order using the following strategy:

(i) Write the linear equation for w_0, depending on parameters λ, μ; this gives the linear part of the vector field on $\mathcal{M}_+^c$.

(ii) Calculate the quadratic (in w_0) expansion of $\mathcal{M}_+^c$ depending on time; this is zero, due to the absence of quadratic terms in the reaction.

(iii) Evaluate the nonlinearity au^3 on $w_0 e^{ik_0\varphi}$.

(iv) Project away noncritical Fourier modes.

(v) Project on $\langle w_0 e^{ik_0\varphi}\rangle$ along $\langle w_1 e^{ik_0\varphi}\rangle$.

Carrying out the necessary calculations gives first, by projecting away the noncritical Fourier modes,

$$\begin{aligned}\triangle_{r,k}u^k &= (i-\alpha)u^k + \beta v^k + i\mu u^k + au^k|u^k|^2,\\ \nu\triangle_{r,k}v &= -\beta u^k + (i+\alpha v^k) + i\mu v^k - \lambda v^k,\end{aligned}$$

and therefore

$$\triangle_{r,k}w_0 = i\mu\beta u^k + \beta a u^k|u^k|^2 - (i-\alpha)\lambda v^k + (i-\alpha)i\mu v^k.$$

Transforming back

$$\begin{aligned}u^k &= \frac{i+\alpha}{\beta(i+\alpha+\nu(i-\alpha))}w_0 + O(w_1),\\ v^k &= \frac{1}{i+\alpha+\nu(i-\alpha)}w_0 + O(w_1)\end{aligned}$$

gives, on $\mathcal{M}_+^c$, up to third order, the second-order-in-time ODE

$$\triangle_{r,k}w_0 = \frac{-2\mu-(i-\alpha)\lambda}{i+\alpha+\nu(i-\alpha)}w_0 + \frac{a}{\beta^2}\frac{1}{1+\nu\frac{i-\alpha}{i+\alpha}}\left|\frac{1}{1+\nu\frac{i-\alpha}{i+\alpha}}\right|^2 w_0|w_0|^2.$$

The fifth-order terms might of course destroy the second-order structure of this equation, although keeping the structure of a local nonautonomous differential equation in $\mathbb{C}^2$.

We write new parameters $\lambda', a' \in \mathbb{C}$ such that the truncated equation takes the form

$$(w_0)_{rr} + \frac{1}{r}(w_0)_r - \frac{k^2}{r^2}w_0 = \lambda' w_0 + a' w_0|w_0|^2. \tag{10.3}$$

Disregarding all of our efforts in reducing and simplifying the problem, this equation is in general still hard to solve analytically. In the following section we study this problem, obtaining existence of bounded solutions $w(r)$ (and thereby solutions $(u(r,\varphi), v(r,\varphi))$ to (10.1)), when a' is almost real. This is actually the approach taken by [7, 11], who deal with a similar system.

By our explicit calculations, the imaginary part of a' will be small if the diffusion rate ν or the parameter α is close to zero.

The first condition has an interesting interpretation as the limit $d_2 \to 0$ is exactly the interesting limit in excitable media, although we admit that our equation is different from the typical models for excitable media (the null-clines of $\alpha u - au^3 - v$ are symmetric to the origin whereas this is not the usual assumption for excitable media, modeled for example by the FitzHugh–Nagumo equation). We refer the reader to the interesting, although formal, work on spiral waves in excitable media reviewed in [22].

The second, alternative condition is merely an assumption on the location of equilibria in the pure reaction system, which are situated approximately at $u \sim \pm\sqrt{b/(a\alpha)}$ and zero.

The important point to notice at this stage is that the full equation on $\mathcal{M}_+^c$ is a small perturbation of the truncated equation close to the bifurcation point; that is, close to $\operatorname{Re}\lambda' = 0$. Indeed, scaling $|\operatorname{Re}\lambda'|r^2 = \tilde{r}^2$ and $w_0^2 = |\operatorname{Re}\lambda'|\tilde{w_0}^2$ makes the higher-order terms $\mathrm{O}(|\operatorname{Re}\lambda'|)$. Structurally stable dynamics of the truncated equation persist for the full system on $\mathcal{M}_+^c$ for sufficiently small $|\operatorname{Re}\lambda'|$.

In this sense, we have established a rigorous proof of the validity of approximations of reaction-diffusion systems by λ-ω systems, at least when we restrict to the question of existence of rotating wave solutions. This was proved up to now only using formal multiscale methods. The advantage of our approach is that it gives rigorous proofs and information on the domain of validity in parameter space of such kinds of approximations.

Furthermore, we should comment on the symmetry. The equation possesses, as announced, an $\mathcal{SO}(2)$-symmetry $w_0 \to w_0 e^{i\theta}$, $\theta \in S^1$. The additional reflection $(u, v) \to (-u, -v)$ in the original reaction-diffusion system does not yield any more symmetry in the bifurcation equation.

At $\operatorname{Im} a' = \operatorname{Im}\lambda' = 0$, there is the additional reflectional symmetry $w_0 \to \overline{w_0}$, fixing the real subspace. Note also that $\operatorname{Im}\lambda' = 0$ can be achieved by adjusting the wave speed c.

11. The bifurcation equations. During this section we omit the primes of λ and a. We begin with a study of possible asymptotic states of (10.3) at $r = \infty$. The limit equation

$$w'' = \lambda w + aw|w|^2, \quad {}' = \frac{d}{dr},$$

can be simplified by dividing out the symmetry with the new coordinates $z = z_R + iz_I = w'/w \in \bar{\mathbb{C}}$ and $R = |w| \in \mathbb{R}_+$:

$$\begin{aligned} R' &= z_R R, \\ z' &= -z^2 + \lambda + aR^2. \end{aligned} \tag{11.1}$$

Reversibility of the w-equation ($r \to -r$) is translated into reversibility with respect to the reflection $z \to -z$ (and, of course, $r \to -r$). Any equilibrium of (11.1) corresponds to a periodic orbit of the w-equation, which we call a rotating wave, as it is a relative equilibrium, for the dynamics in r, with respect to rotational symmetry $\mathcal{SO}(2)$. The asymptotic shape of a spiral wave behaving like such a rotating wave for large r is just a one-dimensional periodic wave-train, translation invariant in one space-direction. There are two types of equilibria. Type I has $R = 0$ and corresponds to the origin of the w-equation, and $z = \pm\sqrt{\lambda}$ are the blown up invariant manifolds of the equilibrium $w = w' = 0$. Type II has necessarily $z_R = 0$ and

$$R^2 = -\frac{\lambda_I}{a_I}, \quad z_I^2 = -\lambda_R + a_R\frac{\lambda_I}{a_I}.$$

A linear stability analysis gives that the type I equilibrium with $\operatorname{Re}\sqrt{\lambda} > 0$ is stable in $\{R = 0, z \in \bar{\mathbb{C}}\}$ and unstable in the direction of R. The equilibrium with $\operatorname{Re}\sqrt{\lambda} < 0$ is unstable in $\{R = 0, z \in \bar{\mathbb{C}}\}$ but stable in the direction of R. Along the type II

equilibria the linearization is

$$L = \begin{pmatrix} 0 & R & 0 \\ 2a_R R & 0 & 2z_I \\ 2a_I R & -2z_I & 0 \end{pmatrix}, \qquad \det L = -4\lambda_I z_I, \quad \operatorname{trace} L = 0,$$

such that one equilibrium is two-dimensional unstable and the other is two-dimensional stable.

Bifurcations occur at $\operatorname{Re}\sqrt{\lambda} = 0$, where type I equilibria coalesce, the origin becomes a center, and when $a\bar{\lambda} \in \mathbb{R}$, where a reversible saddle-node bifurcation of the type II equilibria occurs.

For the nonautonomous system, we can interpret the manifold $\mathcal{M}$ as a shooting manifold, which is two-dimensional in (w, w')-space at any fixed time r, invariant under the symmetry and therefore yields a one-dimensional shooting curve in the reduced phase space (z, R). We focus here on asymptotically stationary behavior, where the shooting curve intersects the stable manifold of an equilibrium of (11.1). These are possibly not the only asymptotic shapes at large distances from the center of rotation, but they seem to be of sufficient physical relevance, making such a restriction reasonable.

In the following, we distinguish two different cases which we refer to as the subcritical case, when $a_R > 0$, and the supercritical case, when $a_R < 0$. These terms are justified by the branching of equilibria of (11.1) at $\lambda_I = a_I = 0$. In our model problem of the preceding section, these two cases are distinguished by the sign of $a(1 - \nu(\alpha^2 - 1)/\beta^2)$.

We now study the real subsystem in the nonautonomous setting.

LEMMA 6 (supercritical [5, 11]). *Suppose $a_R < 0$ and $\lambda_R > 0$. Then for any wave number $k_0 \in \mathbb{N}$, there exists a heteroclinic orbit $w(r) > 0$, with $\lim_{r\to 0} w(r) = 0$ and $\lim_{r\to\infty} w(r) = \sqrt{-\lambda_R/a_R}$. Moreover the heteroclinic orbit is transverse in the real subsystem: the center-manifold $\mathcal{M}$ intersects transversely the stable manifold of $\sqrt{-\lambda_R/a_R}$.*

Proof. The proof of this lemma can be found in [11], where the necessary modifications to the proof of a similar statement in [5] are indicated. □

LEMMA 7 (subcritical). *Suppose $a_R > 0$ and $\lambda_R < 0$. Then for any $k_0 \in \mathbb{N}$, there exists a heteroclinic orbit $w(r)$, with $\lim_{r\to 0} w(r) = \lim_{r\to\infty} w(r) = 0$. Moreover the heteroclinic orbit is transverse in the real subsystem: the manifold $\mathcal{M}$ intersects transversely the stable manifold of the origin at $r = \infty$.*

Proof. The proof, together with a more detailed description of such solutions, can be found in [20]. □

We next examine the nondegenerate system with $a_I \neq 0$.

PROPOSITION 8 (supercritical, $a_I \neq 0$). *Suppose $a_R < 0$ and $\lambda_R > 0$ and fix any wave number $k_0 \in \mathbb{N}$. Then for any a_I, λ_I sufficiently small and $\lambda_I/a_I - 1 \gg a_I^2$ there exists a heteroclinic orbit $w(r)$, with $\lim_{r\to 0} w(r) = 0$ and tending to a type* II *equilibrium as $r \to \infty$. The heteroclinic orbit is transverse. Moreover there exists a unique value $\lambda_I^0 = \mathrm{O}(a_I)$ such that the heteroclinic orbit tends to the other type* II *equilibrium as $r \to \infty$. This heteroclinic orbit is transversely unfolded by the parameter λ_I.*

Proof. We suppose $a_R = -1$ and $\lambda_R = 1$. We use singular perturbation methods in order to establish the existence of heteroclinic orbits for the perturbed system. At $a_I = \lambda_I = 0$, there is a curve of type II equilibria for the asymptotic equations at $r = \infty$, given by $z_I^2 = 1 - R^2$, which intersects transversely the real subspace at the

equilibrium $z = 0, R = 1$. Therefore the center-stable manifold of this line of equilibria intersects transversely the shooting manifold $\mathcal{M}$ in (z, R, τ)-space. In the perturbed system, the line of equilibria persists as a normally hyperbolic slow manifold (see [3]). The heteroclinic as a transverse intersection persists as the intersection with a strong stable fiber of the slow manifold for a_I, λ_I small enough. On the slow manifold there are two equilibria $z_I^2 = -1 - \lambda_I/a_I$, which are close to the real subspace $\{z_I = 0\}$ if $-\lambda_I/a_I$ is close to but bigger than one. By the above stability analysis, the equilibrium which is stable within the slow manifold has $\det L > 0$ and thereby $\lambda_I z_I < 0$. We now have to examine the perturbation of the shooting manifold $\mathcal{M}$ by the complex perturbation terms involving λ_I and a_I. The derivative along the real heteroclinic at $\lambda_I = a_I = 0$ of the nonautonomous equation for z_I with respect to λ_I and a_I gives

$$z_I' = \lambda_I + a_I R^2 = a_I(\lambda_I/a_I + R^2).$$

Thereby the Melnikov integral along the heteroclinic gives a contribution $\mathrm{O}(a_I)$, which shows that the shooting manifold $\mathcal{M}$ intersects transversely a stable fiber of a point on the slow manifold with $z_I^0 = \mathrm{O}(a_I)$.

With these ingredients we can establish the existence of the desired connections. First choosing λ_I as a parameter, the shooting manifold $\mathcal{M}$ crosses transversely the strong stable fibers of the slow manifold. The type II equilibria on the slow manifold are located at $\mathrm{O}(\sqrt{|\lambda_I/a_I + 1|})$. If $|z_I^0| < \sqrt{|\lambda_I/a_I - 1|}$, there is a heteroclinic trajectory connecting to the type II equilibrium, which is stable within the slow manifold. If $(z_I^0)^2 = -\lambda_I/a_I - 1$, the heteroclinic trajectory connects to the type II equilibrium, which is unstable on the slow manifold. This proves the proposition. □

PROPOSITION 9 (subcritical, $a_I \neq 0$). *Suppose $a_R > 0$ and $\lambda_R < 0$ and fix any wave number $k_0 \in \mathbb{N}$. Then for any a_I sufficiently small, there exists a smooth function $\lambda_I = \lambda_I(a_I)$ such that there exists a heteroclinic orbit $w(r)$, with $\lim_{r\to 0} w(r) = \lim_{r\to\infty} w(r) = 0$. The heteroclinic orbit is transversely unfolded by the parameter λ_I.*

Proof. We suppose $a_R = 1$ and $\lambda_R = -1$. In the real subspace at $\lambda_I = a_I = 0$, the heteroclinic orbit joining the origin at $r = 0$ to the origin at $r = \infty$ is transverse by Lemma 7. Transverse to the real subspace, the origin is unstable at both $r = 0$ and $r = \infty$: the heteroclinic is nontransverse in full-space. We now need the parameter λ_I (alias the speed of rotation) in order to obtain connections for specific values of the parameter $\lambda_I = \lambda_I(a_I)$. For this it is sufficient to show that the Melnikov integral with respect to the parameter λ_I along the heteroclinic does not vanish. The adjoint variational equation along the heteroclinic has a unique (up to scalar multiples) bounded solution which lies strictly in the half space $z_I > 0$, because $z_I = 0$ is invariant. The derivative of the vector field with respect to λ_I in the direction of this half space is just 1, which proves that the Melnikov integral is nonzero. In other words we can push through the stable and unstable manifolds with the help of λ_I with nonzero speed. This proofs the proposition. □

12. Conclusions. For a large class of reaction-diffusion systems we have shown the existence of spiral wave solutions. In contrast to the previous results on λ-ω systems, our reduction to a nonautonomous ODE is not based on the *assumption* that Fourier modes decouple. We merely show that, close to the threshold of instability of a homogeneous equilibrium, there is some kind of decoupling. The interaction between critical modes is in a smooth sense of higher order than the projection on the critical modes. Compared to similar reduction methods, technical complications arise here

because the problem is nonautonomous, even in the principal part (from a regularity point of view).

As another advantage of our method, we are able to determine explicitly coefficients in our bifurcation equations. These are in general still hard to analyze analytically—we considered a simple but interesting model problem in the last section—but can easily be studied numerically.

The reduction procedure can be applied to other problems, possibly involving higher-dimensional center-manifolds. A systematic treatise of such equations (as known for elliptic problems in infinite cylinders, exploiting reversibility, integrability, and normal forms of the reduced bifurcation equations) would be interesting.

The rotating waves we discover are of various shape, depending on the nature of the bifurcation. In supercritical bifurcations, they are approximately archimedian spirals at large distances from the tip. Indeed, the derivative of the phase of u is given by z_I and approaches for large values of the radius r a constant but nonzero value. As a subtle difference we noticed that in the supercritical case there are two different types of asymptotic states, given by the two different types of equilibria $z_I = \pm\sqrt{-\lambda_I/a_I - 1}$ (see the preceding section). For the first type, z_I approaches its limit value exponentially at a uniform rate with respect to λ_I, whereas for the other type the exponential rate is close to zero. The sign of z_I has another important interpretation. If z_I is positive, then the arms turn in the sense of the rotation of the spiral; at a fixed ray, under time evolution of the reaction-diffusion system, the arms move toward the center of rotation. Similarly, $z_I < 0$ corresponds to an outwards movement of the arms. Therefore the waves, appearing for discrete wave speeds, move outward if $a_I < 0$ and inward if $a_I > 0$.

The rotating waves bifurcating subcritically are isolated as rotating waves and appear for distinguished speeds of rotation. Their shape at large distances from the center of rotation is determined by the phase varying according to $\varphi \sim e^{-\mathrm{const}\cdot r}$ and their amplitude decaying to zero exponentially.

Although we do not carry out here a stability analysis, we comment on the difference between sub- and supercritical bifurcation. Linearizing the reaction-diffusion system along the subcritical waves in, say, $L^2(\mathbb{R}^2, \mathbb{R}^2)$ gives us a linearized operator for the period map whose continuous spectrum is strictly contained in the left half plane, bounded away from the imaginary axis. Zero is (at least) a triple eigenvalue due to the Euclidean symmetry, generated by rotation and translations. An analysis of secondary bifurcations from this type of spiral wave, including meandering and drifting waves, has been carried out in [17, 18] and [4].

The linearized period map along supercritical waves has zero in the essential spectrum and rigorous stability proofs seem to be hard. Hagan [7] showed that one-armed spiral waves might be stable whereas multiarmed waves ($k_0 \neq 1$) should be unstable.

REFERENCES

[1] D.S. Cohen, J.C. Neu, and R.R. Rosales, *Rotating spiral wave solutions of reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 536–547.

[2] W.A. Coppel, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, 1978.

[3] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[4] B. Fiedler, B. Sandstede, A. Scheel, and C. Wulff, *Bifurcation from relative equilibria with non-compact group actions: Skew products, meanders and drifts*, Doc. Math. J. DMV, 1 (1996), pp. 479–505.
[5] J.M. Greenberg, *Spiral waves for λ-ω systems,* SIAM J. Appl. Math., 39 (1980), pp. 301–309.
[6] J.M. Greenberg, *Spiral waves for λ-ω systems,* II, Adv. Appl. Math., 2 (1989), pp. 450–455.
[7] P.S. Hagan, *Spiral waves in reaction-diffusion equations*, SIAM J. Appl. Math., 42 (1982), pp. 762–786.
[8] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 804, Springer-Verlag, New York, 1981.
[9] K. Kirchgässner, *Wave solutions of reversible systems and applications*, J. Differential Equations, 45 (1982), pp. 113–127.
[10] N. Kopell and L.N. Howard, *Plane wave solutions to reaction-diffusion equations*, Stud. Appl. Math., 52 (1973), pp. 291–328.
[11] N. Kopell and L.N. Howard, *Target patterns and spiral solutions to reaction-diffusion equations with more than one space dimension*, Adv. Appl. Math., 2 (1981), pp. 417–449.
[12] I. Melbourne, *Steady-State Bifurcation with Euclidean Symmetry*, Research Report, University of Houston/MD-214; Trans. Amer. Math. Soc., to appear.
[13] A. Mielke, *A reduction principle for non-autonomous systems in infinite-dimensional spaces,* J. Differential Equations, 65 (1986), pp. 68–88.
[14] A. Mielke, *Reduction of PDEs on domains with several unbounded directions: A first step towards modulation equations,* Z. Angew. Math. Phys., 43 (1992), pp. 449–470.
[15] K. J. Palmer, *Exponential dichotomies and transversal homoclinic points,* J. Differential Equations, 55 (1984), pp. 225–256.
[16] D. Peterhof, B. Sandstede, and A. Scheel, *Exponential dichotomies for solitary-wave solutions of semilinear elliptic equations on infinite cylinders,* J. Differential Equations, 140 (1997), pp. 266–308.
[17] B. Sandstede, A. Scheel, and C. Wulff, *Center-manifold reduction for spiral waves*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 153–158.
[18] B. Sandstede, A. Scheel, and C. Wulff, *Dynamics of spiral waves on unbounded domains using center-manifold reductions*, J. Differential Equations, 141 (1997), pp. 122–149.
[19] A. Scheel, *Existence of fast travelling waves for some parabolic equations – a dynamical systems approach,* J. Dynam. Differential Equations, 8 (1996), pp. 469–548.
[20] A. Scheel, *Subcritical bifurcation to infinitely many rotating waves,* J. Math. Anal. Appl., 215 (1997), pp. 252–261.
[21] G. Schneider, *Hopf-bifurcation in spatially extended reaction-diffusion systems,* Preprint 1996.
[22] J. J. Tyson and J. P. Keener, *Singular perturbation theory of traveling waves in excitable media (a review),* Phys. D, 32 (1988), pp. 327–361.
[23] A. Vanderbauwhede, *Center-manifolds, normal forms and elementary bifurcations*, Dynam. Report. Expositions Dynam. Systems, 2 (1989), pp. 89–169.
[24] A. Vanderbauwhede and G. Iooss, *Center-manifold theory in infinite dimensions*, Dynam. Report. Expositions Dynam. Systems (N.S.), 1 (1992), pp. 125–163.
[25] G. N. Watson, *Theory of Bessel Functions*, Cambridge University Press, London, 1922.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1419–1433, November 1998

006

THE SURFACE DIFFUSION FLOW FOR IMMERSED HYPERSURFACES*

JOACHIM ESCHER†, UWE F. MAYER‡, AND GIERI SIMONETT‡

Abstract. We show existence and uniqueness of classical solutions for the motion of immersed hypersurfaces driven by surface diffusion. If the initial surface is embedded and close to a sphere, we prove that the solution exists globally and converges exponentially fast to a sphere. Furthermore, we provide numerical simulations showing the creation of singularities for immersed curves.

Key words. surface diffusion, mean curvature, free boundary problem, immersed hypersurfaces, center manifolds, maximal regularity, numerical simulations

AMS subject classifications. 35R35, 35K55, 35S30, 65C20, 80A22

PII. S0036141097320675

1. Introduction.

In this paper we study the motion of a family of immersed hypersurfaces whose normal velocity is equal to its surface diffusion. More precisely, let Γ_0 be a compact closed immersed orientable hypersurface in $\mathbb{R}^n$ of class $C^{2+\beta}$. We are looking for a family $\Gamma = \{\Gamma(t); t \geq 0\}$ of smooth immersed orientable hypersurfaces satisfying the following evolution equation:

$$V(t) = \Delta_{\Gamma(t)} H_{\Gamma(t)}\,, \qquad \Gamma(0) = \Gamma_0\,. \tag{1.1}$$

Here $V(t)$ denotes the velocity in the normal direction of Γ at time t, while $\Delta_{\Gamma(t)}$ and $H_{\Gamma(t)}$ stand for the Laplace–Beltrami operator and the mean curvature of $\Gamma(t)$, respectively. Both the normal velocity and the curvature depend on the local choice of the orientation; however, (1.1) does not, and so we are free to choose whichever one we like. In particular, if $\Gamma(t)$ is embedded and encloses a region $\Omega(t)$, we always choose the *outer* normal, so that V is positive if $\Omega(t)$ grows and so that $H_{\Gamma(t)}$ is positive if $\Gamma(t)$ is convex with respect to $\Omega(t)$. Due to the local nature of the evolution, we may assume the hypersurface Γ_0 to be connected.

In order to give precise results, let us introduce the following notation. Given an open set $U \subset \mathbb{R}^n$, let $h^s(U)$ denote the little Hölder spaces of order $s > 0$, that is, the closure of $BUC^\infty(U)$ in $BUC^s(U)$, the latter space being the Banach space of all bounded and uniformly Hölder continuous functions of order s. If Σ is a (sufficiently) smooth submanifold of $\mathbb{R}^n$ then the spaces $h^s(\Sigma)$ are defined by means of a smooth atlas for Σ.

THEOREM 1.1. *Assume that $0 < \beta < 1$, and let Γ_0 be a compact closed immersed orientable hypersurface in $\mathbb{R}^n$ belonging to the class $h^{2+\beta}$.*

(a) *The surface diffusion flow* (1.1) *has a unique local classical solution $\Gamma = \{\Gamma(t); t \in [0,T)\}$ for some $T > 0$. Each hypersurface $\Gamma(t)$ is of class C^∞ for $t \in (0,T)$. Moreover, the mapping $[t \mapsto \Gamma(t)]$ is continuous on $[0,T)$ with respect to the $h^{2+\beta}$-topology and smooth on $(0,T)$ with respect to the C^∞-topology.*

*Received by the editors April 23, 1997; accepted for publication (in revised form) December 8, 1997; published electronically August 3, 1998.
http://www.siam.org/journals/sima/29-6/32067.html
†Mathematical Institute, University of Basel, CH-4051 Basel, Switzerland (escher@math.unibas.ch).
‡Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (mayer@math.vanderbilt.edu, simonett@math.vanderbilt.edu).

(b) *Suppose that the initial hypersurface* Γ_0 *is a* $h^{2+\beta}$*-graph in normal direction over some smooth immersed hypersurface* Σ*. Then the mapping* $\varphi := [(t, \Gamma_0) \mapsto \Gamma(t)]$ *induces a smooth local semiflow on an open subset of* $h^{2+\beta}(\Sigma)$*.*

Remark. The assumption that the initial surface be orientable is not necessary for Theorem 1.1(a) to hold true.

This follows by evolving the double cover in case the initial hypersurface is not orientable. The double cover remains a double cover by uniqueness of smooth solutions, and (1.1) is invariant with respect to the local orientation, as noted before. Hence, one can go back to the quotient space which therefore also evolves according to the surface diffusion flow. □

The motion given by (1.1) has some interesting geometrical features. Assume that Γ is a smooth orientable solution to (1.1) and let $\mathcal{A}(t)$ denote the area of $\Gamma(t)$. Then the function $\mathcal{A}$ is smooth and we find for its derivative (e.g., see [22, Theorem 4] or [15, p. 70])

$$\begin{aligned} \frac{1}{n-1}\frac{d}{dt}\mathcal{A}(t) &= \int_{\Gamma(t)} V(t) H_{\Gamma(t)}\, d\sigma = \int_{\Gamma(t)} [\Delta_{\Gamma(t)} H_{\Gamma(t)}] H_{\Gamma(t)}\, d\sigma \\ &= -\int_{\Gamma(t)} |\mathrm{grad}_{\Gamma(t)} H_{\Gamma(t)}|^2_{\Gamma(t)} d\sigma \le 0\,. \end{aligned} \tag{1.2}$$

Hence the motion driven by surface diffusion is area decreasing.

It is also possible to identify the surface diffusion flow as an H^{-1}-gradient flow for the area functional; see [10, 25]. The notion of such gradient flows was proposed by Fife [19, 20].

Assume that Γ is a smooth solution to (1.1) consisting of embedded hypersurfaces which enclose a region $\Omega(t)$, and let $\mathrm{Vol}(t)$ denote the volume of $\Omega(t)$. The derivative of the smooth function Vol is then given by

$$\frac{d}{dt}\mathrm{Vol}(t) = \int_{\Gamma(t)} V(t)\, d\sigma = \int_{\Gamma(t)} \Delta_{\Gamma(t)} H_{\Gamma(t)}\, d\sigma = 0\,,$$

thus the motion driven by surface diffusion is also volume preserving in the embedded case. Every Euclidean sphere is an equilibrium for (1.1), and it follows from Alexandrov's characterization [1] of embedded constant mean curvature surfaces that spheres are the only equilibria. However, none of these equilibria is isolated, since in every neighborhood of a fixed sphere there is a continuum of further spheres. Thus the dynamics of the flow generated by (1.1) is even locally quite copious.

THEOREM 1.2. *Let* S *be a fixed Euclidean sphere and let* $\mathcal{M}$ *denote the set of all spheres which are sufficiently close to* S*. Then* $\mathcal{M}$ *attracts all embedded solutions which are* $h^{2+\beta}(S)$*-close to* $\mathcal{M}$ *at an exponential rate. In particular, if* Γ_0 *is sufficiently close to* S *in* $h^{2+\beta}(S)$*, then* Γ *exists globally and converges exponentially fast to some sphere in* $\mathcal{M}$ *enclosing the same volume as* Γ_0*. The convergence is in the* C^k*-topology for every initial hypersurface* Γ_0 *which is in a sufficiently small* $h^{2+\beta}(S)$*-neighborhood* $W = W(k)$ *of* S*, where* $k \in \mathbb{N}$ *is a fixed number.*

The surface diffusion flow (1.1) was first proposed by Mullins [26] to model surface dynamics for phase interfaces when the evolution is only governed by mass diffusion in the interface. It has also been examined in a more general mathematical and physical context by Davì and Gurtin [13], and by Cahn and Taylor [9]. More recently, Cahn, Elliott, and Novick-Cohen [8] showed by formal asymptotics that the surface diffusion flow is the singular limit of the zero level set of the solution to the Cahn–Hilliard

equation with a concentration-dependent mobility. In the case of constant mobility in the Cahn–Hilliard equation, Alikakos, Bates, and Chen [2] proved that the motion of the singular limit is governed by the Mullins–Sekerka model (also called the Hele–Shaw model with surface tension), rigorously establishing a result that was formally derived by Pego [27]. The Mullins–Sekerka model shares many properties with the surface diffusion flow (1.1). They both preserve the enclosed volume and decrease the area of the interface, and for both the invariant manifold $\mathcal{M}$ of spheres is exponentially attracting; see [16, 17, 18].

In two dimensions and for strip-like domains, the surface diffusion flow was investigated by Baras, Duchon, and Robert [7]. They prove global existence of weak solutions. Also in two dimensions, the surface diffusion flow for closed embedded curves was analytically investigated by Elliott and Garcke [14]. They show local existence and regularization for C^4-initial curves, and global existence for small perturbations of circles. Furthermore, assuming global existence, they show that any closed curve will become circular under this evolution. They do not obtain uniqueness of solutions. Recently, Giga and Ito [21] established the existence of unique (local) solutions for immersed H^4-initial curves. Moreover, they prove that the surface diffusion flow can drive an initially embedded curve to a self intersection. The techniques in [14, 21] seem to be restricted to two dimensions.

Our methods work in any dimension, and we obtain existence and uniqueness for immersed hypersurfaces. This is of particular interest since embedded hypersurfaces can become immersed under the surface diffusion flow, which is in clear contrast to the mean curvature flow where smooth solutions remain embedded if their initial surface is embedded. Our numerical simulations show that an immersed curve can develop singularities under the surface diffusion flow. Our example consists of a curve with a loop within a loop where the inner loop tightens and then contracts to a point. This situation has been analyzed in great detail by Angenent [6] for the mean curvature flow. In case of surface diffusion we do not have an analytical proof for the occurrence of singularities. We also give an example showing that an immersed curve evolves towards a stable limiting configuration which is not an embedded circle, but a multiply covered immersed circle. Finally, we provide evidence that the surface diffusion flow shrinks a figure eight to a point in finite time. Our approach for proving existence and uniqueness of solutions can be used to set up the numerical scheme for our simulations.

In case the initial hypersurface has several components, it is clear that some components may collide under the surface diffusion flow. This is most easily seen by choosing any nonstationary initial hypersurface, and then placing a stationary sphere in its path.

Theorem 1.1 constitutes a precise local existence and uniqueness result for classical solutions to (1.1) starting out as immersed hypersurfaces. In particular, the results disclose a parabolic regularization of the flow φ since we are allowed to choose initial surfaces Γ_0 of class $h^{2+\beta}$, although $\Delta_{\Gamma_0} H_{\Gamma_0}$ is for such Γ_0 in general not a classical function. This parabolic structure also provides the foundation for the study of the qualitative behavior of the semiflow φ. Our approach for proving existence, uniqueness, and regularity of solutions is based on the general theory of Amann [3, 4] for quasi-linear parabolic evolution equations.

The proof of Theorem 1.2 consists of two steps. We first show that the semiflow φ admits a stable $(n+1)$-dimensional local center manifold $\mathcal{M}^c$. This means, in particular, that $\mathcal{M}^c$ is a locally invariant manifold and that $\mathcal{M}^c$ contains all small global

solutions of φ. In a second step we then prove that $\mathcal{M}^c$ coincides with the manifold $\mathcal{M}$ of the theorem. It is well-known that local center manifolds are generally not unique. However, since each local center manifold of the surface diffusion flow consists only of equilibria, this forces uniqueness. Under suitable spectral assumptions for the linearization, the existence of center manifolds is well known for finite-dimensional dynamical systems. The corresponding construction for quasi-linear infinite-dimensional semiflows (e.g., for φ) is considerably more involved. The basic technical tool here is the theory of maximal regularity, due to Da Prato and Grisvard [11]; see also [4, 5, 23]. In particular, these results allow to treat (1.1) as a fully nonlinear perturbed linear evolution equation; see [12, 23, 29].

2. Existence and uniqueness. In this section we introduce the mathematical setting in order to reformulate (1.1) as a quasi-linear parabolic evolution equation. Let Σ be a smooth compact closed immersed oriented hypersurface in $\mathbb{R}^n$, and assume that Γ_0 is close to this fixed reference manifold Σ. Let ν be the unit normal field on Σ commensurable with the chosen orientation. Choose $a > 0$ and an open covering $\{U_l\,;\, l = 1, \dots, m\}$ of Σ such that

$$X_l : U_l \times (-a, a) \to \mathbb{R}^n\,, \qquad X_l(s, r) := s + r\nu(s)$$

is a smooth diffeomorphism onto its image $\mathcal{R}_l := \mathrm{im}(X_l)$, that is,

$$X_l \in \mathit{Diff}^\infty(U_l \times (-a, a), \mathcal{R}_l)\,, \qquad 1 \le l \le m\,.$$

This can be done by choosing the open sets $U_l \subset \Sigma$ in such a way that they are embedded in $\mathbb{R}^n$ instead of only immersed, and then taking $a > 0$ sufficiently small so that each of the U_l has a tubular neighborhood of radius a. It is convenient to decompose the inverse of X_l into $X_l^{-1} = (S_l, \Lambda_l)$, where

$$S_l \in C^\infty(\mathcal{R}_l, U_l) \qquad \text{and} \qquad \Lambda_l \in C^\infty(\mathcal{R}_l, (-a, a))\,.$$

Note that $S_l(x)$ is the nearest point on U_l to $x \in \mathcal{R}_l$, and that $\Lambda_l(x)$ is the signed distance from x to U_l (that is, to $S_l(x)$). Moreover, the union of the sets $\mathcal{R}_l$, $1 \le l \le m$, consists exactly of those points in $\mathbb{R}^n$ with distance less than a to Σ.

Let $T > 0$ be a fixed number. We assume that $\Gamma := \{\Gamma(t),\, t \in [0, T)\}$ is a family of immersed graphs in normal direction over Σ. To be precise, we ask that there is a globally defined function

$$\rho : \Sigma \times [0, T) \to (-a, a)$$

such that for fixed $t \in [0, T)$, a manifold $\Gamma(t)$ is locally given by the images of the maps $[s \mapsto X_l(s, \rho(s, t))]$, $1 \le l \le m$.

Conversely, given any (sufficiently) smooth function $\rho : \Sigma \times [0, T) \to (-a, a)$, let

$$\Phi_{l,\rho} : \mathcal{R}_l \times [0, T) \to \mathbb{R}\,, \qquad \Phi_{l,\rho}(x, t) := \Lambda_l(x) - \rho(S_l(x), t)\,, \quad 1 \le l \le m\,. \tag{2.1}$$

Then for each $t \in [0, T)$, the zero-level set $\Phi_{l,\rho}^{-1}(0, t) \subset \mathcal{R}_l$ defines a smooth hypersurface, and the hypersurfaces $\Phi_{l,\rho}^{-1}(0, t)$ can be glued together to constitute a compact closed immersed orientable hypersurface $\Gamma_{\rho(t)}$. It is then easy to see that

$$\Gamma_{\rho(t)} = \Gamma(t) = \bigcup_{l=1}^{m} \mathrm{Im}\left(X_l : U_l \to \mathbb{R}^n\,,\ [s \mapsto X_l(s, \rho(s, t))]\right).$$

In addition, the normal velocity V of $\Gamma := \{\Gamma_{\rho(t)} \,;\, t \in [0,T)\}$ at time t and at the point $x = X_l(s,\rho(s,t))$, expressed as a function over U_l, is given by

$$V(s,t) = -\left.\frac{\partial_t \Phi_{l,\rho}(x,t)}{|\nabla_x \Phi_{l,\rho}(x,t)|}\right|_{x=X_l(s,\rho(s,t))} = \left.\frac{\partial_t \rho(s,t)}{|\nabla_x \Phi_{l,\rho}(x,t)|}\right|_{x=X_l(s,\rho(s,t))}$$

for $(s,t) \in U_l \times (0,T)$. In the following, we fix $t \in [0,T)$ and drop it in our notation. Moreover, we fix $0 < \alpha < \beta < 1$ and define

$$\mathfrak{A} := \{\rho \in h^{2+\alpha}(\Sigma) \,;\, \|\rho\|_{C(\Sigma)} < a\}\,.$$

Then for any $\rho \in \mathfrak{A}$,

$$\theta_\rho \,:\, \Sigma \to \Gamma_\rho, \qquad \theta_\rho(s) := X_l(s,\rho(s)) \text{ for } s \in U_l\,,$$

is a well-defined global $(2+\alpha)$-diffeomorphism. We write Δ_{Γ_ρ} for the Laplace–Beltrami operator of Γ_ρ and H_{Γ_ρ} for the mean curvature of Γ_ρ. Finally, let

$$G(\rho) := -L_\rho \theta_\rho^*(\Delta_{\Gamma_\rho} H_{\Gamma_\rho}) \quad \text{for} \quad \rho \in h^{4+\alpha}(\Sigma) \cap \mathfrak{A}\,,$$

where $L_\rho(s) := \theta_\rho^* |\nabla_x \Phi_{l,\rho}|(s)$ for $s \in U_l$, $1 \le l \le m$. On the phase space

$$\mathcal{V} := h^{2+\beta}(\Sigma) \cap \mathfrak{A},$$

we are now considering the following evolution equation for the distance function ρ:

$$\partial_t \rho + G(\rho) = 0\,, \qquad \rho(0) = \rho_0\,, \tag{2.2}$$

where ρ_0 is a function on Σ determined by Γ_0. More precisely, given $\rho \in \mathcal{V}$, we call a family $\rho \,:\, [0,T) \to \mathcal{V}$ a classical solution of (2.2) if

$$\rho \in C([0,T),\mathcal{V}) \cap C^\infty((0,T),C^\infty(\Gamma))$$

and if ρ satisfies (2.2) pointwise for $t \in (0,T)$. It is not difficult to see that the surface diffusion flow (1.1) and the evolution equation (2.2) are equivalent on $\mathcal{R} := \cup_{l=1}^m \mathcal{R}_l$. That is, if $\Gamma := \{\Gamma(t) \,;\, t \in [0,T)\}$ is a classical solution of (1.1) such that $\Gamma(t) \subset \mathcal{R}$ for $t \in [0,T)$, then the above construction yields a classical solution of (2.2) and vice-versa; if $\rho \,:\, [0,T) \to \mathcal{V}$ is a classical solution of (2.2), then $\Gamma := \{\Gamma_{\rho(t)} \,;\, t \in [0,T)\}$ is a classical solution of (1.1).

In order to state our next result, let E_1 and E_0 be Banach spaces with $E_1 \hookrightarrow E_0$, and let $\mathcal{H}(E_1,E_0)$ be the set of all $A \in \mathcal{L}(E_1,E_0)$ such that $-A$, considered as an unbounded operator in E_0, generates a strongly continuous analytic semigroup on E_0. It can be shown that $\mathcal{H}(E_1,E_0)$ is open in $\mathcal{L}(E_1,E_0)$; cf. [4, Theorem 1.3.1]. We always assume that $\mathcal{H}(E_1,E_0)$ carries the corresponding relative topology. Recall that we already fixed $0 < \alpha < \beta < 1$. Now, in addition, pick $\beta_0 \in (\alpha,\beta)$ and let

$$\mathcal{U} := h^{2+\beta_0}(\Sigma) \cap \mathfrak{A}\,.$$

LEMMA 2.1. *There exist*

$$P \in C^\infty(\mathcal{U},\mathcal{H}(h^{4+\alpha}(\Sigma),h^\alpha(\Sigma)))\,, \quad F \in C^\infty(\mathcal{U},h^{\beta_0}(\Sigma))\,,$$

such that

$$G(\rho) = P(\rho)\rho + F(\rho), \qquad \rho \in h^{4+\alpha}(\Sigma) \cap \mathfrak{A}\,.$$

Proof. (a) The first step is to define a metric on Σ that lends itself well for computations in local coordinates. We choose a system of local coordinates on Σ, where we can assume that the sets U_l, $1 \le l \le m$, are exactly the domains of the charts. We fix some index $l \in \{1, \dots, m\}$ and we let η be the restriction of the Euclidean metric on $\mathcal{R}_l \subset \mathbb{R}^n$. Now define the pull-back metric

$$g_l := X_l^* \eta \quad \text{on} \quad T(U_l \times (-a,a)) \,.$$

The mapping X_l is exactly translation into the normal direction of U_l, and hence it is easily seen that the metric g_l splits along the fibers of $U_l \times (-a,a)$, i.e., $g_l = w_l(r) + dr \otimes dr$. Here r denotes the coordinate in the normal direction of U_l, and $w_l(r)$ is a metric on the tangent space to $U_l \times \{r\} \equiv U_l$. Given $\rho \in \mathcal{U}$, we set

$$g(\rho) := w(\rho) + dr \otimes dr := g_l|_{(s,\rho(s))} \quad \text{on} \quad T_{(s,\rho(s))}(U_l \times (-a,a)) \,.$$

In particular $w(\rho)$ constitutes a metric on $T(U_l)$ with components $w_{jk}(\rho)$. Furthermore, let $w^*(\rho)$ be the induced metric on the cotangent bundle $T^*(U_l)$, that is, $w^*(\rho)(\xi,\zeta) := w^{jk}(\rho)\xi_j\zeta_k$ for $\xi, \zeta \in T^*(U_l)$, where $w^{jk}(\rho)$ are the entries of the inverse matrix of $[w_{jk}(\rho)]$. Note that the metric $w(\rho)$ is not the same as $\theta_\rho^*\eta$. In particular, $w(\rho)$ does not involve any derivatives of ρ, whereas $\theta_\rho^*\eta$ does. We define $U_{l,\rho} := (U_l, w(\rho))$ and $\Xi_l := (U_l \times (-a,a), g_l)$. As a consequence of the special form (2.1) of $\Phi_{l,\rho}$, we have $\widehat{\Phi}_{l,\rho}(s,r) := \Phi_{l,\rho}(X_l(s,r)) = r - \rho(s)$ on $\mathcal{R}_l$, and hence

$$\nabla_{\Xi_l} \widehat{\Phi}_{l,\rho}(s,r) = \tfrac{\partial}{\partial r} - \nabla_{U_{l,\rho}} \rho(s) \,, \qquad (s,r) \in U_l \times (-a,a) \,.$$

Therefore we get for $s \in U_l$

$$\begin{aligned} L_\rho^2(s) = \; & |\nabla_{\mathbb{R}^n} \Phi_{l,\rho}|^2 \Big|_{x = X_l(s,\rho(s))} = g_l(\nabla_{\Xi_l} \widehat{\Phi}_{l,\rho}, \nabla_{\Xi_l} \widehat{\Phi}_{l,\rho}) \Big|_{(s,\rho(s))} \\ = \; & 1 + w(\rho)(\nabla_{U_{l,\rho}} \rho, \nabla_{U_{l,\rho}} \rho) \Big|_s = 1 + w^*(\rho)(d\rho, d\rho) \Big|_s \,, \end{aligned} \tag{2.3}$$

where $d\rho := \partial_j \rho dx^j \in T^*(\Sigma)$ denotes the exterior differential of any $\rho \in C^1(\Sigma)$. We did not label the metrics $g(\rho)$ and $w(\rho)$ with an index l as they can be defined globally on Σ.

To simplify the notation we set $H_\rho := \theta_\rho^* H_{\Gamma_\rho}$. It is known that the mean curvature operator H_ρ is a second order quasi-linear elliptic operator acting on functions defined on Σ; see, for instance, [18, Lemma 3.1]. Moreover, it follows from the proof of that lemma presented in [18] that

$$H_\rho = P_1(\rho)\rho + F_1(\rho), \qquad \rho \in \mathcal{U} \,.$$

$P_1(\rho)$ and $F_1(\rho)$ are represented in local coordinates by

$$\begin{aligned} P_1(\rho) = \tfrac{1}{(n-1)L_\rho^3} \Big[\; & \big(- L_\rho^2 w^{jk}(\rho) + w^{jl}(\rho) w^{km}(\rho) \partial_l \rho \partial_m \rho \big) \partial_j \partial_k \\ & + \big(L_\rho^2 w^{jk}(\rho) \Gamma_{jk}^i(\rho) + w^{jl}(\rho) w^{ki}(\rho) \Gamma_{jk}^n(\rho) \partial_l \rho \\ & \quad + 2 w^{km}(\rho) \Gamma_{nk}^i(\rho) \partial_m \rho - w^{jl}(\rho) w^{km}(\rho) \Gamma_{jk}^i(\rho) \partial_l \rho \partial_m \rho \big) \partial_i \Big], \\ F_1(\rho) = -\frac{1}{(n-1)L_\rho} & w^{jk}(\rho) \Gamma_{jk}^n(\rho). \end{aligned}$$

Here the summation runs from 1 to $(n-1)$ for all repeated indices. Moreover, Γ_{jk}^i are the Christoffel symbols of the metric g_l and

$$\Gamma_{jk}^i(\rho) := \Gamma_{jk}^i \Big|_{(s,\rho(s))} \quad \text{on} \quad T_{(s,\rho(s))}(\Xi_l) \,.$$

An important observation here is that $w^{jk}(\rho)$ and $\Gamma^i_{jk}(\rho)$ are all independent of the derivatives of ρ, and hence the above equations together with (2.3) give complete information on how derivatives of ρ go into the operators $P_1(\rho)$ and $F_1(\rho)$.

Given $\xi \in T^*(\Sigma)$, let $p_1^\pi(\rho)(\xi)$ denote the symbol of the principal part of $P_1(\rho)$. Then (2.3) and the Cauchy–Schwarz inequality yield

$$\begin{aligned} p_1^\pi(\rho)(\xi) &= \frac{1}{(n-1)L_\rho^3}\big[w^*(\rho)(\xi,\xi) + w^*(\rho)(d\rho,d\rho)w^*(\rho)(\xi,\xi) - (w^*(\rho)(d\rho,\xi))^2\big] \\ &\geq \frac{w^*(\rho)(\xi,\xi)}{(n-1)L_\rho^3} \end{aligned}$$

for any $\xi \in T^*(\Sigma)$.

(b) Let us now turn to the operator $\theta_\rho^*\Delta_{\Gamma_\rho}$. Since θ_ρ is a diffeomorphism between Σ and Γ_ρ, we obtain that

$$\theta_\rho^*\Delta_{\Gamma_\rho} = \Delta_\rho\theta_\rho^*\,, \quad \rho \in \mathcal{U}\,,$$

where Δ_ρ is the Laplace–Beltrami operator on $(\Sigma, \theta_\rho^*\eta)$. Here η is the Euclidean metric on the immersed manifold Γ_ρ and $\theta_\rho^*\eta$ denotes the Riemannian metric that is induced by θ_ρ on the manifold Σ. To simplify the notation we set $\sigma(\rho) := \theta_\rho^*\eta$. Let $\sigma_{jk}(\rho)$ be the components of $\sigma(\rho)$ in local coordinates and let $\sigma^*(\rho)$ be the induced metric on $T^*(\Sigma)$, that is, $\sigma^*(\rho)(\xi,\zeta) := \sigma^{jk}(\rho)\xi_j\zeta_k$ for $\xi,\zeta \in T^*(\Sigma)$. As usual, $\sigma^{jk}(\rho)$ are the entries of the inverse matrix of $[\sigma_{jk}(\rho)]$. Finally, $\gamma^i_{jk}(\rho)$ denote the Christoffel symbols of $\sigma(\rho)$. Using local coordinates, we find

$$\Delta_\rho = \sigma^{jk}(\rho)\big(\partial_j\partial_k - \gamma^i_{jk}(\rho)\partial_i\big)\,, \qquad \rho \in \mathcal{U}\,.$$

(c) Let $\rho \in \mathcal{U}$ be given. Then we define $P^\pi(\rho) \in \mathcal{L}(h^{4+\alpha}(\Sigma), h^\alpha(\Sigma))$ by

$$P^\pi(\rho) := -\frac{1}{(n-1)L_\rho^2}\,\sigma^{rs}(\rho)\big[-L_\rho^2 w^{jk}(\rho) + w^{jl}(\rho)w^{km}(\rho)\partial_l\rho\partial_m\rho\big]\partial_r\partial_s\partial_j\partial_k\,.$$

We show that there exists a mapping $Q(\rho) \in \mathcal{L}(h^{3+\alpha}(\Sigma), h^\alpha(\Sigma))$ such that

$$-L_\rho\Delta_\rho P_1(\rho)\rho - P^\pi(\rho)\rho = Q(\rho)\rho, \quad \rho \in \mathcal{U} \cap h^{4+\alpha}(\Sigma)\,.$$

Using the representations of Δ_ρ and $P_1(\rho)\rho$ in local coordinates we see that fourth order derivatives of ρ can only occur when $\partial_r\partial_s$ falls on $\partial_j\partial_k\rho$, and these terms are collected exactly in the operator $P^\pi(\rho)$. Third order derivatives of ρ can only enter in a linear way. For this recall that $w^{jk}(\rho)$ and $\Gamma^i_{jk}(\rho)$ do depend on ρ, but not on its derivatives. So $\partial_r\partial_s$ applied to these functions will only generate second order derivatives of ρ. Hence

$$\partial_r\partial_s(w^{jl}(\rho)w^{km}(\rho)\partial_l\rho\partial_m\rho)$$

will, for instance, produce a third order derivative of ρ exactly when $\partial_r\partial_s$ falls on $\partial_l\rho$ or on $\partial_m\rho$. The result is clearly linear in the third order derivatives. Next observe that L_ρ is represented in local coordinates by

$$L_\rho = \sqrt{1 + w^{jk}(\rho)\partial_j\rho\partial_k\rho}\,,$$

see (2.3). It is then easily seen that $\partial_r\partial_s L_\rho^{-1}$ generates, once again, third order derivatives which enter linearly. Similar arguments apply to all of the remaining

terms. Finally, in case that an expression does not contain third order derivatives we can always split off a linear term $\partial_j\partial_k\rho$ or $\partial_i\rho$. By similar arguments we also conclude that there exists a mapping $R(\rho) \in \mathcal{L}(h^{3+\alpha}(\Sigma), h^\alpha(\Sigma))$ such that

$$-L_\rho \left(\Delta_\rho \frac{1}{L_\rho}\right) L_\rho F_1(\rho) = R(\rho)\rho\,, \quad \rho \in \mathcal{U} \cap h^{3+\alpha}(\Sigma)\,.$$

We set

$$\begin{aligned} P(\rho) &:= P^\pi(\rho) + Q(\rho) + R(\rho), \quad \rho \in \mathcal{U}\,, \\ F(\rho) &:= -L_\rho\Delta_\rho F_1(\rho) - R(\rho)\rho, \quad \rho \in \mathcal{U} \cap h^{3+\alpha}(\Sigma)\,. \end{aligned}$$

It follows from the above considerations and from the representation of $F_1(\rho)$ in local coordinates that

$$P \in C^\infty(\mathcal{U}, \mathcal{L}(h^{4+\alpha}(\Sigma), h^\alpha(\Sigma)))\,, \quad F \in C^\infty(\mathcal{U}, h^{\beta_0}(\Sigma))$$

and that $G(\rho) = P(\rho)\rho + F(\rho)$ for $\rho \in h^{4+\alpha}(\Sigma) \cap \mathfrak{A}$.

(d) It remains to show that $P(\rho) \in \mathcal{H}(h^{4+\alpha}(\Sigma), h^\alpha(\Sigma))$ for $\rho \in \mathcal{U}$. Given $\rho \in \mathcal{U}$, let $p^\pi(\rho)$ denote the symbol of $P^\pi(\rho)$. Then the results in steps (a)–(c) yield

$$p^\pi(\rho)(\xi) = L_\rho\, \sigma^*(\rho)(\xi,\xi)\, p_1^\pi(\rho)(\xi) \geq \frac{1}{(n-1)L_\rho^2}\sigma^*(\rho)(\xi,\xi)\, w^*(\rho)(\xi,\xi)$$

for all $\xi \in T^*(\Sigma)$. Hence, for any fixed $\rho \in \mathcal{U}$, the operator $P^\pi(\rho)$ is a uniformly elliptic fourth order operator acting on functions over the compact manifold Σ. Consequently, $-P^\pi(\rho)$ generates a strongly continuous analytic semigroup on $h^\alpha(\Sigma)$, that is, we have that

$$P^\pi(\rho) \in \mathcal{H}(h^{4+\alpha}(\Sigma), h^\alpha(\Sigma)), \quad \rho \in \mathcal{U}\,.$$

Since $Q(\rho)$ and $R(\rho)$ are lower order perturbations, we can now conclude that $-P(\rho)$ generates an analytic semigroup on $h^\alpha(\Sigma)$ as well. □

Now we are in a position to apply the general theory of quasi-linear evolution equations developed by H. Amann providing a unique classical solution of problem (2.2). More precisely, we have the following theorem.

THEOREM 2.2. *Given any $\rho \in \mathcal{V}$, there exists a unique classical solution*

$$\rho \in C([0,t^+), \mathcal{V}) \cap C^\infty((0,t^+), C^\infty(\Sigma))$$

of problem (2.2). *Here, $t^+ := t^+(\rho_0) > 0$ stands for the maximal time of existence. The map $[(t,\rho_0) \mapsto \rho(t,\rho_0)]$ defines a smooth local semiflow on $\mathcal{V}$.*

Proof. Set $E_0 := h^\alpha(\Sigma)$ and $E_1 := h^{4+\alpha}(\Sigma)$ and let

$$E_\theta := (E_0, E_1)^0_{\theta,\infty}, \qquad \theta \in (0,1)\,,$$

denote the continuous interpolation spaces between E_1 and E_0; see [23] or [4]. Next we fix

$$\theta_1 := \frac{2+\beta-\alpha}{4}\,, \quad \theta_0 := \frac{2+\beta_0-\alpha}{4}\,, \quad \theta := \frac{\beta_0-\alpha}{4}\,.$$

Since the little Hölder spaces are stable under continuous interpolation, we get the following identities

$$E_{\theta_1} = h^{2+\beta}(\Sigma)\,, \quad E_{\theta_0} = h^{2+\beta_0}(\Sigma)\,, \quad E_\theta = h^{\beta_0}(\Sigma)\,.$$

Hence Lemma 2.1 and [3, Theorem 12.1] imply that there exists a unique solution in the class

$$C([0,t^+),\mathcal{V})\cap C((0,t^+),h^{4+\alpha}(\Sigma))\cap C^1((0,t^+),h^{\alpha}(\Sigma))\,.$$

The additional regularity in the assertion follows from a bootstrapping argument in the scale of Banach spaces $h^s(\Sigma)$, cf. the proof of [17, Theorem 1]. Moreover, the results in [3, section 12] also show that the map $[(t,\rho_0)\mapsto\rho(t,\rho_0)]$ defines a smooth local semiflow on $\mathcal{V}$. □

3. Global existence. To prove Theorem 1.2, we fix a Euclidean sphere S and set $\Sigma=S$ in the construction of section 2. Without loss of generality we may assume that S is the unit sphere centered at 0. Observe that Lemma 2.1 implies that

$$G\ :\ \mathcal{U}\cap h^{4+\alpha}(S)\to h^{\alpha}(S)\,,\qquad \rho\mapsto G(\rho)$$

is smooth. Let $A:=\partial G(0)$ be the Fréchet derivative of G at 0. Then we have the following representation of A.

LEMMA 3.1.

$$A=\frac{1}{n-1}\Delta_S^2+\Delta_S\,,$$

where Δ_S denotes the Laplace–Beltrami operator on S.

Proof. Recall that $G(\rho)=-L_\rho\Delta_\rho H_\rho$ for $\rho\in\mathcal{U}\cap h^{4+\alpha}(S)$. Thus we get

$$Ah=\partial G(0)h=-\partial\big(L_\rho\Delta_\rho\big)\big|_{\rho=0}[h,H_0]-L_0\Delta_0\partial H_\rho\big|_{\rho=0}h \tag{3.1}$$

for $h\in h^{4+\alpha}(S)$. Furthermore, observe that

$$L_0\equiv 1,\qquad \Delta_0=\Delta_S,\qquad H_0\equiv 1\,. \tag{3.2}$$

In particular, given $\rho\in\mathcal{U}$, we have that $L_\rho\Delta_\rho H_0=0$. Hence

$$\partial\big(L_\rho\Delta_\rho\big)\big|_{\rho=0}[h,H_0]=\frac{d}{d\varepsilon}\big(L_{\varepsilon h}\Delta_{\varepsilon h}\big)\big|_{\varepsilon=0}H_0=0\,. \tag{3.3}$$

Finally, it was shown in [18, Lemma 3.1] that

$$\partial H_\rho\big|_{\rho=0}h=-\frac{1}{n-1}\big(n-1+\Delta_S\big)h, \tag{3.4}$$

and the assertion follows from (3.1)–(3.4). □

LEMMA 3.2. *The spectrum of $-A$ consists of a sequence of real eigenvalues*

$$\cdots<\mu_{k+1}<\mu_k<\mu_{k-1}<\cdots<\mu_1<\mu_0=0\,.$$

In addition, μ_0 is an eigenvalue of geometric multiplicity $(n+1)$.

Proof. (a) Due to the compact embedding of $h^{4+\alpha}(S)$ in $h^{\alpha}(S)$ it is clear that the spectrum of $-A$ consists only of eigenvalues.

(b) Assume that

$$Ah = \frac{1}{n-1}\Delta_S(n-1+\Delta_S)h = 0$$

for some $h \in h^{4+\alpha}(S)$. Then

$$(n-1+\Delta_S)h = c \tag{3.5}$$

for some constant c. Observe that $g_0 = c/(n-1)$ is a solution of (3.5). Consequently, we find that

$$(n-1+\Delta_S)(h-g_0) = 0\,.$$

On the other hand it is well known that $(n-1)$ is an eigenvalue of $-\Delta_S$ of multiplicity n and that the spherical harmonics $\{Y_k; 1 \le k \le n\}$ of degree 1 span the corresponding eigenspace. Let $Y_0 = 1$ and set $N := \operatorname{span}\{Y_k\,;\, 0 \le k \le n\}$. We have shown that 0 is an eigenvalue of A of geometric multiplicity $(n+1)$ with eigenspace N.

(c) Suppose that $\lambda \in \mathbb{C} \setminus \{0\}$ and $h \in h^{4+\alpha}(S)$ satisfy the equation $(\lambda + A)h = 0$. Then h belongs to $N^\perp$, where the orthogonal complement has to be taken in $L_2(S)$. Indeed, given $k \in \{0, \dots, n\}$, we have that

$$0 = \big((\lambda + A)h|Y_k\big) = \lambda\,(h|Y_k)\,,$$

showing that $h \in N^\perp$. Next observe that there are positive constants c_1 and c_2 such that

$$\big((\Delta_S)^{-1}g|g\big) \le -c_1\big(g|g\big), \qquad \big((n-1+\Delta_S)g|g\big) \le -c_2\big(g|g\big)$$

for all $g \in h^{2+\alpha}(S) \cap N^\perp$, where $(\,\cdot\,|\,\cdot\,)$ denotes the inner product in $L_2(S)$. Now, multiplying the equation $(\lambda + A)h = 0$ in $L_2(S)$ with $(\Delta_S)^{-1}h$, we get

$$\lambda\,(h|(\Delta_S)^{-1}h) + \frac{1}{n-1}\big((n-1+\Delta_S)h|h\big) = 0\,.$$

It follows that $\lambda < 0$ and this completes the proof. □

We are now ready to prove our Theorem 1.2. Here we follow [18].

(i) In a first step we sketch the construction of a center manifold $\mathcal{M}^c$ over N. For $g \in h^r(S)$, $r > 0$, let $Pg := \sum_{k=0}^n (g|Y_k)Y_k$. Then it is easily verified that P is a continuous projection of $h^r(S)$ onto $N = \{Y_k\,;\, 0 \le k \le n\}$, the kernel of the operator A. Moreover, P commutes with A, that is, $PAg = APg = 0$ for all $g \in h^{4+\alpha}(S)$. Therefore, N and $h_s^{4+\alpha}(S) := \ker(P)$ provide topologically complementary subspaces of $h^{4+\alpha}(S)$, which reduce the operator A. We conclude that $\sigma(-\pi^c A) = \{0\}$ and $\sigma(-\pi^s A) \subset (-\infty, \mu_1]$ with $\mu_1 < 0$, where $\pi^c = P$ and $\pi^s = id - P$ denote the projections onto N and $h_s^{4+\alpha}(S)$, respectively, the center subspace and the stable subspace of $-A$. It is now clear that the eigenvalue 0 also has algebraic multiplicity $(n+1)$. We can now apply [29, Theorem 4.1]; see also [23, Theorem 9.2.2]. These results imply that, given $m \in \mathbb{N}^*$, there exists an open neighborhood U of 0 in N and a mapping

$$\gamma \in C^m(U, h_s^{4+\alpha}(S)) \quad \text{with} \quad \gamma(0) = 0,\ \ \partial\gamma(0) = 0$$

such that $\mathcal{M}^c := \operatorname{graph}(\gamma)$ is a locally invariant manifold for the semiflow generated by the solutions of (2.2). $\mathcal{M}^c$ is an $(n+1)$-dimensional submanifold of $h^{4+\alpha}(S)$

with $T_0(\mathcal{M}^c) = N$. In addition, the manifold $\mathcal{M}^c$ is exponentially attractive. More precisely, it follows from [29, Theorem 5.8] that given $\omega \in (0, -\mu_1)$, there exist a positive constant c and a neighborhood W of 0 in $h^{2+\beta}(S)$ such that

$$\|\pi^s\rho(t,\rho_0) - \gamma(\pi^c\rho(t,\rho_0))\|_{h^{4+\alpha}(S)} \le \frac{c}{t^{1-\theta}} e^{-\omega t}\|\pi^s\rho_0 - \gamma(\pi^c\rho_0)\|_{h^{2+\beta}(S)} \tag{3.6}$$

for each ρ_0 in W. Estimate (3.6) is valid for all $t \in (0, t^+(\rho_0))$ with $\pi^c\rho(t,\rho_0) \in U$. Moreover, $\theta := (2+\beta-\alpha)/4$.

(ii) Step (i) implies that $\mathcal{M}^c$ contains all small equilibria of (2.2). We show that $\mathcal{M}^c$ and $\mathcal{M}$ coincide near 0. Suppose that S' is a sphere which is sufficiently close to S. Let $(z_1, \dots, z_n)$ be the coordinates of its center and r be its radius. Recall that S is the unit sphere in $\mathbb{R}^n$ and let $z_0 := 1 - r$. If ρ measures the distance from S to S' in normal direction with respect to S, we get the identity

$$(1+z_0)^2 = \sum_{k=1}^n \big((1+\rho)Y_k - z_k\big)^2. \tag{3.7}$$

Here we used that the spherical harmonics Y_k, $k = 1, \dots, n$, are just the restrictions of the harmonic polynomials $[x \mapsto x_k]$. Solving (3.7) for ρ we obtain that S' can be parameterized over S by the distance function

$$\rho(z) = \sum_{k=1}^n z_k Y_k - 1 + \sqrt{\left(\sum_{k=1}^n z_k Y_k\right)^2 + (1+z_0)^2 - \sum_{k=1}^n z_k^2}\,, \tag{3.8}$$

where $z := (z_0, \dots, z_n) \in \mathbb{R}^{n+1}$. If O is a sufficiently small neighborhood of 0 in $\mathbb{R}^{n+1}$, then it is clear that any sphere S' which is close to S can be characterized by (3.8) with $z \in O$. Furthermore, the mapping $[z \mapsto \rho(z)] : O \to h^{4+\alpha}(S)$ is smooth and its derivative at 0 is given by

$$\partial\rho(0)h = \sum_{k=0}^n h_k Y_k, \qquad h \in \mathbb{R}^{n+1}\,. \tag{3.9}$$

Now, let $\{F_0(z), \dots, F_n(z)\}$ be the coordinates of $\pi^c\rho(z)$ with respect to the basis $\{Y_0, \dots, Y_n\}$ of N. Then (3.9) yields that $\partial F(0) = \mathrm{id}_{\mathbb{R}^{n+1}}$. Consequently, the inverse function theorem implies that F is a smooth diffeomorphism from O onto its image $V := \mathrm{im}(F)$, provided O is small enough. Let $\mathcal{M} := \{\rho(z)\,;\, z \in O\}$. Then it follows that $\pi^c\mathcal{M}$ is an open neighborhood of 0 in N which can be assumed to coincide with the open neighborhood U of 0 in N obtained in step (i). Hence we conclude that $\mathcal{M} = \mathcal{M}^c$.

(iii) It follows from step (ii) that the reduced flow of (2.2) on $\mathcal{M}^c$ consists of equilibria. Therefore, 0 is a stable equilibrium for the reduced flow and we conclude that 0 is also stable for the evolution equation (2.2); see [28, Theorem 3.3]. In particular, there exists a neighborhood W of 0 in $h^{2+\beta}(S)$ such that solutions of (2.2) exist globally for every initial value $\rho_0 \in W$ and such that estimate (3.6) is satisfied for all $t > 0$.

(iv) As in [18, Theorems 6.5 and 6.6] one can show the following result. Given $k \in \mathbb{N}$ and $\omega \in (0, -\mu_1)$, there exists a neighborhood $W = W(k,\omega)$ of 0 in $h^{2+\beta}(S)$

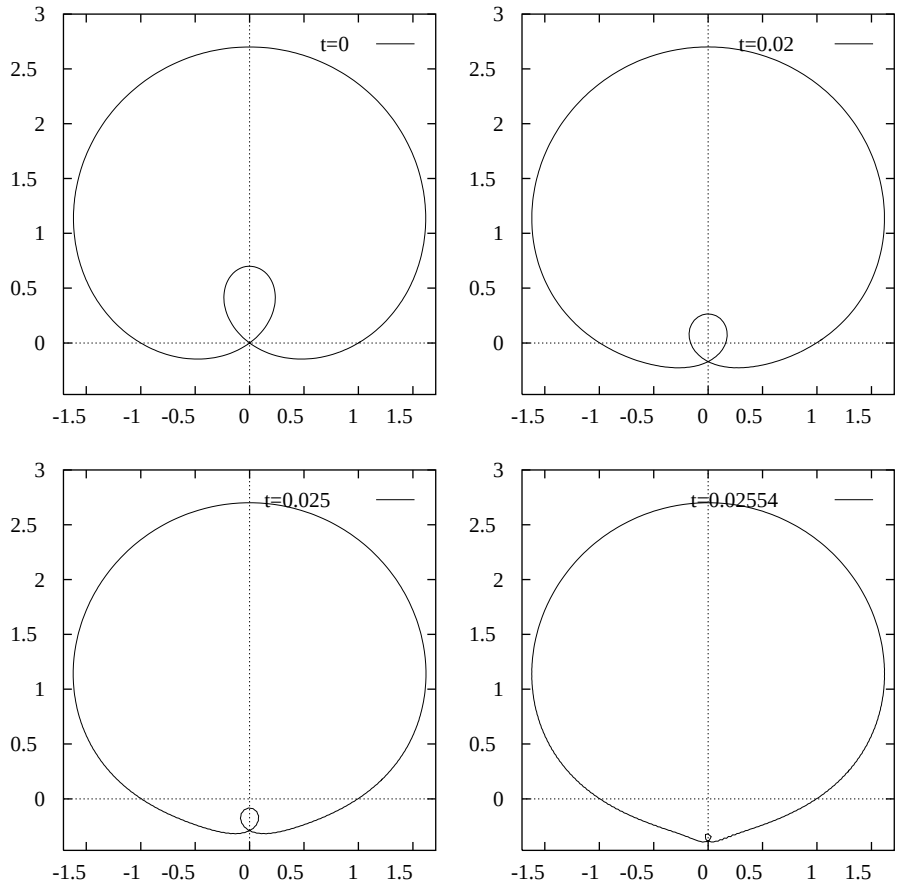

FIG. 1. *The limaçon* $r(\theta) = 1 + 1.7\sin(\theta)$.

with the following property. Given $\rho_0 \in W$, the solution $\rho(\cdot, \rho_0)$ of (2.2) exists globally and there exist $c = c(k, \omega) > 0$ and a unique $z_0 = z_0(\rho_0) \in U$ such that

$$\|\big(\pi^c\rho(t,\rho_0), \pi^s\rho(t,\rho_0)\big) - \big(z_0, \gamma(z_0)\big)\|_{C^k(S)} \le ce^{-\omega t}\|\pi^s\rho_0 - \gamma(\pi^c\rho_0)\|_{h^{2+\beta}(S)}$$

for $t \ge 1$. According to step (ii), $(z_0, \gamma(z_0)) \in \mathcal{M}^c$ is a sphere. Hence we have proved that given $\rho_0 \in W$ the solution $\rho(t, \rho_0)$ of (2.2) exists globally and converges to the sphere $(z_0, \gamma(z_0))$ exponentially fast in the C^k-topology as $t \to \infty$. And so, the proof of Theorem 1.2 is now completed.

4. Numerical simulations. The general theory from the previous sections can be used to set up a numerical scheme as well. The idea is to discretize in time and to use an implicit scheme for stability reasons. Linearization of the dependence on the next time step leads to a semi-implicit scheme. Discretization of the interface leads then to a front-tracking method. We implement this here for two space dimensions; for three space dimensions, and for further details see [24, 25].

4.1. A limaçon. The example of a limaçon shows that the surface diffusion flow can produce singularities. This is not unlike the mean curvature flow, for which Angenent [6] has investigated the singularities arising from the evolution of this shape.

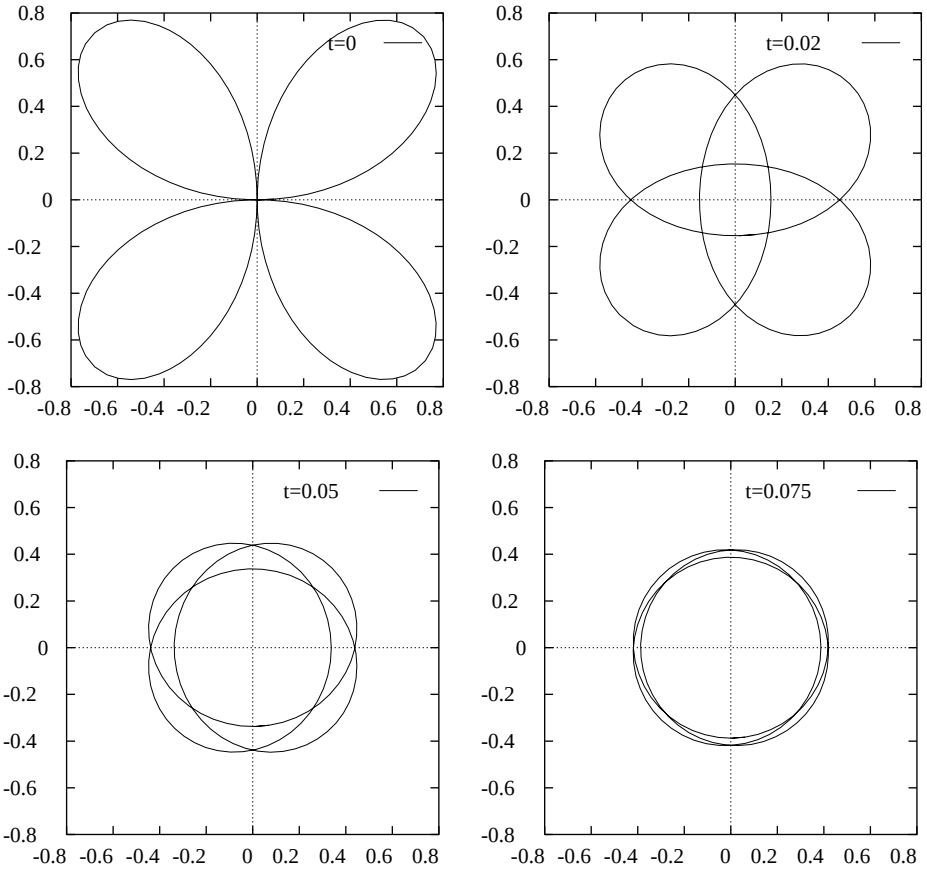

FIG. 2. *The rose* $r(\theta) = \sin(2\theta)$.

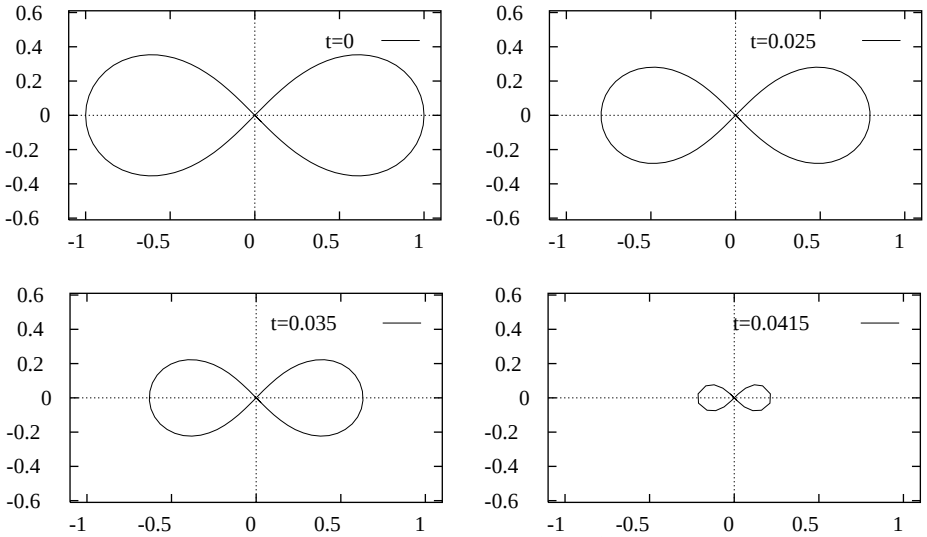

FIG. 3. *The figure eight* $r(\theta) = \sqrt{\cos(\theta)}$.

The smaller loop tightens, having a maximum for the curvature. Therefore the curvature increases for the smaller loop. This leads to a blow-up of the curvature in finite time (Figure 1.)

4.2. A four-leafed rose. The rose exhibits the phenomenon that the stable limiting configuration need not necessarily be an embedded circle, it can also be a multiply covered immersed circle. For positive time the winding number of the curve with respect to the origin does not change, and hence the limiting curve is a triply covered circle (Figure 2.)

4.3. A figure eight. One can make perfect sense of the enclosed signed area of a figure eight, which is for a symmetric figure eight equal to zero. As the evolution decreases the length of the curve and preserves the enclosed area, it can be expected that the limiting figure has zero area and zero length. This is exactly what happens, the figure eight shrinks in finite time to a point. As the curve shortens it is necessary to remove vertices from the numerical simulation to maintain the ratio of temporal versus spatial resolution. In other words, as the length of the curve decreases one needs to remove vertices to maintain a given lower bound on the distance between any two consecutive points of the discretized curve. This is somewhat visible in the last picture where the curve has shrunk so much that because of the increased curvature one can discern faint corners (Figure 3.)

Acknowledgment. We thank the anonymous referee for valuable suggestions.

REFERENCES

[1] A. D. Alexandrov, *Uniqueness theorems for surfaces in the large I*, Vestnik Leningrad Univ. 11, Math. Rev. 19, 167 (1956), pp. 5–7.

[2] N. Alikakos, P.W. Bates, and X. Chen, *Convergence of the Cahn–Hilliard equation to the Hele–Shaw model*, Arch. Rational Mech. Anal., 128 (1994), pp. 164–205.

[3] H. Amann, *Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems*, in Function Spaces, Differential Operators and Nonlinear Analysis, H. J. Schmeisser and H. Triebel, eds., Teubner, Stuttgart, Leipzig, 1993, pp. 9–126.

[4] H. Amann, *Linear and Quasilinear Parabolic Problems,* Vol. I, Birkhäuser, Basel, 1995, Vol. II, III, in preparation.

[5] S. B. Angenent, *Nonlinear analytic semiflows*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 91–107.

[6] S. B. Angenent, *On the formation of singularities in the curve shortening flow*, J. Differential Geom., 33 (1991), pp. 601–633.

[7] P. Baras, J. Duchon, and R. Robert, *Évolution d'une interface par diffusion de surface*, Comm. Partial Differential Equations, 9 (1984), pp. 313–335.

[8] J. W. Cahn, C. M. Elliott, and A. Novick-Cohn, *The Cahn–Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math., 7 (1996), pp. 287–301.

[9] J. W. Cahn and J. E. Taylor, *Surface motion by surface diffusion*, Acta Metall. Mater., 42 (1994), pp. 1045–1063.

[10] J. W. Cahn and J. E. Taylor, *Linking anisotropic sharp and diffuse surface motion laws via gradient flows*, J. Statist. Phys., 77 (1994), pp. 183–197.

[11] G. Da Prato and P. Grisvard, *Équations d'évolution abstraites nonlinéaires de type parabolique*, Ann. Mat. Pura Appl., 120 (1979), pp. 329–396.

[12] G. Da Prato and A. Lunardi, *Stability, instability and center manifold theorem for fully nonlinear autonomous parabolic equations in Banach space*, Arch. Rational Mech. Anal., 101 (1988), pp. 115–144.

[13] F. Davì and M. E. Gurtin, *On the motion of a phase interface by surface diffusion*, Z. Angew. Math. Phys., 41, (1990), pp. 782–811.

[14] C. M. Elliott and H. Garcke, *Existence results for geometric interface models for surface diffusion*, Adv. Math. Sci. Appl., 7 (1997), pp. 467–490.

[15] M. Gage and R. Hamilton, *The shrinking of convex curves by the heat equation*, J. Differential Geom., 23 (1986), pp. 69–96.
[16] J. Escher and G. Simonett, *On Hele-Shaw models with surface tension*, Math. Res. Lett., 3 (1996), pp. 467–474.
[17] J. Escher and G. Simonett, *Classical solutions for Hele-Shaw models with surface tension*, Adv. Differential Equations, 2 (1997), pp. 619–642.
[18] J. Escher and G. Simonett, *A center manifold analysis for the Mullins–Sekerka model*, J. Differential Equations, 143 (1998), pp. 267-292.
[19] P. Fife, *Dynamical aspects of the Cahn–Hilliard equations*, Barret Lectures, University of Tennessee, Knoxville, TN, 1991.
[20] P. Fife, *Seminar notes on curve shortening*, University of Utah, Salt Lake City, UT, 1991.
[21] Y. Giga and K. Ito, *On Pinching of Curves Moved by Surface Diffusion*, preprint, 1997.
[22] H. B. Lawson, *Lectures on Minimal Submanifolds*, Publish or Perish, Berkeley, CA, 1980.
[23] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, 1995.
[24] U. F. Mayer, *A Numerical Scheme for Moving Boundary Problems that are Gradient Flows for the Area Functional*, preprint.
[25] U. F. Mayer, *Numerical Solutions for the Surface Diffusion Flow in Three Space Dimensions*, in preparation.
[26] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.
[27] R. L. Pego, *Front migration in the nonlinear Cahn–Hilliard equation*, Proc. Roy. Soc. London Ser. A, 422 (1989), pp. 261–278.
[28] G. Simonett, *Invariant manifolds and bifurcation for quasilinear reaction-diffusion systems*, Nonlinear Anal., 23 (1994), pp. 515–544.
[29] G. Simonett, *Center manifolds for quasilinear reaction-diffusion systems*, Differential Integral Equations, 8 (1995), pp. 753–796.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1434–1446, November 1998

007

BLOWUP AND LIFE SPAN OF SOLUTIONS FOR A SEMILINEAR PARABOLIC EQUATION*

NORIKO MIZOGUCHI† AND EIJI YANAGIDA‡

Abstract. This paper is concerned with the Cauchy problem

$$\begin{cases} u_t = \Delta u + |u|^{p-1}u & \text{in } \mathbf{R}^N \times (0,\infty), \\ u(x,0) = u_0(x) & \text{in } \mathbf{R}^N, \end{cases}$$

where $p > 1$. Let Ω be a set in $\mathbf{R}^N$ given by

$$\Omega \equiv \left\{ (r,\omega) \in \mathbf{R}^+ \times S^{N-1} \ : \ r > R, \ d(\omega,\omega_0) < cr^{-\mu} \right\}$$

for some $R > 0$, $c > 0$, $\omega_0 \in S^{N-1}$, and $0 \le \mu < 1$, where $d(\cdot,\cdot)$ denotes the standard distance on S^{N-1}. It is shown that if u_0 decays like $|x|^{-\alpha}$ as $|x| \to \infty$ in Ω with $0 < \alpha < 2(1-\mu)/(p-1)$, then the solution blows up in finite time regardless of the behavior of u_0 outside Ω. Moreover the life span of such a solution with $u_0 = \lambda\varphi$ is estimated from above for small $\lambda > 0$ in terms of p, α, and μ. The optimality of these results is also studied.

Key words. blowup, life span, semilinear parabolic equations

AMS subject classifications. 35K15, 35K57

PII. S0036141097324934

1. Introduction.

In this paper, we consider the Cauchy problem for the semilinear parabolic equation

$$\begin{cases} u_t = \Delta u + |u|^{p-1}u & \text{in } \mathbf{R}^N \times (0,\infty), \\ u(x,0) = u_0(x) & \text{in } \mathbf{R}^N, \end{cases} \tag{1.1}$$

where $p > 1$. Since the pioneering work of Fujita [1] on the critical exponent for global existence and blowup, many results concerning the blowup of solutions have been obtained by many authors. Among them, Lee and Ni [7] proved that any nonnegative solution of (1.1) blows up in finite time if the nonnegative initial data u_0 decays more slowly than $|x|^{-2/(p-1)}$ as $|x| \to \infty$. Later, Gui and Wang [5] obtained some related results. See the survey paper by Levine [8] and the book by Samarskii et al. [12] for detailed information on this subject.

On the other hand, it seems that little is known about the blowup of sign-changing solutions. In fact, the methods of the above authors cannot be applied to such solutions. Recently, the result of Fujita was extended in [9, 10] for a one-dimensional problem when the number of sign changes of initial data is prescribed. It was also shown in [10] that any (sign-changing) solution of (1.1) in one-dimensional space blows up in finite time provided that u_0 decays more slowly than $|x|^{-2/(p-1)}$ as $x \to +\infty$ or $-\infty$. This implies that in this case, the number of sign changes of initial data are irrelevant to the blowup or global existence of solutions of (1.1). Since the existence

*Received by the editors July 23, 1997; accepted for publication (in revised form) December 11, 1997; published electronically August 3, 1998.
http://www.siam.org/journals/sima/29-6/32493.html

†Department of Mathematics, Tokyo Gakugei University, Koganei, Tokyo 184, Japan (mizoguti@u-gakugei.ac.jp). This author was partially supported by the Grant-in-Aid for Scientific Research (09440075), Ministry of Education, Science, Sports and Culture.

‡Graduate School of Mathematical Sciences, University of Tokyo, Meguro-ku, Tokyo 153, Japan (yanagida@ms.u-tokyo.ac.jp).

of forward self-similar solutions of (1.1) in $\mathbf{R}^N$ decaying like $|x|^{-2/(p-1)}$ is shown in [6, 11, 13], $|x|^{-2/(p-1)}$ is the critical decay order of initial data for the blowup in the one-dimensional case.

Our first purpose in this paper is to generalize the results of Lee and Ni [7] and obtain a sufficient condition on the decay order of initial data, which may change sign, such that the solution of (1.1) blows up in finite time. Let (r, ω) be the polar coordinate in $\mathbf{R}^N$ with $r > 0$ and $\omega \in S^{N-1}$, and $d(\cdot, \cdot)$ denotes the usual distance on S^{N-1}. We denote by $\Omega = \Omega(R, c, \omega_0, \mu)$ a set in $\mathbf{R}^N$ defined by

$$\Omega \equiv \left\{(r, \omega) \in \mathbf{R}^+ \times S^{N-1} \ : \ r > R, \ d(\omega, \omega_0) < cr^{-\mu}\right\}$$

for some $R > 0$, $c > 0$, $\omega_0 \in S^{N-1}$, and $0 \le \mu < 1$. We note that if $\mu > 0$, Ω can be regarded as a small set in the sense that the Lebesgue measure of the set

$$\{\omega \in S^{N-1} \ : \ (r, \omega) \in \Omega\}$$

tends to zero as $r \to \infty$. We also note that if $\mu = 0$, then Ω is a cone. Let φ be any bounded function on $\mathbf{R}^N$ that decays like $|x|^{-\alpha}$ as $|x| \to \infty$ in Ω with $0 < \alpha < 2(1-\mu)/(p-1)$. We will show that the solution with initial data φ blows up in finite time, regardless of the value of $p > 1$ and the behavior of φ outside Ω. Moreover we will show that the condition on α is sharp if $p \ge (N+1)/(N-1)$.

Our second purpose is to estimate the life span of solutions. We call the maximal existence time of a solution u in the classical sense the life span (or blowup time) of u. We denote by $T(\lambda)$ for $\lambda > 0$ the life span of the solution u of (1.1) with initial data $u_0 = \lambda\varphi$. Under the same assumption on φ as above, we will give an upper estimate of $T(\lambda)$ in terms of α, μ, and p, and study the best possibility of this estimate.

The difficulty in studying the blowup when solutions are allowed to change sign lies in the fact that the standard comparison theorem is not useful, because sign-changing solutions have two directions of blowup, that is, $+\infty$ and $-\infty$. In order to overcome it, the nonincrease (in time) of intersection number of two solutions was essentially used in [10] in one-dimensional problems. However, this tool cannot be applied to higher-dimensional problems. In this paper, we make use of the transformed equation and the corresponding energy which were effectively used by Giga and Kohn [2, 3, 4] to study the backward self-similarity of solutions of (1.1) near a blowup point. We also modify the method of Lee and Ni [7] to study the best possibility.

This paper is organized as follows. In the next section, we introduce precise assumptions on initial data and state our main results. In section 3, we derive a sufficient condition on initial data for the blowup of solutions and obtain an upper estimate of the life span. In section 4, we study the optimality of these results.

2. Main results. We begin this section by introducing precise assumptions. We will impose the following conditions on initial data:

(A1) $\varphi \ge K_1 r^{-\alpha}$ in Ω for some $\alpha > 0$ and $K_1 > 0$.

(A2) $|\nabla\varphi| \le K_2 r^{-\alpha-1+\mu}$ in Ω for some $\alpha > 0$ and $K_2 > 0$.

It should be noted that these conditions impose no restriction on the behavior of φ outside Ω.

In our first result, we show that if the initial data decay slowly only in Ω, then the solution blows up in finite time.

THEOREM 2.1. *Suppose that* $\varphi \in W^{1,\infty}(\mathbf{R}^N)$ *satisfies* (A1) *and* (A2) *with*

$$0 < \alpha < \frac{2(1-\mu)}{p-1}.$$

Then the solution of (1.1) *with initial data* $u_0 = \varphi$ *blows up in finite time.*

For nonnegative initial data, the condition above can be weakened somewhat. In fact, by using the comparison theorem, we can obtain the same result under the assumption that $\varphi \in L^\infty(\mathbf{R}^N)$ is nonnegative in $\mathbf{R}^N$ and satisfies (A1) with $0 < \alpha < 2(1-\mu)/(p-1)$.

It is shown in [6, 11, 13] that there exists a global self-similar solution with radial symmetry that decays like $|x|^{-2/(p-1)}$ as $|x| \to \infty$. Therefore the condition on α in Theorem 2.1 is sharp in case $\mu = 0$.

In the next theorem, we show that if α is large, then there exists a global solution of (1.1) with initial data satisfying (A1) and (A2).

THEOREM 2.2. *Let*

$$\alpha > \begin{cases} \dfrac{2(1-\mu)}{p-1} & \text{if } \ p \geq \dfrac{N+1}{N-1}, \\ \dfrac{2}{p-1} - (N-1)\mu & \text{if } \ 1 < p < \dfrac{N+1}{N-1}. \end{cases}$$

Then there exists $\varphi \in W^{1,\infty}(\mathbf{R}^N)$ *satisfying* (A1) *and* (A2) *such that the solution of* (1.1) *with initial data* $u_0 = \varphi$ *exists globally in time.*

We note that the above theorem implies that if $p \geq (N+1)/(N-1)$, then the condition on α in Theorem 2.1 is sharp. However, since

$$\frac{2(1-\mu)}{p-1} < \frac{2}{p-1} - (N-1)\mu$$

if $1 < p < (N+1)/(N-1)$, there is a gap between the ranges of α in Theorems 2.1 and 2.2. It seems that this gap comes from a technical reason (see Remark 4.1).

Next we consider the life span of solutions of (1.1) with initial data $u_0 = \lambda\varphi$.

THEOREM 2.3. *Suppose that* $\varphi \in W^{1,\infty}(\mathbf{R}^N)$ *satisfies* (A1) *and* (A2) *with*

$$0 < \alpha < \frac{2(1-\mu)}{p-1}.$$

Then for any $\varepsilon > 0$, *there exists* $\lambda_\varepsilon > 0$ *such that*

$$T(\lambda) \leq \lambda^{-\left(\frac{1}{p-1} - \frac{\alpha}{2(1-\mu)}\right)^{-1} - \varepsilon} \tag{2.1}$$

for $0 < \lambda \leq \lambda_\varepsilon$.

As noted above, the condition can be weakened somewhat for nonnegative initial data.

In the next result, we provide some lower bounds that complement (2.1) in Theorem 2.3.

THEOREM 2.4. *Let*

$$0 < \alpha < \frac{2(1-\mu)}{p-1}.$$

Then there exists $\varphi \in W^{1,\infty}(\mathbf{R}^N)$ *satisfying* (A1) *and* (A2) *such that for any* $\varepsilon > 0$, $T(\lambda)$ *satisfies*

$$T(\lambda) \geq \begin{cases} \lambda^{-\left(\frac{1}{p-1} - \frac{\alpha}{2(1-\mu)}\right)^{-1} + \varepsilon} & \text{if } \ \alpha \leq (1-\mu)(N-1), \\ \lambda^{-\left(\frac{1}{p-1} - \frac{\alpha+(N-1)\mu}{2}\right)^{-1} + \varepsilon} & \text{if } \ \alpha > (1-\mu)(N-1) \end{cases}$$

for $0 < \lambda \le \lambda_\varepsilon$ *with some* $\lambda_\varepsilon > 0$.

We note that if $p \ge (N+1)/(N-1)$ and $0 < \alpha < 2(1-\mu)/(p-1)$, then

$$\alpha < \frac{2(1-\mu)}{p-1} \le (1-\mu)(N-1).$$

This implies that if $p \ge (N+1)/(N-1)$, then the estimate of $T(\lambda)$ in Theorem 2.3 is sharp. However, if $1 < p < (N+1)/(N-1)$ and $\alpha > (1-\mu)(N-1)$, there is a gap between the exponents in Theorems 2.3 and 2.4. Again the authors conjecture that the gap comes from a technical reason, and the estimate of $T(\lambda)$ in Theorem 2.3 is optimal (see Remark 4.1).

Finally we remark that a sharp estimate from below for positive solutions was obtained in [7], but one cannot expect a similar one in our situation. Indeed, the solution can blow up independently of the assumption in Ω, because no restriction is imposed on the behavior of initial data outside Ω.

3. Blowup and life span. Let $\tau > 0$ and $a \in \mathbf{R}^N$ be fixed. For a solution u of (1.1), we set

$$w(y,s) = (\tau - t)^{1/(p-1)} u(x,t)$$

with

$$y = (\tau - t)^{-1/2}(x - a), \quad s = -\log(\tau - t).$$

Then (1.1) is written as

$$\text{(3.1)} \qquad \begin{cases} w_s = \Delta w - \dfrac{y}{2}\nabla w - \dfrac{1}{p-1} w + |w|^{p-1} w & \text{in } \mathbf{R}^N \times (s_0, \infty), \\ w(y, s_0) = \tau^{1/(p-1)} u_0(\tau^{1/2} y + a) & \text{in } \mathbf{R}^N, \end{cases}$$

where $s_0 = -\log \tau$. We define

$$I_{a,\tau}(u_0) \equiv \int_{\mathbf{R}^N} \left\{ \frac{1}{2}|\nabla u_0|^2 - \frac{1}{p+1}|u_0|^{p+1} \right\} \cdot \exp\left(-\frac{|x-a|^2}{4\tau} \right) dx + \tau^{-1} \int_{\mathbf{R}^N} \frac{1}{2(p-1)} u_0^2 \cdot \exp\left(-\frac{|x-a|^2}{4\tau} \right) dx.$$

The following lemma is obtained by a minor modification of the argument in [3].

LEMMA 3.1. *If* $I_{a,\tau}(u_0) < 0$ *for some* $a \in \mathbf{R}^N$ *and* $\tau > 0$, *then the solution of* (1.1) *blows up at some* $t < \tau$.

Proof. We define the energy associated with (3.1) by

$$E(w) \equiv \int_{\mathbf{R}^N} \left\{ \frac{1}{2}|\nabla w|^2 + \frac{1}{2(p-1)} w^2 - \frac{1}{p+1}|w|^{p+1} \right\} \rho dy,$$

where

$$\rho(y) = \exp\left(-\frac{|y|^2}{4} \right).$$

Multiplying (3.1) by $w_s \rho$ and integrating it by parts, we obtain

$$\frac{d}{ds} E(w(y,s)) = -\int_{\mathbf{R}^N} w_s^2 \rho dy \le 0.$$

Hence $E(w(y,s))$ is nonincreasing in s. Multiplying (3.1) by $w\rho$ and integrating it by parts, we get

$$\begin{aligned}\frac{1}{2}\frac{d}{ds}\int_{\mathbf{R}^N} w(y,s)^2\rho dy &= -2E(w(y,s)) + \frac{p-1}{p+1}\int_{\mathbf{R}^N} |w(y,s)|^{p+1}\rho dy \\ &\geq -2E(w(y,s_0)) + \frac{p-1}{p+1}\int_{\mathbf{R}^N} |w(y,s)|^{p+1}\rho dy.\end{aligned}$$

Assume here $E(w(y,s_0)) < 0$. Then, by Jensen's inequality, there is $C_1 > 0$ such that

$$\frac{1}{2}\frac{d}{ds}\int_{\mathbf{R}^N} w(y,s)^2\rho dy \geq C_1\left(\int_{\mathbf{R}^N} w(y,s)^2\rho dy\right)^{(p+1)/2}.$$

Hence $\int_{\mathbf{R}^N} w(y,s)^2\rho dy$ diverges to ∞ at some finite s, which implies the blowup of $w(y,s)$. Thus, if $E(w(y,s_0)) < 0$, the corresponding solution u of (1.1) blows up at some $t < \tau$. Since

$$E(w(y,s_0)) = \tau^{\frac{2}{p-1}-\frac{N}{2}+1} I_{a,\tau}(u_0),$$

the proof is complete. □

We set

$$\begin{aligned}I_1(u_0) &= \int_{\mathbf{R}^N} |\nabla u_0|^2 \cdot \exp\left(-\frac{|x-a|^2}{4\tau}\right) dx, \\ I_2(u_0) &= \int_{\mathbf{R}^N} |u_0|^{p+1} \cdot \exp\left(-\frac{|x-a|^2}{4\tau}\right) dx, \\ I_3(u_0) &= \int_{\mathbf{R}^N} u_0^2 \cdot \exp\left(-\frac{|x-a|^2}{4\tau}\right) dx.\end{aligned}$$

LEMMA 3.2. *Suppose that $\varphi \in W^{1,\infty}(\mathbf{R}^N)$ satisfies* (A1) *and* (A2) *with $0 < \alpha < 2(1-\mu)/(p-1)$. If $a = (\tau^{\frac{\beta}{1-\mu}}, \omega_0) \in \Omega$ with some $\beta > 1/2$, then the following hold:*

(a) $\lim_{\tau\to\infty} I_1(\varphi)/I_2(\varphi) = 0$.

(b) $\lim_{\tau\to\infty} I_1(\varphi)/\tau^{-1} I_3(\varphi) = 0$.

(c) $\limsup_{\tau\to\infty} \tau^{-\frac{\alpha\beta(p-1)}{1-\mu}} I_3(\varphi)/I_2(\varphi) < \infty$.

Proof. We define

$$\Omega_\beta = \left\{ (r,\omega) \in \mathbf{R}^N \ : \ |r - \tau^{\frac{\beta}{1-\mu}}| \leq \frac{1}{2}\tau^\beta, \ d(\omega,\omega_0) \leq \frac{c}{2}\tau^{\frac{-\mu\beta}{1-\mu}} \right\}.$$

Then $\Omega_\beta \subset \Omega$ for sufficiently large $\tau > 0$. Since

$$\exp\left(-\frac{|x-a|^2}{4\tau}\right) = O\left(\exp\left(-\frac{\tau^{2\beta-1}}{4}\right)\right) \quad \text{in } \mathbf{R}^N\backslash\Omega_\beta$$

as $\tau \to \infty$, the contribution from $\mathbf{R}^N\backslash\Omega_\beta$ is smaller than any power of τ.

By (A1) and (A2), we have

$$\frac{|\nabla\varphi|^2}{|\varphi|^{p+1}} \leq \frac{K_2^2 r^{2(-\alpha-1+\mu)}}{K_1^{p+1} r^{-(p+1)\alpha}} \quad \text{in } \Omega_\beta.$$

Since

$$2(-\alpha-1+\mu) < -(p+1)\alpha$$

by assumption on α, we obtain (a). Similarly we have

$$\frac{\tau|\nabla\varphi|^2}{\varphi^2} \le \frac{\tau K_2^2 r^{2(-\alpha-1+\mu)}}{K_1^2 r^{-2\alpha}} = \frac{\tau K_2^2 r^{2(-1+\mu)}}{K_1^2} \le C_1\tau^{-2\beta+1} \quad \text{in } \Omega_\beta$$

for some $C_1 > 0$ and large $\tau > 0$. Since $\beta > 1/2$, (b) is proved. Finally, we have

$$\frac{\tau^{-\frac{\alpha\beta(p-1)}{1-\mu}}\varphi^2}{|\varphi|^{p+1}} \le \frac{\tau^{-\frac{\alpha\beta(p-1)}{1-\mu}}}{K_1^{p+1} r^{-\alpha(p-1)}} \le C_2 \quad \text{in } \Omega_\beta$$

for some $C_2 > 0$ and large $\tau > 0$. Thus (c) holds. □

We are now in a position to prove Theorems 2.1 and 2.3.

Proof of Theorem 2.1. By assumption on α, we can take β satisfying

$$\frac{1}{2} < \beta < \frac{1-\mu}{\alpha(p-1)}. \tag{3.2}$$

Let $a = (\tau^{\frac{\beta}{1-\mu}}, \omega_0) \in \Omega$. Then, by Lemma 3.2 and the above inequality, $I_1(\varphi)$ and $\tau^{-1}I_3(\varphi)$ are negligible compared with $I_2(\varphi)$. Hence

$$I_{a,\tau}(\varphi) = \frac{1}{2}I_1(\varphi) - \frac{1}{p+1}I_2(\varphi) + \frac{1}{2(p-1)}\tau^{-1}I_3(\varphi) < 0$$

if $\tau > 0$ is sufficiently large. Thus, by virtue of Lemma 3.1, the solution of (1.1) with initial data $u_0 = \varphi$ must blow up in finite time. □

Proof of Theorem 2.3. Let $\overline{u}(t)$ be a spatially homogeneous solution of (1.1) with initial data $\overline{u}(0) = \lambda|\varphi|_{L^\infty}$. Then $\overline{u}(t)$ is explicitly obtained as

$$\overline{u}(t) = \left\{\lambda^{1-p}|\varphi|_{L^\infty}^{1-p} - (p-1)t\right\}^{-1/(p-1)},$$

which blows up at $t = \lambda^{1-p}|\varphi|_{L^\infty}^{1-p}/(p-1)$. By the comparison theorem, the solution of (1.1) with initial data $u_0 = \lambda\varphi$ satisfies

$$-\overline{u}(t) \le u(x,t) \le \overline{u}(t)$$

for $0 \le t < \lambda^{1-p}|\varphi|_{L^\infty}^{1-p}/(p-1)$. Hence we obtain

$$T(\lambda) \ge \frac{\lambda^{1-p}|\varphi|_{L^\infty}^{1-p}}{p-1}$$

for all λ, which implies that $T(\lambda) \to \infty$ as $\lambda \to 0$.

By Lemma 3.1, we have

$$I_{a,T(\lambda)}(\lambda\varphi) = \frac{1}{2}\lambda^2 I_1(\varphi) - \frac{1}{p+1}\lambda^{p+1}I_2(\varphi) + \frac{1}{2(p-1)}\lambda^2 T(\lambda)^{-1}I_3(\varphi) \ge 0$$

for any $a \in \mathbf{R}^N$. (Otherwise the solution must blow up at some $t < T(\lambda)$, contradicting the definition of $T(\lambda)$.) Let $a = (T(\lambda)^{\frac{\beta}{1-\mu}}, \omega_0) \in \Omega$ with β as in (3.2). Since

$T(\lambda) \to \infty$ as $\lambda \to 0$, it follows from Lemma 3.2 (b) that $I_1(\varphi)$ is negligible compared with $T(\lambda)^{-1} I_3(\varphi)$. Hence, by Lemma 3.2 (c), we obtain

$$\begin{aligned} T(\lambda) &\le \frac{C_1 \lambda^2 I_3(\varphi)}{\lambda^{p+1} I_2(\varphi)} \\ &= \frac{C_1}{\lambda^{p-1} T(\lambda)^{-\frac{\alpha\beta(p-1)}{1-\mu}}} \cdot \frac{T(\lambda)^{-\frac{\alpha\beta(p-1)}{1-\mu}} I_3(\varphi)}{I_2(\varphi)} \\ &\le \frac{C_2}{\lambda^{p-1} T(\lambda)^{-\frac{\alpha\beta(p-1)}{1-\mu}}} \end{aligned}$$

for some positive constants C_1 and C_2. Namely, we have

$$T(\lambda) \le C_3 \lambda^{-\left(\frac{1}{p-1} - \frac{\alpha\beta}{1-\mu}\right)^{-1}}$$

for some $C_3 > 0$ and small $\lambda > 0$. Since $\beta > 1/2$ can be chosen arbitrarily close to $1/2$, the proof is complete. □

4. Optimality of blowup results. This section is devoted to proofs of Theorems 2.2 and 2.4. To do that, we employ the technique of Lee and Ni [7], who constructed and effectively used positive supersolutions with radial symmetry. Since we are concerned with solutions that are not necessarily positive, we need to modify their method.

We rewrite $x = (x_1, x_2, \dots, x_N) \in \mathbf{R}^N$ and $y = (y_1, y_2, \dots, y_N) \in \mathbf{R}^N$ as

$$x = (\xi \cos\theta, \xi \sin\theta, x'), \quad y = (\eta \cos\kappa, \eta \sin\kappa, y'),$$

where x', $y' \in \mathbf{R}^{N-2}$. We put

$$\theta_0 = \pi/m$$

for some positive integer m, and define a domain D by

$$D = \left\{ x = (\xi \cos\theta, \xi \sin\theta, x') \ : \ \xi > 0, \ -\frac{\theta_0}{2} < \theta < \frac{\theta_0}{2} \right\}.$$

Let $G(x, y, t)$ be a fundamental solution of the heat equation on D with Dirichlet condition on ∂D. It is known that $G(x, y, t) = G(y, x, t) > 0$ on D. Set

$$g(x, y, t) = \frac{1}{(4\pi t)^{N/2}} \exp\left(-\frac{|x - y|^2}{4t} \right).$$

Since $g(x, y, t)$ is a fundamental solution of the heat equation on $\mathbf{R}^N$, it follows from the comparison theorem that

$$G(x, y, t) < \frac{1}{(4\pi t)^{N/2}} \exp\left(-\frac{|x - y|^2}{4t} \right) \tag{4.1}$$

for $(x, y, t) \in D \times D \times (0, \infty)$. Moreover the following estimate is obtained.

LEMMA 4.1. *For any positive integer m, there exists a constant $C > 0$ such that*

$$0 < G(x, y, t) \le C t^{-\frac{N}{2} - m} |x|^m |y|^m$$

for $(x, y, t) \in D \times D \times (0, \infty)$.

Proof. For $y = (\eta \cos\kappa, \eta \sin\kappa, y') \in \mathbf{R}^N$, we put

$$\kappa_j^+ = \kappa + 2(j-1)\theta_0, \quad \kappa_j^- = -\kappa + (2j-1)\theta_0, \quad j = 1, 2, \ldots, m,$$

and define $2m$ points in $\mathbf{R}^N$ by

$$y_j^+ = (\eta \cos\kappa_j^+, \eta \sin\kappa_j^+, y'), \quad y_j^- = (\eta \cos\kappa_j^-, \eta \sin\kappa_j^-, y').$$

Then $G(x, y, t)$ is explicitly written by the superposition of g as

$$G(x, y, t) = \sum_{j=1}^{m} g(x, y_j^+, t) - \sum_{j=1}^{m} g(x, y_j^-, t).$$

Set

$$\begin{aligned} h(x, y, t) &\equiv \frac{1}{g(x, y, t)} \cdot \left\{ \sum_{j=1}^{m} g(x, y_j^+, t) - \sum_{j=1}^{m} g(x, y_j^-, t) \right\} \\ &= \sum_{j=1}^{m} \exp\left(\frac{\langle\, x, y_j^+ \,\rangle - \langle\, x, y \,\rangle}{2t} \right) - \sum_{j=1}^{m} \exp\left(\frac{\langle\, x, y_j^- \,\rangle - \langle\, x, y \,\rangle}{2t} \right), \end{aligned}$$

where $\langle\, \cdot, \cdot \,\rangle$ denotes the usual inner product of two vectors in $\mathbf{R}^N$. It is easy to see that

$$h(t^{-1/2}x, t^{-1/2}y, 1) \equiv h(x, y, t).$$

Hence, if there exists a constant $C_1 > 0$ such that

$$h(x, y, 1) \le C_1 |x|^m |y|^m \tag{4.2}$$

for $(x, y) \in \mathbf{R}^N \times \mathbf{R}^N$, then

$$\begin{aligned} G(x, y, t) &= g(x, y, t) h(x, y, t) \\ &= g(x, y, t) h(t^{-1/2}x, t^{-1/2}y, 1) \\ &\le \frac{1}{(4\pi t)^{N/2}} \cdot C_1 |t^{-1/2}x|^m |t^{-1/2}y|^m \quad \text{in } D, \end{aligned}$$

which proves the lemma.

Let us prove (4.2). Since $h(\gamma x, \gamma^{-1} y, t) \equiv h(x, y, t)$ for any $\gamma \neq 0$, it suffices to show that (4.2) holds for any $|x| > 1$ and $|y| = 1$. By the location of the points y_j^+ and y_j^-, we have $h(x, y, t) = 0$ if x is on the plane $\theta = (j - 1/2)\theta_0$ for some j. Namely, $h(x, y, t) = 0$ if $a_j x_1 + b_j x_2 = 0$, where

$$a_j = \sin\left(j - \frac{1}{2}\right)\theta_0, \quad b_j = -\cos\left(j - \frac{1}{2}\right)\theta_0, \quad j = 1, 2, \ldots, m.$$

Since $h(x, y, t)$ is analytic in x and y and $h(x, y, t) \equiv h(y, x, t)$, we can write $h(x, y, 1)$ as

$$h(x, y, 1) = \left\{ \prod_{j=1}^{m} (a_j x_1 + b_j x_2) \right\} \left\{ \prod_{j=1}^{m} (a_j y_1 + b_j y_2) \right\} h_0(x, y),$$

where $h_0(x,y)$ is analytic in $(x,y) \in \mathbf{R} \times \mathbf{R}$. Thus, if we take

$$C_1 > \sup \{ \, |h_0(x,y)| \ : \ |x| \le 1, \ |y| = 1\},$$

then (4.2) holds for $|x| \le 1$ and $|y| = 1$, because $|a_j x_1 + b_j x_2| \le |x|$ and $|a_j y_1 + b_j y_2| \le |y|$.

It remains to show (4.2) in the case where $|x| > 1$ and $|y| = 1$. Since

$$\begin{aligned} \langle \, x, y_j^+ \, \rangle - \langle \, x, y \, \rangle &= \xi \cos\theta \cdot \eta \cos\kappa_j^+ + \xi \sin\theta \cdot \eta \sin\kappa_j^+ \\ &\quad -\xi \cos\theta \cdot \eta \cos\kappa - \xi \sin\theta \cdot \eta \sin\kappa \\ &= \xi\eta \, \{ \cos(\theta - \kappa_j^+) - \cos(\theta - \kappa) \, \} \\ &= -2\xi\eta \sin(\theta - \kappa - (j-1)\theta_0) \sin(-(j-1)\theta_0) \\ &\le 0, \end{aligned}$$

we have

$$h(x,y,1) = \sum_{j=1}^m \exp\left(\frac{\langle \, x, y_j^+ \, \rangle - \langle \, x, y \, \rangle}{2} \right) - \sum_{j=1}^m \exp\left(\frac{\langle \, x, y_j^- \, \rangle - \langle \, x, y \, \rangle}{2} \right) \le m.$$

Hence, if we take $C_1 \ge m$, then (4.2) holds for $|x| > 1$ and $|y| = 1$. Thus the proof is complete. □

We assume $\omega_0 = (1, 0, \dots, 0)$ and choose a nonnegative function $\psi \in W_0^{1,\infty}(D)$ such that

(i) $\psi(r,\omega)$ satisfies (A1), (A2) and $\psi(r,\omega) \le K_3 r^{-\alpha}$ in Ω for some K_3;
(ii) $\psi(r,\omega)$ is rotationally symmetric with respect to ω_0;
(iii) $\psi(r,\omega)$ is nonincreasing in $d(\omega,\omega_0) \in [0, 2cr^{-\mu}]$ for each $r > R/2$;
(iv) $\psi(r,\omega) \equiv 0$ if $r \le R/2$ or $d(\omega,\omega_0) \ge 2cr^{-\mu}$.

It is easy to verify that such a function $\psi(r,\omega)$ exists.

Now we consider the following auxiliary problems:

$$\text{(4.3)} \qquad \begin{cases} w_t = \Delta w + |w|^{p-1} w & \text{in } D \times (0,\infty), \\ w(x,t) = 0 & \text{on } \partial D \times (0,\infty), \\ w(x,0) = \lambda\psi(x) & \text{in } D \end{cases}$$

and

$$\text{(4.4)} \qquad \begin{cases} U_t = \Delta U & \text{in } D \times (0,\infty), \\ U(x,t) = 0 & \text{on } \partial D \times (0,\infty), \\ U(x,0) = \psi(x) & \text{in } D. \end{cases}$$

LEMMA 4.2. *Let $\overline{w}(x,t) = h(t)U(x,t)$, where $U(x,t)$ is the solution of* (4.4) *and*

$$h(t) = \left\{ \lambda^{1-p} - (p-1) \int_0^t |U(x,t)|_{L^\infty(D)}^{p-1} dt \right\}^{1/(1-p)}.$$

Then $\overline{w}(x,t)$ is a supersolution of (4.3) *with initial data $u_0 = \lambda\psi$ as long as $h(t)$ is positive.*

Proof. If $h(t) > 0$, then we have

$$\begin{aligned} &\overline{w}_t(x,t) - \Delta w_t(x,t) - w^p \\ &= h_t(t)U(x,t) + h(t)U_t(x,t) - h(t)\Delta U(x,t) - \{h(t)U(x,t)\}^p \\ &= h(t)^p |U(x,t)|_{L^\infty(D)}^{p-1} U(x,t) - \{h(t)U(x,t)\}^p \\ &\ge 0. \end{aligned}$$

Since $\overline{w}(x,t) = 0$ on ∂D, $\overline{w}(x,t)$ is a supersolution. The equality $\overline{w}(x,0) = \lambda\psi(x)$ is obvious. □

In order to apply Lemma 4.2 to our problem, we need an upper estimate of $|U(x,t)|_{L^\infty(D)}$. The next lemma is crucial for the proofs of Theorems 2.2 and 2.4.

LEMMA 4.3. *Let m be a positive integer satisfying*

$$m - \alpha + (1-\mu)(N-1) > 0,$$

and let $U(x,t)$ be the solution of (4.4). Then for any $\varepsilon > 0$, there exists t_0 such that

$$|U(x,t)|_{L^\infty(D)} \le t^{-\frac{N}{2}+\frac{\nu}{2}+\varepsilon}$$

for $t \ge t_0$, where

$$\nu = \max\left\{1 - \alpha + (1-\mu)(N-1),\ -\frac{\alpha}{1-\mu} + N\right\}.$$

Proof. We first note that $U(x,t)$ is given by

$$U(x,t) = \int_D G(x,y,t)\psi(y)dy.$$

We see from (ii) and (iii) that

$$|U(x,t)|_{L^\infty(D)} = \sup_{x\in\Gamma} U(x,t)$$

for $t > 0$, where $\Gamma = \{(r,\omega_0) : r > 0\}$. Therefore it suffices to estimate $U(x,t)$ for $x \in \Gamma$ and $t > 0$.

We take a constant $\beta > 1/2$ arbitrarily, and set

$$D_\beta = \{x \in D \ :\ |x| \le 3t^\beta\}.$$

Let $x \in \Gamma$ with $|x| \le 2t^\beta$. Then, by (4.1), Lemma 4.1, and the assumption on m, we have

$$\begin{aligned}
U(x,t) &\le C\int_{D_\beta} G(x,y,t)\psi(y)dy \\
&\le Ct^{-\frac{N}{2}-m}\int_{D_\beta} |x|^m|y|^m\psi(y)dy \\
&\le Ct^{-\frac{N}{2}-m+m\beta}\int_0^{3t^\beta} r^m(1+r)^{-\alpha}r^{(1-\mu)(N-1)}dr \\
&\le Ct^{-\frac{N}{2}-m+m\beta}\int_0^{3t^\beta} r^{m-\alpha+(1-\mu)(N-1)}dr \\
&= Ct^{-\frac{N}{2}+\beta\{1-\alpha+(1-\mu)(N-1)\}+m(2\beta-1)}
\end{aligned}$$

for sufficiently large $t > 0$. (Here and hereafter, C denotes a generic positive constant which may vary from line to line.)

Next, let $x \in \Gamma$ with $2t^\beta < |x| \le t^{\beta/(1-\mu)}$. Then

$$\frac{1}{2}|x| \le |x| - t^\beta \le |x| + t^\beta \le 2|x|.$$

Hence we have

$$\begin{aligned} U(x,t) &\le Ct^{-\frac{N}{2}} \int_{|x|-t^\beta}^{|x|+t^\beta} (1+r)^{-\alpha} r^{(1-\mu)(N-1)} dr \\ &\le Ct^{-\frac{N}{2}} \int_{|x|-t^\beta}^{|x|+t^\beta} r^{-\alpha+(1-\mu)(N-1)} dr \\ &\le Ct^{-\frac{N}{2}+\beta} |x|^{-\alpha+(1-\mu)(N-1)} \end{aligned}$$

for sufficiently large $t > 0$. Putting $|x| = 2t^\beta$ or $|x| = t^{\beta/(1-\mu)}$ in the right-hand side of the last inequality, $U(x,t)$ is estimated as

$$U(x,t) \le Ct^{-\frac{N}{2}+\beta\nu}.$$

Finally, for $x \in \Gamma$ with $|x| > t^{\beta/(1-\mu)}$ (or $|x|^{1-\mu} > t^\beta$), we have

$$\begin{aligned} U(x,t) &\le Ct^{-\frac{N}{2}} \int_{|x|-t^\beta}^{|x|+t^\beta} r^{-\alpha} t^{\beta(N-1)} dr \\ &\le Ct^{-\frac{N}{2}} |x|^{-\alpha} t^{\beta(N-1)} t^\beta \\ &\le Ct^{-\frac{N}{2}+\beta\left\{-\frac{\alpha}{(1-\mu)}+N\right\}} \end{aligned}$$

for sufficiently large $t > 0$.

Thus $U(x,t)$ is estimated as

$$U(x,t) \le Ct^{-\frac{N}{2}+\beta\nu+m(2\beta-1)}$$

for sufficiently large $t > 0$. Since $\beta > 1/2$ can be arbitrarily close to $1/2$, the proof is complete. □

Now we are in a position to prove Theorems 2.2 and 2.4.

Proof of Theorem 2.2. Without loss of generality, we may assume $\omega_0 = (1, 0, \ldots)$. For $x = (\xi\cos\theta, \xi\sin\theta, x')$ with $\theta \in [-\theta_0/2, \theta_0/2)$, we put

$$\theta_j^+ = \theta + 2(j-1)\theta_0, \quad \theta_j^- = -\theta + (2j-1)\theta_0, \quad j = 1, 2, \ldots, m,$$

and

$$x_j^+ = (\xi\cos\theta_j^+, \xi\sin\theta_j^+, x'), \quad x_j^- = (\xi\cos\theta_j^-, \xi\sin\theta_j^-, x').$$

We define a function $\varphi \in W^{1,\infty}(\mathbf{R}^N)$ by

$$\varphi(x_j^+) = \psi(x), \quad \varphi(x_j^-) = -\psi(x), \quad j = 1, 2, \ldots, m. \tag{4.5}$$

It is clear that $\lambda\varphi$ satisfies (A1) and (A2) for any $\lambda > 0$. Then the solution of (1.1) with initial data $u_0 = \lambda\varphi$ satisfies

$$u(x_j^+, t) = w(x,t), \quad u(x_j^-, t) = -w(x,t), \quad j = 1, 2, \ldots, m,$$

for all $x \in D$ and $t > 0$. Hence u exists globally in time if and only if w exists globally in time.

Let $\overline{w}$ be a supersolution of (4.3) given as in Lemma 4.2. Suppose that

$$\int_0^\infty |U(x,t)|_{L^\infty(D)}^{p-1} dt < \infty.$$

Then, by taking

$$0 < \lambda < \left\{ (p-1) \int_0^\infty |U(x,t)|^{p-1}_{L^\infty(D)} dt \right\}^{1/(1-p)},$$

$h(t)$ is positive for all $t \geq 0$ so that $\overline{w}$ exists globally in time. Then, by the comparison theorem, the solution w of (1.1) with initial data $u_0 = \lambda\varphi$ exists globally.

Therefore, by Lemma 4.3, it suffices to show that

$$\left(-\frac{N}{2} + \frac{\nu}{2} \right) (p-1) < -1.$$

By definition of ν, this inequality holds if and only if

$$\alpha > \frac{2}{p-1} - (N-1)\mu \quad \text{and} \quad \alpha > \frac{2(1-\mu)}{p-1}.$$

Since

$$\frac{2}{p-1} - (N-1)\mu > \frac{2(1-\mu)}{p-1}$$

is equivalent to $p < (N+1)/(N-1)$, the proof is complete. □

Proof of Theorem 2.4. Let φ be given as in (4.5). By Theorem 2.1, the solution u of (1.1) blows up in finite time. Let $\overline{T}(\lambda)$ be the life span of $\overline{w}$ given in Lemma 4.2. Then it follows from the comparison theorem that $\overline{T}(\lambda) \leq T(\lambda) < \infty$. On the other hand, we have

$$(p-1)^{-1}\lambda^{1-p} = \int_0^{\overline{T}(\lambda)} |U(x,t)|^{p-1}_{L^\infty(D)} dt.$$

Thus, making use of Lemma 4.3, we obtain

$$\begin{aligned}
(p-1)^{-1}\lambda^{1-p} &\leq \int_0^{T(\lambda)} |U(x,t)|^{p-1}_{L^\infty(D)} dt \\
&\leq \int_0^{t_0} |U(x,t)|^{p-1}_{L^\infty(D)} dt + \int_{t_0}^{T(\lambda)} t^{\left(-\frac{N}{2}+\frac{\nu}{2}+\varepsilon\right)(p-1)} dt \\
&\leq C_0 + \int_{t_0}^{T(\lambda)} t^{\left(-\frac{N}{2}+\frac{\nu}{2}+\varepsilon\right)(p-1)} dt
\end{aligned}$$

for some positive constants t_0 and C_0. Since

$$\left(-\frac{N}{2} + \frac{\nu}{2} \right) (p-1) > -1$$

by our assumption on α, $T(\lambda)$ satisfies

$$\lambda^{1-p} \leq T(\lambda)^{\left(-\frac{N}{2}+\frac{\nu}{2}+2\varepsilon\right)(p-1)+1}$$

for sufficiently small $\lambda > 0$, or equivalently,

$$T(\lambda) \geq \lambda^{-\left(-\frac{N}{2}+\frac{\nu}{2}+\frac{1}{p-1}\right)^{-1}+\varepsilon}.$$

Since the inequality

$$1-\alpha+(1-\mu)(N-1) \le -\frac{\alpha}{1-\mu}+N$$

is equivalent to

$$\alpha \le (1-\mu)(N-1),$$

the proof is complete. □

Remark 4.1. As we noted in section 2, there is a gap between the ranges of α in Theorems 2.1 and 2.2 if $1 < p < (N+1)/(N-1)$. Also, there is a gap between the exponents of λ in Theorems 2.3 and 2.4 if $1 < p < (N+1)/(N-1)$ and $\alpha > (1-\mu)(N-1)$. We notice that $p = (N+1)/(N-1)$ is Fujita's exponent [1] for the blowup of positive solutions in $\mathbf{R}^{N-1}$. We suspect that the gaps arise from the assumption that ψ is nonnegative in D.

REFERENCES

[1] H. Fujita, *On the blowing up of solutions of the Cauchy problem for* $u_a = \Delta u + u^{1+\alpha}$, J. Fac. Sci. Univ. Tokyo, 13 (1966), pp. 109–124.

[2] Y. Giga and R. V. Kohn, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.

[3] Y. Giga and R. V. Kohn, *Characterizing blowup using similarity variables*, Indiana Univ. Math. J., 36 (1987), pp. 1–40.

[4] Y. Giga and R. V. Kohn, *Nondegeneracy of blowup for semilinear heat equations*, Comm. Pure Appl. Math., 17 (1989), pp. 845–884.

[5] C. Gui and X. Wang, *Life span of solutions of the Cauchy problem for a semilinear heat equation*, J. Differential Equations, 115 (1995), pp. 166–172.

[6] M. Hirose and E. Yanagida, *Global Structure of Self-similar Solutions in a Semilinear Parabolic Equation,* preprint.

[7] T.-Y. Lee and W.-M. Ni, *Global existence, large time behavior and life span of solutions of a semilinear parabolic Cauchy problem*, Trans. Amer. Math. Soc., 333 (1992), pp. 365–378.

[8] H. A. Levine, *The role of critical exponents in blowup theorems*, SIAM Rev., 32 (1990), pp. 262–288.

[9] N. Mizoguchi and E. Yanagida, *Critical exponents for the blow-up of solutions with sign changes in a semilinear parabolic equation*, Math. Ann., 307 (1997), pp. 663–675.

[10] N. Mizoguchi and E. Yanagida, *Critical exponents for the blowup of solutions with sign changes in a semilinear parabolic equation* II, J. Differential Equations, 145 (1998), pp. 295–331.

[11] L. A. Peletier, D. Terman, and F. B. Weissler, *On the equation* $\Delta u + (x\cdot\nabla u)/2 + f(u) = 0$, Arch. Rational Mech. Anal., 121 (1986), pp. 83–99.

[12] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Blow-up in Quasilinear Parabolic Equations*, de Gruyter Exp. Math., 19, Walter de Gruyter & Co., Berlin, 1995.

[13] F. B. Weissler, *Asymptotic analysis of an ordinary differential equation and nonuniqueness for a semilinear partial differential equation*, Arch. Rational Mech. Anal., 91 (1986), pp. 231–245.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1447–1458, November 1998

008

INTERIOR BLOWUP IN A CONVECTION-DIFFUSION EQUATION*

CHRISTOPHER P. GRANT†

Abstract. This paper addresses the qualitative behavior of a nonlinear convection-diffusion equation on a smooth bounded domain in $\mathbf{R}^n$, in which the strength of the convection grows superlinearly as the density increases. While the initial-boundary value problem is guaranteed to have a local-in-time solution for smooth initial data, it is possible for this solution to be extinguished in finite time. We demonstrate that the way this may occur is through finite-time "blow up," i.e., the unboundedness of the solution in arbitrarily small neighborhoods of one or more points in the closure of the spatial domain. In special circumstances, such as the presence of radial symmetry, the set of blowup points can be identified; these points may be either on the boundary or on the interior of the domain. Furthermore, criteria can be established that guarantee that blowup occurs. In this paper, such criteria are presented, involving the dimension of the space, the growth rate of the nonlinearity, the strength of the imposed convection field, the diameter of the domain, and the mass of the initial data. Furthermore, the temporal rate of blowup is estimated.

Key words. blowup, convection-diffusion equations

AMS subject classifications. 35B30, 35B40, 35K20, 35K60

PII. S0036141097327458

1. Introduction.

Let $\Omega \subset \mathbf{R}^n$ be a bounded domain with smooth boundary, and consider the initial-boundary value problem

$$\begin{cases} u_t = \operatorname{div}(\nabla u - u^q \vec{a}), & (x,t) \in \Omega \times (0,T), \\ u_\nu = u^q \vec{a} \cdot \vec{\nu}, & (x,t) \in \partial\Omega \times (0,T), \\ u(x,0) = u_0(x), & x \in \Omega, \end{cases} \tag{1.1}$$

where $\vec{\nu}$ is the unit outward normal on $\partial\Omega$ and u_ν is the corresponding directional derivative, and (for simplicity) the initial data u_0 are assumed to be extendable to a smooth, positive function on $\overline{\Omega}$ that satisfies the boundary conditions.

The unknown quantity $u = u(x,t)$ represents the local density of some material that is evolving in response to diffusion and the effects of an imposed convection field $\vec{a} : \overline{\Omega} \to \mathbf{R}^n$. The exponent q (which we shall call the *exponent of the nonlinearity*) is assumed to be greater than 1, so the magnitude of the convection term is superlinear in u.

Although (1.1) is similar in many respects to equations appearing in the scientific literature (see, e.g., [6] and [10]), the motivation for studying it has less to do with its applicability to any specific physical problem than with its role as an archetype for interacting diffusion and nonlinear convection. In this sense, it plays a role similar to the one that the nonlinear heat equation of Fujita [5],

$$u_t = \Delta u + u^p, \tag{1.2}$$

has come to play for the study of the interaction between diffusion and nonlinear reaction. (For a survey of (1.2) and related equations, see [8].)

*Received by the editors September 17, 1997; accepted for publication December 8, 1997; published electronically August 26, 1998.
http://www.siam.org/journals/sima/29-6/32745.html
†Department of Mathematics, Brigham Young University, Provo, UT 84602 (grant@math.byu.edu).

The Cauchy problem corresponding to (1.1) in the special case that the convection field is constant has been studied in a series of papers [3], [4], [12]. In that context, solutions decay to zero as $t \to \infty$, with the asymptotic profile of the decaying solution depending on the particular value of the exponent q.

The one-dimensional version of (1.1) with constant convection

$$\begin{cases} u_t = (u_x + u^q)_x, & (x,t) \in (0,1) \times (0,T), \\ u_x + u^q = 0, & (x,t) \in \{0,1\} \times (0,T), \\ u(x,0) = u_0(x), & x \in (0,1), \end{cases} \tag{1.3}$$

was studied in [1] and [2]. While solutions to (1.3) stabilize as $t \to \infty$ if $q < 2$, if $q > 2$ the convection is strong enough that its concentrating force, in combination with the obstacle posed by the boundary, can overwhelm the diffusion's homogenizing force on the boundary, and solutions can *blow up* in finite time. By this it is meant that there is a time T such that (1.3) has a classical solution for $t \in (0,T)$, but

$$\limsup_{t \to T} \|u(\cdot,t)\|_\infty = +\infty. \tag{1.4}$$

(Here and throughout this paper $\|\psi\|_p$ will represent the L^p norm on the domain of ψ.) The value of T for which (1.4) holds, if it exists, is called the *blowup time* of the solution.

The purpose of this paper is to address questions about the qualitative behavior of solutions to (1.1): Do solutions exist globally (i.e., for $T = +\infty$)? Can solutions blow up in finite time? At what points in $\overline{\Omega}$ does blowup occur? What can be said about the asymptotic form of u as the blowup time is approached?

Throughout the remainder of this paper, we will confine our attention to the case that $\vec{a}$ is conservative. More precisely, we shall assume the existence of a continuous function $f : \overline{\Omega} \to \mathbf{R}$ such that $\nabla f = \vec{a}$. Note that this case includes the case of constant convection.

The first thing that we will show is that the only way solutions to (1.1) can cease to exist is by blowing up.

THEOREM 1.1. *Given u_0 and sufficiently smooth $\vec{a}$,* (1.1) *has a (unique) positive classical solution for some $T > 0$. Furthermore, if T^* is the supremum of all $T > 0$ for which* (1.1) *has a solution and $T^* < +\infty$, then* $\limsup_{t \to T^*} \|u(\cdot,t)\|_\infty = +\infty$.

In order to obtain more detailed information about the way that blowup can occur, we will focus on the special case of radial symmetry. Consider the following conditions:

(R1) $\Omega = B(0,R) := \{x \in \mathbf{R}^n : |x| < R\}$.

(R2) $\vec{a}(x) = -g(r)x/r$ for some function $g : [0,R] \to [0,\infty)$. (Here and throughout this paper, $r := |x|$.)

(R3) $\vec{a}(x) = g(r)x/r$ for some function $g : [0,R] \to [0,\infty)$.

(R4) $u_0(x) = U_0(r)$.

The following theorem limits the set of possible blowup points for (1.1). (We will call a point $x \in \overline{\Omega}$ a blowup point if there is not a neighborhood N of x in $\overline{\Omega}$ such that u remains bounded on $N \cap \Omega$.)

THEOREM 1.2. *Under conditions* (R1), (R2), *and* (R4), *the only possible blowup point for* (1.1) *is the origin. Under conditions* (R1), (R3), *and* (R4), *all blowup points lie on the sphere $\partial\Omega$.*

In keeping with our interpretation of u as a density, we define the *mass* M of u by

$$M = M(u) = \int_\Omega u\,dx.$$

Note that (1.1) conserves mass, since

$$\begin{aligned} M_t &= \left(\int_\Omega u\,dx\right)_t = \int_\Omega u_t\,dx = \int_\Omega \operatorname{div}(\nabla u - u^q\vec{a})\,dx \\ &= \int_{\partial\Omega} (\nabla u - u^q\vec{a})\cdot\vec{\nu}\,d\sigma = 0. \end{aligned}$$

If, in addition to assuming that the convection is unidirectional and radially oriented, we assume that the convection is inward at the origin and not too weak and that the mass is not too small, we can show that finite-time blowup does occur. More precisely, we have the following result.

THEOREM 1.3. *Suppose* $q > 2$, (R1) *and* (R2) *hold, and there exist constants* $C > 0$ *and* $p < n(q-1)-1$ *such that* $g(r) \geq Cr^p$ *for all* $r \in [0, R]$. *Then there exists a constant* $M_0 = M_0(C, p, q, n, R)$ *such that if the mass* $M(u_0) > M_0$, *then the solution* u *of* (1.1) *blows up in finite time.*

Note that Theorem 1.3 makes no assumption about the radial symmetry of the initial data (and, therefore, of the solution). If this assumption is added, then Theorem 1.2 identifies the point where the blowup takes place.

In situations where finite-time blowup does occur and the strength of the convection field can be appropriately bounded, the rate at which blowup occurs can be estimated.

THEOREM 1.4. *Suppose hypotheses* (R1) *and* (R2) *hold and the solution* u *of* (1.1) *blows up at the origin at time* $T \in (0, \infty)$. *If* $g(r) \leq Cr^p$ *for some constants* $C > 0$ *and* $p < n(q-1) - 1$ *and all* $r \in [0, R]$, *then there exists a constant* K *such that, for all* $t < T$,

$$\|u(\cdot, t)\|_\infty \geq \frac{K}{(T-t)^\beta}, \tag{1.5}$$

where

$$\beta = \frac{p+1}{(n+p+1)(q-1)}. \tag{1.6}$$

If, in addition, hypothesis (R4) *holds, then* (1.5) *holds for*

$$\beta = \frac{p+1}{n(q-1)-(p+1)}. \tag{1.7}$$

We will proceed as follows. In section 2 we will prove that a unique solution of (1.1) exists locally for smooth initial data and that the only way this solution can fail to be global is for it to blow up. In section 3 we will prove the specific results about radially symmetric systems.

2. Existence until blowup. For the proof of Theorem 1.1, as well as for other reasons, it will be helpful to understand the set $\mathcal{E}$ of positive equilibrium solutions, i.e., functions $u_0 : \overline{\Omega} \to (0, \infty)$ for which $u(x,t) \equiv u_0(x)$ solves (1.1).

Suppose $u = u(x)$ is a positive equilibrium solution, and let $w = u^{(1-q)}/(1-q) - f$. A calculation shows that

$$\operatorname{div}(u^q \nabla w) = 0 \tag{2.1}$$

in Ω and

$$u^q \frac{\partial w}{\partial \nu} = 0 \tag{2.2}$$

on $\partial\Omega$. Multiplying (2.1) by w and using the generalized divergence theorem and the boundary condition (2.2) yields

$$\int_\Omega u^q |\nabla w|^2 dx = 0.$$

Thus, $\nabla w \equiv 0$, which implies that

$$u(x) = \frac{1}{[(q-1)(C - f(x))]^{1/(q-1)}} \tag{2.3}$$

for some constant C. It is easy to check that (2.3) defines an equilibrium solution of (1.1) for every $C > \|f\|_\infty$.

Among the facts about $\mathcal{E}$ that follow immediately from this explicit formula is the one that we state without proof in the following lemma.

LEMMA 2.1. *If $\mathcal{E}$ is the set of positive equilibrium solutions of* (1.1) *and $k \in (0, \infty)$, then*

$$\#\{u \in \mathcal{E} : \|u\|_\infty = k\} = 1.$$

We now proceed to prove that the only way that solutions can fail to be global is for them to blow up in finite time.

Proof of Theorem 1.1. Let positive initial data u_0 be given. Pick $b, c > 0$ such that $u_0(x) \in (b, c)$ for every $x \in \overline{\Omega}$, and pick $b' \in (0, b)$ and $c' \in (c, \infty)$. Now, choose $h : \mathbf{R} \to \mathbf{R}$ to be a C^∞ function, such that $h(\sigma) = \sigma^q$ for $\sigma \in (b', c')$ and $h(\sigma) = 0$ if $\sigma < b'/2$ or if $\sigma > 2c'$. Let (1.1′) be the initial-boundary value problem that results when u^q in (1.1) is replaced by $h(u)$. From Theorem 7.4 in [7], we know (1.1′) has a unique solution $\tilde{u}$ and it exists globally. (In fact, for given $\theta > 0$, existence of a solution in $C^{2+\theta, 1+\theta/2}$ is guaranteed if $\vec{a} \in C^{1+\theta}$ and $\partial\Omega \in C^{2+\theta}$.) Note that for some $T > 0$, $u = \tilde{u}$ is also a solution of (1.1) for $t < T$. This proves the first half of the theorem.

Let T^* be as defined in the statement of the theorem, and suppose

$$\limsup_{t \to T^*} \|u(\cdot, t)\|_\infty < +\infty.$$

Let b and c be as above. By Lemma 2.1, we can choose $u_1 \in \mathcal{E}$ with $\|u_1\|_\infty < b$. Pick $b' > 0$ small enough that $u_1(x) > b'$ for every $x \in \overline{\Omega}$. Pick the value $c' > c$ large enough that $u(x,t) < c' - 1$ for all $t < T^*$. Now, consider the solution $\tilde{u}$ to (1.1′). Note that u_1 is an equilibrium solution for the modified equation as well, so by the

strong maximum principle $\tilde{u} > u_1$. Thus, $\tilde{u}$ never leaves the interval (b', c') where the cutoff function h agrees with $u \mapsto u^q$. Hence, $u = \tilde{u}$ is a global solution to (1.1), which implies that $T^* = +\infty$. This proves the second half of the theorem. □

Throughout the remainder of this paper we shall assume that $\vec{a} \in C^{1+\theta}$ for some $\theta > 0$.

3. Radially symmetric systems. The proof of Theorem 1.2 hinges on the existence of quantities which remain bounded and which tie u to u_r. Such quantities are described in the following lemma.

LEMMA 3.1. *Suppose hypotheses* (R1) *and* (R4) *are satisfied. Suppose also that initial data* u_0 *are given and* u *is the corresponding solution to* (1.1). *Then*

1. *if hypothesis* (R2) *is satisfied, then* u *is radially symmetric and the quantity* $Q(x,t) := r^{n-1}(u_r + u^q g(r))$ *is bounded on* $\Omega \times (0,T)$*;*
2. *if hypothesis* (R3) *is satisfied, then* u *is radially symmetric and the quantity* $Q(x,t) := r^{n-1}(u_r - u^q g(r))$ *is bounded on* $\Omega \times (0,T)$.

Proof. The radial symmetry of u in either case is a consequence of the uniqueness of solutions to (1.1) (which was established in Theorem 1.1).

Suppose hypotheses (R1), (R2), and (R4) are satisfied, and let

$$Q(x,t) = r^{n-1}(u_r + u^q g(r)).$$

A calculation shows that Q satisfies the initial-boundary problem

$$\begin{cases} Q_t = \Delta Q + q u^{q-1} g(r) Q_r, & (x,t) \in \Omega \times (0,T), \\ Q = 0, & (x,t) \in \partial\Omega \times (0,T), \\ Q(x,0) = Q_0(x,t) := r^{n-1}(u_{0r} + u_0^q g(r)), & x \in \Omega. \end{cases}$$

By the weak maximum principle for parabolic equations (see [9]),

$$-\|Q_0\|_\infty \le Q(x,t) \le \|Q_0\|_\infty \tag{3.1}$$

for all $x \in \Omega$ and all $t > 0$. This completes the proof of part 1.

Now, suppose hypotheses (R1), (R3), and (R4) are satisfied, and let $Q(x,t) = r^{n-1}(u_r - u^q g(r))$. By an argument similar to the one before, we find, once again, that (3.1) holds for all $x \in \Omega$ and all $t > 0$, where, now, $Q_0(x) = r^{n-1}(u_{0r} - u_0^q g(r))$. This completes the proof of part 2. □

Proof of Theorem 1.2. Under hypotheses (R1), (R2), and (R4), part 1 of Lemma 3.1 states that the quantity $Q := r^{n-1}(u_r + u^q g(r))$ is bounded, say, by C. Thus,

$$u_r \le \frac{C}{r^{n-1}} - u^q g(r) \tag{3.2}$$

for all nonzero $x \in \Omega$ and all $t > 0$.

If u had a blowup point $x_0 \neq 0$, then (3.2) would imply the existence of $\varepsilon > 0$ and a sequence of times $t_1, t_2, t_3, \ldots$ such that $u(x, t_n) > n$ for every $x \in B(0, |x_0|) \setminus B(0, |x_0| - \varepsilon)$. Because of the nonnegativity of u, this would contradict the conservation of mass. Hence, 0 is the only possible blowup point.

If hypothesis (R2) is replaced by (R3), part 2 of Lemma 3.1 states that the quantity $Q := r^{n-1}(u_r - u^q g(r))$ is bounded, say, by C. Thus,

$$-\frac{C}{r^{n-1}} + u^q g(r) \le u_r \tag{3.3}$$

for all nonzero $x \in \Omega$ and all $t > 0$.

If u had a blowup point x_0 with $0 < |x_0| < R$, then (3.2) would imply the existence of $\varepsilon > 0$ and a sequence of times $t_1, t_2, t_3, \ldots$ such that $u(x, t_n) > n$ for every $x \in B(0, |x_0| + \varepsilon) \setminus B(0, |x_0|)$. Because of the nonnegativity of u, this would again contradict the conservation of mass.

Now, note that, from the definition of blowup point, if $x_0 \in \overline{\Omega}$ is not a blowup point, then there exists $\delta > 0$ such that if $x_1 \in \overline{\Omega}$ and $|x_1 - x_0| < \delta$, then x_1 is not a blowup point either. Hence, the complement of the set of blowup points is open (relative to $\overline{\Omega}$) and the set of blowup points is closed (relative to $\overline{\Omega}$). In combination with the result of the previous paragraph, this implies that 0 is not a blowup point. Thus, any blowup points must lie on $\partial\Omega$. □

The crucial idea in proving Theorem 1.3 is focusing on the amount of mass that has accumulated in a neighborhood of a potential blowup point rather than focusing on the density itself.

Proof of Theorem 1.3. Suppose hypotheses (R1) and (R2) hold and that p, q, and C satisfy the conditions in the statement of the theorem. To prove blowup, we measure the concentration of mass near the origin with the variable $w : [0, R^n] \times (0, T)$ defined by

$$w(\rho, t) = \int_{B(0,\rho^{1/n})} u(x, t)\, dx.$$

Straightforward computations reveal that

$$w_t = \int_{S(0,\rho^{1/n})} (u_r + u^q g(\rho^{1/n}))\, d\sigma(x), \tag{3.4}$$

$$w_\rho = \frac{\rho^{(1-n)/n}}{n} \int_{S(0,\rho^{1/n})} u\, d\sigma(x), \tag{3.5}$$

and

$$w_{\rho\rho} = \frac{\rho^{(2-2n)/n}}{n^2} \int_{S(0,\rho^{1/n})} u_r\, d\sigma(x). \tag{3.6}$$

Since $\xi \mapsto \xi^q$ is convex on $(0, \infty)$, applying Jensen's inequality [11] to (3.5) (after scaling $d\sigma(x)$ so that it is a probability measure) and combining it with (3.4) and (3.6) yield

$$w_t \geq n^2 \rho^{2\gamma} w_{\rho\rho} + n^q \omega_n^{1-q} \rho^\gamma g(\rho^{1/n}) w_\rho^q \tag{3.7}$$

for all $\rho \in (0, R^n)$, where ω_n is the area of the unit sphere in $\mathbf{R}^n$ and $\gamma = (n-1)/n$.

Let $p' = \max\{p, n-1\}$. Since, by hypothesis, $g(r) \geq Cr^p$, we have

$$g(r) \geq C' r^{p'}, \tag{3.8}$$

for all $r \in [0, R]$, where $C' = CR^{p-p'} > 0$. Using (3.8) and the fact that, by definition, w_ρ is nonnegative, (3.7) implies that

$$w_t \geq n^2 \rho^{2\gamma} w_{\rho\rho} + C' n^q \omega_n^{1-q} \rho^{\gamma + p'/n} w_\rho^q. \tag{3.9}$$

Now, define

$$M_0 = \omega_n \left(\frac{p'+1}{C'(q-1)} \right)^{1/(q-1)} \frac{R^{n-(p'+1)/(q-1)}}{n - \frac{p'+1}{q-1}}.$$

By hypothesis, $p < n(q-1)-1$ and $q > 2$. The first implies that $n-(p+1)/(q-1) > 0$ and the second implies that $n - ((n-1)+1)/(q-1) > 0$. Since either $p' = p$ or $p' = n-1$, we can conclude that $n - (p'+1)/(q-1) > 0$, so M_0 is well defined, positive, and finite.

Suppose that $M = M(u_0) > M_0$, and set

$$z(\rho, t) = \begin{cases} 0 & \text{if } \rho \in [0, \alpha(t)), \\ M \left(\dfrac{\rho - \alpha(t)}{R^n} \right)^{\lambda} & \text{if } \rho \in [\alpha(t), R^n], \end{cases}$$

where

$$\lambda = 1 - \frac{p'+1}{n(q-1)}$$

and $\alpha(t)$ is a continuous, decreasing function of t yet to be determined. Note that, from the discussion in the previous paragraph, $\lambda \in (0,1)$. Thus, in particular, z is continuous.

If possible, we want to choose α so that z can serve as a comparison function with which we can estimate w and, thereby, u. In particular, we want z to satisfy the opposite inequality to (3.9); i.e., we want z to satisfy

$$z_t \le n^2 \rho^{2\gamma} z_{\rho\rho} + C' n^q \omega_n^{1-q} \rho^{\gamma + p'/n} z_\rho^q \tag{3.10}$$

whenever $\rho \neq \alpha(t)$. Clearly, (3.10) is satisfied whenever $\rho < \alpha(t)$. A straightforward calculation reveals that (3.10) is satisfied for $\rho > \alpha(t)$ if and only if

$$-\alpha'(t) \le (\rho - \alpha(t))^{-\mu} \rho^{2\gamma} (A\rho^{\mu-1} - B(\rho - \alpha(t))^{\mu-1}), \tag{3.11}$$

where $\mu = (p'+1)/n$,

$$A = C' \left(\frac{M\lambda}{\omega_n} \right)^{q-1} n^q R^{n(\mu - q + 1)},$$

and $B = n^2(1-\lambda)$. Another calculation reveals that, because $M > M_0$, $A > B$. Since $\rho > \alpha(t)$, $\rho > \rho - \alpha(t)$, and $\mu - 1 \ge 0$, this means that (3.11) will be satisfied if

$$-\alpha'(t) \le (A-B)(\alpha(t))^{1-2/n}. \tag{3.12}$$

If we take $\alpha(t)$ to be the solution of the initial value problem

$$\begin{cases} \alpha'(t) = -(A-B)(\alpha(t))^{1-2/n}, \\ \alpha(0) = R^n, \end{cases}$$

then (3.12) will be satisfied, and furthermore, there will be some finite T^* such that $\alpha(t) \to 0$ as $t \to T^*$.

Now, suppose that the solution u to (1.1) does not blow up by time T^*. Then, by Theorem 1.1, u is defined for $t \in [0, T^*]$. Consider the function $y : [0, R^n] \times [0, T^*] \to \mathbf{R}$

defined by $y(\rho,t) = e^{-t}(z(\rho,t) - w(\rho,t))$. We claim that $y \le 0$ on its domain D. If it is not, then since y is continuous, it must achieve a positive maximum on D.

Note that

$$y(\rho,0) = z(\rho,0) - w(\rho,0) = -\int_{B(0,\rho^{1/n})} u_0(x)\,dx \le 0,$$

$y(0,t) = e^{-t}(z(0,t) - w(0,t)) = 0$, and

$$y(R^n,t) = e^{-t}(z(R^n,t) - w(R^n,t)) = e^{-t}\left(M\left(\frac{R^n - \alpha(t)}{R^n}\right)^{\lambda} - M\right) \le 0,$$

so y does not achieve a positive maximum on

$$\{(\rho,0) : \rho \in [0,R^n]\} \cup \{(0,t) : t \in [0,T^*]\} \cup \{(R^n,t) : t \in [0,T^*]\}.$$

Also, because $\lambda < 1$,

$$\lim_{\rho \to \alpha(t)^+} z_\rho(\rho,t) = -\infty,$$

so y cannot achieve a positive maximum at a point where $\rho = \alpha(t)$. This implies that at a positive maximum, y must satisfy $y_\rho = 0$, $y_{\rho\rho} \le 0$, and $y_t \ge 0$. But, using (3.9) and (3.10), that would mean that at such a point

$$\begin{aligned} y_t &= -y + e^{-t}(z_t - w_t) \\ &\le -y + e^{-t}(n^2\rho^{2\gamma}(z-w)_{\rho\rho} + C'n^q\omega_n^{1-q}\rho^{\gamma+p'/n}(z_\rho^q - w_\rho^q)) \\ &= -y + n^2\rho^{2\gamma}y_{\rho\rho} < 0, \end{aligned}$$

which is a contradiction.

This contradiction verifies the nonnegativity of y, which means that $w \ge z$ on all of D. But it is not hard to see that $\sup\{z(\rho,t)/\rho : \rho \in (0,R^n)\}$ becomes unbounded as $t \to T^*$, so $\sup\{w(\rho,t)/\rho : \rho \in (0,R^n)\}$ must also become unbounded. From the definition of w and the mean value theorem for integrals, this, in turn, implies that $\|u(\cdot,t)\|_\infty$ becomes unbounded as $t \to T^*$, contradicting the assumption that u does not blow up by time T^*. This completes the proof. □

In order to prove the estimates on the temporal blowup rate, it will be helpful to have the following technical lemma, which estimates the degree to which equilibrium solutions concentrate mass near the origin.

LEMMA 3.2. *Suppose that* (R1) *and* (R2) *hold and that* $g(r) \le Cr^p$ *for some* $p < n(q-1) - 1$ *and all* $r \in [0,R]$, *and let*

$$w_c(\rho) = \int_{B(0,\rho^{1/n})} u_c(x)\,dx,$$

where

$$u_c(x) = \sup\{u(x) : u \in \mathcal{E}\}.$$

Then

$$\rho \mapsto -\int_{B(0,\rho^{1/n})} u(x)\,dx$$

is convex for any $u \in \mathcal{E}$, and so is $-w_c$. Furthermore, there exists a positive constant K' such that

$$\sup\{w_c(\rho) - \lambda\rho : \rho \in [0, R^n]\} \geq K'\lambda^\mu \tag{3.13}$$

for every λ sufficiently large, where

$$\mu = 1 - \frac{n(q-1)}{p+1}. \tag{3.14}$$

Proof. It is not hard to see (e.g., by examining (2.3)) that u_c is radially symmetric and satisfies the same equation as the members of $\mathcal{E}$ but has a singularity at the origin. Furthermore, its radial symmetry implies that equality holds in (3.7) for $w = w_c$ (because Jensen's inequality is an equality in that case). Hence,

$$w_c'' = -K'\rho^{-(n-1)/n} g(\rho^{1/n})(w_c')^q \tag{3.15}$$

for some positive constant K'. (Throughout this proof, K' will represent a positive constant whose value may change from line to line.) Note that (3.15) implies the convexity of $-w_c$, and since

$$\rho \mapsto -\int_{B(0,\rho^{1/n})} u(x)\,dx$$

satisfies a similar equation for any $u \in \mathcal{E}$, it is convex also.

Integrating (3.15) and using the fact that $w_c'(\rho) \to +\infty$ as ρ approaches zero from the right yield

$$w_c'(\rho) = K' \left(\int_0^\rho \sigma^{(1-n)/n} g(\sigma^{1/n})\,d\sigma \right)^{-1/(q-1)}. \tag{3.16}$$

Applying the fact that $w_c(0) = 0$ and the assumption that $g(r) \leq Cr^p$ in (3.16) yields

$$w_c(\rho) \geq K'\rho^\mu, \tag{3.17}$$

where μ is as in (3.14). Using (3.17), it is a straightforward calculus exercise to verify that (3.13) holds for all λ sufficiently large. ∎

Lemma 3.2 provides a crucial estimate for constructing a type of supersolution that will provide an estimate of the blowup rate.

Proof of Theorem 1.4. Assume that the hypotheses (R1) and (R2) hold and that $g(r) \leq Cr^p$ for some constants $C > 0$ and $p < n(q-1) - 1$ and for all $r \in [0, R]$. Assume also that the solution u of (1.1) blows up at the origin at time $T \in (0, \infty)$. Fix $t_0 \in (0, T)$, and let $\lambda = (\omega_n/n)\|u(\cdot, t_0)\|_\infty$.

Pick $\tilde{u} \in \mathcal{E}$ such that $\|\tilde{u}\|_\infty > \|u(\cdot, t_0)\|_\infty$, and let

$$\tilde{w}(\rho) = \int_{B(0,\rho^{1/n})} \tilde{u}(x)\,dx.$$

For y between 0 and

$$G[\tilde{w}] := \sup\{\tilde{w}(\rho) - \lambda\rho : \rho \in [0, R^n]\},$$

let $\varrho(y)$ be the leftmost zero of $\tilde{w}(\rho) - \lambda\rho - y$. For $t \geq t_0$ consider the function

$$z(\rho,t) = \begin{cases} \tilde{w}(\rho) & \text{if } \rho \in [0, \varrho(v(t-t_0))], \\ \lambda\rho + v(t-t_0) & \text{if } \rho \in [\varrho(v(t-t_0)), R^n], \end{cases}$$

where v is a positive constant to be specified later. The intention is that by picking v appropriately, $z(\rho,t)$ will serve as an upper bound for

$$w(\rho,t) := \int_{B(0,\rho^{1/n})} u(x,t)\,dx, \tag{3.18}$$

and therefore, $\|\tilde{u}\|_\infty$ will bound u in a neighborhood of the origin.

Under the assumptions made above, let $v = CR^p n^q \omega_n^{1-q} \lambda^q$. We claim that $w \leq z$ as long as z is well defined (i.e., until $\tilde{w}(\rho) - \lambda\rho - v(t-t_0)$ has no zeros). By the strong maximum principle, it suffices to prove that $W \leq z$, where

$$W(\rho,t) := \int_{B(0,\rho^{1/n})} U(x,t)\,dx,$$

and U is a radially symmetric solution of (1.1) for $t \geq t_0$ that satisfies $u(x,t_0) \leq U(x,t_0) \leq \|u(\cdot,t_0)\|_\infty$.

Suppose that $W(\rho,t) > z(\rho,t)$ for some $\rho \in [0,R^n]$ and some $t \geq t_0$. Note that by the choice of λ, $W(\rho,t_0) \leq z(\rho,t_0)$. Also, $W(0,t) = z(0,t) = 0$ and $W(R^n,t) = W(R^n,t_0) \leq z(R^n,t_0) \leq z(R^n,t)$ for every $t \geq t_0$. Thus, there must be some $t_1 > t_0$ and $\rho_1 \in (0,R^n)$ for which $\zeta := (W-z)e^{-t}$ satisfies $\zeta(\rho_1,t_1) > 0$ and $\zeta(\rho_1,t_1) \geq \zeta(\rho,t)$ for every $t \in [t_0,t_1]$ and every $\rho \in [0,R^n]$. Note that (ρ_1,t_1) cannot be a point $(\varrho(v(t_1-t_0)),t_1)$ at which ζ is not smooth because W is continuously differentiable and the limit of $z_\rho(\rho,t_1)$ as ρ approaches $\varrho(v(t_1-t_0))$ from the left is higher than the corresponding right-hand limit. Hence, it must be the case that

$$\zeta_\rho(\rho_1,t_1) = 0, \tag{3.19}$$

$$\zeta_{\rho\rho}(\rho_1,t_1) \leq 0, \tag{3.20}$$

and

$$\zeta_t(\rho_1,t_1) \geq 0. \tag{3.21}$$

Equation (3.19) implies that

$$W_\rho(\rho_1,t_1) = z_\rho(\rho_1,t_1), \tag{3.22}$$

and (3.20) implies that

$$W_{\rho\rho}(\rho_1,t_1) \leq z_{\rho\rho}(\rho_1,t_1). \tag{3.23}$$

But, because $z = \tilde{w}$ for $\rho < \varrho(v(t_1-t_0))$, and because of the choice of v,

$$z_t \geq n^2\rho^{2\gamma} z_{\rho\rho} + n^q \omega_n^{1-q} \rho^\gamma g(\rho^{1/n}) z_\rho^q \tag{3.24}$$

at (ρ_1,t_1). Also, the radial symmetry of W implies that

$$W_t = n^2\rho^{2\gamma} W_{\rho\rho} + n^q \omega_n^{1-q} \rho^\gamma g(\rho^{1/n}) W_\rho^q. \tag{3.25}$$

Using (3.22)−(3.25) we find that at (ρ_1, t_1)

$$\zeta_t = -\zeta + (W_t - z_t)e^{-t} \leq -\zeta < 0,$$

which contradicts (3.21), so $w(\rho, t) \leq z(\rho, t)$ for all $\rho \in [0, R^n]$ and all $t \geq t_0$ for which z is defined. As was mentioned above, this implies that u does not blow up in this time interval.

Note that the length of this time interval past t_0 on which u is guaranteed not to blow up is proportional to $G[\tilde{w}]$ (because of the convexity of $-\tilde{w}$) and inversely proportional to v. Letting $\tilde{u} \to u_c$ (and, therefore, $\tilde{w} \to w_c$) and using Lemma 3.2 (which estimates $G[w_c]$), we find that

$$T - t_0 \geq \tilde{C}\|u(\cdot, t_0)\|_\infty^{\mu - q}, \tag{3.26}$$

for some constant $\tilde{C}$ (independent of t_0), if $\|u(\cdot, t_0)\|_\infty$ is sufficiently large. It is not hard to see that (by possibly decreasing $\tilde{C}$) we can get (3.26) to hold for t_0 bounded away from T as well, so (3.26) holds for all $t_0 \geq 0$. Replacing t_0 by t and rewriting (3.26) we get (1.5) and (1.6).

Now, suppose that the hypothesis of radial symmetry for u is added. Let w be defined as in (3.18). By part 1 of Lemma 3.1, $Q := r^{n-1}(u_r + u^q g(r))$ is bounded by a constant v, and a calculation shows that $w_t = Q$. Using this fact, an improved lower bound on the size of the remaining interval existence until blowup can be obtained in much the same way as the previous bound. In particular, we can define z as previously but pick v in the definition of z to be equal to the constant that bounds Q. An argument similar to the one used before implies that $w \leq z$ as long as z is defined; the only difference is that for $\rho > \varrho(v(t_1 - t_0))$, the inequality (3.24) no longer holds, so we use the fact that the constant v bounds w_t in place of the combination of (3.24) and (3.25) to conclude directly that $z_t \geq w_t$ for such ρ. Since v is now independent of λ, we obtain

$$T - t_0 \geq \tilde{C}\|u(\cdot, t_0)\|_\infty^{\mu}$$

in place of (3.26). The estimate (1.5) with (1.7) is an immediate consequence. □

REFERENCES

[1] N. D. Alikakos, P. W. Bates, and C. P. Grant, *Blow up for a diffusion-advection equation*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 181–190.

[2] G. R. Conner and C. P. Grant, *Asymptotics of blowup for a convection-diffusion equation with conservation*, Differential Integral Equations, 9 (1996), pp. 719–728.

[3] M. Escobedo, J. L. Vázquez, and E. Zuazua, *A diffusion-convection equation in several space dimensions*, Indiana Univ. Math. J., 42 (1993), pp. 1413–1440.

[4] M. Escobedo and E. Zuazua, *Large time behavior for convection-diffusion equations in* $\mathbf{R}^n$, J. Funct. Anal., 100 (1991), pp. 119–161.

[5] H. Fujita, *On the blowing up of solutions of the Cauchy problem for* $u_t = \delta u^{\sigma+1} + u^\beta$, J. Fac. Sci. Univ. Tokyo Sect. 1A Math., 16 (1966), pp. 105–113.

[6] R. E. Grundy, C. J. van Duijn, and C. N. Dawson, *Asymptotic profiles with finite mass in one-dimensional contaminant transport through porous media: The fast reaction case*, Quart. J. Mech. Appl. Math., 94 (1994), pp. 69–106.

[7] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[8] H. A. Levine, *The role of critical exponents in blowup theorems*, SIAM Rev., 32 (1990), pp. 262–288.

[9] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[10] P. Rosenau and S. Kamin, *Thermal waves in an absorbing and convecting medium*, Phys. D, 8 (1983), pp. 273–283.
[11] W. Rudin, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1987.
[12] E. Zuazua, *Weakly nonlinear large time behavior in scalar convection-diffusion equations*, Differential Integral Equations, 6 (1993), pp. 1481–1491.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1459–1466, November 1998

009

INEQUALITIES FOR THE POLYGAMMA FUNCTIONS*

HORST ALZER† AND JIM WELLS‡

Abstract. Let

$$F_n(x;c) = (\Psi^{(n)}(x))^2 - c\Psi^{(n-1)}(x)\Psi^{(n+1)}(x) \quad (x>0),$$

where Ψ denotes the logarithmic derivative of the gamma function, $n \geq 2$ is an integer, and c is a real number. The authors prove that the function $x \mapsto F_n(x;\alpha)$ is strictly completely monotonic on $(0,\infty)$ if and only if $\alpha \leq (n-1)/n$, while $x \mapsto -F_n(x;\beta)$ is strictly completely monotonic on $(0,\infty)$ if and only if $\beta \geq n/(n+1)$.

Key words. polygamma functions, complete monotonicity, convexity, concavity, superadditivity, inequalities

AMS subject classifications. 33B15, 26D07, 26D15

PII. S0036141097325071

1. Introduction. The psi (or digamma) function Ψ is defined for all positive real numbers x by

$$\Psi(x) = \Gamma'(x)/\Gamma(x) = -C + \sum_{i=0}^{\infty}\left(\frac{1}{1+i} - \frac{1}{x+i}\right),$$

where Γ denotes Euler's gamma function and $C = 0.5772\ldots$ is Euler's constant. Ψ and its derivatives are called polygamma functions.

Polygamma functions arise naturally in the study of beta distributions-probability models for random variables restricted to $[0,1]$. There, geometric properties, such as concavity, assist in determining the uniqueness of maximum likelihood estimates.

Several important inequalities involving the polygamma functions have been obtained in the recent past (cf. [2], [3], [4], [5], [10], [11], [12], [13], [14], [15]). The following inequality was given by S.Y. Trimble, J. Wells, and F.T. Wright [17] in 1989:

$$\frac{1}{2}\Psi'(x)\Psi'''(x) \leq (\Psi''(x))^2 \qquad (x>0). \tag{1.1}$$

The authors, however, did not give a proof for (1.1). They only remarked that "a rather tedious argument" [17, p. 1257] is necessary to establish this inequality. Recently, B.J. English and G. Rousseau [7, Prop. 1] presented a short proof for (1.1) (with "$<$" instead of "$\leq$") and used this result to establish monotonicity properties of certain harmonic sums.

It is natural to look for an extension of (1.1) involving $\Psi^{(n)}$. The direction of our investigation is suggested by the inequality

$$(\Psi^{(n)}(x))^2 \leq \Psi^{(n-1)}(x)\Psi^{(n+1)}(x) \quad (x>0; n=2,3,\ldots), \tag{1.2}$$

*Received by the editors July 30, 1997; accepted for publication (in revised form) January 13, 1998; published electronically August 26, 1998.
http://www.siam.org/journals/sima/29-6/32507.html
†Morsbacher Str. 10, 51545 Waldbröl, Germany.
‡Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027 (wells@ms.uky.edu).

which follows from a result of A.M. Fink [9, Thm. 1]. For $n = 2$, this provides a converse version of inequality (1.1).

In this paper we establish sharp upper and lower bounds, depending only on n, for the ratio

$$\frac{(\Psi^{(n)}(x))^2}{\Psi^{(n-1)}(x)\Psi^{(n+1)}(x)} \qquad (x > 0; n = 2, 3, \ldots).$$

In particular, we obtain a generalization of (1.1) and a refinement of (1.2).

Our approach is based on the observation that many inequalities involving $\Psi^{(n)}$ flow from monotonicity properties of functions connected with the psi function. Indeed, in many cases it is shown that these functions are not only monotonic but actually completely monotonic (cf. [2], [3], [5], [12], [16]).

A function f is said to be completely monotonic on an interval I, if $f \in C^\infty(I)$ and

$$(-1)^k f^{(k)}(x) \geq 0 \tag{1.3}$$

for all $x \in I$ and for all integers $k \geq 0$. If inequality (1.3) is strict for all $x \in I$ and for all $k \geq 0$, then f is said to be strictly completely monotonic on I.

Completely monotonic functions play an important role in physics, numerical analysis, probability theory, and in other fields (cf. [6], [8], [16], [19]). An exposition of the most interesting properties on completely monotonic functions can be found in [18].

The main purpose of this paper is to present new completely monotonic functions which involve the polygamma functions. More precisely, we consider the function

$$F_n(x; c) = (\Psi^{(n)}(x))^2 - c\Psi^{(n-1)}(x)\Psi^{(n+1)}(x) \quad (x > 0),$$

where $n \geq 2$ is an integer and c is a real number, and we determine all real numbers α and β such that $x \mapsto F_n(x; \alpha)$ and $x \mapsto -F_n(x; \beta)$ are strictly completely monotonic on $(0, \infty)$.

2. Main result. Our main result is the following theorem.

THEOREM 2.1. *Let $n \geq 2$ be an integer, and let α and β be real numbers. The function*

$$x \mapsto F_n(x; \alpha)$$

is strictly completely monotonic on $(0, \infty)$ if and only if

$$\alpha \leq (n-1)/n,$$

and

$$x \mapsto -F_n(x; \beta)$$

is strictly completely monotonic on $(0, \infty)$ if and only if

$$\beta \geq n/(n+1).$$

The proof of this theorem needs the following technical lemma.

LEMMA 2.2. *For any fixed integer $n \geq 2$, let*

$$I_n(a) = \int_0^1 [(2n-1)x^2 - 1] f_n(a(1-x)) f_n(a(1+x)) dx,$$

where

$$f_n(t) = t^{n-1}/(1 - e^{-t}).$$

Then $I_n(a) < 0$ for all $a > 0$.

Proof. Let

$$h_n(x) = [(2n-1)x^2 - 1](1-x^2)^{n-2}$$

and

$$u(x;a) = \frac{a(1-x)}{1 - e^{-a(1-x)}} \; \frac{a(1+x)}{1 - e^{-a(1+x)}}.$$

Then

$$I_n(a) = a^{2(n-2)} \int_0^1 h_n(x) u(x;a) dx. \tag{2.1}$$

Now, we show that $x \mapsto u(x;a)$ is strictly decreasing on $(0,1)$. In order to prove that

$$\frac{\partial}{\partial x} u(x;a) < 0 \qquad (0 < x < 1; a > 0),$$

we define

$$v(x;a) = \log u(x;a).$$

Then we have

$$\frac{\partial}{\partial x} v(x;a) = a[w(a(1+x)) - w(a(1-x))],$$

where

$$w(z) = \frac{1}{z} - \frac{1}{e^z - 1}.$$

Since

$$\frac{d}{dz} w(z) = \left[\left(\frac{z}{2}\right)^2 - \left(\sinh \frac{z}{2}\right)^2 \right] \Big/ \left(z \sinh \frac{z}{2}\right)^2 < 0 \qquad \text{for } z > 0,$$

and since $0 < a(1-x) < a(1+x)$ we have

$$w(a(1+x)) < w(a(1-x)),$$

so that

$$\frac{\partial}{\partial x} u(x;a) = u(x,a) \frac{\partial}{\partial x} v(x;a) < 0,$$

and, hence,

$$h_n(x)u(x;a) < h_n(x)u((2n-1)^{-1/2};a) \tag{2.2}$$

for $0 < x < 1$, $x \neq (2n-1)^{-1/2}$, $a > 0$. From (2.1), (2.2), and

$$\int_0^1 h_n(x)dx = \left[-x(1-x^2)^{n-1}\right]_0^1 = 0,$$

it follows that

$$I_n(a) < a^{2(n-2)}u((2n-1)^{-1/2};a)\int_0^1 h_n(x)dx = 0. \qquad \square$$

Proof of Theorem 2.1. First we show that

$$x \mapsto F_n(x;(n-1)/n)$$

is strictly completely monotonic on $(0,\infty)$. From the series representation

$$(-1)^{m+1}\Psi^{(m)}(x) = m!\sum_{i=0}^{\infty}\frac{1}{(x+i)^{m+1}} \qquad (x>0; m=1,2,\ldots) \tag{2.3}$$

we obtain for $x > 0$ and $n \geq 2$:

$$(-1)^n\Psi^{(n-1)}(x) = \int_0^{\infty} e^{-xt}f_n(t)dt, \tag{2.4}$$

where f_n is as in Lemma 2.2 (cf. [1, p. 260]). Differentiation leads to

$$(-1)^{n+1}\Psi^{(n)}(x) = \int_0^{\infty} e^{-xt}tf_n(t)dt, \tag{2.5}$$

and

$$(-1)^{n+2}\Psi^{(n+1)}(x) = \int_0^{\infty} e^{-xt}t^2f_n(t)dt. \tag{2.6}$$

It follows from (2.4), (2.5), and (2.6) that

$$\begin{aligned} F_n(x;(n-1)/n) &= [(-1)^{n+1}\Psi^{(n)}(x)]^2 \\ &\quad - \frac{n-1}{n}[(-1)^n\Psi^{(n-1)}(x)(-1)^{n+2}\Psi^{(n+1)}(x)] \\ &= \int_0^{\infty} e^{-xt}g_n(t)dt, \end{aligned} \tag{2.7}$$

where

$$g_n(t) = (tf_n(t) * tf_n(t)) - \frac{n-1}{n}(f_n(t) * t^2f_n(t))$$

and $*$ denotes the Laplace convolution. We have

$$g_n(t) = \int_0^t \left(t - \frac{2n-1}{n}s\right)sf_n(t-s)f_n(s)ds.$$

Let $s = \frac{t}{2}(1+x)$. Then

$$g_n(t) = \frac{t^3}{8n}\int_{-1}^{1}[1-2(n-1)x-(2n-1)x^2]f_n\left(\frac{t}{2}(1-x)\right)f_n\left(\frac{t}{2}(1+x)\right)dx.$$

Since $x \mapsto xf_n(\frac{t}{2}(1-x))f_n(\frac{t}{2}(1+x))$ is an odd function, we obtain

$$\begin{aligned} g_n(t) &= \frac{t^3}{8n}\int_{-1}^{1}[1-(2n-1)x^2]f_n\left(\frac{t}{2}(1-x)\right)f_n\left(\frac{t}{2}(1+x)\right)dx \\ &= -\frac{t^3}{4n}I_n\left(\frac{t}{2}\right), \end{aligned}$$

where I_n is as in Lemma 2.2. Hence, it follows from Lemma 2.2 and (2.7) that

$$(-1)^k\frac{d^k}{dx^k}F_n(x;(n-1)/n) = \int_0^\infty e^{-xt}t^k g_n(t)dt > 0 \quad (x>0; k=0,1\ldots).$$

This proves that $x \mapsto F_n(x;(n-1)/n)$ is strictly completely monotonic on $(0,\infty)$.

The series representation (2.3) implies that $(-1)^{m+1}\Psi^{(m)}(m \geq 1)$ is strictly completely monotonic on $(0,\infty)$. Since the sum and the product of two strictly completely monotonic functions are also strictly completely monotonic, we get from

$$\begin{aligned} F_n(x;\alpha) &= F_n(x;(n-1)/n) \\ &\quad + ((n-1)/n-\alpha)(-1)^n\Psi^{(n-1)}(x)(-1)^{n+2}\Psi^{(n+1)}(x) \end{aligned}$$

that $x \mapsto F_n(x;\alpha)$ is strictly completely monotonic on $(0,\infty)$ if $\alpha \leq (n-1)/n$.

Conversely, if $x \mapsto F_n(x;\alpha)$ is strictly completely monotonic on $(0,\infty)$, then we obtain

$$\alpha < (\Psi^{(n)}(x))^2/[\Psi^{(n-1)}(x)\Psi^{(n+1)}(x)] \quad (x>0). \tag{2.8}$$

From the asymptotic expansion

$$\Psi^{(m)}(x) \sim (-1)^{m-1}\left[\frac{(m-1)!}{x^m}+\frac{m!}{2x^{m+1}}+\sum_{k=1}^{\infty}B_{2k}\frac{(2k+m-1)!}{(2k)!x^{2k+m}}\right]$$
$$(x\to\infty; m=1,2,\ldots)$$

(cf. [1, p. 260]), we conclude

$$\lim_{x\to\infty}x^m\Psi^{(m)}(x) = (-1)^{m-1}(m-1)!,$$

and, hence,

$$\lim_{x\to\infty}\frac{(\Psi^{(n)}(x))^2}{\Psi^{(n-1)}(x)\Psi^{(n+1)}(x)} = \frac{n-1}{n}, \tag{2.9}$$

so that (2.8) and (2.9) imply $\alpha \leq (n-1)/n$.

Next, we prove that

$$x \mapsto -F_n(x;n/(n+1))$$

is strictly completely monotonic on $(0,\infty)$. From (2.3) and Lagrange's identity (cf. [15, p. 41]), we obtain

$$\begin{aligned}-F_n(x;n/(n+1)) &= (n/(n+1))\Psi^{(n-1)}(x)\Psi^{(n+1)}(x) - (\Psi^{(n)}(x))^2\\ &= (n!)^2\left[\sum_{i=0}^{\infty}(x+i)^{-n}\sum_{i=0}^{\infty}(x+i)^{-n-2}\right.\\ &\qquad\left.-\left(\sum_{i=0}^{\infty}(x+i)^{-n-1}\right)^2\right]\\ &= (n!)^2\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}(j-i)^2[(x+i)(x+j)]^{-n-2}.\end{aligned}$$

Hence, for all integers $k \geq 0$,

$$\begin{aligned}&(-1)^k\frac{d^k}{dx^k}(-F_n(x;n/(n+1)))\\ &\quad= (n!)^2\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}(j-i)^2\sum_{r=0}^{k}\binom{k}{r}(x+i)^{-n-2-r}(x+j)^{-n-2-k+r}\\ &\qquad\times\prod_{s=0}^{r-1}(n+2+s)\prod_{s=0}^{k-r-1}(n+2+s) > 0 \quad (x>0).\end{aligned}$$

The representation

$$\begin{aligned}-F_n(x;\beta) &= -F_n(x;n/(n+1))\\ &\quad+(\beta-n/(n+1))(-1)^n\Psi^{(n-1)}(x)(-1)^{n+2}\Psi^{(n+1)}(x)\end{aligned}$$

implies that $x \mapsto -F_n(x;\beta)$ is strictly completely monotonic on $(0,\infty)$ if $\beta \geq n/(n+1)$.

Finally, we assume that $x \mapsto -F_n(x;\beta)$ is strictly completely monotonic on $(0,\infty)$. Then we obtain

$$(\Psi^{(n)}(x))^2/[\Psi^{(n-1)}(x)\Psi^{(n+1)}(x)] < \beta \quad (x>0). \tag{2.10}$$

Using the identity

$$\Psi^{(m)}(x) = \Psi^{(m)}(x+1) + (-1)^{m+1}m!x^{-m-1} \quad (x>0; m=0,1,\ldots), \tag{2.11}$$

we conclude

$$\lim_{x\to 0}\frac{(\Psi^{(n)}(x))^2}{\Psi^{(n-1)}(x)\Psi^{(n+1)}(x)} = \frac{n}{n+1}. \tag{2.12}$$

From (2.10) and (2.12) we obtain $\beta \geq n/(n+1)$. This completes the proof of the theorem. □

As an immediate consequence of the theorem and the limit relations (2.9) and (2.12) we get sharp bounds for the ratio

$$(\Psi^{(n)}(x))^2/[\Psi^{(n-1)}(x)\Psi^{(n+1)}(x)].$$

COROLLARY 2.3. *Let $n \geq 2$ be an integer. Then for all positive real numbers x,*

$$\frac{n-1}{n} < \frac{(\Psi^{(n)}(x))^2}{\Psi^{(n-1)}(x)\Psi^{(n+1)}(x)} < \frac{n}{n+1}. \tag{2.13}$$

Both bounds are best possible.

Remark 2.4. The series representation (2.3) and the left-hand inequality of (2.13) lead to an inequality for infinite series, which can be considered as a converse of the Cauchy–Schwarz inequality:

$$\frac{n^2-1}{n^2}\sum_{i=0}^{\infty}(x+i)^{-n}\sum_{i=0}^{\infty}(x+i)^{-n-2} < \left(\sum_{i=0}^{\infty}(x+i)^{-n-1}\right)^2 \tag{2.14}$$
$$(x>0; n=2,3,\ldots).$$

From (2.9) we conclude that the constant factor $(n^2-1)/n^2$ is best possible. The special case $n=2$ (with "$\leq$" instead of "$<$") was given in [17] without a proof. We note that slight modifications in the proof of Theorem 2.1 reveal that inequality (2.14) holds for all real numbers $n>1$.

It was mentioned in [17] that inequality (1.1) implies the convexity of $1/\Psi'$ on $(0,\infty)$. This result can be generalized.

COROLLARY 2.5. *Let*

$$f_n(x;c) = ((-1)^{n+1}\Psi^{(n)}(x))^c \quad (x>0),$$

where $n \geq 1$ is an integer and c is a real number. The function $x \mapsto f_n(x;\alpha)$ is strictly convex on $(0,\infty)$ if and only if $\alpha \leq -1/n$ or $\alpha > 0$, while $x \mapsto f_n(x;\beta)$ is strictly concave on $(0,\infty)$ if and only if $-1/(n+1) \leq \beta < 0$.

Proof. Let $\alpha \neq 0$ and $x>0$; we have

$$\begin{aligned}&\frac{1}{\alpha}(\Psi^{(n+1)}(x))^{-2}(f_n(x;\alpha))^{-1+2/\alpha}\frac{\partial^2}{\partial x^2}f_n(x;\alpha)\\ &= \alpha - 1 + \Psi^{(n)}(x)\Psi^{(n+2)}(x)(\Psi^{(n+1)}(x))^{-2}.\end{aligned} \tag{2.15}$$

If $\alpha \leq -1/n$, then we conclude from the first inequality of (2.13) that the expression on the right-hand side of (2.15) is negative. And, if $\alpha > 0$, then the second inequality of (2.13) implies that the right-hand side of (2.15) is positive. In both cases we obtain

$$\frac{\partial^2}{\partial x^2}f_n(x;\alpha) > 0.$$

Now, let $x \mapsto f_n(x;\alpha)$ be strictly convex on $(0,\infty)$. Then we have

$$f_n(\delta x + (1-\delta)y;\alpha) < \delta f_n(x;\alpha) + (1-\delta)f_n(y;\alpha) \tag{2.16}$$
$$(x,y>0, x\neq y; 0<\delta<1).$$

We assume (for a contradiction) that $\alpha \in (-1/n, 0)$. If we multiply both sides of (2.16) by $x^{n\alpha}$ and let x tend to ∞, then we get $\delta^{-n\alpha}((n-1)!)^\alpha \leq \delta((n-1)!)^\alpha$, which implies $\alpha \leq -1/n$.

Similarly, we can prove that $x \mapsto f_n(x;\beta)$ is strictly concave on $(0,\infty)$ if and only if $-1/(n+1) \leq \beta < 0$. We omit the details. □

Remark 2.6. If a function f is strictly convex on $(0,\infty)$ and satisfies $\lim_{x\to 0} f(x) = 0$, then f is strictly superadditive, that is, we have

$$f(x) + f(y) < f(x+y) \quad (x, y > 0)$$

(cf. [17]). From (2.11) we obtain $\lim_{x\to 0}(-1)^{m+1}\Psi^{(m)}(x) = \infty (m \geq 0)$. This implies that the functions $((-1)^{n+1}\Psi^{(n)})^{\alpha} (n \geq 1; \alpha \leq -1/n)$ and $-((-1)^{n+1}\Psi^{(n)})^{\beta} (n \geq 1; -1/(n+1) \leq \beta < 0)$ are strictly superadditive on $(0,\infty)$. Hence, we get the following bounds for $(-1)^{n+1}\Psi^{(n)}(x+y)$:

$$\begin{aligned}
&[((-1)^{n+1}\Psi^{(n)}(x))^{\beta} + ((-1)^{n+1}\Psi^{(n)}(y))^{\beta}]^{1/\beta} \\
&\quad < \ (-1)^{n+1}\Psi^{(n)}(x+y) \\
&\quad < \ [((-1)^{n+1}\Psi^{(n)}(x))^{\alpha} + ((-1)^{n+1}\Psi^{(n)}(y))^{\alpha}]^{1/\alpha} \\
&(x, y > 0; n = 1, 2, \ldots; \alpha \leq -1/n, -1/(n+1) \leq \beta < 0).
\end{aligned}$$

Acknowledgments. The authors thank the referees for helpful comments which improved the presentation of the paper.

REFERENCES

[1] M. Abramowitz and I.A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Dover, New York, 1965.
[2] H. Alzer, *Some gamma function inequalities*, Math. Comp., 60 (1993), pp. 337–346.
[3] H. Alzer, *On some inequalities for the gamma and psi functions*, Math. Comp., 66 (1997), pp. 373–389.
[4] G.D. Anderson, R.W. Barnard, K.C. Richards, M.K. Vamanamurthy, and M. Vuorinen, *Inequalities for zero-balanced hypergeometric functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 1713–1723.
[5] J. Bustoz and M.E.H. Ismail, *On gamma function inequalities*, Math. Comp., 47 (1986), pp. 659–667.
[6] W.A. Day, *On monotonicity of the relaxation functions of viscoelastic materials*, Proc. Cambridge Philos. Soc., 67 (1970), pp. 503–508.
[7] B.J. English and G. Rousseau, *Bounds for certain harmonic sums*, J. Math. Anal. Appl., 206 (1997), pp. 428–441.
[8] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, John Wiley, New York, 1966.
[9] A.M. Fink, *Kolmogorov-Landau inequalities for monotone functions*, J. Math. Anal. Appl., 90 (1982), pp. 252–258.
[10] W. Gautschi, *Some elementary inequalities relating to the gamma and incomplete gamma function*, J. Math. Phys., 38 (1959), pp. 77–81.
[11] L. Gordon, *A stochastic approach to the gamma function*, Amer. Math. Monthly, 101 (1994), pp. 858–865.
[12] M.E.H. Ismail, L. Lorch, and M.E. Muldoon, *Completely monotonic functions associated with the gamma function and its q-analogues*, J. Math. Anal. Appl., 116 (1986), pp. 1–9.
[13] D. Kershaw, *Some extensions of W. Gautschi's inequalities for the gamma function*, Math. Comp., 41 (1983), pp. 607–611.
[14] Y.L. Luke, *Inequalities for the gamma function and its logarithmic derivative*, Math. Balkanica, 2 (1972), pp. 118–123.
[15] D.S. Mitrinović *Analytic Inequalities*, Springer-Verlag, New York, 1970.
[16] M.E. Muldoon, *Some monotonicity properties and characterizations of the gamma function*, Aequationes Math., 18 (1978), pp. 54–63.
[17] S.Y. Trimble, J. Wells, and F.T. Wright, *Superadditive functions and a statistical application*, SIAM J. Math. Anal., 20 (1989), pp. 1255–1259.
[18] D.V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1941.
[19] J. Wimp, *Sequence Transformations and Their Applications*, Academic Press, New York, 1981.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1467–1480, November 1998

010

A METHOD OF CHARACTERISTICS FOR SOME SYSTEMS OF CONSERVATION LAWS*

ARNAUD HEIBIG† AND AMINA SAHEL†

Abstract. We study systems of conservation laws which belong to the Temple class. Some algebraic formulas are derived and used to integrate the Cauchy problem. In particular, the method of characteristics is extended to the case of a system of two coupled equations.

Key words. conservation laws, systems of the Temple class, characteristics

AMS subject classifications. 35L65, 35L99

PII. S0036141096310351

Introduction. This paper deals with $n \times n$ systems of conservation law,

$$u_t + f(u)_x = 0, \qquad t > 0,$$
$$u(x,0) = u_0(x),$$

which belong to the Temple class [19]. A system of the Temple class is defined by the following algebraic requirements:

- It is strictly hyperbolic and diagonalizable.
- Its characteristic hypersurfaces are included in hyperplanes.

Examples of such systems arise in the study of nonlinear motion in elastic strings, or multicomponent chromatography. From a mathematical point of view, the main interest of Temple systems is that no secondary waves can be generated by interaction.

The Temple class has been studied by numerous authors (see [17] for a survey), and most of the scalar theory has been generalized to that frame. The reader may refer to [2], [3], [4], [5], [6], [11], [14], and [15] for existence and uniqueness results. However, unlike the scalar case, no system ($n \geq 2$) has been integrated explicitly. The aim of this paper is to give integration formulas. Our main result can be stated as follows.

for a "general" 2×2 system of the Temple class (we mainly exclude uncoupled equations), if two characteristics issued from two given points of the real axis happen to intersect, then the location of the intersection can be computed by a quadrature of the initial data and the inversion of a linear system.

Our method is valid for piecewise smooth solutions and can therefore be used after the formation of shocks.

This paper is divided into seven parts. The second part is introductory. The third contains the algebraic part of our study (Theorem 2.4). We give some properties related to functions $f(u) - \lambda_i(u)u$ and, in particular, establish by a new proof that systems of the Temple class are semi-Hamiltonian. Various consequences of Theorem 2.4 are presented in part four, including a statement on the generic smoothness of admissible solutions. Next, we give an explicit formula for the infinitesimal displacement of characteristics. The sixth part is devoted to our explicit integration formula

*Received by the editors October 17, 1996; accepted for publication (in revised form) September 11, 1997; published electronically August 26, 1998.
http://www.siam.org/journals/sima/29-6/31035.html
†C.M.I., 39 rue F. Joliot Curie, Château Gombert, 13453 Marseille Cedex 13, France (heibig@gyptis.univ-mrs.fr, sahel@gyptis.univ-mrs.fr).

for the 2×2 systems. Last, we discuss a generalized Lax formula (introduced in [2]) in term of saddle points.

1. Definitions and notations. First results. Let D be a domain of $\mathbb{R}^n$, and let $f : D \to \mathbb{R}^n$ be a smooth function. For $u \in D$, we denote by $A(u)$ the Jacobian matrix of $f(u)$. Throughout this paper, we assume that the system of conservation laws with initial data $u_0 : D \to \mathbb{R}^n$,

$$u_t + f(u)_x = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{1.1}$$

$$u(x,0) = u_0(x), \qquad x \in \mathbb{R}, \tag{1.2}$$

is strictly hyperbolic (i.e., the eigenvalues of $A(u)$ are real and simple). The eigenvalues of $A(u)$ are indexed in a natural fashion: $\lambda_1(u) < \cdots < \lambda_n(u)$. We denote by $r_i(u)$ (resp., $l_i(u)$) $(1 \le i \le n)$ a right eigenvector (resp., a left eigenvector) of $A(u)$ associated with $\lambda_i(u)$. Without loss of generality, we can assume that $(l_1(u), \cdots, l_n(u))$ is the dual basis of $(r_1(u), \cdots, r_n(u))$, i.e., the basis of (co)vectors satisfying $< l_i(u), r_j(u) >= \delta_{ij}$. Last, recall that a function $w_i : D \to \mathbb{R}\,(1 \le i \le n)$ is a (strict) i-Riemann invariant of system 1.1 if

$$\begin{array}{ll} \text{(a)} & \forall u \in D, \qquad dw_i(u) \wedge l_i(u) = 0, \\ \text{(b)} & \forall u \in D, \qquad dw_i(u) \neq 0. \end{array}$$

We now give a definition of the Temple class.

DEFINITION 1.1. (1) *Let $i \in \{1, \ldots, n\}$. We say that the ith characteristic field of system* 1.1 *is a Temple field if*

(*i*) *there exists a (strict) i-Riemann invariant.*

(*ii*) *for any $s \in \mathbb{R}$, the level sets $H_i(s) = \{u \in D / w_i(u) = s\}$ are (linear) affine submanifolds in the u-space D.*

(2) *We say that system* 1.1 *belongs to the Temple class if any of its characteristic fields is a Temple field.*

Example 1. (a) Any conservation law ($n = 1$) belongs to the Temple class.

(b) The system of isotachophoresis can be written as (see [2])

$$\partial_t u_i + \partial_x \left(\frac{a_i u_i}{m} \right) = 0, \qquad m = \sum_{i=1}^{n} u_i,$$

where a_i are positive constants. We assume that $a_1 < a_2 < \cdots < a_n$ and set $D = (\mathbb{R}_+^*)^n$. The linear function $w_1(u) = \sum_{i=1}^n (u_i/a_i)$ is a Riemann invariant associated with the eigenvalue $\lambda_1 = 0$. For any $i \in \{2, \cdots, n\}$, there exists a smooth function w_i with values in $]a_{i-1}, a_i[$ such that

$$\forall u \in D, \qquad \sum_{j=1}^{n} \frac{u_j}{a_j - w_i(u)} = 0.$$

Function w_i is a Riemann invariant associated with the eigenvalue $\lambda_i = w_i/m$. Moreover, $m = w_1(a_1 \ldots a_n)/(w_2 \ldots w_n)$. The characteristic hypersurfaces are given by $H_i(s) = \{u \in D / \sum_{j=1}^n (u_j)/(a_j - s) = 0\}$, with $s \in]a_{i-1}, a_i[$ $(i \in \{2, \ldots, n\})$. Therefore, the general system of isotachophoresis belongs to the Temple class. Last, choosing $l_i(u) = \left((w_i(u) - a_1)^{-1}, ..., (w_i(u) - a_n)^{-1}\right)$ as a left eigenvector $(i \in \{2, \ldots, n\})$, we notice that for any $u \in D$, $l_i(u) \cdot u = 0$ and $l_i(u) \cdot f(u) = -1$ $(i \in \{2, \ldots, n\})$. Hence,

the previous quantities are independent of the w_j ($j \in \{1,\dots,n\}$). See Proposition 2.2 below.

(c) Let $n = 2, \mathcal{U} =]0,+\infty[\times\mathbb{R}$, $\phi \in C^1(\mathcal{U},\mathbb{R})$, and $D = \{u \in \mathcal{U}, \sum_{j=1}^{2}(\partial\Phi/\partial u_j)(u)u_j \neq 0\}$. The Keyfitz–Kranzer system

$$u_t + (\phi(u)u)_x = 0$$

has one Temple field. In fact, the speed $\lambda = \phi$ is associated with the Riemann invariant $\theta = \arctan(u_2/u_1)$. Hence, the characteristic sets $\theta = cte$ are included in straight lines.

From now on, in order to simplify the statements, we assume that system 1.1 belongs to the Temple class. Nevertheless (except in Theorem 3.1 and Proposition 5.2) we may assume that one single field is Temple and state our results "field by field."

2. Systems of the Temple class are semi-Hamiltonian. Systems of the Temple class are conservative and endowed with a complete set of (strict) Riemann invariants, i.e., they are semi-Hamiltonian (see [16] or [20]). Hence, there exist functions $N_i : D \to \mathbb{R}_+^*$ ($i \in \{1,\dots,n\}$) with the following properties:

$$\forall u \in D, \quad \forall j \in \{1,\dots,n\} \setminus \{i\}, \quad \left(N_i \frac{\partial \lambda_i}{\partial w_j}\right)(u) = \left((\lambda_j - \lambda_i)\left(\frac{\partial N_i}{\partial w_j}\right)\right)(u).$$

Expressions of functions N_i are given in [15]. We shall derived these formulae by different means. First, we recall some formal consequences of Definition 1.1.

2.1. Definition of functions m_i and q_i ($i \in \{1,\dots,n\}$). Results of this section are classical (see, for example, [5] or [15]). We fix $i \in \{1,\dots,n\}$ and notice that function l_i (up to a scalar multiplicative function) is constant along each characteristic hyperplane $H_i(a) = \{u \in D / w_i(u) = a\}, a \in \mathbb{R}$ (cf. Definition 1.1). Hence, we may write $l_i(w_i(u))$ instead of $l_i(u)$.

LEMMA 2.1. *Assume that system* 1.1 *belongs to the Temple class. The following assertion holds* ($i \in \{1,\dots,n\}$):

$$\forall (u,v) \in D^2, \qquad l_i(w_i(u)) \cdot (u - v) = 0 \Longrightarrow l_i(w_i(u)) \cdot (f(u) - f(v)) = 0.$$

Proof. Equality $l_i(w_i(v)) \cdot (u - v) = 0$ is nothing but $v \in H_i(w_i(u))$. We also have $u \in H_i(w_i(u))$. Therefore, for any $s \in [0,1], v + s(u - v) \in H_i(w_i(u))$ and

$$\begin{aligned} l_i(w_i(u)) \cdot [f(u) - f(v)] &= \int_0^1 l_i(w_i(u)) \cdot A(v + s(u - v)) \cdot (u - v) ds \\ &= \int_0^1 \lambda_i(v + s(u - v))\, l_i(w_i(u)) \cdot (u - v) ds = 0. \end{aligned}$$ ∎

We deduce from Lemma 2.1 the following proposition.

PROPOSITION 2.2. *Assume that system* 1.1 *belongs to the Temple class. Then, for any* $i \in \{1,\dots,n\}$, *there exist smooth functions* $m_i : w_i(D) \to \mathbb{R}$ *and* $q_i : w_i(D) \to \mathbb{R}$ *such that*

$$\forall u \in D, \quad m_i(w_i(u)) = l_i(w_i(u)) \cdot u \quad \textit{and} \quad q_i(w_i(u)) = l_i(w_i(u)) \cdot f(u).$$

Proof. Due to Definition 1.1, for any $u \in D$ and $v \in H_i(w_i(u))$, $l_i(u)$ is orthogonal to $v - u$, i.e.,

$$w_i(u) = w_i(v) \Longrightarrow l_i(w_i(u)) \cdot u = l_i(w_i(v)) \cdot v,$$

and it follows that $l_i(w_i(u)) \cdot u$ is a function of $w_i(u)$ alone. We set $m_i(w_i(u)) = l_i(w_i(u)) \cdot u$. Similarly, due to Lemma 2.1, $l_i(w_i(u)) \cdot f(u)$ is a function of $w_i(u)$ alone. We set $q_i(w_i(u)) = l_i(w_i(u)) \cdot f(u)$. Functions m_i and q_i are smooth. ∎

From now on, we note $m_i(w_i)$, instead of $m_i(w_i(u))$, $q_i(w_i)$ instead of $q_i(w_i(u))$, etc.

2.2. Functions N_i. Expressions of functions N_i are given in Theorem 2.4. First we need a lemma.

LEMMA 2.3. $\forall u \in D,\ \forall i \in \{1, \dots, n\},\ l_i'(w_i) \cdot u - m_i'(w_i) \neq 0.$

Proof. Assume that $l_i'(w_i(u_0)) \cdot u_0 = m_i'(w_i(u_0))$ for some point $u_0 \in D$ and some index $i \in \{1, \dots, n\}$. Differentiating the expression of m_i, with respect to w_i, we find that

$$\forall u \in D, \qquad l_i'(w_i) \cdot u - m_i'(w_i) + l_i(w_i) \frac{\partial u}{\partial w_i} = 0.$$

Therefore $(l_i(w_i) \cdot (\partial u/\partial w_i))(u_0) = 0$. Next, equality $dw_j(u) \wedge l_j(u) = 0$ implies that $(l_j(w_j) \cdot (\partial u/\partial w_i))(u) = 0$ for any $j \neq i$ and $u \in D$. Finally, $(\partial u/\partial w_i)(u_0) = 0$, which is impossible since $u \mapsto (w_1(u), w_2(u), \cdots, w_n(u))$ is a local diffeomorphism. ∎

From now on, we assume (changing function l_i into $-l_i$ if necessary) that

$$\forall u \in D, \quad \forall i \in \{1, \dots, n\}, \quad m_i'(w_i) - l_i'(w_i) \cdot u > 0.$$

The following theorem is the first step in our method of characteristics.

THEOREM 2.4. (a) $\forall u \in D, \forall i \in \{1, \dots, n\}$,

$$\begin{cases} l_i(w_i)(f(u) - \lambda_i(u)u) &= q_i(w_i) - \lambda_i(u) m_i(w_i), \\ l_i'(w_i)(f(u) - \lambda_i(u)u) &= q_i'(w_i) - \lambda_i(u) m_i'(w_i), \end{cases}$$

(b) $\forall i \in \{1, \dots, n\}, \exists N_i : D \to \mathbb{R}_+^*, \forall j \in \{1, \dots, n\} \setminus \{i\},\ \forall u \in D,$

$$\left(N_i \frac{\partial \lambda_i}{\partial w_j} \right)(u) = \left((\lambda_j - \lambda_i) \left(\frac{\partial N_i}{\partial w_j} \right) \right)(u).$$

Moreover, we can choose $N_i(u) = m_i'(w_i) - l_i'(w_i) \cdot u$.

(c) $\forall u \in D, \forall i \in \{1, \dots, n\}$,

$$\left(N_i \frac{\partial \lambda_i}{\partial w_i} \right)(u) = q_i''(w_i) - \lambda_i(u) m_i''(w_i) - l_i''(w_i) \cdot (f(u) - \lambda_i(u)u),$$

(with $N_i(u) = m_i'(w_i) - l_i'(w_i) \cdot u$).

Proof. (a) The first equality is a consequence of the definition of the functions m_i and the functions q_i. Next, we differentiate this equality with respect to w_i,

$$\begin{aligned} &l_i'(w_i)(f(u) - \lambda_i(u)u) + l_i(w_i) \cdot (A(u) - \lambda_i(u)) \cdot (\partial u/\partial w_i) - (\partial \lambda_i/\partial w_i)(u) \\ &(l_i(w_i) \cdot u) = q'(w_i) - \lambda_i(u) m_i'(w_i) - (\partial \lambda_i/\partial w_i)(u) m_i(w_i). \end{aligned}$$

Therefore, the second formula follows from the definition of m_i and the identity $l_i(w_i) \cdot (A(u) - \lambda_i(u)) = 0$.

(b) We differentiate the second identity of (a) with respect to $w_j, j \neq i$. We get

$$\begin{aligned} & l_i'(w_i) \cdot (\partial f/\partial w_j)(u) - \lambda_i(u) \cdot (\partial u/\partial w_j)(u) - (\partial \lambda_i/\partial w_j)(u)(l_i'(w_i) \cdot u) \\ = \quad & -(\partial \lambda_i/\partial w_j)(u) \cdot m_i'(w_i). \end{aligned}$$

We use the identity $(\partial f/\partial w_j) - \lambda_j(\partial u/\partial w_j) = 0$ and find that

$$(m_i'(w_i) - l_i'(w_i) \cdot u)(\partial \lambda_i/\partial w_j)(u) = (\lambda_j - \lambda_i)(\partial/\partial w_j)(m_i'(w_i) - l_i'(w_i) \cdot u).$$

Now, Lemma 2.3 asserts that $m_i'(w_i) - l_i'(w_i) \cdot u$ doesn't vanish.

(c) Differentiate the second identity of (a) with respect to w_i and use identity $(\partial f/\partial w_i) - \lambda_i(\partial u/\partial w_i) = 0$. □

Remark 1. (a) A proof of Theorem 2.4 (b) is given in [15].

(b) For a general hyperbolic system of conservation laws, the existence of a complete set of (strict) Riemann invariants $(w_1, \ldots, w_n)$ implies the integrability conditions of Theorem 2.4 (b). See [16] and [18].

(c) Functions N_i and $N_i(\partial \lambda_i/\partial w_i)$ were introduced by Lax (cf. [9]) in the 2×2 case. Recall that for a general semi-Hamiltonian system of conservation laws, the following Riccati equation holds (here, the function u is a smooth solutions of (1.1, 1.2)):

$$\forall i \in \{1, \ldots, n\}, \qquad (\partial_t + \lambda_i \partial_x)\left(\frac{\partial_x w_i}{N_i}\right) = -N_i \frac{\partial \lambda_i}{\partial w_i}\left(\frac{\partial_x w_i}{N_i}\right)^2.$$

Example 2. Theorem 2.4 (a) is a generalization of Dafermos–Geng's results. Following [4], we consider the 2×2 system (which can be derived from the 3×3 isotachophoresis system),

$$\begin{cases} \partial_t u_1 - \partial_x(u_2/u_1) & = 0, \\ \partial_t u_2 - \partial_x(1/u_1) & = 0. \end{cases}$$

The domain of strict hyperbolicity is

$$D = \{(u_1, u_2) \in \mathbb{R}^2 / u_1 \neq 0 \text{ and } u_2^2 - 4u_1 > 0\}.$$

We set $\triangle = u_2^2 - 4u_1$. The eigenvalues of $f'(u)$ are $\lambda_i(u) = (u_2 + \varepsilon_i \sqrt{\triangle})/(2u_1^2)$ ($\varepsilon_2 = -\varepsilon_1 = 1$). Corresponding Riemann invariants are given by $w_i(u) = -(u_1 \lambda_i(u))^{-1}$. Corresponding left and right eigenvectors are given by $l_i(w_i) = (1, w_i)$ and $r_i(w_i) = {}^t(-w_j, 1)$ $(j \neq i)$. Hence, the system belongs to the Temple class. One checks that $u_1 = w_1 w_2, u_2 = -(w_1 + w_2), m_i(w_i) = -w_i^2$, and $q_i(w_i) = 1/w_i$. Therefore, for any $u \in D, N_i(u) = \mid w_2(u) - w_1(u) \mid$.

3. A few consequences of Theorem 2.4.

3.1. Generic smoothness of admissible solutions. We assume that $n = 2$, $l_i(w_i) = (1, w_i), i \in \{1, 2\}$, and that $\overline{D}$ is a compact convex subset of $\mathbb{R}^n$. Moreover assume that system (1.1) is strictly hyperbolic on $\overline{D}$. Under the hypothesis of genuine nonlinearity,

$$\forall i \in \{1, 2\}, \quad \forall u \in D, \quad \frac{\partial \lambda_i}{\partial w_i}(u) > 0,$$

the following statement still holds (cf. [4]).

THEOREM 3.1. *Solutions of the Cauchy problem* (1.1, 1.2) *(with bounded variation, constructed as limits in* $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ *of Glimm's approximate solutions) and with initial data in* C^k $(k \geq 4)$*, are generically piecewise* C^k *smooth and do not contain centered compression waves. Solutions of* (1.1, 1.2) *(with bounded variation, constructed as limits in* $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ *of Glimm's approximate solutions) and with real analytic initial data are always piecewise smooth.*

In the previous statement, the expression "solutions of the Cauchy problem (1.1, 1.2) are generically piecewise C^k smooth" means that the set of initial data which does not lead to piecewise C^k smooth solutions of (1.1, 1.2) is a set of the first category.

The proof of Theorem 3.1 is similar to the original proof of [4] in the case of the 2×2 isotachophoresis system. It makes use of Theorem 2.4 and of a one-sided inequality on Riemann invariants (analogous to the Oleinik E-inequality; see [12]) proved in [6] and [13] by means of the Glimm scheme (see also [1]). The restriction $n = 2$ comes from the term $l_i''(w_i)\cdot(f(u)-\lambda_i(u)u)$ in Theorem 2.4 (c). Indeed the function $l_i''(w_i)\cdot(f(u)-\lambda_i(u)u)$ is not identically equal to zero in the general case $n > 2$ and cannot be written nicely as a function of w_i and $\lambda_i(u)$ (see Proposition 3.3 (b)). In our case $(n = 2)$, following Dafermos and Geng, we make use of functions Z_i $(i \in \{1, 2\})$ defined by

$$\begin{aligned}&\forall\ (y,x,t)\in\mathbb{R}^2\times\mathbb{R}_+,\\ &\quad Z_i(y,x,t) = N_i(u_0(y)) + [tq''(w_i(y,0)) - (x-y)m''(w_i(y,0))]w_i'(y,0).\end{aligned}$$

See details in [4] and [13]. See also Proposition 4.2 below.

3.2. The case of a linearly degenerate field. In the case of a linearly degenerate field, we can generalize Theorem 2.4 (a) as follows (Proposition 3.2(b)).

PROPOSITION 3.2. (a) *Let* $n \geq 3$*. Then, for any pair of functions* Q *and* M : $w_i(D)\to\mathbb{R}$*, we have*

$$\forall i\in\{1,...,n\},\quad \exists u\in D,\quad f(u)-\lambda_i(u)u \neq Q(w_i)-\lambda_i(u)M(w_i).$$

(b) *Let* $n \in \mathbb{N}^*$*. Assume that the* i*th characteristic field is linearly degenerate* $(i \in \{1,\dots,n\}$ *is a fixed index). Then*

$$\forall k\in\mathbb{N},\quad \forall u\in D,\quad l_i^{(k)}(w_i)\cdot(f(u)-\lambda_i(u)u) = q_i^{(k)}(w_i)-\lambda_i(u)\cdot m_i^{(k)}(w_i).$$

Moreover, assume that $n \geq 3$*. Then, for any* $u \in D$*, the family* $(l_i(w_i),...,l_i^{(n-1)}(w_i))$ *is not a basis of* $(\mathbb{R}^n)^*$.

Proof. (a) Assume that, for some index $i \in \{1,...,n\}$,

$$\forall u\in D,\qquad f(u)-\lambda_i(u)u = Q(w_i)-\lambda_i(u)M(w_i).$$

Differentiate this expression with respect to w_j $(j \neq i)$ and use identity $(\partial f/\partial w_j)-\lambda_j(\partial u/\partial w_j) = 0$. We get

$$\forall u\in D,\qquad (\partial u/\partial w_j)(u) = (\lambda_j(u)-\lambda_i(u))^{-1}(\partial\lambda_i/\partial w_j)(u)(u-M(w_i)),$$

(strict hyperbolicity ensures that $\lambda_j(u)-\lambda_i(u)\neq 0$). Therefore, for $j \neq k$ and $j \neq i \neq k$,

$$\forall u\in D,\qquad r_j(u)\wedge r_k(u) = 0,$$

contradicting the assumption that $(r_1(u), r_2(u),\dots, r_n(u))$ is a basis.

(b) In the case $k = 0$, or 1, it is just Theorem 2.4 (a). In the case $k = 2$, we use $(\partial\lambda_i/\partial w_i) = 0$ and Theorem 2.4 (c). The general case $k \geq 2$ is similar. Finally, assume that $n \geq 3$. If $(l_i(w_i), \dots, l_i^{(n-1)}(w_i))$ were a basis of $(\mathbb{R}^n)^*$, we could find functions $Q_i : w_i(D) \to \mathbb{R}^n$ and $M_i : w_i(D) \to \mathbb{R}^n$ such that

$$\forall u \in D, \qquad f(u) - \lambda_i(u)u = Q_i(w_i) - \lambda_i(u)M_i(w_i),$$

contradicting (a). □

3.3. The general case. We return to the general case, i.e., we do not assume any longer that system (1.1) is linearly degenerate. The aim of this section is to discuss formula (c) of Theorem 2.4.

PROPOSITION 3.3. (a) *Assume $n = 2$ and $l_i(w_i) = (1, w_i)$ for any $i \in \{1, 2\}$ and $u \in D$. Then,*

$$\left(N_i \frac{\partial \lambda_i}{\partial w_i}\right)(u) = q_i''(w_i) - \lambda_i(u) m_i''(w_i). \tag{3.1}$$

(b) *Let $n = 3$, and let $i \in \{1, 2, 3\}$ be a fixed index. Assume that, for all $u \in D, det(l_i(u), l_i'(u), l_i''(u)) \neq 0$. Then, for any pair of functions g and h: $w_i(D) \to \mathbb{R}$, we have*

$$\exists u \in D, \qquad \left(N_i \frac{\partial \lambda_i}{\partial w_i}\right)(u) \neq h(w_i) - \lambda_i(u) g(w_i).$$

Proof. (a) The proof follows from Theorem 2.4 (c).

(b) Assume that, for all $u \in D, N_i(\partial\lambda_i/\partial w_i)(u) = h(w_i) - \lambda_i(u)g(w_i)$. Theorem 2.4 (c) asserts that

$$\forall u \in D, \qquad l_i''(w_i) \cdot (f(u) - \lambda_i(u)u) = \overline{h}(w_i) - \lambda_i(u)\ \overline{g}(w_i),$$

with $\overline{h} = q_i'' - h$ and $\overline{g} = m_i'' - g$. We now use Theorem 2.4 (a), hypothesis $\det(l_i(u), l_i'(u), l_i''(u)) \neq 0$ and Proposition 3.2 (a) to get a contradiction. □

It is a straightforward computation to check that the isotachophoresis system (Example 1 (b)) satisfies the assumption of Proposition 3.3 (b) $(n = 3, i = 2, 3)$.

Remark 2. Set $n = 2$. For any $i \in \{1, 2\}$, $l_i(w_i) = (\phi_i(w_i), \gamma_i(w_i))$ with $\phi_i : w_i(D) \to \mathbb{R}$ and $\gamma_i : w_i(D) \to \mathbb{R}$. Let $v \in D$, and assume that $\phi_i \neq 0$ in a neighborhood Ω of v. Then the covector $L_i(w_i) = (1, (\gamma_i/\phi_i)(w_i))$ is a left eigenvector of $f'(u)$ in Ω, and function (γ_i/ϕ_i) forms a Riemann invariant. Of course (see Theorem 3.1 and Proposition 3.3), it is easier to assume directly that $l_i(w_i) = (1, w_i)$ in D.

4. The method of characteristics. We now give two formulas which can be viewed as a generalization of the method of characteristics. Following [9], we introduce a function v defined by

$$v(x, t) = \int_\alpha^x u(\xi, t) d\xi - \int_0^t f(u(\alpha, \tau))\, d\tau. \tag{4.1}$$

Here, $\alpha \in \mathbb{R}$ is a fixed number. We set $v_0 = v(., 0)$.

Recall that a function u is piecewise smooth solution if, for any compact $K \subset \mathbb{R} \times \mathbb{R}_+$, the restriction of u on K is smooth on domains D_j $(1 \leq j \leq M_K, M_K \in \mathbb{N})$ separated by smooth (shock) curves $t \mapsto \gamma_j(t)$, with left and right C^1 limits along

γ_j $(1 \leq j \leq M_K)$. If u is a piecewise smooth solution of (1.1, 1.2) on the strip $\mathbb{R} \times [0,T]$, and y is a C^1 regularity point of u_0, we denote by $t \mapsto \xi_i(y,t)$ the ith characteristic with foot $y \in \mathbb{R}$, i.e., the maximal solution of the ordinary differential equation (O.D.E.)

$$\partial_t \xi_i(y,t) = \lambda_i(u(\xi_i(y,t),t)), \tag{4.2}$$

$$\xi_i(y,0) = y. \tag{4.3}$$

In the above system, we admit Lipschitz "solutions" ξ_i such that equation (4.2) may fail across the j-shock curves of u $(j \neq i)$.

We write w_i instead of $w_i(u)$. Obviously, $t \mapsto w_j(\xi_i(y,t),t)$ may not be equal to a constant for $j \neq i$. Hence, the i-characteristics of system (1) may not be straight lines. Nevertheless, we can (in some sense) integrate system (4.2, 4.3).

First we claim that, for any entropy, piecewise smooth solution u of $(1.1, 1.2)$, $w_i(\xi_i(y,t),t) = w_i(y,0)$ along an i-characteristic curve (see [15]). Indeed, any j-wave curve $O_j(v_0)$ $(j \neq i, v_0 \in D)$ is included in the intersection of the $(n-1)$ characteristic hyperplanes $\cap_{k \neq j} H_k(v_0) = \cap_{k \neq j} w_k^{-1}(v_0)$. Hence, $t \mapsto w_i(\xi_i(y,t),t)$ $(i \neq j)$ is constant through a j $(j \neq i)$ shock curve. Of course, an i-characteristic curve never crosses an i-shock curve, due to the Lax shock conditions. This proves the claims.

Finally, for $(x,t) \in \mathbb{R} \times \mathbb{R}_+$, notice that any backward characteristic through (x,t) is defined on the whole interval $[0,t]$ (see [17]).

We are now in a position to prove Proposition 4.1. Earlier statements are given in [15] and [17].

PROPOSITION 4.1. *Let u be an entropy, piecewise smooth solution of* (1.1, 1.2) *and let $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ be a point of continuity of u. Let y be the foot of the ith characteristic $(1 \leq i \leq n)$ through (x,t). Then,*

$$\begin{cases} l_i(w_i(y,0)) \cdot (v(y,0) - v(x,t)) + (x-y) m_i(w_i(y,0)) - t q_i(w_i(y,0)) = 0, \\ l_i'(w_i(y,0)) \cdot (v(y,0) - v(x,t)) + (x-y) m_i'(w_i(y,0)) - t q_i'(w_i(y,0)) = 0. \end{cases}$$

Here, w_i is the i-Riemann invariant associated with the function u, i.e., $w_i = w_i(u)$.

Proof. We integrate equations (1.1, 1.2) over the domain delimited by the x-axis, the vertical axis $\{(z,s) \in \mathbb{R} \times \mathbb{R}_+ / z = x\}$, and the ith characteristic $s \mapsto \xi_i(y,s)$. Using Green's formula (for functions with bounded variation), we have

$$-\int_y^x u_0(\xi) d\xi + \int_0^t f(u(x,s)) ds - \int_0^t (f(u) - \lambda_i(u)u)(\xi_i(y,s),s) ds = 0.$$

Notice that

$$v_0(y) - v(x,t) \quad = \quad -\int_y^x u_0(\xi) d\xi + \int_0^t f(u(x,s)) ds.$$

Hence,

$$v_0(y) - v(x,t) - \int_0^t (f(u) - \lambda_i(u)u)(\xi_i(y,s),s) ds = 0.$$

We multiply this equality by $l_i(w_i(\xi_i(y,s),s)) = l_i(w_i(y,0))$, and we use Theorem 2.4 to get the first equality. Using $l_i'(w_i(\xi_i(y,s),s))$ instead of $l_i(w_i(y,s),s)$, we get the second equality. ∎

From the previous proposition we deduce a formula for $\partial_y \xi_i(y,t)$, which generalizes the classical formula in the scalar case (see Example 3(a) below).

PROPOSITION 4.2. *Let $u \in C^1(\mathbb{R}\times[0,T[)\,(T>0)$ be a solution of system* (1.1). *Then, for any $(y,t)\in\mathbb{R}\times[0,T[$, we have,*

$$N_i(u(\xi_i(y,t),t))\;\partial_y\xi_i(y,t) = [tq''(w_i(y,0)) - l_i''(w_i(y,0))\cdot(v(y,0)-v(\xi_i(y,t),t)) \\ -(\xi_i(y,t)-y)m_i''(w_i(y,0))]w_i'(y,0) + N_i(u_0(y)).$$

Proof. Let $\xi_i : [0,T_1]\to\mathbb{R}, t\mapsto \xi_i(y,t)$ be the ith characteristic curve with foot $y\in\mathbb{R}$. We first prove that the function $y\to\xi_i(y,t)$ $(1\le i\le n, t\in[0,T_1])$ is differentiable. Set

$$H(x,y) = l_i'(w_i(y,0))\cdot(v(y,0)-v(x,t)) + (x-y)m_i'(w_i(y,0)) - tq_i'(w_i(y,0)),$$

and notice that $H(\xi_i(y,t),y)=0$, due to Proposition 4.1. A straightforward computation gives us

$$\begin{aligned}\frac{\partial H}{\partial x}(\xi_i(y,t),y) &= -l_i'(w_i(y,0))\cdot u(\xi_i(y,t),t) + m_i'(w_i(y,0)) \\ &= N_i(u(\xi_i(y,t),t)) \neq 0.\end{aligned}$$

Hence, function $y\mapsto\xi_i(y,t)$ is differentiable. Moreover, if we differentiate equality $H(\xi_i(y,t),y)=0$ with respect to y, we get our formula. ∎

Remark 3. The ith characteristics may not be straight lines. Nevertheless, in the previous formula, intermediate values of ξ_i (say, $\xi_i(y,s)$ with $0<s<t$) are not required.

Example 3. (a) In the scalar case ($n=1$), one may choose $l_1(u)=1$ and $N_1(u)=1$. Proposition 4.2 reads as follows: $\partial_y\xi_1(y,t) = 1+tf''(u_0(y))u_0'(y)$. Since $\xi_1(y,0)=y$, we get $\xi_1(y,t) = y+tf'(u_0(y))$.

(b) For the case of the 3×3 isotachophoresis system, see [4].

5. The case $n=2$. In the case $n=2$, we can derive from Theorem 2.4 (a) an explicit integration formula. We assume that system (1.1) is coupled, in the sense that

$$\forall i\in\{1,2\},\quad \forall u\in D,\quad l_i(w_i(u))\wedge l_i'(w_i(u))\neq 0, \tag{5.1}$$

i.e., the system cannot be reduced to two scalar equations. To be precise, if $l_i\wedge l_i'$ were (identically) equal to zero for some index $i\in\{1,2\}$, we could, up to affine transformation in the u-plane, assume that $l_i(v) = (\delta_{ij})_{j=1,2}$ and $f_i(v) = f_i(v_i)$ identically. Therefore, in order to solve the Cauchy problem (1.1, 1.2), we would have to solve two scalar Cauchy problems (i.e., $\partial_t u_i + \partial_x f_i(u_i)=0$ and $u_i(x,0)=u_{0,i}(x,0)$, and the other equation).

Recall now that (see the proof of Proposition 3.2) the following holds.

LEMMA 5.1. *Assume that system* (1.1) *belongs to the 2×2 Temple class and that property* (5.1) *is satisfied. Then, for any $i\in\{1,2\}$, we can find two smooth functions $Q_i : w_i(D)\to\mathbb{R}^2$ and $M_i : w_i(D)\to\mathbb{R}^2$, such that*

$$\forall u\in D,\qquad f(u)-\lambda_i(u)u = Q_i(w_i) - \lambda_i(u)\cdot M_i(w_i).$$

Proof. It is a consequence of Theorem 2.4 (a) and hypothesis (5.1). ∎

In the following statement, function v is still given by formula (4.1).

PROPOSITION 5.2. *Assume that system* (1.1) *belongs to the* 2×2 *Temple class and that property* (5.1) *is satisfied. Let* $u \in C^1_{ps}(\mathbb{R}\times[0,T[)$ $(T>0)$ *be a piecewise smooth solution of the Cauchy problem* (1.1, 1.2), *and let* $(y_1,y_2)\in\mathbb{R}^2, y_2<y_1$. *Assume that the* i_1th *characteristic with foot* y_1 *intersects the* i_2th *characteristic with foot* y_2 $((i_1,i_2)\in\{1,2\}^2)$, *at a point* (X,T_0) *in* $\mathbb{R}\times[0,T[$. *Then* (X,T_0) *is a solution of the following linear system*

$$[M_{i_1}(w_{i_1}(y_1,0))-M_{i_2}(w_{i_2}(y_2,0))]X+[Q_{i_2}(w_{i_2}(y_2,0))-Q_{i_1}(w_{i_1}(y_1,0))]T_0$$
$$=v_0(y_2)-v_0(y_1)+y_1M_{i_1}(w_{i_1}(y_1,0))-y_2M_{i_2}(w_{i_2}(y_2,0)).$$

Proof. We denote by Ω the domain delimited by the i_1th characteristic with foot y_1, the i_2th characteristic with foot y_2, and the axis $\{(x,t)\in\mathbb{R}\times\mathbb{R}_+, t=0, y_2\le x\le y_1\}$. We integrate equations (1.1, 1.2) over Ω,

$$\int_{y_2}^{y_1}u_0(\xi)d\xi-\int_0^{T_0}(f(u)-\lambda_{i_1}(u)u)(\xi_{i_1}(y,s),s)ds$$
$$+\int_0^{T_0}(f(u)-\lambda_{i_2}(u)u)(\xi_{i_2}(y,s),s)ds=0.$$

On the other hand (see Lemma 5.1),

$$\forall i\in\{1,2\},\qquad f(u)-\lambda_i(u)u=Q_i(w_i)-\lambda_i(u)M_i(w_i).$$

Notice also that $(i\in\{1,2\})$,

$$X-y_j=\int_0^{T_0}\lambda_{i_j}(u)(\xi_{i_j}(y,s),s)ds.$$

Hence,

$$v_0(y_1)-v_0(y_2)+[Q_{i_2}(w_{i_2}(y_2,0))-Q_{i_1}(w_{i_1}(y_1,0))]T_0$$
$$-(X-y_2)M_{i_2}(w_{i_2}(y_2,0))+(X-y_1)M_{i_1}(w_{i_1}(y_1,0))=0.\qquad\square$$

Under the hypothesis $\det\,[M_{i_2}(w_{i_2}(y_2,0))\ -M_{i_1}(w_{i_1}(y_1,0)),\ Q_{i_2}(w_{i_2}(y_2,0))-Q_{i_1}(w_{i_1}(y_1,0))]\neq 0$, one can solve (uniquely) the previous system. Take now $i_1=1$ and $i_2=2$. We get, for any $i\in\{1,2\}$, $w_i(X,T_0)=w_i(y_i,0)$, which are the formulae we were looking for. In the case $i_1=i_2$, the point (X,T_0) lies on a shock curve.

Example 4. The following system (cf. [10]) belongs to the Temple class:

$$\begin{cases}\partial_t u_1+\partial_x(u_1u_2)=0,\\ \partial_t u_2+\partial_x(u_2^2+u_1)=0.\end{cases}$$

Set $\Delta=\{(u_1,u_2)\in\mathbb{R}^2/u_2^2+4u_1>0\}$. One checks that $\lambda_1(u)=w_2+2w_1$ and $\lambda_2(u)=w_1+2w_2$, with $w_1+w_2=u_2$ and $w_1w_2=-u_1$, for $u\in\Delta$. Define D by $D=\{(w_1,w_2)\in\mathbb{R}^2/\alpha_1\le w_1\le\beta_1<\alpha_2\le w_2\le\beta_2\}$ (we assume that the constants α_i and β_i satisfy $\alpha_1\le\beta_1<\alpha_2\le\beta_2$ and $D\subset\subset\Delta$). For $i=1,2$, we choose $l_i(u)=(1,w_i)$, and we check that $m_i(w_i)=w_i^2$, $q_i(w_i)=w_i^3$, $Q_i(w_i)={}^t(-2w_i^3,3w_i^2)$ and $M_i(w_i)={}^t(-w_i^2,2w_i)$. Let $i_1=1, i_2=2$, $y_2<y_1$, and set $\overline{w}_1=w_1(y_1,0)$, $\overline{w}_2=w_2(y_2,0)$. Set also $u_0=(u_{0,1},u_{0,2})$ and let $(\overline{u}_1,\overline{u}_2)\in D$ be the point with Riemann invariants $(\overline{w}_1,\overline{w}_2)$. We have $\det(M_2(\overline{w}_2)-M_1(\overline{w}_1),Q_1(\overline{w}_1)-Q_2(\overline{w}_2))=-(\overline{w}_2-\overline{w}_1)^4\le-(\alpha_2-\beta_1)^4<0$. Assume now that $u_0\in C^\infty(\mathbb{R},D)$ and that functions $(w_i\circ u_0)$ are bounded and

nondecreasing ($i \in \{1,2\}$). Since $(\partial\lambda_i/\partial w_i) > 0$, system (1.1, 1.2) admits a smooth solution. Characteristics are defined globally in time. Moreover, the 2-characteristic with foot y_2 intersects the 1-characteristic with foot $y_1 > y_2$. The intersection point (X,T) is given by $X = y_1 + \Lambda(y_1 - y_2)/(\overline{w}_2 - \overline{w}_1)$ and $T = C(y_1 - y_2)/(\overline{w}_2 - \overline{w}_1)$ with

$$\Lambda = \int_{y_2}^{y_1} \frac{3u_{0,1}\overline{u}_2 + 2(\overline{u}_2^2 + \overline{u}_1)u_{0,2} - \overline{w}_2(4\overline{w}_1^2 + \overline{w}_1\overline{w}_2 + \overline{w}_2^2)}{(\overline{w}_2 - \overline{w}_1)^2}(\xi)\frac{d\xi}{y_1 - y_2},$$

$$C = \int_{y_2}^{y_1} \frac{u_{0,1}\overline{u}_1 + 2(u_{0,1} + \overline{u}_1)}{\overline{u}_2^2 + 4\overline{u}_1}(\xi)\frac{d\xi}{y_1 - y_2}.$$

These formulas can be used in a numerical scheme (see [6]). Finally, notice that, in the case of constant initial data u_0, the previous equalities reduce to $\Lambda = \lambda_1(u_0)$ and $C = 1$, as expected.

6. Lax formula. The smooth case. The aim of this section is to discuss an inf sup formula introduced in [2]. This formula, similar to the well-known Lax formula in the scalar case [9], is recalled below. But first, we need a few notations.

Let $n \in \mathbb{N}^*$ and $i \in \{1,\dots,n\}$ be a fixed index. We assume that $\overline{D}$ is compact, and that system 1.1 is strictly hyperbolic on $\overline{D}$. Moreover, we assume that the ith characteristic field is genuinely nonlinear, i.e.,

$$\forall v \in D, \quad (\partial\lambda_i/\partial w_i)(v) > 0.$$

Throughout this section, we denote by $u \in C^\infty(\mathbb{R} \times [0,T[)$ a smooth solution of the Cauchy problem (1.1, 1.2). In order to simplify our statements, we assume that u is equal to a constant $u_\infty \in D$ outside a compact $K \subset \mathbb{R} \times [0,T]$ $(T > 0)$ and takes values in D. Function v is defined by equality (4.1). We set $D_i = w_i(D)$ and define the function $\mathcal{L}_i$ by

$$\forall\ (y,a,x,t) \in \mathbb{R} \times D_i \times \mathbb{R} \times \mathbb{R}_+,$$

$$\mathcal{L}_i(y,a,x,t) = l_i(a) \cdot [v(y,0) - v(x,t)] + (x-y)m_i(a) - tq_i(a).$$

Following [2], the conjectured representation formula reads as follows:

$$\forall (x,t) \in \mathbb{R} \times \mathbb{R}_+, \qquad \inf_{y\in\mathbb{R}} \sup_{a\in D_i} \mathcal{L}_i(y,a,x,t) = 0. \tag{6.1}$$

As noticed in [2], this formula reduces to the Lax formula in the convex scalar case. Then indeed, equation (6.1) just means that $\inf_{y\in\mathbb{R}} \sup_{a\in\mathbb{R}} \{v_0(y) - v(x,t) + (x-y)a - tf(a)\} = 0$. Which is, for $t > 0$,

$$v(x,t) = \inf_{y\in\mathbb{R}} \left\{ v_0(y) + tf^* \left(\frac{x-y}{t} \right) \right\}.$$

Here, function f^* denotes the Legendre transform of the scalar function f.

Coming back to the general case $n \geq 1$, we now define T^* by

$$T^* = \sup\{t \in [0,T[,\ \text{such that } \partial_y \xi_i(.,t) > 0\}.$$

Inequality $T^* > 0$ is a consequence of $\partial_y \xi_i(.,0) = 1$. We now fix a time $t \in]0,T^*[$ and $\overline{x} \in \mathbb{R}$. Let $\overline{y}$ be the foot of the ith characteristic defined by $\overline{x} = \xi_i(\overline{y},t)$. Our goal is

to prove that $(\overline{y}, w_i(\overline{y},0))$ is a (critical) hyperbolic point of $\mathcal{L}_i(.,.,\overline{x},t)$ (see Theorem 6.2 below). Set $P=(\overline{y}, w_i(\overline{y},0),\overline{x},t)$. We will need the following lemma.

LEMMA 6.1. *Let $u \in C^\infty(\mathbb{R}\times[0,T[)(T>0)$ be a smooth solution of the system (1.1, 1.2). Assume that u is equal to a constant $u_\infty \in D$ outside a compact $K \subset \mathbb{R}\times[0,T]$. Then, for $\overline{y}\in\mathbb{R}$ and $i\in\{1,\dots,n\}$, we have*

(1) $\forall t\in[0,T[$,

$$\int_0^t \left(N_i \frac{\partial \lambda_i}{\partial w_i}\right)(\xi_i(\overline{y},s),s)ds = tq_i''(w_i(\overline{y},0)) - (\xi_i(\overline{y},t)-\overline{y})m_i''(w_i(\overline{y},0)) - l_i''(w_i(\overline{y},0))[v_0(\overline{y}) - v(\xi_i(\overline{y},t),t)].$$

(2) $\forall t\in[0,T[$,

$$\partial_y \xi_i(\overline{y},t) = \left[1+\left(\int_0^t \left(N_i\frac{\partial\lambda_i}{\partial w_i}\right)(\xi_i(\overline{y},s),s)ds\right)\frac{w_i'(\overline{y},0)}{N_i(u_0(\overline{y}))}\right]\frac{N_i(u_0(\overline{y}))}{N_i(u(\xi_i(\overline{y},t),t))}.$$

Proof. (1) Let $\overline{y}\in\mathbb{R}$. We now choose $z\in\mathbb{R}^-$, $|z|$ large enough, and set $\overline{x}=\xi_i(\overline{y},t)$ and $\overline{z}=\xi_i(z,t)$. We want to prove that $\int_0^t (f(u)-\lambda_i(u)u)(\xi_i(\overline{y},s),s)ds = v_0(\overline{y}) - v(\overline{x},t)$. Indeed, integrate equations (1.1, 1.2) over the domain Ω delimited by the two ith characteristics with feet $\overline{y}$ and z, and the lines $\{(\xi,0)\in\mathbb{R}\times\mathbb{R}^+, z\le\xi\le\overline{y}\}$ and $\{(\xi,s)\in\mathbb{R}\times[0,t],\ \xi_i(z,s)\le\xi\le\xi_i(\overline{y},s)\}$. We find

$$\int_z^{\overline{y}} u_0(\xi)d\xi - \int_{\overline{z}}^{\overline{x}} u(\xi,t)d\xi - \int_0^t (f(u)-\lambda_i(u)u)(\xi_i(\overline{y},s),s)ds + \int_0^t (f(u)-\lambda_i(u)u)(\xi_i(z,s),s)ds = 0.$$

A straightforward computation shows that, for $|z|$ large enough,

$$v_0(\overline{y}) - v(\overline{x},t) - f(u_\infty)t + \lambda_i(u_\infty)\,u_\infty\,t - \int_0^t (f(u)-\lambda_i(u)u)(\xi_i(\overline{y},s),s)ds + [f(u_\infty)-\lambda_i(u_\infty)\,u_\infty]t = 0.$$

Hence,

$$\int_0^t (f(u)-\lambda_i(u)u)(\xi_i(\overline{y},s),s)ds = v_0(\overline{y}) - v(\overline{x},t).$$

We now integrate identity (c) of Theorem 2.4, along the ith characteristic with foot $y\in\mathbb{R}$. We use $w_i(\xi_i(\overline{y},s),s)=w_i(\overline{y},0)$ and $\int_0^t \lambda_i(\xi(\overline{y},s),s)ds = \xi_i(\overline{y},t)-\overline{y}$ to obtain our formula.

(2) This follows from equations (1.1, 1.2) and Proposition 4.2. ∎

We deduce the following from Lemma 6.1.

THEOREM 6.2. *Assume $0<t<T^*$. Then, $(\overline{y}, w_i(\overline{y},0))$ is a (critical) hyperbolic point of function $(y,a)\mapsto\mathcal{L}_i(y,a,x,t)$. Moreover, $(\partial^2\mathcal{L}_i/\partial^2 a)(P)<0$ (here, $P=(\overline{y}, w_i(\overline{y},0),\overline{x},t)$).*

Proof. From Proposition 4.1, we have $\mathcal{L}_i(P)=0$. Let $(y,a,x,t)\in\mathbb{R}\times D_i\times\mathbb{R}\times\mathbb{R}_+$. We differentiate function $\mathcal{L}_i$, with respect to the a variable,

$$(\partial\mathcal{L}_i/\partial a)(y,a,x,t) = l_i'(a)[v(y,0)-v(x,t)] + (x-y)m_i'(a) - tq_i'(a).$$

Proposition 4.1 gives us $(\partial\mathcal{L}_i/\partial a)(P)=0$. Next, we differentiate $\mathcal{L}_i$ with respect to the y variable,

$$(\partial\mathcal{L}_i/\partial y)(y,a,x,t) = l_i(a)\cdot u(y,0) - m_i(a).$$

Hence, from definition of m_i, we get $(\partial\mathcal{L}_i/\partial y)(P) = 0$. Now, $(\partial^2\mathcal{L}_i/\partial^2 y)(y,a,x,t) = l_i(a)\cdot u_0'(y)$. Therefore $(\partial^2\mathcal{L}_i/\partial^2 y)(P) = l_i(w_i(\overline{y},0))\cdot u_0'(\overline{y})$. We use identities $m_i(w_i(y,0))\ -(l_i(w_i)\cdot u(y,0)) = 0$ and $N_i(u(y,0)) = m_i'(w_i(y,0)) - l_i'(w_i(y,0))\cdot u$ to conclude that $l_i(w_i(y,0))\cdot u_0'(y) = N_i(u(y,0))\cdot w_i'(y,0)$. Hence,

$$(\partial^2\mathcal{L}_i/\partial^2 y)(P) = N_i(u(\overline{y},0))\cdot w_i'(\overline{y},0).$$

We also have $(\partial^2\mathcal{L}_i/\partial y\partial a)(y,a,x,t) = l_i'(a)\cdot u_0(y) - m_i'(a)$. It follows that

$$(\partial^2\mathcal{L}_i/\partial y\partial a)(P) = -N_i(u(\overline{y},0)).$$

Next, $(\partial^2\mathcal{L}_i/\partial^2 a)(y,a,x,t) = l_i''(a)\cdot[v(y,0)-v(x,t)]+(x-y)m_i''(a)-tq_i''(a)$. Therefore,

$$\begin{aligned}(\partial^2\mathcal{L}_i/\partial^2 a)(P) &= l_i''(w_i(\overline{y},0))\cdot[v(\overline{y},0)-v(\overline{x},t)] + (\overline{x}-\overline{y})m_i''(w_i(\overline{y},0)) - tq_i''(w_i(\overline{y},0))\\ &= -\int_0^t\left(N_i\frac{\partial\lambda_i}{\partial w_i}\right)(\xi_i(\overline{y},s),s)ds,\end{aligned}$$

(cf. Lemma 6.1). In particular, genuine nonlinearity and $N_i > 0$ implies $(\partial^2\mathcal{L}_i/\partial^2 a)(P) < 0$. Lastly,

$$\begin{aligned}\det(\mathcal{L}_i'')(P) &= -N_i(u(\overline{y},0))w_i'(\overline{y},0)\left(\int_0^t N_i\frac{\partial\lambda_i}{\partial w_i}(\xi_i(y,s),s)ds\right) - (N_i(u(\overline{y},0)))^2\\ &= -(N_i(u(\overline{y},0)))^2\left[1+\left(\int_0^t N_i\frac{\partial\lambda_i}{\partial w_i}(\xi_i(y,s),s)ds\right)\left(\frac{w_i'(\overline{y},0)}{N_i(u(\overline{y},0))}\right)\right]\\ &= -N_i(u(\overline{y},0))N_i(u(\overline{x},t))\,\partial_y\xi_i(\overline{y},t),\end{aligned}$$

(cf. Lemma 6.1). But, $t\in]0,T^*[$ implies that $\partial_y\xi_i(\overline{y},t) > 0$. Since N_i is a positive function, we get $\det(\mathcal{L}_i''(P)) < 0$. □

For $y\in\mathbb{R}$, we set $\phi(y) = \sup_{a\in D_i}\mathcal{L}_i(y,a,\overline{x},t)$. Function ϕ has finite values (the set $\overline{D}$ is compact) and is continuous on $\mathbb{R}$. Moreover, we have the following.

COROLLARY 6.3. *Under the assumption (and with the notations) of Theorem* 6.2*, there exists a neighborhood $V(\overline{y})$ of $\overline{y}$ and a bounded interval J (which may depend on $(\overline{x},t)$) such that*

(1)$\forall y\in V(\overline{y}),\ \phi(y)\geq 0$.

(2) $\inf_{y\in V(\overline{y})}\sup_{a\in D_i}\mathcal{L}_i(y,a,\overline{x},t) = 0$.

(3) $\inf_{y\in\mathbb{R}}\sup_{a\in D_i}\mathcal{L}_i(y,a,\overline{x},t) = \inf_{y\in J}\sup_{a\in D_i}\mathcal{L}_i(y,a,\overline{x},t)$.

Proof. (1) This follows from Theorem 6.2 and the Morse lemma.

(2) Inequality $\inf_{y\in V(\overline{y})}\sup_{a\in D_i}\mathcal{L}_i(y,a,\overline{x},t)\geq 0$ follows from (1). The opposite inequality is proved in [2].

(3) Let $A_0\in D$ and set $a_0 = w_i(A_0)$. Obviously $a_0\in D_i$. Since D is an open set and $u_\infty\in D$, we can choose $A_0\in D$ in such a way that $l_i(a_0)\cdot(u_\infty - A_0) > 0$. Next, notice that

$$\mathcal{L}_i(y,a_0,\overline{x},t) = l_i(a_0)\cdot\left[\int_{\overline{x}}^y(u_0(\xi)-A_0)d\xi + \int_0^t(f(u(\overline{x},\tau))-f(A_0))d\tau\right],$$

and assume, for instance, that $y\in\mathbb{R}_+$. We fix $\beta\in\mathbb{R}_+$, β large enough. We have, as $y > \beta$,

$$\begin{aligned}\mathcal{L}_i(y,a_0,\overline{x},t) &= l_i(a_0)\cdot\int_\beta^y(u_\infty - A_0)d\xi + l_i(a_0)\cdot\left[\int_{\overline{x}}^\beta(u_0(\xi)-A_0)d\xi\right.\\ &\qquad\left.\int_0^t(f(u(\overline{x},\tau))-f(A_0))\,d\tau\right].\end{aligned}$$

Hence, $\mathcal{L}_i(y, a_0, \overline{x}, t) \to +\infty$ when $y \to +\infty$. In the case $y \to -\infty$, we argue in the same way. It follows that $\sup_{a\in D_i} \mathcal{L}_i(y, a_0, \overline{x}, t) > 0$ for $|y|$ large enough. □

See [9] for an inf sup equality in the scalar case and [2] for the Temple case for the inequality $\inf_{y\in\mathbb{R}} \sup_{a\in D_i} \mathcal{L}_i(y, a, \overline{x}, t) \leq 0$. In the case of a general solution u (with bounded variations), equality $\inf_{y\in\mathbb{R}} \sup_{a\in D_i} \mathcal{L}_i(y, a, \overline{x}, t) = 0$ is still an open problem.

Acknowledgments. The authors are indebted to the referee for various improvements to the text.

REFERENCES

[1] S. Ayad and A. Heibig, *Global interaction of fields in a system of conservation laws*, Comm. Partial Differential Equations, 23 (1998), pp. 701–725.

[2] S. Benzoni, *On a representation formula for B. Temple systems*, SIAM J. Math Anal., 27 (1996), pp. 1503–1519.

[3] A. Bressan, *A contractive metric for systems of conservation laws with coinciding shocks and rarefaction curves*, J. Differential Equations, 106 (1993), pp. 332–366.

[4] C.M. Dafermos and X. Geng, *Generalized characteristics, uniqueness and regularity of solutions in a hyperbolic system of conservation laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 231–269.

[5] A. Heibig, *Existence and uniqueness for some hyperbolic systems of conservation laws*, Arch. Rational Mech. Anal., 126 (1994), pp. 79–101.

[6] A. Heibig and A. Sahel, *Une méthode des caractéristiques pour certains systèmes de lois de conservation*, C.R. Acad. Sci. Paris Sér. I, 322 (1996), pp. 37–42.

[7] P.D. Lax, *Weak solutions of non linear hyperbolic equations and their numerical computation*, Comm. Pure Appl. Math., 7 (1954), pp. 159–193.

[8] P.D. Lax, *Hyperbolic systems of conservation laws*, II, Comm Pure Appl. Math., 10 (1957), pp. 537–566.

[9] P.D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conference Series in Appl. Math. 11, SIAM, Philadelphia, PA, 1973.

[10] A.Y. Leroux, *Approximation des systèmes hyperboliques*, Cours et Séminaires INRIA: Problèmes hyperboliques, 28 Sept. –2 Octobre 1981, Rocquencourt, p. 171.

[11] R.J. Leveque and B. Temple, *Stability of Godunov's method for a class of* 2×2 *systems of conservations laws*, Trans. AMS, 288 (1985), pp. 115–123.

[12] O. A. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Uspekhi Mat. Nauk, 12 (1957), pp. 3-73. English Transl.: Amer. Math. Soc. Transl. Ser. 2, 26 (1957), pp. 95–172.

[13] A. Sahel, *Étude d'une Classe de Systèmes de Lois de Conservation*, Thèse de Ph.D. Université de Provence, 1997.

[14] D. Serre, *Solutions à variation bornées pour certains systèmes hyperboliques de lois de conservation*, J. Differential Equations, 68 (1987), pp. 137–169.

[15] D. Serre, *Intégrabilité d'une classe de systèmes de lois de conservation*, Forum Math., 4 (1992), pp. 607–623.

[16] D. Serre, *Systèmes hyperboliques riches de lois de conservation*, in Nonlinear Partial Differential Equations and Their Application, Collège de France Seminar, Vol. XI, Paris, 1989–1991, Pitman Res. Notes Math. Ser. 299, H. Brézis and J.L. Lions, eds., Longman Sci. Tech., Harlow, 1994, pp. 248–281.

[17] D. Serre, *Systèmes de lois de conservation* I *et* II, Diderot Éditeur, Arts et Sciences, Paris, 1996.

[18] B. Sévennec, *Géometrie des Systèmes Hyperboliques de Lois de Conservation*, Mémo. Soc. Math. France (N.S.), 56 (1994).

[19] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. Amer. Math. Soc., 280 (1983), pp. 781–795.

[20] S.P. Tsarev, *The geometry of Hamiltonian systems of hydrodynamics type. The generalized hodograph method*, Izv. Akad. Nauk SSSR Ser. Math., 54 (1990), pp. 1048–1068.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1481–1491, November 1998

011

ON THE DIRICHLET PROBLEM FOR VECTORIAL HAMILTON–JACOBI EQUATIONS*

SANDRO ZAGATTI†

Abstract. We give sufficient conditions for the existence of solutions to the Hamilton–Jacobi equations with Dirichlet boundary condition:

$$\begin{cases} g(x, \det Du(x)) = 0, & \text{for a.e. } x \in \Omega, \\ u(x) = \varphi(x), & \text{for } x \in \partial\Omega, \end{cases}$$

obtaining, in addition, an application to the theory of existence of minimizers for a class of nonconvex variational problems.

Key words. Hamilton–Jacobi equations, minimum problem, Jacobian determinant, Baire category method

AMS subject classification. 49A50

PII. S0036141097321279

1. Introduction.

The purpose of this paper is to contribute to the theory of existence of solutions for the Dirichlet problem for Hamilton–Jacobi equations and to give applications of such a theory to the minimum problem in the vectorial case of the calculus of variations.

Consider the problem of minimizing the energy functional (see, for example, [D], [CZ2])

$$G(u) = \int_\Omega g(x, \det Du(x))dx,$$

on the space of maps $u \in W^{1,p}(\Omega, \mathbb{R}^n)$ $(p \geq 1)$ satisfying the boundary condition $u = \varphi$ on $\partial\Omega$, where $g = g(x, \xi)$ is a map from $\Omega(\subseteq \mathbb{R}^n) \times \mathbb{R}$ to $\overline{\mathbb{R}}$. The existence of minimum points for G cannot be deduced by the direct method of the calculus of variations since, in general, the minimizing sequences of G are not weakly compact in $W^{1,p}(\Omega, \mathbb{R}^n)$ and, moreover, if no convexity assumption is made on g (with respect to its second variable), the functional G fails to be weakly lower semicontinuous.

Suppose then $g(x, \xi) \geq 0$ for any $(x, \xi) \in \Omega \times \mathbb{R}$; clearly a solution of the problem

$$(\mathcal{P}) \qquad \begin{cases} g(x, \det Du(x)) = 0, & \text{for a.e. } x \in \Omega, \\ u(x) = \varphi(x), & \text{for } x \in \partial\Omega, \end{cases}$$

which is the vectorial Hamilton–Jacobi equation with Dirichlet boundary condition studied in this paper, is a minimum point for G.

To solve problem $\mathcal{P}$ we follow the idea, contained in a recent paper of Dacorogna and Marcellini ([DMa]), of making use of the Baire category argument. To overcome the difficulties due to the x-dependence of g we impose, instead of global condition on g, suitable compatibility conditions between the set $Z(x)$ in which the map $g(x, \cdot)$ vanishes and the boundary datum φ. More precisely we take φ in $W^{1,\infty}(\Omega, \mathbb{R}^n)$ and

*Received by the editors May 9, 1997; accepted for publication (in revised form) December 11, 1997; published electronically September 3, 1998.
http://www.siam.org/journals/sima/29-6/32127.html
†Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Beirut 2–4, I–34014 Trieste, Italy (zagatti@sissa.it).

assume that for any $x \in \Omega$ there exist real positive numbers $\alpha(x)$ and $\beta(x)$ such that $g(x, \alpha(x)) = g(x, \beta(x)) = 0$ and $\alpha(x) \leq \det D\varphi(x) \leq \beta(x)$ almost everywhere. Hence, problem $\mathcal{P}$ reduces to the differential inclusion:

$$(\mathcal{E}) \qquad \begin{cases} \det Du(x) \in \{\alpha(x), \beta(x)\}, & \text{for a. e. } x \in \Omega, \\ u(x) = \varphi(x), & \text{for } x \in \partial\Omega. \end{cases}$$

By this way no convexity, continuity, or growth condition assumption on g are needed, and the only difficulty consists in finding a suitable complete metric space in which to apply Baire's method. In our main result (Theorem 3.1) we will show that if α and β are (essentially) strictly positive elements of $L^\infty(\Omega, \mathbb{R})$, which can be approximated in $L^1(\Omega, \mathbb{R})$ from above and below, respectively, by continuous functions, then the differential inclusion $\mathcal{E}$ admits solutions in $W^{1,\infty}(\Omega, \mathbb{R}^n)$.

2. Preliminaries and notations. In this paper Ω is an open subset of $\mathbb{R}^n$ ($n \geq 1$), $|\cdot|$ and $\langle \cdot, \cdot \rangle$ denote, respectively, the Euclidean norm and the scalar product in $\mathbb{R}^n$; μ denotes the Lebesgue measure. Given that $E \subseteq \mathbb{R}^n$, $\mathrm{co}(E)$, $\mathrm{extr}(E)$, $\mathrm{int}(E)$, $\mathrm{diam}(E)$, and $\mathrm{Ls}(E)$ are, respectively, the convex hull, the set of extreme points, the interior, the diameter, and the linear span of E, $\mathrm{relint}(E)$ is the interior of E relative to $\mathrm{Ls}(E)$.

Given a set of vectors $V = \{y^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ ($m \leq n+1$), $\mathrm{co}(V)$ is said to be an $(m-1)$-simplex (of $\mathbb{R}^n$) if the dimension of $\mathrm{Ls}(\mathrm{co}(V))$ is $m-1$. Moreover, for $y \in \mathbb{R}^n$ we call $(y)^\perp$ the orthogonal complement of $\mathrm{Ls}(\{y\})$, i.e., $(y)^\perp := \{z \in \mathbb{R}^n : \langle z, y \rangle = 0\}$.

A real $n \times n$ matrix A is written as

$$A = \left(A^i_j\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}} = \begin{pmatrix} A^1_1 & \ldots & A^1_n \\ \vdots & \ddots & \vdots \\ A^n_1 & \cdots & A^n_n \end{pmatrix} = \begin{pmatrix} A^1 \\ \vdots \\ A^n \end{pmatrix} = (A_1, \ldots, A_n)$$

and the space $\mathcal{M}_n$ of real $n \times n$ matrices is endowed with the norm

$$\|A\|_{\mathcal{M}_n} := \max\{|A^i|, i = 1, \ldots, n\}.$$

Following [D] (p. 186 ff.), we introduce the vectors of $\mathbb{R}^n$ $(\mathrm{adj}_{n-1}A)^i$, $\quad i = 1, \ldots n$, defined by $(\mathrm{adj}_{n-1}A)^i_j = (-1)^{(i+j)} (\mathrm{cof}A)^i_j$ where

$$\left((\mathrm{cof}A)^i_j\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

is the adjoint matrix of A and recall that

$$\det A = \langle A^i, (\mathrm{adj}_{n-1}A)^i \rangle \tag{2.1}$$

for any $i = 1, \ldots, n$.

A map $u : \Omega \to \mathbb{R}^n$ is written as

$$u = \begin{pmatrix} u^1 \\ \vdots \\ u^n \end{pmatrix}$$

and its Jacobian matrix will be

$$Du = \begin{pmatrix} Du^1 \\ \vdots \\ Du^n \end{pmatrix}.$$

We shall use the spaces $C^0(\Omega,\mathbb{R})$, $L^\infty(\Omega,\mathbb{R})$, $L^\infty(\Omega,\mathbb{R}^n)$, $L^\infty(\Omega,\mathcal{M}_n)$, $W^{1,\infty}(\Omega,\mathbb{R})$, $W^{1,\infty}(\Omega,\mathbb{R}^n)$, $W_0^{1,\infty}(\Omega,\mathbb{R})$, $W_0^{1,\infty}(\Omega,\mathbb{R}^n)$ endowed with their usual norms.

An element $u \in W^{1,\infty}(\Omega,\mathbb{R}^m)$ ($m = 1$ or $m = n$) is said to be *countably piecewise affine* (CPA) if there exists a countable collection $\{\Omega_j, j \in \mathbb{N}\}$ of open, pairwise disjoint subsets of Ω with Lipschitz boundary, such that $\Omega = (\bigcup_{j\in\mathbb{N}} \Omega_j) \bigcup N$ (where N is a null set) and the restriction of u to Ω_j is affine.

We shall make use of tools and results obtained by Baire category methods in the study of differential inclusions (see, for example, [B], [C1], [DP]). To do this let $\mathcal{I}$ be the set of compact intervals of $\mathbb{R}$ endowed with the Hausdorff topology, following Bressan ([B]), we define the map $h : \mathbb{R} \times \mathcal{I} \to \mathbb{R} \cup \{-\infty\}$ by setting

(2.2)
$$h(\xi, I) = \begin{cases} \sup\left\{ \left(\int_0^1 |\xi - \phi(t)|^2 dt \right)^{\frac{1}{2}}, \phi : [0,1] \to I : \int_0^1 \phi(t)dt = \xi, \right\} & \text{if } \xi \in I, \\ -\infty & \text{if } \xi \notin I. \end{cases}$$

By the properties of the map h listed in [B], we have that, if $I = [a, b]$, then

$$h(\xi, I) = [\xi(a + b) - ab - \xi^2]^{\frac{1}{2}};$$

hence, the following properties are trivial.

PROPOSITION 2.1.

(i) *The map* $\mathbb{R} \times \mathcal{I} \ni (\xi, I) \mapsto h(\xi, I)$ *is upper semicontinuous;*

(ii) *the map* $I \ni \xi \mapsto h(\xi, I)$ *is strictly concave for every* $I \in \mathcal{I}$*;*

(iii) *if* $I = [a, b]$ *then* $h(\xi, I) = 0$ *if and only if* $\xi \in \{a, b\}$ *and, moreover,* $h(\xi, I) \leq \min\{|\xi - a|, |\xi - b|\}$ *for every* $\xi \in I$.

In the proof of our main result we will make use of the *likelihood* functional:

$$L(u) := \int_\Omega h(\det Du(x), J(x))dx,$$

where $u \in W^{1,\infty}(\Omega,\mathbb{R})$ and $J : \Omega \to \mathcal{I}$.

We will need the following tool (see [C2], [CZ1]).

PROPOSITION 2.2. *Let* Ω *be an open subset of* $\mathbb{R}^n$, $\epsilon > 0$ *and let* S *be an* n*-simplex with* $0 \in \operatorname{int}(S)$. *Then there exists a CPA* $u \in W_0^{1,\infty}(\Omega,\mathbb{R})$ *such that*

(i) $Du(x) \in extr(S)S$ *for almost every* $x \in \Omega$*;*

(ii) $\|u\|_{L^\infty(\Omega,\mathbb{R})} \leq \epsilon$.

Proof. Let $S = \operatorname{co}\left(\{s^i, i = 0, 1, \ldots, n\}\right)$ (i.e. $\{s^i, i = 0, 1, \ldots, n\}$ is the set of vertices of S coinciding with the set of extreme points of S) and define the polar set of S, $S^* := \bigcap_{i=0}^n \{x \in \mathbb{R}^n : \langle s^i, x\rangle \leq 1\}$. Applying Lemma 1 of [C2] we construct $v \in W_0^{1,\infty}(S^*,\mathbb{R})$ such that $Dv \in \operatorname{extr}(S)S$ almost everywhere. Then consider the following collection of subsets of Ω:

$$\mathcal{U} := \left\{ z + \rho S^* \subseteq \Omega, z \in \Omega, 0 < \rho < \epsilon(\operatorname{diam}(S))^{-1} \quad \text{subject to} \quad \operatorname{diam}(z + \rho S^*) \leq 1 \right\}.$$

By Vitali covering lemma we may select a countable subfamily of disjoint elements of $\mathcal{U}$, say $\mathcal{V} = \{V_j = z_j + \rho_j S^*, z_j \in \Omega, 0 < \rho_j < \epsilon(\operatorname{diam}(S))^{-1}, j \in \mathbb{N}\}$ and define maps $w_j \in W_0^{1,\infty}(V_j,\mathbb{R})$ by setting

$$w_j(x) := \rho_j v\left(\frac{x - z_j}{\rho_j}\right).$$

Clearly, $Dw_j \in \mathrm{extr}(S)$ almost everywhere and $|w_j(x)| \leq \rho_j(\sup_{V_j}|Dw_j|)\mathrm{diam}(V_j) \leq \epsilon$ almost everywhere.

Setting $u := \lim \sum_{j\in\mathbb{N}} w_j$, (the limit is intended in $W^{1,1}$) we have the thesis. □

We also need the following simple geometric argument.

PROPOSITION 2.3. *Let $a, b \in \mathbb{R}^n$ such that $\langle a, b\rangle \neq 0$. Let $\Sigma = \mathrm{co}(\{\sigma^i, i = 1, \ldots, n\})$ be an $(n-1)$-simplex contained in $(b)^\perp$ such that $0 \in \mathrm{relint}(\Sigma)$. Let $\lambda^-, \lambda^+, \rho \in \mathbb{R}$ such that $\rho > 0$ and $\lambda^- < 0 < \lambda^+$. Set $s^0 = \lambda^+ a$, $s^i = \lambda^- a + \rho\sigma^i$, $i = 1, \ldots, n$, and $S := \mathrm{co}(\{s^i, i = 0, \ldots, n\})$. Then S is an n-simplex and $0 \in \mathrm{int}(S)$.*

Proof. First of all remark that the dimension of $W := \mathrm{Ls}(\{\lambda^- a + \rho\sigma^i, i = 1, \ldots, n\}) = \lambda^- a + (b)^\perp$, is $n-1$. Since $\lambda^+ a \notin W$ it follows that the dimension of $\mathrm{Ls}(S)$ is n. Since $0 \in \mathrm{relint}(\Sigma)$ there exist $\mu_1, \ldots, \mu_n \in]0, 1[$ such that $\sum_{i=1}^n \mu_i = 1$ and $\sum_{i=1}^n \mu_i\sigma^i = 0$ Then

$$\begin{aligned} 0 &= \left(\frac{-\lambda^-}{\lambda^+ - \lambda^-}\right)(\lambda^+ a) + \left(\frac{\lambda^+}{\lambda^+ - \lambda^-}\right)(\lambda^- a) + \sum_{i=1}^n \mu_i\sigma^i \\ &= \left(\frac{-\lambda^-}{\lambda^+ - \lambda^-}\right)(\lambda^+ a) + \left(\frac{\lambda^+}{\lambda^+ - \lambda^-}\right)\left(\sum_{i=1}^n \mu_i\right)(\lambda^- a) + \sum_{i=1}^n \left(\frac{\lambda^+}{\lambda^+ - \lambda^-}\rho\right)\mu_i\sigma^i \\ &= \sum_{i=0}^n \nu_i s^i, \end{aligned}$$

where

$$\nu_0 = \frac{-\lambda^-}{\lambda^+ - \lambda^-}, \quad \nu_i = \frac{\lambda^+}{\lambda^+ - \lambda^-}\mu_i, \quad i = 1, \ldots, n.$$

Since $\nu_0, \ldots, \nu_n \in]0, 1[$ and $\sum_{i=1}^n \nu_i = 1$, this means $0 \in \mathrm{int}(S)$. □

3. Main result. We are interested in the following problem:

$$(\mathcal{P}) \qquad \begin{cases} g(x, \det Du(x)) = 0, & \text{for a. e. } x \in \Omega, \\ u(x) = \varphi(x), & \text{for } x \in \partial\Omega, \end{cases}$$

where $g = g(x, \xi)$ is a map from $\Omega \times \mathbb{R}$ to $\overline{\mathbb{R}}$, and we assume that for every $x \in \Omega$ there exist $\alpha(x), \beta(x) \in \mathbb{R}$ such that $0 < \alpha(x) < \beta(x)$ for every $x \in \Omega$, and $g(x, \alpha(x)) = g(x, \beta(x)) = 0$.

Hence, problem $(\mathcal{P})$ reduces to the solution of the differential inclusion:

$$\begin{cases} \det Du(x) \in \{\alpha(x), \beta(x)\}, & \text{for a. e. } x \in \Omega, \\ u(x) = \varphi(x), & \text{for } x \in \partial\Omega. \end{cases}$$

Before stating our main result we specify the assumption on the map α and β. No more assumptions on g are needed.

DEFINITION 3.1. *Let $\alpha, \beta \in L^\infty(\Omega, \mathbb{R})$. We say that α and β satisfy condition* (A) *if there exist two numbers $\overline{\alpha}, \overline{\beta} \in \mathbb{R}^+$ and two sequences $\{\alpha_l\}_{l\in\mathbb{N}}$, $\{\beta_l\}_{l\in\mathbb{N}}$ in $C^0(\Omega, \mathbb{R})$ such that*

$$0 < \overline{\alpha} \leq \alpha(x) < \beta(x) \leq \overline{\beta} \quad \text{a.e. in } \Omega;$$

$$\alpha_l(x) \geq \alpha(x) \quad \text{and} \quad \beta_l(x) \leq \beta(x) \quad \text{a.e. in } \Omega, \ \forall l \in \mathbb{N};$$

$$\alpha_l \overset{l\to\infty}{\longrightarrow} \alpha, \quad \beta_l \overset{l\to\infty}{\longrightarrow} \beta \quad \text{in } L^1(\Omega, \mathbb{R}).$$

We have the following result.

THEOREM 3.1. *Let Ω be a bounded open subset of $\mathbb{R}^n$ and $\alpha, \beta \in L^\infty(\Omega, \mathbb{R})$ satisfy* (A). *Let $\varphi \in W^{1,\infty}(\Omega, \mathbb{R}^n)$ be a countably piecewise affine map such that $\det D\varphi(x) \in [\alpha(x), \beta(x)]$ for almost every $x \in \Omega$. Then there exists $\overline{u} \in \varphi + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ such that $\det D\overline{u}(x) \in \{\alpha(x), \beta(x)\}$ for almost every $x \in \Omega$.*

Moreover,

$$\|D\overline{u}\|_{L^\infty(\Omega,\mathcal{M}_n)} \le \max\left\{ \|D\varphi\|_{L^\infty(\Omega,\mathcal{M}_n)}, 2\frac{\overline{\beta}\|D\varphi\|^n_{L^\infty(\Omega,\mathcal{M}_n)}}{\overline{\alpha}^{\frac{2n-1}{n}}} \right\}.$$

Proof. Set

$$\gamma = \frac{1}{2}\mathrm{essinf}_{x\in\Omega}\left\{ \min \frac{\det D\varphi(x)}{|D\varphi^i(x)||(\mathrm{adj}_{n-1}D\varphi(x))^i|}, \quad i = 1, \dots, n \right\}.$$

Since, for every $A \in \mathcal{M}_n$, we have that

$$\left((\mathrm{adj}_{n-1}A)^i_j\right)^2 \le \prod_{\substack{k=1\\k\neq i}}^{n} \sum_{\substack{l=1\\l\neq j}}^{n} (A^k_l)^2,$$

it follows that

$$|(\mathrm{adj}_{n-1}A)^i|^2 \le \sum_{j=1}^{n} \prod_{\substack{k=1\\k\neq i}}^{n} \sum_{\substack{l=1\\l\neq j}}^{n} (A^k_l)^2 \le \prod_{\substack{k=1\\l\neq i}}^{n} \sum_{l=1}^{n} (A^k_l)^2 = \prod_{\substack{k=1\\k\neq i}}^{n} |A^k|^2.$$

Hence, recalling that $\det D\varphi \ge \overline{\alpha}$ almost everywhere, we have that

$$\gamma \ge \frac{1}{2}\frac{\overline{\alpha}}{\|D\varphi\|^n_{L^\infty(\Omega,\mathcal{M}_n)}}. \tag{3.1}$$

We set

$$M := \max\left\{ \|D\varphi\|_{L^\infty(\Omega,\mathcal{M}_n)}, \frac{\overline{\beta}}{\overline{\alpha}^{\frac{n-1}{n}}\gamma} \right\} \tag{3.2}$$

and consider the elements u of $W^{1,\infty}(\Omega, \mathbb{R}^n)$ such that

$$\|Du(x)\|_{\mathcal{M}_n} \le M \quad \text{a.e.}, \tag{3.3}$$

$$\det Du(x) \in [\alpha(x), \beta(x)] \quad \text{a.e.}, \tag{3.4}$$

$$\frac{\det Du(x)}{|Du^i(x)||(\mathrm{adj}_{n-1}Du(x))^i|} > \gamma \quad \forall i = 1, \dots, n, \quad \text{a.e.} \tag{3.5}$$

More precisely, we introduce the set

$$V_{M,\gamma} := \left\{ u \in \varphi + W_0^{1,\infty}(\Omega, \mathbb{R}^n) \text{ such that } u \text{ is CPA and satisfies (3.3), (3.4), (3.5)} \right\},$$

remarking that $V_{M,\gamma}$ is nonempty since it contains φ.

Now we call V the completion of $V_{M,\gamma}$ with respect to the $L^\infty(\Omega, \mathbb{R}^n)$ topology and observe that

$$V \subseteq \left\{ u \in \varphi + W_0^{1,\infty}(\Omega, \mathbb{R}^n) : (3.3), (3.4) \text{ hold} \right\}.$$

To prove the last assertion take a sequence $(u_k)_{k\in\mathbb{N}}$ in $V_{M,\gamma}$ converging to some u in $L^\infty(\Omega,\mathbb{R}^n)$. Since $\|Du_k(x)\|_{\mathcal{M}_n} \le M$ a.e. we may extract a subsequence, that we still call (u_k), converging to u in the weak* topology of $W^{1,\infty}$. Hence, clearly, $\|Du(x)\|_{\mathcal{M}_n} \le M$ a.e. Moreover, for every $p > n$, $(u_k) \rightharpoonup u$ in (the weak topology of) $W^{1,p}(\Omega,\mathbb{R}^n)$; hence, (see, for example, [E, p. 31]) $\det Du_k \rightharpoonup \det Du$ in (the weak topology of) $L^{p/n}(\Omega,\mathbb{R})$ consequently, recalling that (3.4) holds for any element of the sequence $(u_k)_{k\in\mathbb{N}}$ we have that $\det Du(x) \in [\alpha(x),\beta(x)]$ for almost every $x \in \Omega$.

Consider now the multifunction $J : \Omega \to \mathcal{I}$, $J(x) = [\alpha(x),\beta(x)]$ and the *likelihood* functional $L : V \to \mathbb{R}$:

$$L(u) := \int_\Omega h(\det Du(x), J(x))dx,$$

where the map h is defined in (2.2). We define the sets

$$V_s := \left\{ u \in V : L(u) < \frac{1}{s} \right\}, \qquad s \in \mathbb{N}.$$

Our aim is now to show that $(V_s)_{s\in\mathbb{N}}$ is a collection of open and dense subsets of V.

Step 1. The sets V_s are open in V.

First of all remark that the map J is continuous with respect to Hausdorff topology; moreover, by the properties of h listed in Proposition 2.1, the map $\mathcal{M}_n \ni A \mapsto h(\det A, J(x))$ is *quasi-concave* for every $x \in \Omega$ (i.e., $\mathcal{M}_n \ni A \mapsto -h(\det A, J(x))$ is quasi-convex in Morrey sense), then L turns out to be upper semicontinuous on V with respect to weak* topology of $W^{1,\infty}$.

Fix $s \in \mathbb{N}$ and take now a sequence $(u_k)_{k\in\mathbb{N}}$ in $V - V_s$ (so that, in particular, $L(u_k) \ge 1/s$) converging in $L^\infty(\Omega,\mathcal{M}_n)$ to some $u \in V$. By the previous argument we may suppose that $(u_k) \overset{*}{\rightharpoonup} u$ in $W^{1,\infty}$; hence,

$$L(u) \ge \limsup_{k\to\infty} L(u_k) \ge \frac{1}{s}.$$

This proves that $V - V_s$ is closed in V and then V_s is open.

Step 2. The sets V_s are dense in V.

Fix $s \in \mathbb{N}$, $u \in V_{M,\gamma}$ and $\epsilon > 0$. We shall construct $v \in V_{M,\gamma}$ such that $L(v) < \frac{1}{s}$ and $\|u - v\|_{L^\infty(\Omega,\mathcal{M}_n)} \le \epsilon$. It will follow that V_s is dense in $V_{M,\gamma}$ and then in V too.

Consider the sequences $\{\alpha_l\}_{l\in\mathbb{N}}$ and $\{\beta_l\}_{l\in\mathbb{N}}$ of Definition 3.1 and take an integer l such that

$$\begin{cases} l \ge \dfrac{1}{10s\mu(\Omega)}, \\ \|\alpha_l - \alpha\|_{L^1(\Omega,\mathbb{R})} \le \dfrac{1}{5s}, \\ \|\beta_l - \beta\|_{L^1(\Omega,\mathbb{R})} \le \dfrac{1}{5s} \end{cases} \tag{3.6}$$

and an open set Λ such that $\overline{\Lambda} \subseteq \Omega$ and

$$\mu(\Omega - \Lambda) \le \frac{1}{5s(\overline{\beta} - \overline{\alpha})}. \tag{3.7}$$

Recall that u is CPA i.e., there exists a collection $\{\Omega_j, j \in \mathbb{N}\}$ of open, pairwise disjoint subsets of Ω such that $\Omega = (\bigcup \Omega_j) \bigcup N$ (N null set) and

$$u = \sum_{j=1}^{\infty} u_j \chi_{\Omega_j},$$

where u_j is affine.

Now set $\Lambda_j := \Omega_j \bigcap \Lambda$ and, for any $j \in \mathbb{N}$, consider the open set

$$\Lambda_j^* := \left\{ x \in \Lambda_j : \det Du_j(x) \in \left(\alpha_l(x) + \frac{1}{l}, \beta_l(x) - \frac{1}{l} \right) \right\}.$$

We call

$$\Lambda^* := \bigcup_{j=1}^{\infty} \Lambda_j^*,$$

set

$$v|_{\Omega - \Lambda^*} := u|_{\Omega - \Lambda^*}, \tag{3.8}$$

and now proceed to define v on each Λ_j^*.

By the uniform continuity of α_l and β_l on Λ^* we may infer the existence of a positive δ such that for any $E \subseteq \Lambda^*$ with $\operatorname{diam}(E) \leq \delta$, each one of the multifunctions

$$E \ni x \mapsto \left[\alpha_l(x), \alpha_l(x) + \frac{1}{l} \right], \qquad E \ni x \mapsto \left[\beta_l(x) - \frac{1}{l}, \beta_l(x) \right]$$

admits at least one constant selection, and by Vitali covering lemma, we may assume that each Λ_j^* has diameter less than δ.

Fix then $j \in \mathbb{N}$ such that $\Lambda_j^* \neq \emptyset$. Let d_j^-, d_j^+ be real numbers such that

$$d_j^- \in \left[\alpha_l(x), \alpha_l(x) + \frac{1}{l} \right], \qquad d_j^+ \in \left[\beta_l(x) - \frac{1}{l}, \beta_l(x) \right] \qquad \forall x \in \Lambda_j^*, \tag{3.9}$$

and, in order to simplify the notations, set

$$Du_j = A = \begin{pmatrix} A^1 \\ \vdots \\ A^n \end{pmatrix},$$

remarking that, by the definition of Λ_j^*,

$$d_j^- < \det A < d_j^+. \tag{3.10}$$

Let $i_0 \in \{1, \ldots, n\}$ be the index such that $|(\operatorname{adj}_{n-1} A)^{i_0}| \geq |(\operatorname{adj}_{n-1} A)^i|$ for $i = 1, \ldots, n$, and assume, to fix the ideas, $i_0 = 1$. We then have

(3.11)

$$\begin{aligned} |(\operatorname{adj}_{n-1} A)^1|^n \geq \prod_{i=1}^{n} |(\operatorname{adj}_{n-1} A)^i| &\geq \det \left(\left((\operatorname{adj}_{n-1} A)_j^i \right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}} \right) \\ &= \det \left(\left((\operatorname{cof} A)_j^i \right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}} \right) = (\det A)^{n-1}. \end{aligned}$$

Now consider the $(n-1)$-dimensional linear subspace of $\mathbb{R}^n$:

$$\left((\mathrm{adj}_{n-1}A)^1\right)^\perp = \{y \in \mathbb{R}^n : \langle y, (\mathrm{adj}_{n-1}A)^1\rangle = 0\}$$

and take an $(n-1)$-simplex $\Sigma = \mathrm{co}(\{\sigma^k, k = 1, \ldots, n\})$ in $\left((\mathrm{adj}_{n-1}A)^1\right)^\perp$ such that zero (of $\mathbb{R}^n$) belongs to its relative interior and $|\sigma^k| = 1$ for every k. Then, for $\rho \in \mathbb{R}$ such that

$$0 < \rho \leq |A^1| \left(1 - \frac{d_j^-}{\det A}\right), \tag{3.12}$$

we define the vectors of $\mathbb{R}^n$:

$$s_\rho^0 := \frac{d_j^+ - \det A}{\det A} A^1, \quad s_\rho^k := \frac{d_j^- - \det A}{\det A} A^1 + \rho\sigma^k, \ k = 1, \ldots, n,$$

and the matrices

$$A(k,\rho) = \begin{pmatrix} A^1 + s_\rho^k \\ A^2 \\ \vdots \\ A^n \end{pmatrix}, k = 0, \ldots, n.$$

Remark that for any $i = 1, \ldots, n$ and $k = 0, 1, \ldots, n$ the map

$$\rho \mapsto g_{k,i}(\rho) := \frac{\det A(k,\rho)}{|A(k,\rho)^i|(\mathrm{adj}_{n-1}A(k,\rho))^i|}$$

is continuous in a neighborhood of $\rho = 0$, and that

$$g_{k,i}(0) = \frac{\det A}{|A^i|(\mathrm{adj}_{n-1}A)^i|} > \gamma.$$

Hence, we can choose a $\overline{\rho}$ sufficiently small so that

$$g_{k,i}(\overline{\rho}) > \gamma \tag{3.13}$$

for any k and i.

Then set $s^k := s_{\overline{\rho}}^k$ for $k = 1, \ldots, n$ and $S := \mathrm{co}\left(\{s^k, k = 0, 1, \ldots, n\}\right)$. By (3.10) and Proposition 2.3, S is an n-simplex and $0 \in \mathrm{int}(S)$. Applying Proposition 2.2 we define a CPA map $w_j \in W_0^{1,\infty}(\Lambda_j^*, \mathbb{R})$ such that $\|w_j\|_{L^\infty(\Lambda_j^*,\mathbb{R})} \leq \epsilon$ and $Dw_j \in \{s^k, k = 0, 1, \ldots, n\}$ a.e. in Λ_j^*. Now define, on Λ_j^*,

$$v_j = \begin{pmatrix} v_j^1 \\ v_j^2 \\ \vdots \\ v_j^n \end{pmatrix} = \begin{pmatrix} u_j^1 + w_j \\ u_j^2 \\ \vdots \\ u_j^n \end{pmatrix}.$$

Clearly, v_j is a CPA element of $u_j + W_0^{1,\infty}(\Lambda_j^*, \mathbb{R}^n)$, $\|u_j - v_j\|_{L^\infty(\Lambda_j^*,\mathbb{R}^n)} \leq \epsilon$, and

$$Dv_j \in \left\{ A(k) = \begin{pmatrix} A^1 + s^k \\ A^2 \\ \vdots \\ A^n \end{pmatrix}, k = 0, \ldots, n \right\}$$

almost everywhere in Λ_j^*. We now show that v_j satisfies (3.3), (3.4), and (3.5).

First of all remark that, by the choice of $\overline{\rho}$ ((3.13)), v_j satisfies (3.5).

Recalling (2.1), we have, by elementary computations,

$$\det A(k) = \langle A^1 + s^k, (\mathrm{adj}_{n-1}A)^1\rangle = \det A + \frac{d_j^{\pm} - \det A}{\det A}\langle A^1, (\mathrm{adj}_{n-1}A)^1\rangle = d_j^{\pm}.$$

Hence,

$$\det Dv_j \in \{d_j^-, d_j^+\} \tag{3.14}$$

almost everywhere in Λ_j^*. Hence, by (3.9), (3.4) holds.

We are left to prove that v_j satisfy (3.3), i.e., that $\|A(k)\|_{\mathcal{M}_n} \le M$ for any k. This means $|A^1 + s^k| \le M$ for $k = 0, 1, \dots, n$. By (3.12) this is trivial for $k = 1, \dots, n$; hence, we estimate $|A^1 + s^0|$. By (3.2), (3.5), (3.11), recalling that u belongs to $V_{M,\gamma}$ and remarking that $d_j^+ \le \overline{\beta}$, we have

$$|A^1 + s^0| = |A^1|\frac{d_j^+}{\det A} = \frac{d_j^+}{|(\mathrm{adj}_{n-1}A)^1|\frac{\det A}{|A^1||(\mathrm{adj}_{n-1}A)^1|}} \le \frac{\overline{\beta}}{(\det A)^{\frac{n-1}{n}}\gamma} \le \frac{\overline{\beta}}{(\overline{\alpha})^{\frac{n-1}{n}}\gamma} \le M.$$

Now we set

$$v|_{\Lambda^*} = \sum_{j=1}^{\infty} v_j|_{\Lambda_j^*}. \tag{3.15}$$

Equations (3.8) and (3.15) define v on the whole Ω. Clearly, v turns out to be a CPA element of $u + W_0^{1,\infty}(\Omega, \mathbb{R}^n) = \varphi + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$; moreover, by construction, conditions (3.3), (3.4), and (3.5) are satisfied locally, and then hold on the whole Ω. Hence v belongs to $V_{M,\gamma}$. Moreover, by construction, $\|u - v\|_{L^\infty} \le \epsilon$ and to end the proof of Step 2 we are left to show that v belongs to V_s.

Let us estimate $L(v)$. By (3.9), (3.14) and by the definition of Λ^*, we have that

$$\det Dv(x) \in \left[\alpha(x), \alpha_l(x) + \frac{1}{l}\right] \bigcup \left[\beta_l(x) - \frac{1}{l}, \beta(x)\right]$$

almost everywhere in Λ. Hence, recalling point (iii) of Proposition 2.1, (3.6) and (3.7) we have

$$\begin{aligned} L(v) &= \int_{\Omega-\Lambda} h(\det Du(x), J(x))dx + \int_{\Lambda} h(\det Dv(x), J(x))dx \\ &\le \mu(\Omega - \Lambda)(\overline{\beta} - \overline{\alpha}) + \int_{\Lambda}\left(\left|\beta(x) - \beta_l(x) - \frac{1}{l}\right| + \left|\alpha(x) - \alpha_l(x) - \frac{1}{l}\right|\right)dx \\ &\le \mu(\Omega - \Lambda)(\overline{\beta} - \overline{\alpha}) + \|\beta - \beta_l\|_{L^1} + \|\alpha - \alpha_l\|_{L^1} + \frac{2}{l}\mu(\Omega) \le \frac{4}{5s} < \frac{1}{s}. \end{aligned}$$

By this way Step 2 is proved.

Now we apply the Baire theorem to conclude that $U := \bigcap_{s\in\mathbb{N}} V_s$ is nonempty. Take any element in $\overline{u} \in U$. Clearly, $\det D\overline{u}(x) \in [\alpha(x), \beta(x)]$ and $\|D\overline{u}(x)\|_{\mathcal{M}_n} \le M$ almost everywhere in Ω; moreover, $L(\overline{u}) = 0$. Hence, $h(x, \det D\overline{u}(x)) = 0$ (a.e.) and this implies, by point (iii) of Proposition 2.1, that $\det D\overline{u}(x) \in \{\alpha(x), \beta(x)\}$ almost everywhere in Ω.

Recalling (3.1), this ends the proof. □

Remarks. 1. In the proof of the theorem we have supposed that Ω is bounded; clearly, such a condition can be removed, considering arbitrary open subsets of $\mathbb{R}^n$ and solving the problem on a countable family of open, bounded, disjoint subsets of Ω.

2. In Theorem 3.1 we impose a condition on the behavior of the boundary value φ in the interior of Ω. This fact is in some sense unnatural. In Corollary 3.1 below we show that under some additional hypothesis such conditions can be removed.

COROLLARY 3.1. *Let Ω be an open bounded connected subset of $\mathbb{R}^n$ with C^∞ boundary $\partial\Omega$. Let $\alpha, \beta : \overline{\Omega} \to \mathbb{R}$ be elements of $L^\infty(\Omega, \mathbb{R})$ satisfying assumption* (A) *and such that $\alpha(x) < \beta(x)$ for almost every $x \in \overline{\Omega}$. Let φ be a C^∞–diffeomorphism of $\overline{\Omega}$ onto $\varphi(\overline{\Omega})$ such that*

$$\int_\Omega \alpha(x)dx < \int_\Omega \det D\varphi(x)dx = \mu(\varphi(\Omega)) < \int_\Omega \beta(x)dx.$$

Then there exists $\overline{u} \in \varphi + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ such that $\det D\overline{u}(x) \in \{\alpha(x), \beta(x)\}$ almost everywhere in Ω.

Proof. Let $f \in C^\infty(\overline{\Omega}, \mathbb{R})$ be such that

$$\alpha(x) < f(x) < \beta(x) \quad \forall x \in \overline{\Omega}$$

and

$$\int_\Omega f(x)dx = \mu(\varphi(\Omega)).$$

By Theorem 5 in [DMo] there exists a C^∞-diffeomorphism ψ of $\overline{\Omega}$ onto itself such that

$$\begin{cases} \det D\varphi(\psi(x)) \det D\psi(x) = f(x), & x \in \Omega, \\ \psi(x) = x, & x \in \partial\Omega. \end{cases}$$

Since $\Phi := \varphi \circ \psi$ is C^∞, by standard approximation arguments (see [ET, Prop 2.1 p. 309]) we may find a sequence $(\Phi_k)_{k\in\mathbb{N}}$ of CPA elements of $\Phi + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ such that $\Phi_k \to \Phi$ in $W^{1,\infty}(\Omega, \mathbb{R}^n)$ as $k \to \infty$. Consequently, we may choose an element $\overline{\Phi}$ of such sequence such that

$$\det\overline{\Phi}(x) \in (\alpha(x), \beta(x)),$$

almost everywhere in Ω. Then, applying Theorem 3.1, we can find $\overline{u} \in \overline{\Phi} + W_0^{1,\infty}(\Omega, \mathbb{R}^n) = \varphi + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ such that $\det D\overline{u}(x) \in \{\alpha(x), \beta(x)\}$. ∎

Acknowledgment. I would like to thank Professor Alberto Bressan for a useful suggestion about the final form of this work.

REFERENCES

[B] A. BRESSAN, *The most likely path of a differential inclusion*, J. Differential Equations, 88 (1990), pp. 155–174.

[C1] A. CELLINA, *On the differential inclusion $x' \in [-1, 1]$*, Atti Accad. Naz Lincei, Rend. Sci. Fis. Mat. Natur., 69 (1980), pp. 1–6.

[C2] A. CELLINA, *On minima of a functional of the gradient: Sufficient condition*, Nonlinear Anal., 20 (1993), pp. 343–347.

[CZ1] A. CELLINA AND S. ZAGATTI, *On a version of Olech's lemma in a problem of the calculus of variations*, SIAM J. Control Optim., 32 (1994), pp. 1114–1127.

[CZ2] A. Cellina and S. Zagatti, *An existence result in a problem of the vectorial case of the calculus of variations*, SIAM J. Control Optim., 33 (1995), pp. 960–970.

[D] B. Dacorogna, *Direct methods in the calculus of variations*, Springer-Verlag, Berlin, 1989.

[DMa] B. Dacorogna and P. Marcellini, *General Existence Theorems for Hamilton–Jacobi Equations in the Scalar and Vectorial Cases*, preprint, 1996.

[DMo] B. Dacorogna and J. Moser, *On a partial differential equation involving the Jacobian determinant*, Ann. Inst. H. Poincaré, 7 (1990), pp. 1–26.

[DP] F.S. De Blasi and G. Pianigiani, *A Baire category approach to the existence of solutions of multivalued differental equations in Banach spaces*, Funkcial Evac., 25 (1982), pp. 153–162.

[E] L.C. Evans, *Weak convergence methods for nonlinear partial differential equations*, CBMS 74, Chicago, IL, 1980.

[ET] I. Ekeland and R. Temam, *Convex analysis and variational problems*, North–Holland, Amsterdam, 1974.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1492–1518, November 1998

012

ON THE STATIONARY CAHN–HILLIARD EQUATION: BUBBLE SOLUTIONS*

JUNCHENG WEI† AND MATTHIAS WINTER‡

Abstract. We study stationary solutions of the Cahn–Hilliard equation in a bounded smooth domain that have an interior spherical interface (bubbles). We show that a large class of interior points (the "nondegenerate peak" points) have the following property: there exists such a solution whose bubble center lies close to a given nondegenerate peak point. Our construction uses, among others, the Liapunov–Schmidt reduction method and exponential asymptotics.

Key words. bubbles, exponential asymptotics, phase transition

AMS subject classifications. Primary, 35B40, 35B45; Secondary, 35J40

PII. S0036141097320663

1. Introduction.

In this paper, we continue our investigation of stationary solutions of the Cahn–Hilliard equation.

The Cahn–Hilliard equation is the simplest model for the separation of a binary mixture in the presence of a mass constraint (see [7]). It can be derived from a Helmholtz-free energy

$$E(u) = \int_\Omega \left[F\big(u(x)\big) + \frac{1}{2}\epsilon^2 |\nabla u(x)|^2 \right] dx \tag{1.1}$$

subject to the constraint $\frac{1}{|\Omega|}\int_\Omega u\,dx = m$. Here Ω is a bounded smooth domain corresponding to the region occupied by the body, $u(x)$ is a conserved order parameter representing, for example, the concentration, ϵ is the range of intermolecular forces, the gradient term is a contribution to the free energy coming from spatial fluctuations of the order parameter, and $F(u)$ is the free energy density which has a double-well structure at low temperatures. The simplest one is $F(u) = \frac{1}{4}(1-u^2)^2$. Hence, $f(u) := F'(u) = u^3 - u$. For the rest of the paper we often write $u^3 - u$ instead of $f(u)$. However, since we are looking for solutions of (1.2) with $\|u\|_{L^\infty(\Omega)} \le C$, we can modify the nonlinearity $f(u) = u^3 - u$ for u large so that the mapping $u \mapsto u^3$, $H^2(\Omega) \to L^2(\Omega)$ is compact regardless of the dimension N. See [32] and [34] for more general nonlinearities.

A stationary solution of $E(u)$ satisfies the following Euler–Lagrange equation:

$$\begin{cases} \epsilon^2 \triangle u - f(u) = \sigma_\epsilon & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \\ \dfrac{1}{\Omega}\displaystyle\int_\Omega u\,dx = m, \end{cases} \tag{1.2}$$

where $f(u) = F'(u)$, σ_ϵ is a constant, and $\nu(x)$ is the unit outer normal at $x \in \partial\Omega$.

*Received by the editors May 2, 1997; accepted for publication (in revised form) December 8, 1997; published electronically September 3, 1998.

http://www.siam.org/journals/sima/29-6/32066.html

†Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). The research of this author was supported by an earmarked grant from RGC of Hong Kong.

‡Mathematisches Institut, Universität Stuttgart, D-70511 Stuttgart, Germany (winter@mathematik.uni-stuttgart.de). The research of this author was done while visiting the Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong.

Equation (1.2) has been studied extensively by many authors. It was first observed by Modica in [19] that global minimizers u_ϵ of $E(u)$ under $m = \frac{1}{|\Omega|}\int_\Omega u\,dx$ have a transition layer. Namely, there exists an open set $\Gamma \subset \Omega$ such that if a sequence u_ϵ converges, then $u_\epsilon \longrightarrow 1$ on $\Omega\backslash\bar{\Gamma}, u_\epsilon \longrightarrow -1$ on Γ as $\epsilon \longrightarrow 0$, and $\partial\Gamma \cap \bar{\Omega}$ is a minimal surface having constant mean curvature. Kohn and Sternberg in [16] studied local minimizers of the functional without mass conservation by using Γ-convergence. Chen and Kowalczyk [9] proved the existence of local minimizers using a geometric approach. The dynamics of the transition layer solution have been studied by many authors, e.g., Chen [8], Alikakos, Bates, and Fusco [3], Alikakos, Bates, and Chen [2], Alikakos, Fusco, and Kowalczyk [4], Pego [25], etc.

The study of the solution set of (1.2) is the key to understanding the global dynamics, as this has been illustrated by Bates and Fife in one dimension [6], Alikakos, Fusco, and Kowalczyk [4], and Grinfeld and Novick-Cohen [13], [14].

In the one-dimensional case, Grinfeld and Novick-Cohen [13], [14] completely determined all stationary solutions and proved some properties of their connecting orbits. In the higher-dimensional case ($N \geqslant 2$), little is known about stationary solutions except for the transition layer solution. In [32], we first established the existence of boundary spike layer solutions, namely, solutions that are "almost" constant and have a spike on the boundary. More precisely, suppose that $\sqrt{\frac{1}{3}} < m < 1$ and $P_0 \in \partial\Omega$ such that $\nabla_{\tau_{P_0}} H(P_0) = 0$, $(\nabla^2_{\tau_{P_0}} H(P_0)) := G_B(P_0)$ is nondegenerate, where $H(P_0)$ is the mean curvature function at P_0 and $\nabla_{\tau_{P_0}}$ is the tangential derivative at P_0. Then for ϵ sufficiently small there exists a solution u_ϵ of (1.2) such that $u_\epsilon(x) \to m$ for $x \in \bar{\Omega}\backslash\{P_0\}$. Moreover, u_ϵ has only one local minimum P_ϵ, where $P_\epsilon \in \partial\Omega, P_\epsilon \longrightarrow P_0$ and $u_\epsilon(P_\epsilon) \longrightarrow \beta < m$. Multiple boundary spikes are also constructed in [33].

In [34], we established the existence of interior spike layer solutions under some geometric conditions on the domain.

We first introduced the following set: For each $P \in \Omega$, we define

$$\Lambda_P := \left\{ d\mu_P(z) \in M(\partial\Omega) \;\middle|\; \begin{array}{l} \exists \epsilon_k \longrightarrow 0 \text{ such that} \\ d\mu_P(z) = \lim\limits_{\epsilon_k \to 0} \dfrac{e^{-(|z-P|/\epsilon_k)}dz}{\int_{\partial\Omega} e^{(-|z-P|/\epsilon_k)}dz} \end{array} \right\}, \tag{1.3}$$

where $M(\partial\Omega)$ are the bounded Borel measures on $\partial\Omega$, and the convergence is the weak convergence of measures.

A point $P_0 \in \Omega$ is called a *nondegenerate peak* point if it satisfies the following conditions:

(1) $\Lambda_{P_0} = \{d\mu_{P_0}(z)\}$.

(2) There exists $a \in R^N$ such that $\int_{\partial\Omega} e^{\langle z-P_0,a\rangle}(z-P_0)d\mu_{P_0}(z) = 0$ and

$$\int_{\partial\Omega} \left\{ \frac{e^{(-|z-P_0|/\epsilon)} e^{\langle z-P_0,a\rangle}}{\int_{\partial\Omega} e^{(-|z-P_0|/\epsilon)}dz} \right\} (z-P_0)\,dz = O(\epsilon^{\alpha_0})$$

for some $\alpha_0 > 0$. Here and throughout the paper $\langle A, B\rangle$ means the inner product of $A \in R^N$ and $B \in R^N$.

(3) The matrix $G(P) := \left(\int_{\partial\Omega} e^{\langle z-P_0,a\rangle}(z-P_0)_i(z-P_0)_j\,d\mu_{P_0}(z)\right)$ is nondegenerate, where a is given in (2).

Remark. The vector $a \in R^N$ in (2) and (3) is unique. A more geometric characterization of a nondegenerate peak point is the following fact: P_0 is a nondegenerate peak point if and only if $P_0 \in$ int (conv(supp $(d\mu_{P_0})$)) where int (conv(supp $(d\mu_{P_0})$)) is the interior of the convex hull of the support of $d\mu_{P_0}$. Moreover, when Ω is strictly

convex, the maximum point of the distance function, $d(x, \partial\Omega)$, is a nondegenerate peak point. See [29]. This is much in line with the formal analysis done in [27] (but here we don't need $N = 2$).

Under conditions (1)–(3), we proved in [34] that if $\sqrt{\frac{1}{3}} < m < 1$, then for ϵ sufficiently small, there exist solutions u_ϵ of (1.2) with the property that u_ϵ has only one local minimum P_ϵ and $u_\epsilon \to m$ for $x \in \overline{\Omega} \setminus \{P_0\}$, $u_\epsilon(P_\epsilon) \to \beta < m$, $P_\epsilon \to P_0$.

In this paper, we shall construct another kind of solution: bubbles. A bubble solution is a transition layer solution with a spherical interface. More precisely, u_ϵ is a bubble solution if there exists an open ball (with center x_0 and radius r_b) $B_{r_b}(x_0) \subset \Omega$ such that $u_\epsilon \to +1$ in $B_{r_b}(x_0)$ and $u_\epsilon \to -1$ in $\overline{\Omega} \setminus \overline{B_{r_b}(x_0)}$.

Bubble-like solutions have been studied recently by some authors. Alikakos and Fusco [5] and Ward [27] studied the dynamics of bubbles. It was proved that bubble solutions are metastable, and the bubble drifts across the domain with exponentially small velocity without changing shape while maintaining a constant radius (to principal order) to conserve mass. In [27], Ward used matched asymptotics expansions to give a careful but formal (*nonrigorous*) analysis on stationary bubbles for equation (1.2) in a strictly convex domain in R^2 and some special domains in R^3. More precisely, it was shown in [27] that for a strictly convex domain Ω in R^2, the center of a bubble is at an $O(\epsilon)$ distance from the center of the largest inscribed circle in Ω. Some special results for R^3 were also contained in [27]. As far as we know, a rigorous proof of the existence of stationary bubbles in general domains has not been given.

The goal of this paper is to give an *explicit* and *rigorous* construction of bubble-like solutions in general domains. Our analysis is based on the Liapunov–Schmidt reduction method which was used in a similar context by Floer and Weinstein [11] and extended by Oh [23], [24] in the study of semiclassical states of the following nonlinear Schrödinger equation

$$\epsilon^2 \Delta u - V(x)u + u^p = 0, \qquad x \in R^N.$$

There they studied the role of the potential $V(x)$ for the existence of concentrated solutions, and the order of the error is algebraic (i.e., $O(\epsilon)$). Here we have to overcome two additional difficulties. First, the error term is exponentially small, and we use the method of viscosity solutions as introduced in [18] and used in [22] to estimate exponentially small terms. Second, the linearized operator, modulo its approximate kernel, is not uniformly invertible with respect to ϵ (it is uniformly invertible in [11], [23], [24], and [34]). We have to estimate the order of small eigenvalues of the linearized operator (modulo its kernel).

The following is the main result of this paper.

THEOREM 1.1. *Let $P_0 \in \Omega$ and $m \in (-1, \frac{2|B_{d(P_0,\partial\Omega)}(P_0)|}{|\Omega|} - 1)$. Suppose P_0 is a "nondegenerate peak" point. Then for ϵ sufficiently small there exists a solution u_ϵ of (1.2) such that $u_\epsilon \to 1$ in $B_{r_b}(P_0)$ and $u_\epsilon \to -1$ in $\overline{\Omega} \setminus \overline{B_{r_b}(P_0)}$, where r_b is such that*

$$|B_{r_b}(P_0)| = \frac{m+1}{2}|\Omega|. \tag{1.4}$$

Examples. (1) A bubble in a dumbbell domain (see Fig. 1.1).

By explicit computation, we know that P_1 and P_2 are nondegenerate peak points. There are two bubble solutions for (1.2).

(2) Let $\Omega \subset R^2$. If the support of $d\mu_{P_0}(z)$ contains more than two points, then P_0 is a nondegenerate peak point (see Fig. 1.2).

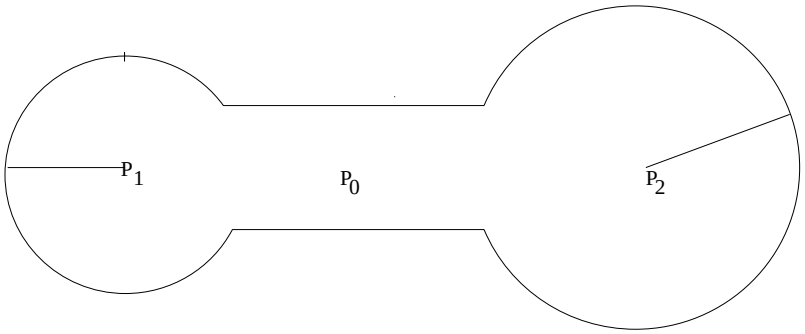

FIG. 1.1. *Dumbbell domain.*

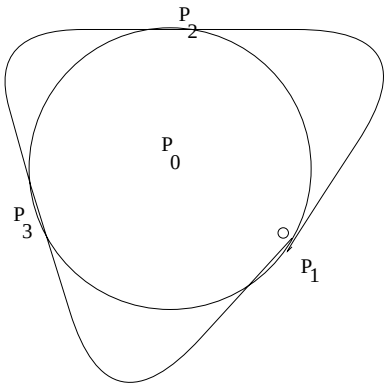

FIG. 1.2. *Support of* $d\mu_{P_0}$ *contains exactly* 3 *points.*

To lay down the proof of Theorem 1.1, we first transform equation (1.2). It is easy to see that equation (1.2) is equivalent to the following:

$$\begin{cases} \epsilon^2 \triangle u + u - u^3 = m - \dfrac{1}{|\Omega|}\displaystyle\int_\Omega u(x)^3\,dx & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \\ \dfrac{1}{|\Omega|}\displaystyle\int_\Omega u\,dx = m. \end{cases} \tag{1.5}$$

We prove Theorem 1.1 in the following steps.

We first study a problem in R^N, namely, the following:

$$\begin{cases} \triangle v + v - v^3 = \sigma & \text{in } R^N, \\ v(0) = \max_{y\in R^N} v(y), v \geqslant \tau_\sigma, v(y) \longrightarrow \tau_\sigma & \text{as } |y| \to +\infty, \end{cases} \tag{1.6}$$

where τ_σ is such that

$$v - v^3 - \sigma = (v - \tau_\sigma)(v - a_\sigma)(b_\sigma - v), \qquad \tau_\sigma < a_\sigma < b_\sigma.$$

Note that as $\sigma \to 0$, $\tau_\sigma \to -1$, $a_\sigma \to 0$, $b_\sigma \to 1$. Moreover, if $\sigma > 0$, we have

$$\int_{\tau_\sigma}^{b_\sigma} [v - v^3 - \sigma]dv > 0.$$

It is well known (see [10] and [26]) that the equation

$$\begin{cases} \triangle w + w(w-a)(b-w) = 0 & \text{in } R^N, \\ w(0) = \max_{z\in R^N} w(z), w(z) > 0, w(z) \to 0 & \text{as } |z| \to \infty \end{cases} \tag{1.7}$$

has a unique solution which is radial if

$$0 < a < b$$

and

$$\int_0^b w(w-a)(b-w)dw > 0.$$

Hence, $\sigma > 0$ fixed and small (1.6) has a unique solution v_σ which is radial.

In section 2, we study the asymptotic behavior of v_σ as $\sigma \to 0$. By a special choice of σ (namely, $\sigma = O(\epsilon)$), we have

$$v_\sigma\left(\frac{|x-P_0|}{\epsilon}\right) \to +1 \text{ in } B_{r_b}(P_0), \qquad v_\sigma\left(\frac{|x-P_0|}{\epsilon}\right) \to -1 \text{ in } \overline{\Omega} \setminus \overline{B_{r_b}(P_0)}$$

for some $r_b > 0$. Hence, v_σ is a bubble solution to (1.6). However, v_σ does not satisfy the boundary condition (which is why we need to introduce the geometric conditions (1)–(3)).

Set

$$\Omega_\epsilon = \{y | \epsilon y \in \Omega\}, \qquad \Omega_{\epsilon,P} = \{y | \epsilon y + P \in \Omega\}.$$

In section 3, we study a function $P_{\Omega_{\epsilon,P}} v_\sigma$ which is a modification of v_σ. It satisfies the Neumann boundary condition on $\partial\Omega_{\epsilon,P}$.

In section 4, we choose σ such that

$$\sigma = m - \frac{1}{|\Omega|}\int_\Omega (P_{\Omega_{\epsilon,P_0}} v_\sigma)^3\, dx. \tag{1.8}$$

We set $P_{\Omega_{\epsilon,P}} v_\sigma = w_{\epsilon,P}$. We use $w_{\epsilon,P}$ as our approximate solution.

In section 5, we set

$$u_\epsilon = w_{\epsilon,P_0+z} + \Phi_{\epsilon,z}, \tag{1.9}$$

where

$$z = \epsilon\left(\frac{1}{2\sqrt{2}} d(P_0, \partial\Omega) a + \tilde{z}\right),$$

and substitute into equation (1.2). We linearize equation (1.2) around w_{ϵ,P_0+z}. The linearized operator is

$$L_\epsilon \Phi = \triangle\Phi + (1 - 3w^2_{\epsilon,P_0+z})\Phi + 3\frac{1}{|\Omega|}\int_\Omega w^2_{\epsilon,P_0+z}\Phi\, dx.$$

The error term $\Phi_{\epsilon,z}$ is exponentially small. We need to obtain the precise exponential asymptotics. This is done in section 5.

In section 6, we use the classical Liapunov–Schmidt reduction procedure. We first define the approximate kernel

$$K_{\epsilon,z} = \text{span}\left\{ \frac{\partial w_{\epsilon,P_0+z}}{\partial z_i} \middle| i = 1, \dots, N \right\} \subset H^2(\Omega_\epsilon)$$

and approximate cokernel

$$C_{\epsilon,z} = \text{span}\left\{ \frac{\partial w_{\epsilon,P_0+z}}{\partial z_i} | i = 1, \dots, N \right\} \subset L^2(\Omega_\epsilon).$$

We solve $\Phi_{\epsilon,z}$ in the approximate kernel. To this end, we need to analyze the small eigenvalues of L_ϵ (modulo $K_{\epsilon,z}$). We will show that these small eigenvalues are of order $O(\epsilon^2)$. Thus $\Phi_{\epsilon,z}$ can be solved. Equation (1.2) is reduced to finite dimensions.

In section 7 we apply a degree-theoretic argument to solve the reduced finite-dimensional problem (in which the nondegeneracy of the peak point P_0 is essential) and complete the proof of Theorem 1.1.

We note that Ward in [27] obtained identities similar to condition (2) about bubbles. In [28], he also derived a similar identity for the location of peaks of localized solutions for a semilinear elliptic equations with Robin boundary conditions. Such kind of identities have also appeared in the analysis of interior spike solutions for the stationary reaction-diffusion equation

$$\begin{cases} \epsilon^2 \triangle u + f(u) = 0 & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 \text{ or } u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.10}$$

See [22], [29], [30], [31], [34], etc.

Throughout this paper, we use C, C_0, C_N, c, etc. to denote various generic constants. The symbols $O(A), o(A)$ mean that $|O(A)| \le C|A|, o(A)/|A| \to 0$, respectively. $A \sim B$ means $A/B \to C$ in some limit. The numbers μ, δ are small positive numbers.

2. Equation in R^N. In this section, we study a parametrized semilinear elliptic equation in R^N.

Let v_σ be the unique solution of the problem

$$\begin{cases} \triangle v + v - v^3 = \sigma & \text{in } R^N, \\ v(0) = \max_{y \in R^N} v(y), v \geqslant \tau_\sigma, v(y) \longrightarrow \tau_\sigma & \text{as } |y| \to +\infty. \end{cases} \tag{2.11}$$

For σ small, let $v - v^3 - \sigma = (v - \tau_\sigma)(v - a_\sigma)(b_\sigma - v)$, where $\tau_\sigma < a_\sigma < b_\sigma$. Then

$$\tau_\sigma = -1 + c_0\sigma + O(\sigma^2), \quad a_\sigma = 0 + c_1\sigma + O(\sigma^2), \quad b_\sigma = 1 + c_2\sigma + O(\sigma^2), \tag{2.12}$$

where c_0, c_1, c_2 are constants.

Let R_σ be the radius such that

$$v_\sigma(R_\sigma) = 0. \tag{2.13}$$

We have the following.

LEMMA 2.1.

$$\sigma R_\sigma = c_b + O(\sigma) \tag{2.14}$$

as $\sigma \to 0$ where $c_b > 0$ is a positive constant.

Proof. We divide the proof into the following steps.

STEP 1. *$R_\sigma \to \infty$ as $\sigma \to 0$.*

We have $v_\sigma \to v_0$ uniformly in any compact set, where v_0 satisfies

$$\begin{cases} \triangle v_0 + v_0 - v_0^3 = 0, \\ v_0(0) = 1, \qquad v_0'(0) = 0. \end{cases} \tag{2.15}$$

This implies $v_0 \equiv 1$ (since v_0 is radial). Therefore, $R_\sigma \to \infty$ as $\sigma \to 0$ and Step 1 is proved.

STEP 2. *$v_\sigma(R_\sigma + s) \to U_0(s)$ in $C^2_{loc}(R)$ as $\sigma \to 0$, where $U_0(s)$ is the unique solution of the ODE*

$$\begin{cases} u'' + u - u^3 = 0, \qquad -\infty < r < +\infty, \\ u(0) = 0, \quad \lim_{r\to-\infty} u(r) = -1, \quad \lim_{r\to+\infty} u(r) = +1. \end{cases} \tag{2.16}$$

Set $\hat{v}_\sigma(|x|) := v_\sigma(x)$ and $\tilde{v}_\sigma(s) := \hat{v}_\sigma(R_\sigma + s)$. Note that $\tilde{v}_\sigma$ satisfies

$$\tilde{v}_\sigma'' + \frac{N-1}{R_\sigma + s}\tilde{v}_\sigma' + \tilde{v}_\sigma - \tilde{v}_\sigma^3 = \sigma. \tag{2.17}$$

Now

$$\frac{1}{R_\sigma + s} \to 0 \tag{2.18}$$

uniformly with respect to s in any compact subset of the real line R since $R_\sigma \to \infty$.

This implies that $\tilde{v}_\sigma \to U_0$ in $C^2_{loc}(R)$, where U_0 satisfies (2.16). Step 2 is thus proved.

STEP 3. *$\sigma R_\sigma = c_b + O(\sigma)$ as $\sigma \to 0$.*

Set $\Phi_\sigma(s) = \tilde{v}_\sigma(s) - U_0(s)$. Then Φ_σ satisfies

$$\Phi_\sigma'' + (1 - 3U_0^2)\Phi_\sigma + O(|\Phi_\sigma|)\Phi_\sigma = \sigma - \frac{N-1}{R_\sigma + s}\tilde{v}_\sigma' \tag{2.19}$$

uniformly in any compact subset of R. This implies

$$\|\Phi_\sigma\|_{C^2_{loc}[-R_\sigma,\infty)} \le C\mathrm{Max}(\sigma, R_\sigma^{-1}). \tag{2.20}$$

Furthermore, U_0' satisfies

$$(U_0')'' + (1 - 3U_0^2)U_0' = 0. \tag{2.21}$$

Multiplying (2.19) by U_0' and (2.21) by Φ_σ, integrating, and taking the difference, we get

$$\begin{aligned}&\Phi_\sigma' U_0' - \Phi_\sigma U_0''|_{-R_\sigma}^{\infty} + \int_{-R_\sigma}^{\infty} O(|\Phi_\sigma|^2) U_0'\, ds \\ &= \sigma \int_{-R_\sigma}^{\infty} U_0'\, ds - \int_{-R_\sigma}^{\infty} \frac{N-1}{R_\sigma + s} \tilde{v}_\sigma' U_0'\, ds. \end{aligned} \tag{2.22}$$

This implies

$$\sigma R_\sigma = \frac{N-1}{2} \int_{-\infty}^{\infty} (U_0')^2\, ds + O(R_\sigma \mathrm{Max}(\sigma^2, R_\sigma^{-2})) \tag{2.23}$$

as $\sigma \to 0$. Therefore, Step 3 is proved and Lemma 2.1 follows. □

Let $U_0(r)$ be the solution of (2.16). We then have the following.

LEMMA 2.2.

$$v_\sigma(r) = U_0(r - R_\sigma) + O(\sigma). \tag{2.24}$$

Proof. Lemma 2.2 follows by Lemma 2.1 and (2.20). □

Next we shall study the eigenvalues associated with the linearized operator

$$L_\sigma \Phi := \triangle \Phi + (1 - 3v_\sigma^2)\Phi,$$

$$L_\sigma : H_N^2(\Omega_{\epsilon,P}) \to L^2(\Omega_{\epsilon,P}),$$

where

$$\Omega_{\epsilon,P} = \{y | \epsilon y + P \in \Omega\}$$

and

$$H_N^2(\Omega_{\epsilon,P}) = \left\{ u \in H^2(\Omega_{\epsilon,P}) \middle| \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega_{\epsilon,P} \right\}.$$

We first consider the operator on R^N:

$$L\Phi := \triangle \Phi + (1 - 3v_\sigma^2)\Phi,$$

$$L : H^2(R^N) \to L^2(R^N).$$

LEMMA 2.3. *For $\sigma > 0$ sufficiently small*

$$\mathrm{Kernel}(L) := X = \mathrm{span} \left\{ \frac{\partial v_\sigma}{\partial y_j} \middle| j = 1, 2, \ldots, N \right\} \subset H^2(R^N).$$

Proof. By [26], L_σ is invertible in the space $H_r^2(R^N) = \{u = u(|y|) \in H^2(R^N)\}$. Similar to the proof of Lemma B.2 in [21], we have Lemma 2.3. □

We now use a perturbation analysis to extend Lemma 2.3 to the operator defined on $\Omega_{\epsilon,P}$. Similar to [32], we introduce a notion of "distance" between two closed subspaces E, F of a Hilbert space $H := L^2(\Omega_\epsilon)$. Following [15], we set

$$\vec{d}\,(E, F) = \sup\{d(x, F) | x \in E, \|x\|_H = 1\}.$$

It is easy to see that $\vec{d}$ is nonsymmetric, $\vec{d}\,(E,F) \leqslant 1$, and that

$$\vec{d}(E,F) = 1 \quad \text{if and only if } E \perp F. \tag{2.25}$$

Moreover, it is not hard to show that

$$\vec{d}\,(E,F) = \vec{d}\,(F^{\perp}, E^{\perp}).$$

Then the following two lemmata are proved in [15].

LEMMA 2.4. *Let A be a self-adjoint operator on a Hilbert space H, I a compact interval in R, and $\{\Psi_1, \dots, \Psi_N\}$ linearly independent normalized elements in $D(A)$. Assume that the following conditions are true:*

(i)

$$\begin{cases} A\Psi_j = \mu_j \Psi_j + r_j, & \|r_j\| < \epsilon', \\ \mu_j \in I, \ j = 1, \dots, N. \end{cases}$$

(ii) *There is a number $a > 0$ such that I is a-isolated in the spectrum of A:*

$$(\sigma(A) \setminus I) \cap (I + (-a, a)) = \emptyset.$$

Then

$$\vec{d}\,(E,F) = \sup\{d(x,F) | x \in E, \|x\|_H = 1\} \leq \frac{N^{1/2}\epsilon'}{a(\lambda_{\min})^{1/2}},$$

where

$$E = \operatorname{span}\{\Psi_1, \dots, \Psi_N\},$$

$$F = \text{closed subspace associated to } \sigma(A) \cap I,$$

$$\lambda_{\min} = \text{the smallest eigenvalue of the matrix } (\langle \Psi_i, \Psi_j \rangle).$$

LEMMA 2.5. *Let $K > 0$, and consider that part of the spectra of two linear operators L and M which lie in $I(\epsilon) = (-\infty, K\epsilon^2)$. Let E and F be the corresponding spectral subspaces. Assume, moreover, that $I(\epsilon)$ is ϵ^2-isolated in $\sigma(L)$ for $\epsilon < \epsilon_0$:*

$$\sigma(L) \cap \left(K\epsilon^2, (K + \overline{a})\epsilon^2\right) = \emptyset$$

for some $\overline{a} > 0$. Then there is a bijection

$$b : \sigma(L) \cap I(\epsilon) \to \sigma(M) \cap I(\epsilon)$$

(counting multiplicities) such that for $\epsilon < \epsilon_0$ the following estimates hold:

$$b(\lambda) - \lambda = O(e^{-C/\epsilon}), \tag{2.26}$$

$$\vec{d}\,(E,F) = O(e^{-C/\epsilon}), \tag{2.27}$$

$$\vec{d}\,(F,E) = O(e^{-C/\epsilon}), \tag{2.28}$$

for some $C > 0$.

The following result gives an approximation of the kernel of the linear operator L_σ defined on $\Omega_{\epsilon,P}$.

LEMMA 2.6. *Suppose that $\sigma = c\epsilon + O(\epsilon^2)$, where $c > 0$ is constant. For $\epsilon > 0$ sufficiently small there exists $C > 0$ such that*

$$\vec{d}(Kernel(L), X_\sigma) = O(e^{-C/\epsilon})$$

and

$$\vec{d}(X_\sigma, Kernel(L)) = O(e^{-C/\epsilon}),$$

where

$$X_\sigma = span\left\{ \frac{\partial v_\sigma}{\partial y_j} \in L^2(\Omega_{\epsilon,P}) \middle| \, j = 1, 2, \ldots, N \right\}$$

is the kernel of L_σ defined on $\Omega_{\epsilon,P}$.

Proof. The lemma is an immediate consequence of Lemma 2.5. □

Now we estimate the eigenvalues of the operator defined on $\Omega_{\epsilon,P}$.

LEMMA 2.7. *Let (τ, Φ_τ) with $\Phi_\tau \in H^2(\Omega_{\epsilon,P})$ be a solution of the following eigenvalue problem*

$$\begin{cases} \Delta\Phi + (1 - 3v_\sigma^2)\Phi = \tau\Phi & \text{in } \Omega_{\epsilon,P}, \\ \dfrac{\partial\Phi}{\partial\nu} = 0 & \text{on } \partial\Omega_{\epsilon,P}. \end{cases} \tag{2.29}$$

Suppose that $\sigma = c\epsilon + O(\epsilon^2)$ and $\Phi_\tau \perp X_\sigma$, where $c > 0$ and

$$X_\sigma := span\left\{ \frac{\partial v_\sigma}{\partial y_j} \in L^2(\Omega_{\epsilon,P}) \middle| \, j = 1, 2, \ldots, N \right\}.$$

Then $|\tau| \geq C\sigma^2$, where C is independent of $\sigma \ll 1$.

Proof. Suppose Lemma 2.7 is not true. Then there exist sequences τ_k and σ_k, $k = 1, 2, \ldots$ such that $\frac{\tau_k}{\sigma_k^2} \to 0$ as $k \to \infty$. Here τ_k is an eigenvalue of L_{σ_k} and $\tau_k \neq 0$, i.e.,

$$L_{\sigma_k}\Phi_k = \tau_k\Phi_k, \qquad \Phi_k \perp X_{\sigma_k},$$

where

$$X_{\sigma_k} = \left\{ \frac{\partial v_{\sigma_k}}{\partial y_j}, \, j = 1, 2, \ldots, N \right\} \subset L^2(\Omega_{\epsilon,P}).$$

Φ_k satisfies

$$\Phi_k'' + \frac{N-1}{r}\Phi_k' + \frac{1}{r^2}\triangle_{S^{N-1}}\Phi_k + (1 - 3v_{\sigma_k}^2)\Phi_k = \tau_k\Phi_k. \tag{2.30}$$

Assume that

$$\|\Phi_k\|_{H^2(\Omega_{\epsilon,P})} = 1. \tag{2.31}$$

Extend Φ_k from $\Omega_{\epsilon,P}$ to a function in R^N such that $\Phi_k = O(e^{-C|y|})$ for $y \in R^N \setminus \Omega_{\epsilon,P}$ and such that the same result holds for the first and second derivatives of Φ_k.

We make the following decomposition:

$$\Phi_k(r) = \sum_{m=1}^{\infty} \Phi_{k,m}(r - R_{\sigma_k}) e_m(\theta), \tag{2.32}$$

where $r = |y|$. Here $e_m(\theta)$ are the eigenfunctions of $\Delta_{S^{N-1}}$, i.e.,

$$\triangle_{S^{N-1}} e_m + \mu_m e_m = 0.$$

Note that $\Phi_k(r) = O(e^{-\delta R_\sigma})$ for $|r - R_\sigma| \geq \beta\delta_0 > 0$. Hence, there exists $\delta > 0$ such that

$$\Phi_{k,m}(r) = \int_{|\theta|=1} \Phi_k(r) e_m(\theta)\, d\theta = O(e^{-\delta R_\sigma}) \quad \text{for } |r - R_\sigma| \geq \beta\delta_0 > 0.$$

It is well known that

$$\mu_0 = 0, \quad \mu_1 = \cdots = \mu_N = N - 1, \quad \mu_{N+1} > N - 1, \quad \mu_m \sim m^2 \text{ as } m \to \infty.$$

Furthermore, $\Phi_{k,m}$ satisfies

$$\Phi''_{k,m} + \frac{N-1}{R_{\sigma_k} + s}\Phi'_{k,m} - \frac{\mu_m}{(R_{\sigma_k} + s)^2}\Phi_{k,m} + (1 - 3\tilde{v}^2_{\sigma_k})\Phi_{k,m} = \tau_k \Phi_{k,m} \tag{2.33}$$

in $[-R_\sigma, \infty)$. Note that $\tilde{v}'_{\sigma_k}$ satisfies

$$(\tilde{v}'_{\sigma_k})'' + \frac{N-1}{R_{\sigma_k} + s}(\tilde{v}'_{\sigma_k})' + (1 - 3\tilde{v}^2_{\sigma_k})\tilde{v}'_{\sigma_k} = \frac{N-1}{(R_{\sigma_k} + s)^2}\tilde{v}'_{\sigma_k} \quad \text{in } [-R_\sigma, \infty). \tag{2.34}$$

We next decompose $\Phi_{k,m}$ into

$$\Phi_{k,m} = C_{k,m}\tilde{v}'_{\sigma_k} + \Phi^2_{k,m},$$

where

$$\Phi^2_{k,m} \perp \tilde{v}'_{\sigma_k}.$$

Multiplying (2.33) by $\tilde{v}'_{\sigma_k}$, multiplying (2.34) by $\Phi_{k,m}$, taking the difference, and integrating we obtain

$$\int_{-R_\sigma}^{\infty} \left(\tau_k + \frac{\mu_m - (N-1)}{(R_{\sigma_k} + s)^2}\right) \Phi_{k,m}\tilde{v}'_{\sigma_k}\, ds = O(e^{-\delta R_\sigma}). \tag{2.35}$$

Since $\tau_k = o(1)\sigma_k^2$, we have

$$C_{k,m} = O\left(\frac{R_\sigma^2 e^{-\delta R_\sigma}}{\mu_m}\right). \tag{2.36}$$

Note that $\Phi^2_{k,m}$ satisfies

$$(\Phi^2_{k,m})'' + \frac{N-1}{R_{\sigma_k} + s}(\Phi^2_{k,m})' + (1 - 3\tilde{v}^2_{\sigma_k})(\Phi^2_{k,m}) = \frac{\mu_m}{(R_{\sigma_k} + s)^2}\Phi^2_{k,m}$$

$$+ \tau_k \Phi^2_{k,m} + \frac{\mu_m - (N-1)}{(R_{\sigma_k} + s)^2} C_{k,m}\tilde{v}'_{\sigma_k} + \tau_k C_{k,m}\tilde{v}'_{\sigma_k} \quad \text{in } [-R_\sigma, \infty). \tag{2.37}$$

Multiplying (2.37) by $\Phi_{k,m}^2$ and integrating by parts, we have

$$\int_{-R_\sigma}^{\infty}\left[((\Phi_{k,m}^2)')^2-(1-3\tilde{v}_{\sigma_k}^2)(\Phi_{k,m}^2)^2+\left(\frac{\mu_m}{(R_{\sigma_k}+s)^2}+\tau_k\right)(\Phi_{k,m}^2)^2\right.$$

$$\left.-\frac{N-1}{R_{\sigma_k}+s}(\Phi_{k,m}^2)'\Phi_{k,m}^2\right]ds=O(e^{-\delta R_\sigma}). \tag{2.38}$$

Since $\Phi_{k,m}^2\perp\tilde{v}_{\sigma_k}'$, we have that

$$\int_{-R_\sigma}^{\infty}\left[((\Phi_{k,m}^2)')^2-(1-3\tilde{v}_{\sigma_k}^2)(\Phi_{k,m}^2)^2+\left(\frac{\mu_m}{(R_{\sigma_k}+s)^2}+\tau_k\right)(\Phi_{k,m}^2)^2\right.$$

$$\left.-\frac{N-1}{R_{\sigma_k}+s}(\Phi_{k,m}^2)'\Phi_{k,m}^2\right]ds\geq\int_{-R_\sigma}^{\infty}\sigma_0[((\Phi_{k,m}^2)')^2+(\Phi_{k,m}^2)^2]\,ds. \tag{2.39}$$

(Suppose not. Then there exists a subsequence, again denoted by $\Phi_{k,m}^2$, such that $\Phi_{k,m}^2\to\Phi_0$ in $H^1(-\infty,\infty)$, where $\int_{-\infty}^{\infty}((\Phi_0)')^2+(\Phi_0)^2=1$ and $\Phi_0\perp U_0'$. Furthermore, Φ_0 satisfies

$$\int_{-\infty}^{\infty}[((\Phi_0)')^2-(1-3(U_0)^2)(\Phi_0)^2]\,ds=0.$$

This is a contradiction since the operator $-\Delta+(1-3U_0^2)$ is positive and has the kernel $\operatorname{span}(U_0')$.)

Hence, combining (2.38) and (2.39),

$$\int_{-R_\sigma}^{\infty}[((\Phi_{k,m}^2)')^2+(\Phi_{k,m}^2)^2]\,ds=O\left(\frac{e^{-\delta R_\sigma}}{R_\sigma^2+\mu_m}\right)=O\left(\frac{e^{-\delta R_\sigma}}{\mu_m}\right),$$

or, in other words,

$$\|\Phi_{k,m}^2\|_{H^1([-R_\sigma,\infty))}^2=O(e^{-\delta R_\sigma}/\mu_m).$$

By elliptic regularity theory we also know that

$$\|\Phi_{k,m}^2\|_{H^2([-R_\sigma,\infty))}=O(e^{-\delta R_\sigma}/\mu_m).$$

Hence,

$$\|\Phi_{k,m}^2\|_{H^2(R^N)}^2=O(R_\sigma^{N-1}e^{-\delta R_\sigma}/\mu_m). \tag{2.40}$$

By (2.36) and (2.40),

$$\|\Phi_k\|_{H^2(\Omega_{\epsilon,P})}^2\leq\sum_{m=N+1}^{\infty}\|\Phi_{k,m}\|_{H^2(R^N)}^2=O(R_\sigma^{N+1}e^{-\delta R_\sigma})\sum_{m=N+1}^{\infty}\frac{1}{\mu_m}=o(1).$$

This is a contradiction! The proof is finished. ∎

COROLLARY 2.1. *For all $\Phi\in H_N^2(\Omega_{\epsilon,P})$, where Φ is orthogonal to the kernel of L_σ, we have*

$$\|L_\sigma\Phi\|_{L^2(\Omega_{\epsilon,P})}\geq C\sigma^2\|\Phi\|_{H^2(\Omega_{\epsilon,P})}, \tag{2.41}$$

where $C>0$ is independent of $\sigma\ll 1$.

Proof. Let $L_\sigma \Phi = \sigma^2 f$. Then by Lemma 2.4, we have

$$\|\sigma^2 f\|_{L^2(\Omega_{\epsilon,P})} \geq C\sigma^2 \|\Phi\|_{L^2(\Omega_{\epsilon,P})}.$$

On the other hand, Φ satisfies

$$\Delta\Phi - 2\Phi = (3v_\sigma^2 - 3)\Phi + \sigma^2 f.$$

Hence, by elliptic regularity estimates, we have

$$\|\Phi\|_{H^2(\Omega_{\epsilon,P})} \leq C(\|\Phi\|_{L^2(\Omega_{\epsilon,P})} + \sigma^2 \|f\|_{L^2(\Omega_{\epsilon,P})})$$

$$\leq C\|f\|_{L^2(\Omega_{\epsilon,P})} \leq C\sigma^{-2}\|L_\sigma \Phi\|_{L^2(\Omega_{\epsilon,P})}.$$

The corollary is thus proved. □

Finally, we study the asymptotic behavior of v_σ.

LEMMA 2.8. *For σ sufficiently small, we have*

$$v_\sigma - \tau_\sigma = C\left(\frac{r}{R_\sigma}\right)^{-\frac{N-1}{2}} e^{\bar\nu_\sigma(R_\sigma - r)}(1 + O(\sigma)) \quad \text{for } r \geq R_\sigma, \tag{2.42}$$

where τ_σ is defined in section 2 (note that $\tau_\sigma \to -1$ as $\sigma \to 0$), $C \neq 0$ is a generic constant, and

$$\bar\nu_\sigma = \sqrt{3\tau_\sigma^2 - 1}.$$

Proof. We use matched asymptotics as in [27], although the proof can be made rigorous by ODE arguments and the maximum principle.

Let $\hat v_\sigma = v_\sigma - \tau_\sigma$. Linearizing (2.1) around τ_σ, we have that $\hat v_\sigma$ satisfies

$$\hat v_\sigma'' + \frac{N-1}{r}\hat v_\sigma' - \bar\nu_\sigma^2 \hat v_\sigma + O(\hat v_\sigma^2) = 0.$$

Note that $\bar\nu_\sigma = \sqrt{2} + O(\sigma)$, and the exact solution of the following problem

$$u'' + \frac{N-1}{r}u' - \bar\nu_\sigma^2 u = 0, \quad u(R_\sigma) = -\tau_\sigma, \quad r \geq R_\sigma, \quad u(r) \to 0 \text{ as } r \to \infty$$

is $(-\tau_\sigma)(\frac{r}{R_\sigma})^{1-N/2} K_m(\bar\nu_\sigma r)(K_m(\bar\nu_\sigma R_\sigma))^{-1}$, where $m = (N-2)/2$ and $K_m(z)$ is the modified Bessel function of the second kind of order m.

Since

$$K_m(z) = \left(1 + O\left(\frac{1}{z}\right)\right)(\pi/(2z))^{1/2} e^{-z}$$

as $z \to \infty$, we have

$$\hat v_\sigma = C_\sigma \left(\frac{r}{R_\sigma}\right)^{1-\frac{N}{2}} \left(\frac{\pi}{2r}\right)^{\frac{1}{2}} e^{-\bar\nu_\sigma r}(1 + O(\sigma)) \quad \text{as } r \to \infty, \tag{2.43}$$

where C_σ may depend on σ. On the other hand, let $r = R_\sigma + s$; then

$$\hat v_\sigma = C_0 e^{-\bar\nu_\sigma s}(1 + O(\sigma)) \tag{2.44}$$

for s large, where $C_0 \neq 0$ is a generic constant. Combining (2.43) and (2.44), we have

$$C_\sigma = C_0 \pi^{-1/2} (2\bar{\nu}_\sigma R_\sigma)^{1/2} e^{\bar{\nu}_\sigma R_\sigma}.$$

Hence Lemma 2.8 is proved. □

In the following, it will be more convenient to rewrite (2.42) as follows:

$$v_\sigma - \tau_\sigma = C\sigma^l r^{-\frac{N-1}{2}} e^{\bar{\nu}_\sigma (R_\sigma - r)}(1 + O(\sigma)) \quad \text{for } r \geq R_\sigma, \tag{2.45}$$

where $l = -(N-1)/2$.

3. The projection of v_σ. In this section, we construct a modified function $P_{\Omega_{\epsilon,P}} v_\sigma$. It is close to v_σ and satisfies the Neumann boundary condition. Furthermore, we provide an error estimate for $\Psi_{\epsilon,P} = v_\sigma - P_{\Omega_{\epsilon,P}} v_\sigma$.

Let $\Psi_{\epsilon,P}$ be the unique solution of

$$\begin{cases} \epsilon^2 \triangle u - \bar{\nu}_\sigma^2 u = 0 & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v_\sigma((x-P)/\epsilon)}{\partial \nu} & \text{on } \partial\Omega. \end{cases} \tag{3.46}$$

Define $P_{\Omega_{\epsilon,P}} v_\sigma := v_\sigma - \Psi_{\epsilon,P}$. Later, in section 4, we will show that for every small $\epsilon > 0$ there exists exactly one $\sigma = \sigma(\epsilon)$ satisfying a certain nonlinear equation, and, furthermore, we have $\sigma(\epsilon) = \gamma_0 \epsilon + O(\epsilon^2)$ as $\epsilon \to 0$, where γ_0 is some positive constant. In this section we will write σ and ϵ with the understanding that this relation holds. We set

$$\nu_\epsilon = \bar{\nu}_{\sigma(\epsilon)}.$$

Note that by (2.45) on the boundary of $\partial\Omega$,

$$v_\sigma\left(\frac{x-P}{\epsilon}\right) = \tau_\sigma + C\sigma^l \left(\frac{|x-P|}{\epsilon}\right)^{-\frac{N-1}{2}} e^{-\nu_\epsilon(|x-P|/\epsilon - R_\sigma)}(1 + O(\sigma)).$$

In particular, we have the following asymptotic expansion of $\Psi_{\epsilon,P}$. A proof can be found in [34].

LEMMA 3.1. *For ϵ sufficiently small, we have*

$$\begin{aligned} \Psi_{\epsilon,P}(x) &= (C_N + O(\epsilon))\epsilon^{l_1} e^{\nu_\epsilon R_\sigma} \\ &\quad \times \int_{\partial\Omega} \left\{ e^{-\nu_\epsilon \frac{|t-P|+|t-x|}{\epsilon}} |t-P|^{-\frac{N-1}{2}} |t-x|^{-\frac{N-1}{2}} \frac{\langle t-x, \nu\rangle}{|t-x|} \right\} dt, \end{aligned} \tag{3.47}$$

where l_1 is a rational number.

Let us introduce the following notation:

$$\tilde{\varphi}_{\epsilon,P}(P) := \left[\int_0^\infty (\tau_\sigma^2 - v_\sigma^2(r)) v_\sigma'(r) u_\sigma'(r) r^{N-1}\, dr \right] \Psi_{\epsilon,P}(P), \tag{3.48}$$

where u_σ is the unique solution of

$$\Delta u - \nu_\epsilon^2 u = 0,\ u(0) = 1,\ u > 0,\ u = u(r) \quad \text{for } r \in [0, \infty). \tag{3.49}$$

We have the following key computations.

LEMMA 3.2. *Let P_0 be a nondegenerate peak point of Ω, and $\alpha_0 > 0$ is given by condition (2) in section 1. Suppose $P_\epsilon = P_0 + \epsilon(\frac{a}{2\sqrt{2}} d(P_0, \partial\Omega) + \tilde{z})$ with $|\tilde{z}| = O(\epsilon^\alpha), 0 < \alpha < \alpha_0$. Then*

$$\begin{aligned} L_j(\epsilon, \tilde{z}) &:= \int_{\Omega_{\epsilon,P_\epsilon}} (\tau_\sigma^2 - v_\sigma^2)\Psi_{\epsilon,P_\epsilon} \frac{\partial v_\sigma}{\partial y_j} \\ &= L_j(\tilde{z})\tilde{\varphi}_{\epsilon,P_\epsilon}(P_\epsilon) + O\big(\tilde{\varphi}_{\epsilon,P_\epsilon}(P_\epsilon)\epsilon^{\min(1,2\alpha,\alpha_0)}\big), \end{aligned} \tag{3.50}$$

where $L(\tilde{z}) := (L_1(\tilde{z}), \dots, L_N(\tilde{z}))$ is a matrix which satisfies

$$L_j(\tilde{z}) = \gamma \frac{\int_{\partial\Omega} e^{\langle t-P_0, a\rangle} \langle t - P_0, \tilde{z}\rangle \big(t_j - P_{0,j}\big)\, d\mu_{P_0}(t)}{\int_{\partial\Omega} e^{\rangle t-P_0, a\langle}\, d\mu_{P_0}(t)},$$

where $\gamma \neq 0$ is a constant depending on N and $d(P_0, \partial\Omega)$ only.

Proof. Since the proof is quite similar to the proof of Lemma 3.4 in [34], we will merely sketch it. Note that

$$L_j(\epsilon, \tilde{z}) = \int_{\Omega_{\epsilon,P_\epsilon}} (\tau_\sigma^2 - v_\sigma^2)\Psi_{\epsilon,P_\epsilon} \frac{\partial v_\sigma}{\partial y_j}$$

$$= \int_0^\infty (\tau_\sigma^2 - v_\sigma^2) v_\sigma' r^{N-1}\, dr \int_{|\theta|=1} \theta_j \Psi_{\epsilon,P_\epsilon}(\epsilon y + P_\epsilon)\, d\theta + O(\tilde{\varphi}^{1+\mu}_{\epsilon,P_\epsilon}(P_\epsilon)).$$

However (let $x = \epsilon y + P_\epsilon$),

$$\begin{aligned} &\Psi_{\epsilon,P_\epsilon}(\epsilon y + P_\epsilon) \\ &= \Psi_{\epsilon,P_\epsilon}(P_\epsilon) \frac{\int_{\partial\Omega} \Big\{ e^{-\nu_\epsilon \frac{|t-P_\epsilon|+|t-x|}{\epsilon}} |t-P_\epsilon|^{-\frac{N-1}{2}} |t-x|^{-\frac{N-1}{2}} (\langle t-x, \nu\rangle/|t-x|) \Big\}\, dt}{\int_{\partial\Omega} \Big\{ e^{-\nu_\epsilon \frac{2|t-P_\epsilon|}{\epsilon}} |t-P_\epsilon|^{-\frac{N-1}{2}} |t-x|^{-\frac{N-1}{2}} (\langle t-x, \nu\rangle/|t-x|) \Big\}\, dt} \\ &= \Psi_{\epsilon,P_\epsilon}(P_\epsilon) \\ &\quad \times \frac{\int_{\partial\Omega} \Big\{ e^{-\nu_\epsilon \frac{2|t-P_\epsilon|}{\epsilon}} e^{\nu_\epsilon \langle \frac{t-P_\epsilon}{|t-P_\epsilon|}, y\rangle} |t-P_\epsilon|^{-\frac{N-1}{2}} |t-x|^{-\frac{N-1}{2}} (\langle t-x, \nu\rangle/|t-x|) \Big\}\, dt}{\int_{\partial\Omega} \Big\{ e^{-\nu_\epsilon \frac{2|t-P_\epsilon|}{\epsilon}} |t-P_\epsilon|^{-\frac{N-1}{2}} |t-x|^{-\frac{N-1}{2}} (\langle t-x, \nu\rangle/|t-x|) \Big\}\, dt} \\ &= \Psi_{\epsilon,P_\epsilon}(P_\epsilon) \int_{\partial\Omega} e^{\langle t-P_0, a\rangle} e^{\nu_\epsilon \langle \frac{t-P_0}{|t-P_0|}, y\rangle}\, d\mu^a_{P_0}(t)(1 + O(\epsilon^{\alpha_0})) \end{aligned}$$

by condition (2) in section 1, where

$$d\mu^a_P(t) = \lim_{\epsilon\to 0} \frac{e^{-2\nu_\epsilon |t-P_\epsilon|/\epsilon}\, dt}{\int_{\partial\Omega} e^{-2\nu_\epsilon |t-P_\epsilon|/\epsilon}\, dt}.$$

Hence,

$$\begin{aligned} L_j(\epsilon, \tilde{z}) &= \left[\int_0^\infty (\tau_\sigma^2 - v_\sigma^2(r)) v_\sigma'(r) u_\sigma'(r) r^{N-1}\, dr \right] \Psi_{\epsilon,P_\epsilon}(P_\epsilon) L_j(\tilde{z}) \\ &\quad + O(\tilde{\varphi}_{\epsilon,P_\epsilon}(P_\epsilon)\epsilon^{\min(1,2\alpha,\alpha_0)}) \\ &= L_j(\tilde{z})\tilde{\varphi}_{\epsilon,P_\epsilon}(P_\epsilon) + O\big(\tilde{\varphi}_{\epsilon,P_\epsilon}(P_\epsilon)\epsilon^{\min(1,2\alpha,\alpha_0)}\big). \qquad \square \end{aligned}$$

4. Choosing σ. In this section we choose σ and give an asymptotic expansion including error estimate for its behavior as $\epsilon \to 0$.

Let $P_{\Omega_{\epsilon,P}} v_\sigma$ be defined as in section 3. Set

$$\sigma = m - \frac{1}{|\Omega|} \int_\Omega P_{\Omega_{\epsilon,P}} v_\sigma \, dx. \tag{4.51}$$

We show that this equation has a unique solution σ if ϵ is small enough.

Note that

$$\int_\Omega (P_{\Omega_{\epsilon,P}} v_\sigma)^3 \, dx = \int_\Omega v_\sigma^3 \, dx + \int_\Omega [(P_{\Omega_{\epsilon,P}} v_\sigma)^3 - v_\sigma^3] \, dx.$$

Now choose R_σ such that, for $r_b = \epsilon R_\sigma$,

$$\frac{|B_{r_b}| - |\Omega \setminus B_{r_b}|}{|\Omega|} = m + O(\sigma) + O(\epsilon) \tag{4.52}$$

as $\sigma, \epsilon \to 0$. This implies

$$\frac{1}{|\Omega|} \int_\Omega v_\sigma^3 \, dx = m + c\sigma + O(\sigma^2)$$

for some constant $c > 0$. Furthermore, there exists $C > 0$ such that

$$\int_\Omega \left[(P_{\Omega_{\epsilon,P}} v_\sigma)^3 - v_\sigma^3 \right] dx \le C \int_\Omega |\Psi_{\epsilon,P}| = O(e^{-C/\epsilon}).$$

Therefore, by the implicit function theorem, if ϵ is small enough, there exists exactly one solution σ of (4.51). Furthermore, this σ satisfies

$$\sigma = \gamma_0 \epsilon + O(\epsilon^2) \tag{4.53}$$

as $\epsilon \to 0$, where $\gamma_0 = c_b / r_b$.

5. Technical framework. In this section, we set up the technical framework to solve equation (1.2). As we mentioned in section 1, this framework was originated by Floer and Weinstein [11] and later used by Oh [23], [24]. We modified their approach to the Cahn–Hilliard equation in [32], [33], and [34]. We shall follow [34].

Without loss of generality, we assume that $P_0 = 0 \in \Omega$ is a nondegenerate peak point, i.e.,

(1) $\Lambda_0 = \{d\mu_0(t)\}$,

(2) $\exists a \in R^N$ such that

$$\int_{\partial\Omega} e^{\langle t,a \rangle} t \, d\mu_0(t) = 0$$

and

$$\int_{\partial\Omega} \left\{ \frac{e^{-\frac{|t|}{\epsilon}} e^{\langle t,a \rangle}}{\int_{\partial\Omega} e^{-\frac{|t|}{\epsilon}} \, dt} \right\} t \, dt = O(\epsilon^{\alpha_0})$$

for some $\alpha_0 > 0$,

(3) the matrix $G(0) := \left(\int_{\partial\Omega} e^{\langle t,a \rangle} (t_i t_j) \, d\mu_0(t) \right)$ is nondegenerate.

Let $z = \epsilon(\frac{a}{2\sqrt{2}} d(0, \partial\Omega) + \tilde{z})$, where $|\tilde{z}| < \epsilon^\alpha$, with $0 < \alpha < 1$ to be chosen later.
We assume that $\sigma = \sigma(\epsilon)$, where $\sigma(\epsilon)$ is defined in section 4.
Define $H_\epsilon : H^2_N(\Omega_\epsilon) \to L^2(\Omega_\epsilon)$ by

$$H_\epsilon(u) := \triangle u + u - u^3 - m + \frac{1}{|\Omega_\epsilon|} \int_{\Omega_\epsilon} u^3 \, dy, \tag{5.54}$$

where

$$H^2_N(\Omega_\epsilon) := \left\{ u \in H^2(\Omega_\epsilon) : \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega_\epsilon \right\}.$$

We are looking for a nontrivial zero of (5.1). We make the ansatz

$$u = P_{\Omega_{\epsilon,z}} v_\sigma + \Phi_\epsilon,$$

where Φ_ϵ is now the unknown. Recall that we set $w_{\epsilon,z} = P_{\Omega_{\epsilon,z}} v_\sigma$. We assume that $\epsilon > 0$ is small and Φ_ϵ is small in $C^2_{loc}(\Omega_\epsilon)$. We shall see that solutions of this particular form correspond to bubble solutions of (1.2), where the center of the bubble is located near zero. Inserting this into the equation gives

$$\triangle\Phi_\epsilon + \Phi_\epsilon + \triangle(P_{\Omega_{\epsilon,z}} v_\sigma) + P_{\Omega_{\epsilon,z}} v_\sigma - (P_{\Omega_{\epsilon,z}} v_\sigma + \Phi_\epsilon)^3$$

$$= m - \frac{1}{|\Omega_\epsilon|} \int_{\Omega_\epsilon} (P_{\Omega_{\epsilon,z}} v_\sigma + \Phi_\epsilon)^3 \, dy.$$

Recall that

$$\triangle(P_{\Omega_{\epsilon,z}} v_\sigma) + P_{\Omega_{\epsilon,z}} v_\sigma = \Delta v_\sigma - \triangle\Psi_{\epsilon,z} + v_\sigma - \Psi_{\epsilon,z}$$

$$= v_\sigma^3 + \sigma - 3\tau_\sigma^2 \Psi_{\epsilon,z}.$$

This implies

$$\triangle\Phi_\epsilon + \Phi_\epsilon + v_\sigma^3 + \sigma - 3\tau_\sigma^2 \Psi_{\epsilon,z} - (P_{\Omega_{\epsilon,z}} v_\sigma + \Phi_\epsilon)^3$$

$$= m - \frac{1}{|\Omega_\epsilon|} \int_{\Omega_\epsilon} (P_{\Omega_{\epsilon,z}} v_\sigma + \Phi_\epsilon)^3 \, dy.$$

By the choice of σ,

$$L_\epsilon \Phi_\epsilon + v_\sigma^3 - 3\tau_\sigma^2 \Psi_{\epsilon,z} - (v_\sigma - \Psi_{\epsilon,z})^3 + N_{\epsilon,z}(\Phi_\epsilon) = 0,$$

where

$$L_\epsilon \Phi_\epsilon := \triangle\Phi_\epsilon + \Phi_\epsilon - 3(P_{\Omega_{\epsilon,z}} v_\sigma)^2 \Phi_\epsilon + 3 \frac{1}{|\Omega_\epsilon|} \int_{\Omega_\epsilon} (P_{\Omega_{\epsilon,z}} v_\sigma)^2 \Phi_\epsilon \, dy$$

and

$$N_{\epsilon,z}(\Phi_\epsilon) = -3 P_{\Omega_{\epsilon,z}} v_\sigma \Phi_\epsilon^2 - \Phi_\epsilon^3 + \frac{1}{|\Omega_\epsilon|} \int_{\Omega_\epsilon} [3 P_{\Omega_{\epsilon,z}} v_\sigma \Phi_\epsilon^2 + \Phi_\epsilon^3] \, dy.$$

Recalling that $\Phi_\epsilon \to 0$ as $\epsilon \to 0$ in $C^2_{loc}(\Omega_\epsilon)$, we finally arrive at

$$L_\epsilon \Phi_\epsilon + 3(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z} + N_{\epsilon,z}(\Phi_\epsilon) + M_{\epsilon,z}(\Psi_{\epsilon,z}) = 0,$$

where

$$M_{\epsilon,z}(\Psi_{\epsilon,z}) = -3v_\sigma \Psi^2_{\epsilon,z} + \Psi^3_{\epsilon,z}.$$

It is easy to see the following.

LEMMA 5.1. *For ϵ sufficiently small,*

$$\|N_{\epsilon,z}(\Phi_\epsilon)\|_{L^2(\Omega_{\epsilon,z})} \leqslant c\|\Phi_\epsilon\|^2_{H^2(\Omega_{\epsilon,z})},$$
$$\|M_{\epsilon,z}(\Psi_{\epsilon,z})\|_{L^2(\Omega_{\epsilon,z})} \leqslant c\|\Psi_{\epsilon,z}\|^2_{L^2(\Omega_{\epsilon,z})} \leq c|\tilde{\varphi}_{\epsilon,z}(z)|.$$

Furthermore,

$$\|N_{\epsilon,z}(\Phi^{(1)}_\epsilon) - N_{\epsilon,z}(\Phi^{(2)}_\epsilon)\|_{L^2(\Omega_{\epsilon,z})} \leq c\|\Phi^{(1)}_\epsilon - \Phi^{(2)}_\epsilon\|^2_{H^2(\Omega_{\epsilon,z})}.$$

It remains then to estimate the term $3(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z}$. We have the following.

LEMMA 5.2. *For ϵ sufficiently small, we have*

$$\|(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z}\|^2_{L^2(\Omega_\epsilon,z)} \leq C|\tilde{\varphi}_{\epsilon,z}(z)|^{1.5}. \tag{5.55}$$

Proof. In fact,

$$(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z} = e^{\nu_\epsilon R_\sigma} u_\sigma (v_\sigma^2 - \tau_\sigma^2) u_\sigma^{-1} e^{-\nu_\epsilon R_\sigma}\Psi_{\epsilon,z},$$

where u_σ is the unique radial solution of $\Delta u - \nu_\epsilon^2 u = 0, u(0) = 1, u > 0$.

Now

$$|u_\sigma(v_\sigma^2 - \tau_\sigma)| \leq e^{(\nu_\epsilon+\delta)R_\sigma}, \tag{5.56}$$

where $\delta > 0$ is small. Furthermore, by Lemma 3.1, (note that $\epsilon y + z = x$),

$$e^{-\nu_\epsilon R_\sigma}\Psi_{\epsilon,z}$$

$$= (C_N + O(\epsilon)) \int_{\partial\Omega} \left\{ e^{-\nu_\epsilon \frac{|t-z|+|t-x|}{\epsilon}} |t-z|^{-\frac{N-1}{2}} |t-x|^{-\frac{N-1}{2}} \frac{\langle t-x, \nu\rangle}{|t-x|} \right\} dt$$

$$\leq e^{\nu_\epsilon R_\sigma} e^{-2\nu_\epsilon d(z,\partial\Omega)/\epsilon} e^{(\nu_\epsilon+\delta)|y|}.$$

Therefore,

$$|u_\sigma^{-1} e^{-\nu_\epsilon R_\sigma}\Psi_{\epsilon,z}| \leq Ce^{-2\nu_\epsilon\, d(z,\partial\Omega)/\epsilon} e^{(\nu_\epsilon+\delta)R_\sigma}. \tag{5.57}$$

Combining (5.56) and (5.57), we obtain

$$|(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z}| \leq Ce^{-2\nu_\epsilon(d(z,\partial\Omega)-\epsilon R_\sigma)+2(\delta+\nu_\epsilon)R_\sigma}$$

$$\leq C(\tilde{\varphi}_{\epsilon,z}(z))^{0.8}.$$

This implies

$$\|(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z}\|^2_{L^2(\Omega_{\epsilon,z})} \leq \tilde{\varphi}_{\epsilon,z}(z)^{1.5}.$$

The lemma is thus proved. □

6. Reduction to finite dimensions: Fredholm inverses. In this section, we show that $H'_\epsilon(w_{\epsilon,z})$, modulo its approximate kernel, is an invertible linear operator if ϵ is small enough. Moreover, we show that the operator norm of the inverse operator is bounded by $C\epsilon^{-2}$. (Note that in [11], [23], [24], and [34] the operator norm of the inverse operator is uniformly bounded.)

Set

$$K_{\epsilon,z} = \text{ span}\left\{ \frac{\partial w_{\epsilon,z}}{\partial z_i} | i = 1, \dots, N \right\} \subset H^2_N(\Omega_\epsilon) \tag{6.58}$$

and

$$C_{\epsilon,z} = \text{ span } \left\{ \frac{\partial w_{\epsilon,z}}{\partial z_i} | i = 1, \dots, N \right\} \subset L^2(\Omega_\epsilon). \tag{6.59}$$

$K_{\epsilon,z}$ is called the approximate kernel, while $C_{\epsilon,z}$ is called the approximate cokernel. Note that a function $\Phi \in$ cokernel of $H'_\epsilon(w_{\epsilon,z})$ if and only if for all $\psi \in H^2_N(\Omega_\epsilon)$ we have

$$\int_{\Omega_\epsilon} \Phi H'_\epsilon(w_{\epsilon,z})\psi \, dy = 0.$$

Integrating by parts, we have

$$\begin{aligned} &\int_{\partial\Omega_\epsilon} \psi \frac{\partial \Phi}{\partial \nu} \, do + \int_{\Omega_\epsilon} [(\Delta\Phi + (1 - 3w^2_{\epsilon,z})\Phi)\psi] \, dy \\ &\quad + 3\frac{1}{|\Omega_\epsilon|} \int_{\Omega_\epsilon} \Phi \, dy \int_{\Omega_\epsilon} w^2_{\epsilon,z}\psi \, dy = 0, \quad \text{for all } \psi \in H^2_N(\Omega_\epsilon). \end{aligned}$$

Hence, $\Phi \in$ cokernel of $H'_\epsilon(w_{\epsilon,z})$ if and only if

$$\begin{cases} \Delta\Phi + (1 - 3w^2_{\epsilon,z})\Phi + 3w^2_{\epsilon,z}\dfrac{1}{|\Omega_\epsilon|}\displaystyle\int_{\Omega_\epsilon} \Phi \, dy = 0 & \text{in } \Omega_\epsilon, \\ \dfrac{\partial \Phi}{\partial \nu} = 0 & \text{in } \partial\Omega_\epsilon. \end{cases}$$

Observe also that $\text{span}\{(\partial v_\sigma/\partial y_i) | i = 1, \dots, N\}$ is the kernel of L, where L is the linear operator defined as

$$L\Phi := \Delta\Phi + \Phi - 3v^2_\sigma\Phi, \qquad \Phi \in H^2(R^N).$$

Our main result in this section can be stated as follows.

PROPOSITION 6.1. *There exist positive constants ϵ_1, λ such that, for all $\epsilon \in (0, \epsilon_1)$,*

$$\|L_{\epsilon,z}\Phi\|_{L^2(\Omega_\epsilon)} \geqslant \lambda\sigma^2\|\Phi\|_{H^2(\Omega_\epsilon)} \tag{6.60}$$

for all $|z| \leq C\epsilon$ and for all $\Phi \in K^\perp_{\epsilon,z}$, where

$$L_{\epsilon,z} = \pi_{\epsilon,z} \circ H'_\epsilon(w_{\epsilon,z}), \tag{6.61}$$

and $\pi_{\epsilon,z}$ is the L^2-orthogonal projection from $L^2(\Omega_\epsilon)$ to $C^\perp_{\epsilon,z}$.

The next proposition gives the surjectivity of $L_{\epsilon,z}$.

PROPOSITION 6.2. *There exists a positive constant ϵ_2 such that, for all $\epsilon \in (0, \epsilon_2)$ and $|z| \leq C\epsilon$, the map*

$$L_{\epsilon,z} = \pi_{\epsilon,z} \circ H'_\epsilon(w_{\epsilon,z}) : K^\perp_{\epsilon,z} \longrightarrow C^\perp_{\epsilon,z}$$

is surjective.

Combining Propositions 6.1 and 6.2 gives us the invertibility of $L_{\epsilon,z}$.

PROPOSITION 6.3.

$$L_{\epsilon,z} : K^{\perp}_{\epsilon,z} \longrightarrow C^{\perp}_{\epsilon,z}$$

is invertible, namely,

$$L^{-1}_{\epsilon,z} : C^{\perp}_{\epsilon,z} \longrightarrow K^{\perp}_{\epsilon,z}$$

exists. Furthermore, $L^{-1}_{\epsilon,z}$ *is bounded in the operator norm by* $C\epsilon^{-2}$.

We now begin to prove Proposition 6.1.

Proof of Proposition 6.1. We use a different strategy than in [32].

Suppose (6.60) is false. Then there exist sequences $\{\epsilon_k\}, \{z_k\}$, and $\{\Phi_k\}$, with $|z_k| \leqslant C\epsilon_k$ and $\epsilon_k \to 0$ as $k \to \infty$ such that

$$\Phi_k \in K^{\perp}_{\epsilon_k, z_k}$$

and

$$\|L_{\epsilon_k,z_k}(\Phi_k)\|_{L^2(\Omega_{\epsilon_k})} = o(1)\epsilon_k^2, \qquad \|\Phi_k\|_{H^2(\Omega_{\epsilon_k})} = 1. \tag{6.62}$$

We denote, for $i = 1, \ldots, N$,

$$e_{k,i} = \frac{(\partial w_{\epsilon_k,z_k}/\partial z_i)}{\|(\partial w_{\epsilon_k,\epsilon_k}/\partial z_i)\|_{L^2(\Omega_{\epsilon_k})}}, \quad e^*_{k,i} = \frac{(\partial v_{\sigma_k}/\partial y_i)}{\|(\partial v_{\sigma_k}/\partial y_i)\|_{L^2(\Omega_{\epsilon_k})}}. \tag{6.63}$$

Note that the difference between $e_{k,i}$ and $e^*_{k,i}$ is exponentially small. Hence, after applying the Gram–Schmidt process to $\{e_{k,i} | i = 1, \ldots, N\}$, we obtain a family of orthonormal functions $\{\tilde{e}_{k,i} | i = 1, \ldots, N\}$ with

$$\tilde{e}_{k,i} = e_{k,i} + \delta_{k,i}, \quad i = 1, \ldots, N,$$

where $\delta_{k,i} = O(e^{-\delta/\epsilon})$ in $L^2(\Omega_{\epsilon_k})$ as $k \to \infty$ for each $i = 1, \ldots, N$.

Hence,

$$L_{\epsilon_k,z_k}\Phi_k = H'_{\epsilon_k}(w_{\epsilon_k,z_k})\Phi_k - \sum_{i=1}^{N-1}\left(\int_{\Omega_{\epsilon_k}} [H'_{\epsilon_k}(w_{\epsilon_k,z_k})\Phi_k]e_{k,i}\,dy\right)e_{k,i} + E_k, \tag{6.64}$$

where E_k is defined by (6.64), and it is easy to see that $\|E_k\|_{L^2(\Omega_{\epsilon_k})} = O(e^{-\delta/\epsilon_k})$ as $k \to \infty$.

Note that

$$\begin{aligned} \|L_{\epsilon_k,z_k}\Phi_k\|^2_{L^2(\Omega_{\epsilon_k})} = \; & \|H'_{\epsilon_k}(w_{\epsilon_k,z_k})\Phi_k\|^2_{L^2(\Omega_{\epsilon_k})} \\ & - \sum_{i=1}^{n}\left(\int_{\Omega_{\epsilon_k}} [H'_{\epsilon_k}(w_{\epsilon_k,z_k})\Phi_k]e_{k,i}\,dy\right)^2 + O(e^{-\delta/\epsilon_k}) \end{aligned} \tag{6.65}$$

as $k \to \infty$.

Let us denote

$$\Delta\Phi_k + (1 - 3w^2_{\epsilon_k,z_k})\Phi_k + 3\frac{1}{|\Omega_{\epsilon_k}|}\int_{\Omega_{\epsilon_k}} w^2_{\epsilon_k,z_k}\Phi_k\,dy = \sigma_k^2 f_k.$$

By Corollary 2.1, we have

$$\left\| f_k - 3\frac{1}{|\Omega_{\epsilon_k}|\sigma_k^2}\int_{\Omega_{\epsilon_k}} w_{\epsilon_k,z_k}^2\Phi_k\,dy\right\|_{L^2(\Omega_{\epsilon_k})} \geq C\|\Phi_k\|_{H^2(\Omega_{\epsilon_k})}. \tag{6.66}$$

Note that since Φ_k satisfies the Neumann boundary condition, we have

$$\left|\int_{\Omega_{\epsilon_k}}\Phi_k\right| = \left|\sigma_k^2\int_{\Omega_{\epsilon_k}} f_k\,dy\right| \leq C\epsilon_k^{2-\frac{N}{2}}\|f_k\|_{L^2(\Omega_{\epsilon_k})}.$$

Hence,

$$3\frac{1}{|\Omega_{\epsilon_k}|\sigma_k^2}\int_{\Omega_{\epsilon_k}} w_{\epsilon_k,z_k}^2\Phi_k\,dy \leq C\epsilon_k^{\frac{N}{2}}\|f_k\|_{L^2(\Omega_{\epsilon_k})}.$$

Thus,

$$\left\|3\frac{1}{|\Omega_{\epsilon_k}|\sigma_k^2}\int_{\Omega_{\epsilon_k}} w_{\epsilon_k,z_k}^2\Phi_k\,dy\right\|_{L^2(\Omega_{\epsilon_k})} \leq C\|f_k\|_{L^2(\Omega_{\epsilon_k})}.$$

The last inequality and (6.66) imply that

$$\|f_k\|_{L^2(\Omega_{\epsilon_k})} \geq C\|\Phi_k\|_{H^2(\Omega_{\epsilon_k})} \geq C.$$

Therefore,

$$\|H'_{\epsilon_k}(w_{\epsilon_k,z_k})\Phi_k\|^2_{L^2(\Omega_{\epsilon_k})} \geq C\sigma_k^2. \tag{6.67}$$

Now we estimate

$$\begin{aligned}
&\int_{\Omega_{\epsilon_k}}[H'_{\epsilon_k}(w_{\epsilon_k,z_k})\Phi_k]e_{k,i}\,dy\\
&\quad= \int_{\Omega_{\epsilon_k}}[H'_{\epsilon_k}(w_{\epsilon_k,z_k})\Phi_k]\frac{\partial w_{\epsilon_k,z_k}}{\partial z_i}\,dy + O(e^{-\delta/\epsilon_k})\\
&\quad= \int_{\Omega_{\epsilon_k}}\left[\Delta\Phi_k + (1-3v_{\sigma_k}^2)\Phi_k + 3\frac{1}{|\Omega_{\epsilon_k}|}\int_{\Omega_{\epsilon_k}} v_{\sigma_k}^2\Phi_k\,dy\right]\frac{\partial v_{\sigma_k}}{\partial y_i}\,dy + O(e^{-\delta/\epsilon_k})\\
&\quad= \int_{\partial\Omega_{\epsilon_k}}\left[\frac{\partial v_{\sigma_k}}{\partial y_i}\frac{\partial\Phi_k}{\partial\nu} - \Phi_k\frac{\partial}{\partial\nu}\left(\frac{\partial v_{\sigma_k}}{\partial y_i}\right)\right]do\\
&\qquad+ 3\frac{1}{|\Omega_{\epsilon_k}|}\int_{\Omega_{\epsilon_k}} v_{\sigma_k}^2\Phi_k\,dy\int_{\Omega_{\epsilon_k}}\frac{\partial v_{\sigma_k}}{\partial y_i}\,dy\\
&\qquad+ O(e^{-\delta/\epsilon_k}) = O(e^{-\delta/\epsilon_k}).
\end{aligned}$$

Therefore, (6.65) implies that

$$o(1)\epsilon_k^2 \geq C\sigma_k^2 - o(e^{-\delta/\epsilon_k}). \tag{6.68}$$

This is a contradiction! Proposition 6.1 is thus proved. □

The following lemma, which can be found in [15], will be needed in the proof of Proposition 6.2.

LEMMA 6.1 (see [15, Lemma 1.3]). *If* $\vec{d}\ (E,F) := \sup\{d(x,F)|x \in E, \|x\|_H = 1\} < 1$, *then* $\pi_{F|E} : E \to F$ *is injective and* $\pi_{E|F} : F \to E$ *has a bounded right inverse, where* π_E (π_F, *respectively*) *is the orthogonal projection from* H *to* E (F, *respectively*). *In particular,* $\pi_{E|F} : F \to E$ *is surjective.*

We are now ready to prove Proposition 6.2.

Proof of Proposition 6.2. Let $CK_{\epsilon,z}$ = cokernel of $H'_\epsilon(w_{\epsilon,z})$. We first claim that

$$\vec{d}(CK_{\epsilon,z}, C_{\epsilon,z}) < 1 \tag{6.69}$$

for all $\epsilon > 0$ sufficiently small.

In fact, suppose (6.69) is not true. Then there exist $\epsilon_k \to 0$ and $\Phi_k \in CK_{\epsilon_k,z_k}$ such that

$$\Delta\Phi_k + (1 - 3w^2_{\epsilon_k,z_k})\Phi_k + 3w^2_{\epsilon_k,z_k}\frac{1}{|\Omega_{\epsilon_k}|}\int_{\Omega_{\epsilon_k}} \Phi_k\, dy = 0 \quad \text{in } \Omega_{\epsilon_k}, \tag{6.70}$$

$$\frac{\partial\Phi_k}{\partial\nu} = 0 \quad \text{on } \partial\Omega_{\epsilon_k}, \tag{6.71}$$

$$\|\Phi_k\|_{L^2(\Omega_{\epsilon_k})} = 1, \tag{6.72}$$

$$\int_{\Omega_{\epsilon_k}} \Phi_k \frac{\partial(w_{\epsilon_k,z_k})}{\partial z_i}\, dy = 0, \qquad i = 1,\dots,N. \tag{6.73}$$

By (6.70) and (6.71), we have

$$\int_{\Omega_{\epsilon_k}} (1 - 3w^2_{\epsilon_k,z_k})\Phi_k\, dy + 3\int_{\Omega_{\epsilon_k}} w^2_{\epsilon_k,z_k}\, dy \frac{1}{|\Omega_{\epsilon_k}|}\int_{\Omega_{\epsilon_k}} \Phi_k\, dy = 0.$$

Note that

$$\int_{\Omega_{\epsilon_k}} w^2_{\epsilon_k,z_k}\, dy = |\Omega_{\epsilon_k}|(1 + O(\epsilon_k)).$$

Hence, we have

$$\int_{\Omega_{\epsilon_k}} \Phi_k\, dy = \int_{\Omega_{\epsilon_k}} (1/3 - w^2_{\epsilon_k,z_k})\Phi_k\, dy(1 + O(\epsilon_k)) \le O(\epsilon_k^{\frac{N+1}{2}})\|\Phi_k\|_{L^2(\Omega_{\epsilon_k})}.$$

Similar to the proof of Proposition 6.1, we conclude that

$$\|\Phi_k\|_{H^2(\Omega_{\epsilon_k})} = o(1). \tag{6.74}$$

This is a contradiction! Hence, (6.69) is true.

Now by the fact that $\vec{d}(E,F) = \vec{d}(F^\perp, E^\perp)$, we have

$$\vec{d}(\overline{C}^\perp_{\epsilon,z}, \overline{CK}^\perp_{\epsilon,z}) < 1,$$

where $\overline{C}^\perp_{\epsilon,z}$ ($\overline{CK}^\perp_{\epsilon,z}$, respectively) is the orthogonal complement of $C_{\epsilon,z}$ ($CK_{\epsilon,z}$, respectively) in $L^2(\Omega_\epsilon)$. Thus, the map

$$\pi_{\overline{C}^\perp_{\epsilon,z}}\big|_{\overline{CK}^\perp_{\epsilon,z}} : \overline{CK}^\perp_{\epsilon,z} \to \overline{C}^\perp_{\epsilon,z} \tag{6.75}$$

is surjective by Lemma 6.1.

Since $\overline{CK}^{\perp}_{\epsilon,z}$ is the range of L_ϵ, it suffices to show that the map in (6.75), when restricted to $CK^{\perp}_{\epsilon,z}$, which is just $\pi_{\epsilon,z}$, is onto $C^{\perp}_{\epsilon,z}$. However, this follows easily from the expression

$$\pi_{\overline{C}^{\perp}_{\epsilon,z}}(\Phi) = \Phi - \pi_{C_{\epsilon,z}}\Phi. \qquad \square$$

Finally, in this section we solve the following equation for $\Phi_\epsilon \in K^{\perp}_{\epsilon,z}$:

$$\pi_{\epsilon,z} \circ H_\epsilon(w_{\epsilon,z})(w_{\epsilon,z} + \Phi_\epsilon) = 0. \tag{6.76}$$

Since $L_{\epsilon,z}\big|_{K^{\perp}_{\epsilon,z}}$ is invertible (and we shall denote its inverse just by $L^{-1}_{\epsilon,z}$) by Proposition 6.3, this is equivalent to solving

$$\Phi_\epsilon = L^{-1}_{\epsilon,z} \circ \pi_{\epsilon,z}(L_\epsilon(\Phi_\epsilon)) = -L^{-1}_{\epsilon,z} \circ \pi_{\epsilon,z}(3(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z} + N_{\epsilon,z}(\Phi_\epsilon) + M_{\epsilon,z}(\Psi_{\epsilon,z}))$$

$$:\equiv Q_{\epsilon,z}(\Phi_\epsilon),$$

where $Q_{\epsilon,z}$ is defined in the last equality for every $\Phi_\epsilon \in H^2_N(\Omega_\epsilon)$.

By Proposition 6.3, we have

$$\|L^{-1}_{\epsilon,z}\| \le C\epsilon^{-2}.$$

Hence,

$$\begin{aligned}\|Q_{\epsilon,z}(\Phi_\epsilon)\|_{H^2(\Omega_\epsilon)} &\leqslant C\epsilon^{-2}(\|(v_\sigma^2 - \tau_\sigma^2)\Psi_{\epsilon,z}\|_{L^2(\Omega_\epsilon)} + \|N_{z,\epsilon}(\Phi_\epsilon)\|_{L^2(\Omega_\epsilon)} \\ &\quad + \|M_{z,\epsilon}(\Psi_{\epsilon,z})\|_{L^2(\Omega_\epsilon)}) \\ &\leqslant c\epsilon^{-2}(\tilde{\varphi}_{\epsilon,z}^{\frac{1}{2}+\tilde{\eta}} + \ \delta\|\Phi_\epsilon\|_{H^2(\Omega_\epsilon)})\end{aligned}$$

for some $\tilde{\eta} > 0$ (in fact, we can take $\tilde{\eta} = \frac{1}{4}$ by Lemma 5.1).

Take $\delta = |\tilde{\varphi}_{\epsilon,z}(z)|^{\frac{1+\eta}{2}}$ for $0 < \eta < 2\tilde{\eta}$. Then we have (since $\delta\epsilon^{-2} = o(1)$)

$$\|Q_{\epsilon,z}(\Phi_\epsilon)\|_{H^2(\Omega_\epsilon)} \leqslant C(\tilde{\varphi}_{\epsilon,z}^{\frac{1+\eta}{2}}(z)). \tag{6.77}$$

Equation (6.77) says that $Q_{\epsilon,z}(\Phi)$ is a continuous map:

$$B_\delta(0) \cap H^2_N(\Omega_\epsilon) \longrightarrow B_\delta(0) \cap H^2_N(\Omega_\epsilon).$$

Furthermore, $Q_{\epsilon,z}(\Phi)$ is a contracting map if ϵ is small by Lemma 5.1. Hence, by the contraction mapping principle we have the following proposition.

PROPOSITION 6.4. *There exists $\epsilon_0 > 0$ such that, for $\epsilon < \epsilon_0, |z| \le C\epsilon$ there is a unique $\Phi_{\epsilon,z} \in K^{\perp}_{\epsilon,z}$ such that*

$$H_\epsilon(w_{\epsilon,z} + \Phi_{\epsilon,z}) \in C_{\epsilon,z}. \tag{6.78}$$

Furthermore,

$$\|\Phi_{\epsilon,z}\|_{H^2(\Omega_\epsilon)} \le C\tilde{\varphi}_{\epsilon,z}^{\frac{1+\mu}{2}}(z). \tag{6.79}$$

7. The reduced problem. In this section, we shall prove our main result, Theorem 1.1.

By Proposition 6.4, for $\epsilon \leqslant \epsilon_0$ and $|z| \leq C\epsilon$, there exists a unique $\Phi_{\epsilon,z}$ such that

$$H_\epsilon(w_{\epsilon,z} + \Phi_{\epsilon,z}) \in C_{\epsilon,z}. \tag{7.80}$$

Therefore, it is enough to show that for some $|z| \leq C\epsilon$, we have

$$H_\epsilon(w_{\epsilon,z} + \Phi_{\epsilon,z}) \perp C_{\epsilon,z}.$$

To this end, we now define a vector field

$$V_{\epsilon,j}(\tilde{z}) := \frac{1}{\epsilon^{\alpha-1}\tilde{\varphi}_{\epsilon,z}(z)}\left[\int_{\Omega_\epsilon} H_\epsilon(w_{\epsilon,z} + \Phi_{\epsilon,z})\frac{\partial w_{\epsilon,z}}{\partial z_j}\,dy\right], \tag{7.81}$$

where $z = \epsilon\frac{a}{2\sqrt{2}}\,d(0,\partial\Omega) + \epsilon^{\alpha+1}\tilde{z}, |\tilde{z}| \leqslant 1$, and $\vec{a}$ is given by conditions (2) and (3) in section 1.

The main estimate of this section is the following.

LEMMA 7.1. *For every $0 < \alpha < \alpha_0$, the vector field V_ϵ converges uniformly to V_0 in $B_1(0)$ as $\epsilon \to 0$, where*

$$V_0 = (V_{0,1}, \ldots, V_{0,N}),$$

$$V_{0,j} = \frac{\gamma}{\int_{\partial\Omega} e^{\langle t - P_0, a\rangle}\,d\mu_{P_0}(t)}\sum_{i=1}^{N}\left(\int_{\partial\Omega} e^{\langle x-P_0,a\rangle}x_i x_j\,d\mu_{P_0}(x)\tilde{z}_i\right), \qquad j = 1,\ldots,N,$$

and γ is given by Lemma 3.2.

Once Lemma 7.1 is proved, then Theorem 1.1 follows easily. In fact, since 0 is a nondegenerate peak point, V_0 has a nondegenerate zero at 0 (with degree different from 0). Then Lemma 7.1 and a simple degree theoretic argument imply that V_ϵ has a zero $\tilde{z}(\epsilon) \in B_{\frac{1}{2}}(0)$ for every ϵ sufficiently small. This solves the equation $H_\epsilon(w_{\epsilon,z}+\Phi_{\epsilon,z}) = 0$ for every ϵ sufficiently small. Setting $z(\epsilon) = \epsilon\frac{a}{2\sqrt{2}}d(0,\partial\Omega) + \epsilon^{\alpha+1}\tilde{z}(\epsilon)$ and

$$v_\epsilon = w_{\epsilon,z(\epsilon)} + \Phi_{\epsilon,z(\epsilon)}$$

for $x \in \Omega$ and ϵ sufficiently small, it then follows that

$$v_\epsilon \not\equiv 0 \quad \text{since } \Phi_{\epsilon,z(\epsilon)} \to 0 \quad \text{in } H^2(\Omega_\epsilon) \quad \text{as } \epsilon \to 0,$$

while $w_{\epsilon,z(\epsilon)}$ remains bounded away from 0 in $H^2(\Omega_\epsilon)$ as $\epsilon \to 0$.

In other words, v_ϵ is a nontrivial solution of (1.2). By the structure of v_ϵ, v_ϵ has all the properties of Theorem 1.1.

It remains to prove Lemma 7.1. To this end, we have

$$\begin{aligned}
&\int_{\Omega_{\epsilon,z}} H_\epsilon(w_{\epsilon,z} + \Phi_{\epsilon,z})\frac{\partial w_{\epsilon,z}}{\partial z_j}\\
&\quad= \int_{\Omega_{\epsilon,z}} [H_\epsilon'(w_{\epsilon,z})\Phi_{\epsilon,z}]\frac{\partial w_{\epsilon,z}}{\partial z_j}\\
&\qquad+ \int_{\Omega_{\epsilon,z}} [N_{\epsilon,z}(\Phi_{\epsilon,z})]\frac{\partial w_{\epsilon,z}}{\partial z_j}\\
&\qquad+ \int_{\Omega_{\epsilon,z}} M_{\epsilon,z}(\Psi_{\epsilon,z})\frac{\partial w_{z,\epsilon}}{\partial z_j}\\
&\qquad+ \int_{\Omega_{\epsilon,z}} 3[v_\sigma^2 - \tau_\sigma^2]\Psi_{\epsilon,z}\frac{\partial w_{\epsilon,z}}{\partial z_j}\\
&\quad= I_1 + I_2 + I_3 + I_4,
\end{aligned}$$

where $I_i, i = 1,2,3,4$ are defined by the last equality.

Note that

$$\begin{aligned}
I_1 &= 3\int_{\Omega_{\epsilon,z}} \big[(P_{\Omega_{\epsilon,z}} v_\sigma)^2 - v_\sigma^2\big]\Phi_{\epsilon,z}\frac{\partial w_{\epsilon,z}}{\partial z_j}\,dy \\
&\quad + 3\int_{\Omega_{\epsilon,z}} \frac{\partial w_{\epsilon,z}}{\partial z_j}\,dy \int_{\Omega_{\epsilon,z}} (P_{\Omega_{\epsilon,z}} v_\sigma)^2\Phi_{\epsilon,z}\,dy \\
&\le C\left\|(P_{\Omega_{\epsilon,z}} v_\sigma - v_\sigma)\frac{\partial w_{\epsilon,z}}{\partial z_j}\right\|_{L^2(\Omega_{\epsilon,z})} \|\Phi_{\epsilon,z}\|_{L^2(\Omega_{\epsilon,z})} \\
&\quad + 3\int_{\Omega_{\epsilon,z}} \frac{\partial w_{\epsilon,z}}{\partial z_j}\,dy \epsilon^{-N/2}\|\Phi_{\epsilon,z}\|_{L^2(\Omega_{\epsilon,z})} \\
&\le C\tilde{\varphi}_{\epsilon,z}(z)^{\frac{1+\mu}{2}}\tilde{\varphi}_{\epsilon,z}(z)^{\frac{1+\mu}{2}} \\
&= O(\tilde{\varphi}_{\epsilon,z}^{1+\mu}(z)),
\end{aligned}$$

where $\mu > 0$ is some small number. By Lemma 5.1 and Proposition 6.4, we have

$$|I_2| \le C|\tilde{\varphi}_{\epsilon,z}(z)|^{1+\mu}$$

and

$$|I_3| \le C|\tilde{\varphi}_{\epsilon,z}(z)|^{1+\mu},$$

since $N_{\epsilon,z}(\cdot)$ and $M_{\epsilon,z}(\cdot)$ depend on their arguments only in the second or higher powers. So we just need to compute I_4. In fact,

$$\begin{aligned}
I_4 &= -\int_{\Omega_{\epsilon,z}} 3\big[\tau_\sigma^2 - v_\sigma^2\big]\Psi_{\epsilon,z}\frac{\partial P_{\Omega_{\epsilon,z}} v_\sigma}{\partial z_j} \\
&= -\epsilon\int_{\Omega_{\epsilon,z}} 3[\tau_\sigma^2 - v_\sigma^2]\Psi_{\epsilon,z}\frac{\partial v_\sigma}{\partial y_j} \\
&\quad + O\big(e^{-\sqrt{\nu_\epsilon}\frac{(2+\mu)d(z,\partial\Omega)}{\epsilon}}\big).
\end{aligned}$$

By Lemma 3.2, we conclude the proof of Lemma 7.1.

Acknowledgment. The first author wishes to thank Professor Wei-Ming Ni for his constant encouragement.

Note Added in Proof. After the paper was accepted for publication we were informed by Professor Alikakos that in [35] an implicit proof of the existence of stationary bubbles in a general domain is given. However, the exact location of stationary bubbles is left open.

REFERENCES

[1] S. Agmon, *Lectures on Elliptic Boundary Value Problems,* Van Nostrand, Princeton, NJ, 1965.
[2] N. Alikakos, P. W. Bates, and X. Chen, *Convergence of the Cahn-Hilliard equation to the Hele-Shaw model*, Arch. Rational Mech. Anal., 128 (1994), pp. 165–205.
[3] N. Alikakos, P. W. Bates, and G. Fusco, *Slow motion for Cahn-Hilliard equation in one space dimension*, J. Differential Equations, 90 (1991), pp. 81–134.
[4] N. Alikakos, G. Fusco, and M. Kowalczyk, *Finite dimensional dynamics and interfaces intersecting the boundary: Equilibria and quasi-invariant manifold*, Indiana Univ. Math. J., 45 (1996), pp. 1119–1155.

[5] N. Alikakos and G. Fusco, *Slow dynamics for the Cahn-Hilliard equation in higher space dimension, Part* I: *Spectral estimates*, Comm. Partial Differential Equations, 19 (1994), pp. 1397–1447.

[6] P. W. Bates and P. C. Fife, *The dynamics of nucleation for the Cahn–Hilliard equation*, SIAM J. Appl. Math., 53 (1993), pp. 990–1008.

[7] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system*, I: *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[8] X. Chen, *Generation and propagation of interfaces in reaction diffusion systems*, J. Differential Equations, 96 (1992), pp. 116–141.

[9] X. Chen and M. Kowalczyk, *Existence of equilibria for the Cahn-Hilliard equation via local minimizers of the perimeter*, Comm. Partial Differential Equations, 21 (1996), pp. 1207–1233.

[10] E. N. Dancer, *A note on asymptotic uniqueness for some nonlinearities which change sign*, Rocky Mountain J. Math., to appear.

[11] A. Floer and A. Weinstein, *Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential*, J. Funct. Anal., 69 (1986), pp. 397–408.

[12] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in* R^n, in Mathematical Analysis and Applications, Part A, Adv. Math. Suppl. Studies 7A, Academic Press, New York, 1981, pp. 369–402.

[13] M. Grinfeld and A. Novick-Cohen, *Counting stationary solutions of the Cahn-Hilliard equation by transversality arguments*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 351–370.

[14] M. Grinfeld and A. Novick-Cohen, *The viscous Cahn-Hilliard equation*: *Morse decomposition and structure of the global attractor*, submitted.

[15] B. Helffer and J. Sjöstrand, *Multiple wells in the semi-classical limit* I, Comm. Partial Differential Equations, 9 (1984), pp. 337–408.

[16] R. V. Kohn and P. Sternberg, *Local minimizers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.

[17] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.

[18] P. L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Res. Notes Math. 69, Pitman, London, 1982.

[19] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 107 (1989), pp. 71–83.

[20] W.-M. Ni and I. Takagi, *On the shape of least energy solution to a semilinear Neumann problem*, Comm. Pure Appl. Math., 41 (1991), pp. 819–851.

[21] W.-M. Ni and I. Takagi, *Locating the peaks of least energy solutions to a semilinear Neumann problem*, Duke Math. J., 70 (1993), pp. 247–281.

[22] W.-M. Ni and J. Wei, *On the location and profile of spike-layer solutions to singularly perturbed semilinear Dirichlet problems*, Comm. Pure Appl. Math., 48 (1995), pp. 731–768.

[23] Y. G. Oh, *Existence of semi-classical bound states of nonlinear Schrödinger equations with potentials of the class* $(V)_a$, Comm. Partial Differential Equations, 13 (1988), pp. 1499–1519.

[24] Y. G. Oh, *On positive multi-bump bound states of nonlinear Schrödinger equations under multiple-well potentials*, Comm. Math. Phys., 131, (1990), pp. 223–253.

[25] R. L. Pego, *Front migration in the nonlinear Cahn-Hilliard equation*, Proc. Roy. Soc. London Ser. A, 422 (1989), pp. 261–278.

[26] L. A. Peletier and J. Serrin, *Uniqueness of positive solutions of semilinear equations in* R^n, Arch. Rational Mech. Anal., 81, (1983), pp. 181–197.

[27] M. J. Ward, *Metastable bubble solutions for the Allen–Cahn equation with mass conservation*, SIAM J. Appl. Math., 56 (1996), pp. 1247–1279.

[28] M. J. Ward, *An asymptotic analysis of localized solutions for some reaction-diffusion models in multi-dimensional domains*, Stud. Appl. Math., 97 (1996), pp. 103–126.

[29] J. Wei, *On the interior spike layer solutions to a singularly perturbed semilinear Neumann problem*, Tohoku Math. J., 50 (1998), in press.

[30] J. Wei, *On the effect of domain geometry in some singular perturbation problems*, Differential Integral Equations, to appear.

[31] J. Wei, *On the interior spike solutions for some singular perturbation problems*, Proc. Roy. Soc. Edinburgh Sect. A., 128A (1998), in press.

[32] J. Wei and M. Winter, *Stationary solutions of the Cahn-Hilliard equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 459–492.

[33] J. Wei and M. Winter, *Multiple spike layer solutions for a wide class of singular perturbation problems*, J. London Math. Soc., to appear.

[34] J. Wei and M. Winter, *On the stationary Cahn-Hilliard equation: Interior spike solutions*, J. Differential Equations, to appear.
[35] N. Alikakos and G. Fusco, *Slow dynamics for the Cahn-Hilliard equation in higher space dimension: The motion of bubbles*, Arch. Rational Mech. Anal., 141 (1998), pp. 1–61.

SIAM J. MATH. ANAL.
Vol. 29, No. 6, pp. 1519–1536, November 1998

013

SINGULAR PERTURBATION APPROACH TO A 3-COMPONENT REACTION-DIFFUSION SYSTEM ARISING IN POPULATION DYNAMICS*

YUKIO KAN-ON† AND MASAYASU MIMURA‡

Abstract. In order to understand theoretically predation-mediated coexistence of competing species which is often observed in ecological systems, we consider a 3-component reaction-diffusion system describing the interaction of one predator and two competing prey species which move by diffusion. It is shown that there exist stable spatially inhomogeneous positive equilibrium solutions of the one-dimensional system under the Neumann boundary condition. This implies ecologically that in the presence of the predator, two competing species coexist with spatially segregating structures. The main tools we use are the singular perturbation technique and the associated singular limit spectral analysis.

Key words. prey-predator model, singular perturbation, stability

AMS subject classification. 35B25

PII. S0036141097318328

1. Introduction.

It is often observed that predation may have a tendency to increase species diversity in competitive communities. This is called *predation-mediated coexistence*. There have been a number of theoretical studies on the possibility of temporal coexistence of competing species under predation pressure by using ODE models with Lotka–Volterra prey-predator interaction (for example, see Takeuchi and Adachi [18]). In this paper, we consider the situation where all species can move by diffusion and study the spatial structure of competing species that may coexist in the presence of predators.

In order to study this situation, we propose here the following 3-component reaction-diffusion system for two prey and one predator species:

$$\text{(1.1a)}\qquad \begin{cases} u_{1t} = d_1\,\Delta u_1 + u_1\,(a_1 - b_1\,u_1 - c_1\,u_2 - k_1\,v), \\ u_{2t} = d_2\,\Delta u_2 + u_2\,(a_2 - c_2\,u_1 - b_2\,u_2 - k_2\,v), \\ v_t = d_v\,\Delta v + v\,(-r + \alpha_1\,k_1\,u_1 + \alpha_2\,k_2\,u_2), \quad x \in \Omega, \quad t > 0, \end{cases}$$

where u_1 and u_2 are the population densities of two competing prey species, and v is that of their predator. d_1, d_2, and d_v are the diffusion rates, r is the death rate for the predator, a_i is the intrinsic growth rate, b_i and c_i are the intraspecific and interspecific competition rates, respectively, k_i is the predation rate, and α_i is the transformation rate of predation ($i = 1, 2$). All of the coefficients are positive constants. Ω is a bounded domain with smooth boundary $\partial\Omega$. We assume that u_1, u_2, and v satisfy the Neumann boundary conditions on the boundary $\partial\Omega$, given by

$$\text{(1.1b)}\qquad \frac{\partial}{\partial\nu}u_1 = 0, \quad \frac{\partial}{\partial\nu}u_2 = 0, \quad \frac{\partial}{\partial\nu}v = 0, \quad x \in \partial\Omega, \quad t > 0,$$

*Received by the editors March 14, 1997; accepted for publication (in revised form) December 9, 1997; published electronically September 3, 1998.
http://www.siam.org/journals/SIMA/29-6/31832.html

†Department of Mathematics, Faculty of Education, Ehime University, Matsuyama 790, Japan (kanon@ed.ehime-u.ac.jp).

‡Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153, Japan (mimura@ms.u-tokyo.ac.jp) and Department of Mathematics, Hiroshima University, 1-3-1 Higashi-Hiroshima 739, Japan (mimura@math.sci.hiroshima-u.ac.jp).

where ν is the outward normal unit vector on $\partial\Omega$.

Let us begin by reviewing the qualitative behaviors of solutions of two 2-species systems deriving from (1.1). The first system is a two-competing-species model for (u_1, u_2) in the absence of the predator ($v = 0$):

$$\begin{cases} u_{1t} = d_1\,\Delta u_1 + u_1\,(a_1 - b_1\,u_1 - c_1\,u_2), \\ u_{2t} = d_2\,\Delta u_2 + u_2\,(a_2 - c_2\,u_1 - b_2\,u_2), \quad x \in \Omega, \quad t > 0, \\ \frac{\partial}{\partial\nu} u_1 = 0, \quad \frac{\partial}{\partial\nu} u_2 = 0, \quad x \in \partial\Omega, \quad t > 0. \end{cases} \tag{1.2}$$

Suppose

$$\frac{b_1}{c_2} < \frac{a_1}{a_2} < \frac{c_1}{b_2}, \tag{1.3}$$

which indicates that the interspecific competition is stronger than the intraspecific one for the competing species. For an arbitrary convex domain Ω, Kishimoto and Weinberger [11] proved that any spatially inhomogeneous positive (SIP) equilibrium solution of (1.2) is unstable even if it exists, that is, any stable positive equilibrium solution of (1.2) has to be spatially homogeneous. The condition (1.3) means that any positive solution (u_1, u_2) of (1.2) generically tends to either $(0, a_2/b_2)$ or $(a_1/b_1, 0)$ as $t \to +\infty$, which ecologically implies that competitive exclusion occurs between two species. On the other hand, Matano and Mimura [13] showed that when the domain Ω is of suitable dumbbell shape, there are stable SIP equilibrium solutions which exhibit regionally segregating coexistence of two competing species. These results indicate that the stability as well as the existence of SIP equilibrium solutions of the competition-diffusion system (1.2) depends on the shapes of the domain Ω.

The second system is a one prey–one predator model in the absence of one of the competing species (either $u_1 = 0$ or $u_2 = 0$), under which (1.1) becomes

$$\begin{cases} u_t = d\,\Delta u + u\,(a - b\,u - k\,v), \\ v_t = d_v\,\Delta v + v\,(-r + \alpha\,k\,u), \quad x \in \Omega, \quad t > 0, \\ \frac{\partial}{\partial\nu} u = 0, \quad \frac{\partial}{\partial\nu} v = 0, \quad x \in \partial\Omega, \quad t > 0, \end{cases} \tag{1.4}$$

where d, a, b, k, and α are positive constants which have the same meanings as d_i, a_i, b_i, k_i, and α_i, respectively, in (1.1). For any domain Ω, it is shown in Rothe [17] that any positive solution of (1.4) tends to a homogeneous equilibrium solution

$$\left(\frac{r}{\alpha\,k},\ \max\left\{ 0, \frac{a\,\alpha\,k - b\,r}{\alpha\,k^2} \right\} \right)$$

as $t \to +\infty$.

Integrating the above results, we address the following questions: (i) Does the system (1.1) possess stable SIP equilibrium solutions when Ω is convex? (ii) If such solutions exist, what is the spatial profile of the two competing species (u_1, u_2)? In a previous paper [15], we numerically answered that solutions stably exist such that u_1 and u_2 exhibit spatially segregating structures for values of parameters in a suitable region. Figure 1 demonstrates that in the absence of the predator ($t < t_0$), the u_1-species occupies almost the whole interval and the u_2-species is extinct due to competition, but when the predator is present ($t \geq t_0$), the spatial segregating coexistence occurs for two competing species. This clearly shows the occurrence of predation-mediated coexistence of two competing species which move by diffusion.

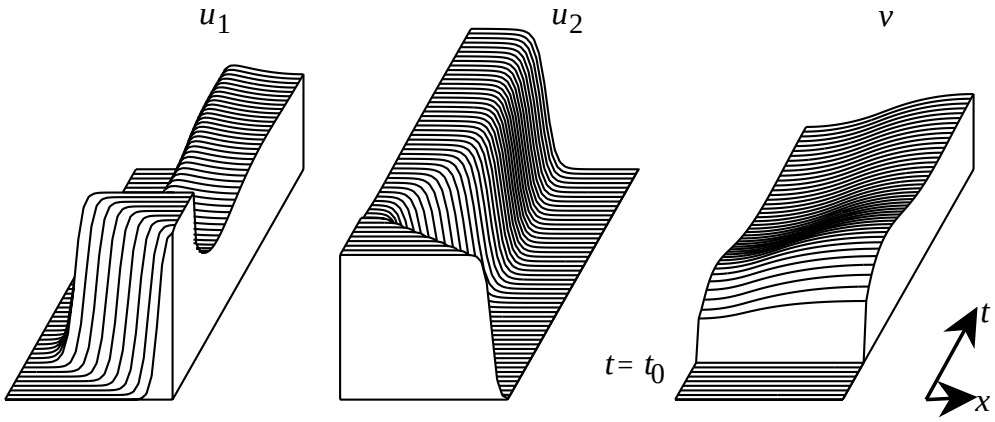

FIG. 1. *Spatially segregating coexistence of u_1 and u_2 in the presence of v. The predator is absent for $0 < t < t_0$ and is present for $t > t_0$.*

In this paper, we restrict our discussion to the one-dimensional problem of (1.1) in the interval $(0, 1)$ and show the existence of stable SIP equilibrium solutions when the diffusion rates of u_1 and u_2 are sufficiently smaller than that of v. To express (1.1) in nondimensional variables, we set

$$\bar{t} = a_1\, t, \quad \bar{u}_1 = b_1\, u_1/a_1, \quad \bar{u}_2 = c_2\, u_2/a_1, \quad \bar{v} = k_1\, v/a_1, \quad \bar{r} = r\, b_1/(a_1\, \alpha_1\, k_1),$$

$$a = a_2/a_1, \quad b = b_2/b_1, \quad c = c_1/c_2, \quad d = d_2/d_1, \quad k = k_2/k_1,$$

$$\varepsilon^2 = d_1/a_1, \quad \sigma = a_1/d_v, \quad \alpha = \alpha_1\, k_1/b_1, \quad \beta = \alpha_2\, b_1/(\alpha_1\, c_2).$$

Dropping the overbar of $\bar{t}$, $\bar{u}_1$, $\bar{u}_2$, $\bar{v}$, and $\bar{r}$, we can rewrite (1.1) as

$$\begin{cases} \boldsymbol{u}_t = \varepsilon^2\, D\, \boldsymbol{u}_{xx} + \boldsymbol{f}(\boldsymbol{u}, v), \\ \sigma\, v_t = v_{xx} + \sigma\, g(\boldsymbol{u}, v), \quad x \in (0,1), \quad t > 0, \\ \boldsymbol{u}_x = 0, \quad v_x = 0, \quad x = 0, 1, \quad t > 0, \end{cases} \tag{1.5}$$

where $\boldsymbol{u} = (u_1, u_2)$, $D = \mathrm{diag}(1, d)$, $\boldsymbol{f}(\boldsymbol{u}, v) = (f_1, f_2)(\boldsymbol{u}, v)$ with

$$\begin{aligned} f_1(\boldsymbol{u}, v) &= u_1\,(1 - u_1 - c\, u_2 - v), \\ f_2(\boldsymbol{u}, v) &= u_2\,(a - b\, u_1 - u_2 - k\, v), \\ g(\boldsymbol{u}, v) &= \alpha\, v\,(-r + u_1 + \beta\, k\, u_2), \end{aligned}$$

and all of the coefficients in the system are positive constants.

For the kinetics $(\boldsymbol{f}, g)$ in (1.5), we first assume

$$k < a < 1/c < b, \tag{H.1}$$

which is ecologically interpreted as follows: Suppose that the predator v is always constant. Then (1.5) is reduced to a two-competing-species model with a parameter v:

$$\begin{cases} \boldsymbol{u}_t = \varepsilon^2\, D\, \boldsymbol{u}_{xx} + \boldsymbol{f}(\boldsymbol{u}, v), \quad x \in (0,1), \quad t > 0, \\ \boldsymbol{u}_x = 0, \quad x = 0, 1, \quad t > 0. \end{cases} \tag{1.6}$$

Let $v_\pm$ and $\boldsymbol{h}_\pm(v)$ be

$$v_- = \frac{1-a\,c}{1-c\,k}, \quad v_+ = \frac{b-a}{b-k}, \quad \boldsymbol{h}_-(v) = (0, a-k\,v), \quad \boldsymbol{h}_+(v) = (1-v, 0)$$

($0 < v_- < v_+$ by (H.1)). If the initial condition $\boldsymbol{u}(x,0)$ is nonnegative but not identically zero on $[0,1]$, then the asymptotic behavior of the solution $\boldsymbol{u}(x,t)$ of (1.6) can be classified into the following three cases (see de Mottoni [3], for instance):

(a) If $v \le v_-$, then $\lim_{t\to+\infty} \boldsymbol{u}(x,t) = \boldsymbol{h}_+(v)$.

(b) If $v \in (v_-, v_+)$, then almost every solution converges to either $\boldsymbol{h}_-(v)$ or $\boldsymbol{h}_+(v)$ as $t \to +\infty$, depending on the initial condition.

(c) If $v \ge v_+$, then $\lim_{t\to+\infty} \boldsymbol{u}(x,t) = \boldsymbol{h}_-(v)$.

Inequalities $k < a < 1/c$ in (H.1) mean that the predator prefers to eat the u_1-species over the u_2-species. Therefore, cases (a), (b), and (c) can be easily interpreted as the following: Suppose that v is constant. Then only the u_1-species always survives and the u_2-species is extinct for smaller $v > 0$, while the situation is the reverse for larger $v > 0$.

First of all, we consider the situation where all of the diffusion rates are very large (that is, $\varepsilon > 0$ and $\sigma^{-1} > 0$ are both very large) in (1.5). Under this situation, it turns out that any solution of (1.5) becomes spatially homogeneous asymptotically so that the asymptotic state of solutions of (1.5) is described by the diffusionless system of (1.5):

$$\boldsymbol{u}_t = \boldsymbol{f}(\boldsymbol{u}, v), \quad v_t = g(\boldsymbol{u}, v), \quad t > 0 \tag{1.7}$$

(see Conway, Hoff, and Smoller [1]). The existence and stability of positive solutions of the ODEs (1.7) have been intensively studied, from viewpoints on the possibility of temporally segregating coexistence of two competing species (see Fujii [5], Hsu [6], and Takeuchi and Adachi [18], for instance). Let E_{+++} be an equilibrium point of (1.7) in the positive quadrant, which indicates the coexistence of two competing species. When r and k are taken as free parameters and the others appropriately fixed, the existence region of E_{+++} is shown in Figure 2. The region consists of two subregions A and B which are surrounded by the curves $r = r_c^-(k)$ and $r = r_c^+(k)$, where

$$r_c^-(k) = \beta\,k\,\frac{a-k}{1-c\,k}, \quad r_c^+(k) = \frac{a-k}{b-k}.$$

In region A, the stability of E_{+++} depends on the values of the parameters a, b, c, α, and β. If E_{+++} is not stable, there exist not only periodic solutions but also chaotic ones (for more information, we refer to [15]). On the other hand, in the region B, E_{+++} is always unstable. In the region not including A, numerical calculations suggest that neither stable positive equilibria nor stable positive periodic solutions exist. In view of these results, we may conclude that there occurs no predation-mediated coexistence for arbitrarily fixed (r,k) in the region not including A, if all of the diffusion rates are very large.

We now propose the following problem: Under the situation where (r,k) is arbitrarily fixed in the region not including A (that is, there occurs no predation-mediated coexistence under spatially homogeneous situations), is it possible for two competing species to coexist in the presence of a predator when one of the diffusion rates is not necessarily large?

In order to study this problem, we assume that

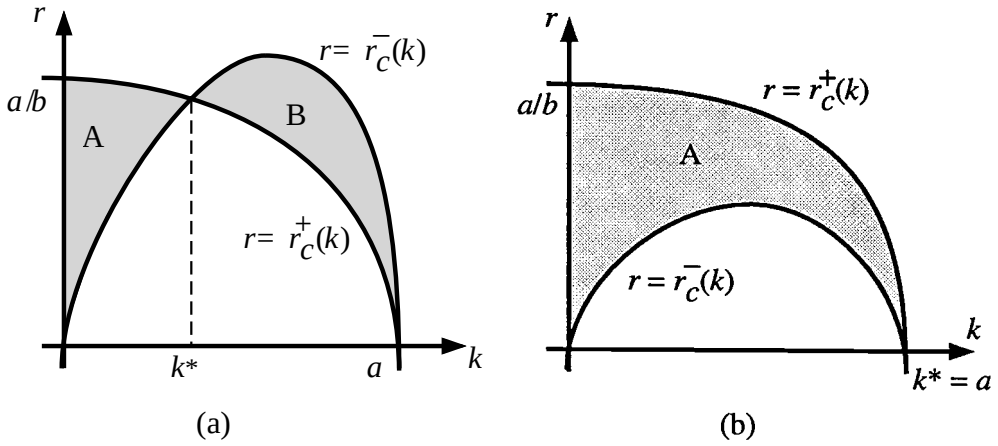

FIG. 2. *Existence region of* E_{+++}.

(H.2) $\varepsilon > 0$ is sufficiently small compared with other parameters,
which indicates that the prey species diffuse very slowly compared to the predator.

(H.3) r and k satisfy

$$(r_t^-(k) \equiv)\, a_0\, \beta\, k\, \frac{a-k}{a_0-k} < r < \frac{a-k}{a_0-k}\, (\equiv r_t^+(k)),$$

where a_0 is given by Lemma 2.1 in the next section. If $k > 0$ is sufficiently small, the inequalities $1/c < a_0 < b$ give the inequalities $r_t^-(k) < r_c^-(k) < r_c^+(k) < r_t^+(k)$. Hence it turns out that (r, k) does not lie in the region A shown in Figure 2 if (r, k) satisfies either

$$r_t^-(k) < r < r_c^-(k) \quad \text{or} \quad r_c^+(k) < r < r_t^+(k).$$

Our main theorem is stated as follows.

THEOREM 1.1. *Under the assumptions* (H.1), (H.2), *and* (H.3), *there exist* $\varepsilon_0 > 0$, $\sigma_0 > 0$, *and* $k_0 > 0$ *such that* (1.5) *has a stable SIP equilibrium solution* $(\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})(x)$ *for any* $(\varepsilon, \sigma, k) \in (0, \varepsilon_0] \times (0, \sigma_0] \times (0, k_0]$. *Furthermore* $(\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})(x)$ *satisfies the following properties:*

(i) $(\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})$ *is bounded in* $\mathcal{Z}_\varepsilon \times C^2([0,1])$,

(ii) $\lim_{(\varepsilon,\sigma)\to(0,0)} \boldsymbol{u}^{\varepsilon,\sigma}(x) = \boldsymbol{h}(x, v_0, x_0)$ *uniformly on* $[0, x_0 - \kappa) \cup (x_0 + \kappa, 1]$ *for any* $\kappa > 0$,

(iii) $\lim_{(\varepsilon,\sigma)\to(0,0)} v^{\varepsilon,\sigma}(x) = v_0$ *uniformly on* $[0,1]$
(*see Figure* 3), *where* v_0 *is given in Lemma* 2.1 *and satisfies*

$$g(\boldsymbol{h}_-(v_0), v_0) < 0 < g(\boldsymbol{h}_+(v_0), v_0),$$

and $\mathcal{Z}_\varepsilon$, x_0, *and* $\boldsymbol{h}(x, v, \tau)$ *are defined by*

$$\mathcal{Z}_\varepsilon = \{\, \boldsymbol{u} \in C^0([0,1], \mathbf{R}^2) \,|\, \|\, \boldsymbol{u} \,\|_{\mathcal{Z}_\varepsilon} < +\infty \,\},$$

$$\|\, \boldsymbol{u} \,\|_{\mathcal{Z}_\varepsilon} = \textstyle\sum_{j=0}^{2} \|\, (\varepsilon\, \frac{d}{dx})^j \boldsymbol{u} \,\|_{C^0([0,x_0)\cup(x_0,1],\mathbf{R}^2)},$$

$$x_0 = \frac{g(\boldsymbol{h}_+(v_0), v_0)}{g(\boldsymbol{h}_+(v_0), v_0) - g(\boldsymbol{h}_-(v_0), v_0)} \quad \textit{and} \quad \boldsymbol{h}(x, v, \tau) = \begin{cases} \boldsymbol{h}_-(v) & \textit{for } x < \tau, \\ \boldsymbol{h}_+(v) & \textit{for } x > \tau, \end{cases}$$

respectively.

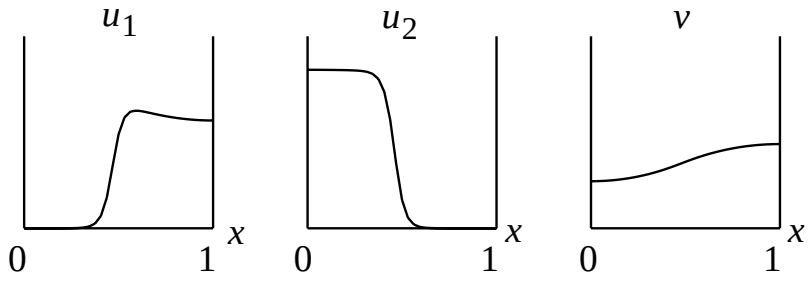

FIG. 3. *Spatial profiles of SIP equilibrium solution* $(u_1, u_2, v)(x)$ *of* (1.5).

The proof is achieved by the singular perturbation technique and the associated singular limit eigenvalue problem (SLEP) method. It is stated in sections 2 and 3.

Figure 3 demonstrates the spatial profiles of the SIP equilibrium solution $(\boldsymbol{u}, v)(x)$, where the solution $\boldsymbol{u}(x)$ clearly exhibits spatial segregation of two competing species, which are separated by an internal layer, while $v(x)$ is distributed smoothly in $(0, 1)$.

2. Construction of equilibrium solutions. In order to obtain SIP equilibrium solutions shown in Figure 3, we use singular perturbation methods for sufficiently small $\varepsilon > 0$. Since the case of multi-internal layers can be treated similarly, we restrict our discussion to the case of a single internal layer.

Let us consider the situation where such an internal layer exists in the vicinity of $x = \tau$, where τ is not known a priori. The resulting problem is the following stationary problem of (1.5):

$$\text{(2.1L)} \qquad \begin{cases} 0 = \varepsilon^2 D\, \boldsymbol{u}_{xx} + \boldsymbol{f}(\boldsymbol{u}, v), \\ 0 = v_{xx} + \sigma\, g(\boldsymbol{u}, v), \quad x \in (0, \tau), \\ \boldsymbol{u}_x(0) = 0, \quad v_x(0) = 0, \\ \boldsymbol{u}(\tau) = \hat{\boldsymbol{u}}, \quad v(\tau) = \hat{v}, \end{cases}$$

$$\text{(2.1R)} \qquad \begin{cases} 0 = \varepsilon^2 D\, \boldsymbol{u}_{xx} + \boldsymbol{f}(\boldsymbol{u}, v), \\ 0 = v_{xx} + \sigma\, g(\boldsymbol{u}, v), \quad x \in (\tau, 1), \\ \boldsymbol{u}(\tau) = \hat{\boldsymbol{u}}, \quad v(\tau) = \hat{v}, \\ \boldsymbol{u}_x(1) = 0, \quad v_x(1) = 0, \end{cases}$$

$$\text{(2.1B)} \qquad \boldsymbol{u}_x(\tau-) = \boldsymbol{u}_x(\tau+), \quad v_x(\tau-) = v_x(\tau+),$$

where $\hat{\boldsymbol{u}}$, $\hat{v}$, and τ are constants to be determined appropriately, and $u(x-)$ and $u(x+)$ are denoted by

$$u(x-) = \lim_{\delta \uparrow 0} u(x + \delta), \quad u(x+) = \lim_{\delta \downarrow 0} u(x + \delta),$$

respectively. Our strategy to construct SIP equilibrium solutions is stated as follows:

(i) Construct an inner (respectively, outer) approximated solution of (2.1L), (2.1R) in a neighborhood of (respectively, outside of) $x = \tau$ which is of C^1-class in $(\varepsilon, \hat{\boldsymbol{u}}, \hat{v})$ (respectively, $(\sigma, \hat{v}, \tau)$).

(ii) Assume $\varepsilon > 0$ and $\sigma > 0$ to be sufficiently small, and seek a solution of (2.1L), (2.1R) by using the inner and outer approximated solutions constructed in (i).

(iii) Determine $\hat{\boldsymbol{u}}$, $\hat{v}$, and τ as C^1-class functions of (ε, σ) such that the solution given in (ii) satisfies the boundary condition (2.1B).
This is a modified version of the usual singular perturbation method (for instance, see Fife [4]).

2.1. Inner approximation. In this subsection, we obtain an approximated solution of (2.1) in a neighborhood of $x = \tau$. To do this, we use the stretched variable $\xi = (x - \tau)/\varepsilon$ to rewrite (2.1) as

$$0 = D\,\boldsymbol{u}_{\xi\xi} + \boldsymbol{f}(\boldsymbol{u}, v), \quad 0 = v_{\xi\xi} + \varepsilon^2\,\sigma\,g(\boldsymbol{u}, v).$$

When $\varepsilon \to 0$, the second equation formally becomes $v_{\xi\xi} = 0$, so that v becomes constant because of the Neumann boundary condition, which implies $v(.) \to \hat{v}$ as $\varepsilon \to 0$. Hence as $\varepsilon \to 0$, the above equation can be approximated by

$$\begin{cases} 0 = D\,\boldsymbol{u}_{\xi\xi} + \boldsymbol{f}(\boldsymbol{u}, \hat{v}), \quad \xi \in R_- \cup R_+, \\ \boldsymbol{u}(0) = \hat{\boldsymbol{u}}, \end{cases} \tag{2.2a}$$

where $R_- = (-\infty, 0)$ and $R_+ = (0, +\infty)$. For the boundary conditions of $\boldsymbol{u}$ at $\xi = \pm\infty$, we may take

$$\boldsymbol{u}(-\infty) = \boldsymbol{h}_-(\hat{v}), \quad \boldsymbol{u}(+\infty) = \boldsymbol{h}_+(\hat{v}). \tag{2.2b}$$

Let us define the order relations $\preceq_s$ and $\preceq_o$ in the following manner:

$$(u_1, u_2) \preceq_s (\bar{u}_1, \bar{u}_2) \iff u_1 \le \bar{u}_1,\ u_2 \le \bar{u}_2,$$
$$(u_1, u_2) \preceq_o (\bar{u}_1, \bar{u}_2) \iff u_1 \le \bar{u}_1,\ u_2 \ge \bar{u}_2.$$

Similarly to the above definition, $\prec_s$ and $\prec_o$ are also defined by replacing $\le$ with $<$. We shall say that $\boldsymbol{u}(\xi)$ is *monotone* if $\boldsymbol{u}_\xi(\xi) \succ_o (0, 0)$ holds for any $\xi \in \mathbf{R}$.

LEMMA 2.1 (Theorem A2 in [14]). *Under the assumption* (H.1)*, there exists* $a_0 = a_0(b, c) \in (1/c, b)$ *such that* (2.2) *with*

$$\hat{v} = \frac{a_0 - a}{a_0 - k}\,(\equiv v_0 \in (v_-, v_+))$$

has a monotone solution $\boldsymbol{u}_0(\xi) = (u_{01}, u_{02})(\xi) \in C^2(\mathbf{R}, \mathbf{R}^2)$.

Let us define the linear operators $\mathcal{L}_0$ and $\mathcal{L}_0^*$ by $\mathcal{L}_0\,\boldsymbol{u} = D\,\boldsymbol{u}_{\xi\xi} + \boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{u}_0(\xi), v_0)\,\boldsymbol{u}$ and its formal adjoint operator, respectively. Setting

$$\gamma_1^\pm = \sqrt{-f_{1u_1}(\boldsymbol{h}_\pm(v_0), v_0)}, \quad \gamma_2^\pm = \sqrt{-f_{2u_2}(\boldsymbol{h}_\pm(v_0), v_0)/d},$$
$$\Gamma_1^\pm = \min\{\,\gamma_1^\pm, \gamma_2^\pm\,\}, \quad \Gamma_2^\pm = \max\{\,\gamma_1^\pm, \gamma_2^\pm\,\}, \quad m_\pm = 2 - {}^{\#}\{\,\gamma_1^\pm, \gamma_2^\pm\,\},$$

where ${}^{\#}A$ is the number of elements of the set A, we obtain the following lemmas.

LEMMA 2.2 (section 3.2 in [7]). *There exists a fundamental set* $\{\,\boldsymbol{U}_j(\xi)\,\}_{j=1}^4$ *of solutions of* $\mathcal{L}_0\,\boldsymbol{u} = 0$ *such that the limits as* $\xi \to \pm\infty$ *given by*

$$\lim_{\xi\to-\infty} \left| \boldsymbol{U}_1(\xi)\,\xi^{-m_-}\,e^{\Gamma_2^-\,\xi} \right|, \qquad \lim_{\xi\to+\infty} \left| \boldsymbol{U}_1(\xi)\,e^{\Gamma_2^+\,\xi} \right|,$$
$$\lim_{\xi\to-\infty} \left| \boldsymbol{U}_2(\xi)\,e^{\Gamma_1^-\,\xi} \right|, \qquad \lim_{\xi\to+\infty} \left| \boldsymbol{U}_2(\xi)\,e^{-\Gamma_1^+\,\xi} \right|,$$
$$\lim_{\xi\to-\infty} \left| \boldsymbol{U}_3(\xi)\,\xi^{-m_-}\,e^{-\Gamma_1^-\,\xi} \right|, \quad \lim_{\xi\to+\infty} \left| \boldsymbol{U}_3(\xi)\,\xi^{-m_+}\,e^{\Gamma_1^+\,\xi} \right|,$$
$$\lim_{\xi\to-\infty} \left| \boldsymbol{U}_4(\xi)\,e^{-\Gamma_2^-\,\xi} \right|, \qquad \lim_{\xi\to+\infty} \left| \boldsymbol{U}_4(\xi)\,\xi^{-m_+}\,e^{-\Gamma_2^+\,\xi} \right|$$

exist and all of the limits are positive. Furthermore, $\boldsymbol{U}_1(\xi)$*,* $\boldsymbol{U}_3(\xi)$*, and* $\boldsymbol{U}_4(\xi)$ *can be chosen to satisfy*

$$\boldsymbol{U}_1(\xi) \succ_s (0,0), \quad \boldsymbol{U}_3(\xi) = \boldsymbol{u}_{0\xi}(\xi), \quad \boldsymbol{U}_4(\xi) \succ_s (0,0)$$

for any $\xi \in \mathbf{R}$.

LEMMA 2.3 (Theorem A.2 in [10]). *Any nontrivial solution* $(u_1^*, u_2^*)(\xi)$ *of* $\mathcal{L}_0^* \boldsymbol{u} = 0$ *satisfies* $u_1^*(\xi)\, u_2^*(\xi) < 0$ *for any* $\xi \in \mathbf{R}$.

LEMMA 2.4 (Theorem 3.6 in [9]). *There exists* $\mu_0 > 0$ *such that*

$$\sigma(\mathcal{L}_0) \subset \{\, 0 \,\} \cup \{\, \lambda \in \mathbf{C} \,|\, \mathrm{Re}\ \lambda \le -\mu_0 \,\}$$

holds, where $\sigma(\mathcal{L}_0)$ *is the set of spectra of* $\mathcal{L}_0$ *relative to the space of bounded uniformly continuous functions from* $\mathbf{R}$ *to* $\mathbf{R}^2$ *with the supremum norm. Furthermore* $0 \in \sigma(\mathcal{L}_0)$ *is a simple eigenvalue.*

Putting $\hat{v} = v_0 + \omega$, we show that (2.2) has a positive solution for any ω in a neighborhood of $\omega = 0$. To do this, we use the new variable $\boldsymbol{y}$ with the form $\boldsymbol{y}\,(\equiv (y_1, y_2, y_3, y_4)) = (\boldsymbol{u}, \boldsymbol{u}_\xi)$ so that

$$\begin{cases} \frac{d}{d\xi}\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{y}, \omega), \quad \xi \in R_- \cup R_+, \\ \boldsymbol{y}(\pm\infty) = \boldsymbol{H}_\pm(\omega), \quad (y_1, y_2)(0) = \hat{\boldsymbol{u}}, \end{cases} \tag{2.3}$$

where $\boldsymbol{F}(\boldsymbol{y}, \omega) = (\boldsymbol{u}_\xi, -D^{-1}\, \boldsymbol{f}(\boldsymbol{u}, v_0 + \omega))$ and $\boldsymbol{H}_\pm(\omega) = (\boldsymbol{h}_\pm(v_0 + \omega), 0, 0)$. Without loss of generality, we may assume $\hat{\boldsymbol{u}} = \boldsymbol{u}_0(0) + (0, \eta)$, where η will be determined later. Setting

$$\boldsymbol{y} = \boldsymbol{y}_0(\xi) + \boldsymbol{H}_\pm(\omega) - \boldsymbol{H}_\pm(0) + \boldsymbol{z} \quad \text{ on } R_\pm,$$

we rewrite (2.3) as

$$\begin{cases} \frac{d}{d\xi}\boldsymbol{z} = A(\xi)\, \boldsymbol{z} + \boldsymbol{N}(\xi, \boldsymbol{z}, \omega), \quad \xi \in R_- \cup R_+, \\ \boldsymbol{z}(\pm\infty) = 0, \quad \boldsymbol{z}(0\pm) = \boldsymbol{H}_\pm(0) - \boldsymbol{H}_\pm(\omega) + (0, \eta, \boldsymbol{q}_\pm), \end{cases} \tag{2.4}$$

where

$$A(\xi) = \boldsymbol{F}_{\boldsymbol{y}}(\boldsymbol{y}_0(\xi), 0), \quad \boldsymbol{N}(\xi, \boldsymbol{z}, \omega) = \boldsymbol{F}(\boldsymbol{y}, \omega) - \boldsymbol{F}(\boldsymbol{y}_0(\xi), 0) - A(\xi)\, \boldsymbol{z}$$

for $\boldsymbol{y}_0(\xi) = (\boldsymbol{u}_0, \boldsymbol{u}_{0\xi})(\xi)$, and η, ω, and $\boldsymbol{q} = (\boldsymbol{q}_-, \boldsymbol{q}_+)$ will be determined later.

We define $X(\xi)$ by

$$X(\xi) = \begin{pmatrix} \boldsymbol{U}_3 & \boldsymbol{U}_4 & \boldsymbol{U}_1 & \boldsymbol{U}_2 \\ \boldsymbol{U}_{3\xi} & \boldsymbol{U}_{4\xi} & \boldsymbol{U}_{1\xi} & \boldsymbol{U}_{2\xi} \end{pmatrix}(\xi),$$

where $\{\, \boldsymbol{U}_j(\xi) \,\}_{j=1}^4$ is a fundamental set of solutions of $\mathcal{L}_0\, \boldsymbol{u} = 0$ given in Lemma 2.2. We easily see that $X(\xi)$ is a fundamental matrix of $\frac{d}{d\xi}\boldsymbol{x} = A(\xi)\, \boldsymbol{x}$. It follows from Lemma 2.2 and the definition of Coppel [2] that $\frac{d}{d\xi}\boldsymbol{x} = A(\xi)\, \boldsymbol{x}$ has an exponential dichotomy on R_- (respectively, R_+) with the projection matrix $P_- = \mathrm{diag}(0, 0, 1, 1)$ (respectively, $P_+ = \mathrm{diag}(1, 0, 1, 0)$). Let $\boldsymbol{x}_i^*(\xi)$ and $x_{ij}^*(\xi)$ be the ith row vector and the (i, j)th element, respectively, of $X(\xi)^{-1}$. Using the estimates in Lemma 2.2, we can calculate $X(\xi)^{-1}$ directly, and then obtain

$$\boldsymbol{x}_4^*(\xi) = \begin{cases} O(\xi^{m_-}\, e^{\Gamma_1^-\, \xi}) & \text{as } \xi \to -\infty, \\ O(\xi^{m_+}\, e^{-\Gamma_1^+\, \xi}) & \text{as } \xi \to +\infty, \end{cases}$$

which means that $(u_{01}^*, u_{02}^*)(\xi) \equiv (x_{43}^*, x_{44}^*/d)(\xi)$ is a bounded solution of $\mathcal{L}_0^* \boldsymbol{u} = 0$. By Lemma 2.3, we have $u_{01}^*(\xi)\, u_{02}^*(\xi) < 0$ for any $\xi \in \mathbf{R}$.

LEMMA 2.5 (Lemma 3.2 in Kokubu [12]). *$\boldsymbol{z}(\xi, \boldsymbol{q}, \eta, \omega)$ is a solution of (2.4) if and only if $\boldsymbol{z}(\xi, \boldsymbol{q}, \eta, \omega)$ satisfies*

$$E(\boldsymbol{q}, \eta, \omega) \equiv \begin{pmatrix} \boldsymbol{x}_3^*(0)\, \boldsymbol{z}(0-, \boldsymbol{q}, \eta, \omega) - \int_{R_-} \boldsymbol{x}_3^*(\xi)\, \boldsymbol{N}(\xi, \boldsymbol{z}(\xi, \boldsymbol{q}, \eta, \omega), \omega)\, d\xi \\ \boldsymbol{x}_4^*(0)\, \boldsymbol{z}(0-, \boldsymbol{q}, \eta, \omega) - \int_{R_-} \boldsymbol{x}_4^*(\xi)\, \boldsymbol{N}(\xi, \boldsymbol{z}(\xi, \boldsymbol{q}, \eta, \omega), \omega)\, d\xi \\ \boldsymbol{x}_2^*(0)\, \boldsymbol{z}(0+, \boldsymbol{q}, \eta, \omega) + \int_{R_+} \boldsymbol{x}_2^*(\xi)\, \boldsymbol{N}(\xi, \boldsymbol{z}(\xi, \boldsymbol{q}, \eta, \omega), \omega)\, d\xi \\ \boldsymbol{x}_4^*(0)\, \boldsymbol{z}(0+, \boldsymbol{q}, \eta, \omega) + \int_{R_+} \boldsymbol{x}_4^*(\xi)\, \boldsymbol{N}(\xi, \boldsymbol{z}(\xi, \boldsymbol{q}, \eta, \omega), \omega)\, d\xi \end{pmatrix} = 0.$$

Let $B = (b_{ij})$ be an arbitrary $n \times n$ matrix. We denote by $\Delta_{j_1, \dots, j_\ell}^{i_1, \dots, i_\ell}(B)$ the determinant of the $(n-\ell) \times (n-\ell)$-matrix obtained by eliminating the i_1th, …, i_ℓth rows and j_1th, …, j_ℓth columns in B. By definition, we have $b_{ij} = (-1)^{i+j}\, \Delta_i^j(B^{-1})\, \det B$ for each i and j, if the inverse matrix B^{-1} exists. We know the following formula (for example, see [19, p. 379]):

$$\Delta_{j_1, j_2}^{i_1, i_2}(B)\, \det B = \Delta_{j_1}^{i_1}(B)\, \Delta_{j_2}^{i_2}(B) - \Delta_{j_2}^{i_1}(B)\, \Delta_{j_1}^{i_2}(B) \tag{2.5}$$

holds for any $i_1 < i_2$ and $j_1 < j_2$.

Clearly we know that $E(0,0,0) = 0$ because $\boldsymbol{z}(\xi, 0, 0, 0) = 0$ for any $\xi \in \mathbf{R}$. Since $\boldsymbol{N}_{\boldsymbol{z}}(\xi, 0, 0) = 0$ on $\mathbf{R}$ and $\boldsymbol{N}_\omega(\xi, 0, 0) = A(\xi)\, \boldsymbol{H}_\pm'(0) + \boldsymbol{F}_\omega(\boldsymbol{y}_0(\xi), 0)$ on $R_\pm$ are satisfied, we obtain

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{q}} \int_0^\xi \boldsymbol{x}_j^*(\zeta)\, \boldsymbol{N}(\zeta, \boldsymbol{z}(\zeta, \boldsymbol{q}, 0, 0), 0)\, d\zeta \bigg|_{\boldsymbol{q}=0} = 0, \\ &\frac{\partial}{\partial \eta} \int_0^\xi \boldsymbol{x}_j^*(\zeta)\, \boldsymbol{N}(\zeta, \boldsymbol{z}(\zeta, 0, \eta, 0), 0)\, d\zeta \bigg|_{\eta=0} = 0, \\ &\frac{\partial}{\partial \omega} \int_0^\xi \boldsymbol{x}_j^*(\zeta)\, \boldsymbol{N}(\zeta, \boldsymbol{z}(\zeta, 0, 0, \omega), \omega)\, d\zeta \bigg|_{\omega=0} \\ &\qquad = (\boldsymbol{x}_j^*(0) - \boldsymbol{x}_j^*(\xi))\, \boldsymbol{H}_\pm'(0) + \int_0^\xi \boldsymbol{x}_j^*(\zeta)\, \boldsymbol{F}_\omega(\boldsymbol{y}_0(\zeta), 0)\, d\zeta, \end{aligned}$$

and then we have

$$E_{\boldsymbol{q}}(0,0,0) = \begin{pmatrix} \hat{x}_{33}^* & \hat{x}_{34}^* & 0 & 0 \\ \hat{x}_{43}^* & \hat{x}_{44}^* & 0 & 0 \\ 0 & 0 & \hat{x}_{23}^* & \hat{x}_{24}^* \\ 0 & 0 & \hat{x}_{43}^* & \hat{x}_{44}^* \end{pmatrix}, \quad \frac{\partial E}{\partial(\eta, \omega)}(0,0,0) = \begin{pmatrix} \hat{x}_{32}^* & -F_3^- \\ \hat{x}_{42}^* & -F_4^- \\ \hat{x}_{22}^* & F_2^+ \\ \hat{x}_{42}^* & F_4^+ \end{pmatrix},$$

where $\hat{x}_{ij}^* = x_{ij}^*(0)$ and

$$F_j^\pm = \int_{R_\pm} \boldsymbol{x}_j^*(\xi)\, \boldsymbol{F}_\omega(\boldsymbol{y}_0(\xi), 0)\, d\xi = \int_{R_\pm} (x_{j3}^*(\xi)\, u_{01}(\xi) + k\, x_{j4}^*(\xi)\, u_{02}(\xi)/d)\, d\xi.$$

By (2.5) and Lemma 2.2, we have

$$\det E_{\boldsymbol{q}}(0,0,0) = -\frac{\det(\boldsymbol{U}_3, \boldsymbol{U}_4)(0)\, \det(\boldsymbol{U}_3, \boldsymbol{U}_1)(0)}{(\det X(0))^2} \neq 0.$$

Consequently it follows from the implicit function theorem that there exists a C^1-class function $\boldsymbol{q}(\eta,\omega)=(\boldsymbol{q}_-,\boldsymbol{q}_+)(\eta,\omega)$ defined in a neighborhood of $(\eta,\omega)=(0,0)$ such that $E(\boldsymbol{q}(\eta,\omega),\eta,\omega)=0$ is satisfied for each (η,ω). Since

$$\begin{aligned}\frac{\partial(\boldsymbol{q}_+-\boldsymbol{q}_-)}{\partial\eta}(0,0)=&\begin{pmatrix}\hat{x}^*_{33}&\hat{x}^*_{34}\\ \hat{x}^*_{43}&\hat{x}^*_{44}\end{pmatrix}^{-1}\begin{pmatrix}\hat{x}^*_{32}\\ \hat{x}^*_{42}\end{pmatrix}-\begin{pmatrix}\hat{x}^*_{23}&\hat{x}^*_{24}\\ \hat{x}^*_{43}&\hat{x}^*_{44}\end{pmatrix}^{-1}\begin{pmatrix}\hat{x}^*_{22}\\ \hat{x}^*_{42}\end{pmatrix}\\ =&\frac{u_{01\xi}(0)}{\det X(0)\,\det E_{\boldsymbol{q}}(0,0,0)}\begin{pmatrix}-\hat{x}^*_{44}\\ \hat{x}^*_{43}\end{pmatrix}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial(\boldsymbol{q}_+-\boldsymbol{q}_-)}{\partial\omega}(0,0)=&-\begin{pmatrix}\hat{x}^*_{23}&\hat{x}^*_{24}\\ \hat{x}^*_{43}&\hat{x}^*_{44}\end{pmatrix}^{-1}\begin{pmatrix}F^+_2\\ F^+_4\end{pmatrix}-\begin{pmatrix}\hat{x}^*_{33}&\hat{x}^*_{34}\\ \hat{x}^*_{43}&\hat{x}^*_{44}\end{pmatrix}^{-1}\begin{pmatrix}F^-_3\\ F^-_4\end{pmatrix}\\ =&\left(\frac{F^+_2}{\Delta^{1,3}_{1,2}(X(0)^{-1})}+\frac{F^-_3}{\Delta^{1,2}_{1,2}(X(0)^{-1})}\right)\begin{pmatrix}-\hat{x}^*_{44}\\ \hat{x}^*_{43}\end{pmatrix}\\ &+\frac{F^+_4}{\Delta^{1,3}_{1,2}(X(0)^{-1})}\begin{pmatrix}\hat{x}^*_{24}\\ -\hat{x}^*_{23}\end{pmatrix}+\frac{F^-_4}{\Delta^{1,2}_{1,2}(X(0)^{-1})}\begin{pmatrix}\hat{x}^*_{34}\\ -\hat{x}^*_{33}\end{pmatrix}\end{aligned}$$

hold by virtue of (2.5), we have

$$\begin{aligned}\det\frac{\partial(\boldsymbol{q}_+-\boldsymbol{q}_-)}{\partial(\eta,\omega)}(0,0)=&\frac{u_{01\xi}(0)}{\det X(0)\,\det E_{\boldsymbol{q}}(0,0,0)}\\ &\times\int_{\mathbf{R}}(u_{01}(\xi)\,u^*_{01}(\xi)+k\,u_{02}(\xi)\,u^*_{02}(\xi))\,d\xi\,(\equiv J_1(k)).\end{aligned}$$

By combining the above argument with the comparison principle, we obtain the following lemma.

LEMMA 2.6. *There exists a C^1-class function $\boldsymbol{u}^I(.,\eta,\omega)$ defined on a neighborhood of $(\eta,\omega)=(0,0)$ such that (2.2) has a positive solution $\boldsymbol{u}^I(\xi,\eta,\omega)$ for each (η,ω), which satisfies $\boldsymbol{u}^I(.,\eta,\omega)\to\boldsymbol{u}_0$ in the C^2-sense as $(\eta,\omega)\to(0,0)$, and furthermore*

$$\det\left.\frac{\partial(\boldsymbol{u}^I_\xi(0+,\eta,\omega)-\boldsymbol{u}^I_\xi(0-,\eta,\omega))}{\partial(\eta,\omega)}\right|_{(\eta,\omega)=(0,0)}=J_1(k).$$

2.2. Outer approximation. In this subsection, we shall construct an approximate solution of (2.1) outside of $x=\tau$. As $\varepsilon\to0$, (2.1) is formally reduced to

$$\begin{cases}0=\boldsymbol{f}(\boldsymbol{u},v),\\ 0=v_{xx}+\sigma\,g(\boldsymbol{u},v),\quad x\in(0,\tau)\cup(\tau,1),\\ v_x(0)=0,\quad v(\tau)=v_0+\omega,\quad v_x(1)=0.\end{cases}\tag{2.6}$$

By the definition of $\boldsymbol{h}(x,v,\tau)$, we easily have $\boldsymbol{f}(\boldsymbol{h}(x,v,\tau),v)=0$ for any $x\in[0,\tau)\cup(\tau,1]$. Substituting $\boldsymbol{u}=\boldsymbol{h}(x,v,\tau)$ into the second equation of (2.6), we obtain the following equation for v only:

$$\begin{cases}0=v_{xx}+\sigma\,g(\boldsymbol{h}(x,v,\tau),v),\quad x\in(0,\tau)\cup(\tau,1),\\ v_x(0)=0,\quad v(\tau)=v_0+\omega,\quad v_x(1)=0.\end{cases}\tag{2.7}$$

Setting $J_2(v,\tau)=\int_0^1 g(\boldsymbol{h}(x,v,\tau),v)\,dx$, we find $J_2(v_0,x_0)=0$ and

$$J_{2\tau}(v_0,x_0)=g(\boldsymbol{h}_-(v_0),v_0)-g(\boldsymbol{h}_+(v_0),v_0)<0.$$

LEMMA 2.7. *Under the assumptions* (H.1) *and* (H.3)*, there exist* $\sigma_1>0$*,* $\delta_1>0$*, and* $\omega_1>0$ *such that* (2.7) *with* $\tau=x_0+\delta$ *has a unique solution* $v^O(x,\sigma,\delta,\omega)$ *for any* $\sigma\in[0,\sigma_1]$*,* $|\,\delta\,|\le\delta_1$*, and* $|\,\omega\,|\le\omega_1$ *which satisfies the following properties:*

(i) $v^O(x,\sigma,\delta,\omega)$ *is of* C^1*-class with respect to* (σ,δ,ω)*;*

(ii) $\frac{\partial}{\partial\delta}[v_x^O(x_0+\delta-,\sigma,\delta,\omega)-v_x^O(x_0+\delta+,\sigma,\delta,\omega)]/\sigma\to-J_{2\tau}(v_0,x_0)$ *as* $(\sigma,\delta,\omega)\to(0,0,0)$.

Proof. Setting $v=v_0+\omega+\sigma V$, we see that $V(x)$ satisfies

$$\begin{cases}0=V_{xx}+g(\boldsymbol{h}(x,v,x_0+\delta),v), & x\in(0,x_0+\delta)\cup(x_0+\delta,1),\\ V_x(0)=0, \quad V(x_0+\delta)=0, \quad V_x(1)=0.\end{cases}\tag{2.8}$$

Since $J_2(v_0,x_0)=0$ holds and the variational problem of (2.8) with respect to V at $(\sigma,\delta,\omega)=0$ is represented as

$$\begin{cases}0=V_{xx}, & x\in(0,x_0)\cup(x_0,1),\\ V_x(0)=0, \quad V(x_0)=0, \quad V_x(1)=0,\end{cases}$$

property (i) can be shown by virtue of the implicit function theorem. Furthermore by (2.7), we obtain

$$\begin{aligned}&\frac{1}{\sigma}\,\frac{\partial}{\partial\delta}[v_x^O(x_0+\delta-,\sigma,\delta,\omega)-v_x^O(x_0+\delta+,\sigma,\delta,\omega)]\\ &\qquad=-\frac{\partial}{\partial\delta}\int_0^1 g(\boldsymbol{h}(x,v^O(x,\sigma,\delta,\omega),x_0+\delta),v^O(x,\sigma,\delta,\omega))\,dx\to-J_{2\tau}(v_0,x_0)\end{aligned}$$

as $(\sigma,\delta,\omega)\to(0,0,0)$, which proves property (ii). ∎

2.3. Construction of solutions. In this subsection, we shall prove the existence of a C^2-class solution $(\boldsymbol{u},v)(x)$ of (2.1) on the whole interval $[0,1]$ by matching the inner and outer approximated solutions which were constructed in the previous subsections. In order to do this, we define $\boldsymbol{u}^a(x,\rho)$ and $v^a(x,\rho)$ with $\rho=(\varepsilon,\sigma,\delta,\eta,\omega)$ by

$$\begin{aligned}\boldsymbol{u}^a(x,\rho)&=\boldsymbol{h}(x,v^O(x,\sigma,\delta,\omega),x_0+\delta)\\ &\qquad+\theta(x)\,(\boldsymbol{u}^I((x-x_0-\delta)/\varepsilon,\eta,\omega)-\boldsymbol{h}(x,v_0+\omega,x_0+\delta)),\\ v^a(x,\rho)&=v^O(x,\sigma,\delta,\omega),\end{aligned}$$

where

$$\theta(x)=\begin{cases}\theta_0(x/(x_0+\delta)) & \text{for } x\le x_0+\delta,\\ \theta_0((1-x)/(1-x_0-\delta)) & \text{for } x>x_0+\delta\end{cases}$$

with a C^∞-cutoff function $\theta_0(x)$ which satisfies $0\le\theta_0(x)\le1$ for any $x\ge0$ and

$$\theta_0(x)=\begin{cases}0 & \text{for } 0\le x\le1/4,\\ 1 & \text{for } x\ge3/4.\end{cases}$$

We shall find a solution $(\boldsymbol{u}, v)$ of (2.1) in the following form:

$$(\boldsymbol{u}, v) = (\boldsymbol{u}^a, v^a)(x, \rho) + t, \quad t = (\boldsymbol{r}, s).$$

Here we define $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ by

$$\mathcal{T}(t, \rho) = \begin{pmatrix} \varepsilon^2 D\, \boldsymbol{u}_{yy} + \chi(., \delta)\, \boldsymbol{f}(\boldsymbol{u}, v) \\ v_{yy} + \sigma\, \chi(., \delta)\, g(\boldsymbol{u}, v) \end{pmatrix},$$

where

$$y = \begin{cases} \dfrac{x_0\, x}{x_0 + \delta} & \text{for } x < x_0 + \delta, \\ x_0 + \dfrac{(1 - x_0)\,(x - x_0 - \delta)}{1 - x_0 - \delta} & \text{for } x > x_0 + \delta, \end{cases}$$

$$\chi(y, \delta) = \begin{cases} \dfrac{(x_0 + \delta)^2}{x_0^2} & \text{for } y < x_0, \\ \dfrac{(1 - x_0 - \delta)^2}{(1 - x_0)^2} & \text{for } y > x_0, \end{cases}$$

$$\mathcal{X} = \mathcal{X}_u \times (H^2(0,1) \cap \mathcal{Z}_0([0,1], \mathbf{R})), \quad \mathcal{X}_u = \mathcal{Z}_\varepsilon \cap \mathcal{Z}_0([0,1], \mathbf{R}^2),$$

$$\mathcal{Y} = \mathcal{Y}_u \times L^2(0,1), \quad \mathcal{Y}_u = C^0([0,1], \mathbf{R}^2),$$

$$\mathcal{Z}_0([0,1], \mathbf{R}^\ell) = \{\, u \in C^1([0,1], \mathbf{R}^\ell) \,|\, u_y(0) = 0,\ u(x_0) = 0,\ u_y(1) = 0 \,\}.$$

LEMMA 2.8. *The following properties are satisfied in a neighborhood of $\rho = 0$:*
(i) *$\mathcal{T}$ is of C^1-class with respect to t, and there exists $C_1 > 0$ such that*

$$\| \mathcal{T}_t(t_1, \rho) - \mathcal{T}_t(t_2, \rho) \|_{\mathcal{L}(\mathcal{X}, \mathcal{Y})} \le C_1 \| t_1 - t_2 \|_{\mathcal{X}}$$

for any t_1, $t_2 \in \mathcal{X}$.
(ii) *$\mathcal{T}_t(0, \rho)$ has an inverse.*
(iii) *$\| \mathcal{T}(0, \rho) \|_{\mathcal{Y}} \to 0$ as $\varepsilon \to 0$ uniformly in $(\sigma, \delta, \eta, \omega)$.*

Proof. Using the arguments in Fife [4], we may only show that there exist $C_2 > 0$, $\varepsilon_1 > 0$, and $\sigma_2 > 0$ such that $\| (\hat{\mathcal{L}}^{\varepsilon,\sigma})^{-1} \|_{\mathcal{L}(\mathcal{X}_u, \mathcal{Y}_u)} \le C_2$ holds for any $(\varepsilon, \sigma) \in (0, \varepsilon_1] \times (0, \sigma_2]$, where $\boldsymbol{f}^a_{\boldsymbol{u}}(y, \varepsilon, \sigma) = \boldsymbol{f}_{\boldsymbol{u}}((\boldsymbol{u}^a, v^a)(y, \varepsilon, \sigma, 0, 0, 0))$ and $\hat{\mathcal{L}}^{\varepsilon,\sigma}\, \boldsymbol{u} = \varepsilon^2 D\, \boldsymbol{u}_{yy} + \boldsymbol{f}^a_{\boldsymbol{u}}(y, \varepsilon, \sigma)\, \boldsymbol{u}$.

Contrary to the conclusion, suppose that there exists $\{\, (\hat{\boldsymbol{u}}_n(y), \hat{\varepsilon}_n, \hat{\sigma}_n) \,\}_{n=1}^\infty$ such that $\lim_{n\to\infty}(\hat{\varepsilon}_n, \hat{\sigma}_n) = (0, 0)$ and $1 = \| \hat{\boldsymbol{u}}_n \|_{\mathcal{X}_u} \ge n \| \mathcal{L}^{\hat{\varepsilon}_n, \hat{\sigma}_n}\, \hat{\boldsymbol{u}}_n \|_{\mathcal{Y}_u}$ for any $n \in \mathbf{N}$. Since

$$\boldsymbol{f}^a_{\boldsymbol{u}}(x_0 + \varepsilon\, \xi, \varepsilon, \sigma) \to \boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{u}_0(\xi), v_0) \quad \text{as } (\varepsilon, \sigma) \to (0, 0)$$

uniformly on any compact set of $\mathbf{R}$ in the C^0-sense, it follows from the Ascoli–Arzelà theorem that there exists $\hat{\boldsymbol{u}}_0(\xi)$ such that $\hat{\boldsymbol{u}}_0(\xi)$ is a solution of $\mathcal{L}_0\, \boldsymbol{u} = 0$ on $R_- \cup R_+$ and satisfies $\hat{\boldsymbol{u}}_0(0) = 0$ and $\| \hat{\boldsymbol{u}}_0 \|_{C^2(R_- \cup R_+, \mathbf{R}^2)} = 1$. By Lemma 2.2, we have $\hat{\boldsymbol{u}}_0(\xi) = 0$ for any $\xi \in \mathbf{R}$. This contradiction implies that the desired result holds. ∎

By the above lemma and the implicit function theorem [4, Theorem 3.4], we see that there exists a solution family $(\boldsymbol{u}, v)(x, \rho)$ of (2.1L) and (2.1R) in a neighborhood of $\rho = 0$ such that

$$\| \boldsymbol{u}(., \rho) - \boldsymbol{u}^a(., \rho) \|_{Z_\varepsilon} + \| v(., \rho) - v^a(., \rho) \|_{H^2(0,1)} \to 0 \tag{2.9}$$

as $\varepsilon \to 0$.

We now define $(\Phi, \Psi)(\rho)$ by

$$(\Phi, \Psi)(\rho) = (\varepsilon\, \boldsymbol{u}_x, v_x/\sigma)(x_0 + \delta+, \rho) - (\varepsilon\, \boldsymbol{u}_x, v_x/\sigma)(x_0 + \delta-, \rho).$$

By Lemma 2.7 and (2.9), it is obvious that $(\Phi, \Psi)(0) = 0$,

$$\frac{\partial \Phi}{\partial \delta}(0) = 0, \quad \text{and} \quad \frac{\partial \Psi}{\partial \delta}(0) = J_{2\tau}(v_0, x_0)\, (< 0)$$

hold. Lemma 2.3 implies that there exists $k_0 > 0$ such that

$$\int_{\mathbf{R}} (u_{01}(\xi)\, u_{01}^*(\xi) + k\, u_{02}(\xi)\, u_{02}^*(\xi))\, d\xi \neq 0$$

holds for any $k \in [0, k_0]$. Thus we find

$$\det \frac{\partial(\Phi, \Psi)}{\partial(\delta, \eta, \omega)}(0) = -J_1(k)\, J_{2\tau}(v_0, x_0) \neq 0$$

for $k \in [0, k_0]$. By using the implicit function theorem, we find that there exist $\varepsilon_2 > 0$ and $\sigma_3 > 0$ such that (2.1) has a solution $(\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})(x)$ for any $(\varepsilon, \sigma) \in (0, \varepsilon_2] \times (0, \sigma_3]$.

We thus find that the spatial profile of $\boldsymbol{u}^{\varepsilon,\sigma}(x)$ indicates that two competing species exhibit spatial segregation with a small overlapping zone in $[0, 1]$, as was shown in Figure 3.

3. Stability. In this section, we shall discuss the stability of the equilibrium solution $(\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})(x)$ of (1.5) which was given in the previous section. To do this, it is enough to study the distribution of eigenvalues of the following linearized eigenvalue problem of (1.5) around $(\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})(x)$ for $(\varepsilon, \sigma) \in (0, \varepsilon_2] \times (0, \sigma_3]$:

$$\begin{cases} \lambda\, \boldsymbol{w} = \varepsilon^2\, D\, \boldsymbol{w}_{xx} + \boldsymbol{f}_{\boldsymbol{u}}^{\varepsilon,\sigma}(x)\, \boldsymbol{w} + \boldsymbol{f}_{v}^{\varepsilon,\sigma}(x)\, z, \\ \lambda\, \sigma\, z = z_{xx} + \sigma\, g_{\boldsymbol{u}}^{\varepsilon,\sigma}(x)\, \boldsymbol{w} + \sigma\, g_{v}^{\varepsilon,\sigma}(x)\, z, \quad x \in (0, 1), \\ \boldsymbol{w}_x = 0, \quad z_x = 0, \quad x = 0, 1, \end{cases} \tag{3.1}$$

where $\boldsymbol{f}_{\boldsymbol{u}}^{\varepsilon,\sigma}(x) = \boldsymbol{f}_{\boldsymbol{u}}((\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})(x))$, and the other functions $\boldsymbol{f}_{v}^{\varepsilon,\sigma}(x)$, $g_{\boldsymbol{u}}^{\varepsilon,\sigma}(x)$, and $g_{v}^{\varepsilon,\sigma}(x)$ are similarly defined.

We define the operators $\mathcal{L}^{\varepsilon,\sigma}$ and $\mathcal{L}^{*\,\varepsilon,\sigma}$ by

$$\mathcal{L}^{\varepsilon,\sigma}\, \boldsymbol{w} = \varepsilon^2\, D\, \boldsymbol{w}_{xx} + \boldsymbol{f}_{\boldsymbol{u}}^{\varepsilon,\sigma}(x)\, \boldsymbol{w} \quad \text{and} \quad \mathcal{L}^{*\,\varepsilon,\sigma}\, \boldsymbol{w} = \varepsilon^2\, D\, \boldsymbol{w}_{xx} + {}^t\boldsymbol{f}_{\boldsymbol{u}}^{\varepsilon,\sigma}(x)\, \boldsymbol{w},$$

respectively. Let $\{\, \zeta_n^{\varepsilon,\sigma} \,\}_{n=0}^{\infty}$ be the set of eigenvalues of $\mathcal{L}^{\varepsilon,\sigma}$ with the Neumann boundary condition, and let $\phi_n^{\varepsilon,\sigma}$ (respectively, $\phi_n^{*\,\varepsilon,\sigma}$) be an eigenfunction of $\mathcal{L}^{\varepsilon,\sigma}$ (respectively, $\mathcal{L}^{*\,\varepsilon,\sigma}$) corresponding to the eigenvalue $\zeta_n^{\varepsilon,\sigma}$ (respectively, $\overline{\zeta_n^{\varepsilon,\sigma}}$) for each n. Here we normalize $\phi_n^{\varepsilon,\sigma}$ and $\phi_n^{*\,\varepsilon,\sigma}$ as $\| \phi_n^{\varepsilon,\sigma} \|_{L^2(0,1)} = 1$ and $\langle\, \phi_n^{\varepsilon,\sigma}, \phi_n^{*\,\varepsilon,\sigma} \,\rangle = 1$ for each $n \geq 0$, where $\langle ., . \rangle$ is the inner product in $L^2(0, 1)$. Without loss of generality, we may assume that $\zeta_n^{\varepsilon,\sigma}$ satisfies $\operatorname{Re} \zeta_n^{\varepsilon,\sigma} \leq \operatorname{Re} \zeta_0^{\varepsilon,\sigma}$ for any $n \geq 1$. In a similar manner to the proof of Lemma 1.4 in Nishiura and Fujii [16], the following lemma can be proved by using Lemma 2.4.

LEMMA 3.1. *$\{\, \zeta_n^{\varepsilon,\sigma} \,\}_{n=0}^{\infty}$ satisfies the following properties:*

(i) $\operatorname{Re} \zeta_0^{\varepsilon,\sigma} = o(\varepsilon)$ *as* $(\varepsilon, \sigma) \to (0, 0)$.

(ii) *There exist* $\varepsilon_3 > 0$, $\sigma_4 > 0$, *and* $\mu_1 > 0$ *such that* $\operatorname{Re}\, \zeta_n^{\varepsilon,\sigma} \leq -\mu_1$ *for any* $(\varepsilon, \sigma) \in (0, \varepsilon_3] \times (0, \sigma_4]$ *and* $n \geq 1$.

We denote by $\Sigma^{\varepsilon,\sigma}$ the set of eigenvalues of (3.1) for (ε,σ). Let the projection $P_{\varepsilon,\sigma}$ be $P_{\varepsilon,\sigma}\,\boldsymbol{u} = \boldsymbol{u} - \langle\, \boldsymbol{u}, \phi_0^{*\,\varepsilon,\sigma}\,\rangle\,\phi_0^{\varepsilon,\sigma}$. By setting $\mu_2 = \mu_1/2$ and $\Lambda(\mu) = \{\,\lambda \in \mathbf{C} \,|\, \mathrm{Re}\ \lambda \geq -\mu\,\}$, Lemma 3.1 shows that $(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\,P_{\varepsilon,\sigma}$ has a uniformly L^2-bounded inverse for any $(\varepsilon,\sigma) \in (0,\varepsilon_3] \times (0,\sigma_4]$ and $\lambda \in \Lambda(\mu_2)$, i.e.,

$$\| \,[(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\,P_{\varepsilon,\sigma}]^{-1}\, \| \leq \frac{C_3}{|\,\mu_2 + \lambda\,|}$$

for some $C_3 > 0$. Hence we see that the solution $\boldsymbol{w}$ of $(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\,\boldsymbol{w} = -\boldsymbol{f}_v^{\varepsilon,\sigma}\,z$ is represented as

$$\boldsymbol{w} = -\frac{\langle\, \boldsymbol{f}_v^{\varepsilon,\sigma}\,z, \phi_0^{*\,\varepsilon,\sigma}\,\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda}\,\phi_0^{\varepsilon,\sigma} - [(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\,P_{\varepsilon,\sigma}]^{-1}\,P_{\varepsilon,\sigma}\,[\boldsymbol{f}_v^{\varepsilon,\sigma}\,z].$$

Substituting the above formula into the second equation of (3.1), we obtain the following eigenvalue problem:

$$\text{(3.2)}\qquad \begin{cases} 0 = \dfrac{1}{\sigma}\,z_{xx} - \dfrac{\langle\, \boldsymbol{f}_v^{\varepsilon,\sigma}\,z, \phi_0^{*\,\varepsilon,\sigma}\,\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda}\,g_{\boldsymbol{u}}^{\varepsilon,\sigma}\,\phi_0^{\varepsilon,\sigma} \\ \qquad\qquad -g_{\boldsymbol{u}}^{\varepsilon,\sigma}\,[(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\,P_{\varepsilon,\sigma}]^{-1}\,P_{\varepsilon,\sigma}\,[\boldsymbol{f}_v^{\varepsilon,\sigma}\,z] \\ \qquad\qquad +g_v^{\varepsilon,\sigma}\,z - \lambda\,z, \quad x \in (0,1), \\ z_x = 0, \quad x = 0,1, \end{cases}$$

so that

$$\begin{aligned} 0 =& K^{\varepsilon,\sigma}(z,\lambda) \equiv \frac{\langle\, \boldsymbol{f}_v^{\varepsilon,\sigma}\,z, \phi_0^{*\,\varepsilon,\sigma}\,\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda}\,\langle\, g_{\boldsymbol{u}}^{\varepsilon,\sigma}\,\phi_0^{\varepsilon,\sigma}, 1\,\rangle \\ &+ \langle\, g_{\boldsymbol{u}}^{\varepsilon,\sigma}\,[(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\,P_{\varepsilon,\sigma}]^{-1}\,P_{\varepsilon,\sigma}\,[\boldsymbol{f}_v^{\varepsilon,\sigma}\,z], 1\,\rangle - \langle\, g_v^{\varepsilon,\sigma}\,z, 1\,\rangle + \lambda\,\langle\, z, 1\,\rangle \end{aligned}$$

is satisfied for any eigenvalue $\lambda \in \Sigma^{\varepsilon,\sigma}$ and its eigenfunction $z(x)$.

Let B_δ be a closed ball with center at the origin and radius δ in the complex plane $\mathbf{C}$. We set $\Lambda_\delta(\mu) = \Lambda(\mu) \setminus B_\delta$ for an arbitrarily fixed small constant $\delta > 0$. For a given (ε,σ)-dependent function $u^{\varepsilon,\sigma}$, we define $u^{0,0}$ by $u^{0,0} = \lim_{(\varepsilon,\sigma)\to(0,0)} u^{\varepsilon,\sigma}$. We obtain the following lemma.

LEMMA 3.2 (Sublemma 2.1 and Corollary 2.1 in [16]). *There exists $C_4 > 0$ such that for each $(\varepsilon,\sigma) \in (0,\varepsilon_3] \times (0,\sigma_4]$, $|\,\lambda\,| \leq C_4$ and $1 \leq \|\,z\,\|_{H^1(0,1)} \leq \sqrt{1 + \sigma\,C_4}$ hold for any eigenvalue $\lambda \in \Lambda_\delta(\mu_2)$ and its eigenfunction $z(x)$ with $\|\,z\,\|_{L^2(0,1)} = 1$.*

LEMMA 3.3 (Lemmas 2.2 and 2.3 in [16]). *As $(\varepsilon,\sigma) \to (0,0)$, the following assertion holds: For $z \in L^2(0,1) \cap L^\infty(0,1)$,*

(i) $[(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\,P_{\varepsilon,\sigma}]^{-1}\,P_{\varepsilon,\sigma}\,[\boldsymbol{f}_v^{\varepsilon,\sigma}\,z] \to (\boldsymbol{f}_{\boldsymbol{u}}^{0,0} - \lambda)^{-1}\,[\boldsymbol{f}_v^{0,0}\,z]$ *strongly in the $L^2(0,1)$-sense;*

(ii) $\langle\, \boldsymbol{f}_v^{\varepsilon,\sigma}\,z, \phi_0^{*\,\varepsilon,\sigma}\,\rangle\,\langle\, g_{\boldsymbol{u}}^{\varepsilon,\sigma}\,\phi_0^{\varepsilon,\sigma}, 1\,\rangle/\varepsilon \to C_0\,z(x_0)$,

where

$$\begin{aligned} C_0 =& (g(\boldsymbol{h}_-(v_0), v_0) - g(\boldsymbol{h}_+(v_0), v_0)) \\ &\times \frac{\int_{\mathbf{R}} (u_{01}(\xi)\,u_{01}^*(\xi) + k\,u_{02}(\xi)\,u_{02}^*(\xi))\,d\xi}{\int_{\mathbf{R}} (u_{01\xi}(\xi)\,u_{01}^*(\xi) + u_{02\xi}(\xi)\,u_{02}^*(\xi))\,d\xi}. \end{aligned}$$

The convergence in (i) *and* (ii) *is uniform for $\lambda \in \Lambda(\mu_2)$. Furthermore if $z \in H^1(0,1)$, then the convergence in* (i) *is also uniform on a bounded set in $H^1(0,1)$.*

By (H.3) and Lemma 2.3, we easily find $C_0 < 0$ for $k \in (0,k_0]$.

We set

$$\Sigma_1 = \cap_{(p,q)\in(0,\varepsilon_3]\times(0,\sigma_4]}\overline{\cup_{(\varepsilon,\sigma)\in(0,p]\times(0,q]}\Sigma^{\varepsilon,\sigma}},$$

$$Q_1(x) = \begin{cases} f_{2u_2}(\boldsymbol{h}_-(v_0), v_0)\,(<0) & \text{for } x < x_0, \\ f_{1u_1}(\boldsymbol{h}_+(v_0), v_0)\,(<0) & \text{for } x > x_0, \end{cases}$$

$$Q_2(x) = \begin{cases} f_{2v}(\boldsymbol{h}_-(v_0), v_0)\, g_{u_2}(\boldsymbol{h}_-(v_0), v_0)\,(<0) & \text{for } x < x_0, \\ f_{1v}(\boldsymbol{h}_+(v_0), v_0)\, g_{u_1}(\boldsymbol{h}_+(v_0), v_0)\,(<0) & \text{for } x > x_0, \end{cases}$$

$$\mu_3 = \min\left\{ \mu_2,\, \langle\, Q_2/(2\,Q_1), 1\,\rangle,\, \inf_{x\in[0,x_0)\cup(x_0,1]} |\, Q_1(x)/4\,| \right\}\ (>0),$$

and assume that λ satisfies $\lambda \in (\Sigma_1 \cap \Lambda(\mu_3)) \setminus \{\,0\,\}$. By the definition of Σ_1, we see that there exists $\{\,(\hat\varepsilon_n, \hat\sigma_n, \hat\lambda_n)\,\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty}(\hat\varepsilon_n, \hat\sigma_n, \hat\lambda_n) = (0,0,\lambda)$ and $\hat\lambda_n \in \Sigma^{\hat\varepsilon_n,\hat\sigma_n}$ for each n. Let $\hat z_n(x)$, which is normalized as $\|\,\hat z_n\,\|_{L^2(0,1)} = 1$, be an eigenfunction of (3.2) corresponding to $\hat\lambda_n$. We know $\langle\, g_v^{0,0}, 1\,\rangle = 0$ by construction. Since $\hat z_n(x) \to 1$ as $n \to \infty$ by virtue of Lemma 3.2, we obtain

$$0 = K^{\hat\varepsilon_n,\hat\sigma_n}(\hat z_n, \hat\lambda_n) \to K_0(\lambda) \equiv \left\langle \frac{Q_2}{Q_1 - \lambda} + \lambda, 1 \right\rangle \quad \text{as } n \to \infty.$$

Since $Q_1(x) \le Q_1(x) - \lambda < 0$ holds for any $x \in [0, x_0) \cup (x_0, 1]$ and $\lambda \in (-\mu_3, 0]$, we have

$$0 = K_0(\lambda) \begin{cases} > 0 & \text{if } \lambda > 0, \\ \ge \langle\, Q_2/Q_1, 1\,\rangle + \lambda > 0 & \text{if } \lambda \in (-\mu_3, 0], \end{cases}$$

$$0 = \text{Re } K_0(\lambda) - \frac{\text{Re }\lambda}{\text{Im }\lambda}\,\text{Im } K_0(\lambda)$$
$$= \int_0^1 \frac{Q_2(x)\,(Q_1(x) - 2\,\text{Re }\lambda)}{|\,Q_1(x) - \lambda|^2}\,dx > 0 \quad \text{if } \lambda \in \Lambda(\mu_3) \setminus \mathbf{R}.$$

This is a contradiction, which leads to $\Sigma_1 \cap \Lambda(\mu_3) \subset \{\,0\,\}$.

We define $\mathcal{K}^{\varepsilon,\sigma,\lambda}$ by

$$\mathcal{K}^{\varepsilon,\sigma,\lambda}\, z = \frac{1}{\sigma}\, z_{xx} - g_{\boldsymbol{u}}^{\varepsilon,\sigma}\,[(\mathcal{L}^{\varepsilon,\sigma} - \lambda)\, P_{\varepsilon,\sigma}]^{-1}\, P_{\varepsilon,\sigma}\,[\boldsymbol{f}_v^{\varepsilon,\sigma}\, z] + g_v^{\varepsilon,\sigma}\, z - \lambda\, z.$$

Then we have the following lemma.

LEMMA 3.4 (Lemma 3.1 in [16]). *There exist $\varepsilon_4 > 0$, $\sigma_5 > 0$, and $\mu_4 > 0$ such that $\mathcal{K}^{\varepsilon,\sigma,\lambda}$ has a uniformly bounded inverse $(\mathcal{K}^{\varepsilon,\sigma,\lambda})^{-1} : H^{-1}(0,1) \to H_N^1(0,1)$ for any $(\varepsilon,\sigma) \in (0,\varepsilon_4] \times (0,\sigma_5]$ and $\lambda \in \Lambda(\mu_4)$, which is continuous on (ε,σ) and analytical on λ in the operator norm sense, where $H_N^1(0,1) = \{\, z \in H^1(0,1) \,|\, z_x(0) = 0 = z_x(1)\,\}$. Furthermore, $(\mathcal{K}^{\varepsilon,\sigma,\lambda})^{-1}\, z \to -\langle\, z, 1\,\rangle / K_0(\lambda)$ as $(\varepsilon,\sigma) \to (0,0)$ holds for any $\lambda \in \Lambda(\mu_4)$.*

Let $\lambda^{\varepsilon,\sigma}$ be an eigenvalue of (3.1) for (ε,σ) which satisfies $\lambda^{\varepsilon,\sigma} \to 0$ as $(\varepsilon,\sigma) \to (0,0)$, and let $(\boldsymbol{w}^{\varepsilon,\sigma}, z^{\varepsilon,\sigma})(x)$ be an eigenfunction of (3.1) corresponding to $\lambda^{\varepsilon,\sigma}$. Since $z^{\varepsilon,\sigma}(x)$ is represented as

$$z^{\varepsilon,\sigma} = \frac{\langle\, \boldsymbol{f}_v^{\varepsilon,\sigma}\, z^{\varepsilon,\sigma}, \phi_0^{*\,\varepsilon,\sigma}\,\rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda^{\varepsilon,\sigma}} \left(\mathcal{K}^{\varepsilon,\sigma,\lambda^{\varepsilon,\sigma}}\right)^{-1} [g_{\boldsymbol{u}}^{\varepsilon,\sigma}\, \phi_0^{\varepsilon,\sigma}],$$

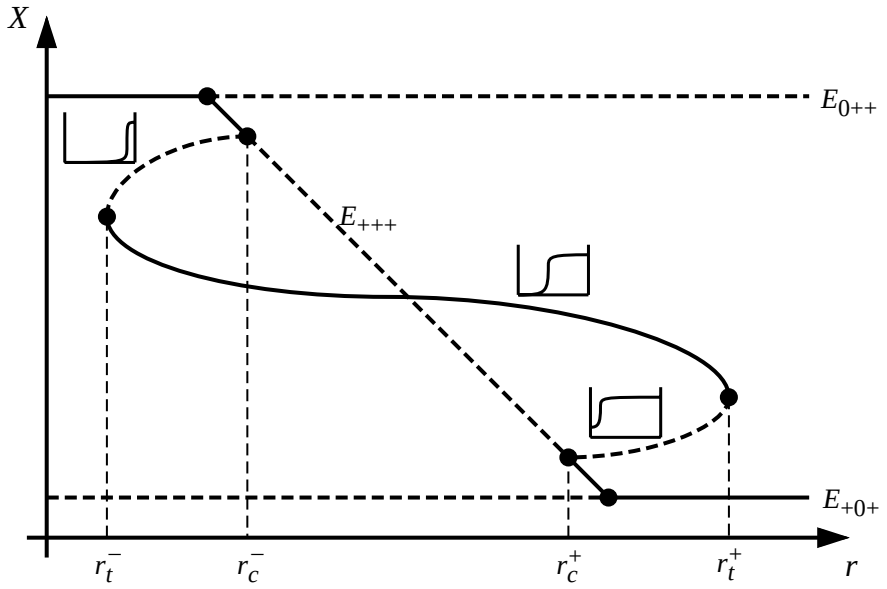

FIG. 4. *Bifurcation picture of positive equilibrium solutions. Solid curves indicate the stable branch, and broken ones the unstable branch. The profile of the u_1-component on the branch is shown. E_{0++} and E_{+0+} indicate the equilibrium points $(0, r/(\beta k), (a\beta k - r)/(\beta k^2))$ and $(r, 0, 1-r)$, respectively.*

we find

$$\langle \boldsymbol{f}_v^{\varepsilon,\sigma}\, z^{\varepsilon,\sigma}, \phi_0^{*\,\varepsilon,\sigma} \rangle \left\{ 1 - \frac{\langle \boldsymbol{f}_v^{\varepsilon,\sigma}\, \left(\mathcal{K}^{\varepsilon,\sigma,\lambda^{\varepsilon,\sigma}}\right)^{-1} [g_{\boldsymbol{u}}^{\varepsilon,\sigma}\, \phi_0^{\varepsilon,\sigma}], \phi_0^{*\,\varepsilon,\sigma} \rangle}{\zeta_0^{\varepsilon,\sigma} - \lambda^{\varepsilon,\sigma}} \right\} = 0.$$

If $\langle \boldsymbol{f}_v^{\varepsilon,\sigma}\, z^{\varepsilon,\sigma}, \phi_0^{*\,\varepsilon,\sigma} \rangle = 0$ holds, then $(\boldsymbol{w}^{\varepsilon,\sigma}, z^{\varepsilon,\sigma})(x) = 0$ is satisfied for any $x \in [0,1]$. Therefore we obtain $\langle \boldsymbol{f}_v^{\varepsilon,\sigma}\, z^{\varepsilon,\sigma}, \phi_0^{*\,\varepsilon,\sigma} \rangle \neq 0$. From $K_0(0) > 0$ and Lemmas 3.1, 3.3, and 3.4, we have

$$\frac{\lambda^{\varepsilon,\sigma}}{\varepsilon} = \frac{1}{\varepsilon} \left\{ \zeta_0^{\varepsilon,\sigma} - \left\langle \boldsymbol{f}_v^{\varepsilon,\sigma} \left(\mathcal{K}^{\varepsilon,\sigma,\lambda^{\varepsilon,\sigma}}\right)^{-1} [g_{\boldsymbol{u}}^{\varepsilon,\sigma}\, \phi_0^{\varepsilon,\sigma}], \phi_0^{*\,\varepsilon,\sigma} \right\rangle \right\} \to C_0/K_0(0) < 0$$

as $(\varepsilon, \sigma) \to (0,0)$ for any $k \in (0, k_0]$.

By summarizing the above results, we arrive at the fact that $(\boldsymbol{u}^{\varepsilon,\sigma}, v^{\varepsilon,\sigma})(x)$ is stable for sufficiently small $\varepsilon > 0$ and $\sigma > 0$.

4. Concluding remarks. In this paper, we have proved that there are stable SIP equilibrium solutions of the 3-component reaction-diffusion system which describes the interaction of one predator and two competing prey species. This result is ecologically stated in the following: Consider the situation where coexistence of two competing species never occurs even in the presence of a predator if the diffusion rates of the three species are very large. Then, if the diffusion rates of two competing species are much smaller than that of the predator ($\varepsilon > 0$ and $\sigma > 0$ are sufficiently small), predator-mediated coexistence possibly occurs so that two competing species coexist with spatial segregation.

For the existence of stable SIP equilibrium solutions, we assumed r satisfied the inequality (H.3) for fixed $k > 0$. It would be interesting to consider the case when (H.3) is violated. In Figure 4, the global structure of positive equilibrium solutions is

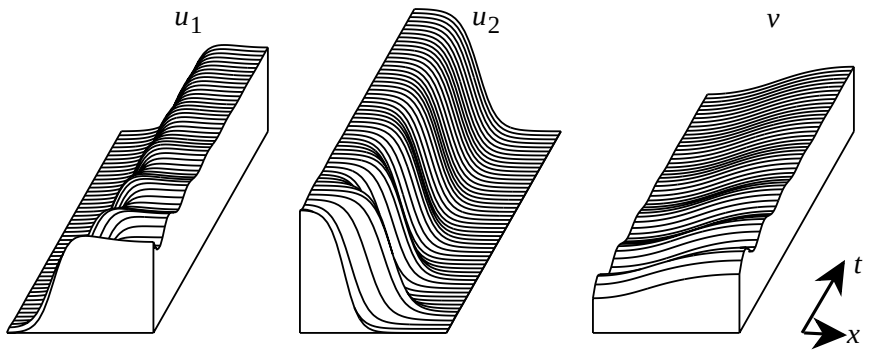

Figure 5a. $a = 0.85$, $\alpha = 0.1$ and $\sigma = 100$.

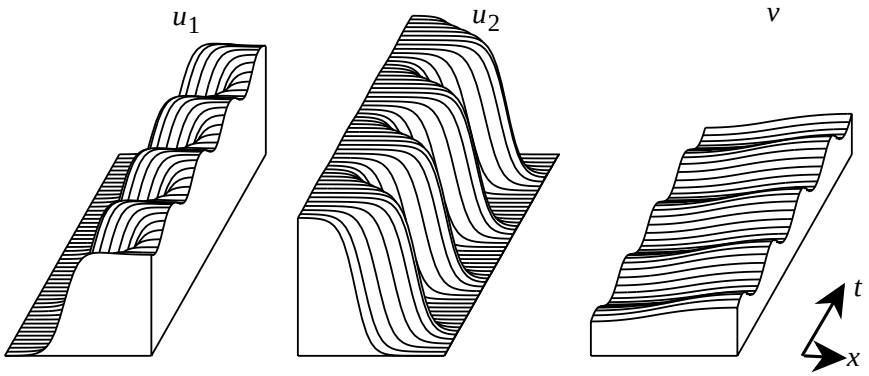

Figure 5b. $a = 0.95$, $\alpha = 0.05$ and $\sigma = 100$.

FIG. 5. *Numerical simulations of equilibrium solutions* (a) *and periodic solutions* (b), *where* $\varepsilon = 0.02$, $d = 1$, $b = 1.5$, $c = 1$, $k = 0.1$, $r = 0.4$, *and* $\beta = 1$.

numerically shown when r is globally varied. The branch of SIP equilibrium solutions with a single layer bifurcates from E_{+++} at $r = r_c^-$ and $r = r_c^+$ and consists of three parts: (i) an upper branch ($r_t^- < r < r_c^-$), (ii) a middle branch ($r_t^- < t < r_t^+$), and (iii) a lower branch ($r_c^+ < r < r_t^+$). The stable solutions which we discussed in the previous section correspond to the middle branch. On the other hand, there are SIP equilibrium solutions with a boundary layer in a neighborhood of $x = 1$ or $x = 0$ which correspond to the upper and lower branches, respectively. Quite recently the existence and instability of such solutions have been proved in [8].

In this paper, we restricted our discussion to the case when $\sigma > 0$ is small. When $\sigma > 0$ is not so small, the situation is changed and our equilibrium solutions might be destabilized through Hopf bifurcation so that there are periodic solutions as shown in Figure 5. These periodic solutions demonstrate spatially oscillating segregation of two competing species. This will be discussed elsewhere.

REFERENCES

[1] E. Conway, D. Hoff, and J. Smoller, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.

[2] W. A. Coppel, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, 1978.

[3] P. de Mottoni, *Qualitative analysis for some quasi-linear parabolic systems*, Inst. Math. Pol. Acad. Sci. Zam. 11/70, 190 (1979).

[4] P. C. Fife, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497–521.

[5] K. Fujii, *Complexity-stability relationship of two-prey-one-predator species system model: Local and global stability*, J. Theoret. Biol., 69 (1977), pp. 613–623.

[6] S. B. Hsu, *Predator-mediated coexistence and extinction*, Math. Biosci., 54 (1981), pp. 231–248.

[7] Y. Kan-on, *Parameter dependence of propagation speed of travelling waves for competition-diffusion equations*, SIAM J. Math. Anal., 26 (1995), pp. 340–363.

[8] Y. Kan-on, *Instability of Neumann layer solutions for a 3-component Lotka-Volterra model with diffusion*, in preparation.

[9] Y. Kan-on and Q. Fang, *Stability of monotone travelling waves for competition-diffusion equations*, Japan J. Indust. Appl. Math., 13 (1996), pp. 343–349.

[10] Y. Kan-on and E. Yanagida, *Existence of non-constant stable equilibria in competition-diffusion equations*, Hiroshima Math. J., 23 (1993), pp. 193–221.

[11] K. Kishimoto and H. F. Weinberger, *The spatial homogeneity of stable equilibria of some reaction-diffusion systems on convex domains*, J. Differential Equations, 58 (1985), pp. 15–21.

[12] H. Kokubu, *Homoclinic and heteroclinic bifurcations of vector fields*, Japan J. Appl. Math., 5 (1988), pp. 455–501.

[13] M. Matano and M. Mimura, *Pattern formation in competition-diffusion systems in nonconvex domains*, Publ. Res. Inst. Math. Sci., 19 (1983), pp. 1049–1080.

[14] M. Mimura and P. C. Fife, *A 3-component system of competition and diffusion*, Hiroshima Math. J., 16 (1986), pp. 189–207.

[15] M. Mimura and Y. Kan-on, *Predation-mediated coexistence and segregation structures*, in Patterns and Waves, T. Nishida et al., eds., Kinokuniya/North-Holland, Amsterdam, 1986, pp. 129–155.

[16] Y. Nishiura and H. Fujii, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 1726–1770.

[17] F. Rothe, *Global Solutions of Reaction-Diffusion Systems*, Lecture Notes in Math. 1072, Springer-Verlag, Berlin, New York, 1984.

[18] Y. Takeuchi and N. Adachi, *Existence and bifurcation of stable equilibrium in two-prey, one-predator communities*, Bull. Math. Biol., 45 (1983), pp. 877–900.

[19] The Mathematical Society of Japan, *Encyclopedic Dictionary of Mathematics*, K. Ito, ed., MIT Press, Cambridge, MA, 1987.